

BÁO CÁO BÀI TẬP THỰC HÀNH

Môn học: Big Data (Chương trình SIC của Samsung)

The logo for Samsung Innovation Campus, featuring the words "Samsung", "Innovation", and "Campus" stacked vertically in a blue, sans-serif font, centered within a light gray square.

**Samsung
Innovation
Campus**

Sinh viên thực hiện: Nguyễn Thị Thu Trang

Lớp: TLU_BD01

Mục lục

I. Tổng quan về cách giải quyết bài tập	3
II. Kết quả	3
Câu 1: Tạo sơ đồ quan hệ ERD và cơ sở dữ liệu trong hệ thống RDBMS – MariaDB	3
Câu 2: Sử dụng Sqoop để nhập dữ liệu từ RDBMS sang hệ thống HDFS	5
Câu 3: Sử dụng Hive để tạo bảng với định dạng Parquet	9
Câu 4 & 5: Phân tích dữ liệu và lưu kết quả	11
Lời cảm ơn	13

Danh mục hình ảnh

Hình 1: Ảnh truy vấn SQL tạo bảng	4
Hình 2: Ảnh sơ đồ ERD	4
Hình 3: Ảnh quá trình import dữ liệu	5
Hình 4: Ảnh quá trình khởi động Docker	6
Hình 5: Ảnh khởi động container thành công	6
Hình 6: Ảnh lệnh thực thi import bảng employees	7
Hình 7: Ảnh dữ liệu bảng employees được chuyển thành công	7
Hình 8: Ảnh lệnh import bảng salaries	8
Hình 9: Ảnh dữ liệu bảng employees được chuyển thành công	8
Hình 10: Ảnh kiểm tra thư mục chứa file parquet	9
Hình 11: Ảnh lệnh tạo cơ sở dữ liệu Hive	9
Hình 12: Ảnh lệnh tạo bảng employees	9
Hình 13: Ảnh lệnh tạo bảng salaries	10
Hình 14: Ảnh bảng employee đã được map	10
Hình 15: Ảnh bảng salaries đã được map	10
Hình 16: Ảnh kiểm tra dữ liệu trong bảng salaries	11
Hình 17: Ảnh lệnh truy vấn và ghi file	11
Hình 18: Ảnh kết quả đã truy vấn và lưu thành công	12

I. Tổng quan về cách giải quyết bài tập

Câu 1: Làm việc với hệ quản trị cơ sở dữ liệu RDBMS (MariaDB)

Trong câu này, em đã thực hiện thao tác tạo cơ sở dữ liệu và các bảng trong hệ quản trị cơ sở dữ liệu MariaDB trên máy host. Để trực quan hóa dữ liệu và dễ dàng thiết kế sơ đồ quan hệ thực thể (ERD), em sử dụng công cụ DBeaver để kết nối đến MariaDB. Việc này hỗ trợ trực quan hóa cấu trúc bảng, khóa chính - khóa ngoại, và giúp xác định mối quan hệ giữa các bảng.

Câu 2, 3, 4: Di chuyển và phân tích dữ liệu bằng hệ thống Big Data

- Em đã sử dụng Docker với các image gồm: Sqoop, Hadoop, và Hive.
- Thực hiện kết nối từ MariaDB trên máy host tới container Docker để sử dụng Sqoop trích xuất dữ liệu từ RDBMS sang HDFS. Kết nối này được thực hiện thông qua JDBC.
- Dữ liệu sau khi được nhập vào hệ thống lưu trữ phân tán HDFS được lưu ở định dạng parquetfile nhằm tối ưu hóa việc lưu trữ và xử lý.
- Tiếp theo, em sử dụng Hive để tạo bảng và xử lý dữ liệu từ HDFS. Trong Hive, dữ liệu cũng được lưu trữ dưới định dạng parquetfile.
- Sau đó, thực hiện truy vấn trong Hive để phân tích dữ liệu, cụ thể: liệt kê các trường first_name, last_name, birth_date của những nhân viên có mức lương > 55000.

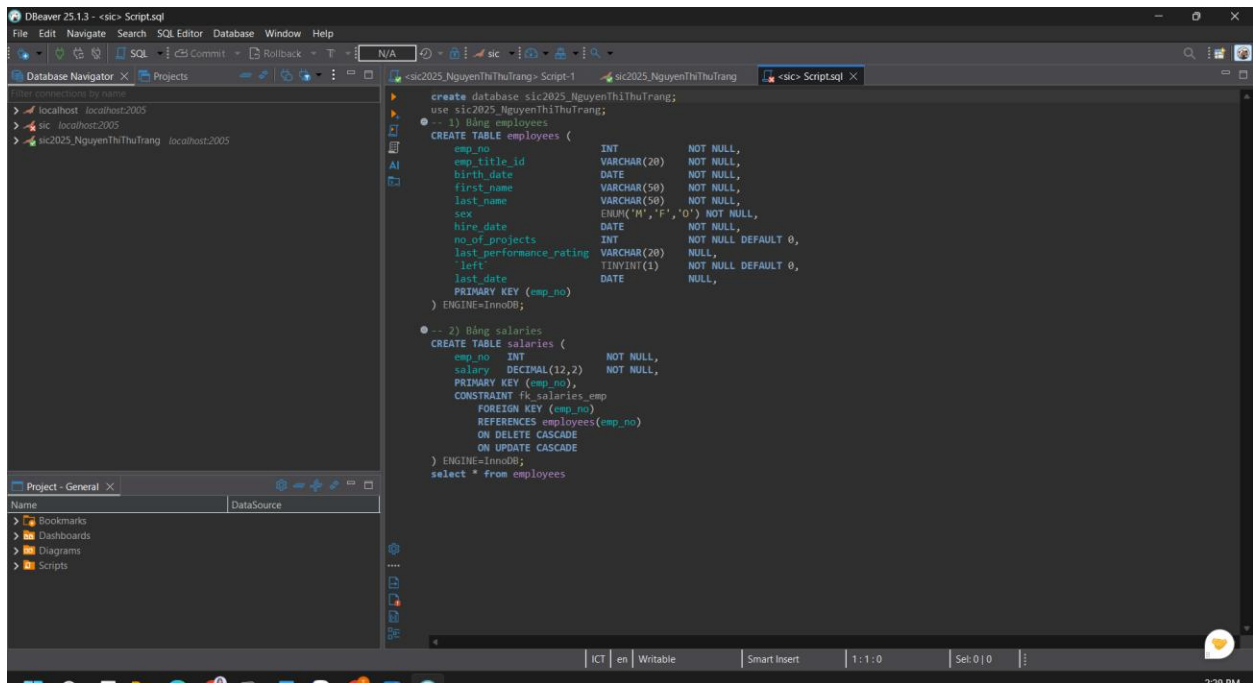
II. Kết quả

Toàn bộ mã nguồn và file kết quả em đã đẩy lên GitHub tại đường dẫn sau:
<https://github.com/trangtretrau2005/SIC2025.git>

Câu 1: Tạo sơ đồ quan hệ ERD và cơ sở dữ liệu trong hệ thống RDBMS – MariaDB

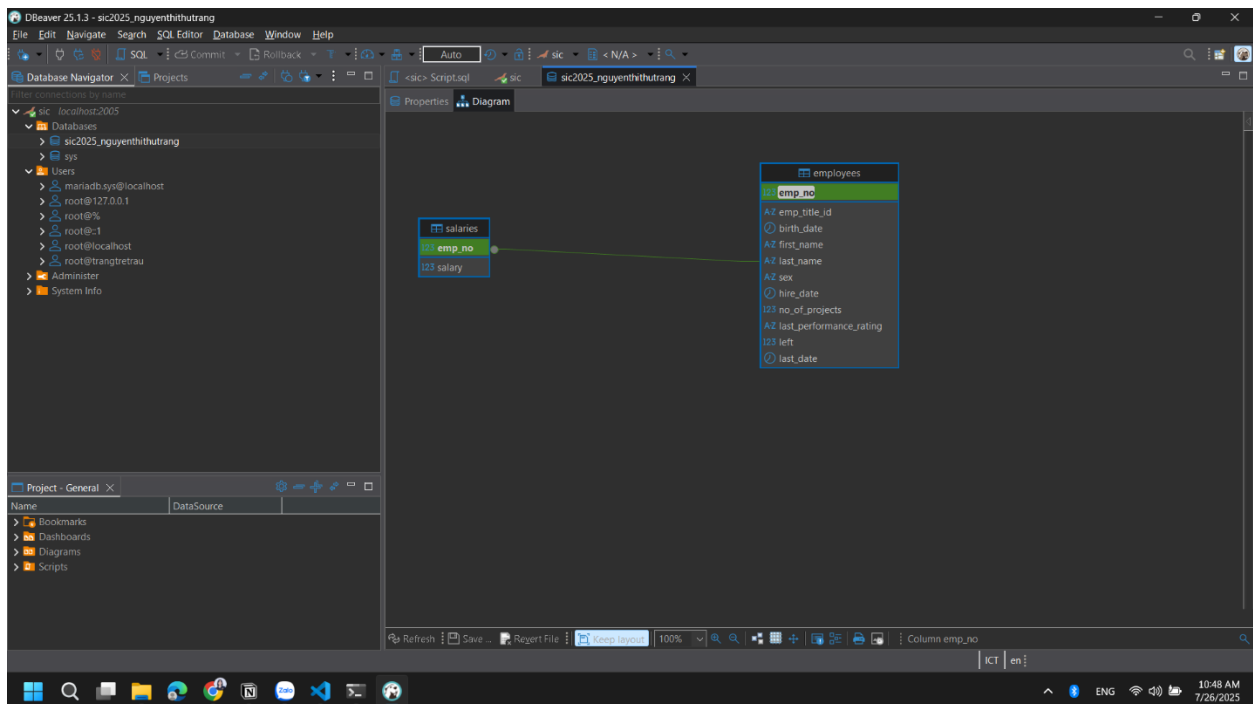
Bước 1: Tạo cơ sở dữ liệu và các bảng

- Thực hiện viết các câu lệnh SQL để:
 - Tạo cơ sở dữ liệu có tên sic2025_HoVaTenSinhVien
 - Tạo các bảng employees và salaries
 - Thiết lập các ràng buộc khóa chính - khóa ngoại giữa các bảng



Hình 1: Ảnh truy vấn SQL tạo bảng

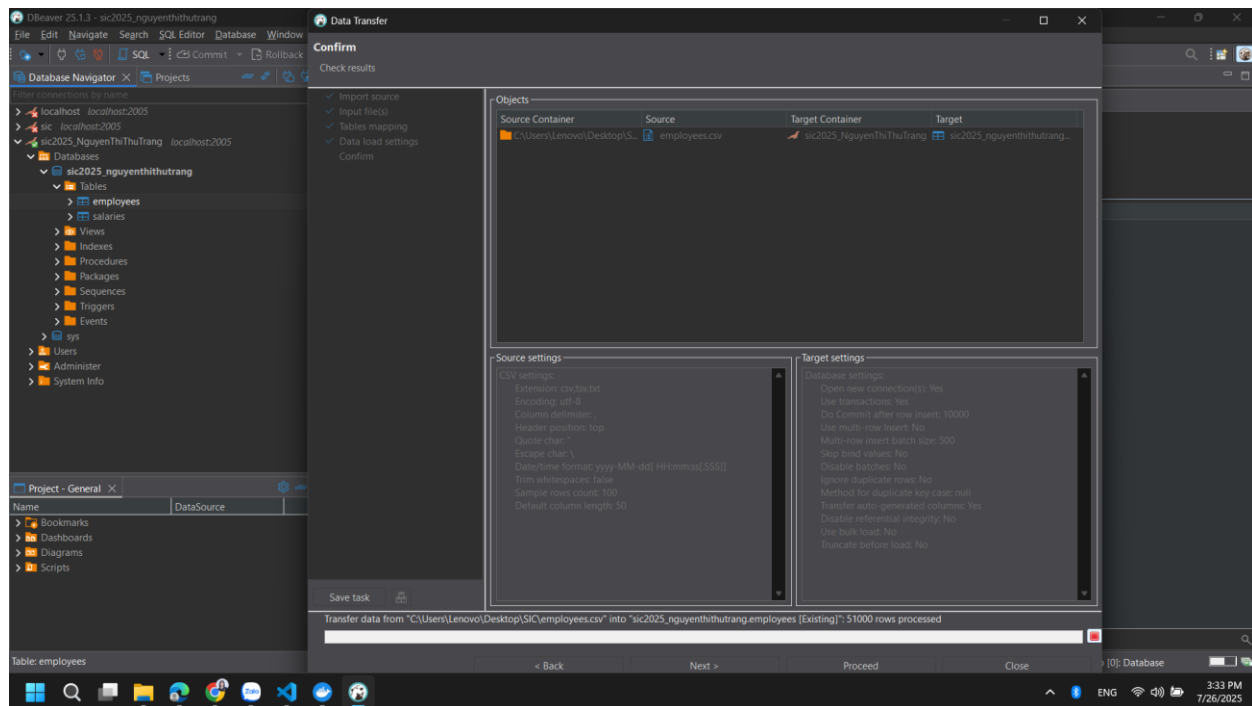
Kết quả: Sơ đồ ERD thể hiện mối quan hệ giữa các bảng được tạo thành công qua DBeaver.



Hình 2: Ảnh sơ đồ ERD

Bước 2: Nhập dữ liệu từ CSV vào MariaDB

- Sử dụng chức năng import CSV trên DBeaver để nạp dữ liệu từ employees.csv và salaries.csv vào 2 bảng tương ứng.



Hình 3: Ảnh quá trình import dữ liệu

Câu 2: Sử dụng Sqoop để nhập dữ liệu từ RDBMS sang hệ thống HDFS

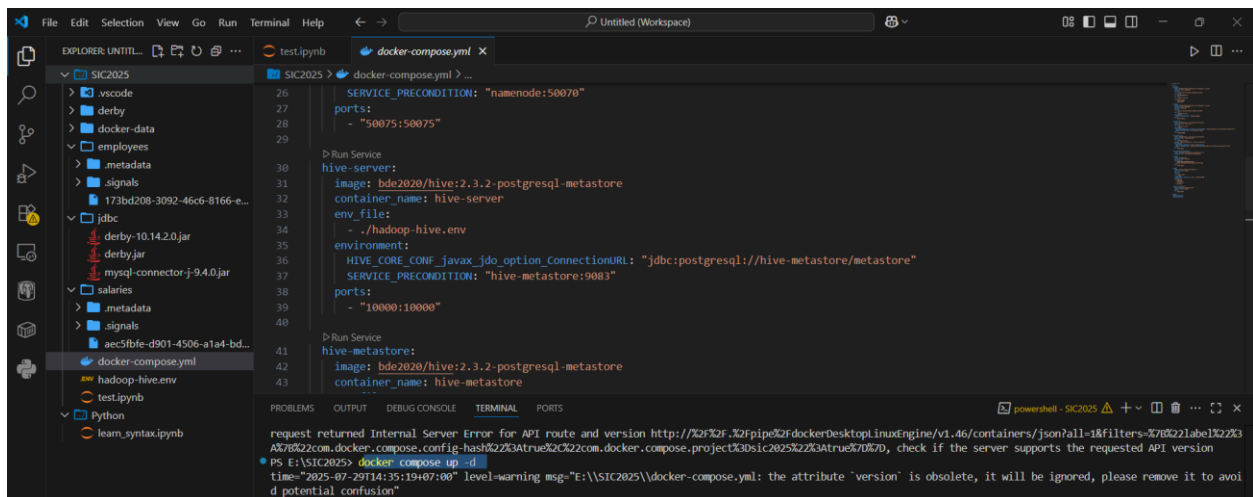
Đường dẫn lưu trữ HDFS:

/user/sic2025_HoVaTenSinhVien/hive/warehouse/Capstonesic2025_HoVaTenSinhVien

Định dạng tệp: Parquet

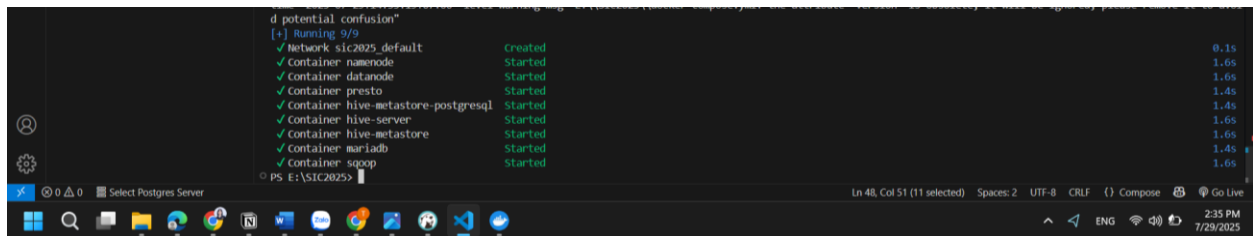
Bước 1: Khởi động hệ thống Big Data

- Viết file docker-compose.yaml cấu hình các image: Sqoop, Hadoop, Hive
- Thiết lập biến môi trường trong file .env
- Khởi động hệ thống bằng Docker



Hình 4: Ảnh quá trình khởi động Docker

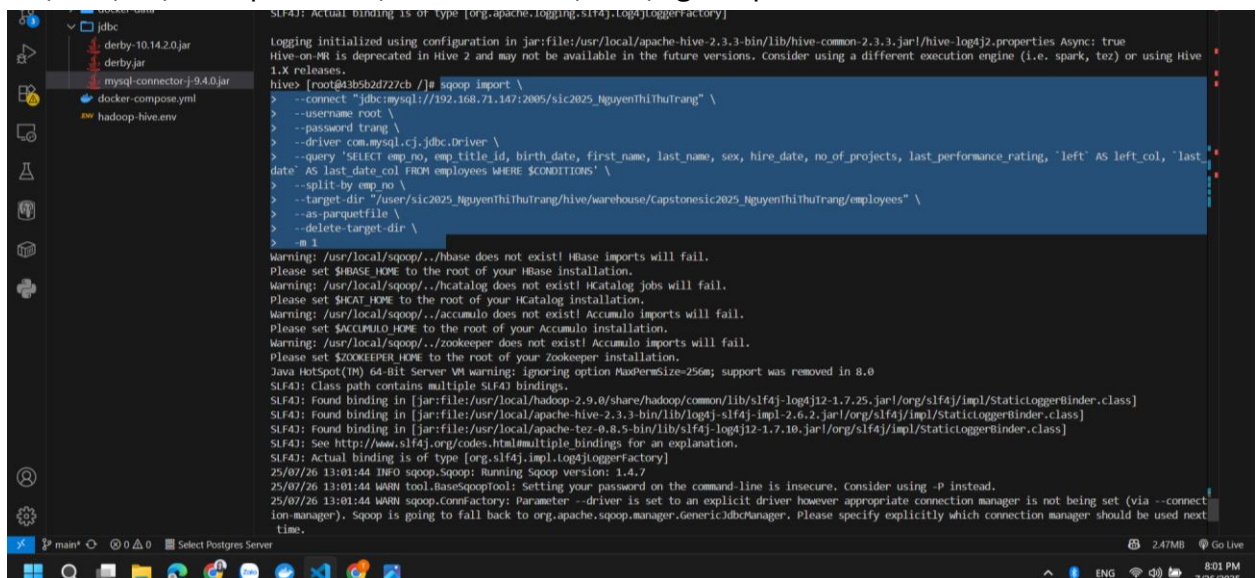
Kết quả: Docker container khởi chạy thành công



Hình 5: Ảnh khởi động container thành công

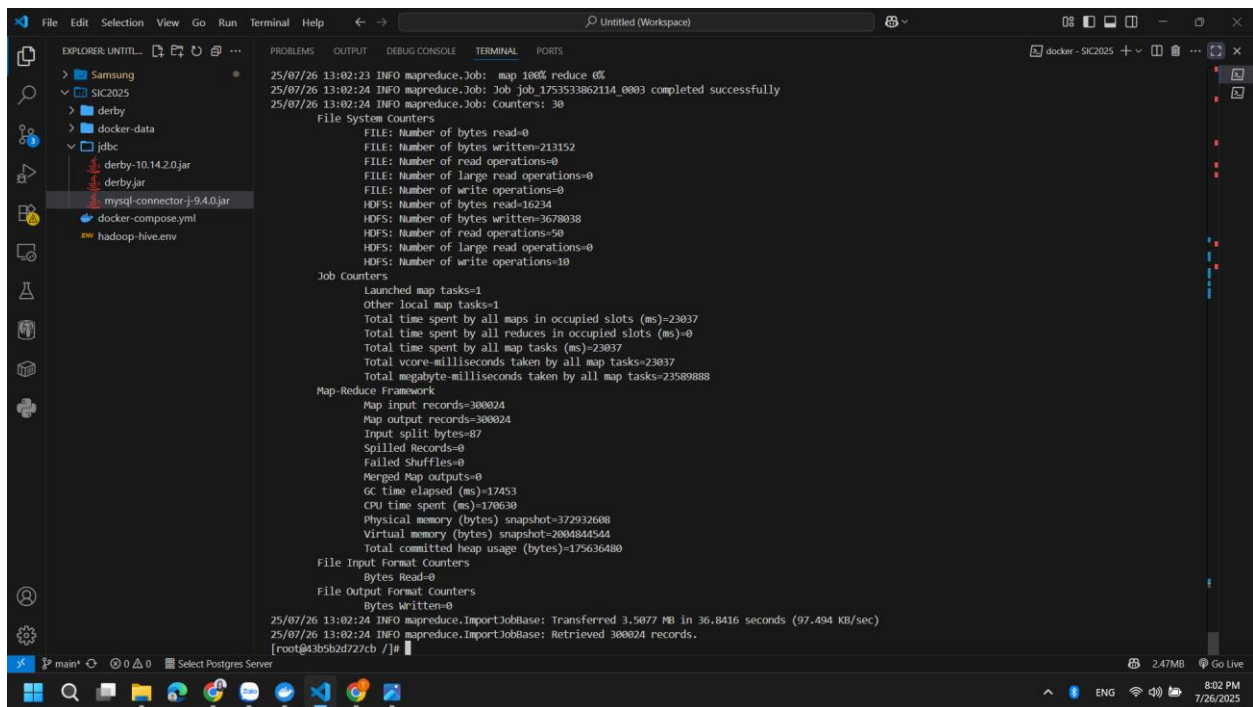
Bước 2: Dùng Sqoop để trích xuất bảng employees sang HDFS

- Kết nối từ Sqoop terminal tới MariaDB bằng JDBC connector
- Thực hiện lệnh import dữ liệu vào HDFS định dạng Parquet



Hình 6: Ảnh lệnh thực thi import bảng employees

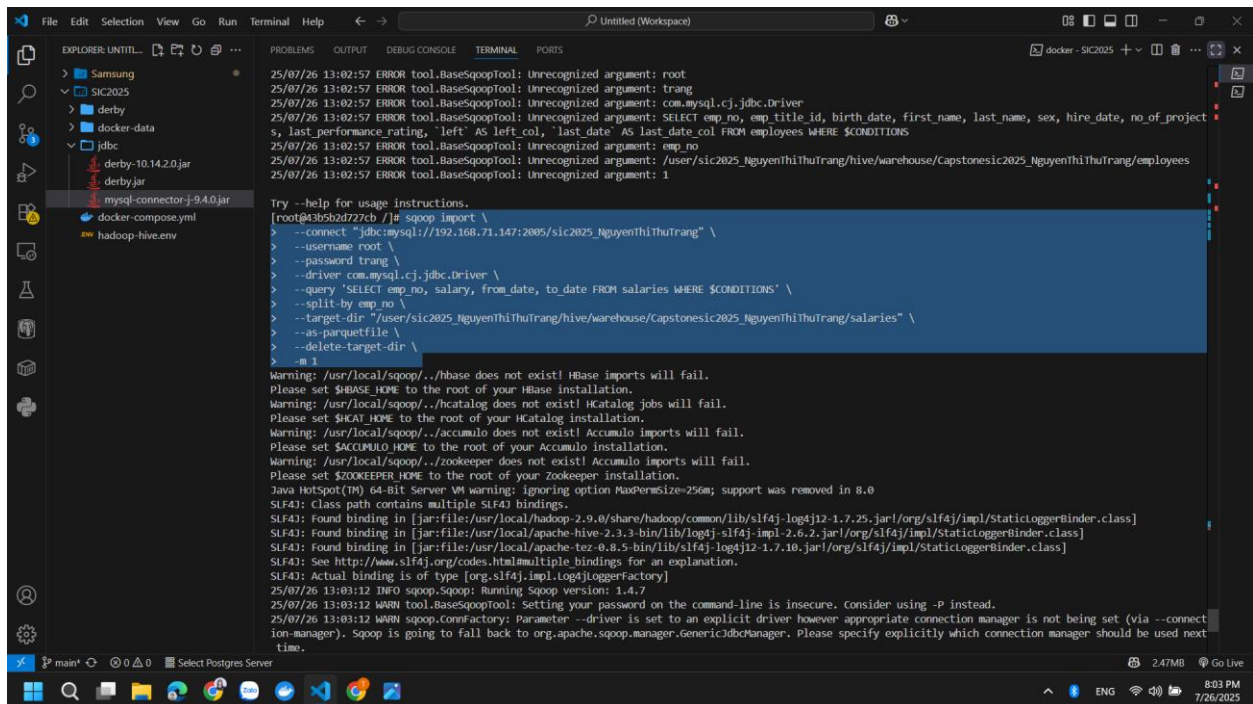
Kết quả: Dữ liệu bảng employees đã được chuyển thành công



```
25/07/26 13:02:23 INFO mapreduce.Job: map 100% reduce 0%
25/07/26 13:02:24 INFO mapreduce.Job: Job job_1753533862114_0003 completed successfully
25/07/26 13:02:24 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=213152
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=16234
  HDFS: Number of bytes written=3678938
  HDFS: Number of read operations=50
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=10
Job Counters
  Launched map tasks=1
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=23037
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=23037
  Total vcore-milliseconds taken by all map tasks=23037
  Total megabyte-milliseconds taken by all map tasks=23589888
Map-Reduce Framework
  Map input records=300024
  Map output records=300024
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=17453
  CPU time spent (ms)=170630
  Physical memory (bytes) snapshot=372932608
  Virtual memory (bytes) snapshot=2004844544
  Total committed heap usage (bytes)=175636480
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
25/07/26 13:02:24 INFO mapreduce.ImportJobBase: Transferred 3.5877 MB in 36.8416 seconds (97.494 KB/sec)
25/07/26 13:02:24 INFO mapreduce.ImportJobBase: Retrieved 300024 records.
[root@3b5b2d727cb /]#
```

Hình 7: Ảnh dữ liệu bảng employees được chuyển thành công

Bước 3: Tương tự với bảng salaries

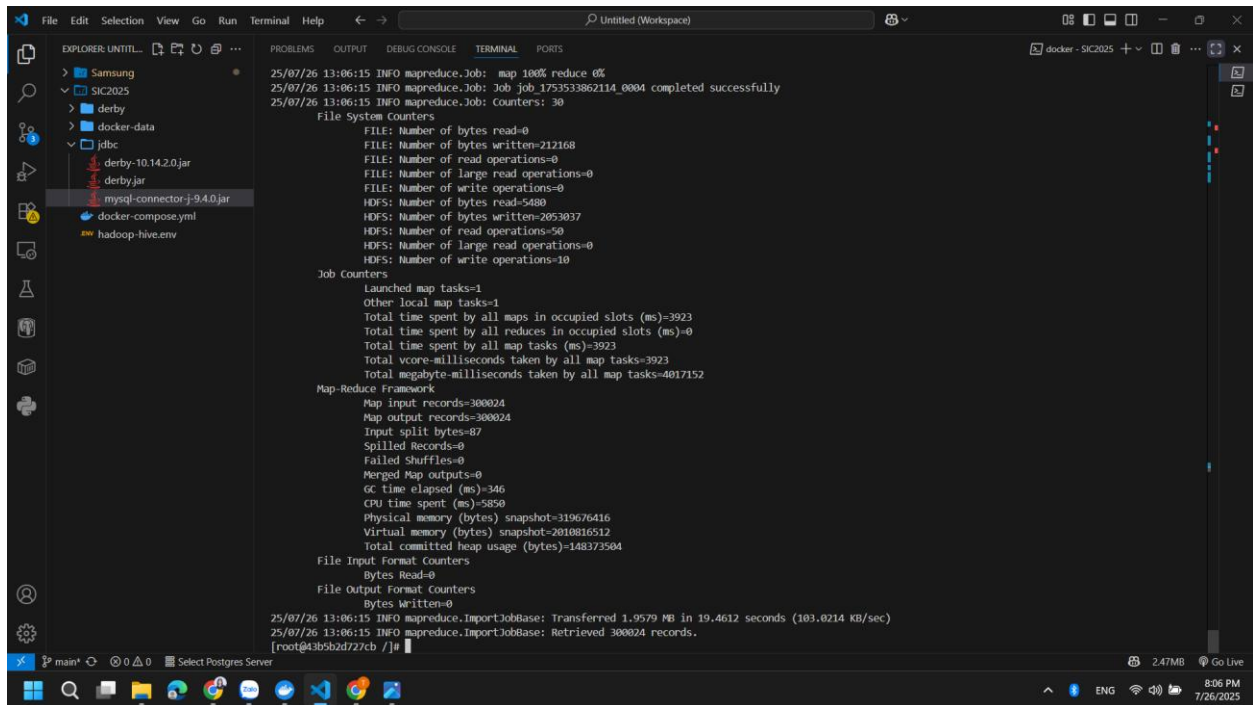


```
Try --help for usage instructions.
[root@3b5b2d727cb /]# sqoop import \
--connect 'jdbc:mysql://192.168.71.147:2005/sic2025_NguyenthithuTrang' \
--username root \
--password trang \
--driver com.mysql.cj.jdbc.Driver \
--query 'SELECT emp_no, salary, from_date, to_date FROM salaries WHERE $CONDITIONS' \
--split-by emp_no \
--target-dir '/user/sic2025_NguyenthithuTrang/hive/warehouse/Capstonesic2025_NguyenthithuTrang/salaries' \
--as-parquetfile \
--delete-target-dir \
-m 1

Warning: /usr/local/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/local/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=250m; support was removed in 8.0
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.9.0/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.3-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-tez-0.8.5-bin/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
25/07/26 13:03:12 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
25/07/26 13:03:12 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
25/07/26 13:03:12 WARN sqoop.ConnectionFactory: Parameter --driver is set to an explicit driver however appropriate connection manager is not being set (via --connect
ion-manager). Sqoop is going to fall back to org.apache.sqoop.manager.GenericJdbcManager. Please specify explicitly which connection manager should be used next
time.
```

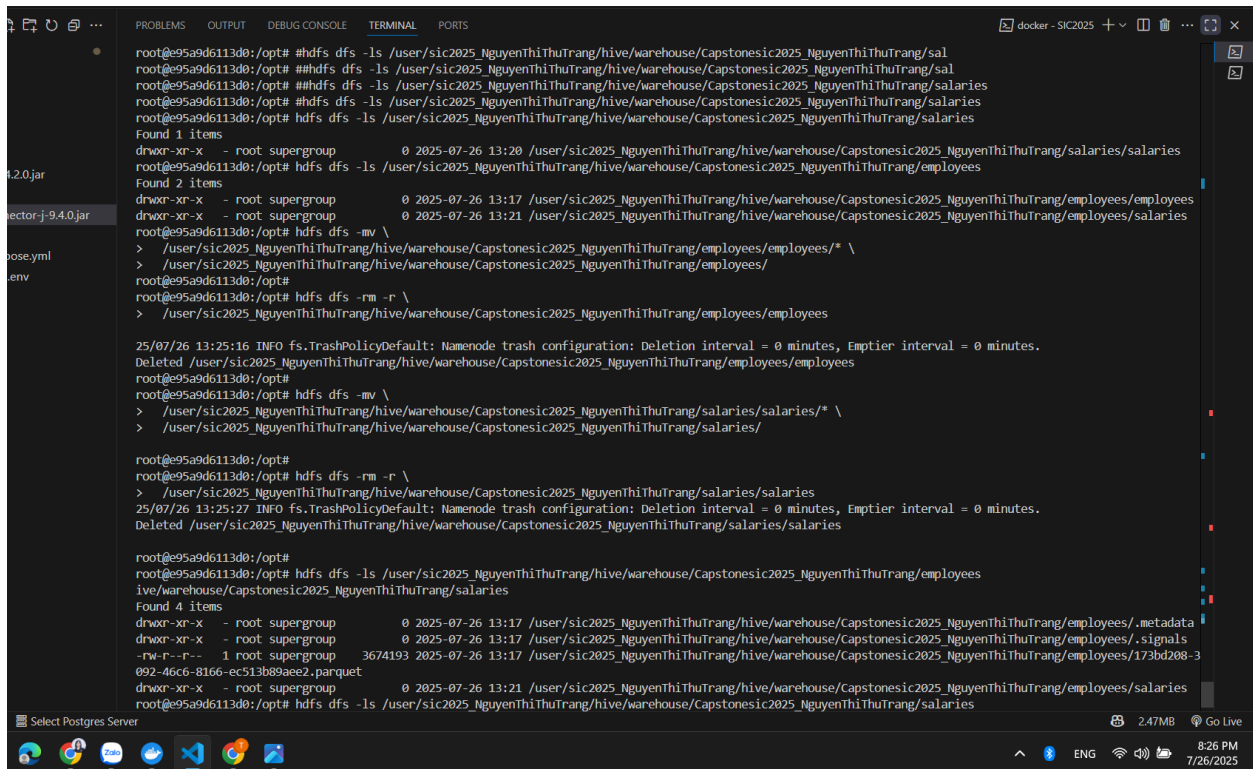

Hình 8: Ảnh lệnh import bảng salaries

Kết quả: Dữ liệu bảng salaries đã được import thành công vào HDFS



```
25/07/26 13:06:15 INFO mapreduce.Job: map 100% reduce 0%
25/07/26 13:06:15 INFO mapreduce.Job: Job job_1753533862114_0004 completed successfully
25/07/26 13:06:15 INFO mapreduce.Job: Counters: 30
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=212168
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=5480
  HDFS: Number of bytes written=2853937
  HDFS: Number of read operations=50
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=10
Job Counters
  Launched map tasks=1
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=3923
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=3923
  Total vcore-milliseconds taken by all map tasks=3923
  Total megabyte-milliseconds taken by all map tasks=4017152
Map-Reduce Framework
  Map input records=300024
  Map output records=300024
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=346
  CPU time spent (ms)=5850
  Physical memory (bytes) snapshot=319676416
  Virtual memory (bytes) snapshot=2010816512
  Total committed heap usage (bytes)=148373584
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
25/07/26 13:06:15 INFO mapreduce.ImportJobBase: Transferred 1.9579 MB in 19.4612 seconds (103.0214 KB/sec)
25/07/26 13:06:15 INFO mapreduce.ImportJobBase: Retrieved 300024 records.
[root@b3b2d727cb /]#
```

Hình 9: Ảnh dữ liệu bảng employees được chuyển thành công



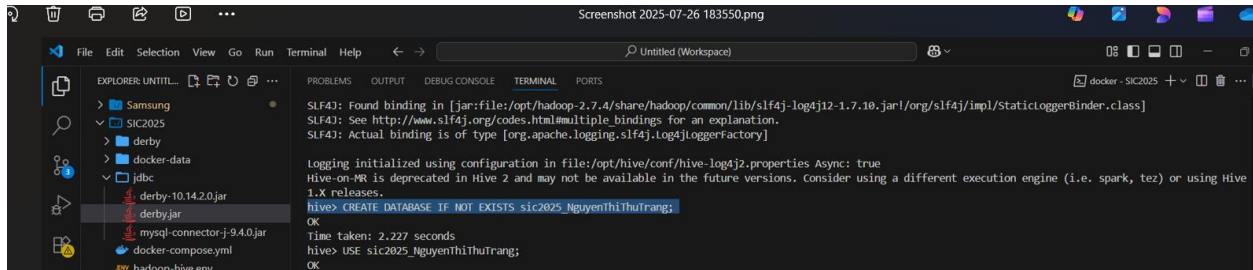
```
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/sal
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/sal
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries
Found 1 items
drwxr-xr-x - root supergroup 0 2025-07-26 13:20 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries/salaries
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees
Found 2 items
drwxr-xr-x - root supergroup 0 2025-07-26 13:17 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/employees
drwxr-xr-x - root supergroup 0 2025-07-26 13:21 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/salaries
root@e95a9d6113d0:/opt# hdfs dfs -mv \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/employees/* \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/
root@e95a9d6113d0:/opt# hdfs dfs -rm -r \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/employees
25/07/26 13:25:16 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/employees
root@e95a9d6113d0:/opt# hdfs dfs -mv \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries/salaries/* \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries/
root@e95a9d6113d0:/opt# hdfs dfs -rm -r \
> /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries/salaries
25/07/26 13:25:27 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries/salaries
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees
Found 4 items
drwxr-xr-x - root supergroup 0 2025-07-26 13:17 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/.metadata
drwxr-xr-x - root supergroup 0 2025-07-26 13:17 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/.signals
-rw-r--r-- 1 root supergroup 3674193 2025-07-26 13:17 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/173bd208-3
092-46c6-8166-ec513b89aee2.parquet
drwxr-xr-x - root supergroup 0 2025-07-26 13:21 /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/employees/salaries
root@e95a9d6113d0:/opt# hdfs dfs -ls /user/sic2025_NguyenThiThuTrang/hive/warehouse/Capstonesic2025_NguyenThiThuTrang/salaries
```


Hình 10: Ảnh kiểm tra thư mục chứa file parquet

Câu 3: Sử dụng Hive để tạo bảng với định dạng Parquet

Do dữ liệu đã được lưu trữ sẵn trong HDFS từ bước trước, các bảng Hive cần được ánh xạ chính xác đến các tệp tương ứng.

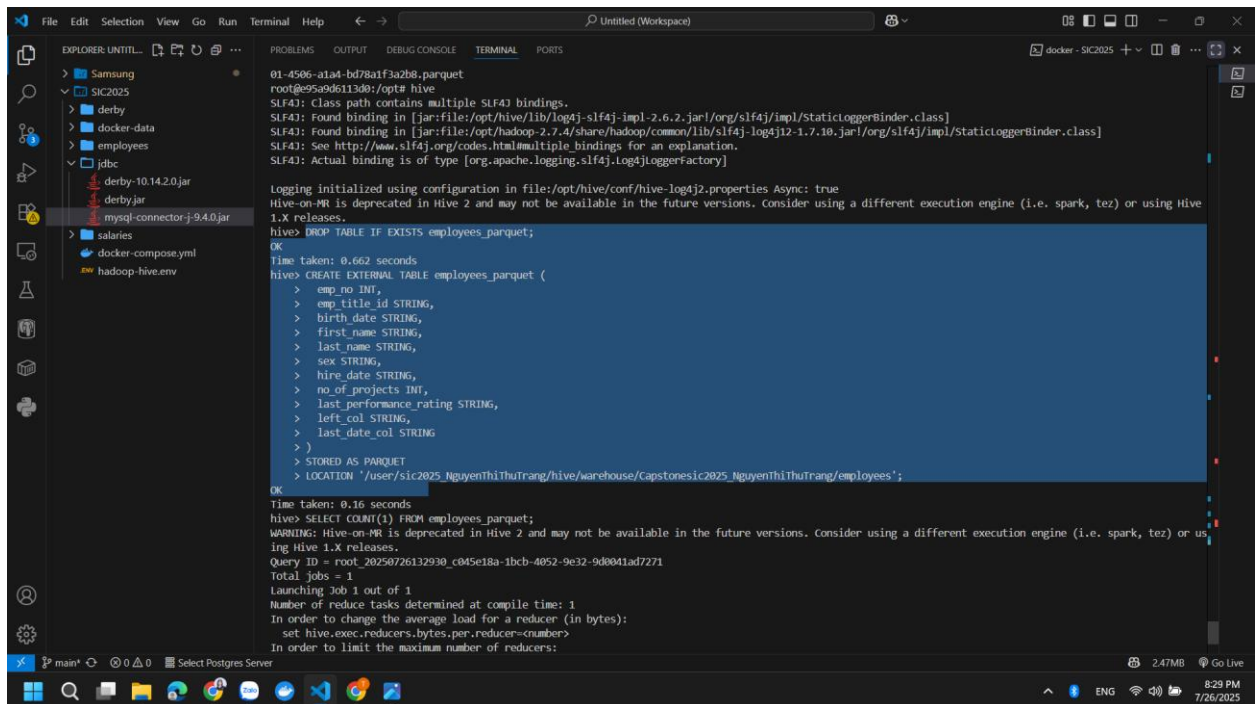
Bước 1: Tạo cơ sở dữ liệu trong Hive



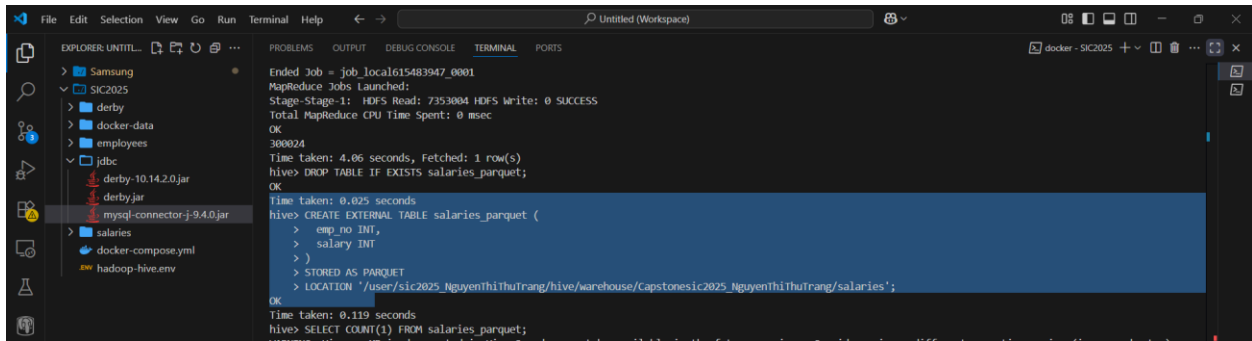
Hình 11: Ảnh lệnh tạo cơ sở dữ liệu Hive

Bước 2: Tạo bảng employees và salaries trong Hive

- Xác định đúng kiểu dữ liệu và đường dẫn HDFS tương ứng

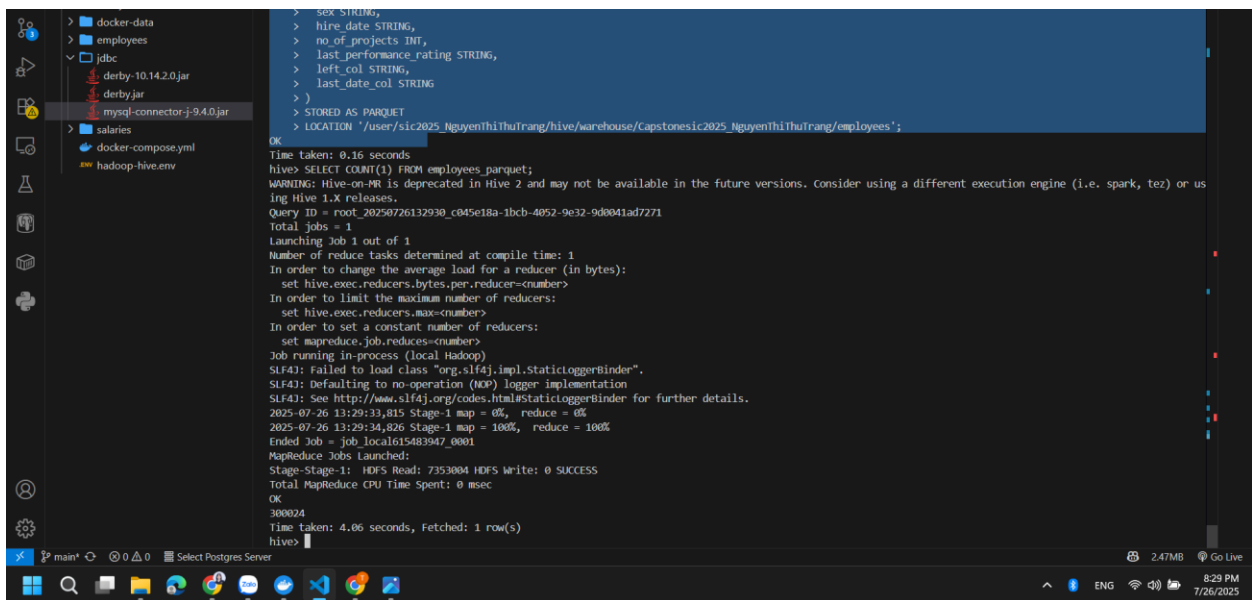


Hình 12: Ảnh lệnh tạo bảng employees

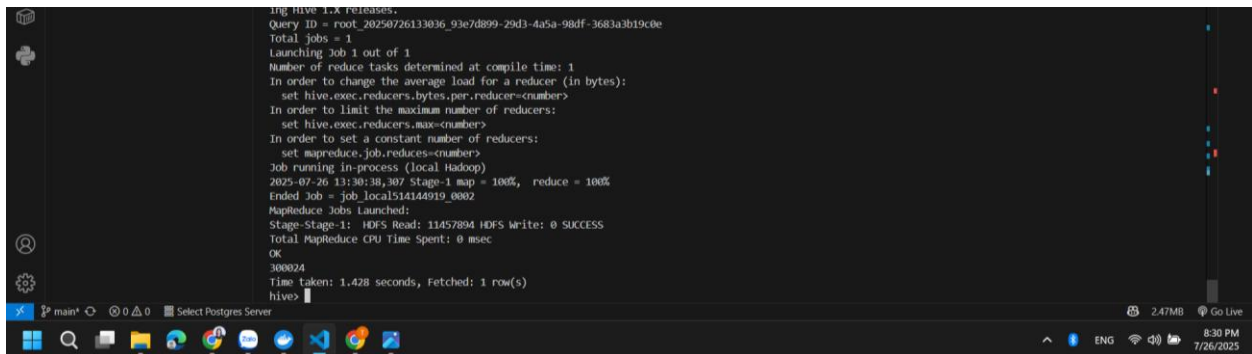


Hình 13: Ảnh lệnh tạo bảng salaries

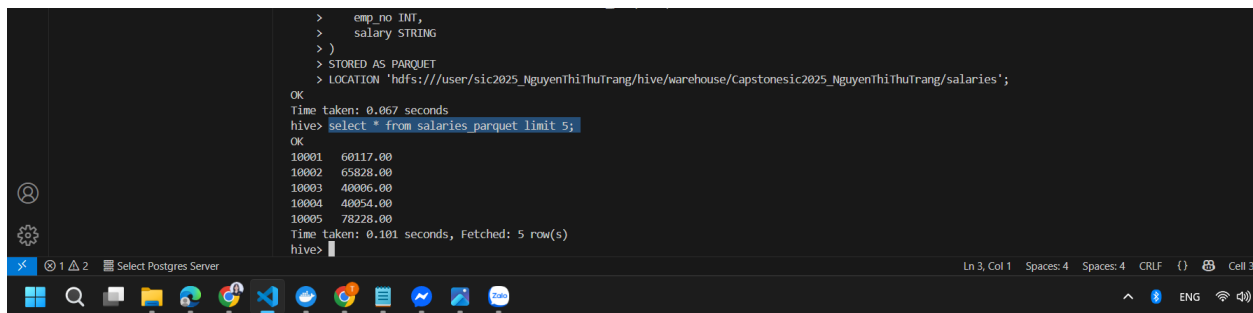
Kết quả: 2 bảng đã map dữ liệu thành công



Hình 14: Ảnh bảng employee đã được map



Hình 15: Ảnh bảng salaries đã được map

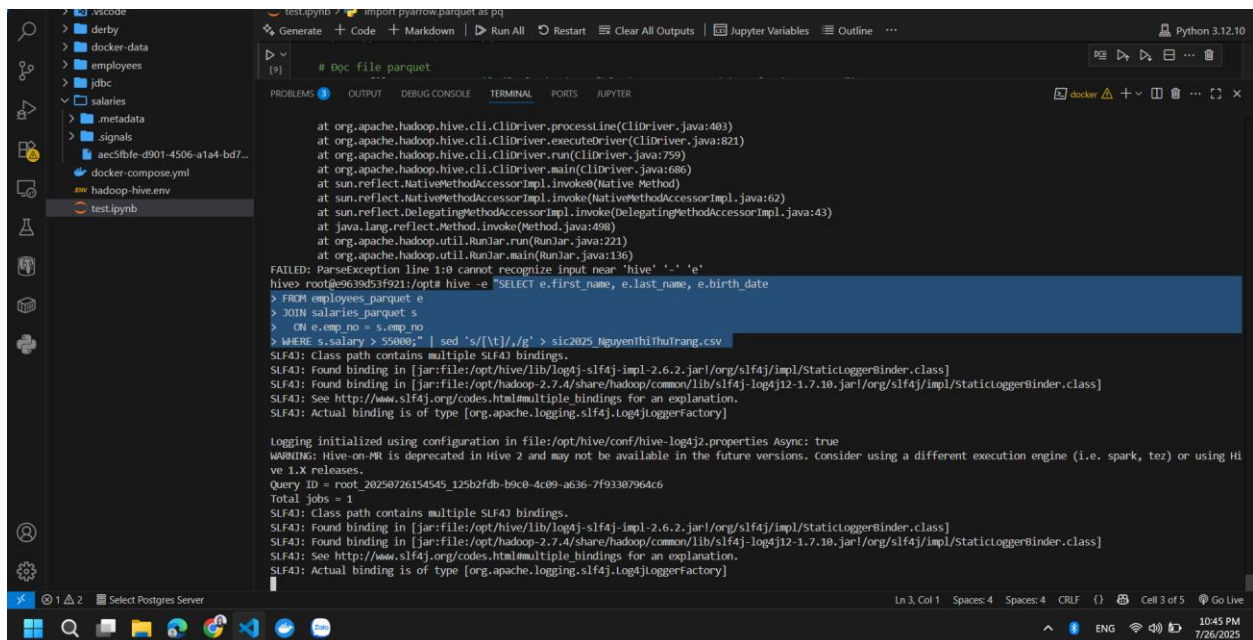


Hình 16: Ảnh kiểm tra dữ liệu trong bảng salaries

Câu 4 & 5: Phân tích dữ liệu và lưu kết quả

Yêu cầu: Liệt kê các thông tin first_name, last_name, birth_date của nhân viên có mức lương > 55,000 và lưu kết quả vào file sic2025_HoVaTenSinhVien.csv

- Thực hiện truy vấn SQL trên Hive
- Lưu kết quả vào file csv



Hình 17: Ảnh lệnh truy vấn và ghi file

Kết quả: Tổng cộng **110,684 bản ghi** thỏa mãn điều kiện đã được xuất ra file CSV thành công

The screenshot shows a Jupyter Notebook environment with a file explorer on the left and a terminal at the bottom. The notebook contains a single cell with the following code:

```
# Đọc file parquet
import pyarrow.parquet as pq
```

The terminal output shows the execution of the code, including the following log messages:

```
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.log4j.Log4jLoggerFactory]
Execution log at: /tmp/root/root_20250726154545_125b2fdb-b9c0-4c09-a636-7f9330796dc6.log
2025-07-26 15:45:53 Starting to launch local task to process map join; maximum memory = 477626368
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
2025-07-26 15:45:56 Dump the side-table for tag: 1 with group count: 110684 into file: file:/tmp/root/e09ee813-54b7-496e-9271-ce5f2791ec79/hive_2025-07-26_15-45-56_7094436408686159799-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2025-07-26 15:45:56 Uploaded 1 File to: file:/tmp/root/e09ee813-54b7-496e-9271-ce5f2791ec79/hive_2025-07-26_15-45-45_674_7094436408686159799-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (2315454 bytes)
2025-07-26 15:45:56 End of local task; Time Taken: 2.395 sec.
Execution completed successfully
MapReduce local task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
2025-07-26 15:45:58,572 Stage-3 map = 0%, reduce = 0%
2025-07-26 15:45:59,582 Stage-3 map = 100%, reduce = 0%
Ended Job = job_local154552406_0001
MapReduce Jobs Launched:
Stage-Stage-3: HDFS Read: 2572528 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 13.949 seconds, Fetched: 110684 row(s)
root@e9639d53f921:/opt#
```

Hình 18: Ảnh kết quả đã truy vấn và lưu thành công

Lời cảm ơn

Trên đây là toàn bộ quá trình thực hiện bài kiểm tra thực hành môn Big Data. Em xin chân thành cảm ơn thầy đã tạo điều kiện để em được thực hành và củng cố kiến thức. Nếu có bất kỳ sai sót hoặc thiếu sót nào trong bài, em rất mong nhận được sự góp ý và hướng dẫn thêm từ thầy để hoàn thiện hơn trong tương lai.