upGrad

Data Warehousing and ETL

Course: Data Engineering - I

Lecture On: Data Warehousing and ETL

Instructor: Ganesh Nerale

Session 3 | ETL, ELT and Data Lakes

Learning Objectives of the Session

Segment 02 Description of the ETL process.

Segment 03 Elaboration of the data extraction process and its various types.

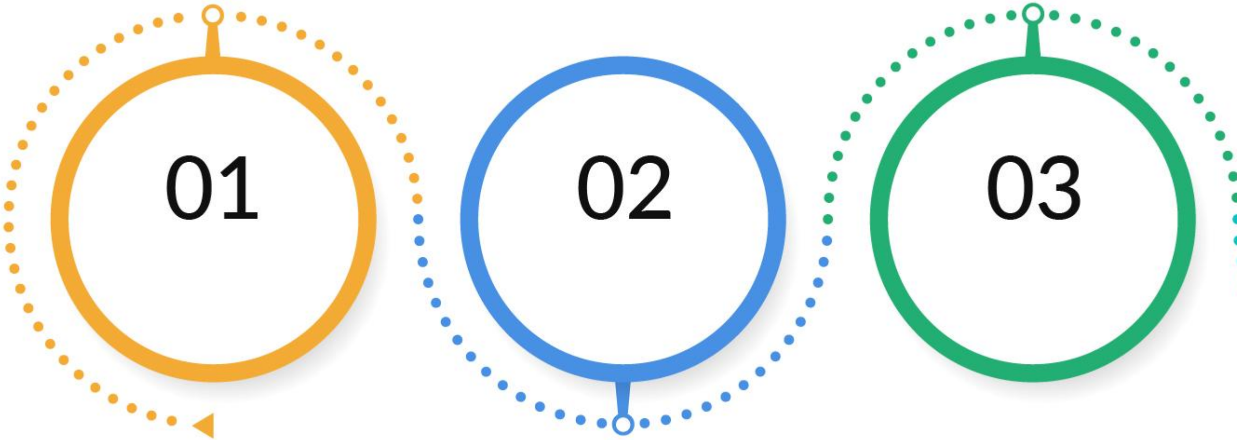
Segment 04 Understanding the need of cleaning, structuring and enrichment of data before loading it to data warehouses

Segment 05 Understanding the use of ELT process to handle unstructured data.

Segment 06 Understanding the use of Data Lakes to run analytics on unstructured data.

Segment 2 | ETL Process

Learning Objectives

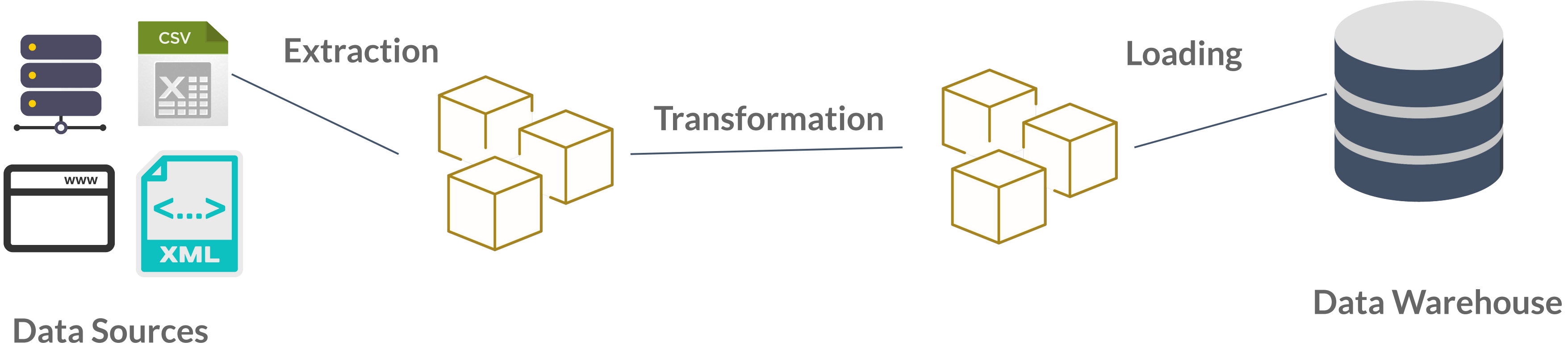


Introduction to the ETL process

Why Business Requirements must be known before designing the ETL processes?

Important concepts for designing the ETL processes

The ETL Process



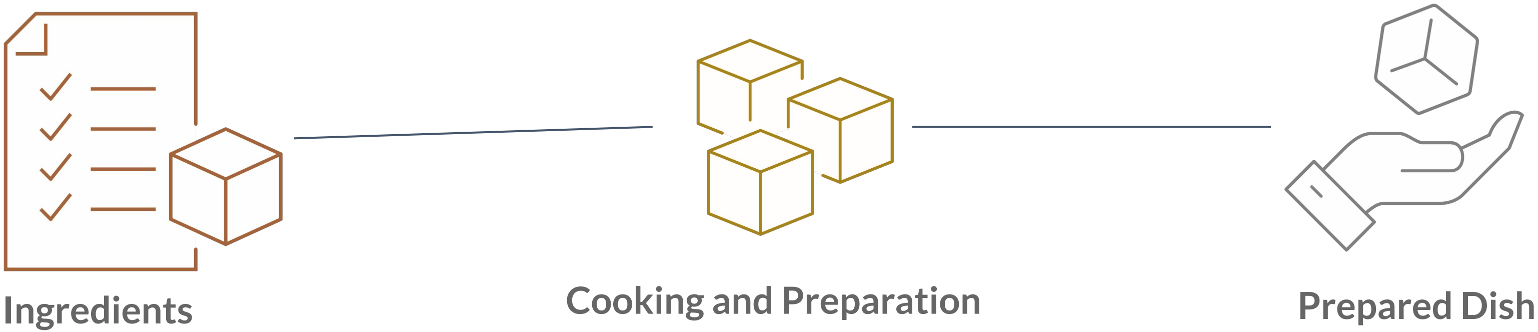
Data Sources

Extraction

Transformation

Loading

Data Warehouse



Ingredients

Cooking and Preparation

Prepared Dish

ETL: Important Concepts

upGrad

1 Business Requirements

01 A discussion with the users of this entire system is important

02 A discussion with the data modelling team working on the designing dimensions and facts is important

03 Once you know the end product, you can trace the sources that would be required and the processes to be applied

2 Data Profiling

01 Examining different sources for their quality and context of use

02 In case of a good data source, the data from the source is put into the data warehouse using the least amount of data transformation

03 A dirty data source requires the removal of corrupt values and columns that have no scope of use

ETL: Important Concepts

upGrad

3 Exception Handling

01 If any of the ETL processes throws an exception, then there must be a record to tell which process has thrown which exception

02 The ETL processes have to be designed around the concept that they are able to recover and restart if any such event happens.

4 Data Security

01 You don’t want the business users to access the data at the different stages of data warehouse development

02 Who has the right to access the data?

Important Concepts

upGrad

5 Data Archiving

01 Data from sources may require some changes during the process

02 Maintain a separate staging area for recording the changes to the data

03 If you have to reprocess some data, then you don’t have to start from the data source again. It is required to provide proof of transactional flow of changes to the data

Summary | ETL

01 The data is first extracted, transformed and then loaded into data warehouses.

02 Business Requirements define the data sources and processes to be used for ETL.

03 Data Profiling is checking for good and dirty data sources.

04 Data Security is securing the ETL process data from business users.

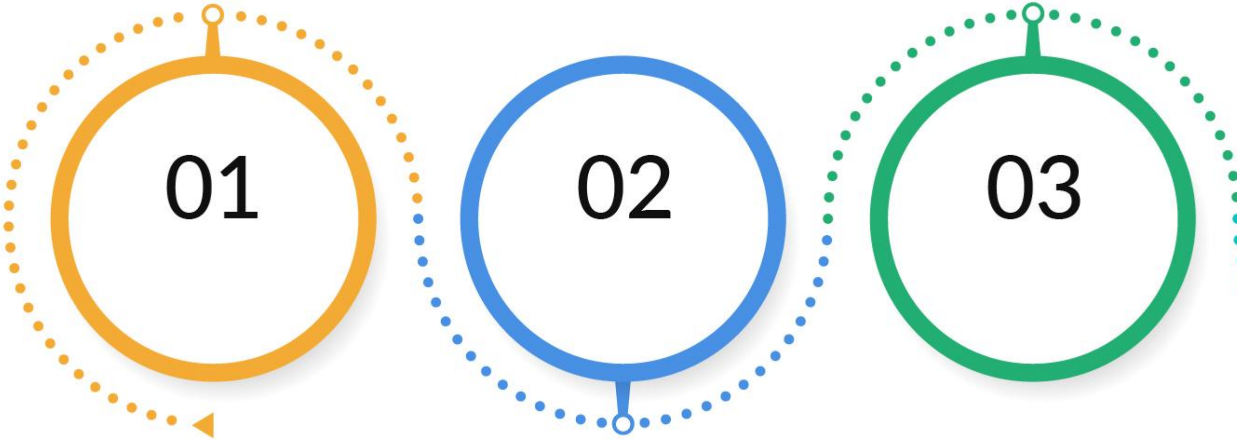
05 Exception Handling is checking for any exceptions that are thrown by any ETL process.
Data Archiving is keeping track of changes to a data source from extraction to loading.

Segment 3 | Data Extraction

Learning Objectives

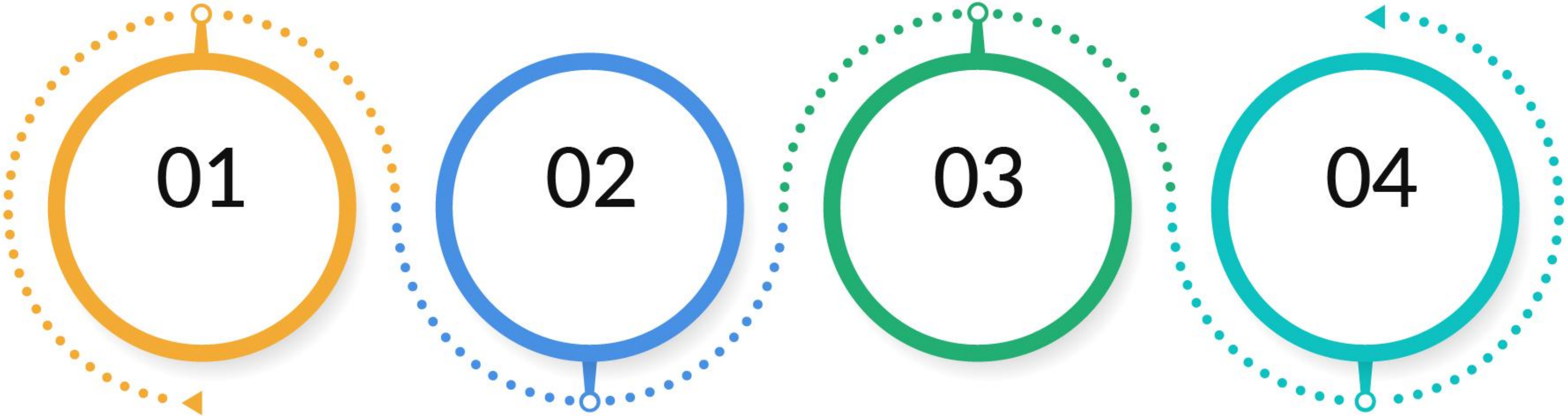
What is the process of data extraction? How to choose data sources?

Different types of data extraction processes



Different types of data sources

The Process of Extraction



Choosing data sources
The right data sources for the data warehouse

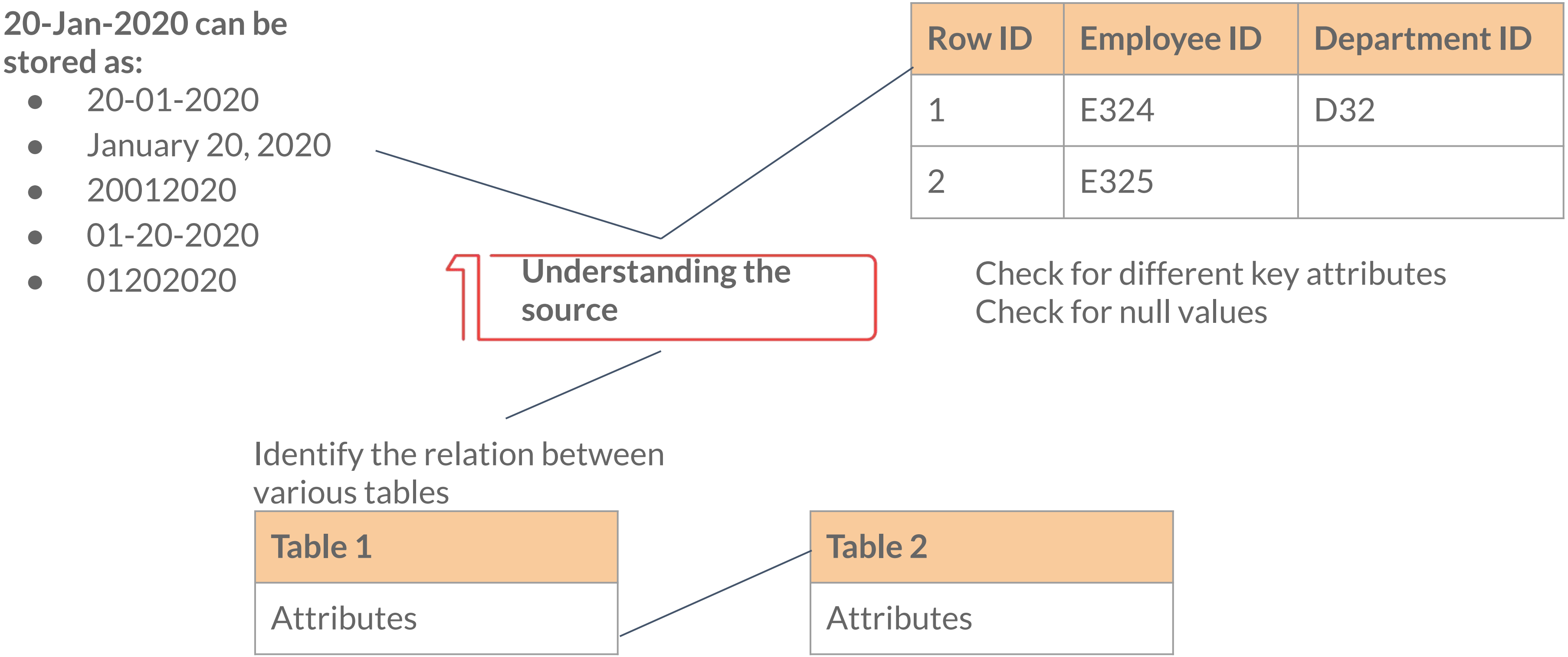
Data Profiling
Finding out good and bad data sources

Data Auditing
Which key metrics are useful in every data source?

Bringing the required data from the data sources to the main processing stream

Relational Databases as Data Sources

Since the database is already implemented, a relational model based on the implementation is drawn.



20-Jan-2020 can be stored as:

- 20-01-2020
- January 20, 2020
- 20012020
- 01-20-2020
- 01202020

Understanding the source

Row ID	Employee ID	Department ID
1	E324	D32
2	E325	

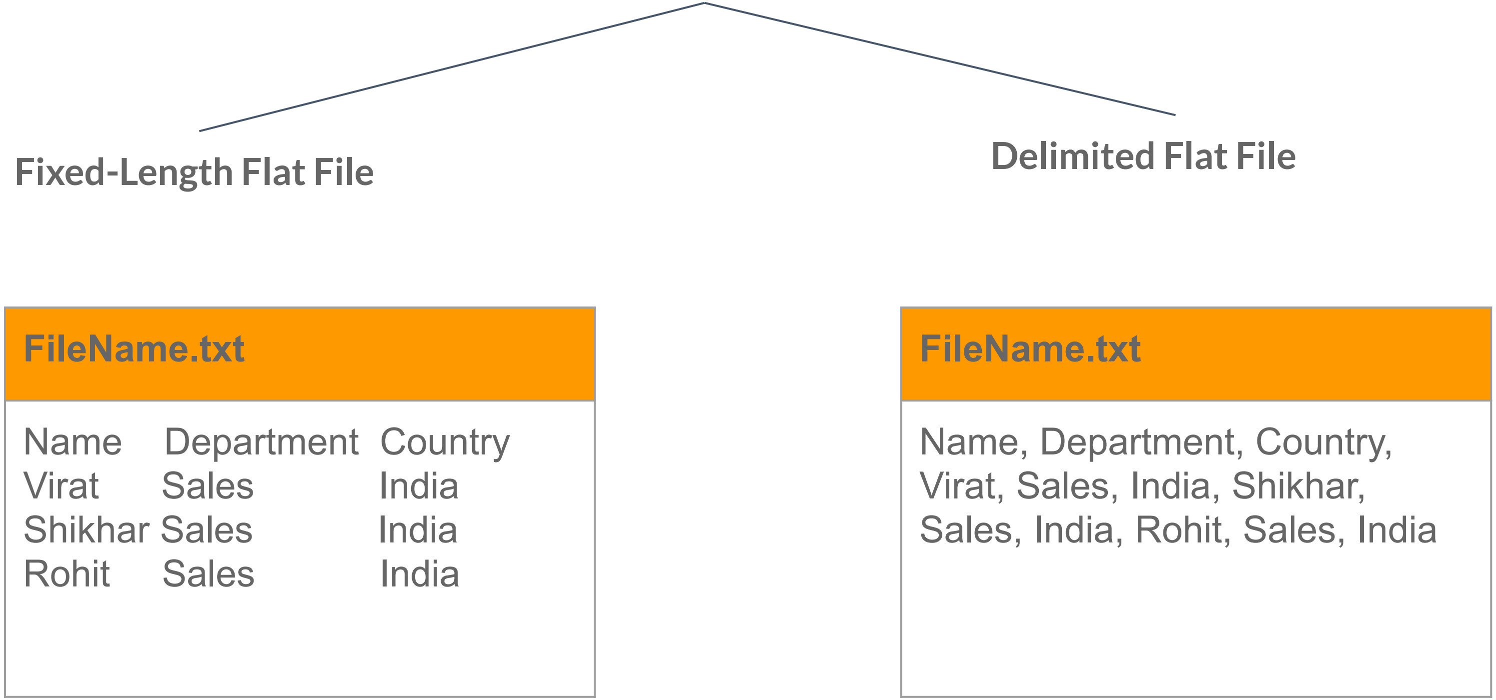
Check for different key attributes
Check for null values

Identify the relation between various tables

Table 1
Attributes

Table 2
Attributes

Flat Files as Data Sources



Fixed-Length Flat File

Delimited Flat File

FileName.txt

```
Name    Department  Country
Virat   Sales       India
Shikhar Sales       India
Rohit   Sales       India
```

FileName.txt

```
Name, Department, Country,
Virat, Sales, India, Shikhar,
Sales, India, Rohit, Sales, India
```

Other Data Sources

Different Extraction Processes

upGrad

1 Full Extraction

01 The entire table is extracted as it is; the data is transformed and is loaded into the data warehouse

02 While building data warehouses, a full extraction is performed on the data sources

2 Incremental Extraction

01 Only when there is an update in the source is the new data extracted and loaded into the data warehouse

02 Incremental extraction can be automated. One way to do this is to extract new data daily

03 Various triggers are used to extract data from data sources only when there are changes to the data in those sources

Summary | Data Extraction

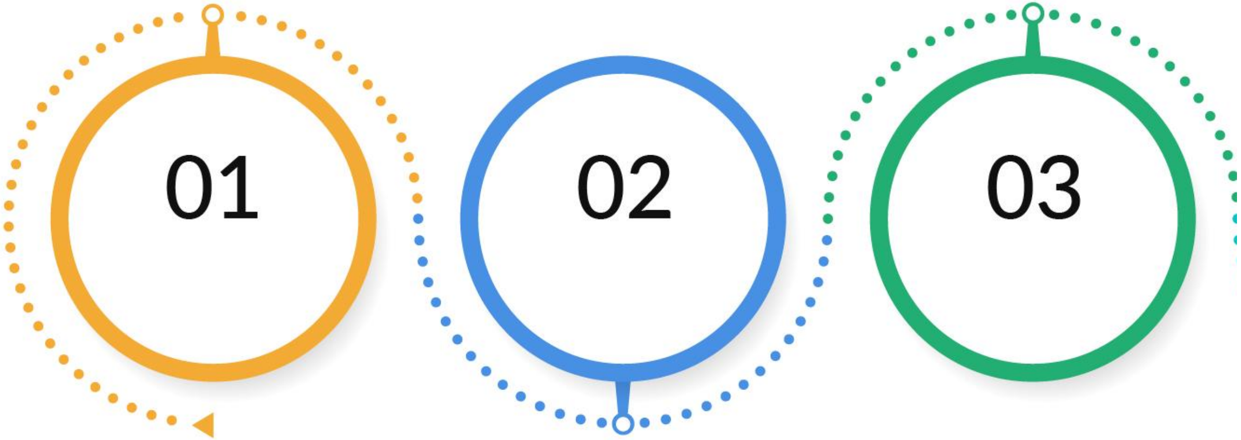
01 Data Extraction involves defining the sources, understanding the format in which the data is stored and extract the data into main processing stream.

02 Relational Databases, XML files, CSV files, Flat Files and Weblogs are various different data sources.

03 Full Extraction involves extracting the entire data from a data source.
Incremental Extraction involves extracting the data step by step from a data source.

Segment 4 | Data Transformation and Loading

Learning Objectives



What is the process of data transformation?

Data transformation on various data sources

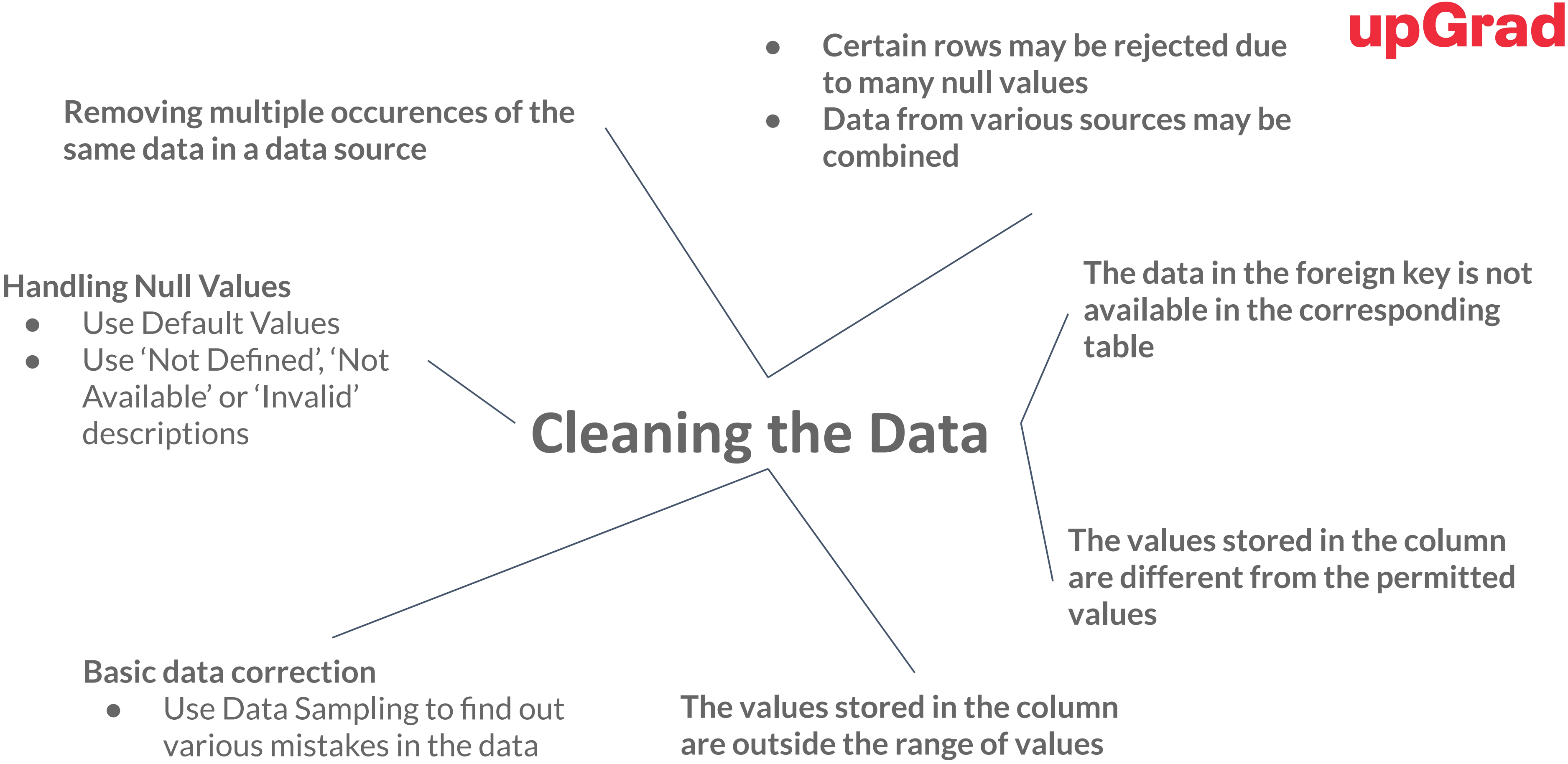
The process of data loading

The Process of Data Transformation

The motive of this step is to clean the data and give it a structure to ensure Data Quality

Data quality checks are done at two stages on the extracted data. One when the data is extracted and, next, when the data is loaded

In the extraction process, you understand the various sources. Here, you clean and make changes to the data



Cleaning the Data

Removing multiple occurences of the same data in a data source

- Certain rows may be rejected due to many null values
- Data from various sources may be combined

Handling Null Values

- Use Default Values
- Use ‘Not Defined’, ‘Not Available’ or ‘Invalid’ descriptions

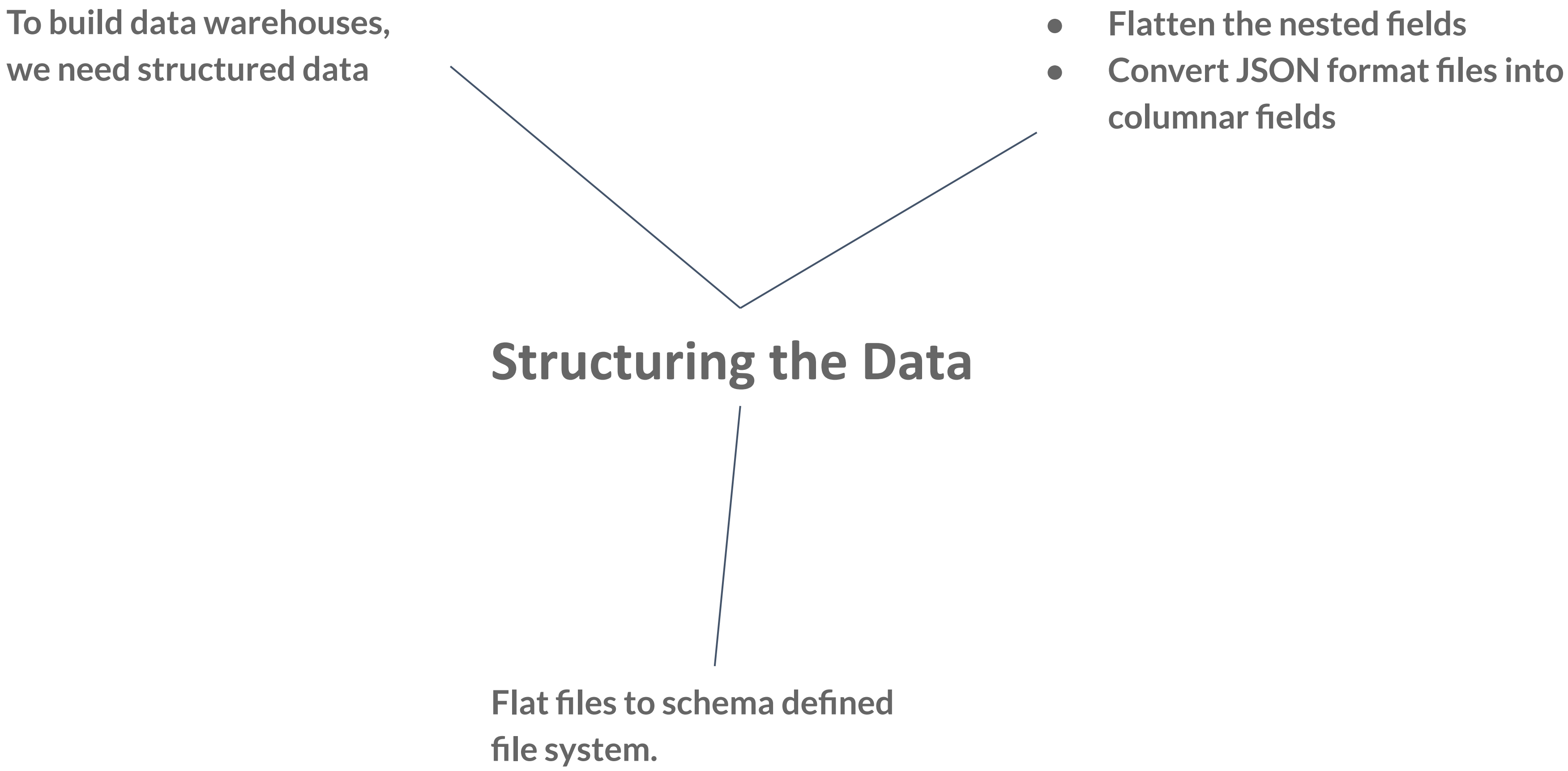
The data in the foreign key is not available in the corresponding table

The values stored in the column are different from the permitted values

Basic data correction

- Use Data Sampling to find out various mistakes in the data

The values stored in the column are outside the range of values



To build data warehouses, we need structured data

- Flatten the nested fields
- Convert JSON format files into columnar fields

Structuring the Data

Flat files to schema defined file system.

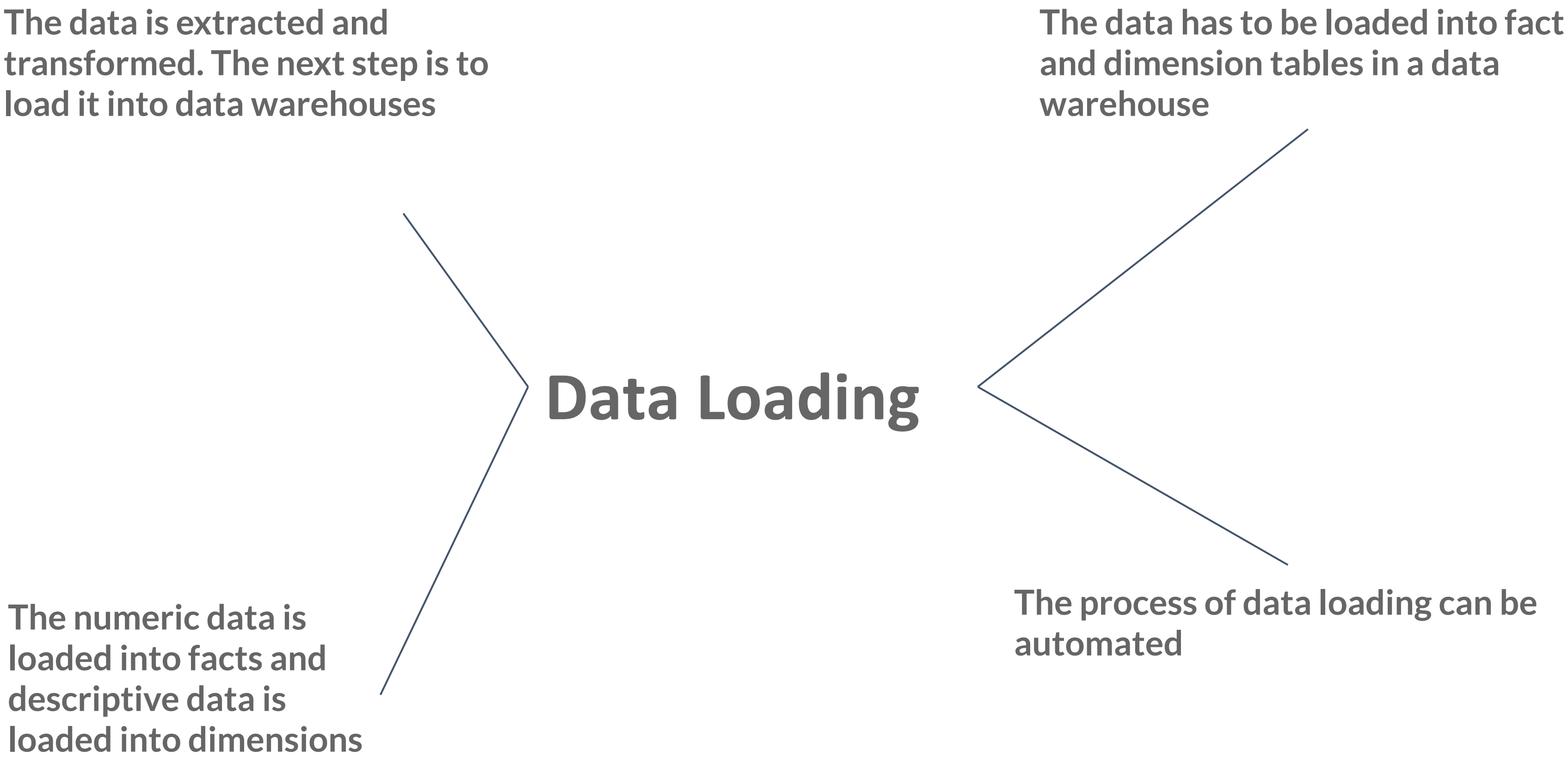
Data Enrichment

01 Apply Business rules that we captured after discussions with Business.

02 Apply join and aggregation logic to denormalize your source data

03 Derive key attributes based on business requirement

04 Data can be made descriptive according to the business rules.



Data Loading

The data is extracted and transformed. The next step is to load it into data warehouses

The data has to be loaded into fact and dimension tables in a data warehouse

The numeric data is loaded into facts and descriptive data is loaded into dimensions

The process of data loading can be automated

Summary | Data Transformation and Loading

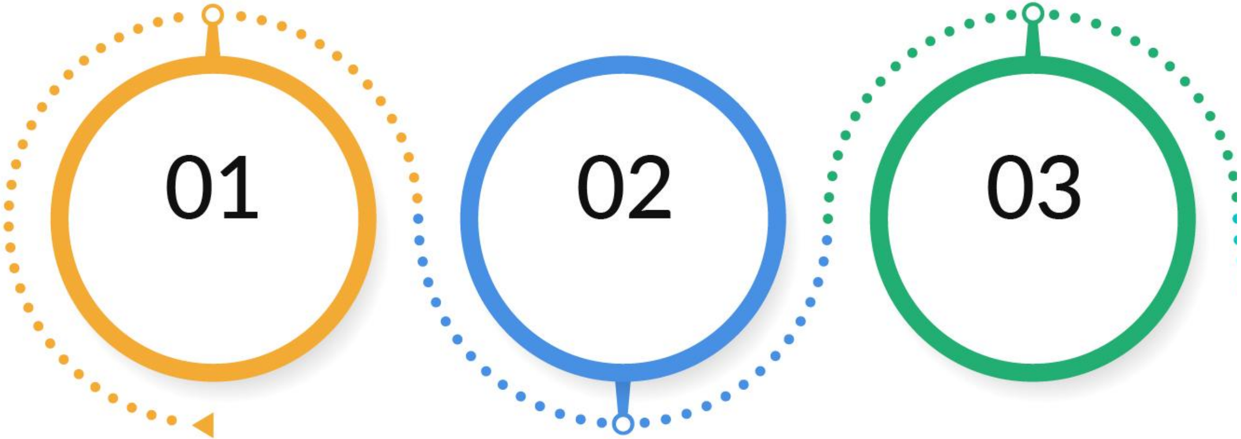
01 Data Transformation includes cleaning and structuring the data extracted.

02 The process of Data Transformation also involves checking for any incorrect data in the data sources.

03 Data Loading involves loading the data into facts and dimension tables of data warehouses.

Segment 5 | ELT

Learning Objectives



What are some issues with the ETL process?

What is the ELT process?

Use cases of the ELT process

Issues with ETL

01

Specify the Business Subject First

Although it is useful to specify the subject first, data extraction and transformation according to every particular business subject are not fast **when data is available in huge volumes in different formats**.

02

Need for Servers to Perform Data Transformation

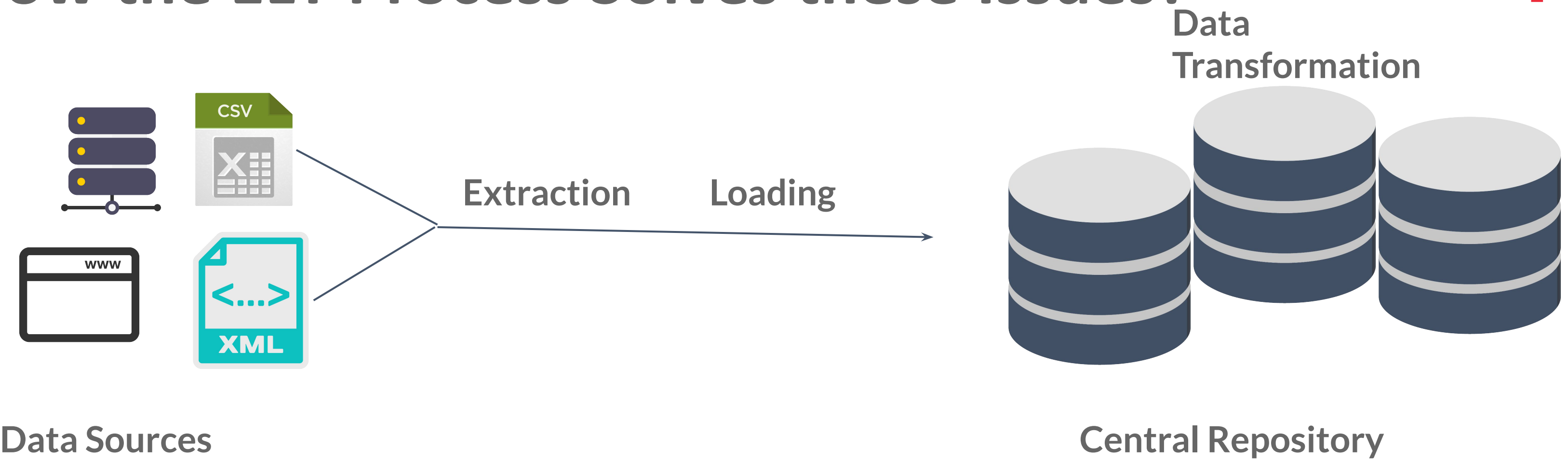
To perform transformation processes before loading, **you have to building a staging area** where data transformations can be performed.

03

Less Flexible

The data sources have to be cleaned and, first, data transformation has to be performed according to the business process defined.
Data from every source has to be transformed first, even though it is not used frequently.

How the ELT Process Solves these Issues?

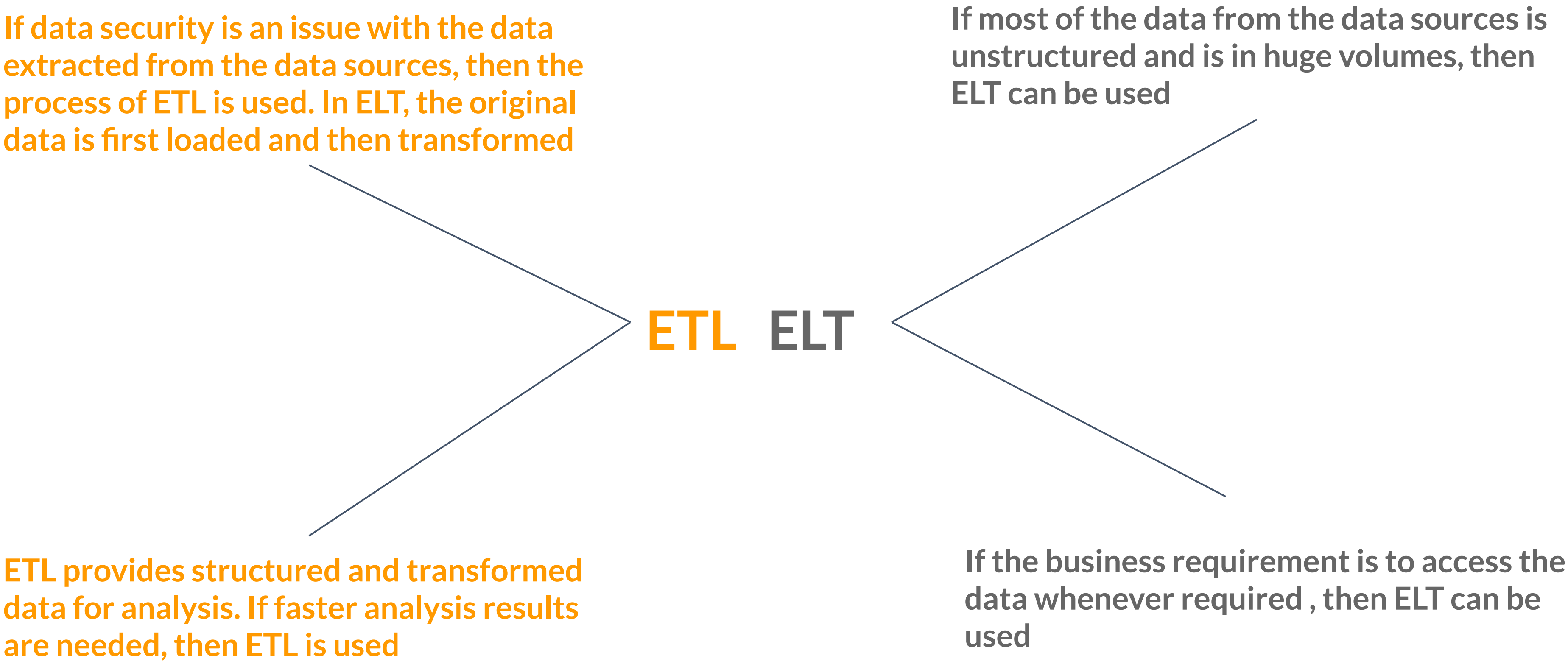


Extract all the data
With this process, you can load all kinds of data from various data sources without transforming any data

Specific data transformation
If you need any data for analysis, then you can transform it at that time and perform your analysis

Transformation can be done at the central repository
With this process, you do not need separate servers for data transformation

Use Cases for ETL and ELT



ETL ELT

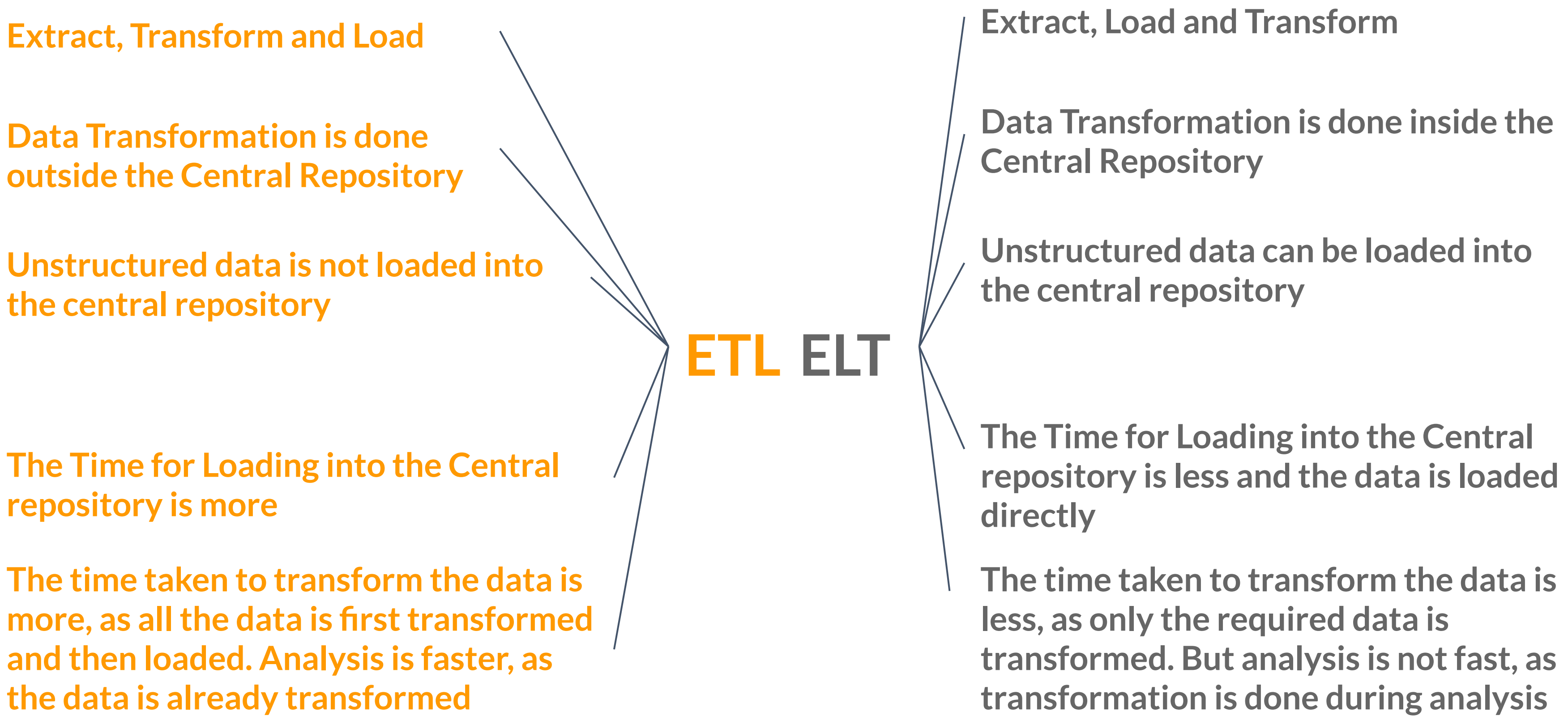
If data security is an issue with the data extracted from the data sources, then the process of ETL is used. In ELT, the original data is first loaded and then transformed

ETL provides structured and transformed data for analysis. If faster analysis results are needed, then ETL is used

If most of the data from the data sources is unstructured and is in huge volumes, then ELT can be used

If the business requirement is to access the data whenever required , then ELT can be used

ETL vs ELT



ETL

Extract, Transform and Load

Data Transformation is done outside the Central Repository

Unstructured data is not loaded into the central repository

The Time for Loading into the Central repository is more

The time taken to transform the data is more, as all the data is first transformed and then loaded. Analysis is faster, as the data is already transformed

ELT

Extract, Load and Transform

Data Transformation is done inside the Central Repository

Unstructured data can be loaded into the central repository

The Time for Loading into the Central repository is less and the data is loaded directly

The time taken to transform the data is less, as only the required data is transformed. But analysis is not fast, as transformation is done during analysis

Summary | ELT

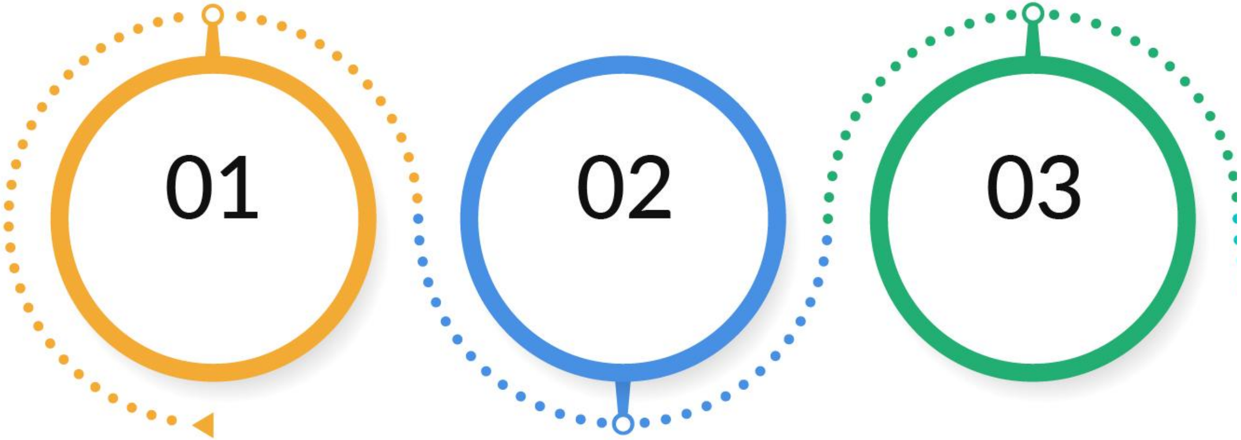
01 In ETL, the data transformation step requires separate servers. The process is also less flexible as the data is transformed first and then loaded into data warehouses.

02 In ELT, the process of transformation is done during data analysis. All the data is loaded and only the required data is transformed for analysis.

03 Analysis is fast for ETL as the transformed data is available during analysis.

Segment 6 | Data Lakes

Learning Objectives



What are the issues with data warehouses?

What are Data Lakes?

Use Cases of Data Lakes

Issues with Data Warehouses

upGrad

01

Schema-Specific

A data warehouse schema with facts and dimensions has to be defined before loading the data into the warehouse.

02

Highly Structured

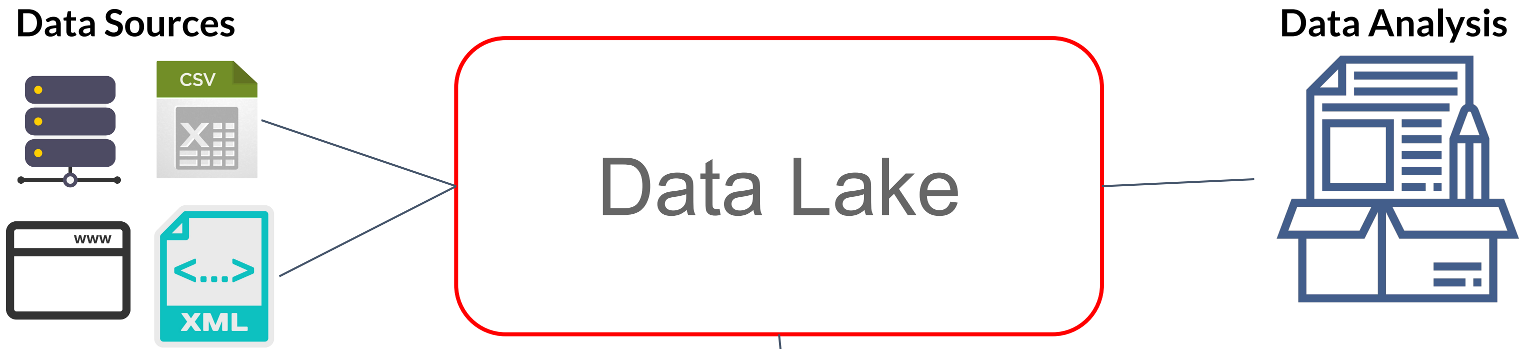
Only structured data can be loaded into the data warehouse. If there is any unstructured data source, then it has to be structured first before loading the data into the data warehouse.

03

Less Flexible

There has been an increase in data formats. Weblogs, XML formats, and many other forms of semi-structured and unstructured data are prominent sources of data for analysis. Each form of unstructured data has to be transformed into a structured format for loading into data warehouses.

How does a Data Lake Solve these Issues?



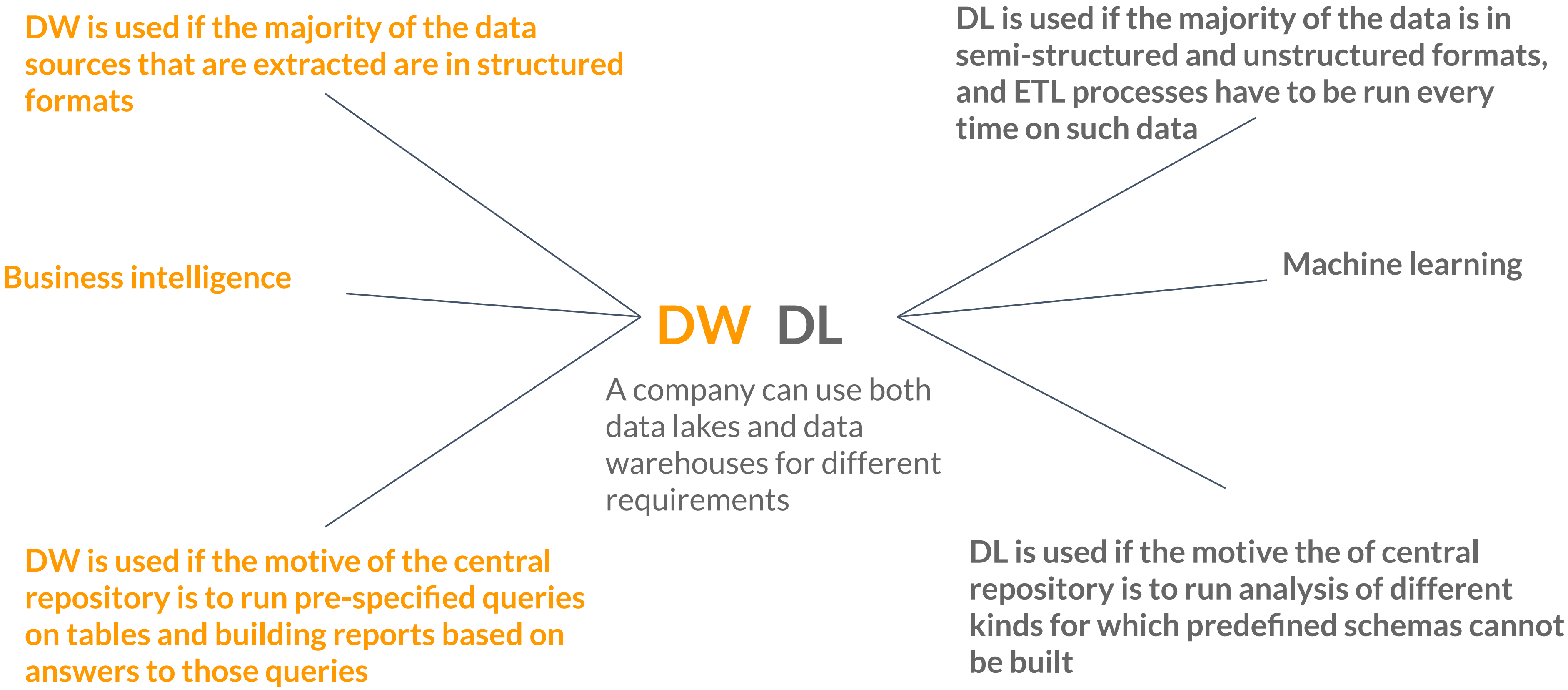
A data lake is central repository of all structured, semi-structured, unstructured data of an enterprise.

Data Lakes can be used for analytics on unstructured data

There is no need for data transformation before loading the data into data lakes. You can apply structuring after loading the data into data lakes

All sources of data can be stored in their original formats in data lakes

Use Cases of Data Warehouses and Data Lakes



DW DL

A company can use both data lakes and data warehouses for different requirements

DW is used if the majority of the data sources that are extracted are in structured formats

Business intelligence

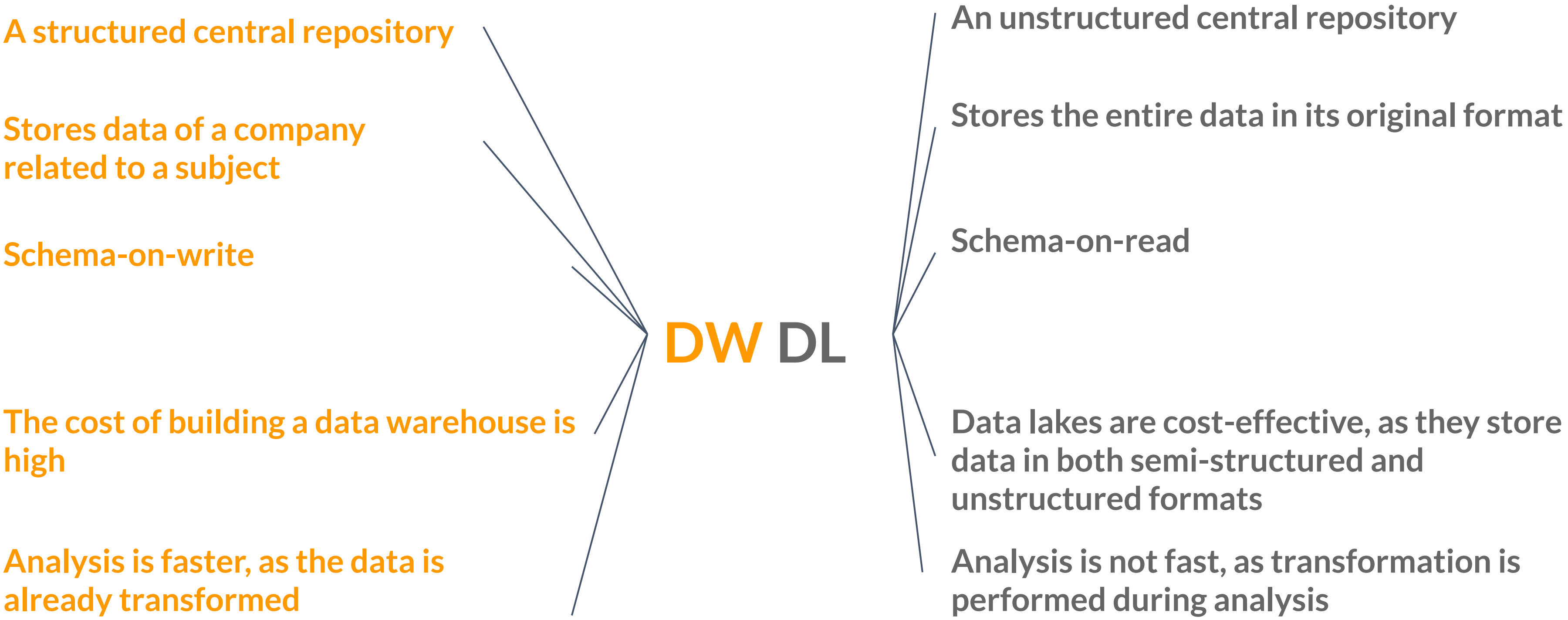
DW is used if the motive of the central repository is to run pre-specified queries on tables and building reports based on answers to those queries

DL is used if the majority of the data is in semi-structured and unstructured formats, and ETL processes have to be run every time on such data

Machine learning

DL is used if the motive the of central repository is to run analysis of different kinds for which predefined schemas cannot be built

Data Warehouse vs Data Lake



A structured central repository

Stores data of a company related to a subject

Schema-on-write

The cost of building a data warehouse is high

Analysis is faster, as the data is already transformed

DW DL

An unstructured central repository

Stores the entire data in its original format

Schema-on-read

Data lakes are cost-effective, as they store data in both semi-structured and unstructured formats

Analysis is not fast, as transformation is performed during analysis

Summary | Data Lakes

01 Data Warehouses are schema-specific and contain only structured data.

02 Data Lakes can be used if most of the data sources are semi-structured or unstructured.

03 Data Warehouses are used for analytics reporting. Data Lakes are used for machine learning and analysing data in various different ways.

Session Summary

upGrad

01 **ETL** stands for Extraction, Transformation and Loading.

02 **Data extraction** involves understanding and selecting data sources.

03 The **ETL** process is performed keeping in mind the business requirements, data profiling, data archiving, exception handling and data security.

04 **Relational databases, XML files, CSV files, flat files, WEBlogs and web pages** are various sources of data.

05 **Data transformation involves cleaning and structuring the data.** The data is then loaded into the fact and dimension tables of a data warehouse.

06 The **ETL** process is less flexible if the data is unstructured. Therefore, the transformation step has to be performed first in ETL. ELT is used to solve this issue.

07 **ETL** is used when data security is important. ELT is used when the majority of of the data is unstructured.

08 **Data warehouses are schema-specific, and only store structured data**. If the majority of the data is unstructured, then data lakes are used as a central repository.

09 **Data lakes** provide a cost-effective storage for unstructured data.

10 Unstructured data can be loaded directly into **data lakes**.

Thank You