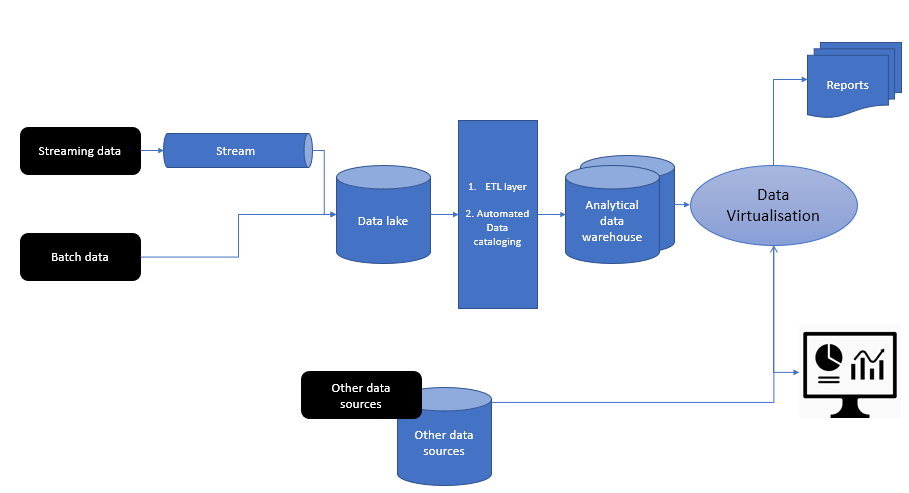
# High level architecture of Data processing system to handle both batch and streaming data



# Data ingestion

* Data ingestion for both Batch and Streaming data into a Data Lake, which is meant to house all data before cleansing. One example would be Hadoop.
* Additional pipe for streaming data to be used. Some of the popular libraries for handling streaming include Apache-Kafka / Spark-Streaming.

# Data Processing and Cleansing

* Data ingested to the data lake is considered ‘unprocessed’.
* There will be an ETL layer to Transform the data into usable format.
* During this transformation, the data can be catalogued as well, with the data transformations captured. Some popular enterprise tools include Informatica. Informatica Big Data Management (BDM) provides support for ETL jobs and provides a tool for Data Catalog and management, with the entire linage of the ETL captured.

# Analytical Data Warehouse

* Processed and cleaned data is stored in the one of more Data Warehouses.
* Multiple Data Warehouses could be required for a number of reasons, including segregation by the purpose of the data, the consumers of the data, security, and also performance.
* The main purpose of these data warehouses would be to serve the reporting needs of the various consumers.

# Data Virtualization (optional)

* There may be a need for a data virtualization layer depending on the number of data sources.
* Data virtualization provides a connection to one or many data sources to provide a single location for building out reports and dashboards. One such tool would be Denodo.

# Reports and Dashboards

* Finally the data is presented to the organization in the form of reports and dashboards.
* To optimize performance, the heavy lifting and data crunching should have been done in the ETL layer to minimize the transformations required when dashboards are loaded.