BIG DATA

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

- **Big data** is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.

3 V's of Big Data

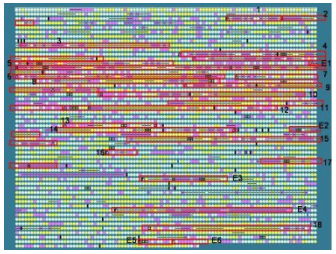
1-Scale (Volume)

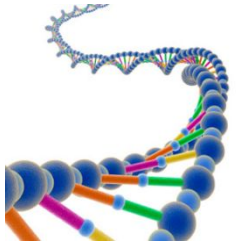
- Volume refers to amount of data ,volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

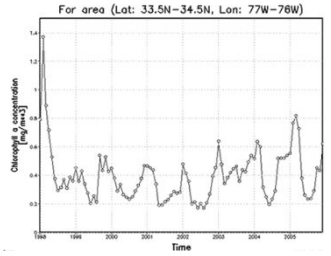
- Data volume is increasing exponentially

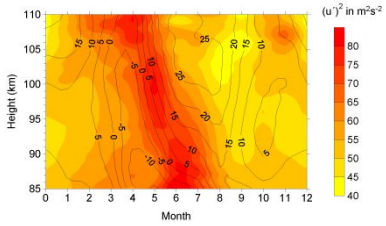
2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









To extract knowledge all these types of data need to linked together

3-Speed (Velocity)

- Data is being generated fast and need to be processed fast.
- The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.
- Online Data Analytics.
- Late decisions missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like send promotions
  - **Healthcare monitoring:** sensors monitoring your activities and body any abnormal measurements require immediate reaction

5 V's of Big data:

- Velocity, Volume, Value, Variety, and Veracity.
- **1)Velocity-**velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every day the number of emails, twitter messages, photos, video clips, etc. increases at lighting speeds around the world. Every second of every day data is increasing. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access to website, credit card verification and instant messaging. Big data technology allows us now to analyze the data while it is being generated, without ever putting it into databases.
- **2)Volume**-Volume refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc. The vast amounts of data have become so large in fact that we can no longer store and analyze data using traditional database technology.

- **3)Value-**When we talk about value, we're referring to the worth of the data being extracted. Having endless amounts of data is one thing, but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data.
- **4)Variety-**Variety is defined as the different types of data we can now use. Today's data is unstructured. In fact, 80% of all the world's data fits into this category, including photos, video sequences, social media updates, etc. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.
- **5)Veracity-**Veracity is the quality or trustworthiness of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content.

Pillars of Big Data

- **Big table:** relational,tabular format
- **Big Text:** all kinds of unstructured data,natural language,semantic data.
- **Big metadata:** data about data,taxonomies, glossary,concepts
- **Big Graphs:** Object connections,semantic discovery,degree of separation etc.

Infrastructure requirements in Big data

- **1)Data Acquisition in Big data**

-data will be in distributed environment, infrastructure must support to carry out high volume of data.

-NOSQL are often used in Big data.

- **2)Data organization in Big data:**

-organizing means data integration

-requires good infrastructure so that processing and manipulating data in the original storage location can be done easily.

-Hadoop-handles large volume of data and keeps data on the original data storage cluster.

-HDFS used to store web logs.

-MapReduce on cluster

- **3)Data analysis in Big data:**

-the infrastructure must be able to integrate analysis on the combination of Big data and traditional enterprise data.

-the infrastructure required for analyzing big data must be able to support deeper analytics such as statistical analysis and data mining on variety of data stored in systems.

Benefits of Big Data Processing

- Ability to process 'Big Data' brings in multiple benefits, such as-

- **Businesses can utilize outside intelligence while taking decisions**

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

- **Improved customer service**

Traditional customer feedback systems are getting replaced by new systems designed with 'Big Data' technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

- **Early identification of risk to the product/services, if any .**

- **Better operational efficiency**

'Big Data' technologies can be used for creating staging area or landing zone for new data before identifying what data should be moved to the data warehouse.

References

- https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity

END