**Part A** (**Data Modelling)**

* Database:

We are using MongoDB to store and retrieve the SEC filings of various companies extracted from their Form 10-k and related financial documents.

* Collection name used in MongoDB:

sec10k

* Document details:

Under the sec10k collection, multiple documents are stored in JSON format into the MongoDB. The data includes, financial data – consisting of details such as the cost of goods sold (COGS), inventories, total assets, total current liabilities, long term debt, operating income, revenue and gross profit; form 10k html; exhibits; year; and company details. Some of the data is stored in an embedded document structure. This single document structure facilitates the data manipulation, retrieval of the related data to be performed in a single data base fetch operation.

* Pseudo data model:

{

*"company"*:{

*"name"*:*<String>*,

*"secIdentifier"*:*<String>*,

*"address"*:{

*"street1:"*: *<String>*,

*"city"*:*<String>*,

*"state"*:*<String>*,

*"zip"*: *<String>*,

*"BUSINESS PHONE"*:*<String>*

}

},

*"year"*:*<String>*,

*"form10k"*:*<String>*,

*"exhibit"*:*<Array>*,

*"financialData"*:{

*"costofgoodssold"*:<Int32>,

*"inventories"*:<Int32>,

*"totalassets"*:<Int32>,

*"totalcurrentliabilities"*:<Int32>,

*"longtermdebt"*:<Int32>,

*"operationalincome"*:<Int32>,

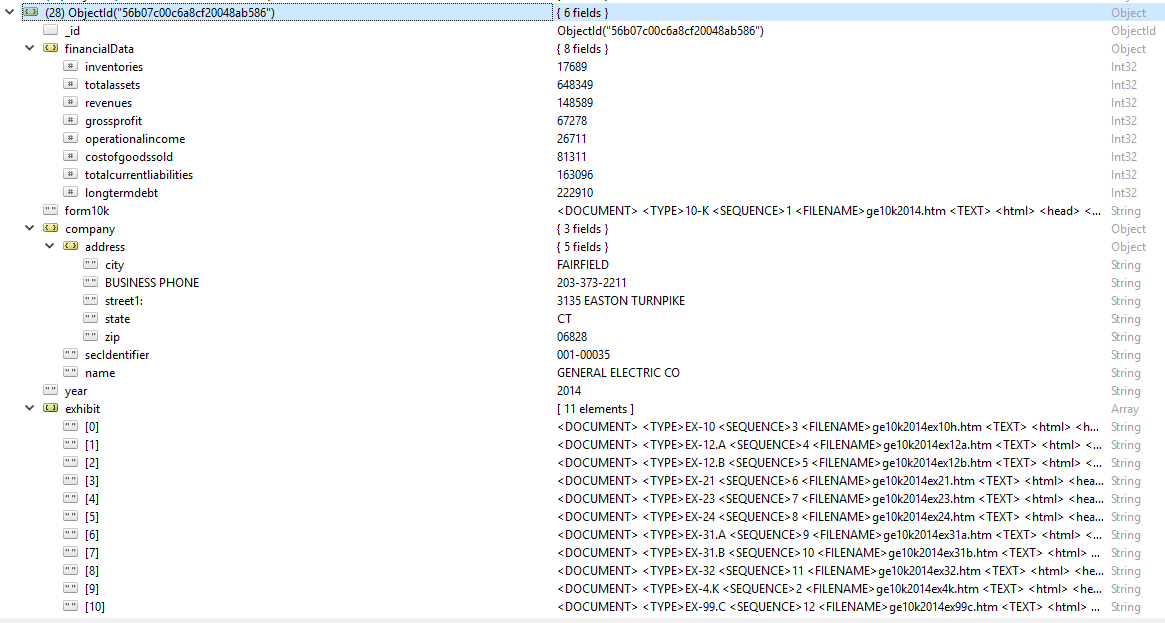
*"revenues"*:<Int32>,

*"grossprofit"*:<Int32>

}

}

* Actual structure of the JSON document:

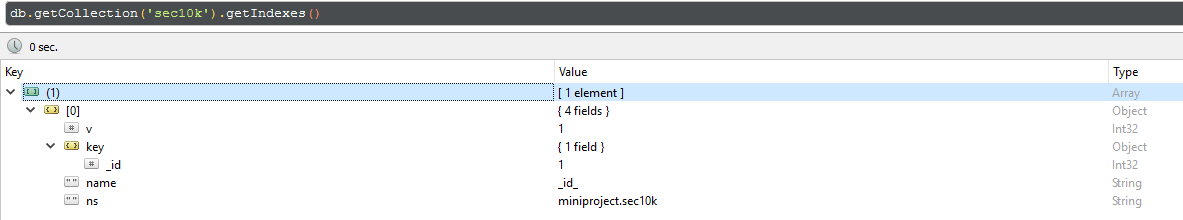


* Constraints:

MongoDB is a NoSQL (non-relational) database that does not have constraints i.e., we don’t have primary key and foreign key relationships as in relational database. MongoDB is considered a “schema less” database, so the only constraints could be on the application side, and not on the database side. The only constraint we are faced with is the limitation of the document size. The JSON document should not be greater than 16 MB size.

* Indexes:

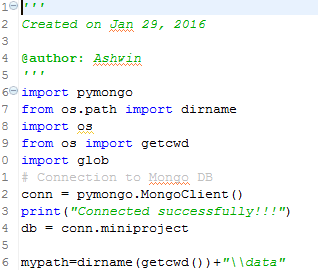
Since the size of this collection (sec10k) is not large indexing will not significantly improve the performance of searching. But as the collection grows in size we can make use of indexing. For now a default index is on the auto-generated id as shown below.



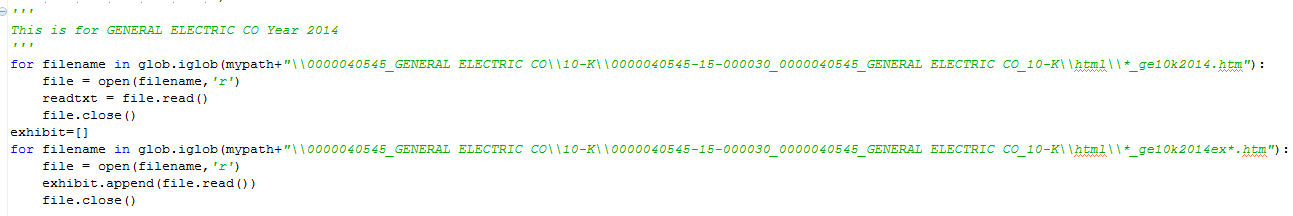
**Part B (Data Migration)**

* Step 1:

Create a connection to the MongoDB using pymongo library.



* Step 2:

Read the html files using python script and store the form 10k file and its exhibits in strings. Below is the code snippet. 

* The extracted data is then inserted into MongoDB by using the pseudo structure described in part (a). Once we have the JSON insert the data using insert function as shown below.



An alternate way of performing the same task would be using mongoimport. It is faster and much efficient compared to inserting the data using pymongo module. However, we are not using it because of the following reasons:

1. The amount of data we are migrating to the database is not significant in size.
2. By using pymongo module migrating the data is a one step process. If mongoimport is to be used, all the components must be synchronized. i.e., the program should not pull the data unless it is exported to the database using mongoimport utility.

* Challenges faced:

The structure of the form 10k could have been broken down into sub-parts for easier access to the information required, but we decided not to split the form 10k data. Our alternate choice was to use BeautifulSoup to parse the html file and convert it into multiple JSON files. The challenges we faced with this alternative was that, the main part of the form 10k apart from the text was the financial data. And each company used different ways to represent the same data in different formats. So there was no fixed template that we could come up with to extract the required output.

* Financial data:

The financial data could not be extracted because of the reason stated above. So we manually entered the parameters we were going to work with into the JSON structure.

**Part C (Company Health)**

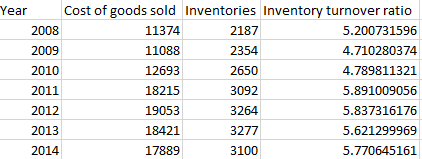
**Inventory turnover ratio:** The ratio indicates the pace of the inventory movement, i.e., how fast the inventory moves.

Rationale:

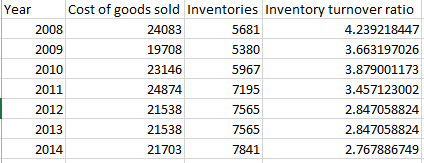
Based on the inventory turnover rate the ability of the company to make more revenue in the given period can be identified (a company with higher inventory rate means the company has short operating cycle and makes more revenues in the given period of time).

**Inventory turnover ratio = Cost of goods sold/Average Inventory**

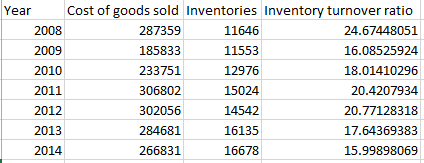
**COCA-COLA**



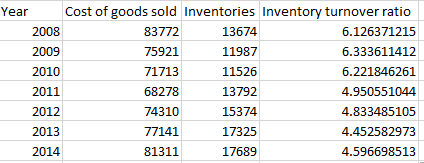
**DUPONT**

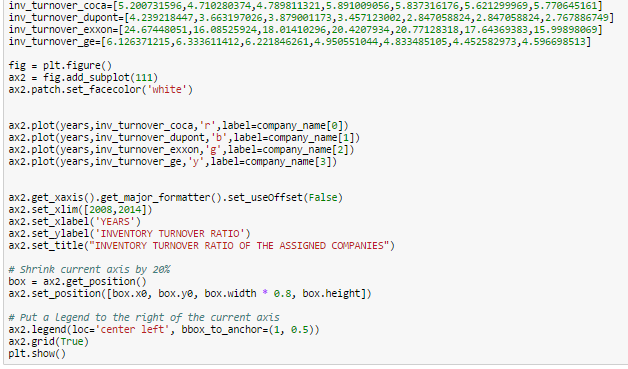


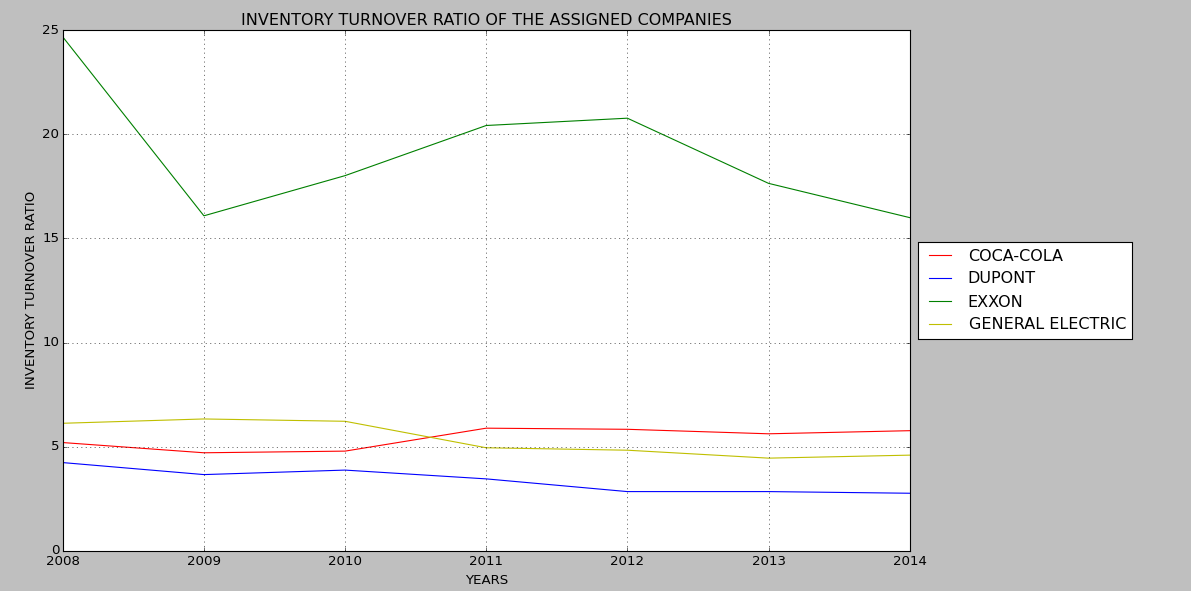
**EXXON MOBL**



**GENERAL ELECTRICS**



Python program to produce graph for inventory turnover ratio:



**Operating Margin:** The ratio indicates the impact of a company’s pricing strategy and operating efficiency.

Rationale: Operating margin can be used to identify the company’s cash position whether strong or weak. i.e., if revenue goes up but margin declines it means the financial position of the company is getting weaker. A healthy operating margin differs for companies in different industries. An increasing operating margin indicates that the company is able to control cost and increase profit as it grows.

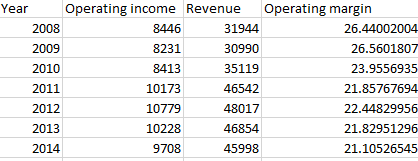
**Operating Margin = Operating income/Revenue**

(Comparison of operating margin for the current and previous year can be done to determine

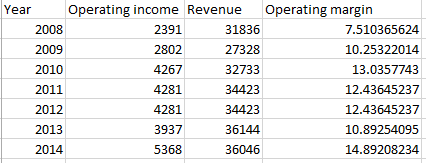
Whether there has occurred a growth or decline and also the cash position depending on the

Value of Revenue Vs Operating margin.)

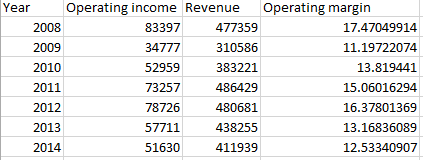
**COCA-COLA**



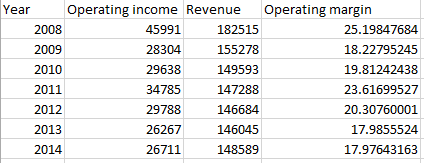
**DUPONT**

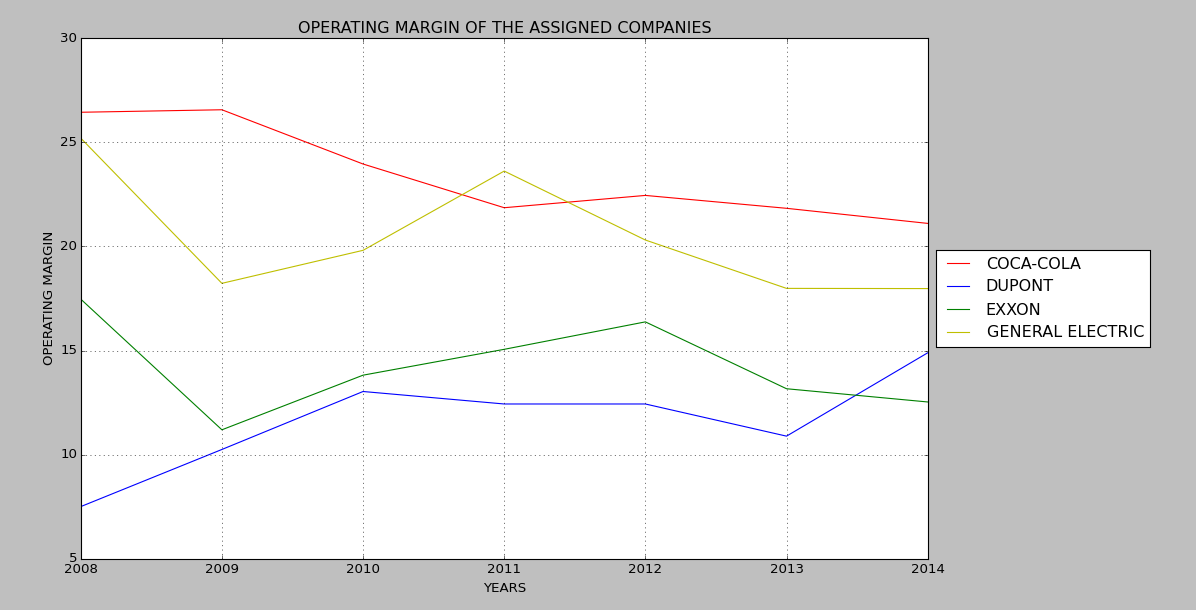


**EXXON MOBL**

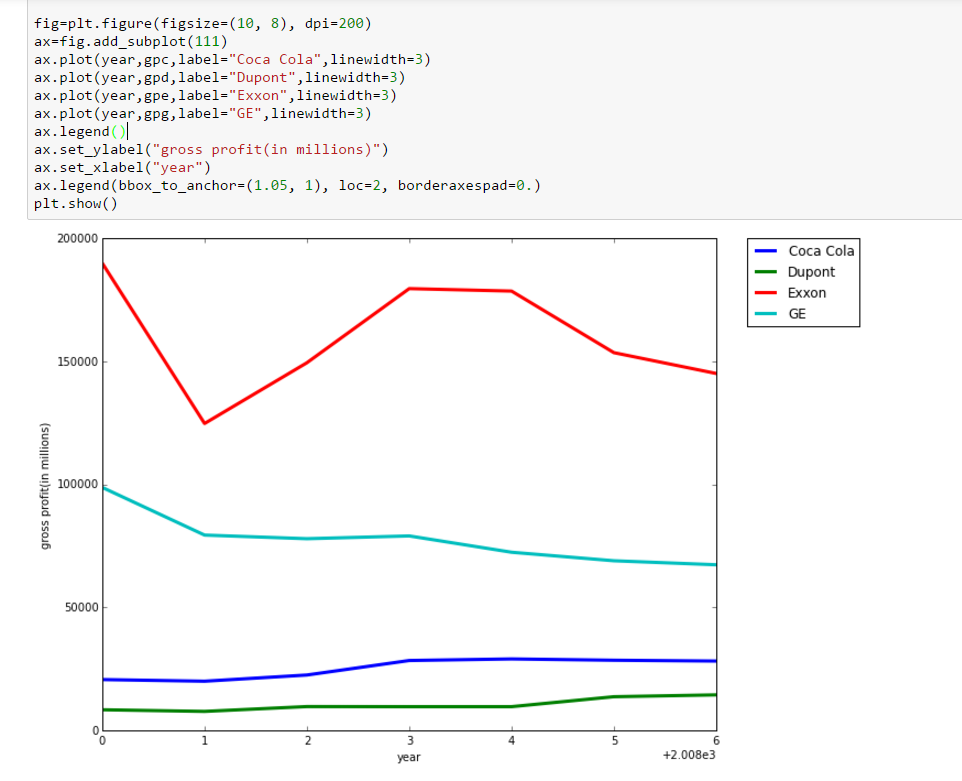


**GENERAL ELECTRICS**

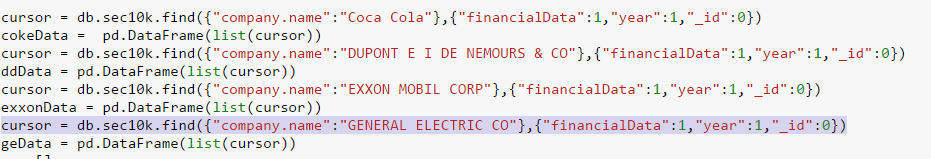


Python program to Generate graphs for the company metric:

**Gross Profit**:



Query used:



Summary:

As it can be seen from the graph that the operating margin of Coca-Cola Company has been in a down trend over the period of seven years (2008-2014) which implies that the company’s cash position is not in a strong phase.

The DuPont company operating margin has been subject to fluctuations in the taken period of time but has never cut its starting low which was on 2008.The company’s cash position is growing gradually which is a good sign of the company health.

For the Exxon Mobil the operating margin which was on a growing phase between 2009-2012 has started losing its grip on revenue growth as well as the operating margin.

General electric company facing a down trend in the operating margin between 2011 and 2013 indicates a weak cash position but it has managed not to drop below the previous low, i.e., there is no significant growth in the company but it has managed to maintain the current margin.

From the last plot, we can see the variation in the gross profit from each company. In 2009 most of the companies had a low profit probably because of recession. But you can observe that the business boomed from there on.

**Part D**

Our initial approach was to use tf\_idf to determine the frequency of occurrence of the word “acquisition”. Based on that we would use n gram algorithm to determine if there is a company name adjacent to this word. But materializing this concept is not an easy task and fetching the numerical values would be a different task altogether.

To get over this problem, one approach could be using clustering (SVC) of the sentences and using google news to identify the name of the companies and the dollar amounts. Or simply do text mining on the sentences within the clustered group to identify the solution.