CIT650: Introduction to Big Data

Lab #4

Lab Goals

- Apache Hive on Docker
- Getting Started with Hive
- Creating and Managing Tables
- Querying Data with HiveQL
- Task

Apache Hive on Docker

- Clone the following GitHub repo. to get the docker-compose file of Apache Hive.
- Change directory to the repo. git clone https://github.com/big-data-europe/docker-hive
- Run the following command: docker-compose up -d.
- Once building is done, open the hive server container using the following command: docker exec -it docker-hive-hive-server-1 bash

Getting Started with Hive

- Launch the Hive CLI.
- Show databases.
- Create a Hive database.
- Switch to the newly created database.
- Show tables of a database.

```
cd hive/bin
hive
```

Creating and Managing Tables

- Move the data file to the container using docker cp.
- Create a Hive table with appropriate data types.
- Load data into the table.
- Describe the table structure.
- Verify the data in the table.

```
docker cp auto-mpg.csv
docker-hive-hive-server-1:/home
```

Querying Data with HiveQL [Examples]

- Filtering using WHERE clause
- Aggregate using GROUP BY
- Aggregation with Filtering
- Grouping and Counting
- Sorting with ORDER BY
- Partitioning
- Bucketing
- Subquery with IN
- Conditional Aggregation with CASE
- Date Transformation - Convert Model Year to Production Year

Lab Task

- Refer to the **Task** directory in **Lab #4 Material** on Moodle/Drive
- Deadline: Wednesday, 24 April. 2024 @ 11:59 PM.

Thanks