IPLData_DataEngineering_Project_Pyspark_Databricks

IPL Data Analytics with AWS S3 + Databricks + PySpark

🏏 IPL Data Analytics using AWS S3, Databricks, and PySpark

This project demonstrates a complete end-to-end data engineering and analytics pipeline for analyzing **Indian Premier League (IPL)** cricket data. It utilizes cloud-native services and big data technologies to process, clean, and prepare rich cricket datasets for future analytical and machine learning use cases.

📁 Dataset Source

The dataset was downloaded from [Data.World – IPL Dataset and includes the following CSV files:

File Name	Description
`Ball_By_Ball.csv`	Ball-by-ball level data for each IPL match
`Match.csv`	Match-level details, including venue, results
`Player.csv`	Player demographic and skill information
`Player_match.csv`	Player-level statistics for each match
`Team.csv`	Team identifiers and names

☁️ Cloud Infrastructure

- **AWS S3**: Used as the storage layer to upload and host all raw CSV files.
- **Databricks**: Used as the compute and data processing environment.
- **PySpark**: Used for schema definitions, data ingestion, transformation, and exploration.

🔧 Technologies Used

- 🐍 Python
- 🔥 Apache Spark (PySpark)
- ☁️ AWS S3
- 🧠 Databricks Notebook
- 📊 Structured Streaming (optional for real-time extensions)

📌 Project Structure

IPL_Data_Analysis_Spark.ipynb Read.me Data
data-pipeline.pdfdata-pipeline.pdf