Data Engineering Infrastructure and Work Progress Report

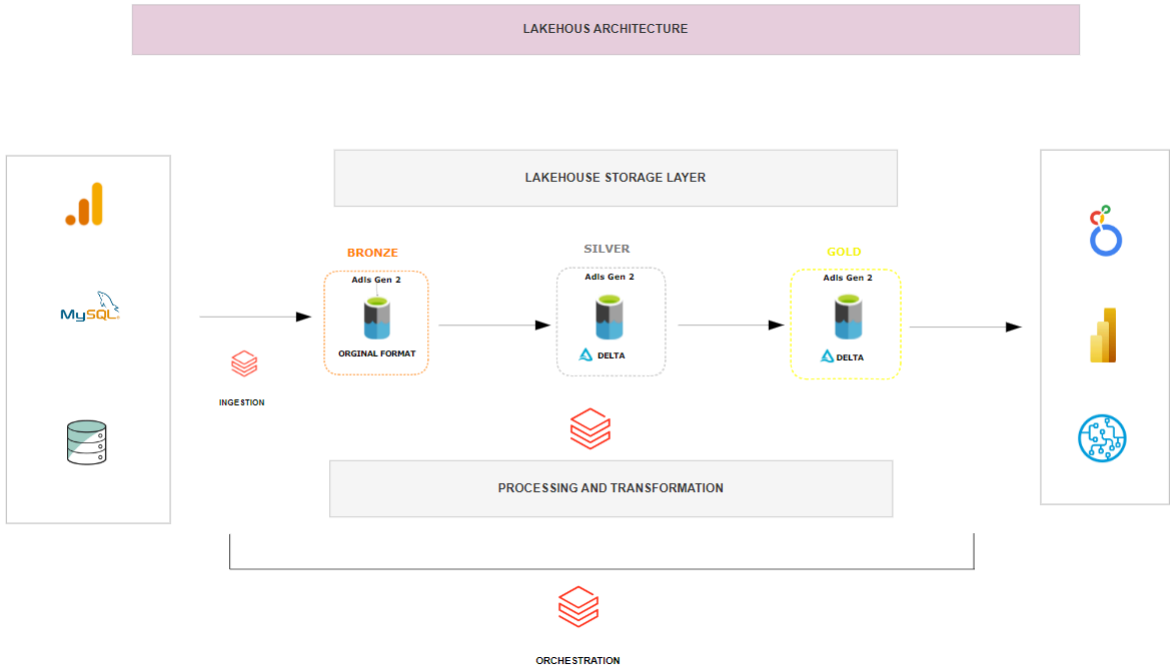
---

1. Executive Summary

This report outlines our ongoing data engineering efforts to enhance business intelligence and forecasting capabilities. We are capturing data from an AWS-hosted MySQL database, processing it using Azure Databricks, and storing it in Azure Data Lake Storage. The data is structured in three layers—Bronze, Silver, and Gold—with the Gold layer already integrated with our business intelligence tools.

Key achievements include successful data ingestion, transformation, and preparation of key tables for business analysis. Current challenges involve ensuring data freshness. Next steps include automating data updates, improving processing efficiency, and developing sales and demand forecasting models for both mobile & web platforms.

Our Architecture



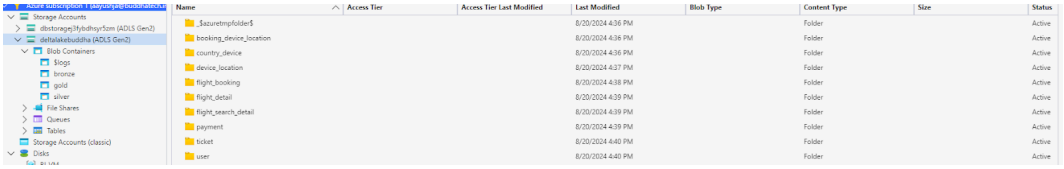
---

2. Infrastructure Overview

- **Data Source**:
  - ★ **Location**: The original data comes from a MySQL database hosted in the AWS cloud.

- **Data Processing**:
  - ★ **Tool**: We use Azure Databricks, a powerful platform that allows us to process large amounts of data efficiently.
  - ★ **Programming Language**: The data processing is done using PySpark, a programming language designed for big data tasks.

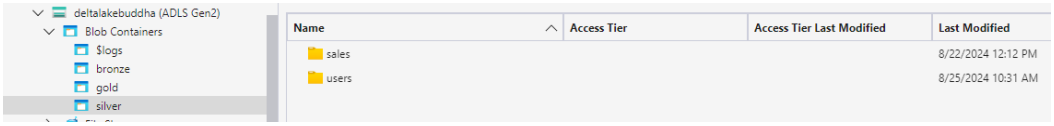
- **Data Storage**:
  - ★ **Bronze Layer**:



    - **Function**: This is where we store the raw data exactly as it is received from the source.

    - **Purpose**: The raw data is preserved for historical analysis or reprocessing if needed.

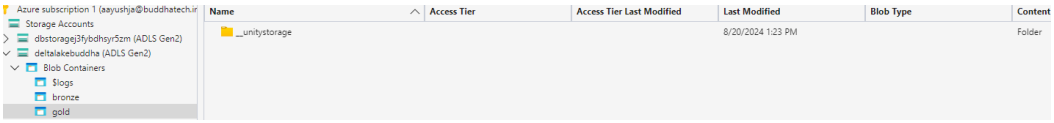
  - ★ **Silver Layer**:



    - **Function**: Here, we clean and structure the data, making it easier to work with.

    - **Purpose**: The Silver layer prepares the data for more in-depth analysis.

  - ★ **Gold Layer**:



- **Function**: This layer contains the final, refined data that is ready for analysis.
- **Purpose**: Data from the Gold layer is used in business intelligence tools to generate insights and reports.

Data Preview

	Identifier	Repetition	LTV	NPassenger	RecentPurchase	Currency	Sectors	UserType	UserStatus	AgeGroup	ConQuantile	Score	Category
1	007abhiman@gmail.com	1	3849.03	1	2023-09-20	NPR	["KTM-SIF"]	Individual	Login	Unknown	112	1	Lost
2	007arai@gmail.com	2	12753.98	2	2024-04-08	NPR	["KTM-BDP...	Individual	Login	Unknown	333	3	High Potential
3	007binodp@gmail.com	2	8433.45	2	2024-02-21	NPR	["PKR-BWA...	Individual	Login	Unknown	233	2	High Potential
4	00ankit00@gmail.com	12	73653.48	13	2024-07-07	NPR	["KTM-PKR...	Individual	Login	Unknown	555	5	Loyal
5	0163294@gmail.com	1	4289.34	1	2023-12-14	NPR	["KTM-JKR"]	Individual	Guest	Unknown	112	1	Lost
6	01deathly.hallows@gmail.com	1	17558.44	2	2023-12-08	NPR	["KTM-PKR...	Individual	Login	35-44	312	3	Hybernating
7	0201shercbij@gmail.com	1	4377	1	2024-04-05	NPR	["KTM-PKR"]	Individual	Login	Unknown	113	1	New
8	03soruthapa@gmail.com	1	5176.99	1	2024-03-23	NPR	["KTM-BW...	Individual	Guest	Unknown	113	1	New
9	0415678184	3	15400	2	2023-08-02	NPR	["KTM-DHI...	Business	Guest	Unknown	342	3	Hybernating
10	0507abd@gmail.com	4	25537.86	4	2024-07-29	NPR	["KTM-RJB...	Individual	Login	Unknown	445	4	Loyal
11	0610anil@gmail.com	2	7586.99	1	2023-11-05	NPR	["KTM-BIR"]	Individual	Guest	Unknown	232	2	Hybernating

- **BI Tools Integration**:
  - ★ **Current Status**: Data in the Gold layer is already linked with our BI tools, providing us with real-time insights and enabling data-driven decision-making.

---

3. Implementation Details

- **Completed Work**:
  - ★ **Data Ingestion**:
    - **Process**: We have successfully set up a process to capture and store data from the MySQL database into the Bronze layer in Azure Storage.
    - **Outcome**: This ensures that data is securely stored and can be accessed for further processing.
  - ★ **Data Transformation**:
    - **Process**: The data from the Bronze layer is cleaned and organized in the Silver layer, making it easier to analyze.

        - **Outcome**: We have completed the main table creation for Sales and the users tables which are being stored in the Bronze layer, which are already being used for further analysis.
    - ★ **Final Data Preparation**:
        - **Process**: The bronze layer data is further refined and stored in the Gold layer.

- **Outcome**: The Gold layer now holds key tables like salesAgg and averagebookingtime, RMF, sector sales tables, monitor tables like range_rmf, dynamic calendar and others which are already being used for business intelligence.

- **In Progress**:
    - ★ **Automating Data Updates**:
        - **Objective**: Set up a pipeline that automatically updates the raw bronze and gold layer with new data.
        - **Benefit**: This will ensure that our data remains current without manual intervention.
        - **Current Status**: We are in the process of developing this pipeline which is an iterative process.
    - ★ **Improving Efficiency**:
        - **Objective**: Streamline our data processing to reduce time and resource usage.
        - **Benefit**: This will allow us to process data faster and more cost-effectively.
        - **Current Status**: Optimization efforts are underway.
    - ★ **Preparing for Forecasting**:
        - **Objective**: Develop models to predict sales and demand for both mobile and web platforms.

- **Benefit**: Accurate forecasts will help us better plan and allocate resources.
- **Current Status**: Data is being prepared for this analysis (Ongoing discussion)

---

4. Challenges and Risks

- **Data Freshness**:
  - ★ **Challenge**: Keeping the data in the Bronze layer up-to-date.
  - ★ **Solution**: Implement an automated data loading pipeline.
- **Resource Efficiency**:
  - ★ **Challenge**: Reducing the time and cost of processing data.
  - ★ **Solution**: Optimizing the current data workflows.
- **Forecast Accuracy**:
  - ★ **Challenge**: Developing models that accurately predict future sales and demand due to inadequate data and data quality
  - ★ **Solution**: Using historical data and advanced algorithms to improve prediction accuracy and creating external tables such as an event calendar.

---

5. Resources Used

- **Cloud Services**:
  - ★ **AWS**: Hosts the MySQL database that stores the original data.
  - ★ **Azure**: Provides the infrastructure for data processing and storage.

---

6. Performance Metrics

- **Pipeline Automation**:
  - ★ **Goal**: Ensure that data updates are automated and seamless.

★ **Current Progress**: Pipeline is being developed.

- **Efficiency Gains**:
  - ★ **Goal**: Reduce processing time and cost.
  - ★ **Current Progress**: Optimization is in progress.

- **Forecast Accuracy**:
  - ★ **Goal**: Achieve a high level of accuracy in sales and demand predictions.
  - ★ **Current Progress**: Data preparation is underway.

---

7. Next Steps

- **Immediate Actions**:
  - ★ Complete the development of the automated data loading pipeline.
  - ★ Implementation of replication from functional (production db) to transit database on a daily basis (Which will store the data not older than 3 days). That enables us to ingest data from the transit database to the bronze layer on a daily basis.
  - ★ Creation of Event calendar, which contains the major events within the national and international territory such as election, major holidays, disaster, festival, national and international events.
  - ★ Finalize the optimization of current workflows.
  - ★ Integration of Reservation data source*

- **Future Actions**:
  - ★ Develop and validate forecasting models for sales and demand in case of web and mobile applications transaction.
  - ★ Develop and validate Dynamic Pricing models.*
  - ★ Continue integrating additional data sources as needed.

Cost Estimation:

★ **Data Engineering**

The tools and technology that are being utilized by the data engineering team are data storage account, databricks, virtual machine and disks. The total estimated cost for these infrastructure is estimated to be within the range of **200 - 350** USD per month, assuming we'll utilize the resources **8hrs/day** for on average **20 days/month**. As you can see the cost is estimated on the basis of usage hence as the usage and the data grows cost will increase as well.

★ **Business Intelligence (Visualization)**

The tools and technology that are on plan to be implemented are Power BI for data visualization and the inhouse application where the reports will be hosted and can be accessed from for the strategic decision and planning.

In a nutshell, we will be utilizing Power BI and SKU to host the reports and that costs us around **1.4$/hr** usage. So assuming the business will view the report on average **6 hrs/day 24 days/month** and the cost will be around **200-250** USD excluding the licensing cost that will be around (**$40/month**) for 2 licenses.

Now, The estimated cost for BI is around **$300/month** on an average.

Hence, our research tells us that the total cost for the DE and BI Infra should not be exceeding **$700/month.**

8. Appendices

- **Glossary**:
  - ➔ **Bronze Layer**: Raw data storage
  - ➔ **Silver Layer**: Cleaned and structured data and analysis-ready data
  - ➔ **Gold Layer**: Refined, aggregated data
  - ➔ **PySpark**: Python API for Apache Spark, used for data processing

---