**Day 6: Sakina\_HW**

**Task 1: BigQuery Project Use Case – E-commerce Sales Analytics**

**Industry:** E-commerce  
**Problem Statement:**  
An online retail company wants to analyze its customer transactions, product performance, and marketing campaigns across millions of records in real-time to make data-driven decisions.

**Data Sources:**

* Daily transaction logs stored in Google Cloud Storage (CSV/JSON).
* Product catalog and customer data in Google Cloud SQL.
* Marketing campaign data pulled from Google Ads API and stored in BigQuery.

**Solution with BigQuery:**

* Load structured data into **BigQuery native tables** using batch ingestion from GCS and Cloud SQL.
* Use **SQL** queries to generate insights such as:
  + Top-performing products by region.
  + Customer lifetime value (CLV).
  + Sales trends by time period and marketing channel.
* Create **dashboards in Looker Studio** directly connected to BigQuery for real-time reporting.
* Run **machine learning models using BigQuery ML** for churn prediction and personalized recommendations.
* Benefit from **scalability** to handle seasonal traffic spikes and **fast query performance** over petabytes of data.

**Key BigQuery Features Used:**

* Serverless, scalable infrastructure.
* Seamless integration with GCS, Google Ads, and Looker.
* SQL support with powerful OLAP capabilities.
* Support for partitioning and clustering to improve performance.

**Task 2: BigLake Project Use Case – Healthcare Data Lake for Research**

**Industry:** Healthcare  
**Business Challenge:**  
A hospital research department needs a unified platform to store, secure, and analyze diverse datasets from multiple sources to support public health studies.

**Data Variety & Sources:**

* **Structured data**: Electronic Health Records (EHRs) in Parquet format from hospital systems.
* **Semi-structured data**: Lab reports and prescriptions in JSON.
* **Unstructured data**: Medical images (X-rays, MRIs) and doctor notes (PDFs, DOCX).
* All stored in **Google Cloud Storage (GCS)** across multiple buckets.

**Solution with BigLake:**

* Use BigLake to create **external tables** pointing to Parquet and JSON files in GCS without moving data.
* Apply **fine-grained access control** via IAM and Dataplex to ensure researchers only see anonymized, relevant data.
* Enable **cross-team collaboration** by allowing data scientists to use Spark or Presto for ML model development and exploration.
* Ensure **data governance** through a centralized metadata catalog and data classification.

**Key BigLake Features Used:**

* Supports **multi-format** (JSON, Parquet, PDF).
* Allows **unified access** across warehouse (BigQuery) and lake (GCS).
* Compatible with **open-source engines** like Spark and Trino.
* Enables **centralized governance** and role-based security with Dataplex.