Assignment: Data Engineering

Marina Nazeh Mikhael

Chosen company: Talabat

Talabat is the leading on-demand delivery platform in the Middle East & North Africa (MENA), offering online ordering and delivery for food, groceries, and other essentials from restaurants and retailers via its app and website. Founded in Kuwait in 2004, it's now a subsidiary of Delivery HeroSE, operates across 8 countries (UAE, Kuwait, Qatar, Bahrain, Egypt, Oman, Jordan, Iraq), and provides a digital marketplace connecting customers, partners, and riders.

1. Data Sources

   The data being collected by talabat is basically: User information like name, age, location, email,...etc, Order records for each user, and also the partners information like Restaurants' meu, offers and prices by keeping a database with these information. They also get data from users' behavior on the app itself, most time spent in which page of the app, how many clicks till the order is in checkout from things similar to the cookies I guess, another thing is the realtime location of the user and the location of the restaurants and the optimal route between them in terms of time taken and distance from things like GPS.

2. Data Ingestion Layer

   Talabat may be using 3 methods for its data to enter the system:
   a. Streaming for realtime data like: order state, GPS, payment and the app clicks for the users. Tools used for this: Apache Kafka or Amazon Kinesis.
   b. Batch: for the daily tasks like updating a restaurant's menu. Tools used for example: Apache Airflow.
   c. API Gateway: used for external data requests from partners. Tools used: AWS API Gateway.

3. Storage layer

   Data Lake -> stores all raw data with all its types wither structured or not. Cloud storage: Amazon S3.

   Data Warehouse -> used to store only structured data (cleaned, transformed, and modeled data) with its information ready to be used and apply queries on. Using AWS Redshift.

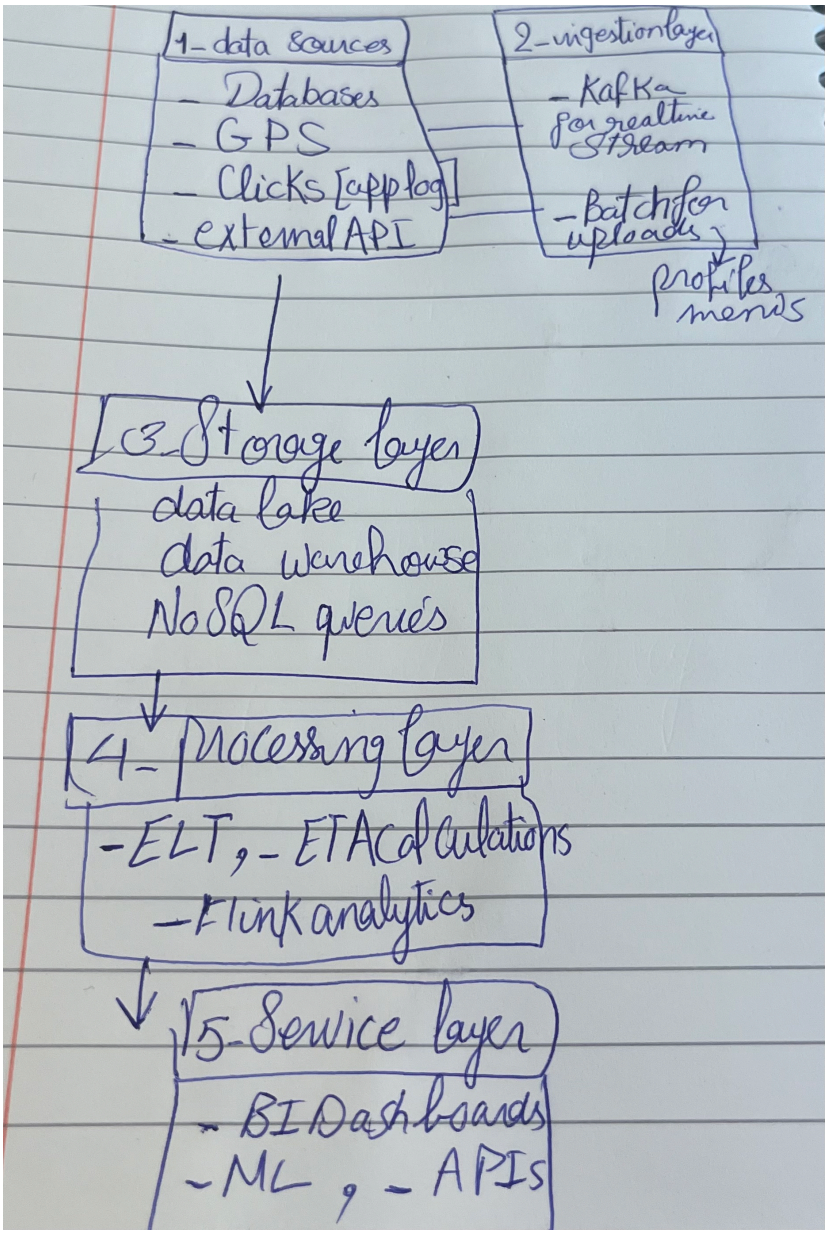
   No SQL -> low latency storage for real-time features and caching. DynamoDB.

4. Processing Layer

   ELT and data build tool: it is used for the output required to use in BI dashboards and machine learning models.
   Stream Processing: used for the realtime data using Apache Flink, Kafka Streams.

5. Serving layer
   They use BI dashboards using tools like power BI and Tableau.
   They use ML models on the realtime data to know how to improve their services.
   APIs are used also to feed the user app with realtime data and to do things like auto calculations and the estimated time of arrival and so on.

   pipelined diagram:



Used Gemini in answering some questions