## Team 20 Code Sample doc

## Environment setup

1. Spark installation: MAC:

<https://github.com/charlie-ph/BigDataAnalytics/blob/master/Installations-HowTos/How-To-Install-Spark-On-MACOS.md>

Windows:

<https://github.com/charlie-ph/BigDataAnalytics/blob/master/Installations-HowTos/How-To-Install-Spark-On-Windows.md>

1. PostgreSQL installation:

<https://www.postgresql.org/docs/current/tutorial-install.html>

## How to run the code

Part 1: PostgreSQL

1. run code/Postgre/create\_file.sql to create Data-warehouse in Postgre
2. import all csv files in /data into your database
3. run code/Postgre/SQL\_basic\_search.sql to get SQL results

Part2: Spark

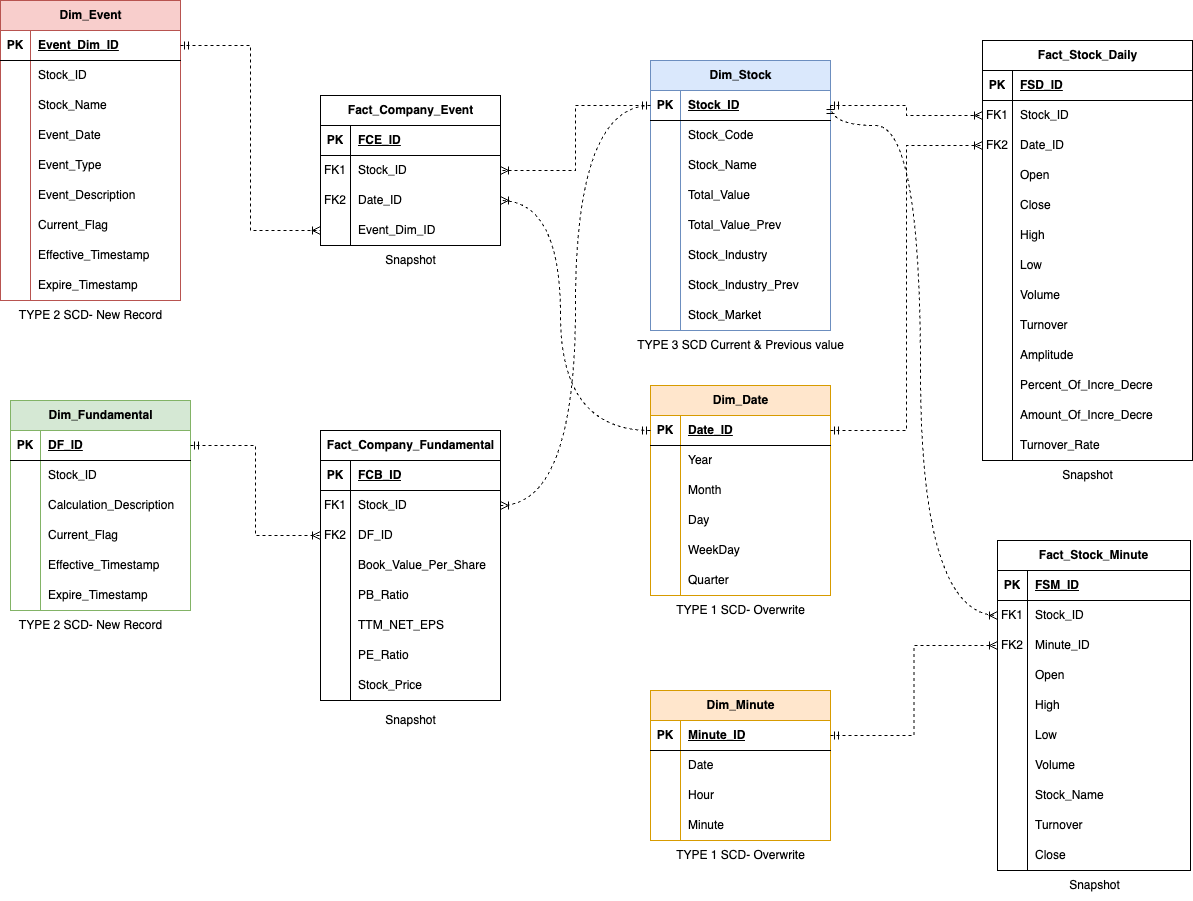
1. run code/Spark/Example.ipynb in your notebook.
2. make sure you have correct address to access all csv files.

## Dataset Explanation

I designed this database according to some basic requirements of my own trading system.



Database ERD:

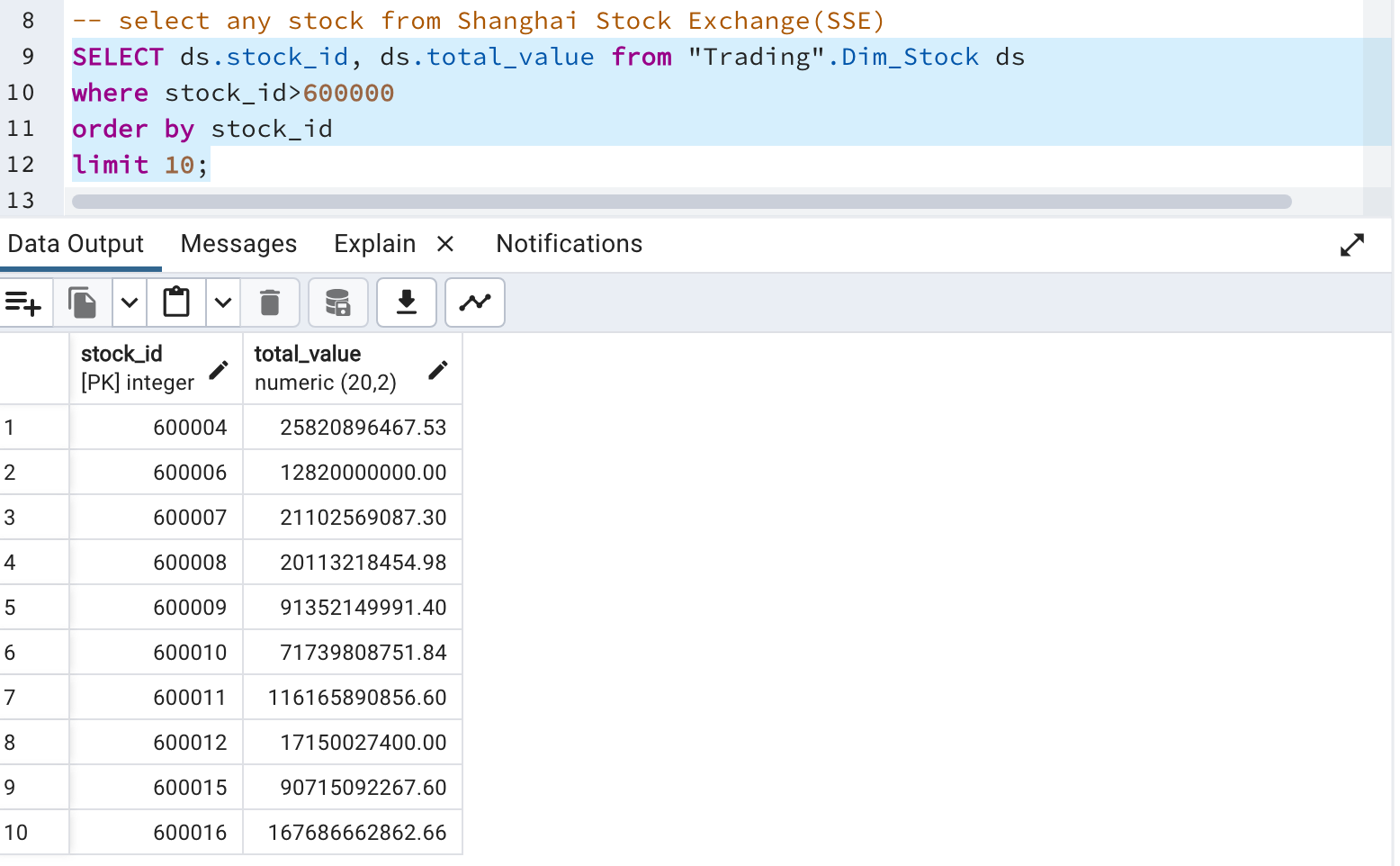


For each table, we provide one notebook to finish the ETL process(code/Database/\*.ipynb). All the data is from this API: akshare(https://akshare.akfamily.xyz/)

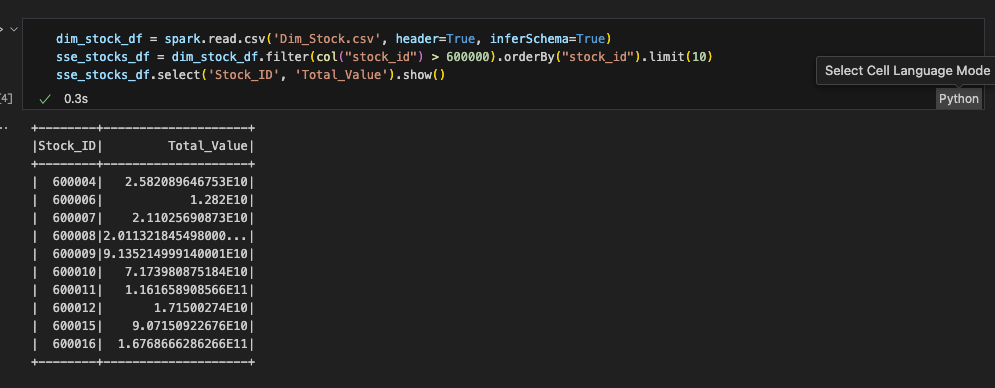
**Results of running the code with data**

**Seq1: select 10 stocks from Shanghai Stock Exchange(SSE)**

PostgreSQL:

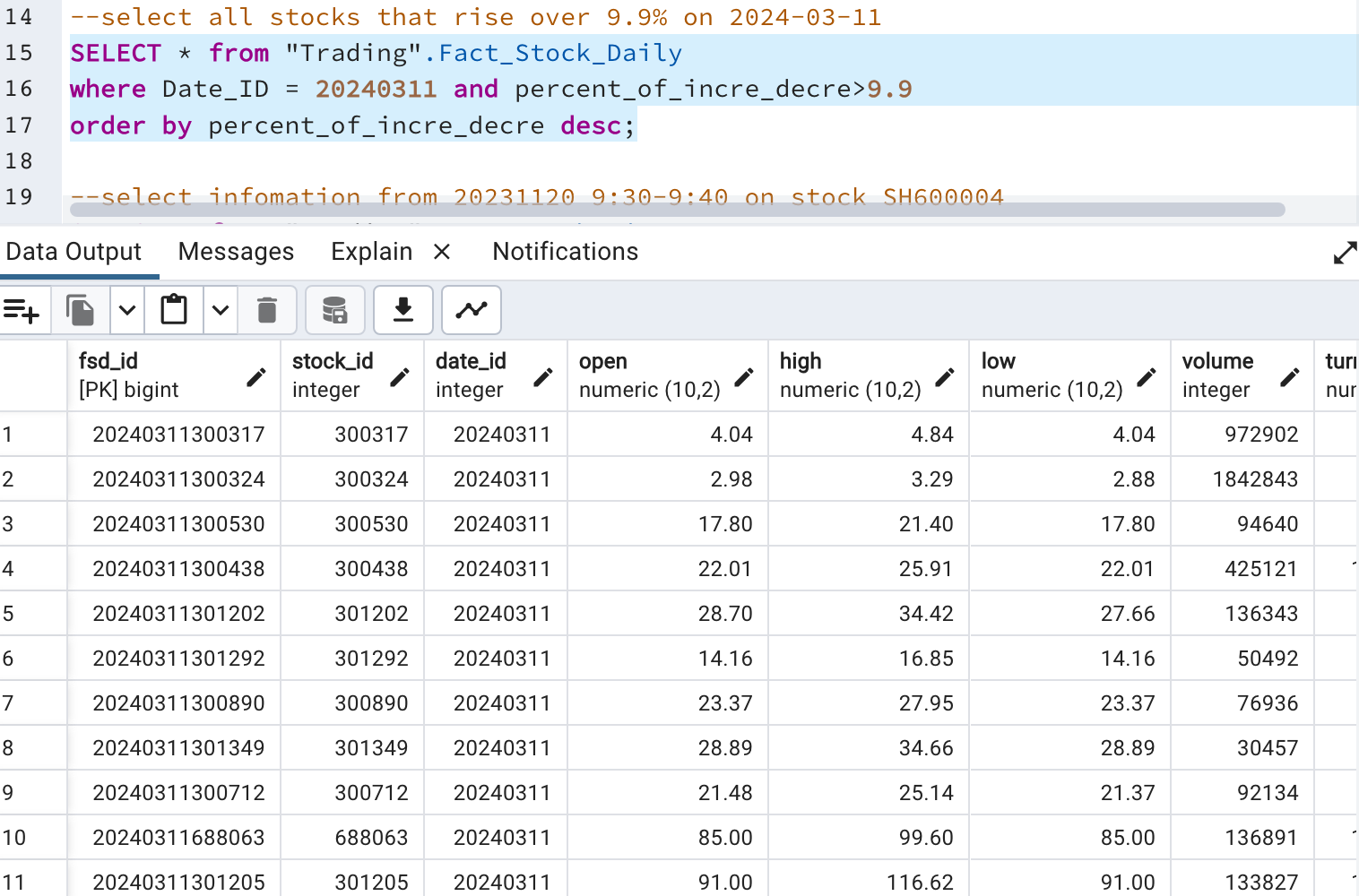


Spark:

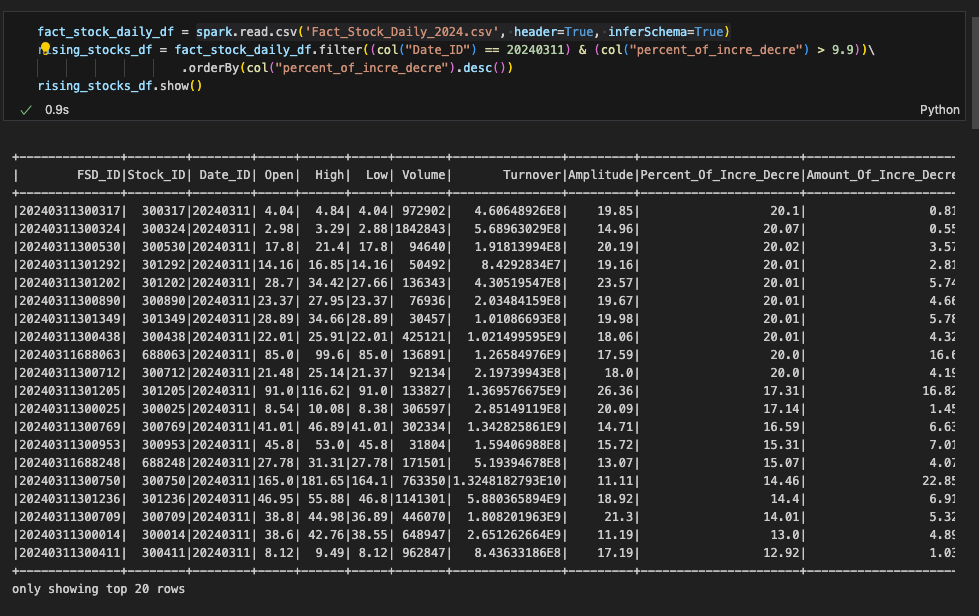


**Seq2: select all stocks that rise over 9.9% on 2024-03-11**

PostgreSQL:

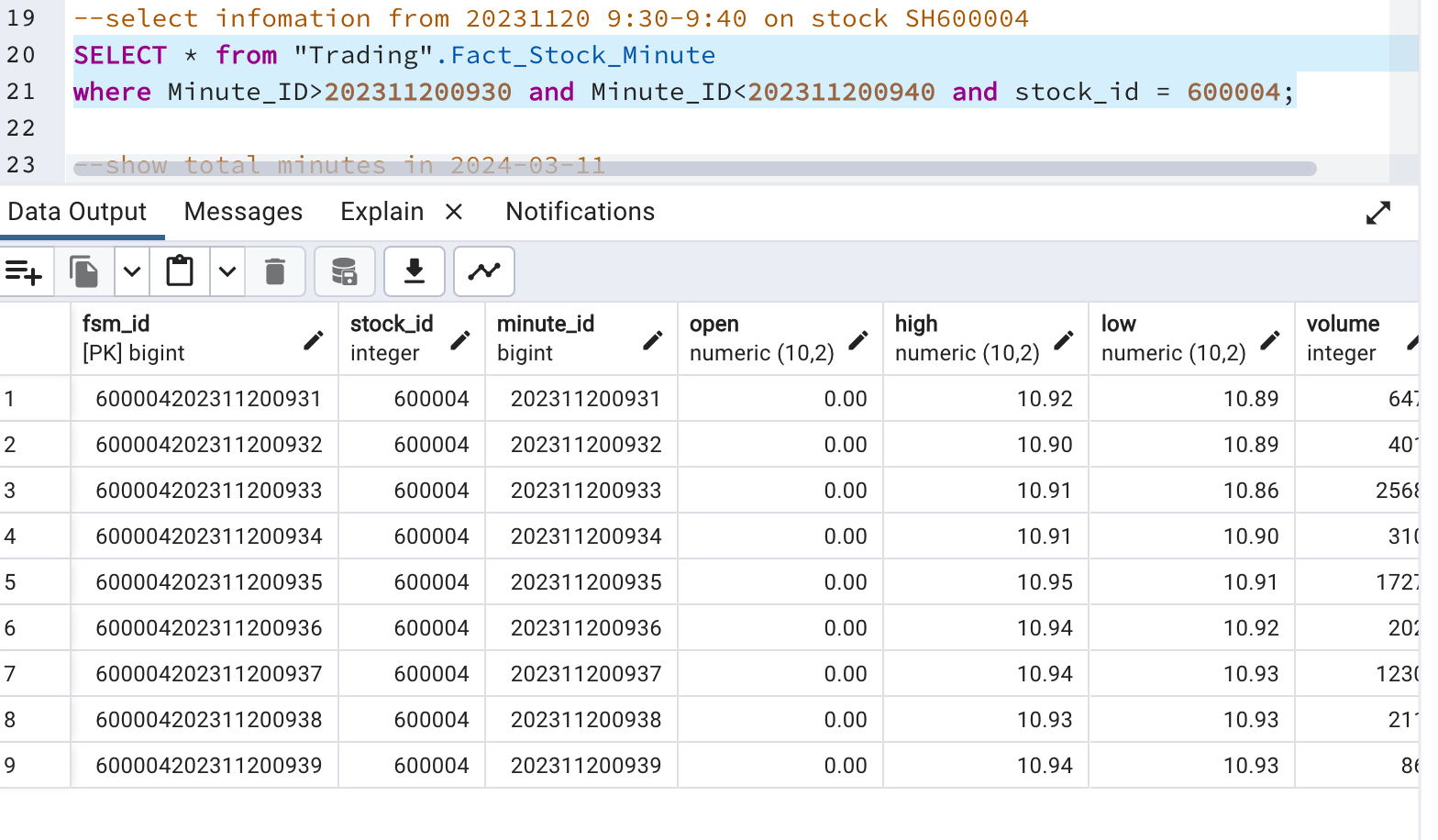


Spark:

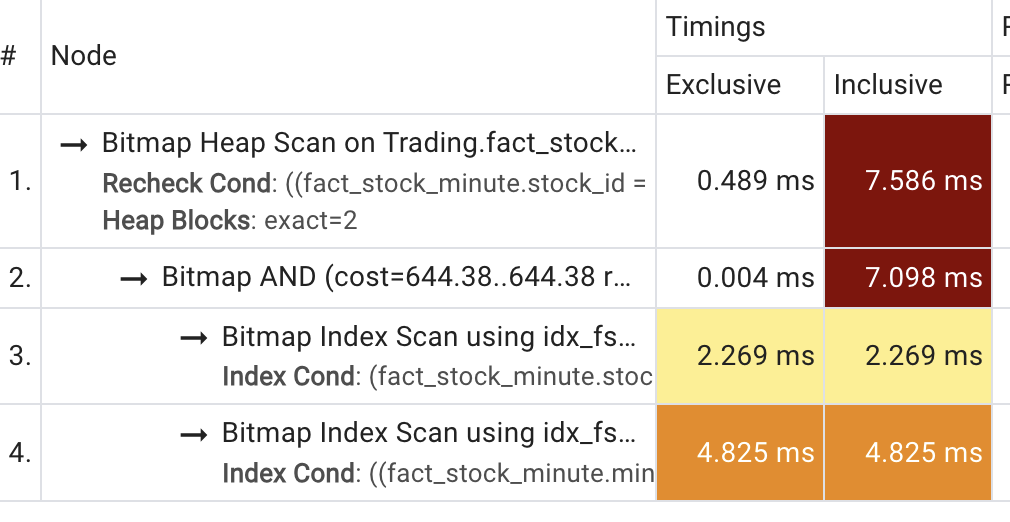


**Seq3: select information from 20231120 9:30-9:40 on stock SH600004**

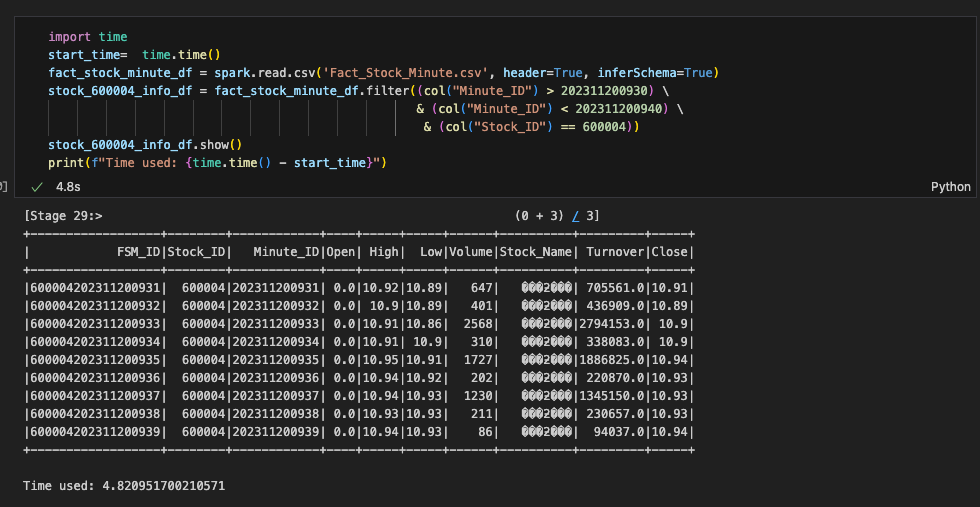
PostgreSQL:



Time Cost for PostgreSQL:

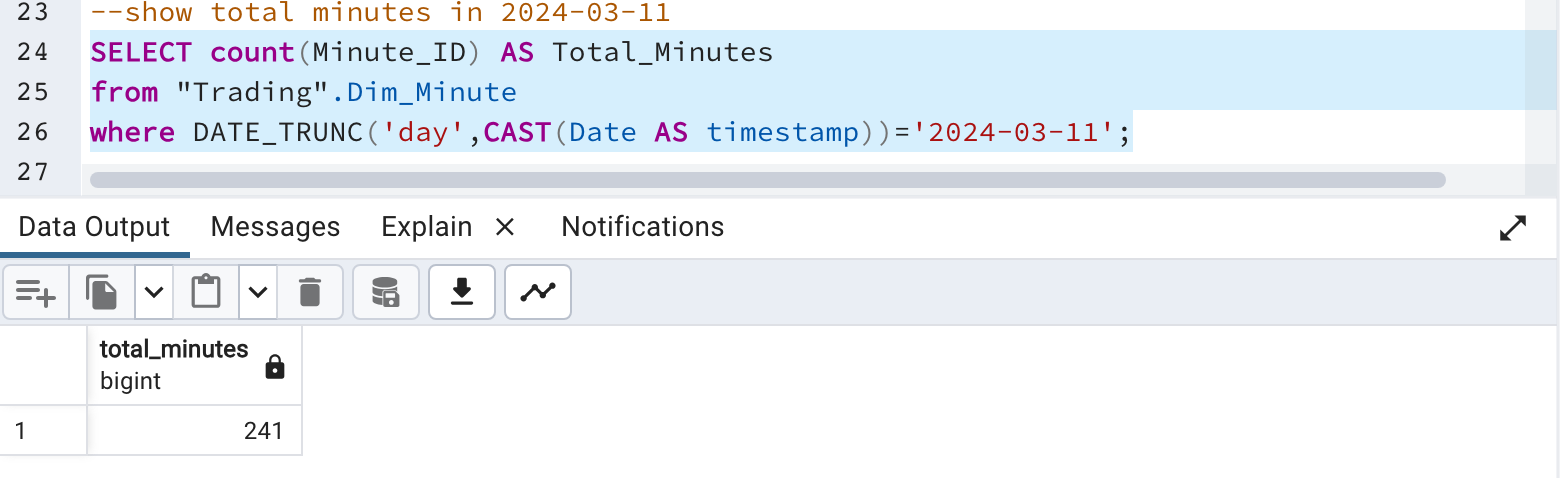


Spark:

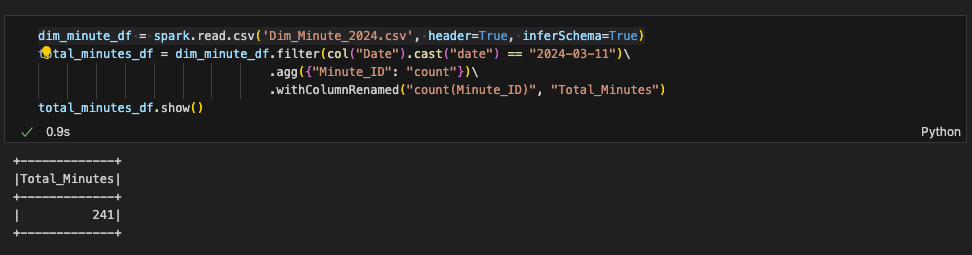


**Seq4: show total minutes in 2024-03-11**

PostgreSQL:

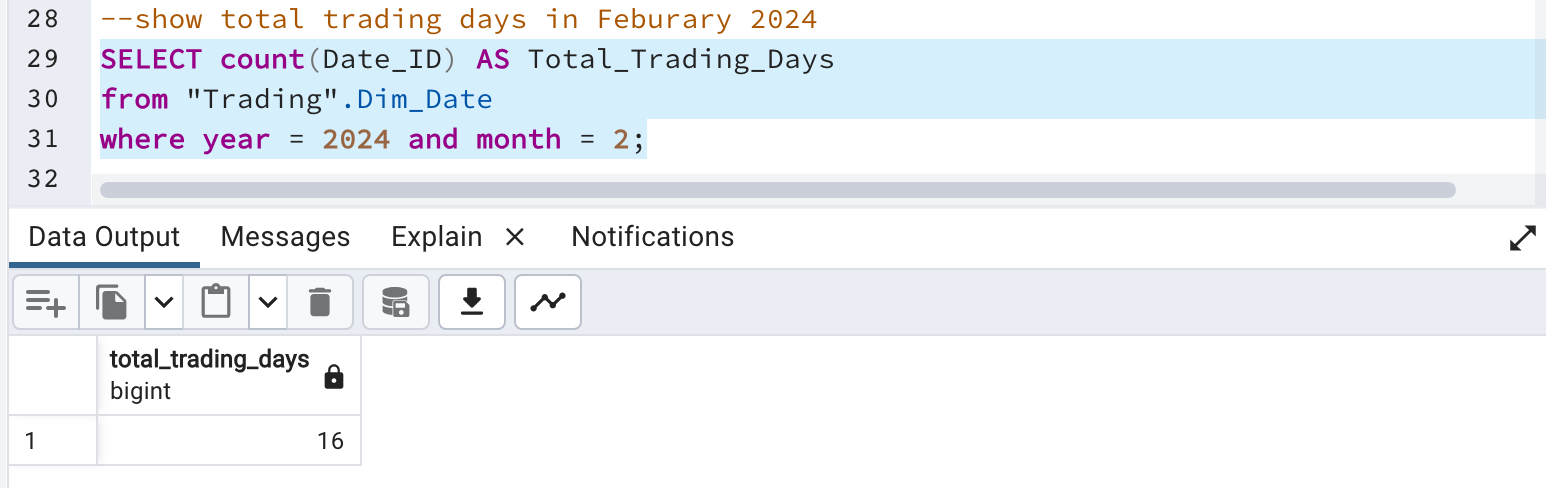


Spark:

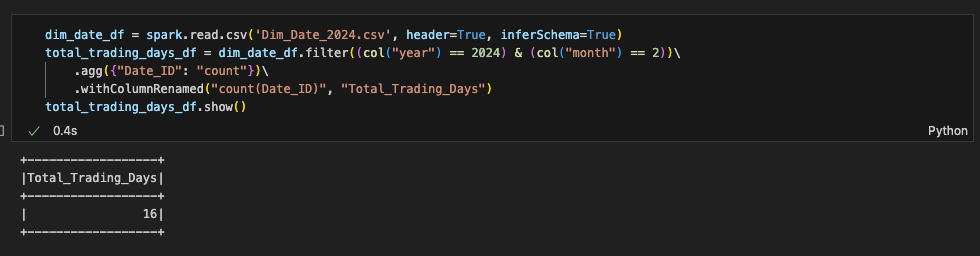


**Seq5: show total trading days in October**

PostgreSQL:

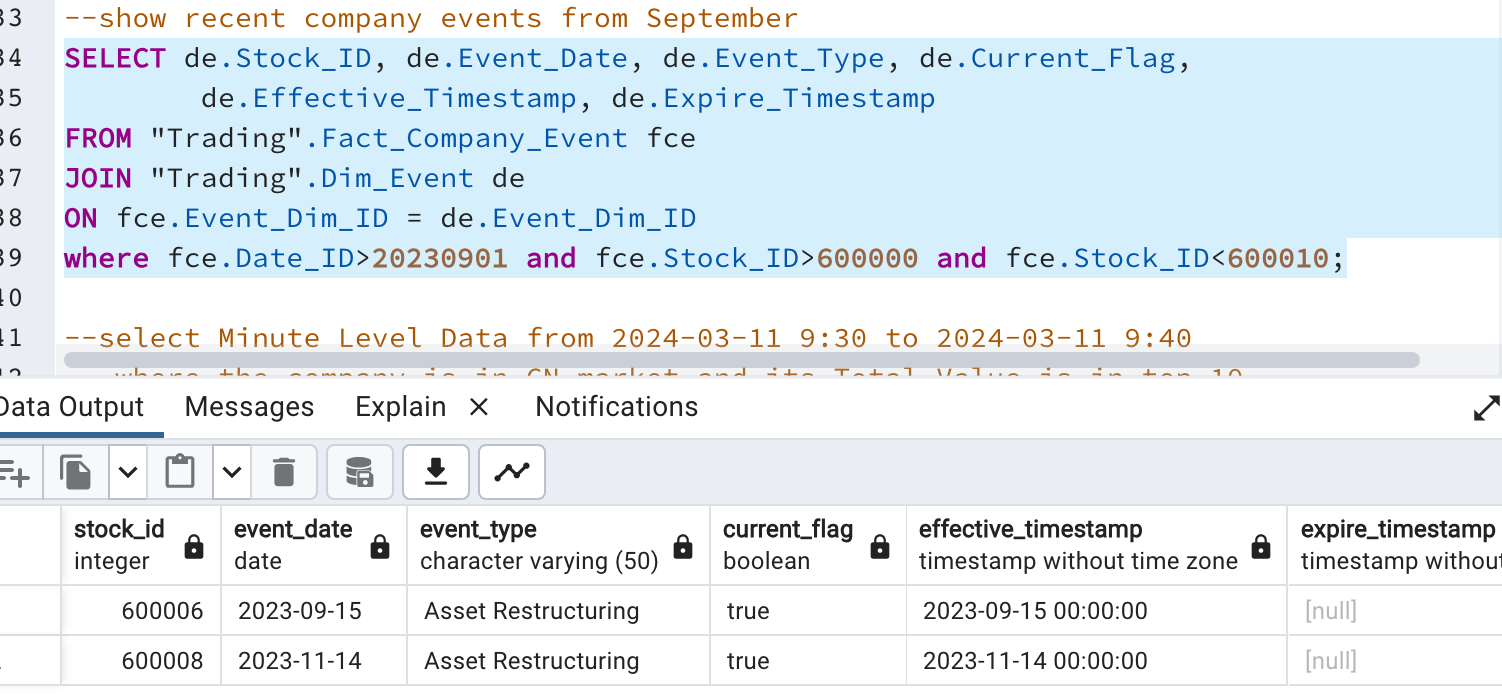


Spark:

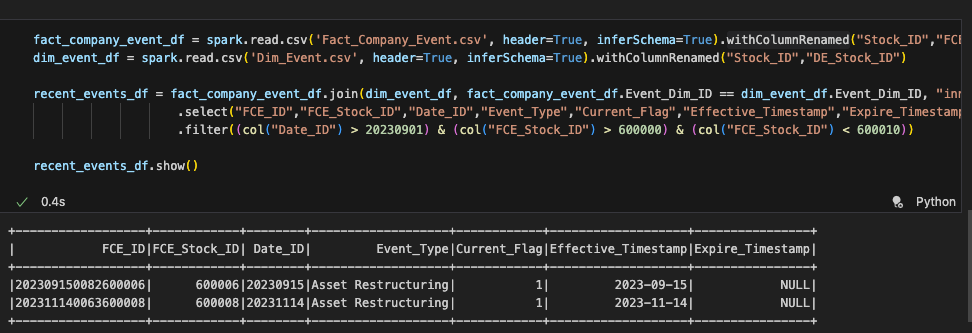


**Seq6: recent company events from September**

PostgreSQL:



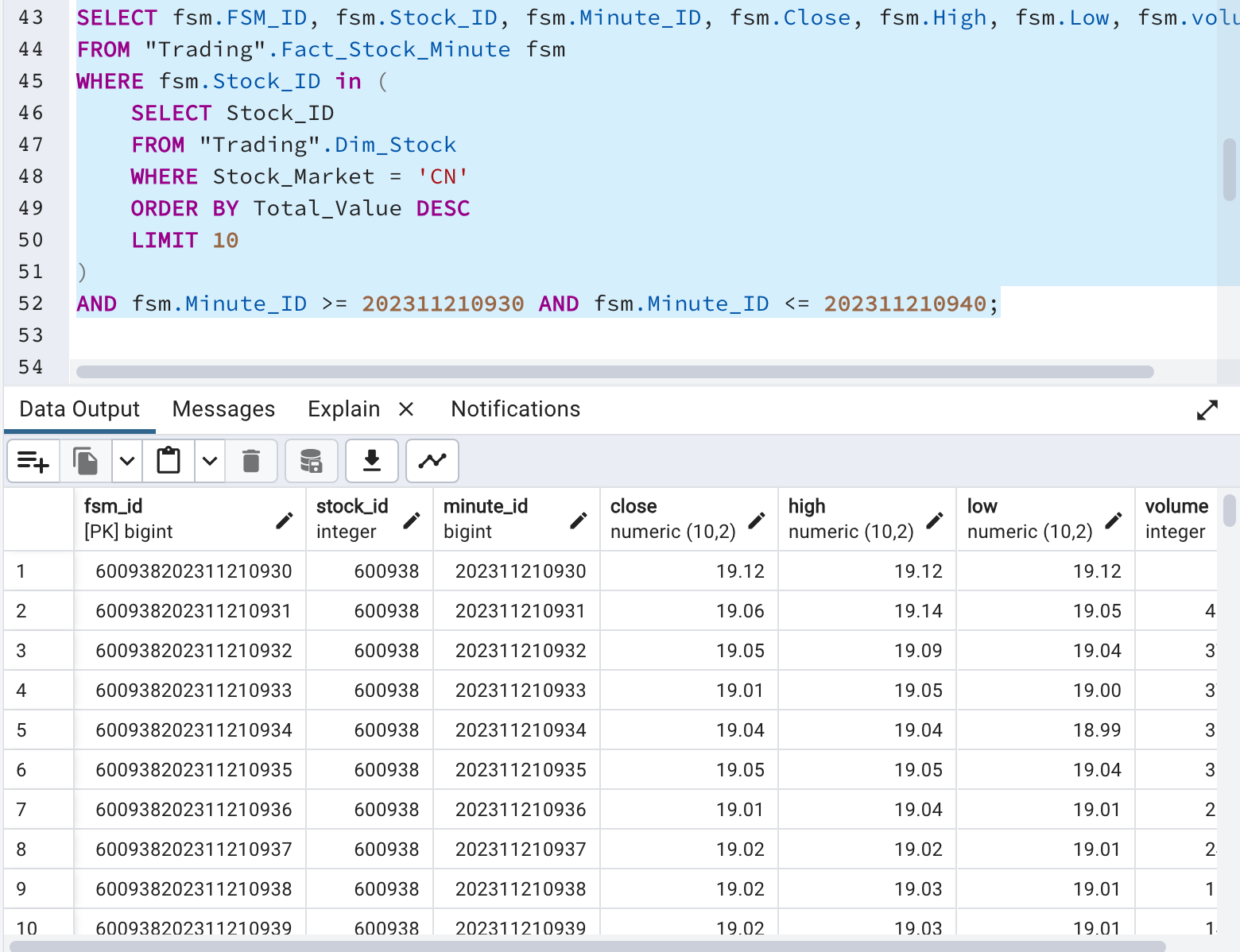
Spark:



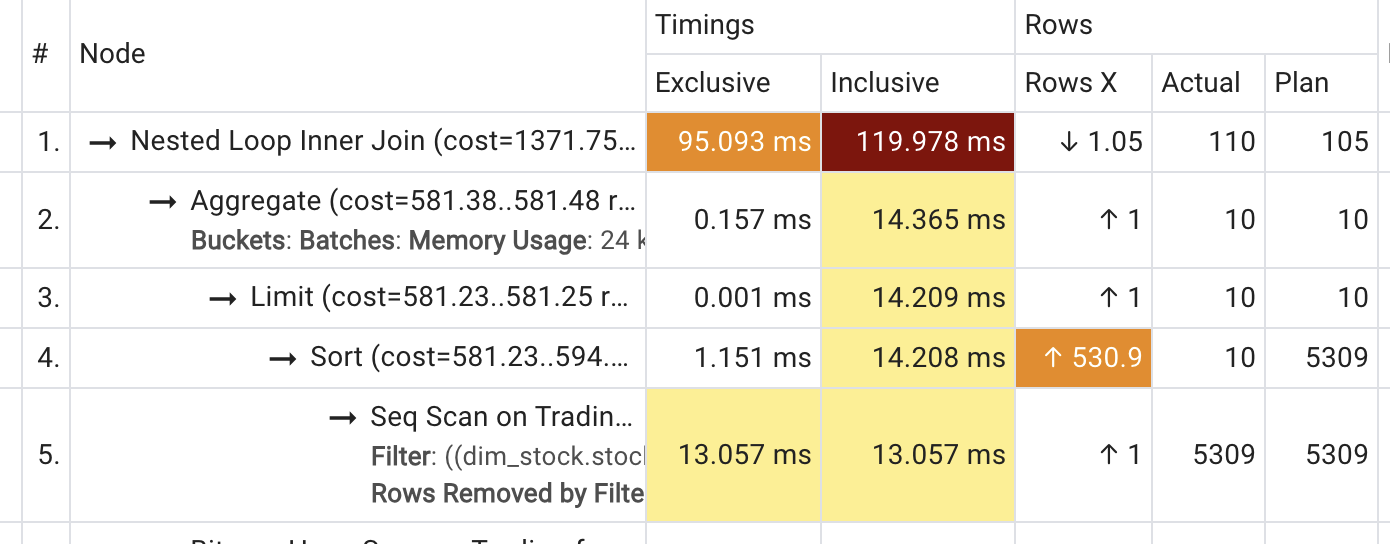
**Seq7 select Minute Level Data from 2023-11-21 9:30 to 2023-11-21 9:40**

**where the company is in CN market and its Total\_Value is in top 10**

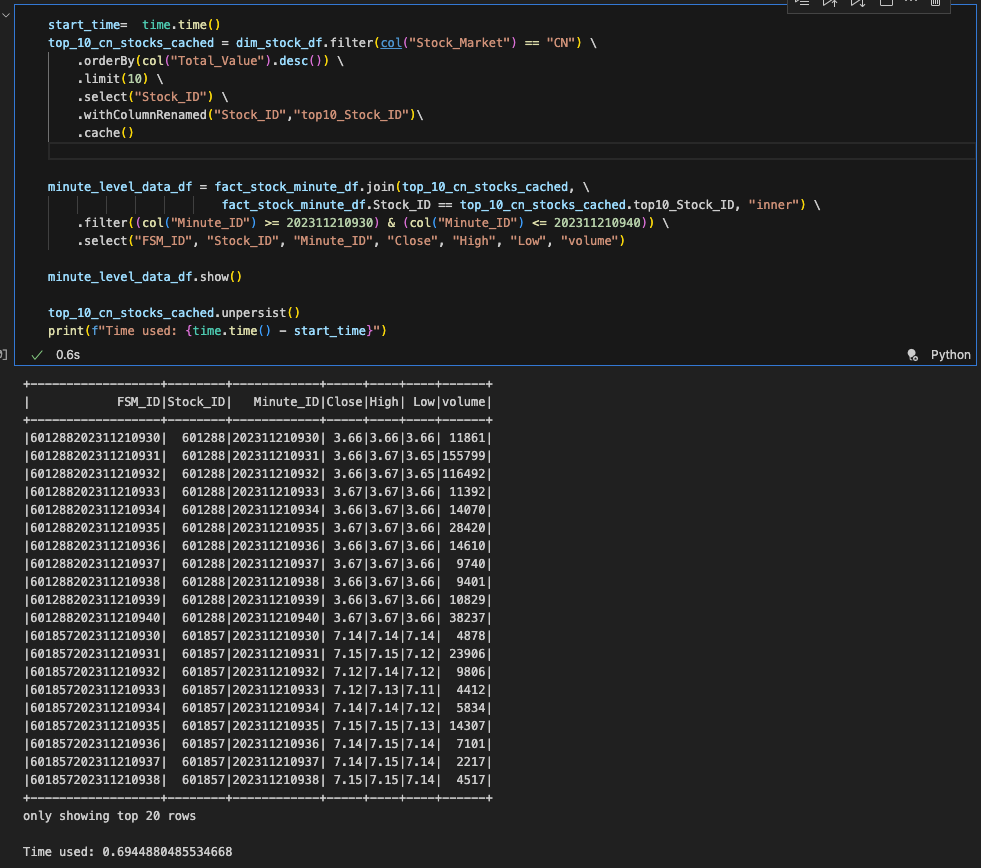
PostgreSQL:



Time Cost: 119.978ms



Spark:



Time Cost: 0.69ms