Big Data

M. H. Rahman

Associate Professor
Department of Statistics and Data Science
Jahangirnagar University
Bangladesh
E-mail: habib.drj@juniv.edu

Twenty-Fifty

Big Data

Course Code: Stat-510 and WMASDS-19
Course Title: Big Data
Department of Statistics and Data Science
Jahangirnagar University

Name of the Topics of this Course

- Introduction of Big Data
- Big Data Related Technologies
- Data Collection, Sampling, and Preprocessing
- Big Data Storage
- Big Data Architecture
- Big Data Processing Algorithms
- Big Data Search and Mining
- Big Data Applications

Book References

Texts

1. Min Chen, Shiwen Mao, Yin Zhang and Victor C.M. Leung, (2014), Big Data: Related Technologies, Challenges and Future Prospects, Springer.
2. Hrushikesha Mohanty, Prachet Bhuyan and Deepak Chenthati, (2015), Big Data a Primer, Springer.
3. Bart Baesens, (2014), Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, Wiley

References

1. Zikopoulos, P.C., Eaton, C., Deroos, D., Deutsch, T. and Lapis, G., (2012), Understanding Big Data, McGraw-Hill, New York.

Syllabus for WM-ASDS 19 Big Data

Introduction of Big Data: Dawn of the Big Data Era, Definition and Features of Big Data, Big Data Value, the Development of Big Data, Challenges of Big Data.

Big Data Related Technologies: Cloud Computing, Relationship between Cloud Computing and Iot, Relationship between Iot and Big Data, Data Center.

Data Collection, Sampling and Preprocessing: Types of Data Sources, Sampling, Types of Data Elements, Big Data Generation, Enterprise Data, Iot Data, Internet Data, Bio-Medical Data, Data Collection, Data Transportation, Data Pre-Processing, Visual Data Exploration and Exploratory Statistical Analysis.

Big Data Storage: Storage System for Massive Data, Distributed Storage System, Storage Mechanism for Big Data, Database Technology, Design Factors, Database Programming Model.

Syllabus for WM-ASDS 19 Big Data

Big Data Processing Algorithms: Multi-Core Versus Distributed Systems, Distributed Algorithms, Distributed Hash Tables, Bulk Synchronous Parallel (Bsp), Mapreduce Paradigm, Input Reader and output Writers, Putting all Together.

Big Data Search and Mining: Big Data Search and Retrieval, K-Means Clustering, Social Network Clustering—Topology Discovery, Clustering Algorithm to find Network Topologies, Social Network Condensation, Text Sentiment Mining, Big Data Mining and Analysis Tools.

Big Data Applications: Structured Data Analysis, Text Data Analysis, Web Data Analysis, Multimedia Data Analysis, Network Data Analysis, Mobile Traffic Analysis, Application of Big Data in Enterprises, Application of Online Social Network-Oriented Big Data, Applications of Healthcare and Medical Big Data.

Assignment andor a mini project to be completed on the basis of the above topics by SPSS and or SASR.

Syllabus for WM-ASDS 19 Big Data

Marks Distribution for WMASDS-19 Big Data

Description	Marks
Class Attendance	05
Class Performance	05
Quiz/Class Test	10
Assignment/Short Report/Presentation	10
Mid-term Examination	30
Final Examination	40
Total	**100**

Syllabus for Stat-510 Big Data

Introduction of Big Data: Dawn of the Big Data Era, Definition and Features of Big Data, Big Data Value, the Development of Big Data, Challenges of Big Data

Big Data Related Technologies: Cloud Computing, Relationship between Cloud Computing and Big, Iot, Relationship between IoT and Big Data, Data Center, Hadoop, Relationship between Hadoop and Big Data.

Data Collection, Sampling and Preprocessing: Types of Data Sources, Sampling, Types of Data Elements, Big Data Generation, Enterprise Data, Iot Data, Internet Data, Bio-Medical Data, Data Generation from other Fields, Big Data Acquisition, Data Collection, Data Transportation, Data Pre-Processing, Visual Data Exploration And Exploratory Statistical Analysis, Missing Values, Outlier Detection and Treatment, Standardizing Data, Categorization, Weights of Evidence Coding, Variable Selection, Segmentation.

Syllabus for Stat-510 Big Data

Big Data Storage: Storage System for Massive Data, Distributed Storage System, Storage Mechanism for Big Data, Database Technology, Design Factors, Database Programming Model.

Big Data Architecture: Space of Big Data, Characteristics of Big Data, Data-Driven Decision Making, Deriving Value from Data, Data R&D—the Fertile Ground for Innovation, Building the Data Architecture, Putting it all Together, Architecture Futures.

Big Data Processing Algorithms: Multi-Core Versus Distributed Systems, Distributed Algorithms, Distributed Hash Tables, Bulk Synchronous Parallel (Bsp), MapReduce Paradigm, Hdfs, MapReduce Computing Platform, Job Tracker, Task Trackers, Yarn, Partitioners and Combiners, Input Reader and output Writers, Putting all Together, Hadoop Streaming, Distributed Cache, Multiple Outputs, Iterative MapReduce, Machine Learning with MapReduce.

Syllabus for Stat-510 Big Data

Big Data Search and Mining: Big Data Search and Retrieval, K-Means Clustering, Social Network Clustering—Topology Discovery, Clustering Algorithm to find Network Topologies, Social Network Condensation, Text Sentiment Mining, Big Data Mining and Analysis Tools.

Big Data Applications: Application Evolution, Big Data Analysis Fields, Structured Data Analysis, Text Data Analysis, Web Data Analysis, Multimedia Data Analysis, Network Data Analysis, Mobile Traffic Analysis, Application of Big Data in Enterprises, Application of Iot Based Big Data, Application of Online Social Network-Oriented Big Data, Applications of Healthcare and Medical Big Data, Collective Intelligence, Smart Grid.

Syllabus for Stat-510 Big Data

Text Books:

1. Min Chen, Shiwen Mao, Yin Zhang and Victor C.M. Leung, (2014), Big Data: Related Technologies, Challenges and Future Prospects, Springer.

2. Hrushikesha Mohanty, Prachet Bhuyan and Deepak Chenthati, (2015), Big Data a Primer, Springer.

3. Bart Baesens, (2014), Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, Wiley

Reference

1. Zikopoulos, P.C., Eaton, C., Deroos, D., Deutsch, T. and Lapis, G., (2012), Understanding Big Data, Mcgrawhill, New York.

Marks Distribution for Stat-510 Big Data

Description	Marks
Class Attendance	05
Tutorial Examination	10
Final Examination	35
Total	**50**

Chapter 1 Big Data

Big Data

What is Big Data?

Big data is an abstract concept. In general, big data refers to the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time.

In 2010, Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope.

Big data shall mean datasets that classic database software could not acquire, store, and manage. This definition includes two connotations: First, the dataset volumes that conform to the standard of big data are changing, and may grow over time or with technological advances;
Second, the dataset volumes that conform to the standard of big data in different applications differ from each other.

Big Data

Big data has been defined as early as 2001. Doug Laney was an analyst of 'META' and 'Gartner' defined challenges and opportunities brought about by the increased data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety, in a research report.

Although such a model was not originally used to define big data, Gartner and many other enterprises, including IBM and some research departments of Microsoft still used the "3Vs" model to describe big data within the following 10 years.
In the "3Vs" model, Volume means, with the generation and collection of massive data, data scale becomes increasingly huge; Velocity means the timeliness of big data, specifically, data collection and analysis, etc., must be rapidly and timely conducted, so as to maximumly utilize the commercial value of big data; Variety indicates the various types of data, which include semi-structured and unstructured data such as **audio, video, webpage, and text, as well as traditional structured data**.

Big Data

International Data Corporation (IDC) is one of the most influential leaders in big data and its research fields. In 2011, an IDC report defined big data as "big data technologies describes a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and analysis".

With this definition, characteristics of big data can be summarized as **four Vs**, i.e., Volume (great volume), Variety (various modalities), Velocity (rapid generation), and Value (huge value but very low density).

Big Data

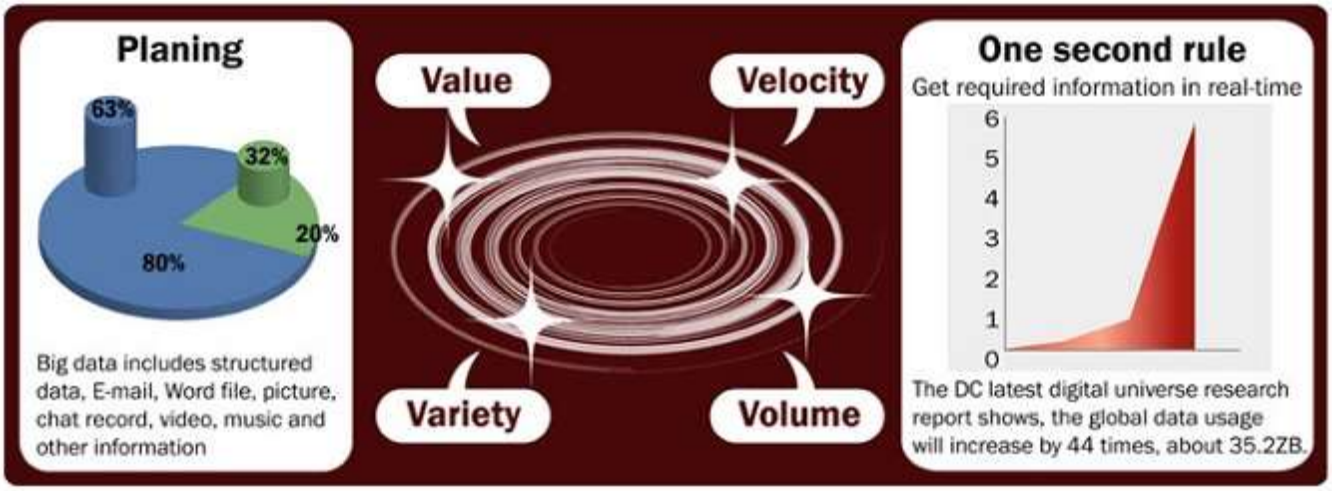


Figure: The 4Vs feature of big data

Big Data

In addition, the US National Institute of Standards and Technology (NIST) defines big data as "Big data shall mean the data of which the data volume, acquisition speed or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies," which focuses on the technological aspect of big data. It indicates that efficient methods or technologies need to be developed and used to analyze and process big data.

Big Data

Value of Big Data

Through research on the five core industries that represent the global economy, the McKinsey report pointed out that big data may give full play to the economic function, improve the productivity and competitiveness of enterprises and public sectors, and create huge benefits for consumers.

Development of Big Data

In the late **1970s**, the concept of "database machine" emerged, which is a technology specially used for storing and analyzing data. With the increase in data volume, the storage and processing capacity of a single mainframe computer system has become inadequate.

In the **1980s**, people proposed "share nothing," a parallel database system, to meet the demand of the increasing data volume. The share-nothing system architecture is based on the use of cluster and every machine has its own processor, storage, and disk. **Teradata** system was the first successful commercial parallel database system.

Big Data

Disk space, or data storage space

1 Bit	Binary Digit
8 Bits	1 Byte
1024 Bytes	1 Kilobyte
1024 Kilobytes	1 Megabyte
1024 Megabytes	1 Gigabyte
1024 Gigabytes	1 Terabyte
1024 Terabytes	1 Petabyte
1024 Petabytes	1 Exabyte
1024 Exabytes	1 Zettabyte
1024 Zettabytes	1 Yottabyte
1024 Yottabytes	1 Brontobyte
1024 Brontobytes	1 Geopbyte

Table: Kilo, mega, giga, tera, peta, exa, zetta and all that.

Big Data

However, many challenges on big data arose. With the development of Internet services, indexes and queried contents were rapidly growing. Therefore, search engine companies had to face the challenges of handling such big data. Google created Google File System (GFS) and MapReduce programming models to cope with the challenges brought about by data management and analysis at the Internet scale.
GFS is a distributed system to be run on clusters.
Over the past few years, nearly all major companies, including EMC, Oracle, IBM, Microsoft, Google, Amazon, Facebook, etc., have started their big data projects. Taking IBM as an example, since 2005, IBM has invested USD 16 billion on 30 acquisitions related to big data. In academia, big data was also under the spotlight. In 2008, Nature published the big data special issue At the beginning of 2012, a report titled Big Data, Big Impact presented at the Davos Forum in Switzerland, announced that big data has become a new kind of economic assets, just like currency or gold.

What is MapReduce?
It is a simple methodology to process large-sized data by distributing it across a large number of servers or nodes. The master node will first partition the input into smaller subproblems which are then distributed to the slave nodes which process the portion of the problem that they receive. This step is known as the Map step.

In the Reduce step, a master node takes the answers from all the subproblems and combines them in such a way as to get the output that solves the given application problem. Such parallel processing requires that there are no dependencies in the data.

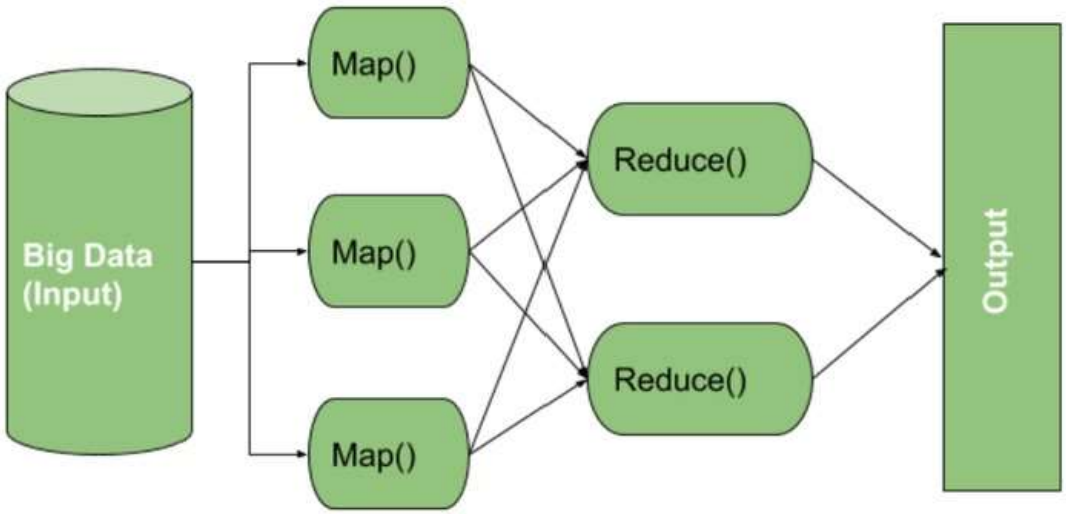


Figure: Map Reduce

Big Data

For example, if daily temperature data in different locations in different months is required to be processed to find out the maximum temperature among all of them, the data for each location for each month can be processed parallelly, and finally the maximum temperature for all the given locations can be combined together to find out the global maximum temperature. The first phase of sending different locations of data to different nodes is called Map Phase, and the final step of integrating all the results received from different nodes into the final answer is called Reduce Phase.

Map Reduce: It is a framework in which we can write applications to run huge amounts of data in parallel and in large clusters of commodity hardware in a reliable manner. Different Phases of MapReduce: The MapReduce model has three major and one optional phase.
Mapping, Shuffling and Sorting, Reducing, and Combining

Big Data

Google File System
Google developed the Google File System (GFS), a scalable distributed file system (DFS), to meet the company's growing data processing needs. GFS offers fault tolerance, dependability, scalability, availability, and performance to big networks and connected nodes. GFS is made up of a number of storage systems constructed from inexpensive commodity hardware parts. The search engine, which creates enormous volumes of data that must be kept, is only one example of how it is customized to meet Google's various data use and storage requirements.

More than 1,000 nodes with 300 TB of disc storage capacity make up the largest GFS clusters. This is available for constant access by hundreds of clients.

Components of GFS GFS Clients, GFS Master Server, GFS Chunk Servers.

Big Data

Challenges of Big Data

- Data Representation
- Redundancy Reduction and Data Compression
- Data Life Cycle Management
- Analytical Mechanism
- Data Confidentiality
- Energy Management
- Expendability and Scalability
- Cooperation

V's of Big Data

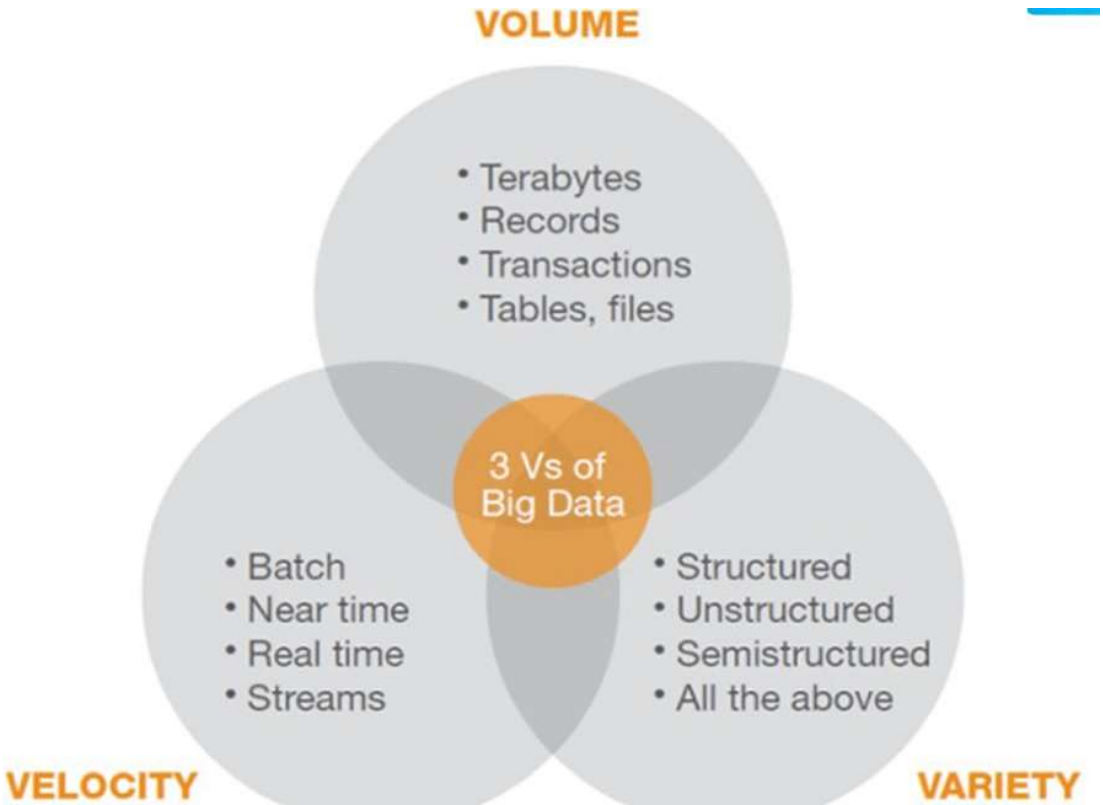


Figure: 3V's of Big Data

V's of Big Data

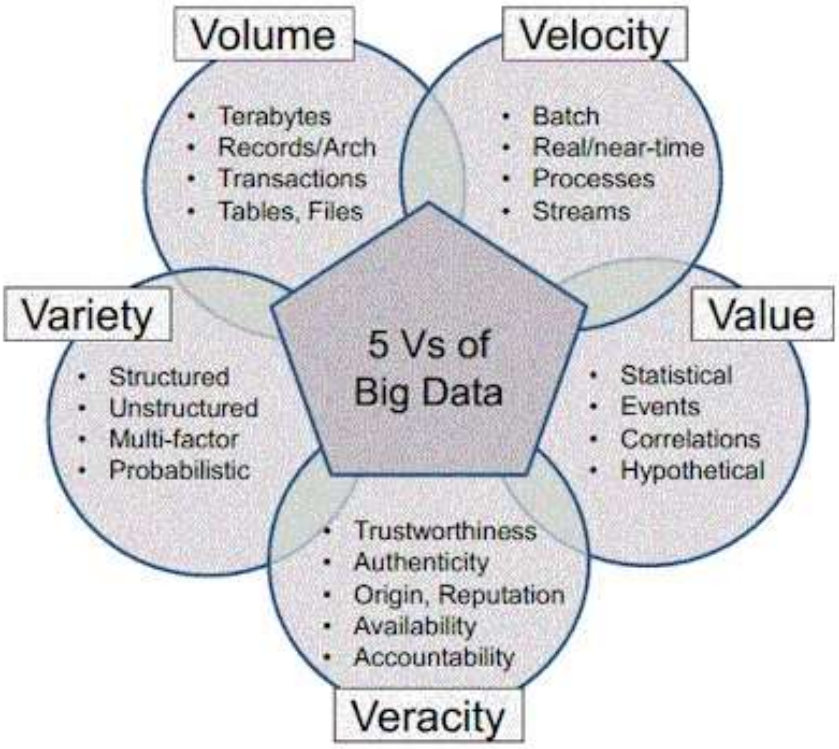


Figure: 5V's of Big Data

V's of Big Data

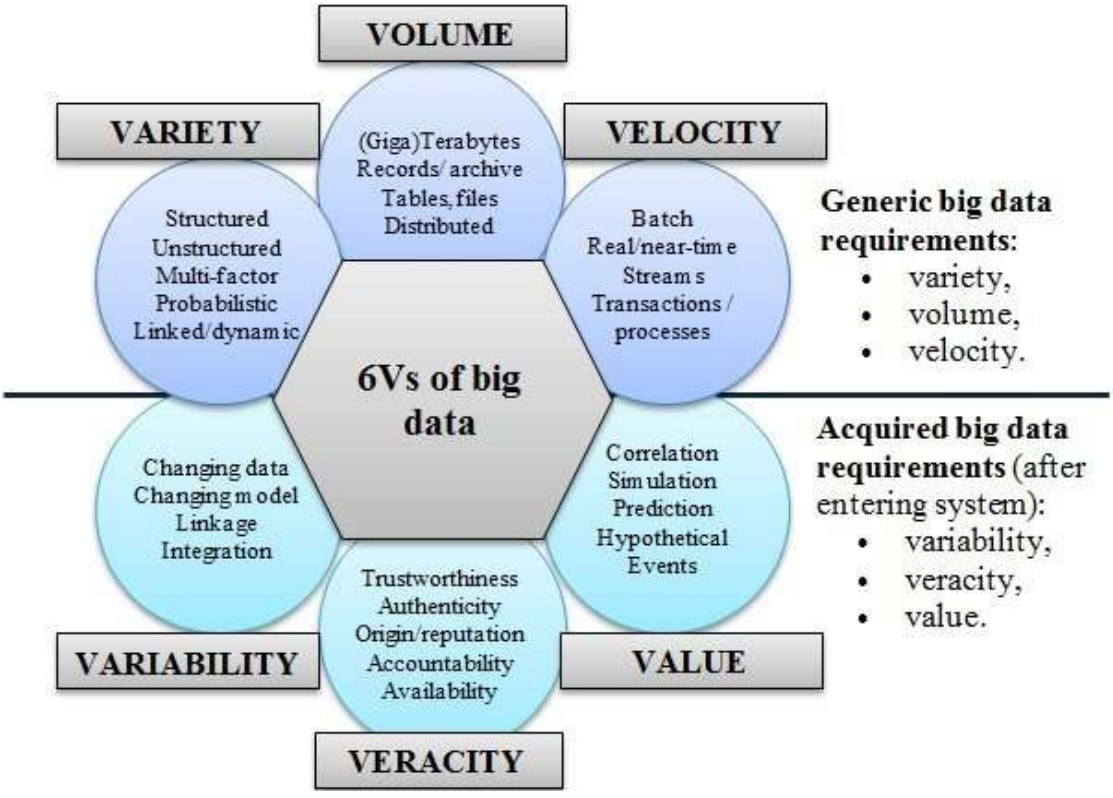
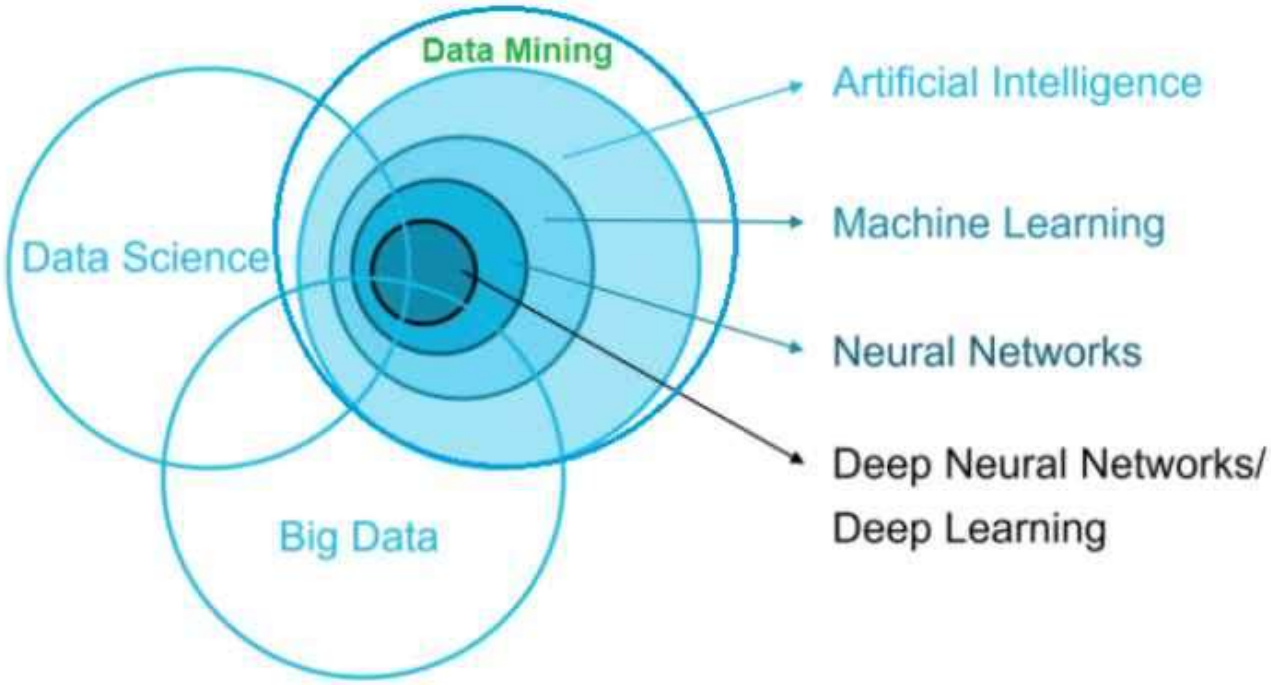


Figure: 6V's of Big Data

Data Mining vs Big Data vs Data Science



Big Data

Difference between Traditional Data and Big Data

Parameters	Traditional Data	Big Data
Structure of data	Structures are defined	Mix of Structured, semi-structured and unstructured data
Data Volume	Based on business volumes and extent of digitization	Very high, in petabytes and even more
Variety of Data Sources	Data source from database systems	Besides data from business information systems, text (emails, documents), weblogs, sensors, RFID, etc.
Velocity	Low to moderate based on volume of business	High velocity
Flow	Fixed	Continuous round the clock accumulation of data
Structured Data	Structured Data	Structured, Semi-structured and Unstructured data
Sources of data	Organizational data, trading partners data	Organizational data, RFID, Sensor data, Google searches, Social media (Linked in, Facebook, Twitter, Whatsapp, etc.)
Analytics	Provide historical view, status reports	Real-time, direct feedback from the consumer, sentiment analysis, opinions
Functions	Advises senior executives on internal business decisions, focused on analyzing data for	Customer facing functions get direct market feedback which can be used for planning market strategies, planning etc.,

Big Data

Difference between Traditional Data and Big Data

Traditional Data	Big Data
Traditional data is generated at the enterprise level.	Big data is generated outside the enterprise level.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Traditional database system deals with structured data.	Big data system deals with structured, semi-structured, database, and unstructured data.
Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per second.
Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.

Big Data

Difference between Traditional Data and Big Data....

Traditional Data	Big Data
Data integration is very easy.	Data integration is very difficult.
Normal system configuration is capable of processing traditional data.	High system configuration is required to process big data.
Traditional database tools are required to perform any database operation.	Special kinds of database tools are required to perform any database schema-based operation.
Normal functions can manipulate data.	Special kinds of functions can manipulate data.
Its data sources include ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data, etc.	Its data sources include social media, device data, sensor data, video, images, audio, etc.

Statistics vs Data Mining vs Big Data

	Statistics	Data Mining	Big Data
Structure	structured	structured	unstructured
Size	small	large	very large
Generation	planned	transactional	behavioral
Aim	understand	optimize business	generate business
Privacy Issues	non	minor	huge
Founded On	concepts & theory	technology & tool	technology & tools
Marketing	bad	good	perfect

https://www.quotemaster.org/images/8b/8b46fc588072674a0f1ca7483d
png