Data Engineering

204426

Overview of Data Engineering

Outline

- What is Data Engineering?
- Required skills and knowledge
- Data Engineering & Data Science
- Data Engineering Process
  - ETL Process
  - ELT Process
  - Data Pipeline

What is Data Engineering?

- A part of the big data ecosystem
- Closely linked to data science
- **Extract** - Query data from a source
- **Transform** - Perform some modifications to the data
- **Load** - Put that data in a location

What is Data Engineering?

Example An online retailer has a website where you can purchase widgets in a variety of colors. The website is backed by transactional databases. There may be a database at different geographical locations. How many blue widgets did the retailer sell in the last quarter?

To answer the preceding question, a data engineer would:

- Create connections to all of the transactional databases
- Extract the data
- Load it into a data warehouse.
- Count the number of all the blue widgets sold.

What is Data Engineering?

To answer the preceding question, a data engineer would:

- Create connections to all of the transactional databases
- Extract the data from each database
- Add a field to tag the location for each transaction in the data
- Transform the date from local time to ISO 8601
- Load the data into the data warehouse
- Count the number of all the blue widgets sold

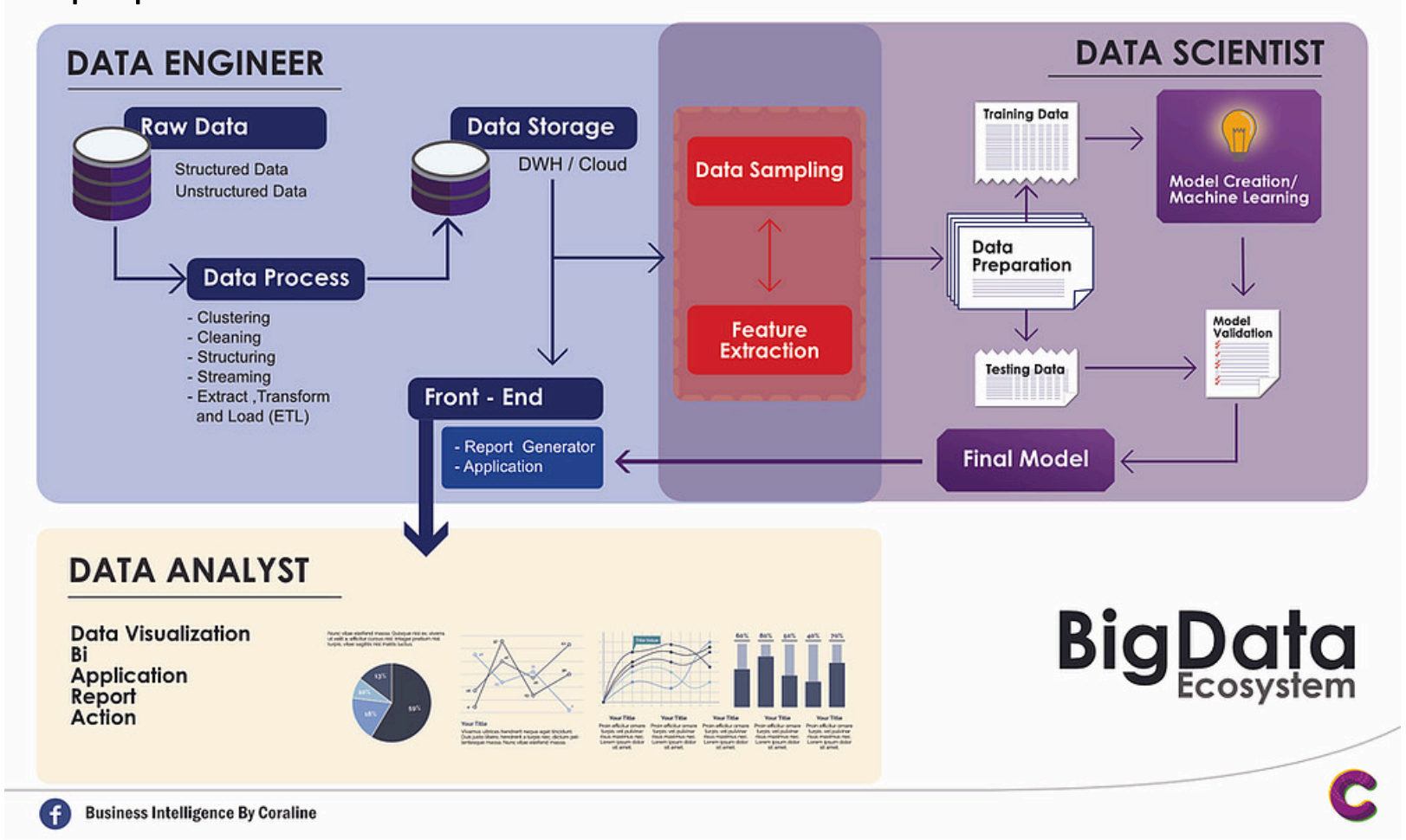
Required Skills and Knowledge

- How to extract data from files in different formats or different types of databases.
  - several languages used to perform many different tasks, such as SQL and Python
- Data modeling and structures
- Understand the business and what knowledge and insight they are hoping to extract from the data
- Data Warehouse, Data Lake
- How to manage servers, as well as how to install and configure software.
- Infrastructure on the cloud platform

Data Engineering & Data Science

Data engineers prepare data for data scientists.

Data scientists use the data for analysis.



Source: https://www.coraline.co.th/single-post/data-engineer-in-organization

Data Engineering & Data Science

- **Data engineers** need to understand
  - data formats
  - data flow
  - data structures
  - data models
  - server and security

  to efficiently transport data.
- **Data scientists** utilize them for building statistical models and mathematical computation.

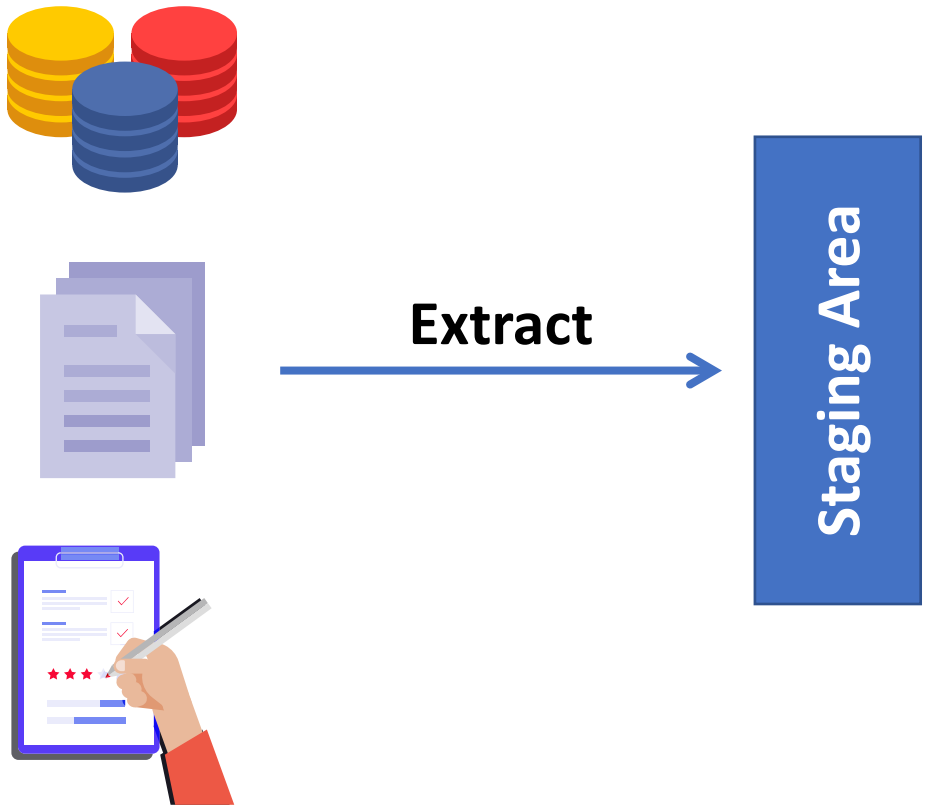
Data Engineering Process

Extract, Transform, Load (ETL) Process is an automated process which include:

1. Gathering raw data
2. Extracting information needed for reporting and analysis
3. Cleaning standardizing, and transforming data into usable format
4. Loading data into a data repository

Data Engineering Process

Extract, Transform, Load (ETL) Process is an automated process which include:

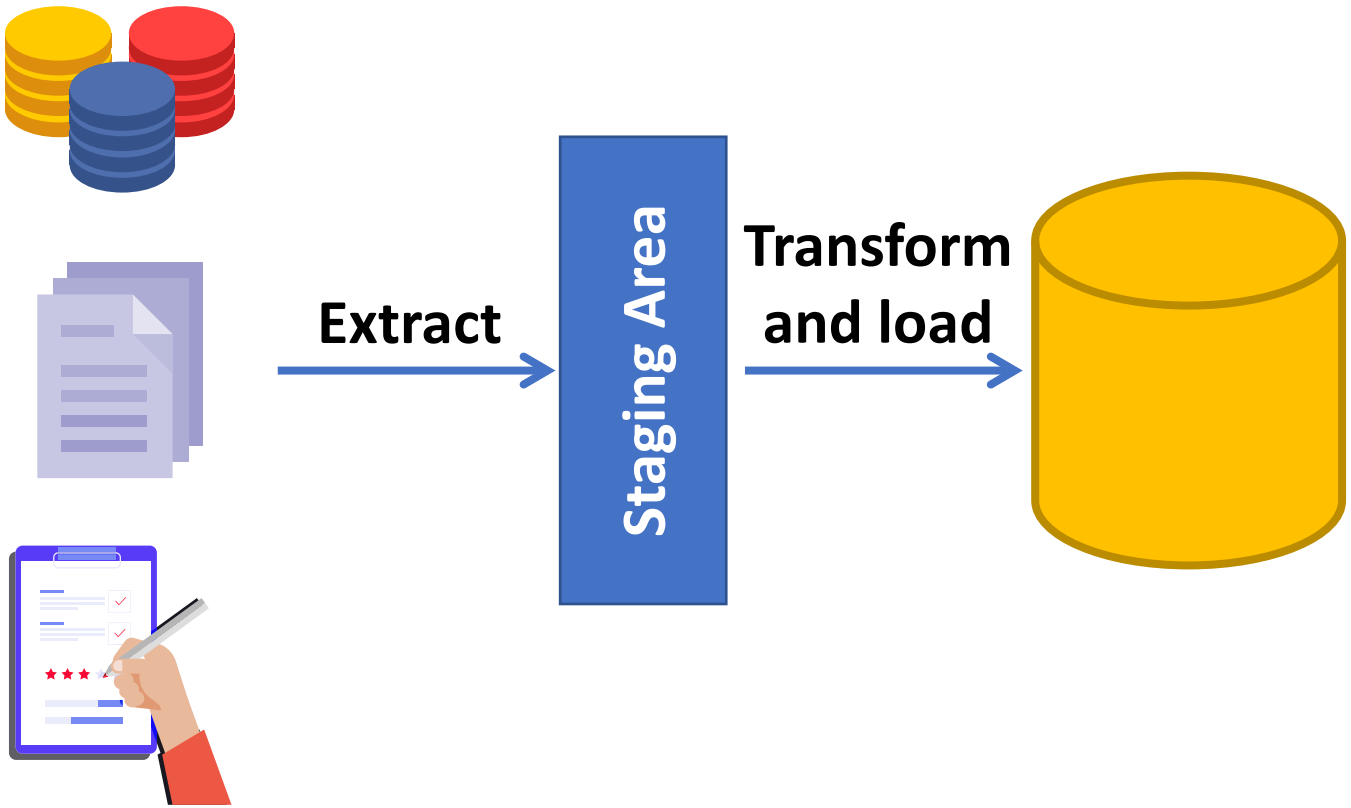


Extraction can be through:

- **Bath processing** – large chunks of data moved from source to destination at scheduled intervals.
- **Stream processing** – data pulled in real-time from source, transformed in transit, and loaded into data repository

Data Engineering Process

Extract, Transform, Load (ETL) Process is an automated process which include:

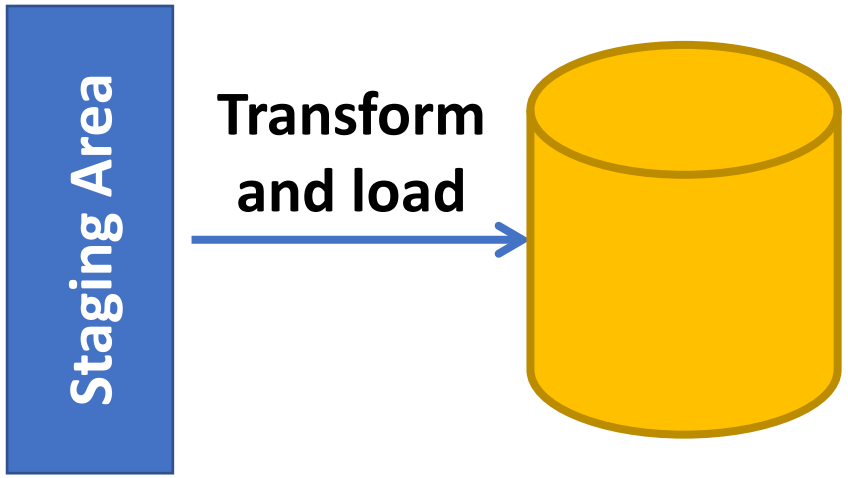


Transforming data:

- Standardizing data formats and units of measurement
- Removing duplicate data
- Filtering out data that is not required
- Fill missing data
- Enriching data
- Establishing key relationships across tables
- Applying business rules and data validations

Data Engineering Process

Extract, Transform, Load (ETL) Process is an automated process which include:

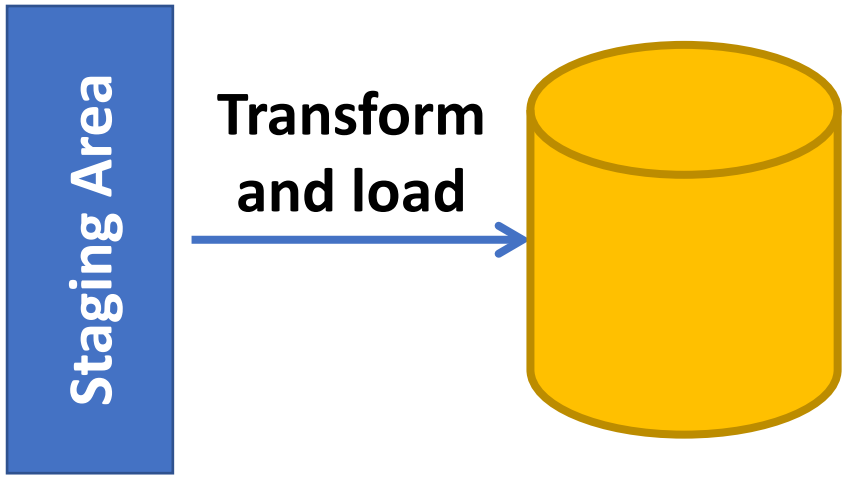


Loading is the transportation of processed data into a data repository. It can be

- Initial loading – populating all of the data in the repository
- Incremental loading – applying updates and modifications periodically
- Full refresh – erasing a data table and reloading fresh data

Data Engineering Process

Extract, Transform, Load (ETL) Process is an automated process which include:

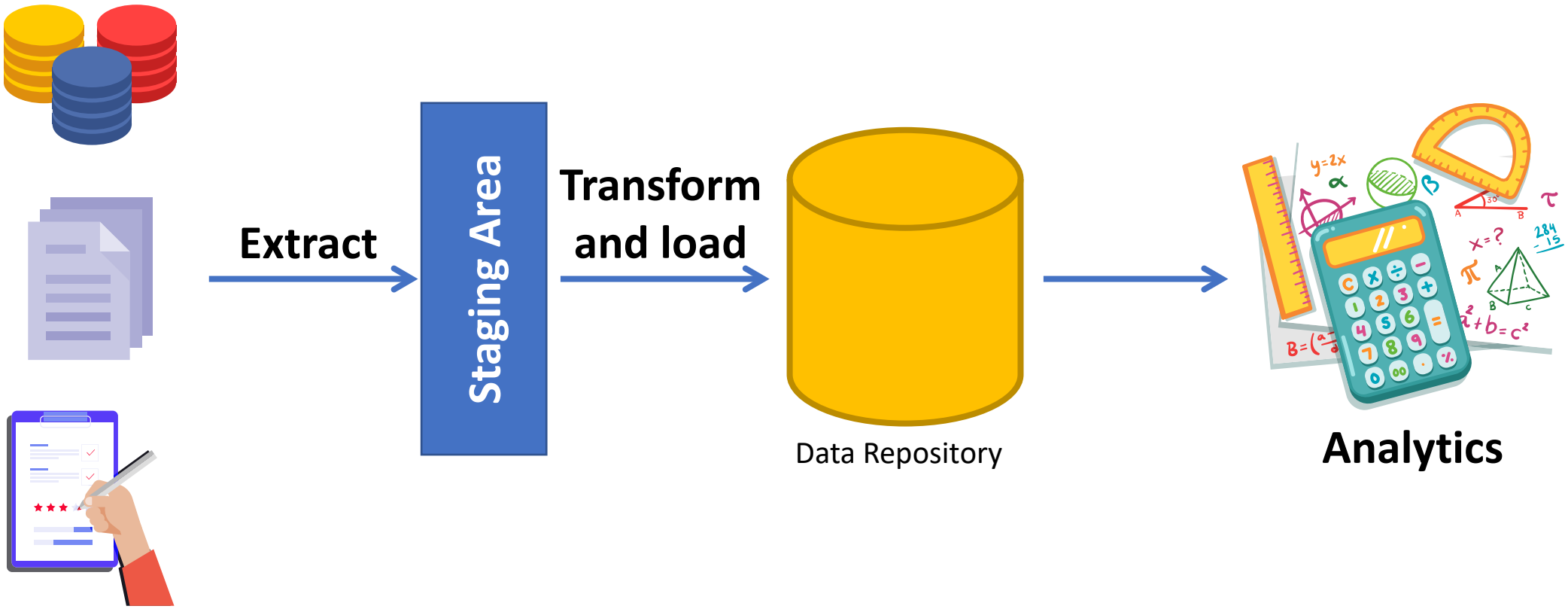


Load verification includes checks for:

- Missing or null values
- Server performance
- Load failures

Data Engineering Process

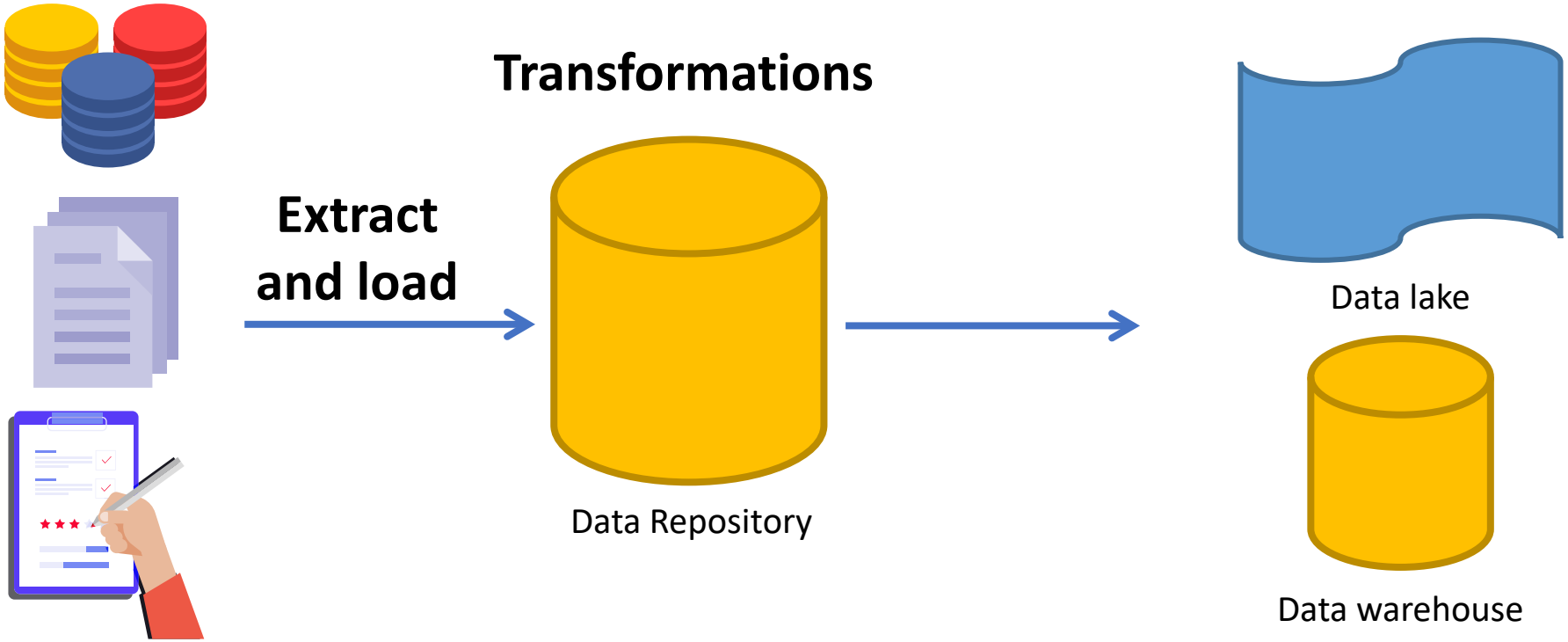
Extract, Transform, Load (ETL) Process is an automated process which include:



Data Engineering Process

Extract, Load, Transform (ELT) Process

- Helps process large sets of unstructured data and non-relational data
- Is ideal for data lakes



Data Engineering Process

Advantages of ELT process

- Shortens the cycle between extraction and delivery
- Allows you to ingest volumes of raw data as immediately as the data becomes available
- Affords greater flexibility to analysts and data scientists for exploratory data analysis
- Transforms only that data which is required for a particular analysis so it can be leveraged for multiple use cases
- Is more suited to work with Big Data

Data Engineering Process

Data Pipeline

- Encompasses the entire journey of moving data from one system to another, including the ETL/ELT process
- Combining a transactional database, a programming language, a processing engine, and a data warehouse results in a pipeline.
- Can be used for both batch and streaming data
- Supports both long-running batch queries and smaller interactive queries
- Typically loads data into a data lake but can also load data into a variety of target destinations – including other application and visualization tools

References

- Paul Crickard (2020). *Data Engineering with Python*. Birmingham, UK: Packt Publishing Ltd.
- https://www.coursera.org/lecture/introduction-to-data-engineering/etl-elt-and-data-pipelines-xbxfN