Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Αγρονομων και Τοπογραφων Μηχανικων
Δ.Π.Μ.Σ. ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ
Τομεας Τοπογραφιας

Ανάλυση Μεγάλων Γεωχωρικών Δεδομένων σε Περιβάλλον Cluster Υπολογιστών για Εφαρμογές Παρατήρησης της Γης

Geospatial Big Data Analysis in a Computer Cluster Environment for Earth Observation Applications

Μεταπτυχιακη Εργασια

του

ΑΘΑΝΑΣΙΟΥ Κ. ΚΑΡΜΑ

Εργαστηριο Τηλεπισκοπησης
Αθήνα, Οκτώβριος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών
Δ.Π.Μ.Σ. ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ
Τομέας Τοπογραφίας
Εργαστήριο Τηλεπισκόπησης

Ανάλυση Μεγάλων Γεωχωρικών Δεδομένων σε Περιβάλλον Cluster Υπολογιστών για Εφαρμογές Παρατήρησης της Γης

Geospatial Big Data Analysis in a Computer Cluster Environment for Earth Observation Applications

Μεταπτυχιακη Εργασια

του

ΑΘΑΝΑΣΙΟΥ Κ. ΚΑΡΜΑ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 27η Οκτωβρίου 2016.

.....	.....	.....
Δημήτριος Αργιαλάς	Κώστας Καράντζαλος	Μαρίνος Κάβουρας
Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Επιβλέπων: Δημήτριος Αργιαλάς
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2016

(Υπογραφή)

.....

Αθανασιος Κ. Καρμας
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών
Δ.Π.Μ.Σ. ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ
Τομέας Τοπογραφίας
Εργαστήριο Τηλεπισκόπησης

Η εργασία αυτή
αφιερώνεται
στους γονείς μου.

Ευχαριστίες

Θα ήθελα να πω ένα μεγάλο ευχαριστώ στην οικογένειά μου και στους φίλους μου για την αγάπη, την υποστήριξη και όλα όσα μου προσέφεραν όλα αυτά τα χρόνια. Χωρίς αυτά τίποτα δεν θα ήταν δυνατό.

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτριο Αργιαλά για την επίβλεψη αυτής της μεταπτυχιακής εργασίας και για την ευκαιρία που μου έδωσε να την εκπονήσω στο εργαστήριο Τηλεπισκόπησης.

Ευχαριστώ επίσης το Διδάκτορα του Εργαστηρίου Τηλεπισκόπησης του Ε.Μ.Π. κ. Άγγελο Τζώτσο για όλες τις τεχνικές συμβουλές, τη βοήθεια που τόσο απλόχερα μου προσέφερε καθώς και για τις μακροσκελείς τεχνικές συζητήσεις που είχαμε κατά καιρούς.

Τέλος, ένα μεγάλο ευχαριστώ στον κ. Κωνσταντίνο Καράντζαλο για την καθοδήγηση, τις ιδέες, τις τεχνικές και όχι μόνο συμβουλές που μου παρείχε κατά την εκπόνηση της παρούσας εργασίας.

Αθήνα,
Οκτώβριος 2016.

Περίληψη

Τα δορυφορικά δεδομένα παρατήρησης γης αυξάνονται με ιλιγγιώδη ρυθμό σε όγκο, ποικιλομορφία και πολυπλοκότητα με συνέπεια να δημιουργούνται νέες προκλήσεις σχετικά με την πρόσβαση, αρχειοθέτηση, επεξεργασία και ανάλυσή τους. Για την άμεση εξυπηρέτηση αυτού του πρωτοφανούς όγκου δεδομένων και την εξαγωγή γνώσης, νέες τεχνολογίες υπολογιστικών συστημάτων και αρχιτεκτονικών όπως επίσης και εργαλεία διαχείρισης αναπτύσσονται διαρκώς. Για να είναι δυνατόν να εκμεταλλευθούμε αυτή την πληθώρα δεδομένων, υπολογιστικών μηχανών, προγραμματιστικών μοντέλων, βιβλιοθηκών και εργαλείων τα οποία είναι διαθέσιμα χρειαζόμαστε συντονισμένες, προσαρμοστικές και ολοκληρωτικές προσπάθειες ώστε να συνδυάσουμε αρμονικά τις υπάρχουσες τεχνολογίες. Προς την κατεύθυνση αυτή παρουσιάζεται σε αυτή την εργασία μια πλατφόρμα ανάλυσης μεγάλων γεωχωρικών δεδομένων για εφαρμογές παρατήρησης της γης η οποία για την επίτευξη των στόχων της αποδοτικής ανάλυσης, αποθήκευσης και διαχείρισης μεγάλων δεδομένων ενσωματώνει μια σειρά από εργαλεία, βιβλιοθήκες και υπολογιστικά μοντέλα σε ένα κατανεμημένο περιβάλλον cluster υπολογιστών. Ειδικότερα, η βασική λειτουργικότητα συνίσταται από το Geotrellis, μια μηχανή επεξεργασίας γεωχωρικών δεδομένων για εφαρμογές υψηλών επιδόσεων για την αποθήκευση, διαχείριση και επεξεργασία των δεδομένων και το framework Apache Spark για την κατανομή υπολογισμών και δεδομένων κατά μήκος του cluster. Διάφοροι αλγόριθμοι υλοποιήθηκαν στη γλώσσα προγραμματισμού Scala τόσο για εργασίες ETL (Extract, Transform, Load) πάνω στα raw δεδομένα όσο και για την πρόσβαση και την κατανεμημένη επεξεργασία πολυφασματικών δορυφορικών εικόνων. Το ανεπτυγμένο σύστημα στην τρέχουσα μορφή του καλύπτει μερικώς τον Ελλαδικό χώρο με πολυφασματικά δεδομένα τα οποία προέρχονται από το δορυφόρο Landsat 8, τα οποία αποθηκεύονται και προ-επεξεργάζονται αυτόματα στο υλικό το οποίο έχουμε στη διάθεσή μας, για σκοπούς επίδειξης. Τα ανεπτυγμένα ερωτήματα επεξεργασίας των δεδομένων εστιάζουν σε αγροτικές εφαρμογές και παράγουν τόσο τεχνητά έγχρωμα σύνθετα με βάση καλώς ορισμένους δείκτες, τα οποία και παρέχουν πληροφορία σχετικά με την κατάσταση των καλλιεργειών διαχρονικά ανά pixel και ανά έτος, όσο και χάρτες αποτύπωσης της εποχικότητας από τους οποίους είναι δυνατόν να εξαχθούν μοτίβα για τον κύκλο ζωής των υπό παρακολούθηση καλλιεργειών.

Λέξεις Κλειδιά

Γεωχωρικά Δεδομένα, Νέφος υπολογιστών, Πλατφόρμες Μεγάλων Δεδομένων, Αναλύσεις Μεγάλων Δεδομένων, Γεωχωρικές Υπηρεσίες, Δέσμες Υπολογιστών

Abstract

Earth Observation satellite data are increasing with tremendous rate in volume, variety and complexity. As a consequence new challenges are emerging in regard to their access, archiving, processing and analysis. To serve rapidly this unprecedented data volume and extract knowledge through its analysis, new computing systems technologies and architectures as well as various administration tools are constantly developed. In order to exploit effectively this plethora of data, computing engines, programming models, libraries and tools that are available, coordinated, adaptive and integrated efforts are needed so as to seamlessly combine the existing technologies. Towards this direction a platform capable of analysing geospatial big data for earth observation applications is presented. The developed platform, in order to achieve the goals of efficient analysis, storage and handling of big data, integrates a number of tools, libraries and processing models in a distributed computer cluster environment. In particular, the core functionality consists of Geotrellis, a geospatial data processing engine for high performance applications, responsible for the storage, handling and processing of the data as well as the Apache Spark framework for distributing computations and data across the cluster. Various algorithms were implemented in Scala programming language both for applying ETL (Extract, Transform, Load) procedures on the raw data and for the access and distributed analysis of multispectral satellite data. The developed system in its current form covers partially the Greek territory with multispectral satellite data which are acquired by Landsat 8 satellite and that are stored and processed in an automated way in the infrastructure which is in our disposal for demonstration purposes. The developed data processing queries are focused mainly in agricultural applications and create artificial color composites, based on well defined indices, that provide valuable information regarding the crops' state over time per pixel and per year as well as value-added products that map seasonality from which it is possible to extract patterns for the life cycle of the monitored crops.

Keywords

Geospatial data, Cloud computing, Big Data Platforms, Big Data Analytics, Geospatial Services, Computer Cluster

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της διπλωματικής

Η σύγχρονη γενιά αισθητήρων οι οποίοι βρίσκονται στο διάστημα παράγει σχεδόν συνεχείς ροές μαζικών δεδομένων παρατήρησης της γης. Σύντομα, υψηλής χωρικής ανάλυσης πολυφασματικές εικόνες που καλύπτουν όλη τη γη θα είναι διαθέσιμες μία φορά κάθε εβδομάδα και σε κάποιες περιοχές δύο φορές κάθε εβδομάδα. Επιπρόσθετα με τις αποστολές εθνικών και ευρωπαϊκών οργανισμών (κυρίως των Η.Π.Α. και της Ε.Ε.) οι οποίες παρέχουν ανοιχτά δεδομένα, η βιομηχανία του διαστήματος, μέσω startup εταιρειών, ολοένα και διογκώνεται, καθώς οι εκτοξεύσεις δορυφόρων γίνονται φθηνότερες και η τεχνολογία ολοένα και πιο προσιτή. Πολλοί μικροί και φθηνοί δορυφόροι σε τροχιά (θα) καθιστούν τα γεωχωρικά δεδομένα ευρέως διαθέσιμα για ένα πολύ μεγάλο εύρος εφαρμογών. Αυτή η νέα γενιά διασυνδεδεμένων δορυφόρων, καθημερινά, συλλέγουν και μεταβιβάζουν τηλεπισκοπικά δεδομένα με υπό του μέτρου χωρική ανάλυση, επιτρέποντας με αυτό τον τρόπο την παρακολούθηση των αλλαγών στην επιφάνεια της με συχνότητα μεγαλύτερη από ποτέ άλλοτε. Η βιομηχανία παρακολούθησης της γης η οποία βασίζεται σε δορυφόρους υφίσταται μια εντυπωσιακή ανάπτυξη, δεδομένου ότι περίπου 260 εκτοξεύσεις δορυφόρων αναμένονται μέσα στην επόμενη δεκαετία. Επιπλέον, η πρόσφατη χαλάρωση της νομοθεσίας σχετικά με την πώληση εικόνων πολύ υψηλής χωρικής ανάλυσης έχει ανοίξει το δρόμο για ακόμα μεγαλύτερη ανάπτυξη καθώς νέοι εταιρικοί παίκτες μπορούν να εισέλθουν στην αγορά και να παρέχουν κατάπαίτηση ευρέος φάσματος υπηρεσίες οι οποίες θα οδηγήσουν στη δημιουργία νέων οικονομικών δραστηριοτήτων καθώς και εφαρμογών, προϊόντων, υπηρεσιών καθώς και νέων επιχειρηματικών μοντέλων.

Ο όγκος αυτών των, τεραστίων σε απαιτήσεις, ροών δεδομένων παρατήρησης της γης οι οποίες λαμβάνονται από δορυφορικά κανάλια κατερχόμενης ζεύξης με ρυθμούς gigabit, αυξάνεται με τρομακτικούς ρυθμούς, αγγίζοντας αυτή τη στιγμή την τάξη των πολλών Petabyte σε πολλά αρχεία δορυφορικών δεδομένων [74],[69],[22],[60]. Σύμφωνα με στατιστικές του οργανισμού Open Geospatial Consortium (OGC), ο συνολικός όγκος των αρχειοθετημένων δεδομένων παρατήρησης της γης θα ξεπεράσει το 1 Exabyte (=1000 Petabytes) κατά τη διάρκεια του έτους 2015.

Παρ'ολα αυτά, εκτιμάται ότι τα περισσότερα σύνολα δεδομένων στα υπάρχοντα αρχεία

δορυφορικών δεδομένων δεν έχουν ποτέ τύχει πρόσβασης και επεξεργασίας [68] εκτός συγκεκριμένων κέντρων υπερυπολογιστών (supercomputers). Συνεπώς, για να καταφέρουμε να αξιοποιήσουμε στο μέγιστο βαθμό αυτά τα μαζικά σύνολα δεδομένων παρατήρησης της γης καθώς και τα περιβαλλοντικά σύνολα δεδομένων, προηγμένες μέθοδοι για την οργάνωση και ανάλυση τους απαιτούνται [83],[69],[29], [50] καθώς και η επεξεργαστική ισχύς, το ανθρώπινο δυναμικό και τα διαθέσιμα εργαλεία επεξεργασίας είναι αναγκαίο να έρθουν πιο "κοντά" στις αποθήκες δεδομένων [10],[58],[30],[51].

Η διαδικασία της εξόρυξης πολύτιμης γνώσης και πληροφορίας από μεγάλα σύνολα δεδομένων παρατήρησης της γης, παρουσιάζει σημαντικές τεχνολογικές προκλήσεις [60],[61],[72],[54], ενώ ο ολοένα και αυξανόμενος όγκος των αποθηκευμένων δεδομένων δεν είναι ο μόνος παράγοντας που καθιστά απαιτητική τη διαδικασία. Καθώς ο πλούτος των δεδομένων αυξάνεται οι προκλήσεις της δημιουργίας κατάλληλων ευρετηρίων (indexing), της αναζήτησης και της μεταφοράς των δεδομένων αυξάνονται αντίστοιχα και μάλιστα σε υπερθετικό βαθμό. Ανοιχτά ζητήματα περιλαμβάνουν την αποτελεσματική αποθήκευση, χειρισμό, διαχείριση και μεταφορά των δεδομένων καθώς και την επεξεργασία διαφορετικών τύπων και πολυδιάστατων συνόλων δεδομένων όπως επίσης και τις ολοένα και αυξανόμενες απαιτήσεις για πραγματικού χρόνου ή σχεδόν πραγματικού χρόνου επεξεργασία για πολλές κρίσιμες γεωχωρικές εφαρμογές [74], [73],[83].

Επιπλέον τα τηλεπισκοπικά δεδομένα είναι πολύτυπα (multi-modal) καθώς συλλέγονται από πολλούς διαφορετικούς αισθητήρες όπως για παράδειγμα είναι οι πολυφασματικοί αισθητήρες, οι αισθητήρες radar και lidar κ.α. Εκτιμάται ότι τα αρχεία της NASA περιλαμβάνουν σχεδόν 7000 διαφορετικούς τύπους συνόλων δεδομένων παρατήρησης της γης. Όσον αναφορά στα πολυδιάστατα σύνολα δεδομένων (π.χ. υπερφασματικές εικόνες) αυτά περιέχουν πληροφορία σε εκατοντάδες διαφορετικά μήκη κύματος και συνεπώς μεγάλος όγκος πληροφορίας είναι αναγκαίο να αποθηκευτεί, να αναλυθεί, να μεταδοθεί και να επεξεργασθεί προς την κατεύθυνση της αξιοποίησης αυτών ετερογενών και πολυδιάστατων συνόλων δεδομένων.

Η ανάπτυξη καινοφανών διαδικτυακών υπηρεσιών γεωχωρικών δεδομένων [84],[81], [54] για την ανάλυση, κατάπαίτηση, τηλεπισκοπικών δεδομένων αποτελεί ένα θεμελιώδες ζήτημα. Οι διαδικτυακές υπηρεσίες γεωχωρικών δεδομένων δίνουν τη δυνατότητα στους χρήστες να μεγιστοποιήσουν την αξιοποίηση των διανεμημένων γεωχωρικών δεδομένων καθώς και των υπολογιστικών πόρων κατά μήκος του διαδικτύου ώστε να επιτευχθεί η αυτοματοποίηση των ενεργειών αφενός της ενσωμάτωσης των γεωχωρικών δεδομένων και αφετέρου των αναλυτικών διαδικασιών. Είναι επιτακτική ανάγκη οι υπηρεσίες αυτές να είναι διαλειτουργικές και να επιτρέπουν τη συνεργατική επεξεργασία των γεωχωρικών συνόλων δεδομένων προς την κατεύθυνση της εξόρυξης πληροφορίας και γνώσης. Τα προαναφερθέντα χαρακτηριστικά είναι δυνατόν να επιτευχθούν μέσω της χρησιμοποίησης των τεχνολογιών των υπολογιστικών υπηρεσιών (service computing technologies) και των υπηρεσιών ροής εργασιών (workflow technologies) [82],[40].

Όλες οι προαναφερθείσες πτυχές του ζητήματος της διαχείρισης και χρήσης των γεωχωρικών δεδομένων τυγχάνουν ενεργής και εντατικής ερευνητικής δραστηριότητας στους κύκλους της επιστημονικής και βιομηχανικής κοινότητας προς την κατεύθυνση της εξεύρεσης καινοτό-

μων και πρωτότυπων νέων τεχνολογιών.

Στο πλαίσιο αυτό η εκπόνηση της παρούσας διπλωματικής εργασίας είχε σαν επακόλουθο τα εξής αποτέλεσματα. Άρχικά πραγματοποιήθηκε μια κριτική ανασκόπηση των τωρινών, state-of-the-art συστημάτων για μεγάλα δεδομένα με τις ικανότητες της διαχείρισης, επεξεργασίας, ανάλυσης και διαμοιρασμού μεγάλων συνόλων δεδομένων καθώς και προϊόντων προστιθέμενης αξίας. Πραγματοποιήθηκε επίσης ανάλυση πρόσφατα ανεπτυγμένων εργαλείων ειδικά για γεωχωρικά δεδομένα όπως για παράδειγμα είναι οι δέσμες (clusters) υπολογιστικών συστημάτων υψηλών επιδόσεων, οι πλατφόρμες νέφους υπολογιστών, τα παράλληλα συστήματα αρχείων και οι βάσεις δεδομένων. Όλη η παραπάνω πληροφορία αξιοποιήθηκε στο έπακρο και οδήγησε στο σχεδιασμό, τη δημιουργία και την υλοποίηση ενός συστήματος διαχείρισης και επεξεργασίας μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών γενικού σκοπού το οποίο μπορεί να αξιοποιηθεί από εφαρμογές παρατήρησης της Γης έως υδρολογικές και δημογραφικές εφαρμογές. Τέλος, με βάση τις τρέχουσες συνθήκες, τη γνώση που αποκτήθηκε στο πλαίσιο αυτής της εργασίας αλλά και τις αναδυόμενες προκλήσεις, έγινε μια προσπάθεια να παρουσιασθούν συγκεκριμένα ζητήματα, ιδέες και μελλοντικές κατευθύνσεις προς την αποτελεσματική εκμετάλλευση των μεγάλων δεδομένων παρατήρησης της γης για σημαντικές μηχανικές, περιβαλλοντικές και αγροτικές εφαρμογές.

1.2 Συνεισφορά

Η συνεισφορά της παρούσας διπλωματικής εργασίας συνίσταται στα παρακάτω σημεία :

- Στην εργασία αυτή έγινε η επίδειξη της διασύνδεσης σύγχρονων και καινοτόμων τεχνολογιών αποθήκευσης, διαχείρισης και ανάλυσης μεγάλων γεωχωρικών δεδομένων (τηλεπισκοπικά δεδομένα) σε περιβάλλον cluster υπολογιστών. Για την επιλογή των τεχνολογιών που χρησιμοποιήθηκαν στο σύστημα πραγματοποιήθηκε σύγκριση μεταξύ διάφορων συστημάτων, εργαλείων και αρχιτεκτονικών με συνέπεια να διασφαλίζεται ότι η προτεινόμενη αρχιτεκτονική συστήματος είναι ικανή να εξυπηρετήσει τις κολοσσιαίες απαιτήσεις της ανάλυσης μεγάλων γεωχωρικών δεδομένων.

- Επίσης πραγματοποιήθηκε ο σχεδιασμός και η υλοποίηση ενός επιχειρησιακού συστήματος διαχείρισης και επεξεργασίας μεγάλων γεωχωρικών δεδομένων το οποίο είναι ικανό να χρησιμοποιηθεί για πολύ μεγάλο εύρος εφαρμογών ανάλυσης τόσο raster δεδομένων όσο και vector δεδομένων ή και συνδυασμό τους. Οι εφαρμογές αυτές περιλαμβάνουν τόσο μη διαδραστικές εφαρμογές (batch processing) όσο και εφαρμογές πραγματικού χρόνου (real-time processing). Άρα, η συνεισφορά της εργασίας για το σημείο αυτό, έγκειται στη δημιουργία ενός συστήματος διαχείρισης μεγάλων γεωχωρικών δεδομένων γενικού σκοπού το οποίο μπορεί να χρησιμοποιηθεί ανάλογα με τις υπάρχουσες ανάγκες και δυνατότητες και αποκρύπτει από τον τελικό χρήστη (προγραμματιστή ανάπτυξης εφαρμογών ανάλυσης μεγάλων γεωχωρικών δεδομένων) όλες τις λεπτομέρειες σχεδίασης και υλοποίησης ενός υπολογιστικού συστήματος μεγάλης κλίμακας προσφέροντάς του τη δυνατότητα να ασχοληθεί απρόσκοπτα και απροβλημάτιστα με την ανάπτυξη και

υποβολή αλγορίθμων επεξεργασίας στο σύστημα καθώς και τη συλλογή των αποτελεσμάτων.

1.3 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια.

Στο **Κεφάλαιο 2** παρουσιάζονται οι βασικές τεχνολογίες που σχετίζονται με την εργασία αυτή. Αρχικά περιγράφονται οι διαθέσιμες επιλογές σχετικά με τις πλατφόρμες επεξεργασίας μεγάλων γεωχωρικών δεδομένων, στη συνέχεια παρουσιάζονται οι πιο δημοφιλείς και υποσχόμενες από αυτές ενώ τέλος γίνεται μια σύντομη αναφορά σε σύγχρονα ερευνητικά θέματα επεξεργασίας μεγάλων γεωχωρικών δεδομένων.

Στο **Κεφάλαιο 3** παρουσιάζεται η θεματική βάση των εφαρμογών ανάλυσης τηλεπισκοπικών δεδομένων οι οποίες υλοποιήθηκαν στα πλαίσια αυτής της εργασίας. Η βάση αυτή είναι η σύγχρονη τάση για την επεξεργασία και την εξαγωγή γνώσης από ολόκληρα αρχεία τηλεπισκοπικών δεδομένων μέσω της διενέργειας διαχρονικών αναλύσεων με τη βοήθεια κατανεμημένων συστημάτων επεξεργασίας υψηλών επιδόσεων. Επιχείρειται η προσέγγιση αυτής της τάσης μέσω της παρουσίασης διάφορων εφαρμογών για την παρατήρηση της γης και κυρίως για περιβαλλοντικές και αγροτικές εφαρμογές.

Στο **Κεφάλαιο 4** παρουσιάζεται η ανάλυση και η σχεδίαση του συστήματος που δημιουργήθηκε ως αποτέλεσμα αυτής της εργασίας και δίνεται η περιγραφή των υποσυστημάτων και των εφαρμογών του. Γίνεται επίσης εκτενής παρουσίαση της υλοποίησης του συγκεκριμένου συστήματος για τη διενέργεια αναλύσεων και την εξαγωγή γνώσης από μεγάλα γεωχωρικά δεδομένα καθώς και παρέχονται λεπτομέρειες σχετικά με τις πλατφόρμες και τα προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν καθώς και με την ανάλυση των βασικών αλγορίθμων του και της δομής του κώδικα.

Στο **Κεφάλαιο 5** γίνεται η παρουσίαση των αποτελεσμάτων των αλγορίθμων που αναπτύχθηκαν καθώς και πραγματοποιείται αξιολόγησή τόσο των αποτελεσμάτων όσο και του συστήματος που χρησιμοποιήθηκε για την παραγωγή τους.

Τέλος στο **Κεφάλαιο 6** δίνονται τα συμπεράσματα αυτής της διπλωματικής εργασίας, καθώς και μελλοντικές επεκτάσεις της.

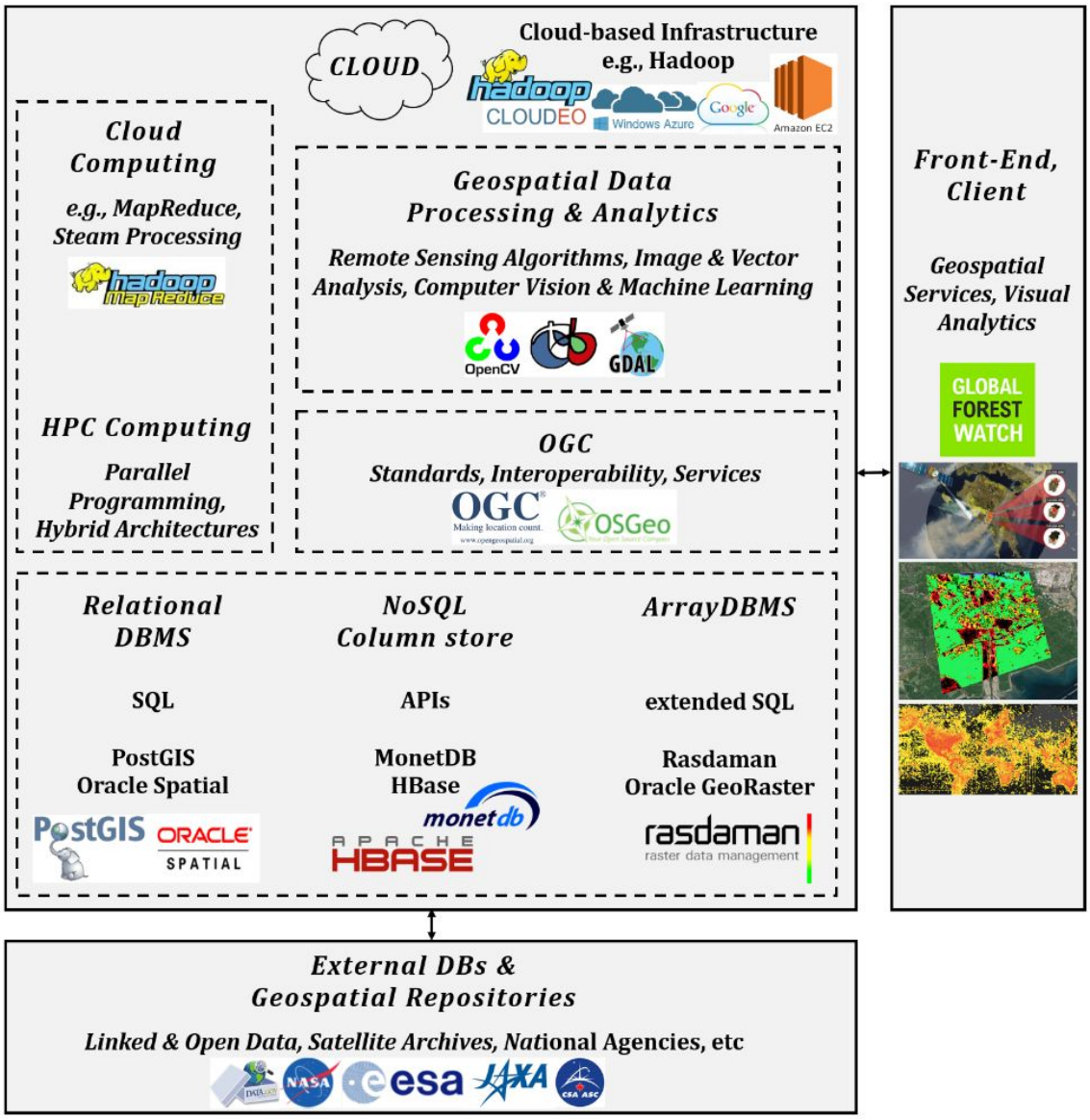
Κεφάλαιο 2

Πλατφόρμες επεξεργασίας Μεγάλων Γεωχωρικών Δεδομένων

Στο κεφάλαιο αυτό θα γίνει προσπάθεια ώστε να περιγραφεί [55] το τεχνολογικό υπόβαθρο εκείνο το οποίο είναι ζωτικής σημασίας για την κατανόηση των πιο βασικών ζητημάτων που αφορούν τις πλατφόρμες επεξεργασίας μεγάλων γεωχωρικών δεδομένων.

Ο ολοένα και αυξανόμενος αριθμός συνόλων γεωχωρικών δεδομένων [56] ξεπερνά τις δυνατότητες των τωρινών συστημάτων να τα εξερευνήσουν και να τα ερμηνεύσουν. Οι εργασίες της αποθήκευσης, του χειρισμού, της ανάκτησης, της ανάλυσης και της δημοσίευσης μεγάλων γεωχρικών δεδομένων [62] θέτουν σημαντικές προκλήσεις και δημιουργούν ευκαιρίες για πολλές πτυχές των state-of-the-art συστημάτων επεξεργασίας μεγάλων γεωχωρικών δεδομένων (Σχήμα 2.1). Διάφορα συστατικά στοιχεία συμπεριλαμβάνονται όπως για παράδειγμα είναι οι υποδομές νέφους υπολογιστών, τα πρότυπα διαλειτουργικότητας, τα παράλληλα συστήματα και προγραμματιστικά μοντέλα, τα Συστήματα Διαχείρισης Βάσεων Δεδομένων (DBMSs) σε πολυεπίπεδες ιεραρχίες μνήμης και η χρονοδρομολόγηση εργασιών. Επιπροσθέτως, η ολοένα και αυξανόμενη ποσότητα και ποιότητα των γεωχωρικών δεδομένων που προέρχονται από τηλεπισκοπικές πλατφόρμες, από παραδοσιακές τεχνολογίες GIS συστημάτων όπως επίσης και γεωχωρικά δεδομένα που προέρχονται από μια σειρά νέων πηγών όπως είναι τα μέσα κοινωνικής δικτύωσης και τα σύνολα δεδομένων του λεγόμενου "Internet of Things" παρέχουν μεγάλες ευκαιρίες προς την απάντηση νέων και μεγαλύτερων ερωτήματων από γεωχωρικής άποψης.

Η εμφάνιση των μεγάλων γεωχωρικών συνόλων δεδομένων έχει δημιουργήσει επανάσταση στις τεχνικές για την ανάλυση και την εξαγωγή πολύτιμων πληροφοριών από αυτές τις ολοένα και αυξανόμενες ροές γεωχωρικών συνόλων δεδομένων. Η ταχεία επεξεργασία μεγάλων γεωχωρικών δεδομένων που χαρακτηρίζονται από ολοένα και αυξανόμενο όγκο και πολυπλοκότητα αποτελεί μια μεγάλη πρόκληση για τα τωρινά συστήματα επεξεργασίας μεγάλων δεδομένων. Υπάρχει μια επείγουσα ανάγκη για την εξεύρεση καινοτόμων λύσεων στην αρχιτεκτονική συστημάτων και ειδικά προς την επίτευξη της εγγενούς κλιμακωσιμότητας της



Σχήμα 2.1: Η κυρίαρχη αρχιτεκτονική και τεχνολογίες για τη διαχείριση μεγάλων γεωχωρικών δεδομένων και τη διενέργεια analytics.(Πηγή: [55])

βασικής αρχιτεκτονικής του υλικού και του λογισμικού. Ειδικότερα, αυτά τα συστήματα που κάνουν έντονη χρήση δεδομένων (data-intensive systems) πρέπει να επιδεικνύουν δυνατότητες γραμμικής κλιμάκωσης ώστε να μπορούν να φιλοξενήσουν την επεξεργασία γεωχωρικών δεδομένων οποιουδήποτε όγκου.

Για τον σκοπό της εκπλήρωσης της απαίτησης της επεξεργασίας των δεδομένων σε πραγματικό χρόνο για κάποιες γεωχωρικές εφαρμογές, εύκολες διαδικασίες για την προσθήκη επιπλέον υπολογιστικών πόρων σε υπάρχοντα συστήματα είναι επίσης αναγκαίες. Από την άποψη της αποτελεσματικότητας σε επιδόσεις είναι κρίσιμης σημασίας για data-intensive πλατφόρμες να συμμορφώνονται με την αρχή "της μετακίνησης των επεξεργασιών στα δεδομένα" [35] ώστε να ελαχιστοποιείται κατά το δυνατόν η μετακίνηση των δεδομένων κατά μήκος του δικτύου. Συνεπώς, η ιεραρχία αποθήκευσης ενός συστήματος βελτιστοποιημένου για data-intensive

υπολογισμούς πιθανότατα θα τοποθετούσε τα δεδομένα τοπικά ώστε να μειωθούν οι καθυστερήσεις από το δίκτυο και το σύστημα οι οποίες και εισάγονται από την μεταφορά των δεδομένων.

Οι μαζικές απαιτήσεις αποθήκευσης από τις βάσεις δεδομένων, η ανάγκη για πινακοποιημένα μοντέλα αποθήκευσης για μεγάλης κλίμακας επιστημονικούς υπολογισμούς και μεγάλα αρχεία εξόδου καθώς επίσης και η ανάγκη για την επίτευξη υψηλού ταυτοχρονισμού και ρυθμοαπόδοσης ανά server [24] είναι βασικές απαιτήσεις για εφαρμογές που τρέχουν σε υψηλής κλιμακωσιμότητας υπολογιστικές δέσμες.

Επί του παρόντος, ποικίλες πλατφόρμες υψηλών επιδόσεων επιστρατεύονται σε μια προσπάθεια να εκπληρωθούν οι προαναφερθείσες απαιτήσεις και να πραγματοποιηθεί επεξεργασία αυτών των μεγάλων γεωχωρικών δεδομένων. Οι κυρίαρχες επιλογές για αυτές τις πλατφόρμες επικεντρώνονται κυρίως σε καινοτόμες πλατφόρμες βάσεων δεδομένων, σε Cluster-Based HPC (*i.e.* high performance computing) συστήματα ή αλλιώς supercomputers όπως επίσης και σε πλατφόρμες οι οποίες βασίζονται σε νέφη υπολογιστών (Cloud-based platforms).

Όλες οι προαναφερθείσες επιλογές για τις πλατφόρμες επεξεργασίας θα συζητηθούν στην ενότητα 2.1. Σε συνέχεια της συζήτησης αυτής θα γίνει μια προσπάθεια (ενότητες 2.2 - 2.6) ώστε να παρουσιασθούν οι πιο δημοφιλείς όπως επίσης και οι πιο υποσχόμενες πλατφόρμες για τη διαχείριση και την επεξεργασία μεγάλων γεωχωρικών δεδομένων. Τέλος, στην ενότητα 2.7 θα πραγματοποιηθεί μια σύντομη αναφορά σε σύγχρονα ερευνητικά θέματα στον τομέα της επεξεργασίας μεγάλων γεωχωρικών δεδομένων ώστε να δοθεί στον αναγνώστη μια αίσθηση για την κατεύθυνση που θα ακολουθήσει η τεχνολογική εξέλιξη στο πεδίο των υπολογιστικών συστημάτων επεξεργασίας δεδομένων μεγάλης κλίμακας.

2.1 Databases, HPC systems, Cloud-based architectures

Παρ'ολο που το μέγεθος των μεγάλων δεδομένων συνεχίζει να αυξάνεται εκθετικά, οι τρέχουσες δυνατότητες για την επεξεργασία μεγάλων γεωχωρικών συνόλων δεδομένων είναι μόνο στα σχετικά χαμηλά επίπεδα των τάξεων μεγεθών των petabytes, exabytes και zettabytes δεδομένων. Επιπλέον, τα παραδοσιακά εργαλεία βάσεων δεδομένων και οι πλατφόρμες [24] δεν δύνανται να επεξεργαστούν γεωχωρικά σύνολα δεδομένων τα οποία αυξάνονται τόσο πολύ σε μέγεθος και πολυπλοκότητα. Προς την αντιμετώπιση αυτής της πρόκλησης πρωτότυπες και καινοτόμες λύσεις έχουν προταθεί και χρησιμοποιούνται επί του παρόντος. Οι δύο πιο δημοφιλείς, ισχυρές και εύρωστες είναι τα DBMSs τα οποία υλοποιούν το πινακοποιημένο μοντέλο δεδομένων από τη μία πλευρά και οι NoSQL (Not Only SQL) πλατφόρμες διαχείρισης βάσεων δεδομένων από την άλλη.

Τα πινακοποιημένα συστήματα διαχείρισης βάσεων δεδομένων (Array DBMSs) έχουν κατασκευαστεί ειδικά για την εξυπηρέτηση raster συνόλων δεδομένων. Καθως τα raster σύνολα δεδομένων αποτελούνται από pixel τα οποία βρίσκονται σε μία, δύο ή και περισσότερες διαστάσεις, η δομή δεδομένων τύπου πίνακα είναι η πιο κατάλληλη για τη μοντελοποίηση μεγάλων γεωχωρικών raster δεδομένων. Τα Array DBMSs υλοποιούν το πινακοποιημένο μοντέλο το οποίο υποστηρίζει τους πίνακες ως κατηγορήματα πρώτης τάξης. Το μοντέλο βασίζεται σε

μια άλγεβρα πινάκων [15] η οποία αναπτύχθηκε για τους σκοπούς των βάσεων δεδομένων και εισάγει τον πίνακα σαν ένα νέο τύπο ιδιότητας στο σχεσιακό μοντέλο. Σε συνδυασμό με το πινακοποιημένο μοντέλο μια νέα γλώσσα ερωτημάτων, η οποία επεκτείνει την τυπική SQL, έχει υλοποιηθεί και έχει εμπλουτισθεί με τελεστές ειδικά για πίνακες για την ανάκτηση και επεξεργασία raster δεδομένων. Τα Array DBMSs στην ουσία αποθηκεύουν και διαχειρίζονται δομημένα δεδομένα. Γνωστές υλοποιήσεις Array DBMSs περιλαμβάνουν τα συστήματα *Rasdaman*, MonetDB/SciQL, PostGIS, Oracle GeoRaster και SciDB.

Η NoSQL [38] είναι μια σύγχρονη προσέγγιση για μεγάλης κλίμακας, κατανεμημένη διαχείριση δεδομένων και σχεδιασμό βάσεων δεδομένων. Μια NoSQL βάση δεδομένων προσφέρει ένα μηχανισμό για την αποθήκευση και την ανάκτηση δεδομένων που είναι μοντελοποιημένα με διαφορετικό τρόπο από τις σχέσεις μεταξύ πινάκων των σχεσιακών βάσεων δεδομένων. Τα NoSQL συστήματα είναι είτε εξόλοκλήρου μη σχεσιακά είτε απλά αποφεύγουν να υλοποιήσουν επιλεγμένη σχεσιακή λειτουργικότητα όπως είναι τα αυστηρώς καθορισμένα σχήματα πινάκων και οι λειτουργίες join. Ο λόγος για τον οποίο πολλές mainstream πλατφόρμες μεγάλων δεδομένων υιοθετούν την προσέγγιση NoSQL είναι για να σπάσουν και να ξεπεράσουν την ακαμψία των κανονικοποιημένων σχημάτων των σχεσιακών DBMSs. Παρ'ολα αυτά, πολλές NoSQL πλατφόρμες βάσεων δεδομένων χρησιμοποιούν την SQL στα συστήματά τους καθώς η SQL είναι μια αξιόπιστη και απλή γλώσσα ερωτημάτων η οποία επιδεικνύει υψηλή επίδοση κατά τη διενέργεια αναλύσεων σε πραγματικό χρόνο, μεγάλων δεδομένων συνεχούς ροής.

Οι NoSQL πλατφόρμες βάσεων δεδομένων αποθηκεύουν και διαχειρίζονται αδόμητα δεδομένα με έναν τρόπο ο οποίος έρχεται σε αντίθεση με τις σχεσιακές βάσεις δεδομένων καθώς διαχωρίζει την αποθήκευση και τη διαχείριση των δεδομένων σε δύο ανεξάρτητα τμήματα αντί να αντιμετωπίζει τα δύο αυτά ζητήματα ταυτόχρονα. Αυτή η σχεδιαστική επιλογή παρέχει στα NoSQL συστήματα βάσεων δεδομένων μια σειρά από πλεονεκτήματα. Τα κυριότερα αυτών είναι η δυνατότητα για την επίτευξη της κλιμακωσιμότητας της αποθήκευσης των δεδομένων με υψηλές επιδόσεις όπως επίσης και η ευελιξία κατά την μοντελοποίηση των δεδομένων, την ανάπτυξη και την εγκατάσταση εφαρμογών [38]. Οι περισσότερες NoSQL βάσεις δεδομένων είναι απαλλαγμένες σχήματος. Αυτή η ιδιότητα δίνει τη δυνατότητα στις εφαρμογές να αλλάζουν γρήγορα τη δομή των δεδομένων χωρίς να είναι αναγκαίο να γράφονται από την αρχή οι πίνακες στους οποίους τα δεδομένα είναι αποθηκευμένα. Επιπλέον, επιτρέπει μεγαλύτερη ευελιξία στην περίπτωση που δομημένα δεδομένα αποθηκεύονται με ετερογενείς τρόπους. Γνωστές υλοποιήσεις NoSQL βάσεων δεδομένων οι οποίες εξυπηρετούν γεωχωρικά δεδομένα περιλαμβάνουν τα συστήματα MongoDB, Google BigTable, Apache HBASE και Cassandra. Εταιρείες οι οποίες χρησιμοποιούν NoSQL πλατφόρμες βάσεων δεδομένων περιλαμβάνουν τις Google, Facebook, Twitter, LinkedIn και NetFlix.

Εκτός από τις πλατφόρμες βάσεων δεδομένων μια μεγάλη ποικιλία από cluster-based HPC συστήματα συμπεριλαμβανομένων των grid computing, cluster computing και ubiquitous computing χρησιμοποιούνται για την επεξεργασία μεγάλων συνόλων δεδομένων και την εξαγωγή χρήσιμων πληροφοριών. Μια cluster πλατφόρμα [74], [73] συνήθως αντιμετωπίζει ένα μεγάλο υπολογιστικό πρόβλημα μέσω της συνεργατικής εργασίας πολλαπλών υπολογιστικών κόμβων οι οποίοι όμως προσφέρουν την εικόνα ενός μοναδικού συστήματος. Την

τρέχουσα περίοδο, οι cluster πλατφόρμες είναι η κυρίαρχη αρχιτεκτονική όσον αναφορά τους υπολογισμούς υψηλών επιδόσεων και μεγάλης κλίμακας επιστημονικές εφαρμογές. Σημαντικοί οργανισμοί και επιχειρήσεις όπως η NASA [57] και η Google [11] έχουν αναπτύξει μεγάλα cluster συστήματα επεξεργασίας τα οποία αποτελούνται από πολλούς υπολογιστικούς κόμβους για την επεξεργασία γεωχωρικών δεδομένων. Αυτού του είδους τα συστήματα [62] έχουν τη δυνατότητα να προσφέρουν υψηλού επιπέδου χωρητικότητα δεδομένων, ρυθμοαπόδοση και διαθεσιμότητα μέσω της ανεκτικότητας λαθών από την πλευρά του λογισμικού, τη δημιουργία αντιγράφων των πρωτότυπων δεδομένων και βελτιστοποιημένη διαχείριση συστήματος.

Τα HPC συστήματα εξελίσσονται προς υβριδικές αρχιτεκτονικές και αρχιτεκτονικές με επιταχυντές περιέχοντας όχι μόνο πολυπύρηνες CPUs αλλά και GPUs [37]. Επιπλέον, εξοπλίζονται με υψηλής επίδοσης δικτύωση τύπου Infiniband για την επίτευξη πολύ υψηλού εύρους ζώνης και συνεπώς πολύ μικρή καθυστέρηση δικτύου για την επικοινωνία μεταξύ των υπολογιστικών κόμβων του συστήματος. Τα HPC συστήματα είναι προσανατολισμένα σε υπολογιστικά έντονες εφαρμογές (compute-intensive oriented) και σαν συνέπεια και η αρχιτεκτονική συστήματος και τα διάφορα εργαλεία δεν είναι βελτιστοποιημένα για την υποστήριξη data-intensive εφαρμογών όπου η διαθεσιμότητα των δεδομένων είναι το κύριο ζητούμενο. Συνεπώς, παρά τις τεράστιες υπολογιστικές δυνατότητες που έχουν τα HPC συστήματα, η αποτελεσματική επεξεργασία μεγάλων γεωχωρικών δεδομένων μέσω των υπαρχόντων cluster-based HPC συστημάτων παραμένει ακόμα μια μεγάλη πρόκληση. Για την αντιμετώπιση αυτού του ζητήματος συντελείται μια μετάβαση προς συστήματα που είναι δομημένα σε πολλαπλά ιεραρχικά επίπεδα. Τα συστήματα αυτά έχουν μεγαλύτερης διαστασιμότητας τοπολογία σύνδεσης όπως επίσης και πολυεπίπεδη αρχιτεκτονική αποθήκευσης. Όσον αναφορά την αποδοτικότητα της επεξεργασίας των δεδομένων είναι κρίσιμης σημασίας να ληφθεί υπ'οψη η τοπικότητα των δεδομένων. Ο προγραμματισμός εφαρμογών σε αυτά τα πολλαπλά επίπεδα τοπικότητας και σε μεγάλης διαστασιμότητας τοπολογίες συνδέσεων αποτελεί ακόμα μία σημαντική πρόκληση και πρέπει να διανυθεί ακόμα πολύς δρόμος προς την κατεύθυνση της εύρεσης βιώσιμων λύσεων.

Η ανάπτυξη των εικονικών τεχνολογιών [62] έχει καταστήσει τους υπερυπολογιστές πολύ πιο προσιτούς δια μέσου της χρήσης κοινού και σχετικά φθηνού υλικού αντί των πολύ ακριβών HPC συστημάτων. Ισχυρές υπολογιστικές υποδομές αποκρύπτονται πίσω από λογισμικό εικονικών μηχανών το οποίο δίνει τη δυνατότητα σε αυτά τα συστήματα να συμπεριφέρονται σαν ένας αληθινός και φυσικός H/Υ εμπλουτισμένος με τη δυνατότητα της ευελιξίας κατά τον καθορισμό των προδιαγραφών της εικονικής μηχανής όπως για παράδειγμα είναι ο αριθμός των επεξεργαστικών πυρήνων, η μνήμη, το μέγεθος του δίσκου και το λειτουργικό σύστημα. Η χρήση αυτών των εικονικών H/Υ είναι γνωστή σαν νέφος υπολογιστών (cloud computing) [33] και αποτελεί μια από τις πιο εύρωστες τεχνικές κατά την χρήση μεγάλων δεδομένων.

Για εφαρμογές μεγάλων γεωχωρικων δεδομένων πραγματικού χρόνου, η έγκαιρη απόκριση του συστήματος όταν ο όγκος των δεδομένων είναι πολύ μεγάλος αποτελεί κορυφαία προτεραιότητα. Η τεχνική cloud computing ([78], [3], [31]) ενσωματώνει λογισμικό, υπολογισμούς και δεδομένα χρηστών ώστε να παρέχει απομακρυσμένες υπηρεσίες δια μέσου της άθροισης πολλαπλών διαφορετικών ροών εργασιών σε μία μεγάλη δέσμη επεξεργαστικών κόμβων. Το cloud computing [24] όχι μόνο διανέμει εφαρμογές και υπηρεσίες κατά μήκος του διαδικτύου

αλλά επίσης έχει επεκταθεί ώστε να παρέχει τις υποδομές σαν υπηρεσία (infrastructure as a service (IaaS)), όπως για παράδειγμα είναι το Amazon EC2, τις πλατφόρμες σαν υπηρεσία (platform as a service (PaaS)), όπως για παράδειγμα είναι το Google AppEngine και το Microsoft Azure καθώς και το λογισμικό σαν υπηρεσία (software as a service (SaaS)). Επιπλέον, η αποθήκευση σε υποδομές cloud computing παρέχει ένα εργαλείο για την αποθήκευση μεγάλων δεδομένων με πολύ καλές προοπτικές κλιμακωσιμότητας.

Το cloud computing είναι μια πολύ εφικτή τεχνολογία και έχει προσελκύσει ένα μεγάλο αριθμό ερευνητών για να το αναπτύξουν και να δοκιμάσουν να εφαρμόσουν τις λύσεις του σε προβλήματα μεγάλων δεδομένων. Είναι αναγκαίο να αναπτυχθούν πλατφόρμες λογισμικού και αντίστοιχα προγραμματιστικά μοντέλα τα οποία θα αξιοποιούν στο μέγιστο τις αρχές και τις δυνατότητες του cloud computing για την αποθήκευση και την επεξεργασία μεγάλων δεδομένων. Προς αυτήν την κατεύθυνση, το Apache Hadoop είναι μια από τις πιο καλώς ανεπτυγμένες πλατφόρμες λογισμικού η οποία υποστηρίζει data-intensive, κατανεμημένες και παράλληλες εφαρμογές. Υλοποιεί το υπολογιστικό πρότυπο το οποίο ονομάζεται Map/Reduce. Η πλατφόρμα Apache Hadoop αποτελείται από τον Hadoop kernel, το Map/Reduce και το κατανεμημένο σύστημα αρχείων του συστήματος Hadoop το Hadoop distributed file system (HDFS) το οποίο προσφέρει στρατηγικές και αντιγραφή δεδομένων ώστε να επιτευχθεί η ανεκτικότητα του συστήματος σε αστοχίες υλικού και καλύτερες επιδόσεις πρόσβασης στα δεδομένα όπως επίσης και ένα σημαντικό αριθμό από σχετιζόμενα projects τα οποία τρέχουν "πάνω" από το Hadoop όπως είναι για παράδειγμα τα Apache Hive, Apache HBase, Apache Spark, κ.α.

Το Map/Reduce [27] είναι ένα προγραμματιστικό μοντέλο και ένα σχήμα εκτέλεσης για την επεξεργασία και δημιουργία ενός μεγάλου όγκου από σύνολα δεδομένων. Αρχικά η εισαγωγή του και η ανάπτυξη του πραγματοποιήθηκε από την Google και μετά την απελευθέρωση του στο κοινό αναπτύχθηκε περαιτέρω από τη Yahoo καθώς και από άλλες εταιρείες. Το Map/Reduce βασίζεται στο αλγοριθμικό σχεδιαστικό πρότυπο "διαίρει και κυρίευε" και εργάζεται αναδρομικά μέσω της διάσπασης ενός σύνθετου προβλήματος σε πολλά υπο-προβλήματα (βήμα Map), μέχρι να έχουν το απαιτούμενο μέγεθος ώστε να μπορούν να επιλυθούν απευθείας. Έπειτα τα υπο-προβληματα επιλύονται ξεχωριστά και παράλληλα (βήμα Reduce). Οι λύσεις των υπο-προβλημάτων στη συνέχεια συνδυάζονται ώστε να δώσουν την ολοκληρωμένη λύση του αρχικού προβλήματος.

Όλες οι πολύ γνωστές εταιρείες χρησιμοποιούν το cloud computing ως μέσο για να παρέχουν τις υπηρεσίες τους. Εκτός από την Google, πρόσφατα και η Yahoo έχει αναπτύξει τη δική της μηχανή αναζήτησης σε έναν Hadoop cluster. Επιπλέον, και το Facebook αλλά και το eBay έχουν αναπτύξει τις δικές τους μεγάλες εφαρμογές της τάξης των Exabytes με τη βοήθεια του Hadoop. Επιπροσθέτως, για μεγάλης κλίμακας επεξεργασίες γεωχωρικών δεδομένων, το πλαίσιο επεξεργασίας Hadoop GIS [5] έχει επίσης αναπτυχθεί με τη βοήθεια του συστήματος Hadoop.

2.2 Rasdaman

Το *Rasdaman* είναι ένα καθολικό (ανεξάρτητο πεδίου εφαρμογής) Array DBMS [14], [16], [13] το οποίο προσφέρει δυνατότητες για την αποθήκευση, διαχείριση και επεξεργασία μεγάλων raster δεδομένων. Ανεξάρτητο πεδίου σημαίνει ότι το *Rasdaman* μπορεί να λειτουργήσει ως η βασική πλατφόρμα βάσης δεδομένων σε ένα μεγάλο εύρος από εφαρμογές βάσεων δεδομένων συμπεριλαμβανομένων του online analytical processing (OLAP), των στατιστικών, των επιστημών της γης και του διαστήματος, των ιατρικών απεικονίσεων, των αεροσηράγγων, των προσομοιώσεων και των multimedia.

Το *Rasdaman* υποστηρίζει πολυδιάστατους πίνακες πολύ μεγάλου μεγέθους και αυθαίρετου αριθμού διαστάσεων οι οποίοι μπορούν να περιέχουν ένα αξιοσημείωτα πλούσιο εύρος πληροφορίας. Από χρονοσειρές μίας διάστασης και δυσ-διάστατες εικόνες σε κύβους δεδομένων OLAP με πολύ μεγάλο πλήθος διαστάσεων. Εξαιτίας των σχεδιαστικών επιλογών και των δυνατοτήτων του, το σύστημα μπορεί εγγενώς να χειριστεί μεγάλα δορυφορικά εικονιστικά δεδομένα. Η αρχιτεκτονική του Το *Rasdaman* βασίζεται στη διαφανή κατάτμηση πινάκων η οποία ονομάζεται tiling. Εννοιολογικά, δεν υπάρχει περιορισμός μεγέθους στους πίνακες που εξυπηρετεί το *Rasdaman* σαν το κεντρικό DBMS για raster σύνολα δεδομένων. Παρέχει μια πλούσια και ισχυρή γλώσσα ερωτημάτων (RasQL) η οποία μοιάζει με την SQL αλλά έχει σχεδιαστεί και υλοποιηθεί ειδικά για την εξυπηρέτηση raster συνόλων δεδομένων. Η (RasQL) είναι μια γενικού σκοπού δηλωτική γλώσσα ερωτημάτων η οποία έχει εμπλουτισθεί με εσωτερική εκτέλεση ερωτημάτων καθώς και βελτιστοποιήσεις κατά την αποθήκευση και τη μεταφορά δεδομένων. Επιπροσθέτως, το *Rasdaman* χαρακτηρίζεται από παράλληλη αρχιτεκτονική server η οποία προσφέρει ένα κλιμακώσιμο, κατανεμημένο περιβάλλον για την αποδοτική επεξεργασία ενός μεγάλου αριθμού από ταυτόχρονα αιτήματα χρηστών καθώς και την εξυπηρέτηση κατανεμημένων συνόλων δεδομένων κατά μήκος του διαδικτύου.

Το *Rasdaman* έχει αποδείξει [1] ([69], [52], [12], [53]) την αποδοτικότητα και την αποτελεσματικότητά του που οφείλεται στην ισχυρή γλώσσα ερωτημάτων του, στη διαφανή κατάτμηση πινάκων η οποία ακυρώνει τους περιορισμούς μεγέθους ενός μεμονωμένου αντικειμένου και δίνει τη δυνατότητα για την επίτευξη της κλιμακωσιμότητας όπως επίσης και στο χαρακτηριστικό της υποστηριζόμενης από το σύστημα συμπίεσης των tiles για την επίτευξη της μείωσης του αποθηκευτικού χώρου που χρειάζεται η βάση δεδομένων.

Επιπλέον, το *Rasdaman* υλοποιεί πολλά από τα πρότυπα του οργανισμού OGC προς την επίτευξη της διαλειτουργικότητας με άλλα συστήματα. Ειδικότερα, για το πρότυπο διεπαφής WCPS το *Rasdaman* αποτελεί την υλοποίηση αναφοράς [17]. Το πρότυπο διεπαφής WCPS ορίζει μια γλώσσα ερωτημάτων η οποία επιτρέπει την ανάκτηση, το φιλτράρισμα, την επεξεργασία και τη γρήγορη εξαγωγή υποσυνόλων από πολυδιάστατα raster δεδομένα τύπου coverage όπως για παράδειγμα είναι τα δεδομένα από αισθητήρες, προσομοιώσεις και στατιστικές.

Τα ερωτήματα γραμμένα στην WCPS υποβάλλονται στον εξυπηρετητή βάσεων δεδομένων του *Rasdaman* δια μέσου του δομικού συστατικού λογισμικού PetaScope [4]. Το PetaScope είναι ένα πακέτο από java servlets το οποίο υλοποιεί τα πρότυπα διεπαφών του OGC

[1] `http://www.copernicus-masters.com/index.php?kat=winners.html&anzeige=winner_t-systems2014.html`

με συνέπεια να επιτρέπει την κατάπαίτηση υποβολή ερωτημάτων τα οποία επιτελούν εργασίες αναζήτησης, ανάκτησης και επεξεργασίας πολυδιάστατων πινάκων πολύ μεγάλου μεγέθους. Επιπλέον, προσθέτει υποστήριξη στο σύστημα για γεωγραφικά και χρονικά συστήματα αναφοράς συντεταγμένων και έτσι μετατρέπει το *Rasdaman* σε έναν πλήρη και εύρωστο εξυπηρετητή μεγάλων γεωχωρικών δεδομένων. Η κοινότητα που αναπτύσσει το *Rasdaman* διανέμει το λογισμικό υπό την άδεια *Rasdaman Community* η οποία διανέμει το λογισμικό του εξυπηρετητή υπό την άδεια GPL και όλα τα υπόλοιπα μέρη υπό την άδεια LGPL, οπότε και είναι δυνατή η χρήση του συστήματος σε οποιοδήποτε περιβάλλον αδειών λογισμικού.

2.3 MonetDB

Ένα άλλο πλαίσιο το οποίο έχει επιτυχώς χρησιμοποιηθεί σε εφαρμογές δεδομένων παρατήρησης της γης είναι η MonetDB η οποία είναι ένα ανοιχτού κώδικα column-oriented DBMS. Η MonetDB [65] σχεδιάστηκε ώστε να επιτυγχάνει υψηλή επίδοση κατά την εκτέλεση περίπλοκων ερωτημάτων σε πολύ μεγάλου όγκου βάσεις δεδομένων. Για παράδειγμα κατά το συνδυασμό πινάκων που έχουν εκατοντάδες στήλες και εκατομμύρια γραμμές. Η MonetDB έχει εφαρμοστεί σε ένα μεγάλο εύρος εφαρμογών υψηλών επιδόσεων όπως είναι το OLAP, η εξόρυξη δεδομένων, τα GIS, η επεξεργασία δεδομένων συνεχούς ροής, η ανάκτηση κειμένου καθώς και οι επεξεργασίες ευθυγράμμισης ακολουθιών. Χρησιμοποιήθηκε επιτυχώς ως η βάση δεδομένων του οπίσθιου τμήματος ενός συστήματος που παρέχει υπηρεσίες παρακολούθησης πυρκαγιών σε πραγματικό χρόνο και εκμεταλλεύεται δορυφορικές εικόνες και διασυνδεδεμένα ανοικτά γεωχωρικά δεδομένα.

Η αρχιτεκτονική της MonetDB [43] αναπαρίσταται από 3 επίπεδα, καθένα από τα οποία παρέχει το δικό του σύνολο βελτιστοποιήσεων. Το εμπρόσθιο τμήμα βρίσκεται στο υψηλότερο επίπεδο και παρέχει διεπαφές ερωτημάτων για τις γλώσσες προγραμματισμού γενικού σκοπού SQL, SciQL και SPARQL. Τα ερωτήματα αναλύονται σε αναπαραστάσεις εξαρτώμενες από το πεδίο, όπως για παράδειγμα είναι η σχεσιακή άλγεβρα για την SQL και στη συνέχεια βελτιστοποιούνται. Τα παραγόμενα λογικά σχέδια εκτέλεσης στη συνέχεια μεταφράζονται σε εντολές της γλώσσας MonetDB Assembly Language (MAL) και περνάνε στο επόμενο επίπεδο. Το μεσαίο ή αλλιώς το επίπεδο οπίσθιου τμήματος παρέχει έναν αριθμό από βελτιστοποιητές βασιζόμενους στο κόστος για τη γλώσσα MAL. Το κατώτερο επίπεδο αποτελείται από τον πυρήνα της βάσης δεδομένων ο οποίος παρέχει πρόσβαση στα δεδομένα τα οποία αποθηκεύονται σε πίνακες Binary Association Tables (BATs). Κάθε πίνακας BAT αποτελείται από ένα μοναδικό αναγνωριστικό αντικειμένου και τιμές στηλών τα οποία και αναπαριστούν μία και μόνη στήλη η οποία είναι αποθηκευμένη στη βάση δεδομένων.

Η εσωτερική αναπαράσταση των δεδομένων της MonetDB επαφίεται επίσης στα εύρη διευθυνσιοδότησης των σύγχρονων κεντρικών επεξεργαστικών μονάδων οι οποίες χρησιμοποιούν κατάπαίτηση σελιδοποίηση των χαρτογραφημένων στη μνήμη αρχείων. Αυτό έχει σαν συνέπεια την απαγκίστρωση από το σχεδιασμό των παραδοσιακών DBMSs τα οποία επιτελούν σύνθετη διαχείριση μεγάλων συνόλων δεδομένων σε σχετικά περιορισμένη μνήμη.

Η αρχιτεκτονική της είναι πράγματι πρωτοποριακή και επίσης ενσωματώνει τη λογική της

ανακύκλωσης των ερωτημάτων(query recycling). Η ανακύκλωση ερωτημάτων [45] είναι μια αρχιτεκτονική για την επαναχρησιμοποίηση των υποπροϊόντων του προτύπου "operator-at-a-time" σε ένα column store DBMS. Η ανακύκλωση κάνει χρήση της γενικευμένης ιδέας της αποθήκευσης και επαναχρησιμοποίησης των αποτελεσμάτων ακριβών υπολογισμών και χρησιμοποιεί ένα βελτιστοποιητή ο οποίος θα επιλέγει τις εντολές οι οποίες θα αποθηκεύονται. Η τεχνική λειτουργεί με έναν αυτο-οργανώμενο τρόπο και έχει σχεδιαστεί ώστε να βελτιώνει τους χρόνους απόκρισης των ερωτημάτων και τη ρυθμοαπόδοση του συστήματος.

Επιπλέον, η MonetDB ήταν μία από τις πρώτες βάσεις δεδομένων που εισήγαγε την έννοια του Database Cracking. Το Database Cracking [44] είναι ένα αθροιστικό μερικό ευρετήριο και/ή πραγματοποιεί ταξινόμηση των δεδομένων. Εκμεταλλεύεται απέυθείας τη columnar φύση της MonetDB. Το Cracking είναι μία τεχνική η οποία μετατοπίζει το κόστος της συντήρησης του ευρετηρίου από τις ενημερώσεις στην εκτέλεση των ερωτημάτων. Οι βελτιστοποιητές της διοχέτευσης των ερωτημάτων χρησιμοποιούνται για να εντοπίσουν τα μοτίβα των ερωτημάτων και να διαδώσουν αυτήν την πληροφορία. Η τεχνική αυτή δίνει τη δυνατότητα για βελτιωμένους χρόνους πρόσβασης και αυτο-οργανούμενη συμπεριφορά.

Επιπροσθέτως, η MonetDB επιδεικνύει το τμήμα λογισμικού MonetDB/SQL/GIS το οποίο υλοποιεί μια διεπαφή στην προδιαγραφή Simple Feature του OGC και συνεπώς υποστηρίζει όλα τα αντικείμενα και τις συναρτήσεις που ορίζονται στην προδιαγραφή. Αυτό το χαρακτηριστικό ανοίγει το δρόμο για την εξυπηρέτηση γεωχωρικών δεδομένων και την ανάπτυξη γεωχωρικών εφαρμογών. Η υλοποίηση της προδιαγραφής Simple Feature δίνει τη δυνατότητα στη MonetDB να λειτουργεί ως ένας εξυπηρετητής γεωχωρικών δεδομένων.

2.4 MrGeo

Η υπηρεσία National Geospatial-Intelligence Agency (NGA) [66] σε συνεργασία με την εταιρεία DigitalGlobe[2], πρόσφατα εξέδωσαν ελεύθερα στο κοινό μία εφαρμογή ανοικτού κώδικα η οποία απλοποιεί και καθιστά οικονομικά εφικτή την αποθήκευση και επεξεργασία μεγάλης κλίμακας raster δεδομένων μέσω της μείωσης του χρόνου που χρειάζονται οι αναλυτές για να αναζητήσουν, να μεταφέρουν, να εκτελέσουν εργασίες προ-επεξεργασίας και να μορφοποιήσουν κατάλληλα τα δεδομένα για περαιτέρω ανάλυση.

Η εφαρμογή του προτύπου MapReduce για τα γεωχωρικά δεδομένα ή αλλιώς MrGeo, είναι ένα γεωχωρικό εργαλείο το οποίο έχει σχεδιαστεί ώστε να παρέχει δυνατότητες (αποθήκευση και επεξεργασία) για τη διαχείριση raster δεδομένων οι οποίες θα εκτελούνται σε μεγάλη κλίμακα κάνοντας χρήση της δύναμης και της λειτουργικότητας των αρχιτεκτονικών cloud computing. Τα δικαιώματα χρήσης, μετατροπής και διανομής του λογισμικού αυτού καθορίζονται πλήρως από την άδεια Apache 2.0. Το όραμα της NGA αναφορικά με το MrGeo είναι αυτό να γίνει το πρότυπο για την αποθήκευση και ανάλυση μεγάλων ποσοτήτων raster δεδομένων σε ένα κατανεμημένο περιβάλλον cloud.

Το MrGeo έχει τη δυνατότητα της εισαγωγής και αποθήκευσης μεγάλων συνόλων δεδομένων σε υποδομές cloud σε μία μορφή έτοιμη για χρήση, η οποία εξαλείφει πολλά βήματα

[2]https://www.digitalglobe.com

προ-επεξεργασίας δεδομένων από ροές εκτέλεσης εργασιών παραγωγής με συνέπεια να απελευθερώνει τους χρήστες από πολλά χρονοβόρα βήματα τα οποία απαιτούνταν προηγουμένως κατά την ανάκτηση και προ-επεξεργασία των δεδομένων. Αυτό επιτρέπει στους χρήστες να επικεντρώνονται στη διενέργεια αναλύσεων των δεδομένων στο cloud και να λαμβάνουν τις υπολογισμένες απαντήσεις μόνο για τις περιοχές ενδιαφέροντός τους αντί να πρέπει να προ-επεξεργαστούν όλα τα αποθηκευμένα δεδομένα για να αποκτήσουν το αποτέλεσμα.

Το MrGeo παρέχει μία γενική αλλά ταυτόχρονα εύρωστη μηχανή διενέργειας αναλύσεων τύπου MapReduce για την επεξεργασία γεωαναφερμένων raster δεδομένων όπως είναι τα ψηφιακά μοντέλα εδάφους και οι πολυφασματικές/υπερφασματικές δορυφορικές εικόνες και αεροφωτογραφίες. Επίσης παρέχει μία φιλική προς το χρήστη σύνταξη γραμμής εντολών η οποία καλείται διεπαφή Map Algebra και δίνει τη δυνατότητα για την ανάπτυξη προσαρμοσμένων αλγορίθμων με ένα απλό API γραφής σεναρίων το οποίο και επιτρέπει τη συγγραφή αλγεβρικών υπολογισμών, κεντρικών υπολογισμών (π.χ υπολογισμός κλίσης) και πράξεις γραφημάτων ώστε να οδηγήσει σε αλυσίδες βασικών πράξεων και να δημιουργήσει προϊόντα ανάλυσης υψηλού επιπέδου.

Το MrGeo [67] έχει υλοποιηθεί πάνω από το οικοσύστημα της πλατφόρμας Hadoop ώστε να εκμεταλλευθεί τις δυνατότητες επεξεργασίας και αποθήκευσης εκατοντάδων μονάδων κοινού υλικού. Ένα επίπεδο αφαίρεσης ανάμεσα στη διενέργεια αναλύσεων MapReduce και στις μεθόδους αποθήκευσης παρέχει ένα ευρύ σύνολο επιλογών αποθήκευσης στο cloud όπως είναι τα συστήματα HDFS, Accumulo, HBASE, κ.α. Λειτουργικά το MrGeo αποθηκεύει μεγάλα σύνολα raster δεδομένων σαν μία συλλογή ανεξάρτητων tiles τα οποία αποθηκεύονται στο σύστημα Hadoop ώστε να είναι δυνατή η διενέργεια μεγάλης κλίμακας υπηρεσιών δεδομένων και αναλύσεων. Το μοντέλο αποθήκευσης των δεδομένων το οποίο διατηρεί την τοπικότητα μέσω χωρικών ευρετηρίων μαζί με την συνύπαρξη δεδομένων και διαδικασιών ανάλυσης προσφέρει το πλεονέκτημα της ελαχιστοποίησης της μετακίνησης των δεδομένων υπέρ της μεταφοράς των υπολογισμών στα δεδομένα, μία τυπική αρχή κατά το σχεδιασμό συστημάτων επεξεργασίας μεγάλων δεδομένων. Η αρχιτεκτονική του MrGeo διευκολύνει την τμηματοποιημένη ανάπτυξη λογισμικού καθώς και τις διάφορες στρατηγικές εγκατάστασης νέων συστημάτων. Οι δυνατότητες που παρέχει για το χειρισμό δεδομένων και τη διενέργεια αναλύσεων είναι αυτές που προβλέπονται από τον οργανισμό OGC καθώς και από τα πρωτόκολλα τύπου REST.

Το MrGeo έχει χρησιμοποιηθεί για την αποθήκευση, το indexing, το tiling, καθώς και για τη δημιουργία πυραμίδων εικονιστικών συνόλων δεδομένων κλίμακας πολλών Terabytes. Από τη στιγμή που έχουν αποθηκευτεί, αυτά τα δεδομένα γίνονται διαθέσιμα μέσω απλών υπηρεσιών Tiled Map Services (TMS) και/ή Web Mapping Services (WMS). Αν και το MrGeo έχει αναπτυχθεί κυρίως για την εξυπηρέτηση raster συνόλων δεδομένων, νέα χαρακτηριστικά έχουν πρόσφατα προστεθεί τα οποία και επιτρέπουν την αποθήκευση και επεξεργασία και vector συνόλων δεδομένων.

2.5 CartoDB

Υπάρχει μια τρέχουσα ανάγκη για ευέλικτους και διαισθητικούς τρόπους δημιουργίας online δυναμικών χαρτών καθώς και για το σχεδιασμό διαδικτυακών γεωχωρικών εφαρμογών. Το CartoDB είναι ένα εργαλείο ανοικτού κώδικα το οποίο επιτρέπει την αποθήκευση και οπτικοποίηση γεωχωρικών δεδομένων στο διαδίκτυο και στοχεύει στο να εξελιχθεί στη χαρτογραφική πλατφόρμα νέας γενιάς για μεγάλα δεδομένα τα οποία ακολουθούν το vector μοντέλο δεδομένων.

Το CartoDB [23] είναι μια πλατφόρμα cloud computing της μορφής Software as a Service (SaaS), η οποία και προσφέρει εργαλεία GIS και εργαλεία διαδικτυακής χαρτογραφίας για την απεικόνιση vector δεδομένων σε ένα φυλλομετρητή ιστού. Το CartoDB αναπτύχθηκε με τη βοήθεια λογισμικού ανοικτού κώδικα συμπεριλαμβανομένης της βάσης δεδομένων PostGIS/ PostgreSQL. Χρησιμοποιεί εκτενώς τη γλώσσα προγραμματισμού Javascript για το εμπρόσθιο τμήμα της διαδικτυακής εφαρμογής, στο οπίσθιο τμήμα δια μέσου των APIs τα οποία βασίζονται στη βιβλιοθήκη Node.js όπως επίσης και για την υλοποίηση βιβλιοθηκών που έχουν ανάγκη οι πελάτες του συστήματος. Η πλατφόρμα CartoDB προσφέρει ένα σύνολο από APIs και βιβλιοθήκες οι οποίες βοηθούν τους χρήστες να δημιουργήσουν χάρτες, να διαχειρισθούν τα δεδομένα τους, να εκτελέσουν εργασίες γεωχωρικών αναλύσεων καθώς και επιπρόσθετες αναλύσεις δια μέσου υπηρεσιών τύπου REST ή βιβλιοθηκών αναπτυγμένων από τους χρήστες.

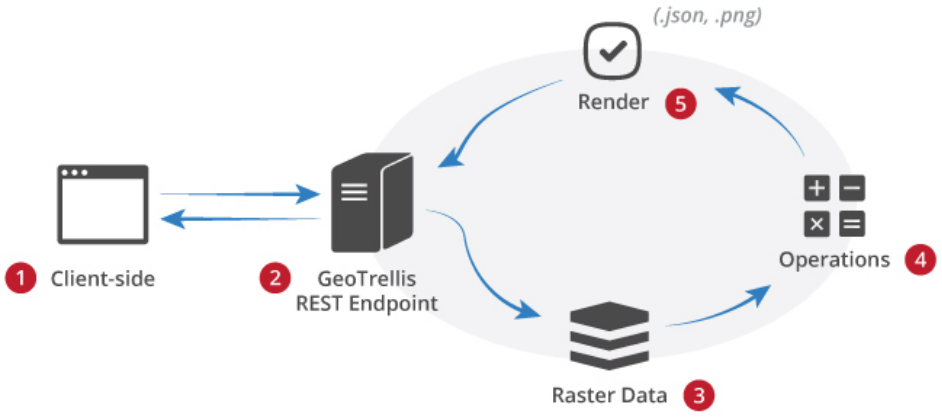
Οι χρήστες του συστήματος CartoDB μπορούν να χρησιμοποιήσουν την δωρεάν πλατφόρμα της εταιρείας ή να εγκαταστήσουν το δικό τους στιγμιότυπο του λογισμικού ανοικτού κώδικα. Με το CartoDB είναι εύκολη η διαχείριση γεωχωρικών δεδομένων τα οποία βρίσκονται σε πολλούς διαφορετικούς τύπους (π.χ. Shapefiles, GeoJSON, κ.α.) κάνοντας χρήση μιας διαδικτυακής διεπαφής. Γνωστοί χρήστες της πλατφόρμας CartoDB για την παραγωγή online δυναμικών χαρτών περιλαμβάνουν σημαντικούς οργανισμούς και ισχυρούς παίκτες του τεχνολογικού κόσμου όπως είναι η NASA, η Nokia [34] και το Twitter.

2.6 Geotrellis

Η azavea[3], μια αμερικάνικη εταιρεία ανάπτυξης λογισμικού και GIS εφαρμογών είναι υπεύθυνη για τη δημιουργία, το σχεδιασμό, την υλοποίηση, τη διατήρηση και διανομή του Geotrellis μιας μηχανής επεξεργασίας γεωχωρικών δεδομένων για εφαρμογές υψηλών επιδόσεων. Το Geotrellis είναι ένα project ανοικτού κώδικα το οποίο διανέμεται υπό την άδεια Apache 2 και αναπτύχθηκε για να υποστηρίξει την επεξεργασία γεωχωρικών δεδομένων τόσο σε διαδικτυακή κλίμακα όσο και σε περιβάλλοντα cluster υπολογιστών. Έχει υλοποιηθεί εξ΄ ολοκλήρου στη γλώσσα προγραμματισμού Scala, μια υψηλού επιπέδου γλώσσα προγραμματισμού γενικού σκοπού η οποία παρέχει πλήρη υποστήριξη για αντικειμενοστρεφή και συναρτησιακό προγραμματισμό καθώς και ένα πολύ ισχυρό στατικό σύστημα τύπων.

Το Geotrellis σχεδιάστηκε με σκοπό να επιλύσει 3 βασικά προβλήματα με μια αρχική εστίαση στην επεξεργασία raster δεδομένων. Τα προβλήματα αυτά είναι τα εξής:

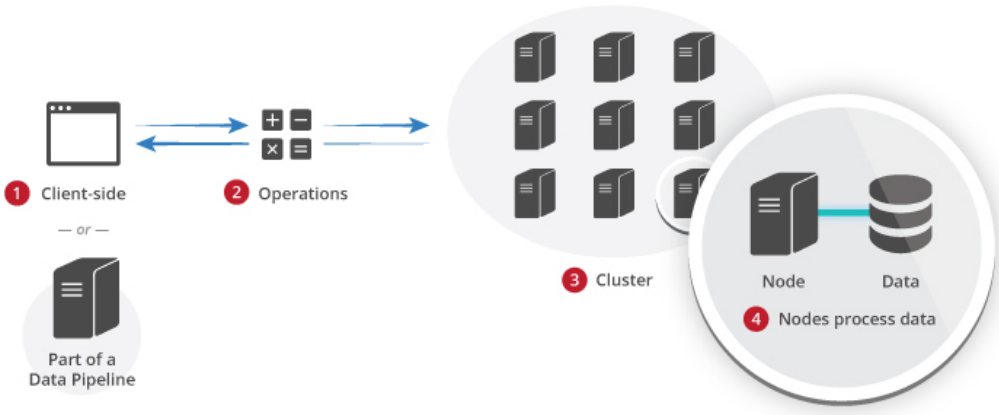
[3] https://www.azavea.com/



Σχήμα 2.2: Αρχιτεκτονική υπηρεσίας πραγματικού χρόνου βασισμένη στην πλατφόρμα Geotrellis (Πηγή: http://geotrellis.io/)

- Η δημιουργία χαμηλής καθυστέρησης, κλιμακώσιμων γεωχωρικών διαδικτυακών υπηρεσιών.
- Η δημιουργία batch υπηρεσιών επεξεργασίας μεγάλων γεωχωρικών δεδομένων οι οποίες μπορούν να λειτουργήσουν σε κατανεμημένες αρχιτεκτονικές.
- Η παραλληλοποίηση γεωχωρικών επεξεργασιών ώστε να είναι δυνατή η πλήρης αξιοποίηση των πολυ-πύρηνων αρχιτεκτονικών.

Ο βασικός πυρήνας του Geotrellis παρέχει τη δυνατότητα επεξεργασίας είτε μεγάλων είτε μικρών συνόλων δεδομένων με πολύ μικρή καθυστέρηση μέσω της κατανομής του υπολογιστικού φορτίου κατά μήκος πολλαπλών νημάτων, πυρήνων, επεξεργαστών και μηχανημάτων. Η azavea αφ΄ ότου αξιολόγησε πολλές γλώσσες προγραμματισμού και αρχιτεκτονικές προσεγγίσεις, επέλεξε τη Scala ως τη γλώσσα υλοποίησης του Geotrellis και το Spark ως το υπολογιστικό πλαίσιο εκτέλεσης. Το Spark είναι ένα ανοικτού κώδικα υπολογιστικό πλαίσιο εκτέλεσης σε περιβάλλον cluster υπολογιστών το οποίο παρέχει κατάλληλες διεπαφές για τον προγραμματισμό cluster υπολογιστών με υπονοούμενη και δεδομένη την παράλληλη πρόσβαση στα δεδομένα και την ανεκτικότητα σε σφάλματα κατά την εκτέλεση των προγραμμάτων. Το Geotrellis επεκτείνει το μοντέλο δεδομένων του Spark για την επίτευξη της ταχείας και κατανεμημένης επεξεργασίας τόσο raster όσο και vector δεδομένων. Παρέχει κατάλληλους τύπους δεδομένων για να είναι δυνατή η επεξεργασία γεωχωρικών δεδομένων στο πλαίσιο της γλώσσας Scala όπως επίσης και η ταχύτατη ανάγνωση και εγγραφή αυτών των τύπων δεδομένων στο δίσκο. Το σύστημα επιτρέπει επίσης την ενσωμάτωση μιας πληθώρας εξωτερικών εργαλείων για τις διάφορες διαδικασίες μετατροπής και εισαγωγής των raster δεδομένων στο Geotrellis για την επίτευξη βελτιστοποιημένων δομών δεδομένων. Το Geotrellis μπορεί να λειτουργήσει συμπληρωματικά ως προς διάφορα άλλα projects ανοικτού κώδικα, όπως είναι ο GeoServer, το OpenLayers και η PostGIS.



Σχήμα 2.3: Αρχιτεκτονική batch υπηρεσίας βασισμένη στην πλατφόρμα Geotrellis (Πηγή: http://geotrellis.io/)

Έχει ωριμάσει ως μια συνεχώς επεκτεινόμενη σουίτα διαφορετικών τμημάτων λογισμικού τα οποία βοηθούν τους χρήστες ώστε να αναπτύξουν γεωχωρικές εφαρμογές με έμφαση στην υψηλή επίδοση και στους κατανεμημένους υπολογισμούς. Σχεδιάστηκε ώστε να βοηθήσει τους χρήστες να αναπτύξουν τόσο διαδικτυακές υπηρεσίες πραγματικού χρόνου τύπου REST (Σχήμα 2.2) όσο και ταχύτατες μη διαδραστικές εφαρμογές (batch processing) (Σχήμα 2.3), για την επεξεργασία και την εξαγωγή πληροφορίας από μεγάλα σύνολα raster δεδομένων. Η αυτόματη παραλληλοποίηση και βελτιστοποίηση των γεωχωρικών υπολογισμών αποτελεί εγγενή λειτουργικότητα, όπου αυτό είναι δυνατόν.

Όπως τα περισσότερα ανταγωνιστικά συστήματα λογισμικού έτσι και το Geotrellis έχει τη δυνατότητα της εισαγωγής και αποθήκευσης μεγάλων συνόλων δεδομένων σε υποδομές cloud σε μία μορφή έτοιμη για χρήση, η οποία εξαλείφει πολλά βήματα προ-επεξεργασίας δεδομένων από ροές εκτέλεσης εργασιών παραγωγής με συνέπεια να απελευθερώνει τους χρήστες από πολλά χρονοβόρα βήματα τα οποία απαιτούνταν προηγουμένως κατά την ανάκτηση και προ-επεξεργασία των δεδομένων. Από τη στιγμή που έχουν αποθηκευτεί, αυτά τα δεδομένα γίνονται διαθέσιμα μέσω απλών υπηρεσιών Tiled Map Services (TMS) και/ή Web Mapping Services (WMS) οι οποίες έχουν τη δυνατότητα να προβάλλουν πολλαπλά layers πληροφορίας.

Το Geotrellis, όπως και το MrGeo για το οποίο μιλήσαμε σε προηγούμενη ενότητα, έχει υλοποιηθεί πάνω από το οικοσύστημα της πλατφόρμας Hadoop ώστε να εκμεταλλευθεί τις δυνατότητες επεξεργασίας και αποθήκευσης εκατοντάδων μονάδων κοινού υλικού, όπως άλλωστε "προστάζουν" οι σύγχρονες πρακτικές σχεδιασμού κλιμακώσιμων αρχιτεκτονικών συστημάτων. Ένα επίπεδο αφαίρεσης ανάμεσα στη διενέργεια αναλύσεων με τη βοήθεια του Apache Spark και στις μεθόδους αποθήκευσης παρέχει ένα ευρύ σύνολο επιλογών αποθήκευσης στο cloud όπως είναι τα συστήματα HDFS, Accumulo, HBASE, Apache Cassandra κ.α. Λειτουργικά το Geotrellis αποθηκεύει μεγάλα σύνολα raster δεδομένων σαν μία συλλογή ανεξάρτητων tiles τα οποία αποθηκεύονται στο κατανεμημένο σύστημα αρχείων HDFS της πλατφόρμας Hadoop ώστε να είναι δυνατή η διενέργεια μεγάλης κλίμακας υπηρεσιών δε-

δομένων και αναλύσεων. Το μοντέλο αποθήκευσης των δεδομένων το οποίο διατηρεί την τοπικότητα μέσω χωρικών ευρετηρίων μαζί με την συνύπαρξη δεδομένων και διαδικασιών ανάλυσης προσφέρει το πλεονέκτημα της ελαχιστοποίησης της μετακίνησης των δεδομένων υπέρ της μεταφοράς των υπολογισμών στα δεδομένα, μία τυπική αρχή κατά το σχεδιασμό συστημάτων επεξεργασίας μεγάλων δεδομένων. Η αρχιτεκτονική του Geotrellis διευκολύνει την τμηματοποιημένη ανάπτυξη λογισμικού καθώς και τις διάφορες στρατηγικές εγκατάστασης νέων συστημάτων.

Το Geotrellis είναι μία σχετικά νέα πλατφόρμα επεξεργασίας γεωχωρικών δεδομένων, της οποίας η δημοτικότητα ανεβαίνει συνεχώς εξαιτίας των πολλών δυνατοτήτων που προσφέρει στο πλαίσιο της κατανεμημένης ανάλυσης μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών, της μεγιστοποίησης της αξιοποίησης της διαθέσιμης υποδομής υλικού καθώς και της αύξησης της παραγωγικότητας των προγραμματιστών/χρηστών μέσω των ολοκληρωμένων εργαλείων για την ανάπτυξη γεωχωρικών υπηρεσιών που παρέχονται.

2.7 Ερευνητικά θέματα επεξεργασίας Μεγάλων Γεωχωρικών Δεδομένων

Στην ενότητα αυτή θα συζητηθούν τρέχουσες προκλήσεις στο πεδίο των μεγάλων γεωχωρικών δεδομένων καθώς και το ζήτημα της αποτελεσματικής διαχείρισής τους για τη διενέργεια αναλύσεων.

Όγκος δεδομένων (Data volume): Πως θα πραγματοποιηθεί η διαχείριση ενός ολοένα και αυξανόμενου όγκου γεωχωρικών δεδομένων ; Αυτή η πρόκληση είναι η πιο θεμελιώδης στο πεδίο των μεγάλων δεδομένων, καθώς αυτή είναι η προσδιοριστική ιδιότητα για αυτό το είδος των γεωχωρικών δεδομένων. Οι πιο κοινοί τρόποι για την αντιμετώπιση αυτού του ζητήματος είναι είτε μέσω της τμηματοποίησης των δεδομένων είτε μέσω της απομακρυσμένης επεξεργασίας των γεωχωρικών δεδομένων. Κατά την πρώτη περίπτωση τα γεωχωρικά δεδομένα χωρίζονται σε μικρότερα, κατανεμημένα σύνολα δεδομένων και υφίστανται επεξεργασία μέσα σε ένα παράλληλο περιβάλλον. Τα αποτελέσματα στη συνέχεια συλλέγονται και συνδυάζονται για τη διεξαγωγή της τελικής ανάλυσης. Κατά τη δεύτερη περίπτωση, όπου τα δεδομένα δεν μπορούν να διαχωριστούν σε μικρότερα σύνολα (δηλαδή σε περιπτώσεις όπου το πρόβλημα υπό μελέτη περιλαμβάνει παγκόσμια χαρακτηριστικά τα οποία πρέπει να υπολογιστούν και να γίνουν διαθέσιμα σε όλες τις υπολογιστικές διεργασίες, π.χ. κατά την ταξινόμηση βασισμένη στη γνώση, κατά την κατανόηση εικόνων δια μέσου πληροφοριών περιεχομένου κ.α.) η λύση είναι να μεταφερθούν οι υπολογιστικές διεργασίες στα δεδομένα μέσω της απομακρυσμένης εκτέλεσης υψηλού επιπέδου, βελτιστοποιημένων αλγορίθμων.

Ποικιλομορφία δεδομένων (Data variety): ;Οταν τα δεδομένα είναι αδόμητα, πόσο γρήγορα είναι δυνατόν να εξαχθούν χρήσιμες πληροφορίες περιεχομένου από αυτά ; Πως θα πραγματοποιηθεί ο συνδυασμός και η συσχέτιση δεδομένων συνεχούς ροής (streaming data) από πολλαπλές πηγές ; Εξαιτίας της μεγάλης ποικιλίας των πρωτογενών δεδομένων οι επαγγελματίες οι οποίοι ειδικεύονται στα μεγάλα δεδομένα συνήθως καταφεύγουν σε μεθόδους ανακάλυψης γνώσης (knwoledge discovery) οι οποίες αναφέρονται σε ένα σύνολο ενεργειών

οι οποίες έχουν σχεδιαστεί ώστε να εξάγουν νέα γνώση από σύνθετα σύνολα δεδομένων. Μέθοδοι όπως είναι η συγχώνευση multi-modal δεδομένων (multi-modal data fusion) για raster δεδομένα, βοηθούν τους επαγγελματίες στο χώρο της τηλεπισκόπησης να εξάγουν πληροφορία από εικονιστικά δεδομένα πολλών διαφορετικών αισθητήρων. Με παρόμοιο τρόπο, προχωρημένες μέθοδοι χωρικών ευρετηρίων (όπως για παράδειγμα είναι τα προσαρμοστικά σχήματα για το tiling χαρτών μέσα σε ν-διάστατους πίνακες δεδομένων) μπορούν να βοηθήσουν σημαντικά στην εκτέλεση σύνθετων ερωτημάτων και σε raster αλλά και σε vector σύνολα δεδομένων.

Αποθήκευση Δεδομένων (Data storage): Πως θα πραγματοποιηθεί η αποτελεσματική αναγνώριση και αποθήκευση σημαντικής πληροφορίας η οποία εξάγεται από αδόμητα δεδομένα ; Πως θα πραγματοποιηθεί η αποθήκευση μεγάλου όγκου πληροφορίας με ένα τρόπο ο οποίος θα επιτρέψει την έγκαιρη ανάκτησή της ; Είναι τα τρέχοντα συστήματα αρχείων βελτιστοποιημένα για την εξυπηρέτηση του όγκου και της ποικιλομορφίας των δεδομένων που απαιτείται από εφαρμογές ανάλυσης ; Αν όχι ποιες νέες δυνατότητες είναι απαραίτητες ; Πως θα πραγματοποιηθεί η αποθήκευση της πληροφορίας με ένα τρόπο ο οποίος θα επιτρέπει την εύκολη μετανάστευση δεδομένων μεταξύ μεγάλων κέντρων δεδομένων και παρόχων υπηρεσιών Cloud ; Από τη στιγμή που η αποθήκευση των δεδομένων γίνεται ολοένα και πιο φθηνή και πιο αποτελεσματική, τα συστήματα αποθήκευσης γίνονται ολοένα και πιο αποδοτικά. Όμως για να είναι δυνατή η επίλυση τέτοιων προβλημάτων αποθήκευσης μεγάλων δεδομένων, εύρωστα κατανεμημένα συστήματα αποθήκευσης είναι απαραίτητα με τις ικανότητες της αποδοτικής κατάτμησης των δεδομένων, της αντιγραφής πληροφορίας καθώς και της αναζήτησης και ανάκτησης δεδομένων με πολύ μεγάλες ταχύτητες. Τέτοια state-of-the-art συστήματα συνεχώς εξελίσσονται αλλά ειδικά για τα γεωχωρικά σύνολα δεδομένων (και ακόμα πιο ειδικά για τα raster σύνολα δεδομένων) τα συστήματα αυτά δεν είναι ακόμα βελτιστοποιημένα ώστε να διαχειρίζονται πολύ μεγάλα αρχεία σε ένα κατανεμημένο περιβάλλον. Πρόσφατα ανεπτυγμένα συστήματα τα οποία εμπλέκονται στην αποθήκευση εικονιστικών δεδομένων τα οποία προέρχονται από τους δορυφόρους Landsat 8 και MODIS δείχνουν ότι η κλιμακωσιμότητα της αποθήκευσης δεν βρίσκεται ακόμα σε ικανοποιητικά επίπεδα. Ιδανικά, τα συστήματα αποθήκευσης νέφους υπολογιστών όπως για παράδειγμα είναι το Rados και το GlusterFS θα γίνουν πολύ πιο αποδοτικά στο κοντινό μέλλον και θα εφοδιασθούν με χωρικές επεκτάσεις ώστε να κατανέμουν έξυπνα χωρική πληροφορία στις μονάδες αποθήκευσης.

Ενσωμάτωση Δεδομένων (Data integration): Νέα πρωτόκολλα και διεπαφές για την ενσωμάτωση των δεδομένων οι οποίες θα είναι ικανές να διαχειρίζονται δεδομένα διαφορετικής φύσης (δομημένα, αδόμητα, ημι-δομημένα) και διαφορετικής προέλευσης. Σε αυτό το πεδίο οι εργασίες οι οποίες διεξάγονται από τα ανοιχτά πρότυπα γεωχωρικών δεδομένων βρίσκονται στην καλύτερη δυνατή κατεύθυνση. Δια μέσου των ανοιχτών διαδικασιών προτυποποίησης, νέα αποτελεσματικά πρότυπα αναδύονται όπως είναι τα GeoJSON, Vector Tiles, Irregular Tiles σε ν-διάστατους πίνακες τα οποία βρίσκονται υπό την αιγίδα είτε του OGC είτε μιας άδεια ανοιχτού κώδικά στο σύστημα GitHub. Δεδομένου του γεγονότος ότι υπάρχει ενεργή δραστηριότητα στον τομέα των χωρικών ευρετηρίων ακόμα και για αδόμητα δεδομένα (όπως για παράδειγμα είναι το ευρετήριο SOLR ή οι χωρικές επεκτάσεις του συστήματος

CouchDB), η πρόκληση της ενσωμάτωσης των μεγάλων γεωχωρικών δεδομένων φαίνεται να βρίσκεται σε μια καλή τεχνολογική κατεύθυνση.

Επεξεργασία δεδομένων και διαχείριση πόρων (Data Processing and Resource Management): Απαιτούνται νέα προγραμματιστικά μοντέλα ειδικά βελτιστοποιημένα για δεδομένα συνεχούς ροής (streaming data) και/ή για πολυδιάστατα δεδομένα, νέες μηχανές συστημάτων οι οποίες θα διαχειρίζονται βελτιστοποιημένα συστήματα αρχείων καθώς και μηχανές οι οποίες θα έχουν τη δυνατότητα να συνδυάζουν εφαρμογές από πολλά προγραμματιστικά μοντέλα (π.χ. MapReduce, workflows και bag-of-tasks) σε μία ενιαία λύση/αφαίρεση. Πως θα πραγματοποιηθεί η βελτιστοποίηση της χρήσης πόρων σε τόσο σύνθετες αρχιτεκτονικές συστήματος ; Αυτό το πρόβλημα είναι πιθανόν το πιο δυσεπίλυτο στον κόσμο των μεγάλων δεδομένων. Η λύση έγκειται είτε στην πρόβλεψη της φύσης των δεδομένων είτε σε εύρωστους αλγορίθμους οι οποίοι μπορούν να εξάγουν αποδοτικά πληροφορία από γεωχωρικά δεδομένα, τρόποι οι οποίοι θα δώσουν τη δυνατότητα σε αρχιτέκτονες συστημάτων να βελτιστοποιήσουν πραγματικά τις ροές εκτέλεσης εργασιών για την ανάκτηση των δεδομένων, την προ-επεξεργασία τους, την κατανομή της αποθήκευσης και τελικά της χρησιμοποίησης των υπηρεσιών καθώς και της παροχής χρήσιμων αποτελεσμάτων τα οποία προέρχονται από χωρικές αναλύσεις.

Οπτικοποίηση και αλληλεπίδραση με τους χρήστες (Visualisation and user interaction): Υπάρχουν πολλές ερευνητικές προκλήσεις στο πεδίο της οπτικοποίησης των μεγάλων δεδομένων και ειδικά των γεωχωρικών δεδομένων εξαιτίας της διαστασιμότητάς τους. Αρχικά, αποτελεσματικότερες τεχνικές επεξεργασίας δεδομένων απαιτούνται ώστε να είναι δυνατόν να επιτευχθεί η οπτικοποίηση δεδομένων σε πραγματικό χρόνο. Οι Choo και Park [25] ορίζουν κάποιες τεχνικές οι οποίες μπορούν να εφαρμοσθούν για αυτό το σκοπό όπως είναι η ελάττωση της ακρίβειας των αποτελεσμάτων, η μείωση των απαιτήσεων σύγκλισης καθώς και ο περιορισμός της κλίμακας των δεδομένων. Οι μέθοδοι οι οποίες κάνουν χρήση αυτών των τεχνικών μπορούν να ερευνηθούν και να βελτιστοποιηθούν περαιτέρω. Η οπτικοποίηση για τους σκοπούς της διαχείρισης δικτύων υπολογιστών καθώς και αναλυτικών διαδικασιών λογισμικού [64] είναι επίσης μια περιοχή η οποία προσελκύει την προσοχή των ερευνητών καθώς παρουσιάζει πολύ μεγάλη συνάφεια με τη διαχείριση υποδομών μεγάλης κλίμακας (όπως είναι τα συστήματα νέφους υπολογιστών) και λογισμικού με επιπτώσεις στη συνεργατική ανάπτυξη λογισμικού, στην ανάπτυξη λογισμικού ανοιχτού κώδικα καθώς και στη βελτιστοποίηση της ποιότητας του λογισμικού.

Υπάρχουν ακόμα, ωστόσο, πολλές ανοικτές προκλήσεις όσον αναφορά τα παραπάνω ζητήματα [6], [28], [18]. Η παραπάνω λίστα δεν είναι εξαντλητική και καθώς διεξάγεται όλο και περισσότερη έρευνα σε αυτό το πεδίο, ολοένα και περισσότερο ενδιαφέροντα και απαιτητικά ζητήματα θα προκύπτουν.

Κεφάλαιο 3

Εφαρμογές Ανάλυσης Μεγάλων Γεωχωρικών Δεδομένων

Στο κεφάλαιο αυτό παρουσιάζεται η θεματική βάση των εφαρμογών ανάλυσης τηλεπισκοπικών δεδομένων οι οποίες υλοποιήθηκαν στα πλαίσια αυτής της εργασίας. Η βάση αυτή είναι η σύγχρονη τάση για την επεξεργασία και την εξαγωγή γνώσης από ολόκληρα αρχεία τηλεπισκοπικών δεδομένων μέσω της διενέργειας διαχρονικών αναλύσεων με τη βοήθεια κατανεμημένων συστημάτων επεξεργασίας υψηλών επιδόσεων. Επιχειρείται η προσέγγιση αυτής της τάσης μέσω της παρουσίασης διάφορων εφαρμογών για την παρατήρηση της γης και κυρίως για περιβαλλοντικές και αγροτικές εφαρμογές.

Οι εφαρμογές ανάλυσης γεωχωρικών δεδομένων βρίσκονται τα τελευταία χρόνια στο επίκεντρο της επιχειρηματικής και ερευνητικής δραστηριότητας. Αυτό συμβαίνει γιατί η συστηματική παρακολούθηση του φλοιού της γης μέσω δορυφορικών δεδομένων χαμηλής, μεσαίας αλλά και υψηλής χωρικής ανάλυσης είναι πλέον μια πραγματικότητα η οποία δεν περιορίζεται ούτε από τα υψηλά κόστη των δεδομένων αλλά ούτε και από τις περιορισμένες υπολογιστικές δυνατότητες των συστημάτων του παρελθόντος. Ολόκληρα αρχεία τηλεπισκοπικών δεδομένων είναι διαθέσιμα ελεύθερα στο ευρύ κοινό και το κόστος των υπολογιστικών πόρων που απαιτούνται για την επεξεργασία τους καθώς και των συστημάτων/μέσων αποθήκευσής τους συνεχίζει να μειώνεται καθιστώντας πλήρως εφικτή την κατανεμημένη επεξεργασία και ανάλυση δεδομένων σε παράλληλα περιβάλλοντα υψηλών επιδόσεων. Αυτό δίνει τη δυνατότητα για τη δημιουργία και την παροχή προϊόντων και υπηρεσιών με ρυθμό μεγαλύτερο από ποτέ άλλοτε.

Οι ερευνητικές προσπάθειες που ολοένα και εντείνονται προς την κατεύθυνση των αναλύσεων μεγάλων γεωχωρικών δεδομένων για την παροχή έξυπνων προϊόντων καθώς και οι αναλύσεις αγορών καταδεικνύουν και συμφωνούν στο ότι τα δεδομένα παρατήρησης της γης, τα προϊόντα των αναλύσεων τους καθώς και τα συστήματα τα οποία αναπτύσσονται συνεχώς με στόχο την αποδοτική αποθήκευση, διαχείριση, επεξεργασία και αναζήτηση τους θα παραμείνουν στο τεχνολογικό επίκεντρο για πολλά ακόμα χρόνια. Η κατάσταση αυτή διαμορφώνει ένα πλαίσιο στο οποίο οι καινοτόμες προσεγγίσεις των διάφορων επιστημονικών και τεχνολογικών προκλήσεων για την ανάλυση των μεγάλων γεωχωρικών δεδομένων είναι το κλειδί προς

την επίτευξη βιώσιμων και πρακτικών λύσεων οι οποίες μπορούν να ωφελήσουν την κοινωνία με πολλούς τρόπους και σε πολλά διαφορετικά επίπεδα. Στο κεφάλαιο αυτό μελετάμε τέτοιες προσεγγίσεις οι οποίες στοχεύουν στο να επεκτείνουν τις τυπικές μεθόδους ανάλυσης γεωχωρικών δεδομένων προς την επίτευξη ποιοτικών, βιώσιμων και χρήσιμων τηλεπισκοπικών προϊόντων.

3.1 Διαχρονική σύνθεση εικόνων ανά pixel

Η πληροφορία για τη συνεχώς μεταβαλλόμενη κατάσταση της γήινης επιφάνειας απαιτείται, πλέον, σε υψηλές χωρικές αναλύσεις καθώς πολλές εφαρμογές και ζητήματα δεν μπορούν να επιλυθούν μέσω της χρήσης δεδομένων χαμηλής χωρικής ανάλυσης. Η εξαγωγή τέτοιας πληροφορίας για πολύ μεγάλες εκτάσεις για τα δεδομένα τύπου Landsat όμως ακόμα ενέχει σημαντικές προκλήσεις.

Παραδείγματος χάριν, προϊόντα και πληροφορία [32] βασισμένα σε τεχνικές τηλεπισκόπησης σχετικά με την παρακολούθηση και τον προσδιορισμό αλλαγών για τις χρήσεις/καλύψεις γης απαιτούνται σε χωρικές αναλύσεις υψηλότερες από αυτές που είναι σήμερα διαθέσιμες από τα υπάρχοντα προϊόντα κάλυψης γης για όλη την υδρόγειο. Αυτό συμβαίνει γιατί πολλές διαδικασίες οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης, όπως είναι η υλοτομία, η αποψίλωση, η εγκατάλειψη εκτάσεων ή η εξάπλωση των πόλεων αποτελούν κρίσιμους οδηγούς για την παγκόσμια περιβαλλοντική αλλαγή αλλά συμβαίνουν σε χωρικές κλίμακες για τις οποίες τα δορυφορικά δεδομένα χαμηλής ανάλυσης δεν αρκούν για τον προσδιορισμό, την παρακολούθησή και την ανάλυσή τους σε πολλές περιοχές του πλανήτη. Ειδικότερα, το παραπάνω είναι αληθές για πολλές περιοχές της Αφρικής ή της νοτιο-ανατολικής Ασίας στις οποίες παρατηρούνται τα ίδια μοτίβα στις διαδικασίες οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης.

Παρά τον πολύ μεγάλο όγκο δεδομένων παρατήρησης της γης τα οποία είναι διαθέσιμα σε χωρικές αναλύσεις οι οποίες κυμαίνονται από τα 10 έως τα 50 μέτρα (οι οποίες θα αναφέρονται από εδώ και πέρα ως υψηλές χωρικές αναλύσεις), τυπικά προϊόντα κάλυψης γης καθώς και παρακολούθησης των διαδικασιών οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης δεν είναι ευρέως διαθέσιμα για πάρα πολύ μεγάλες εκτάσεις. Ο λόγος για το γεγονός αυτό είναι ότι η χαρτογράφηση και η παρακολούθηση της κάλυψης γης πολύ μεγάλων εκτάσεων σε υψηλές χωρικές αναλύσεις ακόμη ενέχει μοναδικές προκλήσεις [71]. Αυτές οι προκλήσεις, εν μέρει, συνδέονται με τη χωρική λεπτομέρεια αυτού του τύπου των δορυφορικών δεδομένων για τα οποία ενδιαφερόμαστε, καθώς η βελτιωμένη χωρική ανάλυση συνοδεύεται από το κόστος των μικρών λωρίδων σάρωσης. Συνεπώς, ένα μεγάλο πλήθος δορυφορικών αποτυπωμάτων (footprints) συχνά απαιτείται για να είναι δυνατή η πλήρης κάλυψη μεγάλων περιοχών. Επιπροσθέτως, φαινολογικά και ραδιομετρικά σταθερά σύνολα δεδομένων απαιτούνται [59] κατά την ανάλυση των διαδικασιών οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης καθώς και περίπλοκες αλλαγές στην κάλυψη της βλάστησης. Αυτό το χαρακτηριστικό δεν είναι εύκολο να επιτευχθεί καθώς οι μικρές συχνότητες επανεπισκεψιμότητας των δορυφόρων (αν και η κατάσταση έχει βελτιωθεί κάπως και με τα δεδομένα του δορυφόρου Sentinel) οδηγούν

στην διαθεσιμότητα λίγων σκηνών χωρίς σύννεφα ανά καλλιεργητική περίοδο σε πολλές περιοχές [48]. Η διαθεσιμότητα των δεδομένων επιδεινώνεται επιπλέον και από ασυνέχειες στα αρχεία των εικονιστικών δεδομένων όπως επίσης και από σφάλματα σχετικά με τα δεδομένα ή τους αισθητήρες λήψης τους (π.χ. η αστοχία της γραμμής σάρωσης του δορυφόρου Landsat 7 μετά το Μάιο του 2003). Συνεπώς, εννοιολογικά πιο εξελιγμένες προσεγγίσεις πρέπει να αναπτυχθούν ώστε να καταστεί δυνατή η εύρωστη χαρτογράφηση και παρακολούθηση πολύ μεγάλων εκτάσεων σε υψηλές χωρικές αναλύσεις.

Σε πολλές μελέτες όπου η κάλυψη γης έχει χαρτογραφηθεί σε υψηλές χωρικές αναλύσεις σε πολύ μεγάλες εκτάσεις, μη επιβλεπόμενες ή/και επιβλεπόμενες μέθοδοι χρησιμοποιήθηκαν και οι οποίες απαιτούσαν σημαντική παρέμβαση από τους χρήστες με συνέπεια να περιορίζονται σε πολύ μεγάλο βαθμό οι δυνατότητες για αυτοματοποίηση των διαδικασιών [20]. Προϊόντα κάλυψης γης στην κλίμακα της χωρικής ανάλυσης του δορυφόρου Landsat είναι διαθέσιμα και για την Ευρώπη (CORINE Land Cover). Ωστόσο, τα δεδομένα του CORINE τα οποία προέρχονται μέσω της ερμηνείας και της ψηφιοποίησης των δεδομένων του Landsat έχουν πάρα πολύ μεγάλα κόστη παραγωγής και περιορίζονται σε μια ελάχιστη περιοχή χαρτογράφησης των 25 εκταρίων [46]. Ένας σημαντικός αριθμός από επιχειρησιακές προσεγγίσεις για την παρακολούθηση της κάλυψης γης κάνοντας χρήση δεδομένων Landsat, σε κλίμακα από μικρές περιοχές έως ηπειρωτική έχουν ήδη αναπτυχθεί στην Αυστραλία πριν την απελευθέρωση του αρχείου δεδομένων του δορυφόρου Landsat από τη USGS [70, 76]. Κάποιες μελέτες έχουν εξερευνήσει την δυνατότητα της μεταφοράς μοντέλων ταξινόμησης κατά μήκος του χώρου ή τη δημιουργία δεδομένων εκπαίδευσης για μια μη ταξινομημένη εικόνα από την επικάλυψη με μια υπάρχουσα ταξινόμηση [47, 42]. Προσφάτως διαφορετικές επιστημονικές ομάδες ερεύνησαν τη δυνατότητα της εκμετάλλευσης ετήσιων χρονο-σειρών από δεδομένα Landsat προς την βελτίωση της κατανόησης των διαδικασιών οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης σε δασικά οικοσυστήματα [21]. Ωστόσο, οι υπάρχουσες προσεγγίσεις για τη χαρτογράφηση πολύ μεγάλων εκτάσεων με δεδομένα τύπου Landsat έχουν πολύ περιορισμένες δυνατότητες για πλήρη αυτοματοποίηση, λειτουργούν μόνο για προβλήματα διαχωρισμού πολύ απλών κλάσεων ή είναι θεματικά περιορισμένες σε συγκεκριμένα περιβάλλοντα. Συνεπώς, επιπρόσθετες βελτιστοποιήσεις απαιτούνται στις μεθόδους για την αποτύπωση πάρα πολύ μεγάλων εκτάσεων [26].

Η σύνθεση εικονιστικών προϊόντων ανά pixel προσφέρει μεγάλες δυνατότητες [79] προς την παράκαμψη και αντιμετώπιση των προαναφερθέντων προκλήσεων. Αυτό συμβαίνει γιατί η προσέγγιση αυτή διευκολύνει τη δημιουργία ραδιομετρικά σταθερών, εποχιακών και ελεύθερων από σύννεφα συνόλων προϊόντων τα οποία προέρχονται από δορυφορικά δεδομένα τύπου Landsat. Επιπροσθέτως, προσφέρει δυνατότητες προς την υπερκέραση των περιορισμών στη διαθεσιμότητα των δεδομένων και στη βελτίωση της χαρτογράφησης και της παρακολούθησης πολύ μεγάλων περιοχών ταυτόχρονα. Η σύνθεση εικόνων, αρχικά, αναπτύχθηκε για αισθητήρες με ευρεία λωρίδα σάρωσης οι οποίοι παρέχουν παγκόσμια κάλυψη με πολύ μεγάλη συχνότητα. Νέα σύνολα εικονιστικών δεδομένων δημιουργούνται μέσω της επιλογής μια συγκεκριμένης παρατήρησης από πολυάριθμες καταγραφές ή από τον υπολογισμό μέσων όρων από τις φασματικές τιμές. Για αισθητήρες χαμηλής χωρικής ανάλυσης, όπως είναι ο αισθη-

τήρας AVHRR, ο κύριος στόχος ήταν η μείωση της επιρροής της νεφοκάλυψης στο σήμα. Τις περισσότερες φορές, απλοί κανόνες απόφασης εφαρμόζονται, όπως είναι η επιλογή της μέγιστης/ελάχιστης τιμής κάποιου καναλιού ή του δείκτη NDVI [9, 26]. Πιο προχωρημένες προσεγγίσεις σύνθεσης λαμβάνουν, επιπρόσθετα, υπ΄ όψη τους και τη γωνία λήψης του κάθε pixel. Εν τούτοις, η σύνθεση εικόνων δεν λογιζόταν στις περιπτώσεις όπου τα δεδομένα ήταν υψηλής ανάλυσης, κυρίως λόγω του υψηλού κόστους των δεδομένων και των περιορισμών που εισήγαγαν οι περιορισμένες δυνατότητες των υπολογιστικών συστημάτων. Οι πρόσφατες εξελίξεις όμως ενθαρρύνουν τη σύνθεση εικόνων για δορυφορικά δεδομένα τύπου Landsat. Οι εξελίξεις αυτές περιλαμβάνουν αλλαγή στην πολιτική διάθεσης των δεδομένων (δηλαδή ελεύθερη, πλέον, διάθεση), βελτιωμένοι αλγόριθμοι προ-επεξεργασίας οι οποίοι οδηγούν σε βελτίωση της τυπικής ποιότητας των εικονιστικών δεδομένων καθώς και εξελίξεις στις τεχνολογίες αποθήκευσης και στους υπολογιστικούς πόρους [2].

Η σύνθεση δορυφορικών εικόνων υψηλής ανάλυσης ανά pixel (pixel-based compositing) σε αντίθεση με τη σύνθεση ή τη δημιουργία μωσαϊκών ανά σκηνή, προσφέρει σημαντικά πλεονεκτήματα για την ανάλυση των διαδικασιών οι οποίες οδηγούν σε αλλαγές στις χρήσεις/καλύψεις γης για πολύ μεγάλες εκτάσεις. Παγκόσμια μοντέλα ανάλυσης μπορούν να βασιστούν πάνω σε ένα ενιαίο, ομογενές και ελεύθερο από σύννεφα σύνολο δεδομένων, το οποίο ιδεατά παρέχει σταθερή ραδιομετρική απόκριση κατά μήκος πολύ μεγάλων εκτάσεων. Καθώς όλες οι παρατηρήσεις οι οποίες είναι ελεύθερες από σύννεφα εξάγονται κατά τη διάρκεια της ανάλυσης ανά pixel, διαφορετικές εικονιστικές μετρικές μπορούν να παραχθούν ως πολύτιμα υποπροϊόντα της διαδικασίας σύνθεσης. Τέτοιες μετρικές μπορούν για παράδειγμα να παραχθούν ώστε να συλλέξουν σχετικές φαινολογικές καταστάσεις κατά τον εποχιακό κύκλο των καλλιεργειών ή μπορούν να αντιστοιχούν σε περιγραφικά στατιστικά τα οποία παρέχουν ένα μέτρο για την απόκριση μέσου όρου ή την φασματική μεταβλητότητα [39]. Πάνω από όλα, η αλλαγή στην προσέγγιση από μια ανάλυση βασισμένη στην κάθε σκηνή ξεχωριστά σε μια ανάλυση βασισμένη σε κάθε pixel κάθε σκηνής αποφέρει πολλές και σημαντικές βελτιώσεις:

- Οι αναλύσεις δεν περιορίζονται, πλέον, σε λίγες εικόνες στις οποίες η χαμηλή νεφοκάλυψη ήταν πολλές φορές το κριτήριο, πάνω από εποχιακά πιο ταιριαστές καταγραφές.
- Ολόκληρα αρχεία τηλεπισκοπικών δεδομένων μπορούν να εκμεταλλευθούν αποδοτικά όπως επίσης και μερικώς χρήσιμες εικόνες μπορούν εύκολα να συμπεριληφθούν.
- Η συχνότητα των παρατηρήσεων αυξάνεται καθώς τα pixel τα οποία είναι ανεπηρέαστα από τα σύννεφα μπορούν να ενσωματωθούν στις αναλύσεις ακόμα και από σκηνές οι οποίες έχουν πολύ μεγάλη νεφοκάλυψη και με τυπικές μεθόδους θα απορρίπτονταν.
- Η συχνότητα των παρατηρήσεων αυξάνεται επίσης και από την εκμετάλλευση των επικαλύψεων κατά μήκος της γραμμής πτήσης του δορυφόρου.

Λαμβάνοντας όλα τα παραπάνω υπ΄ όψη, η διαχρονική σύνθεση εικόνων ανά pixel αναδύεται ως ένα πολύτιμο εργαλείο για εφαρμογές οι οποίες στοχεύουν στο να καλύψουν πάρα πολύ μεγάλες εκτάσεις κάνοντας χρήση οπτικών δεδομένων υψηλής χωρικής ανάλυσης. Αρκετές εφαρμογές έχουν αναδυθεί για τη σύνθεση διαχρονικών εικονιστικών προϊόντων ανά

pixel, όπως για παράδειγμα είναι μια σύγχρονη εφαρμογή βασισμένη σε δεδομένα Landsat για μεγάλης κλίμακας αποτύπωση της κάλυψης γης [36].

3.2 Τηλεπισκοπικές Εφαρμογές σε περιβάλλον cluster υπολογιστών

Όπως συζητήθηκε και στην προηγούμενη ενότητα η σημερινή ευρεία διαθεσιμότητα υπολογιστικών πόρων σε χαμηλό, σχετικά, κόστος και η ταχεία ανάπτυξη καινοτόμων και αξιόπιστων πλαισίων διαχείρισής τους έχει δώσει την απαιτούμενη ώθηση για την ανάπτυξη εφαρμογών γεωχωρικών δεδομένων υψηλών επιδόσεων, την ανάπτυξη μεθόδων και προσεγγίσεων ανάλυσης για τη δημιουργία νέων προϊόντων οι οποίες δεν ήταν δυνατόν να υλοποιηθούν παλαιότερα καθώς και για την ανάλυση ολόκληρων αρχείων τηλεπισκοπικών δεδομένων για την παροχή αξιόπιστων και βιώσιμων λύσεων. Η μεγάλη αλλαγή έγκειται στο ότι πλέον είναι δυνατόν μέσω κατάλληλου ελεύθερου λογισμικού να πραγματοποιηθούν αναλύσεις (analytics) των δεδομένων σε κατανεμημένα περιβάλλοντα υψηλών επιδόσεων μέσω της εκμετάλλευσης κοινού υλικού (hardware).

Υπάρχει η σύγχρονη τάση για τη μεταφορά των επεξεργαστικών pipelines από τη συμβατική επεξεργασία σε ένα μόνο μηχάνημα σε κατανεμημένα περιβάλλοντα cloud computing, για παράδειγμα τα περιβάλλοντα clusters υπολογιστών, ώστε να είναι δυνατή η παράλληλη και κατανεμημένη επεξεργασία μεγάλων γεωχωρικών δεδομένων για εφαρμογές υψηλών επιδόσεων.

Αυτό συμβαίνει καθώς η ολοένα και αυξανόμενη διαθεσιμότητα τηλεπισκοπικών δεδομένων πολύ υψηλής ανάλυσης τις τελευταίες δεκαετίες έχει δώσει τη δυνατότητα για την ανάπτυξη νέων εφαρμογών αλλά και έχει αποκαλύψει και νέες προκλήσεις οι οποίες σχετίζονται με το υπολογιστικό κόστος της επεξεργασίας πολύ μεγάλου όγκου δεδομένων. Στην πραγματικότητα, η σημαντική αύξηση στις χωρικές, φασματικές και χρονικές αναλύσεις των τηλεπισκοπικών δεδομένων τα τελευταία χρόνια έχει σταδιακά εκθέσει τους περιορισμούς πολλών συμβατικών αλλά και παράλληλων λύσεων οι οποίες εφαρμόζονται μέχρι σήμερα κατά την ανάλυση εικόνων. Συνεπώς, υπάρχει μια νέα απαίτηση για κλιμακώσιμες λύσεις διαχείρισης τηλεπισκοπικών δεδομένων.

Την τελευταία δεκαετία [77, 7], οι αναλύσεις μεγάλων δεδομένων (big data analytics) έχουν εξελιχθεί σε σημαντικό παράγοντα για τη λήψη αποφάσεων στις βιομηχανίες και ένας σημαντικός αριθμός νέων τεχνολογιών και συστημάτων έχει αναπτυχθεί ώστε να επιτευχθεί τόσο η επεξεργασία πολύ μεγάλου όγκου δεδομένων όσο και η δυνατότητα για την υποστήριξη πολλών ετερογενών τύπων δεδομένων. Το πιο αξιόπιστο, επιτυχημένο και δημοφιλές ανάμεσα σε αυτά τα συστήματα είναι το Apache Hadoop, ένα ανοικτού κώδικα πλαίσιο λογισμικού για την αποθήκευση και την επεξεργασία σε μεγάλη κλίμακα μεγάλων συνόλων δεδομένων σε περιβάλλοντα τα οποία αποτελούνται από clusters κοινού υλικού υπολογιστών τα οποία μπορούν με αξιόπιστο τρόπο να κλιμακώσουν σε χιλιάδες υπολογιστικούς κόμβους καθώς και σε petabytes σε όγκο δεδομένων. Παράλληλα με τον βασικό πυρήνα του συστήματος Hadoop έχει αναπτυχθεί ένα οικοσύστημα τεχνολογιών το οποίο περιλαμβάνει μεταξύ άλλων: τεχνο-

λογίες workflows (Apache Oozie), βάσεις δεδομένων (HBase), μηχανές μηχανικής μάθησης (Apache Mahout), εργαλεία ανάλυσης logs (Apache Flume) καθώς και πλαίσια παραγωγικότητας.

Το Hadoop χρησιμοποιεί το προγραμματιστικό μοντέλο MapReduce ως τον υπολογιστικό πυρήνα της εκτέλεσης προγραμμάτων. Το μοντέλο αυτό επιτρέπει τον καθορισμό των υπολογισμών με ένα τρόπο ο οποίος, ναι μεν, είναι εύκολο να κατανεμηθεί κατά μήκος ενός cluster υπολογιστών ο οποίος φιλοξενεί μεγάλα δεδομένα αλλά βασίζεται στις χαμηλού επιπέδου έννοιες των συναρτήσεων map και reduce. Αυτό καθιστά δύσκολη την εφαρμογή περίπλοκων μετασχηματισμών οι οποίοι απαιτούν πολλαπλά βήματα Map και Reduce τα αποτελέσματα των οποίων αλληλοεξαρτώνται. Ένας σημαντικός αριθμός από πλαίσια λογισμικού έχει αναπτυχθεί ώστε να επιτρέψει έναν ευκολότερο καθορισμό τέτοιων περίπλοκων pipelines επεξεργασίας δεδομένων καθώς επίσης και για την αύξηση της παραγωγικότητας κατά την ανάπτυξη εφαρμογών MapReduce. Τα frameworks αυτά απομονώνουν το έργο των προγραμματιστών από τις χαμηλού επιπέδου εργασίες MapReduce μέσω της προσφοράς διάφορων αφαιρέσεων για τους υπολογισμούς. Τα πιο δημοφιλή από αυτά τα εργαλεία είναι το Apache Pig, το Cascading καθώς και το Crunch.

Κάποια από αυτά τα εργαλεία έχουν υλοποιηθεί σε γλώσσες προγραμματισμού οι οποίες υποστηρίζουν το συναρτησιακό προγραμματιστικό μοντέλο και κατά συνέπεια στοχεύουν στην παροχή μιας ακόμα πιο αφηρημένης (και υψηλού επιπέδου) συναρτησιακής προσέγγισης για την έκφραση των υπολογισμών στο framework Hadoop. Αυτά τα frameworks περιλαμβάνουν τα εργαλεία Scoobi, Scrunch και Scalding τα οποία βασίζονται στη γλώσσα προγραμματισμού Scala καθώς και το εργαλείο Cascalog το οποίο βασίζεται στη γλώσσα προγραμματισμού Clojure.

Τα frameworks τα οποία βασίζονται στη γλώσσα προγραμματισμού Scala είναι τα πιο πολυπληθή εξαιτίας του σχετικά υψηλού βαθμού υιοθέτησης της γλώσσας Scala από εφαρμογές analytics σε σύνολα δεδομένων big data καθώς και της μεγάλης ευελιξίας της γλώσσας αυτής. Αν και η εσωτερική διαδικασία μετατροπής αφηρημένων pipelines σε πραγματικές ακολουθίες εργασιών MapReduce διαφέρει από framework σε framework τα APIs τα οποία εκθέτουν στους χρήστες τους είναι αρκετά κοντά μεταξύ τους. Αν και το Hadoop χρησιμοποιείται κυρίως για την ανάλυση και επεξεργασία δεδομένων σε μορφή κειμένου (textual data), υπάρχουν αρκετά παραδείγματα χρήσης του Hadoop για την επεξεργασία δυαδικών δεδομένων συμπεριλαμβανομένων και εικόνων, σε επιστημονικά πεδία όπως αυτά της αστρονομίας, της βιολογίας και άλλων. Σε πολλές από αυτές τις εφαρμογές, ωστόσο, το Hadoop χρησιμοποιείται για την εκτέλεση εξαιρετικά απλών παράλληλων εργασιών (οι οποίες περιλαμβάνουν μόνο το κομμάτι Map), οι οποίες εφαρμόζουν τον ίδιο ανεξάρτητο μετασχηματισμό σε ένα σύνολο από εικόνες εισόδου.

Τα τελευταία χρόνια έχουν γίνει προσπάθειες στον τομέα της επεξεργασίας τηλεπισκοπικών δεδομένων ώστε να μεταφερθούν pipelines επεξεργασίας και υπολογισμών σε κατανεμημένες αρχιτεκτονικές τύπου clusters υπολογιστών μέσω της πλατφόρμας Hadoop. Εξαιτίας της φύσης και της πολυπλοκότητας αυτού του τύπου δεδομένων, η ανάλυσή τους έχει προκαλέσει νέα ζητήματα και προκλήσεις και για το λόγο αυτό γίνονται σημαντικές προσπάθειες

από κάθε κατεύθυνση (επιστημονικές ομάδες και εταιρείες) για την ανάπτυξη εργαλείων τα οποία χτίζοντας επάνω στην πλατφόρμα Hadoop δίνουν / θα δώσουν όλα εκείνα τα εφόδια που απαιτούνται για την ανάλυση μεγάλων τηλεπισκοπικών δεδομένων.

Ήδη κάποιες από αυτές τις προσπάθειες για την ανάλυση μεγάλων γεωχωρικών δεδομένων σε περιβάλλοντα clusters υπολογιστών έχουν αρχίσει να δείχνουν τα πρώτα αποτελέσματα των προσπαθειών τους. Μερικά παραδείγματα τα οποία δείχνουν ότι μια τέτοια προσέγγιση λόγω του όγκου των σημερινών δεδομένων εκτός από αναγκαία είναι και εφικτή είναι τα ακόλουθα. Τεχνικές μηχανικής μάθησης καθώς και προσεγγίσεις εξόρυξης δεδομένων έχουν δοκιμαστεί σε περιβάλλοντα clusters υπολογιστών για τηλεπισκοπικά δεδομένα. Η δημιουργία μεγάλων μωσαϊκών [77] είναι ένας τομέας στον οποίο η κατανεμημένη ανάλυση έχει πολλά πλεονεκτήματα όπως επίσης και η κατάτμηση εικόνας (image segmentation) [41].

Ιδιαίτερη μνεία χρειάζεται για την ταξινόμηση, μια από τις κύριες εφαρμογές της τηλεπισκόπησης. Η επίδοση και η κλιμακωσιμότητα των ταξινομήσεων μπορεί να ωφεληθεί σημαντικά αν υλοποιηθεί σε κατανεμημένα περιβάλλοντα clusters υπολογιστών [8, 7]. Αυτό συμβαίνει γιατί η διαδικασία μιας επιβλεπόμενης ταξινόμησης έχει δύο κύρια στάδια. Το πρώτο είναι το στάδιο της εκπαίδευσης στο οποίο χτίζεται το μοντέλο ταξινόμησης. Το δεύτερο είναι το ίδιο το στάδιο της ταξινόμησης, το οποίο εφαρμόζει το εκπαιδευμένο μοντέλο για να αναθέσει τα " άγνωστα " δεδομένα σε μία από το δεδομένο σύνολο κλάσεων. Το στάδιο της εκπαίδευσης συσσωρεύει την περισσότερη ερευνητική δραστηριότητα αλλά συνήθως βασίζεται σε ένα μικρό αντιπροσωπευτικό δείγμα δεδομένων οπότε δεν συνιστά ενδιαφέρον ζήτημα από τη σκοπιά των εφαρμογών μεγάλων δεδομένων. Από την άλλη πλευρά η παραλληλοποίηση του βήματος της ταξινόμησης είναι μείζονος σημασίας για τη βελτίωση της επίδοσης αυτής της εργασίας όταν εφαρμόζεται σε πάρα πολύ μεγάλες εκτάσεις και η προσέγγιση της εκτέλεσής της σε ένα περιβάλλον cluster υπολογιστών δείχνει να είναι στη σωστή κατεύθυνση [8].

Η παραπάνω συνοπτική ανάλυση καταδεικνύει πόσο σημαντική και αναγκαία είναι η προσπάθεια για την ανάλυση μεγάλων γεωχωρικών δεδομένων σε περιβάλλοντα clusters υπολογιστών κάνοντας χρήση σύγχρονων εργαλείων ανάλυσης. Η προσέγγιση αυτή είναι ο μόνος, ίσως, τρόπος ώστε να αξιοποιηθούν αποτελεσματικά τα ολοένα και αυξανόμενα σύνολα τηλεπισκοπικών δεδομένων προς την παραγωγή καινοτόμων γεωχωρικών προϊόντων τα οποία θα δώσουν νέες δυνατότητες για την ανάπτυξη μεθόδων για τη διαχείριση του περιβάλλοντος και την παρακολούθηση των αλλαγών στη επιφάνεια της γης.

Κεφάλαιο 4

Σχεδιασμός και Υλοποίηση

Στο κεφάλαιο αυτό παρουσιάζεται η μελέτη που έγινε για την υλοποίηση ενός συστήματος για την ανάλυση μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών για εφαρμογές παρατήρησης της γης στο πλαίσιο εκπόνησης αυτής της διπλωματικής εργασίας. Αρχικά αφιερώνεται μία ενότητα ώστε να περιγραφούν οι επιστημονικές προκλήσεις οι οποίες αποτέλεσαν κίνητρο για την ανάπτυξη του συστήματος μας. Στη συνέχεια περιγράφεται η αρχιτεκτονική του συστήματος και γίνεται ο διαχωρισμός του στα επιμέρους υποσυστήματα από τα οποία αυτό συνίσταται. Τέλος, περιγράφονται οι βασικές εφαρμογές του συστήματος στην τρέχουσα μορφή του.

4.1 Επιστημονικές Προκλήσεις

Προς την κατεύθυνση της εκμετάλλευσης του ολοένα και αυξανόμενου όγκου των γεωχωρικών δεδομένων (geospatial big data), ο οποίος είναι της τάξεως πολλών petabyte, υπάρχει μια τρέχουσα ανάγκη για εντατική έρευνα και ανάπτυξη καθώς και για αποτελεσματικές τεχνολογικές λύσεις. Σε συνδυασμό με τον ολοένα και αυξανόμενο αριθμό καθώς και την αξιοπιστία των δορυφορικών, εναέριων και UAV αισθητήρων καθώς και των αισθητήρων εγγύτητας παρατήρησης της γης, η ανάγκη για υψηλής επίδοσης συστήματα άμεσης επεξεργασίας και ανάλυσης μεγάλων γεωχωρικών δεδομένων, τα οποία θα είναι σε θέση να μοντελοποιήσουν και να προσομοιάσουν γεωχωρικό περιεχόμενο, είναι μεγαλύτερη από ποτέ.

Η απελευθέρωση του αρχείου δεδομένων των αποστολών Landsat [80] του οργανισμού United States Geological Survey (USGS), η πρόσφατη έναρξη της αποστολής Landsat Data Continuity Mission [75] με το δορυφόρο Landsat 8, η ήδη προγραμματισμένη συνέχιση της αποστολής με το δορυφόρο Landsat 9 (2023), το πρόγραμμα EU Sentinel [63] με τις αποστολές Sentinel 1/2/3/4/5/6 καθώς και η πολιτική σε σχέση με τα ανοιχτά δεδομένα της Ευρωπαϊκής Ένωσης [1] έχουν δώσει τη δυνατότητα για την εύκολη πρόσβαση σε ένα πρωτοφανή όγκο δεδομένων και σχετικών μελετών πάνω στην παρακολούθηση αλλαγών στην κάλυψη και χρήση της γης. Βοηθούν επίσης στην ενημέρωση των εθνικών χαρτών κάλυψης γης, στην ανίχνευση χωρο−χρονικών δυναμικών καθώς επίσης και στην εξέλιξη των μεθόδων ανίχνευσης αλλαγών στην επιφάνεια της γης.

Εξαιτίας αυτής της τεράστιας διαθεσιμότητας δεδομένων, σε συνδυασμό με τις πολιτικές ανοιχτών δεδομένων για την πρόσβαση και χρήση αυτών των δεδομένων και στις Η.Π.Α αλλά και στην Ευρωπαϊκή Ένωση, οι προσπάθειες έρευνας και ανάπτυξης θα πρέπει να ανταναικλούν τις τρέχουσες απαιτήσεις για τη βελτίωση των υπαρχουσών δυνατοτήτων για την άμεση επεξεργασία big data, καθώς και να διαμορφώσουν το πλαίσιο για την αποτελεσματική χωροχρονική μοντελοποίησή τους. Συνεπώς, η ανάπτυξη αποτελεσματικών τεχνολογιών για τον άμεσο χειρισμό και την επεξεργασία big data στην πλευρά των εξυπηρετητών δεδομένων (server-side) δια μέσου ετερογενών συστημάτων αποθήκευσης, μηχανημάτων, μοντέλων επεξεργασίας και αρχιτεκτονικών εν γένει είναι θεμελιώδους σημασίας. Ειδικότερα, οι σύγχρονες τεχνολογικές εξελίξεις στα πεδία της επεξεργαστικής ισχύος, των υπολογιστικών πόρων και της ταχύτητας του διαδικτύου (Internet) έχουν ανοίξει το δρόμο για την online ανάλυση και επεξεργασία μεγάλων γεωχωρικών δεδομένων. Η πρόοδος αυτή παρέχει τη βάση πάνω στην οποία είναι δυνατή η ανάπτυξη εφαρμογών με σημαντικό επιστημονικό και βιομηχανικό ενδιαφέρον [49] όπως είναι η εξόρυξη γνώσης και η διενέργεια αναλύσεων (data analytics) μέσω της διαχρονικής ανάλυσης ολόκληρων αρχείων τηλεπισκοπικών δεδομένων.

Το πεδίο της συστηματικής υπολογιστικής ανάλυσης δεδομένων (data analytics) συμπεριλαμβάνει τεχνικές, αλγορίθμους και εργαλεία για την επεξεργασία συλλογών δεδομένων με σκοπό την εξαγωγή μοτίβων, γενικεύσεων και άλλων χρήσιμων πληροφοριών. Η ανάλυση μεγάλων δεδομένων (big data analytics) έχει καταστεί πλέον αναγκαία στην πλειοψηφία των βιομηχανιών με συνέπεια μηχανικοί, εμπειρογνώμονες πεδίου και επιστήμονες να εργάζονται ταυτόχρονα για την εκμετάλλευση τεράστιων ποσοτήτων δεδομένων τα οποία είναι κρίσιμης σημασίας τόσο για τη διενέργεια επιχειρηματικής δραστηριότητας όσο και για την πρόοδο της επιστήμης. Η επιτυχία και η αποτελεσματικότητα μιας τέτοιας ανάλυσης εξαρτάται από πολυάριθμες προκλήσεις οι οποίες σχετίζονται τόσο με τα ίδια τα δεδομένα όσο και τη φύση των αναλυτικών διεργασιών καθώς και με το υπολογιστικό περιβάλλον στο οποίο τέτοιες διεργασίες εκτελούνται. Αυτά τα ζητήματα έχουν δώσει την ώθηση για την εμφάνιση πολλών ετερογενών προγραμματιστικών μοντέλων, περιβαλλόντων εκτέλεσης και αποθηκών δεδομένων ώστε να καταστεί δυνατή η διαχείριση δεδομένων σε μεγάλη κλίμακα. Ενώ οι περισσότερες οικογένειες αυτών των συστημάτων έχουν μεγάλη επιτυχία, ακόμα, προβάλλουν τα πλεονεκτήματά τους σε ένα περιορισμένο υποσύνολο εφαρμογών και τύπων δεδομένων. Παραδείγματος χάριν, οι πλατφόρμες επεξεργασίας γράφων περιορίζουν το βαθμό ελευθερίας στους υπολογισμούς σε κάθε κόμβο (ή τμήμα) του γράφου με συνέπεια να αποτυγχάνουν να εκμεταλλευτούν πλήρως τις δυνατότητες για παραλληλισμό των υπολογισμών. Επιπροσθέτως, οι τωρινές ροές εργασιών των αναλύσεων (analytics workflows) είναι τρομερά περίπλοκες καθώς οι πηγές των δεδομένων είναι ετερογενείς και κατανεμημένες. Επιπλέον, οι υπολογιστικές διεργασίες μπορεί να είναι αυθαίρετης χρονικής διάρκειας και να απαιτούν διαφορετικές λεπτομέρειες στην εκτέλεσή τους ανάλογα και με το ρόλο του εκάστοτε χρήστη και των τεχνικών γνώσεων και δεξιοτήτων που αυτός κατέχει. Τέτοιες διεργασίες μπορεί να κυμαίνονται από απλές ή και περίπλοκες εργασίες και ερωτήματα πάνω στα δεδομένα έως αλγοριθμική επεξεργασία όπως είναι η εξόρυξη δεδομένων και να απαιτούν πολλαπλές μηχανές ερωτημάτων.

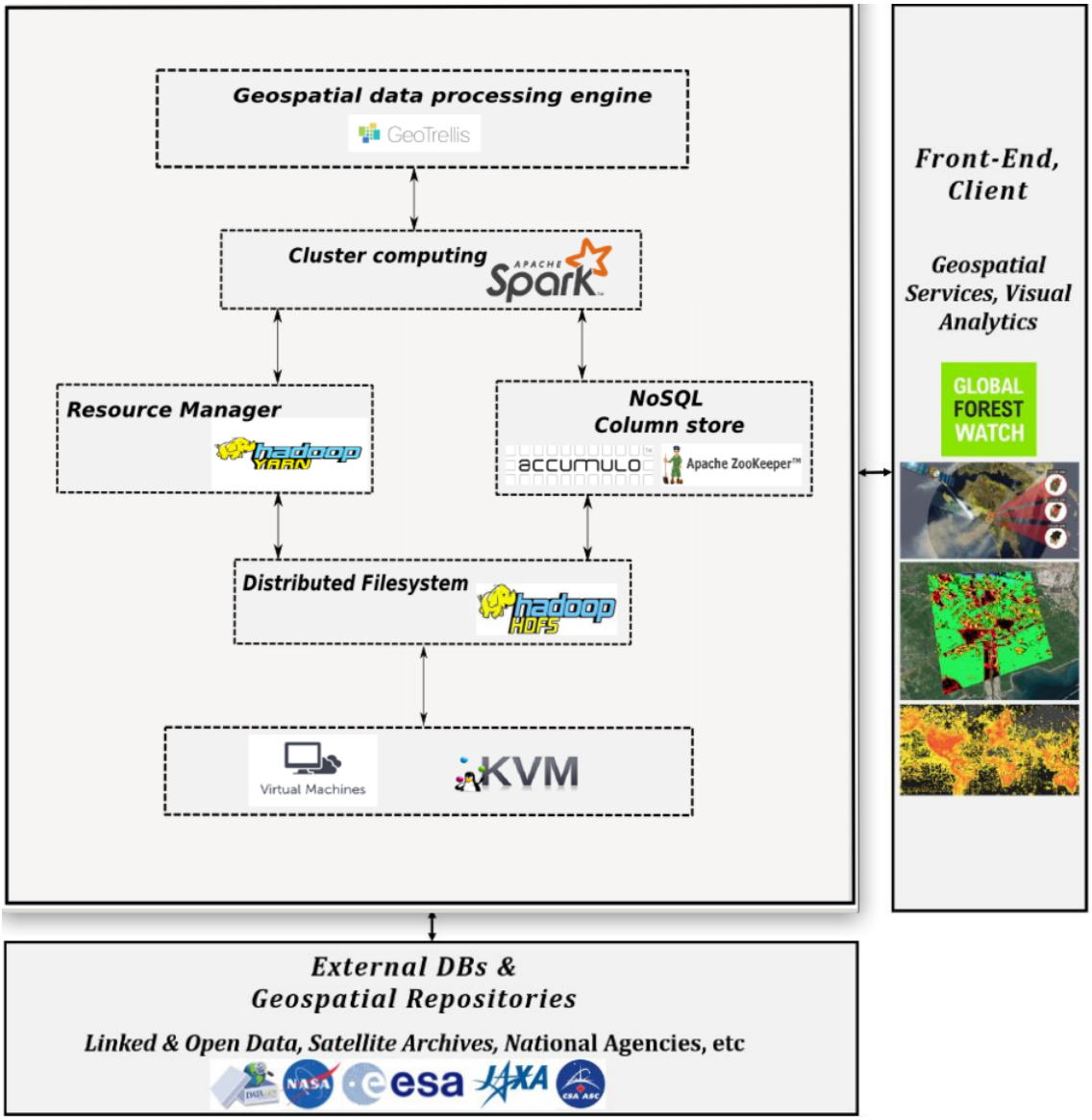
Για να είναι δυνατόν να εκμεταλλευθούμε αυτή την πληθώρα δεδομένων, υπολογιστικών

μηχανών, προγραμματιστικών μοντέλων, βιβλιοθηκών και εργαλείων τα οποία είναι διαθέσιμα χρειαζόμαστε συντονισμένες, προσαρμοστικές και ολοκληρωτικές προσπάθειες ώστε να συνδυάσουμε αρμονικά τις υπάρχουσες τεχνολογίες. Αυτές οι προσπάθειες είναι ο κεντρικός στόχος της παρούσας μεταπτυχιακής εργασίας. Τέτοιες προσπάθειες περιλαμβάνουν τον ορισμό ευέλικτων προγραμματιστικών μοντέλων, την μοντελοποίηση και αξιολόγηση των επιδόσεων των υπολογιστικών μηχανών, τη βελτιστοποίηση αλγορίθμων, την κατανεμημένη εκτέλεση αλγορίθμων επεξεργασίας καθώς επίσης και τη διαχείριση των ροών εργασιών και τεχνικές οπτικοποίησης για τη διενέργεια αναλύσεων μεγάλης πολυπλοκότητας πάνω σε μεγάλα, ετερογενή και πιθανώς αδόμητα δεδομένα κατά μήκος πολυποίκιλων υπολογιστικών οικοσυστημάτων.

Προς την κατεύθυνση αυτή παρουσιάζεται σε αυτή την εργασία μια πλατφόρμα ανάλυσης μεγάλων γεωχωρικών δεδομένων για εφαρμογές παρατήρησης της γης η οποία για την επίτευξη των στόχων της αποδοτικής ανάλυσης, αποθήκευσης και διαχείρισης μεγάλων δεδομένων ενσωματώνει μια σειρά από εργαλεία, βιβλιοθήκες και υπολογιστικά μοντέλα σε ένα κατανεμημένο περιβάλλον cluster υπολογιστών. Ειδικότερα, η βασική λειτουργικότητα συνίσταται από το Geotrellis, μια μηχανή επεξεργασίας γεωχωρικών δεδομένων για εφαρμογές υψηλών επιδόσεων για την αποθήκευση, διαχείριση και επεξεργασία των δεδομένων και το framework Apache Spark για την κατανομή υπολογισμών και δεδομένων κατά μήκος του cluster. Διάφοροι αλγόριθμοι υλοποιήθηκαν στη γλώσσα προγραμματισμού Scala τόσο για εργασίες ETL (Extract, Transform, Load) πάνω στα raw δεδομένα όσο και για την πρόσβαση και την επεξεργασία πολυφασματικών (multispectral) δορυφορικών εικόνων. Αναπτύχθηκε επίσης και ένα απλό πρόγραμμα πελάτη του συστήματος (WebGIS) για να εξυπηρετήσει τις ανάγκες προβολής των αποτελεσμάτων το οποίο και βασίζεται στη βιβλιοθήκη Leaflet η οποία είναι γραμμένη στην γλώσσα προγραμματισμού javascript. Το ανεπτυγμένο σύστημα στην τρέχουσα μορφή του καλύπτει μερικώς τον Ελλαδικό χώρο με πολυφασματικά δεδομένα τα οποία προέρχονται από το δορυφόρο Landsat 8, τα οποία αποθηκεύονται και προ-επεξεργάζονται αυτόματα στο υλικό το οποίο έχουμε στη διάθεσή μας, για σκοπούς επίδειξης. Τα ανεπτυγμένα ερωτήματα επεξεργασίας των δεδομένων εστιάζουν σε αγροτικές εφαρμογές και παράγουν τόσο τεχνητά έγχρωμα σύνθετα με βάση καλώς ορισμένους δείκτες, τα οποία και παρέχουν πληροφορία σχετικά με την κατάσταση των καλλιεργειών διαχρονικά ανά pixel και ανά έτος, όσο και χάρτες αποτύπωσης της εποχικότητας από τους οποίους είναι δυνατόν να εξαχθούν μοτίβα για τον κύκλο ζωής των υπό παρακολούθηση καλλιεργειών.

4.2 Σχεδιασμός - Περιγραφή Αρχιτεκτονικής

Ο κύριος σκοπός αυτής της εργασίας αυτής ήταν ο σχεδιασμός και η υλοποίηση ενός πλαισίου για την ανάλυση μεγάλων γεωχωρικών δεδομένων (πολυφασματικών δορυφορικών εικόνων) σε περιβάλλον cluster υπολογιστών με απώτερο στόχο τη διενέργεια διαχρονικών αναλύσεων ολόκληρων αρχείων τηλεπισκοπικών δεδομένων για αγροτικές εφαρμογές. Ποικίλες συνιστώσες και υπολογιστικά βήματα εμπλέκονται στη ρύθμιση, στη λειτουργία και στη χρησιμοποίηση του ανεπτυγμένου συστήματος.



Σχήμα 4.1: Η αρχιτεκτονική του ανεπτυγμένου συστήματος.

Όπως αναφέρθηκε, η βασική λειτουργικότητα του ανεπτυγμένου πλαισίου συνίσταται από το Geotrellis για την αποθήκευση, διαχείριση και επεξεργασία των τηλεπισκοπικών δεδομένων και το framework Apache Spark για την κατανομή υπολογισμών και δεδομένων κατά μήκος του cluster . Το Geotrellis επιλέχθηκε ως το βασικό σύστημα της υλοποίησής μας εξαιτίας των δυνατοτήτων του στην επεξεργασία μεγάλων γεωχωρικών δεδομένων. Παρ΄ όλο που είναι ένα σχετικά νέο σύστημα, με συνέπεια να μην έχει ωριμάσει αρκετά, βασίζεται στην πλέον σύγχρονη και ευρέως δοκιμασμένη στοίβα λογισμικού κατανεμημένης επεξεργασίας και αποθήκευσης μεγάλων δεδομένων γενικού σκοπού (οικοσύστημα Hadoop) και παρέχει τα κατάλληλα προγραμματιστικά εργαλεία για την επεξεργασία μεγάλων γεωχωρικών δεδομένων. Έκτος από τα ολοκληρωμένα APIs τα οποία παρέχει στους προγραμματιστές εφαρμογών δίνει τη δυνατότητα για την ενσωμάτωση πολλών εξωτερικών βιβλιοθηκών, με συνέπεια να δομεί μια εύρωστη και ταυτόχρονα ευέλικτη πλατφόρμα επεξεργασίας. Το Geotrellis, αυτή τη στιγμή εί-

ναι από τα πιο ανερχόμενα και πλήρη NoSQL συστήματα στον τομέα των μηχανών γεωχωρικής επεξεργασίας, εμφανίζει σημαντικά καλύτερη δυνατότητα έκφρασης παράλληλων υπολογισμών από συστήματα που ακολουθούν το μοντέλο δεδομένων πινάκων (Array Databases) όπως επίσης και σημαντικά μεγαλύτερες δυνατότητες από συστήματα συγγενούς προσέγγισης και αρχιτεκτονικής.

Την τρέχουσα χρονική περίοδο, τα τηλεπισκοπικά δεδομένα τα οποία είναι διαθέσιμα στο σύστημα για επεξεργασία και διενέργεια αναλύσεων, προέρχονται από την αποστολή Landsat Data Continuity Mission (LDCM) [1]. Οι αισθητήρες του δορυφόρου Landsat 8, OLI και TIRS, καταγράφουν πολυχρονικά (multitemporal), πολυφασματικά (multispectral) δεδομένα αρκετά καλής χωρικής ανάλυσης. Τα ακατέργαστα (raw) δεδομένα του δορυφόρου Landsat 8 λαμβάνονται, αποθηκεύονται και προ-επεξεργάζονται αυτόματα από το σύστημα μας, σε μια διαδικασία η οποία θα περιγραφεί αναλυτικά παρακάτω.

Η διαχρονική ανάλυση των αποθηκευμένων δεδομένων, η παραγωγή τεχνητών έγχρωμων σύνθετων τα οποία αποτελούν προϊόντα των διαχρονικών αναλύσεων με βάση καλώς ορισμένους δείκτες καθώς και η παραγωγή χαρτών αποτύπωσης της εποχικότητας από τους οποίους είναι δυνατόν να εξαχθούν μοτίβα για τον κύκλο ζωής των υπό παρακολούθηση καλλιεργειών είναι οι λειτουργικότητες κλειδιά του ανεπτυγμένου συστήματος στην τωρινή έκδοσή του. Αυτές οι λειτουργικότητες, οι οποίες έχουν τη μορφή προγραμμάτων γραμμένων στη γλώσσα προγραμματισμού Scala, χρησιμοποιούν το σύνολο δεδομένων του δορυφόρου Landsat 8 και τα APIs του Geotrellis, ώστε να αντλήσουν πληροφορία με τηλεπισκοπικές μεθόδους και να παράξουν τους αντίστοιχους χρωματικούς χάρτες (color maps), οι οποίοι και περιέχουν την αποκτηθείσα πληροφορία.

Το τωρινό υλικό (hardware) το οποίο υποστηρίζει το ανεπτυγμένο σύστημα είναι ένας υπολογιστής (host) με 8 διαθέσιμους πυρήνες επεξεργασίας και 32 GB RAM, στον οποίο είναι εγκατεστημένα το λειτουργικό σύστημα openSUSE Leap 42.1 καθώς και η υποδομή για τη δημιουργία εικονικών μηχανών KVM. Μέσω του KVM δημιουργήθηκαν 3 εικονικά μηχανήματα πάνω στο host μηχάνημα σε μια διάταξη που θα περιγραφεί αναλυτικά σε επόμενη ενότητα, τα οποία και αποτέλεσαν μια μικρογραφία ενός cluster υπολογιστών για τις ανάγκες υλοποίησης της παρούσας εργασίας. Σε αυτό το περιβάλλον επίδειξης, τα αποθηκευμένα και προ-επεξεργασμένα τηλεπισκοπικά δεδομένα, καλύπτουν πλήρως την Ελληνική επικράτεια παρέχοντας κάθε -περίπου- 16 μέρες δορυφορικές εικόνες από το ξεκίνημα της αποστολής του δορυφόρου Landsat 8 (Φεβρουάριος 2013). Η μεταφορά του συστήματος σε περιβάλλον παραγωγής έχει ήδη δρομολογηθεί.

Οι βασικές συνιστώσες του ανεπτυγμένου συστήματος παρουσιάζονται διαγραμματικά στο **σχήμα 4.1** και περιγράφονται ενδελεχώς στις παρακάτω ενότητες. Όσον αναφορά στην αρχιτεκτονική του συστήματος μας, όπως ίσως κανείς συμπεραίνει και από το σχήμα 4.1, αυτή ακολουθεί το πρότυπο της αρθρωτής σχεδίασης (modular design) καθώς χαρακτηρίζεται από ανεξάρτητες, καλώς ορισμένης λειτουργικότητας μονάδες λογισμικού για την επίτευξη της κατανεμημένης επεξεργασίας γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών.

Υπάρχει μονάδα λογισμικού υπεύθυνη για την αυτόματη συλλογή των πρόσφατα καταγε-

[1] http://landsat.usgs.gov/

γραμμένων συνόλων δεδομένων, της αποσυμπίεσης και αρχειοθέτησης των raw τηλεπισκοπικών δεδομένων. Κατάλληλη μονάδα εκτελεί όλες τις διαδικασίες της προ-επεξεργασίας οι οποίες είναι απαραίτητες για την εισαγωγή των δεδομένων στο σύστημα ώστε αυτά να είναι σε μορφή έτοιμη για επεξεργασία οποτεδήποτε το επιθυμούν οι χρήστες του συστήματος. Επιπρόσθετα, υπάρχουν δομικές μονάδες οι οποίες είναι επιφορτισμένες με τη διαχείριση τόσο του cluster όσο και του κατανεμημένου συστήματος αρχείων καθώς και όλων των επιμέρους διεργασιών οι οποίες είναι κρίσιμες για την εύρυθμη λειτουργία του συστήματος.

Τέλος, υπάρχει το πρόγραμμα πελάτη του συστήματος (Web Client) το οποίο είναι υπεύθυνο για τη διεπαφή με το χρήστη, την αποστολή ερωτημάτων επεξεργασίας δεδομένων ή αιτημάτων προβολής αναλύσεων στον εξυπηρετητή καθώς και για την παραλαβή και προβολή των αποτελεσμάτων.

Το διάγραμμα ροής του σχήματος 4.1 εκτός από τις βασικές μονάδες λογισμικού του ανεπτυγμένου συστήματος παρουσιάζει σιωπηρά και τις ροές δεδομένων ή/και υπολογισμών κατά μήκος του συστήματος καθώς και τα σημεία αλληλεπίδρασης μεταξύ των διάφορων δομικών μονάδων. Οποιαδήποτε στιγμή, οι μονάδες λογισμικού υπεύθυνες για τις διάφορες εργασίες πάνω στα δεδομένα βρίσκονται σε λειτουργία και ανακτούν, αποθηκεύουν, προ-επεξεργάζονται και εισάγουν στο Geotrellis νέα δεδομένα. Με την εισαγωγή τα δεδομένα είναι αμέσως διαθέσιμα στους χρήστες για επεξεργασία. Ταυτόχρονα, μέσω του Web Client, οι χρήστες του συστήματος μπορούν να υποβάλουν ερωτήματα επεξεργασίας στο σύστημα για τα υπάρχοντα δεδομένα και να λαμβάνουν τα ζητούμενα αποτελέσματα.

4.2.1 Διαχωρισμός Υποσυστημάτων

Όπως φαίνεται και από το διάγραμμα ροής του **σχήματος 4.1** τα υποσυστήματα από τα οποία αποτελείται το ανεπτυγμένο σύστημα για την επεξεργασία μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών είναι τα ακόλουθα:

- Αυτόματη ανάκτηση, αποθήκευση και προ-επεξεργασία δεδομένων
- Virtual Machines
- Hadoop
- Apache Spark
- Accumulo
- Apache Zookeeper
- Geotrellis
- Web Client

4.2.2 Περιγραφή Υποσυστημάτων

Παρακάτω δίνεται η λεπτομερής περιγραφή για καθένα από τα υποσυστήματα που παρουσιάστηκαν. Η περιγραφή αυτή γίνεται με βάση το διάγραμμα ροής δεδομένων (σχήμα 4.1) το οποίο και περιγράφει τα βασικά συστατικά του συστήματος που σχεδιάστηκε και υλοποιήθηκε στην παρούσα εργασία.

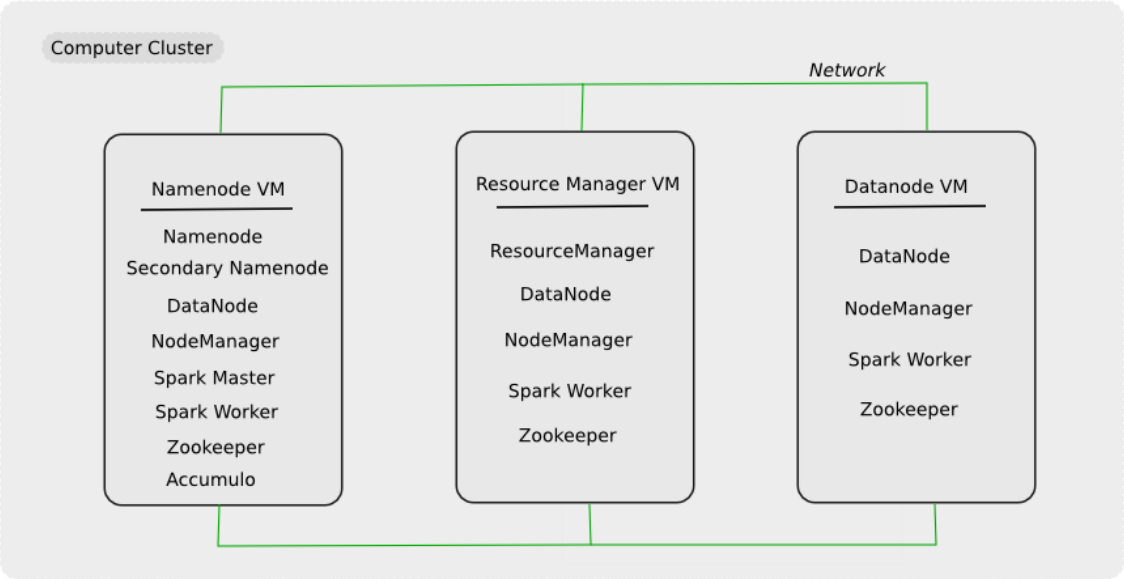
Αυτόματη συλλογή, αποθήκευση και προ-επεξεργασία των δεδομένων

Όσον αναφορά τα στάδια της αυτόματης συλλογής, αποθήκευσης και προ-επεξεργασίας των τηλεπισκοπικών δεδομένων από το σύστημά μας, ένας σημαντικός αριθμός από προγράμματα γραμμένα στη γλώσσα προγραμματισμού Python (Python scripts) καθώς και στη γλώσσα προγραμματισμού Scala αναπτύχθηκαν για τον έλεγχο, τη διευκόλυνση και την αυτοματοποίηση ολόκληρου του εγχειρήματος. Η ανάπτυξη του προαναφερθέντος λογισμικού ήταν απολύτως αναγκαία και καταλυτική για την πρόοδο της εργασίας, καθώς θα ήταν αδύνατη η χειροκίνητη εκτέλεση των παραπάνω εργασιών λόγω της φύσης των δεδομένων (big data). Ακόμα και για την απλή περίπτωση της διαχείρισης μίας μόνο δορυφορικής εικόνας για σκοπούς ανάπτυξης του συστήματος, ο χρόνος και ο κόπος που απαιτούνταν για τη χειροκίνητη εκτέλεση ήταν εκθετικά μεγαλύτερος. Η ανάπτυξη επομένως του εν λόγω υποσυστήματος ήταν κρίσιμης σημασίας.

Εστιάζοντας στον αυτοματισμό της διαδικασίας, αρχικά ένα Python script το οποίο είναι υπεύθυνο για τον έλεγχο του αρχείου δεδομένων του δορυφόρου Landsat 8, εξετάζει για το αν υπάρχουν νέα σύνολα δεδομένων και συλλέγει οποιαδήποτε καινούργια σύνολα βρεθούν.

Από τη στιγμή στην οποία όλα τα νέα σύνολα δεδομένων έχουν γίνει διαθέσιμα, ένα άλλο Python script τα αποσυμπιέζει, καθώς αυτά παρέχονται σε συμπιεσμένη μορφή από το αρχείο δεδομένων του δορυφόρου Landsat 8 και στη συνέχεια τα αρχειοθετεί. Αυτό σημαίνει ότι τα νέα σύνολα δεδομένων μεταφέρονται στους κατάλληλους φακέλους στο κατανεμημένο σύστημα αρχείων (Hadoop HDFS), ώστε να είναι διαθέσιμα μέσω του cluster υπολογιστών για επεξεργασία μέσω αλγορίθμων τηλεπισκόπησης από το Geotrellis.

Από τη στιγμή που ολοκληρώνονται τα παραπάνω στάδια σχετικά με τη διαχείριση των συλλογών δεδομένων και είναι γνωστή η διαδρομή στο σύστημα αρχείων (μέσω url) στην οποία βρίσκεται η κάθε εικόνα, ένα script γραμμένο στη γλώσσα προγραμματισμού Scala αναλαμβάνει τη διαδικασία του reprojection των δεδομένων στο επιθυμητό σύστημα αναφοράς, το tiling των δεδομένων ώστε να είναι δυνατή η κατανεμημένη επεξεργασία τους καθώς και τον υπολογισμό διάφορων τηλεπισκοπικών δεικτών από τα αποθηκευμένα δεδομένα (π.χ. NDVI). Τα παραπάνω βήματα έχουν ως κύριο στόχο τη μετατροπή των δεδομένων στο μοντέλο δεδομένων που ακολουθεί το Geotrellis με συνέπεια μετά το τέλος των παραπάνω διαδικασιών τα δεδομένα να είναι έτοιμα για τη διενέργεια αναλύσεων οποτεδήποτε ζητηθεί μέσω του ανεπτυγμένου συστήματος.



Σχήμα 4.2: Σχήμα του υλοποιημένου cluster εικονικών μηχανών συμπεριλαμβανομένων και των διεργασιών που κάθε εικονικό μηχάνημα φιλοξενεί για σκοπούς διαχείρισης του cluster .

Virtual Machines (VMs)

Η βασική επιδίωξη της παρούσας εργασίας ήταν η κατανεμημένη ανάλυση μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών. Τι εννοούμε όμως όταν λέμε cluster υπολογιστών; Ένα cluster υπολογιστών [19] αποτελείται από ένα σύνολο από "χαλαρά" ή "στενά" συνδεδεμένους, μέσω δικτύου, υπολογιστές (κόμβοι) οι οποίοι συνεργάζονται ώστε, από πολλά σημεία, να μπορούν να θεωρηθούν ως ένα ενιαίο σύστημα. Ανόμοια, από τα πλέγματα υπολογιστών (grid computers), στα clusters υπολογιστών κάθε κόμβος ρυθμίζεται ώστε να εκτελεί την ίδια εργασία με τους υπόλοιπους κόμβους, ρύθμιση η οποία ελέγχεται μέσω κατάλληλου λογισμικού. Οι κόμβοι ενός cluster συνήθως συνδέονται μεταξύ τους μέσω γρήγορων τοπικών δικτύων, με κάθε κόμβο να "τρέχει" το δικό του λειτουργικό σύστημα. Στις περισσότερες περιπτώσεις, όλοι οι κόμβοι χρησιμοποιούν το ίδιο υλικό και τρέχουν το ίδιο λειτουργικό σύστημα. Ένα cluster υπολογιστών χρησιμοποιείται για να βελτιώσει την επίδοση και τη διαθεσιμότητα πόρων για κάποια εργασία σε σχέση με την εκτέλεση της ίδιας εργασίας σε ένα και μόνο μηχάνημα ενώ ταυτόχρονα είναι πολύ πιο αποδοτική λύση σε σχέση με τη χρήση πολλών ξεχωριστών υπολογιστών οι οποίοι όμως εργάζονται ανεξάρτητα. Τα clusters υπολογιστών εμφανίστηκαν ως το αποτέλεσμα της σύγκλισης ενός σημαντικού αριθμού υπολογιστικών τάσεων όπως είναι η ευρεία διαθεσιμότητα μικροεπεξεργαστών χαμηλού κόστους, τα δίκτυα υψηλών ταχυτήτων καθώς και λογισμικό κατάλληλο για την εκτέλεση κατανεμημένων υπολογισμών υψηλών επιδόσεων. Έχουν ένα πολύ μεγάλο εύρος εφαρμογών το οποίο κυμαίνεται από μικρά εταιρικά clusters τα οποία περιλαμβάνουν λίγους μόνο κόμβους (< 10) έως κάποιους από τους μεγαλύτερους υπερ-υπολογιστές του κόσμου όπως είναι ο υπερ-υπολογιστής Sequoia της IBM (> 90.000 κόμβους).

Με βάση τα παραπάνω επιλέξαμε να δημιουργήσουμε και να πειραματιστούμε με ένα cluster

για την κατανεμημένη ανάλυση μεγάλων γεωχωρικών δεδομένων στα πλαίσια της παρούσας εργασίας. Καθώς δεν ήταν δυνατή η δημιουργία ενός cluster από φυσικά μηχανήματα καταλήξαμε στη λύση της δημιουργίας και σύνδεσης μέσω δικτύου 3 εικονικών μηχανημάτων (Virtual Machines - VMs), τα οποία αποτέλεσαν ένα μικρό cluster επίδειξης για τους σκοπούς της εργασίας, πάνω από ένα φυσικό μηχάνημα. Για τη δημιουργία των εικονικών μηχανημάτων χρησιμοποιήθηκε κατάλληλη υποδομή εικονικών μηχανών. Η υποδομή αυτή ήταν το KVM (Kernel-based Virtual Machine), μέσω του οποίου πραγματοποιήθηκε η δημιουργία, η ρύθμιση και η διαχείριση των εικονικών μηχανών. Τα 3 VMs ήταν πανομοιότυπα καθώς όλα από πλευράς υλικού είχαν 1 πυρήνα επεξεργασίας και 6 GB RAM και από πλευράς λειτουργικού συστήματος όλα έτρεχαν τη διανομή Ubuntu 14.04.4 LTS.

Για τη λειτουργία και τη διαχείριση του cluster ήταν απαραίτητο το λογισμικό εκείνο το οποίο θα έδινε την εικόνα στο χρήστη μιας ενιαίας μονάδας αποθήκευσης και επεξεργασίας πάνω από τα 3 VMs. Το λογισμικό το οποίο επιλέχθηκε ήταν το λογισμικό Hadoop, το οποίο θα αναλυθεί στην επόμενη υπο-ενότητα και είναι υπεύθυνο για τη δημιουργία του κατανεμημένου συστήματος αρχείων πάνω από τα 3 VMs, καθώς και τη διαχείριση των πόρων του cluster.

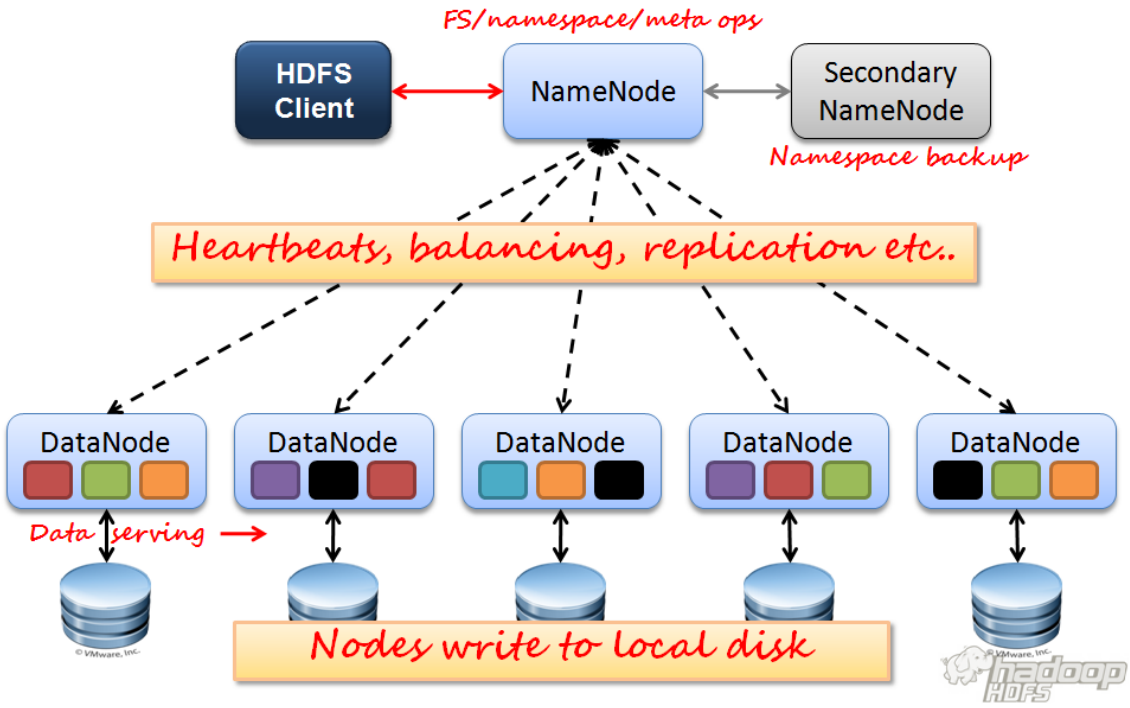
Στο Σχήμα 4.2 παρουσιάζεται ένα απλοποιημένο διάγραμμα του ανεπτυγμένου cluster που χρησιμοποιήθηκε στα πλαίσια της παρούσας εργασίας. Για κάθε εικονικό μηχάνημα αναγράφονται και οι διεργασίες που αυτό φιλοξενεί ώστε να είναι δυνατή η δημιουργία, λειτουργία και συντήρηση του cluster.

Hadoop

Το Apache Hadoop[2] είναι ένα πλαίσιο λογισμικού ανοικτού κώδικα για την κατανεμημένη αποθήκευση και επεξεργασία πολύ μεγάλων συνόλων δεδομένων σε περιβάλλοντα clusters υπολογιστών τα οποία φτιάχνονται από hardware κοινών δυνατοτήτων. Όλες οι μονάδες λογισμικού οι οποίες αποτελούν το Hadoop έχουν σχεδιαστεί πάνω στην υπόθεση ότι οι αστοχίες υλικού είναι κοινότυπη κατάσταση και πρέπει να αντιμετωπίζονται αυτόματα από το σύστημα.

Ο πυρήνας του Hadoop αποτελείται από το κομμάτι της αποθήκευσης, το οποίο είναι γνωστό ως το Hadoop Distributed File System (HDFS) και το κομμάτι της επεξεργασίας το οποίο καλείται MapReduce. Το Hadoop κατατμεί τα αρχεία σε μεγάλα (σε σχέση με άλλα συστήματα αρχείων) μπλοκ και τα κατανέμει κατά μήκος των κόμβων σε ένα cluster. Για την επεξεργασία των δεδομένων, το Hadoop μεταφέρει τον κώδικα σε πακέτα σε όλους τους κόμβους ώστε να είναι δυνατή η παράλληλη επεξεργασία των δεδομένων από όλους τους κόμβους. Αυτή η προσέγγιση εκμεταλλεύεται την τοπικότητα των δεδομένων - δηλαδή κάθε κόμβος επεξεργάζεται δεδομένα τα οποία είναι αποθηκευμένα σε αυτόν - ώστε να είναι δυνατή η ταχύτερη και αποδοτικότερη επεξεργασία των δεδομένων από ότι θα ήταν σε μια πιο συμβατική αρχιτεκτονική υπερ-υπολογιστών η οποία βασίζεται σε ένα παράλληλο σύστημα αρχείων στο οποίο δεδομένα και υπολογισμοί διανέμονται στους κόμβους δια μέσου δικτύων

[2] http://hadoop.apache.org/



Σχήμα 4.3: Η αρχιτεκτονική του συστήματος Hadoop. (Πηγή: hadoop.apache.org)

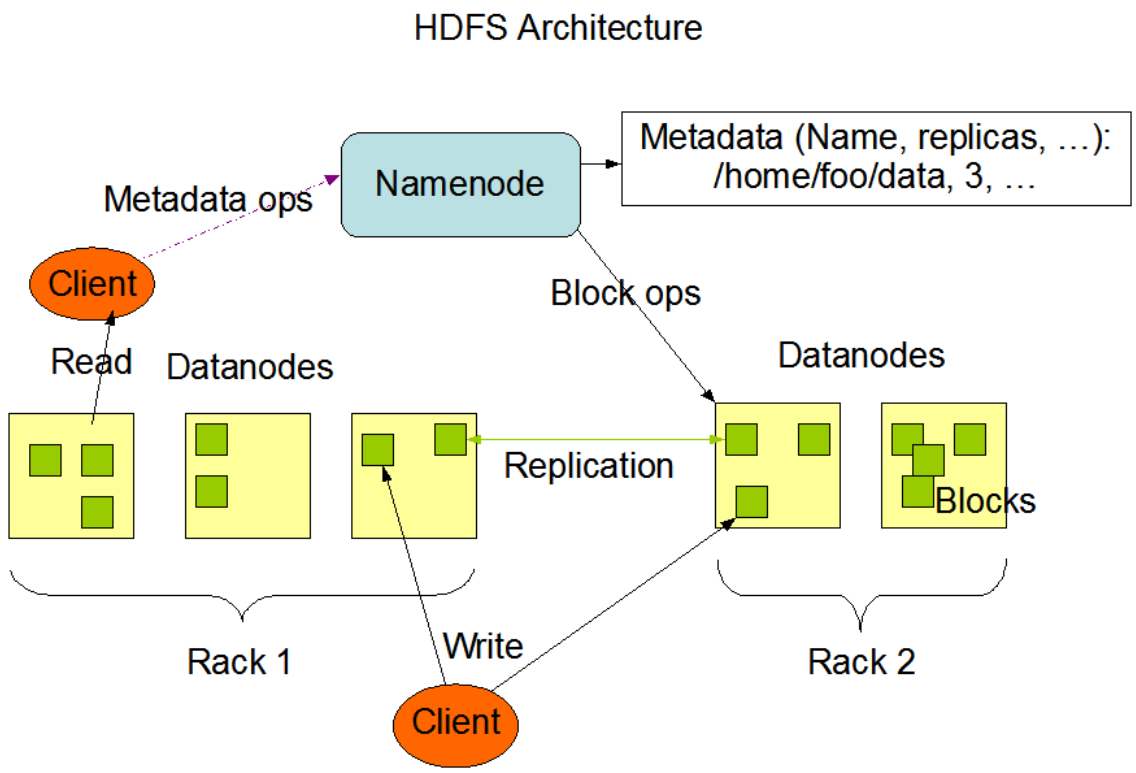
υψηλών ταχυτήτων.

Τη βάση του Apache Hadoop framework αποτελούν οι παρακάτω μονάδες:

- **Hadoop Common** - βιβλιοθήκες και εργαλεία τα οποία απαιτούν οι υπόλοιπες μονάδες του Hadoop καθώς παρέχουν αφαιρέσεις στο επίπεδο του συστήματος αρχείων καθώς και του λειτουργικού συστήματος.
- **HDFS** - ένα κατανεμημένο σύστημα αρχείων το οποίο αποθηκεύει δεδομένα σε "κοινά" μηχανήματα, παρέχοντας πολύ υψηλό συνδυασμένο εύρος ζώνης κατά μήκος του cluster.
- **Hadoop YARN** - μια πλατφόρμα υπεύθυνη για τη διαχείριση των πόρων του cluster καθώς και τη χρονοδρομολόγηση των εφαρμογών των χρηστών.
- **Hadoop MapReduce** - μια υλοποίηση του προγραμματιστικού μοντέλου MapReduce για επεξεργασία δεδομένων σε μεγάλη κλίμακα, είτε στην μορφή MapReduce/MR1 είτε στη μορφή YARN/MR2.

Ο όρος Hadoop, πλέον, δεν αφορά μόνο στις προαναφερθείσες μονάδες λογισμικού αλλά και σε ολόκληρη τη συλλογή των επιπρόσθετων πακέτων λογισμικού που μπορούν να εγκατασταθούν επάνω ή μαζί με το Hadoop όπως, ενδεικτικά, είναι οι πλατφόρμες Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache ZooKeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache Oozie και Apache Storm.

Οι μονάδες MapReduce και HDFS του Hadoop άντλησαν έμπνευση από τις επιστημονικές εργασίες της εταιρείας Google σχετικά με τα δικά της λογισμικά για το προγραμματιστικό

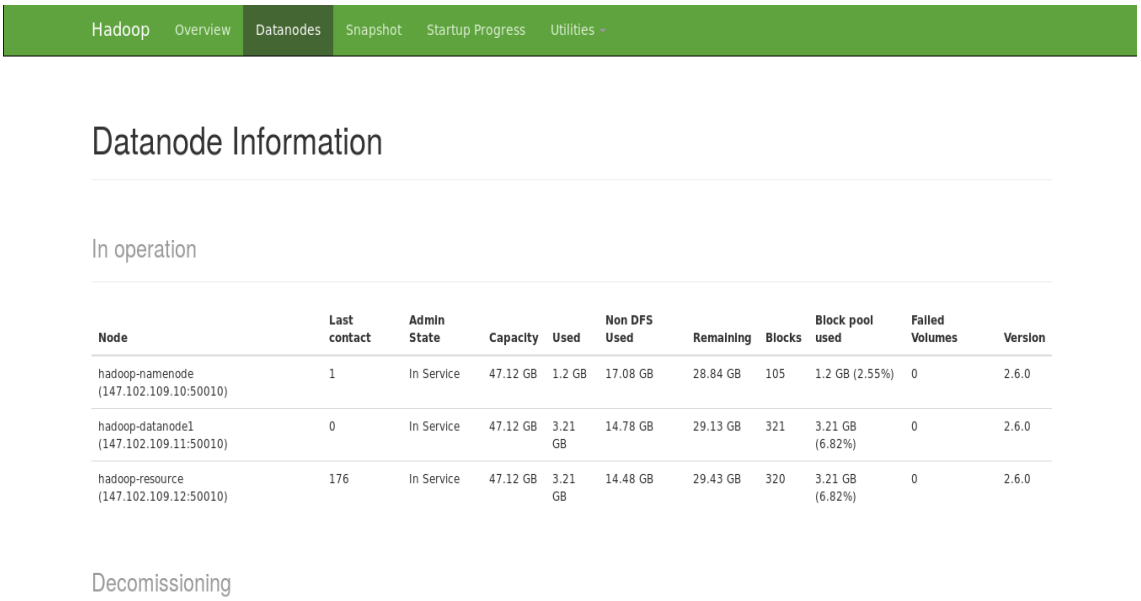


Σχήμα 4.4: Η αρχιτεκτονική του κατανεμημένου συστήματος αρχείων HDFS. (Πηγή: hadoop.apache.org)

μοντέλο MapReduce καθώς και το Google File System.

Το Hadoop έχει υλοποιηθεί κατά κύριο λόγο στη γλώσσα προγραμματισμού Java, με κάποια εγγενή τμήματα κώδικα να έχουν υλοποιηθεί στη γλώσσα προγραμματισμού C και τα εργαλεία γραμμής εντολών να έχουν υλοποιηθεί ως shell scripts. Ενώ υπάρχει εγγενής υποστήριξη από το σύστημα για προγράμματα χρηστών σύμφωνα με το μοντέλο MapReduce γραμμένα στη γλώσσα προγραμματισμού Java, οποιαδήποτε γλώσσα μπορεί να χρησιμοποιηθεί μέσω του βοηθητικού εργαλείου Hadoop Streaming.

Για την αποτελεσματική χρονοδρομολόγηση των εργασιών μέσα στο cluster κάθε σύστημα αρχείων συμβατό με το Hadoop πρέπει να παρέχει ενημερότητα σχετικά με την τοποθεσία του κάθε κόμβου και πιο συγκεκριμένα το όνομα του rack στο οποίο κάθε κόμβος επεξεργασίας ανήκει. Με αυτόν τον τρόπο, οι εφαρμογές μπορούν να χρησιμοποιήσουν αυτή την πληροφορία για να εκτελέσουν τον κώδικά τους στον κόμβο στον οποίο βρίσκονται τα δεδομένα τα οποία πρέπει να επεξεργαστούν και σε περίπτωση αποτυχίας να συμβεί αυτό να επεξεργαστούν τα δεδομένα μέσα στο ίδιο rack στο οποίο βρίσκονται ώστε να ελαττωθεί κατά το δυνατόν η μετακίνηση δεδομένων και συνεπώς η δικτυακή καθυστέρηση. Το HDFS χρησιμοποιεί και αυτό την παραπάνω πληροφορία όταν κρατάει αντίγραφα των αποθηκευμένων δεδομένων για να υπάρχει πλεονασμός δεδομένων (data redundancy) κατά μήκος πολλαπλών racks. Η προσέγγιση αυτή μειώνει την επίδραση της πτώσης του ρεύματος σε κάποιο rack ή της καταστροφής κάποιου switch, καθώς ακόμα και αν κάποια από αυτές τις αστοχίες υλικού



Σχήμα 4.5: Στιγμιότυπο από τη διεπιφάνεια χρήστη για την παρακολούθηση της κατάστασης του κατανεμημένου συστήματος αρχείων HDFS.

συμβεί τα δεδομένα θα εξακολουθούν να παραμένουν διαθέσιμα.

Όπως βλέπουμε και στο Σχήμα 4.3 ένα μικρό Hadoop cluster περιλαμβάνει ένα μόνο κύριο (master) κόμβο και πολλούς κόμβους σκλάβους (worker). Ο master κόμβος εκτελεί τις διεργασίες του Resource Manager, του NodeManager, του NameNode, του Secondary NameNode και του DataNode. Ένας slave ή worker κόμβος εκτελεί τις διεργασίες του NodeManager και του DataNode. **Για τη δική μας αρχιτεκτονική (Σχήμα 4.2), στα πλαίσια της παρούσας εργασίας, επιλέξαμε να έχουμε ένα κόμβο ως τον master (NameNode, Secondary NameNode, DataNode, NodeManager), έναν κόμβο τόσο για τη διενέργεια υπολογισμών όσο και για την διαχείριση των πόρων του cluster τον κόμβο Resource Manager (ResourceManager, NodeManager, DataNode) και τέλος ένα τρίτο κόμβο μόνο για τη διενέργεια υπολογισμών τον Datanode (NodeManager, DataNode).**

Σε ένα μεγαλύτερο cluster, οι κόμβοι του HDFS διαχειρίζονται δια μέσου ενός NameNode server ο οποίος χρησιμοποιείται μόνο για να φιλοξενεί το ευρετήριο του συστήματος αρχείων καθώς και ενός δευτερεύων NameNode ο οποίος μπορεί να δημιουργεί στιγμιότυπα των δομών μνήμης του κύριου namenode, με συνέπεια να αποφεύγεται η αλλοίωση του συστήματος αρχείων καθώς και η απώλεια δεδομένων. Παρόμοια, ένας αυτόνομος Resource Manager διαχειρίζεται την χρονοδρομολόγηση εργασιών κατά μήκος των κόμβων.

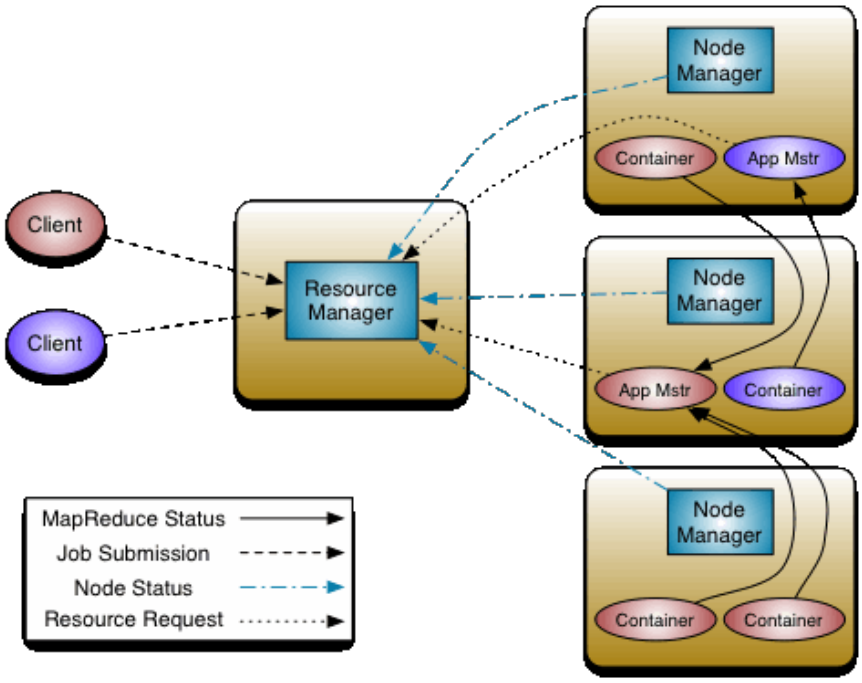
Τώρα όσον αναφορά στο HDFS, αυτό είναι ένα κατανεμημένο, κλιμακώσιμο και μεταφέρσιμο σύστημα αρχείων το οποίο έχει υλοποιηθεί στη γλώσσα προγραμματισμού Java για το πλαίσιο Hadoop. Κάθε Hadoop cluster έχει ένα namenode συν μια δέσμη από datanodes, αν και διάφορες επιλογές για πλεονασμό των δεδομένων είναι διαθέσιμες για τον namenode εξαιτίας της ζωτικής σημασίας που αυτός έχει για την εύρυθμη λειτουργία όλου του συστή-

ματος. Κάθε datanode σερβίρει μπλοκ δεδομένων πάνω από το δίκτυο χρησιμοποιώντας ένα πρωτόκολλο μπλοκ ειδικό για το HDFS. Το σύστημα αρχείων χρησιμοποιεί υποδοχείς (sockets) TCP/IP για την επικοινωνία μεταξύ των κόμβων. Το λογισμικό των κόμβων (clients) χρησιμοποιεί την τεχνική RPC (remote procedure call) για την επίτευξη της επικοινωνίας.

Το HDFS χρησιμοποιείται για την αποθήκευση μεγάλων αρχείων (τυπικά στο εύρος των gigabytes έως terabytes) κατά μήκος πολλαπλών μηχανημάτων. Η αξιοπιστία του συστήματος επιτυγχάνεται μέσω της αντιγραφής των δεδομένων κατά μήκος πολλαπλών κόμβων με συνέπεια θεωρητικά να μην χρειάζεται αποθήκευση τύπου RAID στους κόμβους (αν και για την αύξηση της επίδοσης των λειτουργιών I/O κάποιες ρυθμίσεις RAID μπορεί να είναι χρήσιμες). Με την τυπική τιμή αντιγραφής των δεδομένων να είναι 3, τα δεδομένα αποθηκεύονται σε 3 κόμβους, εκ των οποίων οι 2 κόμβοι βρίσκονται στο ίδιο rack και ένας σε διαφορετικό rack. Οι κόμβοι datanode μπορούν να επικοινωνούν μεταξύ τους ώστε να εξισορροπούν την αποθήκευση των δεδομένων, να μεταφέρουν αντίγραφα των δεδομένων από τον ένα κόμβο στον άλλο καθώς και να κρατάνε αντίγραφα των δεδομένων σε μεγάλο βαθμό. Το HDFS δεν είναι πλήρως συμβατό με το standard POSIX (POSIX-compliant) καθώς οι απαιτήσεις για ένα POSIX σύστημα αρχείων διαφέρουν από τους στόχους μιας εφαρμογής Hadoop. Το ισοζύγιο της μη ύπαρξης ενός πλήρους συμβατού με το standard POSIX συστήματος αρχείων είναι η αυξημένη επίδοση κατά τη μετακίνηση των δεδομένων καθώς και η υποστήριξη λειτουργιών που δεν καθορίζονται από το POSIX.

Η αρχιτεκτονική του HDFS φαίνεται στο Σχήμα 4.4. Στο cluster ο NameNode είναι ο server ο οποίος διαχειρίζεται το σύστημα αρχείων του HDFS και ρυθμίζει την πρόσβαση των πελατών στα αρχεία. Το HDFS εκθέτει ένα σύστημα αρχείων το οποίο επιτρέπει στα δεδομένα των χρηστών να αποθηκεύονται σε αρχεία. Εσωτερικά στο σύστημα, κάθε αρχείο κατατμείται σε ένα ή περισσότερα μπλοκ και αυτά τα μπλοκ αποθηκεύονται σε ένα σύνολο από DataNodes. Ο NameNode εκτελεί τυπικές εργασίες του χώρου ονομάτων του συστήματος αρχείων όπως είναι το άνοιγμα, το κλείσιμο, η αλλαγή ονόματος καθώς και η δημιουργία φακέλων αρχείων. Καθορίζει επίσης την αντιστοίχιση των μπλοκ με τους DataNodes. Οι DataNodes είναι υπεύθυνοι για την εξυπηρέτηση αιτήσεων ανάγνωσης και εγγραφής από τους πελάτες του συστήματος αρχείων. Επίσης πραγματοποιούν τις λειτουργίες της δημιουργίας, διαγραφής και αντιγραφής μπλοκ όπως αυτές καθορίζονται από τον NameNode. Ο NameNode και ο DataNode είναι τμήματα λογισμικού σχεδιασμένα για να "τρέχουν" σε κοινών δυνατοτήτων μηχανήματα. Αυτά τα μηχανήματα τυπικά τρέχουν ένα λειτουργικό σύστημα της οικογένειας GNU/Linux. **Το HDFS σχεδιάστηκε κυρίως για την εξυπηρέτηση αμετάβλητων αρχείων και μπορεί να μην είναι κατάλληλο για συστήματα τα οποία απαιτούν ταυτόχρονες λειτουργίες εγγραφής. Για να αντιμετωπιστεί το ζήτημα αυτό μπορεί να γίνει χρήση επιπλέον λογισμικού τεχνολογίας BigTable**, όπως είναι το Apache Accumulo, το οποίο λειτουργεί επάνω από το HDFS και λύνει το ζήτημα των συνεχών ενημερώσεων και εγγραφών των ίδιων αρχείων δεδομένων. **Στα πλαίσια της παρούσας εργασίας είχαμε αυτό το ζήτημα** και για το λόγο αυτό χρησιμοποιήσαμε το σύστημα Accumulo το οποίο και θα περιγραφεί σε επόμενη υπο-ενότητα.

Στο **Σχήμα 4.5** μπορούμε να δούμε ένα στιγμιότυπο από τη διεπιφάνεια χρήστη που πα-



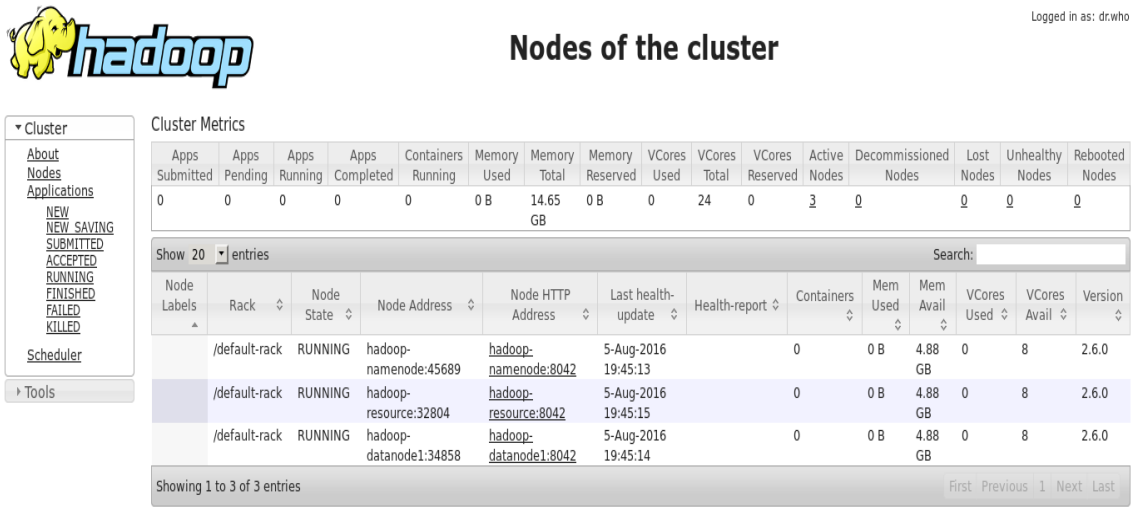
Σχήμα 4.6: Η αρχιτεκτονική του YARN, του συστήματος διαχείρισης πόρων του Hadoop.(Πηγή: hadoop.apache.org)

ρέχει το σύστημα Hadoop για την παρακολούθηση της κατάστασης του HDFS σε πραγματικό χρόνο, οποιαδήποτε στιγμή το απαιτεί ο διαχειριστής του cluster.

Η άλλη βασική μονάδα του Hadoop είναι ο YARN (Yet Another Resource Negotiator), μια πλατφόρμα υπεύθυνη για τη διαχείριση των πόρων του cluster καθώς και τη χρονοδρομολόγηση των εφαρμογών των χρηστών (**Σχήμα 4.6**). Η θεμελιώδης ιδέα του YARN είναι ο διαχωρισμός των λειτουργιών της διαχείρισης των πόρων του cluster και της χρονοδρομολόγησης/παρακολούθησης των εργασιών των χρηστών. Ο διαχωρισμός πραγματοποιείται μέσω της ύπαρξης δύο ξεχωριστών διεργασιών (daemons) οι οποίες διεκπεραιώνουν τις παραπάνω λειτουργίες. Η ιδέα είναι να υπάρχει ένας γενικός διαχειριστής πόρων - ResourceManager (RM) και ένας, για κάθε εφαρμογή, διαχειριστής εφαρμογής - ApplicationMaster (AM). Μια εφαρμογή είναι είτε μία και μόνο εργασία είτε ένα DAG (Directed acyclic graph - κατευθυνόμενο άκυκλο γράφημα) εργασιών.

Ο ResourceManager και ο NodeManager δομούν το υπολογιστικό πλαίσιο πάνω στα δεδομένα. Ο ResourceManager έχει την πλήρη εξουσία για την κατανομή των πόρων σε όλες τις εφαρμογές που τρέχουν στο σύστημα. Ο NodeManager τρέχει σε κάθε κόμβο και είναι το μέσο το οποίο είναι υπεύθυνο για τις εφαρμογές (containers) καθώς παρακολουθεί τη χρησιμοποίηση των πόρων (cpu, μνήμη, δίσκος, δίκτυο) και αναφέρει στον ResourceManager/Scheduler.

Ο ApplicationMaster είναι μια εξειδικευμένη βιβλιοθήκη και επιφορτίζεται με το να ζητά πόρους για την εκάστοτε εφαρμογή από τον ResourceManager και να συνεργάζεται με τον NodeManager για την εκτέλεση και την παρακολούθηση των εργασιών.



Σχήμα 4.7: Στιγμιότυπο από τη διεπιφάνεια χρήστη για την παρακολούθηση της κατάστασης του συστήματος διαχείρισης πόρων YARN.

Ο ResourceManager έχει δύο κύριες συνιστώσες: τον Scheduler και τον ApplicationsManager. Ο Scheduler είναι υπεύθυνος για την κατανομή των πόρων στις διάφορες εφαρμογές που τρέχουν και οι οποίες υπόκεινται στους συνήθεις περιορισμούς της χωρητικότητας, των ουρών κ.τ.λ. Ο Scheduler είναι ένας αμιγής χρονοδρομολογητής με την έννοια ότι δεν πραγματοποιεί παρακολούθηση της κατάστασης της εκάστοτε εφαρμογής. Επιπροσθέτως, δεν προσφέρει εγγυήσεις για την επανεκκίνηση των εφαρμογών που απέτυχαν να τερματίσουν είτε για λόγους της ίδιας της εφαρμογής είτε εξαιτίας αστοχίας υλικού. Ο Scheduler πραγματοποιεί της λειτουργίες της δρομολόγησης βασιζόμενος στις απαιτήσεις σε πόρους που έχουν οι εφαρμογές. Ο τρόπος που συμβαίνει αυτό στην πράξη είναι μέσω της αφηρημένης έννοιας ενός δοχείου πόρων (resource Container) το οποίο ενσωματώνει στοιχεία όπως είναι η μνήμη, η cpu, ο δίσκος, το δίκτυο κ.α.

Ο Scheduler έχει μια πολιτική συνδεόμενων μονάδων λογισμικού (plug-ins) η οποία είναι υπεύθυνη για την διαίρεση των πόρων του cluster ανάμεσα στις διάφορες ουρές, εφαρμογές κ.τ.λ. Οι τρέχοντες δρομολογητές όπως είναι ο CapacityScheduler και ο FairScheduler είναι τέτοια παραδείγματα plug-ins.

Ο ApplicationsManager είναι υπεύθυνος για την αποδοχή υποβολής εργασιών από τους χρήστες, την έναρξη ενός Container για την εκτέλεση ενός ApplicationMaster υπεύθυνου για μια συγκεκριμένη εφαρμογή καθώς και την παροχή της υπηρεσίας της επανέναρξης του container του ApplicationMaster σε περίπτωση αποτυχίας ολοκλήρωσης της εφαρμογής. Ο ApplicationMaster ο οποίος είναι υπεύθυνος για μια συγκεκριμένη εφαρμογή έχει την ευθύνη της λήψης containers επαρκών πόρων από το χρονοδρομολογητή για να είναι δυνατή η εκτέλεσή της και στη συνέχεια έχει την ευθύνη για την παρακολούθηση της κατάστασής της εφαρμογής και της προόδου της.

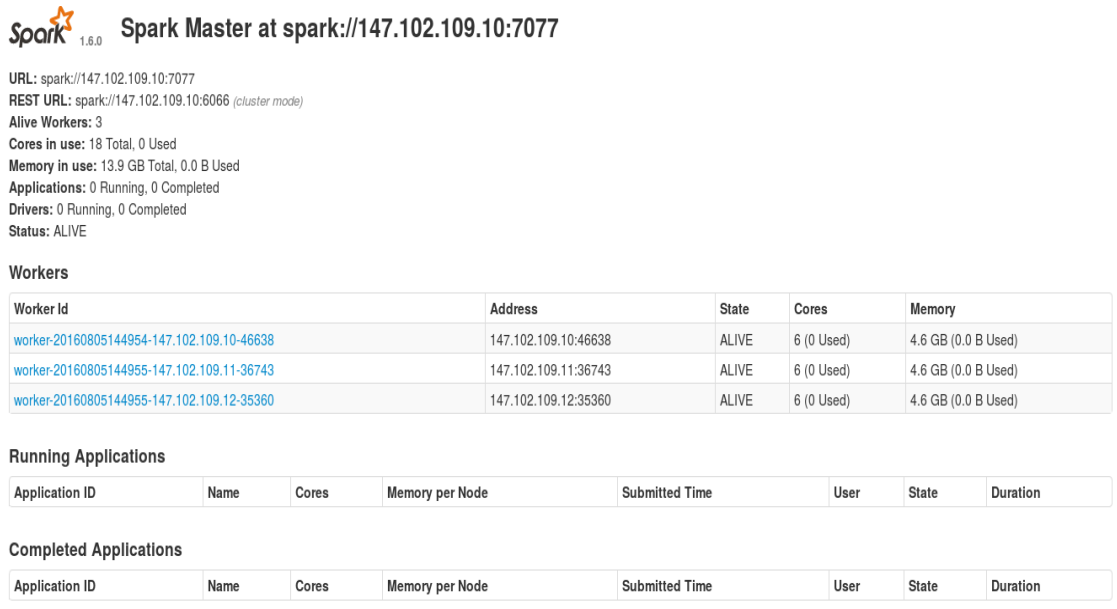
Ο YARN (ResourceManager + ApplicationsManager) λειτουργεί συνεπώς "έπάνω" από το σύστημα αρχείων (HDFS) και είναι το υπο-σύστημα εκείνο στο οποίο οι εφαρμογές χρηστών υποβάλλουν ερωτήματα επεξεργασίας πάνω στα αποθηκευμένα δεδομένα. Με ένα

σύστημα αρχείων το οποίο γνωρίζει ακριβώς σε πιο μηχάνημα είναι αποθηκευμένο κάθε αρχείο, ο ResourceManager γνωρίζει ποιος κόμβος περιέχει τα δεδομένα που πρέπει να αναλυθούν και ποιοι κόμβοι βρίσκονται - δικτυακά - κοντά σε αυτόν. Σε περίπτωση που κάποια εργασία δεν μπορεί να φιλοξενηθεί στον κόμβο στον οποίο βρίσκονται τα δεδομένα, δίνεται προτεραιότητα για την εκτέλεση της εργασίας σε κόμβους οι οποίοι βρίσκονται στο ίδιο rack. Αυτό μειώνει την κίνηση του δικτύου στο κύριο backbone δίκτυο του cluster. Εάν ένας NodeManager αποτύχει ή περάσει αρκετή ώρα χωρίς να επικοινωνήσει τότε εκείνο το τμήμα της εργασίας που απέτυχε επανεκτελείται. **Η κατάσταση του ResourceManager καθώς και του NodeManager εκτίθεται μέσω web browser και μπορεί να τη δεί σε πραγματικό χρόνο ο διαχειριστής του cluster (Σχήμα 4.7)**.

Apache Spark

Το Apache Spark[3] είναι ένα πλαίσιο ανοικτού κώδικα για τη διενέργεια υπολογισμών στο περιβάλλον ενός cluster υπολογιστών. Αρχικά αναπτύχθηκε στο AMPLab του Πανεπιστημίου Berkeley της Καλιφόρνια και στη συνέχεια ο πυρήνας του κώδικα δωρήθηκε στον οργανισμό Apache Software Foundation ο οποίος και έχει αναλάβει τη συντήρησή του έκτοτε. Το Spark παρέχει στους προγραμματιστές μια προγραμματιστική διεπαφή εφαρμογής για τον προγραμματισμό ολόκληρων clusters με υπονοούμενη παραλληλία δεδομένων και ανεκτικότητα σε σφάλματα.

Η προγραμματιστική διεπαφή που παρέχει το Spark στους χρήστες είναι επικεντρωμένη γύρω από μία δομή δεδομένων η οποία αποκαλείται Resilient Distributed Dataset (RDD) και μοντελοποιεί ένα πολυσύνολο αντικειμένων δεδομένων, προσβάσιμων μόνο για ανάγνωση, τα



Σχήμα 4.8: Στιγμιότυπο από τη διεπιφάνεια χρήστη για την παρακολούθηση της κατάστασης του cluster computing framework Spark.

[3]http://spark.apache.org/

οποία είναι κατανεμημένα κατά μήκος ενός cluster υπολογιστών το οποίο συντηρείται με ένα τρόπο ανεκτικό στις αστοχίες. Αναπτύχθηκε ως απάντηση στους περιορισμούς που εισάγει στο υπολογιστικό μοντέλο MapReduce, το οποίο επιβάλλει μια συγκεκριμένη γραμμική δομή στη ροή των δεδομένων στα κατανεμημένα προγράμματα. Τα προγράμματα γραμμένα με το μοντέλο MapReduce διαβάζουν δεδομένα εισόδου από τον δίσκο, εφαρμόζουν μια συνάρτηση κατά μήκος όλων των δεδομένων εισόδου (στάδιο Map) και στη συνέχεια εφαρμόζουν μια άλλη συνάρτηση (στάδιο Reduce) στο αποτέλεσμα του σταδίου Map και αποθηκεύουν τα αποτελέσματα του σταδίου Reduce στο δίσκο. Αυτό το μοτίβο είναι αρκετά δύσκαμπτο καθώς δεν μπορεί να εφαρμοστεί άμεση επεξεργασία στα αποτελέσματα του σταδίου Reduce αλλά θα πρέπει να γραφεί νέος κώδικας τύπου MapReduce και επιπλέον στερείται ταχύτητας καθώς όλα τα αποτελέσματα αποθηκεύονται στο δίσκο και όχι στην κύρια μνήμη με συνέπεια να έχουμε σημαντική καθυστέρηση ανάγνωσης σε περίπτωση που θέλουμε συνεχή/επαναληπτική επεξεργασία κάποιου συνόλου δεδομένων.

Σε αντίθεση, τα RDDs του Spark λειτουργούν ως ένα "ζωντανό" σύνολο εργασίας για τα κατανεμημένα προγράμματα τα οποία δεν υπόκεινται στους περιορισμούς του μοντέλου MapReduce και παρέχουν, σκοπίμως, μια κατανεμημένη μοιραζόμενη μνήμη περιορισμένου τύπου η οποία και αποθηκεύει προσωρινά (caches) τα αποτελέσματα στην κύρια μνήμη. Αυτή η διαθεσιμότητα των RDDs διευκολύνει την υλοποίηση τόσο επαναληπτικών αλγορίθμων, οι οποίοι επισκέπτονται το σύνολο δεδομένων τους πολλές φορές μέσα σε ένα βρόχο όσο και διαδραστική/εξερευνητική ανάλυση δεδομένων, δηλαδή, την επαναλαμβανόμενη εκτέλεση ερωτημάτων πάνω στα δεδομένα με ένα τρόπο όπως τα ερωτήματα στις βάσεις δεδομένων. Η καθυστέρηση των εφαρμογών γραμμένων στο Spark είναι πολλές τάξεις μεγέθους μικρότερη από τις ίδιες εφαρμογές υλοποιημένες στο Apache Hadoop μια δημοφιλής υλοποίηση του προγραμματιστικού μοντέλου MapReduce. Ανάμεσα στις τάξεις των επαναληπτικών αλγορίθμων οι οποίες υλοποιούνται ιδανικά στο Spark βρίσκεται και η κλάση των αλγορίθμων εκπαίδευσης για συστήματα μηχανικής μάθησης, τα οποία και έδωσαν την αρχική ώθηση για την ανάπτυξη του Spark.

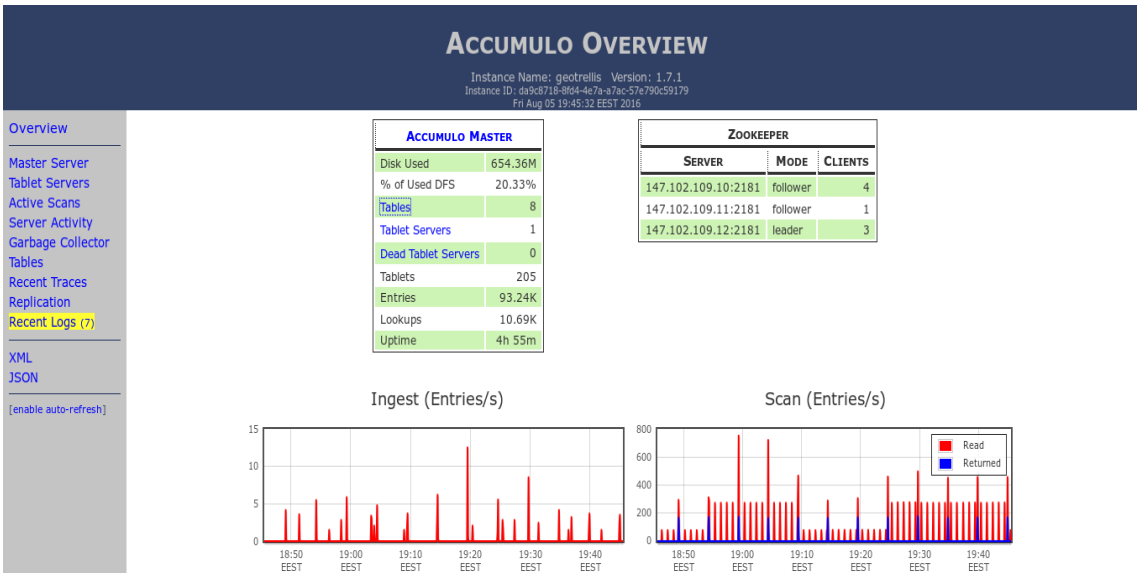
Το Spark απαιτεί για τη λειτουργία του έναν διαχειριστή των πόρων του cluster στο οποίο είναι εγκατεστημένο καθώς επίσης και ένα κατανεμημένο σύστημα αποθήκευσης. Σχετικά με τη διαχείριση των πόρων του cluster το Spark υποστηρίζεται είτε από την αυτόνομη διαχείριση του cluster από το ίδιο το Spark, είτε από τη μονάδα λογισμικού Hadoop YARN είτε από το σύστημα Apache Mesos. Για την κατανεμημένη αποθήκευση των δεδομένων το Spark έχει διεπαφές προς μια μεγάλη ποικιλία συστημάτων συμπεριλαμβανομένων του HDFS, του MapR File System (MapR-FS), του Cassandra, του OpenStack Swift, του Amazon S3, του Kudu και είναι επίσης δυνατόν να υλοποιηθεί μια προσαρμοσμένη λύση. Το Spark υποστηρίζει επίσης και μια ψευδο-κατανεμημένη τοπική κατάσταση, η οποία χρησιμοποιείται συνήθως μόνο για σκοπούς ανάπτυξης ή ελέγχου και στην οποία δεν απαιτείται κατανεμημένη αποθήκευση και το τοπικό σύστημα αρχείων μπορεί να χρησιμοποιηθεί εναλλακτικά. Σε ένα τέτοιο σενάριο το Spark τρέχει σε ένα μόνο μηχάνημα με δυνατότητα δρομολόγησης ενός μόνο νήματος εκτέλεσης ανά πυρήνα CPU.

Στην παρούσα εργασία χρησιμοποιήσαμε το σύστημα Apache Spark "έπάνω" από το σύ-

στημα Hadoop για να εκμεταλλευτούμε τις δυνατότητες (ευκολία στην εγκατάσταση, αποδεδειγμένη αξιοπιστία) και παροχές (Hadoop YARN, HDFS) του Hadoop στη διαχείριση ενός cluster υπολογιστών και να τις συνδυάσουμε με τις υπολογιστικές δυνατότητες του πλαισίου Spark για ταχύτατη επεξεργασία μεγάλου όγκου δεδομένων για τη διενέργεια analytics. Το Spark ακολουθεί μια αρχιτεκτονική master/slave στην οποία ένας κόμβος είναι ο master και οι υπόλοιποι (ή και ο master) είναι οι slaves - workers, στους οποίους και εκτελούνται οι υπολογισμοί. Από το **Σχήμα 4.2** μπορούμε να δούμε ότι στην αρχιτεκτονική του υλοποιημένου cluster στα πλαίσια της παρούσας εργασίας ο κόμβος Namenode είναι ταυτόχρονα και master και worker ενώ οι υπόλοιποι δύο κόμβοι (ReasourceManager, Datanode) είναι μόνο workers. Το Spark παρέχει επίσης μια διαδικτυακή διεπαφή μέσω της οποίας ο διαχειριστής του cluster μπορεί να παρακολουθεί σε πραγματικό χρόνο την κατάσταση των κόμβων του Spark, τη χρησιμοποίηση των πόρων του cluster και την πορεία των υπό εκτέλεση υπολογισμών (**Σχήμα 4.8**).

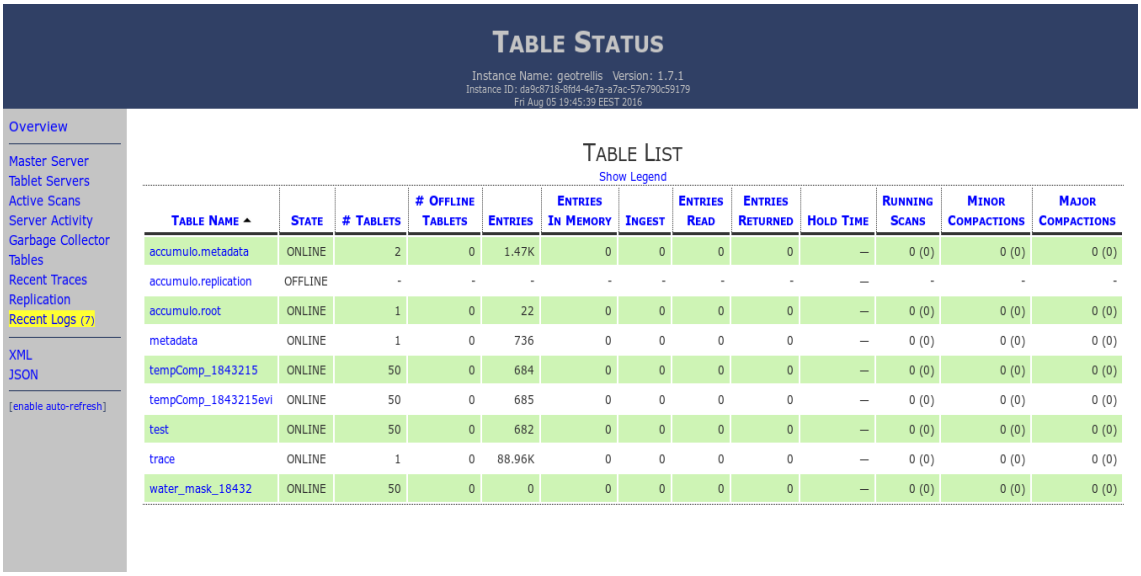
Accumulo

Το Apache Accumulo[4] είναι ένα project λογισμικού ηλεκτρονικών υπολογιστών το οποίο ανέπτυξε ένα οργανωμένο, κατανεμημένο key/value store το οποίο βασίζεται στην τεχνολογία BigTable που προέρχεται από την εταιρεία Google. Είναι ένα wide column store σύστημα διαχείρισης βάσεων δεδομένων το οποίο και αυτό χτίζεται πάνω από το Apache Hadoop, το Apache ZooKeeper και το Apache Thrift. Είναι υλοποιημένο στη γλώσσα προγραμματισμού Java και υποστηρίζει μηχανισμούς προγραμματισμού στο server-side. Το Accumulo είναι το τρίτο πιο δημοφιλές σύστημα NoSQL του τύπου wide column store σύμφωνα με την



Σχήμα 4.9: Στιγμιότυπο από τη διεπιφάνεια χρήστη για την παρακολούθηση της κατάστασης του συστήματος Accumulo.

[4]https://accumulo.apache.org/



Σχήμα 4.10: Στιγμιότυπο από τη διεπιφάνεια χρήστη για την παρακολούθηση της κατάστασης των αποθηκευμένων πινάκων του συστήματος Accumulo.

αξιολόγηση DB-Engines[5] πίσω από τα συστήματα Apache Cassandra και HBase κατά τον Οκτώβριο του 2016.

Στα πλαίσια της παρούσας εργασίας κρίθηκε αναγκαία η χρήση του στη στοίβα λογισμικού του ανεπτυγμένου συστήματος καθώς ως κατανεμημένο σύστημα αρχείων χρησιμοποιήθηκε το HDFS. Όπως αναφέρθηκε και σε προηγούμενη ενότητα το HDFS σχεδιάστηκε κυρίως για την εξυπηρέτηση αμετάβλητων αρχείων και δεν είναι κατάλληλο για συστήματα τα οποία απαιτούν ταυτόχρονες λειτουργίες εγγραφής. Για να αντιμετωπιστεί το ζήτημα αυτό κοινή πρακτική αποτελεί η χρήση επιπλέον λογισμικού τεχνολογίας BigTable, όπως είναι το Apache Accumulo, το οποίο λειτουργεί επάνω από το HDFS και λύνει το ζήτημα των συνεχών ενημερώσεων και εγγραφών των ίδιων αρχείων δεδομένων. Στα πλαίσια της παρούσας εργασίας είχαμε αυτό το ζήτημα καθώς οι ανεπτυγμένοι αλγόριθμοι επεξεργασίας ασκούν συνεχόμενα ερωτήματα επεξεργασίας πάνω στα δεδομένα, δημιουργούν πολλά ενδιάμεσα layers πληροφορίας και διαβάζουν και γράφουν συνεχώς στο δίσκο οπότε και για το λόγο αυτό χρησιμοποιήσαμε το σύστημα Accumulo.

Το Accumulo λειτουργεί με βάση και επάνω από το HDFS με σκοπό την γρήγορη αναζήτηση και ανάκτηση των αποθηκευμένων προς επεξεργασία δεδομένων. Χρησιμοποιείται στις περιπτώσεις στις οποίες η απόκριση σε σχεδόν πραγματικό χρόνο είναι αναγκαία καθώς το χαρακτηρίζει η εγγενής υλοποίηση του μοντέλου MapReduce, οι ταχύτερες σαρώσεις κατά μήκος των πινάκων, ο ευέλικτος ορισμός σχήματος και η καλύτερη ασφάλεια των δεδομένων.

Επιπροσθέτως, ο πυρήνας των υλοποιημένων αλγορίθμων μας αφορά layers πληροφορίας τα οποία ενημερώνονται συνεχώς κατά τη διάρκεια του κύκλου ζωής τους. Μέσω του HDFS δεν είναι δυνατή η ενημέρωση ενός layer πληροφορίας καθώς το HDFS είναι ένα copy on write σύστημα αρχείων και επιπλέον το HDFS δεν υποστηρίζει σε ικανοποιητικό βαθμό λει-

[5] http://db-engines.com/en/ranking

τουργικότητες CRUD (Create, Read, Update and Delete), σε αντίθεση με το Accumulo που επιλύει τέτοια θέματα σε μεγάλο βαθμό.

Συμπερασματικά, ήταν επιτακτική η ανάγκη χρήσης ενός συστήματος διαχείρισης βάσεων δεδομένων τεχνολογίας BigTable επάνω από το κατανεμημένο σύστημα αρχείων HDFS τόσο για λόγους αδυναμίας υποστήριξης της φύσης των εφαρμογών μας από το HDFS (συνεχόμενες ενημερώσεις layers πληροφορίας) όσο και για λόγους αυξημένης επίδοσης - απόκρισης του συστήματος. Ο λόγος που χρησιμοποιήθηκε το Accumulo και όχι κάποιο πιο δημοφιλές σύστημα τύπου column-store, π.χ. το Apache Cassandra, είναι επειδή η μηχανή επεξεργασίας μεγάλων γεωχωρικών δεδομένων Geotrellis (υπο-ενότητα 4.2.2) που χρησιμοποιήθηκε στα πλαίσια αυτής της εργασίας υποστήριζε καλύτερα από οποιοδήποτε άλλο παρόμοιο σύστημα το Accumulo την περίοδο υλοποίησης του ανεπτυγμένου συστήματος. Σε προσπάθεια περαιτέρω ανάπτυξης του ανεπτυγμένου συστήματος θα παρουσίαζε εξαιρετικό ενδιαφέρον η χρήση ενός διαφορετικού συστήματος από το Accumulo και η διενέργεια συγκρίσεων μεταξύ των δύο συστημάτων.

Το Accumulo παρέχει μια διαδικτυακή διεπαφή για την παρακολούθηση, από το διαχειριστή του cluster, της κατάστασης του συστήματος σε πραγματικό χρόνο (**Σχήματα 4.9 - 4.10**). Η διεπαφή παρέχει διάφορα στατιστικά και περιληπτικές πληροφορίες για τα αποθηκευμένα δεδομένα αλλά και πιο λεπτομερείς πληροφορίες σχετικά με την κατάσταση των αποθηκευμένων πινάκων.

Apache Zookeeper

Το Apache Zookeeper[6] είναι και αυτό ένα project λογισμικού του οργανισμού Apache Software Foundation. Είναι κατ΄ ουσίαν ένα κατανεμημένο, ιεραρχικό, key/value store το οποίο χρησιμοποιείται ώστε να παρέχει κατανεμημένες υπηρεσίες ρύθμισης, συγχρονισμού, συστήματος ενημέρωσης καθώς και μητρώου ονομάτων σε μεγάλα κατανεμημένα συστήματα όπως είναι το ανεπτυγμένο σύστημά μας υπό το Apache Accumulo. Το Zookeeper ξεκίνησε ως ένα τμήμα του Hadoop αλλά πλέον έχει εξελιχθεί σε ένα αυτόνομο υψηλού επιπέδου project.

Η αρχιτεκτονική του Zookeeper υποστηρίζει την υψηλή διαθεσιμότητα των παρεχόμενων υπηρεσιών μέσω της ύπαρξης πλεοναζώντων πόρων. Οι πελάτες του συστήματος μπορούν συνεπώς να ζητούν υπηρεσίες από κάποιον άλλον Zookeeper leader αν ο κύριος αποτύχει να ανταποκριθεί. Οι κόμβοι που τρέχουν το Zookeeper αποθηκεύουν τα δεδομένα τους σε έναν ιεραρχικό χώρο ονομάτων όπως συμβαίνει και με ένα σύστημα αρχείων ή μία δενδρική δομή δεδομένων. Οι πελάτες μπορούν να διαβάζουν και να γράφουν από/στους κόμβους και έτσι με αυτόν τον τρόπο επιτυγχάνεται μια κοινή υπηρεσία ρύθμισης. Οι ενημερώσεις των δεδομένων είναι απόλυτα διατάξιμες.

Το Zookeeper χρησιμοποιείται από πολλές μεγάλες εταιρείες συμπεριλαμβανομένων των Rackspace, Yahoo!, Reddit καθώς και του eBay όπως και από πολλά επιχειρησιακά συστήματα αναζήτησης ανοικτού κώδικα όπως για παράδειγμα είναι το Solr.

[6] https://zookeeper.apache.org/

Geotrellis

Το Geotrellis[7] αποτελεί μια καινοτόμα μηχανή επεξεργασίας γεωχωρικών δεδομένων για εφαρμογές υψηλών επιδόσεων σε ένα κατανεμημένο περιβάλλον ενός cluster υπολογιστών. Κατ΄ ουσίαν είναι μια βιβλιοθήκη γραμμένη στη γλώσσα προγραμματισμού Scala η οποία αφενός χρησιμοποιεί και επεκτείνει το μοντέλο δεδομένων του Spark (RDDs) για να επιτύχει την κατανεμημένη αποθήκευση και επεξεργασία γεωχωρικών δεδομένων και αφετέρου παρέχει μια πραγματικά μεγάλη εργαλειοθήκη για να διευκολύνει το έργο των προγραμματιστών για τη διενέργεια αναλύσεων των αποθηκευμένων δεδομένων και την εξαγωγή γνώσης από αυτά.

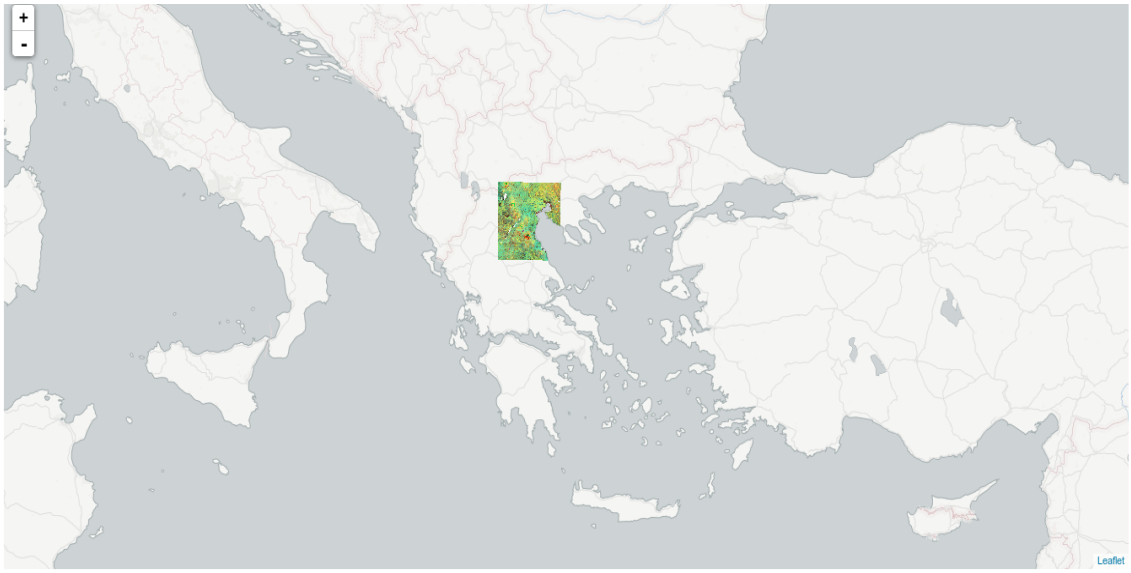
Η γεωχωρική επεξεργασία μέσω του Geotrellis οργανώνεται σε λειτουργίες. Ακολουθώντας την τυπική ονοματολογία της άλγεβρας χαρτών (Map Algebra), όπως αυτή ορίστηκε από τον Charles Dana Tomlin, οι λειτουργίες αυτές περιλαμβάνουν λειτουργικότητες και αναλύσεις του τύπου Local, Focal, Zonal και Global για raster δεδομένα καθώς και λειτουργικότητες για δίκτυα και vector δεδομένα. Πολλαπλές λειτουργίες μπορούν να συνθέσουν ένα «Model». Ένα γεωχωρικό μοντέλο στο πλαίσιο του Geotrellis αποτελείται από μικρότερες γεωχωρικές λειτουργίες με καλώς ορισμένες εισόδους και εξόδους. Το Geotrellis παρέχει επίσης μεγάλο αριθμό διαχειριστικών εργαλείων για τα raster δεδομένα συμπεριλαμβανομένων δυνατοτήτων cropping, μετασχηματισμού προβολής, κατασκευής μωσαϊκών, δημιουργία πυραμίδων και συμπίεση, κατάτμηση σε tiles, δημιουργία ευρετηρίων, πλήθος λειτουργιών Map Algebra, λειτουργίες μετατροπής vector δεδομένων σε raster, όπως είναι η διαδικασία Kernel Density καθώς επίσης και διανυσματοποίηση raster δεδομένων.

Όσον αναφορά στο μοντέλο δεδομένων του Geotrellis, αυτό επεκτείνει τα RDDs του framework Spark σε μια δομή δεδομένων που ονομάζεται RasterRDD[Key, Value]. Αυτή η δομή έχει όλες τις ιδιότητες που έχει η τυπική RDD δομή αλλά έχει επεκταθεί για να υποστηρίζει την κατανεμημένη επεξεργασία και αποθήκευση raster δεδομένων σε περιβάλλον cluster υπολογιστών. Αυτή η επέκταση λαμβάνει χώρα κατά την εισαγωγή νέων δεδομένων στο σύστημα. Σε αυτό το στάδιο μια αρχική εικόνα μέσω ενός προσαρμοσμένου partitioner "κόβεται" σε tiles και στο κάθε tile ανατίθεται ένα μοναδικό κλειδί (Key) ώστε να είναι δυνατόν να προσδιορίζεται από το σύστημα. Το κλειδί μπορεί να είναι είτε ένα χωρικό κλειδί (SpatialKey) το οποίο προκύπτει από το χωρικό εύρος του tile, είτε ένα χρονικό κλειδί (TemporalKey) το οποίο προκύπτει από την ημερομηνία λήψης των δεδομένων του tile, είτε ένα χωρο-χρονικό κλειδί (SpaceTimeKey) το οποίο προκύπτει ως συνδυασμός του χωρικού εύρους του tile και της ημερομηνίας λήψης των δεδομένων του tile. Μέσω του SpaceTimeKey είναι δυνατόν στα προγράμματά μας στη γλώσσα προγραμματισμού Scala να προσομοιάσουμε την κατασκευή layers πληροφορίας του τύπου πολυδιάστατων κύβων δεδομένων, όπως συμβαίνει στα Array συστήματα, με συνέπεια να είναι δυνατή η εύκολη και διαισθητική διενέργεια διαχρονικών αναλύσεων πάνω στα δεδομένα. Στη δομή RasterRDD[Key, Value] το ίδιο το tile παίρνει τη θέση της τιμής (Value). Με αυτόν τον τρόπο μια αρχική εικόνα είναι ένα σύνολο από RasterRDDs. Με αυτόν τον τρόπο τα tiles που συνθέτουν την εικόνα αποθηκεύονται σε ξεχωριστούς κόμβους του cluster και το σύστημα είναι υπεύθυνο να γνωρίζει

[7] http://geotrellis.io/

που βρίσκεται το κάθε tile που συνθέτει ένα layer πληροφορίας. Κατά την επεξεργασία το σύστημα κατανέμει τους υπολογισμούς μέσω του Spark στους κόμβους του συστήματος με σκοπό να επεξεργαστεί παράλληλα και ταυτόχρονα τα tiles τα οποία αφορά κάποιο ερώτημα επεξεργασίας/ανάλυσης των δεδομένων.

Σε αυτό το στάδιο πρέπει πλέον να είναι ξεκάθαρες οι σχεδιαστικές επιλογές για την υλοποίηση του ανεπτυγμένου συστήματος σε αυτήν την εργασία. Με μια οπτική από κάτω προς τα πάνω (bottom-up), για τη διαχείριση του cluster εικονικών μηχανών που είχαμε στη διάθεσή μας χρησιμοποιήσαμε το σύστημα Hadoop. Μέσω του συστήματος Hadoop είχαμε στη διάθεσή μας ένα κατανεμημένο σύστημα αρχείων πάνω από το cluster καθώς και έναν διαχειριστή πόρων - χρονοδρομολογητή εργασιών του cluster. Επάνω από αυτές τις δύο υπηρεσίες χρησιμοποιήσαμε από τη μία πλευρά το cluster computing framework Spark για την κατανομή των υπολογισμών πάνω από το cluster και την γρήγορη εκτέλεση των υπολογισμών σε αυτό καθώς και τα συστήματα Apache Accumulo - Zookeeper για να ξεπεράσουμε διάφορα μειονεκτήματα του HDFS τόσο όσον αναφορά στην ταχύτητα της ανάκτησης των αποθηκευμένων δεδομένων όσο και στις δυνατότητες των συνεχόμενων ενημερώσεων υπαρχόντων layers πληροφορίας στο σύστημα. Με όλη αυτήν την υποδομή εγκατεστημένη και διαθέσιμη ήταν δυνατόν να χρησιμοποιήσουμε τις δυνατότητες της μηχανής Geotrellis για την κατανεμημένη επεξεργασία γεωχωρικών δεδομένων στο περιβάλλον ενός cluster υπολογιστών μέσω προγραμμάτων γραμμένων στη γλώσσα προγραμματισμού Scala. Τα προγράμματα αυτά χρησιμοποιούν τις δομές δεδομένων του Spark για να μοντελοποιήσουν δεδομένα και υπολογισμούς και χρησιμοποιούν κατάλληλες συναρτήσεις για τη διαχείριση των δεδομένων μέσω του συστήματος Accumulo και τελικά του HDFS.



Σχήμα 4.11: Στιγμιότυπο της διεπιφάνειας χρήστη του Web Client για την online δημοσίευση των αποτελεσμάτων επεξεργασίας του ανεπτυγμένου συστήματος.

Web Client

Στα πλαίσια της παρούσας εργασίας αναπτύχθηκε ένα πολύ απλό πρόγραμμα πελάτη (Web Client) του ανεπτυγμένου συστήματος για την δημοσίευση των αποτελεσμάτων της επεξεργασίας μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών δια μέσου του Internet. Το πρόγραμμα πελάτη βασίζεται σε πολύ μεγάλο βαθμό στη βιβλιοθήκη Leaflet[8] για την παραγωγή διαδραστικών χαρτών, η οποία είναι γραμμένη στη γλώσσα προγραμματισμού javascript όπως επίσης και στη τυπική markup γλώσσα HTML για τη δημιουργία ιστοσελίδων. Η διεπαφή με το χρήστη είναι πάρα πολύ απλή, καθώς αυτός μπορεί μόνο να δει τα αποτελέσματα της επεξεργασίας που πραγματοποιείται μέσω του συστήματος. Στο **Σχήμα 4.11** παρουσιάζεται ένα στιγμιότυπο της διεπιφάνειας χρήστη του Web Client του ανεπτυγμένου συστήματος για την online δημοσίευση των αποτελεσμάτων της επεξεργασίας μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών.

4.3 Υλοποίηση - Περιγραφή αλγορίθμων

Στην παρούσα εργασία το κύριο κίνητρο ήταν η ανάλυση μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών για εφαρμογές παρατήρησης της γης. Ποικίλες συνιστώσες και υπολογιστικά βήματα εμπλέκονται στη ρύθμιση, στη λειτουργία και στη χρησιμοποίηση του ανεπτυγμένου συστήματος. Σε κάθε βήμα αυτής της εργασίας όλες οι σχεδιαστικές επιλογές καθώς και οι αποφάσεις όσον αναφορά στην υλοποίηση και στην επιλογή του κάθε υποσυστήματος λήφθηκαν υπό το πρίσμα της επίτευξης μιας εξαιρετικά κλιμακώσιμης λύσης από πλευράς αρχιτεκτονικής αλλά και μηχανής υψηλών επιδόσεων από πλευράς υλοποίησης. Η ενδελεχής μελέτη της διαθέσιμης βιβλιογραφίας, η συνεχής παρακολούθηση των τεχνολογικών εξελίξεων καθώς και ο προσωπικός πειραματισμός με εργαλεία διαχείρισης μεγάλων γεωχωρικών δεδομένων οδήγησαν στην τελική αρχιτεκτονική του ανεπτυγμένου συστήματος όπως αυτή παρουσιάστηκε αναλυτικά στην προηγούμενη ενότητα και στο **Σχήμα 4.1**.

Από πάνω προς τα κάτω (top-down) οι αλγόριθμοι επεξεργασίας των τηλεπισκοπικών δεδομένων υλοποιημένοι στη γλώσσα προγραμματισμού Scala, μέσω του Geotrellis υποβάλλονται στο Spark το οποίο αναλαμβάνει την κατανομή των υπολογισμών στους κόμβους του cluster, μέσω της υποβολής εργασιών στον Resource Manager Hadoop YARN. Επίσης γίνεται χρήση διάφορων συναρτήσεων του Geotrellis για τη διαχείριση των δεδομένων μέσω του συστήματος Accumulo και κατ επέκταση του κατανεμημένου συστήματος αρχείων HDFS. Η ίδια μεθοδολογία ισχύει αν αντί για αλγορίθμους επεξεργασίας έχουμε κώδικα ο οποίος είναι υπεύθυνος για την έναρξη ενός Web Server για τη δημοσίευση των αποτελεσμάτων των αλγορίθμων επεξεργασίας.

Εφοδιασμένοι με τη συγκεκριμένη αρχιτεκτονική και το τεκμηριωθέν πλαίσιο υλοποίησης, για την περάτωση της παρούσας εργασίας πραγματοποιήθηκε σε επίπεδο δημιουργίας εφαρμογών η ανάπτυξη διάφορων μονάδων λογισμικού στη γλώσσα προγραμματισμού Scala αρχικά για τις διαδικασίες της προ-επεξεργασίας και της εισαγωγής των τηλεπισκοπικών δεδομένων

[8] http://leafletjs.com/

στο σύστημα Geotrellis (ingestion module), στη συνέχεια για την επεξεργασία των αποθηκευμένων δεδομένων και τη διενέργεια αναλύσεων (processing module) και τέλος για τη δημοσίευση των αποτελεσμάτων μέσω του διαδικτύου (serve module).

Όσον αναφορά στο κομμάτι της προ-επεξεργασίας, όπως περιγράφηκε, τα διάφορα δεδομένα παρατήρησης της γης συλλέγονται με αυτοματοποιημένο τρόπο από εξωτερικά repositories. Μόλις ολοκληρωθεί η λήψη τους και γίνουν διαθέσιμα στο τοπικό σύστημα αρχείων πρέπει να υποστούν κατάλληλους μετασχηματισμούς και επεξεργασίες προκειμένου να κατανεμηθούν στους κόμβους του cluster και κατά συνέπεια να γίνουν διαθέσιμα από το Geotrellis ώστε να μπορούν να εφαρμοστούν σε αυτά ερωτήματα επεξεργασίας γραμμένα από τους χρήστες. Για το σκοπό αυτό, ένα συγκεκριμένο τμήμα λογισμικού (ingestion module) το οποίο και αναπτύχθηκε στη γλώσσα προγραμματισμού Scala είναι υπεύθυνο να ανεβάσει τα νεοληφθέντα δεδομένα στο HDFS και με την ολοκλήρωση της διαδικασίας αυτής να ξεκινήσει την εισαγωγή των δεδομένων (ingestion process) στο σύστημα Geotrellis. Το συγκεκριμένο τμήμα λογισμικού για κάθε εικόνα εισόδου πραγματοποιεί τις εξής διαδικασίες:

- Διάβασμα εικόνας από το HDFS
- Μετατροπή σε RasterRDD
- Διάβασμα μεταδεδομένων εικόνας και "κόψιμο" σε tiles σύμφωνα με την προδιαγραφή TMS.
- Επαναπροβολή στο επιθυμητό σύστημα αναφοράς
- Indexing με βάση χωρο-χρονικό κλειδί για τη δημιουργία προσομοιασμένων πολυδιάστατων κύβων δεδομένων
- Υπολογισμός τηλεπισκοπικών δεικτών οι οποίοι είναι αναγκαίοι για τη διενέργεια της ανάλυσής μας (NDVI, EVI2)
- Αποθήκευση ως layers στο Geotrellis - Accumulo για τη διενέργεια αναλύσεων

Παρακάτω βρίσκεται ένα απόσπασμα κώδικα από το ingestion module στο οποίο φαίνονται όλες οι παραπάνω διαδικασίες. Παρατηρούμε πόσο συνοπτικός και καθαρός είναι ο κώδικας, με συνέπεια να χρειάζονται πολύ λίγες γραμμές κώδικα ακόμα και για αρκετά περίπλοκες διαδικασίες.

```
// Ingestion module code snippet

// Διάβασμα εικόνας από το  HDFS
 val image_path = mergePaths(inputPath, new Path("/input_final.tif"))

 val imconf: Configuration = sc.hadoopConfiguration.withInputPath(image_path)
     TemporalGeoTiffInputFormat.setTimeTag(imconf, "TIFFTAG_DATETIME")
     TemporalGeoTiffInputFormat.setTimeFormat(imconf, "YYYY-MM-dd HH:mm:ss")
```

```
 val source: RDD[(TemporalProjectedExtent, MultibandTile)] =
       sc.newAPIHadoopRDD(imconf, classOf[TemporalMultibandGeoTiffInputFormat],
       classOf[TemporalProjectedExtent],
       classOf[MultibandTile])
// Διάβασμα μεταδεδομένων εικόνας και κόψιμο σε  tiles  σύμφωνα
// με την προδιαγραφή  TMS
 val (_, rasterMetaData) =
       TileLayerMetadata.fromRdd[TemporalProjectedExtent, MultibandTile, SpaceTimeKey]
       (source, FloatingLayoutScheme(512))
 val interm: RDD[(SpaceTimeKey, MultibandTile)] =
       source.tileToLayout(rasterMetaData.cellType, rasterMetaData.layout,
                           Bilinear)
 val tiled: RDD[(SpaceTimeKey, MultibandTile)] = interm.repartition(50)
 val layoutScheme = ZoomedLayoutScheme(WebMercator, tileSize = 256)

// Επαναπροβολή στο επιθυμητό σύστημα αναφοράς
 val (zoom, reprojected): (Int, RDD[(SpaceTimeKey, MultibandTile)]
       with Metadata[TileLayerMetadata[SpaceTimeKey]]) =
         MultibandTileLayerRDD(tiled, rasterMetaData).reproject(WebMercator,
           layoutScheme, Bilinear)

// Υπολογισμός τηλεπισκοπικών δεικτών  (NDVI, EVI2)
//  κρατώντας ταυτόχρονα τα μεταδεδομένα από τις εικόνες εισόδου

 val result = reprojected.withContext(_.mapValues { tile =>
     tile.convert(DoubleCellType).combineDouble(0, 1) { (r, ir) =>
       if(isData(r) && isData(ir) && (idx == "ndvi")) {
         val ndvi = (ir - r) / (ir + r)
         if(ndvi > 0.0) {ndvi} else {0.0}
       } else if (isData(r) && isData(ir) && (idx == "evi2")) {
         val evi2 = 2.5 * ((ir - r) / (ir + 2.4 * r + 1))
         if(evi2 > 0.0) {evi2} else {0.0}
       } else {
         //Double.NaN
         0.0
       }
     }
  })
```

Μόλις ολοκληρωθούν όλα τα βήματα προ-επεξεργασίας (ingestion module) και έχουν δημιουργηθεί οι κατάλληλες δομές στο σύστημα Geotrellis τα δεδομένα είναι σε κατάσταση έτοιμη για επεξεργασία οποτεδήποτε ζητηθεί από τους χρήστες. Για την επίδειξη των δυνατοτήτων του συστήματος και της ωφέλειας που παρέχει η κατανεμημένη ανάλυση σε σχέση με συμβατικές αναλύσεις επιλέξαμε να υλοποιήσουμε 2 αλγορίθμους διαχρονικής ανάλυσης ολόκληρων αρχείων τηλεπισκοπικών δεδομένων για την εξαγωγή χρήσιμης πληροφορίας και τη δημιουργία πρωτότυπων και διαισθητικών χαρτών.

Στην ενότητα αυτή (υπο-ενότητες 4.3.1 και 4.3.2) θα παρουσιαστούν τα ερωτήματα επεξεργασίας (processing module) των τηλεπισκοπικών δεδομένων που υλοποιήθηκαν, γραμμένα στη γλώσσα προγραμματισμού Scala, τα οποία και σχετίζονται με συγκεκριμένες αγροτικές εφαρμογές. Σε σχέση με την κάθε εφαρμογή θα παρουσιαστεί και το θεωρητικό υπόβαθρο πίσω από αυτή το οποίο επεξηγεί βασικά στοιχεία του κώδικα υλοποίησης του κάθε ερωτήματος.

Τέλος, μετά και την επεξεργασία των εικόνων μέσω των ανεπτυγμένων ερωτημάτων (processing module) χρησιμοποιείται επιλεκτικά κώδικας από το τμήμα λογισμικού υπεύθυνο για τη δημοσίευση των αποτελεσμάτων μέσω του διαδικτύου (serve module) για την έναρξη ενός web server αφοσιωμένου στο να εξυπηρετεί αιτήματα των χρηστών μέσω του Web Client για την προβολή των tiles των αποτελεσμάτων.

Ακολουθεί η ανάλυση των ανεπτυγμένων αλγοριθμικών διαδικασιών για την ανάλυση των αποθηκευμένων δεδομένων.

4.3.1 Παραγωγή διαχρονικών έγχρωμων σύνθετων ανα pixel

Η ελεύθερη και ανοικτή πρόσβαση στα δεδομένα τα οποία συλλέγονται εδώ και 40 χρόνια από την αποστολή Landsat καθώς και στα δεδομένα της πρόσφατης Ευρωπαϊκής αποστολής Sentinel σε συνδυασμό με τις βελτιώσεις στην προτυποποίηση των εικονιστικών προϊόντων όπως επίσης και στις ολοένα και αυξανόμενες δυνατότητες για υπολογισμούς και αποθήκευση έχουν δώσει τη δυνατότητα για την παροχή προϊόντων προστιθέμενης αξίας τα οποία δημιουργούνται από τη σύνθεση εικόνων ανά pixel (pixel-based image composites) χωρίς σύννεφα, ανακλαστικότητας στο έδαφος (surface reflectance) και τα οποία καλύπτουν πολύ μεγάλες περιοχές. Τέτοια εικονιστικά σύνθετα, τα οποία παράγονται με ξεχωριστή ανάλυση για κάθε pixel στο χρόνο (Temporal Compositing), αποτελούν ένα νέο μοτίβο στην τηλεπισκόπηση καθώς δεν βασίζονται πλέον στην τυπική ανάλυση ανά σκηνή. Μια διαχρονική παρουσίαση τέτοιων έγχρωμων σύνθετων δίνει τη δυνατότητα για καινοφανείς ευκαιρίες στην παραγωγή πληροφορίας η οποία θα χαρακτηρίζει την κάλυψη γης, τη δυναμική αλλαγή των καλύψεων γης καθώς και χαρακτηριστικά της δομής των δασών με ένα τρόπο ο οποίος είναι δυναμικός, διαφανής, συστηματικός, επαναλαμβανόμενος και θα εφαρμόζεται σε πάρα πολύ μεγάλες εκτάσεις.

Οι πληροφορίες σχετικά με τη μεταβαλλόμενη κατάσταση στην γήινη επιφάνεια πρέπει να δίνονται σε αρκετά καλές χωρικές αναλύσεις για την επίτευξη αποτελεσματικών μεθόδων διαχείρισης του περιβάλλοντος. Οι δορυφόροι Landsat και Sentinel παρέχουν δεδομένα τα οποία μπορούν να χρησιμοποιηθούν σε τέτοιου είδους μεθόδους, εφαρμογές και αναλύσεις. Παρ΄ όλα

αυτά η χρησιμοποίησή τους ενέχει αρκετές προκλήσεις. Η σύνθεση εικόνων με ξεχωριστή ανάλυση για κάθε pixel στο χρόνο παρέχει μεγάλες δυνατότητες για να ξεπεραστούν διάφορες ελλείψεις και δυσκολίες.

Στην υπο-ενότητα αυτή θα παρουσιάσουμε μια αλγοριθμική διαδικασία ποικίλων βημάτων για τη σύνθεση Temporal Composites τα οποία διευκολύνουν τη δημιουργία συνόλων δεδομένων από δεδομένα παρατήρησης της γης τα οποία είναι εποχιακά, χωρίς σύννεφα, ανκλαστικότητας εδάφους. Για την ανάλυση αυτή και τη σύνθεση των τελικών εικονιστικών προϊόντων χρησιμοποιήθηκαν οι τιμές των τηλεπισκοπικών δεικτών NDVI και EVI2, όπως αυτοί έχουν υπολογιστεί για κάθε αποθηκευμένη δορυφορική σκηνή στο σύστημά μας κατά το στάδιο της εισαγωγής των εικόνων στο σύστημα Geotrellis.

Οι τιμές του δείκτη NDVI (Normalized Difference Vegetation Index), ο οποίος είναι ο πιο ευρέως χρησιμοποιούμενος δείκτης βλάστησης, κυμαίνονται από -1 έως 1, με την τιμή -1 να υποδεικνύει την πλήρη απουσία βλάστησης και την τιμή 1 να υποδεικνύει την πυκνή παρουσία υγιούς βλάστησης. Η μαθηματική σχέση που οδηγεί στον υπολογισμό του δείκτη NDVI περιλαμβάνει δύο κανάλια του ηλεκτρομαγνητικού φάσματος, το υπέρυθρο (NIR) και το κόκκινο (Red) και είναι η ακόλουθη:

$$NDVI = \frac{NIR-Red}{NIR+Red}$$

Με τον υπολογισμό του δείκτη NDVI ενάντια σε μια τιμή κατωφλίωσης είμαστε σε θέση να ξεχωρίσουμε τη βλάστηση καθώς για μία δεδομένη τιμή κατωφλίωσης, όλα τα εικονοστοιχεία μιας εικόνας μιας περιοχής τα οποία έχουν τιμή στο δείκτη NDVI μεγαλύτερη από την τιμή κατωφλίωσης υποδεικνύουν την παρουσία βλάστησης με πολύ μεγάλη πιθανότητα.

Ο δείκτης EVI2, ο οποίος συνιστά έναν δείκτη EVI (Enhanced Vegetation Index) με μόνο 2 κανάλια του φάσματος να εμπλέκονται στον υπολογισμό του, αποτελεί έναν βελτιστοποιημένο δείκτη βλάστησης ο οποίος έχει σχεδιαστεί για να έχει μεγάλη απόκριση στην ανακλαστικότητα της βλάστησης σε περιοχές με σημαντική βιομάζα με ταυτόχρονη μείωση των ατμοσφαιρικών επιρροών. Η μαθηματική σχέση που οδηγεί στον υπολογισμό του δείκτη EVI2 περιλαμβάνει δύο κανάλια του ηλεκτρομαγνητικού φάσματος, το υπέρυθρο (NIR) και το κόκκινο (Red) και είναι η ακόλουθη:

$$EVI2 = 2.5 * \frac{NIR-Red}{NIR+2.4*Red+1}$$

Ενώ ο δείκτης NDVI είναι ευαίσθητος στην χλωροφύλλη των φυτών ο δείκτης EVI2 είναι περισσότερο ευαίσθητος στις δομικές διαφορές της κόμης φυλλώματος (canopy), συμπεριλαμβανομένων του Leaf Area Index (LAI), του τύπου της κόμης φυλλώματος καθώς και της αρχιτεκτονικής της. Οι δύο αυτοί δείκτες βλάστησης συμπληρώνουν ο ένας τον άλλον σε μελέτες της παγκόσμιας βλάστησης και παρέχουν ικανές απαντήσεις σε ερωτήματα σχετικά με την αλλαγή στην κατάσταση των καλλιεργειών καθώς και με την εξαγωγή βιοφυσικών παραμέτρων της κόμης φυλλώματος. Για τους λόγους αυτούς επιλέξαμε να χρησιμοποιήσουμε αυτούς τους δύο δείκτες ως την βάση των υλοποιημένων αλγοριθμικών διαδικασιών για την δημιουργία διαχρονικών σύνθετων ανά pixel στα πλαίσια της παρούσας εργασίας για εφαρμογές παρατήρησης της γης με μια εστίαση στις αγροτικές εφαρμογές.

Η ανίχνευση της βλάστησης και ο διαχωρισμός της από άλλα αντικείμενα και κατηγορίες του εδάφους είναι η πρωταρχική εργασία σε πολλές τηλεπισκοπικές εφαρμογές. Ωστόσο, δεν είναι πάντα εύκολη υπόθεση η αναγνώρισή της. Η χωρική ανομοιογένεια είναι μια σημαντική ιδιότητα των φυσικών τοπίων, η οποία και περιγράφει τη μεταβλητότητα των ιδιοτήτων της παρατηρούμενης επιφανείας στο χώρο. Οι απλές μέθοδοι ανίχνευσης της βλάστησης οι οποίες βασίζονται σε δείκτες βλάστησης καθώς και σε απλές πράξεις διαίρεσης μεταξύ διαφορετικών φασματικών ζωνών, οι οποίες πραγματοποιούνται ανά δορυφορική σκηνή, όχι μόνο δεν ξεχωρίζουν τους διαφορετικούς τύπους βλάστησης αλλά και μπερδεύουν διαφορετικούς τύπους επιφανειών.

Για το λόγο αυτό, στην παρούσα εργασία επιλέξαμε προς τη μελέτη των δυναμικών αλλαγών στην γήινη επιφάνεια, τη σύνθεση τεχνητών έγχρωμων σύνθετων τα οποία προκύπτουν από τη διαχρονική ανάλυση ξεχωριστά για κάθε pixel στην περιοχή και στο χρονικό διάστημα ενδιαφέροντος. Επειδή, ο κύριος προσανατολισμός μας είναι στις εφαρμογές γεωργίας ακριβείας επιλέξαμε τη σύνθεση αυτών των έγχρωμων σύνθετων με βάση τους τηλεπισκοπικούς δείκτες NDVI και EVI2. Έτσι, εργαστήκαμε προς την παραγωγή ενός τεχνητού έγχρωμου σύνθετου με 3 κανάλια πληροφορίας ήτοι κόκκινο, πράσινο και μπλε.

Για το κόκκινο κανάλι υπολογίστηκε για κάθε pixel της περιοχής ενδιαφέροντος, ο συντελεστής μεταβλητότητας (coefficient of variation, CV) των τιμών του κάθε pixel για τους τηλεπισκοπικούς δείκτες NDVI και EVI2, για το χρονικό διάστημα ενδιαφέροντος το οποίο ορίζεται από το χρήστη. Ο συντελεστής μεταβλητότητας αποτελεί ένα τυπικό μέτρο διασποράς μιας κατανομής συχνοτήτων στη θεωρία πιθανοτήτων και στατιστική. Εκφράζεται συχνά ως ποσοστό και ορίζεται ο λόγος της τυπικής απόκλισης σ προς τη μέση τιμή μ :

$$CV = \frac{\sigma}{\mu}$$

Ο συντελεστής μεταβλητότητας δείχνει το εύρος της μεταβλητότητας των τιμών του παρατηρούμενου μεγέθους (εδώ σύνολο τιμών για κάθε pixel του εκάστοτε τηλεπισκοπικού δείκτη) σε σχέση με τη μέση τιμή του υπό εξέταση πληθυσμού τιμών. Όσο μεγαλύτερη η μεταβλητότητα τόσο πιο πιθανό είναι να υπάρχει κάποια σημαντική μεταβολή (αλλαγή στην κάλυψη γης) στην υπό εξέταση περιοχή για παράδειγμα, στα πλαίσια της παρούσας εργασίας.

Για το πράσινο κανάλι υπολογίστηκε για κάθε pixel της περιοχής ενδιαφέροντος, η μέγιστη τιμή (max value) των τιμών του κάθε pixel για τους τηλεπισκοπικούς δείκτες NDVI και EVI2 για το χρονικό διάστημα ενδιαφέροντος το οποίο ορίζεται από το χρήστη. Η μέγιστη τιμή για τους δείκτες αυτούς δείχνει το μέγιστο σημείο του κύκλου ανάπτυξης της κόμης φυλλώματος των παρακολουθούμενων καλλιεργειών.

Για το μπλε κανάλι υπολογίστηκε για κάθε pixel της περιοχής ενδιαφέροντος, η μέση τιμή (mean value) των τιμών του κάθε pixel για τους τηλεπισκοπικούς δείκτες NDVI και EVI2 για το χρονικό διάστημα ενδιαφέροντος το οποίο ορίζεται από το χρήστη. Η μέση τιμή για τους δείκτες αυτούς δίνει μια ένδειξη για τον τύπο των παρατηρούμενων καλλιεργειών.

Ο σκοπός της διαδικασίας σύνθεσης λοιπόν είναι η πλήρωση του τελικού έγχρωμου σύνθετου με τις τιμές των 3 προαναφερθέντων καναλιών οι οποίες έχουν προκύψει από την διαχρονική ανάλυση των τιμών ανακλαστικότητας εδάφους των αποθηκευμένων δορυφορικών

δεδομένων. Παρακάτω περιγράφονται αναλυτικά τα βήματα της αλγοριθμικής διαδικασίας για την παραγωγή διαχρονικών έγχρωμων σύνθετων ανα pixel (Temporal Composites) σε ένα κατανεμημένο περιβάλλον ενός cluster υπολογιστών.

Όπως έχει ήδη αναφερθεί, κατά την εισαγωγή των δεδομένων στο σύστημα Geotrellis υπολογίζονται επί τόπου και αποθηκεύονται μαζί με τα raw δορυφορικά δεδομένα για κάθε εικόνα οι τιμές των τηλεπισκοπικών δεικτών NDVI και EVI2 σαν δύο επιπλέον layers πληροφορίας. Η βασική δομή του Geotrellis, δηλαδή τα RasterRDDs τα οποία δημιουργούνται επεκτείνοντας το μοντέλο δεδομένων του Spark χρησιμοποιεί key-value pairs για την αποθήκευση των δεδομένων με την τιμή στο ζεύγος να την αποτελεί το tile το οποίο αποθηκεύεται στο κατανεμημένο σύστημα αρχείων και περιέχει τις πραγματικές τιμές των δεδομένων (ενώ το κλειδί περιέχει μεταδεδομένα). Όπως εξηγήσαμε ήδη, κατά την εισαγωγή τα δεδομένα προς αποθήκευση κόβονται σε tiles. Έχουμε επιλέξει συγκεκριμένο τρόπο με τον οποίο κόβουμε τα tiles αυτά καθώς και τον τρόπο με τον οποίο δημιουργούμε μεταδεδομένα στο Geotrellis για αυτά, με συνέπεια να διασφαλίζουμε τόσο ότι κάθε δορυφορική παρατήρηση μιας συγκεκριμένης περιοχής της γήινης επιφάνειας αντιστοιχεί πάντα σε tile που έχει κοινά μεταδεδομένα χωρικής έκτασης με τα tiles που απεικονίζουν την ίδια περιοχή (σαν να έχουμε κάποιο κάναβο δηλαδή και να "χτίζουμε" tiles πάνω από αυτόν) όσο και ότι δημιουργούμε - μέσω των μεταδεδομένων - την ψευδαίσθηση κατασκευής ενός πολυδιάστατου κύβου δεδομένων για παρατηρήσεις οι οποίες απεικονίζουν την ίδια έκταση με συνέπεια να είναι δυνατή η εκτέλεση χωρο-χρονικών ερωτημάτων ανάλυσης πάνω στα αποθηκευμένα δεδομένα.

Με αυτόν τον τρόπο για τον υπολογισμό ενός διαχρονικού έγχρωμου σύνθετου ανα pixel (Temporal Composite) με βάση τον δείκτη NDVI (ακριβώς ίδια διαδικασία για τον δείκτη EVI2 και όποιον άλλο δείκτη) εκτελείται ένας κώδικας γραμμένος στη γλώσσα προγραμματισμού Scala ο οποίος έχει 5 βασικά στάδια. Στο πρώτο στάδιο, από τα δεδομένα εισόδου που δίνει ο χρήστης (path-row του δορυφόρου, δείκτης που θα χρησιμοποιηθεί, χρονικό διάστημα ενδιαφέροντος) ανακτάται, με τη βοήθεια του API του Geotrellis και του Accumulo, το σύνολο των δεδομένων το οποίο πρέπει να αναλυθεί για να προκύψει το τελικό αποτέλεσμα. Το σύνολο αυτό αποτελείται από όλα εκείνα τα tiles με τα μεταδεδομένα τους, τα οποία πληρούν τα κριτήρια του ερωτήματος ανάκτησης. Αφού έχει προσδιορισθεί το σύνολο δεδομένων επεξεργασίας σειρά παίρνουν τα βήματα της επεξεργασίας. Αρχικά, για κάθε pixel της τελικής εικόνας υπολογίζεται κατανεμημένα, παράλληλα και στον εκάστοτε κόμβο στον οποίο βρίσκονται τα δεδομένα (μεταφέρονται οι υπολογισμοί στα δεδομένα) η μέγιστη τιμή στο χρόνο για κάθε pixel της εικόνας για τις τιμές του ζητούμενου δείκτη με τη βοήθεια έτοιμων συναρτήσεων χαμηλού επιπέδου τις οποίες παρέχει το Geotrellis. Έπειτα, υπολογίζεται με την ίδια μεθοδολογία η μέση τιμή στο χρόνο για κάθε pixel της εικόνας για τις τιμές του ζητούμενου δείκτη και εδώ με τη βοήθεια έτοιμων συναρτήσεων χαμηλού επιπέδου τις οποίες παρέχει το Geotrellis . Έτσι, έχουν υπολογιστεί τα 2 πρώτα κανάλια της εικόνας του τελικού αποτελέσματος. Ο υπολογισμός του συντελεστή μεταβλητότητας λαμβάνει χώρα σε 2 βήματα καθώς για τον υπολογισμό του χρειαζόμαστε τόσο την τυπική απόκλιση όσο και τη μέση τιμή για κάθε pixel της εικόνας. Η μέση τιμή είναι διαθέσιμη από προηγούμενο στάδιο. Για την εύρεση της τυπικής απόκλισης χρησιμοποιούνται έτοιμες συναρτήσεις χαμηλού επιπέδου τις

οποίες παρέχει το Geotrellis και εκτελούνται κατανεμημένα και παράλληλα. Μόλις η τυπική απόκλιση γίνει διαθέσιμη πραγματοποιείται ο τελικός υπολογισμός για τον προσδιορισμό του συντελεστή μεταβλητότητας. Αξίζει να σημειωθεί ότι και για τα τρία κανάλια του αποτελέσματος, αφού πραγματοποιηθεί ο υπολογισμός τους, οι τιμές τους κανονικοποιούνται (image normalization) και πραγματοποιείται κατάλληλο φιλτράρισμα ώστε να τοποθετηθούν τιμές NoData όπου υπάρχει υδάτινο σώμα. Στη συνέχεια λαμβάνει χώρα το 4ο στάδιο κατά το οποίο πραγματοποιείται η σύνθεση, με τη βοήθεια των μεταδεδομένων των tiles που διατηρεί το Geotrellis, των 3 επιμέρους υπολογισμένων μεγεθών σε ένα ενιαίο σύνολο δεδομένων τύπου RasterRDD το οποίο περιλαμβάνει πολυδιάστατα tiles για να είναι δυνατή η προβολή τους ως ένα τεχνητό έγχρωμο σύνθετο. Κατά το πέμπτο στάδιο της εκτέλεσης προσδιορίζονται τα μεταδεδομένα του τελικού προϊόντος και πραγματοποιείται η εγγραφή του στο σύστημα ώστε να είναι δυνατή είτε η δημοσίευση του μέσω του διαδικτύου είτε η εξαγωγή του από το σύστημα σαν ένα geotiff για περαιτέρω ανάλυση μέσω λογισμικού GIS.

Ακολουθούν τα βασικά σημεία του κώδικα στη γλώσσα προγραμματισμού Scala του υλοποιημένου ερωτήματος για την παραγωγή διαχρονικών έγχρωμων σύνθετων ανα pixel (Temporal Composites):

```
// Υπολογισμός  Temporal Composite

// Διάβασμα  layer  εισόδου

 val irdd: RDD[(SpaceTimeKey, Tile)] with Metadata[TileLayerMetadata[SpaceTimeKey]] =
   reader.query[SpaceTimeKey, Tile, TileLayerMetadata[SpaceTimeKey]](layerId)
   .where(Between(time1, time2)).result

// Διάβασμα  layer  μάσκας
 val mask: RDD[(SpatialKey, Tile)] with Metadata[TileLayerMetadata[SpatialKey]] =
   reader.read(maskLayer)

// Υπολογισμός της μέγιστης τιμής στο χρόνο για κάθε  pixel  της εικόνας
// κατανεμημένα και ταυτόχρονα

 val maxValues: RDD[(SpatialKey, Tile)] =
   irdd.map { case (key, tile) =>
     // Get the spatial component of the SpaceTimeKey, which turns it into SpatialKey
     (key.getComponent[SpatialKey], tile)
   }
   // Now we have all the tiles that cover the same area with the same key.
   // Simply reduce by the key with a localMax
   .reduceByKey(_.localMax(_))
```

```
 val normMax:  RDD[(SpatialKey, Tile)] =
   maxValues.map { case (key, tile) =>
     (key, tile.normalize(0.0, 1.0, 0.0, 255.0).convert(IntConstantNoDataCellType))
   }

 var combined = normMax.cogroup(mask)
 val maskedNormMax: RDD[(SpatialKey, Tile)] =
   combined.map{ case(key,tiles) =>
     val c = tiles._1.head;
     val d = tiles._2.head;
     (key, (c * d).convert(IntConstantNoDataCellType))
   }

// Υπολογισμός της μέσης τιμής στο χρόνο για κάθε  pixel  της εικόνας
// κατανεμημένα και ταυτόχρονα

 val meanValues: RDD[(SpatialKey, Tile)]=
   irdd.map { case (key, tile) =>
     (key.getComponent[SpatialKey], tile)
   }.groupBy(_._1)
   .map{ case(key,tiles) =>
     val localMean =
       tiles
       .map(_._2) // Maps to the 2nd element of the tuple, i.e. the tiles
       .toSeq
       .localMean
     (key, localMean)
   }

 val normMean:  RDD[(SpatialKey, Tile)] =
   meanValues.map { case (key, tile) =>
     (key, tile.normalize(0.0, 1.0, 0.0, 255.0).convert(IntConstantNoDataCellType))
   }

 combined = normMean.cogroup(mask)
 val maskedNormMean: RDD[(SpatialKey, Tile)] =
   combined.map{ case(key,tiles) =>
     val c = tiles._1.head;
     val d = tiles._2.head;
     (key, (c * d).convert(IntConstantNoDataCellType))
   }
```

```
// Υπολογισμός της τυπικής απόκλισης στο χρόνο για κάθε  pixel  της εικόνας
// κατανεμημένα και ταυτόχρονα

 val sdValues: RDD[(SpatialKey, Tile)]=
   irdd.map { case (key, tile) =>
     (key.getComponent[SpatialKey], tile)
   }.groupBy(_._1)
   .map{ case(key,tiles) =>
     val localVariance =
       tiles
       .map(_._2) // Maps to the 2nd element of the tuple, i.e. the tiles
       .toSeq
       .localVariance
     (key, Sqrt(localVariance))
  }

// Υπολογισμός του συντελεστή μεταβλητότητας στο χρόνο για κάθε  pixel  της εικόνας
// κατανεμημένα και ταυτόχρονα

 val cv = sdValues/meanValues

 val normCV:  RDD[(SpatialKey, Tile)] =
   cv.map { case (key, tile) =>
     (key, tile.normalize(0.0, 1.0, 0.0, 255.0).convert(IntConstantNoDataCellType))
   }

 combined = normCV.cogroup(mask)
 val maskedNormCV: RDD[(SpatialKey, Tile)] =
   combined.map{ case(key,tiles) =>
     val c = tiles._1.head;
     val d = tiles._2.head;
     (key, (c * d).convert(IntConstantNoDataCellType))
   }

// Σύνθεση του διαχρονικού έγχρωμου σύνθετου της περιοχής ενδιαφέροντος
// για το χρονικό διάστημα ενδιαφέροντος κατανεμημένα

 val b3 = maskedNormCV.cogroup(maskedNormMax).cogroup(maskedNormMean)
```

```
val temporalComposite: RDD[(SpatialKey, MultibandTile)] =
  b3.map{ case(key,tiles) =>
    val c = tiles._1.head._1.head;
    val d = tiles._1.head._2.head;
    val e=tiles._2.head;
    (key, MultibandTile(Traversable(c, d, e)))
  }

var meta = irdd.metadata
val min = meta.bounds.get.minKey.getComponent[SpatialKey]
val max = meta.bounds.get.maxKey.getComponent[SpatialKey]
val bounds: KeyBounds[SpatialKey] = {
  KeyBounds(min, max)
}
val new_metadata =
  new TileLayerMetadata(meta.cellType, meta.layout, meta.extent, meta.crs, bounds)

val result: RDD[(SpatialKey, MultibandTile)] with Metadata[TileLayerMetadata[SpatialKey]] =
  MultibandTileLayerRDD(temporalComposite, new_metadata)

// Write layer
val layoutScheme = ZoomedLayoutScheme(WebMercator, tileSize = 256)
Pyramid.upLevels(result, layoutScheme, 13) { (rdd, z) =>
  val newLayer = LayerId(outputLayerName, z)
  if(attributeStore.layerExists(newLayer)) {
    deleter.delete(newLayer)
  }
  writer.write(newLayer, rdd, ZCurveKeyIndexMethod)
}
```

4.3.2 Αποτύπωση της εποχικότητας

Το ερώτημα για την παραγωγή χαρτών αποτύπωσης της εποχικότητας είναι σημαντικά πιο απλό από το ερώτημα της παραγωγής Temporal Composites τόσο από πλευράς υλοποίησης όσο και από πλευράς θεωρητικού υποβάθρου και ερμηνείας του αποτελέσματος. Με τον όρο αποτύπωση της εποχικότητας αναφερόμαστε στη μέρα του έτους για την οποία επικρατεί κάποια συγκεκριμένη συνθήκη/συνθήκες σε μία περιοχή και ενδιαφερόμαστε να τη γνωρίζουμε για να προχωρήσουμε σε διάφορες αναλύσεις.

Στην παρούσα εργασία, και με μια εστίαση σε αγροτικές εφαρμογές, με τον όρο αποτύπωση της εποχικότητας ορίζουμε το ενδιαφέρον μας ώστε να γνωρίζουμε τη μέρα του έτους

για την οποία σε μια περιοχή (pixel) συμβαίνει ο δείκτης NDVI (ή ο δείκτης EVI2) να παίρνει τη μέγιστη τιμή του. Με την εξαγωγή μιας τέτοιας πληροφορίας είμαστε σε θέση να εξάγουμε μοτίβα για την ανάπτυξη παρακολουθούμενων καλλιεργειών, να παρατηρούμε ομοιότητες, να διαχωρίζουμε ανομοιογένειες και να παρακολουθούμε επακριβώς την κατάσταση που επικρατεί σε μια περιοχή που καλύπτει ένα συγκεκριμένο pixel για κάθε κύκλο ανάπτυξης των καλλιεργειών.

Επομένως, στο υλοποιημένο ερώτημα για την αποτύπωση της εποχικότητας στα πλαίσια της παρούσας εργασίας, στόχος μας ήταν για κάθε path-row (ή tile) των δορυφόρων, δεδομένα των οποίων αποθηκεύονται στο σύστημά μας, να εντοπίσουμε μέσω της διαχρονικής ανάλυσης όλων των σκηνών κάθε path-row (ή tile) τη μέρα του έτους για την οποία συμβαίνει ο δείκτης NDVI (ή ο δείκτης EVI2) να παίρνει τη μέγιστη τιμή του σε κάθε pixel που περιέχει ένα συγκεκριμένο path-row (ή tile) και αφού γίνει ο εν λόγω εντοπισμός να παράγεται ο αντίστοιχος χάρτης. Παρακάτω περιγράφονται αναλυτικά τα βήματα της αλγοριθμικής διαδικασίας για τον υπολογισμό της εποχικότητας σε ένα κατανεμημένο περιβάλλον ενός cluster υπολογιστών.

Η γενικότερη φιλοσοφία της διαδικασίας είναι παρόμοια με την παραγωγή διαχρονικών έγχρωμων σύνθετων όσον αναφορά τόσο στον προσδιορισμό του συνόλου δεδομένων της επεξεργασίας όσο και στη δημιουργία των μεταδεδομένων και στην εγγραφή του τελικού προϊόντος. Η βασική διαφορά έγκειται στο ότι η αλγοριθμική διαδικασία για την παραγωγή των διαχρονικών έγχρωμων σύνθετων υποβάλλεται ως μία ενιαία εργασία για εκτέλεση στο διαθέσιμο cluster υπολογιστών. Για τον υπολογισμό της εποχιακής κατάλληλότητας ακολουθήθηκε μια διαφορετική προσέγγιση τόσο λόγω των απαιτήσεων του ερωτήματος αλλά και της εκφραστική δύναμης της γλώσσας Scala. Η προσέγγιση διαφέρει στο ότι κάθε διαφορετικό στάδιο της επεξεργασίας, εδώ, υποβάλλεται ως μια ξεχωριστή εργασία για εκτέλεση στο cluster με συνέπεια να επιτυγχάνουμε ακόμα μεγαλύτερη παραλληλία και κατά συνέπεια επίδοση.

Για να προσδιοριστεί η μέρα του έτους για την οποία συμβαίνει ο δείκτης NDVI (ή ο δείκτης EVI2) να παίρνει τη μέγιστη τιμή του, σε κάθε pixel, αρχικά είναι αναγκαίο με μια εργασία να υπολογιστεί η μέγιστη τιμή των τιμών του τηλεπισκοπικού δείκτη για κάθε pixel στο έτος ενδιαφέροντος το οποίο έχει ζητήσει ο χρήστης του συστήματος. Στη συνέχεια, αφού έχει προσδιορισθεί το layer με τις μέγιστες τιμές, για κάθε εικόνα που υπάρχει διαθέσιμη για το έτος ενδιαφέροντος και για την περιοχή ενδιαφέροντος υποβάλλεται μια εργασία σύγκρισης της εικόνας με το layer της μέγιστης τιμής ώστε για κάθε εικόνα να παραχθεί ένα ενδιάμεσο προϊόν το οποίο έχει τιμή 0 αν η τιμή του pixel δεν είναι ίση με τη μέγιστη τιμή ή την τιμή της μέρας του έτους που αναζητούμε αν η τιμή του pixel είναι ίση με τη μέγιστη τιμή. Στο στάδιο αυτό όλες οι υποβαλλόμενες εργασίες μπορούν να εκτελούνται παράλληλα αν υπάρχουν διαθέσιμοι πυρήνες επεξεργασίας και μνήμη καθώς όλες οι εργασίες είναι ανεξάρτητες μεταξύ τους. Στο τελικό στάδιο, μια εργασία αναλαμβάνει την συγχώνευση όλων των ενδιάμεσων προϊόντων ήτοι να διαβάσει όλα τα ενδιάμεσα προϊόντα και να προσθέσει τις τιμές τους ώστε να προκύψει το τελικό προϊόν του χάρτη αποτύπωσης της εποχικότητας που αναζητούμε. Και εδώ στο τελικό προϊόν εφαρμόζεται κατάλληλη μάσκα για την απομάκρυνση των υδάτινων σωμάτων από το τελικό αποτέλεσμα.

Ακολουθούν τα βασικά σημεία του κώδικα στη γλώσσα προγραμματισμού Scala του υλο-

ποιημένου ερωτήματος για τον υπολογισμό του χάρτη αποτύπωσης της εποχικότητας:

```
// Υπολογισμός χάρτη αποτύπωσης της εποχικότητας

// Υπολογισμός της μέγιστης τιμής για τον εκάστοτε τηλεπισκοπικό δείκτη για κάθε
// pixel  της εικόνας κατανεμημένα και ταυτόχρονα

 val maxValues: RDD[(SpatialKey, Tile)] =
   irdd.map { case (key, tile) =>
     // Get the spatial component of the SpaceTimeKey, which turns it into SpatialKey
     (key.getComponent[SpatialKey], tile)
   }
   // Now we have all the tiles that cover the same area with the same key.
   // Simply reduce by the key with a localMax
   .reduceByKey(_.localMax(_))

 val ppmaxValues : RDD[(SpatialKey, Tile)] =
   maxValues.map { case (key, tile) =>
     (key, IfCell(tile, { (x: Double) => x == 0.0 }, Double.NaN, 1.0) * tile)
   }

 // Σύγκριση των τιμών κάθε εικόνας στο χρονικό διάστημα ενδιαφέροντος
 // με τη μέγιστη τιμή ώστε να βρούμε για κάθε  pixel  τη μέρα
 // του χρόνου που συμβαίνει η μέγιστη τιμή

 // Ανάκτηση τιμών τηλεπισκοπικού δείκτη για μια συγκεκριμένη ημερομηνία
 val irdd: RDD[(SpaceTimeKey, Tile)] with Metadata[TileLayerMetadata[SpaceTimeKey]] =
   reader.query[SpaceTimeKey, Tile, TileLayerMetadata[SpaceTimeKey]](layerId)
     .where(Between(time2, time3)).result
   // construct a RDD[(SpatialKey, Tile)] from input
   val in: RDD[(SpatialKey, Tile)] =
     irdd.map { case (key, tile) =>
       (key.getComponent[SpatialKey], tile)
     }

 // Σύγκριση με τη μέγιστη τιμή για να πάρουμε τη μέρα του χρόνου που
 // συμβαίνει η μέγιστη τιμή

 val combined = maxValues.cogroup(in)

 val mapDOY: RDD[(SpatialKey, Tile)] =
```

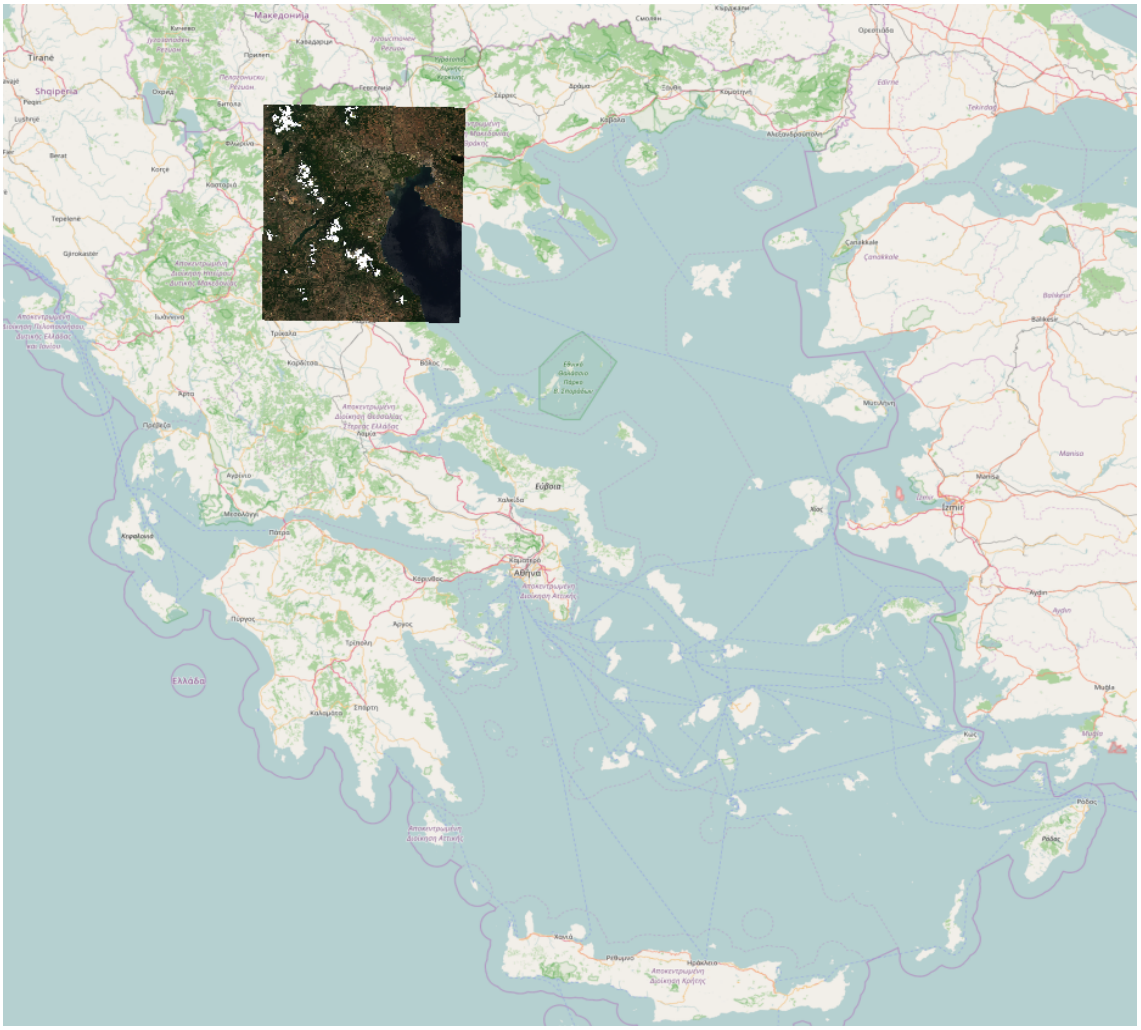
```
  combined.map{ case(key,tiles) =>
    val c = tiles._1.head;
    val d = tiles._2.head;
    (key, IfCell(c, d, { (x: Double, y: Double) => x == y }, dvalue, 0.0))
  }

// Συγχώνευση όλων των αποτελεσμάτων για τη σύνθεση του τελικού χάρτη

var resrdd: RDD[(SpatialKey, Tile)] with Metadata[TileLayerMetadata[SpatialKey]] =
  reader.read[SpatialKey, Tile, TileLayerMetadata[SpatialKey]](layerId)

for( i <- 2 to inlayers_num) {
  val layerId = LayerId(layerName + i.toString, 13)
  val irdd: RDD[(SpatialKey, Tile)] with Metadata[TileLayerMetadata[SpatialKey]] =
    reader.read[SpatialKey, Tile, TileLayerMetadata[SpatialKey]](layerId)

  val combined = resrdd.cogroup(irdd)
  val interm: RDD[(SpatialKey, Tile)] =
    combined.map{ case(key,tiles) =>
      val c = tiles._1.head;
      val d = tiles._2.head;
      (key, c + d)
    }

  resrdd = TileLayerRDD(interm, resrdd.metadata)
}
// Μασκάρισμα του αποτελέσματος για να απομακρυνθούν οι υδάτινες επιφάνειες

val mask: RDD[(SpatialKey, Tile)] with Metadata[TileLayerMetadata[SpatialKey]] =
  reader.read(maskLayer)
val combined = resrdd.cogroup(mask)
val maskedresrdd: RDD[(SpatialKey, Tile)] =
  combined.map{ case(key,tiles) =>
    val c = tiles._1.head;
    val d = tiles._2.head;
    (key, (c * d).convert(DoubleConstantNoDataCellType))
  }
val maskedOut: RDD[(SpatialKey, Tile)] with  Metadata[TileLayerMetadata[SpatialKey]] =
  TileLayerRDD(maskedresrdd, resrdd.metadata)
```

Κεφάλαιο 5

Αποτελέσματα και Αξιολόγηση

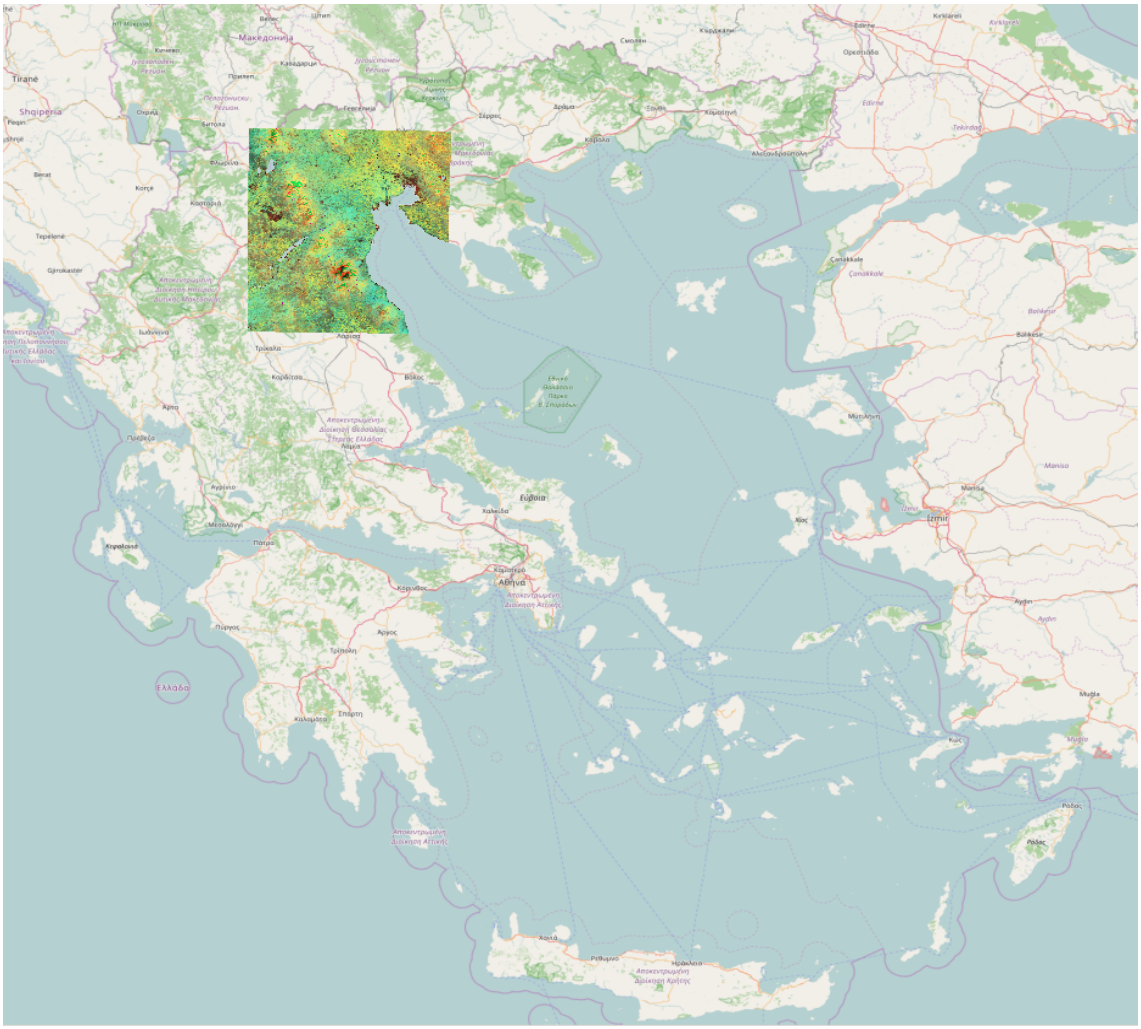
Στο κεφάλαιο αυτό θα παρουσιασθούν τα αποτελέσματα των υλοποιημένων αλγορίθμων στην ανεπτυγμένη υποδομή του συστήματος μας, για την παραγωγή παραγωγή διαχρονικών έγχρωμων σύνθετων ανα pixel όπως επίσης και την αποτύπωση της εποχικότητας για κάθε pixel μιας περιοχής ενδιαφέροντος. Ως περιοχή ενδιαφέροντος επιλέχθηκε να είναι μια σκηνή του δορυφόρου Landsat 8 και συγκεκριμένα αυτή που ορίζεται από το path-row 184/032. Τα αποτελέσματα που παρουσιάζονται στις επόμενες ενότητες προέκυψαν από τη διαχρονική ανάλυση όλων των εικόνων για τη σκηνή αυτή κατά τα έτη 2014 - 2015. Επιλέχθηκαν δύο χρονικά διαστήματα ενδιαφέροντος. Το διάστημα το οποίο ορίζεται από 1/1/2014 - 31/12/2014 (έτος 2014) και το διάστημα το οποίο ορίζεται από 1/1/2015 - 31/12/2015 (έτος 2015). Η σκηνή για την οποία πραγματοποιείται η ανάλυση καθώς και το χρονικό διάστημα ενδιαφέροντος δίνονται σαν ορίσματα της επεξεργασίας από το χρήστη του συστήματος. Αυτό έχει σαν συνέπεια να είναι δυνατή η επεξεργασία οποιασδήποτε σκηνής από το ανεπτυγμένο σύστημα για οποιοδήποτε διάστημα ενδιαφέροντος. Επιλέχθηκε το ένα έτος ως μεμονωμένο διάστημα ενδιαφέροντος με σκοπό να δημιουργηθούν όσο το δυνατόν πιο βαριές υπολογιστικά εργασίες (επιλογή των περισσότερων δυνατών εικόνων για ανάλυση) με την απαίτηση οι εργασίες να έχουν νόημα (σε ένα έτος μπορούμε να μελετήσουμε τον κύκλο ζωής των καλλιεργειών αλλά αν επιλέγαμε μεγαλύτερο χρονικό διάστημα τα αποτελέσματα των αναλύσεων δεν θα είχαν κάποιο νόημα).

Πριν προχωρήσουμε στην παρουσίαση των αποτελεσμάτων ας επαναλάβουμε εν συντομία κάποια κρίσιμα χαρακτηριστικά της αρχιτεκτονικής του ανεπτυγμένου συστήματος τα οποία έχουν ιδιαίτερη σημασία όσον αναφορά στην αξιολόγηση των επιδόσεων του. Την καρδιά του cluster υπολογιστών το οποίο αποτελεί τον πυρήνα του ανεπτυγμένου συστήματος δομούν 3 εικονικά μηχανήματα συνδεδεμένα δικτυακά μεταξύ τους. Κάθε ένα από αυτά τα μηχανήματα έχει **1 πυρήνα επεξεργασίας (CPU), 6 GB RAM καθώς και 50 GB συνολικό αποθηκευτικό χώρο** (με περίπου 30 GB να μένουν διαθέσιμα μετά από το χώρο που καταλαμβάνει το λειτουργικό σύστημα και τα διάφορα άλλα προγράμματα). Σε αυτό το πειραματικό περιβάλλον οι επιδόσεις που πετυχαίνει το ανεπτυγμένο σύστημα δεν είναι καθόλου άσχημες (το αντίθετο συμβαίνει). Κάθε νέα πολυφασματική εικόνα η οποία φθάνει στο σύστημα υφίσταται όλα τα στάδια της προ-επεξεργασίας όπως αυτά περιγράφηκαν στο κεφάλαιο 4 και



Σχήμα 5.1: Φυσικό έγχρωμο σύνθετο της **περιοχής ενδιαφέροντος (path-row: 184_032 του δορυφόρου Landsat 8)** για την οποία πραγματοποιήθηκαν οι ανεπτυγμένες διαχρονικές αναλύσεις ανά pixel οι οποίες και παρουσιάζονται στο κεφάλαιο αυτό. (Ιούνιος 2015)

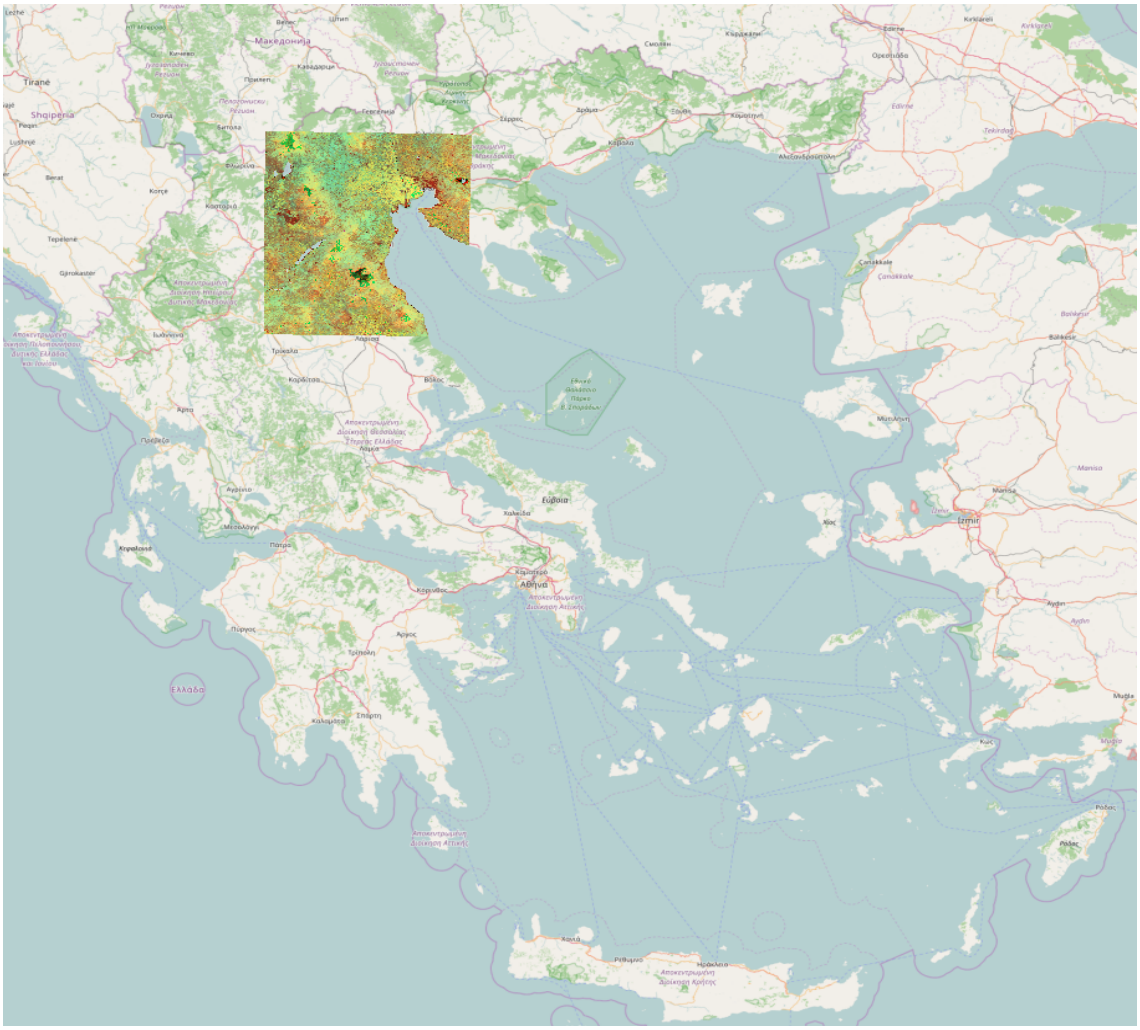
εισάγεται στο σύστημα μέσα σε 3 μόλις λεπτά. Κάθε διαχρονικό έγχρωμο σύνθετο το οποίο προκύπτει από την ανάλυση όλων των εικόνων ενός έτους ανά pixel παράγεται για μία σκηνή του δορυφόρου Landsat 8 σε 15 λεπτά. Και κάθε χάρτης αποτύπωσης της εποχικότητας ο οποίος προκύπτει από όμοιων απαιτήσεων ανάλυση παράγεται σε περίπου 10 λεπτά (5 λεπτά για να υπολογιστεί η μέγιστη τιμή στο χρόνο για κάθε pixel, 2 λεπτά χρειάζεται η κάθε εικόνα για κάθε pixel για να συγκριθεί με τη μέγιστη τιμή ενώ σε αυτό το στάδιο η κάθε εργασία μπορεί να εκτελείται ταυτόχρονα με τις υπόλοιπες και τέλος 3 λεπτά χρειάζονται για να συγχωνευθούν τα επιμέρους αποτελέσματα για κάθε εικόνα στον τελικό χάρτη). Οι επιδόσεις αυτές είναι ιδιαίτερα εντυπωσιακές αν αναλογιστεί κανείς ότι τα προτεινόμενα χαρακτηριστικά και ρυθμίσεις **για ένα τέτοιο cluster υπολογιστών απαιτούν 4 φυσικά μηχα-**



Σχήμα 5.2: Temporal Composite (TC) της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2014**.

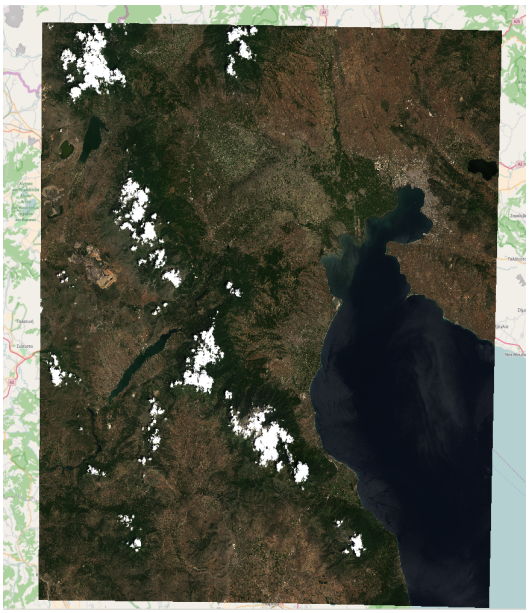
νήματα με το καθένα να έχει 4 ή 8 πυρήνες επεξεργασίας, τουλάχιστον 16 GB RAM και περισσότερα από 500 GB συνολικό αποθηκευτικό χώρο. Με δεδομένη την κλιμακωσιμότητα του συστήματος είμαστε σε θέση να ισχυριστούμε ότι με βελτίωση της υποδομής στα χαρακτηριστικά που αναφέρθηκαν οι επιδόσεις του συστήματος μπορούν να αυξηθούν τουλάχιστον 4 φορές ή και περισσότερο ανάλογα με τις αναβαθμίσεις σε υλικό οι οποίες θα προκριθούν.

Αξίζει να σημειωθεί ότι από την εμπειρία που αποκτήθηκε στις αναλύσεις μεγάλων δεδομένων οδηγηθήκαμε στο συμπέρασμα ότι η διενέργεια τέτοιου είδους αναλύσεων σε συμβατικά υπολογιστικά περιβάλλοντα θα ήταν τρομερά δυσχερής. Χωρίς το προγραμματιστικό μοντέλο του Geotrellis στη φαρέτρα μας και την κατανεμημένη υποδομή Hadoop θα έπρεπε να συντο-

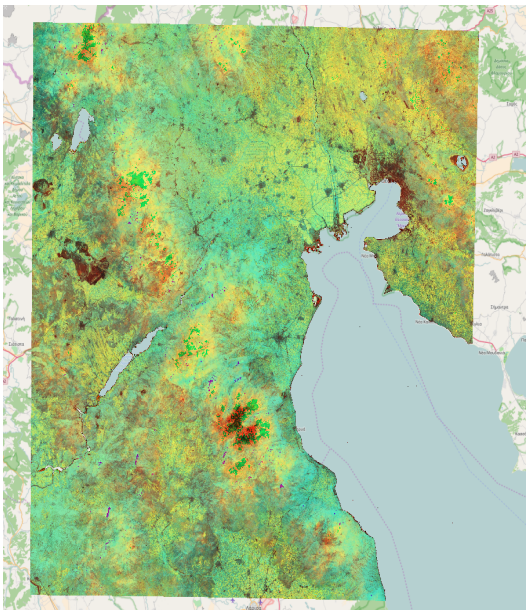


Σχήμα 5.3: Temporal Composite (TC) της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2015**.

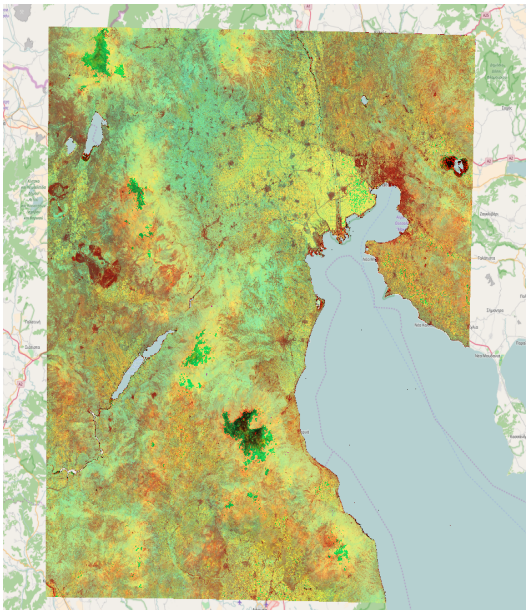
νίζουμε πολύ μεγάλες ροές εργασιών χειροκίνητα, να επιβλέπουμε τις ροές των δεδομένων για τυχόν λάθη, να γράφουμε μεγάλης έκτασης δυσανάγνωστο κώδικα για πολύ απλές εργασίες καθώς και να διαχειριζόμαστε σε πολύ χαμηλό επίπεδο τη διαθέσιμη υποδομή υλικού (hardware). Η λύση και η αρχιτεκτονική η οποία προτείνεται και εφαρμόστηκε κρίθηκε ότι πληροί τις απαιτήσεις για τη διενέργεια analytics σε μεγάλα γεωχωρικά δεδομένα όπως επίσης και ότι η προσέγγιση του cluster υπολογιστών είναι η ενδεδειγμένη για τη δημιουργία κατανεμημένων συστημάτων υψηλών επιδόσεων από κοινό υλικό το οποίο είναι το πιθανότερο ότι θα έχουμε στη διάθεση μας στις περισσότερες των περιπτώσεων για τη δημιουργία τέτοιων συστημάτων.



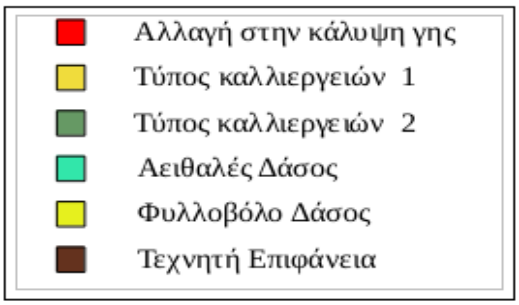
i. RGB (2015)



ii. TC (έτος 2014)



iii. TC (έτος 2015)



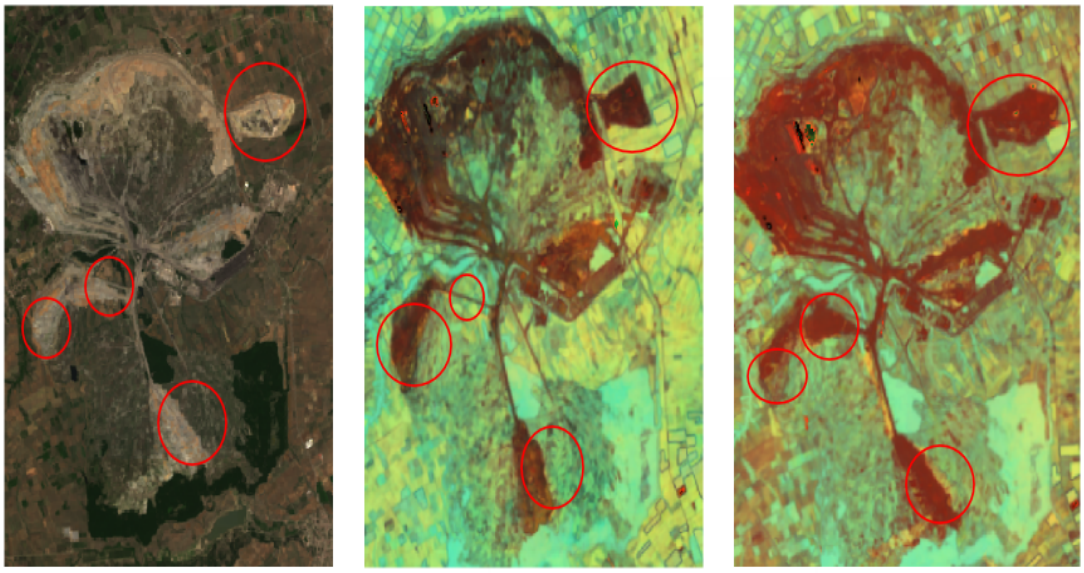
iv. Υπόμνημα

Σχήμα 5.4: Εστίαση στην περιοχή ενδιαφέροντος. Πάνω αριστερά (i) παρουσιάζεται ένα **φυσικό έγχρωμο σύνθετο** της περιοχής ενδιαφέροντος, πάνω δεξιά (ii) παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2014**, κάτω αριστερά (iii) παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2015** και τέλος παρουσιάζεται ένα ενδεικτικό **υπόμνημα των TCs**.

5.1 Παραγωγή διαχρονικών έγχρωμων σύνθετων ανά pixel

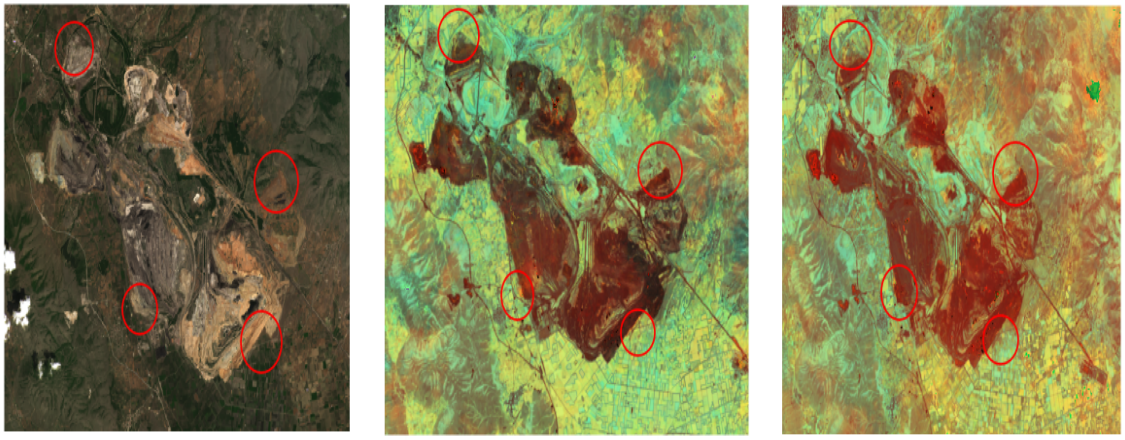
Η ανάλυση για την παραγωγή διαχρονικών έγχρωμων σύνθετων ανά pixel (Temporal Composites (TCs)) έδωσε τη δυνατότητα για την εξαγωγή συμπαγών δομών (αστικός ιστός, δασικές δομές, δρόμοι, όρια καλλιεργειών) καθώς και μεταβολών τους οι οποίες δεν είναι ορατές από τα αντίστοιχα φυσικά έγχρωμα σύνθετα του δορυφόρου Landsat 8. Η διαχρονική ανάλυση "αποκαλύπτει" τις ιδιότητες του κάθε pixel με συνέπεια οι τελικοί χάρτες ενώ έχουν την ίδια χωρική ανάλυση με τα raw δεδομένα, (30 μέτρα) να δίνουν την αίσθηση καλύτερης ανάλυσης και πιο "καθαρής" εικόνας. Αυτό οφείλεται στο γεγονός ότι κάθε κανάλι του TC προκύπτει με στατιστική ανάλυση.

Στο σχήμα 5.1 παρουσιάζεται το φυσικό έγχρωμο σύνθετο της **περιοχής ενδιαφέροντος (path-row: 184_032 του δορυφόρου Landsat 8)** για την οποία πραγματοποιήθηκαν οι ανεπτυγμένες διαχρονικές αναλύσεις ανά pixel οι οποίες και παρουσιάζονται στο κεφάλαιο αυτό. **Στο σχήμα 5.2** παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2014** και τέλος **στο σχήμα 5.3** παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο



i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

Σχήμα 5.5: Εστίαση στο **λιγνιτωρυχείο του Αμύνταιου.** Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **αύξηση των δραστηριοτήτων** του ορυχείου κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.



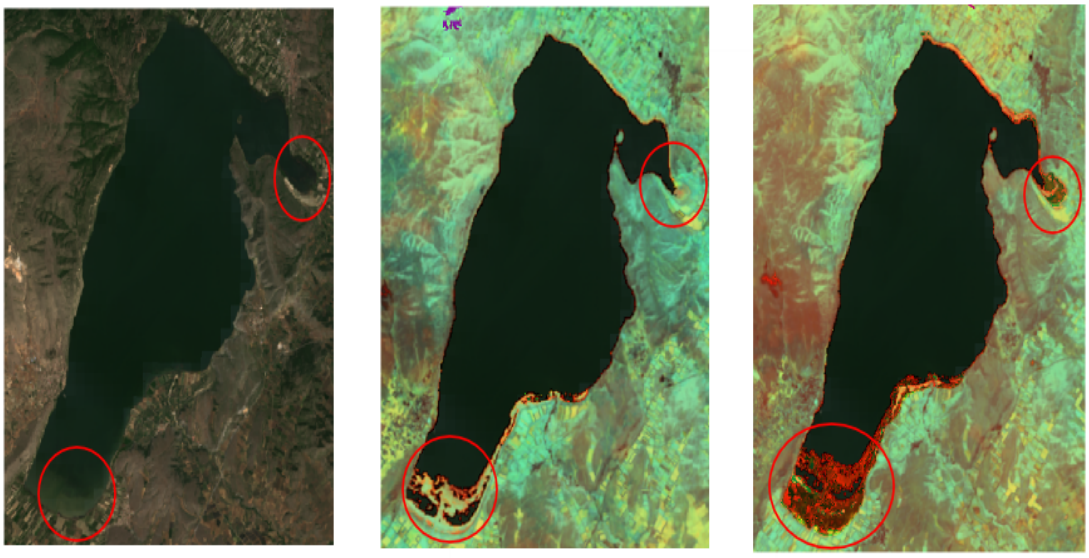
i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

Σχήμα 5.6: Εστίαση στο **λιγνιτωρυχείο της Πτολεμαΐδας**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **αύξηση των δραστηριοτήτων** του ορυχείου κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2015**.

Στο σχήμα 5.4 παρουσιάζονται εστιασμένες εικόνες στην περιοχή ενδιαφέροντος. Πάνω αριστερά (i) παρουσιάζεται ένα **φυσικό έγχρωμο σύνθετο** της περιοχής ενδιαφέροντος, πάνω δεξιά (ii) παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2014**, κάτω αριστερά (iii) παρουσιάζεται το TC της περιοχής ενδιαφέροντος βασισμένο στο δείκτη NDVI, το οποίο προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 **κατά το έτος 2015** και τέλος παρουσιάζεται ένα ενδεικτικό **υπόμνημα των TCs**.

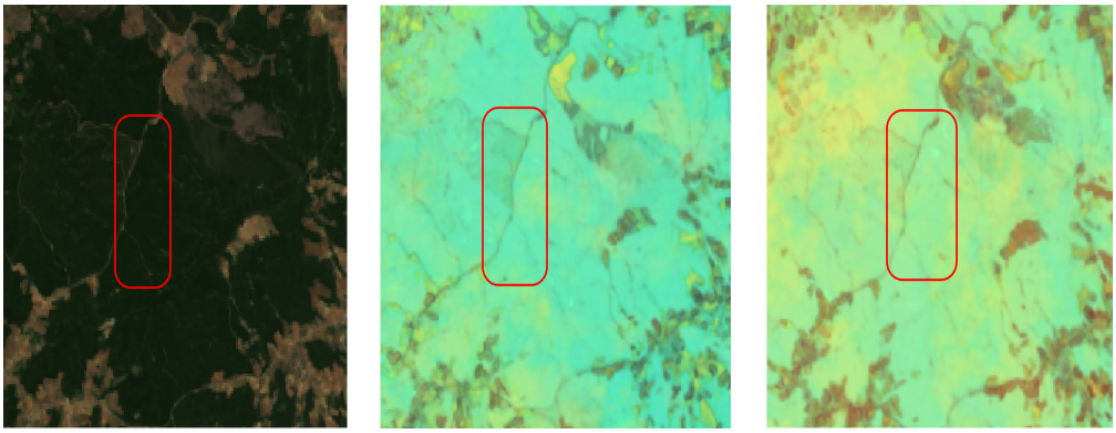
Στο υπόμνημα σχετικά με τα χρώματα των απεικονιζόμενων TCs παρατηρούμε μια ψευδοταξινόμηση η οποία προκύπτει για κάθε pixel ανάλογα με το χρώμα το οποίο έχει. Αυτή η ταξινόμηση προέκυψε με φωτοερμηνεία αλλά και λογική σύνδεση του τρόπου με τον οποίο προκύπτει το κάθε χρώμα από τις τιμές τις διαχρονικής ανάλυσης. Το κάθε TCδημιουργείται όπως είπαμε και στο κεφάλαιο 4 από τιμές στο κόκκινο, στο πράσινο και στο μπλε κανάλι κατ΄ αντιστοιχία με τα υπολογισμένα μεγέθη του συντελεστη μεταβλητότητας, της μέγιστης τιμής του δείκτη NDVI/EVI2 καθώς και της μέσης τιμής του δείκτη NDVI/EVI2 για κάθε pixel. Αυτός ο τρόπος κατασκευής του TC έχει σαν συνέπεια οι ιδιότητες που έχει το κάθε pixel, ανάλογα με την επιφάνεια την οποία απεικονίζει, να διαμορφώνουν σε διαφορετικά ύψη



i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

Σχήμα 5.7: Εστίαση στη **λίμνη της Βεγορίτιδας**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015.** Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **μεταβολή στην έκταση που καλύπτει το νερό** της λίμνης κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

τις τιμές στο κάθε κανάλι ώστε να είναι δυνατόν να διαμορφωθούν μοτίβα τα οποία ακολουθούν ομοειδείς επιφάνειες. Για παράδειγμα όπως βλέπουμε στο υπόμνημα του **σχήματος 5.4** με κόκκινο χρώμα στο TC φαίνονται οι **Αλλαγές στην κάλυψη γης**. Κόκκινο χρώμα σημαίνει υψηλή τιμή στο κόκκινο κανάλι και πολύ χαμηλές τιμές στα υπόλοιπα δηλαδή σημαίνει υψηλή πολύ υψηλή τιμή στο συντελεστή μεταβλητότητας και αρκετά χαμηλές τιμές στη μέγιστη και στη μέση τιμή του NDVI/EVI2, συνεπώς δηλαδή υποδηλώνει έντονη αλλαγή στις απεικονιζόμενες επιφάνειες. Με παρόμοια συλλογιστική ο **Τύπος καλλιεργειών 1** είναι κοντά στο ανοικτό κίτρινο με σχετικά υψηλή τιμή στο κόκκινο κανάλι (συντελεστής μεταβλητότητας), σχετικά υψηλή τιμή στο πράσινο κανάλι (μέγιστη τιμή του NDVI/EVI2) και χαμηλή τιμή στο μπλε κανάλι (μέση τιμή του NDVI/EVI2). Τιμές οι οποίες υποδηλώνουν φυτό το οποίο καλλιεργείται μια φορά το χρόνο και έχει πολύ σύντομο κύκλο ανάπτυξης. Συνεχίζοντας, ο **Τύπος καλλιεργειών 2** είναι κοντά στο πράσινο με σχετικά χαμηλή τιμή στο κόκκινο κανάλι (συντελεστής μεταβλητότητας), υψηλή τιμή στο πράσινο κανάλι (μέγιστη τιμή του NDVI/EVI2) και σχετικά υψηλή τιμή στο μπλε κανάλι (μέση τιμή του NDVI/EVI2). Τιμές οι οποίες υποδηλώνουν φυτό το οποίο καλλιεργείται μια φορά το χρόνο και έχει αρκετά μεγάλο κύκλο ανάπτυξης. Ακολουθεί η κλάση **Αειθαλές Δάσος** η οποία είναι κοντά στο σκούρο πράσινο με πολύ χαμηλή τιμή στο κόκκινο κανάλι (συντελεστής μεταβλητότητας),

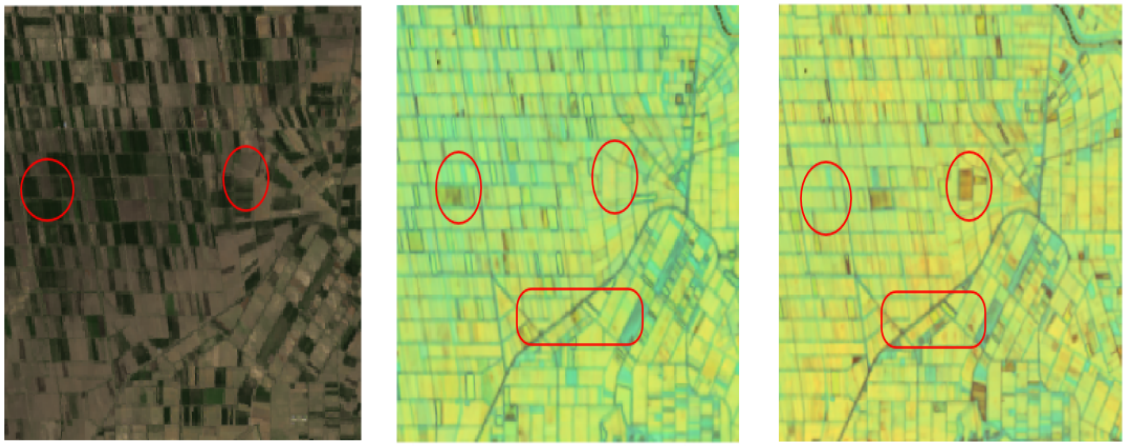


i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

Σχήμα 5.8: Εστίαση σε μία ορεινή περιοχή η οποία επιδεικνύει **πυκνή δασική δομή**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με το κόκκινο ορθογώνιο τονίζεται το πόσο ξεκάθαρα διαφαίνεται ο κύριος δρόμος, ο οποίος διασχίζει την περιοχή, στις εικόνες των TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Επίσης, όπως βλέπουμε και από το **υπόμνημα των TCs (σχήμα 5.4), η περιοχή περιέχει ως επί των πλείστων αειθαλή φυτά.**

υψηλή τιμή στο πράσινο κανάλι (μέγιστη τιμή του NDVI/EVI2) και υψηλή τιμή στο μπλε κανάλι (μέση τιμή του NDVI/EVI2). Αντίθετα η κατηγορία **Φυλλοβόλο Δάσος** είναι κοντά στο κίτρινο με υψηλή τιμή στο κόκκινο κανάλι (συντελεστής μεταβλητότητας), υψηλή τιμή στο πράσινο κανάλι (μέγιστη τιμή του NDVI/EVI2) και χαμηλή τιμή στο μπλε κανάλι (μέση τιμή του NDVI/EVI2). Τέλος, η κλάση **Τεχνητή Επιφάνεια** είναι κοντά στο σκούρο καφέ με αρκετά χαμηλές τιμές σε όλα τα κανάλια, δηλαδή μια επιφάνεια η οποία δεν αλλάζει στο χρόνο, δεν έχει υψηλή τιμή στο δείκτη NDVI/EVI2 και έχει και χαμηλή μέση τιμή στο δείκτη NDVI/EVI2. Με παρόμοιο τρόπο μπορούμε να αξιολογήσουμε το χρώμα του κάθε pixel του TC και να αποφασίσουμε για το είδος της επιφάνειας το οποίο απεικονίζει καθώς οι στατιστικές ιδιότητες του κάθε pixel οι οποίες αποκαλύπτονται με τη διαχρονική ανάλυση την οποία διενεργήσαμε κάνουν φανερή την ταυτότητά του.

Στο σχήμα 5.5 πραγματοποιείται εστίαση στο **λιγνιτωρυχείο του Αμύνταιου.** Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **αύξηση των δραστηριοτήτων** του ορυχείου κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs. **Στο σχήμα 5.6** πραγματοποιείται εστίαση στο **λιγνιτωρυχείο της Πτολεμαΐδας**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το



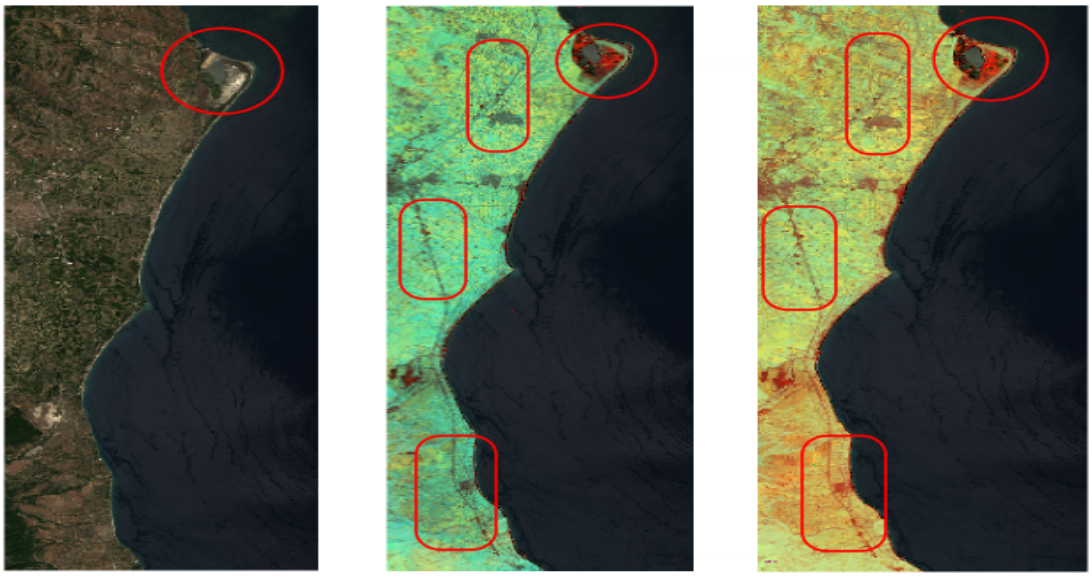
i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

Σχήμα 5.9: Εστίαση σε **αγροτεμάχια** στην περιοχή του Αξιού ποταμού, ποικίλων τύπων καλλιέργειας. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με το κόκκινο ορθογώνιο τονίζεται το πόσο ξεκάθαρα διαφαίνεται ο κύριος δρόμος, ο οποίος διασχίζει την περιοχή, στις εικόνες των TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Με τους κόκκινους κύκλους τονίζονται, ενδεικτικά, κάποιες περιοχές στις οποίες παρατηρείται **έντονη διαφοροποίηση στην απόκριση της βλάστησης** όπως προκύπτει από τη διαχρονική ανάλυση για την παραγωγή των TCs. Η έντονη διαφοροποίηση πιθανώς να σημαίνει αλλαγή στο είδος του φυτού που καλλιεργήθηκε από έτος σε έτος.

έτος 2014 και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **αύξηση των δραστηριοτήτων** του ορυχείου κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

Στο σχήμα 5.7 πραγματοποιείται εστίαση στη **λίμνη της Βεγορίτιδας**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015.** Με τους κόκκινους κύκλους τονίζονται οι περιοχές εκείνες στις οποίες παρατηρείται σημαντική **μεταβολή στην έκταση που καλύπτει το νερό** της λίμνης κατά τα δύο αυτά έτη όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

Στο σχήμα 5.8 πραγματοποιείται εστίαση σε μία ορεινή περιοχή η οποία επιδεικνύει **πυκνή δασική δομή**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με το κόκκινο ορθογώνιο τονίζεται το πόσο ξεκάθαρα διαφαίνεται ο κύριος δρόμος, ο οποίος διασχίζει την περιοχή, στις εικόνες των TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Επίσης, όπως βλέπουμε και από το **υπόμνημα των TCs (σχήμα 5.4), η περιοχή περιέχει ως επί των πλείστων αειθαλή φυτά.**



i. RGB (2015) ii. TC (έτος 2014) iii. TC (έτος 2015)

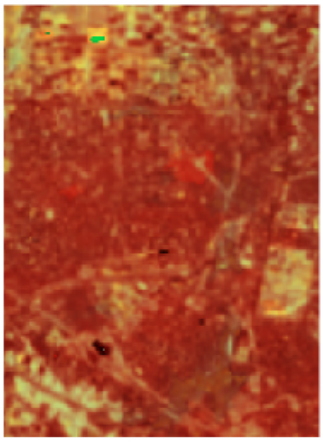
Σχήμα 5.10: Εστίαση στην ευρύτερη περιοχή της **Κατερίνης**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015.** Με το κόκκινα ορθογώνια τονίζεται το πόσο **ξεκάθαρα διαφαίνεται η Εθνική Οδός** Αθηνών-Θεσσαλονίκης σε όλο σχεδόν το μήκος της, η οποία διασχίζει την περιοχή, στις εικόνες των TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Με τον κόκκινο κύκλο τονίζεται η περιοχή της Αλυκής στην οποία φαίνεται η **διαρκής αλλαγή στην κάλυψη γης** κατά τα δύο αυτά έτη εξαιτίας των **δραστηριοτήτων των αλυκών** όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

Στο σχήμα 5.9 πραγματοποιείται εστίαση σε **αγροτεμάχια** στην περιοχή του Αξιού ποταμού, ποικίλων τύπων καλλιέργειας. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015**. Με το κόκκινο ορθογώνιο τονίζεται το πόσο ξεκάθαρα διαφαίνεται ο κύριος δρόμος, ο οποίος διασχίζει την περιοχή, στις εικόνες των TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Με τους κόκκινους κύκλους τονίζονται, ενδεικτικά, κάποιες περιοχές στις οποίες παρατηρείται **έντονη διαφοροποίηση στην απόκριση της βλάστησης** όπως προκύπτει από τη διαχρονική ανάλυση για την παραγωγή των TCs. Η έντονη διαφοροποίηση πιθανώς να σημαίνει αλλαγή στο είδος του φυτού που καλλιεργήθηκε από έτος σε έτος.

Στο σχήμα 5.10 πραγματοποιείται εστίαση στην ευρύτερη περιοχή της **Κατερίνης**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το **έτος 2014** και δεξιά παρουσιάζεται το TC για το **έτος 2015.** Με το κόκκινα ορθογώνια τονίζεται το πόσο **ξεκάθαρα διαφαίνεται η Εθνική Οδός** Αθηνών-Θεσσαλονίκης σε όλο σχεδόν το μήκος της, η οποία διασχίζει την περιοχή, στις εικόνες των

i. RGB (2015)

ii. TC (έτος 2014)



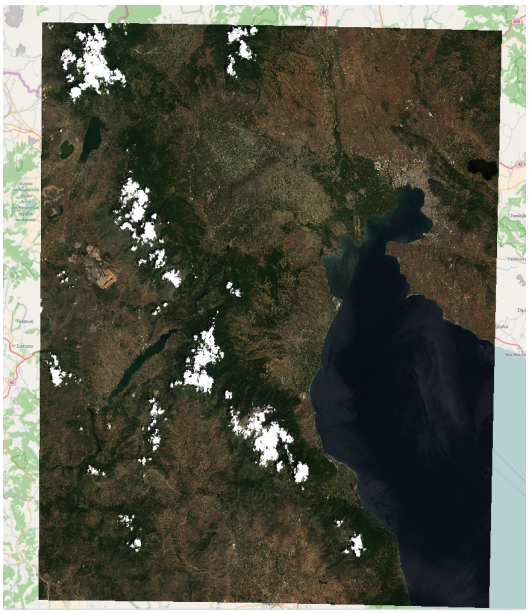
iii. TC (έτος 2015)

Σχήμα 5.11: Εστίαση σε αστική περιοχή στο **κέντρο της Θεσσαλονίκης**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το έτος 2014 και δεξιά παρουσιάζεται το TC για το έτος 2015. Δεν προκύπτουν σημαντικές διαφοροποιήσεις στο πλαίσιο της ανάλυσης μας, κάτι που είναι και αναμενόμενο καθώς δεν συμβαίνουν συχνά μεγάλες αλλαγές στα κέντρα μεγάλων αστικών περιοχών.

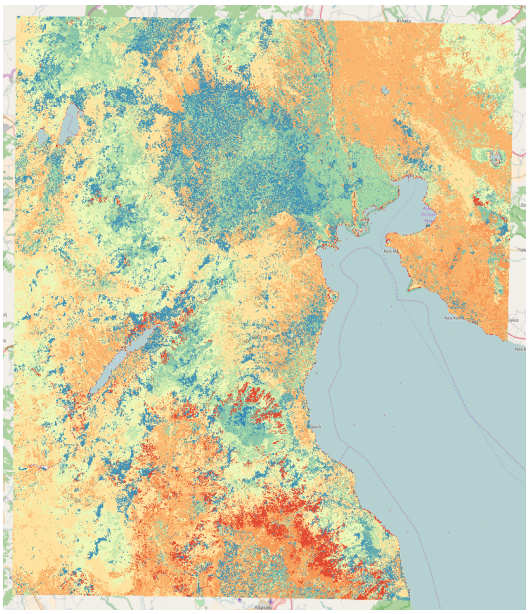
TCs σε σχέση με το φυσικό έγχρωμο σύνθετο. Με τον κόκκινο κύκλο τονίζεται η περιοχή της Αλυκής στην οποία φαίνεται η **διαρκής αλλαγή στην κάλυψη γης** κατά τα δύο αυτά έτη εξαιτίας των **δραστηριοτήτων των αλυκών** όπως ξεκάθαρα φαίνεται από τη διαχρονική ανάλυση η οποία πραγματοποιήθηκε για την παραγωγή των TCs.

Στο σχήμα 5.11 πραγματοποιείται εστίαση σε αστική περιοχή στο **κέντρο της Θεσσαλονίκης**. Αριστερά παρουσιάζεται ένα φυσικό έγχρωμο σύνθετο της περιοχής, στη μέση παρουσιάζεται το TC για το έτος 2014 και δεξιά παρουσιάζεται το TC για το έτος 2015. Δεν προκύπτουν σημαντικές διαφοροποιήσεις στο πλαίσιο της ανάλυσης μας, κάτι που είναι και αναμενόμενο καθώς δεν συμβαίνουν συχνά μεγάλες αλλαγές στα κέντρα μεγάλων αστικών περιοχών.

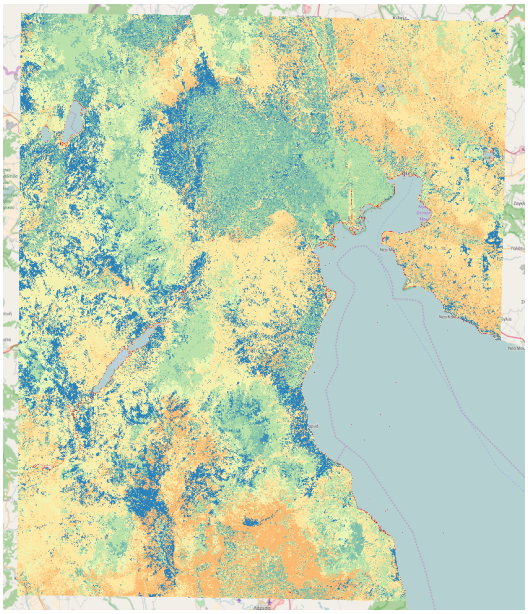
Τα αποτελέσματα καταδεικνύουν πόσο πολύτιμες μπορούν να αποδειχθούν τέτοιου είδους αναλύσεις. Παρατηρούμε πόσο εύκολο είναι με απλή φωτοερμηνεία, να εντοπίσουμε αλλαγές στη γήινη επιφάνεια αλλά και ποσοτικά να εντοπίσουμε το μέγεθος των αλλαγών αυτών. Τα τελικά διαχρονικά έγχρωμα σύνθετα μπορούν να χρησιμοποιηθούν τόσο ως μεμονωμένοι χάρτες για την απεικόνιση και τη μελέτη των μεταβολών στην κάλυψη γης όσο και ως είσοδοι σε σύνθετους αλγορίθμους ταξινόμησης με τη μορφή περιορισμών τους οποίους θα πρέπει να πληροί ένα pixel για να ανατεθεί σε κάποια κλάση με τη μεγαλύτερη δυνατή ασφάλεια.



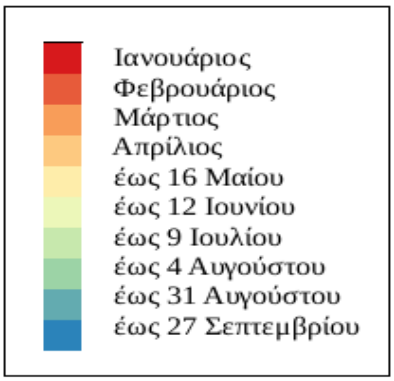
i. RGB (2015)



ii. DOY for max NDVI/EVI2 (2014)



iii. DOY for max NDVI/EVI2 (2015)



iv. Υπόμνημα

Σχήμα 5.12: Σχήμα το οποίο επιδεικνύει τους τελικούς χάρτες οι οποίοι παρήχθησαν για την **αποτύπωση της εποχικότητας** στην περιοχή ενδιαφέροντος (path-row: 184_032 του δορυφόρου Landsat 8). Πάνω αριστερά (i) παρουσιάζεται ένα **φυσικό έγχρωμο σύνθετο** της περιοχής ενδιαφέροντος, πάνω δεξιά (ii) παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας της περιοχής ενδιαφέροντος ο οποίος είναι κοινός για τους δείκτες NDVI/EVI2, ο οποίος προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 κατά το **έτος 2014**, κάτω αριστερά (iii) παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας της περιοχής ενδιαφέροντος ο οποίος είναι κοινός για τους δείκτες NDVI/EVI2, ο οποίος προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 κατά το **έτος 2015** και τέλος παρουσιάζεται το **υπόμνημα** των χαρτών αποτύπωσης της εποχικότητας.

5.2 Αποτύπωση της εποχικότητας

Όπως σχολιάσαμε αναλυτικά και στο κεφάλαιο 4 με τον όρο αποτύπωση της εποχικότητας αναφερόμαστε στη μέρα του έτους για την οποία επικρατεί κάποια συγκεκριμένη συνθήκη/συνθήκες σε μία περιοχή και ενδιαφερόμαστε να τη γνωρίζουμε για να προχωρήσουμε σε διάφορες αναλύσεις.

Στην παρούσα εργασία, και με μια εστίαση σε αγροτικές εφαρμογές, με τον όρο αποτύπωση της εποχικότητας ορίζουμε το ενδιαφέρον μας ώστε να γνωρίζουμε τη μέρα του έτους για την οποία σε μια περιοχή (pixel) συμβαίνει ο δείκτης NDVI (ή ο δείκτης EVI2) να παίρνει τη μέγιστη τιμή του. Με την εξαγωγή μιας τέτοιας πληροφορίας είμαστε σε θέση να εξάγουμε μοτίβα για την ανάπτυξη παρακολουθούμενων καλλιεργειών, να παρατηρούμε ομοιότητες, να διαχωρίζουμε ανομοιογένειες και να παρακολουθούμε επακριβώς την κατάσταση που επικρατεί σε μια περιοχή που καλύπτει ένα συγκεκριμένο pixel για κάθε κύκλο ανάπτυξης των καλλιεργειών.

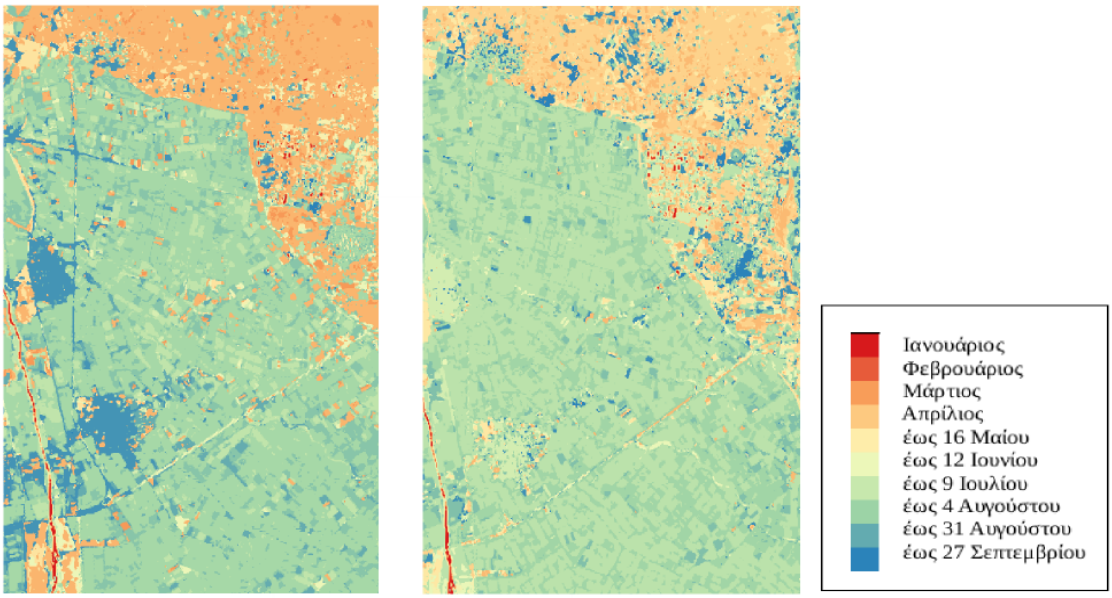
Επομένως, στο υλοποιημένο ερώτημα για την αποτύπωση της εποχικότητας στα πλαίσια της παρούσας εργασίας, στόχος μας ήταν για κάθε path-row (ή tile) των δορυφόρων, δεδομένα των οποίων αποθηκεύονται στο σύστημά μας, να εντοπίσουμε μέσω της διαχρονικής ανάλυσης όλων των σκηνών κάθε path-row (ή tile) τη μέρα του έτους για την οποία συμβαίνει ο δείκτης NDVI (ή ο δείκτης EVI2) να παίρνει τη μέγιστη τιμή του σε κάθε pixel που περιέχει ένα συγκεκριμένο path-row (ή tile) και αφού γίνει ο εν λόγω εντοπισμός να παράγεται ο αντίστοιχος χάρτης. Σημειώνεται ότι οι τιμές του δείκτη NDVI παίρνουν την ίδια μέρα τη μέγιστη τιμή τους με τις τιμές του δείκτη EVI2. Αυτό συμβαίνει γιατί και οι δύο δείκτες προκύπτουν ως αποτελέσματα πράξεων του κόκκινου με το υπέρυθρο κανάλι του δορυφόρου Landsat 8 μέσω παρεμφερών γραμμικών σχέσεων και αυτό έχει σαν συνέπεια ναι μεν οι τιμές τους καθώς και η ερμηνεία των τιμών τους να είναι διαφορετική αλλά οι μέγιστες τιμές τους να προκύπτουν τις ίδιες μέρες. Επομένως, στα διαγράμματα παρουσιάζεται ένας χάρτης αποτύπωσης της εποχικότητας ο οποίος είναι κοινός και για τους δύο δείκτες.

Στο σχήμα 5.12 παρουσιάζονται οι τελικοί χάρτες οι οποίοι παρήχθησαν για την **αποτύπωση της εποχικότητας** στην περιοχή ενδιαφέροντος (path-row: 184_032 του δορυφόρου Landsat 8). Πάνω αριστερά (i) παρουσιάζεται ένα **φυσικό έγχρωμο σύνθετο** της περιοχής ενδιαφέροντος, πάνω δεξιά (ii) παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας της περιοχής ενδιαφέροντος ο οποίος είναι κοινός για τους δείκτες NDVI/EVI2, ο οποίος προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 κατά το **έτος 2014**, κάτω αριστερά (iii) παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας της περιοχής ενδιαφέροντος ο οποίος είναι κοινός για τους δείκτες NDVI/EVI2, ο οποίος προέκυψε μέσω της διαχρονικής ανάλυσης για κάθε pixel κάθε εικόνας η οποία λήφθηκε από το δορυφόρο Landsat 8 κατά το **έτος 2015** και τέλος παρουσιάζεται το **υπόμνημα** των χαρτών αποτύπωσης της εποχικότητας.

Στο σχήμα 5.13 πραγματοποιείται εστίαση στην ευρύτερη **περιοχή του Αξιού** ποταμού. Αριστερά παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2014**, στη μέση παρουσιάζεται ο χάρτης αποτύπωσης της εποχι-

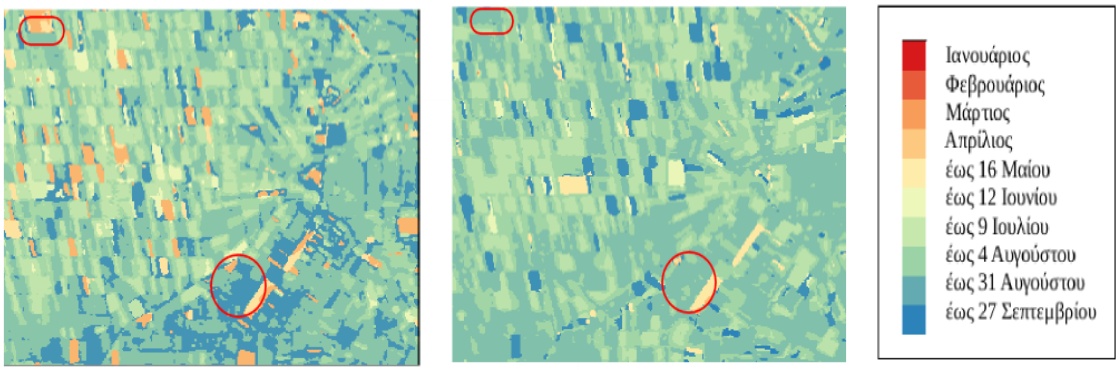
κότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2015** ενώ δεξιά έχουμε το **υπόμνημα** των δύο χαρτών. Παρατηρείται διαφοροποίηση μεταξύ των δύο χαρτών στη μεγαλύτερη έκταση της απεικονιζόμενης περιοχής. **Η διαφοροποίηση, ποσοτικά, είναι της τάξης των 15-20 ημερών** κατά πλειοψηφία στις καλλιεργήσιμες εκτάσεις. Η **κορύφωση του καλλιεργητικού κύκλου ζωής συνέβη περίπου 15-20 ημέρες νωρίτερα** κατά το έτος 2015 από ότι το έτος 2014. Αυτό μπορεί να οφείλεται τόσο σε κλιματολογικές συνθήκες όσο και σε καλλιεργητικές πρακτικές.

Στο σχήμα 5.14 πραγματοποιείται εστίαση σε **καλλιεργήσιμες εκτάσεις**. Αριστερά παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2014**, στη μέση παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2015** ενώ δεξιά έχουμε το **υπόμνημα** των δύο χαρτών. Παρατηρείται **διαφοροποίηση** μεταξύ των δύο χαρτών στη μεγαλύτερη έκταση της απεικονιζόμενης περιοχής. Η διαφοροποίηση είναι εντονότερη σε κάποια αγροτεμάχια και



i. DOY,max ND/E[VI] (2014) ii. DOY,max ND/E[VI] (2015) iii. Υπόμνημα

Σχήμα 5.13: Εστίαση στην ευρύτερη **περιοχή του Αξιού** ποταμού. Αριστερά παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2014**, στη μέση παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2015** ενώ δεξιά έχουμε το **υπόμνημα** των δύο χαρτών. Παρατηρείται διαφοροποίηση μεταξύ των δύο χαρτών στη μεγαλύτερη έκταση της απεικονιζόμενης περιοχής. **Η διαφοροποίηση, ποσοτικά, είναι της τάξης των 15-20 ημερών** κατά πλειοψηφία στις καλλιεργήσιμες εκτάσεις. Η **κορύφωση του καλλιεργητικού κύκλου ζωής συνέβη περίπου 15-20 ημέρες νωρίτερα** κατά το έτος 2015 από ότι το έτος 2014. Αυτό μπορεί να οφείλεται τόσο σε κλιματολογικές συνθήκες όσο και σε καλλιεργητικές πρακτικές.



i. DOY,max ND/E[VI] (2014) ii. DOY,max ND/E[VI] (2015) iii. Υπόμνημα

Σχήμα 5.14: Εστίαση σε **καλλιεργήσιμες εκτάσεις**. Αριστερά παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2014**, στη μέση παρουσιάζεται ο χάρτης αποτύπωσης της εποχικότητας για τους δείκτες NDVI/EVI2 κατά το **έτος 2015** ενώ δεξιά έχουμε το **υπόμνημα** των δύο χαρτών. Παρατηρείται **διαφοροποίηση** μεταξύ των δύο χαρτών στη μεγαλύτερη έκταση της απεικονιζόμενης περιοχής. Η διαφοροποίηση είναι εντονότερη σε κάποια αγροτεμάχια και ηπιότερη σε κάποια άλλα. Με το κόκκινο ορθογώνιο τονίζεται μια περιοχή έντονης διαφοροποίησης (τάξη μεγέθους μήνες) ενώ με τον κόκκινο κύκλο τονίζεται μια περιοχή ηπιότερης διαφοροποίησης (τάξη μεγέθους 1 μήνας). Οι διαφοροποιήσεις αυτές μπορεί να οφείλονται τόσο σε κλιματολογικές συνθήκες όσο και σε καλλιεργητικές πρακτικές (για παράδειγμα αλλαγή του τύπου/είδους του φυτού το οποίο καλλιεργείται).

ηπιότερη σε κάποια άλλα. Με το κόκκινο ορθογώνιο τονίζεται μια περιοχή έντονης διαφοροποίησης (τάξη μεγέθους μήνες) ενώ με τον κόκκινο κύκλο τονίζεται μια περιοχή ηπιότερης διαφοροποίησης (τάξη μεγέθους 1 μήνας). Οι διαφοροποιήσεις αυτές μπορεί να οφείλονται τόσο σε κλιματολογικές συνθήκες όσο και σε καλλιεργητικές πρακτικές (για παράδειγμα αλλαγή του τύπου/είδους του φυτού το οποίο καλλιεργείται).

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές επεκτάσεις

Οι τωρινές τεχνολογίες μεγάλων γεωχωρικών δεδομένων στοχεύουν προς την επίτευξη της κλιμακωσιμότητας των συστημάτων, στην αποτελεσματικότητα της αποθήκευσης και στην παροχή γεωχωρικών υπηρεσιών ανάλυσης ώστε να είναι δυνατόν να προσφέρουν πληροφορίες χωρικής ανάλυσης όχι μόνο σε ένα μεγάλο κοινό από επαγγελματίες στο χώρο των GIS αλλά και σε χρήστες εφαρμογών κινητών συσκευών, χρήστες μέσων κοινωνικής δικτύωσης καθώς και πρωτοβουλιών ανοιχτών δεδομένων. Το κύριο πεδίο ενδιαφέροντος είναι η μετακίνηση από την απλή παρακολούθηση και αναφορά γεγονότων και φαινομένων σε μια συλλογή από υψηλού επιπέδου υπηρεσίες οι οποίες θα μπορούν να προσφέρουν χρήσιμα αποτελέσματα ανάλυσης για την κάλυψη καθημερινών αναγκών των επαγγελματιών οι οποίοι εμπλέκονται στα διάφορα πεδία εφαρμογών χωρικής ανάλυσης όπως για παράδειγμα είναι η γεωργία, το περιβάλλον, η δασοκομία κ.α. Πιο συγκεκριμένα, περιμένουμε μια μετακίνηση από τη διενέργεια αναλύσεων στα δεδομένα και τις αναλύσεις πραγματικού χρόνου στην παρακολούθηση μεταβολών σε πραγματικό χρόνο αλλά και στην πρόβλεψη διάφορων μεταβολών μέσω εύρωστων μοντελοποιήσεων και τεχνικών τεχνητής νοημοσύνης.

Επιπροσθέτως, όσον αναφορά τα γεωχωρικά δεδομένα, αναμένεται στα επόμενα χρόνια μια αλλαγή προσέγγισης καθώς θα είναι δυνατή η καθημερινή παρακολούθηση ολόκληρου του πλανήτη μέσω πολλών, μικρών δορυφόρων σε τροχιά, με χωρική ανάλυση μερικών μέτρων (για τον κόσμο των raster δεδομένων), αλλά και με προχωρημένες τεχνολογίες υπερφασματικής φωτογραφίας πάνω σε drones και δορυφόρους. Επιπλέον, έχει ξεκινήσει η ενασχόληση με την αναδυόμενη τεχνολογία του βίντεο συνεχούς ροής (video streaming) πάνω σε δορυφόρους για την παρακολούθηση της γης, η οποία θα προσφέρει μια τεράστια εξέλιξη στα διαθέσιμα χωρικά δεδομένα που προέρχονται από το διάστημα, με ταυτόχρονη έκρηξη στις ανάγκες αποθηκευτικού χώρου καθώς και ρυθμών συμπίεσης από νέους αλγορίθμους.

Στο πεδίο της τηλεπισκόπησης, η εισαγωγή drones χαμηλού κόστους αναμένεται να δώσει μεγάλη ώθηση στην συλλογή τηλεπισκοπικών δεδομένων και ιδιαίτερα δεδομένων πάρα πολλής μεγάλης χωρικής ανάλυσης της τάξης των 10 cm ή και ακόμα μικρότερης για συγκεκριμένες εφαρμογές. Είναι εύκολο να γίνει αντιληπτό ότι ο τεράστιος υπολογιστικός φόρτος που θα

απαιτείται για την επεξεργασία αυτών των δεδομένων θα απαιτήσει ερευνητική δραστηριότητα στην αυτοματοποίηση τεχνικών και ιδιαίτερα στο μετασχηματισμό πρωτογενών δεδομένων σε καλιμπραρισμένη και επικυρωμένη πληροφορία. Αναμένεται ότι οι τεχνικές μηχανικής μάθησης (machine learning) θα εμπλακούν προς αυτήν την κατεύθυνση ώστε να προσφέρουν εύρωστες εφαρμογές για την ταξινόμηση εικόνων, την αναγνώριση αντικειμένων όπως επίσης και την εγγραφή multi-modal δεδομένων (multi-modal data registration).

Σημαντικές τεχνολογικές προκλήσεις προέρχονται από την εξέλιξη και στη χρήση των χαρτών. Ένας χάρτης συνήθως ήταν κάτι που βρισκόταν σε ένα βιβλίο και περιέγραφε τοποθεσίες εξωτερικών χώρων. Πλέον ο χάρτης είναι κάτι που οι άνθρωποι έχουν συνεχώς μαζί τους σε ένα smartphone ή σε ένα tablet pc και εξαρτώνται σε μεγάλο βαθμό από τις πληροφορίες που υπηρεσίες βασισμένες στους χάρτες προσφέρουν. Επίσης η νέα τάση βρίσκεται στους χάρτες που περιγράφουν εσωτερικούς χώρους όπως για παράδειγμα είναι ένα μουσείο ή ένα πολυκατάστημα. Επειδή οι φορητές συσκευές με γνώση της εσωτερικής τοποθεσίας προσφέρουν μεγάλες επιχειρηματικές ευκαιρίες, περιμένουμε ότι η εκρηκτική διαθεσιμότητα χαρτών εσωτερικών χώρων θα ανοίξει ένα νέο μέτωπο στις υπηρεσίες βασισμένες στην εσωτερική τοποθεσία (indoor location based services). Επιπλέον, η βελτίωση στην τεχνολογία έχει καταστήσει να είναι εφικτή η γρήγορη συλλογή τρισδιάστατης πληροφορίας και για το λόγο αυτό οι τρισδιάστατοι (3D) χάρτες γίνονται όλο και πιο κοινοί. Η ικανότητα για παροχή 3D χαρτών σε περιηγητές ιστού σε συνδυασμό με το ολοένα και αυξανόμενο ενδιαφέρον για το αστικό περιβάλλον, δίνουν την ώθηση για την ανάπτυξη ενός νέου εύρους εφαρμογών.

Επιπροσθέτως, η επαυξημένη πραγματικότητα (Augmented Reality (AR)) συγχωνεύει τους χάρτες με τον πραγματικό κόσμο. Η AR βρίσκεται ακόμα σε πρωταρχικό στάδιο αλλά είναι σε θέση να αλλάξει τον τρόπο με τον οποίο βλέπουμε τον κόσμο. Ένας τομέας με τον οποίο ασχολείται η AR είναι η εξατομίκευση των χαρτών για κάθε χρήστη. Η εξατομίκευση βασίζεται στην ακαριαία και σε πραγματικό χρόνο αλληλεπίδραση των εφαρμογών των χρηστών με γεωχωρική πληροφορία. Τέτοια υπολογιστική χαρτογραφία με σκοπό τη δημιουργία εξειδικευμένων υπηρεσιών για τους χρήστες εξαρτάται από το αν η δόμηση της γεωχωρικής πληροφορίας είναι τέτοια που να επιτρέπει την αυτόματη επεξεργασία της. Η εξέλιξη στο επιστημονικό πεδίο της σημασιολογίας (semantics domain) είναι δυνατόν να προσφέρει μια βασική οντολογία για την εκτέλεση τέτοιων αυτοματοποιημένων γεωχωρικών υπολογισμών.

Στη συνέχεια παρουσιάζουμε κάποια συμπεράσματα που προκύπτουν από την εκπόνηση της μεταπτυχιακής εργασίας καθώς και προτάσεις για θέματα μελλοντικής έρευνας και περαιτέρω μελέτης.

6.1 Σύνοψη και συμπεράσματα

Στην εργασία αυτή παρουσιάστηκε ο σχεδιασμός και η υλοποίηση ενός πλήρους επιχειρησιακού συστήματος διαχείρισης και επεξεργασίας μεγάλων γεωχωρικών δεδομένων το οποίο είναι ικανό να χρησιμοποιηθεί για πολύ μεγάλο εύρος εφαρμογών ανάλυσης τόσο raster δεδομένων όσο και vector δεδομένων ή και συνδυασμό τους. Οι εφαρμογές αυτές περιλαμβάνουν τόσο μη διαδραστικές εφαρμογές (batch processing) όσο και εφαρμογές πραγματικού χρό-

νου (real-time processing). Η κατανεμημένη αρχιτεκτονική του συστήματος, σχεδιασμένη με τις πιο παραδεκτές πρακτικές και χρησιμοποιώντας τα πιο εύρωστα, δημοφιλή και σύγχρονα προγραμματιστικά εργαλεία και πλατφόρμες εγγυάται την ταχύτατη επεξεργασία και ανάλυση των αποθηκευμένων δεδομένων δίνοντας έμφαση στην ολοκληρωτική χρησιμοποίηση των διαθέσιμων υπολογιστικών πόρων, στη μαζική παραλληλοποίηση των αλγοριθμικών βημάτων καθώς και στην πρόσβαση στα δεδομένα και ταυτόχρονα εκπληρώνοντας τις απαιτήσεις για απρόσκοπτη και συνεχόμενη λειτουργία του συστήματος, προστασία των δεδομένων από ενδεχόμενες αστοχίες υλικού και γρήγορη ανάνηψη σε περίπτωση που συμβούν σοβαρές αστοχίες υλικού. Τα αποτελέσματα της εφαρμογής των υλοποιημένων αλγορίθμων στο σύστημα για τη διαχρονική ανάλυση των αποθηκευμένων δεδομένων κατέδειξαν πώς η λειτουργία και η χρήση τέτοιων συστημάτων αποτελεί μονόδρομο για οργανισμούς και εταιρείες που βασίζουν τη δραστηριότητα τους στην ανάλυση μεγάλων γεωχωρικών δεδομένων.

6.2 Μελλοντικές Επεκτάσεις

Παρακάτω παρουσιάζονται προτάσεις για θέματα μελλοντικής έρευνας και περαιτέρω μελέτης, όπως αυτά προκύπτουν από τις σύγχρονες τεχνολογικές προκλήσεις και ανάγκες και έγιναν φανερά στο πλαίσιο εκπόνησης αυτής της διπλωματικής εργασίας.

6.2.1 Βελτίωση και επέκταση του συστήματος

Οι κύριοι μελλοντικοί στόχοι είναι η βελτίωση και η επέκταση του ανεπτυγμένου συστήματος. Πρώτα από όλα κύριο μέλημα αποτελεί η ενσωμάτωση και άλλων εργαλείων στην αρχιτεκτονική του συστήματος για ακόμα πιο γρήγορη και αποδοτική επεξεργασία των δεδομένων ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής. Για παράδειγμα στην παρούσα φάση χρησιμοποιείται το ίδιο ακριβώς pipeline επεξεργασίας τόσο για τις batch εφαρμογές όσο και για τις εφαρμογές πραγματικού χρόνου. Με χρήση ακόμα πιο εξειδικευμένων εργαλείων για την κάθε εφαρμογή μπορούμε να πετύχουμε ταχύτερη επεξεργασία και στις δύο περιπτώσεις.

Η δυνατότητα χρήσης αλγορίθμων υλοποιημένων σε διαφορετικές γλώσσες προγραμματισμού από αυτές που τώρα υποστηρίζει το σύστημα είναι θεμελιώδους σημασίας τόσο από πλευράς επιδόσεων, καθώς κάποιες γλώσσες προγραμματισμού είναι ταχύτερες από κάποιες άλλες για συγκεκριμένες εφαρμογές όσο και από πλευράς εύκολης χρήσης του συστήματος από χρηστές οι οποίοι δεν είναι εξοικειωμένοι με τις υποστηριζόμενες από το σύστημα γλώσσες.

Επιπλέον, η δυνατότητα σύνδεσης του συστήματος με υπάρχουσες εξωτερικές βιβλιοθήκες είναι θέμα που πρέπει να εξεταστεί ώστε να είναι δυνατή η πλήρης αξιοποίηση λειτουργικότητας που έχει ήδη αναπτυχθεί χωρίς να χρειάζεται να υλοποιηθεί λογισμικό ίδιας λειτουργικότητας από την αρχή κάτι το οποίο αποτελεί άσκοπη σπατάλη πόρων.

Τέλος, η ανάπτυξη ή/και η ενσωμάτωση ενός υψηλού επιπέδου διαχειριστικού εργαλείου για τη διαχείριση και την παρακολούθηση των υποβαλλόμενων στο σύστημα υπολογιστικών εργασιών είναι θεμελιώδους σημασίας αν θέλουμε να επιτύχουμε μέσω ενός καλώς ορισμένου API την αποδοτική και συνδυασμένη χρήση του συστήματος τόσο από ειδικούς όσο και από αρχάριους χρήστες αλλά και από προγράμματα εφαρμογών και διαδικτυακές υπηρεσίες.

Αξίζει να σημειωθεί ότι η συνεχής ενημέρωση και ενσωμάτωση στο σύστημα, των πιο πρόσφατων επικυρωμένων πρακτικών ραδιομετρικής και ατμοσφαιρικής διόρθωσης των δεδομένων, αποτελεί κυρίαρχη προτεραιότητα για τις μελλοντικές εργασίες ανάπτυξης του συστήματος σχετικά με τα στάδια της προ-επεξεργασίας των δεδομένων και τα ερωτήματα επεξεργασίας πάνω στα δεδομένα.

6.2.2 Ενσωμάτωση και αξιοποίηση GPUs

Στην παρούσα εργασία αναπτύχθηκε ένα σύστημα επεξεργασίας μεγάλων γεωχωρικών δεδομένων σε περιβάλλον cluster υπολογιστών για εφαρμογές παρατήρησης της Γης. Ως συνέπεια της αρχιτεκτονικής του συστήματος, των τμημάτων λογισμικού και των προγραμματιστικών εργαλείων που χρησιμοποιήθηκαν ήταν δυνατή η παράλληλη εκτέλεση των υπολογιστικών εργασιών τόσο κατά μήκος των διαφορετικών υπολογιστικών κόμβων του cluster περιβάλλοντος όσο και κατά μήκος των διαθέσιμων CPU cores κάθε κόμβου του cluster . Η αξιοποίηση όμως των γραφικών επεξεργαστικών μονάδων (GPUs) δεν είναι ακόμα δυνατή από το υπάρχον σύστημα. Οι GPUs είναι σε θέση να βελτιώσουν σημαντικά τους χρόνους εκτέλεσης των υπολογιστικών εργασιών καθώς ενδείκνυνται για εφαρμογές επεξεργασίας εικόνων οπότε η δημιουργία του πλαισίου για τη διαφανή προς τον χρήστη αξιοποίησή τους από συστήματα μεγάλης κλίμακας είναι καθοριστικής σημασίας. Υπάρχουν διάφορα frameworks τα οποία μπορούν να χρησιμοποιηθούν προς την κατεύθυνση αυτή, οπότε και ο πειραματισμός με αυτά για την ενσωμάτωση τους στο υπάρχον σύστημα αποτελεί σίγουρα σημαντική πτυχή της επέκτασης του ανεπτυγμένου συστήματος.

6.2.3 Benchmarking

Σημείο κλειδί αποτελεί επίσης η εγκατάσταση του συστήματος σε περιβάλλον παραγωγής ώστε να αποκτηθεί εικόνα και εμπειρία για τη συμπεριφορά του σε πραγματικές συνθήκες και απαιτήσεις. Σε τέτοιο περιβάλλον έχει νόημα η διενέργεια benchmarkings καθώς επίσης και stress tests αρχικά σε επίπεδο αξιολόγησης του ίδιου του συστήματος για να διαπιστωθούν οι δυνατότητες του πέρα από θεωρητικές εκτιμήσεις και συνθήκες σε περιβάλλον ελέγχου όπως επίσης και τα bottlenecks ενός τέτοιου συστήματος. Έπειτα θα ακολουθήσει η αξιολόγηση του συστήματος σε σχέση και σε σύγκριση με άλλα συστήματα διαχείρησης και επεξεργασίας μεγάλων γεωχωρικών δεδομένων για να διαπιστωθεί πιο σύστημα παρέχει περισσότερες δυνατότητες, ταχύτερες επιδόσεις, καλύτερη κλιμάκωση και για ποιες υπολογιστικές εργασίες.

Κατάλογος Σχημάτων

Βιβλιογραφία

[1] COMMISSION DECISION of 12 december 2011 on the reuse of Commission documents (2011/833/eu). *Official Journal of the European Union*, χ.χ.

[2] Y. L. Chang A. Plaza, Q. Du και R. L. King. High performance computing for hyperspectral remote sensing. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 4:528;544, 2011.

[3] A. Adamov. Distributed file system as a basis of data-intensive computing. Στο *Application of Information and Communication Technologies (AICT), 2012 6th International Conference on*, σελίδες 1–3, 2012.

[4] Andrei Aiordachioaie και Peter Baumann. Petascope: An open-source implementation of the ogc wcs geo service standards suite. Στο *Scientific and Statistical Database Management*Michael Gertz και Bertram Ludascher, επιμελητές, τόμος 6187 στο *Lecture Notes in Computer Science*, σελίδες 160–168. Springer Berlin Heidelberg, 2010.

[5] Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang και Joel Saltz. Hadoop gis: A high performance spatial data warehousing system over mapreduce. *Proc. VLDB Endow.*, 6(11):1009–1020, 2013.

[6] Marcos D. Assuncao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto και Rajkumar Buyya. Big data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, (0):–, 2014.

[7] V. A. Ayma, R. S. Ferreira, P. Happ, D. Oliveira, R. Feitosa, G. Costa, A. Plaza και P. Gamba. Classification algorithms for big data analysis, a map reduce approach. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, ΞΛ-3/Ω2:17–21, 2015.

[8] V. A. Ayma, R. S. Ferreira, P. N. Happ, D. A. B. Oliveira, G. A. O. P. Costa, R. Q. Feitosa, A. Plaza και P. Gamba. On the architecture of a big data classification tool based on a map reduce approach for hyperspectral image analysis. Στο *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, σελίδες 1508–1511, 2015.

[9] N. E. Digirolamo B. J. Choudhury και T. J. Dorman. A comparison of reflectances and vegetation indices from three methods of compositing the avhrr-gac data over northern africa. *Remote Sens. Rev.*, 10:245;263, 1994.

[10] M. Babaee, M. Datcu και G. Rigoll. Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization. Στο *Big Data, 2013 IEEE International Conference on*, σελίδες 1–6, 2013.

[11] L.A. Barroso, J. Dean και U. Holzle. Web search for a planet: The google cluster architecture. *Micro, IEEE*, 23(2):22–28, 2003.

[12] P. Baumann. rasdaman: Array databases boost spatio-temporal analytics. Στο *Computing for Geospatial Research and Application (COM.Geo), 2014 Fifth International Conference on*, σελίδες 54–54, 2014.

[13] P. Baumann, A. Dehmel, P. Furtado, R. Ritsch και N. Widmann. The multidimensional database system rasdaman. Στο *In Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, σελίδες 575–577. ACM Press, 1998.

[14] Peter Baumann. Management of multidimensional discrete data. *The VLDB Journal*, 4(3):401–444, 1994.

[15] Peter Baumann. A database array algebra for spatio-temporal data and beyond. Στο *In Next Generation Information Technologies and Systems*, σελίδες 76–93. Springer, 1999.

[16] Peter Baumann. Array databases and raster data management. Στο *In: T. Ozsu, L. Liu (eds.), Encyclopedia of Database Systems*. Springer, 2009.

[17] Peter Baumann. The OGC web coverage processing service (WCPS) standard. *GeoInformatica*, 14(4):447–479, 2010.

[18] Edmon Begoli και James Horey. Design principles for effective knowledge discovery from big data. Στο *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on*, σελίδες 215–218. IEEE, 2012.

[19] Rajkumar Buyya. *High Performance Cluster Computing: Architectures and Systems, Vol. 1*. Prentice Hall, 1999.

[20] J. Fry M. Coan N. Hossain C. Larson N. Herold A. McKerrow J. N. VanDriel C. Homer, J. Dewitz και J. Wickham. National land cover database for the conterminous united states. *Photogramm. Eng. Remote Sens.*, 73:337;341, 2007.

[21] J. G. Masek N. Thomas Z. Zhu C. Huang, S. N. Coward και J. E. Vogelmann. An automated approach for reconstructing recent forest disturbance history using dense landsat time series stacks. *Remote Sens. Environ.*, 114:183;198, 2010.

[22] P. Cappelaere, S. Sanchez, S. Bernabe, A. Scuri, D. Mandl και A. Plaza. Cloud implementation of a full hyperspectral unmixing chain within the nasa web coverage processing service for EO-1. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(2):408–418, 2013.

[23] CartoDB. https://cartodb.com/platform. Ρετριεεδ 2015.

[24] Jun Chen, Jin Chen, Anping Liao, Xin Cao, Lijun Chen, Xuehong Chen, Chaoying He, Gang Han, Shu Peng, Miao Lu, Weiwei Zhang, Xiaohua Tong και Jon Mills. Global land cover mapping at 30m resolution: A POK-based operational approach. *International Journal of Photogrammetry and Remote Sensing*, 2014.

[25] Jaegul Choo και Haesun Park. Customizing computational methods for visual analytics with big data. *Computer Graphics and Applications, IEEE*, 33(4):22–28, 2013.

[26] J. Cihlar. Land cover mapping of large areas from satellites: Status and research priorities. *Int. J. Remote Sens.*, 21:1093;1114, 2000.

[27] Jeffrey Dean και Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[28] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono και Cees De Laat. Addressing big data challenges for scientific data infrastructure. Στο *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, σελίδες 614–617. IEEE, 2012.

[29] D. Espinoza-Molina και M. Datcu. Earth-observation image retrieval based on content, semantics, and metadata. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(11):5145–5159, 2013.

[30] Konstantinos Evangelidis, Konstantinos Ntouros, Stathis Makridis και Constantine Papatheodorou. Geospatial services in the cloud. *Computers & Geosciences*, 63(0):116 – 122, 2014.

[31] I. Foster, Yong Zhao, I. Raicu και Shiyong Lu. Cloud computing and grid computing 360-degree compared. Στο *Grid Computing Environments Workshop, 2008. GCE '08*, σελίδες 1–10, 2008.

[32] Steffen Fritz, Linda See, Ian McCallum, Christian Schill, Michael Obersteiner, Marijnvan der Velde, Hannes Boettcher, Petr Havlnk και Fridiric Achard. Highlighting continued uncertainty in global land cover maps for the user community. *Environmental Research Letters*, 6(4):044005, 2011.

[33] Borko Furht και Armando Escalante. *Handbook of Cloud Computing*. Springer, 2011.

[34] gigaom.com. Can you predict future traffic patterns? Nokia thinks it can. Στο *https://gigaom.com/2013/07/02/living-cities-lights-up-traffic-in-5-cities-with-interactive-data-visualization/*, Ρετριεεδ 2015.

[35] Jim Gray. Distributed computing economics. *Queue*, 6(3):63–68, 2008.

[36] P. Griffiths, S.van der Linden, T. Kuemmerle και P. Hostert. A pixel-based landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5):2088–2101, 2013.

[37] Salman Habib, Vitali Morozov, Nicholas Frontiere, Hal Finkel, Adrian Pope και Katrin Heitmann. Hacc: Extreme scaling and performance across diverse architectures. Στο *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '13, σελίδες 6:1–6:10, New York, NY, USA, 2013. ACM.

[38] Jing Han, E. Haihong, Guan Le και Jian Du. Survey on nosql database. Στο *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on*, σελίδες 363–366, 2011.

[39] M.C Hansen, R.S DeFries, J.R.G Townshend, R Sohlberg, C Dimiceli και M Carroll. Towards an operational {MODIS} continuous field of percent tree cover algorithm: examples using {AVHRR} and {MODIS} data. *Remote Sensing of Environment*, 83(1;2):303 – 319, 2002.

[40] Weiguo Han, Zhengwei Yang, Liping Di και Peng Yue. A geospatial web service approach for creating on-demand cropland data layer thematic maps. *Transactions of the ASABE*, 57(1):239–247, 2014.

[41] P. N. Happ, R. S. Ferreira, G. A. O. P. Costa, R. Q. Feitosa, C. Bentes και P. Gamba. Towards distributed region growing image segmentation based on mapreduce. Στο *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, σελίδες 4352–4355, 2015.

[42] C. Butson I. Olthof και R. Fraser. Signature extension through space for northern landcover classification: A comparison of radiometric correction methods. *Remote Sens. Environ.*, 95:290;302, 2005.

[43] Stratos Idreos, Fabian Groffen, Niels Nes, Stefan Manegold, Sjoerd Mullender και Martin Kersten. Monetdb: Two decades of research in column-oriented database architectures. *IEEE Data Eng. Bull*, 2012.

[44] Stratos Idreos, Martin L. Kersten και Stefan Manegold. Database cracking. Στο *CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 7-10, 2007, Online Proceedings*, σελίδες 68–78, 2007.

[45] Milena G. Ivanova, Martin L. Kersten, Niels J. Nes και Romulo A.P. Gonçalves. An architecture for recycling intermediates in a column-store. *ACM Trans. Database Syst.*, 35(4):24:1–24:43, 2010.

[46] S. Christensen J. Feranec, G. Hazeu και G. Jaffrain. Corine land cover change detection in europe (case studies of the netherlands and slovakia). *Land Use Policy*, 24:234;247, 2007.

[47] V. C. Radeloff T. Kuemmerle J. Kozak J. Knorn, A. Rabe και P. Hostert. Land cover mapping of large areas using chain classification of neighboring landsat satellite images. *Remote Sens. Environ.*, 113:957;964, 2009.

[48] J. Ju και D. P. Roy. The availability of cloud-free landsat etm plus data over the conterminous united states and globally. *Remote Sens. Environ.*, 112:1196–1211, 2008.

[49] Karthikand Kambatla, Giorgos Kollias, Vipin Kumar και Ananth Grama. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 2013.

[50] K. Karantzalos, D. Bliziotis και A. Karmas. A Scalable Web Geospatial Service for Near Real-Time, High-Resolution Land Cover Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Σπεσιαλ Ισσυε ον Ίβιγ Δατα ιν Ρεμοτε Σενσινγ′, (ιν πρεσς), 2015.

[51] K. Karantzalos, A. Karmas και A. Tzotsos. RemoteAgri: Processing Online Big Earth Observation Data for Precision Agriculture. Στο *European Conference on Precision Agriculture*, 2015.

[52] A. Karmas, K. Karantzalos και S. Athanasiou. Online analysis of remote sensing data for agricultural applications. Στο *OSGeo's European Conference on Free and Open Source Software for Geospatial*, 2014.

[53] A. Karmas, A. Tzotsos και K. Karantzalos. Scalable Geospatial Web Services through Efficient, Online and Near Real-time Processing of Earth Observation Data. Στο *(BigData Service 2015) IEEE International Conference on Big Data Computing Service and Applications*, 2015.

[54] Athanasios Karmas, Angelos Tzotsos και Konstantinos Karantzalos. Scalable Geospatial Web Services through Efficient, Online and Near Real-time Processing of Earth Observation Data. Στο *IEEE Int. Conf. on Big Data Computing Service and Applications*. IEEE, 2015.

[55] Athanasios Karmas, Angelos Tzotsos και Konstantinos Karantzalos. Big Geospatial Data for Environmental and Agricultural Applications. Στο *Big Data Concepts, Theories and Applications*. Springer International Publishing, 2016.

[56] R.T. Kouzes, G.A. Anderson, S.T. Elbert, I. Gorton και D.K. Gracio. The changing paradigm of data-intensive computing. *Computer*, 42(1):26–34, 2009.

[57] C.A. Lee, S.D. Gasster, A. Plaza, Chein I Chang και Bormin Huang. Recent developments in high performance computing for remote sensing: A review. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 4(3):508–527, 2011.

[58] Bingwei Liu, E. Blasch, Yu Chen, Dan Shen και Genshe Chen. Scalable sentiment classification for big data analysis using Naive Bayes Classifier. Στο *Big Data, 2013 IEEE International Conference on*, σελίδες 99–104, 2013.

[59] J. G. Masek, E. F. Vermote, N. E. Saleous, R. Wolfe, F. G. Hall, K. F. Huemmrich, Feng Gao, J. Kutler και Teng Kui Lim. A landsat surface reflectance dataset for north america, 1990-2000. *IEEE Geoscience and Remote Sensing Letters*, 3(1):68–72, 2006.

[60] Y. Ma, L. Wang, P. Liu και R. Ranjan. Towards building a data-intensive index for big data computing - a case study of remote sensing data processing. *Information Sciences*, 2014.

[61] Yan Ma, Lizhe Wang, A.Y. Zomaya, Dan Chen και R. Ranjan. Task-tree based large-scale mosaicking for massive remote sensed imageries with dynamic dag scheduling. *Parallel and Distributed Systems, IEEE Transactions on*, 25(8):2126–2137, 2014.

[62] Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya και Wei Jie. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, (0):–, 2014.

[63] Susanne Mecklenburg. GMES Sentinel Data Policy - An overview. Στο *GENESI-DR (Ground European Network for Earth Science Interoperations - Digital Repositories) workshop, ESAC, Villafranca, Spain*, 2009.

[64] T. Menzies και T. Zimmermann. Software analytics: So what? *Software, IEEE*, 30(4):31–37, 2013.

[65] MonetDB. https://www.monetdb.org/home/features. Ρετριεεδ 2015.

[66] NGA. Digitalglobe application a boon to raster data storage, processing. 2014.

[67] NGA. https://github.com/ngageoint/mrgeo/wiki. Ρετριεεδ 2015.

[68] C. Nikolaou, K. Kyzirakos, K. Bereta, K. Dogani, S. Giannakopoulou, P. Smeros, G. Garbis, M. Koubarakis, D.E. Molina, O.C. Dumitru, G. Schwarz και M. Datcu. Big, linked and open data: Applications in the german aerospace center. Στο *The Semantic Web: ESWC 2014 Satellite Events*, Lecture Notes in Computer Science, σελίδες 444–449. Springer International Publishing, 2014.

[69] J.H.P. Oosthoek, J. Flahaut, A.P. Rossi, P. Baumann, D. Misev, P. Campalani και V. Unnithan. Planetserver: Innovative approaches for the online analysis of hyperspectral satellite data from Mars. *Advances in Space Research*, σελίδες 219–244, 2013.

[70] J. O?Connell J. Wallace P. Caccetta, S. Furby και X. Wu. Continental monitoring: 34 years of land cover change using landsat imagery. Στο *Proc. 32nd Int. Symp. Remote Sens. Environ.*, 2007.

[71] Mary Pax-Lenney, Curtis E. Woodcock, Scott A. Macomber, Sucharita Gopal και Conghe Song. Forest mapping with a generalized classifier and landsat tm data. *Remote Sensing of Environment*, 77(3):241–250, 2001.

[72] Bryan C. Pijanowski, Amin Tayyebi, Jarrod Doucette, Burak K. Pekin, David Braun και James Plourde. A big data urban growth simulation at a national scale: Configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. *Environmental Modelling & Software*, 51(0):250–268, 2014.

[73] Antonio J. Plaza. Special issue on architectures and techniques for real-time processing of remotely sensed images. *J. Real-Time Image Processing*, 4(3):191–193, 2009.

[74] Antonio J. Plaza και Chein I. Chang. *High Performance Computing in Remote Sensing*. Chapman & Hall/ CRC Press, 2007.

[75] D.P. Roy, M.A. Wulder, T.R. Loveland, C.E. Woodcock, R.G. Allen, M.C. Anderson, D. Helder, J.R. Irons, D.M. Johnson, R. Kennedy, T.A. Scambos, C.B. Schaaf, J.R. Schott, Y. Sheng, E.F. Vermote, A.S. Belward, R. Bindschadler, W.B. Cohen, F. Gao, J.D. Hipple, P. Hostert, J. Huntington, C.O. Justice, A. Kilic, V. Kovalskyy, Z.P. Lee, L. Lymburner, J.G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R.H. Wynne και Z. Zhu. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

[76] J. F. Wallace X. Wu J. O?Connell S. Collings A. Traylen S. L. Furby, P. Caccetta και D. Devereaux. Recent developments in landsat-based continental scale land cover change monitoring in australia. Στο *XXI Congr. Int. Soc. Photogrammetry and Remote Sensing*, 2008.

[77] Piotr Szul και Tomasz Bednarz. Productivity frameworks in big data image processing computations - creating photographic mosaics with hadoop and scalding. *Procedia Computer Science*, 29:2306 – 2314, 2014.

[78] M.A. Vouk. Cloud computing 2014; issues, research and implementations. Στο *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on*, σελίδες 31–40, 2008.

[79] J. C. White, M. A. Wulder, G. W. Hobart, J. E. Luther, T. Hermosilla, P. Griffiths, N. C. Coops, R. J. Hall, P. Hostert, A. Dyk και L. Guindon. Pixel-based image compositing for large-area dense time series applications and science. *Canadian Journal of Remote Sensing*, 40(3):192–212, 2014.

[80] Michael A. Wulder, Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland και Curtis E. Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2–10, 2012.

[81] Peng Yue, Liping Di, Yaxing Wei και Weiguo Han. Intelligent services for discovery of complex geospatial features from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* , 83(0):151 – 164, 2013.

[82] Peng Yue, Jianya Gong, Liping Di, Jie Yuan, Lizhi Sun, Ziheng Sun και Qian Wang. Geopw: Laying blocks for the geospatial processing web. *Transactions in GIS*, 14(6):755–772, 2010.

[83] E. Zell, A.K. Huff, A.T. Carpenter και L.A. Friedl. A user-driven approach to determining critical earth observation priorities for societal benefit. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(6):1594–1602, 2012.

[84] Peisheng Zhao, Theodor Foerster και Peng Yue. The Geoprocessing Web. *Computers & Geosciences*, 47(0):3 – 12, 2012.