*Phạm Thu Thuy – 21002240*

**HOMEWORK**

***Ex1: Online Extraction vs Offline Extraction. Example?***

**Online Extraction:**

Definition: Online extraction involves retrieving data or information in real-time directly from its source, typically over the internet or a network connection. This method is used when data needs to be continuously updated or when immediate access to the most current data is necessary.

Example: Streaming data from social media platforms like Instagram or Facebook to analyze trends or sentiments in real-time.

**Offline Extraction:**

Definition: Offline extraction involves retrieving data from a source that is not actively connected or updated in real-time. Instead, data is gathered from stored sources, such as databases, files, or archives.

Example: Extracting sales data from a company's database at the end of each month to generate monthly reports.

***Ex2: Motivation of ELT model. What's the difference between ETL and ELT?***

**Motivation of ELT Model:**

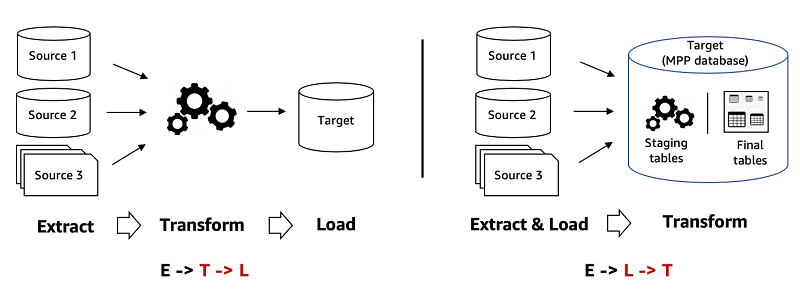
Scalability: ELT models often leverage cloud-based services to handle large volumes of data efficiently.

Cost-Effectiveness: ELT models can take advantage of cloud computing platforms where resources can be provisioned and scaled dynamically based on demand. The pay-as-you-go model of cloud computing platforms can be more cost-effective than investing in on-premises infrastructure.

Data Agility: ELT allows organizations to load raw data into storage systems quickly and then perform transformations on the data as needed. This agility enables faster insights and adaptability to changing business requirements.

Real-Time Analytics: ELT models can facilitate near-real-time or real-time analytics on raw databy loading data quickly and then performing transformations on the fly, enabling faster decision-making.

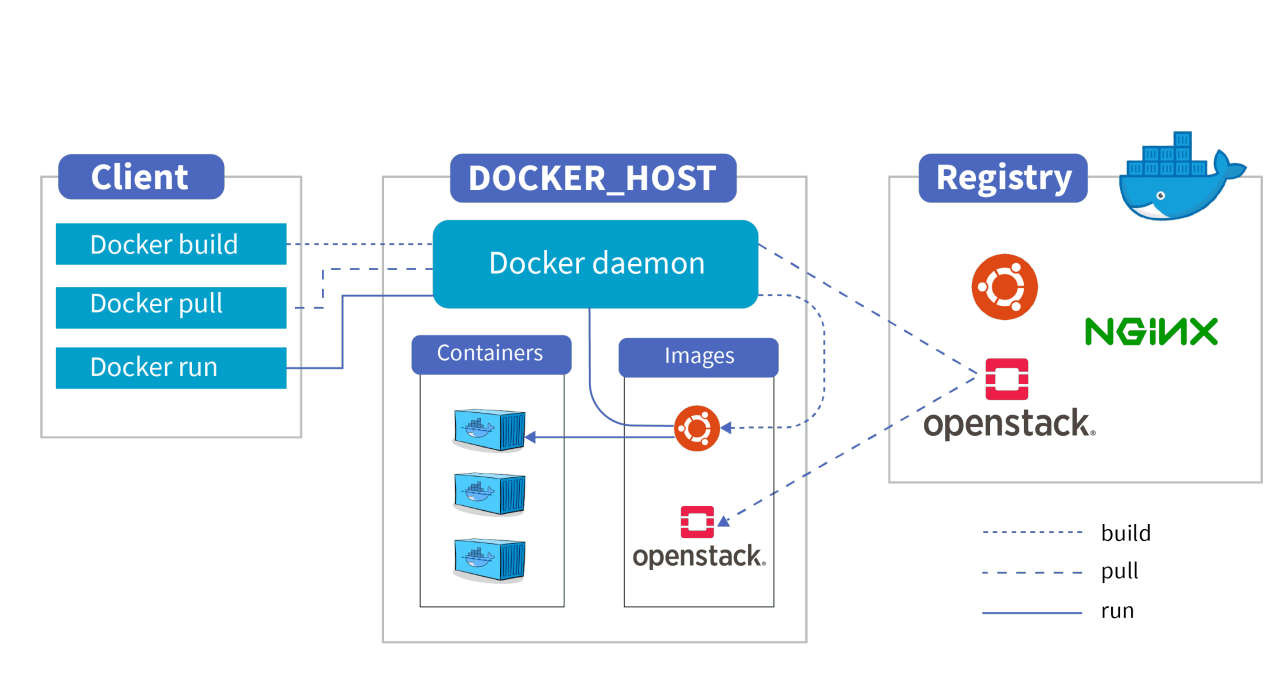
**Difference between ETL and ELT:**



|  |  |  |
| --- | --- | --- |
| **Category** | **ETL (**Extract, convert and load) | **ELT** (Extract, load and convert) |
| Process | 1. Extract raw data from various sources  2. Use a secondary processing server to convert the data  3. Load that data into the target database | 1. Extract raw data from various sources  2. Load data in its native state into a data warehouse or data lake  3. Convert data if necessary while in the target system |
| Procedure | The conversion process takes place in a secondary processing server. | The transformation takes place in the target data warehouse. |
| Switch Position and load | Best suited for structured data. | Can process structured, unstructured, and semi-structured data. |
| Speed | ETL is slower than ELT. | ELT is faster than ETL because it can use the data warehouse's internal resources. |
| Expense | Can be time consuming and costly to set up, depending on the ETL tools used. | More cost-effective, depending on the ELT infrastructure used. |
| Security | It may be necessary to build custom applications to meet data protection requirements. | You can use the built-in features of the target database to manage data protection. |

***Ex3: Investigate the overall architecture of Docker; The differences between Docker Container and Docker Image; How to build a docker image; How to run a container from a self-built docker image; How to run a container from docker hub image.***

Overall Architecture of Docker:



- Docker uses a client-server architecture.

- The Docker client communicates with the Docker daemon, which does the heavy lifting of building, running, and distributing Docker containers.

- The Docker daemon manages Docker objects such as images, containers, networks, and volumes.

- Docker images are stored in a registry, such as Docker Hub or a private registry.

- Containers are instances of Docker images running as isolated processes on the host machine.

Differences between Docker Container and Docker Image:

- Docker Image: A Docker image is a read-only template that contains a set of instructions for creating a Docker container. It includes everything needed to run an application: the code, runtime, libraries, environment variables, and configuration files.

- Docker Container: A Docker container is a runnable instance of a Docker image. It represents a lightweight and portable encapsulation of an application environment, including its code and all dependencies.

Building a Docker Image:

Build a Docker image:

1. Create a Dockerfile: This file contains instructions for building the Docker image: - specifying a base image

- copying files into the image

- installing dependencies

- defining commands to run when the container starts.

1. Define commands: Use Dockerfile instructions like “RUN”, “COPY”, and “CMD” to configure the image.
2. Build the image: Use the “docker build” command to build the image from the Dockerfile.

Running a Container from a Self-built Docker Image: using the “docker run” command followed by the image name.

- Ex: docker run <image\_name>

Running a Container from Docker Hub Image:

- Docker Hub is a public registry of Docker images.

- To run a container from an image on Docker Hub, use the `docker run` command followed by the image name.

- If the image is not available locally, Docker will automatically pull it from Docker Hub before running the container.

- For example: docker run <username>/<image\_name>

***Ex5: Study SQL, DDL, DML, DQL.***

\* SQL (Structured Query Language) is a standard language for managing relational databases.

- SQL is widely used in various applications and industries where data management and manipulation are crucial, including web development, data analysis, business intelligence, and more. It is supported by most relational database systems such as MySQL, PostgreSQL, SQL Server, Oracle, and SQLite.

- It's commonly divided into several categories: DDL, DML, DQL.

\* DDL (Data Definition Language): DDL is used to define the structure of the database. It includes commands such as:

- CREATE: Used to create database objects like tables, views, indexes, etc.

- ALTER: Modifies the structure of an existing database object.

- DROP: Deletes database objects like tables, views, indexes, etc.

- TRUNCATE: Deletes all records from a table, but keeps the table structure intact.

\* DML (Data Manipulation Language): DML is used to manipulate data within the database. It includes commands such as:

- SELECT: Retrieves data from one or more tables.

- INSERT: Adds new records into a table.

- UPDATE: Modifies existing records in a table.

- DELETE: Removes records from a table.

\* DQL (Data Query Language): DQL is a subset of SQL used specifically for querying data. It primarily consists of the SELECT statement, which is used to retrieve data from the database.

*These categories help organize SQL commands based on their primary purpose within the database management process.*