Assignment - 1

Sai Hemanth Nirujogi
G01065588

BIG DATA:

By the year 2020, 1.7mbs of data will be created every second for every person on the earth. 5 billion people are using their mobile devices for communication, social networking and accessing data worldwide every day. 294 billion emails are sent every day. Facebook has over 2 billion monthly active users with 30 billion bits of matter is being shared every month and 136,000 photos are uploaded, 510,000 comments, 293,000 status updates are posted every 60 seconds. 48 hours of videos uploaded to YouTube every minute [1]. Huge stream of structured and unstructured data volumes accessed at a greater momentum every minute. This data is called Big Data.

Big Data is broken into 4 V's

- Volume
- Variety
- Velocity
- Veracity

1. **Volume:** The big data around the world is relying on massive datasets in huge volumes such as zetabytes and petabytes. With the exponential growth of such massive data, new challenges are born. The data must be analyzed to know the trends and to use the data to its fullest.
2. **Variety:** There are new technologies in today's world generating data in different forms. Organizations are determining new ways to integrate the complex data types from the new systems. The data today is not only just rows and columns, there are hundreds of different data types generated.
3. **Velocity:** It is the measure of how fast the information is being generated everyday.
4. **Veracity:** It is considered as the most important V among all the 4Vs. Veracity is not just the quality and accuracy of the data but how reliable the source of data, type of data and processing. [3]

NETFLIX:

Netflix has more than 100 millions users and it is one of the biggest data-driven companies that has mastered big data to its demand and climbed to the top as the world's prime internet television network. Netflix defines big data in terms of data visualization by measuring and leveraging the data to the growth of the company. Using data visualization tools frequently, Netflix found a significant way of representing the data in interactive way to users.

Netflix has become the global leader in delivering internet video at a huge scale by fulfilling their philosophies [4] :

- Accessible data to everyone at any location.
- Making an interactive way of accessing data to users by collecting large amounts of data and analyzing.
- Value of the data ebbs with the longer times took to fetch the data.

By 2016, Netflix has 80+ million members, 125+ millions hours of video streaming per day and more than ⅓ of internet traffic in North America. Data volumes of 600 billion events a day, processing 3 petabytes of data per data and writing 300 terabytes of data per day. By using big data to such an impressive extent Netflix has succeeded in giving the viewers what they want when they wanted it.

IBM:

IBM defines big data as “The ability to achieve superior value from analytics on data at higher volumes, velocities, varieties and veracities”. [4]

[Image on the right is taken from IBM corp, depicting how IBM defines big data].

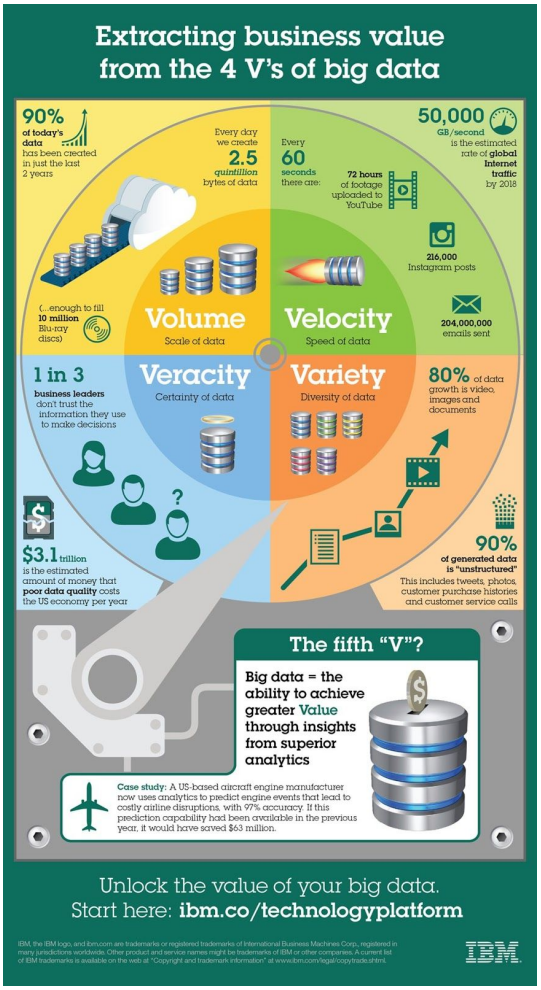

Figure-1: Big Data Definition by IBM [4]

- **Volume** is quantified on scale of data.

2.5 quintillion bytes of data is being created everyday. Volumes of data is erupting and an estimate 90% of data has been created in the last 2 years than in the entire technology era and a massive 40 zettabytes of data will be visualized by the 2020 with a exploding rate of 1.7 megabytes of new information every second.

- **Velocity** is the speeds at which data is accessed and stored.

Data is created every second, for example on Google alone we are performing 40,000 search queries every second making it 3.5 billion searches a day and 1.5 trillion a year. There are many time-sensitive processes in the present world than ever.

- **Variety** is different kinds of data.

Massive growth in media (photos, videos, gifs, etc.,). There are 1.4 billion smartphones used in this world which can create and collect the diverse data. With over 1 Trillion photos taken every year more than billion images are being shared online, surprisingly 80% of them which are taken with a mobile device.

- **Veracity** is certainty and uncertainty in data.

Poor data quality is the main cause for uncertainty in data. Data quality is compromised by the mistakes made in an organisation by an individual, some of them include human errors, Data migration and conversion, mixed entries of data and machine made errors. An estimate of 3.1 trillion dollars a year was costed to the US due to poor quality in data.

PANDORA RADIO:

Pandora is a popular music streaming radio company which uses its industry leading data analysis algorithms to generate personalized music. This big data driven company was first originated from a Music Genome Project in the year 1999, with an idea to create a complex mathematical based system that categorises music using analytical tools and real-time adaption of data.

Pandora believes in processing large amounts of data, distributing the data to each machine and then aggregate small computations to reduce the factors to create an individual stations as an end result tailored with user specific songs. By using pandora's algorithm, interactions like thumbs up/thumbs down can help the behavior of the data.

"Data is the key element in virtually every stage of Pandora. From the playlist personalization engine to the advertisement targeting capability, they have pivoted towards data science" [5].

CITATIONS:

1. Monnappa, Avantika. "How Facebook Is Using Big Data - The Good, the Bad, and the Ugly." *Simplilearn.com*, Simplilearn, 6 July 2018, accessed on September 8h, 2018 www.simplilearn.com/how-facebook-is-using-big-data-article.
2. "What is big data?" by Amazon Web Services. Accessed on Sep 8th, 2018. https://aws.amazon.com/big-data/what-is-big-data
3. McNeill, Cassandra, et al. "Veracity: The Most Important 'V' of Big Data." *GutCheck*, 27 July 2018, accessed on September 8th 2018. www.gutcheckit.com/blog/veracity-big-data-v/.
4. "Big Data Lessons from Netflix" by Wired, Paper Content by Phil Simon. Accessed on Sep 9th, 2018. https://www.wired.com/insights/2014/03/big-data-lessons-netflix
5. "Extracting business value from 4 Vs of Big Data" by IBM hub Accessed on Sep 9th,2018. http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data
6. "How Pandora uses Data to Improve its Service and Music Stations" by Kissmetrics Blog, May 15th, 2017. Accessed on Sep 9th, 2018. https://blog.kissmetrics.com/how-pandora-uses-data