upGrad

Data Ingestion with Apache Sqoop and Apache Flume - Session 1

upGrad

Course: DS - DE

Lecture On: Introduction to Data Ingestion

Instructor: Hitesh Hasija

Subject Matter Expert

Hitesh Hasija

Senior Data Engineer: Intuit

Skilled in softwares such as Hadoop, Hive, Sqoop, Spark, Kafka, Flume, Cassandra and MongoDB. I am working in the Big Data domain from last five years.

Segment - 01
Module Introduction

Module Introduction

Session 1

- What is data ingestion?
- Challenges faced in data ingestion
- Key steps in data ingestion
- Tools used for data ingestion
- Types of data and file formats

Session 2

- Introduction to Sqoop and its advantages and architecture
- Case study introduction
- Setup of Apache Sqoop and Database
- Sqoop export and import
- Various arguments of the Sqoop import command

Module Introduction

upGrad

Session 3

- Additional arguments and options of Apache Sqoop import commands
- Support of SQL queries in Sqoop
- Incremental import in Sqoop
- Sqoop Jobs
- Tuning Sqoop

Session 4

- Introduction to Apache Flume as well as its components and characteristics
- Case study and installation of Flume
- Flume Configuration files and Flume flows
- Tuning Flume and Sqoop vs Flume

upGrad

Segment - 02
Session Overview

Session Overview

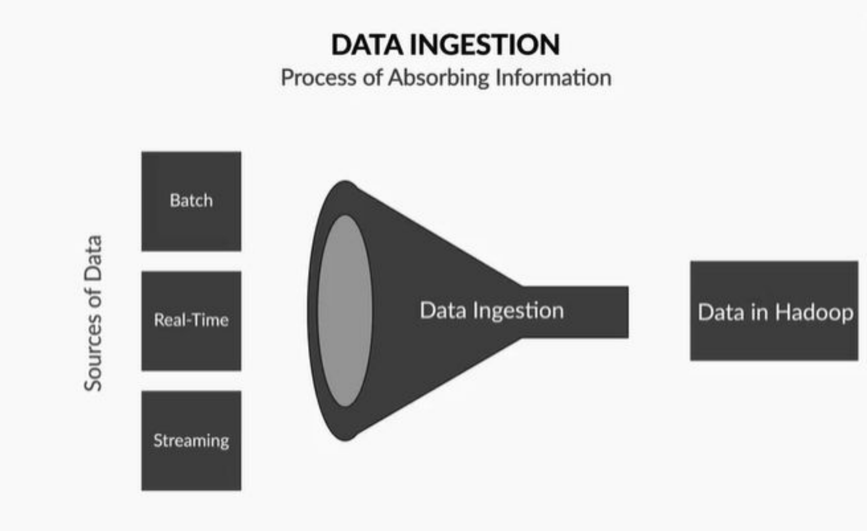
1. What is data ingestion and why it is needed?
2. Challenges faced while designing systems for data ingestion
3. Key steps of data ingestion
4. Tools used for data ingestion in the industry
5. Types of data and file formats dealt with in data ingestion

Segment - 03
What is Data Ingestion?

Learning Objectives

1. What is Data Ingestion?
2. Why is Data Ingestion needed?

What is Data Ingestion?



- How is data transferred to the systems in which it can be used in the first place?
- **Data ingestion** is **the process of absorbing data for immediate use or storage.**
- It acts as a bridge between the **source** and the **destination such as the Hadoop Distributed File System (HDFS)**, where it can be used efficiently.
- Data can be of one of the following types:
  - Batch
  - Real-time
  - Streaming

Segment Summary

1. Learnt in brief about the process of Data Ingestion

2. Learnt why Data Ingestion is important at the industry level

upGrad

Segment - 04
Challenges in Data Ingestion

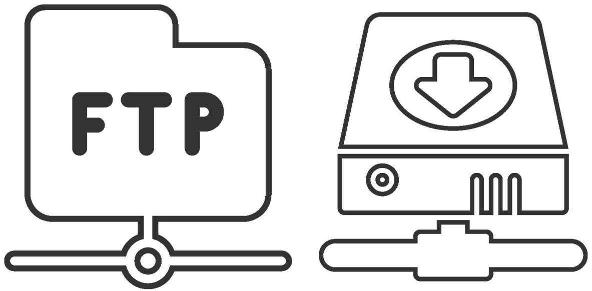
Learning Objectives

1. Challenges faced in the process of Data Ingestion
2. Massive growth of data in today’s era

Challenges in Data Ingestion

Challenges

Multiple Data Sources



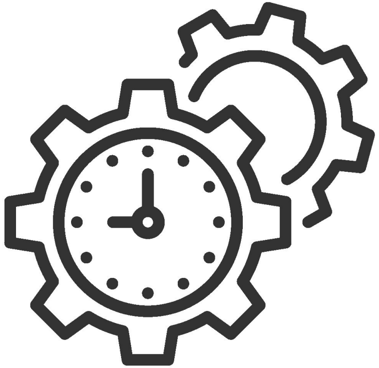
Numerous Data Types and File Formats of Data

HTML
EPS
PNG
RAR
PDF
PPT
CSV
W
DOC
TXT
JPG
ZIP
MOV
MPG
MP3
PSD
AI
XLS
RAW
<head>
<...>
</head>
CSS

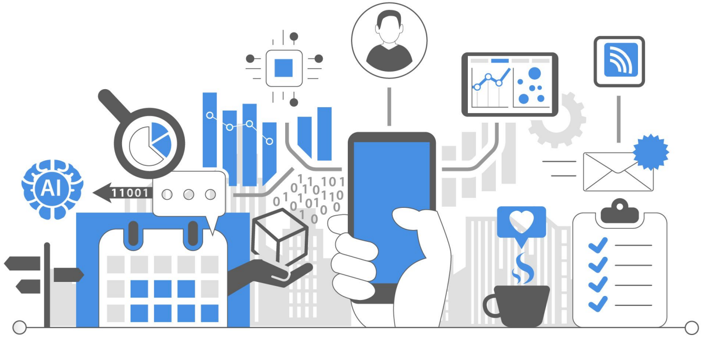
Challenges in Data Ingestion

Challenges

Processing Time



Data Generated at a High Rate and a Huge Scale



Challenges in Data Ingestion

Challenges

Network Performance

Network Security

Challenges in Data Ingestion

Facts about data generated worldwide

175 ZB
of total data created by 2025

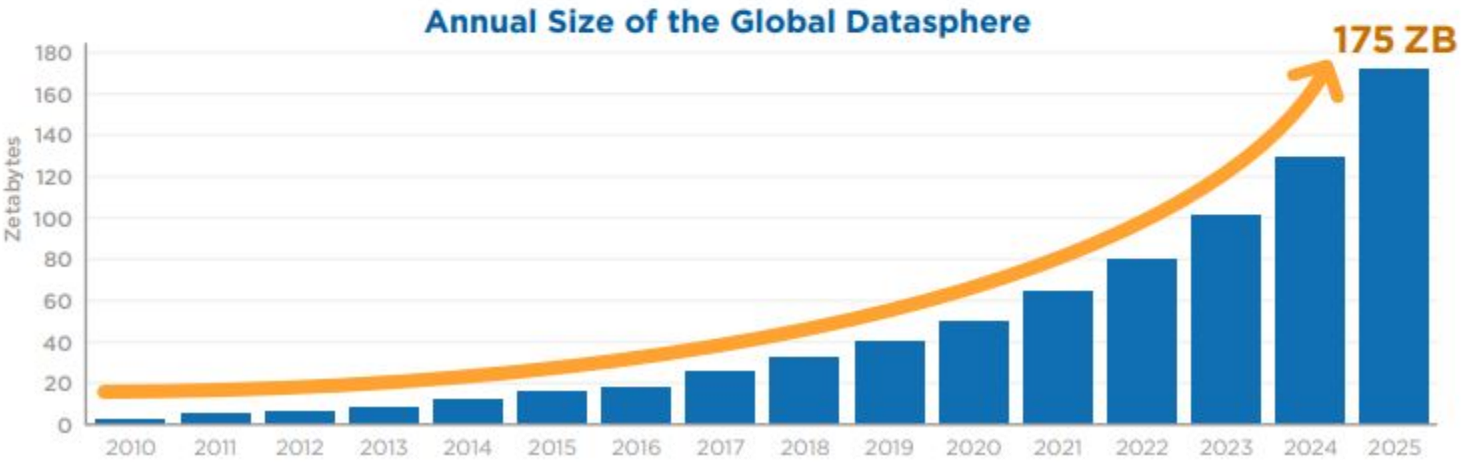
463 EB
of data created each day by 2025

200
million DVDs worth of data created each day by 2025

500 hrs
of video content uploaded to Youtube each minute today

Challenges in Data Ingestion

- Today, many tools for data ingestion such as Sqoop, Flume, Kafka and Gobblin have been developed that help in managing data ingestion tasks.
- The following infographic from the ‘Data Age 2025’ white paper shows the predicted data growth until 2025.



Source: Data Age 2025 sponsored by Seagate with data from the IDC Global DataSphere, November 2018

Segment Summary

1. Learnt about the different challenges faced in Data Ingestion

2. Learnt about the massive growth of data in today’s era

Segment - 05
Key Steps of Data Ingestion

Learning Objectives

1. Key steps of Data Ingestion

2. Demonstration of these steps using an example

Key Steps of Data Ingestion

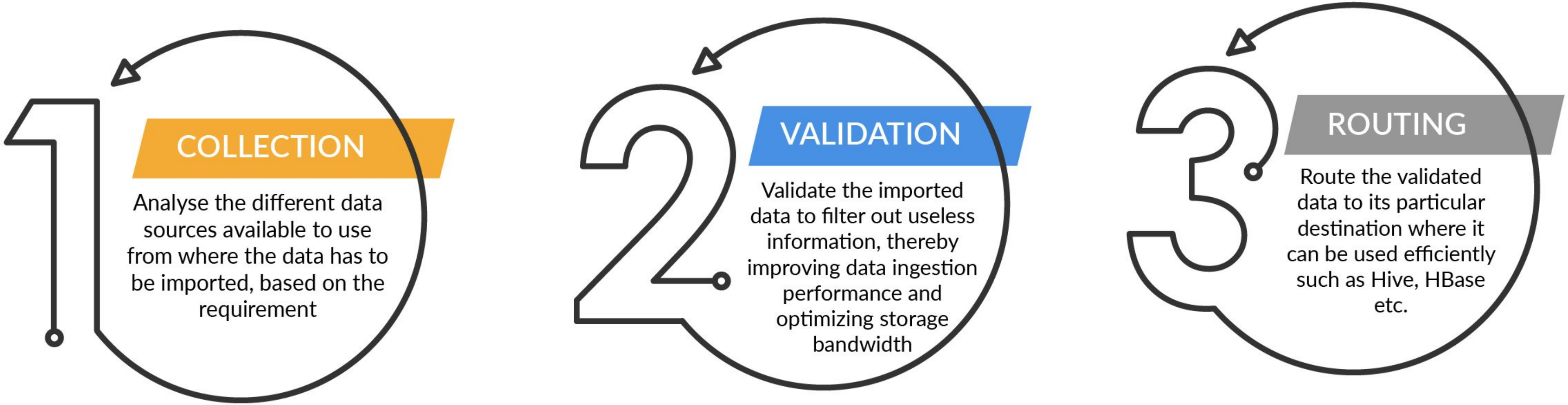
Main Steps Followed in Data Ingestion

Let’s understand how data ingestion takes place in a scenario in which you have to design a recommender system for Netflix:

- Which data sources to use?
  - User profile information
  - User browsing history
  - Survey data from emails and forms
- Which data is useless for our system?
  - Data could be duplicate.
  - Data, such as the user name and the user ID, could be useless.
- How does data reach its destination?
  - Many databases and other systems such as HBase, Hive and HDFS

Key Steps of Data Ingestion

Broadly, three main steps are being carried out, which are as follows:



1 COLLECTION

Analyse the different data sources available to use from where the data has to be imported, based on the requirement

2 VALIDATION

Validate the imported data to filter out useless information, thereby improving data ingestion performance and optimizing storage bandwidth

3 ROUTING

Route the validated data to its particular destination where it can be used efficiently such as Hive, HBase etc.

Segment Summary

1. Learnt the key steps of Data Ingestion
2. Steps were demonstrated with the help of an example

upGrad

Segment - 06
Tools for Data Ingestion

Learning Objectives

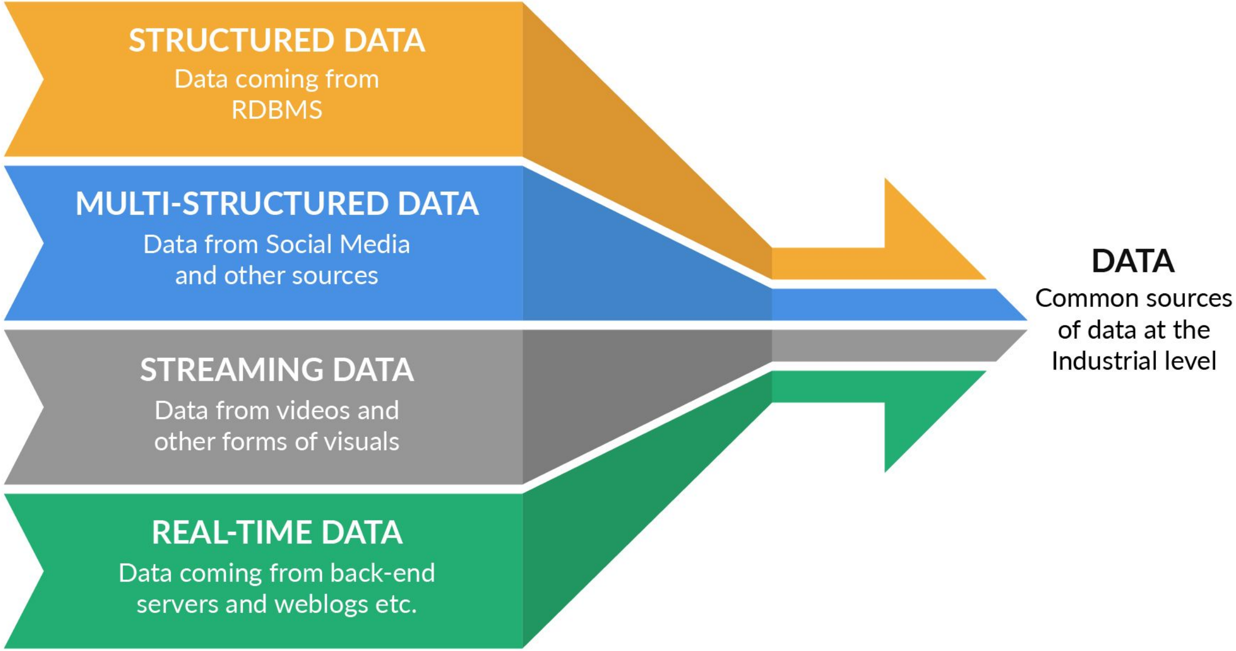
1. Brief overview of the different data sources

2. Introduction to some of the tools used for Data Ingestion

Tools for Data Ingestion

upGrad

Common Data Sources at the Industry Level



STRUCTURED DATA
Data coming from RDBMS

MULTI-STRUCTURED DATA
Data from Social Media and other sources

STREAMING DATA
Data from videos and other forms of visuals

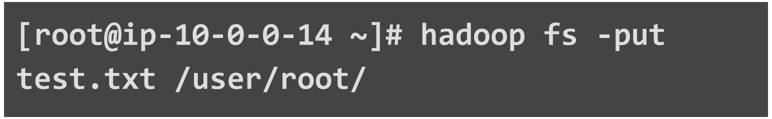
REAL-TIME DATA
Data coming from back-end servers and weblogs etc.

DATA
Common sources of data at the Industrial level

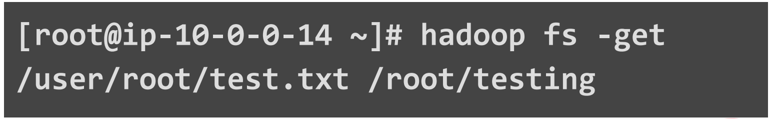
Tools for Data Ingestion

File transfer using commands

- ‘distcp’: copy large data sets between two clusters
- ‘put’ and ‘get’: copy files from the local file system to HDFS and vice versa, respectively



```
[root@ip-10-0-0-14 ~]# hadoop fs -put
test.txt /user/root/
```



```
[root@ip-10-0-0-14 ~]# hadoop fs -get
/user/root/test.txt /root/testing
```

Apache Sqoop

- Short for SQL to Hadoop
- Used for importing data from RDBMS to a Big Data Ecosystem (Hive, HBase, etc.) and exporting data back to RDBMS after it is processed.

Tools for Data Ingestion

Apache Flume

- Distributed data collection service for collecting, aggregating and transporting large amounts of real-time data from various sources to a centralised place, where it can be processed

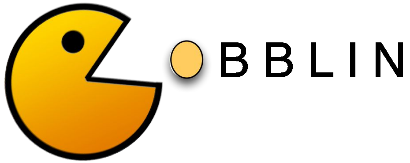
Apache Kafka

- Kafka is a fast, scalable distributed system that can handle a high volume of data.
- It enables programmers to pass messages from one point to another.

Tools for Data Ingestion

Apache Gobblin

- Gobblin is an open source data ingestion framework for extracting, transforming and loading a large volume of data from different data sources. It supports both streaming and batch data ecosystems.



Segment Summary

1. Introduced the various sources of data

2. Introduced some tools used for Data Ingestion

Segment - 07
Types of Data and File Formats

Learning Objectives

1. Different types of data handled in Data Ingestion

2. Different types of file formats handled in Data Ingestion

Types of Data and File Formats

Types of Data

- **Structured data**:
    - Organised data is generally stored in databases
    - Can be easily stored, entered, queried and analysed efficiently using SQL
    - Can be easily read by machines
    - Examples: Financial data, user identification data, etc.

Types of Data and File Formats

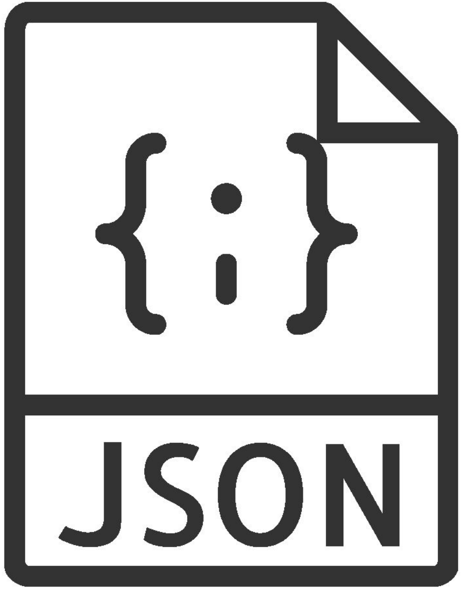
Types of Data

- **Unstructured data**:
    - Opposite of structured data: Cannot be easily stored and organised in databases
    - NoSQL databases can be used for this type of data
    - Approximately 80% of data being created today is unstructured in nature.
    - Examples: Images, audio, video, chat messages, etc.

Types of Data and File Formats

Types of Data

- **Semi-structured data**:
  - No predefined scheme unlike structured data
  - May have an internal structure and markings to identify separate data elements, but its schema does not constrain the data as in an RDBMS such as SQL tables.
  - Example: XML and JSON files



Types of Data and File Formats

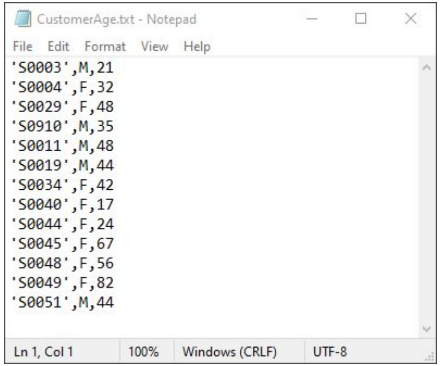
File Formats

Factors of data ingestion that vary depending on the file format:

- Processing power
- Network bandwidth
- Available storage

The file formats that are commonly dealt with in Data Ingestion are as follows:

- **Text/CSV**:
  - CSV: Comma-separated values
  - The most commonly used file format for exchanging large datasets between Hadoop and external systems
  - Limited support for schema evolution
  - Does not support block compression

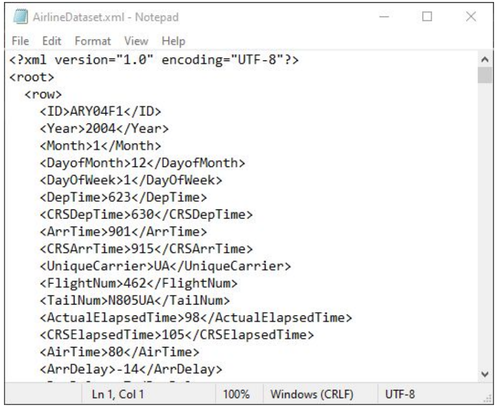


```
'S0003',M,21
'S0004',F,32
'S0029',F,48
'S0910',M,35
'S0011',M,48
'S0019',M,44
'S0034',F,42
'S0040',F,17
'S0044',F,24
'S0045',F,67
'S0048',F,56
'S0049',F,82
'S0051',M,44
```

Figure: Sample text/CSV file

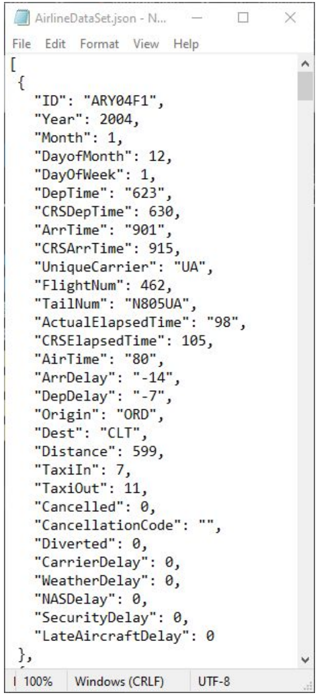
Types of Data and File Formats

- **XML and JSON**:
  - **XML**: Extensible Markup Language
    JSON: JavaScript Object Notation
  - **XML**: It defines a set of rules, using which documents can be encoded in a machine- and human-readable format.
  - **JSON**: Open-standard file format consisting key-value pairs
  - **Essentially text files**: Do not support block compression and are not compact
  - Splitting is hard and cannot be easily processed parallelly because no in-built InputFormat is present for either of the two formats in Hadoop.



```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <row>
    <ID>ARY04F1</ID>
    <Year>2004</Year>
    <Month>1</Month>
    <DayofMonth>12</DayofMonth>
    <DayOfWeek>1</DayOfWeek>
    <DepTime>623</DepTime>
    <CRSDepTime>630</CRSDepTime>
    <ArrTime>901</ArrTime>
    <CRSArrTime>915</CRSArrTime>
    <UniqueCarrier>UA</UniqueCarrier>
    <FlightNum>462</FlightNum>
    <TailNum>N805UA</TailNum>
    <ActualElapsedTime>98</ActualElapsedTime>
    <CRSElapsedTime>105</CRSElapsedTime>
    <AirTime>80</AirTime>
    <ArrDelay>-14</ArrDelay>
```

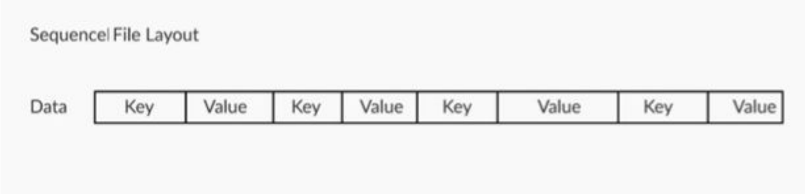
Figure: AirLine data set used as an example for showcasing XML and JSON files



```
[
 {
   "ID": "ARY04F1",
   "Year": 2004,
   "Month": 1,
   "DayofMonth": 12,
   "DayOfWeek": 1,
   "DepTime": "623",
   "CRSDepTime": 630,
   "ArrTime": "901",
   "CRSArrTime": 915,
   "UniqueCarrier": "UA",
   "FlightNum": 462,
   "TailNum": "N805UA",
   "ActualElapsedTime": "98",
   "CRSElapsedTime": 105,
   "AirTime": "80",
   "ArrDelay": "-14",
   "DepDelay": "-7",
   "Origin": "ORD",
   "Dest": "CLT",
   "Distance": 599,
   "TaxiIn": 7,
   "TaxiOut": 11,
   "Cancelled": 0,
   "CancellationCode": "",
   "Diverted": 0,
   "CarrierDelay": 0,
   "WeatherDelay": 0,
   "NASDelay": 0,
   "SecurityDelay": 0,
   "LateAircraftDelay": 0
 },
```

Types of Data and File Formats

- **Sequence file**:
  - Store data as binary key-value pairs in a binary format
  - More compact than text files
  - Supports block compression and can be easily processed parallely
- **Avro**:
  - A language-neutral data serialisation system developed with Apache’s Hadoop project
  - Can be easily read after creation, even in a language different from the one used to write the file.
  - Compact: A type of binary file
  - Self-describing, compressible and splittable and, hence, suitable for MapReduce Jobs
  - Supports scheme evolution



Sequence File Layout

Data	Key	Value	Key	Value	Key	Value	Key	Value

Figure: First, Sequence File Format
Second, Avro Logo

Segment Summary

1. Discussed the different types of data handled in Data Ingestion

2. Discussed various file formats handled in Data Ingestion

Session Summary

1. Discussed about data ingestion and why it is needed
2. Discussed the various challenges faced in Data Ingestion
3. Learnt about the three key steps of Data Ingestion, i.e., data collection, validation and routing
4. Looked at various tools used for Data Ingestion such as Apache Sqoop and Apache Flume
5. Discussed the types of data and file formats that are usually dealt with in Data Ingestion

Thank You