**24030012 Syed Feroz Raza**

**24280042 Muhammad Naveed Ashfaq**

**Group 39**

**Assignment 2: Building a Batch Analytics Pipeline on HDFS &**

**Hive**

**6: Short Write-Up: Batch Analytics Pipeline for MediaCo**

**Overview:**

The batch analytics pipeline processes streaming platform logs and metadata, storing data in **HDFS** and querying it via **Hive**. The pipeline includes **Ingestion**, **Data Transformation**, and **Querying** stages.

**1. Ingestion (ingest\_logs.sh)**

The script loads raw logs into HDFS, organized by **date**. It accepts a date parameter and stores data in partitioned directories (/raw/logs/<year>/<month>/<day>).

* **Design**: Date-based partitioning for efficient querying.
* **Performance**: Execution time depends on data size, typically minutes for small datasets, hours for larger ones.

**2. Raw Tables in Hive (raw\_logs\_table\_ddl.sql)**

Hive external tables are created for raw logs and metadata, partitioned by **year**, **month**, and **day** for optimized queries.

* **Design**: Partitioning by date enhances query performance.
* **Performance**: Speeds up date-based queries by filtering partitions.

**3. Star Schema (star\_schema\_ddl.sql)**

Two Hive tables are created: a **fact table** for user actions and a **dimension table** for content metadata, both stored in **Parquet** for efficiency.

* **Design**: Fact and dimension tables in **Parquet** for fast queries.
* **Performance**: Parquet reduces I/O, speeding up queries.

**4. Data Transformation (data\_transformation.sql)**

Data is transformed from raw tables into the star schema with timestamp conversion and data population.

* **Design**: Data is joined and inserted into the star schema.
* **Performance**: Transformation is efficient due to partitioning and parallelization.

**5. Analytical Queries (sample\_queries.sql)**

Three analytical queries provide insights on user activity, content popularity, and session lengths.

* **Design**: Queries use **JOIN**, **GROUP BY**, and **partition filters**.
* **Performance**: Simple queries execute in **seconds**; complex ones take **minutes to an hour**.

**6. Execution Times**