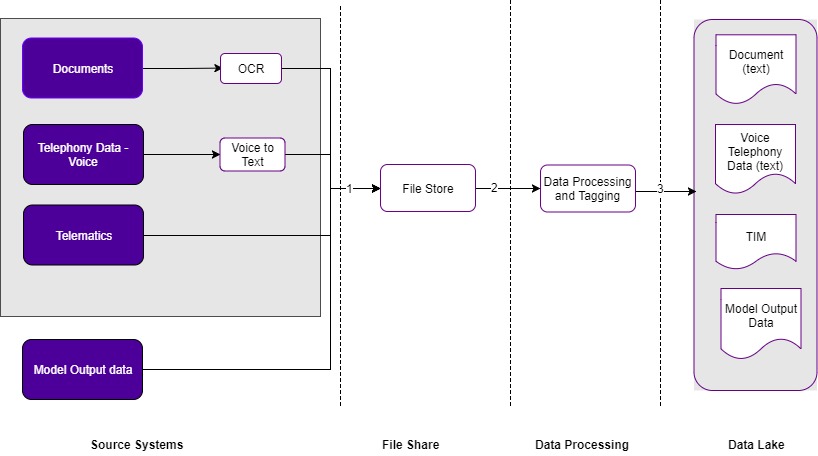
# Technical Description

This section describes how data will be loaded in the Data Lake from multiple sources, including structured and unstructured data sources.

Data will be ingested onto the data lake at a frequency to be determined based on business requirement An intermediate process will tag the metadata regarding the source and time of ingestion. Tagging will help in faster search and filtering of data.

**Data Lake Inbound Feed**

The following high-level steps will be performed to load the data into the Data Lake:

1. Ingest data from multiple sources via a job scheduler and put it into Storage Area Network.
2. The data from Storage Are network is ingested into the Landing Layer.
3. A data transformation process will be performed on the Landing Layer data to be used by downstream Systems. Intermediate processes in the data lake will include the following on a high level:

* Ensuring no interference amongst data while ingesting from multiple sources
* Validating the integrity of the data
* Tagging metadata
* Converting the data into a common file format
  1. Data Lake

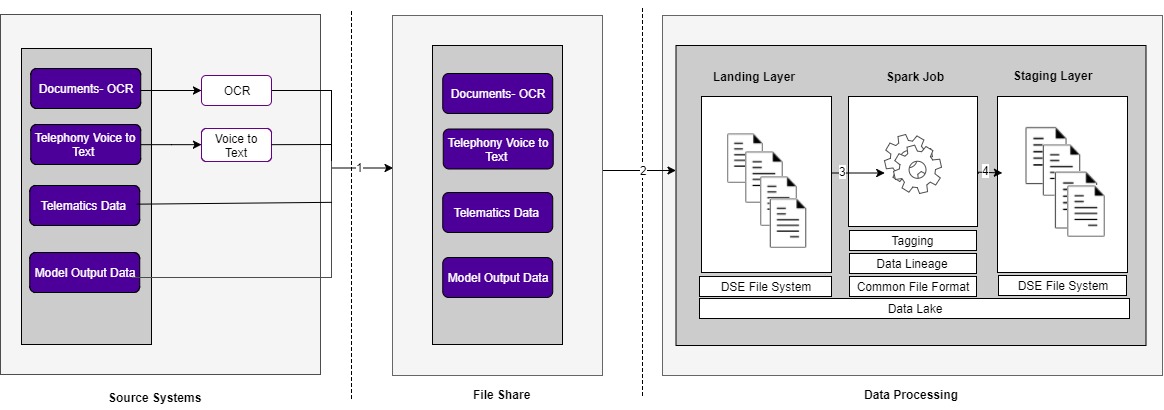
The types of data stored in the data lake are:

* Structured data
* Unstructured data( out of scope for Release 1)

The data sources that will be stored in the data lake are as follows:

1. Documents – Optical Character Recognition (OCR) data. Documents which have been converted to text and persist in the data lake.
2. Telephony data – Voice to text data. The telephony voice files which have been converted to text and persist in the data lake.
3. Telematics data – (TIM data) which is received from the Telematics service provider on a daily basis.
4. Model Output data. The outputs of the analytical models which will be used for model refinement.
5. Unstructured data such as video, audio, image and other sources of data that need to be fed into the Analytics Sandbox.

The first four types of data listed above will persist in the data lake are structured in nature. Additional validation on the data ingested into the data lake will be done to ensure data integrity.

* + 1. STRUCTURED Data

The key steps in the data load process to Data Lake are as follows:

Data from multiple data sources will be exported from the external source systems and transferred to \*\*\* and then \*\*\* will convert the data into the format required for processing. The outputs of this conversion process will be moved to a file share on the SAN for ingestion onto the data lake.

1. A DataStax file system API job will be invoked to ingest the raw data from the file share on the SAN to the landing area folder. Input data files are not changed during this process. Data is kept in a logical format that is one-to-one to the source system. Raw data is kept immutable, and it is not updated or deleted. Processed new incoming data is appended to the existing one.
2. A Spark ETL job will be executed to extract the files from Landing Layer to the Staging layer. Data successfully moved to the Staging Layer is removed from the Landing Layer.

While Data is retrieved from the landing area to the staging area, the Spark ETL process will convert the raw data file format into a format more suitable for data manipulation within the distributed data architecture .The expected common file format is Avro/ Parquet but will be decided during low level design. In the staging layer, data is kept in a physical format that is one-to-one to the originating source system data structures; therefore, modification of existing records is performed as appends. During this process additional metadata such as source and time of ingestion will also be captured to facilitate querying and retrieval of data.

Refer below to know how to add metadata to your Azure Data lake using python

<https://towardsdatascience.com/how-to-add-metadata-to-your-azure-data-lake-f8ec2022f50>

Waterline :

<http://docs4211.waterlinedata.com/azure-marketplace$WLD_newHDI>