Q company data platform

Q company, specializing in retail, operates branches across various regions and utilizes an E-commerce platform. In the business day-to-day operations, we may have new products, customers, branches and salespeople. The company offers multiple offers for customers with a rule for offer redemption. Customer can redeem only one offer from regular offers (1-5). Offer discount rates from 1 to 5 (5 – 10 – 15 – 20 - 25) % from price.

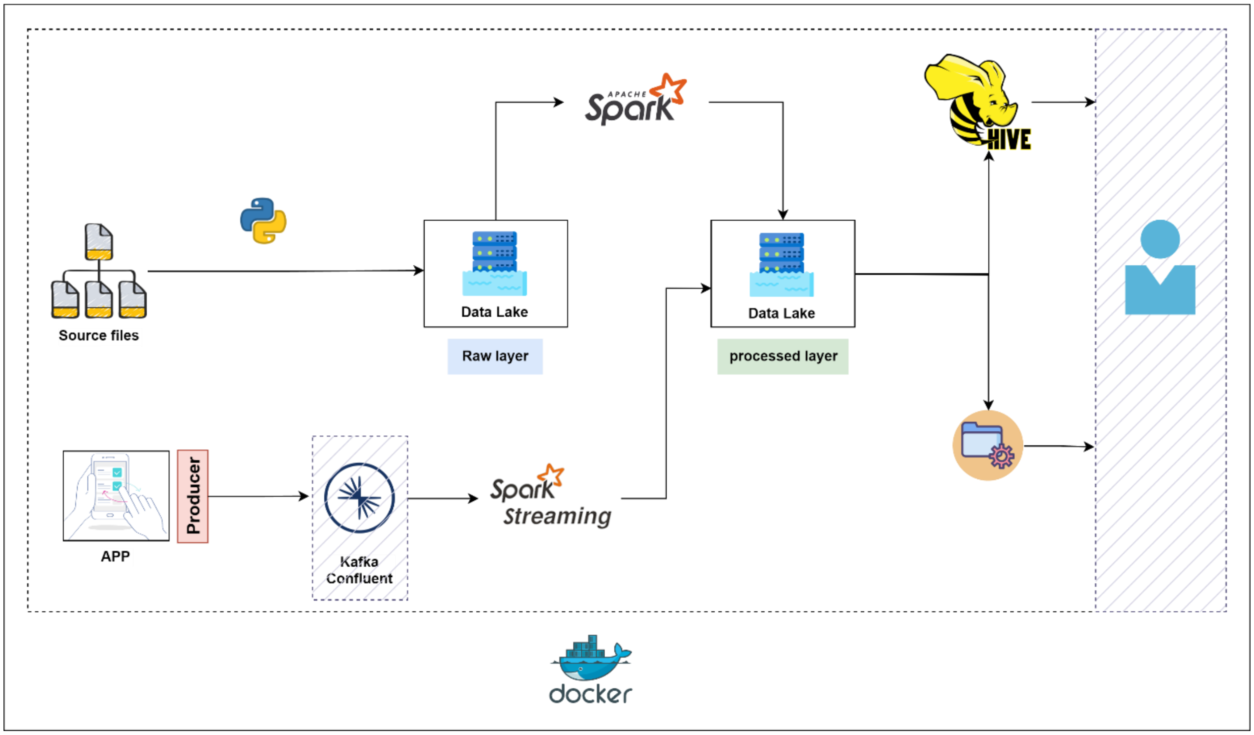
Data nature:

Source system pushes 3 files every hour (branches file, sales agents file, sales transactions file). For the sales transactions file, it contains both branches and online sales. For branches and sales agents' files, may contain new entries at any point of time.

schema of shipping address in online sales transactions: address/city/state/postal code

company app push logs to Kafka cluster to be processed later.

Q company Data platform



Batch Part:

Technical description:

*** Not all mentioned tools/stack are strict, except spark and hive which should be used in your setup

We have 6 groups of files; to simulate it as one group of files comes every hour you need to put one group of files in your pipeline starting point (local file system) with respect to order.

- Put one group of files every hour in the local file system (manually or automate it :)).
- Files in LFS should be ingested to data lake as raw files. (**Python**)
- As files are being pushed every hour on LFS, we should identify how to store data in data lake to be able to track it. (partitioning)
- After ingesting raw files in the data lake, we need to process these files to meet some/all business requirements and put them again in the data lake but as **Hive** tables to serve as DWH. (**spark**)
- Data should be cleaned well.
- business will require some insights will cover it later.

Business requirements:

- You're representing the business team! read your data carefully and build your own DWH model with respect to remaining requirements.
- Most query condition will be used from teams (transaction date)
- Total paid price after discount should be added as column in fact table.
- The marketing team needs to know most selling products, most redeemed offers from customers and most redeemed offers per product.
- The marketing team needs to know which lowest cities in online sales to run more campaigns.
- B2B team needs a daily dump (csv file) that contains (sales_agent_name ,product_name, total_sold_units) and this file should be sent to local file system

Streaming part:

Technical description:

We have Kafka python producer that sends app logs to Kafka cluster, logs have dynamic schema so you need to read producer code (data generating part) to detect all possible data elements that will be sent to Kafka.

- Start Kafka producer to run in background after configuring your topic.

  ```
  >>python /script/location/script.py
  ```

  ```
  Or you can run the producer from jupyter!
  ```

- Create spark streaming job that receive data from Kafka, process it and store data on HDFS.

Business requirements:

- You're representing the business team! read your data carefully and store it in a way that helps teams get most value from data.
- At least write two queries/report from this data.