My Data Science Journey

Undergraduate School

My Journey has been exciting and started in the year 2009 when I was introduced to SQL in my undergraduate Computer Science Engineering. At that time, books were the most recommended source and w3 schools.

SQL Resource:

- https://www.w3schools.com/sql/default.asp
- Textbook: Learning SQL: Master SQL Fundamentals

On the job

I worked on SQL for various projects. I even worked on CouchDB which was used to store user information. I have extensively used Excel to perform analysis. I have even used Excel for validating and cleaning the data. Java was the programming language I used for almost all my projects. Following are good resources to learn Java.

Java Resource

- w3cschool
- tutorialspoint
- Oracle
- learnjavaonline

Graduate School

At Indiana University I started with Python, and R. That was the best choice because python is used extensively. The libraries mostly used were numpy, pandas, matplotlib. Then I took Statistical Analysis in Effective Decision Making which was good and had referred Practice problems given by Prof. SQL/NoSQL course introduced me to MongoDB, MySQL, Neo4j, Cassandra, HBase, and Redis. Social Media Mining was the course in which I was introduced to Jupyter Notebook which made the tasks easier to manage and it could be even used as a report so no need of taking screenshots and saving the images. Used Machine learning algorithms SVC, Naïve Bayes, Logistic Regression, RandomForest Classifier, and DecisionTree Classifier for predictive analysis.

DB Resource:

- MongoDB (https://www.mongodb.com/)
- Neo4j (https://neo4j.com/)
- Cassandra (http://cassandra.apache.org/)
- HBase (https://hbase.apache.org/)
- Redis (https://redis.io/)

Then I took the Search course which was good for understanding the mechanism of the most important and up-to-date retrieval theories and model as well as to design and implement search engines using retrieval models.

Management of Big data course introduced HDFS (Hadoop's distributed file system).

Hadoop is used for the storage and processing of big data applications running in clusters of commodity servers. It is written in Java. Hadoop is freely available, robust, fast, and cost-efficient which makes it special.

The 2 main components of Hadoop are:

1] HDFS - Hadoop distributed file system (HDFS) is used to store big data. It is fault-tolerant. It follows the master-slave configuration where the master is the namenode and slaves are the datanodes. Namenode is used to handle the namespace of the file system. Datanodes are used for read-write operations.

2] MapReduce - MapReduce is a programming model used to deal with a large volume of data. MapReduce has two components map function and reduce function.

Map – The map function takes input and splits it into two components key and value and data would be sorted based on key.

Reduce – Reduce function accepts the data based on the key and data would be split into mountable proportion and each specific computer will focus on only specific sub-data that share the same key.

Applications in what it is called the Hadoop ecosystem are YARN, Hive, Pig, HBase and HBase components, HCatalog, Avro, Thrift, Drill, Mahout, Sqoop, Ambari, Zookeeper. YARN (Yet Another Resource Negotiator) is used to manage resources and job scheduling of the tasks. Hive is used for querying, summarizing, and analysis of large datasets.

Apache Spark is an open-source framework for distributed cluster computing. It provides an interface for programming and working with data that is distributed over a cluster of machines. Used data-lake for storage since datasets can be stored in their native formats.

Resources:

Apache Spark:

- http://spark.apache.org/docs/latest/quick-start.html
- https://spark.apache.org/docs/latest/spark-standalone.html
- https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

HDFS

https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html

The data Visualization course introduced Jupyter Lab. There were various core Python data packages used scikit-learn, matplotlib and seaborn. Then even used altair, vega_datasets, bokeh, datashader, holoviews, wordcloud, and spacy. These are good cloud Jupyter notebook options as well like Google colaboratory, Azure notebooks, and CoCalcS which comes handy if there are package support related issues with the system we are using.