Pipeline for Turbine Data Ingestion and Analytics

Prerequisites

- ### Software Dependencies

- Apache Spark: Version 3.3.0 or higher
- Apache Kafka: Version 3.4.0 or higher
- Docker
- PostgreSQL: Version 42.2.0 or higher

- ### Data Sources

- Csv File

Workflow

- ### Data Producer

- The Producer program is responsible for reading the raw data from CSV files and streaming the data to Kafka.
- Spark is used to create a streaming Data Frame from the CSV files located at data path.
- The Data Frame is transformed to convert the data to JSON format and select the required columns.
- The Kafka configuration specifies the Kafka bootstrap servers and the checkpoint location for fault-tolerance.

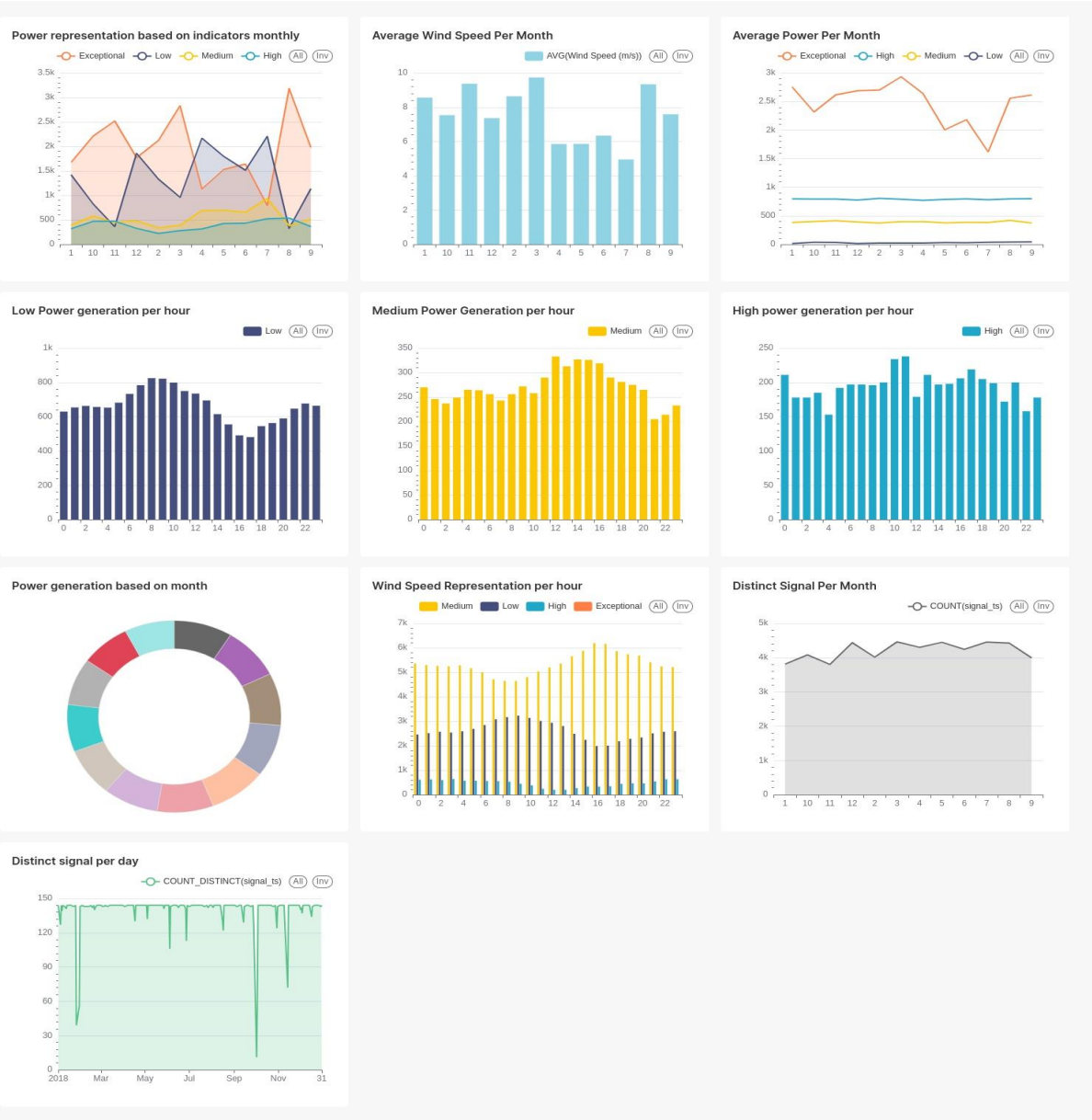
- ### Data Subscriber

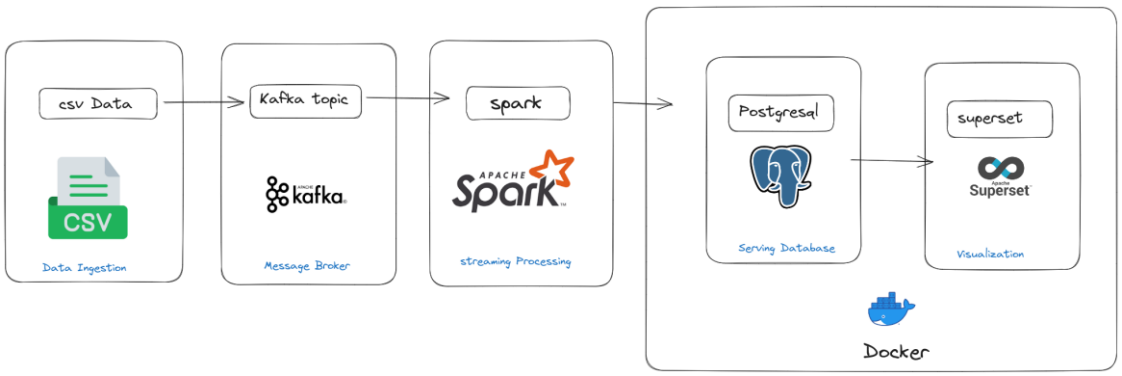
- The Subscriber program acts as a consumer of the data streamed from Kafka and performs further transformations.
- The program reads the data from Kafka using the topic.
- The data is parsed from JSON format and the required columns are selected.
- Additional transformations are applied to clean and validate the data.
- The data is transformed into the desired format, including the addition of derived columns.
- The transformed data is written to a PostgreSQL database using JDBC, with the specified connection properties and table name.

- ## Data Visualization (Superset)

- Superset is a separate tool used for data visualization and exploration.
- The data stored in the PostgreSQL table can be accessed and visualized using Superset.
- Connect Superset to the PostgreSQL database using the provided connection details.
- Configure Superset to create dashboards, charts, and reports based on the turbine data, allowing users to gain insights and perform analytics.

Screenshots





Pipeline Architecture