Data Science Course Exercise Workbook

Hive: Course Material

Adam Kershaw, April 2017

Data Science – Hive

Adam Kershaw, April 2017

08

**Fall**

Contents

[Summary and Intro 3](#_Toc483493709)

[Section 1 - Stocks Data: 3](#_Toc483493710)

[1.1 Database &Table Creation 3](#_Toc483493711)

[1.2 Querying 101: 4](#_Toc483493712)

[1.3 Further Querying: 4](#_Toc483493713)

[1.4 Querying across datasets: 4](#_Toc483493714)

[Section 2 – San Francisco Salaries Mini Project: 5](#_Toc483493715)

[2.1 Scenario & Data: 5](#_Toc483493716)

[2.2 Task: 5](#_Toc483493717)

[Additional Research Task 5](#_Toc483493718)

[Summary 5](#_Toc483493719)

## Summary and Intro

This workbooks aim is to aid and facilitate Data Science trainees learning. This workbook will cover aspects about using apache Hive, to manage and explore large datasets residing in distributed storage.

1. Stocks – Hive Exercises
2. San Fran Salaries - Hive Mini Project

This workbook and subsequent workbooks will act as a record of the trainees work, learnings and understanding.

You should aim to complete all questions. Sections relate to course slides, so the workbook can be completed in stages or once all theory has been delivered.

# 

# Section 1 - Stocks Data:

This section will focus on the Stock data files, please complete all exercises.

**Scenario:**

Come up with some sort of case for this other than the basic “getting to grips”.

**Data:**

For the following exercises in this section we will be using the Stocks data. This data is in 6 CSV files: appl\_daily, face\_daily, nasdaq\_daily, nflx\_daily, twtr\_daily, yhoo\_daily.

Each of these files has the same following attributes: Date, Open, High, Low, Close, Volume, Adj\_Close.

It is recommended to use ‘Geaney’ (a unix alternative to notepad++) to develop your hiveQL code, rather than develop in the command line.

## 1.1 Database &Table Creation

Write a hiveQL statement to create database location for the stock data.

1. Create a Table inside of your database for apple data and load in the appl\_daily .csv file.
2. You can create a Hive Managed or External Table, It is up to you but make sure you can justify your decision
3. In order to test your table is set up correctly; use your SQL knowledge to retrieve the data.
4. Once completed write a script to create tables for the remaining stock datasets (6 Tables in total) **Ensure to use hive Command Line functions to execute your script.**

## 1.2 Querying 101:

This section will test your ability to write efficient SELECT statements in hiveQL, making use of the WHERE clause.

Write a hiveQL query to determine:

1. How the Highs of apple change over time (The entire data set)
2. Apple’s stock data for the last 100 days recorded
3. Apple released the 1st generation IPhone in September 2007, extract the stock data for this month.
4. Apple’s earliest recorded stock data is from 1980, extract the data for 1980’s last day of trading.
5. Steve Jobs joined Apple as CEO in 1977, and left Apple in January 2011, to focus on his health. Extract the stock data the end of January 2011. Comment on how the stock price changed under his lead (1980 – 2011)?

## 

## 1.3 Further Querying:

This section will test your ability to use Sub queries, aggregations & GROUP BY in hiveQL.

Write a hiveQL query to determine:

1. **What was apples stock when it was 1st traded (Earliest existing record)**
2. **When was apples highest close stock value**
3. **What date was apple's greatest volume traded**
4. **What is the average difference between high and adj close for apple**
5. **Determine the total volume traded each month for apple for 2015.**
6. **Break this out by years 2011-15, is there any kind of year on year trend/ pattern? or anomalies? Look into apples history to suggest why these occur?**

## 1.4 Querying across datasets:

This section will test your ability to write hiveQL queries across tables; you may need to use a combination of UNION, JOINS and built in hive functions.

You have been tasked with comparing the historic stock data for two of the social media industries biggest players (Facebook & Twitter).

Use a single query to answer the following questions:

Note: You will need to use joins or union logic

1. Retrieve and output to a file Facebook and Twitters stock price data for the calendar year 2015.
2. Use a single query to identify the month of the year with the highest volume of trades for each company in 2015?
3. Use a single query to identify the month of the year with the lowest volume of trades for each company in 2015?
4. Alter your query to output the average volume of trades for each company & month of the year. Order the result set by company, average volume of trades.
5. Output question 4’s result set to a file & ensuring the query runs in silent mode
6. How do the results differ between the two companies?
7. Adapt your query to include the years 2013 & 2014. Is there any year or year trend/ pattern? Or anomalies? (Hint: Look for typically high/ low trading months)

# 

# **Section 2 – San Francisco Salaries Mini Project:**

## 2.1 Scenario & Data:

One way to understand how a city government works is by looking at who it employs and how its employees are compensated.

The San\_Fran\_Salaries dataset contains the names, job title, and annual income data for San Francisco City employees from 2011 to 2014.

## 2.2 Task:

Put your new hive skills into practise to explore the San Francisco salaries dataset. You can focus on a specific idea, or keep it general. The angle for analysis is up to you!

The dataset is not perfect so you’ll have to watch out for and work around missing values and duplicate records.

To help get you started, here are some data exploration ideas:

• How have salaries changed over time between different groups of people (Job Titles/ Regions/ Job Type)?

• How are base pay, overtime pay, and benefits allocated between different groups?

• Is there any evidence of pay discrimination based on gender in this dataset?

Prepare a 10 minute presentation to present to the class at the end of the day.

To Include:

* Introduction to your presentation, and chosen angle for analysis
* Analysis plan (what do you want to find out? how are you going to it?)
* Findings (With accompanying screenshots of the results & hiveQL code)
* Any challenges/ issues with the data or hive? And how did you resolve them?
* Conclusion

This task is intended to be completed individually, unless otherwise stated by the trainer. It will be graded on both content and professional practise.

## Additional Research Task

Note: This is only to be attempted if you have completed all previous tasks

In hive it’s possible to organise tables into **partitions and buckets**, to allow for more efficient querying.

* Research the use cases of partitioned tables, and how to create them
* Test out partitioning by re-creating the San\_Fran\_Salaries table
  + Partitions can be added when a table is created or altered

# Summary

Upon completion of this workbook’s exercises and hive mini project:

* Created Databases and Tables in Hive
* Projected structure upon data held in Hadoops Distributed File System
* Used hiveQL queries to analyse two large datasets from two large real-world datasets (Stocks, San Francisco Salaries)
* Put into practice hiveQL’s aggregation, filter, joins and union logic functionality.