Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής
ΠΜΣ ΣΤΗΝ ΑΝΑΠΤΥΞΗ & ΑΣΦΑΛΕΙΑ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
Τεχνολογίες Ψηφιακών Υποδομών
Εαρινό Εξάμηνο 2023-2024
Διδάσκουσα: Β. Καλογεράκη
Παραδοτέο Α

Μέλη ομάδας

Βασιλική Κατσαντώνη (f3312308)
Άννα Μωραΐτου (f3312312)
Ευάγγελος Γκίνης (f3312303)

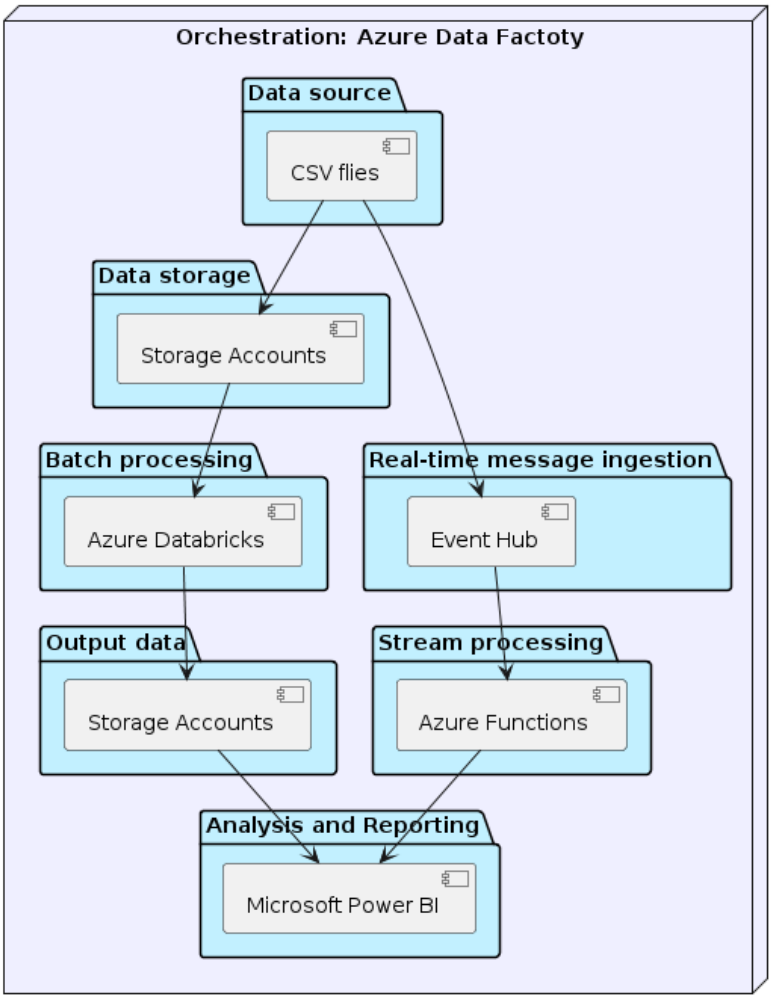
Εργαλεία

Για την υλοποίηση του έργου θα χρησιμοποιηθούν τα ακόλουθα εργαλεία:

- **Storage Accounts:** Το Azure Blob Storage θα χρησιμοποιηθεί για την αποθήκευση των αρχείων CSV που περιέχουν τα δεδομένα των διαδρομών. Επιπλέον ο μηχανισμός αποθήκευσης του είναι ιδανικός για τη κλιμακωσιμότητα και ανθεκτικότητα μεγάλων, σε ποσότητα, μη δομημένων δεδομένων. Επιπλέον, υποστηρίζει triggers για την έναρξη της επεξεργασίας όταν μεταφορτώνονται νέα δεδομένα.
- **Azure Databricks:** Είναι μια πλατφόρμα ανάλυσης δεδομένων που βασίζεται στο Apache Spark. Παρέχει δυνατότητες autoscaling, auto-termination, fast cluster start time. Υποστηρίζει σύνδεση με το Azure Blob Storage καθώς και real time processing. Τέλος υλοποιεί μηχανισμούς με του οποίους επιτυγχάνει κλιμακωσιμότητα και ανοχή σε σφάλματα.
- **Microsoft Power BI:** Παρέχει μια ισχυρή, ευέλικτη και επεκτάσιμη πλατφόρμα για την οπτικοποίηση δεδομένων, την ανάλυση και τη συνεργασία, επιτρέποντας να αντληθούν αξιοποιήσιμες πληροφορίες από τα δεδομένα τους.
- **Azure Data Factory:** Μας επιτρέπει να δημιουργούμε, να προγραμματίζουμε και να διαχειριζόμαστε δεδομένα για την εισαγωγή, την προετοιμασία, το μετασχηματισμό και την ορχήστρωση (συνδυασμό) δεδομένων από διάφορες πηγές και προορισμούς.
- **Azure Functions:** Το Azure Functions θα χρησιμοποιηθεί για την υλοποίηση serverless υπολογισμών για το ερώτημα που απαιτεί επεξεργασία ροών δεδομένων. Θα παραμετροποιηθεί το blob storage ώστε με ένα trigger να επιτρέπεται η αυτόματη επεξεργασία δεδομένων αμέσως μετά τη μεταφόρτωσή τους. Αυτό παρέχει δυναμική κλιμάκωση των υπολογιστικών πόρων βάσει του φόρτου εργασίας και ανοχή σε σφάλματα (fault tolerance) με τον αυτόματο χειρισμό της διαχείρισης της υποδομής.
- **Azure Event Hubs:** Το Azure Event Hubs είναι μια υπηρεσία που παρέχει τη δυνατότητα λήψης, αποθήκευσης και επεξεργασίας μεγάλου όγκου συμβάντων από πολλές πηγές. Αποτελεί ένα κεντρικό σημείο για τη συγκέντρωση και τη δρομολόγηση

δεδομένων από πολλά συστήματα, επιτρέποντας την ανάλυση σε πραγματικό χρόνο, τη ροή εργασιών εργασίας και άλλες εφαρμογές όπως η αναγνώριση προτύπων και η πρόβλεψη. Είναι ιδανικό για περιπτώσεις όπου χρειάζεται υψηλή κλιμακωσιμότητα και αξιοπιστία στην επεξεργασία συμβάντων.

Αρχιτεκτονική του Συστήματος



Το διάγραμμα δείχνει τη ροή των δεδομένων μέσω των διαφορετικών στοιχείων του συστήματος:

1. Η πηγή των δεδομένων μας είναι αρχεία CSV.
2. Τα δεδομένα από τα CSV αν είναι:
    a. **Batch Data**: αποθηκεύονται στο Azure Storage Accounts σε μορφή CSV και η επεξεργασία τους θα γίνει με το Azure Databricks.
    b. **Streaming Data**: εισάγονται στο Azure Event Hubs και η επεξεργασία τους θα γίνει με το Azure Functions και τη βοήθεια ενός trigger. Θα οριστεί ένα trigger το οποίο με το που ανέβουν τα αρχεία στο Storage Accounts θα προωθούνται κατευθείαν στο Azure Functions για επεξεργασία.
3. Τα αποτελέσματα των προηγούμενων διαδικασιών θα αποθηκεύονται στο Storage Accounts.
4. Στη συνέχεια, τα αποτελέσματα του προηγούμενου βήματος θα υποβάλλονται στο Microsoft Power BI όπου θα γίνει οπτικοποίηση και ανάλυση των συμπερασμάτων.

Την όλη διαδικασία επεξεργασίας των δεδομένων επιβλέπει και διαχειρίζεται το Azure Data Factory.

Κλιμακωσιμότητα και Ανοχή σε σφάλματα

Η εφαρμογή εξασφαλίζει τόσο στην κλιμακωσιμότητά της στις αυξανόμενες απαιτήσεις όσο και ανοχή σε σφάλματα λόγω των στοιχείων που χρησιμοποιούνται σε αυτή. Κάθε εργαλείο έχει μηχανισμούς ανταπόκρισης στις εξής απαιτήσεις. Συγκεκριμένα:

- Το Azure Blob Storage αντιγράφει αυτόματα τα δεδομένα για να εξασφαλίσει υψηλή διαθεσιμότητα και ανθεκτικότητα. Από προεπιλογή, τα δεδομένα αντιγράφονται τρεις φορές εντός της ίδιας περιοχής και προαιρετικά μπορούν να αντιγραφούν σε δευτερεύουσα περιοχή για σκοπούς ανάκαμψης από καταστροφή. Αυτή η αντιγραφή εξασφαλίζει ότι τα δεδομένα παραμένουν διαθέσιμα ακόμη και σε περίπτωση αποτυχιών hardware ή διακοπών του Data Center.
- Το Azure Blob Storage είναι σχεδιασμένο να κλιμακωθεί αυτόματα βάσει της ζήτησης. Όσο αυξάνεται το ποσό δεδομένων που αποθηκεύονται στο Blob Storage, μπορούν να προστεθούν δυναμικά επιπλέον κόμβοι αποθήκευσης για να εξυπηρετήσουν το αυξημένο φορτίο εργασίας. Αυτή η ελαστική κλιμάκωση επιτρέπει στο Blob Storage να χειρίζεται μεγάλα όγκο δεδομένων χωρίς απώλεια απόδοσης.
- Το Azure Blob Storage είναι χτισμένο σε μια κατανεμημένη αρχιτεκτονική που καλύπτει πολλά κέντρα δεδομένων και περιοχές. Αυτή η αρχιτεκτονική επιτρέπει τον οριζόντιο εκτεταμένο κλιμάκωσης διανέμοντας τα δεδομένα σε πολλούς κόμβους αποθήκευσης.
- Το Azure Databricks διαχειρίζεται την εκτέλεση των tasks σε clusters από διάφορα components. Αυτό επιτρέπει την αυτόματη κλιμάκωση και τη διαχείριση των ανωμαλιών μεταξύ των συστατικών.
- Το Azure Databricks επιτρέπει τη δυναμική προσαρμογή των πόρων υπολογισμού σύμφωνα με τη ζήτηση. Αυτό σημαίνει ότι μπορεί να κλιμακωθεί για να αντιμετωπίσει μεγάλους φόρτους εργασίας και να μειώσει το κόστος όταν η ζήτηση είναι χαμηλή.
- Το Azure Data Factory μπορεί αυτόματα να επεκτείνει ή μείωση των πόρων ανάλογα με το φόρτο εργασίας. Δυναμικά διαθέτει πόρους όπως απαιτείται, εξασφαλίζοντας αποδοτική χρήση πόρων και βέλτιστη απόδοση.
- Η Azure Data Factory εφαρμόζει πολιτικές επανάληψης για τις δραστηριότητες και τις λειτουργίες μεταφοράς δεδομένων. Εάν μια λειτουργία αποτύχει λόγω προσωρινών σφαλμάτων, επαναλαμβάνει αυτόματα τη λειτουργία σύμφωνα με την πολιτική επανάληψης που έχει διαμορφωθεί, μειώνοντας τον αντίκτυπο των αποτυχιών στις διαδικασίες επεξεργασίας δεδομένων.
- Η Azure Data Factory παρέχει δυνατότητες παρακολούθησης και ειδοποίησης για να παρακολουθεί την κατάσταση και την απόδοση των αγωγών επεξεργασίας δεδομένων. Δημιουργεί ειδοποιήσεις για οποιεσδήποτε ανωμαλίες ή αποτυχίες που ανιχνεύονται κατά τη διάρκεια της επεξεργασίας δεδομένων, επιτρέποντας στους διαχειριστές να λάβουν άμεσα διορθωτικά μέτρα.
- Το Azure Functions ακολουθεί serverless αρχιτεκτονική, όπου η διαχείριση της υποδομής δεν απασχολεί από τους προγραμματιστές. Κλιμακώνει αυτόματα ανάλογα με τη ζήτηση, εξασφαλίζοντας βέλτιστη χρήση πόρων και απόδοσης.
- Τα Azure Functions σχεδιάστηκαν να είναι ασύγχρονα, που σημαίνει ότι κάθε εκτέλεση λειτουργίας είναι ανεξάρτητη και δεν εξαρτάται από την κατάσταση των προηγούμενων εκτελέσεων. Αυτή η αρχιτεκτονική επιτρέπει στις λειτουργίες να

κλιμακώνονται οριζόντια με την εκκίνηση πρόσθετων instances για την επεξεργασία εισερχομένων αιτημάτων χωρίς να επηρεάζονται τα υπάρχοντα.