What is Big Data?

"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

That's a formal Definition that you can study anywhere.

is Big Data a platform ?solution? or something Else

We can refer to term that Big Data is a "*problem*".

As the term Big means something that is of anonymously of large size and If anything whether materialistic or not it start causes problem that's why you can refer it as "*Problem*".

Growth of Data

Cisco report says smartphone traffic will exceed PC traffic by 2020

IP traffic will grow in a massive way as 10 billion new devices come online over the next five years

Advancements in the Internet of Things (IoT) are continuing to drive IP traffic and tangible growth in the market

Graphic: *NetworkWorld*, 7 Jun 2016

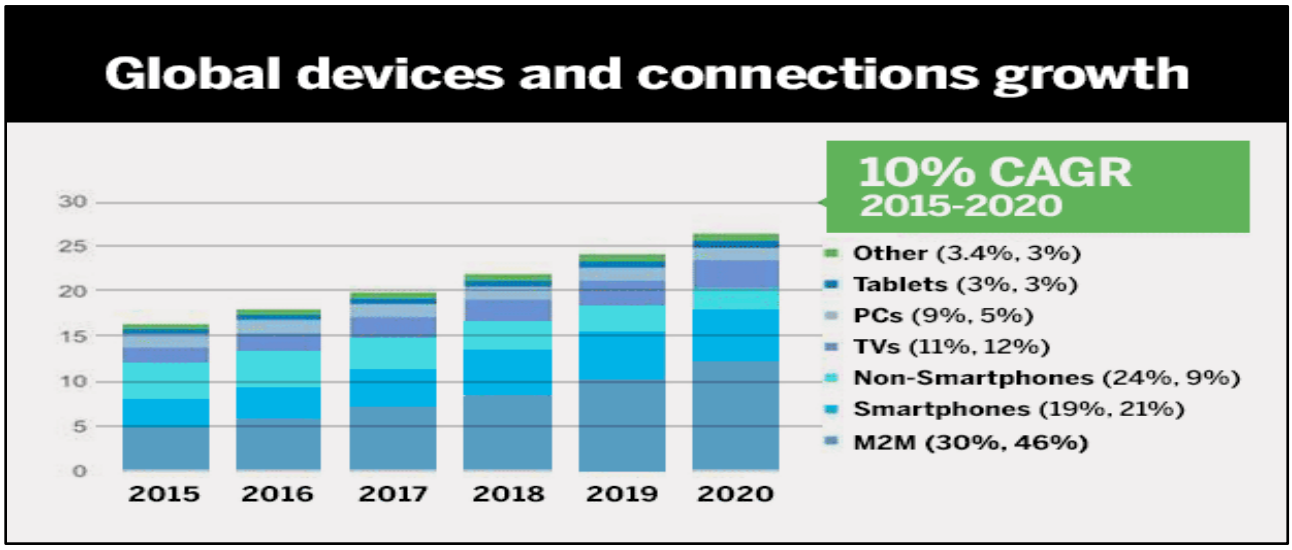
Reference:

- http://www.networkworld.com/article/3080001/lan-wan/cisco-ip-traffic-will-surpass-the-zettabyte-level-in-2016.html

Applications such as video surveillance, smart meters, digital health monitors and a host of other Machine-to-Machine services are creating new network requirements and incremental traffic increases.

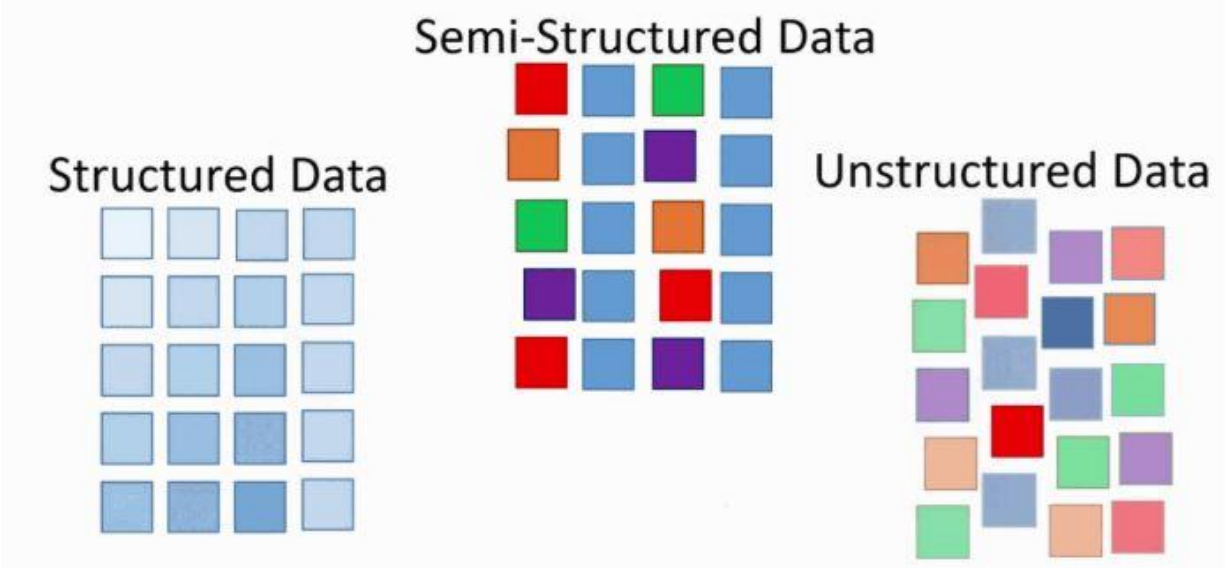
Annual global IP traffic will surpass the zettabyte (ZB; 1000 exabytes) threshold in 2016, and will reach 2.3 ZB by 2020. Global IP traffic will reach 1.1 ZB per year or 88.7 EB (one billion gigabytes [GB]) per month in 2016. By 2020, global IP traffic will reach 2.3 ZB per year, or 194 EB per month.

The number of devices connected to IP networks will be three times as high as the global population in 2020. There will be 3.4 networked devices per capita by 2020, up from 2.2 networked devices per capita in 2015. Accelerated in part by the increase in devices and the capabilities of those devices, IP traffic per capita will reach 25 GB per capita by 2020, up from 10 GB per capita in 2015.



Types of Big Data

- **Structured**
    - Data that can be stored
      and processed in a
      fixed format, aka schema
- **Semi-structured**
    - Data that does not have a formal structure of a data model, i.e. a table definition in a relational DBMS, but nevertheless it has some organizational properties like tags and other markers to separate semantic elements that makes it easier to analyze, aka XML or JSON
- **Unstructured**
    - Data that has an unknown form and cannot be stored in RDBMS and cannot be analyzed unless it is transformed into a structured format is called as unstructured data
    - Text Files and multimedia contents like images, audios, videos are example of unstructured data - unstructured data is growing quicker than others, experts say that 80 percent of the data in an organization is unstructured



Big data stories of big companies

Check what Walmart, Nestlé, PepsiCo, JPMorgan Chase, Rolls-Royce, and Uber have to say about their big data experience.

“Over time, the need for more insights has resulted in over 100 petabytes of analytical data that needs to be cleaned, stored, and served with minimum latency through our Hadoop-based big data platform. Since 2014, we have worked to develop a big data solution that ensures data reliability, scalability, and ease-of-use, and are now focusing on increasing our platform’s speed and efficiency.”

Reza Shiftehfar, Hadoop Platform Team Leader at Uber

“Walmart relies on big data to get a real-time view of the workflow in the pharmacy, distribution centers and throughout our stores and e-commerce.”

Walmart Staff

“[About their big data platform Pep Worx] We were able to launch the product [Quaker Overnight Oats] using very targeted media, all the way through targeted in-store support, to engage those most valuable shoppers and bring the product to life at retail in a unique way. These priority customers drove 80% of the product’s sales growth in the first 12 weeks after launch.”

Jeff Swearingen, Senior Vice President of Marketing at PepsiCo

“Artificial intelligence, big data and machine learning are helping us reduce risk and fraud, upgrade service, improve underwriting and enhance marketing across the firm.”

Jamie Dimon, Chairman and Chief Executive Officer at JPMorgan Chase

"We have huge clusters of high-power computing which are used in the design process. We generate tens of terabytes of data on each simulation of one of our jet engines. We then have to use some pretty sophisticated computer techniques to look into that massive dataset and visualize whether that particular product we've designed is good or bad. Visualizing big data is just as important as the techniques we use for manipulating it."

Paul Stein, Chief Scientific Officer at Rolls-Royce

"The projects we're undertaking using big data aren't one-off experiments. They're truly driving business decisions in finance, human resources, sales, and our supply chain."

Shan Collins, Chief Analytics Officer at Nestlé USA

Big data Usecases

Fraud Prevention

For credit card holders, fraud prevention is one of the most familiar use cases for big data. Even before advanced big data analytics became popular, credit card issuers were using rules-based systems to help them flag potentially fraudulent transactions. So, for example, if a credit card were used to rent a car in Hawaii, but the customer lived in Omaha, a customer service agent might call to confirm that the cardholder was on vacation and that someone hadn't stolen the card.

Thanks to big data analytics and machine learning, today's fraud prevention systems are orders of magnitude better at detecting criminal activity and preventing false positives. In the example already mentioned, for instance, a sophisticated fraud prevention system might be able to see that the customer had recently purchased airline tickets, sunscreen and a new swimsuit before the rental car purchase. Based on historical patterns, a predictive analytics or machine learning system would be able to tell that the rental car was thus less likely to be a fraudulent purchase.

But fraud prevention systems can get even more sophisticated than that. According to Experian, fraud tends to be concentrated in certain geographic regions—often near airports, which make it easy for criminals to move stolen goods. However, which zip codes are riskiest tends to change over time. Big data analytics can look at past records of fraudulent transaction and quickly identify changing trends. Credit card companies and retailers can then pay more attention to transactions in zip codes that are emerging as hotbeds for criminal activity.

Credit card issuers are understandably hesitant about disclosing all the advanced analytic techniques that they use to detect and prevent fraud. However, many credit card firms and other consultants offer technology, advice and services to other firms to help them set up systems to stop criminal transactions.

Price Optimization

Both business-to-consumer (B2C) and business-to-business (B2B) enterprises are also using big data analytics to optimize the prices that they charge their customers. For any company, the goal is to set prices so that they maximize their income. If the price is too high, they will sell fewer products, decreasing their net returns. But if the price is too low, they may leave money on the table.

Big data analytics allows companies to see which price points have yielded the best overall results under various historic market conditions. Businesses that are more sophisticated with their pricing analytics may also employ variable or dynamic pricing strategies. They use their big data solutions to segment their customer base and build models that show how much different types of customers will be willing to pay under different circumstances. B2C companies that have attempted this approach have met with mixed results, but it is fairly standard among B2B companies.

Recommendation Engines

Speaking of popularity, one of the most familiar use cases for big data is the recommendation engine. When you are watching a movie at Netflix or shopping for products from Amazon, you probably now take it for granted that the website will suggest similar items that you might enjoy. Of course, the ability to offer those recommendations arises from the use of big data analytics to analyze historical data.

These recommendation engines have become so commonplace on the Web that many customers now expect them when they are shopping online. And organizations that haven't taken advantage of their big data in this way may lose customers to competitors or may lose out on upsell or cross-sell opportunities.

Internet of Things

And enterprises in every industry are beginning to see the possibilities of the Internet of Things (IoT). As in the preventive maintenance example, they are using sensors to collect data that they can then analyze to achieve actionable insights. They might track customer or product movement, monitor the weather or keep an eye on security camera footage.

As with big data itself, the number of ways in which analytics can be applied to IoT solutions seems to be endless.

How Big Data Helps Bar Owners Sell More Beer

the FMCG industry can leverage the enormous potential of Big Data analytics. With each product, massive amounts of data are generated ranging from data in the production process to consumer generated data. One of those products is beer. It may be obvious that large multinationals use Big Data analytics to discover new insights, such as Heineken for example. Heineken is using Big Data in several way, ranging from knowing where in a Walmart store a sixpack was picked-up, to smart social media campaign and a smart beer bottle that dances on the rhythm of the music at parties. But Big Data has also become available for the smaller companies and even for the local bar owner.

The local bar owner can now start using data generated by serving beer and this data reveals great insights and drives profit for the bar. Sensor in the brewing equipment generates all sorts of data and this data is

transmitted via WiFi to a computer that analyses all the data in real-time. The data is then visualized on an App that can be used by the bar owner. This helps to cut waste and boost profit.

The algorithm analyses which beer should be on a discount and at what period, resulting in an increased profit of up to 80%. The product is developed by an Israeli startup called Weissbeerger. They collect vast amounts of beer-related data that helps the bar owner sell more beer and making more profit. Insights derived from all that beer data includes for example understanding which brands are more popular on what dates and what the timings are that the most beer is consumed on any given day. It also revealed that for example larger beers are drunk more often in the beginning of the evening and strong ales typically later at night.

Such insights help the bar owner to better anticipate on what his/her guests are looking for, resulting in happier customers and more profit. This video made by Bloomberg, reveals how English pub owners benefit from these insights.

Link for Video: *https://youtu.be/-W3lSOdn5hw*

Introduction to cloudera:

(Reference to official website of cloudera)

The world's first enterprise data cloud:

Finally, a platform for both IT and the business, Cloudera Data Platform (CDP) is:

- On-premises and public cloud
- Multi-cloud and multi-function
- Simple to use and secure by design
- Manual and automated
- Open and extensible
- For data engineers and data scientists

Apache Hadoop Ecosystem:

Hadoop is an ecosystem of open source components that fundamentally changes the way enterprises store, process, and analyze data. Unlike traditional systems, Hadoop enables multiple types of analytic workloads to run on the same data, at the same time, at massive scale on industry-standard hardware. CDH, Cloudera's open source platform, is the most popular distribution of Hadoop and related projects in the world (with support available via a Cloudera Enterprise subscription).

Link for Cloudera Documentation:

https://docs.cloudera.com/documentation/enterprise/5-14-x/topics/cloudera_quickstart_vm.html

Link for Cloudera Download:

https://www.cloudera.com/downloads/cdh.html