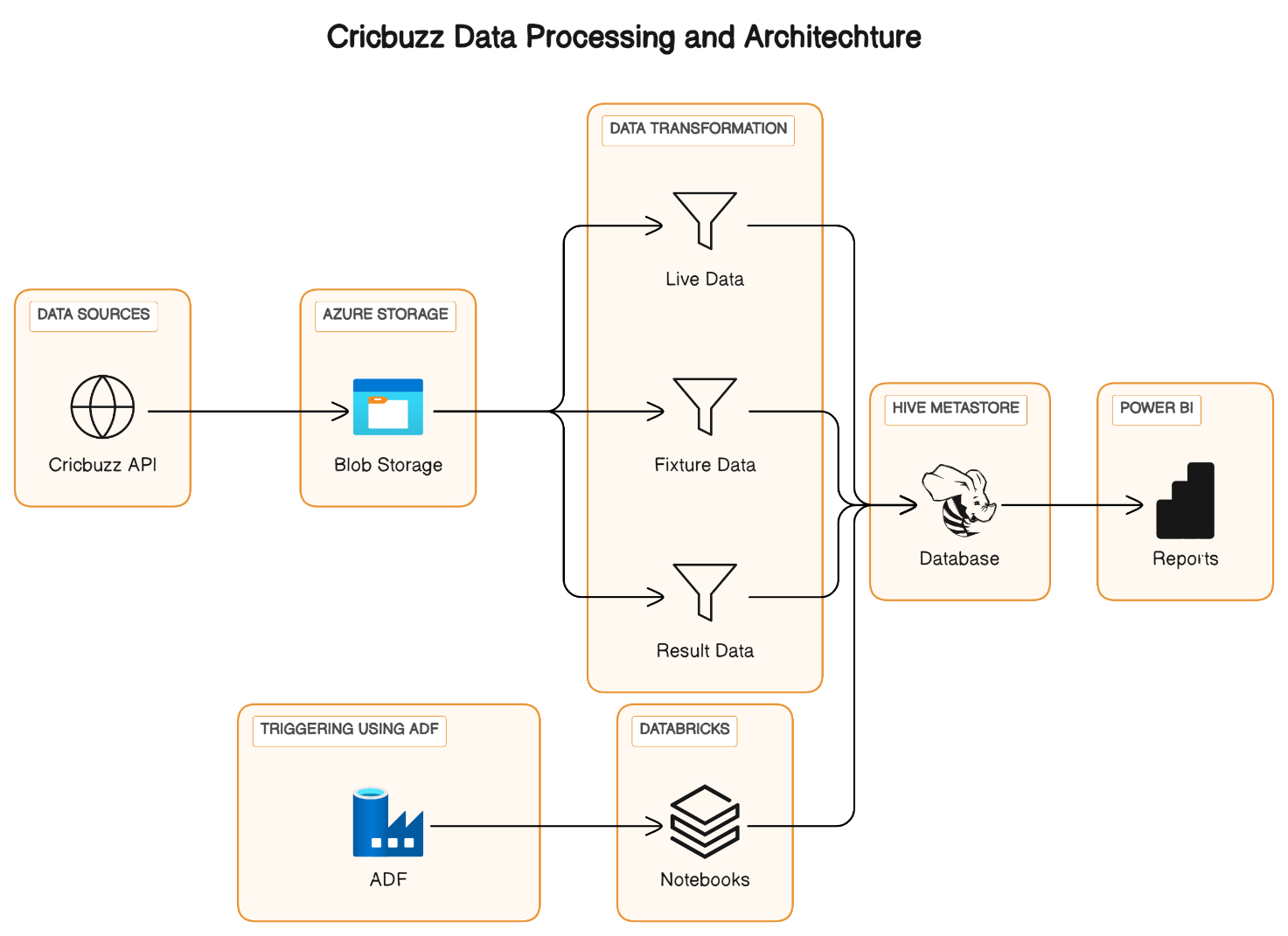
**Data Engineering Project**

**Problem Statement**

* The ultimate goal is to turn this semi-structured JSON data into a valuable asset that can be easily analyzed and understood by data scientists, analysts, and decision-makers.
* To optimize our data cleaning process to make it more efficient, reducing the time and resources required to prepare the data for analysis.
* Fetching the latest JSON data from API and saving as a JSON file in a blob storage for analysis in the future.

****

**Stage 1:** Fetch Cricbuzz Data from APIs and save it in Azure Blob Storage.

Click on this link to see the API Data 🡪 [API Data link](https://cricketdata.org/member-test.aspx#cricScore) (need to register to see this data if you are a new user)

**Notebook 1 Code** --> [Click on this](https://adb-6774691551246281.1.azuredatabricks.net/?o=6774691551246281#notebook/2413873811504032/command/2157360730140533)

**Stage 2:** Read Data directly from Azure Blob Storage.

**Notebook 2 Code** --> [Click on this](https://adb-6774691551246281.1.azuredatabricks.net/?o=6774691551246281#notebook/1188308785938895/command/1188308785938896)

Azure DBFS is a platform which we use for performing ETL process, like reading the file, flattening, and saving the flattened data into Hive storage as tables.

**Stage 3:**

**Transformation:**

Separate Live, Fixture, and Result data, and rename columns as needed. (Refer Notebook 2)

**Stage 4:**

**Create Hive Metastore**: Establish a database and store the **Live, Fixture, and Result data** into respective tables.

**Stage 5:**

**Scheduled Notebook to run every 3 hours to get latest data:**

Use Azure Data Factory (ADF) to trigger data refresh every 3 hours by calling Databricks notebooks.

**Stage 6:** Generating Power-BI reports.

First, establish a connection to Azure Databricks with Power BI through direct query. (Refer to this link for establishing connection --> [Click on this link](https://learn.microsoft.com/en-us/azure/databricks/partners/bi/power-bi))

Perform a Data Modelling in Power-BI to establish an Entity Relationship between Multiple tables using **one-to-many/many-to-one** etc.**,** relationship based on the requirement.

Start building Reports.

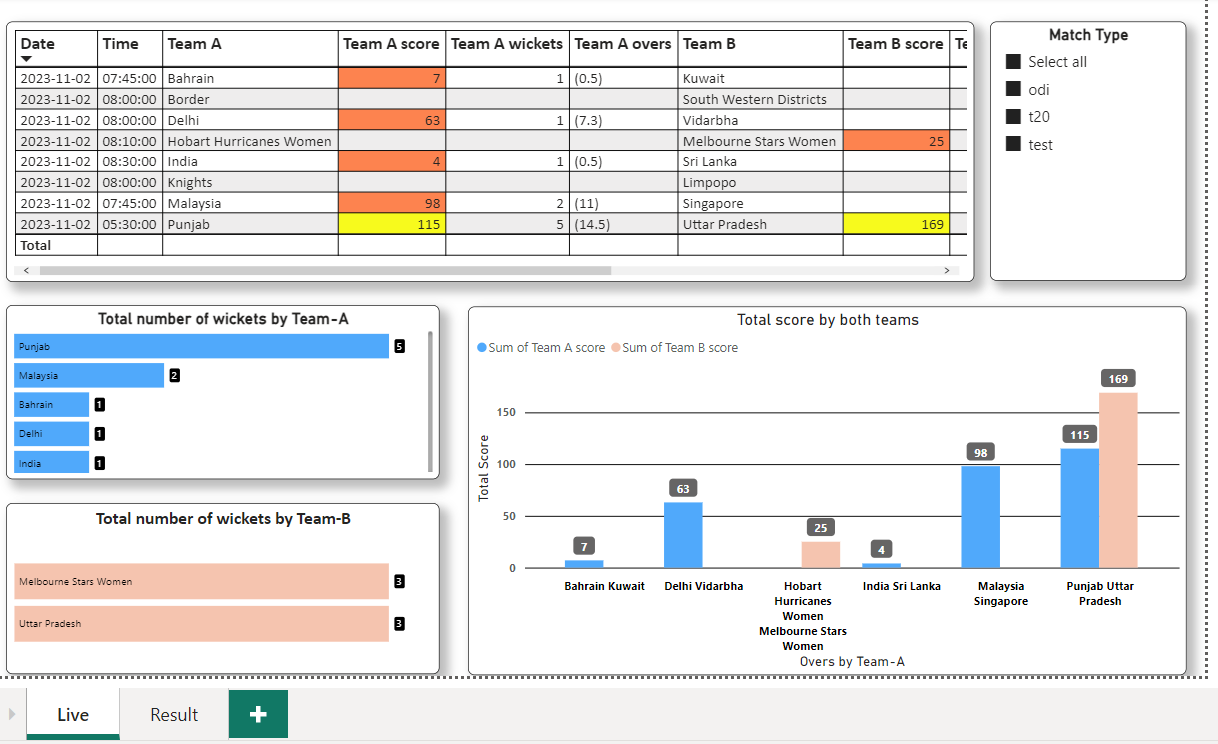
**Stage 7:** Automated Data Refresh.

Use gateways to enable automatic data refresh in Power BI, ensuring up-to-date information in the published report.

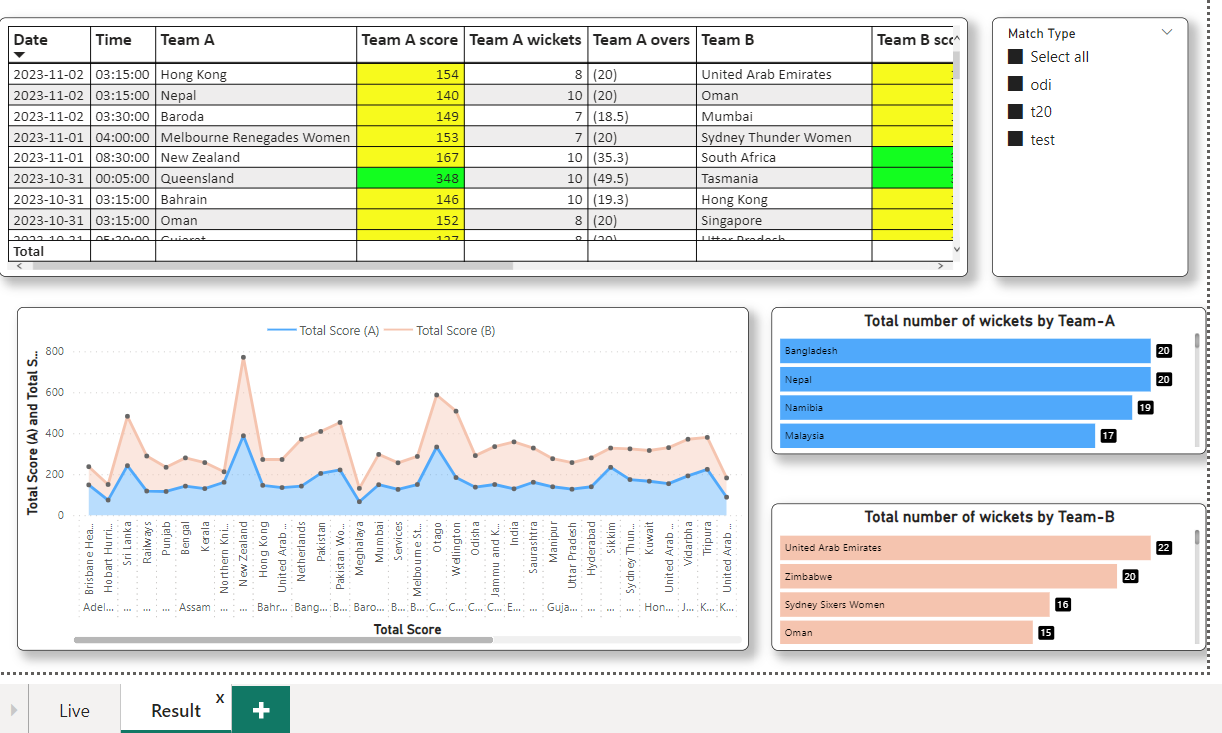
Refer this Microsoft Documentation for Gateway connectivity 🡪 [Click on this link](https://learn.microsoft.com/en-us/power-bi/connect-data/service-gateway-onprem)

**Power BI Report**

**Live Data:**

****

**Completed Matches Data:**

****