The Sqoop Challenge

Overview

- Create a process to mode data from an MSSQL database deployed on Linux into HDFS & Hive using Sqoop.
- **The challenge involves the following software:**
    - Linux
    - MSSQL
    - Sqoop
    - HDFS
    - Hive
- It is recommended to perform all of the challenge operations in the course VM/Docker.
- The challenge is based on the Sqoop tutorial examples and holds both infrastructure elements (Linux) as well as logical elements (SQL).

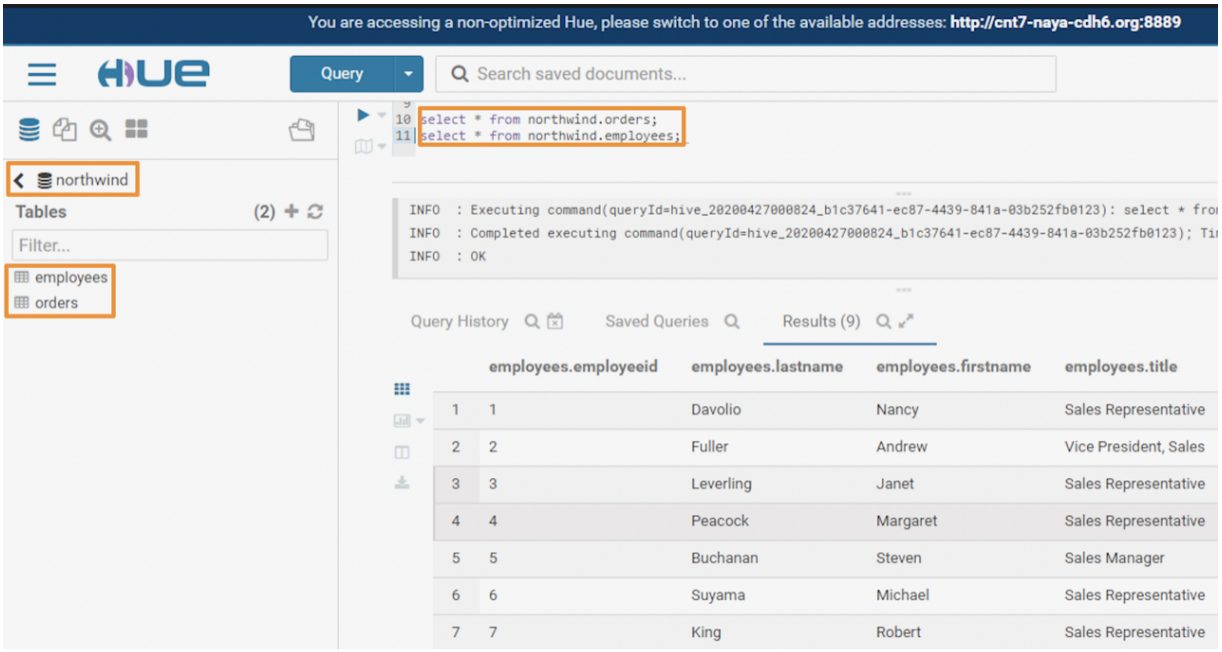
Task Main Steps:

- Install MSSQL (version 2019 was tested and verified) on Linux operating system - use the course VM/Docker).
- Load the Northwind sample database into MSSQL installed on Linux.
    - Data can be loaded via running a SQL file or performing restore.
    - Using Adventureworks is also an option.
- Create a Sqoop process to import the MSSQL sample database tables into HDFS and Hive.
- Perform data validation by retrieving the same data from MSSQL and Hive (simple SQL queries should be sufficient).
- At the end of the task remove the MSSQL installation as it will use some of VM/Docker resources that might cause performance issues later in the course.

Notes:

- Although this challenge holds only a few main steps, it hides many challenges.
- In case that the VM/Docker will be corrupted during the process (e.g. critical OS files were deleted by mistake), recover the environment by re-deploying:
    - **VM** - Open a new VM from the extracted .rar file.
    - **GCP Image** - Re-launch the GCE (Compute Engine - VM).
    - **Docker** (Local/GCP) - Delete the corrupted container and load the Course Docker image again.

Screenshot - MSSQL Northwind Data on HDFS & Hive:



Have Fun