

## Overview of Data Platform Process: Ordered Steps (for detailed instructions, see ETL Report)

1. Gather desired dataset CSVs on local machine.
2. Upload uncleaned CSVs to a designated Azure Data Lake.
3. Read unclean data CSVs into a Databrick.
4. Perform cleaning in PySpark and Pandas.
5. Write cleaned datasets to designated Azure Blob Storage.
  - a. Utilize one dataset for Kafka streaming via producer and consumer.
  - b. Load streamed dataset and non-streamed datasets into SQL Table Schema.
  - c. Transmit cleaned machine learning data to Python script.
6. Connect SQL tables and Python script to PowerBI for visualization.
7. Perform written analysis on visualizations.

**Data Platform Diagram**

