Big Data Concepts and Terminology

Agenda

- ***Big Data concept***
  - ***Is Big Data a Volume or a Technology?***
  - ***Examples***
  - **Why Are Big Data Systems Different**
- **Characteristics**
  - **Volume**
  - **Velocity**
  - **Variety**
  - **VARIABILITY**
  - **VERACITY**
  - **VISUALISATION**
  - **VALUE**
- **Clustered Computing**
- **Ingesting Data into the System**
- **Persisting the Data in Storage**
- **Computing and Analyzing Data**

Big Data

- ***Big Data*** is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

Is Big Data a Volume or a Technology?

- While the term may seem to reference the volume of data, that isn't always the case. The term big data, especially when used by vendors, may refer to the technology (which includes the tools and processes), that an organization requires to handle the large amounts of data and storage facilities. The term is believed to have originated with web search companies who needed to query very large distributed aggregations of loosely-structured data.

An Example

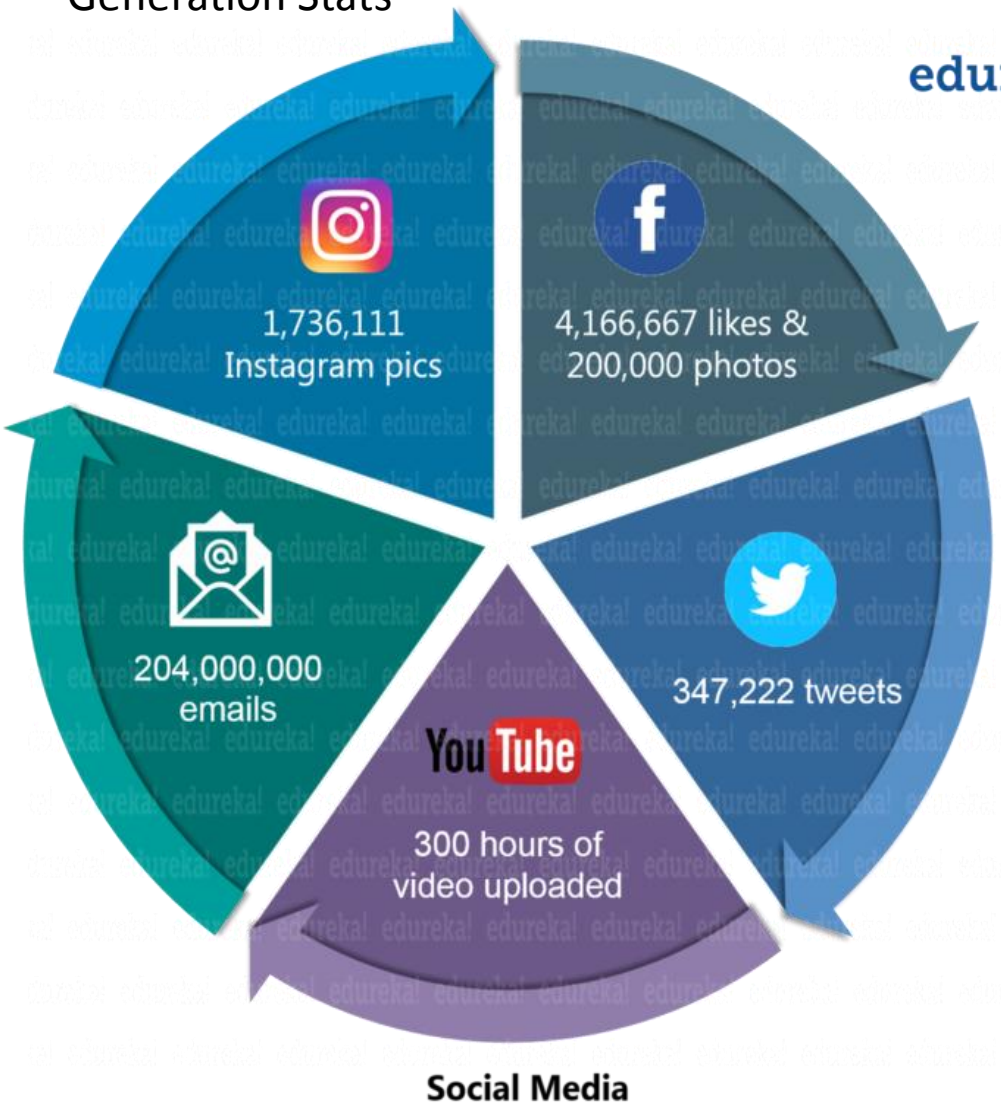
- An example of big data might be petabytes (1,024 terabytes) orexabytes(1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible.

For example

- Earlier we had landline phones, but now we have shifted to smartphones. Similarly, how many of you remember floppy drives that were extensively used back in 90's? These Floppy drives have been replaced by hard disks because these floppy drives had very low storage capacity and transfer speed. Thus, this makes floppy drives insufficient for handling the amount of data with which we are dealing today. In fact, now we can store terabytes of data on the cloud without being bothered about size constraints.
- Now, let us talk about various drivers that contribute to the generation of data.
- Have you heard about IoT? IoT connects your physical device to the internet and makes it smarter. Nowadays, we have smart air conditioners, televisions etc. Your smart air conditioner constantly monitors your room temperature along with the outside temperature and accordingly decides what should be the temperature of the room. Now, in order to do this, it first collects the data of the temperature outside the room from the internet. It continuously stores the data received from its sensors. Finally, with the help of these two data, it infers the required change in room temperature. Now imagine how much data would be generated in a year by smart air conditioner installed in tens & thousands of houses. By this you can understand how IoT is contributing a major share to Big Data.

For example counti…

Fig: Hadoop Tutorial – Social Media Data Generation Stats



- Now, let us talk about the largest contributor to the Big Data which is, nothing but, social media. Social media is actually one of the most important factors in the evolution of Big Data as it provides information about the people's behavior. You can look at the figure below and get an idea how much data is getting generated every minute:
- Apart from the rate at which the data is getting generated, the second factor is the lack of proper format or structure in these data sets that makes processing a challenge.

Why Are Big Data Systems Different?

- The basic requirements for working with big data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions.
- The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

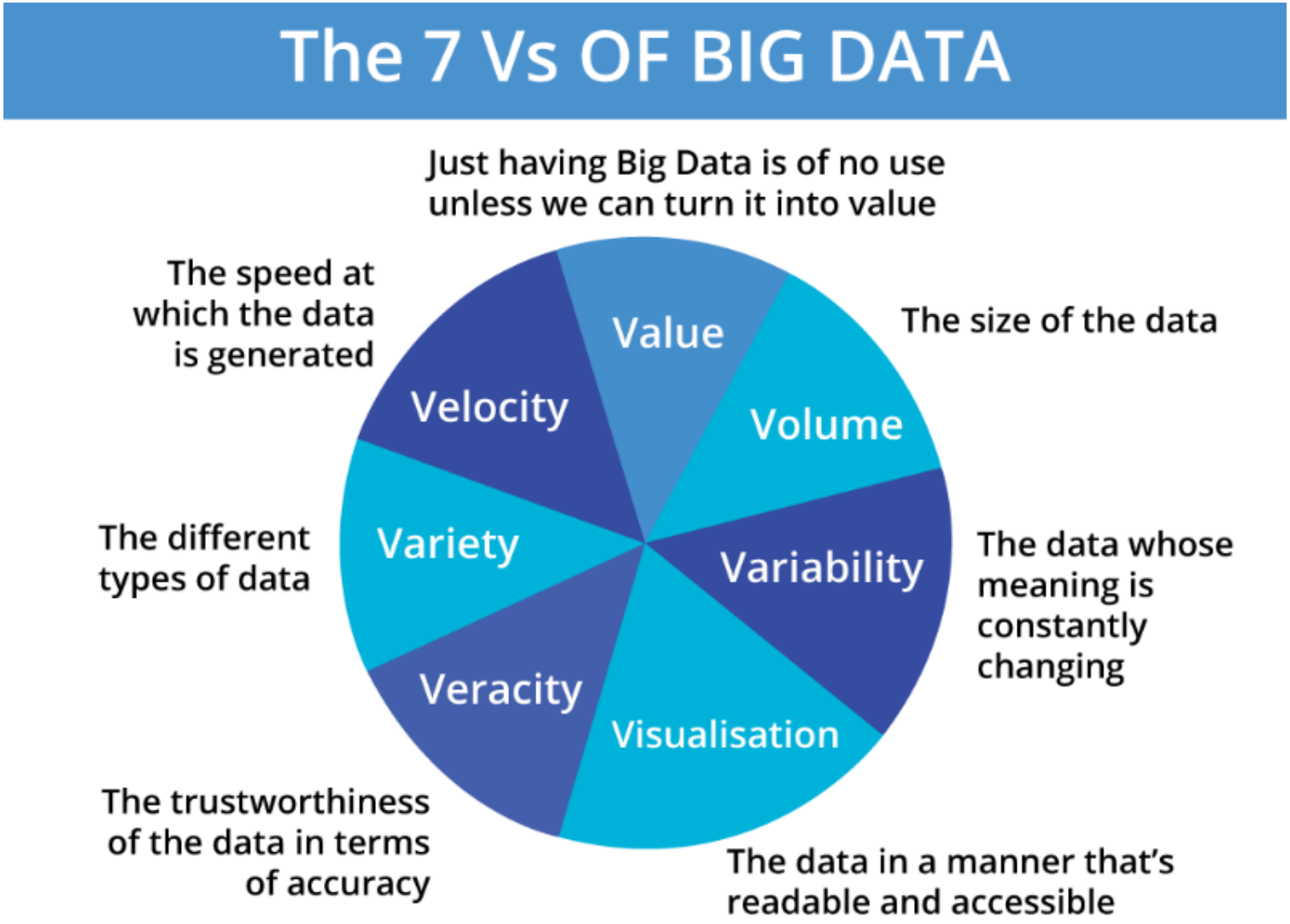
How do you define big data?

- In 2001, Gartner's Doug Laney first presented what became known as the “7 Vs of big data" to describe some of the characteristics that make big data different from other data processing
- How do you define big data? The seven V’s sum it up pretty well –
  Volume, **V**elocity, **V**ariety, **V**ariability, **V**eracity, **V**isualization, and **V**alue.

Big data Defines three D2D's

- Data-to-Decisions
- Data-to-Discovery
- Data-to-Dollars

Big Data Technology with 8V´s



7 V's- Volume

- Volume is how much data we have – what used to be measured in Gigabytes is now measured in Zettabytes (ZB) or even Yottabytes (YB)
- The IoT (Internet of Things) is creating exponential growth in data.
- **Example**. 100 terabytes of data are uploaded daily to Facebook, Akamai analyses 75 million events a day to target online ads, Walmart handles 1 million customer transactions every single hour. 90% of all data ever created was generated in the past 2 years.

- Scale is certainly a part of what makes Big Data big. The internet-mobile revolution, bringing with it a torrent of social media updates, sensor data from devices and an explosion of e-commerce, means that every industry is swamped with data- which can be incredibly valuable, if you know how to use it.

7 V's-Velocity

- Velocity is the speed in which data is accessible. I remember the days of nightly batches, now if it's not real-time it's usually not fast enough.
- Example In 1999, Wal-Mart's data warehouse stored 1,000 terabytes (1,000,000 gigabytes) of data. In 2012, it had access to over 2.5 petabytes (2,500,000 gigabytes) of data.
  Every minute of every day, we upload 100 hours of video on Youtube, send over 200 million emails and send 300,000 tweets.
- 'Velocity' refers to the increasing speed at which this data is created, and the increasing speed at which the data can be processed, stored and analyzed by relational databases.
- The possibilities of processing data in real-time is an area of particular interest, which allows companies to do things like display personalized ads on the web pages you visit, based on your recent search, viewing and purchase history.
- Google alone processes on average more than "40,000 search queries every second," which roughly translates to more than 3.5 billion searches per day.

7 V's- Variety

•Variety describes one of the biggest challenges of big data. It can be unstructured and it can include so many different types of data from XML to video to SMS. Organizing the data in a meaningful way is no simple task, especially when the data itself changes rapidly.

Gone are the days when a company's data could be neatly slotted into a table and analyzed. 90% of data generated is 'unstructured', coming in all shapes and forms- from geo-spatial data, to tweets which can be analyzed for content and sentiment, to visual data such as photos and videos.

The '3 V's' certainly give us an insight into the almost unenvisionable scale of data, and the break-neck speeds at which these vast datasets grow and multiply. But only 'Variety' really begins to scratch the surface of the depth- and crucially, the challenges- of Big Data.

•When it comes to big data, we don't only have to handle structured data but also semi structured and mostly unstructured data as well. As you can deduce from the Velocity's examples, most big data seems to be unstructured, but besides audio, image, video files, social media updates, and other text formats there are also log files, click data, machine and sensor data, etc.

7 V's- Variability

- Variability is different from variety. A coffee shop may offer 6 different blends of coffee, but if you get the same blend every day and it tastes different every day, that is variability. The same is true of data, if the meaning is constantly changing it can have a huge impact on your data homogenization.
- Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
- Brian Hopkins (a principal analyst at Forrester) cited the supercomputer Watson as a prime example of this. To participate in the game show Jeopardy, Watson had to "dissect an answer into its meaning and to figure out what the right question was". Words don't have static definitions, and their meaning can vary wildly in context.
  Say a company was trying to gauge sentiment towards a cafe using these 'tweets':

7 V's- Variability .. Cont

- Example
- Say a company was trying to gauge sentiment towards a cafe using these 'tweets'
  - *"Delicious muesli from the @imaginarycafe- what a great way to start the day!"*
    "Greatly disappointed that my local Imaginary Cafe have stopped stocking BLTs."
    "Had to wait in line for 45 minutes at the Imaginary Cafe today. Great, well there's my lunchbreak gone…"
- Evidently, "great" on its own is not a sufficient signifier of positive sentiment. Instead, companies have to develop sophisticated programs which can 'understand' context and decode the precise meaning of words through it. Although challenging, it's not impossible; Bloomberg, for instance, launched a program that gauged social media buzz about companies for Wall Street last year.

7 V's- Veracity

- Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems. The simplest example is contacts that enter your marketing automation system with false names and inaccurate contact information. How many times have you seen Mickey Mouse in your database? It's the classic "garbage in, garbage out" challenge.
- Although there's widespread agreement about the potential value of Big Data, the data is virtually worthless if it's not accurate. This is particularly true in programmes that involve automated decision-making, or feeding the data into an unsupervised machine learning algorithm. The results of such programmes are only as good as the data they're working with.

7 V’s- Veracity Example

- For example, consider a data set of statistics on what people purchase at restaurants and these items' prices over the past five years. You might ask: Who created the source? What methodology did they follow in collecting the data? Were only certain cuisines or certain types of restaurants included? Did the data creators summarize the information? Has the information been edited or modified by anyone else?
- Answers to these questions are necessary to determine the veracity of this information. Knowledge of the data's veracity in turn helps us better understand the risks associated with analysis and business decisions based on this particular data set.

7 V's- Visualization

- Visualization is critical in today's world. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.
- Once it's been processed, you need a way of presenting the data in a manner that's readable and accessible- this is where visualization comes in. Visualizations can contain dozens of variables and parameters- a far cry from the x and y variables of your standard bar chart- and finding a way to present this information that makes the findings clear is one of the challenges of Big Data.
- Current big data visualization tools face technical challenges due to limitations of in-memory technology and poor scalability, functionality, and response time. You can't rely on traditional graphs when trying to plot a billion data points, so you need different ways of representing data such as data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees.
- Combine this with the multitude of variables resulting from big data's variety and velocity and the complex relationships between them, and you can see that developing a meaningful visualization is not easy.

7 V's- Value

- Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization – which takes a lot of time, effort and resources – you want to be sure your organization is getting value from the data.
- The potential value of Big Data is huge. Speaking about new Big Data initiatives in the US healthcare system last year, "McKinsey" estimated if these initiatives were rolled out system-wide, they "could account for $300 billion to $450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US health-care costs". However, the cost of poor data is also huge- it's estimated to cost US businesses $3.1 trillion a year. In essence, data on its own is virtually worthless. The value lies in rigorous analysis of accurate data, and the information and insights this provides.
- So what does all of this tell us about the nature of Big Data? Well, it's massive and rapidly-expanding, but it's also noisy, messy, constantly-changing, in hundreds of formats and virtually worthless without analysis and visualization.
- In essence, when the media talk about Big Data, they're not just talking about vast amounts of data that are potential treasure troves of information. They're also talking about the business of analyzing this data- the way we pick the lock to the treasure trove. In the world of Big Data, data and analysis are totally interdependent- one without the other is virtually useless, but the power of them combined is virtually limitless.

Clustered Computing

Because of the grades of big data, individual experts are often inadequate for handling the data at most levels. To good addresses the high storage and computational needs of big data, computer agglomerations are a good fit.
Big data clustering program combines the resources of many small gadgets, seeking to give a number of advantages:
Resource sharing : Combining the accessible storage space to hold data is a clear merit, but CPU and memory sharing is also extremely all-important. Processing large datasets requires large amounts of all three of these resources.
High convenience: agglomerations can give varying stages of fault allowance and convenience guarantees to prevent hardware or program failures from affecting accesses to data and processing. This becomes increasingly all-important as we continue to emphasize the value of real-time analytics.
Simple Scalability: agglomerations make it uncomplicated to scale horizontally by increasing more gadgets to the team. This means the system can react to actions in resource requirements without diversifying the animal resources on a device.

Clustered Computing Continu…..

- Using agglomerations requires a success for overseeing agglomeration body, coordinating resource overlapping , and planning effective work on individual nodes. agglomeration body and resource distribution can be handled by app like Hadoop's YARN (which stands for Yet Another Resource communicator) or Apache Mesos.
- The assembled computing agglomeration often acts as a foundation which other app interfaces with to process the data. The gagdets involved in the computing agglomeration are also typically involved with the management of a distributed storage system.

Ingesting Data into the System

- Data ingestion is the process of taking raw data and increasing it to the system. The standard of this operation depends heavily on the format and standard of the data sources and how far the data is from the desired attribute prior to processing.
- One route that data can be increased to a big data system are dedicated ingestion equipment. Technologies like Apache Sqoop can take existing data from relative databases and increase it to a big data system. Similarly, Apache Flume and Apache Chukwa are projects designed to collective and import application and server logs. Queuing systems like Apache Kafka can also be used as an interface between different data generators and a big data system. Ingestion frameworks like Gobblin can aid to collective and normalize the production of these equipment at the end of the ingestion pipeline.
- During the ingestion process, some stage of analysis, sorting, and labeling usually takes place. This process is sometimes labeled ETL, which stands for extract, transform, and load. While this term conventionally refers to gift data warehousing processes, some of the same concepts registry to data entering the big data system. Typical operations might include altering the incoming data to format it, reasoning and labeling data, separating out unneeded or evil data, or potentially validating that it adheres to definite requirements.
- With those aptitudes in mind, ideally, the caught data should be kept as raw as viable for large trait further on down the pipeline.

Persisting the Data in Storage

- The ingestion processes typically hand the data off to the elements that oversee storage, so that it can be reliably continued to disk. While this seems like it would be an uncomplicated operation, the volume of incoming data, the requirements for convenience, and the distributed computing place make more complex storage systems necessary.
- This usually means supplementing a distributed register system for raw data storage. successes like Apache Hadoop's HDFS filesystem allow large quantities of data to be written across aggregate nodes in the agglomeration. This ensures that the data can be accessed by reason resources, can be loaded into the agglomeration's thrust for in-memory operations, and can gracefully handle element failures. Other distributed filesystems can be used in place of HDFS including Ceph and GlusterFS.
- Data can also be imported into other distributed systems for more structured accesses. Distributed databases, especially NoSQL databases, are well-suited for this role because they are often designed with the same fault tolerant considerations and can handle heterogeneous data. There are many dissimilar symbols of distributed databases to select from being on how you want to organize and present the data. To learn more about some of the actions and what purpose they best serve, read our NoSQL comparison lead.

Computing and Analyzing Data

- Once the data is accessible, the system can commence processing the data to surface effective information. The mathematic place is perhaps the most dissimilar part of the system as the requirements and best come can vary significantly being on what symbol of perceptions desired.
- Data is often processed repeatedly, either iteratively by a single equipment or by using a number of equipment to surface dissimilar symbols of perceptions.
- Batch processing is one mode of computing over a large dataset. The process involves breaking work up into small pieces, planning each piece on an individual appliance, reshuffling the data based on the intermediate results, and then reasoning and assembling the closing result. These stages are often referred to individually as splitting, mapping, walking , reducing, and assembling, or collectively as a distributed map reduce algorithm.
- This is the strategy used by Apache Hadoop's MapReduce. Batch processing is most helpful when dealing with very large datasets that demand quite a bit of mathematic.

Computing and Analyzing Data --- Counti…

- Batch processing is a good fit for definite symbols of data and math, other workloads demand more real-time processing. Real-time processing requires that information be processed and made prepared immediately and requires the system to react as brand-new information becomes accessible. One route of earning this is stream processing, which operates on a continuous stream of data composed of individual symbols. Another communal characteristic of real-time processors is in-memory computing, which works with representations of the data in the cluster's memory to evade having to write back to disk.
- Apache assault, Apache Flink, and Apache Spark give non-identical ways of earning real-time or near real-time processing.
- There are trade-offs with each of these technologies, which can affect which come is best for any individual difficulty. In general, real-time processing is best suited for analyzing small agglomerations of data that are changing or being increased to the system rapidly.

- https://www.webopedia.com/TERM/B/big_data.html
- https://www.youtube.com/watch?v=lleDElPTMvg
- https://www.youtube.com/watch?v=ziqx2hJY8Hg