**LOW LEVEL DESIGN**

**Project 1**

**Load and Process IPL match data**

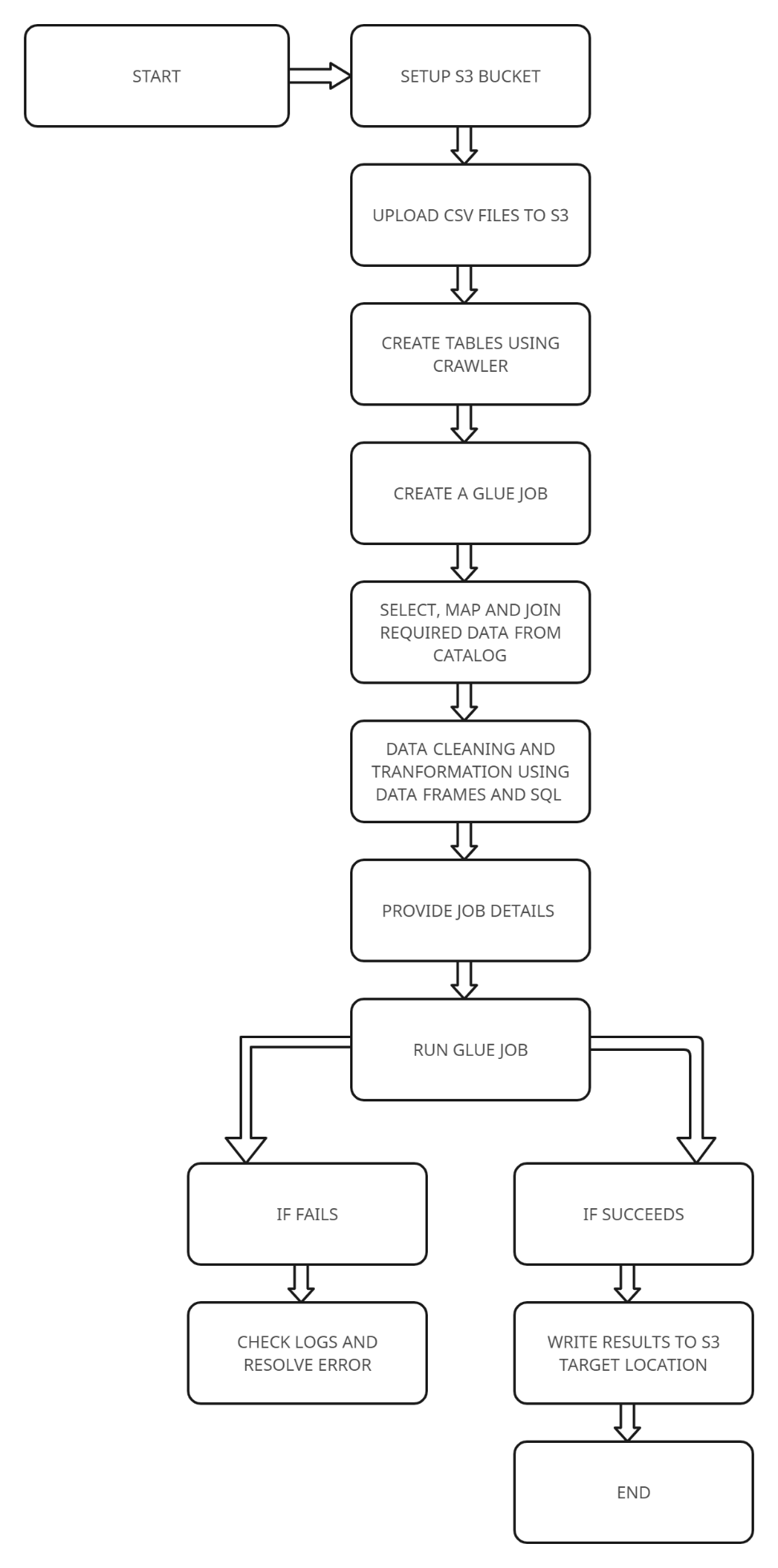
**Introduction:**

**This document serves as a vital resource encompassing a comprehensive understanding and workflow of Project 1. Its primary objective is to provide you with the necessary knowledge and guidelines to effectively implement and execute this project. From elucidating the project's requirements and sources to defining the key targets, this documentation acts as your compass throughout the project journey.**

**Data Ingestion:**

* Set up AWS S3 as the data storage platform.
* Configure data ingestion pipelines to receive and store the IPL match data files in the S3 buckets.
* Implement error handling and data validation mechanisms to ensure data integrity during ingestion.

**Project Flow Block Diagram:**



**Data Processing:**

* Utilize AWS Glue for data processing and transformation.
* Create Glue jobs to extract data from the ingested files and perform necessary transformations based on the specific requirements.
* Use Glue Dynamic Frames to handle semi-structured data and apply schema transformations.
* Leverage Glue Crawlers to automatically discover the schema of the ingested data files and create corresponding metadata.

**Season-wise Data Partitioning:**

* Partition the data based on the season to improve query performance and enable efficient data retrieval.
* Implement a partitioning strategy to store the processed data in S3, organizing it by season.

**REQUIREMENTS:**

1. **Fastest Centuries:**

* Develop a Glue job to identify the fastest centuries scored by players in each season.
* Apply the required filters and sorting logic to determine the players with the fastest centuries.
* Generate separate CSV files for each season, listing the players and their respective centuries in ascending order of the number of balls required.

1. **Most Fours Innings:**

* Create a Glue job to identify the players with the highest number of fours in a single inning for each season.
* Implement the necessary logic to rank the players based on the number of fours.
* Generate separate CSV files for each season, listing the players and their respective innings with the most fours, in descending order.

1. **Most Runs:**

* Develop a Glue job to calculate and compile the statistics of players with the most runs in each season.
* Implement the required calculations, such as total runs, highest score, average, strike rate, etc.
* Generate separate CSV files for each season, listing the players and their corresponding statistics, in descending order of total runs.

1. **Most Wickets:**

* Create a Glue job to calculate and compile the statistics of bowlers with the most wickets in each season.
* Implement the necessary calculations, such as total wickets, economy rate, 4-wicket hauls, 5-wicket hauls, etc.
* Generate separate CSV files for each season, listing the bowlers and their corresponding statistics, in descending order of total wickets.

1. **Most Runs Conceded In Innings:**

* Develop a Glue job to identify the bowlers who conceded the most runs in a single inning for each season.
* Implement the logic to rank the bowlers based on the runs conceded and the number of wickets taken.
* Generate separate CSV files for each season, listing the bowlers and their respective innings with the most runs conceded, in descending order.

1. **Fastest Fifties:**

* Create a Glue job to identify the players who scored the fastest fifties in each season.
* Apply the required filters and sorting logic to determine the players with the fastest fifties.
* Generate separate CSV files for each season, listing the players and their respective fifties in ascending order of the number of balls required.

1. **Most Sixes Innings:**

* Develop a Glue job to identify the players who hit the most sixes in a single inning for each season.
* Implement the necessary logic to rank the players based on the number of sixes hit.
* Generate separate CSV files for each season, listing the players and their respective innings with the most sixes, in descending order.

**Data Export:**

* Store the processed data as separate CSV files in S3 for each analysis.
* Set appropriate access controls and permissions for the generated CSV files.
* Ensure data consistency and accuracy during the export process.

This LLD provides a detailed plan for implementing the IPL Data Processing Project, outlining the specific steps involved in data ingestion, processing, partitioning, and analysis. It incorporates the use of AWS Glue for data processing and transformation tasks.