
# Statement of Work (SoW) for IPL Team Data Pipeline Using a Different Approach

By Shritej Chavan

INT-30

Intern

(2025-26)

**Table of Contents**

# 1 Introduction

**"IPL Gold Layer Analytics – Powered by Azure & Power BI,"** focuses on building an end-to-end data pipeline that enables advanced insights into Indian Premier League (IPL) team performance. Leveraging the **Medallion Architecture** approach, the pipeline is designed to process, transform, and analyze IPL data using **Azure Databricks (PySpark), SQL Server (SSMS), Azure Data Factory, and Power BI**.

The core objective is to ingest raw IPL datasets (such as team info, player roles, match performance, and stadiums) into the **Bronze layer**, perform cleaning and transformation operations to form the **Silver layer**, and aggregate this data into meaningful metrics stored in the **Gold layer** for reporting and visualization purposes.

Throughout the pipeline, structured operations are carried out using PySpark for ETL processes and SQL Server for data storage and aggregation. All processed insights are visualized through a dynamic and interactive **Power BI Dashboard**, which features KPIs like **team win percentage, top player performance, and venue-wise comparisons**.
This project not only demonstrates strong technical capability in cloud-based data engineering and visualization but also ensures **scalability**, **reusability**, and **auditability** across all layers of data processing, fulfilling modern data pipeline standards.

## 1.1 Basic Concepts

**ETL (Extract, Transform, Load):** the process of extracting data from source systems, transforming and refining it to satisfy specific requirements, and finally sending it to a specified destination for further use.

**Data pipeline:** an automated operation that executes and coordinates data transformation and transportation across multiple systems seamlessly.

**Parquet:** Parquet is a columnar data format optimized for storing and accessing data efficiently, specifically tailored to handle large-scale data workloads effectively.

**Azure:** Microsoft's cloud platform offering cloud computing and storage services.

**Databricks:** a cloud computing and analytics platform. Databricks provisions compute for data processing, provides user interface for data operations.

**Azure Data Factory (ADF):** is one of the services in Azure cloud platform. ADF is used to extract data from source systems and orchestrate data pipelines.

**Power BI:** a business analytics tool developed by Microsoft that allows users to visualize data, share insights, and create interactive dashboards and reports.

**PySpark**: A python API for Apache Spark, enabling Python developers to leverage the power of Spark for large-scale distributed data processing, machine learning, and real-time analytics.

**MySQL:** an open-source relational database management system that enables users to store, organize, and retrieve data efficiently.

**Azure Blob Storage:** a Microsoft cloud-based solution designed for storing large amounts of unstructured data, such as text, images, videos, and log files.

**Azure Data Lake Storage Gen2**: Microsoft's cloud-based solution for big data analytics, combining high-performance hierarchical file storage with the scalability of Azure Blob Storage. It is designed to handle both structured and unstructured data, enabling efficient data processing and integration for advanced analytics workflows.

**Automation in Azure Data Factory (ADF):** it refers to the process of streamlining and orchestrating data workflows, enabling the execution of ETL tasks without manual intervention. It leverages triggers, schedules, and pipelines to automate data movement, transformation, and integration across various sources and destinations efficiently.

### 1.2 Medallion Architecture

Medallion Architecture is a data design pattern with its core idea being to logically organize data into layers. Classic Medallion Architecture consists of three layers: bronze, silver, gold. Data is processed, cleaned, and moved between layers using data pipelines. The quality and structure of data increases from layer to layer.

- **Bronze Layer**: Stores raw data directly ingested from the source system.
- **Silver Layer**: Contains cleaned and transformed data from the bronze layer.
- **Gold Layer**: Holds refined, use-case-specific data for business needs.

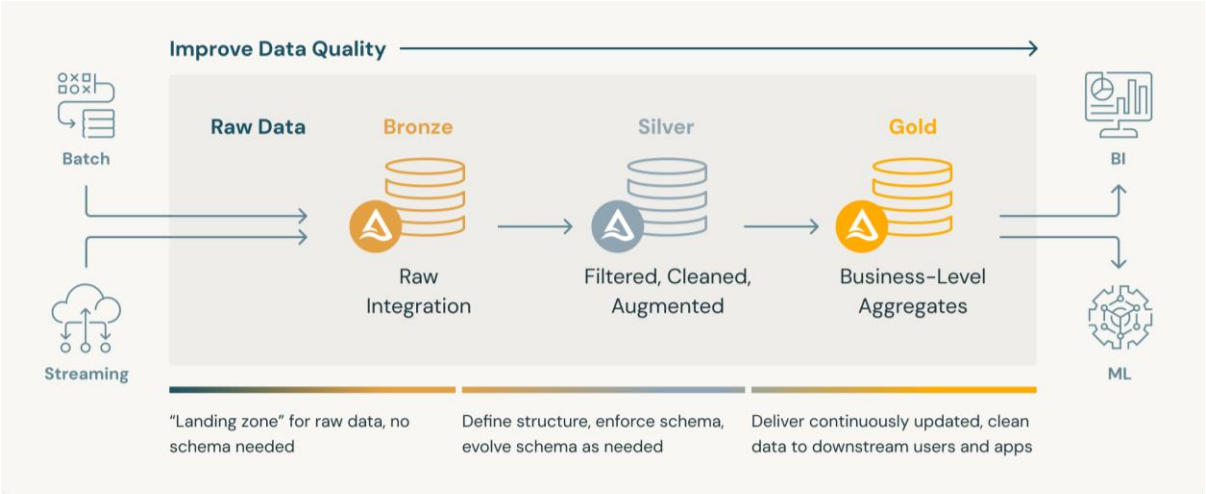


Figure 1. Medalion architecture.

# 2 Architecture Overview

This project uses the Medallion Architecture concept as discussed above, which divides the data pipeline into three fundamental layers: bronze, silver, and gold. Each layer represents a stage of data refinement, enabling scalable, maintainable, and traceable data processing.

In this project the Medallion architecture is in the Azure Data Lake Storage Gen2 storage but since the source file was of csv and the raw file needed to be of parquet for better processing, the source file was stored in a new Azure blob storage and was transformed and brought into ADLS container bronze using PySpark Databricks notebook.

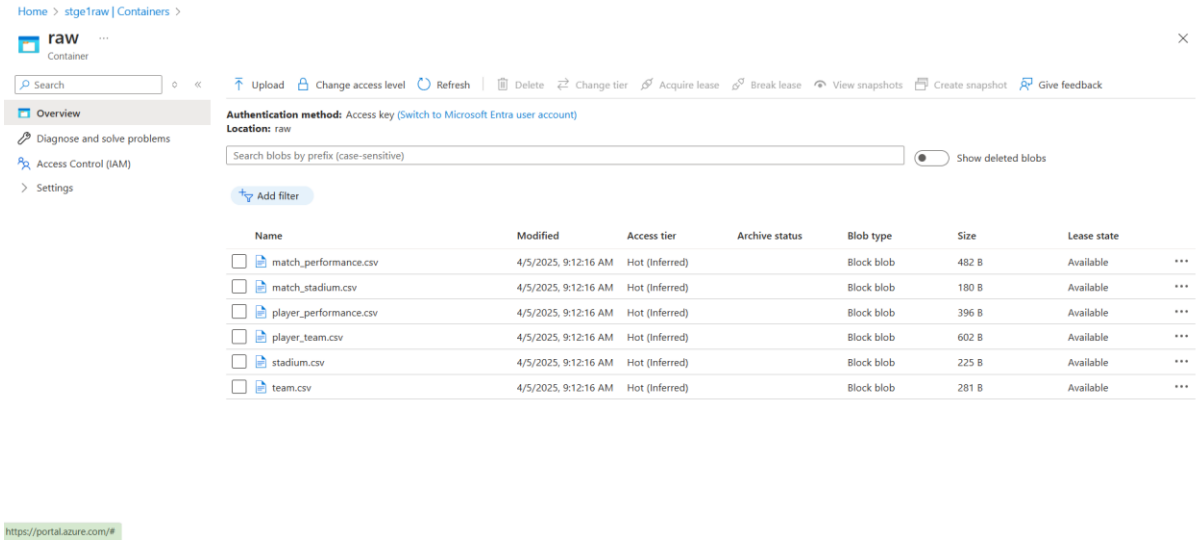


Figure 2. raw container with source csv files

- **Bronze Layer:**
  The Bronze layer contains raw, unaltered data. There are no data types established, and the data is identical to the source system except that it is in parquet format.
  The bronze layer data is used to recreate all subsequent levels. This layer frequently contains meta-data, such as the timestamp when the data was entered, the original file name, or the streaming source name.
  The bronze layer can contain numerous tables, each with data at a different transformation step.

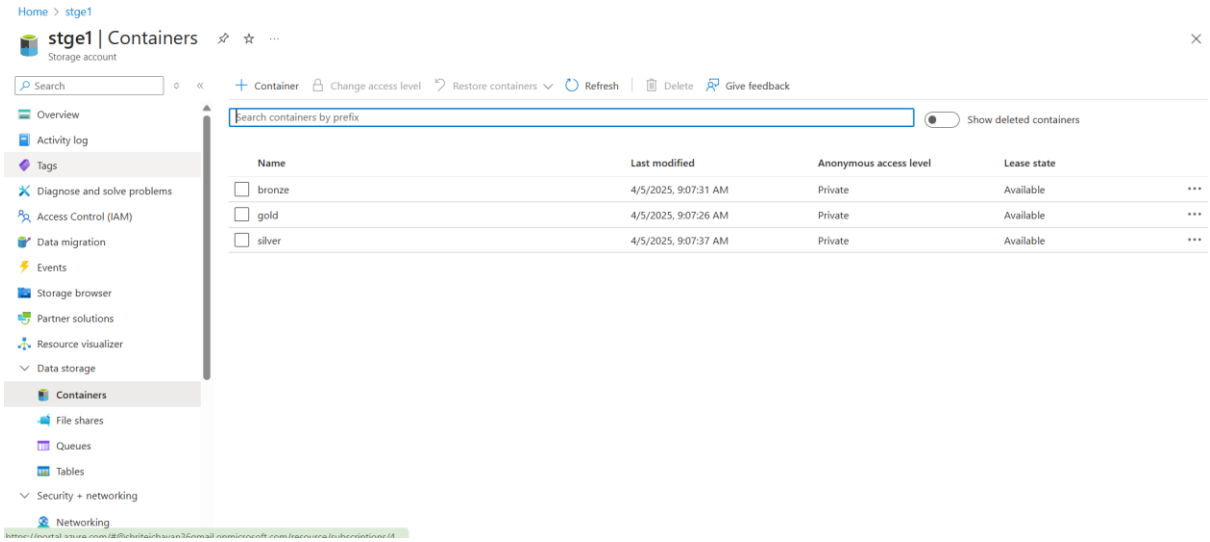


Figure 3. ADLS Gen2 Storage

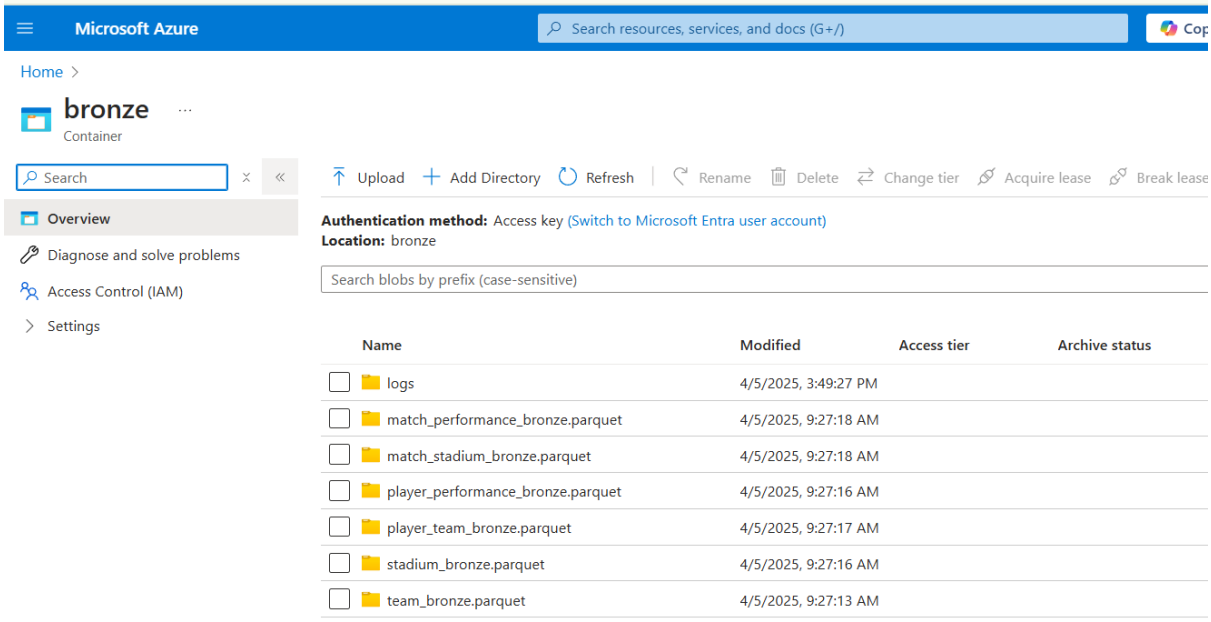


Figure 4. ADLS bronze container

- **Silver Layer:**
  The silver layer contains data that has been conformed, cleansed, and converted in PySpark Databricks. This includes transformation and cleaning processes such as deleting duplicates, resolving faulty or incorrect data and assigning data types and the separation of data into columns. Silver layer has data in two formats one is parquet and the other is SQL as Deatricks notebook writes the transformed data as

parquet file to silver ADLS container and to MySQL database silver _db as delivery_data_silver table.

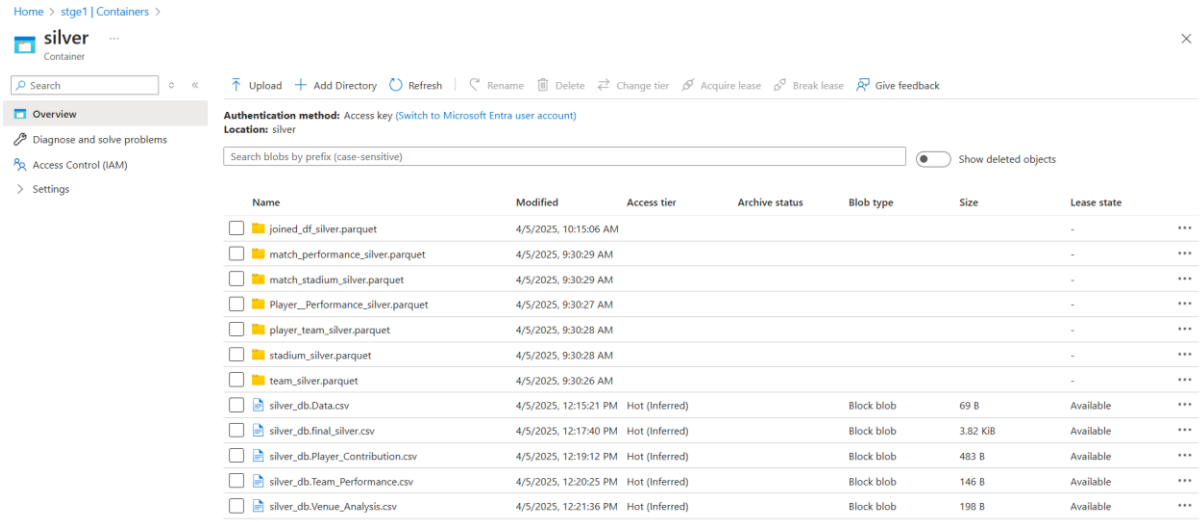


Figure 5. Silver container

- **Gold Layer:**

  The gold layer contains final aggregated dataset in MySQL for Power BI reports, computing key business metrics.

  **Operations Performed in MySQL (Using Silver Layer as Source):**
  - Joining Related Tables
  - Calculating Team Performance Metrics
  - Player Contribution
  - Venue Analysis
  - Player Efficiency Metrics with AVG(strike_rate), SUM(batting_runs), SUM(wickets_taken) by player.

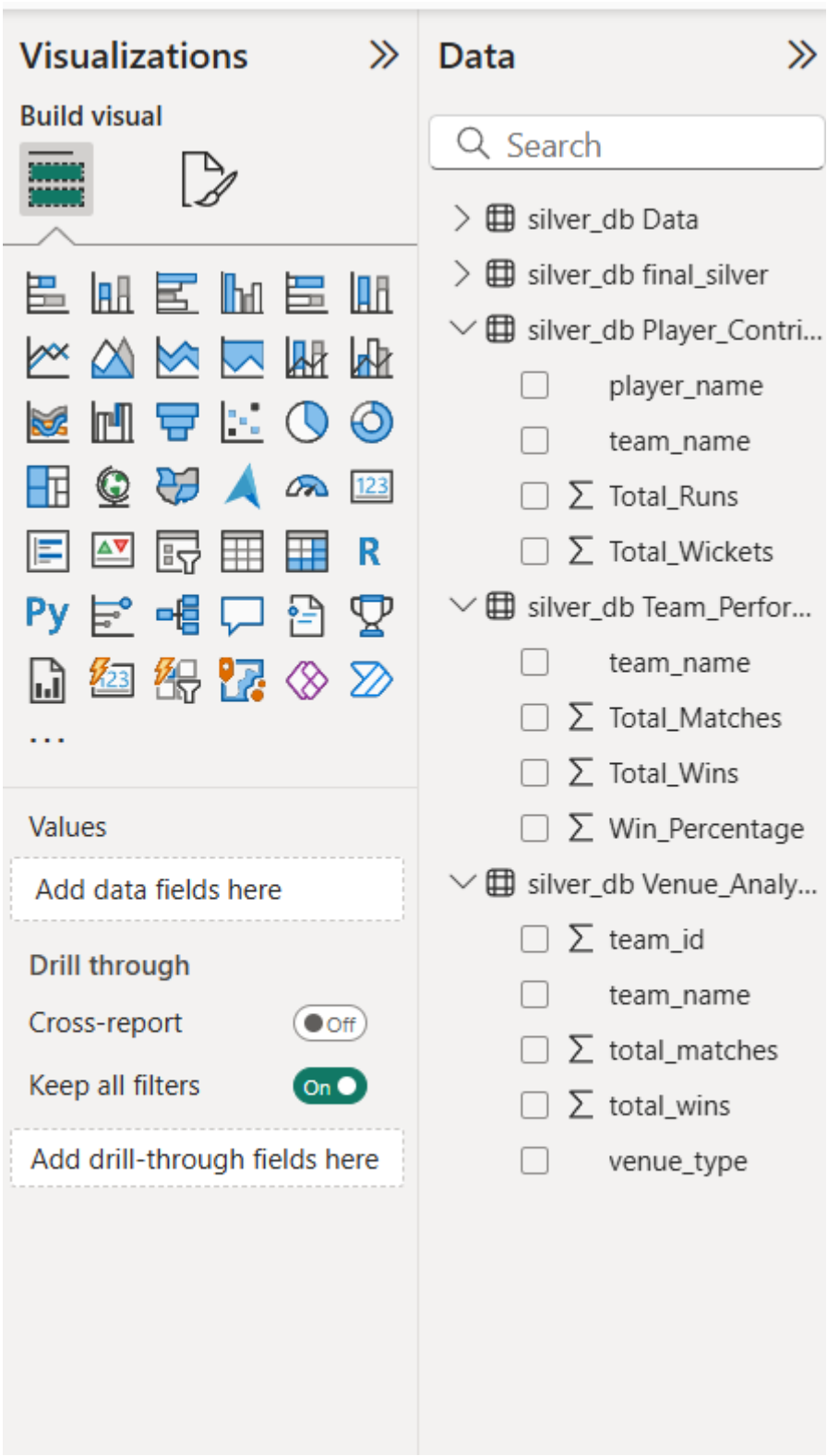


Figure 6. Loaded the required data in bi platforms

# 3 Solution Implementation

The Medallion Architecture was used to build the project solution. The entire process was orchestrated with Azure tools and big data technologies to enable scalability and automation.

### 3.1 Data Ingestion -Bronze Layer

The raw data files were provided in CSV format, representing various aspects of a logistics operation, such as delivery_data.csv, driver_data.csv, vehicle_data.csv, route_data.csv.

These files were ingested using PySpark in Azure Databricks and stored in ADLS the Bronze layer from raw in Azure blob storage.
During ingestion:

- The schema was inferred.
- Additional metadata fields like ingestion date and source file were added to maintain an audit trail.
- Data was written in Parquet format to ensure efficient storage and querying.
  This ingestion step laid the foundation for reliable, traceable data processing.

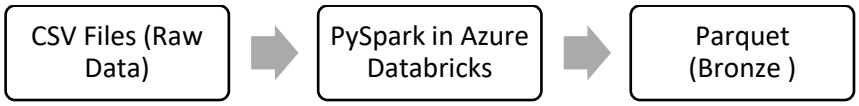


Figure 11. Ingestion Flow (bronze layer)

### 3.2 Data Processing -Silver Layer

In the Silver layer, the goal was to produce clean, enriched datasets ready for business analysis. This involved:

- Null Handling: Records with missing critical fields were filtered out.
- Data Type Casting: Fields such as dates and numerical values were correctly cast for consistency.
-

The processed silver data was written both to Parquet files in ADLS and to MySQL table using JDBC for easy BI access.

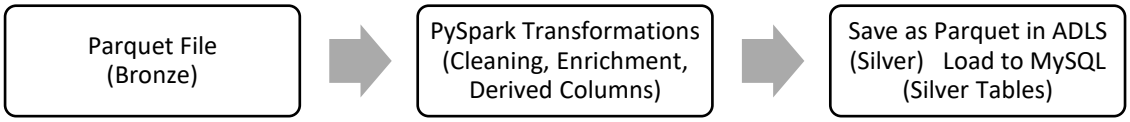


Figure 7. Processing Flow (silver layer)

### 3.3 Aggregation and Business Logic – Gold Layer

The Gold layer focused on generating business-level aggregates . This was achieved through SQL queries executed in MySQL on top of the silver table.

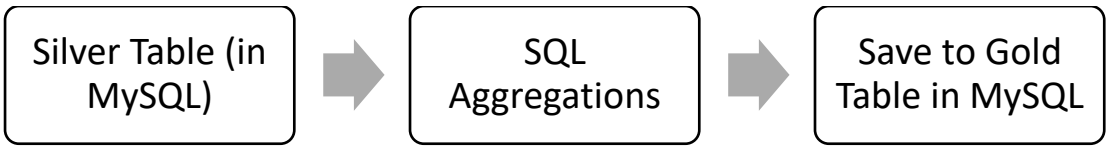


Figure 8 . Aggregation Flow (gold layer)

### 3.4 Dashboarding with Power BI

The final step was to present these insights through a professional, interactive Power BI dashboard, connected to the Gold MySQL table. The dashboard included:

**KPIs:**

- Team Performance Overview
- Player Contributions
- Venue Insights

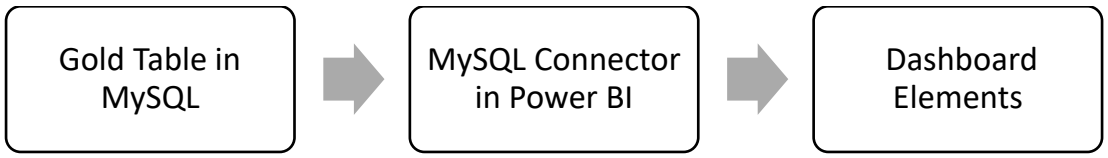


Figure 9. Dashboard Flow (Power BI)

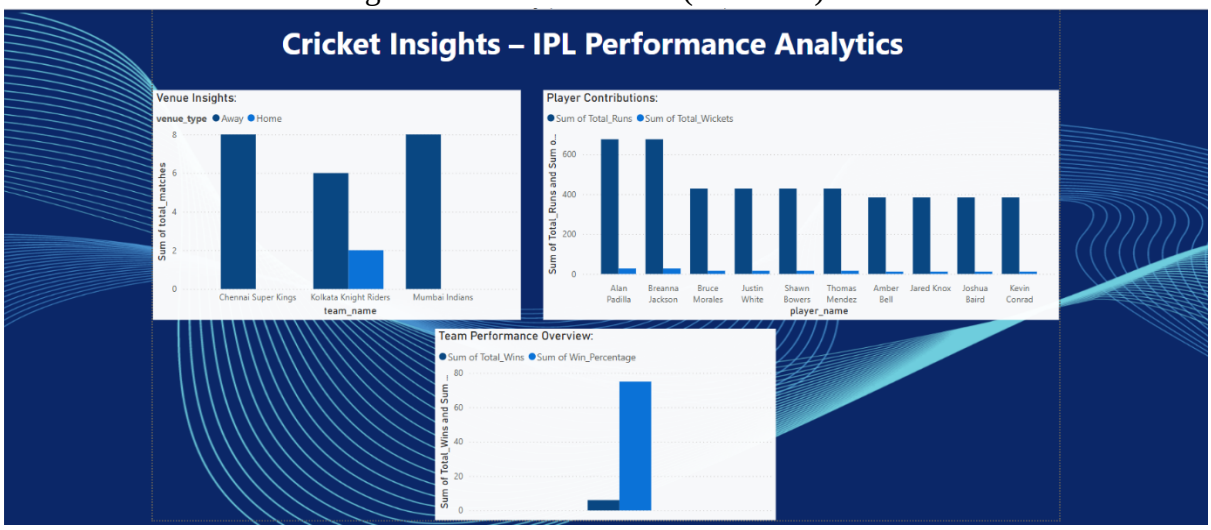


Figure 10. Final Dashboard

### 3.5 Automation using Azure Data Factory (ADF)

To ensure end-to-end automation, Azure Data Factory pipelines were created to orchestrate:

1. Trigger data ingestion from CSVs into Bronze.
2. Launch Databricks notebooks for silver and gold processing.
3. Load final gold data into MySQL.

This made the pipeline fully automated, requiring no manual intervention once deployed.


Figure 11. Overall Pipeline Overview

## 4 Git Repository

All source files, scripts, notebooks, and SQL queries used during the implementation of this project are stored and version-controlled in a Git repository. This ensures collaboration, reproducibility, and code backup throughout the project.
transportation-logistics-project/

Git Rep https://github.com/ShritejC/IPL_Data

## 5 Conclusion

The **IPL Gold Layer Analytics** project successfully demonstrates the implementation of a modern, scalable, and automated **end-to-end data pipeline** using the Medallion Architecture framework. By integrating tools such as **Azure Data Lake (Gen2)**, **Azure Databricks**, **SQL Server (SSMS)**, **Azure Data Factory**, and **Power BI**, this project showcases the seamless flow of data from raw ingestion to insightful visualization.
Key achievements of this project include:

- **Effective data ingestion** into the **Bronze layer**, ensuring secure and centralized storage.
- **Cleansing and transformation** in the **Silver layer** using PySpark and SQL, enabling structured and enriched datasets.
- **Aggregation and business metric generation** in the **Gold layer**, optimized for analytics and reporting.

- **Interactive Power BI dashboard** providing real-time insights into IPL team performance, top player contributions, and venue-based analytics.