

# Data Management and Acquisition

# Course Outline

- Introduction
  - What is data management and why is it important
  - Data formats
  - Data sources
  - Data engineer skills and tools

# Introduction

- What is Data management ?
- Why is it important ?

The constant increase in data processing speeds and bandwidth, the nonstop invention of new tools for creating, sharing, and consuming data, and the steady addition of new data creators and consumers around the world, ensure that data growth continues unabated. Data begets more data in a constant virtuous cycle.

Forbes 2020

# Data Management

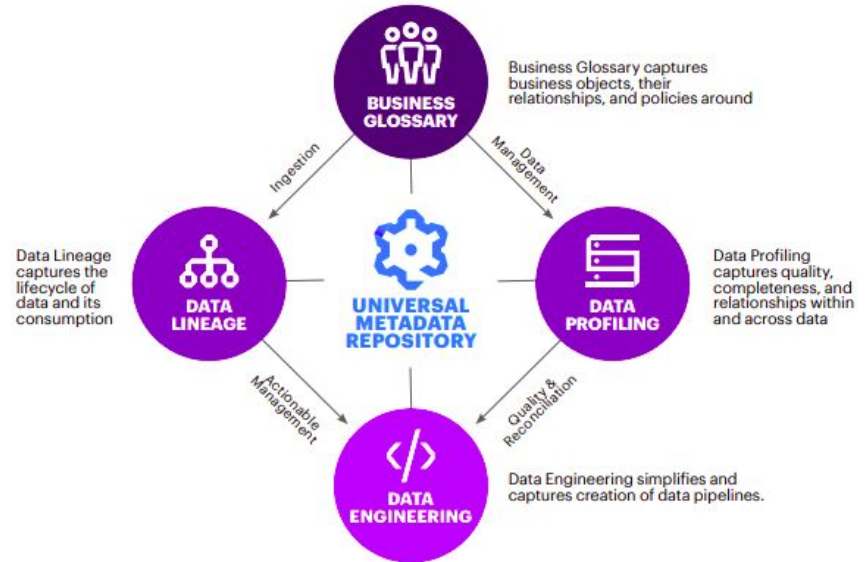
According to DAMA International's Guide to the Data Management Body of Knowledge (DAMA-DMBOK)

***“ data management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their life cycles. ”***

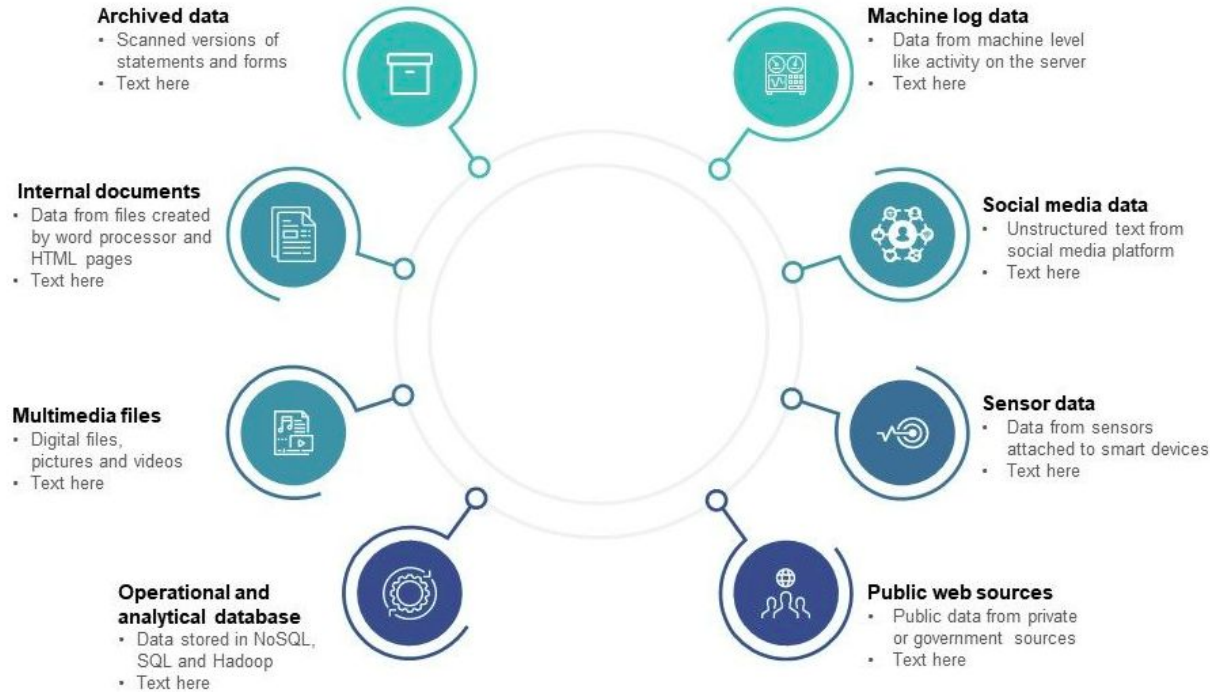
# Why is Data management important?

## The Modern Data Ecosystem and Infrastructure.

- Data from different sources
- Useful insight from the business side
- Collaboration among different stake holders
- Tools, process, infrastructure



# Data Sources



- Structured
- Unstructured

# Data Sources

First pull a copy of the data source.

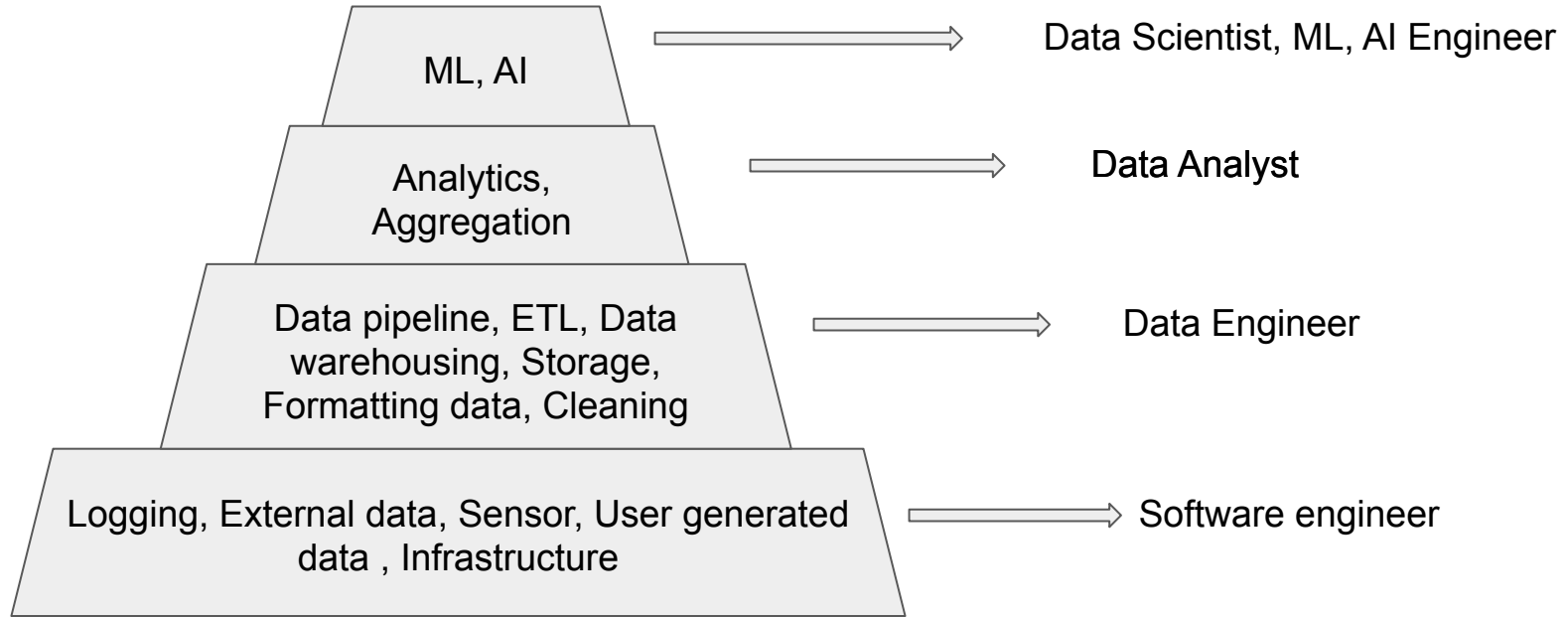
Challenges :

- Reliability
- Security
- Integrity

Secondly, Organise and clean up

Conform to the organization's standard

# Where does the subject fit in?





# Players in the Ecosystem

Software Engineer

Data Engineer

Data Scientist

Data Analyst

Business Intelligence Analyst

# Data Engineer

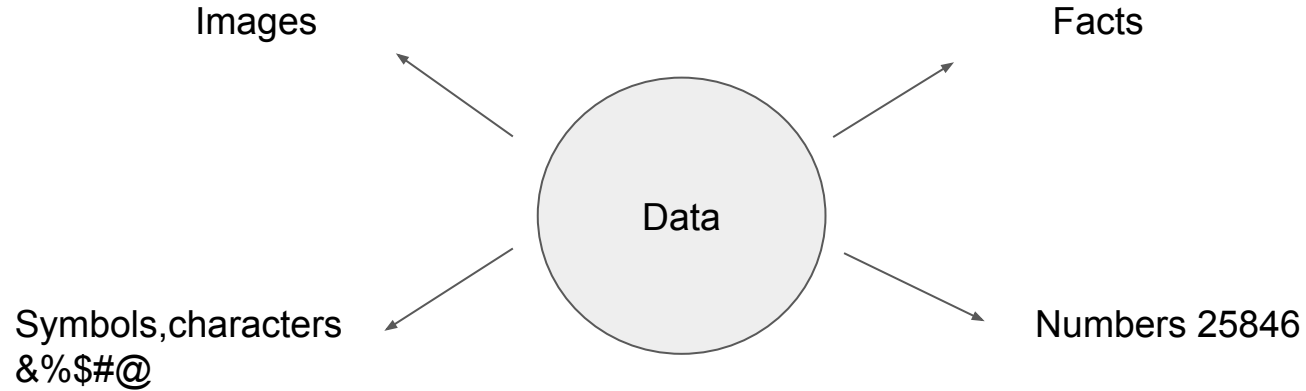
- Roles

- develop and maintain data architectures
- make data available for business operations and analysis.
- extract, integrate, and organize data from disparate sources; clean, transform, and prepare data
- design, store, and manage data in data repositories.
- enable data accessible in adequate format for different applications and stakeholders

- Skills

- Good knowledge of programming,
- Good knowledge of systems architectures,
- and in-depth understanding of relational and non-relational datastores.

# Data Types



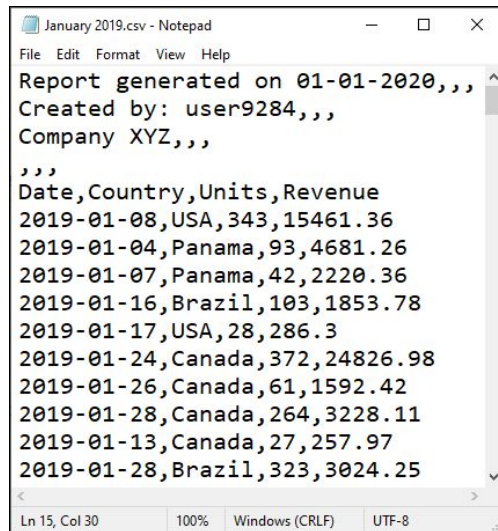
# Data Types

- **Structured**
  - Organised rows and column
  - SQL
  - OLTP
  - Excel, spreadsheet, Google sheet
- **Semi Structured**
  - Contains tags, metadata
  - Emails, xml, markup language integrated with JSON
- **Unstructured**
  - No identifiable structure
  - Web pages, social media feeds, images
  - No SQL

# Different file formats

There are many file formats

- **CSV** - delimited text
- Text Files
- **JSON**
- **XML**
- **Microsoft Excel File**
- SAS
- **SQL**
- Python Pickle File
- Stata
- HDF5
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery



```
January 2019.csv - Notepad
File Edit Format View Help
Report generated on 01-01-2020,,,
Created by: user9284,,,
Company XYZ,,,
,,,
Date,Country,Units,Revenue
2019-01-08,USA,343,15461.36
2019-01-04,Panama,93,4681.26
2019-01-07,Panama,42,2220.36
2019-01-16,Brazil,103,1853.78
2019-01-17,USA,28,286.3
2019-01-24,Canada,372,24826.98
2019-01-26,Canada,61,1592.42
2019-01-28,Canada,264,3228.11
2019-01-13,Canada,27,257.97
2019-01-28,Brazil,323,3024.25
Ln 15, Col 30 100% Windows (CRLF) UTF-8
```

# Different file formats

There are many file formats

- **CSV** - delimited text
- Text Files
- **JSON**
- **XML**
- **Microsoft Excel File**
- SAS
- **SQL**
- Python Pickle File
- PDF
- HDF5
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery

```
[
  {
    "date": "2013-11-05",
    "locations": {
      "United States": 4,
      "Germany": 8
    }
  },
  {
    "date": "2013-11-11",
    "locations": {
      "South Africa": 9
    }
  },
  {
    "date": "2013-11-12",
    "locations": {
      "Japan": 6
    }
  }
]
```

# Different file formats

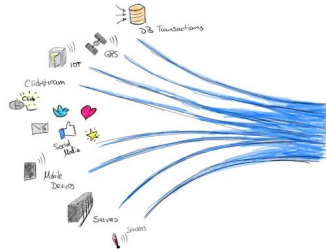
There are many file formats

- **CSV** - delimited text
- Text Files
- **JSON**
- **XML**
- **Microsoft Excel File**
- SAS
- **SQL**
- Python Pickle File
- PDF
- HDF5
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery

```
<?xml version="1.0" encoding="UTF-8"?>
- <EmployeeData>
  - <employee id="34594">
    <firstName>Heather</firstName>
    <lastName>Banks</lastName>
    <hireDate>1/19/1998</hireDate>
    <deptCode>BB001</deptCode>
    <salary>72000</salary>
  </employee>
  - <employee id="34593">
    <firstName>Tina</firstName>
    <lastName>Young</lastName>
    <hireDate>4/1/2010</hireDate>
    <deptCode>BB001</deptCode>
    <salary>65000</salary>
  </employee>
</EmployeeData>
```

# Data Sources

Some common data sources include:



Flat File  
(XML, CSV, etc)

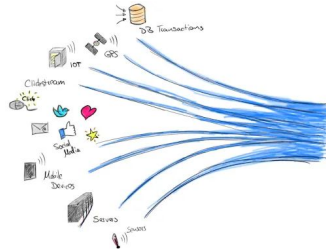


Web Scraping



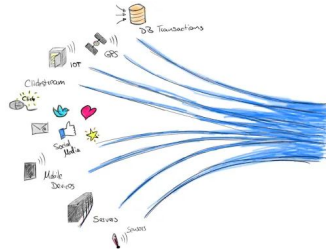
# Data Sources

Some common data sources include:



# Data Sources

Some common data sources include:



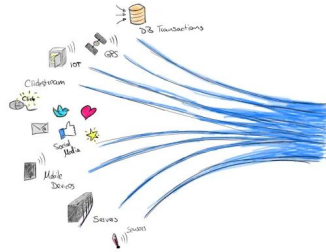
Flat File  
(XML, CSV, etc)



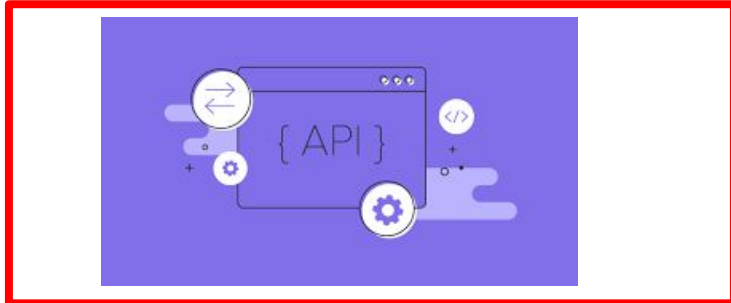
Web Scraping

# Data Sources

Some common data sources include:



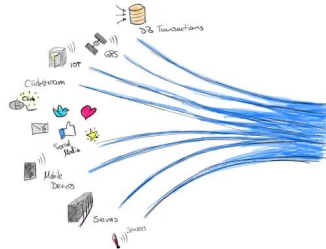
Flat File  
(XML, CSV, etc)



Web Scraping

# Data Sources

Some common data sources include:



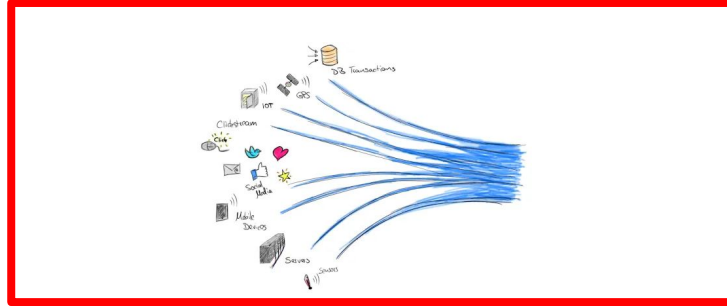
Flat File  
(XML, CSV, etc)



Web Scraping

# Data Sources

Some common data sources include:



Flat File  
(XML, CSV, etc)



Web Scraping

# Essential Language



```
skickar@Dell-3 Nuke % which bash
/bin/bash
skickar@Dell-3 Nuke % ls
expect.exp  trigger.sh
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
Hackers love to learn on Null Byte
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh "I am a computer scientist" "president"
I firmly believe that Vermin is the office of president
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
skickar
skickar@Dell-3 Nuke %
skickar@Dell-3 Nuke %
What is your name?
Kody
Wow, Kody sounds like a pun!
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
bash.sh: line 9: syntax error: unexpected end of file
skickar@Dell-3 Nuke % nano bash.sh
skickar@Dell-3 Nuke % bash bash.sh
What is your name?
Kody
```

**Automate tasks with Bash scripts**



# Common Data Repositories

- Database
  - Relational
  - Non relational
  - Choice depends on many factors
- Data Warehouse
  - Brings data into a point after ETL
  - Data Marts and Data lakes
- Big data stores
  - Distributed data storage system

ETL

Vs

ELT

Vs

Data Pipeline

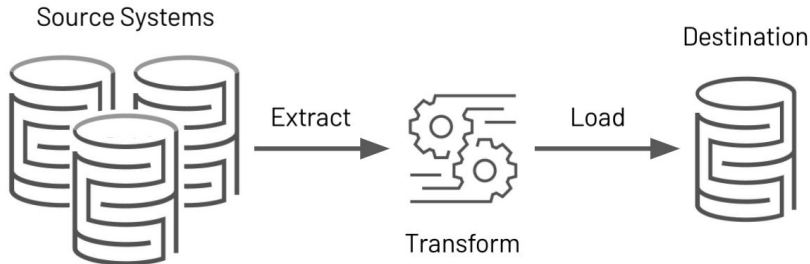


# Extract Transform Load

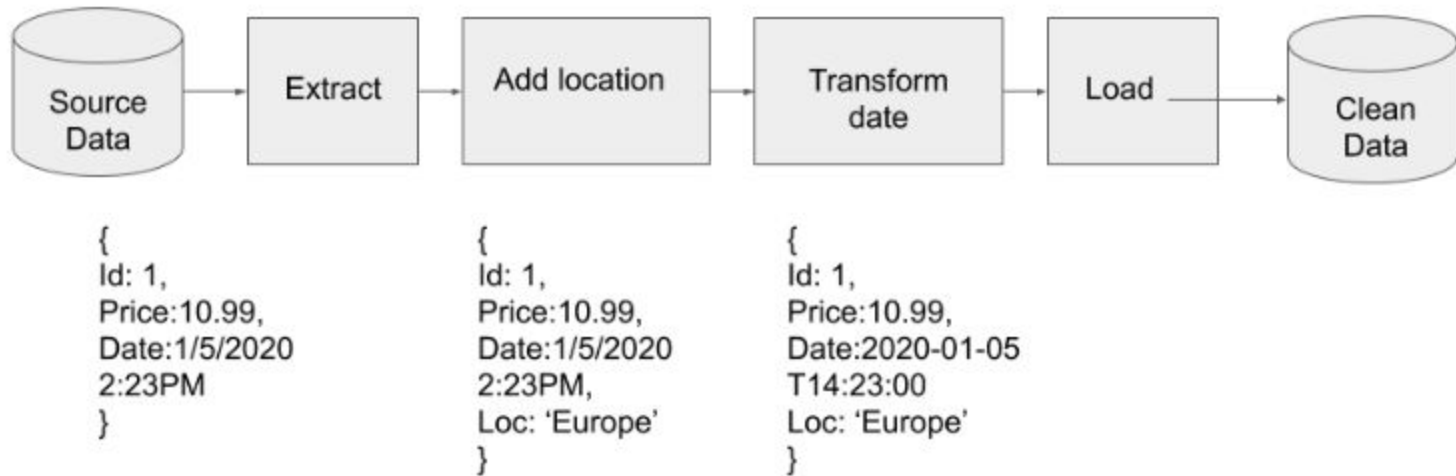
An automated process which comprises of

- Getting data from different source
- Extract relevant information
- Transform the data i.e., clean, standardize, create extra fields
- Load to a destination (a data repository)

## ETL Process



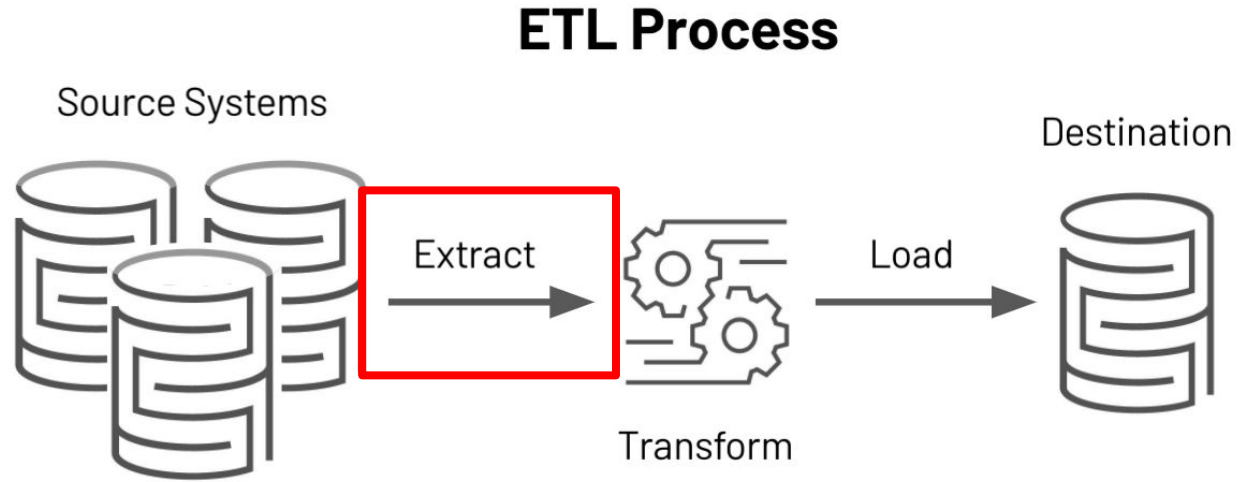
# ETL



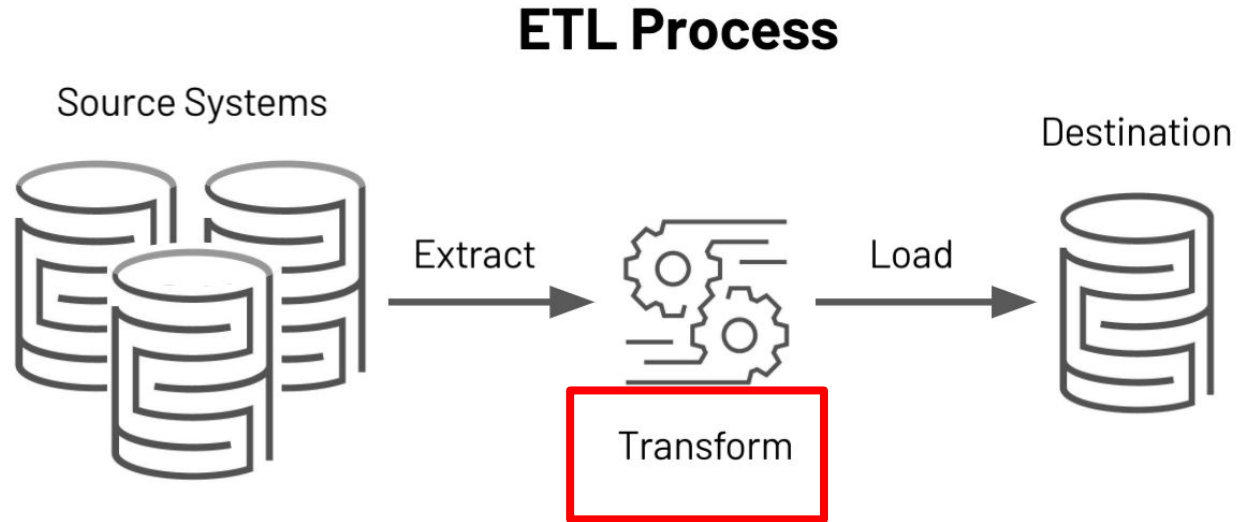
# Extract

Batch or

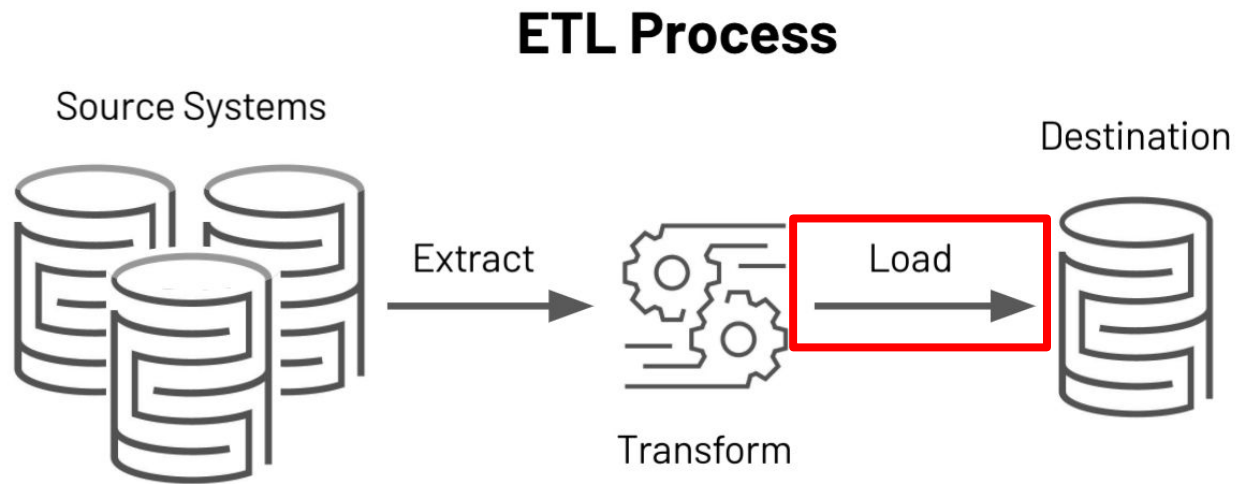
Real time



# Transform



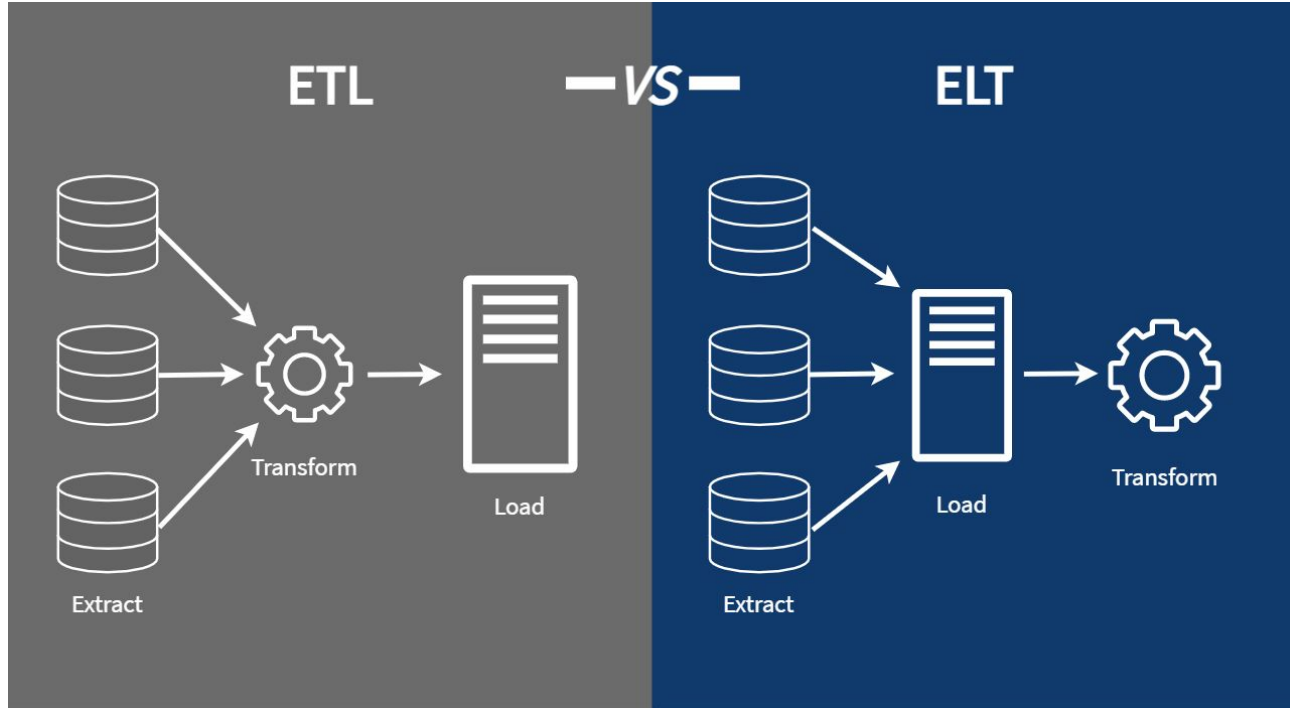
# Load



# Emergence of ELT?

ETL → Challenges with Streaming data

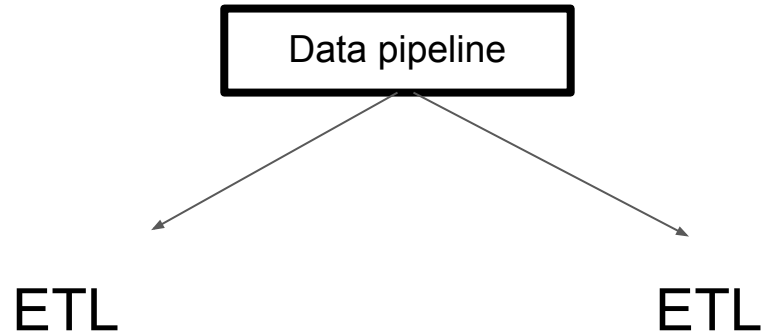
ELT is relatively new



# ETL vs ELT

ETL	ELT
Long cycle before reaching destination	Shorter cycle before reaching destination
Often a batch process	Good for streaming workload as data is available
Transformation is pre-carried out before getting to analyst and data scientists	Gives room for analyst and data scientist to do their own transformations
Transformation is carried out on all the data	Transform only the required data
Works well with small data	Useful for big data

# Data Pipeline



**Data pipeline** is a term used for the migration and transformation of data from the start to the final destination





# About the Assignment

Data Architecture

# Install VM

<https://www.youtube.com/watch?v=x5MhydijWmc>

