Big Data 210 Final Project

Exploring SEC Financial Statement Data Sets

Kirk Force

December 10, 2019

Data Background

- The U.S. Securities and Exchange Commission (SEC) requires publicly traded companies to file earnings reports on a quarterly and annual basis.
- They make available data sets consisting of all financial statements that were filed with the commission using the eXtensible Business Reporting Language (XBRL) on a quarterly basis back to Q1 2009.
- This includes about 50 different types of submission forms.

The Objective

- Original goal: Use the financial statement data to attempt to calculate aggregate quarterly financial statistics and compare it to macroeconomic variables (GDP, unemployment rate, stock market performance, etc.)
- Revised: Filter financial statement data down to members of the Dow Jones Industrial Average to calculate statistics.

Data Structure

- Each quarter's data set consists of four files:
    - sub (Submissions): A file with all form submissions and information about the filing companies.
    - tag (Tags): A file with all documentation labels for numbers presented on the reports (e.g. "SalesRevenueNet" is a tag representing total revenue). This includes both standard accounting tags as well as company specific custom tags.
    - num (Numbers): A file with all numeric facts presented on the financial statements.
    - pre (Presentation of Statements): A file with information on how all tags and numbers were presented in the financial statements.

Data Size

- I downloaded files for Q3 2011 - Q3 2019. This consisted of a total 132 files, where the zipped size of the files was $\approx$ 1.4GB, and the unzipped files turned out to be about 13.5GB.
- There were 243,772 submissions in the full set (6000-7000 submissions per quarter
- There were 87,000,000 numerical entries (1.5 - 4 million per quarter.
- There were 2,638,319 unique tags (much to my dismay).

Challenges

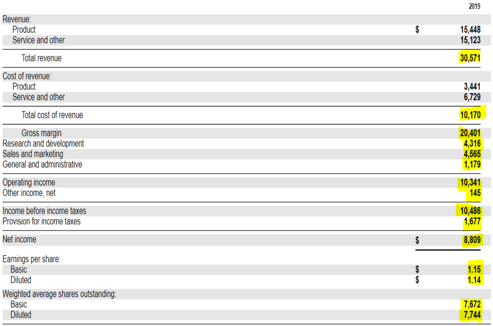
- There were several challenges that made getting clean aggregate numbers non-trivial:
    - There are a large number of tags that represent similar (or the same) concepts. The mix of custom and standard tags only complicated this issue.
    - Accounting standard governing bodies change their tags through time.
    - Companies change the tags they use through time.
    - Companies can revise/restate financial information through time.
    - Companies have a wide variety of reporting dates and fiscal year ends.
    - Companies go in and out of existence, change their names, merge, etc.
    - Companies can change the frequency on which they report data through time.
    - Understanding the data relies on fairly specialized accounting knowledge.

The Approach

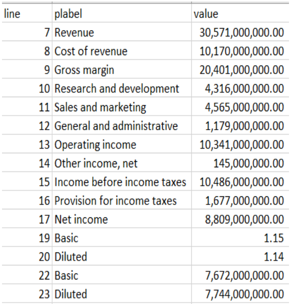
- I sent the 4 sets of files to my VM and the Databricks cluster as full directories, read each into a Spark dataset and re-saved as Parquet.
- In order to simplify the data and solve some of the problems from the last slide, I did the following:
    - Filtered numbers to only values that represented quarterly time periods and annual time periods.
    - Filtered submission forms to only included 10-K and 10-Q forms.
    - Joined num and sub datasets to restrict each remaining numerical value to its most recent reporting date.
    - Filtered data down to only members of the Dow Jones Industrial Index (includes a time component).
    - Attempted to derive quarterly values from yearly values where necessary.
- I used SparkR to analyze the data once filtered down.

Results Part 1: Building a Financial Statement

- I was able to figure out how to build financial statements from a given submission. Below is an example of Microsoft's Q1 2019 income statement:



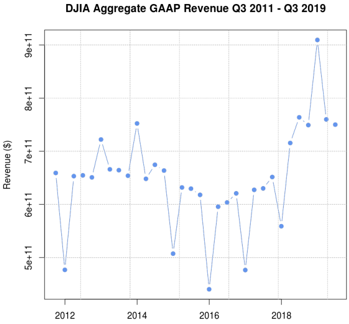
		2019
Revenue:		
Product	$	15,448
Service and other		15,123
Total revenue		30,571
Cost of revenue:		
Product		3,441
Service and other		6,729
Total cost of revenue		10,170
Gross margin		20,401
Research and development		4,316
Sales and marketing		4,565
General and administrative		1,179
Operating income		10,341
Other income, net		145
Income before income taxes		10,486
Provision for income taxes		1,677
Net income	$	8,809
Earnings per share:		
Basic	$	1.15
Diluted	$	1.14
Weighted average shares outstanding:		
Basic		7,672
Diluted		7,744

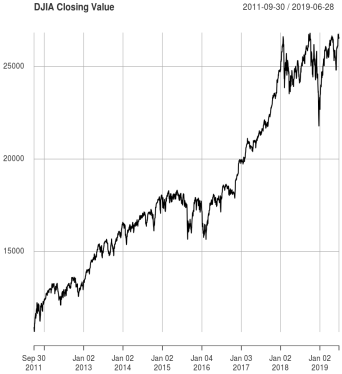


line	plabel	value
7	Revenue	30,571,000,000.00
8	Cost of revenue	10,170,000,000.00
9	Gross margin	20,401,000,000.00
10	Research and development	4,316,000,000.00
11	Sales and marketing	4,565,000,000.00
12	General and administrative	1,179,000,000.00
13	Operating income	10,341,000,000.00
14	Other income, net	145,000,000.00
15	Income before income taxes	10,486,000,000.00
16	Provision for income taxes	1,677,000,000.00
17	Net income	8,809,000,000.00
19	Basic	1.15
20	Diluted	1.14
22	Basic	7,672,000,000.00
23	Diluted	7,744,000,000.00

Results Part 2: Dow Revenues

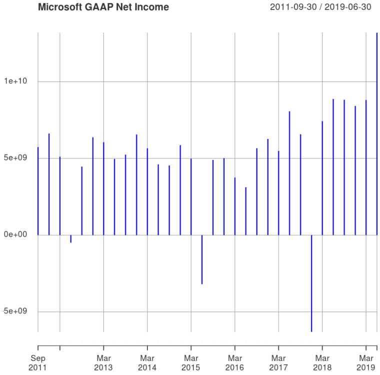
- I made an attempt to arrive at aggregate Dow Jones Industrial Average revenues.

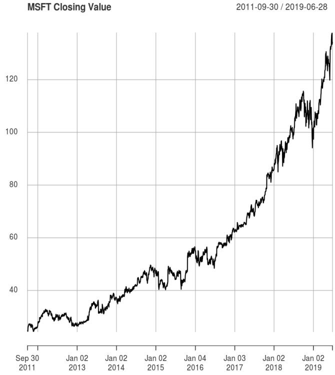




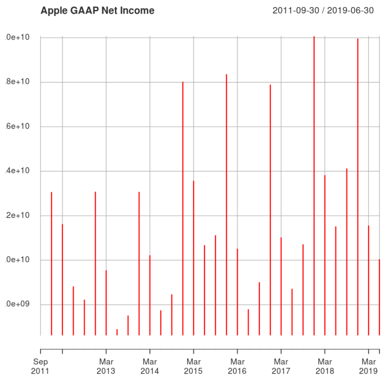
Results Part 3: Individual Earnings vs. Stock Performance

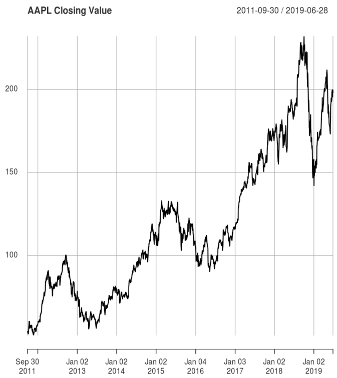
- On an individual basis, I was able to extract company-specific line items. I've included Microsoft and Apple net income numbers and their stock prices.





Results Part 3: Individual Earnings vs. Stock Performance


Apple GAAP Net Income


AAPL Closing Value

Conclusion

- Unable to generate much of interest from this data set in the time provided.
- What I did get was the following:
  - A trial-by-fire lesson on the structure of SEC earnings data.
  - Strong practice with Spark in the Databricks environment.
  - A good look at SparkR.