daaUseful links:

[CitizenChoice](https://citizenchoice.in/course/big-data/Chapter%201/4-drivers-for-big-data)

UNIT-1

Introduction Hadoop: Big Data – Apache Hadoop & Hadoop Eco System – Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.

->**What is Big Data?**

* Big Data literally means large amount of data
* Big Data is the pillar behind the idea that one can make useful inference with a large body of data that wasn’t possible before with smaller datasets.
* Extremely large datasets may be analyzed computationally to reveal patterns, trends, and associations that are not transparent or easy to identify.

->Big Data is Everywhere!

Enormous amounts of data are being collected and stored

* Web data, E-commerce
* Point-of-sale at stores
* Bank Transactions
* Social Network
* Real-time plant data
* Materials Formulation & Design
* Molecular Simulations

How much data is Big Data?

* Google processes 20 Peta Bytes (PB) per day (2008)
* Facebook has 2.5 PB of user data + 15 Tera Bytes (TB)/day (2009)
* eBay has 6.5PB of user data + 50 TB/day (2009)
* CERN’s Large Hydron Collider (LHC) generates 15PB a year.

[The Four V's of Big Data - OpenSistemas](https://opensistemas.com/en/the-four-vs-of-big-data/#:~:text=IBM%20data%20scientists%20break%20it,%2C%20variety%2C%20velocity%20and%20veracity.)

## What is big data?

Having a **big data strategy** in place means: capturing, storing, twisting, polishing, presenting and apprehending large amounts of information unique to each organization, in order to make business decisions.

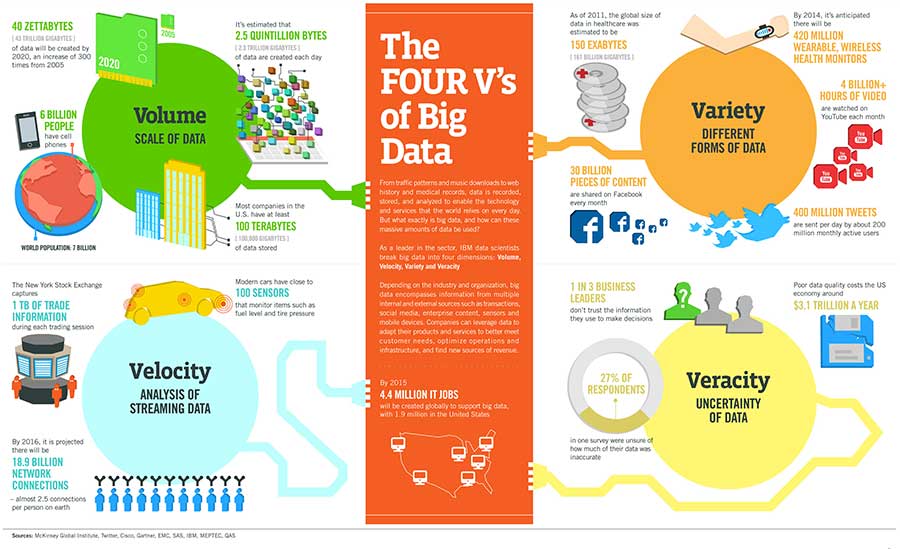
## The 4 V’s of Big Data in infographics

The term is not new.**For many people this term is directly associated with “a lot of data”**. Understanding this technology in this way, however, is not entirely accurate. Big Data technology implies:

* **Compilation.**
* **Storage.**
* **Exploitation.**

…of a large volume of data. However, this **does not necessarily mean that we are talking about “Big Data”.**

IBM data scientists break it into four dimensions: **volume, variety, velocity and veracity**. This infographic explains and gives examples of each.

Find the original infographic [here](https://www.ibmbigdatahub.com/infographic/four-vs-big-data).

## The 4 V’s of Big Data (or) Characteristics of big data

It can be said that the Big Data environment has to have these four basic characteristics:

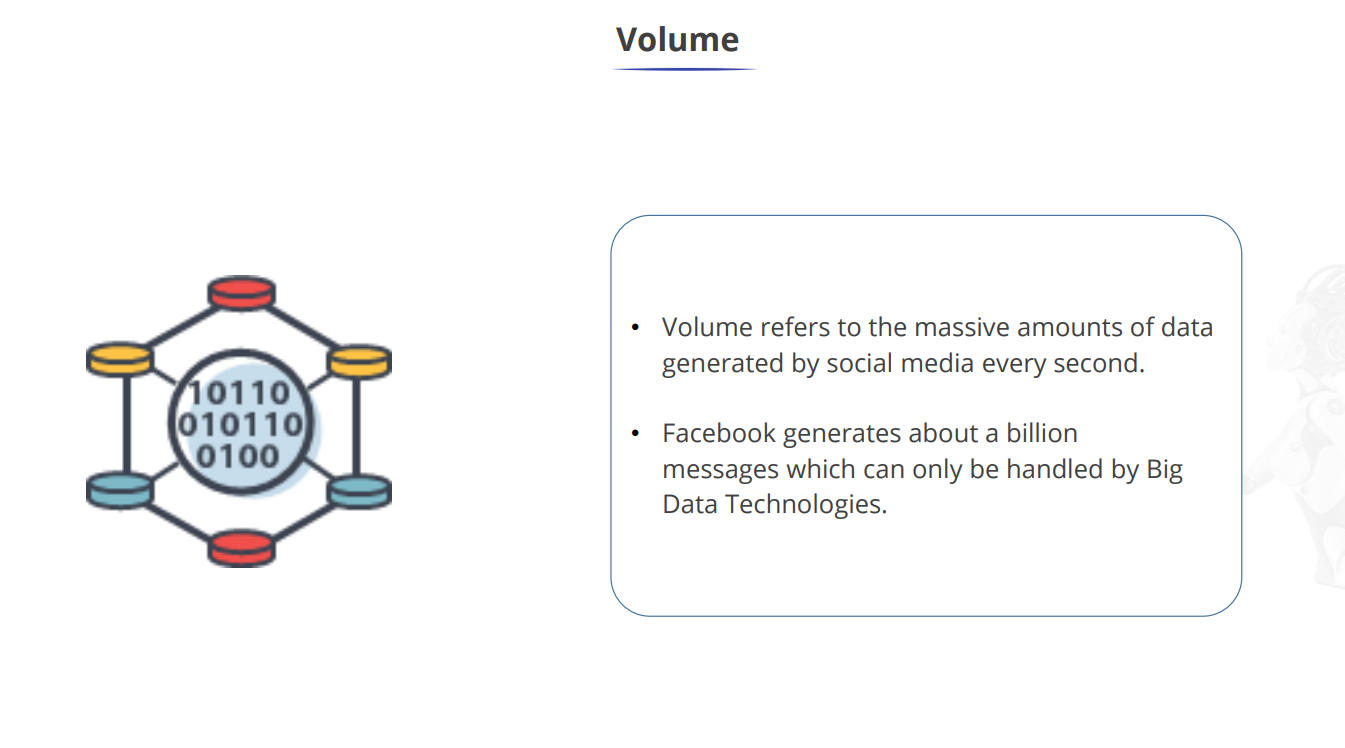
****

Volume

You may have heard on more than**one occasion that Big Data is nothing more than business intelligence,** but in a very large format. More data, however, is not a synonym of it.

Obviously, the Big Data, needs a certain amount of data, but **having a huge amount of data, does not necessarily mean that you are working on this field of data.**

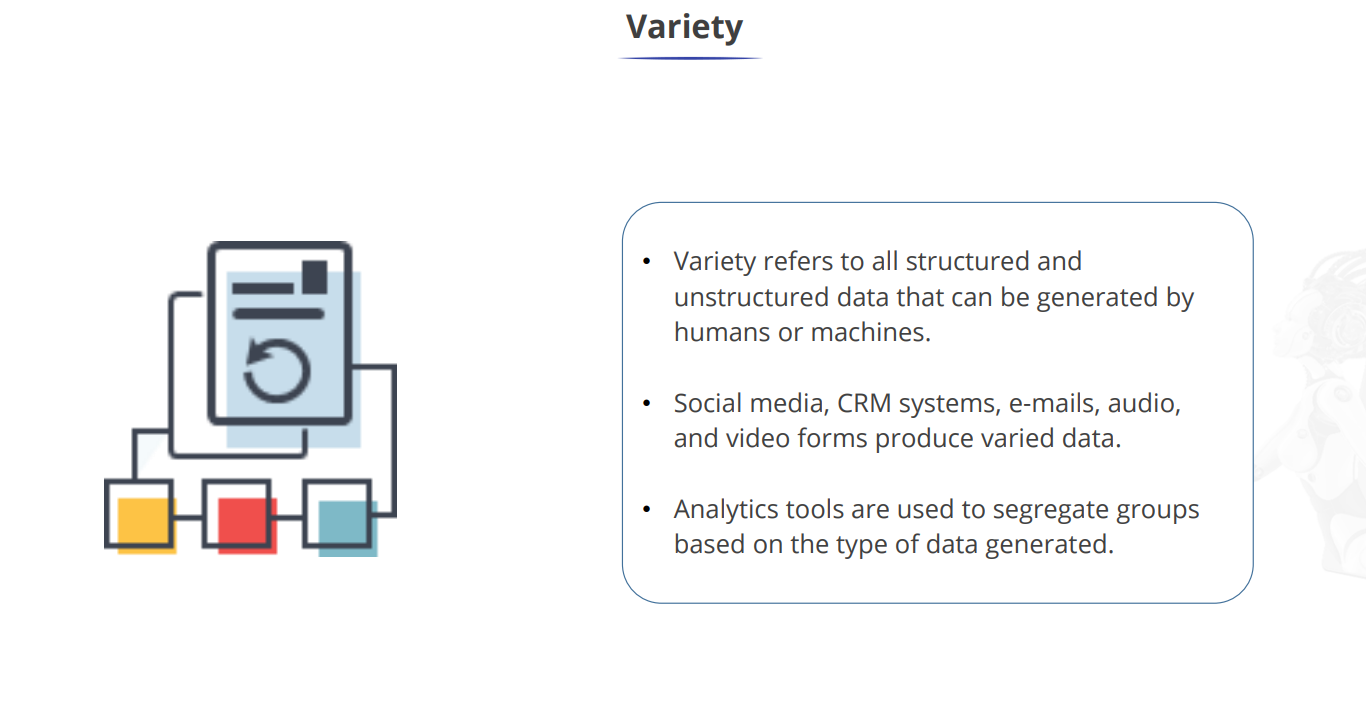
It would also be a mistake to think that all areas of Big Data are business intelligence**since is not limited or defined by the objectives sought with that initiative.**But it will be by the characteristics of the data itself.

****

### Variety

Today, we can base our decisions on the **prescriptive data obtained through this method.** Thanks to this technology, every action of customers, competitors, suppliers, etc, will generate prescriptive information that will range from structured and easily managed data to unstructured information that is difficult to use for decision making.

Each piece of data, or core information, **will require specific treatment.** In addition, each type of data will require specific storage needs (the storage of an e-mail will be much less than that of a video).

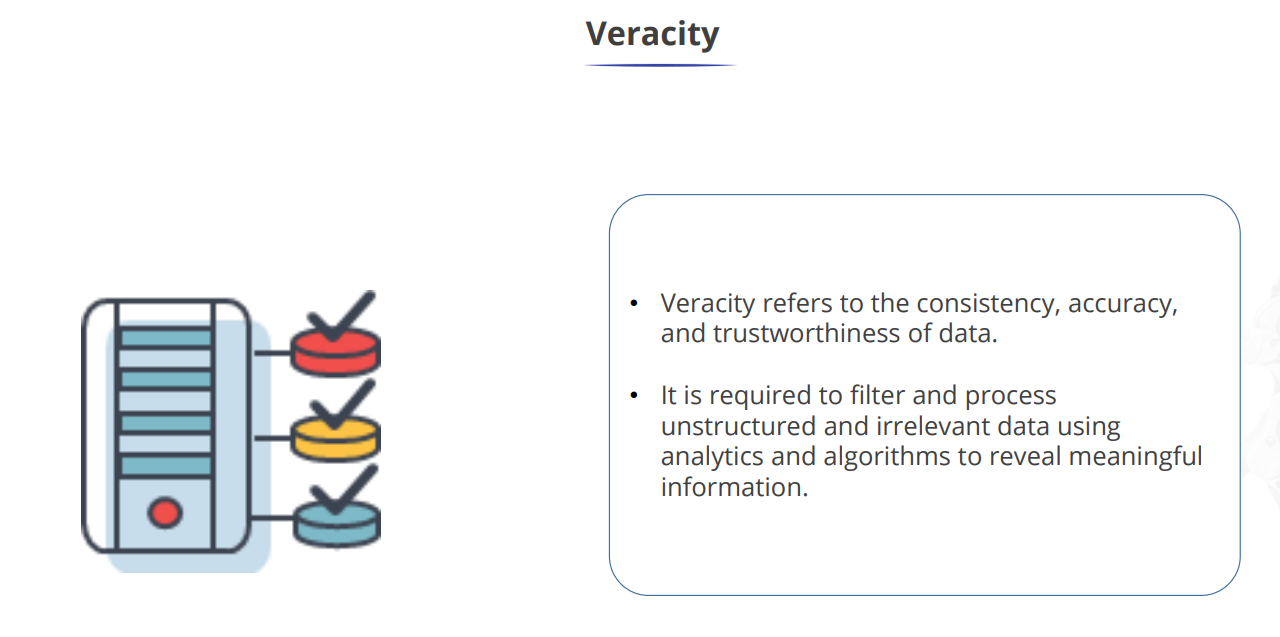
****

### Veracity

This V will refer to both data quality and availability. When it comes to traditional business analytics, the source of the data is going to be much smaller in both quantity and variety.

However, **the organization will have more control over them, and their veracity will be greater.**

When we talk about the Big D, variety is going to mean**greater uncertainty**about the quality of that data and its availability. It will also have implications in terms of the data sources we may have.

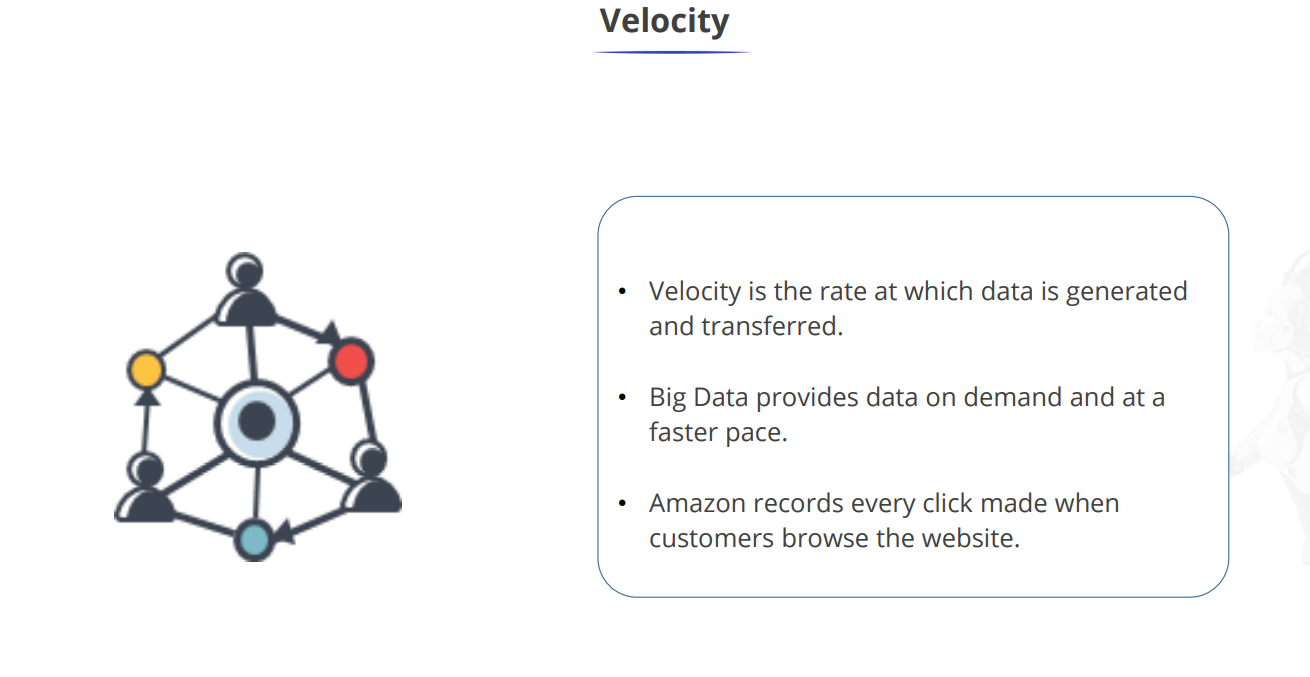
****

### Velocity

It is very possible that Variety and Veracity would not be so relevant and would not be under so much pressure when facing a Big Data initiative if it were not for the **high volume of information that has to be handled and, above all, for the velocity at which the information has to be generated and managed.**

The data will be an input for the technology area (it will be essential to be able to store and digest large amounts of information). And **the output part will be the decisions and reactions that will later involve** the corresponding departments.

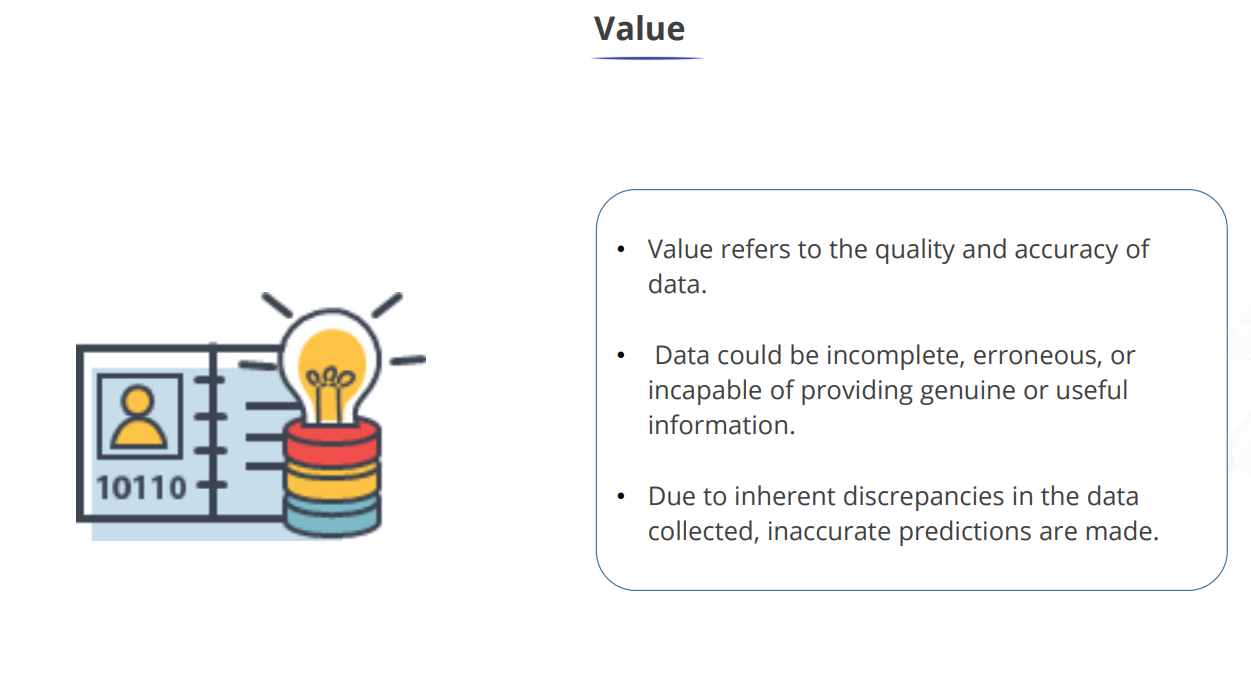
The important thing here is that they are able to react with the necessary speed to boost the business area.

****

To the traditional 4 V’s of Big Data we can add another 3 V’s that are also fundamental in big data technology:

## The value of data

The importance and the possible benefit that can be obtained from the data is one of the fundamental factors of this technology.  
The value that can be obtained from the data depends on many factors, among others the use that is made of it or the combination that is made between the data itself to get the most out of them.

****

## Data visualization

Data visualization is the way in which data is represented for display and explanation to facilitate understanding and decision making. In addition, data visualization tools make data more accessible and understandable and help to find problems and patterns.

## Data Feasibility

This refers to the ability of companies to make effective use of the data they collect. In other words, there is no point in having a large amount of data if the information they provide is not used efficiently.

[The Four V's of Big Data - OpenSistemas](https://opensistemas.com/en/the-four-vs-of-big-data/#:~:text=IBM%20data%20scientists%20break%20it,%2C%20variety%2C%20velocity%20and%20veracity.)

**What Is Big Data?**

* Big data is a massive amount of datasets that cannot be stored, processed and analyzed using traditional tools.
* It is defined as : “[Big Data](https://www.simplilearn.com/tutorials/big-data-tutorial) are high volume, high velocity, or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.”
* The term ‘big data’ is self-explanatory − a collection of huge data sets that normal computing techniques cannot process. The term not only refers to the data, but also to the various frameworks, tools, and techniques involved. Technological advancement and the advent of new channels of communication (like social networking) and new, stronger devices have presented a challenge to industry players in the sense that they have to find other ways to handle the data.
* Big data is an open-source software framework for storing data and executing applications on commodity hardware clusters.
* Big data refers to large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions

**Characteristics Of Big Data:**

1. Big Data is defined as data that is extremely large. It refers to a large set of datasets that are growing rapidly over time and have an enormous scope.
2. Examples of big data analytics include social media platforms, aircraft engines, and financial markets.
3. Structured, unstructured, or semi-structured big data are the three possible types.
4. A few properties of big data are volume, variety, velocity, and variability.
5. A few benefits of big data include better decision-making, improved customer service, and improved operational efficiency.

The Sources of Big Data

1. Black Box Data

This is the data generated by airplanes, including jets and helicopters. Black box data includes flight crew voices, microphone recordings, and aircraft performance information.

1. Social Media Data

This is data developed by such social media sites as Twitter, Facebook, Instagram, Pinterest, and Google+.

1. Stock Exchange Data

This is data from stock exchanges about the share selling and buying decisions made by customers.

1. Power Grid Data

This is data from power grids. It holds information on particular nodes, such as usage information.

1. Transport Data

This includes possible capacity, vehicle model, availability, and distance covered by a vehicle.

1. Search Engine Data

This is one of the most significant sources of big data. Search engines have vast databases where they get their data.

## Types Of Big Data (With Examples)

### Structured Data

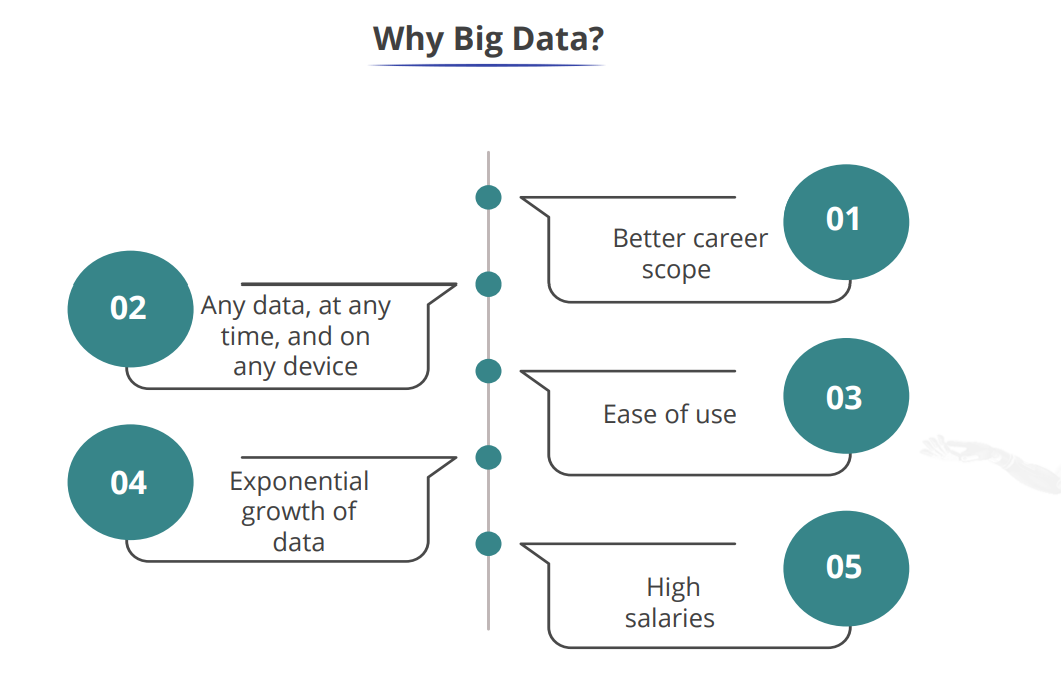
Organised data is easy to evaluate and sort since it has predetermined organisational characteristics and is provided in a structured or tabular schema. Each field is independent and accessible individually or with information from other fields since it is specified. Since of this, structured data is very significant because it enables the rapid data collection from numerous database locations.

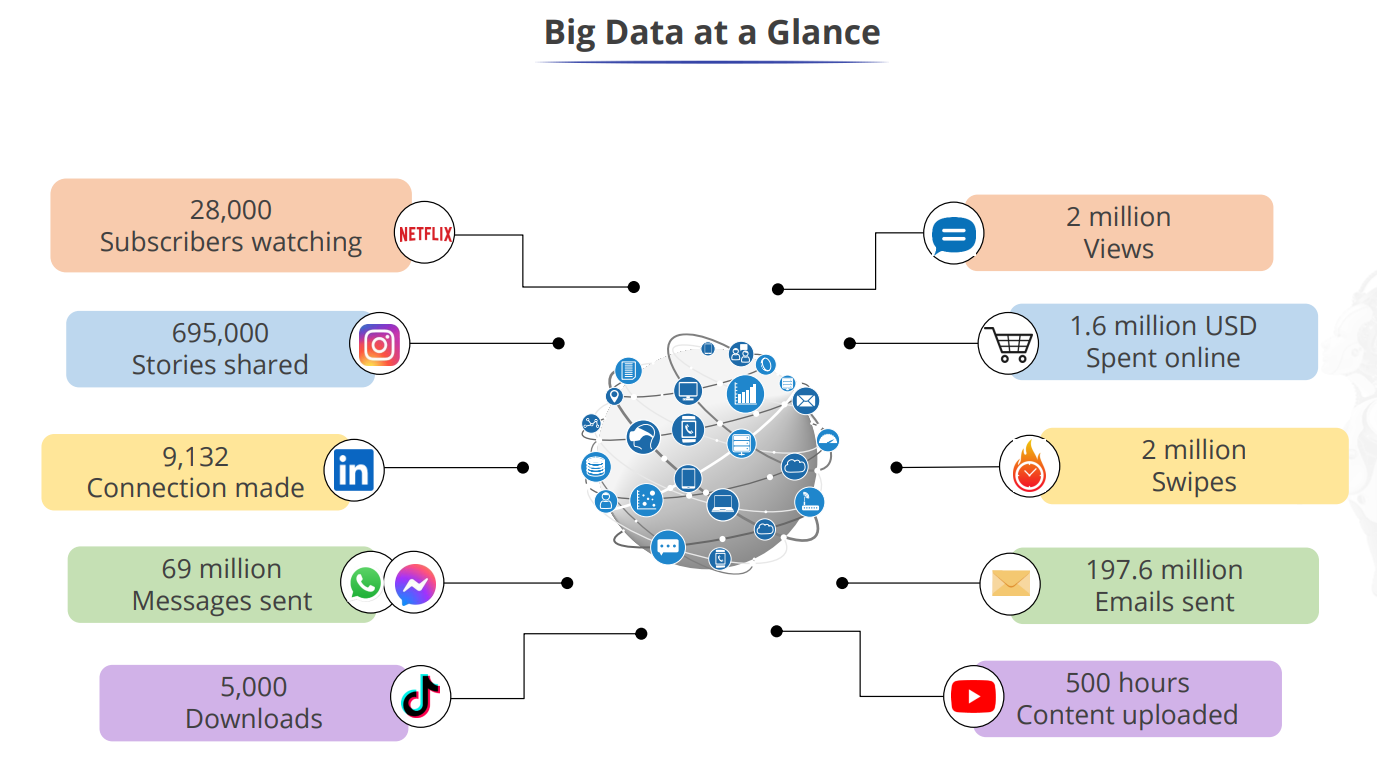
### Unstructured data

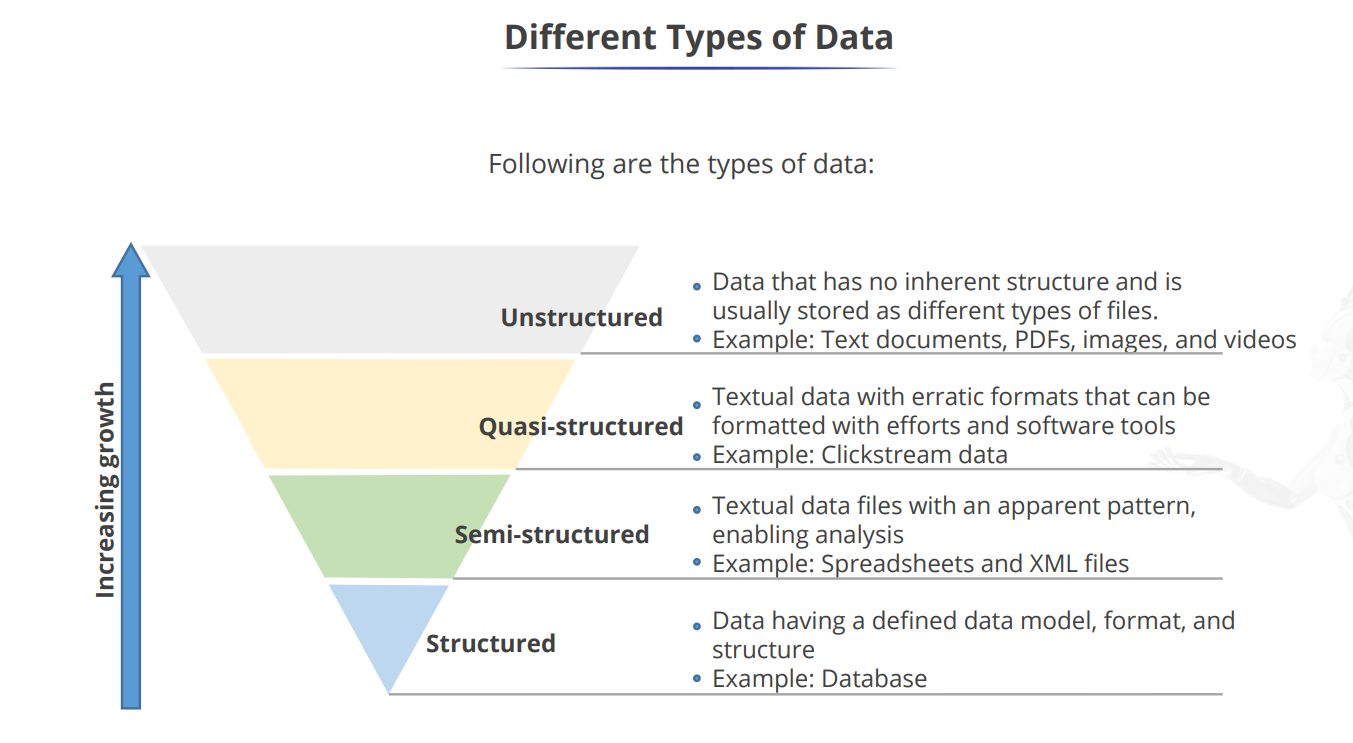
Unstructured data refers to information that lacks predetermined conceptual meanings and is difficult for conventional databases or data models to comprehend or analyse. Most large data is made up of unstructured data, including facts, dates, and numbers. Video and audio files, mobile activities, satellite photos, and other types of big data

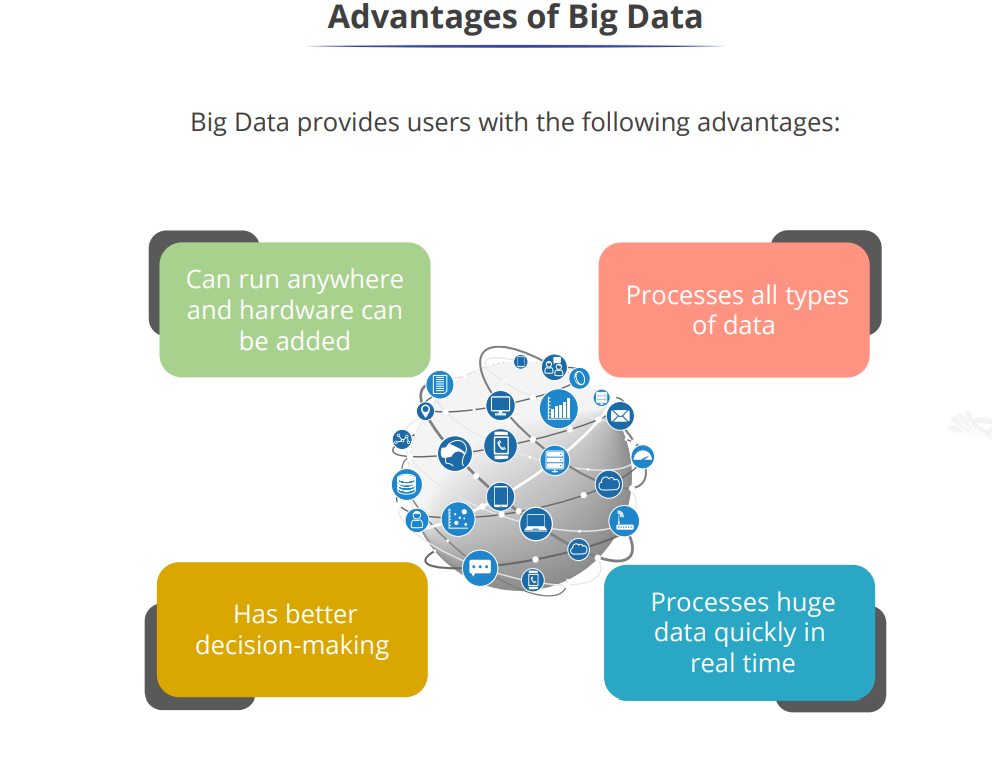
### Semi-Structured Data

A combination of unstructured and structured data is semi-structured data. As a result, it incorporates some characteristics of structured data, but also lacks a clear organisation and does not adhere to relational databases' or data models' official formats. For instance, semi-structured data is frequently seen in JSON and XML.

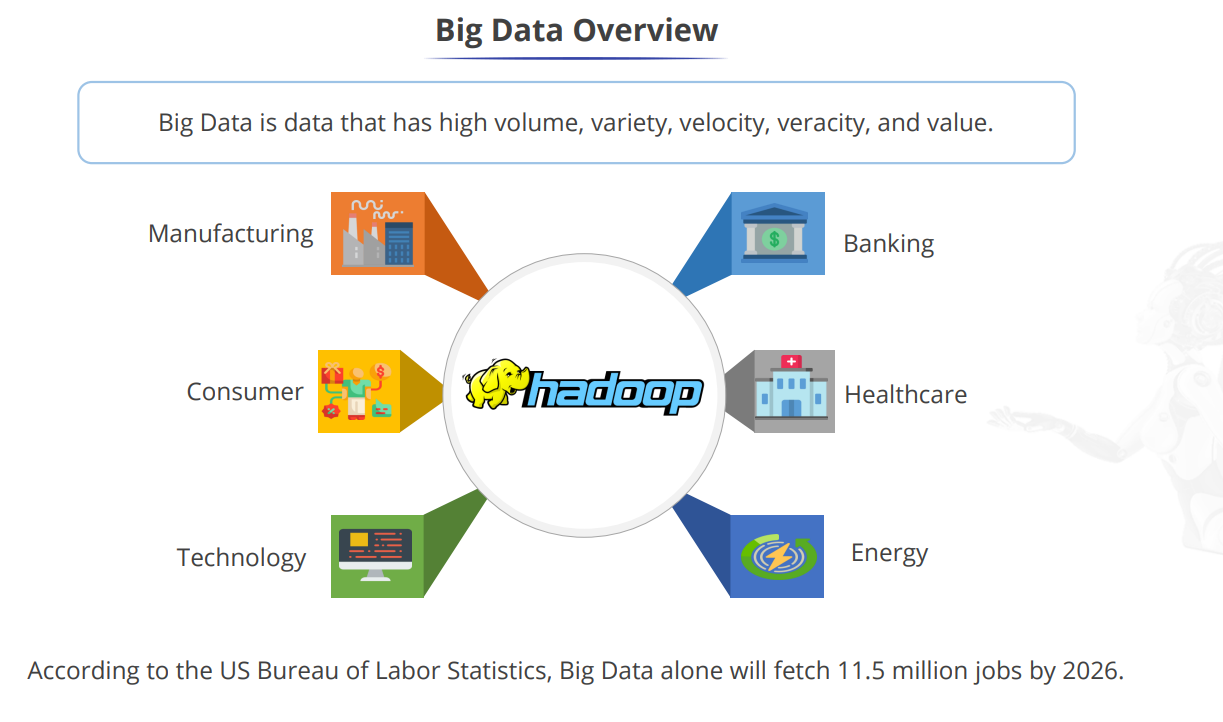




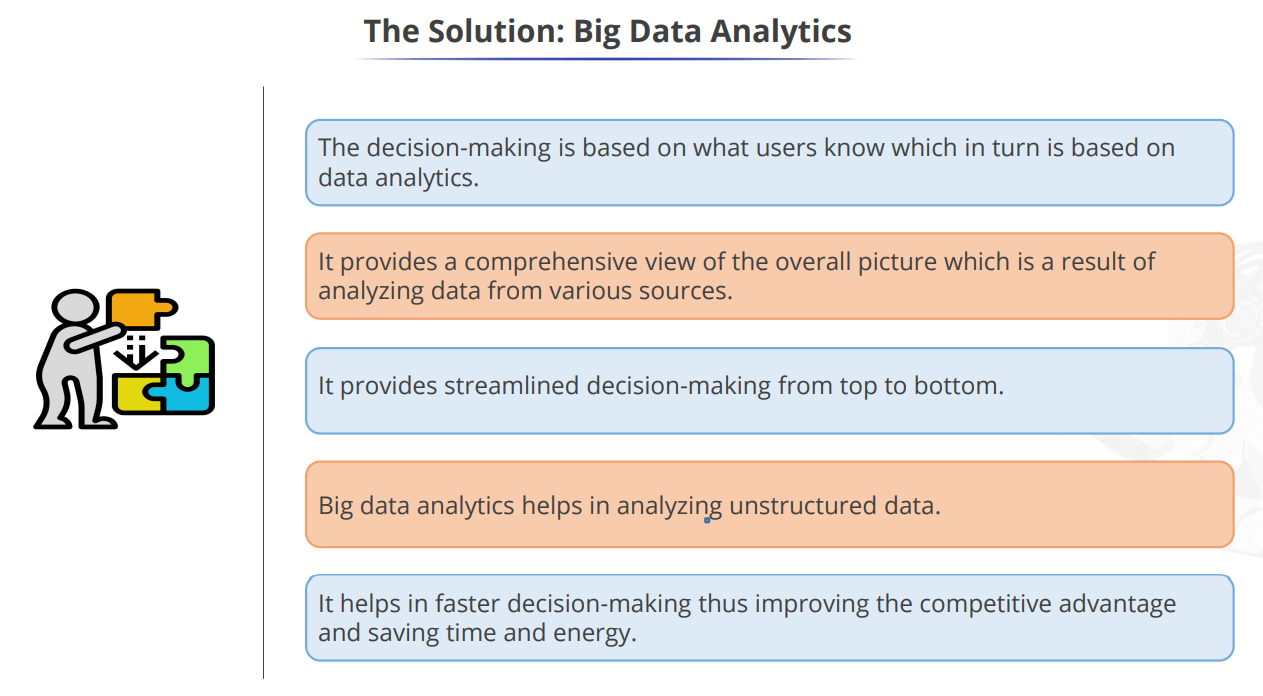


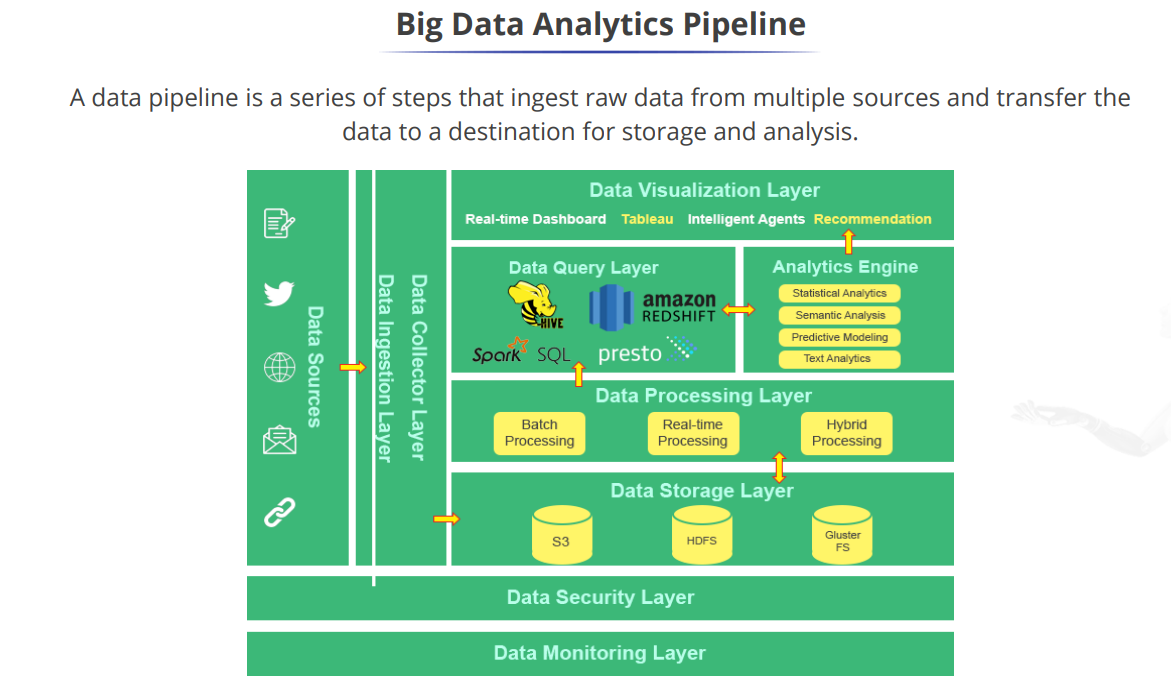
b





**Big data analytics** is a solution to the traditional decision making..



****

**Q) Give Real-time applications of Big Data Analytics.**

**1. Banking and Securities Industry:**

• This industry also heavily relies on Big Data for risk analytics, including; anti-money laundering, demand enterprise risk management, "Know Your Customer," and fraud mitigation.

• The Securities Exchange Commission (SEC) is using Big Data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

**2. Communications, Media and Entertainment Industry:** Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:

* Create content for different target audiences
* Recommend content on demand
* Measure content performance Eg.
* Spotify, an on-demand music service, uses Hadoop Big Data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users.
* Amazon Prime, which is driven to provide a great customer experience by offering video, music, and Kindle books in a one-stopshop, also heavily utilizes Big Data.

**3. Healthcare Sector:**

• Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidencebased medicine as opposed to administering several medical/lab tests to all patients who go to the hospital.

• Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

**4. Education:**

• The University of Tasmania, An Australian university with students has deployed a Learning and Management System that tracks, among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time.

• On a governmental level, the Office of Educational Technology in the U. S. Department of Education is using Big Data to develop analytics to help correct course students who are going astray while using online Big Data certification courses. Click patterns are also being used to detect boredom.

**5. Government:**

• In public services, Big Data has an extensive range of applications, including energy exploration, financial market analysis, fraud detection, health-related research, and environmental protection.

• The Food and Drug Administration (FDA) is using Big Data to detect and study patterns of food-related illnesses and diseases.

**6. Insurance Industry**: Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices, and CCTV footage. The Big Data also allows for better customer retention from insurance companies.

**7. Transportation Industry**: Some applications of Big Data by governments, private organizations, and individuals include:

• Governments use of Big Data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)

• Private-sector use of Big Data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)

**8. Energy and Utility Industry:** Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze the consumption of utilities better, which allows for improved customer feedback and better control of utilities use.

UNIT - 2

**Big Data Technologies: Hadoop’s Parallel World – Data discovery – Open source technology for Big Data Analytics – cloud and Big Data –Predictive Analytics – Mobile Business Intelligence and Big Data.**

**Explain big data technology?**

Technology is radically changing the way data is produced, processed, analyzed, and consumed. On one hand, technology helps evolve new and more effective data sources. On the other, as more and more data gets captured, technology steps in to help process this data quickly, efficiently, and visualize it to drive informed decisions. Now, more than any other time in the short history of analytics, technology plays an increasingly pivotal role in the entire process of how we gather and use data.

**The Elephant in the Room: Hadoop ’s Parallel World**

There are many Big Data technologies that have been making an impact on the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk.

* Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.

The original creators of Hadoop are Doug Cutting and Mike Cafarella.

* The name “Hadoop” itself comes from Doug ’s son, he just made the word up for a yellow plush elephant toy that he has. Yahoo! hired Doug and invested significant resources into growing the Hadoop project, initially to store and index the Web for the purpose of Yahoo! Search. That said, the technology quickly mushroomed throughout the whole company as it proved to be a big hammer that can solve many problems.
* Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve:

■ The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.

■ The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before, at a transformationally lower TCO than proprietary solutions and thereby take advantage of all their data to increase operational efficiency and gain a competitive edge. At one-tenth the cost of traditional solutions, Hadoop excels at supporting complex analyses— including detailed, special-purpose computation—across large collections of data.

■ Hadoop handles a variety of workloads, including search, log processing, recommendation systems, data warehousing, and video/image analysis. Today ’s explosion of data types and volumes means that Big Data equals big opportunities and Apache Hadoop empowers organizations to work on the most modern scale-out architectures using a clean-sheet design data framework, without vendor lock-in.

■ Apache Hadoop is an open-source project administered by the Apache Software Foundation. The software was originally developed by the world ’s largest Internet companies to capture and analyze the data that they generate. Unlike traditional, structured platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data. Hadoop stores terabytes, and even petabytes, of data inexpensively.

It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses.

■ Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

The two critical components of Hadoop are:

1. **The Hadoop Distributed File System (HDFS) --** HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

**2. MapReduce --** Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data.

Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

Both HDFS and MapReduce are designed to continue to work in the face of system failures.

HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster.

Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data. Because of the way that HDFS and MapReduce work.

Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost.

**Old vs. New Approaches:** 

We interviewed data guru Abhishek Mehta to get his perceptions of the differences between the “old” and “new” types of big data analytics. The following are the observations…

The old way is a data and analytics technology stack with different layers “cross-communicating data” and working on “scale-up” expensive hardware. The new way is a data and analytics platform that does all the data processing and analytics in one “layer,” without moving data back and forth on cheap but scalable (“scale out”) commodity hardware.

**The new approach is based on two foundational concepts:**

1. Data needs to be stored in a system in which the hardware is infinitely scalable. In other words, you cannot allow hardware (storage and network) to become the bottleneck.
2. Data must be processed, and converted into usable business intelligence where it sits. Put simply, you must move the code to the data and not the other way around.

That is a fundamental departure and the primary difference between the old way and the new way.

In the old ways, you had the multiple tiers of the stack and in the new way we have what is essentially a horizontal platform for data. The data sits in one place, you never move it around. That ’s the “secret” to big data analytics.

And here’s **another important point to remember:**

The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data.

Hardware and storage are more affordable than ever before, and continuing to get cheaper, which allows for increasingly larger and more ambitious massively parallel architectures. As the sheer quantity and complexity of data increases, our ability to handle complex and unstructured data is also rising.

**For the moment, let’s boil his observations down to three main points:**

1. The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data.

2. Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing.

3. The variety of data is on the rise and the ability to handle unstructured data is on the rise.

Today we can run the algorithm, look at the results, extract the results, and feed the business process—automatically and at massive scale, using all of the data available.

**Data Discovery: Work the Way People’s Minds Work:**

* Data discovery, the term used to describe the new wave of business intelligence that enables users to explore data, make discoveries, and uncover insights in a dynamic and intuitive way versus predefined queries and preconfigured drill-down dashboards.
* This approach has resonated with many business users who are looking for the freedom and flexibility to view Big Data.
* In fact, there are two software companies that stand out in the crowd by growing their businesses at unprecedented rates in this space: Tableau Software and QlikTech International.

**State the drawbacks of BI over the use of Qliktech?**

Both companies’ approach to the market is much different than the traditional BI software vendor. They grew through a sales model that many refer to as “land and expand.” It basically works by getting intuitive software in the hands of some business users to get in the door and grow upward.

In the past, BI players typically went for the big IT sale to be the preferred tool for IT to build reports for the business users to then come and use. In order to succeed at the BI game of the “land and expand model”.

**Tableau** provides a company to use an interactive dashboard to track the critical metrics driving their business. Interactivity is key which Tableau provides: a click on any filter lets the executive look into specific markets or products.

“Business intelligence needs to work the way people’s minds work. Users need to navigate and interact with data any way they want to—asking and answering questions on their own and in big groups or teams.”

**Qliktech** has designed a way for users to leverage direct— and indirect—search. With QlikView search, users type relevant words or phrases in any order and get instant, associative results. With a global search bar, users can search across the entire data set. With search boxes on individual list boxes, users can confidence the search to just that field. Users can conduct both direct and indirect searches. For example, if a user wanted to identify a sales rep but couldn’t remember the sales rep ’s name—just details about person, such as that he sells fish to customers in the Nordic region—the user could search on the sales rep list box for “Nordic” and “fish” to narrow the search results to just the people who meet those criteria.

**Differences between Tableau and Qliktech?**

**Open-Source Technology for Big Data Analytics:**

What is open-source technology/software/Stack? Explain it’s advantages.

Open-source software is computer software that is available in source code form under an open-source license that permits users to study, change, and improve and at times also to distribute the software.

The open-source name came out of a 1998 meeting in Palo Alto in reaction to Netscape’s announcement of a source code release for Navigator (as Mozilla). Although the source code is released, there are still governing bodies and agreements in place.

The most prominent and popular example is the GNU General Public License (GPL), which “allows free distribution under the condition that further developments and applications are put under the same license.” This ensures that the products keep improving over time for the greater population of users.

Some other open-source projects are managed and supported by commercial companies, such as Cloudera, that provide extra capabilities, training, and professional services that support open-source projects such as Hadoop. This is similar to what Red Hat has done for the open-source project Linux.

“One of the key attributes of the open-source analytics stack is that it’s not constrained by someone else ’s predetermined ideas or vision”

“The open-source stack doesn’t put you into a straitjacket. You can make it into what you want and what you need. If you come up with an idea, you can put it to work immediately. That’s the advantage of the open-source stack—flexibility, extensibility, and lower cost.”

“One of the great benefits of open source lies in the flexibility of the adoption model: you download and deploy it when you need it.“ With open source, you can try it and adopt it at your own pace.”

The pace of software development has accelerated dramatically because of open-source software.

The old model ’s end state was a monolithic stack of proprietary tools and systems that could not be swapped out, modified, or upgraded without the original vendor ’s support. This model was largely unchallenged for decades.

The old model was top-down, slow, inflexible and expensive. The new software development model is bottom-up, fast, flexible, and considerably less costly. A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors.

An open-source stack is defined by its community of users and contributors. No one “controls” an open-source stack, and no one can predict exactly how it will evolve. The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

Our hunch is that open-source and proprietary solutions will coexist for a long time, and for many good reasons. In fact, most proprietary vendors have been designing their solutions to plug and play with technology such as Hadoop. For example, Teradata Aster designed SQL-H, which is a seamless way to execute SQL and SQL-MapReduce on Apache Hadoop data.

Tasso Argyros, In a recent blog, explained **the significance of his firm ’s integration with open-source Hadoop**:

This is a significant step forward from what was state-of-the-art until yesterday. This means that [in the past] getting data from Hadoop to a database required a Hadoop expert in the middle to do the data cleansing and the data type translation. If the data was not 100% clean (which is the case in most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn’t directly access and analyze data in Hadoop clusters. SQL-H, an industry-first, solves all those problems.

**Explain** **the significance of big-data analytics for firm’s integration with open-source technology?**

**The Cloud and Big Data:**

It is important to remember that for all kinds of reasons—technical, political, social, regulatory, and cultural—cloud computing has not been a successful business model that has been widely adopted for enterprises to store their Big Data assets. However, there are many who believe that some obvious industry verticals will soon realize that there is a huge ROI opportunity if they do embrace the cloud.

There will be Big Data platforms that companies will build, especially for the core operational systems of the world. Where we continue to have an explosive amount of data come in and because the data is so proprietary that building out an infrastructure in-house seems logical. it’s going to the cloud, it’s just a matter of time!

It is believed that cloud models are inevitable for every industry and it’s just a matter of when an industry will shift to the cloud model.

**Clients** are saying they “don’t have unlimited capital to invest in infrastructure. their data is exploding both structured and unstructured. The models that they use to price products or manage risks are broken. they under immense pressure to streamline their operations and reduce headcount. How can they solve these problems?”

**Market economics** are demanding that capital-intensive infrastructure costs disappear and business challenges are forcing clients to consider newer models. At the crossroads of high capital costs and rapidly changing business needs is a sea change that is driving the need for a new, compelling value proposition that is being manifested in a cloud-deployment model.

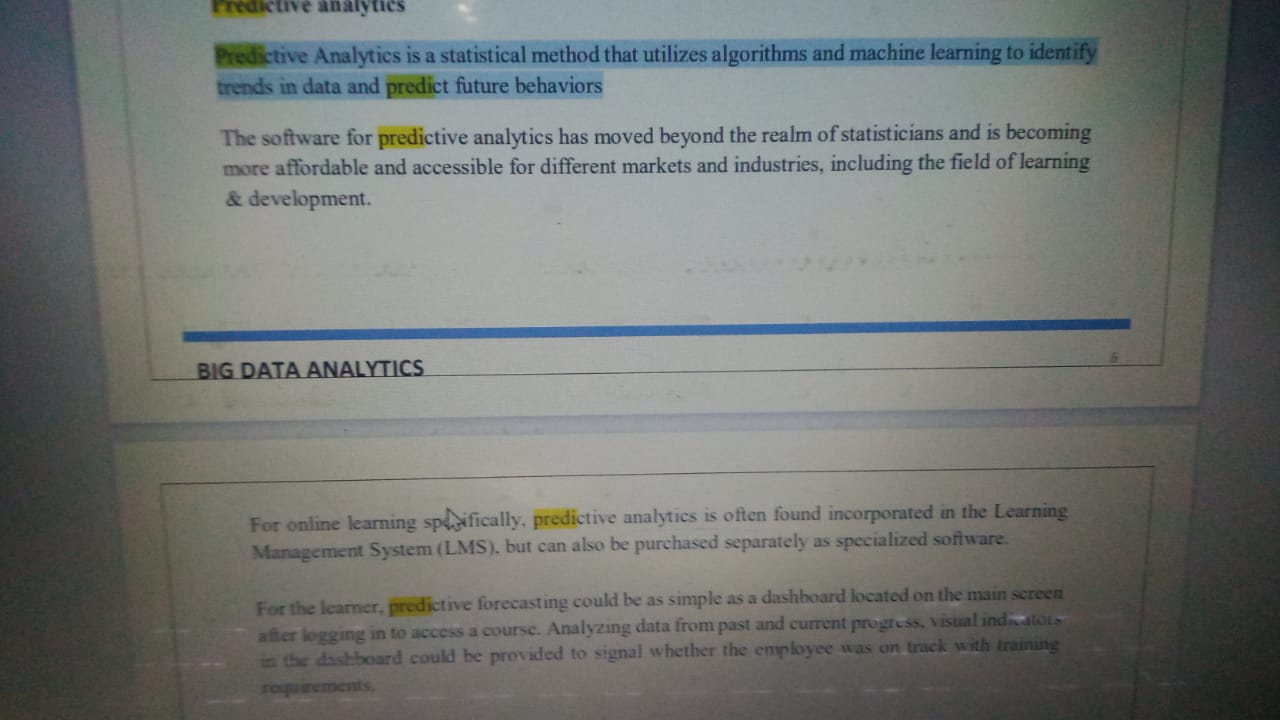
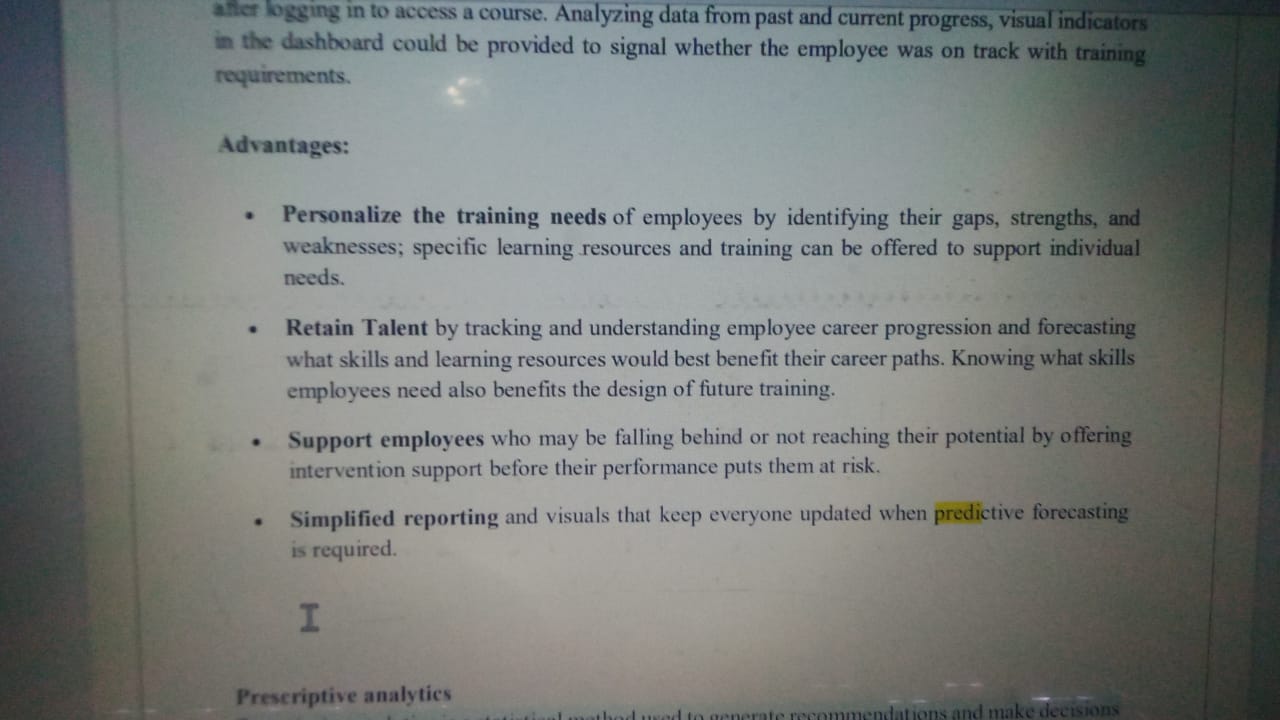
With a **Cloud model**, you pay on a subscription basis with no upfront capital expense. You don’t incur the typical maintenance fees—and all the updates on the platform are automatically available. The traditional cost of value chains is being completely disintermediated by platforms—massively scalable platforms where the marginal cost to deliver an incremental product or service is zero.

The ability to build massively scalable platforms—platforms where you have the option to keep adding new products and services for zero additional cost—is giving rise to business models that weren’t possible before. It is “the next industrial revolution, where the raw material is data and data factories replace manufacturing factories.”

**A few guiding principles that his firm stands by:**

1. **Stop saying “cloud.”** It’s not about the fact that it is virtual, but the true value lies in delivering software, data, and/or analytics in an “as a service” model. Whether that is in a private hosted model or a publicly shared one does not matter. The delivery, pricing, and consumption model matters.
2. **Acknowledge the business issues**. There is no point to make light of matters around information privacy, security, access, and delivery. These issues are real, more often than not heavily regulated by multiple government agencies, and unless dealt with in a solution, will kill any platform sell.
3. **Fix some core technical gaps.** Everything from the ability to run analytics at scale in a virtual environment to ensuring information processing and analytics authenticity are issues that need solutions and have to be fixed.

**Explain the emergence of cloud with growth of big data (to store it)?**

**Predictive Analytics Moves into the Limelight:**

To master analytics, enterprises will move from being in reactive positions (business intelligence) to forward learning positions (predictive analytics). Using all the data available—traditional internal data sources combined with new rich external data sources—will make the predictions more accurate and meaningful.

Algorithmic trading and supply chain optimization are just two typical examples where predictive analytics have greatly reduced the friction in business. Look for predictive analytics to proliferate in every facet of our lives, both personal and business.

**Here are some leading trends that are making their way to the forefront of businesses today:**

**■ Recommendation engines** similar to those used in Netflix and Amazon that use past purchases and buying behaviour to recommend new purchases.

■ **Risk engines** for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.

■ **Innovation engines** for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.

■ **Customer insight engines** that integrate a wide variety of customer related info, including sentiment, behaviour, and even emotions. Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.

■ **Optimization engines** that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs.

Today we are at the tip of the iceberg in terms of applying predictive analytics to real-world problems. With predictive analytics you can realize the uncontested market space [competitive free].

**Mobile Business Intelligence Is Going Mainstream:**

Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one. Mobile devices are a great levelling field where making complicated actions easy is the name of the game. For example, a young child can use an iPad but not a laptop. As a result, this will drive broad-based adoption as much for the ease of use as for the mobility these devices offer.

People have been talking about mobile BI for quite some time, especially since the 1999 release of the good-old BlackBerry. However, it seems as though we have finally hit an inflection point.

Kerzner explains his view on this topic:

Kerzner shares his experience of how mobile BI became mainstream after the introduction of the iPad. He says that touch-driven devices like the iPhone are more about software than hardware, and that this allows for more innovation and usability in mobile analytics.

The combination of multi-touch and having a software oriented device is what has unlocked the potential of these devices to really bring mobile analytics and intelligence to a much wider audience in a productive way.

**Ease of Mobile Application Deployment:**

Another inflection point for the industry is the development and deployment of mobile applications. In the past, that was controlled by the relationship with the carrier. It used to be that if you wanted to push out a mobile application, the only way you could get that application on the phone for the most part was to go through the carriers. That meant there were development environments that were sometimes very proprietary or you had to develop one set of applications for one carrier and another set of applications for a different one, maybe a third for the RIM BlackBerry environment.

Kerzner elaborated:

One of the things that’s happened recently is that with the advent of these app stores and the maturing of the browsers on the devices into something much more powerful, now as a software provider, you can go directly to the end user. I can go to a corporation and say I ’m going to roll out a powerful global reporting application that’s also going to do deal approvals and it’s going to totally change a whole business process. I think something that was previously locked in a desk will now give people insights into the purchasing patterns, as well as the ability to take that action. I can roll out that whole application— I never have to talk to anybody but that customer because the devices that everybody ’s lugging around are really little computers and of course you can put any software you want on a little computer and that really wasn’t the case historically in the mobile space.

**Three elements that have impacted the viability of mobile BI:**

1. Location—the GPS component and location . . . know where you are in time as well as the movement.

2. It’s not just about pushing data; you can transact with your smart phone based on information you get.

3. Multimedia functionality allows the visualization pieces to really come into play.

**Three challenges with mobile BI include:**

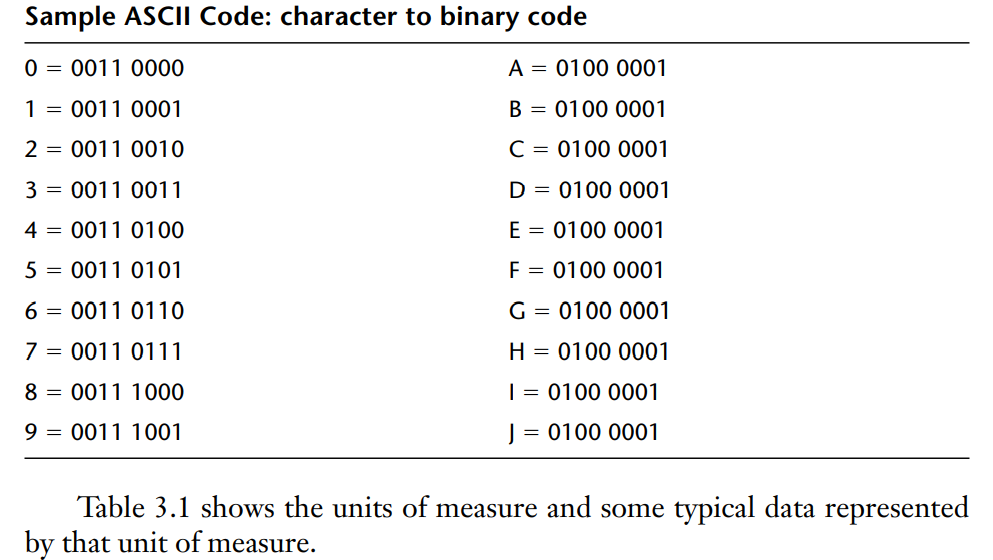
1. Managing standards for rolling out these devices.

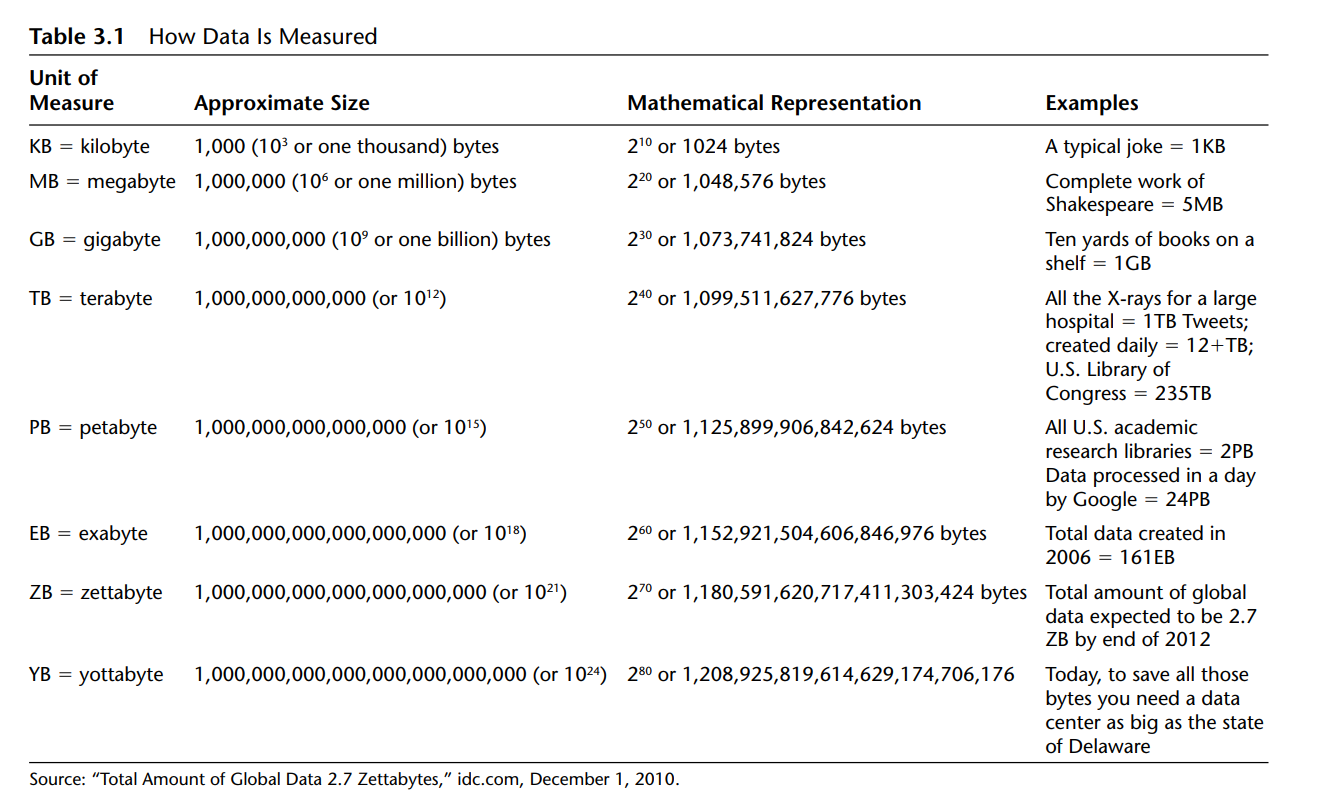
2. Managing security (always a big challenge).

3. Managing “bring your own device,” where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

**Data Size 101**

Data is measured by basic units of measure that work up from a bit. A bit is represented by either a 1 (electricity flowing) or a 0 (no electricity flowing). This is called binary code. The code converts images, text, and sounds into numbers in order to send information from one digital device such as a computer to another. Computers use binary numbers because they are easier to handle. In binary, the digits (read and write) are worth 1, 2, 4, 8, and so on—not units, tens, and hundreds. A byte is a unit of measure and it is 8 bits put together. In ordinary numbers, “1,001” is one unit. But in binary, “1001” is one 1, no 2, no 4, and one 8, which equals 9.



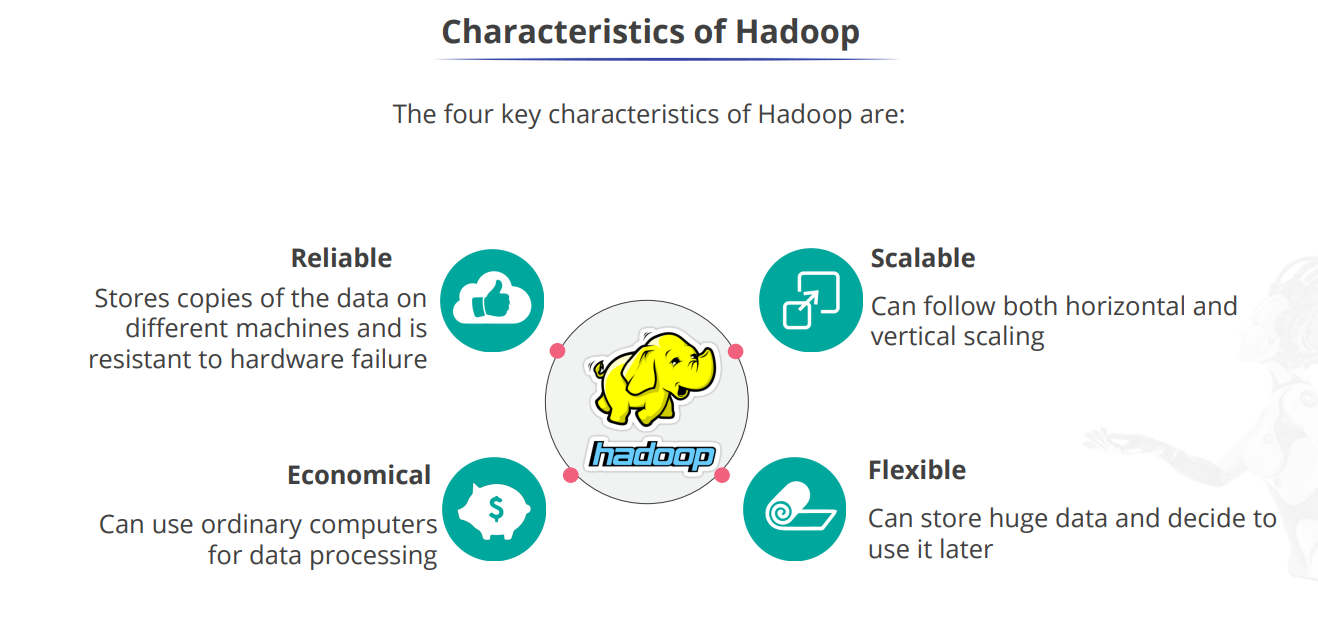


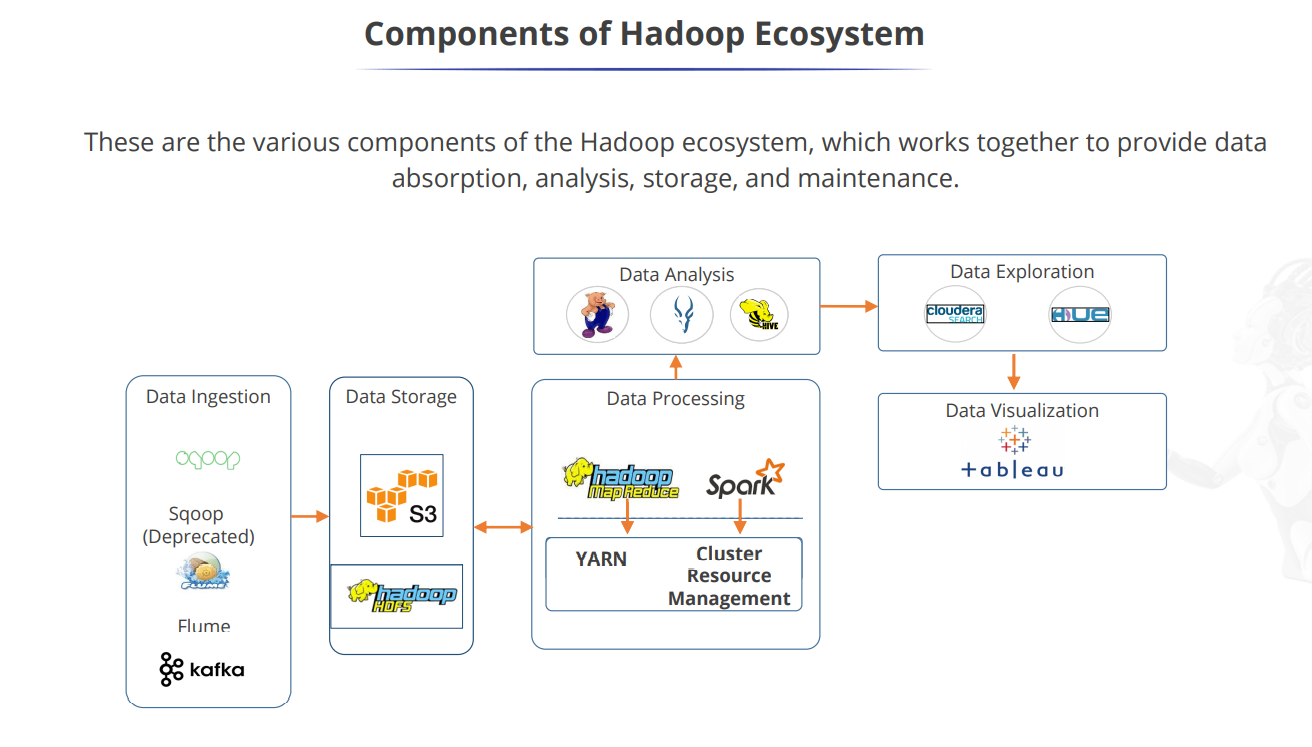
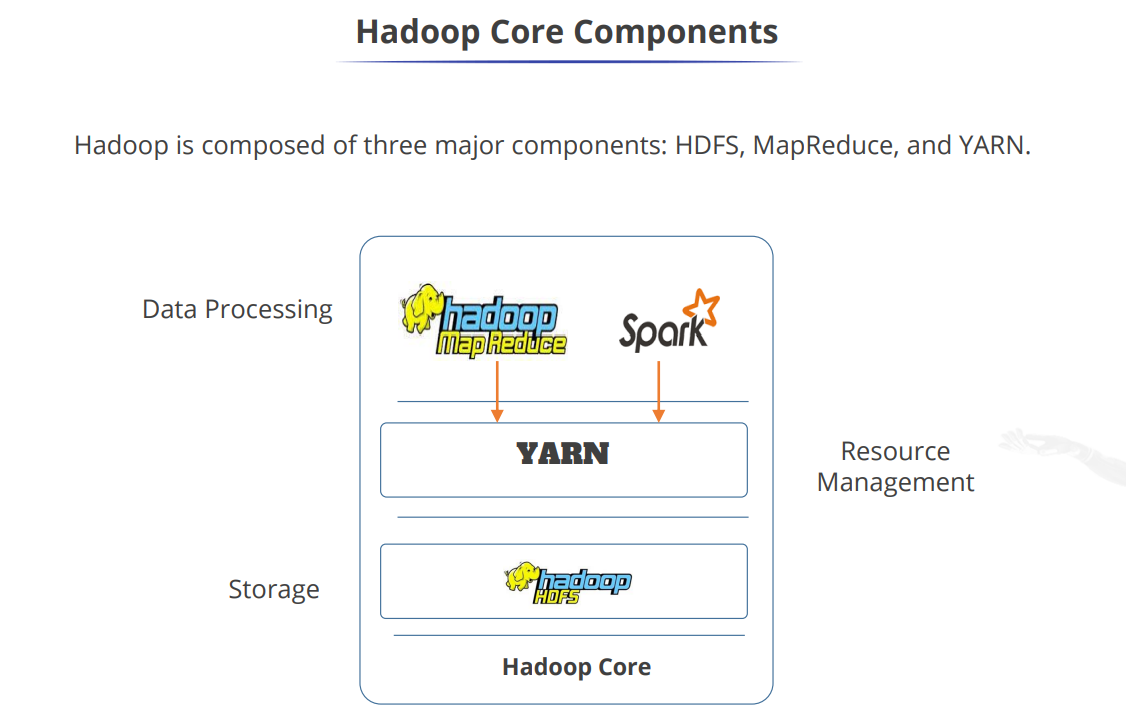
UNIT-3

**Introduction Hadoop: Big Data – Apache Hadoop & Hadoop Eco System – Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce - Data Serialization.**

**What Is Hadoop?**

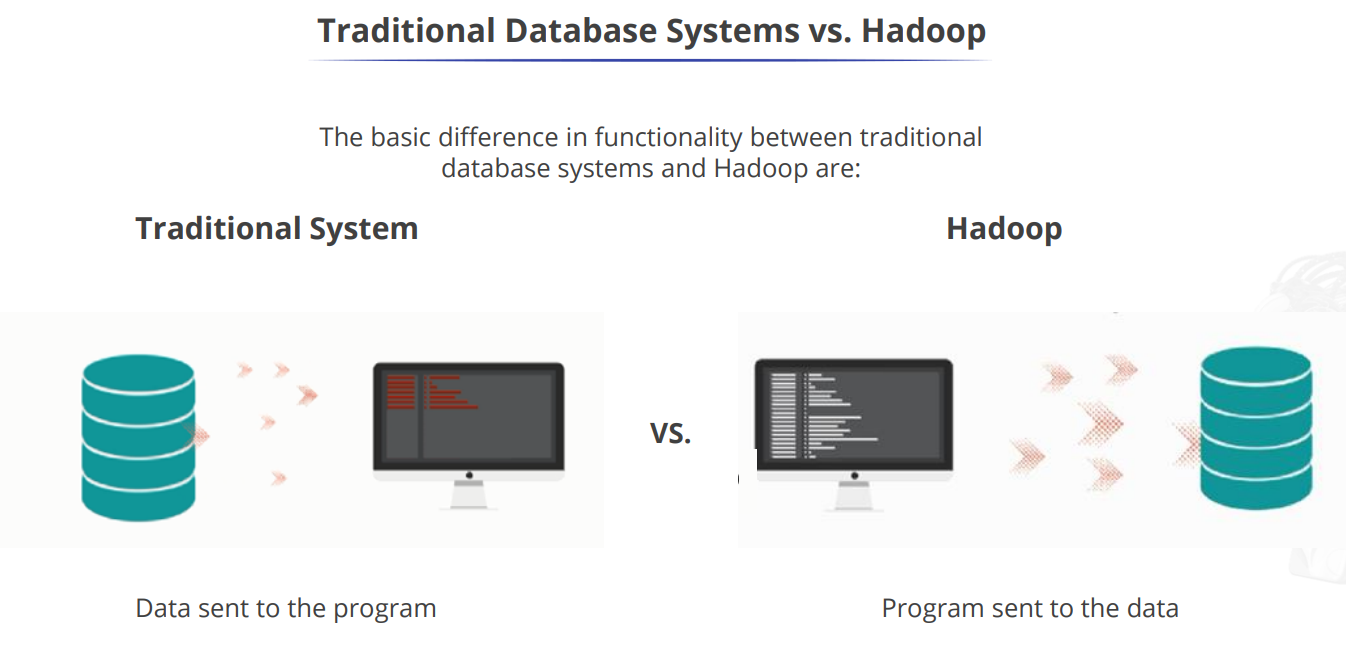
Hadoop is a framework that allows distributed processing of large datasets across clusters of commodity computers using simple programming models.

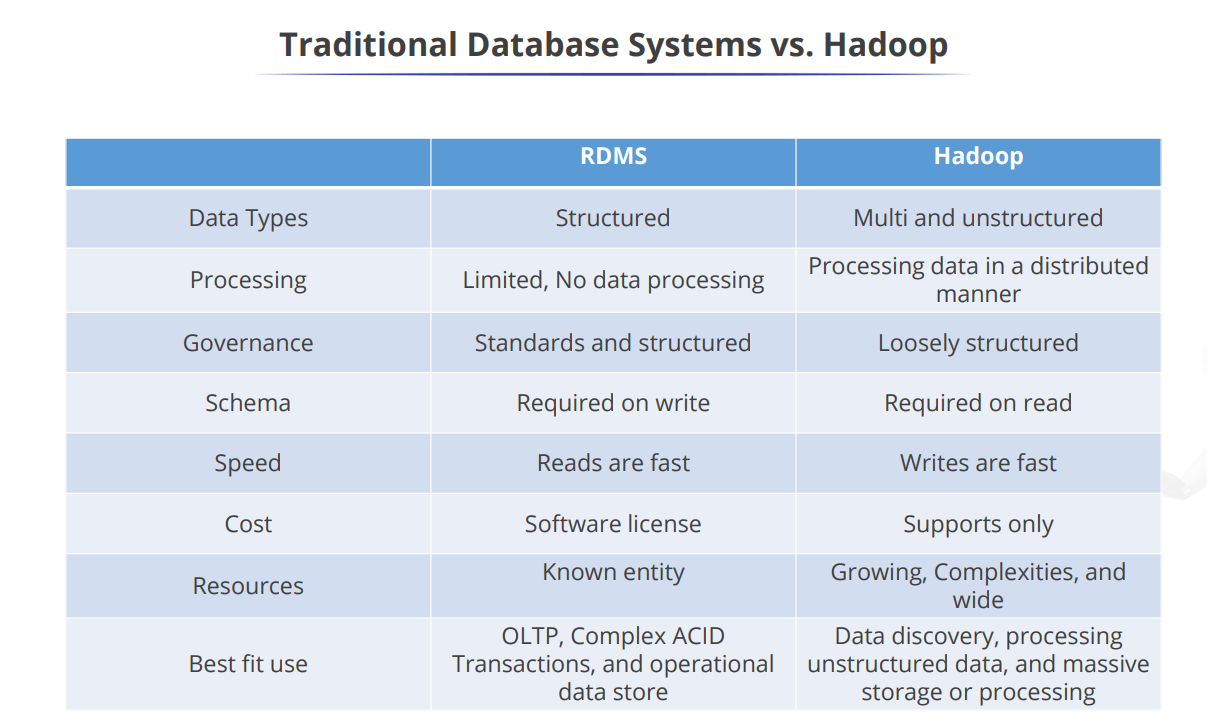


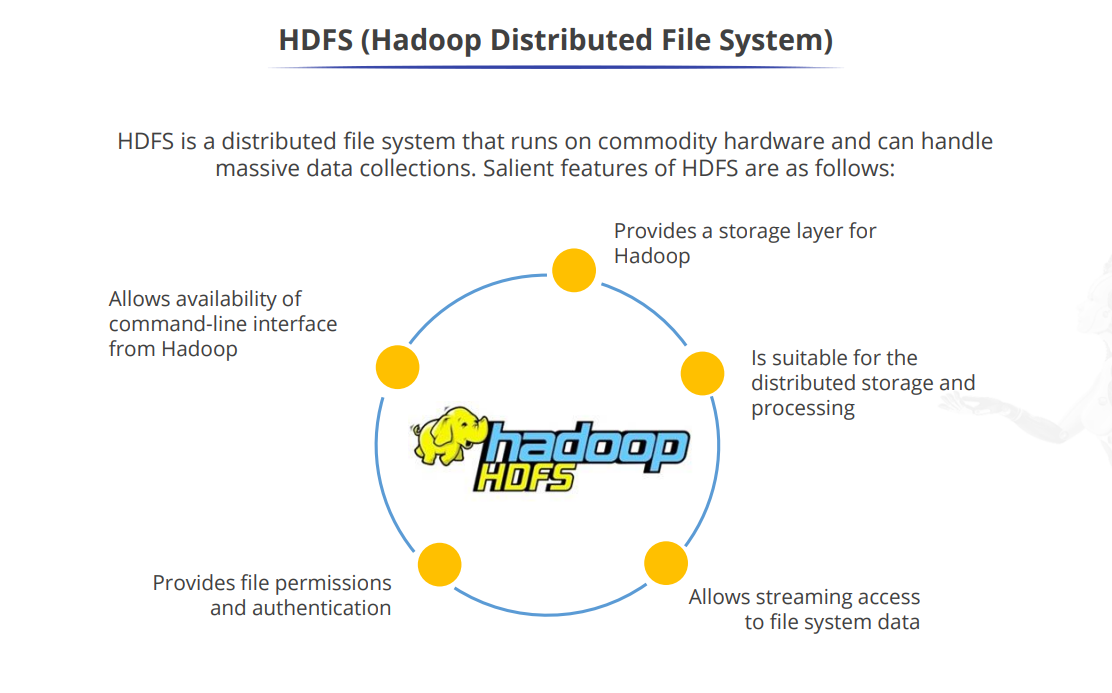


UNIT-4

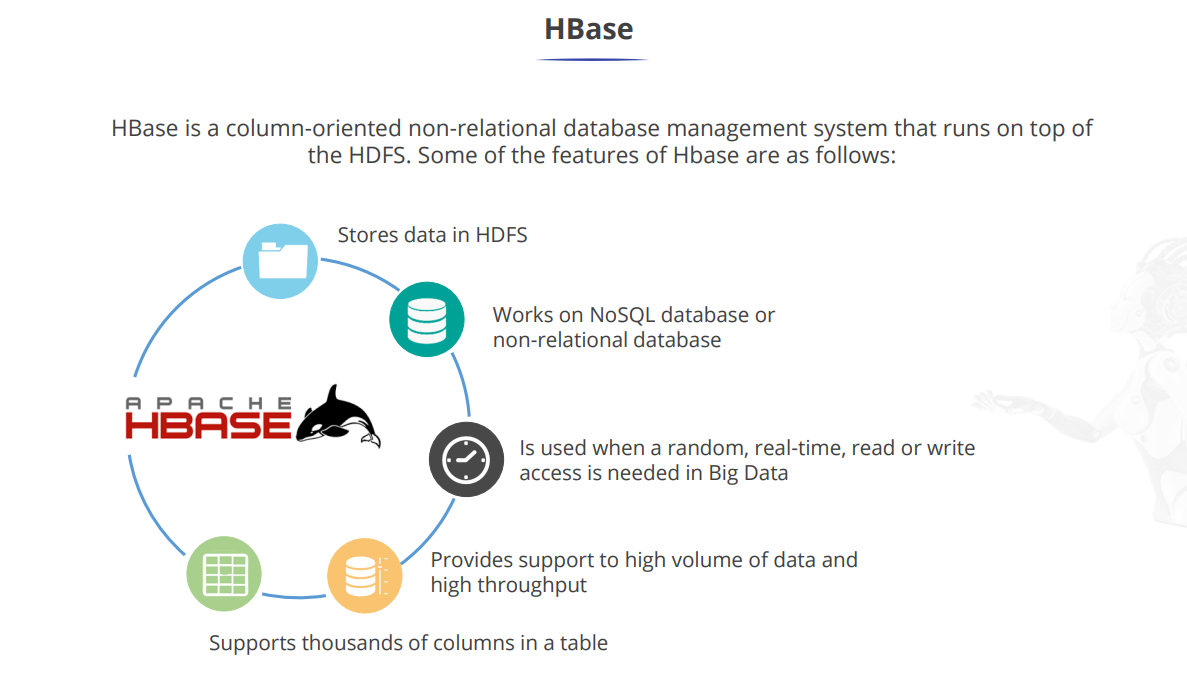
Hadoop Architecture: Hadoop: RDBMS Vs Hadoop, Hadoop Overview, Hadoop distributors, HDFS, HDFS Daemons, Anatomy of File Write and Read., Name Node, Secondary Name Node, and Data Node, HDFS Architecture, Hadoop Configuration, Map Reduce Framework, Role of HBase in Big Data processing, HIVE, PIG.







[Introduction to HBase for Hadoop | HBase Tutorial - MindMajix](https://mindmajix.com/hadoop/introduction-to-hbase)



HBase provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases.

HBase relies on [ZooKeeper](https://www.ibm.com/analytics/hadoop/zookeeper" \o "us-en_analytics_hadoop_SP_zookeeper) for high-performance coordination. ZooKeeper is built into HBase.

HBase works well with [Hive](https://www.ibm.com/analytics/hadoop/hive), a query engine for batch processing of big data, to enable fault-tolerant big data applications.

**Features of HBase –**

1. It is linearly scalable across various nodes as well as modularly scalable, as it divided across various nodes.
2. HBase provides consistent read and writes.
3. It provides atomic read and write means during one read or write process, all other processes are prevented from performing any read or write operations.
4. It provides easy to use Java API for client access.
5. It supports Thrift and REST API for non-Java front ends which supports XML, Protobuf and binary data encoding options.
6. It supports a Block Cache and Bloom Filters for real-time queries and for high volume query optimization.
7. HBase provides automatic failure support between Region Servers.
8. It support for exporting metrics with the Hadoop metrics subsystem to files.
9. It doesn’t enforce relationship within your data.
10. It is a platform for storing and retrieving data with random access.

## ****Storing Big Data with HBase****

[HBase](https://www.codingninjas.com/codestudio/library/hbase-introduction) is highly configurable and gives great flexibility to address massive amounts of data efficiently. Now let's understand how HBase can help address your significant data challenges.

* HBase is a columnar database. Like relational database management systems (RDBMSs), it stores all data in tables with columns and rows.
* The intersection of a column and row is called a cell. Each cell value contains a "version" attribute that is no more than a timestamp, distinctively selecting the cell.
* Versioning tracks swap in the cell and makes it possible to redeem any version of the contents.
* HBase stores the data in cells in decreasing order (using the timestamp), so a reader will always first choose the most current values.
* Columns in HBase belong to a column family. The column family name is used to identify its family members.
* The rows in HBase tables also have a key associated with them. The structure of the key is very flexible. It can be a computed value, a string, or another data structure.
* The key is used to control access to the cells in the row, and they are stored in order from low to high value.
* These features together make up the schema. It can alter new tables and column families after the database is up and running.

Unlike [relational database systems](https://www.ibm.com/analytics/relational-database), HBase does not support a structured query language like SQL; in fact, HBase isn’t a relational data store at all. HBase applications are written in Java™ much like a typical [Apache MapReduce](https://www.ibm.com/analytics/hadoop/mapreduce) application. HBase does support writing applications in [Apache Avro](https://www.ibm.com/analytics/hadoop/avro), REST and Thrift.

It comprises a set of standard tables with rows and columns, much like a traditional database. Each table must have an element defined as a primary key, and all access attempts to HBase tables must use this primary key.

**An example of HBase**

HBase allows for many attributes to be grouped together into column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families.

Just as HDFS has a NameNode and slave nodes, and MapReduce has JobTracker and TaskTracker slaves, HBase is built on similar concepts. In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data. HBase is also sensitive to the loss of its master node.

HBase, or Hadoop Database, is an open-source, distributed, column-oriented NoSQL database that is designed to handle large amounts of structured and semi-structured data. It is part of the Apache Hadoop ecosystem and is specifically built to provide scalable, high-performance storage and random access to big data.

The role of HBase in big data processing can be summarized in the following key points:

1. \*\*Scalable Storage:\*\* HBase is designed to store and manage massive amounts of data. It can handle petabytes of data distributed across a cluster of commodity hardware. It achieves scalability by horizontally partitioning data and distributing it across multiple nodes in the cluster. This enables it to handle the ever-increasing data volumes in big data processing scenarios.

2. \*\*High Performance:\*\* HBase provides low-latency random read and write access to data. It achieves this by storing data in a columnar format and maintaining indexes, called Bloom filters, for efficient data retrieval. HBase also leverages the distributed nature of Hadoop to parallelize data processing across multiple nodes, enabling faster data access and processing.

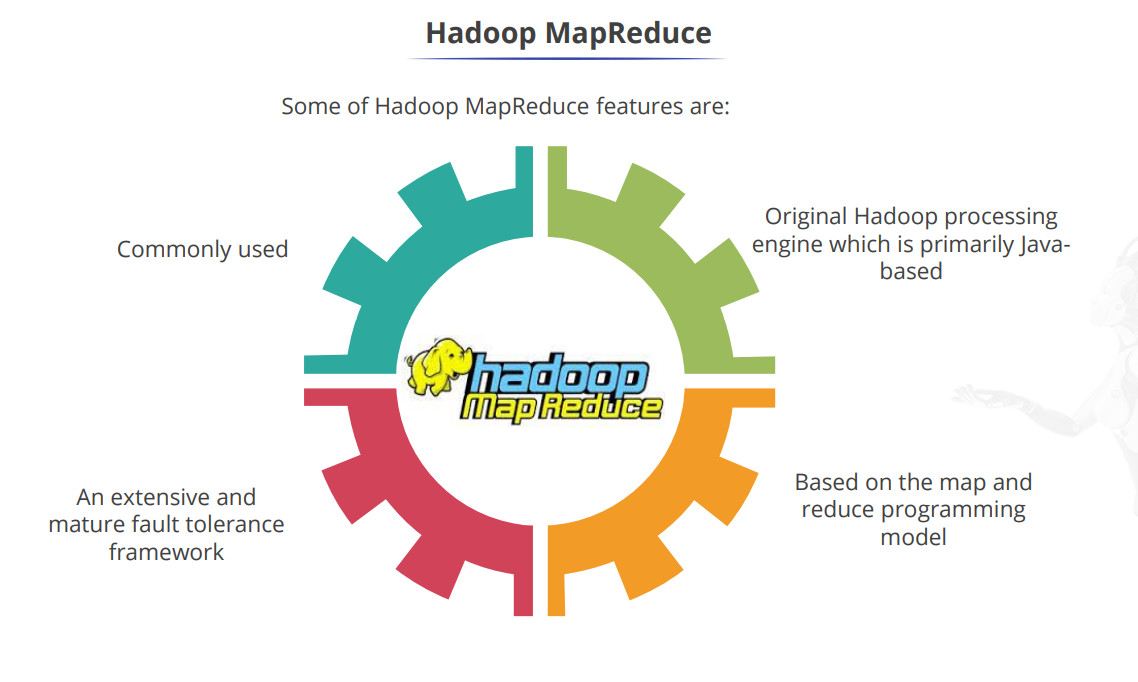
3. \*\*Fault Tolerance:\*\* HBase ensures data reliability and fault tolerance by replicating data across multiple nodes in the cluster. If a node fails, HBase automatically redirects the requests to other available nodes, ensuring continuous availability of data. The replication factor can be configured to provide the desired level of data durability.

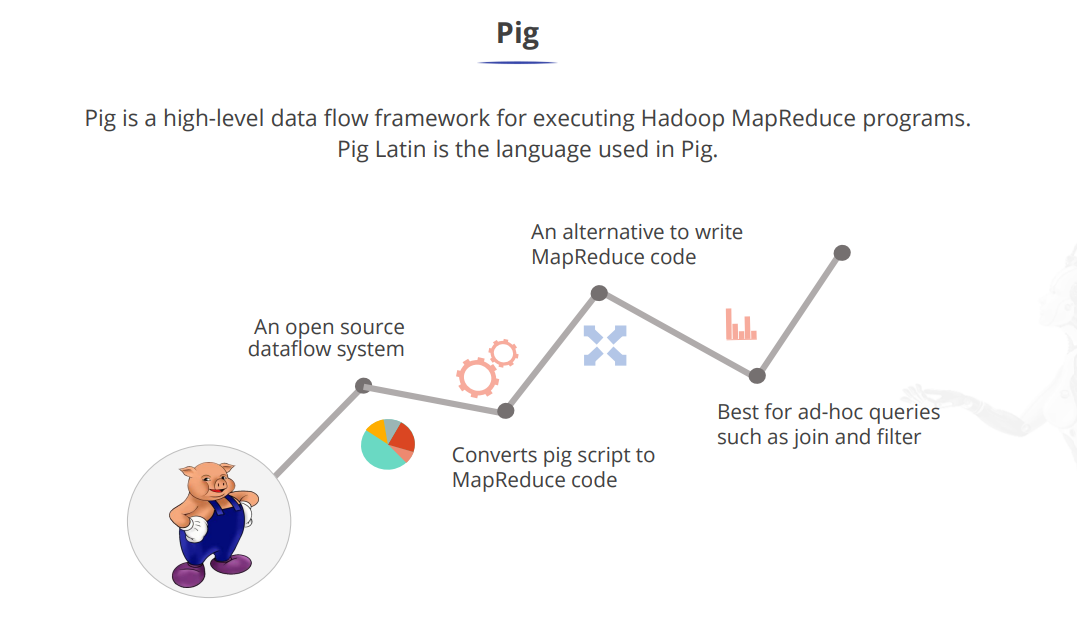
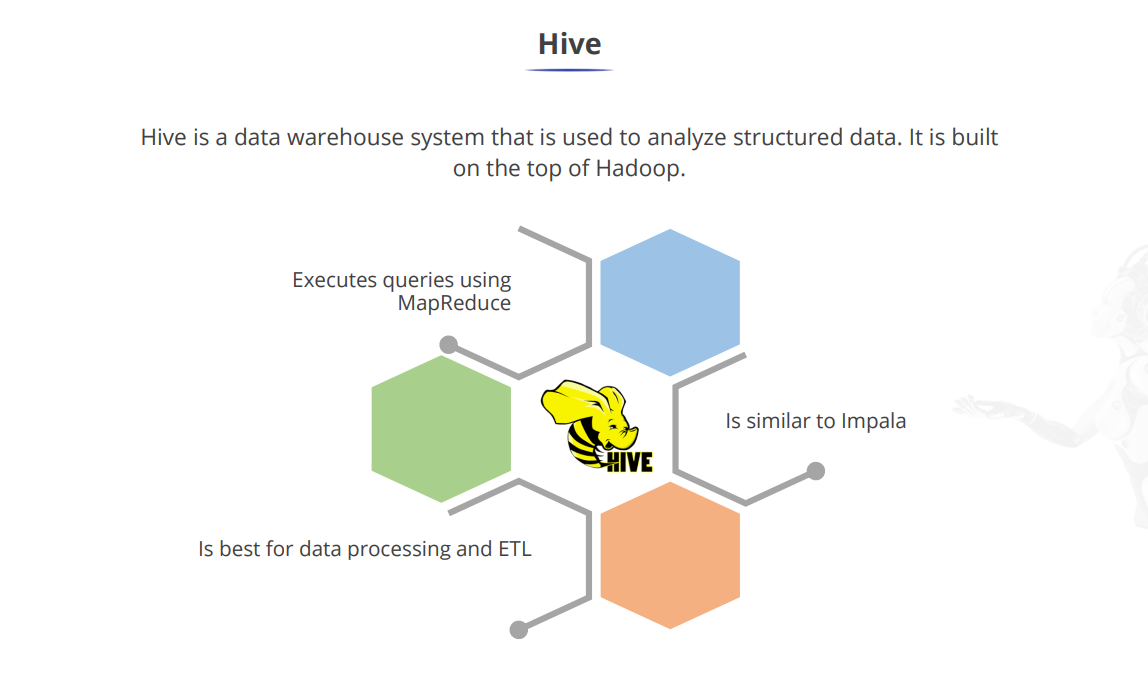
4. \*\*Schema Flexibility:\*\* HBase is schema-less, meaning it can handle data with varying structures and formats. It allows for dynamic column addition or modification without requiring any predefined schema changes. This flexibility is well-suited for scenarios where data formats may evolve or change frequently, common in big data processing.

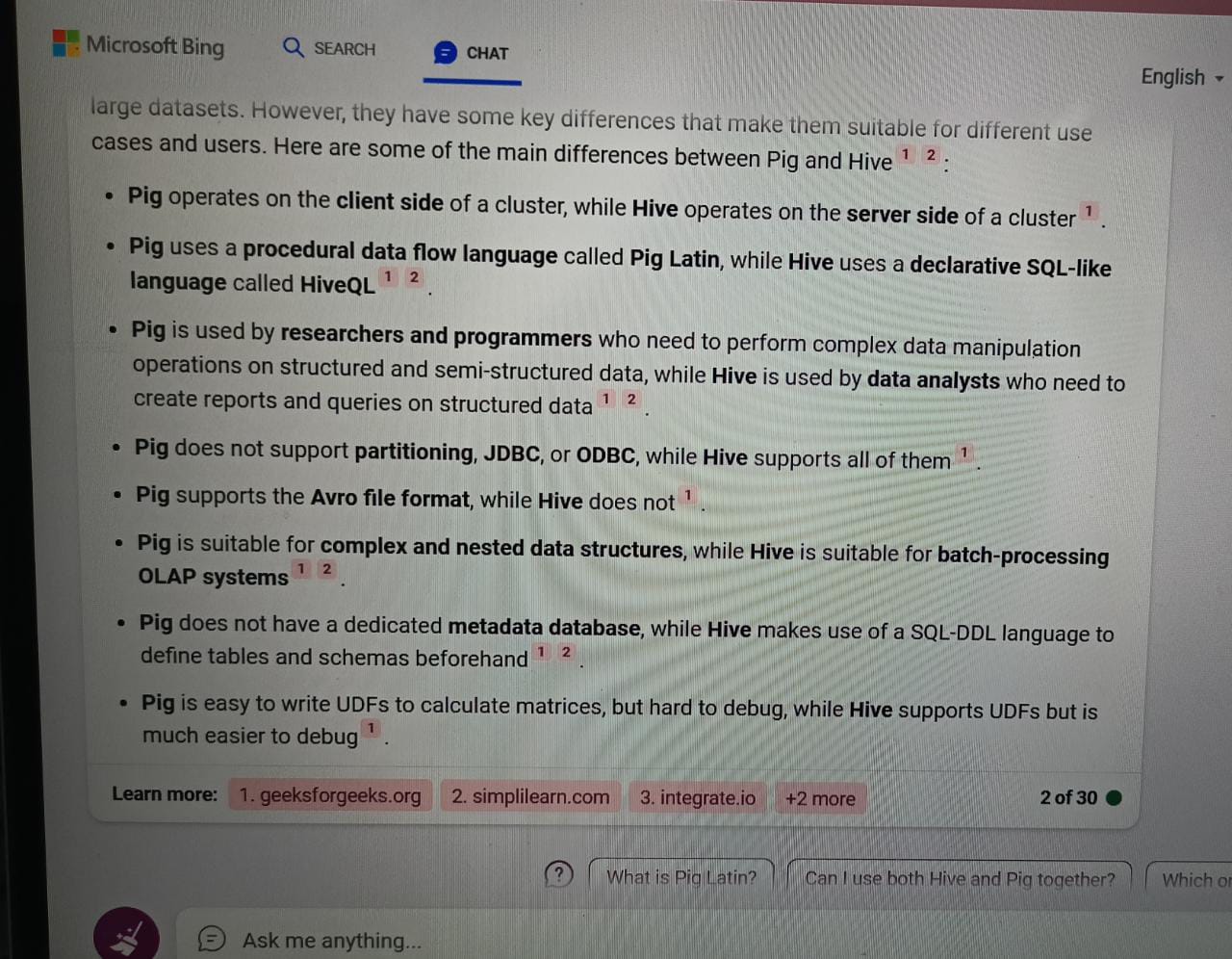
5. \*\*Integration with Hadoop Ecosystem:\*\* HBase seamlessly integrates with other components of the Hadoop ecosystem, such as Apache Hive, Apache Pig, and Apache Spark. This integration enables easy data exchange and processing between different tools, allowing for complex data analytics and processing workflows.

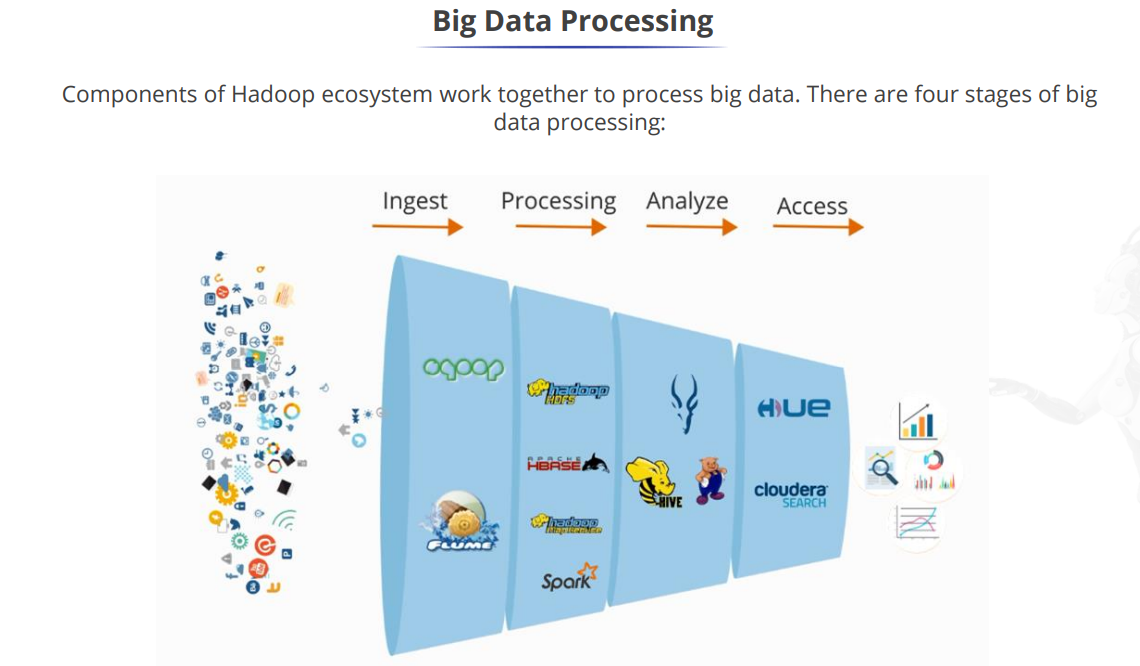
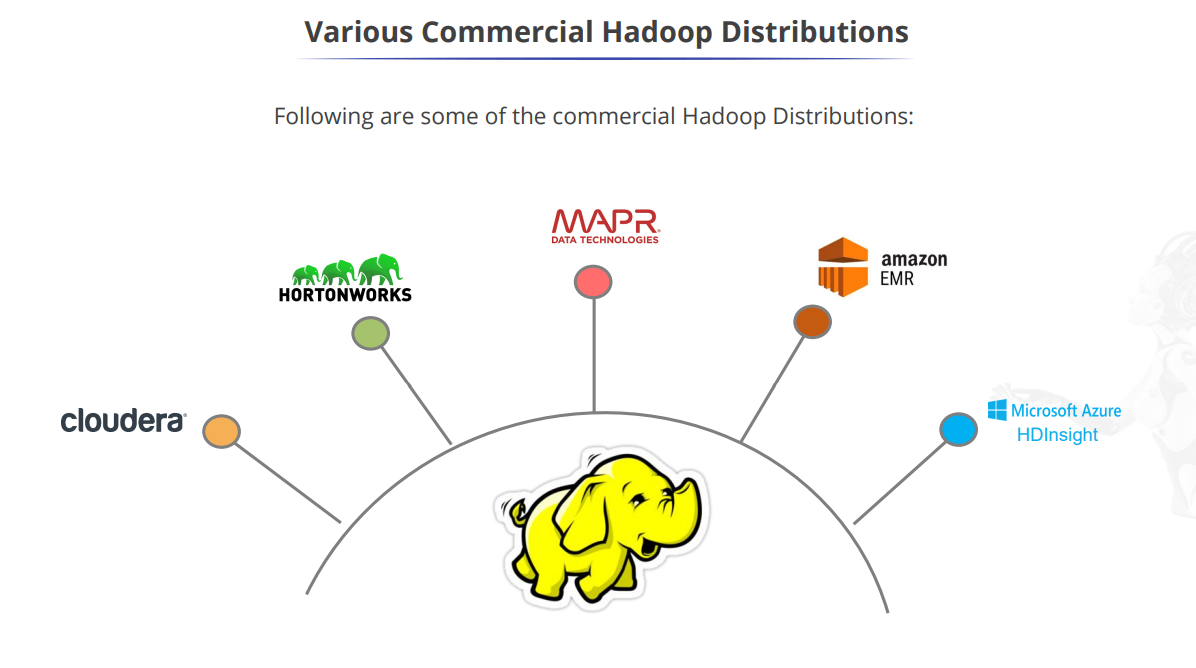
6. \*\*Real-time Data Processing:\*\* HBase supports real-time data ingestion and processing. It allows for efficient writes and updates, making it suitable for use cases that require low-latency data access, such as real-time analytics, fraud detection, monitoring systems, and social media applications.

Overall, HBase plays a crucial role in big data processing by providing a scalable, high-performance, fault-tolerant storage system that can handle vast amounts of structured and semi-structured data. Its integration with the Hadoop ecosystem and support for real-time data processing make it a valuable tool for building large-scale data-intensive applications.









UNIT – V

Data Analytics with R Machine Learning: Introduction, Supervised Learning, Unsupervised Learning, Collaborative Filtering, Social Media Analytics, Mobile Analytics, Big Data Analytics with BigR