**BIG DATA SAMSUNG PROJECT**

Docker running

A screenshot of a computer program

AI-generated content may be incorrect.

Copying data to docker container

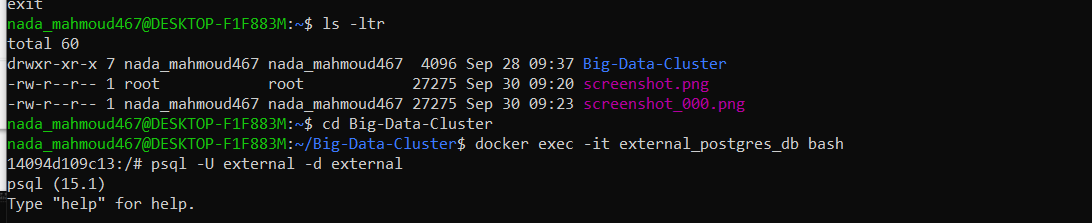
A screen shot of a computer

AI-generated content may be incorrect.

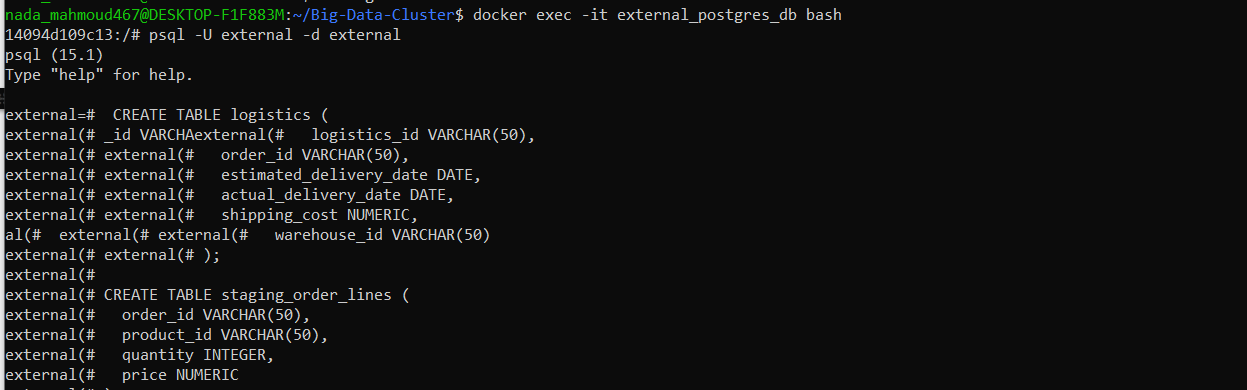
A screen shot of a computer code

AI-generated content may be incorrect.

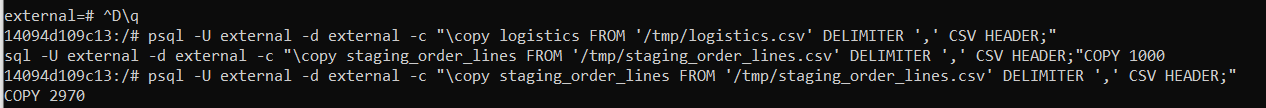
Access Postgres in the container



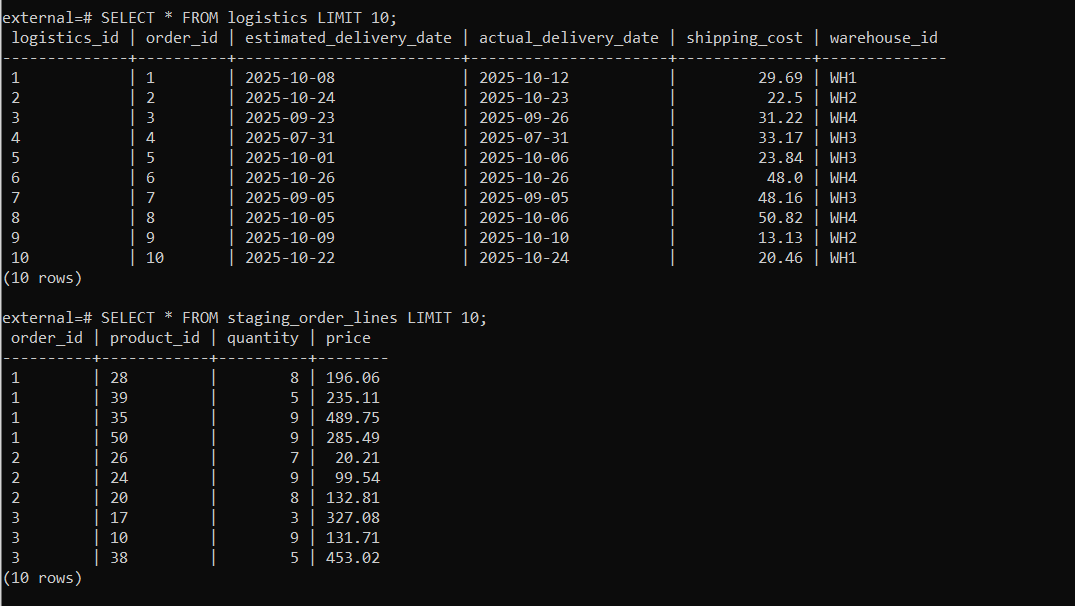
Create tables matching CSV schemas



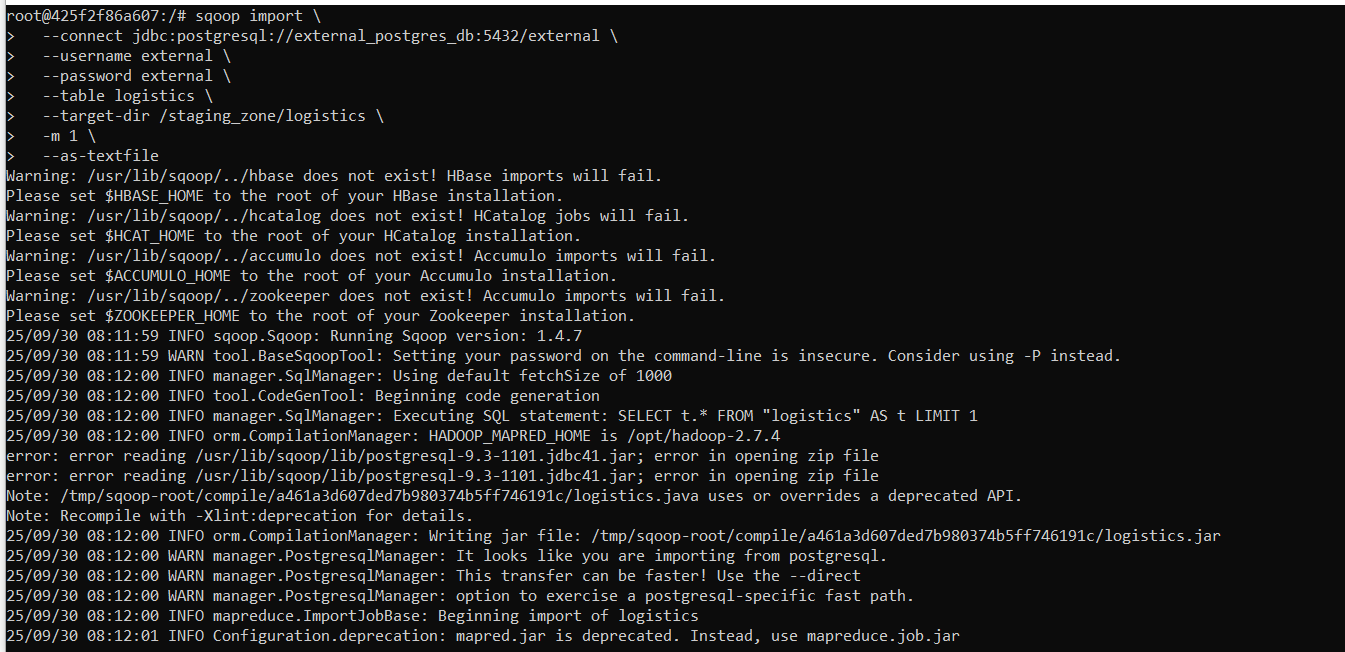
Load CSV files into Postgres tables



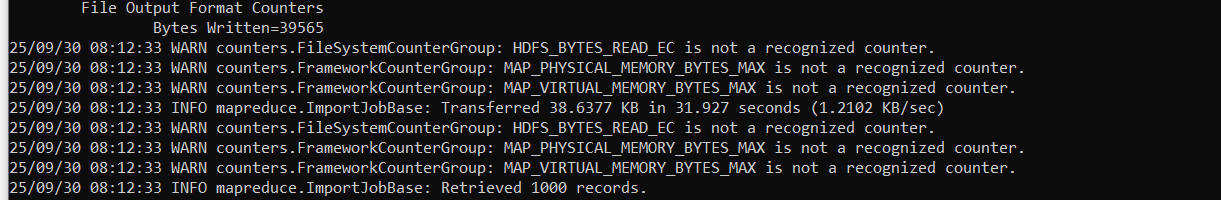
Show the data



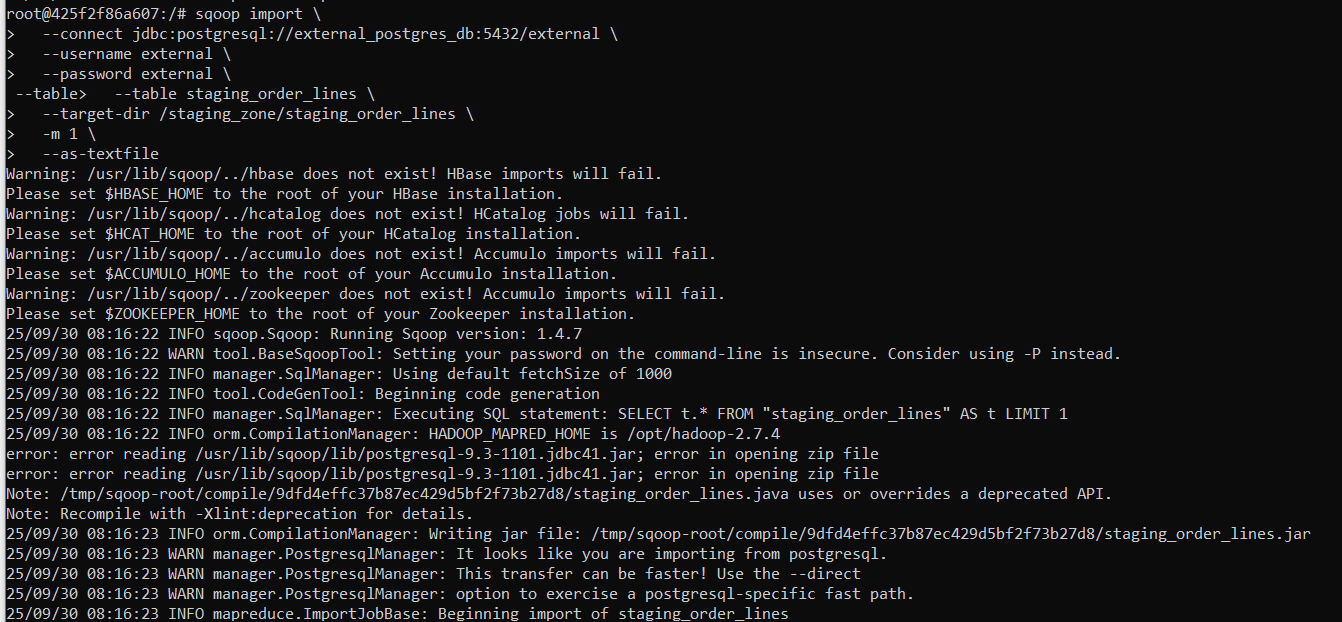
Extract data using Sqoop



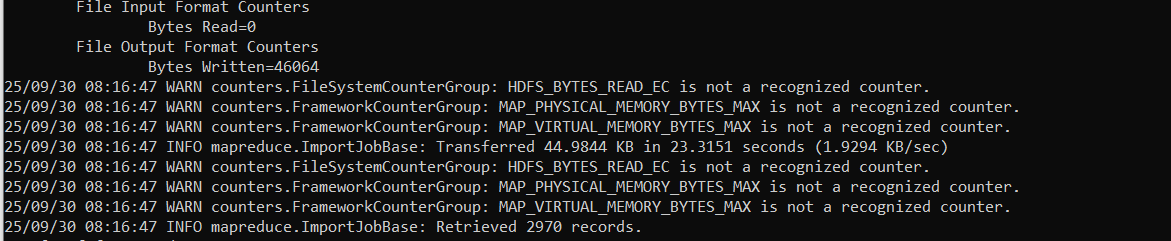
Extract 1000 record from logistics file



Extract data using Sqoop



Extract 2970 record from staging \_order\_lines file



Transformation on zeppelin using Spark

1-Load data and assign column names

A screenshot of a computer code

AI-generated content may be incorrect.

2-Inspect schema and sample data for logistics

A screenshot of a computer

AI-generated content may be incorrect.

3-Inspect schema and sample data for order lines

A screenshot of a computer program

AI-generated content may be incorrect.

3-Clean and transform logistics data

A computer code on a white background

AI-generated content may be incorrect.

4-Create logistics fact table

A screenshot of a computer

AI-generated content may be incorrect.

5-Join logistics fact with order lines to get revenue

A screenshot of a computer

AI-generated content may be incorrect.

6-Check if an order is served by multiple warehouses

A screenshot of a computer

AI-generated content may be incorrect.

7-Aggregate order-level summary

A screenshot of a computer screen

AI-generated content may be incorrect.

8-Aggregate warehouse-level summary

A screenshot of a computer

AI-generated content may be incorrect.

9-Calculate ratio of shipping cost to total revenue

A screenshot of a computer

AI-generated content may be incorrect.

10-Average delivery delay in days

A screenshot of a computer

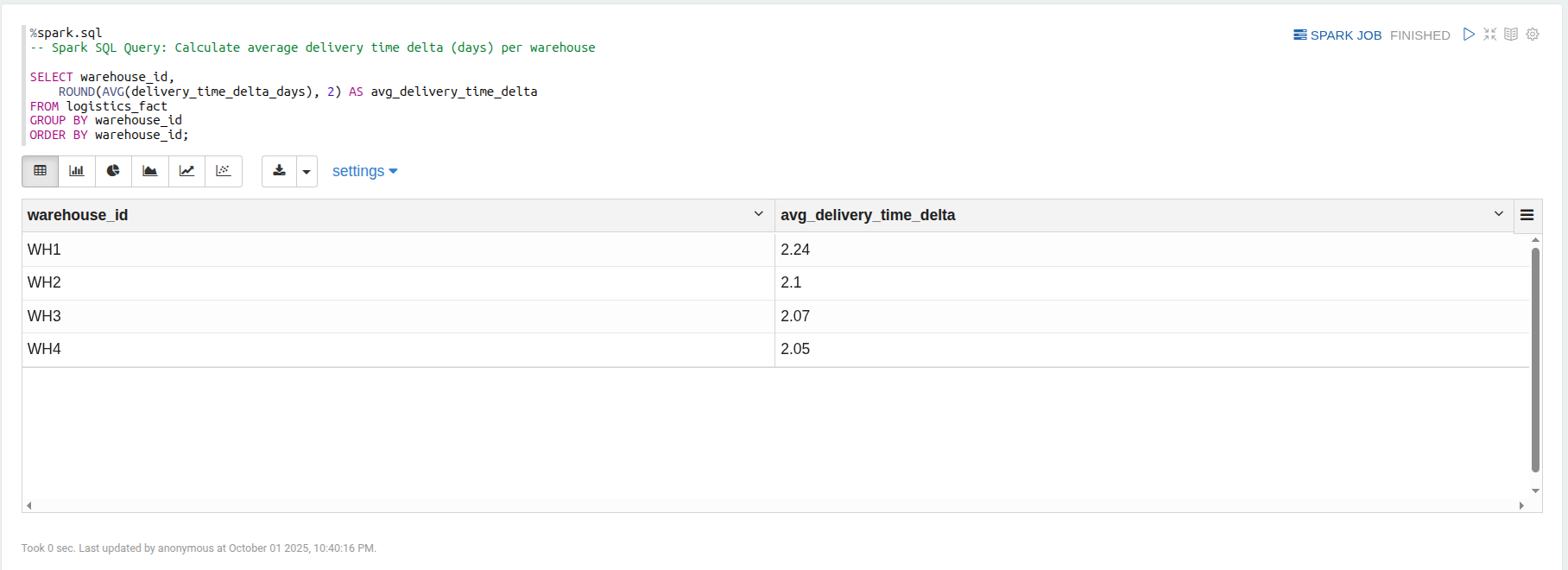
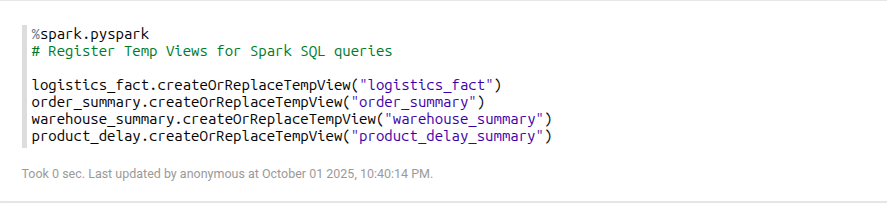
AI-generated content may be incorrect.

11-Product-level analysis of late deliveries

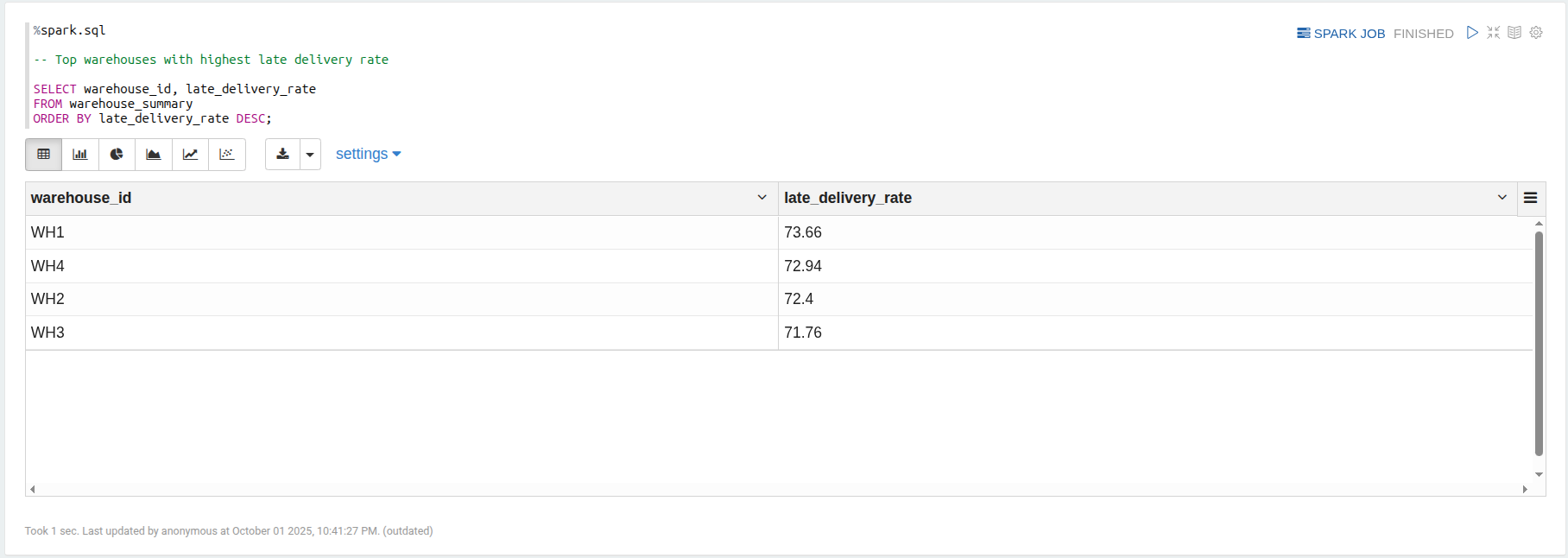
A screenshot of a computer

AI-generated content may be incorrect.

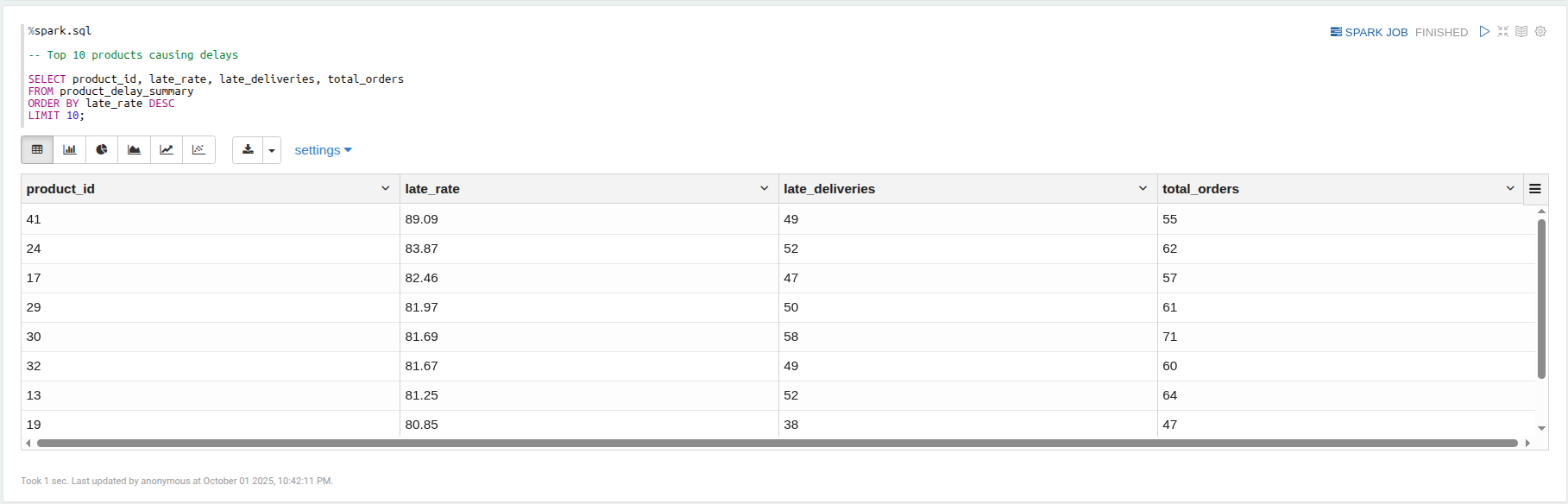
12-Average delivery time delta (days) per warehouse

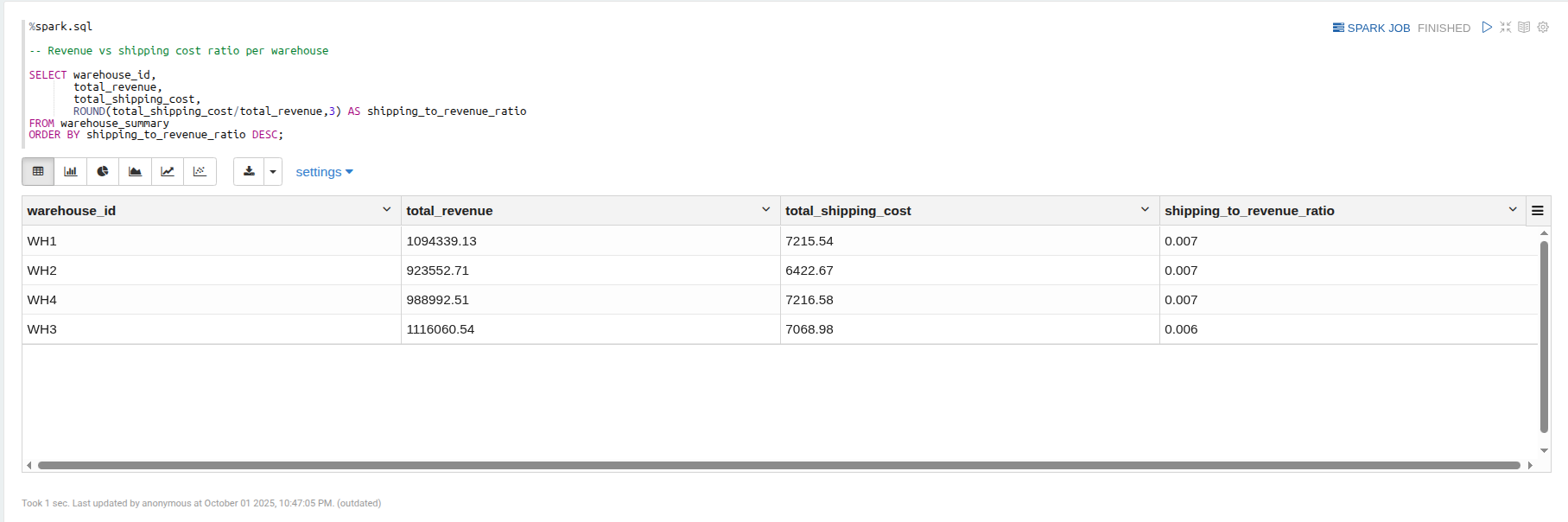


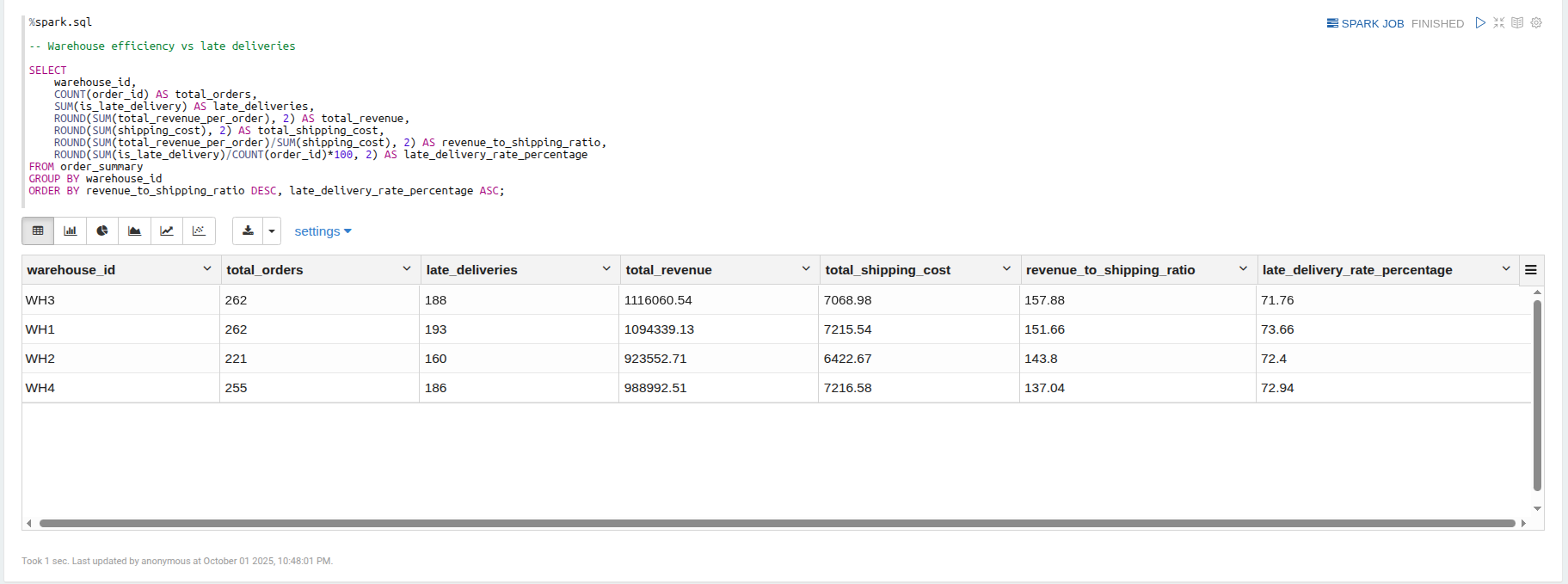
13-Top warehouses with highest late delivery rate



14-Top 10 products causing delay



15- Revenue vs Shipping cost ratio per warehouse

16- Warehouse financial efficiency vs late deliveries

Star Schema for tables:

Fact\_table (logistics\_fact):

* order\_id, delivery\_time\_delta\_days, is\_late\_delivery, shipping\_cost, warehouse\_id

**order\_summary:**

* **order\_id, warehouse\_id, is\_late\_delivery, shipping\_cost, total\_revenue\_per\_order**

**warehouse\_summary:**

* **warehouse\_id, total\_revenue, total\_shipping\_cost, late\_deliveries, total\_orders, late\_delivery\_rate, avg\_shipping\_cost\_per\_order**

**product\_delay:**

* **product\_id, late\_deliveries, total\_orders, late\_rate**