ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Κατηγοριοποίηση Απόδοσης σε Παραγωγικό Περιβάλλον

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΡΥΣΑΝΘΗ, Κ. ΔΟΥΡΟΥ

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Κατηγοριοποίηση Απόδοσης σε Παραγωγικό Περιβάλλον

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΡΥΣΑΝΘΗ, Κ. ΔΟΥΡΟΥ

Επιβλέπων : Στέφανος Κόλλιας
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10η Ιουλίου 2023.

.....	.....	.....
Στέφανος Κόλλιας	Γεώργιος Στάμου	Αθανάσιος Βουλόδημος
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Χρυσάνθη Κ. Δούρου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

. . . . . .

ΧΡΥΣΑΝΘΗ ΔΟΥΡΟΥ
10 Ιουλίου 2023

Περίληψη

Η παρούσα διπλωματική αφορά την διαχείριση big data και εστιάζει στην υλοποίηση ενός συστήματος κατηγοριοποίησης ψυγείων επιχειρήσεων βάσει της απόδοσής τους. Ο σκοπός της έρευνας είναι να αναπτυχθεί μια μεθοδολογία που να επιτρέπει την ανάλυση και την κατάταξη των ψυγείων βάσει δεδομένων πωλήσεων, πορτών και πληρότητας. Η διαδικασία περιλαμβάνει τον καθαρισμό των εισαγόμενων δεδομένων και την επεξεργασία τους με σκοπό την εφαρμογή των απαραίτητων υπολογισμών. Κατόπιν, γίνεται η ομαδοποίηση των ψυγείων σε εννέα διαφορετικές κατηγορίες, βασιζόμενη στη διαθέσιμη πληροφορία για κάθε ψυγείο. Η μέθοδος χρησιμοποιεί τη μέση τιμή των εβδομαδιαίων πωλήσεων, τη μέση τιμή της κινητικότητας των πορτών, την τελευταία τιμή πληρότητας και τους στόχους πωλήσεων και κινητικότητας πορτών. Επιπλέον, η ανάπτυξη γίνεται σε τρία στρώματα πινάκων για απλούστευση και μελλοντική συντήρηση του κώδικα.

Λέξεις Κλειδιά

διαχείριση μεγάλων δεδομένων, κατηγοριοποίηση ψυγείων, ανάλυση απόδοσης, καθαρισμός δεδομένων, προεπεξεργασία δεδομένων, ομαδοποίηση, μαθηματικοί τύποι, οπτικοποίηση δεδομένων, επιχειρησιακές εργασίες και συμπεράσματα

Abstract

This thesis focuses on big data management and specifically on the implementation of a classification system for business refrigerators based on their performance. The purpose of this research is to develop a methodology that allows for the analysis and categorization of refrigerators based on sales, doors, and occupancy data. The process involves cleaning the imported data and processing it to apply the necessary calculations. Subsequently, the refrigerators are grouped into nine different categories, based on the available information for each refrigerator. The method utilizes the average value of weekly sales, the average value of door mobility, the latest occupancy value, and sales and door mobility targets. Additionally, the development is carried out using three layers of matrices for code simplification and future maintenance.

Key Words

big data management, refrigerator categorization, performance analysis, data cleaning, data preprocessing, clustering, mathematical formulas, data visualization, business insights

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε, καθώς και για τη δυνατότητα που μου έδωσε να εκπονήσω την παρούσα διπλωματική εργασία κάτω από την καθοδήγησή του στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Γνώσης του Εθνικού Μετσόβιου Πολυτεχνείου,.

Επιπροσθέτως, θα ήθελα να εκφράσω τις ευχαριστίες μου στην Ευαγγελία Λαμπάκη, την Χρυσούλα Ρίνου καθώς και τον αδερφό μου Ιωάννη Δούρο για τη διάθεση και την υπομονή που επέδειξαν, το χρόνο που αφιέρωσαν και τις γνώσεις και την έμπνευση που μου μετέδωσαν κατά την προσπάθεια μου. Χωρίς την ανεκτίμητη υποστήριξη και βοήθεια τους, αυτή η διπλωματική δεν θα ήταν δυνατό να γραφεί.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, τους φίλους μου, τους δασκάλους μου καθώς και όλους τους κοντινούς μου ανθρώπους για την αμέριστη στήριξή τους σε κάθε μου βήμα.

Χρυσάνθη Δούρου

Ιούλιος, 2023

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1: Εισαγωγή

Η εποχή της τεχνολογίας Big Data έχει επιφέρει μια επανάσταση στον τρόπο που αντιλαμβανόμαστε και χειριζόμαστε τα δεδομένα στη σύγχρονη κοινωνία. Με την ανάπτυξη των υπολογιστικών τεχνολογιών και της συλλογής δεδομένων από διάφορες πηγές, η τεχνολογία Big Data επιτρέπει στις επιχειρήσεις, τους ερευνητές και τους επαγγελματίες να ανακαλύψουν νέα μοτίβα και συσχετίσεις μεταξύ των δεδομένων.

Ο όρος "Big Data" αναφέρεται στην επεξεργασία και ανάλυση μεγάλου όγκου δεδομένων που δημιουργούνται από διάφορες πηγές, όπως κοινωνικά δίκτυα, ιστοσελίδες, αισθητήρες, συστήματα GPS, συστήματα σταθερής τηλεφωνίας και κινητής τηλεφωνίας, και άλλες πηγές. Η αύξηση του αριθμού των συσκευών που συλλέγουν και αναφέρουν δεδομένα έχει οδηγήσει στη δημιουργία μεγάλων όγκων δεδομένων που απαιτούν ειδικές τεχνικές επεξεργασίας και ανάλυσης.

1.1 Πλεονεκτήματα της τεχνολογίας Big Data

Η χρήση της τεχνολογίας Big Data έχει πολλά πλεονεκτήματα για την κοινωνία, τις επιχειρήσεις και τον δημόσιο τομέα. Πρώτον, η τεχνολογία Big Data μπορεί να βελτιώσει την ανταγωνιστικότητα των επιχειρήσεων και να αυξήσει την παραγωγικότητα τους μέσω της βελτίωσης της ανάλυσης δεδομένων. Δεύτερον, η τεχνολογία Big Data μπορεί να βοηθήσει τις επιχειρήσεις να αναγνωρίσουν τις τάσεις της αγοράς και να προβλέψουν τη ζήτηση, βελτιώνοντας έτσι τη διαχείριση τους. Τρίτον, η τεχνολογία Big Data μπορεί να βοηθήσει στη βελτίωση της διαχείρισης της κυκλοφορίας, της ασφάλειας και της ποιότητας της υγείας. Τέταρτον, η τεχνολογία Big Data μπορεί να βοηθήσει στη βελτίωση της διαχείρισης των δημοσίων υπηρεσιών, όπως η παροχή ηλεκτρικού ρεύματος και η διαχείριση του νερού.

1.2 Προκλήσεις της τεχνολογίας Big Data

Παρά τα πλεονεκτήματα που προσφέρει η τεχνολογία Big Data, υπάρχουν και προκλήσεις που πρέπει να αντιμετωπιστούν. Πρώτον, η ασφάλεια των δεδομένων είναι ένα σημαντικό ζήτημα που πρέπει να ληφθεί υπόψη κατά την επεξεργασία και ανάλυση των δεδομένων. Δεύτερον, η επικοινωνία μεταξύ των διαφόρων συστημάτων και πλατφορμών είναι επίσης ένα ζήτημα που πρέπει να αντιμετωπιστεί. Τρίτον, η έλλειψη εκπαίδευσης και τεχνικών δεξιοτήτων στη χρήση της τεχνολογίας Big Data είναι ένα ζήτημα που πρέπει να αντιμετωπιστεί.

1.3 Εφαρμογές Big Data

Η τεχνολογία Big Data χρησιμοποιείται σε διάφορους τομείς, όπως η υγεία, η παιδεία, η έρευνα, οι μεταφορές και ο τουρισμός όπως περιγράφονται συνοπτικά παρακάτω:

- **Υγεία**
  Η τεχνολογία Big Data έχει επανασχεδιάσει τον τρόπο παροχής υπηρεσιών υγείας. Με τη βοήθεια της ανάλυσης μεγάλων δεδομένων, οι πάροχοι υπηρεσιών υγείας μπορούν να προβλέπουν εξάρσεις ασθενειών, να εντοπίζουν ασθενείς με υψηλό κίνδυνο, να εξατομικεύουν τις θεραπείες και να βελτιώνουν τα αποτελέσματα των ασθενών. Τα ιατρικά αρχεία, τα δεδομένα κλινικών δοκιμών, τα γενομικά δεδομένα και τα δεδομένα της φαρμακευτικής βιομηχανίας μπορούν να χρησιμοποιηθούν για να βελτιώσουν την πρόληψη και τη θεραπεία ασθενειών και να βελτιώσουν την υγεία των ασθενών.

- **Εκπαίδευση**
  Η τεχνολογία Big Data μπορεί να βοηθήσει στη βελτίωση της εκπαίδευσης. Με την ανάλυση των δεδομένων της ακαδημαϊκής επίδοσης των μαθητών και των φοιτητών, οι εκπαιδευτικοί μπορούν να προβλέψουν τις ανάγκες των μαθητών και να προσαρμόσουν τα μαθήματα και τη διδασκαλία στις ανάγκες τους. Επιπλέον, η ανάλυση των δεδομένων μπορεί να βοηθήσει στη βελτίωση της απόδοσης των μαθητών και στην αναγνώριση των περιοχών που χρειάζονται περισσότερη προσοχή. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει στην ανάπτυξη προσωποποιημένων μαθημάτων και στη βελτίωση της εμπειρίας των φοιτητών με τη χρήση προηγμένων τεχνολογιών, όπως οι εικονικές πραγματικότητες και η τεχνητή νοημοσύνη.

- **Έρευνα**
  Η τεχνολογία Big Data μπορεί να βοηθήσει στην πρόοδο της έρευνας σε πολλούς τομείς, όπως η ιατρική, η βιολογία, η φυσική, η χημεία και η αστρονομία. Με τη χρήση της τεχνολογίας Big Data, οι ερευνητές μπορούν να αναγνωρίζουν μοτίβα και συσχετίσεις στα δεδομένα, να αναπτύσσουν μοντέλα πρόβλεψης και να ανακαλύπτουν νέες γνώσεις και ανακαλύψεις. Η τεχνολογία Big Data μπορεί επίσης να βοηθήσει τους ερευνητές να κατανοήσουν τις προτιμήσεις και τις συμπεριφορές των ανθρώπων και να βελτιώσουν τα προϊόντα και τις υπηρεσίες τους σύμφωνα με τις ανάγκες τους. Η τεχνολογία Big Data μπορεί επίσης να βοηθήσει στην ανάπτυξη νέων ιατρικών φαρμάκων και θεραπειών, καθώς και στη βελτίωση της πρόληψης και της αντιμετώπισης ασθενειών.

- **Τουρισμός**
  Η τεχνολογία Big Data μπορεί να βοηθήσει στην ανάπτυξη και τη βελτίωση του τουρισμού. Με τη χρήση της τεχνολογίας Big Data, οι επιχειρήσεις μπορούν να αναγνωρίζουν τις τάσεις και τα μοτίβα των ταξιδιωτών, να προσαρμόζουν την προσφορά τους στις ανάγκες των ταξιδιωτών και να βελτιώνουν την εμπειρία των επισκεπτών. Η

τεχνολογία Big Data μπορεί επίσης να βοηθήσει τους ταξιδιωτές να βρουν προσαρμοσμένες προτάσεις για το προορισμό τους, μέσω διαδραστικών χαρτών, συστάσεων και εκδηλώσεων. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει στην πρόβλεψη της κίνησης τουριστών και στον σχεδιασμό καλύτερων στρατηγικών προώθησης και μάρκετινγκ.

- **Μεταφορές**
  Η τεχνολογία Big Data μπορείνα βοηθήσει στη βελτίωση του συστήματος μεταφορών και στη μείωση της κυκλοφοριακής συμφόρησης. Με τη χρήση της τεχνολογίας Big Data, οι μεταφορικές εταιρείες μπορούν να παρακολουθούν τα δεδομένα κυκλοφορίας και να προβλέπουν την κίνηση των οχημάτων, την κατανομή τους και τις συνθήκες των δρόμων. Αυτό μπορεί να βοηθήσει στη βελτίωση της απόδοσης του συστήματος μεταφορών και στην αύξηση της αποτελεσματικότητας του. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει στην ανάπτυξη και χρήση των δημόσιων μεταφορών, καθώς και στην προσαρμογή τους στις ανάγκες των επιβατών. Με την ανάλυση των δεδομένων των επιβατών, οι μεταφορικές εταιρείες μπορούν να βελτιώσουν την εμπειρία των επιβατών, να μειώσουν τους χρόνους αναμονής και να παρέχουν πιο αποτελεσματικές υπηρεσίες.

Η τεχνολογία Big Data αντιπροσωπεύει ένα σημαντικό εργαλείο για την επεξεργασία και ανάλυση των μεγάλων όγκων δεδομένων που δημιουργούνται στη σημερινή κοινωνία. Η χρήση της τεχνολογίας Big Data έχει πολλά πλεονεκτήματα για την κοινωνία, τις επιχειρήσεις και τον δημόσιο τομέα. Παράλληλα, υπάρχουν και προκλήσεις που πρέπει να αντιμετωπιστούν, όπως η ασφάλεια των δεδομένων και η έλλειψη εκπαίδευσης και τεχνικών δεξιοτήτων στη χρήση της τεχνολογίας. Για να αξιοποιηθεί πλήρως η τεχνολογία Big Data, πρέπει να ληφθούν υπόψη οι παραπάνω προκλήσεις και να αναπτυχθούν λύσεις που θα βοηθήσουν στην αντιμετώπισή τους. Πρέπει επίσης να δοθεί έμφαση στην εκπαίδευση και την ανάπτυξη τεχνικών δεξιοτήτων στη χρήση της τεχνολογίας Big Data, ώστε να επιτευχθεί η πλήρης αξιοποίησή της.

Τέλος, η τεχνολογία Big Data παρέχει μια μεγάλη ευκαιρία για τη βελτίωση της απόδοσης των επιχειρήσεων και την ανάπτυξη της οικονομίας της χώρας. Με τη σωστή χρήση της τεχνολογίας Big Data, η κάθε χώρα μπορεί να γίνει πιο ανταγωνιστική και να αυξήσει την παραγωγικότητα των επιχειρήσεων της. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει στη βελτίωση της διαχείρισης των δημοσίων υπηρεσιών και να συμβάλει στη βελτίωση της ποιότητας ζωής των πολιτών.

1.4 Big Data στις επιχειρήσεις

Οι επιχειρήσεις χρησιμοποιούν την τεχνολογία Big Data για να αντλήσουν πολύτιμες πληροφορίες από τα δεδομένα και να βελτιώσουν τις επιχειρηματικές τους δραστηριότητες σε

διάφορους τομείς, όπως η μάρκετινγκ, οι πωλήσεις, οι χρηματοοικονομικές δραστηριότητες, η λογιστική και η διοίκηση.

Ένα από τα πιο κοινά παραδείγματα χρήσης της τεχνολογίας Big Data από τις επιχειρήσεις είναι η ανάλυση των στοιχείων των πελατών. Με τη χρήση δεδομένων από κοινωνικά δίκτυα, ιστοσελίδες και άλλες πηγές, οι επιχειρήσεις μπορούν να αναγνωρίσουν τις προτιμήσεις των πελατών, τις συνήθειες τους και τις ανάγκες τους. Με αυτόν τον τρόπο, οι επιχειρήσεις μπορούν να προσαρμόσουν τις δραστηριότητές τους στις ανάγκες των πελατών τους και να βελτιώσουν την εξυπηρέτησή τους.

Η τεχνολογία Big Data χρησιμοποιείται επίσης για την πρόβλεψη των τάσεων της αγοράς και των πωλήσεων. Με τη χρήση αλγορίθμων μηχανικής μάθησης και ανάλυσης δεδομένων, οι επιχειρήσεις μπορούν να προβλέψουν τα προϊόντα και τις υπηρεσίες που θα έχουν μεγαλύτερη ζήτηση και να προσαρμόσουν αναλόγως το απόθεμά τους και τις τιμές τους. Με αυτόν τον τρόπο, οι επιχειρήσεις μπορούν να βελτιώσουν την απόδοσή τους και να αυξήσουν τα κέρδη τους. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει τις επιχειρήσεις να προβλέψουν πιθανούς κινδύνους και προβλήματα στις δραστηριότητές τους. Με τη συλλογή και την ανάλυση δεδομένων από διάφορες πηγές, όπως συστήματα ασφαλείας, δεδομένα παραγωγής και δεδομένα ενεργειακής κατανάλωσης, οι επιχειρήσεις μπορούν να προβλέψουν τα προβλήματα και να λάβουν μέτρα πριν εκδηλωθούν. Αυτό μπορεί να βοηθήσει στη μείωση των κινδύνων και στην αποφυγή καταστροφικών καταστάσεων.

Η τεχνολογία Big Data μπορεί επίσης να βελτιώσει την αποτελεσματικότητα των διαδικασιών λογιστικής και διοίκησης των επιχειρήσεων. Με τη συλλογή και την ανάλυση των δεδομένων, οι επιχειρήσεις μπορούν να βελτιώσουν τη διαχείριση των αποθεμάτων, τη διαχείριση των χρηματοοικονομικών μετρήσεων και τον εντοπισμό των αδυναμιών και των πιθανών προβλημάτων στις επιχειρηματικές διαδικασίες. Επιπλέον, η τεχνολογία Big Data μπορεί να βελτιώσει τη διαχείριση των χρηματοοικονομικών μετρήσεων των επιχειρήσεων. Με τη χρήση δεδομένων από διάφορες πηγές, όπως οι τραπεζικοί λογαριασμοί και οι πιστωτικές κάρτες, οι επιχειρήσεις μπορούν να αναλύσουν τις χρηματοοικονομικές τους επιδόσεις και να προβλέψουν τις μελλοντικές τους ανάγκες σε κεφάλαιο και χρηματοδότηση. Αυτό μπορεί να βοηθήσει τις επιχειρήσεις να λαμβάνουν πιο ορθές αποφάσεις για τις επενδύσεις, τη χρηματοδότηση και τη διαχείριση των χρηματοοικονομικών κινήσεων τους.

Η χρήση της τεχνολογίας Big Data μπορεί να βοηθήσει στην εντοπισμό αδυναμιών στις αλυσίδες εφοδιαστικής, τις οποίες μπορούν να επιλύσουν οι επιχειρήσεις με τη βελτίωση της παραγωγής, τη διαχείριση των αποθεμάτων και των παραγγελιών. Επίσης, η ανάλυση των δεδομένων μπορεί να βοηθήσει στον εντοπισμό των πιθανών απωλειών, τη βελτίωση του χρόνου παράδοσης και την αύξηση της αποδοτικότητας των επιχειρήσεων. Επιπλέον, μπορεί να βοηθήσει τις επιχειρήσεις να εντοπίζουν και να αντιμετωπίζουν τα προβλήματα της απάτης και της κακοδιαχείρισης δεδομένων. Η ανάλυση των δεδομένων μπορεί να βοηθήσει στην

αναγνώριση των συμπτωμάτων και των δεικτών της απάτης, και στη λήψη των απαραίτητων μέτρων για την αντιμετώπισή της. Επιπλέον, η τεχνολογία Big Data μπορεί να βοηθήσει τις επιχειρήσεις στην τήρηση των διατάξεων των κανονιστικών αρχών και στην προστασία των προσωπικών δεδομένων των πελατών τους.

Συνολικά, η τεχνολογία Big Data παρέχει στις επιχειρήσεις τη δυνατότητα να αναλύουν και να επεξεργάζονται μεγάλες ποσότητες δεδομένων από διάφορες πηγές, ώστε να λαμβάνουν πιο έξυπνες και αποτελεσματικές αποφάσεις. Η ανάλυση των δεδομένων μπορεί να βοηθήσει στη βελτίωση της παραγωγικότητας, της ποιότητας των προϊόντων και των υπηρεσιών, της απόδοσης και της αποτελεσματικότητας των επιχειρήσεων.

1.5 Στόχος

Στόχος της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος ανάλυσης δεδομένων με σκοπό την εκμετάλλευση του από επιχειρήσεις για την αναγνώριση της παραγωγικής αποδοτικότητας τους. Χρησιμοποιώντας αληθινά δεδομένα για τις πωλήσεις των ψυγείων μιας επιχείρησης καθώς και επιπλέον πληροφορίες για τα προϊόντα, μέσω κατάλληλων μετασχηματισμών θα κατηγοριοποιήσουμε τα ψυγεία σε κατηγορίες απόδοσης, με σκοπό να βοηθήσουν τους χρήστες να αντιληφθούν πιθανά προβλήματα ώστε να προβούν έγκαιρα στις απαραίτητες ενέργειες και να βελτιστοποιήσουν την απόδοση και τελικά το κέρδος τους.

Η παρούσα διπλωματική οργανώνεται ως εξής: Στο κεφάλαιο 2 θα περιγράψουμε τα δημοφιλέστερα εργαλεία που χρησιμοποιούνται ευρέως για επεξεργασία και ανάλυση big data, και θα παραθέσουμε συνοπτικά τις διαφορές των δυνατοτήτων τους. Στο κεφάλαιο 3 θα αναλύσουμε τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση του συστήματός μας. Στο κεφάλαιο 4 θα δοθεί μια αναλυτική περιγραφή της λογικής του αλγορίθμου και στη συνέχεια θα περιγραφεί Βήματικά η όλη διαδικασία παραγωγής του τελικού αποτελέσματος. Στη συνέχεια, στο κεφάλαιο 5 θα περιγραφεί η διαδικασία εκτέλεσης των πειραμάτων και η παραγωγή του τελικού output μεσώ του pipeline, ενώ στο κεφάλαιο 6 θα παρουσιαστούν και θα σχολιαστούν τα αποτελέσματα του κώδικα μας και πιθανές επεκτάσεις.

Κεφάλαιο 2: Σχετικά εργαλεία

Για την υλοποίηση ενός ολοκληρωμένου συστήματος ανάλυσης δεδομένων είναι απαραίτητη η χρήση ορισμένων εργαλείων τα οποία να μας παρέχουν συγκεκριμένες δυνατότητες όπως (α) αποθηκευτικό χώρο για την αποθήκευση των δεδομένων εισόδου καθώς και των πινάκων μας, (β) ένα τρόπο αποθήκευσης και διαχείρισης του κώδικά μας έτσι ώστε να είμαστε σε θέση να διατηρηθεί το ιστορικό των αλλαγών και να είναι ορατός ο κώδικας σε κάθε χρήστη, (γ) ένα περιβάλλον με το κατάλληλο cluster για την εκτέλεση του κώδικα και (δ) αυτόματη εκτέλεση του κώδικα και παραγωγή του τελικού αποτελέσματος. Στην επόμενη ενότητα θα περιγραφούν τα δημοφιλέστερα διαθέσιμα εργαλεία που εξυπηρετούν τους ανωτέρω σκοπούς.

2.1 Αποθήκευση Δεδομένων

Υπάρχουν αρκετές εναλλακτικές λύσεις όσον αφορά την αποθήκευση των δεδομένων, ανάλογα με τις ανάγκες της κάθε περίστασης. Μερικές από αυτές τις επιλογές είναι:

- Amazon S3

Το Amazon S3 είναι μια από τις πιο δημοφιλείς υπηρεσίες αποθήκευσης στο cloud, που παρέχει αποθήκευση αντικειμένων, αποθήκευση αρχείων και αποθήκευση block. Αυτή η υπηρεσία παρέχει επίσης δυνατότητες διαχείρισης δεδομένων, εργαλεία ασφάλειας και ενσωματωμένη υποστήριξη CDN.

- Google Cloud Storage

Το Google Cloud Storage είναι μια υπηρεσία αποθήκευσης στο cloud που παρέχει δυνατότητες αποθήκευσης αντικειμένων, αρχείων και block. Αυτή η υπηρεσία παρέχει επίσης δυνατότητες διαχείρισης δεδομένων, εργαλεία ασφάλειας και ενσωματωμένη υποστήριξη CDN.

- IBM Cloud Object Storage

Το IBM Cloud Object Storage είναι μια υπηρεσία αποθήκευσης στο cloud που παρέχει αποθήκευση αντικειμένων και αρχείων, καθώς και δυνατότητες διαχείρισης δεδομένων και εργαλεία ασφάλειας. Αυτή η υπηρεσία προσφέρει επίσης υποστήριξη για πολλαπλές εφαρμογές και περιβάλλον αποθηκευτικών χώρων.

- Backblaze B2

Το Backblaze B2 είναι μια οικονομική υπηρεσία αποθήκευσης στο cloud που παρέχει αποθήκευση αντικειμένων, αρχείων και block. Αυτή η υπηρεσία προσφέρει επίσης δυνατότητες διαχείρισης δεδομένων και εργαλεία ασφάλειας.

- Wasabi

Το Wasabi είναι μια υπηρεσία αποθήκευσης στο cloud που προσφέρει αποθήκευση αντικειμένων με απλή και διαφανή τιμολόγηση. Αυτή η υπηρεσία παρέχει επίσης δυνατότητες διαχείρισης δεδομένων και εργαλεία ασφάλειας.

- Microsoft Azure Storage

Η υπηρεσία αποθήκευσης αντικειμένων της Microsoft παρέχει επίσης αξιόπιστη και ευέλικτη αποθήκευση στο cloud, με ενσωματωμένη υποστήριξη CDN και εργαλεία διαχείρισης δεδομένων και ασφαλείας.

Αν και όλες αυτές οι υπηρεσίες προσφέρουν αποθήκευση αντικειμένων στο cloud, υπάρχουν κάποιες διαφορές μεταξύ τους που θα πρέπει να ληφθούν υπόψη κατά την επιλογή της κατάλληλης υπηρεσίας αποθήκευσης για τις επίτευξη του επιθυμητού στόχου. Για παράδειγμα, το Amazon S3 προσφέρει αξιόπιστη και ασφαλή αποθήκευση αντικειμένων στο cloud, ενώ παρέχει επίσης ενσωματωμένη υποστήριξη CDN, κρυπτογράφηση δεδομένων και διαχείριση πιστοποιητικών SSL. Επιπλέον, οι χρήστες μπορούν να επιλέξουν από διάφορα επίπεδα αντοχής και αξιοπιστίας για να προσαρμόσουν τη λύση αποθήκευσης στις ανάγκες τους. Το Google Cloud Storage προσφέρει εξίσου αξιόπιστη και ασφαλή αποθήκευση αντικειμένων στο cloud και ενσωματώνει επίσης εργαλεία διαχείρισης δεδομένων και ανάλυσης δεδομένων. Επιπλέον, η υπηρεσία παρέχει εργαλεία ασφαλείας, όπως κρυπτογράφηση δεδομένων και διαχείριση πιστοποιητικών SSL, που διασφαλίζουν την ασφαλή αποθήκευση και μετάδοση των δεδομένων των χρηστών. Το IBM Cloud Object Storage προσφέρει αποθήκευση αντικειμένων στο cloud με δυνατότητες προσαρμογής και κλιμακωσιμότητας. Επιπλέον, η υπηρεσία παρέχει ευέλικτες επιλογές κατανομής δεδομένων και διαθέτει εργαλεία ασφαλείας και διαχείρισης. Το Wasabi από τη μεριά του προσφέρει χαμηλό κόστος αποθήκευσης αντικειμένων στο cloud χωρίς κρυπτογράφηση δεδομένων. Επιπλέον, η υπηρεσία παρέχει αξιόπιστη αντοχή και αποδοτικότητα στη διαχείριση των δεδομένων των χρηστών. Το Backblaze B2 προσφέρει επίσης αξιόπιστη και ασφαλή αποθήκευση αντικειμένων στο cloud, με δυνατότητες κλιμάκωσης και εύκολη προσαρμογή στις ανάγκες των χρηστών. Το Azure Storage της Microsoft προσφέρει αξιόπιστη και ευέλικτη αποθήκευση αντικειμένων στο cloud, με ενσωματωμένη υποστήριξη CDN και εργαλεία διαχείρισης δεδομένων και ασφαλείας.

Οι διαφορές μεταξύ αυτές υπηρεσιών αποθήκευσης αντικειμένων στο cloud έγκεινται στις λειτουργίες και τις επιλογές που προσφέρουν. Για παράδειγμα, το Amazon S3 και το Google Cloud Storage παρέχουν εργαλεία ανάλυσης δεδομένων, ενώ το IBM Cloud Object Storage προσφέρει δυνατότητες κατανομής δεδομένων σε πολλαπλές τοποθεσίες. Από την άλλη, το Wasabi προσφέρει χαμηλότερο κόστος αποθήκευσης αντικειμένων σε σχέση με το Amazon S3 και το Google Cloud Storage, ενώ το Backblaze B2 προσφέρει επίσης χαμηλό κόστος αλλά δεν έχει τόσο ευέλικτες επιλογές όσο άλλες υπηρεσίες. Το Azure Storage παρέχει επίσης ευέλικτες επιλογές και εργαλεία διαχείρισης, χωρίς να θυσιάζει τον τομέα της ασφάλειας των δεδομένων.

2.2 Διαχείριση / Αποθήκευση Κώδικα

Υπάρχουν αρκετές επιλογές για τη διαχείριση κώδικα, ανάλογα τις ανάγκες. Κάποιες από τις δημοφιλέστερες είναι:

- GitHub

Η πλατφόρμα GitHub προσφέρει αξιόπιστη και εύκολη διαχείριση κώδικα, με δυνατότητες συνεργασίας και εργαλεία παρακολούθησης των αλλαγών στον κώδικα.

- GitLab

Η πλατφόρμα GitLab προσφέρει επίσης διαχείριση κώδικα με δυνατότητες συνεργασίας και εργαλεία παρακολούθησης των αλλαγών στον κώδικα, καθώς και ενσωματωμένες δυνατότητες CI/CD.

- Bitbucket

Η πλατφόρμα Bitbucket προσφέρει διαχείριση κώδικα με δυνατότητες συνεργασίας και εργαλεία παρακολούθησης των αλλαγών στον κώδικα, καθώς και ενσωματωμένες δυνατότητες CI/CD.

- AWS CodeCommit

Η υπηρεσία CodeCommit της Amazon παρέχει αξιόπιστη διαχείριση κώδικα με δυνατότητες συνεργασίας και εργαλεία παρακολούθησης των αλλαγών στον κώδικα, καθώς και ενσωματωμένες δυνατότητες CI/CD.

- Azure Repos

Το Azure Repos είναι μια υπηρεσία διαχείρισης κώδικα που προσφέρεται από τη Microsoft και διαθέτει επιλογές διαχείρισης κλαδιών και παρακολούθησης των αλλαγών στον κώδικα. Επιπλέον, παρέχει δυνατότητες CI/CD και εργαλεία ελέγχου πρόσβασης για τα μέλη της ομάδας.

Τα παραπάνω εργαλεία διαχείρισης κώδικα που προσφέρουν παρόμοιες δυνατότητες, ωστόσο έχουν κάποιες διαφορές στα χαρακτηριστικά τους και στη λειτουργικότητά τους. Παρακάτω περιγράφονται κάποια από τα βασικά χαρακτηριστικά τους:

GitHub: Το GitHub είναι η πιο δημοφιλής υπηρεσία διαχείρισης κώδικα που προσφέρει εύκολη συνεργασία μεταξύ των μελών μιας ομάδας και παρακολούθηση των αλλαγών στον κώδικα. Διαθέτει επίσης εκτεταμένα εργαλεία για τη διαχείριση του κώδικα, όπως η δυνατότητα διαχείρισης pull requests, η δυνατότητα δημιουργίας wiki και η δυνατότητα προβολής του κώδικα στον πλοηγό.

GitLab: Το GitLab είναι μια παρόμοια υπηρεσία με το GitHub που προσφέρει περισσότερα εργαλεία διαχείρισης κώδικα και δυνατότητες CI/CD. Επίσης, διαθέτει επιλογές για τη

διαχείριση επαφών και τη δημιουργία groups, και επιτρέπει την εγκατάσταση της υπηρεσίας σε ιδιωτικούς διακομιστές.

Bitbucket: Το Bitbucket είναι μια υπηρεσία διαχείρισης κώδικα που ανήκει στην Atlassian και προσφέρει δυνατότητες συνεργασίας και ελέγχου πρόσβασης για ομάδες εργασίας. Διαθέτει επίσης εργαλεία διαχείρισης pull requests και δυνατότητες παρακολούθησης του κώδικα.

CodeCommit: Το CodeCommit είναι μια υπηρεσία διαχείρισης κώδικα που προσφέρεται από την Amazon Web Services και διαθέτει επιλογές διαχείρισης κλαδιών, δυνατότητες παρακολούθησης και εργαλεία ελέγχου πρόσβασης. Επιπλέον, προσφέρει ενσωματωμένες δυνατότητες CI/CD για τη διαχείριση του κώδικα.

Azure Repos: Το Azure Repos είναι μια υπηρεσία διαχείρισης κώδικα που προσφέρεται από τη Microsoft και διαθέτει επιλογές διαχείρισης κλαδιών και παρακολούθησης των αλλαγών στον κώδικα. Επιπλέον, παρέχει δυνατότητες CI/CD και εργαλεία ελέγχου πρόσβασης για τα μέλη της ομάδας μέσω της υπηρεσίας DevOps της Microsoft καθώς επίσης ευέλικτες επιλογές διαχείρισης κώδικα και διαδικασίες ανάπτυξης λογισμικού, και εργαλεία παρακολούθησης και ανάλυσης δεδομένων.

2.3 Επεξεργασία Δεδομένων / Περιβάλλον Εκτέλεσης

Υπάρχουν πολλές διαθέσιμες επιλογές όσον αφορά το περιβάλλον εκτέλεσης του κώδικα, ανάλογα με τις συγκεκριμένες ανάγκες της κάθε περίπτωσης. Μερικές από αυτές τις επιλογές είναι:

- AWS EMR

Το AWS EMR είναι μια υπηρεσία διαχείρισης και επεξεργασίας μεγάλων δεδομένων στο Amazon Web Services (AWS) και παρέχει επαγγελματικές λύσεις για τη δημιουργία και τη διαχείριση cluster για την επεξεργασία μεγάλων δεδομένων.

- Google Cloud Dataproc

Το Google Cloud Dataproc είναι μια υπηρεσία επεξεργασίας μεγάλων δεδομένων που λειτουργεί στο cloud της Google και προσφέρει παρόμοιες δυνατότητες με το Amazon EMR.

- Databricks

Το Databricks είναι μια υπηρεσία επεξεργασίας μεγάλων δεδομένων που λειτουργεί στο cloud και προσφέρει εργαλεία για την ανάλυση και επεξεργασία μεγάλων δεδομένων.

- Cloudera

Το Cloudera είναι μια υπηρεσία επεξεργασίας μεγάλων δεδομένων που προσφέρει επαγγελματικές λύσεις για τη διαχείριση και επεξεργασία μεγάλων δεδομένων.

Οι διαφορές μεταξύ του Azure Databricks, του AWS EMR, του Cloudera και του Google Cloud Dataproc στη διαχείριση και επεξεργασία μεγάλων δεδομένων είναι οι εξής:

Azure Databricks: Το Azure Databricks είναι ένας συνδυασμός του Apache Spark και του Delta Lake και παρέχει επαγγελματικές λύσεις για τη δημιουργία και τη διαχείριση cluster με στόχο την επεξεργασία και ανάλυση μεγάλων δεδομένων στο cloud της Microsoft Azure.

AWS EMR: Το AWS EMR είναι μια υπηρεσία διαχείρισης και επεξεργασίας μεγάλων δεδομένων στο Amazon Web Services (AWS) και παρέχει επαγγελματικές λύσεις για τη δημιουργία και τη διαχείριση cluster για την επεξεργασία μεγάλων δεδομένων.

Cloudera: Το Cloudera είναι μια υπηρεσία διαχείρισης και επεξεργασίας μεγάλων δεδομένων που προσφέρει επαγγελματικές λύσεις για τη διαχείριση και επεξεργασία μεγάλων δεδομένων.

Google Cloud Dataproc: Το Google Cloud Dataproc είναι μια υπηρεσία διαχείρισης και επεξεργασίας μεγάλων δεδομένων στο cloud της Google και παρέχει παρόμοιες δυνατότητες με το AWS EMR.

2.4 Αυτόματη Εκτέλεση / Pipeline

Για την αυτοματοποιημένη εκτέλεση του κώδικα και την απαραίτητη διαχείριση δεδομένων υπάρχει μια ευρεία γκάμα εργαλείων. Ορισμένα από αυτά είναι τα ακόλουθα:

- AWS Glue

Το AWS Glue είναι μια υπηρεσία επεξεργασίας και διαχείρισης δεδομένων της Amazon Web Services που προσφέρει επαγγελματικές λύσεις για την επεξεργασία μεγάλων όγκων δεδομένων.

- Google Cloud Dataflow

Το Google Cloud Dataflow είναι μια υπηρεσία επεξεργασίας δεδομένων στο cloud της Google που προσφέρει λύσεις για τη διαχείριση και επεξεργασία μεγάλων δεδομένων.

- Talend

Το Talend είναι μια ανοιχτού κώδικα υπηρεσία διαχείρισης και επεξεργασίας δεδομένων που προσφέρει λύσεις για την επεξεργασία μεγάλων όγκων δεδομένων σε πολλαπλές πλατφόρμες.

- Apache Nifi

Το Apache Nifi είναι μια υπηρεσία διαχείρισης και επεξεργασίας δεδομένων ανοιχτού κώδικα που προσφέρει λύσεις για την επεξεργασία μεγάλων όγκων δεδομένων.

- Azure Data Factory

Το Azure Data Factory είναι μια υπηρεσία διαχείρισης και μεταφοράς δεδομένων στο Azure. Χρησιμοποιείται για την εξαγωγή, μετατροπή και φόρτωση δεδομένων από διάφορες πηγές δεδομένων στο Azure, καθώς και για τη δημιουργία και την εκτέλεση εργασιών ETL.

Κάθε υπηρεσία έχει τα δικά της πλεονεκτήματα και μειονεκτήματα και η επιλογή πρέπει να γίνεται βάσει των αναγκών της κάθε επιχείρησης. Για παράδειγμα, το Azure Data Factory είναι ένα εύχρηστο εργαλείο κατάλληλο για ETL διαδικασίες. Το AWS Glue και το Google Cloud Dataflow είναι ανταγωνιστές του Azure Data Factory και προσφέρουν παρόμοιες δυνατότητες για τη διαχείριση και την επεξεργασία μεγάλων όγκων δεδομένων. Το Talend είναι μια εναλλακτική λύση ανοιχτού κώδικα, που προσφέρει ευελιξία και προσαρμοστικότητα στις ανάγκες της κάθε επιχείρησης. Το Apache Nifi προσφέρει εξειδικευμένες λύσεις για τη διαχείριση δεδομένων στον τομέα της ασφάλειας και του IoT.

Όλες οι παραπάνω υπηρεσίες είναι σχεδιασμένες για τη διαχείριση και την επεξεργασία μεγάλων όγκων δεδομένων, ωστόσο, υπάρχουν ορισμένες διαφορές στις λειτουργίες και τις δυνατότητες που παρέχουν. Ας δούμε πιο αναλυτικά τις διαφορές ανάμεσα στις υπηρεσίες αυτές:

AWS Glue: Το AWS Glue είναι μια υπηρεσία ETL (Extract, Transform, Load), η οποία επιτρέπει στους χρήστες να δημιουργούν και να εκτελούν εργασίες ETL στο AWS. Το Glue χρησιμοποιείται για τη διασύνδεση και την επεξεργασία μεγάλων όγκων δεδομένων από διάφορες πηγές δεδομένων.

Google Cloud Dataflow: Το Google Cloud Dataflow είναι μια υπηρεσία επεξεργασίας δεδομένων σε πραγματικό χρόνο, η οποία χρησιμοποιείται για τη διαχείριση και την επεξεργασία μεγάλων όγκων δεδομένων σε πραγματικό χρόνο. Το Dataflow επιτρέπει την επεξεργασία των δεδομένων με τη χρήση προγραμματιστικών γλωσσών, όπως η Java και η Python.

Talend: Το Talend είναι μια πλατφόρμα επεξεργασίας δεδομένων ανοιχτού κώδικα, η οποία υποστηρίζει τη διασύνδεση και την επεξεργασία δεδομένων από διάφορες πηγές δεδομένων. Το Talend παρέχει επίσης εργαλεία διαχείρισης δεδομένων και δυνατότητες οπτικοποίησης για την επεξεργασία δεδομένων σε πραγματικό χρόνο.

Apache Nifi: Το Apache Nifi είναι μια πλατφόρμα επεξεργασίας δεδομένων ανοιχτού κώδικα, η οποία χρησιμοποιείται για τη διασύνδεση και την επεξεργασία δεδομένων από διάφορες πηγές δεδομένων. Το Nifi υποστηρίζει τη διασύνδεση μεταξύ διαφορετικών πηγών δεδομένων και την επεξεργασία των δεδομένων σε πραγματικό χρόνο.

Azure Data Factory: Το Azure Data Factory είναι μια υπηρεσία διαχείρισης και μεταφοράς δεδομένων στο Azure. Χρησιμοποιείται για την εξαγωγή, μετατροπή και φόρτωση δεδομένων από διάφορες πηγές δεδομένων στο Azure, καθώς και για τη δημιουργία και την εκτέλεση εργασιών ETL.

Οι τελική επιλογή ανάμεσα στις υπηρεσίες αυτές εξαρτώνται από τις απαιτήσεις του συγκεκριμένου σεναρίου χρήσης και τις προτεραιότητες του χρήστη. Για παράδειγμα, το AWS Glue και το Azure Data Factory είναι κατάλληλα για εργασίες ETL, ενώ το Google Cloud Dataflow είναι κατάλληλο για την επεξεργασία δεδομένων σε πραγματικό χρόνο. Επίσης, το Talend και το Apache Nifi είναι εργαλεία επεξεργασίας δεδομένων ανοιχτού κώδικα που παρέχουν ευελιξία και προσαρμοστικότητα στους χρήστες.

Βλέπουμε ότι παρότι υπάρχει πληθώρα εναλλακτικών εργαλείων για την ανάπτυξη ενός συστήματος ανάλυσης δεδομένων σε cloud, τρεις εταιρείες έχουν αναπτύξει ολοκληρωμένα πακέτα λύσεων για τον συγκεκριμένο σκοπό. Η Microsoft παρέχει μια πλήρη λύση μέσω του Azure Cloud Services, η Google μέσω του Google Cloud Services, καθώς επίσης και η Amazon προσφέρει end to end solution με την AWS (Amazon Web Services). Οι υπηρεσίες του Azure Cloud αποτελεί μια εξαιρετική επιλογή για τους σκοπούς της εργασίας μας αφού προσφέρουν αρκετά πλεονεκτήματα σε σχέση με τις υπηρεσίες της Google Cloud και της AWS, καθιστώντας το μια δημοφιλή επιλογή για επιχειρήσεις που αναζητούν μια πλατφόρμα υπολογιστικού νέφους που ανταποκρίνεται στις συγκεκριμένες ανάγκες τους.

Μερικά από τα πλεονεκτήματα των υπηρεσιών Azure Cloud σε σχέση με τις υπηρεσίες της Google Cloud και της AWS:

1. Ενσωμάτωση με προϊόντα της Microsoft: Οι υπηρεσίες του Azure Cloud ενσωματώνονται απρόσκοπτα με άλλα προϊόντα της Microsoft, όπως τα Windows Server και το SQL Server, καθιστώντας το μια ελκυστική επιλογή για επιχειρήσεις που χρησιμοποιούν ήδη προϊόντα της Microsoft.

2. Δυνατότητες Hybrid Cloud: Οι υπηρεσίες του Azure Cloud επιτρέπουν τη χρήση δυνατοτήτων Hybrid Cloud, δηλαδή επιχειρήσεις μπορούν να χρησιμοποιούν τόσο δημόσιους όσο και ιδιωτικούς χώρους νέφους για να αποθηκεύουν τα δεδομένα και τις εφαρμογές τους. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο για επιχειρήσεις με ευαίσθητα δεδομένα ή εφαρμογές.

3. Ευκολία στην ανάπτυξη εφαρμογών: Οι υπηρεσίες του Azure Cloud προσφέρουν εκτενή εργαλεία για την ανάπτυξη και τη διαχείριση εφαρμογών, καθιστώντας την ευκολότερη και πιο αποτελεσματική διαδικασία για τους προγραμματιστές και τους διαχειριστές συστημάτων.

4. Ευέλικτη τιμολόγηση: Οι υπηρεσίες του Azure Cloud προσφέρουν ευέλικτες επιλογές τιμολόγησης, όπως pay-as-you-go, πακέτα συνδρομής και ανάλυση χρήσης, επιτρέποντας στους χρήστες να προσαρμόζουν τη χρήση τους στις ανάγκες τους και να εξοικονομήσουν χρήματα στο μακροπρόθεσμο.

Βλέπουμε ότι το πακέτο υπηρεσιών που προσφέρει το Azure καλύπτει όλες τις ανάγκες ενός αυτοματοποιημένου συστήματος ανάλυσης δεδομένων, συνεπώς επιλέξαμε αυτό για τους σκοπούς της εργασίας μας. Στο επόμενο κεφάλαιο θα περιγραφούν σε λεπτομέρεια τα εργαλεία του Azure που χρησιμοποιήθηκαν για την υλοποίηση του συστήματός μας.

Κεφάλαιο 3: Εργαλεία που χρησιμοποιήθηκαν

Όπως αναλύσαμε στο προηγούμενο κεφάλαιο, για την υλοποίηση της παρούσας εργασίας θα χρησιμοποιηθούν τα εργαλεία που προσφέρει το Azure. Η εκτενής περιγραφή των συγκεκριμένων υπηρεσιών καθώς και τα επιμέρους τεχνικά χαρακτηριστικά τους δίνονται παρακάτω.

3.1 Azure Storage

Η υπηρεσία Azure Storage είναι μια πλατφόρμα αποθήκευσης δεδομένων στο cloud της Microsoft, που παρέχει υπηρεσίες αποθήκευσης δεδομένων, αρχείων, αντικειμένων και λογαριασμούς αποθήκευσης για εικόνες εικονικών μηχανών. Η υπηρεσία προσφέρει αντοχή σε σφάλματα και αντοχή σε καταστροφές, διαθέτει υψηλή ασφάλεια και δυνατότητες αυξανόμενης κλίμακας.

Η υπηρεσία Azure Storage περιλαμβάνει διάφορες επιλογές αποθήκευσης, συμπεριλαμβανομένων των ακόλουθων:

i. Blob Storage: Πρόκειται για αποθήκευση αντικειμένων σε μορφή blob, που είναι κατάλληλη για την αποθήκευση δεδομένων από εφαρμογές και υπηρεσίες του cloud. Παρέχει δυνατότητες αυξανόμενης κλίμακας και αντοχής σε καταστροφές.
ii. File Storage: Πρόκειται για μια υπηρεσία κοινής χρήσης αρχείων, που επιτρέπει στους χρήστες να αποθηκεύουν και να κοινοποιούν αρχεία. Η υπηρεσία παρέχει δυνατότητες αυξανόμενης κλίμακας και αντοχής σε καταστροφές.
iii. Queue Storage: Πρόκειται για αποθήκευση μηνυμάτων, που χρησιμοποιείται συνήθως για τη διασύνδεση των διαφόρων συστημάτων του cloud. Παρέχει δυνατότητες αυξανόμενης κλίμακας και αντοχής σε καταστροφές.
iv. Table Storage: Πρόκειται για μια υπηρεσία αποθήκευσης δεδομένων NoSQL, που χρησιμοποιείται συνήθως για την αποθήκευση δεδομένων από εφαρμογές IoT και λογαριασμούς μηχανικής μάθησης. Παρέχει δυνατότητες αυξανόμενης κλίμακας και αντοχής σε καταστροφές.

Παράλληλα, η υπηρεσία Azure Storage παρέχει και επιπλέον εργαλεία και δυνατότητες, όπως οι ακόλουθες:

1. Ευέλικτες επιλογές επιλογής τοποθεσίας: Η υπηρεσία Azure Storage παρέχει δυνατότητες επιλογής της τοποθεσίας των δεδομένων σας, που μπορεί να βοηθήσει στην ελαχιστοποίηση των καθυστερήσεων και τη βελτίωση της απόκρισης των εφαρμογών σας.

2. Υπηρεσία διαχείρισης ταυτότητας: Η υπηρεσία Azure Storage επιτρέπει στους χρήστες να διαχειρίζονται την ταυτότητά τους με ασφάλεια και ευκολία.
3. Δυνατότητα ανακτήσεων καταστροφής: Η υπηρσία Azure Storage παρέχει δυνατότητες ανακτήσεων καταστροφής, όπως η επαναφορά δεδομένων από την κορυφή της ιεραρχίας των αντιγράφων ασφαλείας σας.
4. Δυνατότητα κρυπτογράφησης: Η υπηρεσία Azure Storage παρέχει δυνατότητες κρυπτογράφησης των δεδομένων σας σε κατάσταση αναπαυτικής αναμονής και κατά τη μεταφορά τους.
5. Διαθεσιμότητα: Η υπηρεσία Azure Storage είναι σχεδιασμένη να παρέχει υψηλή διαθεσιμότητα, με αναλογία διαθεσιμότητας που ξεπερνά το 99,9%.
6. Αντοχή σε καταστροφές: Η υπηρεσία Azure Storage παρέχει δυνατότητες αντοχής σε καταστροφές, με αντιγράφα ασφαλείας και ανακτήσεις καταστροφής.
7. Υψηλή απόδοση: Η υπηρεσία Azure Storage παρέχει υψηλή απόδοση, με δυνατότητες αποθήκευσης και ανάκτησης δεδομένων σε πραγματικό χρόνο.

Για τα άτομα που χρησιμοποιούν την υπηρεσία Azure Storage, υπάρχουν επίσης διάφορα εργαλεία και λειτουργίες διαχείρισης, όπως η δυνατότητα επιλογής του επιπέδου αποθήκευσης, η δυνατότητα παρακολούθησης και ελέγχου της χρήσης της αποθήκευσης καθώς και η διαχείριση των αδειών πρόσβασης των χρηστών.

3.2 Azure Repos

Η υπηρεσία Azure Repos είναι μια πλατφόρμα διαχείρισης κώδικα που παρέχεται από το Azure Cloud. Επιτρέπει στους προγραμματιστές να αναπτύξουν, να διαχειρίζονται και να συνεργάζονται στην ανάπτυξη κώδικα μέσω της αποθήκευσης του κώδικα στο cloud και της παροχής διαφόρων εργαλείων για τη διαχείριση του κώδικα.

Η υπηρεσία παρέχει διάφορα εργαλεία για τη διαχείριση του κώδικα, όπως έλεγχος των αλλαγών στον κώδικα, διαχείριση των διακλαδώσεων του κώδικα, ανίχνευση conflicts στον κώδικα και συνεργατική ανάπτυξη. Επιπλέον, παρέχει εργαλεία για τη διαχείριση των αναθεωρήσεων και της διαδικασίας κυκλοφορίας του κώδικα. Η υπηρεσία επιτρέπει επίσης την ενσωμάτωση σε πολλά εργαλεία ανάπτυξης, όπως το Visual Studio, το Eclipse και το PyCharm, ενώ παρέχει επίσης εργαλεία για τη διαχείριση των άδειων χρήσης και των δικαιωμάτων πρόσβασης.

Η Azure Repos παρέχεται σε δύο κύριες μορφές: το Azure Repos Git και το Azure Repos Team Foundation Version Control (TFVC). Το Azure Repos Git παρέχει μια διανεμημένη αποθήκη κώδικα για τον κώδικα, ενώ το Azure Repos TFVC παρέχει μια κεντρική αποθήκη κώδικα. Η επιλογή μεταξύ των δύο εξαρτάται από τις ανάγκες της εφαρμογής και της ομάδας ανάπτυξης.

Με τη χρήση του Azure Repos, οι προγραμματιστές μπορούν να αναπτύξουν και να διαχειρίζονται τον κώδικα τους από οποιοδήποτε σημείο στον κόσμο, επιτρέποντας την

ανάπτυξη και συνεργασία σε ομάδες που βρίσκονται σε διαφορετικές γεωγραφικές περιοχές. Με τη χρήση του Azure Repos, οι προγραμματιστές μπορούν επίσης να διαχειρίζονται τις άδειες πρόσβασης και τους ρόλους των χρηστών, να παρακολουθούν τις αλλαγές κώδικα και να διαχειρίζονται τις διακλαδώσεις του κώδικα.

Επιπλέον, η υπηρεσία επιτρέπει την εύκολη πρόσβαση στον κώδικα και τη διαχείριση της ομάδας ανάπτυξης μέσω της πλατφόρμας Azure DevOps, που επιτρέπει την αυτοματοποίηση της διαδικασίας ανάπτυξης λογισμικού με χρήση διαφόρων εργαλείων, όπως το Azure Pipelines και το Azure Artifacts. Επιπλέον, μπορεί να συνεργαστεί με άλλες υπηρεσίες της πλατφόρμας Azure DevOps, όπως το Azure Boards και το Azure Pipelines, προκειμένου να ενσωματώσει τη διαδικασία ανάπτυξης λογισμικού στην ίδια πλατφόρμα.

Στο Azure Repos, ο κώδικας αποθηκεύεται σε ένα απομακρυσμένο αποθηκευτικό χώρο, ο οποίος είναι διαθέσιμος σε οποιαδήποτε στιγμή και από οποιοδήποτε σημείο στον κόσμο. Η αποθήκευση γίνεται σε ένα κεντρικό αποθηκευτικό χώρο, ο οποίος παρέχει ασφάλεια και αντοχή σε σφάλματα, ενώ επιτρέπει τη δημιουργία αντιγράφων ασφαλείας του κώδικα. Η πλατφόρμα επίσης παρέχει λειτουργίες παρακολούθησης του κώδικα, όπως οι δυνατότητες αναζήτησης και σύγκρισης του κώδικα, καθώς και η δυνατότητα δημιουργίας αναφορών και διαγραμμάτων για την παρακολούθηση της προόδου του έργου.

Συνολικά, το Azure Repos είναι μια πλούσια και εύχρηστη πλατφόρμα διαχείρισης κώδικα, που παρέχει πλήθος λειτουργιών για την οργάνωση και τη διαχείριση του κώδικα και της διαδικασίας ανάπτυξης λογισμικού. Επιπλέον, ενσωματώνεται εύκολα με άλλες υπηρεσίες της πλατφόρμας Azure DevOps και παρέχει ευέλικτες επιλογές για τη χρήση διαφόρων τεχνολογιών ελέγχου κώδικα.

3.3 Azure Databricks

Η υπηρεσία Azure Databricks είναι μια εξειδικευμένη υπηρεσία για ανάλυση μεγάλων όγκων δεδομένων σε πραγματικό χρόνο και αποτελεί μια συνεργατική πλατφόρμα ανάπτυξης και εκτέλεσης λειτουργιών για επιστημονικές εφαρμογές και ανάλυση δεδομένων.

Η Azure Databricks χρησιμοποιεί έναν κατανεμημένο και κλιμακωτό κινητήρα ανάλυσης μεγάλων όγκων δεδομένων, το Apache Spark, που επιτρέπει την επεξεργασία μεγάλων όγκων δεδομένων σε πραγματικό χρόνο. Η υπηρεσία χρησιμοποιεί επίσης τη γλώσσα Python, που είναι μια από τις πιο δημοφιλείς γλώσσες προγραμματισμού στον κόσμο, για την ανάπτυξη σεναρίων, ενώ υποστηρίζει επίσης μια ποικιλία γλωσσών προγραμματισμού, συμπεριλαμβανομένων των Scala, R και SQL. Αυτό επιτρέπει την επιλογή της γλώσσας προγραμματισμού που καλύπτει καλύτερα τις εκάστοτε ανάγκες.

Η Azure Databricks ενσωματώνει επίσης εργαλεία για τη διαχείριση της ασφάλειας, τη διαχείριση προσβασιμότητας και την ενσωμάτωση με άλλες υπηρεσίες της Azure, όπως το Azure SQL Database και το Azure Event Hub. Επιπλέον, η Azure Databricks προσφέρει μια ευέλικτη αρχιτεκτονική για τη διαχείριση των δεδομένων, επιτρέποντας στους χρήστες είτε να χρησιμοποιήσουν διαφορετικά συστήματα αποθήκευσης, όπως το Azure Storage, είτε εναλλακτικές πηγές δεδομένων, όπως αρχεία CSV και JSON, δεδομένα από βάσεις δεδομένων SQL και NoSQL και δεδομένα που προέρχονται από υπηρεσίες API.

Η Azure Databricks παρέχει επίσης δυνατότητες επεξεργασίας και ανάλυσης πραγματικού χρόνου, καθώς και δυνατότητες ανάλυσης δεδομένων σε batch mode. Προσφέρει ενσωματωμένη υποστήριξη για πολλαπλούς παράλληλους υπολογιστικούς κόμβους, καθώς και ενσωματωμένη ασφάλεια και παρακολούθηση. Αυτές οι δυναμικές λειτουργίες διασφαλίζουν ότι οι επεξεργασίες δεδομένων θα εκτελούνται με ασφάλεια και αποτελεσματικότητα, ενώ παράλληλα μειώνουν το χρόνο ανταπόκρισης του συστήματος. Παρέχει εξαιρετική ευελιξία στον τρόπο που οι χρήστες μπορούν να εκτελέσουν τις εφαρμογές τους αφού δίνεται η δυνατότητα εκτέλεσής στην ίδια την υπηρεσία Databricks ή σε άλλες υπηρεσίες της Azure ή ακόμα και σε άλλους cloud πάροχους.

Επιπλέον, η Azure Databricks προσφέρει πολλαπλά εργαλεία για την επεξεργασία και την ανάλυση των δεδομένων σας, συμπεριλαμβανομένων των Apache Spark, Delta Lake και MLflow και είναι σχεδιασμένο να λειτουργεί απρόσκοπτα με άλλες υπηρεσίες Azure, όπως το Azure Synapse Analytics και το Azure Machine Learning. Τα παραπάνω επιτρέπουν την εύκολη ενσωμάτωση των διαδικασιών επεξεργασίας δεδομένων καθώς και την εύκολη δημιουργία προβλεπτικών αλγορίθμων και μοντέλων μηχανικής μάθησης. Το Azure Databricks περιλαμβάνει ενσωματωμένες λειτουργίες ασφαλείας, όπως κρυπτογράφηση κατά την αποθήκευση και κατά τη μεταφορά των δεδομένων, και ενσωματώνεται με το Azure Active Directory για τον έλεγχο ταυτότητας χρηστών και τον έλεγχο πρόσβασης. Τέλος, το Azure Databricks περιλαμβάνει επίσης λειτουργίες συνεργασίας και διαμοιρασμού κώδικα, όπως το διαμοιρασμό βιβλιοθηκών και notebooks, τον έλεγχο των αλλαγών στον κώδικα και τη διαχείριση των αδειών πρόσβασης για την ομάδα εργασίας, όπως επίσης και πληθώρα dashboards και reports για την απεικόνιση και ανάλυση των δεδομένων.

Συνολικά, το Azure Databricks είναι μια ισχυρή πλατφόρμα ανάλυσης δεδομένων και επεξεργασίας μεγάλων όγκων δεδομένων, που παρέχει μια ολοκληρωμένη λύση για τη διαχείριση και ανάλυση των δεδομένων σας και τη δημιουργία προχωρημένων αναλύσεων και μοντέλων μηχανικής μάθησης. Το Azure Databricks είναι επίσης ενσωματωμένο με τα υπόλοιπα Azure services, όπως το Azure Storage και το Azure Data Factory, για μια ολοκληρωμένη λύση για τη διαχείριση των δεδομένων σας στο cloud.

3.3.1 Clusters

Στο Azure Databricks δίνεται η δυνατότητα δημιουργίας clusters για την επεξεργασία και ανάλυση των δεδομένων. Ένα cluster είναι ένα σύνολο υπολογιστικών πόρων (επεξεργαστές, μνήμη, αποθηκευτικός χώρος) που εκχωρούνται για την εκτέλεση των εργασιών σας. Κάθε cluster περιλαμβάνει έναν Driver Node, ο οποίος αποτελεί το κεντρικό σημείο ελέγχου και ένα ή περισσότερα Executor Nodes, τα οποία είναι υπεύθυνα για την παράλληλη επεξεργασία των δεδομένων.

Η δημιουργία ενός cluster στο Azure Databricks είναι απλή και γίνεται μέσω της διαδικτυακής πλατφόρμας του Azure. Οι χρήστες μπορούν να ορίσουν το μέγεθος του cluster, ανάλογα με τις ανάγκες τους, και να διαμορφώσουν τις επιλογές επεξεργασίας και αποθήκευσης δεδομένων για το cluster. Η Azure Databricks επιτρέπει επίσης τη δημιουργία προσαρμοσμένων εικόνων cluster, οι οποίες περιλαμβάνουν προεγκατεστημένα πακέτα και βιβλιοθήκες που απαιτούνται για την εκτέλεση των εφαρμογών τους. Στο Azure Databricks μπορείτε να δημιουργήσετε clusters με διαφορετικούς τύπους κόμβων, μεταξύ των οποίων το Standard, το High Concurrency και το GPU-Optimized. Κάθε τύπος κόμβου είναι σχεδιασμένος να παρέχει την καλύτερη δυνατή απόδοση για τις ανάγκες της κάθε εφαρμογής. Το Azure Databricks υποστηρίζει επίσης διάφορες δυνατότητες διαχείρισης cluster, όπως τη δυνατότητα αυτόματης κλιμάκωσης cluster, το δυναμικό διαμόρφωσης cluster και τη διαχείριση της κατάστασης των clusters σας μέσω του Azure Portal ή των APIs.

Συνολικά, η δυνατότητα δημιουργίας clusters στο Azure Databricks προσφέρει μια πολύ ευέλικτη και δυναμική πλατφόρμα για την επεξεργασία των δεδομένων σας. Μπορείτε να δημιουργήσετε πολλαπλά clusters, ανάλογα με τις ανάγκες σας, και να τα διαμορφώσετε κατάλληλα για την εκτέλεση διαφορετικών εργασιών. Επιπλέον, οι χρήστες μπορούν να αποθηκεύουν τα δεδομένα τους σε διαφορετικές πηγές αποθήκευσης, όπως η Azure Blob Storage ή η Azure Data Lake Storage, και να επεξεργαστούν τα δεδομένα σε αυτές τις πλατφόρμες.

3.3.2 Pyspark

Η PySpark είναι ένα από τα πιο δημοφιλή εργαλεία για την ανάπτυξη εφαρμογών ανάλυσης δεδομένων στην πλατφόρμα Apache Spark. Είναι ένας πακέτο Python που παρέχει μια διασύνδεση με το API του Spark και επιτρέπει στους προγραμματιστές να αναπτύξουν εφαρμογές Spark χρησιμοποιώντας τη γλώσσα προγραμματισμού Python. Με τη χρήση του, οι προγραμματιστές μπορούν να εκμεταλλευτούν τις λειτουργίες του Spark, όπως η διανομή και η επεξεργασία μεγάλου όγκου δεδομένων σε παράλληλους υπολογιστικούς κόμβους, καθώς επίσης να χρησιμοποιήσουν τις βιβλιοθήκες Python, όπως NumPy και Pandas, για την επεξεργασία και ανάλυση των δεδομένων σε πραγματικό χρόνο και την κατασκευή μοντέλων

μηχανικής μάθησης με σκοπό την εξαγωγή προβλέψεων από μεγάλα σύνολα δεδομένων.

Η PySpark παρέχει πλούσιες δυνατότητες για την επεξεργασία δεδομένων σε μεγάλη κλίμακα. Μπορεί να χρησιμοποιηθεί για να επεξεργαστεί δεδομένα από πολλαπλές πηγές, όπως αρχεία HDFS, Apache Cassandra, Apache HBase, Amazon S3, Azure Storage και πολλές άλλες. Επιπλέον, οι προγραμματιστές μπορούν να εκτελέσουν διαδικασίες επεξεργασίας δεδομένων χρησιμοποιώντας συναρτήσεις μετασχηματισμού και ενώσεων (SQL), και μπορούν επίσης να εφαρμόσουν συναρτήσεις στοιχείων όπως αριθμητικά, κείμενο και ημερομηνίες.

Η PySpark παρέχει μια σειρά από βιβλιοθήκες Python για την επεξεργασία δεδομένων, συμπεριλαμβανομένων των:

- PySpark SQL: Παρέχει υποστήριξη για δομημένα δεδομένα, όπως SQL βάσεις δεδομένων.
- PySpark Streaming: Παρέχει υποστήριξη για επεξεργασία δεδομένων σε πραγματικό χρόνο.
- MLlib: Παρέχει υποστήριξη για μηχανική μάθηση και ανάλυση δεδομένων.
- GraphX: Παρέχει υποστήριξη για γράφους και δικτυακά δεδομένα.

Η PySpark μπορεί να εκτελείται σε διάφορα περιβάλλοντα, όπως σε επαγγελματικά data centers, στο cloud, σε Hadoop clusters και σε υπολογιστικούς πόρους του συστήματος.

Η βασική δομή της PySpark αποτελείται από τα παρακάτω στοιχεία:

SparkSession: Παρέχει τη σύνδεση με το Spark Cluster και την πρόσβαση στις διαθέσιμες λειτουργίες της Apache Spark.
DataFrame: Είναι μια δομή δεδομένων που υποστηρίζει δομημένα και ημιδομημένα δεδομένα, και μπορεί να εκτελεί λειτουργίες μετασχηματισμού και ανάλυσης.
RDD: Είναι μια δομή δεδομένων που χρησιμοποιείται στην Apache Spark και παρέχει τη δυνατότητα επεξεργασίας δεδομένων σε μεγάλη κλίμακα.
Κατά τη διαδικασία επεξεργασίας δεδομένων με PySpark, οι διακριτές εργασίες επεξεργασίας εκτελούνται σε διαφορετικούς κόμβους του συστήματος, ώστε να επιταχυνθεί η επεξεργασία των δεδομένων. Επιπλέον, η PySpark υποστηρίζει και την παραλληλοποίηση εργασιών σε επίπεδο εφαρμογής, ώστε να εκτελούνται ταυτόχρονα πολλές εργασίες σε διαφορετικά σημεία του κώδικα.

Μερικά από τα βασικά χαρακτηριστικά της PySpark είναι:

1. Υποστήριξη για πολλές πλατφόρμες εκτέλεσης, όπως Hadoop, Kubernetes, Apache Mesos, και Amazon EC2.

2. Υποστήριξη για πολλές πηγές δεδομένων, όπως HDFS, S3, JDBC, Cassandra, και Hive. Επεξεργασία δεδομένων σε μεγάλη κλίμακα με αξιοποίηση του συστήματος κατανεμημένου υπολογισμού της Apache Spark.
3. Υψηλή απόδοση στην επεξεργασία δεδομένων λόγω της χρήσης της γλώσσας προγραμματισμού Python και της υποστήριξης της μνήμης RAM στο σύστημα.
4. Επικοινωνία με άλλες βιβλιοθήκες Python για την επεξεργασία δεδομένων, όπως NumPy και Pandas.
5. Εύκολη διαχείριση των δεδομένων με την χρήση της δομής δεδομένων DataFrame.
6. Επεκτασιμότητα και προσαρμοστικότητα στις ανάγκες των επιχειρήσεων, με την υποστήριξη πολλών προγραμματιστικών γλωσσών, όπως Java, Scala, και R.

Επιπλέον, η χρήση PySpark συνεπάγεται τη διευκόλυνση της ανάπτυξης και συντήρησης του κώδικα, καθώς οι προγραμματιστές μπορούν να χρησιμοποιήσουν τη γλώσσα προγραμματισμού Python και να εκτελούν τις εργασίες τους σε μια κατανεμημένη υπολογιστική πλατφόρμα. Αυτό καθιστά τον κώδικα πιο ευανάγνωστο και πιο ευέλικτο στις αλλαγές και τις επεκτάσεις στο μέλλον. Προσφέρει ευρεία δυνατότητα προσαρμογής και επέκτασης με τη χρήση διαφόρων βιβλιοθηκών Python και τη δυνατότητα επέκτασης της λειτουργικότητας του με τη χρήση custom UDFs (User-Defined Functions). Αυτό το καθιστά ιδανικό για την ανάπτυξη προηγμένων εφαρμογών δεδομένων και ανάλυσης, όπως η πρόβλεψη μελλοντικών τάσεων και η αναγνώριση προτύπων σε μεγάλα σύνολα δεδομένων.

Συνολικά, η PySpark είναι μια πολύ χρήσιμη επιλογή για την επεξεργασία μεγάλου όγκου δεδομένων και την ανάλυση τους σε πραγματικό χρόνο. Με τη χρήση PySpark, οι επιχειρήσεις μπορούν να επεξεργάζονται μεγάλα σύνολα δεδομένων με μεγάλη απόδοση και να εκτελούν πολλαπλές εργασίες ταυτόχρονα σε διαφορετικά σημεία του κώδικα.

3.4 Azure Data Factory

Το Azure Data Factory είναι μια υπηρεσία cloud-based ETL (Extract, Transform, Load) που διαχειρίζεται και εκτελεί εργασίες επεξεργασίας δεδομένων στο cloud. Επιτρέπει στους χρήστες να συνδέσουν διαφορετικές πηγές δεδομένων και να τις μετατρέπουν, ενώ ταυτόχρονα τις μεταφέρουν σε άλλες πηγές.

Η βασική λειτουργικότητα του Azure Data Factory περιλαμβάνει τη σύνδεση με πολλαπλές πηγές δεδομένων, τη διαχείριση και τον έλεγχο της διαδικασίας ETL και την πρόσβαση σε εκατοντάδες προετοιμασμένα στοιχεία δεδομένων (data connectors) για τη συνένωση, μετατροπή και μεταφορά δεδομένων. Οι χρήστες μπορούν να δημιουργήσουν περιπλοκότερα συστήματα ETL με την υποστήριξη του Azure Data Factory, συμπεριλαμβανομένης της δυνατότητας να διαχειρίζονται δεδομένα σε διαφορετικά περιβάλλοντα, να παρακολουθούν και να ελέγχουν τις διεργασίες και να χρησιμοποιούν συστήματα όπως το Hadoop, το Spark και το Azure HDInsight. Παρέχονται εργαλεία διαχείρισης δεδομένων όπως το Data Flow, που

επιτρέπει στους χρήστες να δημιουργήσουν πιο περίπλοκες μετασχηματιστές δεδομένων (data transformations) με χρήση drag-and-drop εργαλείων. Το Data Flow διαθέτει επίσης πολλαπλές επιλογές συνδυασμού μετασχηματιστών δεδομένων για την επίτευξη πιο σύνθετων λειτουργιών.

Με τη βοήθεια του Azure Data Factory, οι χρήστες μπορούν να δημιουργήσουν προγραμματιστικά και εποπτικά σενάρια για την επεξεργασία και μεταφορά δεδομένων στο cloud. Μπορούν επίσης να χρησιμοποιήσουν την υπηρεσία ως ενδιάμεσο επίπεδο μεταξύ διαφορετικών πηγών και προορισμών δεδομένων, προσθέτοντας επιπλέον λειτουργικότητα και ευελιξία στα συστήματα διαχείρισης δεδομένων. Η συγκεκριμένη υπηρεσία παρέχει αυξημένη ασφάλεια δεδομένων με δυνατότητες όπως ο έλεγχος πρόσβασης, η κρυπτογράφηση και η ελεγχόμενη διαμοιρασμός δεδομένων και είναι δυνατή η ρύθμιση προσαρμοστικών πολιτικών επιτήρησης για την ανίχνευση και την αντίδραση σε πιθανά προβλήματα ασφαλείας.

Το Azure Data Factory παρέχει πολλές ακόμη δυνατότητες και χαρακτηριστικά για την επεξεργασία και διαχείριση δεδομένων. Αυτά περιλαμβάνουν:

1. Δυνατότητα παρακολούθησης και αναφοράς διαδικασιών επεξεργασίας δεδομένων: Οι χρήστες μπορούν να παρακολουθούν και να αναφέρουν την εκτέλεση των διαδικασιών επεξεργασίας δεδομένων στο cloud, για να επιβεβαιώσουν ότι οι διαδικασίες έχουν ολοκληρωθεί επιτυχώς ή να εντοπίσουν πιθανά προβλήματα.
2. Υποστήριξη πολλαπλών πηγών και προορισμών δεδομένων: Οι χρήστες μπορούν να συνδέσουν το Azure Data Factory με πολλαπλές πηγές και προορισμούς δεδομένων, συμπεριλαμβανομένων των Azure services, όπως το Azure Blob Storage, το Azure Data Lake Storage, και το Azure SQL Database, καθώς και εξωτερικών πηγών, όπως το Amazon S3 και το Google Cloud Storage.
3. Διαχείριση επιλογών εκτέλεσης διαδικασιών επεξεργασίας δεδομένων: Οι χρήστες μπορούν να επιλέξουν ποιοι υπολογιστικοί πόροι και υπηρεσίες του cloud θα χρησιμοποιηθούν για την εκτέλεση των διαδικασιών επεξεργασίας δεδομένων, προσαρμόζοντας την επιλογή ανάλογα με τις ανάγκες και τον προϋπολογισμό τους.
4. Δυνατότητα προγραμματισμού και διαχείρισης διαδικασιών επεξεργασίας δεδομένων μέσω του REST API: Οι χρήστες μπορούν να χρησιμοποιήσουν το REST API του Azure Data Factory για την προγραμματιστική επέκταση και τη διαχείριση των διαδικασιών επεξεργασίας δεδομένων.
5. Δυνατότητα διαχείρισης διακομιστής ονομάτων συνόλων δεδομένων (data set) και διαδικασιών επεξεργασίας δεδομένων (data pipeline): Οι χρήστες μπορούν να διαχειρίζονται τα δεδομένα σε συνόλα δεδομένων και να δημιουργούν διαδικασίες επεξεργασίας δεδομένων στο Azure Data Factory.
6. Υποστήριξη πολλών είδη δεδομένων: Με χρήση του Azure Data Factory, οι χρήστες μπορούν να εισάγουν δεδομένα από μια γκάμα πηγών, όπως αρχεία κειμένου, αρχεία Excel, αρχεία XML και JSON, καθώς και δεδομένα που φιλοξενούνται σε βάσεις δεδομένων, όπως η Microsoft SQL Server, η Oracle και η MySQL

Συνολικά, το Azure Data Factory παρέχει μια ολοκληρωμένη πλατφόρμα διαχείρισης και επεξεργασίας δεδομένων, που επιτρέπει στους χρήστες να συνδέσουν και να μετακινήσουν δεδομένα από πολλές πηγές και προορισμούς, να προγραμματίσουν και να εκτελέσουν διαδικασίες επεξεργασίας δεδομένων και να παράγουν και να διαχειρίζονται αυτές τις διαδικασίες μέσω ενός ευέλικτου και αξιόπιστου περιβάλλοντος. Επιπλέον, το Azure Data Factory είναι ενσωματωμένο με άλλες υπηρεσίες του Microsoft Azure, όπως η Azure Event Grid και η Azure Monitor, για την ενίσχυση της διαχείρισης και παρακολούθησης των διαδικασιών επεξεργασίας δεδομένων.

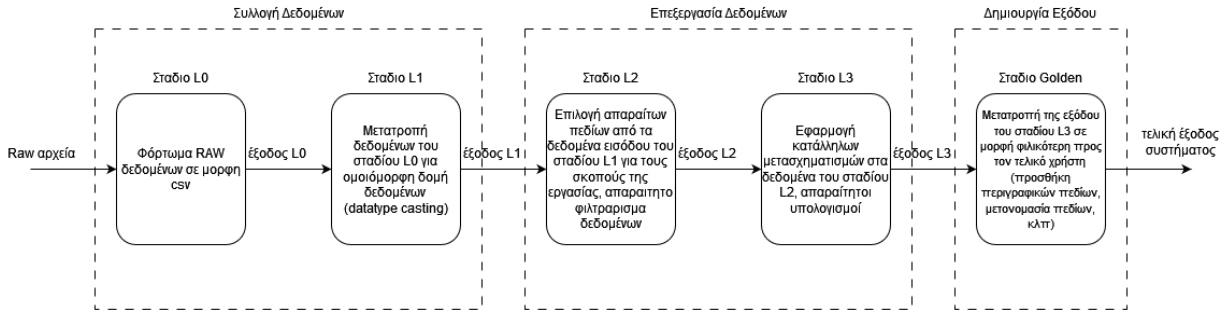
Έχοντας πλέον περιγράψει εκτενώς τις δυνατότητες και τα τεχνικά χαρακτηριστικά των εργαλείων που θα χρησιμοποιηθούν, στο επόμενο κεφάλαιο θα αναλύσουμε σε βάθος την υλοποίηση ενός ολοκληρωμένου συστήματος ανάλυσης δεδομένων με στόχο την αναγνώριση της παραγωγικής αποδοτικότητας μιας επιχείρησης.

Κεφάλαιο 4: Μέθοδοι και Υλοποίηση

Στο συγκεκριμένο κεφάλαιο δίνεται η αναλυτική περιγραφή της υλοποίησης του συστήματος μας για την διαχείριση και επεξεργασία των δεδομένων για την παραγωγή του τελικού αποτελέσματος.

Συνοπτικά, η αρχιτεκτονική του συστήματος είναι η ακόλουθη: Παίρνουμε τα δεδομένα εισόδου από το στάδιο 1, τα οποία είναι κατάλληλα μορφοποιημένα και επιλέγουμε την απαραίτητη πληροφορία για το σύστημά μας στο στάδιο 2. Στη συνέχεια, στο στάδιο 3 εφαρμόζουμε τους απαραίτητους μετασχηματισμούς στα δεδομένα του σταδίου 2 και κάνουμε τους κατάλληλους υπολογισμούς. Στο τελικό στάδιο, μετατρέπουμε τα δεδομένα του σταδίου 3 σε μια μορφή πιο φιλική και κατανοητή προς τους χρήστες, όπου είναι και η τελική έξοδος του συστήματός μας.

Το σχηματικό διάγραμμα της αρχιτεκτονικής του συστήματος είναι το ακόλουθο:



Εικόνα 1: Αρχιτεκτονική συστήματος

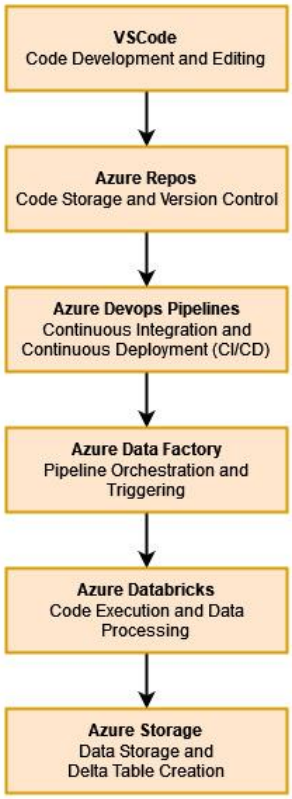
Η υλοποίηση του συστήματος βασίζεται σε τέσσερις πυλώνες, την αποθήκευση των δεδομένων, την ανάπτυξη και διαχείριση του κώδικα, το περιβάλλον εκτέλεσής του κώδικα για την επεξεργασία των δεδομένων και την αυτόματη εκτέλεσή του.

Προκειμένου να δημιουργηθεί το επιθυμητό περιβάλλον επεξεργασίας δεδομένων, γίνεται χρήση των παραπάνω εργαλείων της Azure με την ακόλουθη διαδικασία:

1. Ανάπτυξη κώδικα σε local IDE (στην περίπτωση μας Visual Studio Code), όπου δημιουργούνται scripts και modules για την απαραίτητη επεξεργασία των δεδομένων μας.
2. Μέσω Git, ο κώδικας προωθείται και αποθηκεύεται στο Azure Repos, ένα σύστημα version control που παρέχει κεντρική διαχείριση του πηγαίου κώδικα.

3. Ο κώδικας γίνεται deploy στο προορισμένο περιβάλλον (Develop, Quality or Production) με τη χρήση των Azure DevOps Services, ενός σετ εργαλείων για συνεχή ενσωμάτωση και συνεχή παράδοση (CI/CD). Η διαδικασία αυτοματοποιεί τον έλεγχο και το packaging του κώδικα, καθιστώντας τον έτοιμο για το deployment σε ένα παραγωγικό περιβάλλον.
4. Μετά το deployment, ανάπτυξη ενός pipeline μέσω του Azure Data Factory προκειμένου να γίνει το κατάλληλο orchestration της εκτέλεσης του κώδικα. Επιπλέον δημιουργία trigger για την αυτόματη εκκίνηση της εκτέλεσης του κώδικα σε επιθυμητά time intervals.
5. Μέσω του trigger του Azure Data Factory, ξεκινάει η εκτέλεση του κώδικα με χρήση του Azure Databricks στα επιθυμητά clusters.
6. Το αποτέλεσμα της εκτέλεσης του κώδικα αποθηκεύεται στο Azure Storage σε πίνακες Delta, έναν τύπο αποθήκευσης δεδομένων που παρέχει schema enforce και data versioning.

Μια σχηματική απεικόνιση της παραπάνω διαδικασίας μπορεί να φανεί στο παρακάτω διάγραμμα:



Εικόνα 2: Βήματα διαδικασίας

4.1 Δεδομένα Εισόδου

Παρακάτω περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν ως είσοδος για το σύστημά μας.

L0: Raw Data. Αρχεία της μορφής .csv που δεν έχουν υποστεί καμία επεξεργασία. Περιλαμβάνουν πληροφορία για τις πωλήσεις των πελατών, την πληρότητα των ψυγείων, την κινητικότητα των πορτών των ψυγείων, τους στόχους πωλήσεων της κάθε χώρας για τα διάφορα είδη ψυγείων, πληροφορία για τις μετατροπές μονάδων μέτρησης, την αντιστοιχία μεταξύ πελατών-ψυγείων, καθώς επίσης και πληροφορία για τα έγκυρα ψυγεία, τους έγκυρους πελάτες, τα έγκυρα προϊόντα και τις έγκυρες συσκευασίες της κάθε χώρας.

L1: Curated Data. Στα δεδομένα του προηγούμενου σταδίου εφαρμόζονται ορισμένες μετατροπές (μετονομασίες, Datatype Casting κλπ) με σκοπό την επίτευξη της ομοιομορφίας των δεδομένων Αποτελούν την τελική μορφή των δεδομένων εισόδου της εργασίας.

4.2 Στάδια της Υλοποίησης

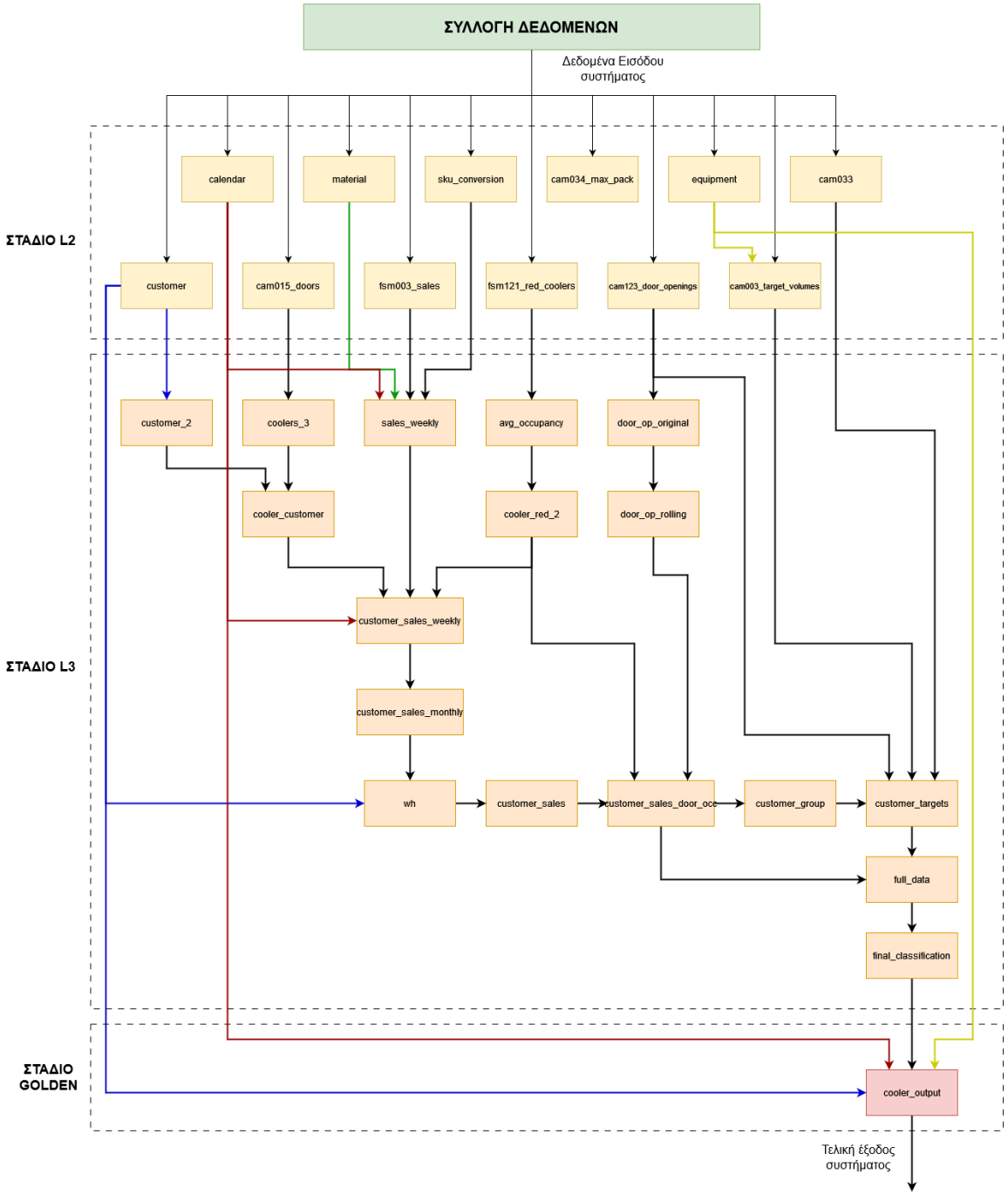
Δεδομένου ότι η εργασία έχει στηθεί σε Azure Data Lake, έχουν υλοποιηθεί τα ακόλουθα στάδια στην επεξεργασία των δεδομένων μας:

L2: Πρώτο στάδιο της εργασίας μας. Σε αυτό το στάδιο εφαρμόζονται τα απαραίτητα φίλτρα στα δεδομένα του L1 μαζί με κάποιες απλές διαδικασίες (selecting ή adding columns) προκειμένου να κρατήσουμε μόνο την πληροφορία που μας είναι απαραίτητη για τους σκοπούς της εργασίας και να προετοιμάσουμε τα datasets μας κατάλληλα για να μπορούν να εφαρμοστούν οι απαραίτητοι υπολογισμοί. Αποτελείται από δύο ειδών πίνακες, country_agnostic πίνακες (πίνακες οι οποίοι είναι κοινοί για όλες τις χώρες π.χ. calendar) και country_specific πίνακες (πίνακες που περιέχουν διαφορετική πληροφορία ανά χώρα π.χ. customer). Αποθηκεύονται σε Delta Tables.

L3: Δεύτερο στάδιο της εργασίας μας. Περιλαμβάνει όλους τους απαραίτητους μετασχηματισμούς των datasets του L2 και τους υπολογισμούς προκειμένου να φτάσουμε στο επιθυμητό τελικό αποτέλεσμα. Αποθηκεύεται σε Delta Tables.

Golden: Τελικό στάδιο της εργασίας μας. Περιέχει τον πίνακα με το τελικό output και αποθηκεύεται σε Delta Table.

Τα παραπάνω στάδια, καθώς και ο συνδυασμός των διαδικασιών που περιλαμβάνουν προκειμένου να παραχθεί το τελικό αποτέλεσμα φαίνεται στο παρακάτω flowchart

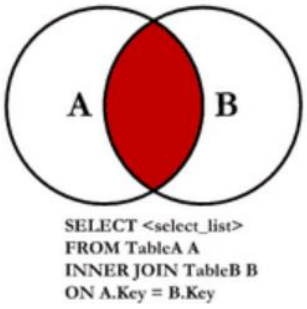


Εικόνα 3: Επιμέρους στάδια υλοποίησης

4.3 Απαραίτητες Συναρτήσεις

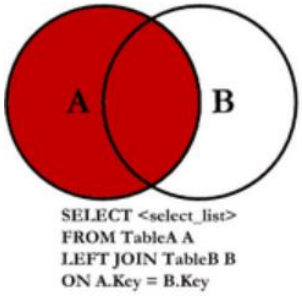
Για την υλοποίηση του συστήματος χρησιμοποιήθηκαν αρκετές συναρτήσεις της Pyspark οι οποίες ανήκουν στην βιβλιοθήκη pyspark.sql, βασίζονται δηλαδή σε εντολές sql. Καθώς χρησιμοποιούνται εκτενώς για τον κατάλληλο μετασχηματισμό των συνόλων δεδομένων για την παραγωγή της τελικής εξόδου και γίνεται συχνή αναφορά τους στο υποκεφάλαιο 4.6, θα δοθεί μια περιγραφή των βασικών συναρτήσεων που χρησιμοποιήθηκαν.

1. Inner Join: Το Inner Join είναι ένας τύπος συνένωσης που επιστρέφει μόνο τις σειρές που έχουν τιμές που ταιριάζουν στους δύο πίνακες που συνδέονται. Δηλαδή, παράγει έναν νέο πίνακα που περιέχει μόνο τις σειρές όπου η καθορισμένη συνθήκη (συνήθως, μια ταιριασμένη τιμή στήλης) είναι αληθής για και τους δύο πίνακες.
   Η σχηματική απεικόνιση του inner join δίνεται παρακάτω:



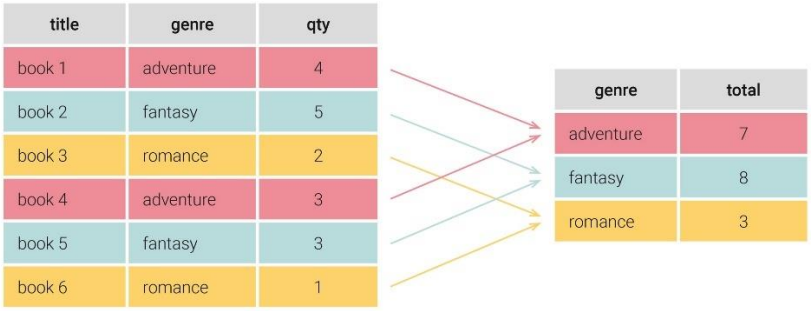
Εικόνα 4: Inner Join Operation

2. Left Join: Το Left Join είναι ένας τύπος συνένωσης που επιστρέφει όλες τις σειρές από τον αριστερό πίνακα και τις αντίστοιχες ταιριαστές σειρές από το δεξί πίνακα. Αν δεν υπάρχουν ταιριαστές σειρές στον δεξί πίνακα, το αποτέλεσμα θα περιέχει τιμές null για αυτές τις στήλες. Αυτός ο τύπος συνένωσης χρησιμοποιείται συνήθως για την εύρεση δεδομένων που υπάρχουν σε έναν πίνακα, αλλά ενδέχεται να μην υπάρχουν σε άλλον.
   Η σχηματική απεικόνιση του left join δίνεται παρακάτω:



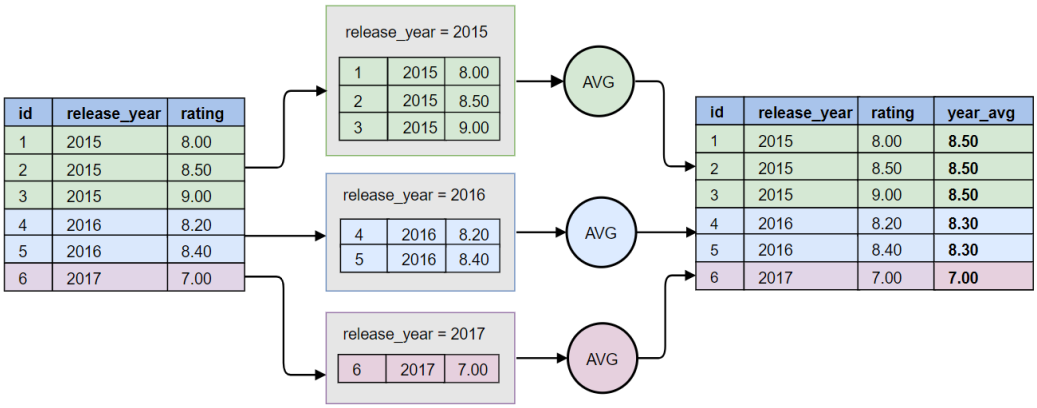
Εικόνα 5: Left Join Operation

3. Group By: Η συνάρτηση group by χρησιμοποιείται για να ομαδοποιήσει τις σειρές σε έναν πίνακα βάσει ενός ή περισσότερων πεδίων. Αυτή η λειτουργία χρησιμοποιείται συχνά με αθροιστικές συναρτήσεις όπως sum, count, avg, κλπ. για να υπολογίσει στατιστικά στοιχεία για κάθε ομάδα. Για παράδειγμα, σε ένα πίνακα πωλήσεων με στήλες για ημερομηνία, προϊόν και έσοδα, χρησιμοποιώντας το group by (ημέρα) τα δεδομένα ομαδοποιούνται ανά ημέρα και υπολογίζονται τα συνολικά έσοδα για κάθε ημέρα.
   Η σχηματική απεικόνιση του group by δίνεται παρακάτω:



Εικόνα 6: Group By Operation

4. Window: Η συνάρτηση Window είναι ένας τύπος συνάρτησης που εκτελεί έναν υπολογισμό σε ένα σύνολο σειρών σε έναν πίνακα, αλλά χωρίς να συγκεντρώνει τις σειρές σε μια μόνο τιμή (σε αντίθεση με τη συνάρτηση group by που περιγράφηκε προηγουμένως). Αυτό σημαίνει ότι μια συνάρτηση παραθύρου μπορεί να επιστρέψει ένα αποτέλεσμα για κάθε σειρά στον πίνακα, βασισμένο στις τιμές άλλων σειρών στο ίδιο παράθυρο. Οι συναρτήσεις παραθύρου χρησιμοποιούνται συχνά για να υπολογίσουν τρέχουσες συνολικές ή μέσες τιμές, να κατατάξουν στοιχεία εντός μιας ομάδας, ή να εφαρμόσουν πολλαπλές συναρτήσεις σε ένα σύνολο δεδομένων χωρίς να χρειάζεται να δημιουργήσουν ενδιάμεσα αποτελέσματα. Για παράδειγμα, χρησιμοποιώντας μια συνάρτηση παραθύρου μπορεί να υπολογιστεί το άθροισμα των τελευταίων τριών συναλλαγών για κάθε πελάτη σε έναν πίνακα πωλήσεων.
   Η σχηματική απεικόνιση της συνάρτησης Window δίνεται παρακάτω:



Εικόνα 7: Window Operation

4.4 Περιγραφή Υψηλού Επιπέδου της Εργασίας

Σκοπός της συγκεκριμένης εργασίας είναι η σχεδίαση και υλοποίηση ενός αυτόματου συστήματος κατηγοριοποίησης της απόδοσης των ψυγείων των πελατών μιας εταιρείας λαμβάνοντας υπόψιν το πόσο κερδοφόρα θεωρούνται. Υπάρχουν 6 διαφορετικές κατηγορίες απόδοσης, των οποίων η σημασία εξηγείται στον ακόλουθο πίνακα:

Κατηγορία Απόδοσης	Απόδοση	Περιγραφή	Action Needed
Α	>=100% + Country Specific threshold	Το ψυγείο αποδίδει πάνω από το target που έχει τεθεί	Καμία ενέργεια
Β	>=100% & <100% + Country Specific threshold	Το ψυγείο έχει κέρδος αλλά δεν πιάνει τον στόχο που έχει τεθεί	Πιθανές ενέργειες για την αύξηση του κέρδους προκειμένου να πιαστεί ο στόχος της εκάστοτε χώρας
Γ	>=50% & <100% του ορίου	Το ψυγείο έχει πωλήσεις οι οποίες όμως έχουν πέσει κάτω από τα λειτουργικά του έξοδα	Ερευνά για τις πιθανές αίτιες/ επικοινωνία με την εκάστοτε χώρα για ενέργειες βελτίωσης με βάση τις οδηγίες της κάθε χώρας
Δ	<50% του ορίου	Το ψυγείο είναι μη κερδοφόρο, ερευνάται η αντικατάσταση του	Ερευνά για τις πιθανές αίτιες/ επικοινωνία με την εκάστοτε χώρα για πιθανή αντικατάσταση
Ε	Μηδαμινές πωλήσεις	Το ψυγείο δεν έχει καθόλου πωλήσεις	Άμεση Αντικατάσταση
Ανεπαρκή Δεδομένα	Ανεπαρκή Δεδομένα	Τα Δεδομένα Εισόδου δεδομένα δεν είναι αρκετά για να αποφανθούμε για την απόδοση	Επικοινωνία με την εκάστοτε χώρα για ενημέρωση των Δεδομένα Εισόδου δεδομένων

Πίνακας 1: Κατηγορίες απόδοσης

Προκειμένου να αποφανθούμε για την κατηγορία του κάθε ψυγείου, πρέπει να συνδυάσουμε την πληροφορία για την απόδοσή τους (εφόσον αυτή είναι διαθέσιμη από τα Δεδομένα Εισόδου δεδομένα μας) στους ακόλουθους τομείς:

1. Πωλήσεις Πελάτη: Εξετάζοντας τις πωλήσεις του κάθε πελάτη μπορούμε να καταλάβουμε πόσες πωλήσεις αντιστοιχούν σε καθένα από τα ψυγεία του και κατά συνέπεια το πόσο κερδοφόρα ήταν.
2. Πληρότητα Ψυγείου: Εξετάζοντας την πληρότητα του κάθε ψυγείου μπορούμε να καταλάβουμε πόσο κερδοφόρο θα μπορούσε να είναι στο μέλλον.
3. Κινητικότητα πορτών Ψυγείου: Εξετάζοντας το πόσο ανοίγονται και κλείνουν οι πόρτες ενός ψυγείου μπορούμε να καταλάβουμε πόσο ενεργό ήταν τον τελευταίο καιρό.

Επιπρόσθετα, υπάρχει και μία εξτρά κατηγορία πελατών, οι οποίοι δεν προμηθεύονται απευθείας από την κεντρική αποθήκη της κάθε χώρας, αλλά προμηθεύονται σε δεύτερο χρόνο από αποθήκες άλλων πελατών. Οι συγκεκριμένοι πελάτες βρίσκονται μεν στο σύστημα της εταιρείας και υπάρχουν τα ψυγεία τους, όμως δεν υπάρχει διαθέσιμη πληροφορία για τα προαναφερθέντα στοιχεία. Αυτή η κατηγορία πελατών ονομάζονται Wholesalers (για συντομία WH), γίνονται identify με συγκεκριμένο τρόπο και η κατηγοριοποίηση της απόδοσης των ψυγείων τους βρίσκεται ακολουθώντας μια ελαφρώς διαφορετική διαδικασία, η οποία θα εξηγηθεί αναλυτικά παρακάτω.

Με βάση τη διαθέσιμη πληροφορία για τα παραπάνω, μπορούμε να διαχωρίσουμε τα ψυγεία των πελατών στις παρακάτω κατηγορίες:

ΔΙΑΘΕΣΙΜΕΣ ΠΩΛΗΣΕΙΣ	ΔΙΑΘΕΣΙΜΗ ΚΙΝΗΤΙΚΟΤΗΤΑ ΠΟΡΤΩΝ	ΔΙΑΘΕΣΙΜΗ ΠΛΗΡΟΤΗΤΑ	WHOLESALER	ΚΑΤΗΓΟΡΙΑ ΨΥΓΕΙΟΥ
ΝΑΙ	ΝΑΙ	ΝΑΙ		1
ΝΑΙ	ΝΑΙ	ΟΧΙ		2
ΝΑΙ	ΟΧΙ	ΝΑΙ		3
ΝΑΙ	ΟΧΙ	ΟΧΙ		4
ΟΧΙ	ΝΑΙ	ΝΑΙ		5
ΟΧΙ	ΝΑΙ	ΟΧΙ		6
ΟΧΙ	ΟΧΙ	ΝΑΙ		7
ΟΧΙ	ΟΧΙ	ΟΧΙ	ΝΑΙ	8
ΟΧΙ	ΟΧΙ	ΟΧΙ	ΟΧΙ	Not assigned

Πίνακας 2: Κατηγορίες διαθέσιμης πληροφορίας ψυγείων

Εφόσον υπολογίσουμε την παραπάνω κατηγοριοποίηση, πρέπει να υπολογίσουμε και τις actual τιμές για τις πωλήσεις, την κινητικότητα και την πληρότητα για κάθε ψυγείο, καθώς και τα λειτουργικά έξοδα που έχει το κάθε ψυγείο και τα targets που έχει θέσει η κάθε χώρα,

προκειμένου να εφαρμόσουμε τους κατάλληλους τύπους για τον υπολογισμό της τελικής απόδοσης. Οι υπολογισμοί αυτοί θα αναλυθούν στο υποκεφάλαιο 4.6.

4.5 Περιγραφή των Πινάκων

Παρακάτω δίνονται οι πίνακες που δημιουργήθηκαν κατά τη διάρκεια της εργασίας, καθώς και μια σύντομη περιγραφή των πινάκων που κατασκευάστηκαν ανά στάδιο.

4.5.1 Πίνακες L2

Calendar

Περιέχει πληροφορία για τα ημερολογιακά στοιχεία από το 2003 έως το 2028, όπως ημέρα, μήνα, εβδομάδα, τελευταία μέρα της εβδομάδας, τελευταία μέρα ενός quarter κλπ. Ακολουθεί το ημερολόγιο 4-4-5, το οποίο χρησιμοποιείται ευρέως στον επιχειρησιακό τομέα και σημαίνει ότι οι δύο πρώτοι μήνες του έτους έχουν 4 εβδομάδες, ο τρίτος 5 κλπ.

Cam033

Περιέχει πληροφορία για τους στόχους πωλήσεων που θέτει η κάθε χώρα.

Cam034_max_pack_size

Περιέχει πληροφορία για τις συσκευασίες προϊόντων που είναι valid.

Equipment

Περιέχει όλη τη διαθέσιμη πληροφορία για όλα τα διαθέσιμα ψυγεία, όπως ημερομηνία εγκατάστασης, ημερομηνία απόκτησης, είδος ψυγείου, αριθμός πορτών κλπ.

Material

Περιέχει όλη τη διαθέσιμη πληροφορία για τα διαθέσιμα προϊόντα προς πώληση που περιέχονται μέσα στα ψυγεία (κωδικοί, είδος, μέγεθος κλπ.).

sku_conversion

Περιέχει πληροφορία για τη μετατροπή μονάδων μέτρησης από unit cases σε physical cases & cs.

Customer

Περιέχει όλη τη διαθέσιμη πληροφορία για τους πελάτες της κάθε χώρας (κωδικούς πελατών, περιοχή δράσης, είδος πελάτη, γκρουπ πελατών κλπ.).

Cam003_target_volumes

Περιέχει πληροφορία για τους breakeven στόχους του κάθε είδους ψυγείου ανάλογα την χώρα, δηλαδή τους στόχους που οφείλει να πέτυχει ένα ψυγείο προκειμένου να καλύψει τα λειτουργικά του έξοδα.

Cam015_doors

Περιέχει πληροφορία για τα ψυγεία υπό μελέτη, την αντιστοιχία του καθενός στον εκάστοτε πελάτη καθώς και άλλες βασικές πληροφορίες (ημερομηνία εγκατάστασης, αν είναι ενεργά η όχι, αριθμό πορτών κλπ.).

Cam123_door_openings

Περιέχει πληροφορία για την κινητικότητα του κάθε ψυγείου κάθε μηνά όσον αφορά τις πόρτες του.

Fsm003_sales_direct_indirect

Περιέχει όλη την πληροφορία για τις πωλήσεις του κάθε πελάτη για κάθε μέρα και για κάθε προϊόν προς πώληση.

Fsm121_occupancy

Περιέχει πληροφορία για την πληρότητα του κάθε ψυγείου κάθε μέρα

4.5.2 Πίνακες L2

Avg_occupancy

Περιέχει την πληροφορία για την μέση πληρότητα των ψυγείων του κάθε πελάτη για τους τελευταίους 3 μήνες σε επίπεδο πελάτη, ψυγείου, ημέρας. Στον συγκεκριμένο πίνακα υπολογίζεται ο δείκτης ο οποίος μας δείχνει αν το ψυγείο έχει διαθέσιμη πληροφορία για την πληρότητα.

Cooler_red_2

Περιέχει τη μέση πληρότητα για τον τελευταίο διαθέσιμο μηνά για κάθε ψυγείο σε επίπεδο πελάτη, ψυγείου, μηνά χρησιμοποιώντας τα δεδομένα που υπολογίστηκαν στον προηγούμενο πίνακα. Μεταφέρεται και ο δείκτης πληρότητας για κάθε ψυγείο.

Coolers_3

Περιέχει πληροφορία για τα ψυγεία των πελατών μόνο για τον τελευταίο μηνά (αντιστοιχία με πελάτη, ημερομηνία εγκατάστασης κλπ.). Αποτελεί το βασική πηγή πληροφορίας μας, όλη η υπόλοιπη πληροφορία προσαρμόζεται στο να περιγραφεί τα ψυγεία που προέρχονται από αυτόν τον πίνακα.

Customer_2

Περιέχει την απαραίτητη πληροφορία για τους πελάτες, συν κάποια εξτρά πεδία που υπολογίζονται και είναι χρήσιμα για την περιγραφή του είδους του κάθε πελάτη.

Cooler_customer

Συνδυάζει την πληροφορία των 2 προηγουμένων πινάκων για να δημιουργήσει ένα ενιαίο dataset με όλη την απαραίτητη πληροφορία για την περιγραφή των ψυγείων και των πελατών.

Door_op_original

Περιέχει την πληροφορία για την κινητικότητα των πορτών του κάθε πελάτη για τους προηγουμένους 12 μήνες σε επίπεδο πελάτη, ψυγείου, μηνά.

Door_op_rolling

Περιέχει την αθροιστική πληροφορία για την κινητικότητα των πορτών του κάθε ψυγείου για τους τελευταίους 12 μήνες χρησιμοποιώντας τα δεδομένα του προηγουμένου πίνακα. Στον συγκεκριμένο πίνακα υπολογίζεται ο δείκτης που μας δείχνει αν υπάρχει διαθέσιμη πληροφορία για τις πόρτες για το κάθε ψυγείο.

Sales_weekly

Περιέχει την πληροφορία για τις πωλήσεις του κάθε πελάτη τους τελευταίους 12 μήνες. Συνδυάζει πληροφορία από πίνακες του L2 προκειμένου να κρατήσει μόνο τις πωλήσεις που έχει το κάθε ψυγείο στα έγκυρα προϊόντα και τις φέρνει σε επίπεδο πελάτη, εβδομάδας.

Customer_sales_weekly

Συνδυάζει την πληροφορία από τους πίνακες cooler_customer, cooler_red_2 και sales_weekly προκειμένου να διαμοιράσει καταλληλά τις εβδομαδιαίες πωλήσεις του κάθε πελάτη στα ψυγεία του ανάλογα την ημερομηνία εγκατάστασης τους, την πληρότητα τους και τον αριθμό των πορτών που έχουν. Στο συγκεκριμένο πίνακα υπολογίζεται και ο δείκτης της διαθέσιμης πληροφορίας για κάθε ψυγείο για κάθε εβδομάδα.

Customer_sales_monthly

Περιέχει την αθροιστική πληροφορία για τις διαμοιρασμένες πωλήσεις του κάθε ψυγείου, καθώς και τις μέσες εβδομαδιαίες του πωλήσεις. Μεταφέρεται και ο δείκτης για την διαθεσιμότητα των πωλήσεων.

Wh

Περιέχει την πληροφορία για το distribution των πωλήσεων των Wholesalers συνδυάζοντας πληροφορία από τον customer (για να αναγνωρίσουμε ποιοι πελάτες ανήκουν σε αυτή την κατηγορία), και τα sales για να βρεθούν οι κατάλληλες πωλήσεις για τους συγκεκριμένους πελάτες. Υπολογίζεται επίσης και ο δείκτης που υποδεικνύει αν το συγκεκριμένο ψυγείο ανήκει σε Wholesaler πελάτη.

Customer_sales

Συνδυάζει την πληροφορία για τις πωλήσεις από τους έμμεσους πελάτες και τους Wholesalers για να δημιουργηθεί ένα ενιαίο dataset με την πληροφορία των πωλήσεων για όλα τα ψυγεία υπό μελέτη.

Customer_sales_door_occ

Συνδυάζει την πληροφορία του από πάνω πίνακα και των πινάκων με την πληρότητα και την κινητικότητα των ψυγείων προκειμένου να δημιουργηθεί ένα συνολικό dataset με όλη την απαραίτητη πληροφορία για τους δείκτες και τις τιμές των πωλήσεων, της κινητικότητας και της πληρότητας του κάθε ψυγείου.

Customer_group

Με βάση το παραπάνω πίνακα υπολογίζονται τα γκρουπ πληροφορίας χρησιμοποιώντας τους δείκτες που έχουν υπολογιστεί.

Customer_targets

Στον προηγούμενο πίνακα προστίθεται η πληροφορία για τους στόχους του ψυγείου ανάλογα τη χώρα και το είδος του κάθε ψυγείου.

Full_data

Στον προηγούμενο πίνακα υπολογίζεται η μέση πληρότητα της κάθε χωράς για να χρησιμοποιηθεί σαν imputed πληροφορία στα ψυγεία που δεν έχουν διαθέσιμη πληροφορία για την πληρότητα τους. Το αποτέλεσμα του συγκεκριμένου dataset έχει όλη την απαραίτητη πληροφορία για να υπολογιστεί η τελική κατηγοριοποίηση της απόδοσης.

Final_classification

Με βάση τα δεδομένα του προηγουμένου πίνακα εφαρμόζονται οι φόρμουλες προκειμένου να υπολογιστεί η τελική κατηγοριοποίηση της απόδοσης του κάθε ψυγείου.

4.5.3 Πίνακας Golden

Cooler_output

Περιέχει το τελικό output του κώδικα μας, combined με κάποια πεδία που έρχονται από το L2 και βοηθούν στην πληρότερη περιγραφή των ψυγείων.

4.6 Αναλυτική Περιγραφή της Υλοποίησης

Παρακάτω δίνεται η λεπτομερής διαδικασία δημιουργίας του κάθε πίνακα

4.6.1 Πίνακες L2

Calendar

Δεδομένα Εισόδου:

fiscal_calendar (L1)

Σκοπός:

Δημιουργία 4-4-5 calendar

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα του calendar από το source και δημιουργείται η στήλη week_relative_counter, η οποία ξεκινάει από την πρώτη εβδομάδα του 2003 και αριθμεί όλες τις επόμενες εβδομάδες με έναν αύξοντα αριθμό (για παράδειγμα, η πρώτη εβδομάδα του 2003 θα έχει week_relative_counter = 1, η δεύτερη 2, η πρώτη εβδομάδα του 2004 θα έχει week_relative_counter = 53 και ούτω καθεξής). Το συγκεκριμένο πεδίο είναι χρήσιμο για τη μετάβαση σε μεταγενέστερη (η προγενέστερη) χρονική περίοδο με χρήση απλής πρόσθεσης (η αφαίρεσης), χωρίς να υπάρχει περιορισμός στην αλλαγή του έτους.

Cam033

Δεδομένα Εισόδου:

nc_targets (L1)

Σκοπός:

Δημιουργία country specific target dataset

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για τους στόχους πωλήσεων της κάθε χώρας.

Cam034_max_pack_size

Δεδομένα Εισόδου:

max_pack_size (L1)

Σκοπός:

Δημιουργία πληροφορίας για τις έγκυρες συσκευασίες της κάθε χώρας

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για τους στόχους πωλήσεων της κάθε χώρας.

Equipment

Δεδομένα Εισόδου:

ca_0_equip_equipment (L1)

Σκοπός:

Δημιουργία cooler master data universe

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για το συνολικό universe των ψυγείων, δηλαδή ποια ψυγεία θεωρεί η εταιρεία ότι είναι έγκυρα. Για κάθε ψυγείο κρατείται μόνο η τελευταία εικόνα που έχουμε διαθέσιμη για αυτό.

Material

Δεδομένα Εισόδου:

ca_0_mat03_material (L1)

Σκοπός:

Δημιουργία material master data universe

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για το συνολικό universe των προϊόντων, δηλαδή ποια προϊόντα θεωρεί η εταιρεία ότι είναι έγκυρα.

sku_conversion

Δεδομένα Εισόδου:

ca_mm_09_sku_conversion (L1)

Σκοπός:

Δημιουργία dataset για τη μετατροπή των μονάδων μέτρησης

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνεται ο πίνακας που μετατρέπει διάφορες μονάδες μέτρησης.

Customer

Δεδομένα Εισόδου:

ca_0_cus03_customer (L1)

Σκοπός:

Δημιουργία customer universe

Περιγραφή:

Στο συγκεκριμένο process φορτώνονται τα δεδομένα για το συνολικό universe των πελατών, δηλαδή ποιοι πελάτες θεωρεί η εταιρεία ότι είναι έγκυροι.

Cam003_target_volumes

Δεδομένα Εισόδου:

ca_co_11_m_cde_target_volumes (L1)

ca_0_equip_equipment (L1)

Σκοπός:

Δημιουργία πίνακα με τα λειτουργικά thresholds της κάθε χώρας

Περιγραφή:

Στο συγκεκριμένο process φορτώνονται τα δεδομένα για τα λειτουργικά thresholds για τα διαθέσιμα ψυγεία της κάθε χώρας ανά έτος και κατηγορία εμπορικού καναλιού πώλησης (Future consumption, Immediate Consumption, Modern Trade). Η πληροφορία αυτή συνδυάζεται με τα δεδομένα του universe των ψυγείων προκειμένου να εξαιρεθεί η οποιαδήποτε πληροφορία για μη έγκυρα ψυγεία. Κρατώνται τα λειτουργικά thresholds του τελευταίου διαθέσιμου έτους για κάθε χώρα, και σε περίπτωση πολλαπλών

καταχωρήσεων λειτουργικών thresholds για κάποιο ψυγείο/κατηγορία, κρατείται η μεγαλύτερη τιμή.

Cam015_doors

Δεδομένα Εισόδου:

ca_cam015_cde_doors (L1)

Σκοπός:

Δημιουργία dataset για τα ψυγεία των πελατών

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για τα ψυγεία των πελατών ανά μήνα. Αποτελεί τη βασική πηγή πληροφορίας του εργασίας, αφού τα ψυγεία του τελικού αποτελέσματος θα είναι τα έγκυρα ψυγεία από τη συγκεκριμένη πηγή δεδομένων για τον τελευταίο μήνα.

Cam123_door_openings

Δεδομένα Εισόδου:

ca_cam123_door_openings (L1)

Σκοπός:

Δημιουργία δεδομένων για την κινητικότητα των πορτών των ψυγείων των πελατών

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για την κινητικότητα των ψυγείων των πελατών. Από τα αρχικά δεδομένα εισόδου η πληροφορία έρχεται ανά πελάτη, ψυγείο και ημερομηνία μέτρησης της κινητικότητας.

Fsm003_sales_direct_indirect

Δεδομένα Εισόδου:

ca_fsm003_sales_direct_indirect (L1)

Σκοπός:

Δημιουργία dataset για τις πωλήσεις όλων των πελατών

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για τις πωλήσεις του κάθε πελάτη ανά μέρα για κάθε προϊόν.

Fsm121_occupancy

Δεδομένα Εισόδου:

ca_fsm121_red_coolers (L1)

Σκοπός:

Δημιουργία δεδομένων για την πληρότητα των ψυγείων των πελατών

Περιγραφή:

Στη συγκεκριμένη διαδικασία φορτώνονται τα δεδομένα για την πληρότητα των ψυγείων των πελατών. Από την πηγή η πληροφορία έρχεται ανά πελάτη, ψυγείο και ημερομηνία μέτρησης της πληρότητας.

4.6.2 Πίνακες L3

Avg_occupancy

Δεδομένα Εισόδου:

fsm121_red_coolers (L2)

Σκοπός:

Δημιουργία δεδομένων για την πληρότητα των ψυγείων των πελατών τους τελευταίους 3 μήνες

Περιγραφή:

Χρησιμοποιώντας την πληροφορία που έρχεται από το L2 σχετικά με την πληρότητα των ψυγείων, εκτελούνται τα ακόλουθα:

1. Φιλτράρονται τα δεδομένα εισόδου ώστε να κρατήσουμε μόνο την πληροφορία για τους 3 τελευταίους μήνες.
2. Μετασχηματίζεται η πληροφορία σε επίπεδο μέρας/πελάτη/ψυγείου προκειμένου να χρησιμοποιηθεί σε επόμενο Βήμα και υπολογίζεται το avg(occupancy) σε αυτό το επίπεδο.
3. When avg(occupancy) >= 0 then 1 else 0 (has_ocupancy_last_3_months)

Cooler_red_2

Δεδομένα Εισόδου:

Avg_occupancy (L3)

Σκοπός:

εξαγωγή την τελευταία εικόνα available για την πληρότητα των ψυγείων.

Περιγραφή:

Με βάση τον πίνακα που υπολογίστηκε στο προηγούμενο Βήμα, εφαρμόζεται ο ακόλουθος υπολογισμός προκειμένου να εξαχθεί η τελευταία εικόνα που υπήρχε διαθέσιμη για κάθε ψυγείο.

```
cooler_red_2 = avg_occupancy.withColumn("rn",
    row_number().over(Window.partitionBy("equipment")
    .orderBy(col("FISCPER").desc(), col("CALDAY").desc))
    .filter(col("rn") == 1)
```

Coolers_3

Δεδομένα Εισόδου:

cam015_doors (L2)

Σκοπός:

Δημιουργία δεδομένων με τις απαραίτητες στήλες για την πληροφορία των ψυγείων.

Περιγραφή:

Στη συγκεκριμένη διαδικασία επιλέγονται μόνο τις απαραίτητες στήλες που χρειάζονται, καθώς επίσης διατηρείται την πληροφορία για τα ψυγεία τα οποία ανήκουν στις ακόλουθες κατηγορίες:

1. placed
2. found
3. not active

Customer_2

Δεδομένα Εισόδου:

customer (L2)

Σκοπός:

Δημιουργία δεδομένων με τις απαραίτητες στήλες για την πληροφορία των πελατών.

Περιγραφή:

Στη συγκεκριμένη διαδικασία επιλέγονται μόνο τις απαραίτητες στήλες, καθώς επίσης κατασκευάζονται συγκεντρωτικές κατηγορίες καναλιών πώλησης χρησιμοποιώντας την πληροφορία των επιμέρους καναλιών

Cooler_customer

Δεδομένα Εισόδου:

Coolers_3 (L3)

Customer_2 (L3)

Σκοπός:

Δημιουργία πίνακα όπου συνδυάζεται η πληροφορία από το σύνολο των ψυγείων με την πληροφορία από τους πελάτες.

Περιγραφή:

Με την πληροφορία των ψυγείων σαν αναφορά, εφαρμόζεται ένα left join με τον πίνακα customer_2 προκειμένου να έρθει πληροφορία όπως τα κανάλια πώλησης του κάθε πελάτη, τη γεωγραφική περιοχή δράσης, κλπ.

Door_op_original

Δεδομένα Εισόδου:

cam123_door_openings (L2)

Σκοπός:

Δημιουργία πίνακα με την πληροφορία για τα door_openings τους τελευταίους 12 μήνες.

Περιγραφή:

Χρησιμοποιώντας την πληροφορία που έρχεται από το L2 σχετικά με την κινητικότητα των πορτών των ψυγείων, εκτελούνται τα ακόλουθα:

1. Τα δεδομένα εισόδου φιλτράρονται ώστε να διατηρηθεί μόνο η πληροφορία για τους 12 τελευταίους μήνες.
2. Μετασχηματίζονται τα δεδομένα του Βήματος 1 σε επίπεδο μήνα/πελάτη/ψυγείου προκειμένου να χρησιμοποιηθεί σε επόμενο Βήμα και υπολογίζονται τα ακόλουθα πεδία σε αυτό το επίπεδο:

Tot_openings = sum(total_openings)

Calendar_days = avg(calendar_days)

Tot_active_days = sum(active_days)

Door_op_rolling

Δεδομένα Εισόδου:

Door_op_original (L3)

Σκοπός:

Δημιουργία δεδομένων με την αθροιστική πληροφορία για τα door openings τους τελευταίους 12 μήνες

Περιγραφή:

Με βάση τον πίνακα που υπολογίστηκε στο προηγούμενο Βήμα, μετασχηματίζεται ο πίνακας εισόδου σε επίπεδο πελάτη/ψυγείου και υπολογίζονται τα ακόλουθα:

tot_openings_rolling = sum(tot_openings)

tot_calendar_days_rolling = sum(calendar_days)

tot_active_days_rolling = sum(tot_active_days)

Sales_weekly

Δεδομένα Εισόδου:

fsm003_sales_direct_indirect (L2)

sku_conversion (L2)

calendar (L2)

customer (L2)

cam034_max_pack_size (L2)

material (L2)

Σκοπός:

Δημιουργία δεδομένων που περιέχουν τις έγκυρες πωλήσεις των πελατών υπό μελέτη ανά βδομάδα για τους τελευταίους 12 μήνες.

Περιγραφή:

Προκειμένου να κατασκευαστεί η απαραίτητη πληροφορία για τις πωλήσεις, οφείλουν να εξαιρεθούν πωλήσεις οι οποίες:

- ανήκουν σε πελάτες που θεωρούνται μη έγκυροι (μόνο οι πελάτες που θεωρούνται έγκυροι, δηλαδή αυτοί που βρίσκονται στον πίνακα customer θα υπάρχουν στο τελικό πίνακα εξόδου)
- αφορούν materials που θεωρούνται έγκυρα (μόνο τα materials που έχουν θεωρηθεί ότι είναι σωστά, δηλαδή αυτούς που βρίσκονται στον πίνακα material)
- αφορούν μεγέθη συσκευασιών που έχει εγκρίνει η κάθε χώρα (συσκευασίες που ανήκουν στον πίνακα cam034)

Αναλυτική Περιγραφή

Βήμα 2: Έχοντας σαν βάση τα δεδομένα του Βήματος 1, εκτελείται ένα inner join με τον πίνακα material προκειμένου να διατηρηθεί η πληροφορία για τα προϊόντα που θεωρούμε έγκυρα (συνεπώς εξαιρούνται μη έγκυρες πωλήσεις λόγω μη έγκυρων προϊόντων). Επιπλέον, προστίθεται η πληροφορία του μεγέθους της συσκευασίας για κάθε έγκυρο προϊόν από τον material για να χρησιμοποιηθεί στο επόμενο Βήμα. Βήμα 3: Εκτελείται ένα inner join με τον πίνακα cam034 προκειμένου να διατηρηθεί η πληροφορία για τις συσκευασίες προϊόντων που έχει εγκρίνει η κάθε χώρα. Δημιουργήθηκε λοιπόν ένα σετ δεδομένων το οποίο περιέχει όλες τις έγκυρες πωλήσεις σε επίπεδο πελάτη, ημέρας και προϊόντος. Μένει να το φέρουμε στο κατάλληλο επίπεδο (πελάτη, εβδομάδας) και να φιλτραριστούν μόνο οι πωλήσεις των τελευταίων 12 μηνών. Αυτό επιτυγχάνεται κάνοντας τα ακόλουθα Βήματα:

Βήμα 4: Inner join των δεδομένων του Βήματος 3 με τον πίνακα calendar προκειμένου να έρθει η απαραίτητη πληροφορία για εβδομάδες, μήνες.

Βήμα 5: Τα δεδομένα φιλτράρονται για τους τελευταίους 12 μήνες

Βήμα 6: group by (customer, week) για να έρθει το dataset μας στο επιθυμητό επίπεδο.

Customer_sales_weekly

Δεδομένα Εισόδου:

cooler_customer (L3)

sales_weekly (L3)

cooler_red_2 (L3)

calendar (L2)

Σκοπός:

Στη συγκεκριμένη διαδικασία υπολογίζουμε 3 απαραίτητα πεδία:

1. Distributed_sales_cs: Οι διαμοιρασμένες πωλήσεις του εκάστοτε πελάτη στα ψυγεία του ανά εβδομάδα για τους τελευταίους 12 μήνες
2. Calendar_weeks_installed: Ο αριθμός των εβδομάδων που ένα ψυγείο είναι εγκατεστημένο
3. Has_sales_last_year: Ο δείκτης που μας δείχνει αν το συγκεκριμένο ψυγείο είχε πωλήσεις κάποια εβδομάδα.

Περιγραφή:

Υπολογισμός Distributed_sales_cs

Δεδομένου ότι ο πίνακας των πωλήσεων περιέχει την πληροφορία για τις συνολικές πωλήσεις του κάθε πελάτη για κάποια συγκεκριμένη εβδομάδα και δεν υπάρχουν κάπου διαθέσιμες οι ακριβείς πωλήσεις του κάθε ψυγείου, θα πρέπει να εφαρμοστούν οι ακόλουθοι κανόνες για να κατανεμηθούν οι πωλήσεις στα ψυγεία:

1. Στην κατανομή των πωλήσεων της κάθε εβδομάδας λαμβάνουν μέρος μόνο τα ψυγεία που ήταν εγκατεστημένα (αφού ψυγεία που εγκαταστάθηκαν αργότερα δεν θα μπορούσαν να έχουν πωλήσεις).
2. Αν κάποιο ψυγείο είχε μηδενική πληρότητα δεν θα συμπεριληφθεί στην κατανομή των πωλήσεων.
3. Ο διαμοιρασμός των πωλήσεων βασίζεται στον αριθμό των πορτών ενός εγκατεστημένου ψυγείου σε σχέση με τις συνολικές πόρτες των εγκατεστημένων ψυγείων του πελάτη τη δεδομένη εβδομάδα.

Για να γίνει πιο κατανοητή η όλη λογική, ας υποθέσουμε το ακόλουθο απλό σενάριο:

Έστω πελάτης Α, ο οποίος έχει για την εβδομάδα 202014 100 πωλήσεις και έχει στην κατοχή του τα ακόλουθα ψυγεία

<u>Ψυγείο Α1</u>:

εβδομάδα εγκατάστασης: 202010

πόρτες: 2

πληρότητα: 80

<u>Ψυγείο Α2</u>:

εβδομάδα εγκατάστασης: 202012

πόρτες: 1

πληρότητα: 50

Ψυγείο Α3:

εβδομάδα εγκατάστασης: 201804

πόρτες: 2

πληρότητα: 0

Ψυγείο Α4:

εβδομάδα εγκατάστασης: 201936

πόρτες: 1

πληρότητα: 20

Ψυγείο Α5:

εβδομάδα εγκατάστασης: 202102

πόρτες: 1

πληρότητα: 100

Με βάση τους παραπάνω κανόνες, τα ψυγεία Α3 και Α5 εξαιρούνται από την κατανομή των πωλήσεων της εβδομάδας 202014 λόγω της μηδενικής πληρότητας και της μεταγενέστερης εγκατάστασης αντίστοιχα. Στα υπόλοιπα 3 ψυγεία, με βάση τον κανόνα 3, γίνεται η ακόλουθη κατανομή:

- Α1: 2 από τις 4 συνολικά active πόρτες εκείνη την εβδομάδα, συνεπώς του αντιστοιχεί το 1/2 των πωλήσεων της εβδομάδας, αρά sales_A1 = 50
- Α2: 1 από τις 4 συνολικά active πόρτες εκείνη την εβδομάδα, συνεπώς του αντιστοιχεί το 1/4 των πωλήσεων της εβδομάδας, αρά sales_A2 = 25
- Α4: 1 από τις 4 συνολικά active πόρτες εκείνη την εβδομάδα, συνεπώς του αντιστοιχεί το 1/4 των πωλήσεων της εβδομάδας, αρά sales_A4 = 25

Αναλυτική Περιγραφή

Προκειμένου να υλοποιηθεί η προαναφερθείσα εικόνα εφαρμόζεται η διαδικασία που περιγράφεται παρακάτω:

Βήμα 1: Υπολογισμός Εβδομάδας εγκατάστασης του κάθε ψυγείου (installed_week)

Από τα δεδομένα εισόδου του πίνακα cooler_customer έχουμε την ημερομηνία εγκατάστασης του κάθε ψυγείου. Κάνοντας ένα left join του πίνακα cooler_customer με τον πίνακα calendar πάνω στην ημερομηνία εξάγεται η εβδομάδα που εγκαταστάθηκε το κάθε ψυγείο (installed_week).

Βήμα 2: Υπολογισμός Install_flag1

Προκειμένου να υπολογιστεί αν το ψυγείο ήταν όντως εγκατεστημένο την εβδομάδα πώλησης πρέπει πρώτα να συνδυαστεί η πληροφορία του πίνακα που υπολογίστηκε στο προηγούμενο Βήμα με τον πίνακα sales_weekly. Με ένα left join του πίνακα του Βήματος 1 με τον πίνακα sales_weekly στο πεδίο customer, κατασκευάζουμε ένα σύνολο δεδομένων το οποίο για κάθε ψυγείο/πελάτη/εβδομάδα θα έχει την εβδομάδα εγκατάστασης του ψυγείου και τις συνολικές πωλήσεις που είχε ο πελάτης εκείνη τη βδομάδα (base_to_cs).

Στη συνέχεια, με χρήση της ακόλουθης συνθήκης:

> When (installed_week <= sales_week) then 1 else 0

Βρίσκεται αν το συγκεκριμένο ψυγείο ήταν εγκατεστημένο την συγκεκριμένη εβδομάδα πωλήσεων (installed_flag1).

Βήμα 3: έλεγχος πληρότητας ψυγείων και τελικό flag των ψυγείων που θα συμπεριληφθούν στην κατανομή πωλήσεων

Με την εφαρμογή ενός left join των ανωτέρων δεδομένων με τον cooler_red_2 (πίνακας τελικής πληρότητας) πάνω στο πεδίο equipment, συνδυάζεται η πληροφορία πωλήσεων και πληρότητας.

Στη συνέχεια, με χρήση της ακόλουθης συνθήκης:

> When (installed_flag1 = 1) & (occupancy != 0) then 1 (το ψυγείο έχει πληροφορία για πληρότητα και δεν είναι μηδενική, οπότε συμπεριλαμβάνεται στην κατανομή)
>
> When (installed_flag1 = 1) & (occupancy is null) then 1 (το ψυγείο δεν έχει διαθέσιμη πληροφορία για πληρότητα, οπότε συμπεριλαμβάνεται στην κατανομή)
>
> else 0 (το ψυγείο είχε πληροφορία για πληρότητα η οποία είναι μηδενική, εξαιρείται από την κατανομή)

Βρίσκονται τα ψυγεία που πρέπει να συμπεριληφθούν στην κατανομή για όλες τις εβδομάδες πωλήσεων (distribution flag).

Βήμα 4: Υπολογισμός active_doors (σύνολο πορτών των ψυγείων του πελάτη που είναι εγκατεστημένα για συγκεκριμένη εβδομάδα)

Αρχικά υπολογίζονται οι ενεργές πόρτες των ψυγείων του κάθε πελάτη για κάθε εβδομάδα.

> Active_doors_prep = n_doors * distribution_flag

Εφαρμόζοντας τον ακόλουθο υπολογισμό

> sum("active_doors_prep").over("customer", "cchbc_week")

βρίσκεται για κάθε πελάτη/εβδομάδα το σύνολο των πορτών από τα ψυγεία που ήταν εγκατεστημένα (active_doors).

Βήμα 5:

Έχοντας όλα τα απαραίτητα στοιχεία υπολογισμένα, εφαρμόζεται ο τελικός υπολογισμός του πεδίου distributed_sales_cs με χρήση του ακολούθου τύπου:

Distributed_sales_cs = base_to_cs * distribution_flag * n_doors / active_doors

Όπου:

Base_to_cs: συνολικές πωλήσεις πελάτη για τη συγκεκριμένη εβδομάδα

Distribution_flag: flag που δείχνει αν το ψυγείο συμπεριλαμβάνεται στο distribution των sales την συγκεκριμένη εβδομάδα

N_doors: αριθμός πορτών του ψυγείου

Active_doors: σύνολο αριθμών πορτών των ψυγείων του πελάτη που συμπεριλαμβάνονται στο distribution των sales την συγκεκριμένη εβδομάδα.

Υπολογισμός calendar_weeks_installed

Έχοντας την εβδομάδα εγκατάστασης του κάθε ψυγείου από το Βήμα 1 μπορεί να υπολογιστεί το σύνολο των εβδομάδων στο οποίο ένα ψυγείο είναι εγκατεστημένο αφαιρώντας την εβδομάδα εγκατάστασης από την εβδομάδα την οποία εκτελέστηκε ο κώδικας. Σε περίπτωση που κάποιο ψυγείο είναι εγκατεστημένο για περισσότερο από 1 χρόνο τότε το συγκεκριμένο πεδίο θέτεται αυτόματα ίσο με 52. Ο λόγος πίσω από αυτό είναι ότι επειδή το συγκεκριμένο πεδίο θα χρησιμοποιηθεί στο επόμενο στάδιο για να βρεθεί ο μέσος ορός των εβδομαδιαίων πωλήσεων και δεδομένου ότι οι πωλήσεις αφορούν το διάστημα των τελευταίων 52 εβδομάδων, αυτή είναι και η μέγιστη τιμή που μπορεί να πάρει το συγκεκριμένο πεδίο.

Υπολογισμός has_sales_last_year

Έχοντας υπολογίσει τις κατανεμημένες πωλήσεις του κάθε ψυγείου στο προηγούμενο στάδιο, υπολογίζεται σε επίπεδο εβδομάδας το απαραίτητο πεδίο χρησιμοποιώντας την ακόλουθη συνθήκη:

When (distributed_sales_cs != NULL) then 1 (αν το ψυγείο έχει διαθέσιμη πληροφορία για πωλήσεις, είτε μηδενική είτε όχι τότε το πεδίο παίρνει την τιμή 1)

else 0 (δεν υπάρχει πληροφορία εκείνη τη βδομάδα για πωλήσεις στο συγκεκριμένο ψυγείο)

Customer_sales_monthly

Δεδομένα Εισόδου:

Customer_sales_weekly (L3)

Σκοπός:

Στη συγκεκριμένη διαδικασία εκτελούνται τα ακόλουθα:

1. Μετατροπή του Δεδομένα Εισόδου πίνακα από επίπεδο εβδομάδας/πελάτη/ψυγείου σε επίπεδο πελάτη/ψυγείου & Υπολογισμός πεδίου has_sales_last_year σε επίπεδο ψυγείου.
2. Υπολογισμός του μ.ό. εβδομαδιαίων πωλήσεων του κάθε ψυγείου (avg_distributed_sales_cs)

Περιγραφή

Βήμα 1: Μετατροπή πίνακα

Εφαρμόζοντας ένα group by (customer, cooler) στα δεδομένα εισόδου και φτιάχνοντας τα ακόλουθα πεδία:

Total_distributed_sales_cs = sum(distributed_sales_cs)

Calendar_weeks_installed = max(calendar_weeks_installed)

Has_sales_last_year = max(has_sales_last_year)

Η πληροφορία εισόδου μετατρέπεται στο σωστό επίπεδο και όλα τα απαραίτητα πεδία του πίνακα έχουν έρθει από επίπεδο εβδομάδας σε επίπεδο ψυγείου πλέον.

Βήμα 2: Υπολογισμός avg_distributed_sales_cs

Εφόσον τα δεδομένα έχουν μετασχηματιστεί ώστε να είναι σε επίπεδο ψυγείου, προκειμένου να υπολογιστούν οι μέσες πωλήσεις ανά εβδομάδα γίνεται ο ακόλουθος υπολογισμός:

Avg_distributed_sales_cs = total_distributed_sales_cs / calendar_weeks_installed

Wh

Δεδομένα Εισόδου:

customer (L2)

sales_weekly (L3)

customer_sales_monthly (L3)

calendar (L2)

Σκοπός:

Στη συγκεκριμένη διαδικασία υπολογίζεται το ακόλουθο πεδίο για τους Wholesalers

cw_avg_distributed_base_to_cs

το συγκεκριμένο πεδίο αντιπροσωπεύει τις μέσες εβδομαδιαίες πωλήσεις των ψυγείων που ανήκουν σε Wholesaler πελάτες.

Περιγραφή:

Όπως έχει προαναφερθεί και σε προηγούμενο υποκεφάλαιο, εκτός από τους πελάτες για τους οποίους δίνεται άμεση πληροφορία για τις πωλήσεις τους, υπάρχει και μια κατηγορία πελατών οι οποίοι προμηθεύονται σε δεύτερο χρόνο προϊόντα και δεν υπάρχει άμεση πληροφορία για τις πωλήσεις τους. Προκειμένου να υπολογιστεί ένας εβδομαδιαίος μέσος ορός πωλήσεων για τους συγκεκριμένους πελάτες θα ακολουθηθεί παρόμοια λογική με αυτή των αμέσων πελατών, με τη διαφορά ότι αφού δεν υπάρχει πληροφορία για τις πωλήσεις των πελατών αυτών θα συμπεριληφθούν οι πωλήσεις του customer_planning[1] στο οποίο ανήκει, όπως επίσης και τις πόρτες όλων των ψυγείων που ανήκουν στο συγκεκριμένο cust_planning επίπεδο.

Αναλυτική Περιγραφή

Η αναλυτική Περιγραφή των Βημάτων δίνεται παρακάτω.

Βήμα 1: Εύρεση των Wholesalers

Χρησιμοποιώντας τον πίνακα customer από το L2, αναγνωρίζονται οι wholesaler πελάτες με χρήση του φίλτρου order_block = ‘CW’. Επιπλέον, προστίθεται η πληροφορία που έχει ο κάθε πελάτης CW για το cust_planning level του (προκειμένου να χρησιμοποιηθεί στο μέλλον για να αναγνωριστούν ποιες πωλήσεις θα πρέπει να θεωρηθούν ότι του αντιστοιχούν).

Βήμα 2: δημιουργία δεδομένων πωλήσεων σε επίπεδο cust_planning

Χρησιμοποιώντας τον πίνακα sales_weekly από το L3, ο οποίος έχει την πληροφορία για τις πωλήσεις των άμεσων πελατών ανά εβδομάδα, εφαρμόζεται ένα group by με σκοπό να υπολογιστούν οι συγκεντρωτικές πωλήσεις σε επίπεδο cust_planning level. Εφαρμόζεται λοιπόν το ακόλουθο:

SALES = (

[1] Κάθε πελάτης, άμεσος ή έμμεσος, ανήκει σε ένα customer_planning γκρουπ. Πρόκειται για μια κατηγοριοποίηση των πελατών ανάλογα τον τομέα τον οποίο δραστηριοποιούνται (εστιατόριο, σουπερμάρκετ, κλπ.). Συνεπώς η συγκεκριμένη υλοποίηση υποθέτει ότι οι έμμεσοι πελάτες θα ακολουθήσουν παρόμοια συμπεριφορά πωλήσεων με τους αντίστοιχους άμεσους που ανήκουν στο ίδιο cust_planning γκρουπ.

```
        sales_weekly.select(
            "customer_planning_level", "sales_cs", "cchbc_week"
        )
        .groupBy("customer_planning_level", "cchbc_week")
        .agg(
            F.sum("sales_cs").alias("phc_sales"),
        )
    )
```

Βήμα 3: Εύρεση των ψυγείων των wholesalers πελατών

Έχοντας αναγνωρίσει τους wholesalers από το Βήμα 1 και έχοντας ως δεδομένα εισόδου τον πίνακα customer_sales_weekly (ο συμπεριλαμβάνει την πληροφορία για όλα τα ψυγεία υπό μελέτη καθώς και πληροφορία για την ημερομηνία εγκατάστασης τους και το πόσες εβδομάδες έχουν εγκατασταθεί), εφαρμόζεται ένα inner join των 2 αυτών συνόλων δεδομένων στο πεδίο customer προκειμένου να εξαχθούν τα ψυγεία τα οποία ανήκουν στους wholesalers (διατηρώντας και την πληροφορία για το cust_planning level του κάθε πελάτη).

Βήμα 4: Κατανομή των πωλήσεων

Μέχρι τώρα, έχουν κατασκευαστεί 2 βασικά σύνολα δεδομένων:

1. Wholesaler cooler dataframe (Βήμα 3) (cw_cooler)
2. Cust_planning_level sales dataframe (Βήμα 2) (cw_sales)

Είναι εμφανές ότι τα παραπάνω σύνολα δεδομένων είναι πανομοιότυπα με αυτά που είχαν κατασκευαστεί για τους άμεσους πελάτες στον πίνακα customer_sales_weekly. Μονή ουσιαστική διαφορά είναι ότι αντί για επίπεδο πελάτη/ψυγείου/εβδομάδας τώρα μιλάμε για επίπεδο cust_planning level/ψυγείο/εβδομάδα. Θα εφαρμοστεί λοιπόν η ιδιά διαδικασία για τον υπολογισμό της τελικής κατανομής των πωλήσεων.

Προκειμένου να υπολογιστεί αν το ψυγείο ήταν όντως εγκατεστημένο την εβδομάδα πώλησης πρέπει να συνδυαστεί η πληροφορία του συνόλου δεδομένων cw_cooler με το σύνολο δεδομένων cw_sales. Με ένα left join του cw_cooler με τον cw_sales στο πεδίο cust_planning level, κατασκευάζεται ένα νέο σύνολο δεδομένων το οποίο για κάθε ψυγείο/cust_planning/εβδομάδα θα έχει την εβδομάδα εγκατάστασης του ψυγείου και τις συνολικές πωλήσεις που είχε το συγκεκριμένο cust_planning εκείνη τη βδομάδα (sales_cs).

Στη συνέχεια, με χρήση της ακόλουθης συνθήκης:

When (installed_week <= sales_week) then 1 else 0

Βρίσκεται αν το συγκεκριμένο ψυγείο ήταν εγκατεστημένο την συγκεκριμένη εβδομάδα πωλήσεων (cw_installed_flag).

Βήμα 5: Υπολογισμός active_doors (σύνολο πορτών των ψυγείων του cust_planning level που είναι εγκατεστημένα για συγκεκριμένη εβδομάδα)

Αρχικά υπολογίζονται οι active πόρτες των ψυγείων του κάθε cust_planning_level για κάθε εβδομάδα.

Cw_active_doors_prep = n_doors * cw_active_doors_flag

Εφαρμόζοντας τον ακόλουθο υπολογισμό

sum("Cw_active_doors_prep").over("cust_planning", "cchbc_week")

υπολογίζεται για κάθε cust_planning/εβδομάδα το σύνολο των πορτών από τα ψυγεία που ήταν εγκατεστημένα (cw_active_doors).

Βήμα 6:

Έχοντας όλα τα απαραίτητα στοιχεία υπολογισμένα, εφαρμόζεται ο τελικός υπολογισμός των cw_distributed_sales_cs με χρήση του ακολούθου τύπου:

Cw_distributed_sales_cs = sales_cs * cw_install_flag * n_doors / cw_active_doors

όπου:

sales_cs: συνολικές πωλήσεις cust_planning για τη συγκεκριμένη εβδομάδα

cw_install_flag: flag που δείχνει αν το ψυγείο συμπεριλαμβάνεται στην κατανομή των πωλήσεων την συγκεκριμένη εβδομάδα

N_doors: αριθμός πορτών του ψυγείου

Cw_active_doors: σύνολο αριθμών πορτών των ψυγείων του cust_planning που συμπεριλαμβάνονται στην κατανομή των πωλήσεων την συγκεκριμένη εβδομάδα.

Βήμα 7: Μετατροπή πίνακα

Εφαρμόζοντας ένα group by (cust_planning, cooler) και φτιάχνοντας τα ακόλουθα πεδία:

Total_cw_distributed_sales_cs = sum(cw_distributed_sales_cs)

Calendar_weeks_installed = max(calendar_weeks_installed)

Έχουν λοιπόν υπολογιστεί όλα τα απαραίτητα πεδία του συνόλου δεδομένων από επίπεδο εβδομάδας σε επίπεδο ψυγείου πλέον.

Βήμα 8: Υπολογισμός avg_cw_distributed_sales_cs

Εφόσον το σύνολο των δεδομένων έχει μετασχηματιστεί σε επίπεδο ψυγείου, προκειμένου να υπολογιστούν οι μέσες πωλήσεις ανά εβδομάδα εφαρμόζεται το ακόλουθο:

Avg_cw_distributed_sales_cs = total_cw_distributed_sales_cs / calendar_weeks_installed

Customer_sales

Δεδομένα Εισόδου:

Customer_sales_monthly (L3)

Wh (L3)

Σκοπός:

Στη συγκεκριμένη διαδικασία δημιουργούνται τα ακόλουθα

1. Συνδυασμός πληροφορίας πωλήσεων για τα ψυγεία των αμέσων και των wholesaler πελατών
2. Υπολογισμός πεδίου wh_flag

Περιγραφή:

Έχοντας πλέον δημιουργήσει 2 ξεχωριστά σύνολα δεδομένων για τις πωλήσεις, δεν μένει πάρα να τους συνδυαστεί η πληροφορία που περιέχουν προκειμένου να ληφθεί τη συνολική εικόνα για όλα τα διαθέσιμα ψυγεία υπό μελέτη. Προκειμένου να επιτευχθεί αυτό, εφαρμόζεται ένα left join μεταξύ του πίνακα customer_sales_monthly και του πίνακα wh πίνακα πάνω στο πεδίο equipment (που συμβολίζει το ψυγείο).

Στη συνέχεια, για να αναγνωριστούν τα ψυγεία που ανήκουν σε wholesaler πελάτες αρκεί να εφαρμοστεί:

When avg_cw_distributed_sales_cs != null then 1 else 0 (wh_flag)

Customer_sales_door_occ

Δεδομένα Εισόδου:

customer_sales (L3)

door_op_rolling (L3)

cooler_red_2 (L3)

Σκοπός:

Σκοπός της συγκεκριμένης διαδικασίας είναι

1. να συνδυάσει την διαθέσιμη πληροφορία για πωλήσεις, την πληρότητα και τις πόρτες για κάθε ψυγείο
2. να υπολογιστούν τα avg_door_op_cal, avg_doop_op_act
3. να υπολογιστούν τα has_doors_last_year, has_occupancy_last_3_months

Περιγραφή:

Από το προηγούμενο Βήμα έχει κατασκευαστεί ένα σύνολο δεδομένων το οποίο περιέχει τις πωλήσεις για όλα τα ψυγεία υπό μελέτη. Από παλαιοτέρα Βήματα έχει επίσης υπολογιστεί η πληρότητα και η πληροφορία για τις πόρτες για κάθε ψυγείο. Αυτό που μένει είναι να συνδυαστεί η πληροφορία για να δημιουργηθεί ένα ενιαίο σύνολο δεδομένων που θα έχει όλη την διαθέσιμη πληροφορία και για τα 3 πεδία τα οποία χρειαζόμαστε προκειμένου να κατηγοριοποιήσουμε τα ψυγεία στα καταλληλά customer_groups.

Έχοντας λοιπόν σαν πίνακα αναφοράς τον πίνακα customer_sales και κάνοντας 2 left join με τους πίνακες door_op_rolling και cooler_red_2 πάνω στα πεδία [customer, equipment] δημιουργούμε το εν λόγω σύνολο δεδομένων.

Στη συνέχεια, χρησιμοποιώντας την πληροφορία των door_openings, υπολογίζονται 2 πεδία με τον ακόλουθο τρόπο:

Avg_doop_op_cal = total_openings / total_calendar_days

Avg_doop_op_act = total_openings / total_active_days

Τέλος, με χρήση της πληροφορίας από τον πίνακα door_op_rolling και τον πίνακα cooler_red_2 υπολογίζονται οι δείκτες has_openings_last_year & has_occupancy_last_3_months αντίστοιχα, με παρόμοιο τρόπο με αυτόν που χρησιμοποιήθηκε για τον υπολογισμό των δεικτών για τις πωλήσεις:

When (Avg_doop_op_act != NULL) then 1 (αν το ψυγείο εχει διαθέσιμη πληροφορία για doors, είτε μηδενική είτε όχι τότε ο δείκτης έχει την τιμή 1)

else 0 (δεν υπάρχει πληροφορία για πόρτες στο συγκεκριμένο ψυγείο)

When (has_occupancy_last_three_months >= 1) then 1 (αν το ψυγείο στον πίνακα cooler_red_2 είχε έστω και μια εγγραφή με πληροφορία για πληρότητα, τότε ο δείκτης παίρνει την τιμή 1)

else 0 (δεν υπάρχει πληροφορία για πληρότητα στο συγκεκριμένο ψυγείο)

Customer_group

Δεδομένα Εισόδου:

Customer_sales_door_occ (L3)

Σκοπός:

Σε αυτή τη διαδικασία υπολογίζεται το customer_group του κάθε ψυγείου με βάση την διαθέσιμη πληροφορία για τα 3 βασικά δεδομένα πηγής (πωλήσεις, πόρτες, πληρότητα).

Περιγραφή:

Έχοντας πλέον ένα ενιαίο σύνολο δεδομένων μπορούν να υπολογιστούν τα customer_groups με βάση τις παρακάτω συνθήκες:

WHEN has_sales_last_year_upd = 1 AND has_openings_last_year_upd = 1 AND has_occupancy_last_three_months = 1 THEN 1

WHEN has_sales_last_year_upd = 1 AND has_openings_last_year_upd = 1 AND has_occupancy_last_three_months <> 1 THEN 2

WHEN has_sales_last_year_upd = 1 AND has_openings_last_year_upd <> 1 AND has_occupancy_last_three_months = 1 THEN 3

WHEN has_sales_last_year_upd = 1 AND has_openings_last_year_upd <> 1 AND has_occupancy_last_three_months <> 1 THEN 4

WHEN has_sales_last_year_upd <> 1 AND has_openings_last_year_upd = 1 AND has_occupancy_last_three_months = 1 THEN 5

WHEN has_sales_last_year_upd <> 1 AND has_openings_last_year_upd = 1 AND has_occupancy_last_three_months <> 1 THEN 6

WHEN has_sales_last_year_upd <> 1 AND has_openings_last_year_upd <> 1 AND has_occupancy_last_three_months = 1 THEN 7

WHEN has_sales_last_year_upd <> 1 AND has_openings_last_year_upd <> 1 AND has_occupancy_last_three_months <> 1 AND wh_flag IS NOT NULL THEN 8

ELSE NULL

Customer_targets

Δεδομένα Εισόδου:

Customer_group (L3)

cam123_door_openings (L2)

cam033 (L2)

cam003_target_volumes (L2)

Σκοπός:

Υπολογισμός των ακολούθων:

1. door_openings target (breakeven_threshold)
2. Προσθήκη των στόχων πωλήσεων και πορτών
3. Υπολογισμός Performance_cs

Περιγραφή:

Προκειμένου να εφαρμοστούν οι κατάλληλοι τύποι για να υπολογιστεί η κατηγορία απόδοσης του κάθε ψυγείου, είναι απαραίτητοι οι στόχοι που έχει θέσει κάθε χώρα για τις πωλήσεις και την κινητικότητα των πορτών. Για να γίνει αυτό ακολουθείται η ακόλουθη διαδικασία:

Βήμα 1: Υπολογισμός door_opening targets

Χρησιμοποιώντας ως Δεδομένα Εισόδου τον πίνακα cam123_door_openings από το L2 και φέρνοντας τον σε επίπεδο πελάτη/ψυγείου, υπολογίζεται για τους τελευταίους 12 μήνες το ακόλουθο:

Breakeven_threshold = avg(breakeven_threshold)

Βήμα 2: Προσθήκη των targets για sales & door_openings

Έχοντας ως πίνακα αναφοράς τον customer_group, κάνοντας left join με τους cam003, cam033 και breakeven_threshold, δημιουργείται ένα σύνολο δεδομένων το οποίο περιέχει τους στόχους πωλήσεων (sales_phc από τον πίνακα cam003 & nc_targets από τον πίνακα cam033) και για door_openings (breakeven_threshold).

Βήμα 3: Υπολογισμός performance_cs

Έχοντας ήδη υπολογίσει τις μέσες εβδομαδιαίες πωλήσεις ανά ψυγείο καθώς και τα targets από το cam003 (sales_phc), υπολογίζουμε το performance_cs ως εξής:

Performance_cs = avg_distributed_sales_cs * 100 / sales_phc

Full_data

Δεδομένα Εισόδου:

customer_sales_doors_occ (L3)

customer_targets (L3)

Σκοπός:

Σκοπός της συγκεκριμένης διαδικασίας είναι να προσθέσει το imputed_occupancy στα ψυγεία τα οποία δεν έχουν πληροφορία για πληρότητα από τα δεδομένα εισόδου της πηγής.

Περιγραφή:

Στα ψυγεία που δεν έχουν πληροφορία για πληρότητα, πρέπει να προστεθεί ο μέσος ορός της πληρότητας της χώρας τους τελευταίους 3 μήνες.

Με χρήση του πίνακα customer_sales_door_occ και διατηρώντας μόνο τις εγγραφές για τις οποίες ισχύει

Has_occupancy_last_3_months = 1

Και βρίσκοντας το avg(occupancy) του παραπάνω συνόλου υπολογίζεται η τιμή ενδιαφέροντος. Με ένα left join του πίνακα customer_targets με το σύνολο δεδομένων που μόλις υπολογίστηκε και θέτοντας

imputed_occupancy = avg(occupancy)

στα ψυγεία τα οποία δεν έχουν πληροφορία για πληρότητα (has_occupancy_last_3_months <> 1) λαμβάνεται ο τελικός πίνακας της διαδικασίας.

Final_classification

Δεδομένα Εισόδου:

Full_data (L3)

Σκοπός:

Τελική κατηγοριοποίηση απόδοσης ψυγείων

Περιγραφή:

Μέχρι το σημείο αυτό έχει δημιουργηθεί ένας πίνακας (full_data) ο οποίος περιέχει όλη την απαραίτητη πληροφορία για τον υπολογισμό της τελικής απόδοσης του κάθε ψυγείου. Για κάθε customer_group χρησιμοποιούνται διαφορετικές συνθήκες για την μέτρηση της απόδοσης, ανάλογα με τη διαθέσιμη πληροφορία. Παρακάτω δίνονται οι συνθήκες που χρησιμοποιούνται ανάλογα τα customer_groups:

Customer_group = 1 or 5

WHEN customer_group IN (1, 5)

AND occupancy <= 10

THEN 'E'

WHEN customer_group IN (1, 5)

```
AND occupancy > 10

AND occupancy <= 50

THEN 'D'


WHEN customer_group IN (1, 5)

AND occupancy > 50

AND (avg_door_op_cal * occupancy / 100) > (breakeven_threshold * (1 +
COALESCE(nc_target, 0) / 100))

THEN 'A'


WHEN customer_group IN (1, 5)

AND occupancy > 50

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold > 1

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold <= (1 +
COALESCE(nc_target, 0) / 100)

THEN 'B'


WHEN customer_group IN (1, 5)

AND occupancy > 50

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold > 0.5

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold <= 1

THEN 'C'


WHEN customer_group IN (1, 5)

AND occupancy > 50

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold > 0

AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold <= 0.5

THEN 'D'
```

```
WHEN customer_group IN (1, 5)
AND occupancy > 50
AND (avg_door_op_cal * occupancy / 100) / breakeven_threshold <= 0
THEN 'E'
```

Customer_group = 3

```
WHEN customer_group = 3
AND occupancy <= 10
THEN 'E'

WHEN customer_group = 3
AND occupancy > 10
AND occupancy <= 50
THEN 'D'

WHEN customer_group = 3
AND occupancy > 50
AND avg_distributed_sales_cs > (sales_phc * (1 + COALESCE(nc_target, 0) / 100))
THEN 'A'

WHEN customer_group = 3
AND occupancy > 50
AND avg_distributed_sales_cs / sales_phc > 1
AND avg_distributed_sales_cs / sales_phc <= (1 + COALESCE(nc_target, 0) / 100)
THEN 'B'

WHEN customer_group = 3
AND occupancy > 50
```

```
AND avg_distributed_sales_cs / sales_phc >  0.5
AND avg_distributed_sales_cs / sales_phc <= 1
THEN 'C'

WHEN customer_group = 3
AND occupancy > 50
AND avg_distributed_sales_cs / sales_phc > 0
AND avg_distributed_sales_cs / sales_phc <=  0.5
THEN 'D'

WHEN customer_group = 3
AND occupancy > 50
AND avg_distributed_sales_cs / sales_phc <= 0
THEN 'E'
```

Customer_group = 7

```
WHEN customer_group = 7
AND occupancy <= 10
THEN 'E'

WHEN customer_group = 7
AND occupancy > 10
AND occupancy <= 50
THEN 'D'

WHEN customer_group = 7
AND occupancy > 50
AND occupancy <= 80
```

```
THEN 'C'

WHEN customer_group = 7
AND occupancy > 80
THEN 'B'
```

Customer_group = 4

```
WHEN customer_group = 4
AND avg_distributed_sales_cs > (sales_phc * (1 + COALESCE(nc_target, 0) / 100))
THEN 'A'

WHEN customer_group = 4
AND avg_distributed_sales_cs / sales_phc > 1
AND avg_distributed_sales_cs / sales_phc <= (1 + COALESCE(nc_target, 0) / 100)
THEN 'B'

WHEN customer_group = 4
AND avg_distributed_sales_cs / sales_phc >  0.5
AND avg_distributed_sales_cs / sales_phc <= 1
THEN 'C'

WHEN customer_group = 4
AND avg_distributed_sales_cs / sales_phc > 0
AND avg_distributed_sales_cs / sales_phc <=  0.5
THEN 'D'

WHEN customer_group = 4
AND avg_distributed_sales_cs / sales_phc <= 0
```

THEN 'E'

Customer_group = 2 or 6

WHEN customer_group IN (2, 6)

AND (avg_door_op_cal * imputed_occupancy / 100) > (breakeven_threshold * (1 + COALESCE(nc_target, 0) / 100))

THEN 'A'

WHEN customer_group IN (2, 6)

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold > 1

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold <= (1 + COALESCE(nc_target, 0) / 100)

THEN 'B'

WHEN customer_group IN (2, 6)

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold > 0.5

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold <= 1

THEN 'C'

WHEN customer_group IN (2, 6)

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold > 0

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold <= 0.5

THEN 'D'

WHEN customer_group IN (2, 6)

AND (avg_door_op_cal * imputed_occupancy / 100) / breakeven_threshold <= 0

THEN 'E'

Customer_group = 8

```
WHEN customer_group = 8
AND cw_avg_distributed_sales_cs > (sales_phc * (1 + COALESCE(nc_target, 0) / 100))
THEN 'A'

WHEN customer_group = 8
AND cw_avg_distributed_sales_cs / sales_phc > 1
AND avg_distributed_sales_cs / sales_phc <= (1 + COALESCE(nc_target, 0) / 100)
THEN 'B'

WHEN customer_group = 8
AND cw_avg_distributed_sales_cs / sales_phc >  0.5
AND avg_distributed_sales_cs / sales_phc <= 1
THEN 'C'

WHEN customer_group = 8
AND cw_avg_distributed_sales_cs / sales_phc > 0
AND avg_distributed_sales_cs / sales_phc <= 0.5
THEN 'D'

WHEN customer_group = 8
AND cw_avg_distributed_sales_cs / sales_phc <= 0
THEN 'E'

ELSE 'not assigned'
```

4.6.3 Πίνακας Golden

Cooler_output

Δεδομένα Εισόδου:

Final_classification (L3)

customer (L2)

equipment (L2)

calendar (L2)

Σκοπός:

Δημιουργία του τελικού πίνακα με την απαραίτητη πληροφορία για τον χρήστη

Περιγραφή:

Εκτός από την τελική κατηγοριοποίηση της απόδοσης των ψυγείων, για τον εκάστοτε χρήστη είναι απαραίτητα και ορισμένα άλλα στοιχεία που περιγραφούν καλυτέρα το σύνολο δεδομένων και κάνουν ευκολότερη την κατανόηση του. Τέτοια πεδία είναι το μοντέλο του ψυγείου, το serial number, το είδος του πελάτη, σε ποια περιοχή δραστηριοποιείται κλπ.

Σε αυτό το Βήμα, επιτυγχάνονται 2 πράγματα

1. Προστίθενται με left join στον τελικό πίνακα του L3 (final_classification) τα απαραίτητα πεδία από τον customer (L2) & equipment (L2), προκειμένου να περιγραφεί με πιο κατανοητό τρόπο στο χρήστη η τελική έξοδος του συστήματος.
2. Διαχωρίζονται τα ψυγεία που είχαν κατηγοριοποιηθεί ως ‘not assigned’ σε κατηγορία Ε ή κατηγορία ‘no data visibility’, ανάλογα με την πληροφορία που έρχεται από τον customer (συγκεκριμένα, αν για ένα ψυγείο ‘not assigned’ ο πελάτης δεν έχει πληροφορία για order_block, τότε θεωρείται ότι δεν έχουν αποσταλεί τα απαραίτητα στοιχεία για την κατηγοριοποίηση του ψυγείου, οπότε δημιουργείται η κατηγορία ‘no data visibility’. Σε περίπτωση που το order_block έχει τιμή σημαίνει ότι έχουν έρθει τα απαραίτητα στοιχεία, απλώς είναι μηδενικά, οπότε το ψυγείο κατηγοριοποιείται στην κατηγορία Ε).

Κεφάλαιο 5: Εκτέλεση Πειραμάτων

Για τους σκοπούς της εργασίας, υλοποιήθηκαν ορισμένα pipelines προκειμένου να είναι ευκολότερη η συνολική εκτέλεση του κώδικα και η δημιουργία του τελικού output. Μέσω του pipeline γίνεται το απαραίτητο orchestration των processes για την ομαλή εκτέλεση της διαδικασίας.

Συνολικά έχουν κατασκευαστεί 6 επιμέρους pipelines, τα οποία συνδυάζονται με τέτοιο τρόπο ώστε να χτιστούν όλα τα δεδομένα για κάθε χώρα και μπορούν να χωριστούν σε 2 κατηγορίες:

- Pipelines που εκτελούν processes, δηλαδή τα code scripts που έχουμε γράψει. Σε αυτή την κατηγορία ανήκουν τα παρακάτω pipelines

  Ca_pipeline

  L2_pipeline

  L3_pipeline

  Gl_pipeline

- Pipelines που καλούν τα pipelines της κατηγορίας 1 με συγκεκριμένο τρόπο για την ομαδοποίηση της εκτέλεσης. Τα ακόλουθα pipelines ανήκουν σε αυτή την κατηγορία

  Cs_pipeline

  Full_execution_pipeline

Παρακάτω δίνεται η Περιγραφή τους.

5.1 Ca_pipeline

Execution parameters:

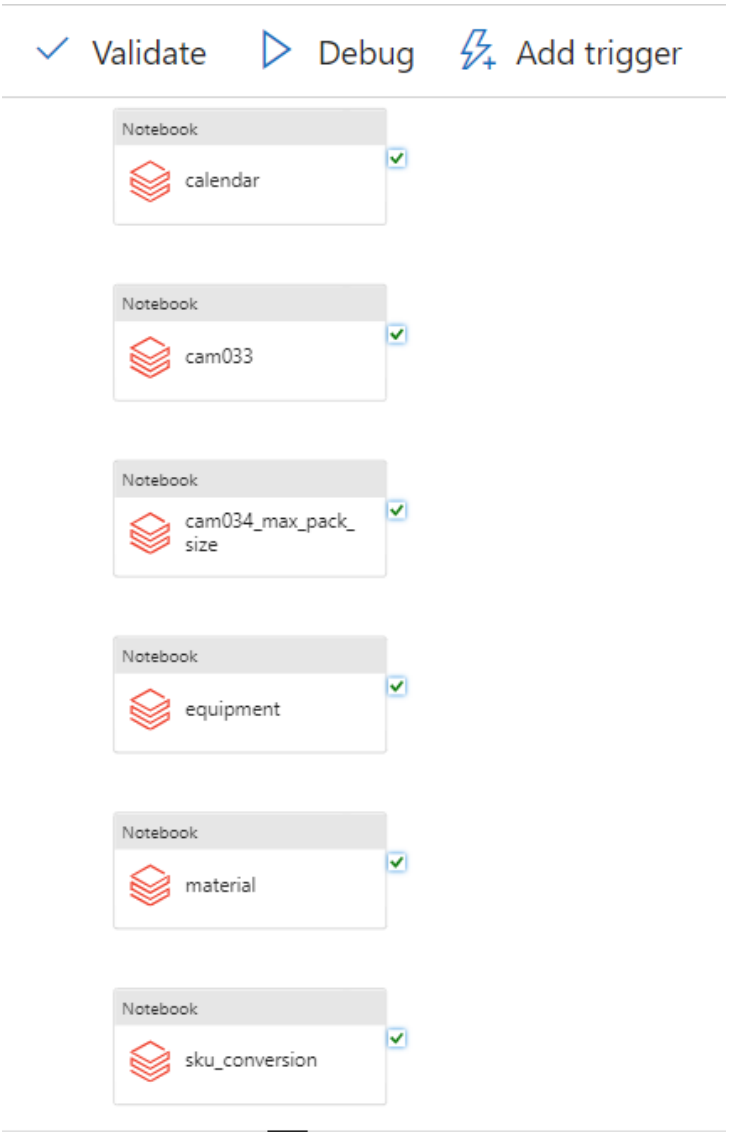
None

Processes executed:

calendar, cam033, cam034_max_pack_size, equipment, material, sku_conversion

Περιγραφή:

Είναι υπεύθυνο για την εκτέλεση των country-agnostic processes του L2, δηλαδή των processes που είναι κοινά για όλες τις χώρες. Εκτελείται μόνο μια φορά και όλα τα processes τρέχουν παράλληλα, δεδομένου ότι δεν υπάρχει κάποια εξάρτηση μεταξύ τους.



Εικόνα 8: Ca pipeline

5.2 L2_pipeline

Execution parameters:

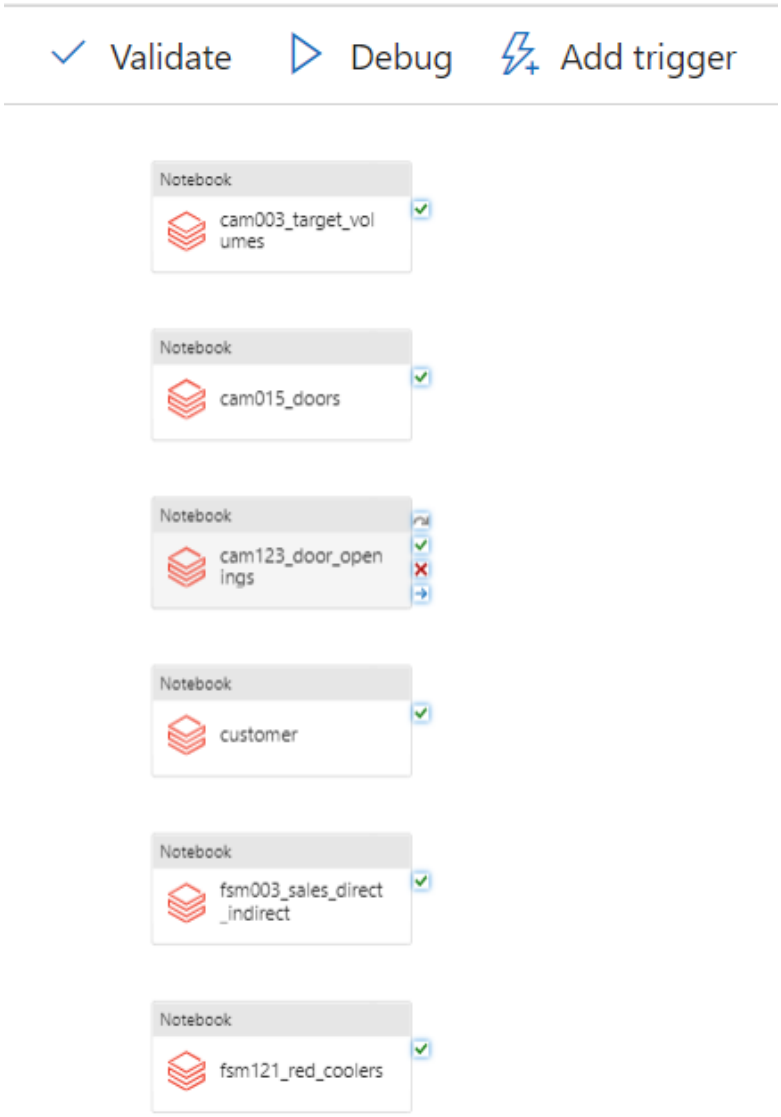
country_code, fiscal_period

Processes executed:

customer, fsm003_sales_direct_indirect, cam003_target_volumes, cam015_doors, cam123_door_openings, fsm121_red_coolers

Περιγραφή:

Είναι υπεύθυνο για την εκτέλεση των country-specific processes του L2, δηλαδή των processes του L2 που περιέχουν πληροφορία για κάθε χώρα χωριστά. Εκτελείται ξεχωριστά για κάθε χώρα και όλα τα processes τρέχουν παράλληλα δεδομένου ότι δεν υπάρχει κάποια εξάρτηση μεταξύ τους.



Εικόνα 9: L2 pipeline

5.3 L3_pipeline

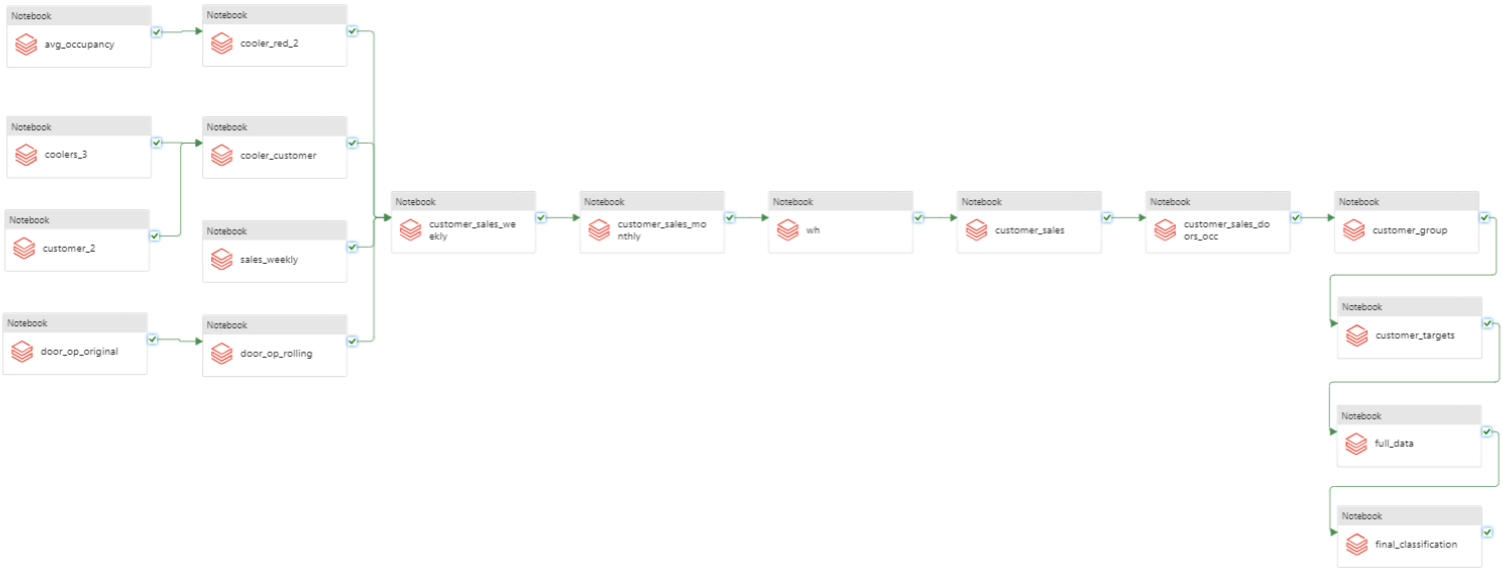
Execution parameters:

country_code, fiscal_period

Processes executed:

Περιγραφή:

Είναι υπεύθυνο για την εκτέλεση των country-specific processes του L3, δηλαδή των processes του L3 που περιέχουν πληροφορία για κάθε χώρα χωριστά. Εκτελείται ξεχωριστά για κάθε χώρα και τα processes εκτελούνται σε συγκεκριμένη σειρά, με βάση τις εξαρτήσεις μεταξύ τους. Η εξάρτηση μεταξύ των processes μπορεί να φανεί στο παρακάτω διάγραμμα:



Εικόνα 10: L3 pipeline

5.4 GL_pipeline

Execution parameters:

country_code, fiscal_period

Processes executed:

cooler_output

Περιγραφή:

Είναι υπεύθυνο για την εκτέλεση του process του golden layer, δηλαδή του process που παράγει την τελική έξοδο του συστήματος. Εκτελείται ξεχωριστά για κάθε χώρα.

5.5 CS_ pipeline

Execution parameters:

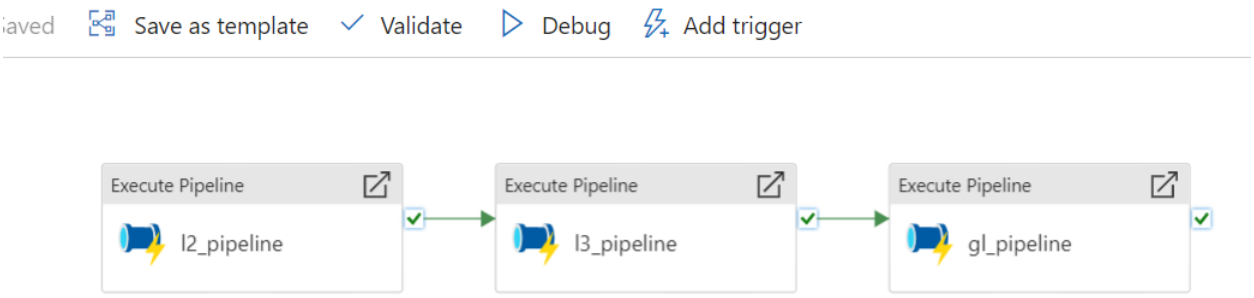
country_code, fiscal_period

Pipelines called:

L2_pipeline, L3_pipeline, GL_pipeline

Περιγραφή:

Καλεί τα country-specific pipelines που Περιγράφησαν σαν προηγουμένως με την κατάλληλη σειρά (L2, L3, GL).Εκτελείται ξεχωριστά για κάθε χώρα.



Εικόνα 11: Cs pipeline

5.6 Full_execution_pipeline

Execution parameters:

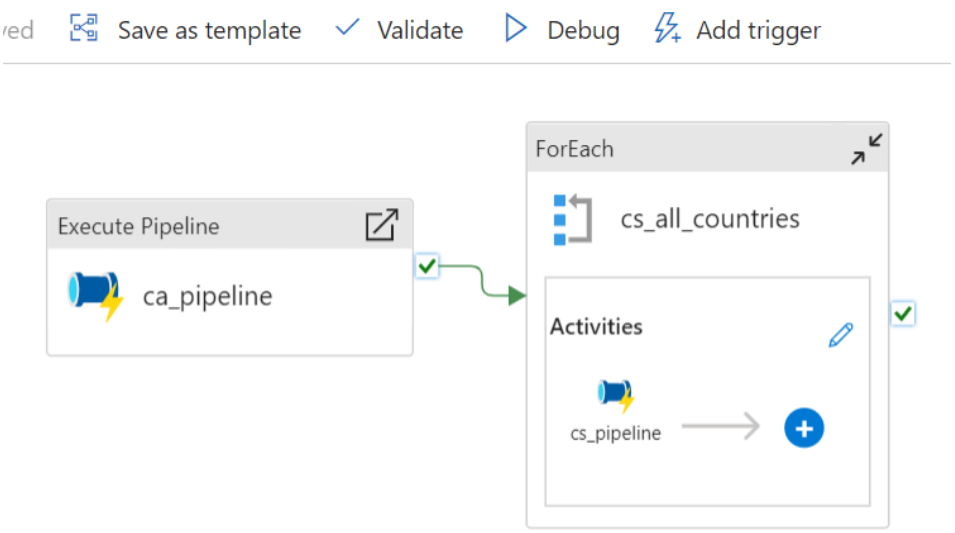
country_code, fiscal_period

Pipelines called:

CA_pipeline (once), CS_pipeline (once for each country)

Περιγραφή:

Αποτελεί το master pipeline του project μας. Το συγκεκριμένο pipeline είναι ο γενικός ενορχηστρωτής και το μοναδικό pipeline που θα πρέπει να καλέσει ο χρήστης. Καλεί μια φορά το CA_PIPELINE προκειμένου να εκτελεστούν τα country-agnostic processes και στη συνέχεια καλεί για κάθε χώρα το CS_PIPELINE προκειμένου να εκτελεστούν όλα τα country-specific processes για κάθε χώρα και να παραχθεί το τελικό αποτέλεσμα.



Εικόνα 12: Full_execution pipeline

Για την αυτοματοποιημένη παραγωγή αποτελεσμάτων χωρίς την συνεχή επέμβαση του χρήστη, έχει δημιουργηθεί ένας trigger ο οποίος ξεκινάει το FULL_EXECUTION_PIPELINE και μπορεί να προγραμματιστεί να ενεργοποιείται σε συγκεκριμένα χρονικά διαστήματα τα οποία επιλέγει ο χρήστης.

Κεφάλαιο 6: Αποτελέσματα

Στον παρακάτω πίνακα φαίνεται η κατανομή των ψυγείων ανά group available πληροφορίας

BG					
December		January		February	
customer_group	cooler_count	customer_group	cooler_count	customer_group	cooler_count
1	9987	1	9989	1	9976
2	3451	2	3385	2	3393
3	18382	3	18244	3	18055
4	24652	4	24602	4	25035
5	352	5	343	5	359
6	344	6	321	6	308
7	1469	7	1420	7	1362
8	54	8	59	8	57
not assigned	6965	not assigned	6650	not assigned	6674

Πίνακας 3: Κατανομή ψυγείων ανά group πληροφορίας

Παρατηρούμε ότι για τους μήνες Δεκέμβριο – Ιανουάριο - Φεβρουάριο ότι η πληροφορία είναι σταθερή και η κατανομή είναι στο ίδιο επίπεδο κατά το πέρασμα των 3 αυτών μηνών.

Βλέπουμε επίσης ότι το 86% των ψυγείων έχουν πληροφορία για τις πωλήσεις, το 30% έχει πληροφορία για το άνοιγμα των πορτών, το 70% έχει πληροφορία για την πληρότητα και το 11% δεν έχει καθόλου πληροφορία για οποιοδήποτε source.

Στον παρακάτω πίνακα φαίνεται η ακριβής μεταβολή των groups από τον Δεκέμβριο στον Ιανουάριο.

BG					
1		2		3	
Jan	cooler_count	Jan	cooler_count	Jan	cooler_count
1	9429	1	423	1	51
2	381	2	2875	2	4
3	64	3	3	3	17065
4	32	4	81	4	955
5	10	5	14	5	0
6	0	6	2	6	0
7	2	7	0	7	73
8	0	8	0	8	0
not assigned	10	not assigned	5	not assigned	27
4		5		6	
Jan	cooler_count	Jan	cooler_count	Jan	cooler_count
1	43	1	26	1	5
2	76	2	6	2	19
3	977	3	3	3	0
4	23028	4	2	4	5
5	1	5	292	5	13
6	1	6	16	6	282
7	42	7	5	7	0
8	0	8	0	8	1
not assigned	134	not assigned	1	not assigned	14
7		8		not_assigned	
Jan	cooler_count	Jan	cooler_count	Jan	cooler_count
1	0	1	0	1	1
2	1	2	0	2	53
3	82	3	0	3	9
4	34	4	0	4	23
5	9	5	0	5	240
6	0	6	0	6	4
7	1235	7	1	7	17
8	5	8	53	8	55
not assigned	74	not assigned	0	not assigned	6266

Πίνακας 4: Μεταβολή πληροφορίας ανά group Δεκ-Ιαν

Παρατηρούμε και σε αυτόν τον πίνακα ότι η συντριπτική πλειοψηφία των ψυγείων του κάθε γκρουπ τον Δεκέμβριο παραμένει στο ίδιο γκρουπ και τον Ιανουάριο, συνεπώς έχουμε αμελητέες μεταβολές στην πληροφορία την οποία λαμβάνουμε για τα ψυγεία.

Στους παρακάτω 2 πίνακες φαίνεται η ιδιά ανάλυση της κατανομής ανά κατηγορία απόδοσης του κάθε ψυγείου.

BG					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	22445	A	22375	A	23147
B	582	B	539	B	538
C	9014	C	9010	C	8730
D	18490	D	18370	D	18070
E	8895	E	8645	E	8516
No Data Visibility	6230	No Data Visibility	6074	No Data Visibility	6218

Πίνακας 5: Κατανομή ψυγείων ανά κατηγορία απόδοσης

BG											
A		B		C		D		E		NDV	
Jan	cooler_count	Jan	cooler_count	Jan	cooler_count	Jan	cooler_count	Jan	cooler_count	Jan	cooler_count
A	21035	A	10	A	472	A	279	A	448	A	33
B	1	B	473	B	27	B	3	B	15	B	16
C	472	C	26	C	7794	C	457	C	201	C	33
D	313	D	11	D	450	D	17055	D	306	D	182
E	436	E	39	E	161	E	305	E	7570	E	24
No Data Visibility	14	No Data Visibility	16	No Data Visibility	26	No Data Visibility	106	No Data Visibility	123	No Data Visibility	5674

Πίνακας 6: Μεταβολή απόδοσης ανά κατηγορία Δεκ-Ιαν

Παρατηρούμε ότι για τους μήνες Δεκέμβριο - Ιανουάριο - Φεβρουάριο ότι η πληροφορία είναι σταθερή και η κατανομή είναι στο ίδιο επίπεδο κατά το πέρασμα των 3 αυτών μηνών.

Επίσης βλέπουμε ότι και σε αυτόν τον πίνακα ότι, όπως και στα γκρουπ πληροφορίας, η συντριπτική πλειοψηφία των ψυγείων της κάθε κατηγορίας απόδοσης τον Δεκέμβριο παραμένει στο ίδιο γκρουπ και τον Ιανουάριο, συνεπώς το κάθε ψυγείο διατηρεί σταθερά νούμερα απόδοσης κατά το πέρασμα των μηνών.

Τέλος, δίνεται η κατανομή των κατηγοριών απόδοσης ανά γκρουπ πληροφορίας για τους 3 μήνες υπό μελέτη (Δεκέμβριο, Ιανουάριο, Φεβρουάριο).

Cluster Allocation Group 1 (sales, doors, occupancy available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	6737	A	6727	A	6752
B	0	B	0	B	0
C	1094	C	1111	C	1083
D	1351	D	1329	D	1329
E	805	E	822	E	812
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group 2 (sales, doors available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	1856	A	1846	A	1847
B	0	B	0	B	0
C	671	C	647	C	660
D	924	D	892	D	886
E	0	E	0	E	0
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group 3 (sales, occupancy available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	7869	A	7820	A	8203
B	0	B	0	B	0
C	2094	C	2059	C	1705
D	2313	D	2365	D	2174
E	6074	E	5969	E	5955
No Data Visibility	32	No Data Visibility	31	No Data Visibility	18

Cluster Allocation Group 4 (sales available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	5588	A	5592	A	5954
B	0	B	0	B	0
C	4894	C	4941	C	5039
D	13687	D	13588	D	13487
E	234	E	206	E	242
No Data Visibility	249	No Data Visibility	275	No Data Visibility	313

Cluster Allocation Group 5 (doors, occupancy available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	195	A	188	A	188
B	0	B	0	B	0
C	35	C	44	C	45
D	37	D	31	D	41
E	85	E	80	E	85
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group 6 (doors available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	146	A	143	A	146
B	0	B	0	B	0
C	68	C	61	C	57
D	130	D	117	D	105
E	0	E	0	E	0
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group 7 (occupancy available)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	0	A	0	A	0
B	582	B	539	B	538
C	158	C	147	C	141
D	48	D	48	D	48
E	681	E	686	E	635
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group 8 (no info available, Wholesalers)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	54	A	59	A	57
B	0	B	0	B	0
C	0	C	0	C	0
D	0	D	0	D	0
E	0	E	0	E	0
No Data Visibility	0	No Data Visibility	0	No Data Visibility	0

Cluster Allocation Group not_assigned (no info, not Wholesalers)					
December		January		February	
profitability_cluster	cooler_count	profitability_cluster	cooler_count	profitability_cluster	cooler_count
A	0	A	0	A	0
B	0	B	0	B	0
C	0	C	0	C	0
D	0	D	0	D	0
E	1016	E	882	E	787
No Data Visibility	5949	No Data Visibility	5768	No Data Visibility	5887

Πίνακας 7: Κατανομή κατηγορίας απόδοσης ψυγείων ανά group πληροφορίας

Και στους συγκεκριμένους πίνακες βλέπουμε ότι για τους μήνες Δεκέμβριο-Ιανουάριο-Φεβρουάριο το κάθε γκρουπ πληροφορίας δεν παρουσιάζει μεταβολές στην κατανομή των ψυγείων στην κάθε κατηγορία απόδοσης.

Συνοψίζοντας, παρατηρούμε ότι το σύστημα μας διατηρεί μια σταθερή συμπεριφορά στο πέρασμα τού χρόνου. Το γεγονός ότι δεν υπάρχει σημαντική μεταβολή στην κατανομή των γκρουπ πληροφορίας μας δείχνει ό,τι δεν υπάρχει κάποια σημαντική ανανέωση των δεδομένων εισόδου όσον αφορά την διαθέσιμη πληροφορία, συνεπώς είναι σπάνιο να έρθει καινούρια πληροφορία για κάποιο ψυγείο. Επιπροσθέτως, η σταθερή κατανομή κατηγοριών απόδοσης τους 3 μήνες μας υποδεικνύει ότι δεν υπάρχουν δραματικές αυξομειώσεις στην απόδοση των επιμέρους ψυγείων.

Κεφάλαιο 7: Συμπεράσματα

7.1 Ανακεφαλαίωση

Όπως περιγράφηκε αναλυτικά στα προηγούμενα κεφάλαια, σκοπός της εργασίας είναι η υλοποίηση ενός συστήματος για την κατηγοριοποίηση των ψυγείων μιας επιχείρησης με βάση την απόδοση τους, ανάλογα την διαθέσιμη πληροφορία που έχουμε για πωλήσεις, πόρτες και πληρότητα. Για την επίτευξη αυτού γίνεται το απαραίτητο καθάρισμα των δεδομένων εισόδου και στη συνέχεια εκτελείται η κατάλληλη επεξεργασία των δεδομένων με σκοπό να έρθουν στο κατάλληλη μορφή για να μπορούν να εφαρμοστούν οι απαραίτητοι υπολογισμοί. Αφού έρθουν τα δεδομένα στο σωστό επίπεδο, υπολογίζεται το γκρουπ πληροφορίας στο οποίο ανήκει το κάθε ψυγείο, ανάλογα τη διαθέσιμη πληροφορία του. Υπάρχουν συνολικά 9 γκρουπ πληροφορίας και το καθένα από αυτά χρησιμοποιεί τον δικό του μαθηματικό τύπο για τον υπολογισμό της τελικής κατηγορίας απόδοσης, τον οποίο εφαρμόζουμε αφότου υπολογίσουμε για κάθε ψυγείο την μέση τιμή των εβδομαδιαίων πωλήσεων για τους τελευταίους 12 μήνες, την μέση τιμή της κινητικότητας των πορτών για τους τελευταίους 12 μήνες, την τελευταία τιμή πληρότητας καθώς και τους στόχους που έχουν τεθεί σε επίπεδο πωλήσεων και κινητικότητας πορτών.

Η διαδικασία έχει χωριστεί σε 3 στρώματα πινάκων, για ευκολότερη κατανόηση και μελλοντική συντήρηση του κώδικα. Στο πρώτο στρώμα (L2) φορτώνονται τα απαραίτητα δεδομένα από τις πηγές και εφαρμόζεται ένα καθάρισμα, στο δεύτερο στρώμα (L3) γίνεται όλη η προεπεξεργασία των δεδομένων, οι κατάλληλοι μετασχηματισμοί και τελικά η εφαρμογή των απαραίτητων τύπων για την τελική κατηγοριοποίηση της απόδοσης, και στο τρίτο και τελικό στρώμα (Golden) ενσωματώνεται επιπλέον πληροφορία για τα ψυγεία και τους πελάτες (τα οποία αν και δεν αξιοποιούνται για τον υπολογισμό των κατηγοριών, είναι χρήσιμα για την κατανόηση των δεδομένων) και τα δεδομένα εξόδου γίνονται ορατά στους χρήστες.

7.2 Περαιτέρω Βελτιώσεις – Επεκτάσεις

Αυτή τη στιγμή ο αλγόριθμος χρησιμοποιεί πληροφορία για τις πωλήσεις, την πληρότητα των ψυγείων και τις πόρτες του κάθε ψυγείου προκειμένου να υπολογιστεί η κατηγορία απόδοσης στην οποία ανήκει. Για να κάνουμε πιο ακριβή τον υπολογισμό της απόδοσης μπορούμε να εισάγουμε επιπλέον πληροφορία από άλλα sources. Πιθανές επιλογές θα μπορούσαν να είναι οι ακόλουθες:

- Τιμές Προϊόντων: Χρησιμοποιώντας πληροφορία για τις τιμές μπορούμε να αντιληφθούμε αν μια πιθανή μείωση των πωλήσεων των ψυγείων οφείλεται σε αύξηση

των τιμών σε καθολικό επίπεδο, οπότε και η οποιαδήποτε μείωση της απόδοσης θα πρέπει να ερμηνευθεί λαμβάνοντας υπόψιν και αυτόν τον παράγοντα.

- Ειδικά γεγονότα/ Γιορτές: Χρησιμοποιώντας πληροφορίες για τις διαφορετικές γιορτές της κάθε χώρας, καθώς και για τυχόν ειδικά γεγονότα (μαραθώνιος, προωθητικές ενέργειες κλπ.), μπορούμε να καταλάβουμε αν η υψηλή απόδοση οφείλεται σε αύξηση των πωλήσεων λόγω εορτών. Είναι δηλαδή ένα συνολικό φαινόμενο και όχι μεμονωμένη βελτίωση της απόδοσης. Συνεπώς, θα πρέπει να προσαρμόσουμε τους υπολογισμούς μας αναλόγως.

- Καιρός: Με δεδομένο ότι σε μήνες με υψηλότερες θερμοκρασίες συνήθως υπάρχει μεγαλύτερη κατανάλωση αναψυκτικών, χρησιμοποιώντας πληροφορίες για τον καιρό μπορούμε να καταλάβουμε αν η υψηλή απόδοση οφείλεται σε τυχόν υψηλές θερμοκρασίες και να προσαρμόσουμε τους υπολογισμούς μας ανάλογα με την εποχικότητα.

Αυτή τη στιγμή, προκειμένου να αποφανθούμε για την κατηγορία απόδοσης του κάθε ψυγείου χρησιμοποιούμε συγκεκριμένους κανόνες (φόρμουλες). Με χρήση αυτών είμαστε σε θέση να κατηγοριοποιήσουμε την απόδοση των ψυγείων για τον προηγούμενο μήνα. Αυτό όμως σημαίνει ότι μπορούμε να κάνουμε υπολογισμούς μόνο για παρελθοντικούς μήνες που έχουμε ήδη πλήρη δεδομένα. Μπορούμε να χρησιμοποιήσουμε τα δεδομένα που έχουμε για τους παρελθοντικούς μήνες προκειμένου να εκπαιδεύσουμε μοντέλα πρόβλεψης για τους μελλοντικούς μήνες. Με αυτόν τον τρόπο θα έχουμε μια εικόνα για την πορεία απόδοσης του κάθε ψυγείου προκειμένου να προβούμε σε ενέργειες για να βελτιώσουμε από πριν τυχόν μειωμένες αποδόσεις. Μια από τις συνηθισμένες προσεγγίσεις για το συγκεκριμένο σκοπό είναι να χρησιμοποιηθούν μοντέλα βαθιάς μηχανικής μάθησης όπως τα δίκτυα Μακράς Βραχύχρονης Μνήμης (Long Short-term Memory - LSTM) και Gated Recurrent Unit (GRU) τα οποία είναι ισχυρά μοντέλα για την πρόβλεψη χρονοσειρών καθώς μπορούν να ανιχνεύουν μακροπρόθεσμες εξαρτήσεις και πολύπλοκα πρότυπα σε χρονοσειριακά δεδομένα.

Τα δίκτυα LSTM χρησιμοποιούν κύτταρα μνήμης για να αποθηκεύουν και να ανακτούν πληροφορίες σε μακρινές ακολουθίες. Αυτά τα κύτταρα μνήμης ενισχύουν τη δυνατότητα του μοντέλου να θυμάται και να χρησιμοποιεί σχετικά πρότυπα από προηγούμενες παρατηρήσεις. Οι μηχανισμοί πύλης στα LSTM, περιλαμβανομένων των πυλών εισόδου, λήθης και εξόδου, ελέγχουν τη ροή των πληροφοριών μέσα στο δίκτυο. Αυτοί οι μηχανισμοί επιτρέπουν την εκλεκτική επεξεργασία και αποθήκευση πληροφοριών, επιτρέποντας στο μοντέλο να αντιμετωπίσει πολύπλοκα και μεταβαλλόμενα πρότυπα σε χρονοσειριακά δεδομένα.
Από την άλλη πλευρά, τα δίκτυα GRU συνδυάζουν τις λειτουργίες ενημέρωσης και επαναφοράς της μνήμης σε μια "πύλη ενημέρωσης", μειώνοντας έτσι τον αριθμό των παραμέτρων και των υπολογισμών. Αυτή η απλότητα οδηγεί σε πιο αποδοτική εκπαίδευση και ταχύτερη σύγκλιση

κατά τη διάρκεια της διαδικασίας εκπαίδευσης. Τα δίκτυα GRU μπορούν να παρουσιάσουν καλή απόδοση ακόμα και με μικρότερα σύνολα εκπαιδευτικών δεδομένων και είναι αποτελεσματικά στην ανίχνευση σύντομων και μεσοπρόθεσμων εξαρτήσεων σε χρονοσειριακά δεδομένα.

Τόσο τα δίκτυα LSTM όσο και τα GRU μπορούν να χειριστούν χρονοσειριακά δεδομένα με μεταβλητό μήκος ακολουθίας, πράγμα που τα καθιστά ευέλικτα για τη μοντελοποίηση ανομοιόμορφων ή αραιών χρονοσειριακών δεδομένων. Έχουν ευρέως χρησιμοποιηθεί σε διάφορες εφαρμογές, όπως η πρόβλεψη των χρηματιστηρίων, η πρόβλεψη του καιρού και η πρόβλεψη της ζήτησης. Τα δίκτυα LSTM ξεχωρίζουν στην ανίχνευση μακροπρόθεσμων εξαρτήσεων και πολύπλοκων προτύπων, ενώ τα δίκτυα GRU προσφέρουν μια απλούστερη αρχιτεκτονική, αποδοτική εκπαίδευση και καλή απόδοση με μικρότερα σύνολα δεδομένων. Η επιλογή μεταξύ των δύο εξαρτάται από τα συγκεκριμένα χαρακτηριστικά των δεδομένων, τους διαθέσιμους υπολογιστικούς πόρους και τον συμβιβασμό μεταξύ της πολυπλοκότητας του μοντέλου και της απόδοσης, συνεπώς χρειάζεται ένας αριθμός πειραμάτων προκειμένου να αποφασίσουμε ποιο μοντέλο εξυπηρετεί καλύτερα τους σκοπούς μας.

Αρχιτεκτονικές βαθιών νευρωνικών δικτύων έχουν υλοποιηθεί και χρησιμοποιηθεί σε διάφορες εφαρμογές από μέλη του Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης του ΕΜΠ. Ειδικότερα επιβλεπόμενες τεχνικές συνελικτικών νευρωνικών δικτύων (Convolutional Neural Network - CNN) και αναδρομικών νευρωνικών δικτύων (CNN-RNN) με μεθοδολογίες LSTM και GRU έχουν εφαρμοστεί στην ιατρική διάγνωση νευροεκφυλιστικών ασθενειών, όπως της νόσου του Πάρκινσον ή της Covid-19. Εμφαση δίδεται στην διαφάνεια και στην προσαρμογή των μοντέλων. Βαθιές ημι- και αυτο-επιβλεπόμενες 3-Δ νευρωνικές αρχιτεκτονικές, αλλά και αρχιτεκτονικές κωδικοποιητή-αποκωδικοποιητή έχουν εφαρμοστεί στην ανίχνευση βλαβών σε πυρηνικούς αντιδραστήρες, στην πρόβλεψη της παραγωγής στον αγροτικό τομέα και στην αναγνώριση και σύνθεση συναισθήματος.

Κεφάλαιο 8: Βιβλιογραφία

1. Amazon S3: https://aws.amazon.com/s3/
2. Google Cloud Storage: https://cloud.google.com/storage
3. IBM Cloud Object Storage: https://www.ibm.com/cloud/object-storage
4. Wasabi: https://wasabi.com/
5. Backblaze B2: https://www.backblaze.com/b2/
6. Azure Repos: https://azure.microsoft.com/en-us/services/devops/repos/
7. GitHub: https://github.com/
8. GitLab: https://about.gitlab.com/
9. Bitbucket: https://bitbucket.org/
10. CodeCommit: https://aws.amazon.com/codecommit/
11. Azure Data Factory: https://azure.microsoft.com/en-us/services/data-factory/
12. AWS Glue: https://aws.amazon.com/glue/
13. Google Cloud Dataflow: https://cloud.google.com/dataflow
14. Talend: https://www.talend.com/
15. Apache NiFi: https://nifi.apache.org/
16. PySpark: https://spark.apache.org/docs/latest/api/python/index.html
17. Apache Spark: https://spark.apache.org/
18. Hadoop: http://hadoop.apache.org/
19. Azure HDInsight: https://azure.microsoft.com/en-us/services/hdinsight/
20. AWS EMR: https://aws.amazon.com/emr/
21. Cloudera: https://www.cloudera.com/
22. Google Cloud Dataproc: https://cloud.google.com/dataproc
23. Azure Storage: https://docs.microsoft.com/en-us/azure/storage/
24. Microsoft Azure: https://azure.microsoft.com/
25. AWS: https://aws.amazon.com/
26. Google Cloud Platform: https://cloud.google.com/
27. "What Is Big Data?" by Bernard Marr: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
28. "Big Data Analytics: What It Is And Why It Matters" by Forbes Technology Council: https://www.forbes.com/sites/forbestechcouncil/2018/09/25/big-data-analytics-what-it-is-and-why-it-matters/?sh=1652abf960ac
29. "Big Data Analytics: A Literature Review Paper" by Mohammad Reza Fahimi et al.: https://www.researchgate.net/publication/316534045_Big_Data_Analytics_A_Literature_Review_Paper
30. "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?" by Shadi Aljawarneh et al.: https://www.mdpi.com/2076-3417/8/5/722/htm
31. "Big Data and the Cloud: A Perfect Match" by TechRadar: https://www.techradar.com/news/big-data-and-the-cloud-a-perfect-match
32. "The Pros and Cons of Cloud Storage" by Harvard Business Review: https://hbr.org/2019/05/the-pros-and-cons-of-cloud-storage

33. "Why Azure is the Best Cloud for Windows" by Cloud Academy: https://cloudacademy.com/blog/why-azure-is-the-best-cloud-for-windows/
34. "A Beginner's Guide to Git and GitHub" by Lawrence McDaniel: https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners
35. "How to Choose the Right Big Data Storage Technology" by Information Age: https://www.information-age.com/choose-right-big-data-storage-technology-123463880/
36. "Big Data Analytics: Applications and Benefits" by Simplilearn: https://www.simplilearn.com/big-data-analytics-applications-and-benefits-article
37. "Big Data Analytics: A Comprehensive Guide" by Analytics Insight: https://www.analyticsinsight.net/big-data-analytics-a-comprehensive-guide/
38. "Azure DevOps Tutorial: Complete Beginner's Guide" by Guru99: https://www.guru99.com/azure-devops-tutorial.html
39. "A Beginner's Guide to Cloud Computing" by Forbes: https://www.forbes.com/sites/forbestechcouncil/2021/02/22/a-beginners-guide-to-cloud-computing/?sh=3e75ef3a3a1c
40. "How to Choose the Best Cloud Storage Service for You" by PCMag: https://www.pcmag.com/picks/the-best-cloud-storage-services
41. "How Big Data is Changing Healthcare" by HealthIT Analytics: https://healthitanalytics.com/features/how-big-data-is-changing-healthcare
42. "The Future of Education: Big Data and Learning Analytics" by EdTech Magazine: https://www.edtechmagazine.com/higher/article/2019/03/future-education-big-data-and-learning-analytics
43. "Big Data in Tourism: Exploring the Opportunities and Challenges" by Journal of Destination Marketing & Management: https://www.sciencedirect.com/science/article/pii/S2212571X20300517
44. "Big Data Analytics in Transportation: Benefits, Challenges, and Future Directions" by IEEE Intelligent Transportation Systems Magazine: https://ieeexplore.ieee.org/document/7376177
45. "A Technical Introduction to Azure Data Factory" by Edureka: https://www.edureka.co/blog/azure-data-factory-tutorial/
46. "What is AWS Glue and How Does it Work?" by TechTarget: https://searchaws.techtarget.com/definition/AWS-Glue
47. "The Benefits and Drawbacks of Apache NiFi" by DZone: https://dzone.com/articles/the-benefits-and-drawbacks-of-apache-nifi
48. "How to Choose the Right Big Data Analytics Tools" by CIO: https://www.cio.com/article/3406524/how-to-choose-the-right-big-data-analytics-tools.html
49. "Getting Started with PySpark" by Towards Data Science: https://towardsdatascience.com/getting-started-with-pyspark-9b7e02bfa3f3
50. "A Comprehensive Guide to Azure HDInsight" by DataFlair: https://data-flair.training/blogs/azure-hdinsight-tutorial/

51. “Deep neural architectures for prediction in healthcare” by Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias & Georgios Tagaris: https://doi.org/10.1007/s40747-017-0064-6
52. “Unified deep learning approach for prediction of Parkinson's disease” by James Wingate, Ilianna Kollia, Luc Bidaut, Stefanos Kollias: https://doi.org/10.1049/iet-ipr.2019.1526
53. “Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis” by Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, Stefanos Kollias: https://doi.org/10.48550/arXiv.2106.07524
54. “A deep neural architecture for harmonizing 3-D input data analysis and decision making in medical imaging” by D. Kollias, A. Arsenos, S. Kollias: https://doi.org/10.48550/arXiv.2303.00175
55. “Adaptation and contextualization of deep neural network models” by Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis, Stefanos Kollias: https://doi.org/10.1109/SSCI.2017.8280975
56. “Transparent Adaptation in Deep Medical Image Diagnosis” by D. Kollias, Y. Vlaxos, M. Seferis, I. Kollia, L. Sukissian, J. Wingate & S. Kollias: http://dx.doi.org/10.1007/978-3-030-73959-1_22
57. “An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth” by Bashar Alhnaity, Stefanos Kollias, Georgios Leontidis, Shouyong Jiang, Bert Schamp, Simon Pearson: https://doi.org/10.48550/arXiv.2012.04041
58. “Deep learning techniques for in-core perturbation identification and localization of time-series nuclear plant measurements” by Antonios Papaoikonomou, James Wingate, Vasudha Verma, Aiden Durrant, George Ioannou, Tasos Papagiannis, Miao Yu, Georgios Alexandridis, Abdelhamid Dokhane, Georgios Leontidis, Stefanos Kollias, Andreas Stafylopatis: https://doi.org/10.1016/j.anucene.2022.109373
59. “Using deep learning to predict plant growth and yield in greenhouse environments” by Bashar Alhnaity, Simon Pearson, Georgios Leontidis and Stefanos Kollias: https://doi.org/10.17660/ActaHortic.2020.1296.55
60. “Abaw: Learning from synthetic data & multi-task learning challenges” by Kollias Dimitrios: https://doi.org/10.48550/arXiv.2207.01138