**Thao tác với Apache Sqoop**

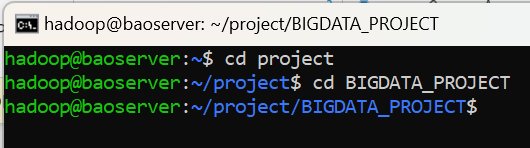
**Chú ý:** bạn pahri chuẩn bị môi trường hadoop, mysql, Apache Sqoop

**1. Chuẩn bị dữ liệu trên hdfs**

Làm việc trong thư mục ~/project/BIGDATA\_PROJECT

$ cd project

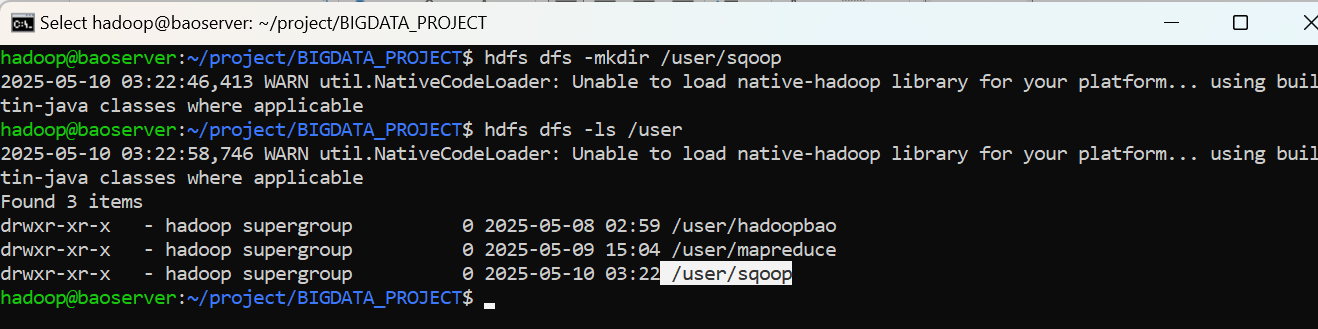
$ cd BIGDATA\_PROJECT



Bước 1: Tạo thư mục /user/sqoop/ trên hdfs

$ hdfs dfs -mkdir /user/sqoop

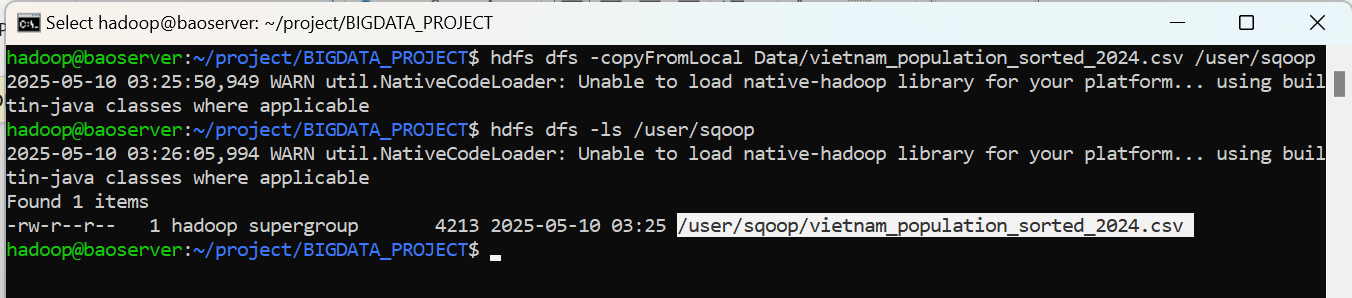
$ hdfs dfs -ls /user



Bước 2: đẩy dữ liệu lên hdfs

$ hdfs dfs -copyFromLocal Data/vietnam\_population\_sorted\_2024.csv /user/sqoop

$ hdfs dfs -ls /user/sqoop



**2. Tạo database và đẩy dữ liệu từ hdfs vào database**

**2.1 Tạo database và table**

Bước 1: Tạo cơ sở dữ liệu

$ sqoop eval \

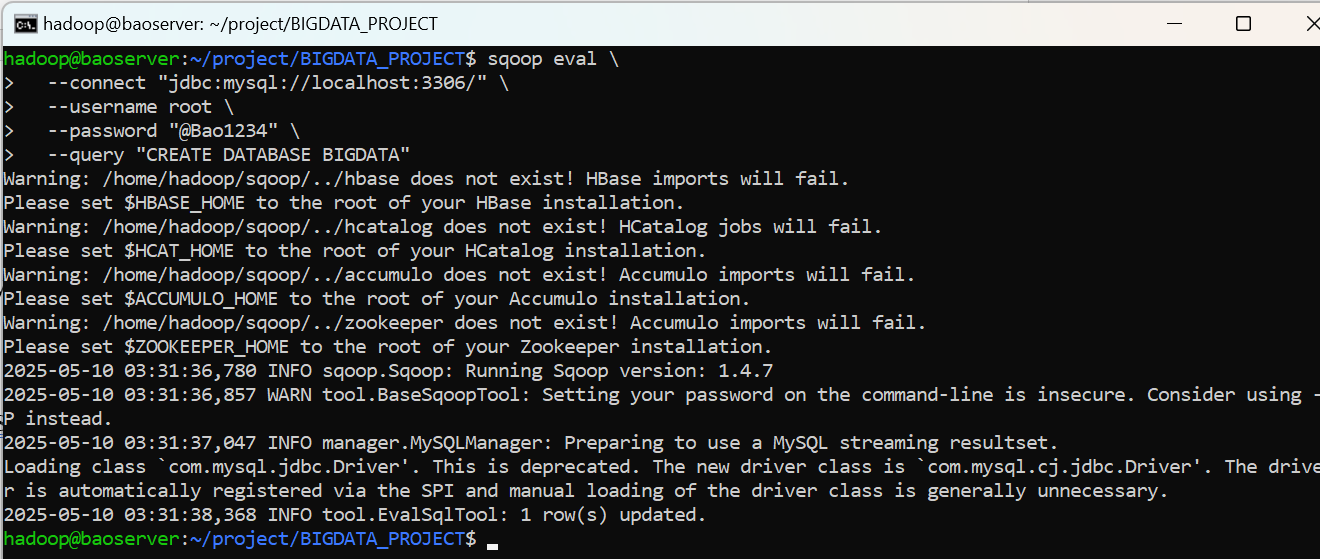
--connect "jdbc:mysql://localhost:3306/" \

--username root \

--password "@Bao1234" \

--query "CREATE DATABASE BIGDATA"

*Chú ý: cần sửa lại mật khẩu user là mật khẩu user của bạn*



Bước 2: Tạo quan hệ

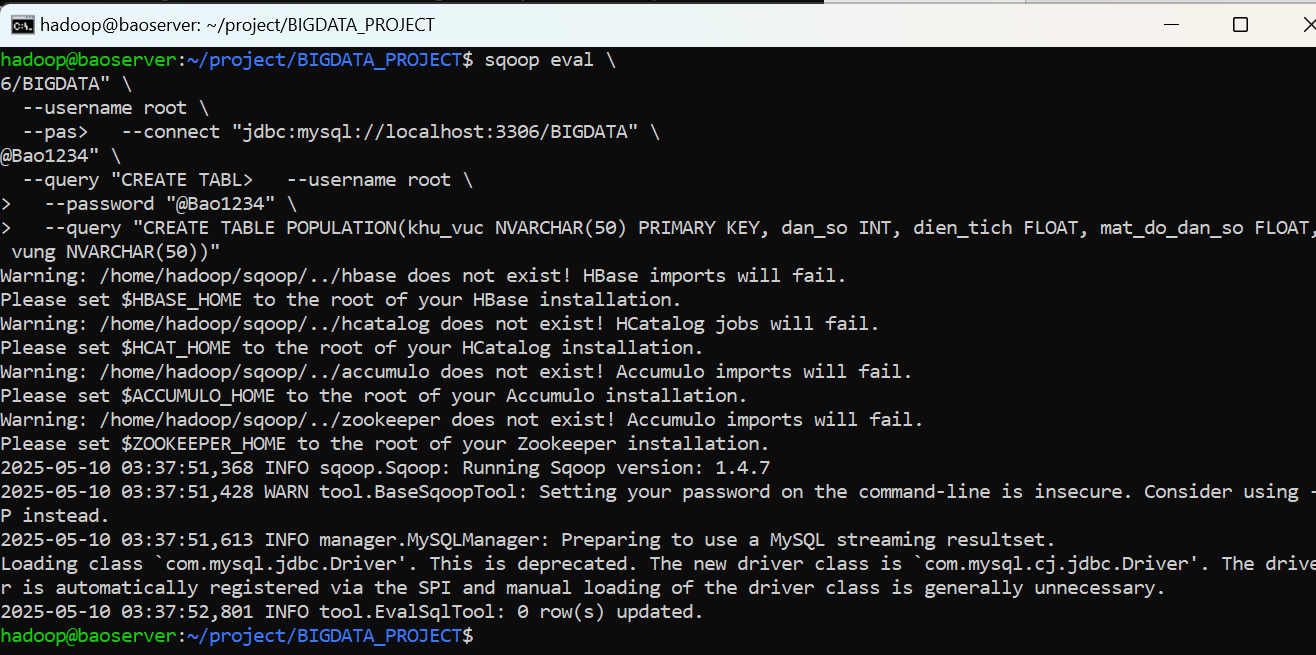
$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/BIGDATA" \

--username root \

--password "@Bao1234" \

--query "CREATE TABLE POPULATION(khu\_vuc NVARCHAR(50) PRIMARY KEY, dan\_so INT, dien\_tich FLOAT, mat\_do\_dan\_so FLOAT, vung NVARCHAR(50))"



Bước 3: Kiểm tra lại database và table đã được tạo đúng hay chưa

Kiểm tra database đã được tạo hay chưa:

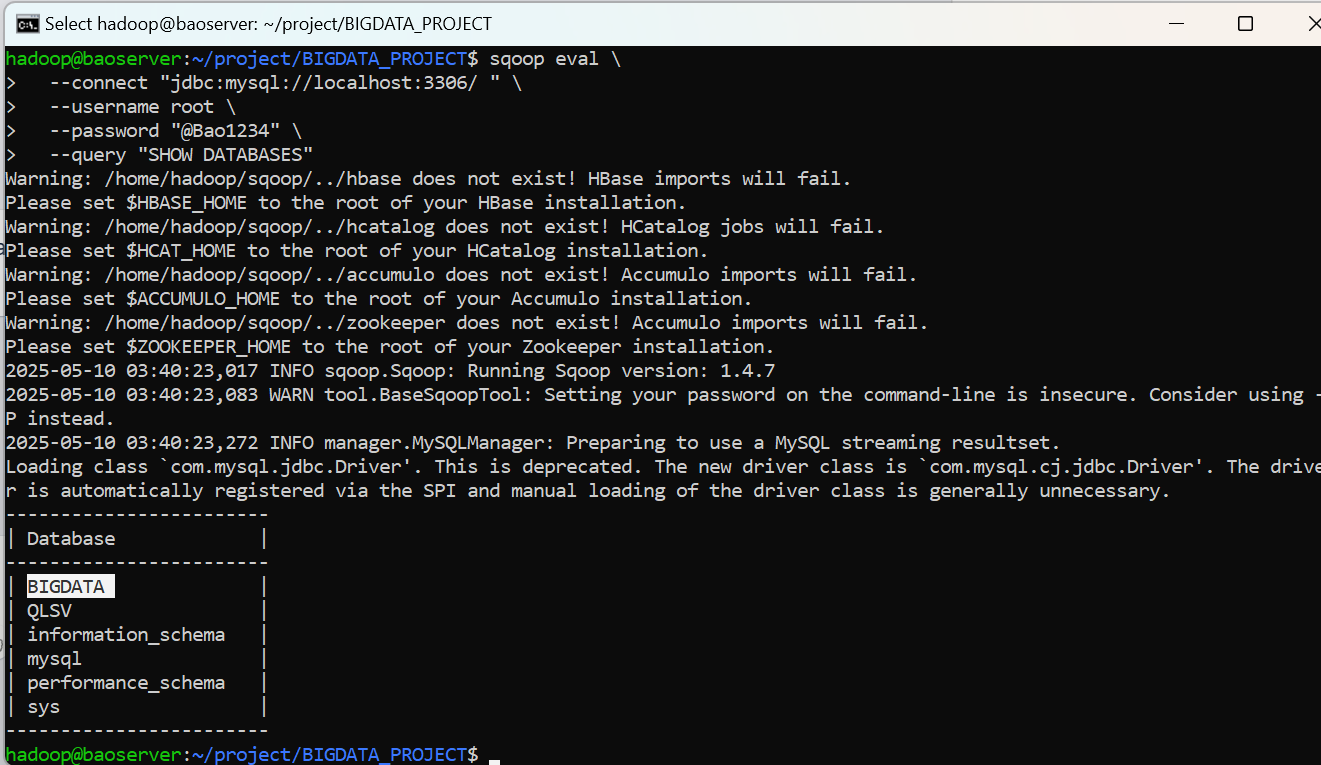
$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/ " \

--username root \

--password "@Bao1234" \

--query "SHOW DATABASES"



Kiểm tra xem table đã được tạo hay chưa

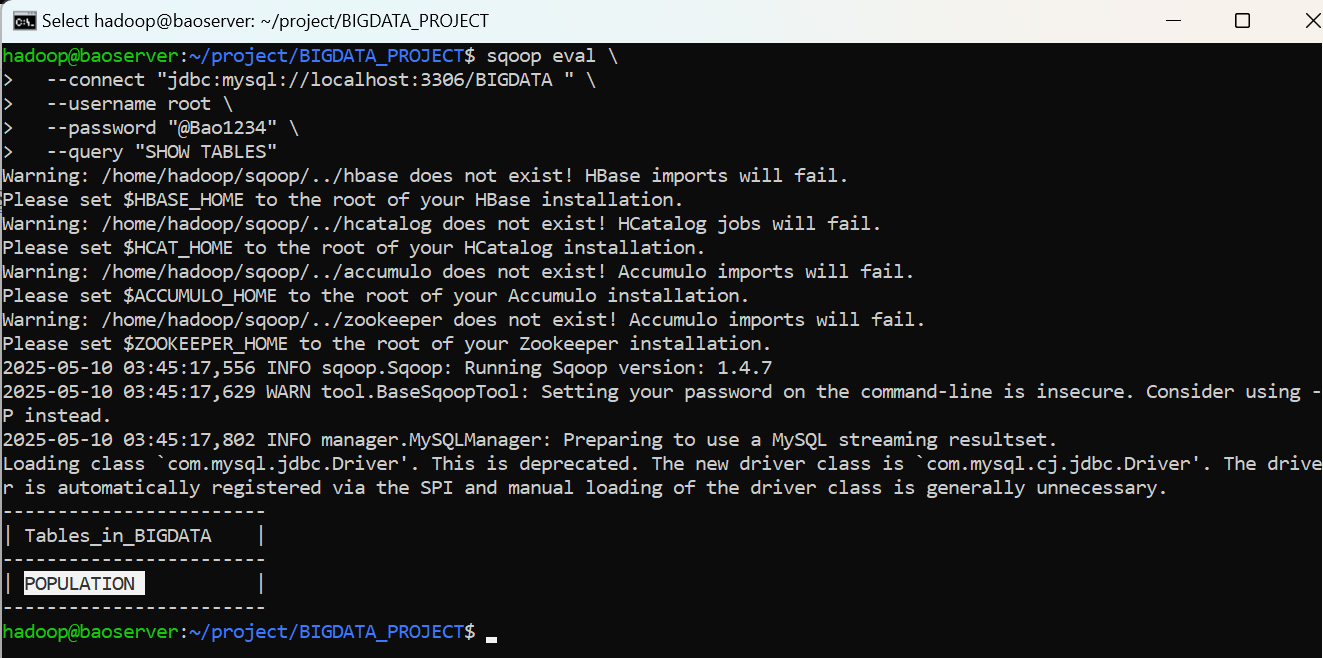
$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/BIGDATA " \

--username root \

--password "@Bao1234" \

--query "SHOW TABLES"

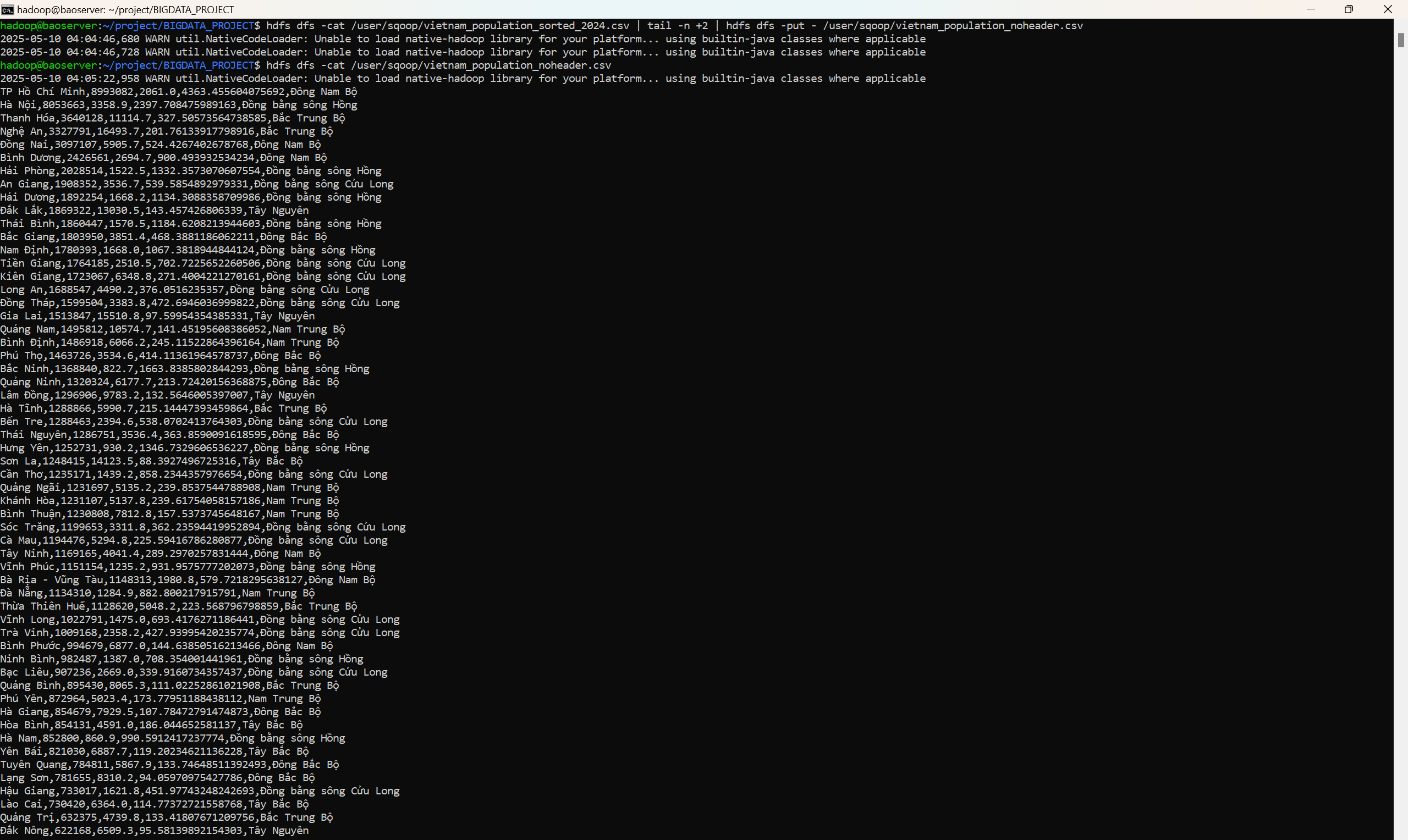


**2.2 Đẩy dữ liệu từ hdfs vào cơ sở dữ liệu**

Bước 1: Bỏ dùng dầu tiên của /user/sqoop/vietnam\_population\_sorted\_2024.csv

$ hdfs dfs -cat /user/sqoop/vietnam\_population\_sorted\_2024.csv | tail -n +2 | hdfs dfs -put - /user/sqoop/vietnam\_population\_noheader.csv

$ hdfs dfs -cat /user/sqoop/vietnam\_population\_noheader.csv



Bước 2: Thực hiện thêm dữ liệu từ file /user/sqoop/vietnam\_population\_noheader.csv vào table population

$ sqoop export \

--connect "jdbc:mysql://localhost:3306/BIGDATA " \

--username root \

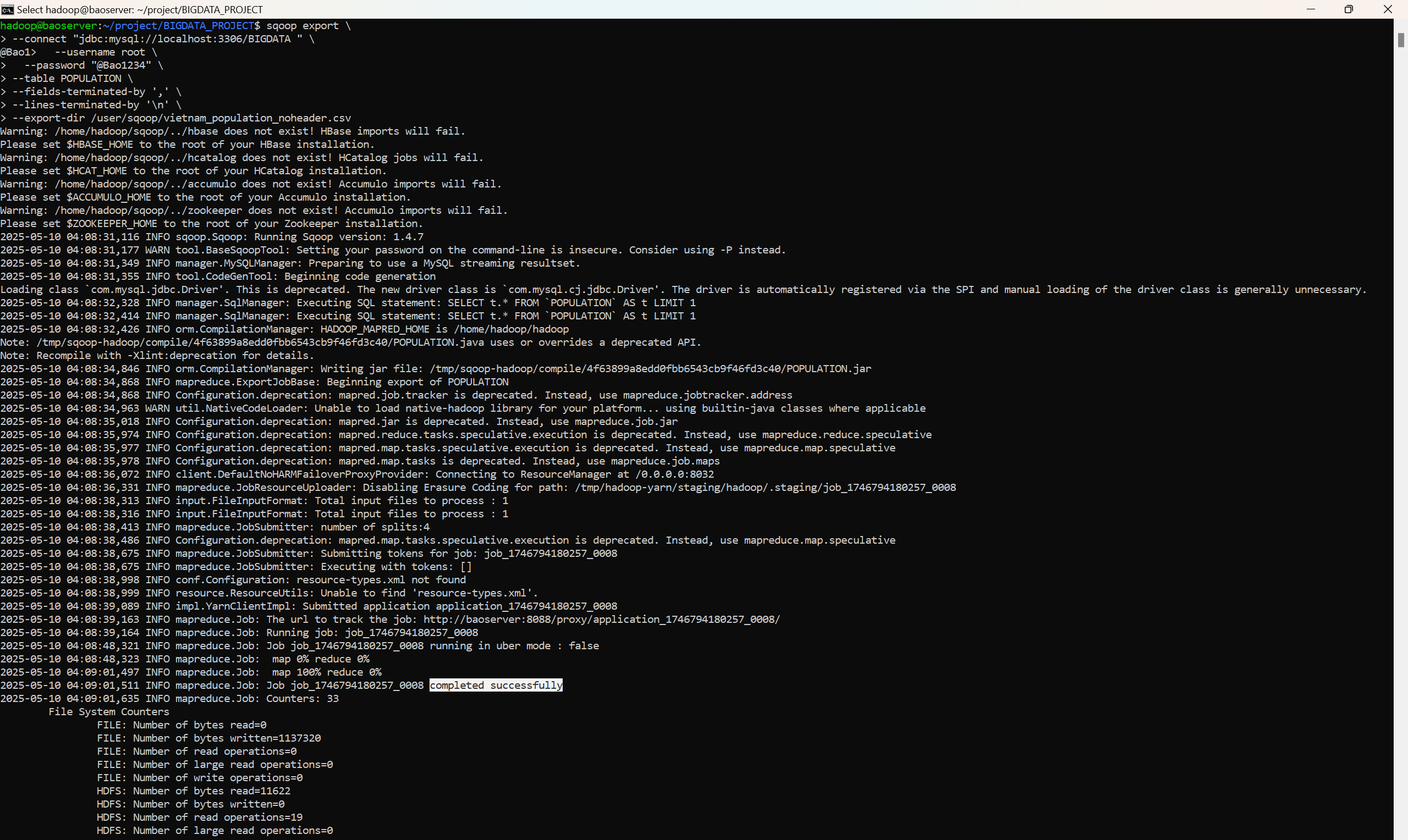
--password "@Bao1234" \

--table POPULATION \

--fields-terminated-by ',' \

--lines-terminated-by '\n' \

--export-dir /user/sqoop/vietnam\_population\_noheader.csv



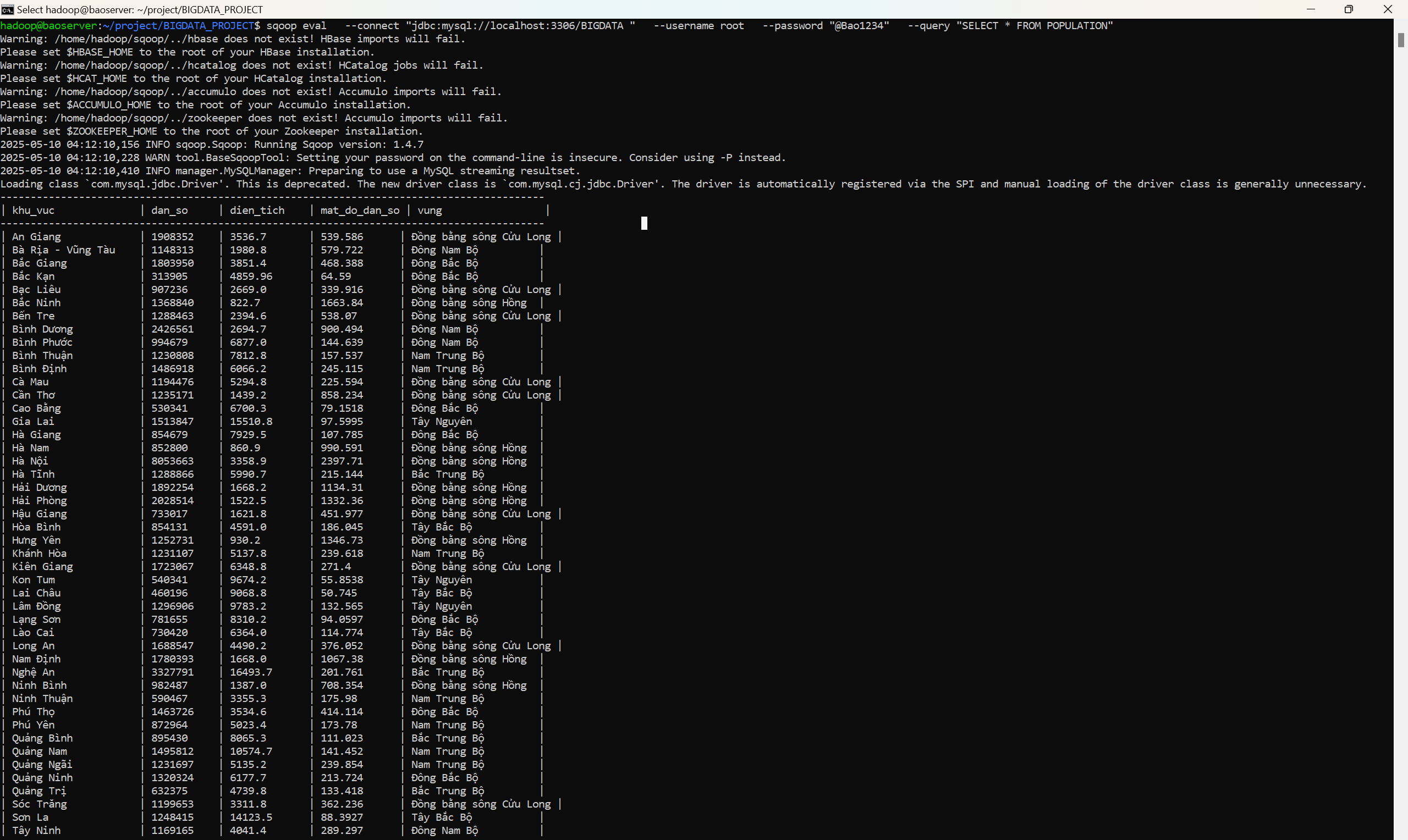
Bước 3: Kiểm tra dữ liệu đã được thêm vào đúng hay chưa

$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/BIGDATA" \

--username root \

--password "@Bao1234" \

--query "SELECT \* FROM POPULATION"

**3. Một số thao tác với sqoop**

**3.1 Chọn ra tỉnh có mật độ dân số cao nhất**

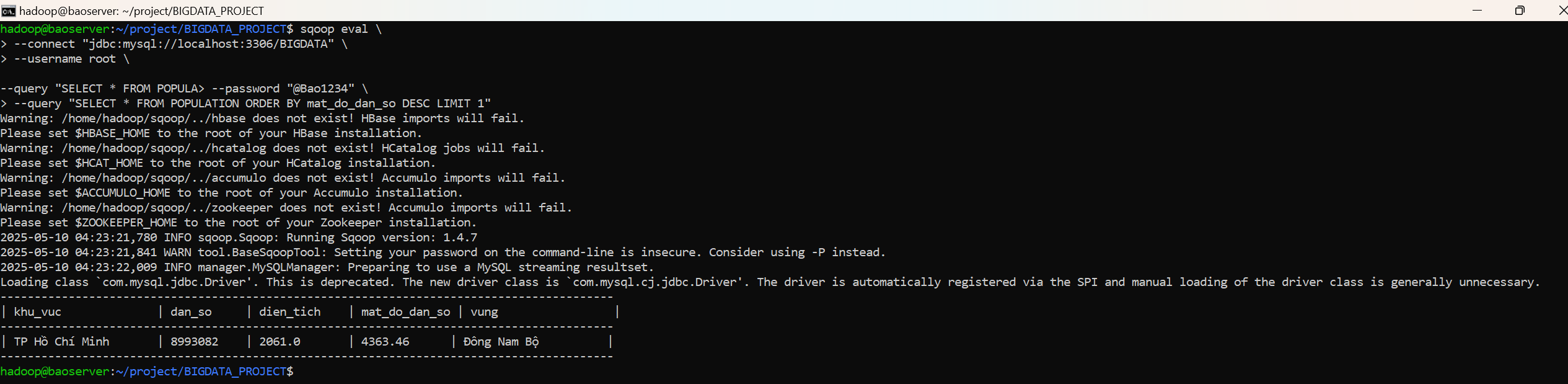
$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/BIGDATA" \

--username root \

--password "@Bao1234" \

--query "SELECT \* FROM POPULATION ORDER BY mat\_do\_dan\_so DESC LIMIT 1"



**3.2 Tìm tỉnh có tỷ trọng đân số nhỏ nhất (tỷ trọng dân số = đân số khu vực/tổng dân số)**

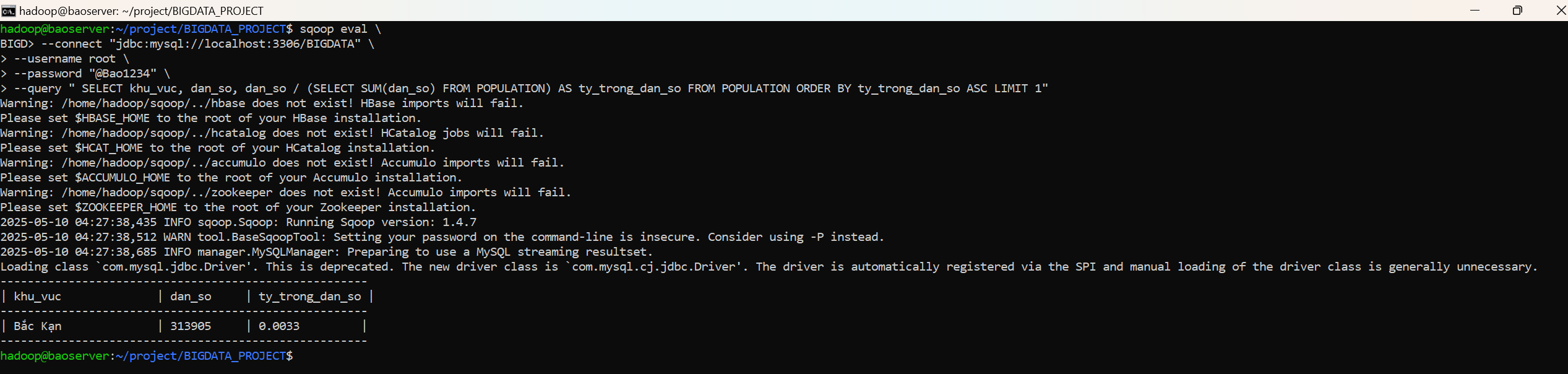
$ sqoop eval \

--connect "jdbc:mysql://localhost:3306/BIGDATA" \

--username root \

--password "@Bao1234" \

--query " SELECT khu\_vuc, dan\_so, dan\_so / (SELECT SUM(dan\_so) FROM POPULATION) AS ty\_trong\_dan\_so FROM POPULATION ORDER BY ty\_trong\_dan\_so ASC LIMIT 1"



**3.3 Tính tỷ trọng dân số của các vùng việt nam lấy các thông tin (vung, dan\_so, dien\_tich, ty\_trong) và đưa dữ liệu đó lên hdfs**

Bước 1: thực hiện câu lệnh sql và đẩy dữ liệu lên hdfs

$ sqoop import \

-Dorg.apache.sqoop.splitter.allow\_text\_splitter=true \

--connect jdbc:mysql://localhost:3306/BIGDATA \

--username root \

--password "@Bao1234" \

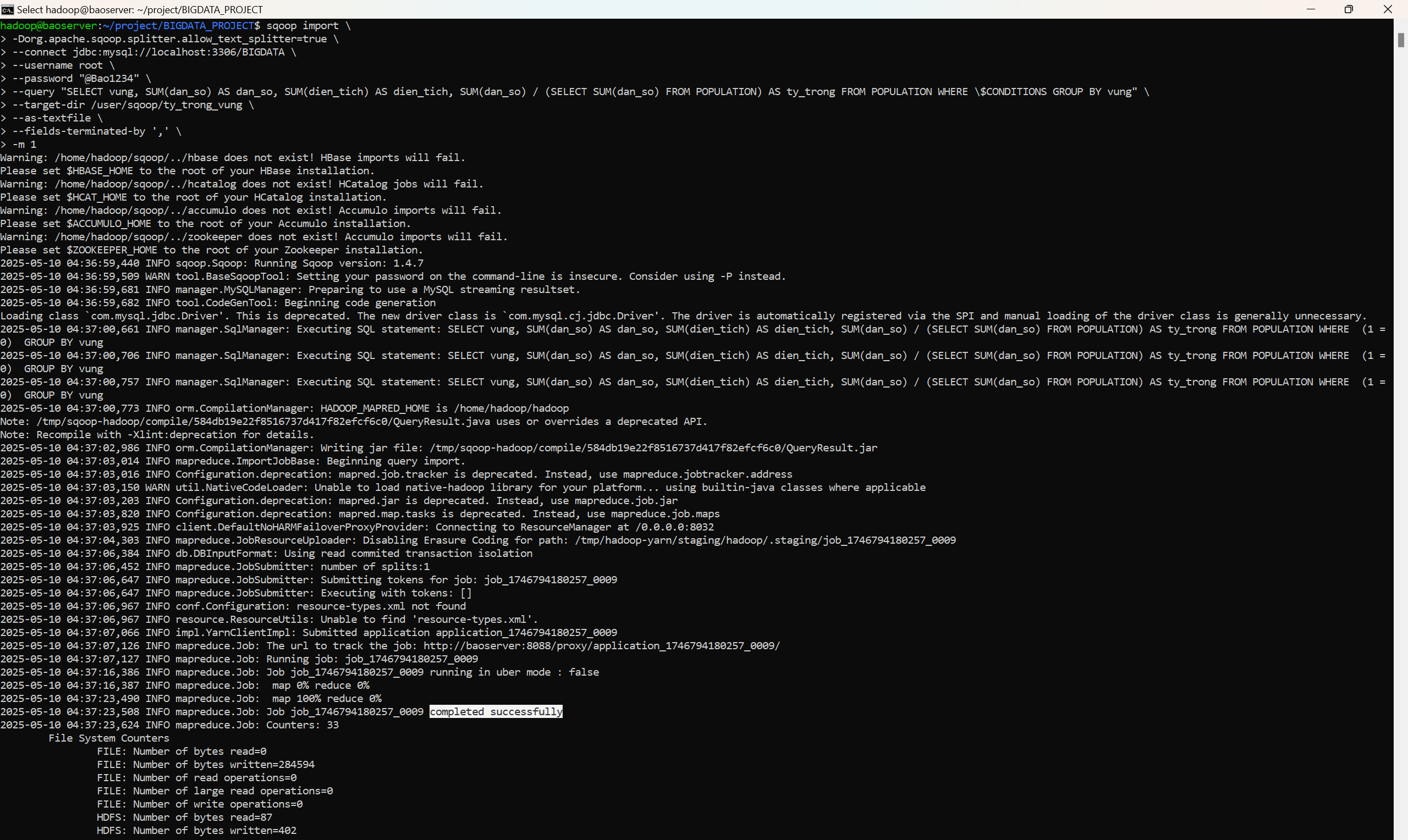
--query "SELECT vung, SUM(dan\_so) AS dan\_so, SUM(dien\_tich) AS dien\_tich, SUM(dan\_so) / (SELECT SUM(dan\_so) FROM POPULATION) AS ty\_trong FROM POPULATION WHERE \$CONDITIONS GROUP BY vung" \

--target-dir /user/sqoop/ty\_trong\_vung \

--as-textfile \

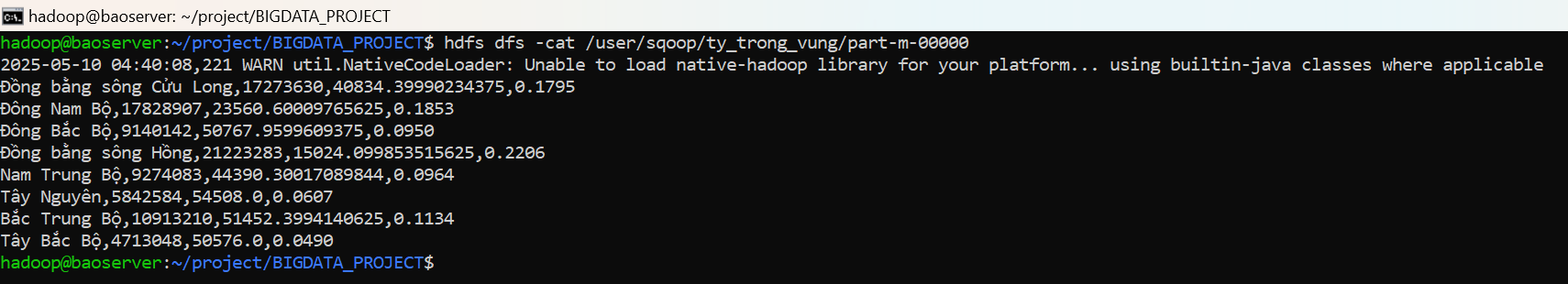
--fields-terminated-by ',' \

-m 1



Bước 2: Kiểm tra dữ liệu trên hdfs

$ hdfs dfs -cat /user/sqoop/ty\_trong\_vung/part-m-00000



**3.4 Lấy thông tin khu vực có đân số đông nhất nhưng lại có diện tích nhỏ nhất và đẩy dữ liệu đó lên hdfs**

Bước 1: Thực hiện câu lệnh SQL và đẩy dữ liệu lên hdfs

$ sqoop import \

-Dorg.apache.sqoop.splitter.allow\_text\_splitter=true \

--connect jdbc:mysql://localhost:3306/BIGDATA \

--username root \

--password "@Bao1234" \

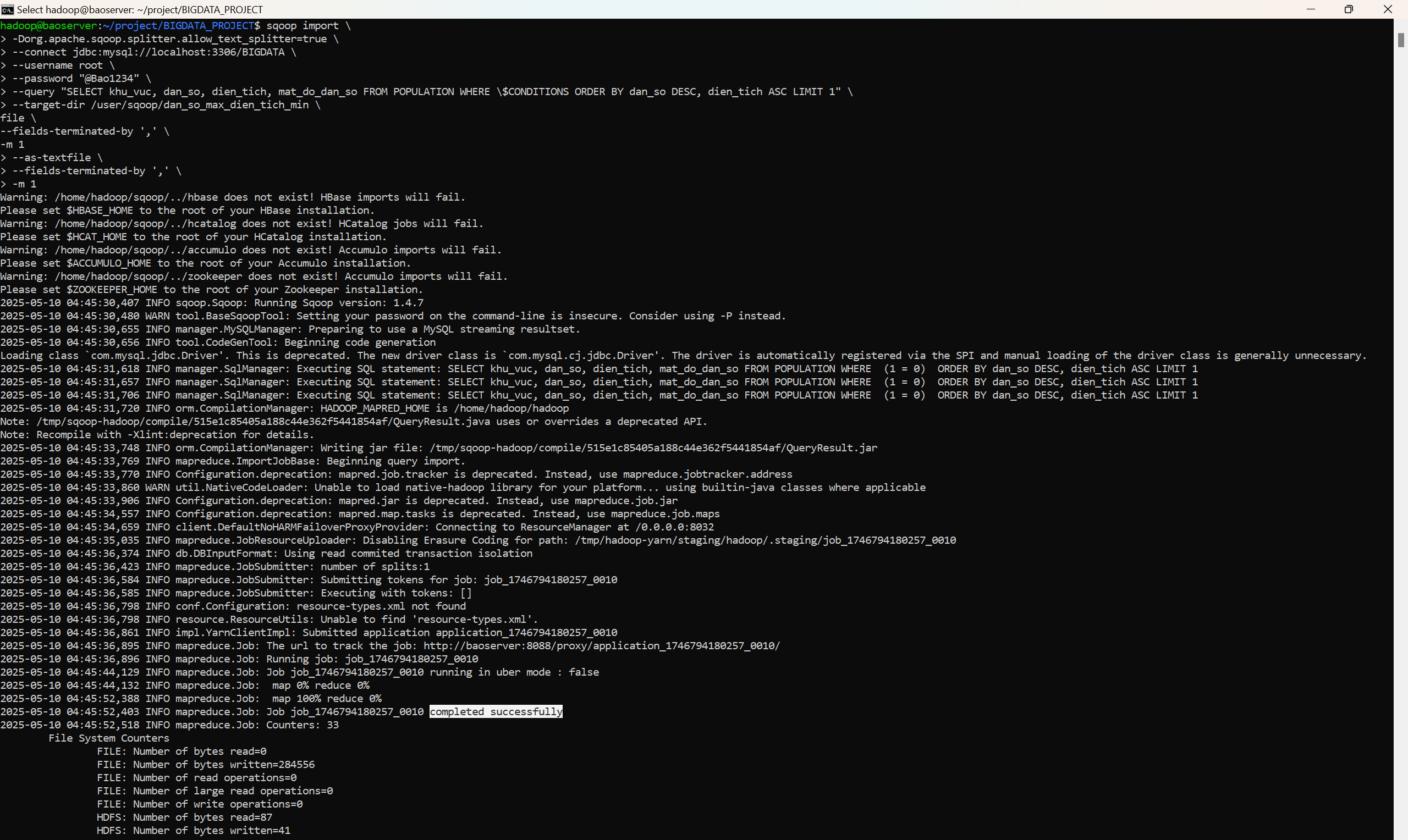
--query "SELECT khu\_vuc, dan\_so, dien\_tich, mat\_do\_dan\_so FROM POPULATION WHERE \$CONDITIONS ORDER BY dan\_so DESC, dien\_tich ASC LIMIT 1" \

--target-dir /user/sqoop/dan\_so\_max\_dien\_tich\_min \

--as-textfile \

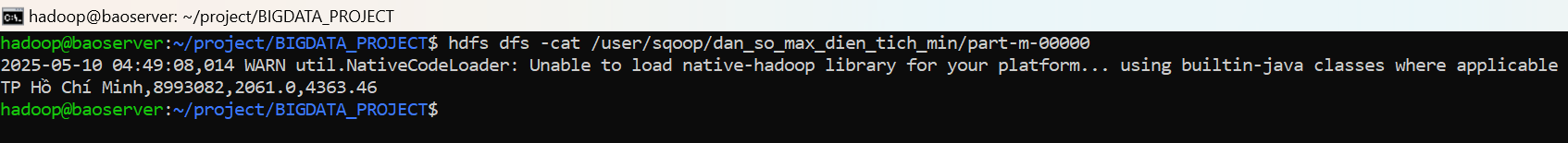
--fields-terminated-by ',' \

-m 1



Bước 2: Kiểm tra dữ liệu trên hdfs

$ hdfs dfs -cat /user/sqoop/dan\_so\_max\_dien\_tich\_min/part-m-00000



**3.5 Xuất dữ liệu từ bảng vào hdfs**

Bước 1: Đẩy dữ liệu từ bảng vào hdfs

$ sqoop import \

--connect jdbc:mysql://localhost:3306/BIGDATA \

--username root \

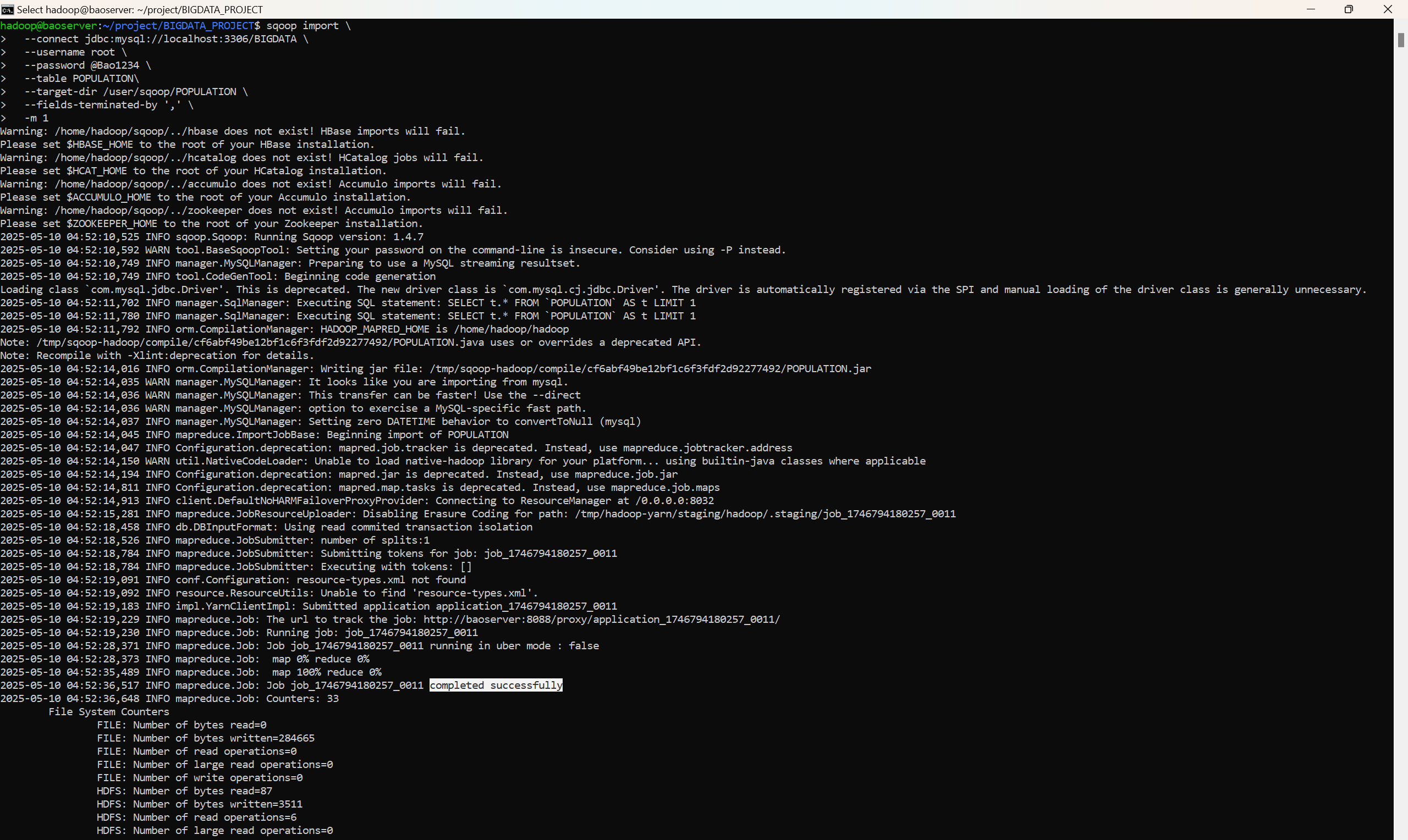
--password @Bao1234 \

--table POPULATION\

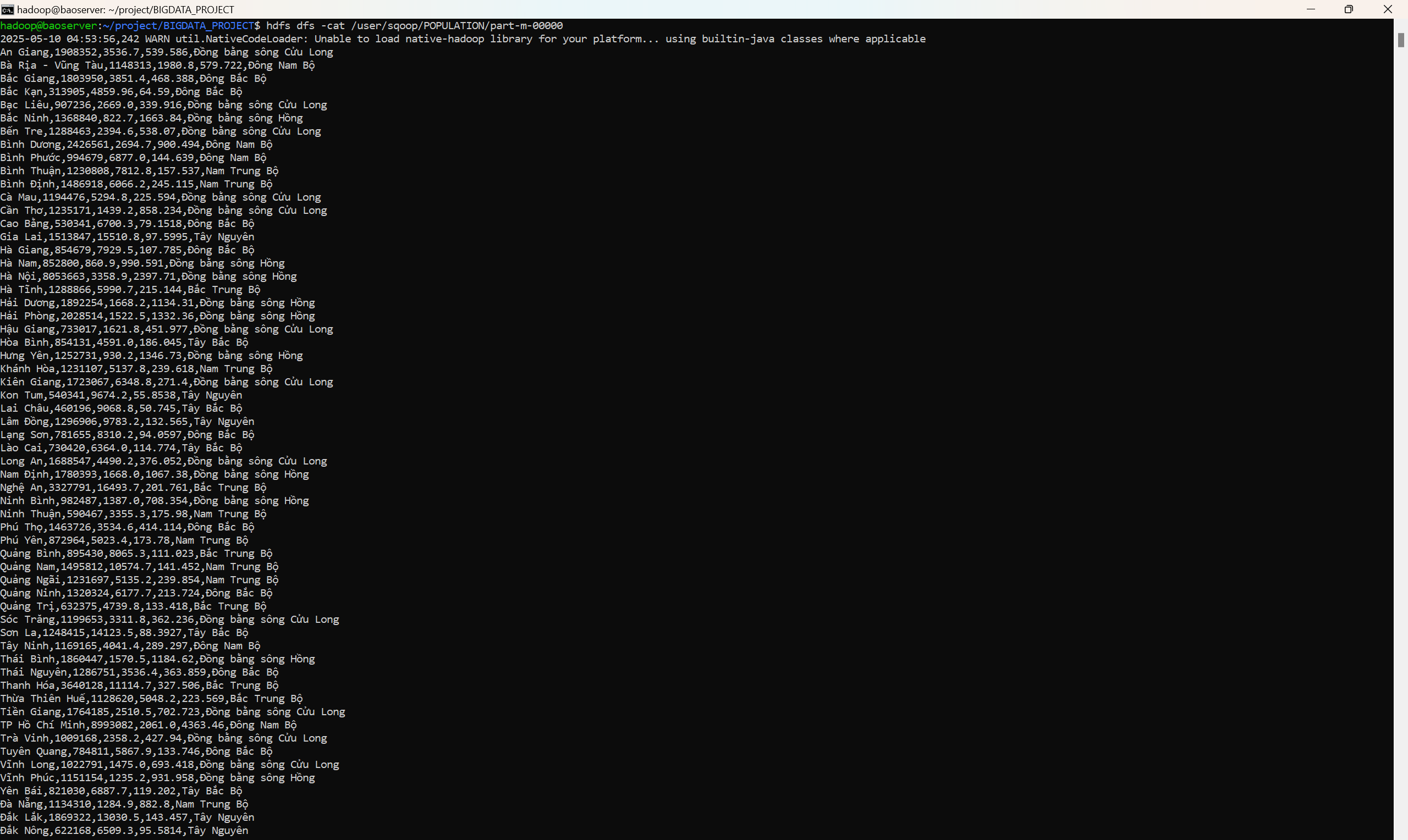
--target-dir /user/sqoop/POPULATION \

--fields-terminated-by ',' \

-m 1



Bước 2: Kiểm tra lại dữ liệu trên hdfs



**3.6 Tạo job cho phép nhập dữ liệu từ các bảng vào hdfs**

Bước 1: Tạo job

$ sqoop job --create job\_population \

-- import \

--connect jdbc:mysql://localhost:3306/BIGDATA \

--username root \

--password @Bao1234 \

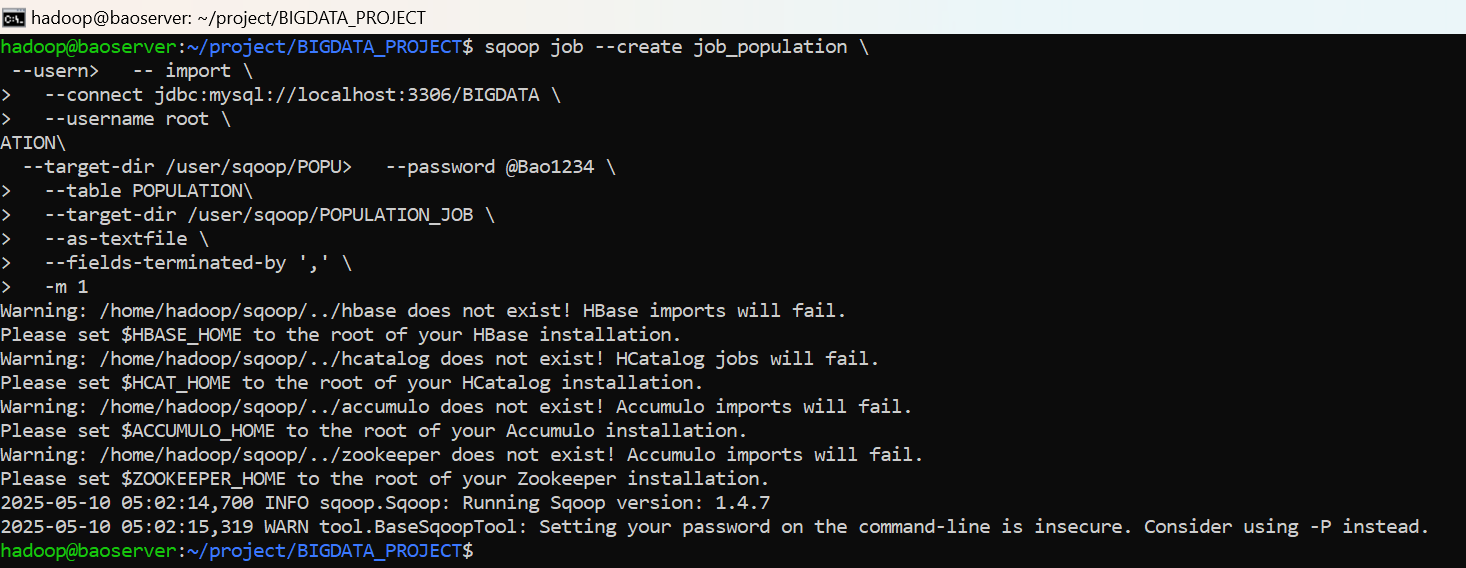
--table POPULATION\

--target-dir /user/sqoop/POPULATION\_JOB \

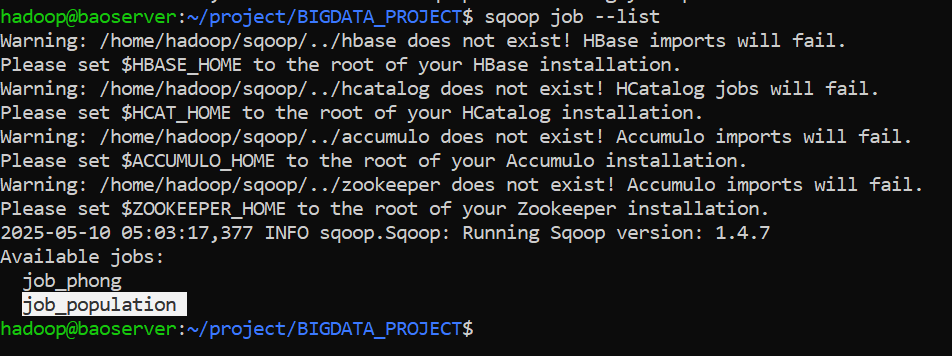
--as-textfile \

--fields-terminated-by ',' \

-m 1

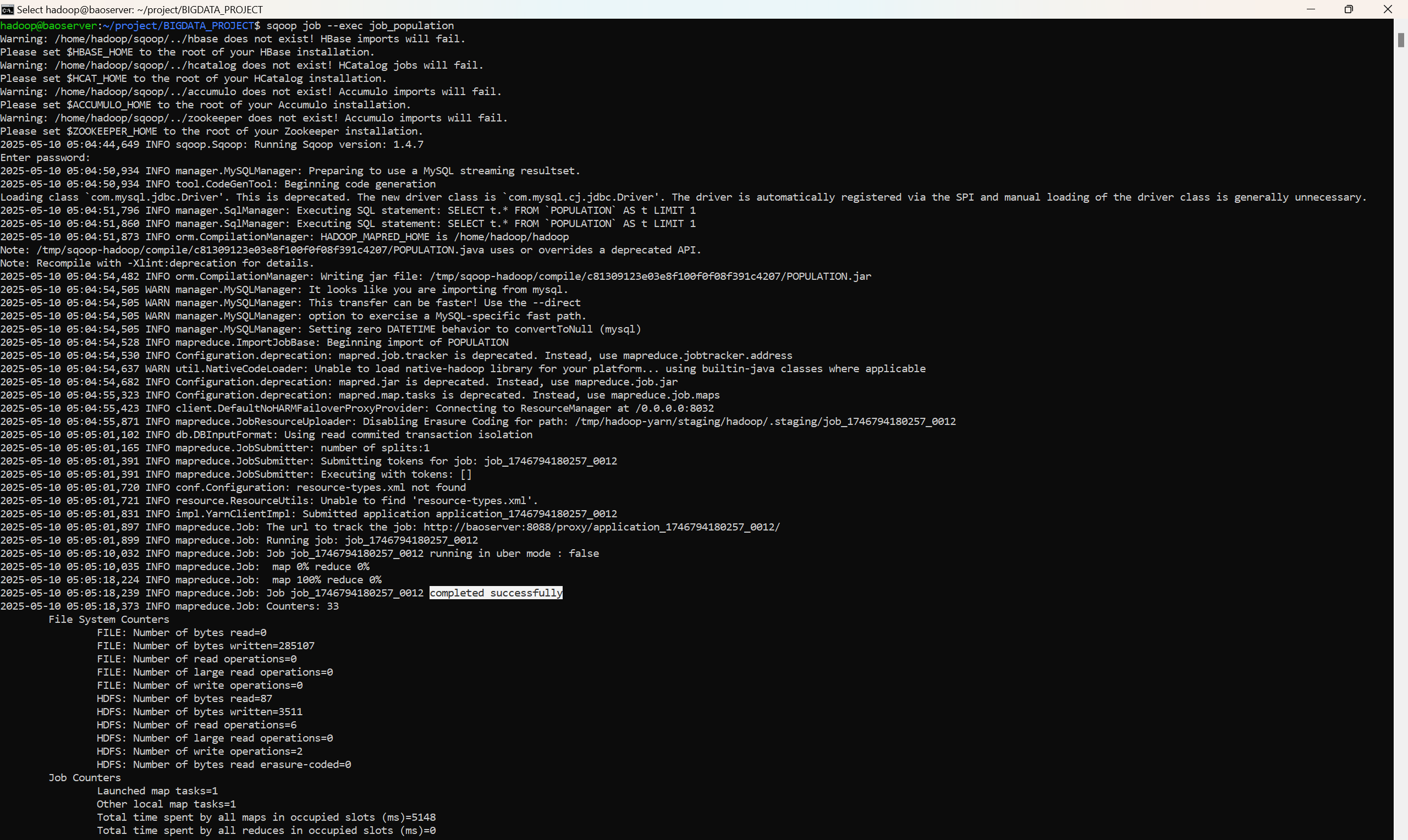


Bước 2: Kiểm tra job có được tạo đúng hya chưa

$ sqoop job --list

Bước 3: Thực thi job

$ sqoop job --exec job\_population



Bước 4: Kiểm tra dữ liệu trên hdfs

$ hdfs dfs -cat /user/sqoop/POPULATION\_JOB/part-m-00000

