Project 2

For this project you'll have a group. Those have been automatically assigned and you can find your groups's information on the assignment link **You should not discuss the project with anyone outside your group. If you are stuck please contact Dr. Post or the TA, Khuzaima Hameed.** If you run into issues with your partners, let me know immediately. I can't help if it is the day before the due date and you tell me you haven't heard from them. You'll be stuck doing it yourself then!

This project is meant to assess your ability to understand big data processes, your use of SQL type code, and basic Spark code. This project has three parts.

- The first part is a report that provides a synopsis of a company's big data pipeline. You'll find the company assigned to you and your partner on the assignment link.
- The second part of the project requires you to code. You'll create a python notebook (`.ipynb` file) and output an `.html` file.
- The third part of the project is to redo the HW 4 NFL Map Reduce part using spark. You'll create a python notebook (`.ipynb` file) and output an `.html` file.
- You should submit five files to wolfware: the report file for part 1 and the `.ipynb` files/corresponding `.html` files for parts two and three.
- **All group members should submit the files.**
- **When submitting, please leave a note about how working with your group went for the writing section (part 1) and for the coding sections (parts 2 and 3).**
- You should each be doing the entire project as a collaboration. Repeat: You should not split the project up and only do certain parts yourself. **If you have no part in writing one section, you will receive a 0 for that part of the project.**

Part 1 Details

I'd like you to research, summarize, and write-up the general process the company you are assigned to uses for their big data pipeline. Each company I've listed also has a reference or two linking a talk or article. You should research further than just the links I've provided.

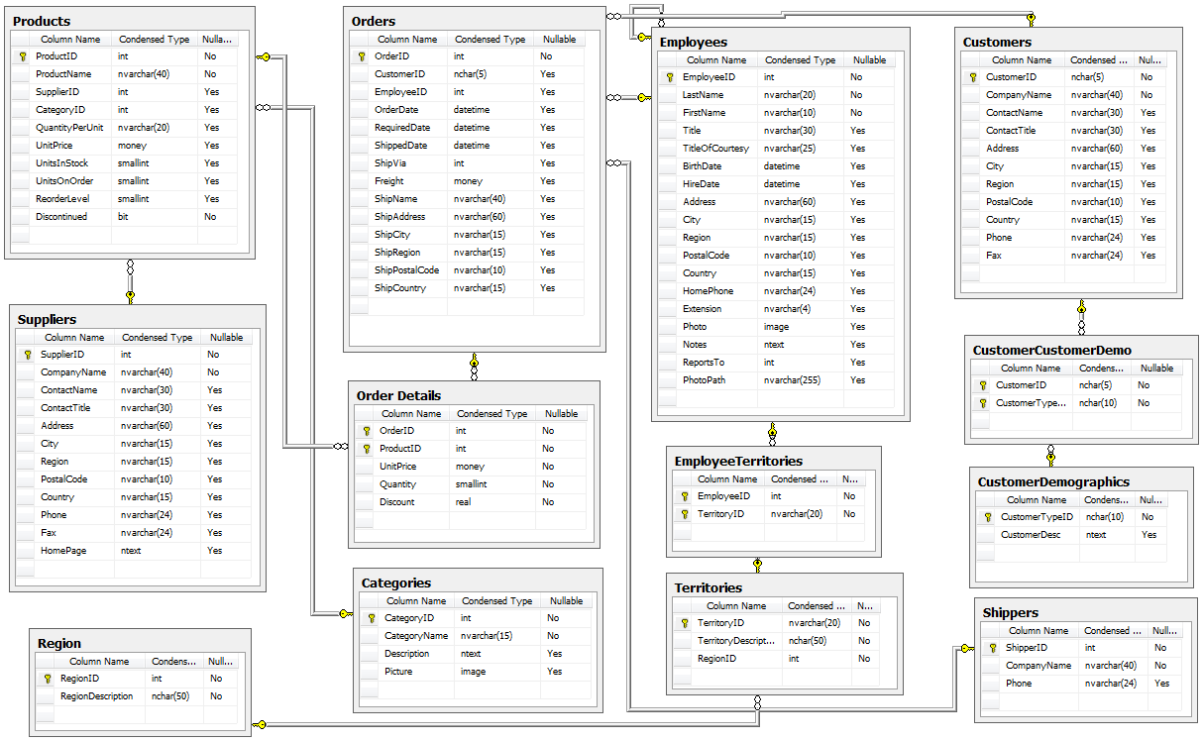
Your report should be 4-6 pages. Discuss the types of problems that company faces with their data and the types of questions they try to answer. Be sure to describe the main purpose of most of the major software you see in their big data pipeline.

Your audience for this report is a fellow student in this course. You can write this in word, open office, or markdown - that is up to you. You should include pictures where possible and cite your information.

Part 2 Details

On the assignment link you'll find an sqlite database called `northwind.sqlite`. This database originally comes from https://github.com/jpwhite3/northwind-SQLite3 but has been modified.

Your task is to write up a report about the employees. That is, you want to understand how much they sell, what products they are able to sell, how they've done across years, how they do in different regions, etc. You'll want to study the database tables a bit to get an idea about the things you could investigate:



You should provide summary statistics and graphs with corresponding interpretation. In the end, for each employee you should describe their main strength and something they can improve upon (backed up by your summary stats and graphs of course). Feel free to use SQL through pandas here to obtain the tables. You can do the joins, summaries, etc. through pandas if you'd like.

Part 3 Details

For part 3, you'll want to take the code you created for homework 4 and do similar things via spark. (This part should be very easy/short if you have spark up and running! You can use the spark SQL or pandas-on-spark functionality rather than writing your own MapReduce type code.) That is, you want to

- read in the full nfl data set into spark a spark data frame or pandas-on-Spark data frame
- use spark SQL or pandas-on-Spark to find the mean and standard deviation for the AQ1, AQ2, AQ3, AQ4, AQFinal, HQ1, HQ2, HQ3, HQ4, and HFinal variables
- repeat the above process but do so for each value of the season variable
- convert the resulting means from the previous part to a pandas data frame (not spark) and plot the means for the 'quarter' variables across season (that is, put season on the x-axis, mean of AQ1, AQ2, . . . , HQ4 on the y-axis, using different colors for each line with a legend).
- **As always**, you should have a basic narrative flowing through what you are doing and an interpretation of any stats/graphs created.

Rubric for Grading (total = 100 points)

Item	Points	Notes
Big Data Report (generally graded for clarity and thoroughness)	35	Worth either 0, 5, 10, . . . , 35
Employee Report (ER)	40	See below
(ER) Reasonable/thoughtful questions investigated	10	Worth either 0, 3, 7, or 10
(ER) Appropriate graphs and summary statistics to answer questions	15	Worth either 0, 3, 7, or 10
(ER) Good discussion of summaries	10	Worth either 0, 3, 7, or 10
(ER) Reasonable conclusions for strengths/improvements for each employee	5	Worth either 0, 3, or 5
NFL on Spark (S)	25	See below
(S) Reading in data appropriately	5	Worth either 0 or 5
(S) Statistics found correctly via spark	12	Worth either 0, 4, 8, or 12
(S) Graph	8	Worth either 0, 4, or 8

Notes on grading:

- For each item in the rubric, your grade will be lowered one level for each each error (syntax, logical, or other) in the code and for each required item that is missing or lacking a description.
- **You should use Good Programming Practices when coding (see wolfware). If you do not follow GPP you can lose up to 25 points on the project.**

The reports should include a narrative throughout, section headings, graphs outputted in appropriate places, etc. To be clear **be sure to include markdown text describing what you are doing, even when not explicitly asked for!** Points will be deducted from appropriate sections as appropriate.