Data Model for Energy Use case:

Table : public_power_data - ingested daily to get last 24 hrs data
Columns :
|-- production_type: string (partitioned column)
|-- net_power_produced: float
|-- timestamp: timestamp

Table: Price - - ingested daily to get last 24 hrs data
Columns:
|-- price: float
|-- timestamp: timestamp

Improvement for future when needed (not implemented) :
We can have a date field created from timestamp and partition table by date for query performance.

Table: installed_power_data – ingested monthly
Columns:
|-- date: date (partitioned column)
|-- production_type: string
|-- installed_power: float

- Time stamp field can be used while querying to get relevant info such as Date, Month etc.
- Dataframes can be joined on the fly to answer Bi queries e.g to get insights about price and install power.

Tech Stack used for local run :
- Data Processing – pyspark , notebooks
- Data storage - Delta format
- BI queries – Spark sql and notebooks (in actual env we can use databricks sql database)
- Logging – Notebook output saved
- ETL – shell script to run notebooks (in actual env we can use airflow , databricks workflows etc.)

Cloud Implementation and Architecture:

We will follow a layered data design pattern which is also know as medallion architecture. By categorizing the data into Bronze, Silver and Gold, the architecture assures streamlined data processing, improved data quality and serves variety of use cases due to its flexible data access methods. This is a type of lakehouse architecture for modern applications.

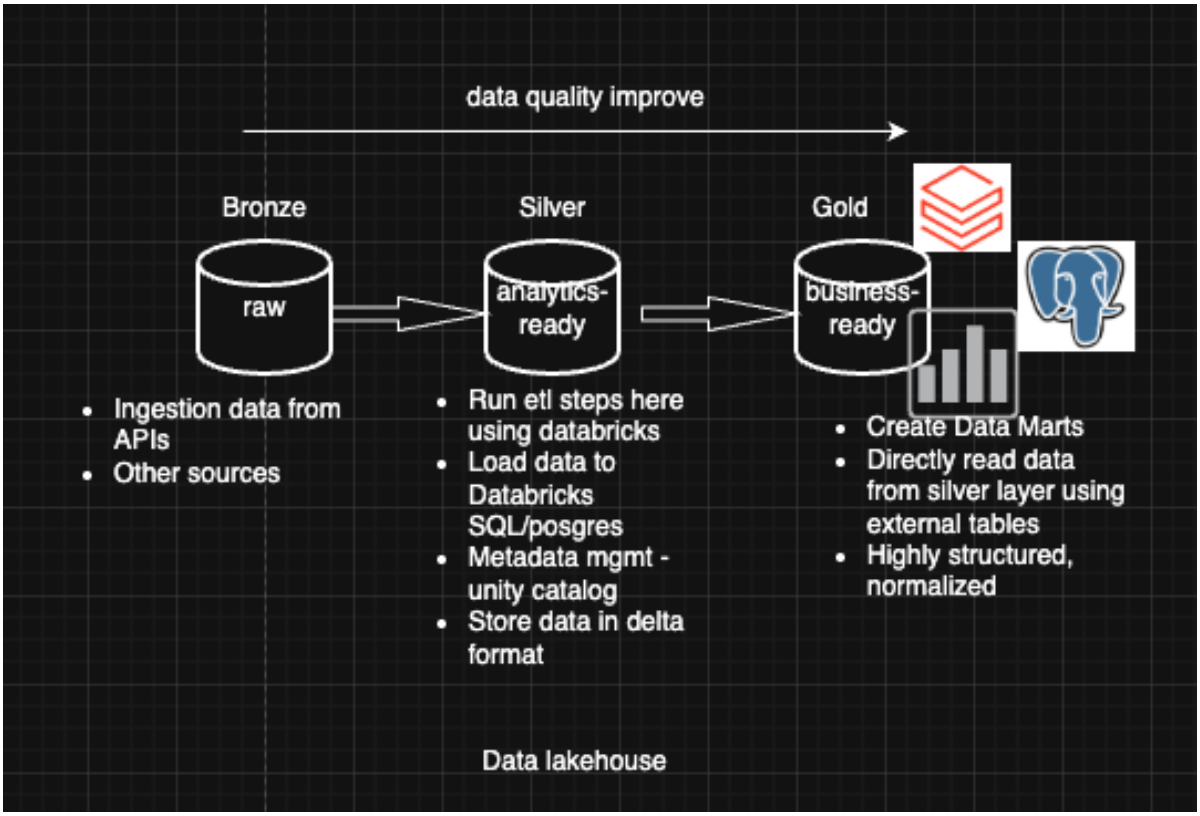
I have noted what each layer means and tech stack to be used

Bronze layer : Data from various sources (structured, unstructured etc.) stored without any modifications to object storage. Any innovation driven initiatives within organization can refer to these datasets.
Data formats – csv, json, parquet etc.

Silver Layer: This layer involves actual preprocessing or data transformation to get a clean/enriched version of data, ready for downstream analytics use cases – e.g. data warehouse application, meta data management.
Data format – mostly column-oriented format e.g parquet/delta

Gold layer: This layer holds the highly curated, business-ready data. It is optimized for reporting, analytics, and business intelligence purposes.
Data format – highly structured, normalized

Tech Stack:
Cloud provider – Azure, Databricks
Storage – ADLS (in bronze layer, silver layer)
Data transformation – Databricks, Unity catalog,
Datawarehouse – Databricks SQL /Postgres
Reporting – PowerBi



Please note :

For this usecase in particular we may not need a bronze layer as we are directly getting data from api and processing it. However, it is recommended to have it a production grade env. Benefits –
-Keeps track of data lineage
-If Api changes, we still have the data
-Reprocess Api data without hitting the API (reduce Api calls)