PolyU COMP4434 Assignment 2

# Introduction

The assignment is to ask you to process an SQL query using Spark.

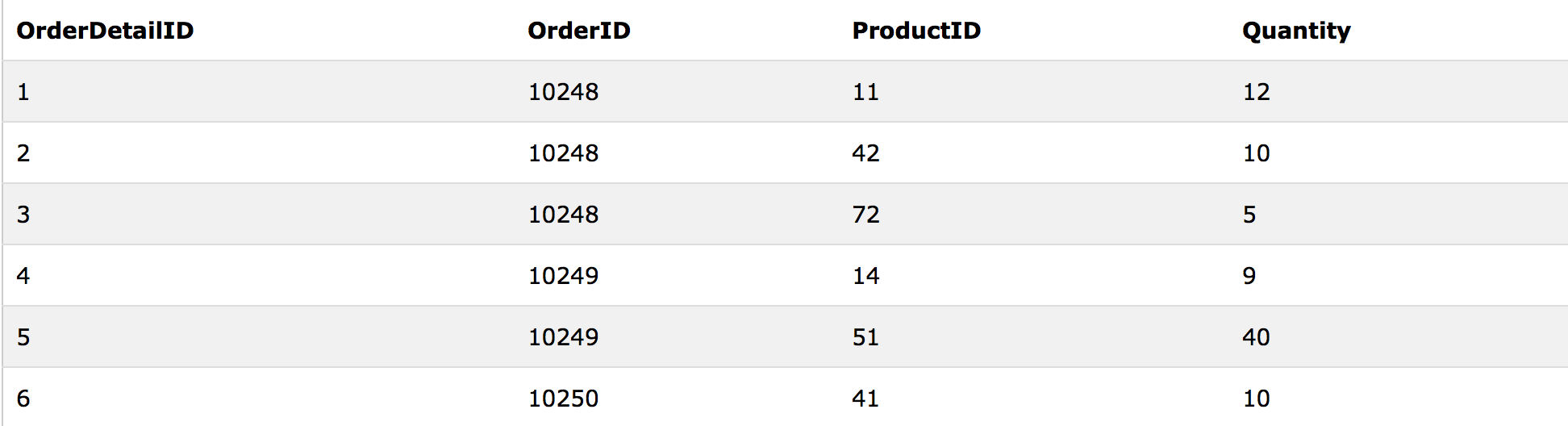
# To begin with

In this assignment, you are supposed to build program to execute an SQL query by Spark, on a given dataset. From the views of and , **three parameters** are set.

* Dataset/table information: Dataset is provided in the folder , so the is **a parameter** in this program. The name of the table is and its schema is fixed:



so the table goes like:



* Query Template: We give a query template with **two parameters: and**.

SELECT , ,

FROM

GROUP BY

HAVING BETWEEN AND

ORDER BY ;

For example, if we set **100** and **120**, then the query is exactly:

SELECT , ,

FROM

GROUP BY

HAVING BETWEEN AND

ORDER BY ;

# You need to program the above SQL query in Spark.

After unzipping our materials, you will see two folders:

1. /Codes

We provide a file called . You need to complete the codes in **.** This method takes and as input, and output the query results.

will print out both the number of records and each element of records that are returned by .

* Please remember NOT to modify any **given** parts of , otherwise your program might not work under our assessment.
* *IMPORTANT REMINDER: you are NOT allowed to use () directly.*

1. /Data

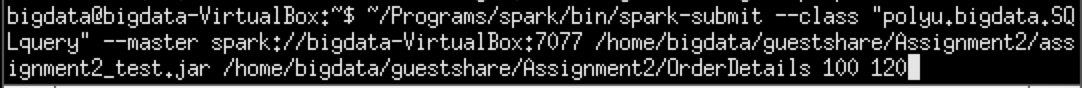
We provide the dataset in this folder. Each line contains a tuple and attributes are delimited by . For example:

in table , it corresponds to:



# Run the program

As you have learned from Spark tutorials, export the “jar” file (e.g., assignment2\_test.jar) of your program then run it with **three parameters**, using spark-submit. For example, in our virtual box environment:



that is:

- …… assignment2\_test.jar

“……” stands for statements when use - and it depends on your environment.

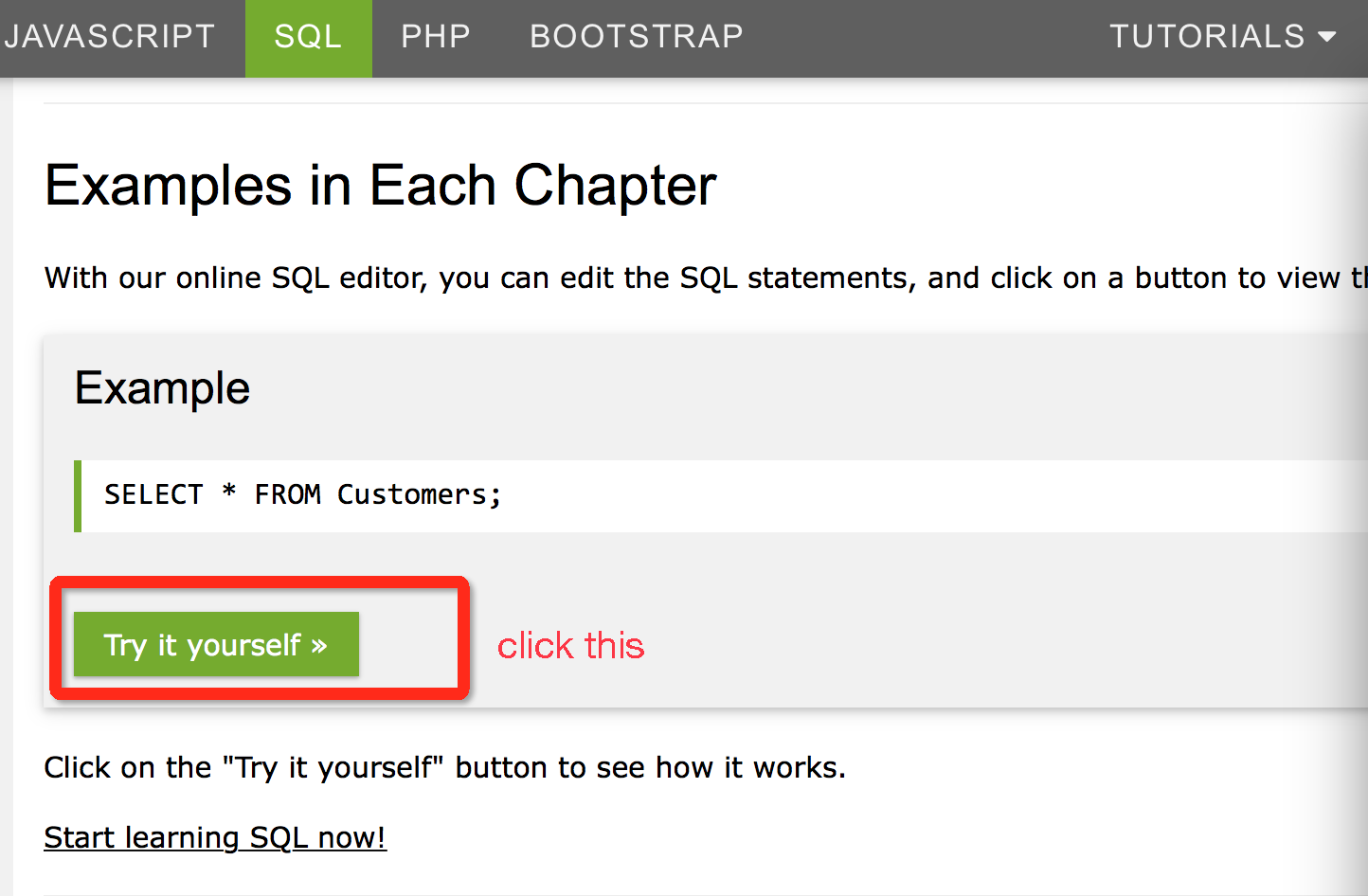
That shall output the following:

**/Users/lizbai/Desktop/Screen Shot 2016-01-30 at 7.02.10 pm.png**

**/Users/lizbai/Desktop/Screen Shot 2016-01-30 at 7.02.03 pm.png**

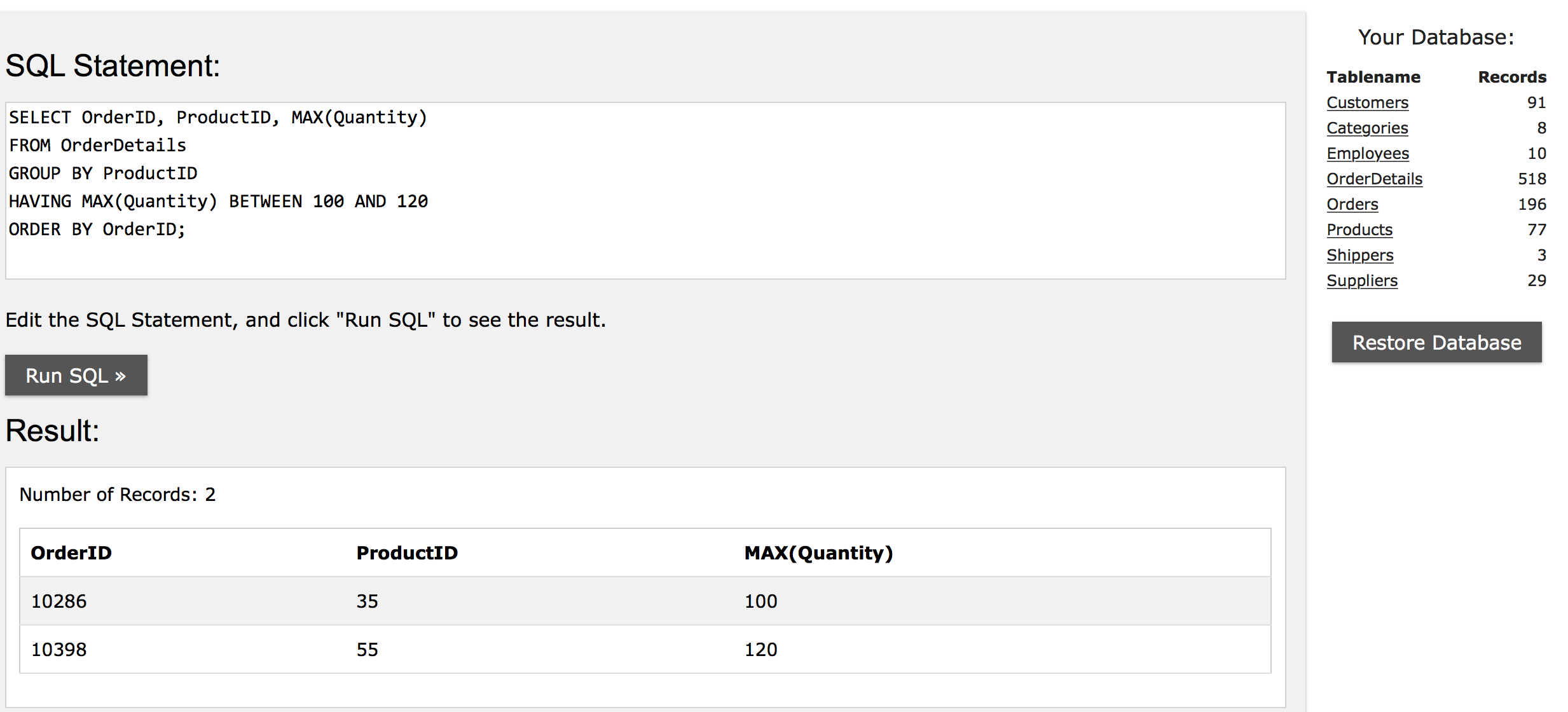
# How to test your program correctness?

* The following website has an online SQL interface to access table
  + Go to <http://www.w3schools.com/sql/>



* + Input the SQL query on , with exact and, run SQL.
  + Compare the results returned from you spark program and W3Schools test.

Take the example in previous section, we set  **100, = 120,** which means the respective SQL query goes as below, and then run the SQL:



The results are the same as we get from spark program, so it works.

# What to submit

Make a **zip file** called **Assignment2.zip** which included the file, submit it through blackboard. You could re-submit your work and we will grade your last attempt.

**WARNING: Submit “Assignment2.zip” instead of scala codes. Use exactly the file name “Assignment2.zip”. Also, use exactly the file name “SQLquery.scala”. You will get 0 marks otherwise.**

# Grading Policy:

We will export a jar file from your source file as jar file and execute it according to “Run the program” section.

We will use several different input files (but in same schema) and set different parameters to test your program, say we use 5 test cases, and if your program only passed 4 of the 5, you get 80.

# DEADLINE: Mar 7 11:59am

# LATE PENALTY

Late X day: your score = raw score \* (100 – 20\* X)%

# Plagiarism

Detecting the similarity of your codes is easy, and cases will be strictly handled according to the University’s regulation, so please don’t risk doing that.

# Questions?

Feel free to ask TA during labs. Or contact [csrbai@comp.polyu.edu.hk](mailto:csrbai@comp.polyu.edu.hk) with email subject “COMP4434 Assignment2 Questions”.