12/1/2021 Murtaza Badshah

Bond Brand Data Question Solution

Assumptions:

1) Each store manages their own inventory (No standardized inventory tracking)
2) Sales from each of the 2000 stores are tracked using a POS system.
3) Online transactions are the only data that is streaming.
4) POS sales are synced in 15 min increments
5) Management wants to be able to track the inventory and synchronize the data across 2000 stores.

Requirements:

1) Handle large write volume: Billions of write events per day.
2) Handle large read/query volume: Millions of merchants wish to gain insight into their business. Read/Query patterns are time-series related metrics.
3) Provide metrics to customers with at most one-hour delay.
4) Run with minimum downtime.
5) Can reprocess historical data in case of bugs in the processing logic

Proposed Solution:

To satisfy the above requirements, we need to propose a 4-layer data platform.

Layer 1: Data Ingestion

With billions of write events are coming per day, we proposed utilizing an open-source tools such as Apache Kafka that can handle data streaming and data ingestion on a large scale. As Apache Kafka utilizes clusters of machines to process data at low latencies. It makes sense to adopt this tool for large writes for the events.

Layer 2: Data Storage and Processing

To store the large streams, we need to develop a separate cloud-based Data Lake for the Pizza company. This will help with handling large read tasks and store time series metrics. For this process Azure Databricks can be used to develop the data lake. Based on the Apache Spark architecture it would be the best solution for the company to store the data at reasonably low cost.

Layer 3: Data Transformation and Modeling

While Databricks SaaS platform provides the modeling and transformation, we can also utilize Apache Airflow for data transformation and modeling to develop the analytics.

Layer 4: BI & Analytics

Lastly, we can connect a BI tool such as Tableau or Power BI to develop the analytics that can be immediately provided to the clients on demand.