

IPL Dataset Analysis Using Google Cloud Platform and Power BI

BY

THENNAVAN.M

Project Overview

This project analyzes an IPL dataset to identify which players lost the most wickets at specific overs during matches. It leverages multiple Google Cloud Platform (GCP) services to process and analyze data in a scalable and efficient pipeline. Power BI is used to visualize the results for deeper insights.

- **Architecture**

- The project uses a cloud-based architecture to handle large datasets with real-time processing and analysis. The key components include:

1. **Google Cloud Storage:** Stores the raw IPL dataset.
2. **BigQuery:** Used for data querying and analysis.
3. **Pub/Sub:** Enables real-time data ingestion (if required for streaming scenarios).
4. **Dataflow:** Processes data in batches or streams, transforming it for analysis.
5. **Dataproc:** Manages and orchestrates any additional distributed data processing.
6. **Cloud Composer:** Orchestrates the entire pipeline to automate data flow from ingestion to analysis.
7. **Power BI:** Visualizes the analysis results for better understanding and decision-making.

Data Analysis Objectives

1. Identify players who lost the most wickets.
2. Determine the overs in which players most frequently lost wickets.
3. Visualize data to highlight patterns and insights.

Visualization

1. Use **Power BI** to connect to BigQuery.
2. Create visualizations, such as bar charts and tables, to show players who lost the most wickets per over.

Conclusion

- This project demonstrates the power of GCP's suite for handling large-scale data analytics, enabling IPL data analysis with real-time capabilities. Power BI provides an accessible platform for non-technical stakeholders to visualize and interpret the results.