**High-Level Design**

**Project 1**

**Load and Process IPL match data**

**Objective:**

The project aims to analyse the Indian Premier League (IPL) cricket dataset and extract valuable insights and statistics related to player performances across multiple seasons.

**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)

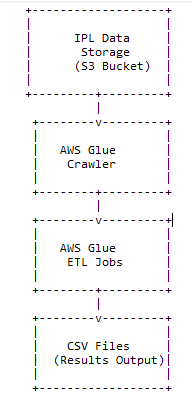
**Context:**

The IPL is a professional men's Twenty20 cricket league contested by ten teams based out of ten Indian cities. It is the most-attended cricket league globally and has significant economic and viewership impact.

**Infrastructure Requirements:**

* S3: Data storage
* Glue: Data processing and transformation

**Infrastructure Setup Block Diagram:**

****

**4. 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

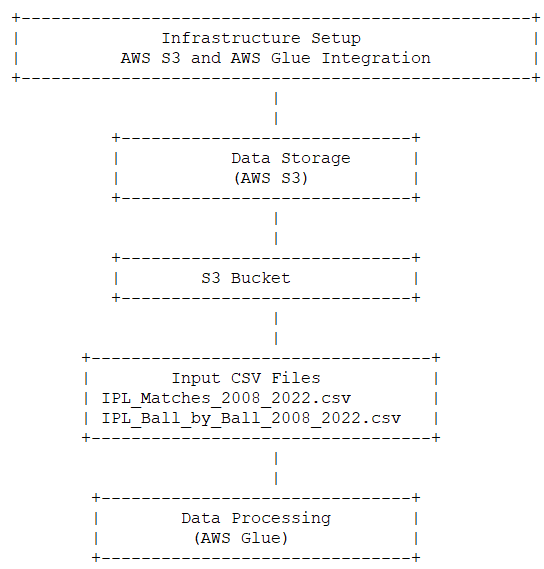
- Requirement No 4: Most Wickets

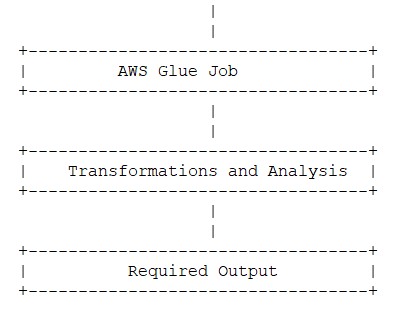
- Requirement No 5: Most Runs Conceded In Innings

- Requirement No 6: Fastest Fifties

- Requirement No 7: Most Sixes Innings

**High Level Workflow:**





**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.

**Benefits:**

* Efficient data processing and transformation using AWS Glue.
* Scalability and handling of large datasets.
* Durability, availability, and scalability of data storage using S3.

The IPL Data Processing Project enables the extraction of valuable insights and records from the IPL dataset, providing stakeholders with a deeper understanding of player performances and facilitating further analysis and presentation.