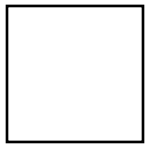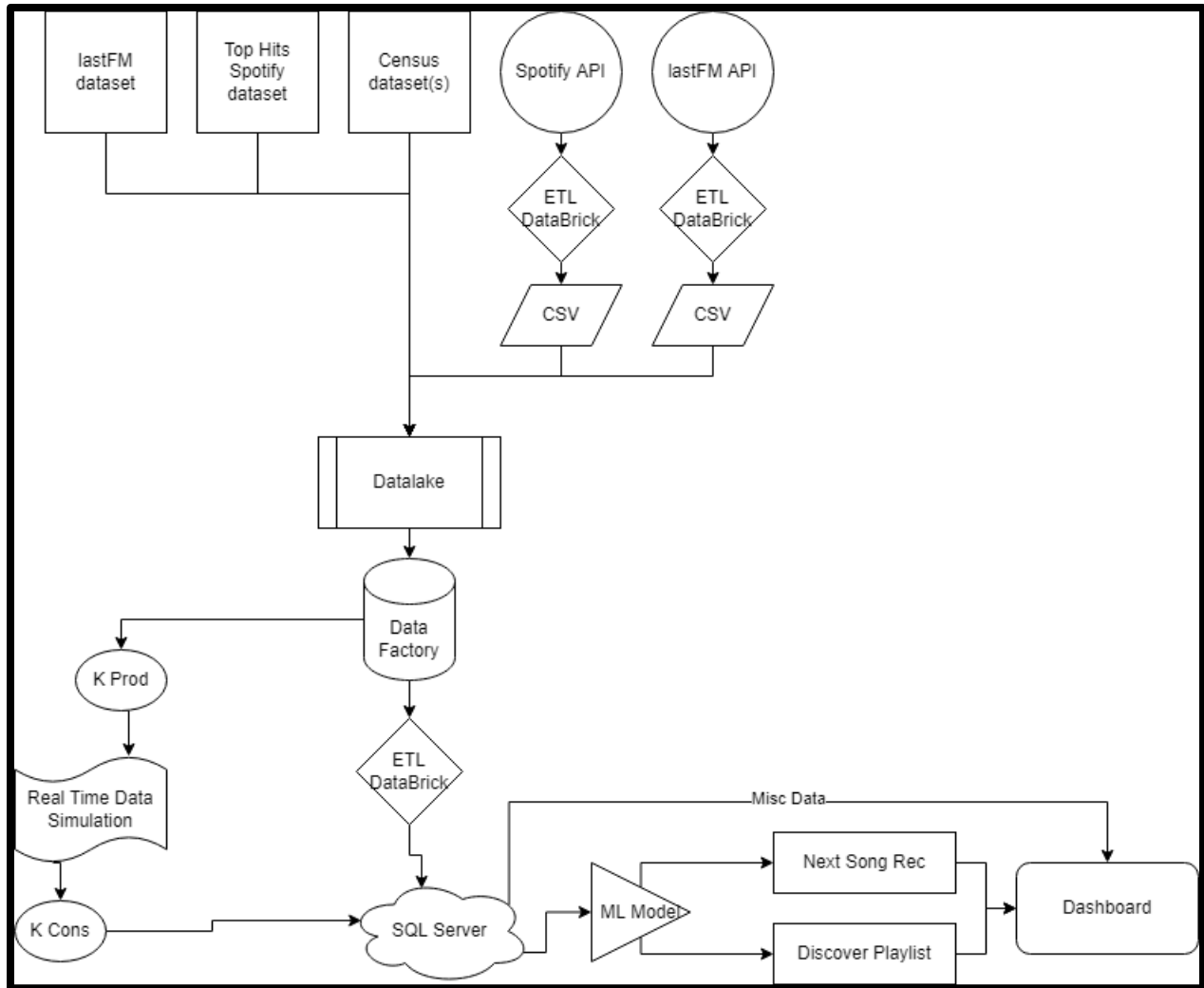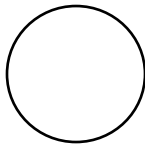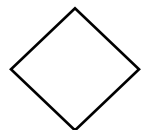# Data Platform Diagram Documentation

[Square]: Represents datasets already in useable formats. These datasets still need to be processed through ETL but can be used with minimal work.

[Circle]: Represent APIs that can be queried to create additional datasets or find supportive data for our main datasets.
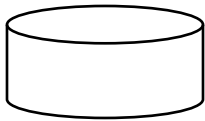
[Diamond]: Represents Azure DataBricks used by the group to move or perform ETL on data. These allow for easier group work and can be used in automation.
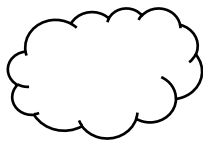
[Parallelogram]: Represents cleaned CSVs ready to be stored onto our groups Azure Blob Container.
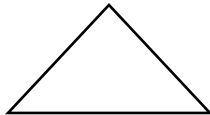
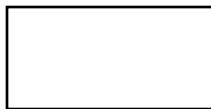[Bordered Rectangle]: Represents an Azure Datalake that contains the group's Blob Container.

[Data Factory]: Represents the group's Azure Data Factory. This contains multiple pipelines and enables automation.

[Cloud]: Represents the group's SQL Database. This is hosted on the cloud off of the cohort's SQL server.
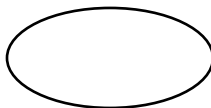
[Triangle]: Represents the group's Machine Learning model. This is planned to be a recommender model.
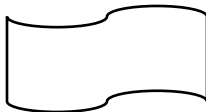
[Long Rectangle]: Represents resulting visualizations / tables from the ML model. Accomplishing these is the group's primary objective.

[Soft-cornered Rectangle]: Represents the group's Power BI dashboard. This is the group's intended final product.

[Oval]: Represents Azure DataBricks intended for use in our groups Kafka producers and consumers.

[Flag]: Represents the group's simulated flow of data using the Confluent cloud. In the context of the project, this is simulating a control group user's music listening history.