

Data Flow Documentation

1. Producer Databricks which is in the Producer pipeline takes the csv file from the Azure blob and sends messages to Kafka. Producer pipeline has storage event trigger which is triggered whenever there is a change in the storage.
2. Consumer Databricks which is in the Consumer pipeline consumes the messages from Kafka. These data are cleaned and stored in Azure Data Lake. Consumer pipeline has schedule trigger which is triggered in specified interval of time.
3. The cleaned data is loaded now to SQL database. Then a dataflow is setup in Power BI to load the data from SQL database in specified interval of time. Machine learning joblib model is created in Python and loaded to Power BI.
4. The dataset in Power BI is used to create visualizations, reports and dashboards.

