**1.What is Big Data? -An Introduction**

Updated: Jan 18, 2020

A Big Data Blog.

**Big data is all over the places, and the term used everywhere. The curiosity about what is Big Data has been soaring in the past few years.**

Diagrama

Descrição gerada automaticamente

**Some facts:**

* Users watch *4.15 million YouTube videos*
* *S*end *456,000 tweets* on Twitter
* Post *46,740 photos* on Instagram
* *510,000 comments* posted on FB
* *293,000 statuses* updated on Facebook

These activities on the Web are the leading formation of Big Data; the business application and weblog data adds additional formation within the organization.

**The gaining of Big Data importance:**

The usage of manual paper, files, floppy, discs, etc. have now become obsolete because of the exponential growth of data. The growth of the data leads to the usage of the database for the storage of the data.

The exponential growth of the data and the usage also lead to new inventions, technologies, applications with a quick response time.

The introduction of the internet leads to the uncontrollable growth of data from all directions. The generation of continuous and massive data can refer to as Big Data. So, the [#InternetOfThings](https://www.abigdatablog.com/blog/hashtags/InternetOfThings)([#IoT](https://www.abigdatablog.com/blog/hashtags/IoT)) is one such technology that plays a significant role in this acceleration. 90% of all data today was generated in the last two years.

Forbes reports that there are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating.

**Database Management Systems:**

[**#Traditional**](https://www.abigdatablog.com/blog/hashtags/Traditional) **vs** [**#Realtime**](https://www.abigdatablog.com/blog/hashtags/Realtime)

“In traditional relational database management systems, data often moved to computational space for processing. In Big Data space, the computation performed at the data location. So, everything is real-time.”

A vital feature of these types of real-time notifications is that they enable real-time actions. However, using such a capability would require you to approach your application and your work differently.

Most Big Data companies have updated their culture to be more real-time action-oriented, refining real-time processes to handle anything from customer relations and fraud detection to system monitoring and control.

Such volumes of real-time data and analytical operations that need to take place requires increased use of scalable computing systems, which need to be a part of the planning for an organizational Big Data strategy.

**Big Data Defined:**

[#BigData](https://www.abigdatablog.com/blog/hashtags/BigData) is not just about the size and the volume of the data. [#BigData](https://www.abigdatablog.com/blog/hashtags/BigData) refers to the massive amounts of data sourced from various data sources and has in different data formats. Data was always captured and stored in databases, but because of the varied nature of this data, the traditional relational database systems are incapable of handling this data.

**Conclusion:**

[#BigData](https://www.abigdatablog.com/blog/hashtags/BigData) is much more than a collection of datasets with different formats; it is an important asset that can get used to obtaining enumerable benefits.

**2. Big Data; What? Why? and How?**

Updated: Jan 18, 2020

**Big data is a powerful tool, as well, necessary for the functioning of modern-day business.**

*Note: If you did not get to read the previous post in this series click the link* : <https://www.abigdatablog.com/post/series-1-what-is-big-data-part-1-an-introduction>

Pessoas na frente de um computador com a imagem de uma pessoa

Descrição gerada automaticamente com confiança baixa

**Welcome to the blog post about Big Data - Series 1.**

***If you did not get to read the previous post in this series click the link* :**

[**https://www.abigdatablog.com/post/series-1-what-is-big-data-part-1-an-introduction**](https://www.abigdatablog.com/post/series-1-what-is-big-data-part-1-an-introduction)

Big data is a powerful tool, as well, necessary for the functioning of modern-day business. Given how competitive the business world is, underestimating the benefits of big data will be a big mistake.

**What is Big Data**

“ The term Big Data refers to a massive volume of data; Data that cannot be stored or processed by traditional data storage or processing technologies. Big Data generated at **an enormous** scale, and it is being used by many companies to **process** and **analyze** to uncover insights and improve the business of many organizations. ”

Given how competitive the business world is, underestimating the benefits of big data will be a big mistake.

**Why is Big-data different from other data?**

The traditional databases are very much structured data with a predefined model and architecture. As well, the conventional databases are slow in storing and processing data. The traditional systems do not have the bandwidth to store and process the data with the required speed.

The Big-data, on the other hand, is the modern data structure; A data structure comes in different formats from multiple types of sources. The form of the data is not specific; some of the data types are text and un-structured most of the time. So, the traditional databases are not an excellent fit to store, process, and analyze the data.

The [#Big](https://www.abigdatablog.com/blog/hashtags/Big)-data is resilient to unstructured data storage and processing capabilities. The Big-data relies on modern architecture; stores data on a distributed and fault-tolerant system.

Big data provides a reliable system and architecture to provide value to the company. Big data replaces the traditional [#ETL](https://www.abigdatablog.com/blog/hashtags/ETL) (Extract, Transform, and load) to [#ELT](https://www.abigdatablog.com/blog/hashtags/ELT) (Extract, Load, and Transform). [#ELT](https://www.abigdatablog.com/blog/hashtags/ELT) removes the complexity and delay in providing the data for the decision-making process.

**Categorizing Big Data**

“ We can categorize Big Data into three; they are **Structured Data, Semi-Structured Data, and Unstructured Data.** ”

**Characterizing Big Data**

“ The massive amount of data, so there is a high volume."

"There is a vastly different format the data is collected, so there is enormous variety."

"The accumulation of the data is high with speed in the collection, so there is high velocity.”

**Advantages of Big Data**

Big Data Technology has given us multiple benefits:

* Big Data has enabled predictive analysis, which can save organizations from operational risks, grow business, understand the customer need, patient records, etc.
* Big data help with root cause analysis of failures in real-time.Understand customer buying habits and generate customer offers in real-time.Improve customer engagement and increasing customer loyalty
* Big-data helps with data-driven marketing.
* Big data provides lots of data and a considerable amount of history for decision making.
* Big data provides data accessible, precise, and timely for decision making.

**Conclusion:**

Big data provides a reliable system and architecture to provide value to the company. Big Data can be intimidating, but the business has an advantage with competitors; when the Big-data technology accompanied by proper data architecture and data management solutions.

**3.Big-data : Traditional Processing VS Real-time.**

Updated: Jan 18, 2020

**Big data; What is the difference between traditional data processing vs real-time processing.**

***Note:*** *If you did not get to read the previous post in this series click the link* : Series 1 Part 2 <https://www.abigdatablog.com/post/big-data-what-how-and-why>

Interface gráfica do usuário, Texto

Descrição gerada automaticamenteInterface gráfica do usuário, Texto

Descrição gerada automaticamente

Welcome to the blog post; Big Data and real-time data processing.

**Database Management Systems**

“In traditional relational database management systems, data often moved to computational space for processing. In Big Data space bringing the computation done at the data location. So, everything is real-time.”

A vital feature of these types of real-time notifications is that they enable real-time actions. However, using such a capability would require you to approach your application and your work differently.

Most Big Data companies have updated their culture to be more real-time action-oriented, refining real-time processes to handle anything from customer relations and fraud detection to system monitoring and control.

Such volumes of real-time data and analytical operations that need to take place requires increased use of scalable computing systems, which need to be a part of the planning for an organizational Big Data strategy.

**4. Big Data - Data Sources and Types**

Updated: Jan 18, 2020

**What are the different sources of Big-data and what are the different types of data?**

**Note:**

**If you did not get to read the previous post in this series, follow the link: Series 1 Part 3**

<https://www.abigdatablog.com/post/big-data-traditional-vs-real-time-processing>

Diagrama

Descrição gerada automaticamente

Welcome to the blog post; sources of Big-data and the source types.

“The primary sources of Big Data are data generated by machines, people, and organizations..”

Big data can be either structured, semi-structured, or unstructured.

**Machine Generated Data:**

Machine-generated data we refer to data generated from [#real](https://www.abigdatablog.com/blog/hashtags/real)-time sensors in industrial machinery or vehicles that logs that track user behavior online, environmental [#sensors](https://www.abigdatablog.com/blog/hashtags/sensors) or personal health trackers, and many other sense data resources.

**Human Generated Data:**

Human-generated data, we refer to the vast amount of social media data, status updates, tweets, photos, and media.

**Organizational Generated Data:**

Organisational generated data we refer to more traditional types of data, including transaction information in databases and structured data open stored in data warehouses.

**Some Facts:**

* Walmart handles more than **1 million** customer transactions every hour.
* Facebook stores, accesses, and analyzes **30+ Petabytes** of user generated data.
* **230+ millions** of tweets are created every day.
* More than **5 billion** people are calling, texting, tweeting and browsing on mobile phones worldwide.
* YouTube users upload **48 hours** of new video every minute of the day.
* Amazon handles **15 million** customer click stream user data per day to recommend products.
* **294 billion** emails are sent every day. Services analyses this data to find the spams.
* Modern cars have close to **100 sensors** which monitors fuel level, tire pressure etc. , each vehicle generates a lot of sensor data

**Conclusion:**

Real value often comes from combining these streams of data sources and analyzing them to generate new insights.

**5. Big Data; The V's Volume, Velocity, Variety..**

**The Volume, Velocity, Variety, Veracity, and Volume considered as the main characteristics of Big Data.**

Note: If you did not get to read the previous post in this series click the link : Series 1 Part 4 <https://www.abigdatablog.com/post/big-data-data-sources-and-types>

Diagrama

Descrição gerada automaticamente

**The introduction to V's in Big Data:**

The V’s are the characteristics of Big Data. The data keeps changing with time; when the data keeps evolving. The industry identified multiple V’s, and the following are some of them.

* Volume: The volume of the data is enormous.
* Velocity: Data is collected and accumulated at a very high speed.
* Variety: The data collected from multiple sources, and the data is in different formats.
* Veracity: There are inconsistencies and uncertainty due to the way the data is collected.
* Value: Identifying the value of the data in the massive pile of Big data is challenging.
* Validity: The correctness of data and validity of the sources
* Variability: The data keeps changing with time; information collected is dynamic by nature.
* Volatility: The tendency to change the data in time is very high.
* Vulnerability: The data is vulnerable to breach or attacks.
* Visualization: Collecting is one thing; visualizing meaningful usage of data is challenging.

**The main** [**#V**](https://www.abigdatablog.com/blog/hashtags/V)**’s are as follows:**

[**#VOLUME**](https://www.abigdatablog.com/blog/hashtags/VOLUME)

Volume refers to the amount of data and growth. The size of data generated by humans, machines, and their interactions on social sites is massive. Researchers have predicted that 40 Zetabytes (40,000 Exabytes).

[**#VELOCITY**](https://www.abigdatablog.com/blog/hashtags/VELOCITY)

Velocity defined as the pace at which different sources generate the data every day. The flow of data is massive and continuous. There are 1.03 billion Daily Active Users of Facebook. If you can handle the velocity, you will be able to generate insights and make decisions based on real-time data.

[**#VARIETY**](https://www.abigdatablog.com/blog/hashtags/VARIETY)

As many sources are contributing to Big Data, the type of data they are generating is different. It can be structured, semi-structured, or unstructured. Hence, there is a variety of data that is getting c every day. Earlier, we used to get the data from excel and database; now, the data are coming in the form of images, audios, videos, sensor data, etc. Hence, this variety of unstructured data creates problems in capturing, storage, data mining, and analyzing the data.

[**#VERACITY**](https://www.abigdatablog.com/blog/hashtags/VERACITY)

Veracity refers to uncertainty, inconsistency, and incompleteness in the data. This inconsistency and incompleteness is Veracity. Data available can sometimes get messy and may be challenging to trust. With many forms of big data, quality and accuracy are difficult to control, like Twitter posts with hashtags, abbreviations, typos, and everyday speech. The volume is often the reason behind the lack of quality and accuracy in the data.

Identified in a survey that 27% of respondents were unsure of how much of their data was inaccurate. Poor data quality costs the US economy around $3.1 trillion a year.

[**#VALUE**](https://www.abigdatablog.com/blog/hashtags/VALUE)

Having access to big data is an excellent opportunity to tap into future business opportunities, but unless we can turn data into profit, it is useless. How much benefit is the drawn data adding to the value to the organizations who are analyzing big data? Any [#ROI](https://www.abigdatablog.com/blog/hashtags/ROI) (Return On Investment)?

**Conclusion:**

Big data provides a reliable system and architecture to provide value to the company. But, the vulnerability is high due to the nature and the characteristics of the data collected in massive quantities.

**6. Big Data; Formats Structured, Unstructured, and Semi-structured**

Updated: Jan 25, 2020

**There are multiple source formats for** [**#BigData**](https://www.abigdatablog.com/blog/hashtags/BigData)**; Some of those formats are** [**#Structured**](https://www.abigdatablog.com/blog/hashtags/Structured)**,** [**#SemiStructured**](https://www.abigdatablog.com/blog/hashtags/SemiStructured)**,** [**#QuasiStructured**](https://www.abigdatablog.com/blog/hashtags/QuasiStructured)**, and** [**#UnStructured**](https://www.abigdatablog.com/blog/hashtags/UnStructured) **data.**

**Note**: If you did not get to read the previous post in this series click the link: Series 1 Part 5 <https://www.abigdatablog.com/post/series-1-part-5-big-data-the-v-s-volume-velocity-variety>

Diagrama

Descrição gerada automaticamente

**The introduction to source data formats in Big Data:**

In the [#BigData](https://www.abigdatablog.com/blog/hashtags/BigData) world, the data comes in different formats: Big Data is capable of handling beyond traditional structured data formats. The data sources like text, streaming, audio, video, and [#IoT](https://www.abigdatablog.com/blog/hashtags/IoT) changed the nature of the data collected for today's business.

**The following are the main data source formats in general for Big Data environment:**

**Structured Data:**

The Data that can be stored and processed in a fixed format called Structured Data. Data stored in a relational database management system RDBMS is one example of 'structured' data. It is easy to process structured data as it has a fixed schema. Structured Query Language (SQL) is often used to manage such kind of Data.

**Semi-Structured Data:**

Semi-Structured Data is a type of data that does not have a formal structure of a data model, i.e., a table definition in a relational [#DBMS](https://www.abigdatablog.com/blog/hashtags/DBMS). Still, nevertheless, it has some organizational properties like tags and other markers to separate semantic elements that make it easier to analyze. [#XML](https://www.abigdatablog.com/blog/hashtags/XML) files or [#JSON](https://www.abigdatablog.com/blog/hashtags/JSON) documents are examples of semi-structured data.

**Unstructured Data:**

The Data which have unknown form and cannot be stored in [#RDBMS](https://www.abigdatablog.com/blog/hashtags/RDBMS) and cannot be analyzed unless it transformed into a structured format called unstructured data. Text Files and multimedia contents like images, audios, videos are an example of unstructured data. The unstructured Data is growing quicker than others; experts say that 80 percent of the data in an organization is unstructured.

**Conclusion:**

Big Data sources data in partially structured and usually un-structured; because the sources of the data are XML, JSON, Text, Audio, Video, and sometimes RDBMS. The nature of the source makes Big Data challenging to collect, process, and store.

**7. Big Data; Challenges**

**It is crucial to have lots of data for the kind of usage the modern companies challenged by the business. But, it is vital to have clean and relevant data to arrive at conclusions.**

**Note**: If you did not get to read the previous post in this series click the link: Series 1 Part 6 <https://www.abigdatablog.com/post/series-1-part-6-big-data-formats-structured-unstructured-and-semi-structured>

Diagrama

Descrição gerada automaticamenteDiagrama

Descrição gerada automaticamente

**Introduction to Big data Challenges:**

Big data opened up to a world of data. One end industry is figuring out how to store and utilize the data; the other end, the industry realized there are multiple challenges with the vast amount of data.

**So, let us see what challenges the industry is facing with the discovery of Big data in regards to Security, Quality, Storage, Analysis, Talent, and Discovery.**

**Challenges with Big Data:**

**Data Quality:**

The data quality problem challenges from multiple ends, even without Big data. Now with Big data, the problem multiplied due to the Veracity. The data here is very messy, inconsistent, and incomplete. Dirty data cost $600 billion to companies every year in the United States.

**Data Discovery:**

Big data sources and stored petabytes of data and finding insight challenges the people working with the data. Big Data is like finding a needle in a haystack. So, analyzing the petabytes of data using powerful algorithms to find patterns and insights are very difficult.

**Data Storage:**

How to store data, always a challenge; Big Data brings massive amounts of data, and storing the data will make things worse in terms of storage. The more data an organization has, the more complex the problem. So, How to store the data? Where to store the data? And the type of storage system? As well, Big Data asks for scale up or down on-demand.

**Data Analytics:**

In the case of Big Data, most of the time, we are unaware of the kind of data we are dealing with one end. We are not sure about the reliability and relevance of the source of the data. Most of the external sources come with messy data, and cleaning such data becomes the priority before usage. So, performing the right analysis is a much tricky challenge.

**Data Security:**

Since the data is enormous, keeping it secure and sage challenges the security team, providing user access and user authentication, restricting access based on a user, recording data access histories, proper use of data encryption, etc.

**Data Professionals:**

There are a lot of Big Data projects in significant organizations. However, a sophisticated team of developers, data scientists, and analysts who also have a sufficient amount of domain knowledge is still a challenge.

**Conclusion:**

There are multiple challenges in the world of Big Data, but the industry made strides from problems and moving in a positive direction. Companies are handling challenges learning through experience and reaping the benefits.

**8. Big Data; Applications**

**Big data considered a buzz word in 2008. At the end of 2018, more than 90 percent of businesses planned to harness big data's growing power.**

**Note**: If you did not get to read the previous post in this series click the link: Series 1 Part 7 <https://www.abigdatablog.com/post/series-1-part-7-big-data-challenges>

Interface gráfica do usuário

Descrição gerada automaticamente

**Introduction to Big data Applications:**

Today, the impact of big data is immense, regardless of the industry. Big V's changed the way companies look at data today; today, the evidence is more measurable, facilitates reasoning, built a culture of analytics, and insights for data.

**Big Data Examples:**

**Consumer shopping habits**

* *Personalized marketing*
* *Amazon shopping*
* *ADs based on your online shopping or behavior.*
* *Facebook and Google ads*

**Geospatial information**

* *liveMaps*
* *Traffic information*
* *Identify mail packages in real-time.*
* *Inform delays using traffic, weather, and other factors.*

**Monitoring health**

* *using wearables*
* *Asthma attacks*
* *Heart issues*
* *Humidity air-quality*

**Autonomous vehicles**

* *Live maps for self-driving cars*
* *Sense surroundings and vehicles*
* *Reach the surrounding where human eyes cannot reach.*
* *warn about the lane closures*

**Media and data streaming**

**Predictive inventory ordering**

**Personalized health plans for patients**

**Real-time cybersecurity**

**Real-time fraud detection**

**Telecommunication Industry:**

Telecom sectors collect information, analyze it, and provide solutions to different problems. By using Big Data applications, telecom companies have been able to significantly reduce data packet loss, which occurs when networks are overloaded, and thus, providing a seamless connection to their customers.

**Retail Business:**

Retail has some of the tightest margins and is one of the greatest beneficiaries of big data. The beauty of using big data in retail is to understand consumer behavior. Amazon's recommendation engine provides a suggestion based on the browsing history of the consumer.

**Traffic control:**

Traffic congestion is a significant challenge for many cities globally. Effective use of data and sensors will be vital in managing traffic better as cities become increasingly densely populated.

**Manufacturing Industry:**

Analyzing big data in the manufacturing industry can reduce component defects, improve product quality, increase efficiency, and save time and money.

**Search Engines:**

Every time we are extracting information from google, we are simultaneously generating data for it. Google stores this data and uses it to improve its search quality.

**Conclusion:**

Today every industry leveraging Big Data applications in multiple ways.