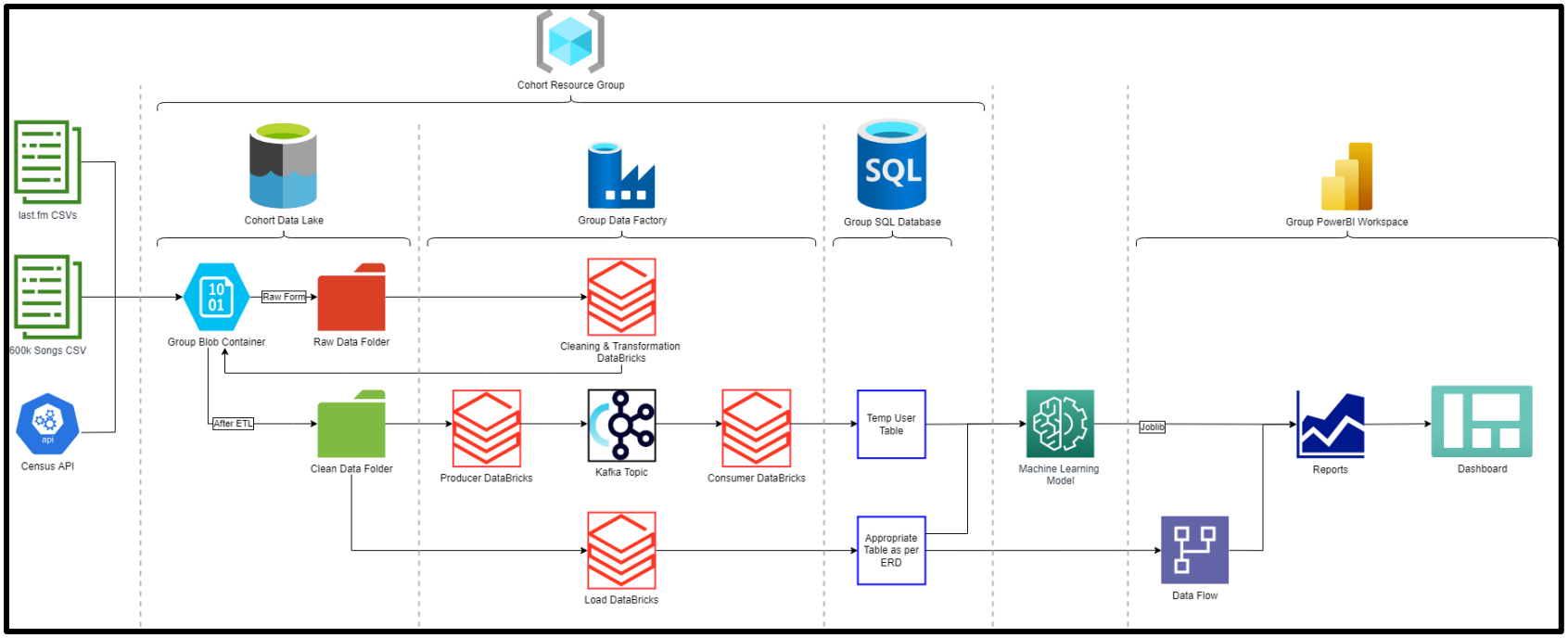
Data Platform Diagram – Olivier Rochaix, Eduard Stalmakov, Vanessa Gleason, and Alistair Marsden



From left to right:

Data sets stored or accessible in CSV format are indicated by this icon. Our group is using two such data sets; the 600k songs data set being one large CSV file and the last.fm data set being a collection of multiple CSVs.

Data that will be acquired via querying an API is indicated by this icon. Our group is using the Census API in order to collect data on audio streaming services' user demographics.

Data Platform Diagram – Olivier Rochaix, Eduard Stalmakov, Vanessa Gleason, and Alistair Marsden

This icon represents an Azure Resource Group. This is used as a visual indicator of where certain processes are taking place or technologies are being used.

This icon represents an Azure Data Lake. This is used as a visual indicator of where certain processes are taking place or technologies are being used. Found within the Resource Group.

This icon represents an Azure Blob Container. Our group is using this Blob Container to store raw and clean data in separate folders. Using an Azure Blob Container allows our group to access this data from multiple machines and automate the use of that data. Stored within the Data Lake.

These icons represent the folders into which we store raw and cleaned data sets. Raw data is represented by the red folder, and cleaned data is represented by the green folder. Stored within the Blob Container.

The group's Azure Data Factory. This is used as a visual indicator of where certain processes are taking place or technologies are being used. Found in the Resource Group.

Azure DataBricks. These DataBricks contain code that can be used for a variety of tasks; in our project we use them for the ETL process and create Kafka producers and consumers. Used in data pipelines within the Data Factory.

Data Platform Diagram – Olivier Rochaix, Eduard Stalmakov, Vanessa Gleason, and Alistair Marsden

The Confluent cloud. This stores the topic with which we produce and consume data from in order to simulate a live stream of data.

The group's SQL database. This is where a record of the simulated live stream data will be stored as well as where the group stores a normalized form of our cleaned data sets.

Tables within the SQL database. Stores values from our data sets and values consumed through Kafka.

The machine learning model(s). Our group is using Machine Learning to create a recommender system for music depending on a user's past music listening history. Feeds into PowerBI reports through JobLib.

PowerBI. Our group plans to use PowerBI in order to create visualizations and our dashboard.

Data Platform Diagram – Olivier Rochaix, Eduard Stalmakov, Vanessa Gleason, and Alistair Marsden

A PowerBI data flow. Used to synch the data in our group's SQL database with the group's PowerBI workstation. This will also ensure that data used for visualizations is always up to date.

The PowerBI reports. Our group will create reports and use them to create meaningful visualizations that can then be pinned to our Dashboard.

The PowerBI dashboard. The group's final product. This will be a mockup of the recommender system, imitating a user's playlist. It will also have additional pages detailing the modelling process and audio streaming user demographics.