HIGH LEVEL DOCUMENTATION -1

PROJECT-1

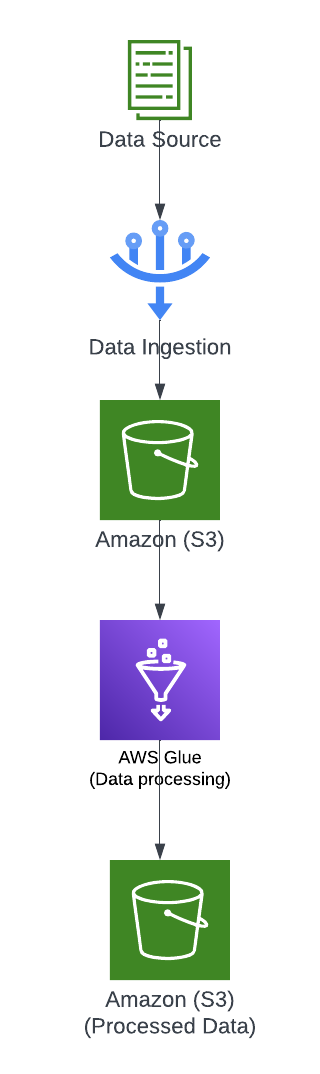
**Project Title: Load and Process IPL Match Data**

**AWS Services:**

* Amazon S3 for Data Storage
* AWS Glue for Data Processing

**Dataset Details:**

* Dataset includes ball-by-ball information for every match played between 2008 and 2022 seasons.
* Tables: Match Details (IPL\_Matches\_2008\_2022.csv), ball by ball details (IPL\_Ball\_by\_Ball\_2008\_2022)

**Infrastructure Setup Block Diagram:**

**Requirements:**

The project has multiple requirements related to IPL match data analysis. Here are the high-level descriptions of the requirements:

- Requirement No 1: Fastest Centuries

- Requirement No 2: Most Fours Innings

- Requirement No 3: Most Runs

- Requirement No 4: Most Wickets

- Requirement No 5: Most Runs Conceded In Innings

- Requirement No 6: Fastest Fifties

- Requirement No 7: Most Sixes Innings

**High-Level Description:**

1. Set up AWS Glue crawler to discover and catalogue the dataset files in the S3 bucket.
2. Create separate Glue ETL jobs for each requirement to extract relevant data from the dataset tables.
3. Perform data transformations and aggregations to calculate the desired statistics.
4. Sort the results based on the specified criteria (e.g., runs scored, balls faced, wickets taken).
5. Write the final results to separate CSV files for each season, following the specified naming convention.

## **Architectural Plan**

## The architecture follows a distributed and scalable approach to handle the IPL match data. It consists of the following key components:

**1. Data Loading Component**

* Identify and integrate with IPL match data sources, such as external APIs or data providers.
* Implement a data loading process to extract data from the sources and transform it into a suitable format for processing.
* Leverage AWS Glue jobs or custom scripts to retrieve and transform the data.
* Load the transformed data into Amazon S3 for storage and further processing.

**2. Data Storage Component**

* Utilize Amazon S3 as the primary data storage solution.
* Organize the IPL match data in structured formats like Parquet CSV, partitioned based on relevant attributes (e.g., season) to optimize data retrieval and processing.

**3. Data Processing Component**

* Utilize AWS Glue for data processing and transformation.
* Develop AWS Glue jobs or ETL scripts to apply transformations, data cleansing, filtering, aggregation, or enrichment as required.
* Implement data quality checks to ensure the accuracy, completeness, and consistency of the processed data.

## **Transactions**

**Transactions**: The different types of transactions or operations that users can perform within the system. This can include:

* **Data Loading:** This is the transaction flow for loading IPL match data into the system. This involves steps such as data source integration, data extraction, and loading the data into Amazon S3.
* **Data Processing:** This is the transaction flow for processing the IPL match data. This involves steps such as transformation and ETL processes, and data quality checks.
* **Data Retrieval:** This is the transaction flow for users to retrieve or access the processed IPL match data. This involves storing the processed data back into S3.

# **Out-of-scope**

# The project does not involve the development of a user interface or visualization components. The focus is on backend data storage and processing, and any frontend or user interface-related development is considered out of scope.

# The project is limited to IPL match data storage and processing. Integrating external data sources beyond the scope of IPL matches, such as weather data or player social media feeds, is considered out of scope.

**High-Level Workflow:**