1. What is Feature Store? What problem it is supposed to solve?

The feature store is a data warehouse of features for machine learning (ML). Architecturally, it differs from the traditional data warehouse in that it is a dual-database, with one database (row-oriented) serving features at low latency to online applications and the other database (column-oriented) storing large volumes of features, used by Data Scientists to create train/test datasets and by batch applications doing offline model scoring.

Benefits:

Helps in Faster development of ML models

Smooth model deployment in production

Increased model accuracy

1. What are building blocks of feature Store? Define & Research on each of the building blocks

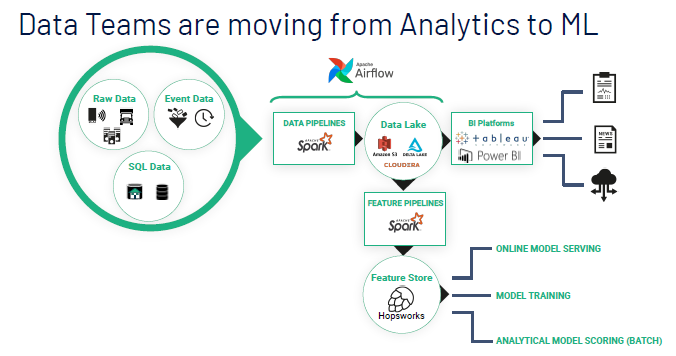
Relational Databases, Data Lake, Spark Pipelines are building blocks of feature store.

Relational Database : A relational database is a collection of data items with pre-defined relationships between them. These items are organized as a set of tables with columns and rows. Tables are used to hold information about the objects to be represented in the database.

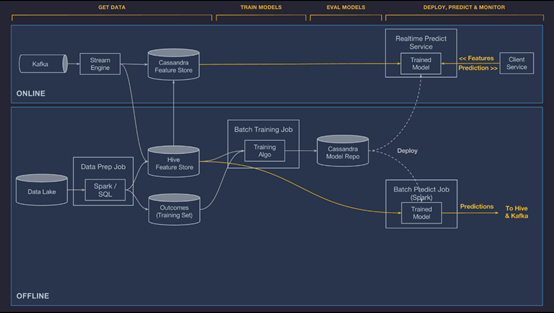
Data Lake : Data Lakes allow you to store relational data like operational databases and data from line of business applications, and non-relational data like mobile apps, IoT devices, and social media. They also give you the ability to understand what data is in the lake through crawling, cataloging, and indexing of data.

Spark Pipelines: Apache Spark is a very demanding and useful Big Data tool that helps to write ETL very easily. You can load the Petabytes of data and can process it without any hassle by setting up a cluster of multiple nodes

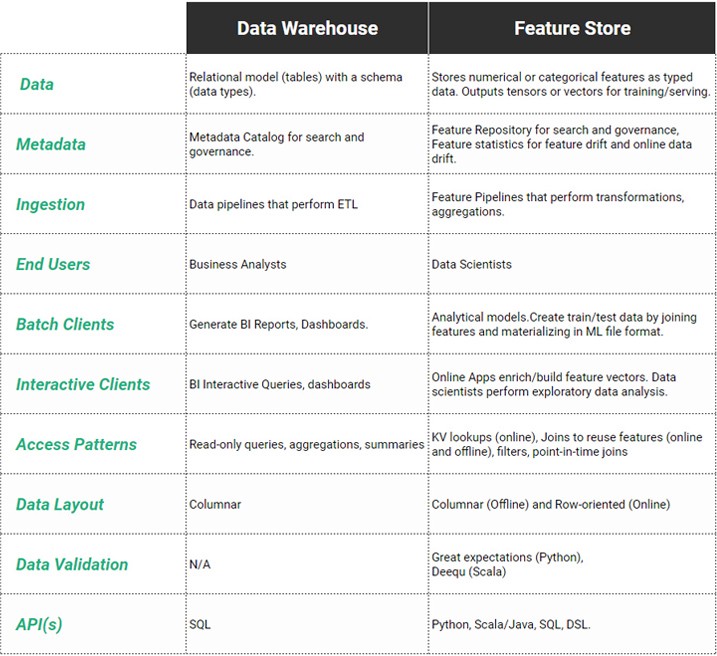
1. Come up with a process flow block diagram for feature Store



Sample Uber feature store



1. Do Competitor Analysis on Feature store and list down their Features & Capabilities



1. Record Your findings in slides & Microsoft word document and attach all the resources that you have referred

Data warehouses can be used to store pre-computed features, but they do not provide much more functionality beyond that for ML pipelines. When Data Scientists need to create train/test data using Python or when online features (for serving features to online models) are needed at low latency, you need a feature store. Similarly, if you want to detect feature drift or data drift, you need support for computing feature statistics and identifying drift.

References:

<https://www.iteblog.com/ppt/sparkaisummit-north-america-2020-iteblog/building-a-feature-store-around-dataframes-and-apache-spark-iteblog.com.pdf>

<https://medium.com/data-for-ai/feature-store-vs-data-warehouse-306d1567c100>

<https://towardsdatascience.com/what-are-feature-stores-and-why-are-they-critical-for-scaling-data-science-3f9156f7ab4>