**SUBWAY DATA ANALYSIS**

The NYC public transportation system – Metro Transit Authority- provides data for download via csv files. Part of information available are data from the subway turnstiles, containing weekly logs for cumulative entries and exits by turnstile and by subway station provided during a timeframe.

GOAL OF THIS PROJECT: is to explore the relationship between data from NYC subway turnstiles and the city weather.

REQUIREMENT: data from the subway and data from the weather in NYC.

LINK used for getting information: <http://web.mta.info/developers/turnstile.html>

Analysis of the above gathered information:

**Learned methods**:

* Basics of BigData
* HDFS and MapReduce – Hadoop Distributed File System and MapReduce
* MapReduce Design Patterns---1. Filtering Patterns

2. Summarization Patterns—Numerical Summarization

3.Structural Patterns

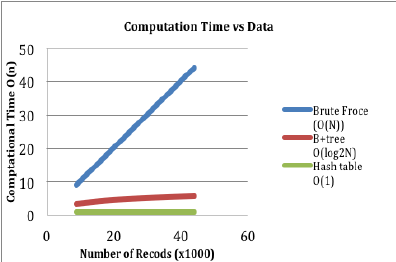
* Deploying Hadoop on Amazon

1. Nodes- NameNode
2. DataNode
3. On demand Hadoop Clusters

**BigData**- Data that is too big to be processed by a single machine.

Data is increasing by the time, numerous organizations generate several bytes of data.

The following data-time graph shows the rate of increase of data production,



Source of the graph: <https://www.researchgate.net/figure/Computational-time-vs-Data-graph-for-range-query_fig2_277262930>

How can we explain BigData?

3V’s define the data size, namely Volume, Variety, and Velocity

Volume: Refers to the size of the data we are dealing with.

Variety: Refers to the fact of the data that it is coming from different sources in different formats.

Velocity: Refers to the speed at which it is generated, speed required to be made available for its processing.

This much amount of data is needed to be stored somewhere, this can be easily stored at SAN (Storage Area Networks), it assures the reliability of data stored but processing and accessing the data from SAN is difficult.

Hence, some amount of data is needed to be discarded. But none of the data can be treated as useless to be discarded hence, some another storing platform is needed to look upon.

Hadoop stores the data in distributed format.

**HDFS:**  Hadoop Distributed File System, each file is divided into smaller chunks called blocks by default of 64 MB. Each one block goes to cluster.

Each cluster have data node and name node. Data node stores the data and some redundant data to recover in case of data node failure.

**MapReduce** is used to process the stored-on Hadoop cluster.

Efficient processing of data is done by mapper and reducer two different groups.

In between mapper and reducer shuffle and sort takes place, sorting refers to the fact that reducer organizes the records in a sorted way and writes the final results.

There can be more than one reducer, each reducer can have any number of keys.