Basic Concepts of Big Data (20 Hours)

- **Session 1-4:**

o Concept and characteristics of Big Data

o History of Big Data

o Jobs in Big Data

o Types of Big data (structured, semi-structured, unstructured)

- **Session 5-9:**

o Big Data Frameworks

o Big Data Programming Paradigms

o Big Data Programming Languages

- **Session 10-11:**

o Introduction to Data Science and Skillset required for working with Big Data

- **Session 12-15:**

o Simplified Overview of Machine Learning Algorithms and Neural Networks

o Types of Machine Learning (Supervised, Un-Supervised, Reinforcement)

- **Session 16-18:**

ACTS, Head Quarters, Pune

o Examples of Big Data and Data Science in Practice (Healthcare, Logistics & Transportation, Manufacturing etc.

- **Session 19-20:**

o Application Examples and Real –World Use Cases (e.g., Healthcare, finance, marketing, etc.)

Types of Big Data Technologies (+ Management Tools)

1. Data storage

- Apache Hadoop
- Apache Spark
- Apache Hive
- Apache Flume
- ElasticSearch
- MongoDB

2. Data mining

- Rapidminer
- Presto

3. Data analytics

- Apache Spark
- Splunk
- KNIME

4. Data visualization

- Tableau
- Power BI

The picture can't be displayed.

Apache Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

MongoDB

MongoDB is a source-available cross-platform document-oriented database program.

Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas

The picture can't be displayed.

Rapidminer

pros:

1. Multiple deployment options based on our preference.
2. Strong visualization.
3. Accurate Preprocessing.
4. Multiple interfaces.
5. Java API available that can be used in programs.

cons:

1. It takes too much memory and so slows down your system.
2. Less forums for support.

Presto

A single Presto query can process data from multiple sources like HDFS, MySQL, Cassandra, Hive and many more data sources. Presto is built in Java and easy to integrate with other data infrastructure components. Presto is powerful, and leading companies like Airbnb, DropBox, Groupon, Netflix are adopting it.

Apache Spark

Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance

Splunk

Splunk is a program that enables the search and analysis of computer data. It analyzes semi-structured data and logs generated by various processes with proper data modeling as per the need of the IT companies. The user produces the data by means of any device like- web apps, sensors, or computers. It has built-in functionality for defining data types, field separators, and search process optimization. For the searched result, it also provides visualization of data.

Tableau can handle huge columns of data and still offer better performance.	Power BI is best for a limited volume of data.
Tableau has better data visualization.	Power BI offers many data points for data visualization.
Tableau works best with huge data.	Power BI is suboptimal with huge data.
Experts and experienced users use Tableau.	Power BI is used by beginners and experienced alike.

1. Airflow
2. Delta Lake
3. Drill
4. Druid
5. Flink
6. Hadoop
7. Hive
8. HPCC Systems
9. Hudi
10. Iceberg
11. Kafka
12. Kylin
13. Pinot
14. Presto
15. Samza
16. Spark
17. Storm
18. Trino
19. MongoDB

1. Data storage

Apache Hadoop

Hortonworks

Data lake

Apache Spark

MongoDB

Apache Cassandra

Presto

Cloudera

Cloud storage

Elastic search

Hybrid storage

Cloud Service Providers

Microsoft Azure

IBM Cloud Services

Adobe Creative Cloud

Kamatera

VMware

Google Cloud Platform

Rackspace

Red Hat

Salesforce

Amazon Web Service (AWS)

Oracle Cloud

SAP

Verizon Cloud

1. Airflow

Airflow in Apache is a popularly used tool to manage the automation of tasks and their workflows. They are also primarily used for scheduling various tasks. Consider that you are working as a data engineer or an analyst and we might need to continuously repeat a task that needs the same effort and time every time. The kind of such tasks might consist of **extracting, loading,** or **transforming data** that need a regular analytical report. We can simply automate such tasks using Airflow in Apache by training your machine learning model to serve these kinds of tasks on a regular interval specified while training it

2. Delta Lake

What is Big Data

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

Sources of Big Data

These data come from many sources like

•**Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.

•**E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.

•**Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.

•**Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.

•**Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

3V's of Big Data

1.Velocity: The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.

2.Variety: Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.

3.Volume: The amount of data which we deal with is of very large size of Peta bytes.

Use case

An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100$ to its top 10 customers who have spent the most in the previous year.Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.

Issues

Huge amount of unstructured data which needs to be stored, processed and analyzed.

Solution

Storage: This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

Processing: Map Reduce paradigm is applied to data distributed over network to find the required output.

Analyze: Pig, Hive can be used to analyze the data.

Cost: Hadoop is open source so the cost is no more an issue.

What is Hadoop

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

Modules of Hadoop

1.HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.

2.Yarn: Yet another Resource Negotiator is used for job scheduling and manage the cluster.

3.Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.

4.Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

Hadoop Architecture

The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.

Using Big Data Analytics Tools

1.Healthcare: Big data analytics technologies and tools are being used in healthcare to predict patient outcomes, identify at-risk patients, and improve population health.

2.Retail: Big data analytics tools are being used by retailers to improve customer experience, target marketing campaigns, and prevent fraud.

3.Manufacturing: Big data analytics tools are being used in manufacturing to improve quality control, reduce downtime, and optimize production processes.

4.Banking: Real time big data analytics tools are being used by banks to detect fraudulent activities, prevent money laundering, and improve customer service.

1.Government: Big data analytics tools are being used by government agencies to improve public services, combat fraud and corruption, and better understand citizen needs.

Limitations of Big Data Analytics Tools

There are several limitations to big data analytics tools, including:

1.They can be expensive and require a lot of resources to implement.

2.They can be complex to use and require skilled staff to get the most out of them.

3.They can require a lot of data to be effective, which can be a challenge to collect.

4.They can be slow and may not be able to keep up with rapidly changing data.

5.They can produce biased results, depending on how they are configured.

Best tool for big data analytics?

Big Data frameworks such as Apache Hadoop are widely used in the market. Clusters of computers can be used to process massive data sets using Hadoop. Scaling up from one server to tens of thousands of commodity computers is one of the best features of this Big Data Tool.

Five types of big data analytics?

The five types of big data analytics are as follows:

Cyber Analytics
Prescriptive Analytics
Descriptive Analytics
Diagnostic Analytics
Predictive Analytics

Data engineering

Data engineering is the process of building robust data architecture that allows for data processing. This includes data transfers between databases and building data warehouses for easy accessibility.

Through data engineering, the following question is answered: “How do I make all the data we collect easier for our data analysts and other stakeholders to wade through?” Data engineering makes the data more reliable, accurate, and ingestible through robust data processing systems.

Data Engineer and a Data Analyst?

The picture can't be displayed.

Data Engineering

The picture can't be displayed.

Requirements To Become a Data Engineer

The picture can't be displayed.

Requirements To Become a Data Analyst

The picture can't be displayed.

Data Analyst, Data Engineer and Data Scientist

The picture can't be displayed.

1. **Data in ____ bytes size is called Big Data.**

Tera
Giga
Peta
Meta

2. **How many V's of Big Data?**

2
3
4
5
Answer: D) 5
Volume, Velocity, Variety, Value and Veracity

3. Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media.

True
False

4. In Big Data environments, Velocity refers –

Data can arrive at fast speed

Enormous datasets can accumulate within very short periods of time

Velocity of data translates into the amount of time it takes for the data to be processed

All of the mentioned above

5. In Big Data environments, Variety of data includes –

Includes multiple formats and types of data

Includes structured data in the form of financial transactions,

Includes semi-structured data in the form of emails and unstructured data in the form of images

All of the mentioned above

6. Which of the following are Benefits of Big Data Processing?

Cost Reduction

Time Reductions

Smarter Business Decisions

All of the mentioned above

7. Data that does not conform to a data model or data schema is known as ______.

Structured data

Unstructured data

Semi-structured data

All of the mentioned above

8. Amongst which of the following is/are not Big Data Technologies?

Apache Hadoop
Apache Spark
Apache Kafka
Apache Pytarch

9.______ involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

Parallel data processing

Single channel processing

Multi data processing

None of the mentioned above

10 Amongst which of the following can be considered as the main source of unstructured data.

Twitter

Facebook

Webpages

All of the mentioned above

11. Amongst which of the following shows an example of unstructured data,

Students roll number, age

Videos

Audio files

Both B and C

12 Scalability, elasticity, resource pooling, self-service, low cost and fault tolerance are the features of,

Cloud computing

Power BI

System development

None of the mentioned above

Big data is a field that treats ways to analyze and systematically extract information from or otherwise deal with data sets. Data can be large or complex to be dealt with by traditional data processing applications software

A large amount of data
It is a popular term used to express the exponential growth of data.
Big data is difficult to store, collect, maintain, analyze and visualize.

Distributed file system: A distributed file system is a file system in which data is stored on a server. The data is accessed and processed as if it were stored on the local client machine. The following are the Characteristics of distributed file system:

Transparency
user mobility
Performance
simplicity and ease of use
Scalability
high availability
high reliability
Big data tools: Apache Hadoop, Apache Storm, Cassandra, Mongo DB, Neo4j. Learn More.

Big data sources:

Amazon, Redshift, Mongo DB

Challenges of big data:

Uncertainty of data management
The talent gap in big data
Getting data into a big data structure
Synchronizing across data sources
Integration

Benefits of big data:

Cost
Time reduction
Speeding up decision-making
Analyze in real-time
Model and Test variation

Characteristics of big data:

Volume
Velocity
Variety

Types of big data:

Structured
unstructured
Semi-structured
hybrid

Use cases of big data:

Recommendation engine
Analyzing call detail records
Fraud detection
sentiment analysis

What is Structured Data?

Structured data is information that has been formatted and transformed into a well-defined data model. The raw data is mapped into predesigned fields that can then be extracted and read through SQL easily. SQL relational databases, consisting of tables with rows and columns, are the perfect example of structured data.

The relational model of this data format utilizes memory since it minimizes data redundancy. However, this also means that structured data is more inter-dependent and less flexible. Now let's look at more examples of structured data.

Examples of Structured Data

This type of data is generated by both humans and machines. There are numerous examples of structured data from machines, such as POS data like quantity, barcodes, and weblog statistics. Similarly, anyone who works on data would have used spreadsheets once in their lifetime, which is a classic case of structured data generated by humans. Due to the organization of structured data, it is easier to analyze than both semi-structured and unstructured data.

What is Semi-Structured Data

We may not always find your data sets to be structured or unstructured. Semi-structured data or partially structured data is another category between structured and unstructured data. Semi-structured data is a type of data that has some consistent and definite characteristics.

It does not confine into a rigid structure such as that needed for relational databases. Businesses use organizational properties like metadata or semantics tags with semi-structured data to make it more manageable. However, it still contains some variability and inconsistency.

Examples of Semi-Structured Data

An example of data in a semi-structured format is delimited files. It contains elements that can break down the data into separate hierarchies. Similarly, in digital photographs, the image does not have a pre-defined structure itself but has certain structural attributes making them semi-structured. F or instance, if you take a photo from a smartphone, it would have some structured attributes like geotag, device ID, and DateTime stamp. After you save them, we can assign tags to images such as ‘pet’ or ‘dog’ to provide a structure. On some occasions, unstructured data is classified as semi-structured data because it has one or more classifying attributes.

What is Unstructured Data?

Unstructured data is defined as data present in absolute raw form. This data is difficult to process due to its complex arrangement and formatting.

Unstructured data includes social media posts, chats, satellite imagery, IoT sensor data, emails, and presentations. Unstructured data management takes this data to organize it in a logical, predefined manner in data storage. Natural language processing (NLP) tools help understand unstructured data that exists in a written format.

In contrast, the meaning of structured data is data that follows predefined data models and is easy to analyze. Structured data examples would include alphabetically arranged names of customers and properly organized credit card numbers. After understanding the definition of unstructured data, let’s look at some examples.

Big Data includes huge volume, high velocity, and extensible variety of data. There are 3 types: Structured data, Semi-structured data, and Unstructured data.

Structured data –

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.

Semi-Structured data –

Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.

Unstructured data –

Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

Examples of Unstructured Data

Unstructured data can be anything that’s not in a specific format. This can be a paragraph from a book with relevant information or a web page. An example of unstructured data could also be Log files that are not easy to separate. Social media comments and posts are also unstructured.

Here is an example of unstructured data from a log file.

```
38,P-R-38636-6-45,P-R-39105-1-11,P-R-38036-1-5,P-R-35697-1-13,P-R-35087-1-27,P-R-34341-1-9,P-R-33341-1-15,P-R-33110-1-29,P-R-31345-1-
693,P-R-29076-1-6,P-R-28767-1-8,P-R-28540-2-8,P-R-28312-1-10,P-R-28069-1-27,P-R-28032-1-9,P-R-26562-1-12,P-R-26527-5-20,P-R-26164-1-
11,P-R-25785-1-30,P-R-25095-9-70,P-R-23504-1-15,P-R-19719-5-41203
Wed Sep 23 2020 05:21:01 GMT+0500
```

Unstructured data is qualitative, not quantitative, so it is mostly categorical and characteristic in nature. For example, data from social media or websites can help predict future buying trends or determine the effectiveness of a marketing campaign. Another unstructured data analytics example is detecting patterns in scam emails and chat, which can be useful for enterprises in monitoring policy compliance. That’s why businesses extract and store unstructured data in data warehouses (also called data lakes) for analysis.

1) Artificial Intelligence is about_____.

1.Playing a game on Computer
2.Making a machine Intelligent
3.Programming on Machine with your Own Intelligence
4.Putting your intelligence in Machine

2) Who is known as the -Father of AI"?

1.Fisher Ada
2.Alan Turing
3.John McCarthy
4.Allen Newell

3)Select the most appropriate situation for that a blind search can be used.

1.Real-life situation
2.Small Search Space
3.Complex game
4. All of the above

4. Ways to achieve AI in real-life are_________.

1.Machine Learning
2.Deep Learning
3.Both a & b
4.None of the abov

5. The main tasks of an AI agent are_______.

1.Input and Output
2.Moment and Humanly Actions
3.Perceiving, thinking, and acting on the environment
4.None of the above

Q.6 The best AI agent is one which____________

1.Needs user inputs for solving any problem
2.Can solve a problem on its own without any human intervention
3.Need a similar exemplary problem in its knowledge base
4.All of the above

Q.7 Which of the given element improve the performance of AI agent so that it can make better decisions?

1.Changing Element
2.Performance Element
3.Learning Element
4.None of the above

Q.8 How many types of Machine Learning are there?

1.1
2.2
3.3
4.4

Q.1 Select the type of data that can be Structured easily ?

Date Of Birth
Profile Photo
Screenshots
directions to the shops

Q.2 Select the unstructured data

Name
shipping time
Price of product
Product description

Q.3 Unstructured data can come from which of the following?

Facebook
Twitter
Presentations
All of these are corr

Q.4 What is structured data?

Structured data is a type of data that is huge in number and has many inaccurate values
Structured data is a type of data that is very less in number and can be stored in proper rows and columns
Structured data is a type of data that has inaccurate values but can be stored in rows and columns

Q.5 An example of structured data is _____.

age information
reason for a customer complaint
customer reviews
pictures of the good/serviceect.

Q.6 What is unstructured data?

Unstructured data is a type of data that is huge in number and has many inaccurate values
Unstructured data is a type of data that is very less in number and can be stored in proper rows and columns
Unstructured data is a type of data that has inaccurate values but can be stored in rows and columns

Q.6 What is semi-structured data?

Semi-structured data is a type of data that is huge in number and has many inaccurate values
Semi-structured data is a type of data that is very less in number and can be stored in proper rows and columns
Semi-structured data is a type of data that has inaccurate values but can be stored in rows and columns
Semi-structured data is a type of data which has contained the data of both types i.e., structured data and semi-structured data

Big Data Frameworks

. Hadoop

There are four components in the Hadoop ecosystem

HDFS: Stands for Hadoop Distributed File System, a file system that stores data on computers in a cluster. In simple words, it is a storage unit of Hadoop.

YARN: An acronym for Yet Another Resource Negotiator, a resource manager. It manages all computing resources in clusters and uses them to schedule user applications.

M**apReduce:** A programming model for processing data.

Hadoop Common: Hosts libraries and utilities and provide them to the above Hadoop components as required.

HDFS is the distributed file system in Hadoop for storing big data. MapReduce is the processing framework for processing vast data in the Hadoop cluster in a distributed manner. YARN is responsible for managing the resources amongst applications in the cluster.

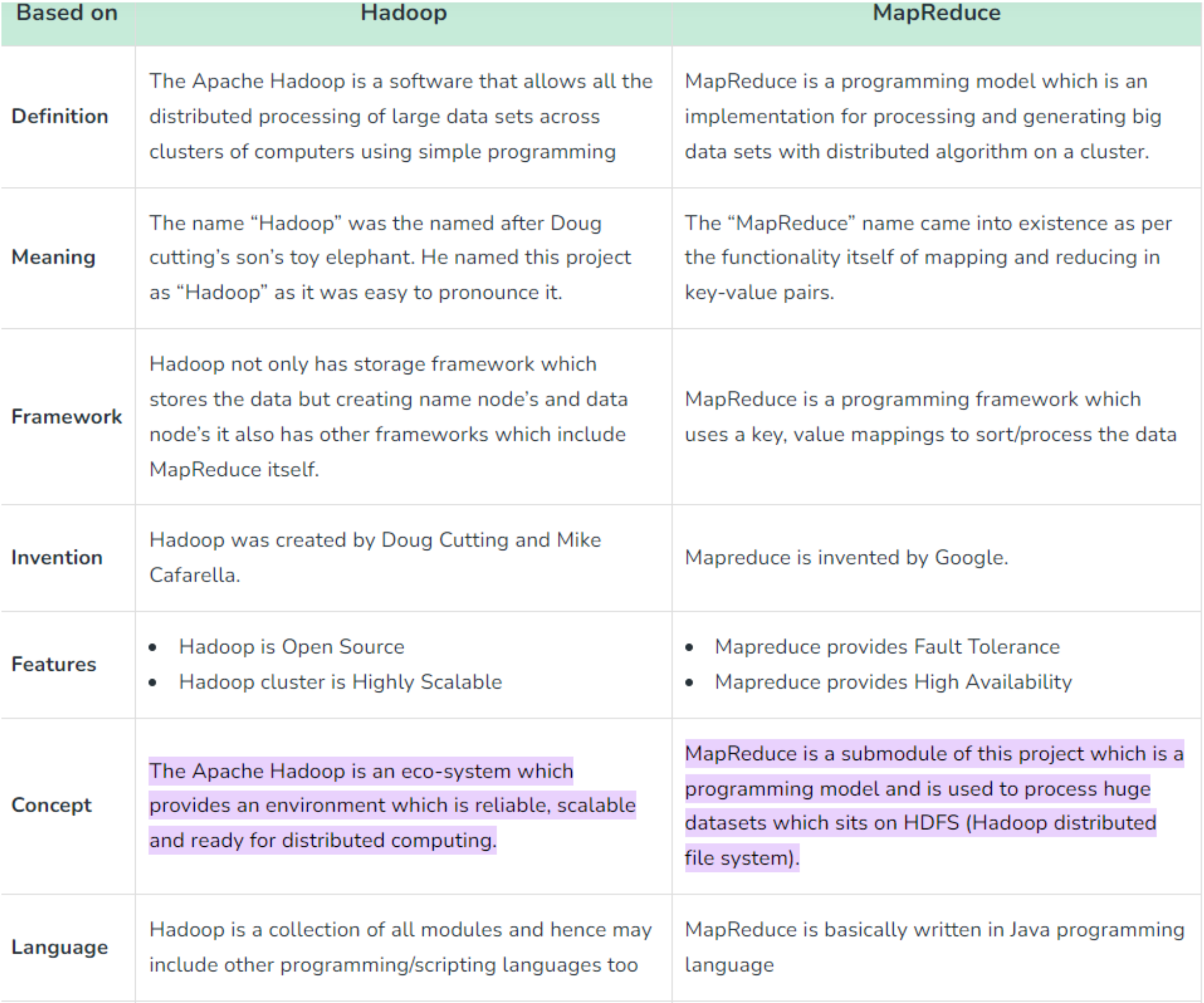
Big Data Frameworks

. 2. Apache Spark

It is a multi-language analytics engine for big data processing. It works well with very massive datasets. Along with batch processing, it supports stream processing. It distributes data across multiple computers itself or with the help of other distributing tools.

Spark uses **in-memory caching**, which makes it a superfast framework than other cluster computing systems, such as Hadoop. The data processing in memory is just one-step — reading data into memory, performing operations, and writing the results back.

Resilient Distributed Dataset (RDD) forms the architectural basis for Spark. It is a read-only multi-set of data items spread across different machines in a cluster.

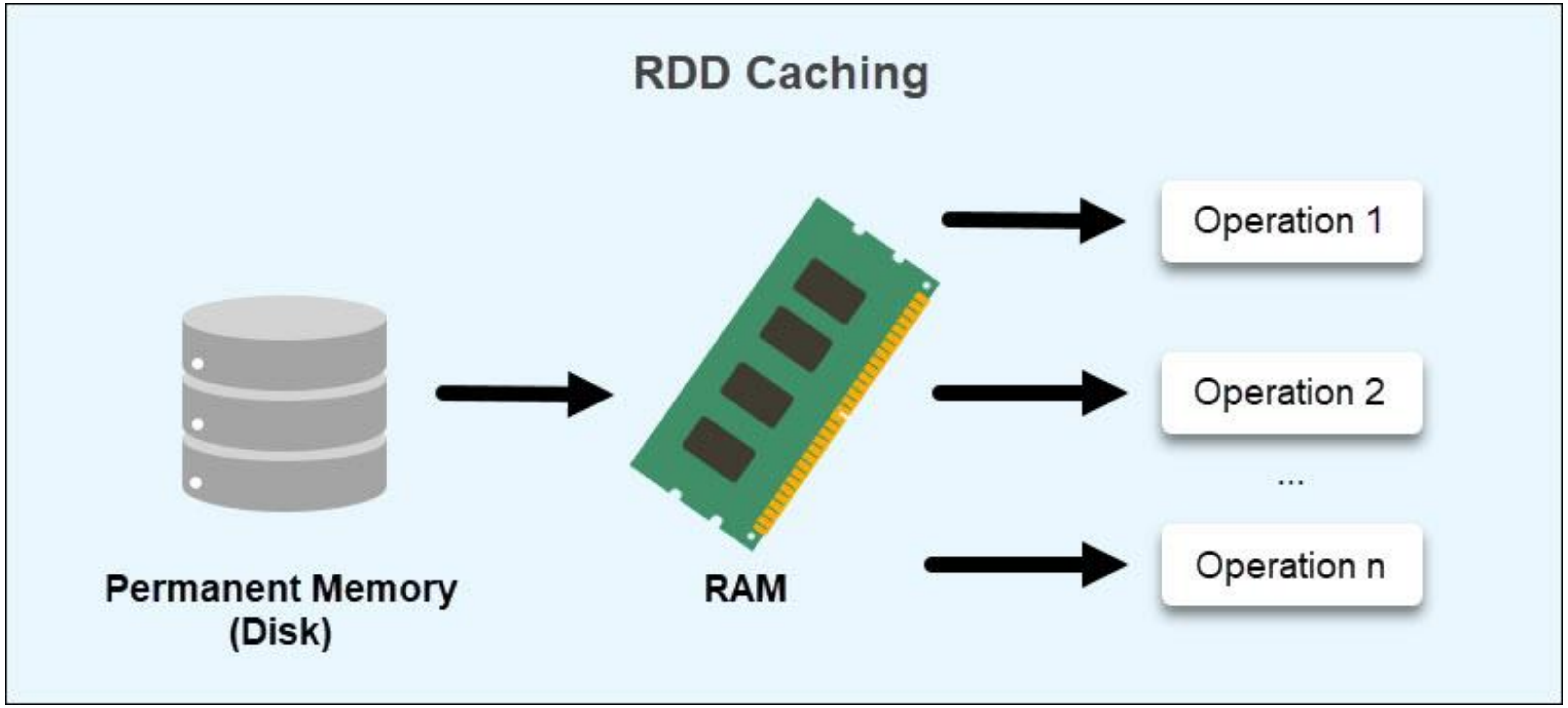


Based on	Hadoop	MapReduce
Definition	The Apache Hadoop is a software that allows all the distributed processing of large data sets across clusters of computers using simple programming	MapReduce is a programming model which is an implementation for processing and generating big data sets with distributed algorithm on a cluster.
Meaning	The name “Hadoop” was the named after Doug cutting’s son’s toy elephant. He named this project as “Hadoop” as it was easy to pronounce it.	The “MapReduce” name came into existence as per the functionality itself of mapping and reducing in key-value pairs.
Framework	Hadoop not only has storage framework which stores the data but creating name node’s and data node’s it also has other frameworks which include MapReduce itself.	MapReduce is a programming framework which uses a key, value mappings to sort/process the data
Invention	Hadoop was created by Doug Cutting and Mike Cafarella.	Mapreduce is invented by Google.
Features	• Hadoop is Open Source • Hadoop cluster is Highly Scalable	• Mapreduce provides Fault Tolerance • Mapreduce provides High Availability
Concept	The Apache Hadoop is an eco-system which provides an environment which is reliable, scalable and ready for distributed computing.	MapReduce is a submodule of this project which is a programming model and is used to process huge datasets which sits on HDFS (Hadoop distributed file system).
Language	Hadoop is a collection of all modules and hence may include other programming/scripting languages too	MapReduce is basically written in Java programming language

Why Do We Need RDDs in Spark?

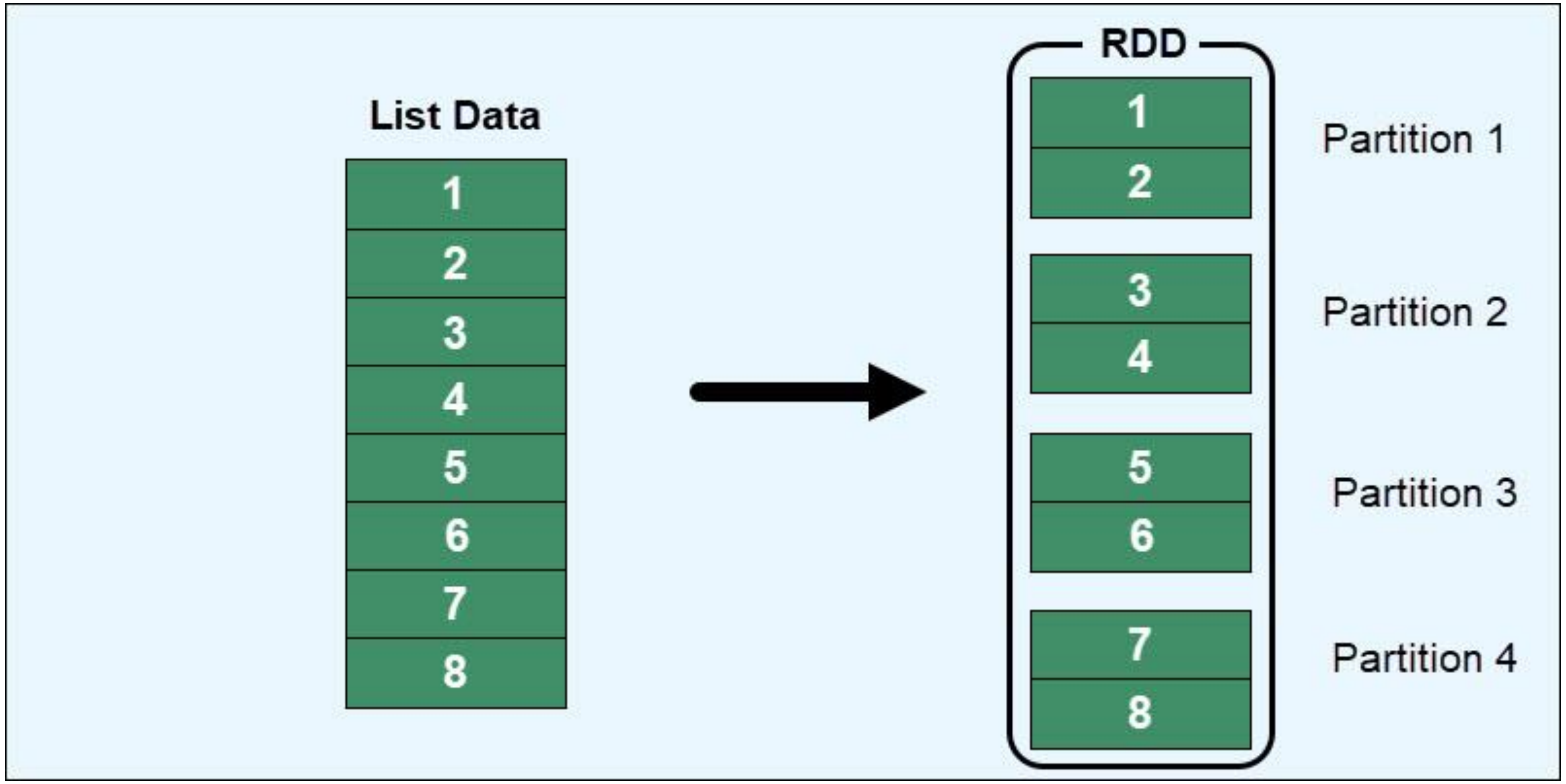
RDDs address MapReduce's shortcomings in data sharing. When reusing data for computations, MapReduce requires writing to external storage (HDFS, Cassandra, HBase, etc.). The read and write processes between jobs consume a significant amount of memory.

Furthermore, data sharing between tasks is slow due to replication, serialization, and increased disk usage.



RDDs aim to reduce the usage of external storage systems by leveraging **in-memory compute operation storage.** This approach improves data exchange speeds between tasks by 10 to 100 times. Speed is critical when working with large data volumes. Spark RDDs make it easier to train [machine learning algorithms]() and handle large amounts of data for analytic

How Does RDD Store Data?



The disadvantages when working with Resilient Distributed Datasets include:

No schematic view of data. RDDs have a hard time dealing with structured data. A better option for handling structured data is through the DataFrames and Datasets APIs, which fully integrate with RDDs in Spark.

Garbage collection. Since RDDs are in-memory objects, they rely heavily on Java's memory management and serialization. This causes performance limitations as data grows.

Overflow issues. When RDDs run out of RAM, the information resides on a disk, requiring additional RAM and disk space to overcome overflow issues.

No automated optimization. An RDD does not have functions for automatic input optimization. While other Spark objects, such as DataFrames and Datasets, use the Catalyst optimizer, for RDDs, optimization happens manually.

Spark consists of the following core components

Spark Core: An execution engine and the heart of Spark, which forms the basis for all other components. It manages task dispatching, I/O operations, and task scheduling.

Spark SQL: Built on top of Spark core, Spark SQL performs distributed processing on data. It provides access to various data sources — HDFS, Hive, etc.

Spark Streaming: A library to process streaming data. It can stream gigabytes per second. It splits data into mini-batches and transforms them into RDDs.

MLlib: A machine learning containing different ML algorithms.

GraphX: A distributed graph-processing framework.

What is the difference between persist and cache table in Spark?

CACHE and PERSIST do the same job to help in retrieving intermediate data used for computation quickly by storing it in memory, while by caching we can store intermediate data used for calculation only in memory , persist additionally offers caching with more options/flexibility

3. Apache Hive

Hive is an open-source distributed data warehouse system built on top of Apache Hadoop. It supports reading, writing, and analyzing petabytes of data stored in distributed storage. We get an SQL-like interface called HiveQL to query data stored in databases and file systems that integrate with Hadoop.

Traditional databases can only process small to medium volumes of data. On the other hand, Hive leverages batch processing, like Hadoop, to process data quickly across a distributed database.

4.Apache Storm

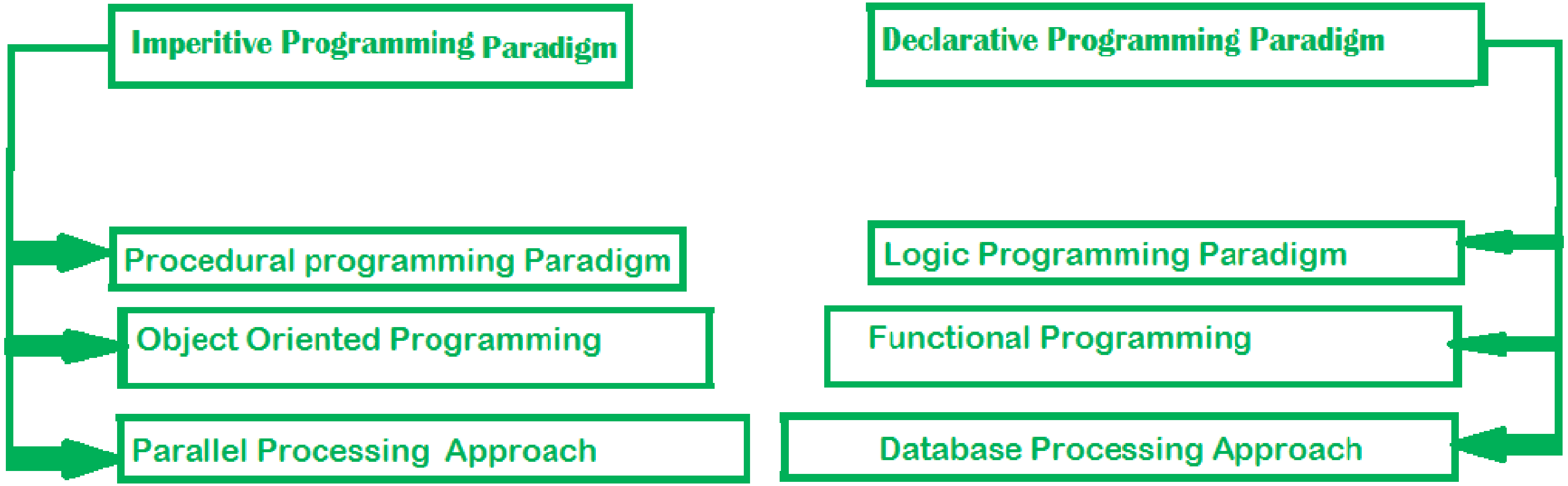
Storm is a free and open-source real-time distributed big data processing system. It makes it easy to process unbounded streams of data. It processes data in a fault-tolerant and horizontally-scalable way. While simple to use, Storm is compatible with any programming language.

5. Apache Samza

It is a distributed stream processing framework, allowing users to build real-time applications that can process data.

Big Data Programming Paradigms

Programming Paradigms



Imperitive Programming Paradigm

Procedural programming Paradigm

Object Oriented Programming

Parallel Processing Approach

Declarative Programming Paradigm

Logic Programming Paradigm

Functional Programming

Database Processing Approach

1. **Imperative programming paradigm:** It is one of the oldest programming paradigm. It features close relation to machine architecture. It is based on Von Neumann architecture. It works by changing the program state through assignment statements. It performs step by step task by changing state. The main focus is on how to achieve the goal. The paradigm consist of several statements and after execution of all the result is stored.

Advantages:

1. Very simple to implement
2. It contains loops, variables etc.

Disadvantage:

1.Complex problem cannot be solved
2.Less efficient and less productive
3.Parallel programming is not possible

Examples of Imperative programming paradigm:

C : developed by Dennis Ritchie and Ken Thompson
Fortran : developed by John Backus for IBM
Basic : developed by John G Kemeny and Thomas E Kurtz

2. Procedural programming paradigm –

This paradigm emphasizes on procedure in terms of under lying machine model. There is no difference in between procedural and imperative approach. It has the ability to reuse the code and it was boon at that time when it was in use because of its reusability.

Examples of Procedural programming paradigm:

C : developed by Dennis Ritchie and Ken Thompson
C++ : developed by Bjarne Stroustrup
Java : developed by James Gosling at Sun Microsystems
ColdFusion : developed by J J Allaire
Pascal : developed by Niklaus Wirth

3. Object oriented programming –

The program is written as a collection of classes and object which are meant for communication. The smallest and basic entity is object and all kind of computation is performed on the objects only. More emphasis is on data rather procedure. It can handle almost all kind of real life problems which are today in scenario.

Advantages:

•Data security
•Inheritance
•Code reusability
•Flexible and abstraction is also present

Examples of Object Oriented programming paradigm:

Simula : first OOP language
Java : developed by James Gosling at Sun Microsystems
C++ : developed by Bjarne Stroustrup
Objective-C : designed by Brad Cox
Visual Basic .NET : developed by Microsoft
Python : developed by Guido van Rossum
Ruby : developed by Yukihiro Matsumoto
Smalltalk : developed by Alan Kay, Dan Ingalls, Adele Goldberg

Parallel processing approach –

Parallel processing is the processing of program instructions by dividing them among multiple processors. A parallel processing system posses many numbers of processor with the objective of running a program in less time by dividing them. This approach seems to be like divide and conquer. Examples are NESL (one of the oldest one) and C/C++ also supports because of some library function.

2. Declarative programming paradigm:

It is divided as Logic, Functional, Database. In computer science the declarative programming is a style of building programs that expresses logic of computation without talking about its control flow. It often considers programs as theories of some logic.It may simplify writing parallel programs. The focus is on what needs to be done rather how it should be done basically emphasize on what code is actually doing. It just declares the result we want rather how it has be produced. This is the only difference between imperative (how to do) and declarative (what to do) programming paradigms. Getting into deeper we would see logic, functional and database.

Logic programming paradigms –

It can be termed as abstract model of computation. It would solve logical problems like puzzles, series etc. In logic programming we have a knowledge base which we know before and along with the question and knowledge base which is given to machine, it produces result. In normal programming languages, such concept of knowledge base is not available but while using the concept of artificial intelligence, machine learning we have some models like Perception model which is using the same mechanism.

In logical programming the main emphasize is on knowledge base and the problem. The execution of the program is very much like proof of mathematical statement, **e.g., Prolog**

```
predicates
  sumoftwonumber(integer, integer).

clauses
  sumoftwonumber(0, 0).
  sumoftwonumber(N, R) :-
    N > 0,
    N1 is N - 1,
    sumoftwonumber(N1, R1),
    R is R1 + N.
```

Functional programming paradigms –

The functional programming paradigms has its roots in mathematics and it is language independent. The key principle of this paradigms is the execution of series of mathematical functions. The central model for the abstraction is the function which are meant for some specific computation and not the data structure. Data are loosely coupled to functions.The function hide their implementation. Function can be replaced with their values without changing the meaning of the program. Some of the languages like perl, javascript mostly uses this paradigm.

Examples of Functional programming paradigm:

JavaScript : developed by Brendan Eich
Haskell : developed by Lennart Augustsson, Dave Barton
Scala : developed by Martin Odersky
Erlang : developed by Joe Armstrong, Robert Virding
Lisp : developed by John Mccarthy
ML : developed by Robin Milner
Clojure : developed by Rich Hickey

Database/Data driven programming approach –

This programming methodology is based on data and its movement. Program statements are defined by data rather than hard-coding a series of steps. A database program is the heart of a business information system and provides file creation, data entry, update, query and reporting functions. There are several programming languages that are developed mostly for database application. For example SQL. It is applied to streams of structured data, for filtering, transforming, aggregating (such as computing statistics), or calling other programs. So it has its own wide application.

Big Data Programming Languages

Big Data Programming Languages

1. Python

Python is a general-purpose programming language that programmers use when integrating data analysis with web applications. Data scientists and technical analysts prefer using this open-source programming language because it offers them multiple data manipulation and plotting libraries, such as Pandas and Matplotlib. Data visualisation and machine learning become easy with Python, as it can support varied analyses and integrate with third-party packages. This high-level programming language may also work as a programming interface for an analytics system.

2. R

R is an open-source programming language that users utilise to work with graphics data visualisation and statistics. There is a wide range of graphical tools in R, along with open-source packages that help users visualise, model, manipulate and load data. Its robust environment allows a technical analyst to perform several types of data analyses. This programming language is highly flexible, as a user can run it on almost all operating systems.

3. Scala

Scala, also referred to as scalable language, is an efficient programming language for processing data quickly. It supports both functional programming and object-oriented programming (OOP), so it becomes easy for users to utilise languages based on these programming models. Several front-end developers prefer using Scala, as it combines compact syntax with impressive development tools. Fintech companies also utilise Scala to work with data architectures and cloud-based technologies. Scala has several features that a user can employ to write algorithms for machine learning and devise solutions for complex analytics.

4. Java

Programmers use Java to write production code that enables them to use big-data algorithms. Java for big data is helpful when programmers are implementing a theoretical model that they have created in Python. Big-data analysis is easy with Java, as it helps data scientists to process big data, manage higher prediction load and resize intricate ecosystems. Java also works as the base for many big-data tools, such as Spark, Storm or Mahout. Java is the foundation of Scala's Apache Spark library, so knowing how to work with it may also help users to write in Scala comfortably.

5. SQL

SQL stands for Structured Query Language, which is useful when working with complex datasets outside a relational-database environment. With SQL, a user can perform various operations, such as modifying tables and updating or removing records. It can help a data scientist to work with structured data. There are big-data platforms, such as Spark and Hadoop, that offer extension for querying using this domain-specific language. SQL is an effective tool with which data scientists can wrangle and prepare data, so they use the language when working with different big-data tools.

6. C++

When technical experts are working with complex machine-learning algorithms, they may often process data sets in terabytes and petabytes. To complete such tasks quickly, they may use C++, as this platform can process data in gigabytes in just a few seconds. Conducting predictive analytics in real time and keeping records consistent are some other benefits of using C++. A data scientist may use the programming language to code libraries and big-data frameworks. There are many deep-learning algorithms and neural networks that data scientists can write in C++.

7. Go

Go, also referred to as Golang, is an open-source programming language that helps developers build simple and efficient software tools. Its presence is usually evident in DevOps and web servers, but it is also a useful language for businesses that make data-driven decisions. A business can use Go to integrate computationally exhaustive algorithms with all its levels of organisational structure. With Go, pre-processing, transforming, analysing, modelling and validating become easy. This language allows users to write controllers for events that occur asynchronously.

8. Julia

Julia's performance is comparable to C++, which means it is fast, reliable and efficient. It is a programming language that offers robust statistical applicability and an interactive command line. C, R, Java and Python form the basis for many of its libraries. These libraries help data scientists to perform artificial intelligence development with ease. Doing high-level statistical work on this platform is unchallenging. It also outperforms some other languages when working with linear algebra, as it supports several machine-learning equations and matrixes.

Introduction to Data Science

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

Better decisions (should we choose A or B)
Predictive analysis (what will happen next?)
Pattern discoveries (find pattern, or maybe hidden information in the data)

Where is Data Science Needed?

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

Examples of where Data Science is needed:

For route planning: To discover the best routes to ship
To foresee delays for flight/ship/train etc. (through predictive analysis)
To create promotional offers
To find the best suited time to deliver goods
To forecast the next years revenue for a company
To analyze health benefit of training
To predict who will win elections

Data Science can be applied in nearly every part of a business where data is available. Examples are:

- Consumer goods
- Stock markets
- Industry
- Politics
- Logistic companies
- E-commerce

How Does a Data Scientist Work?

A Data Scientist requires expertise in several backgrounds:

Machine Learning
Statistics
Programming (Python or R)
Mathematics
Databases

A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.

Here is how a Data Scientist works:

Ask the right questions - To understand the business problem.

Explore and collect data - From database, web logs, customer feedback, etc.

Extract the data - Transform the data to a standardized format.

Clean the data - Remove erroneous values from the data.

Find and replace missing values - Check for missing values and replace them with a suitable value (e.g. an average value).

Normalize data - Scale the values in a practical range (e.g. 140 cm is smaller than 1,8 m.

However, the number 140 is larger than 1,8. - so scaling is important).

Analyze data, find patterns and make future predictions.

Represent the result - Present the result with useful insights in a way the "company" can understand.

Skills for Big Data Jobs

1. SQL

SQL is one of the most important skills that must have. While using SQL, a programmer can have an advantage in working with multiple technologies (such as NoSQL).
SQL is the data-centered language that works as a base for the big data era.

Programmers use SQL for multiple operations such as adding, updating, deleting, or modifying any records or tables, and so on. Besides this, RDBM or Relational database management is a crucial part of the field of data science and a data scientist can only control, manipulate or define and query the DB using SQL commands.

Today, some of the modern big data systems (such as Hadoop and Spark) also use SQL only for maintaining the RDBMS (relational database systems) and processing structured data.

2. Apache Spark

Spark was first introduced by UC Berkeley in 2009 and since then it started gaining popularity in the field of data science. Today, Spark is capable enough to handle data (up to Petabytes) at a time and its data distribution happens across thousands of cluster cooperating servers (both physical and virtual). Spark also comes with an extensive range of libraries (and APIs) that can be commonly used by multiple programming languages (such as R, Scala, and Python).

Besides this, Apache uses Hadoop Distributed File System (HDFS) but can be integrated equally with other data storage systems. Developers prefer Spark often because it enables overlapping the complex technologies (such as MapReduce) and that’s why it is widely being used by data scientists and has been highly adopted by major organizations. People holding such skills possess to bag more lucrative packages than others.

3. ML/AI/DL

Machine Learning, Artificial Intelligence, and Deep Learning are three hot fields of big data. Although the path is way beyond them they are the ones making a significant impact on the field.

Whether it’s your smartphone, car, laptop, home devices, etc. they all are now highly equipped with artificial intelligence that we’re using on daily basis. Whenever you pick your phone up and say aloud “Hey Siri”, it’s likely that you’re using these technologies. The point is, that AI, ML, and DL are everywhere in our surroundings today and data science is the interdisciplinary field of getting that knowledge as per requirement. These technologies are making a huge impact in our day-to-day lives and helping us in making a better future.

That’s why the professionals with the knowledge of machine learning, artificial intelligence, and deep learning are in huge demand irrespective of the business scale (from small to large) and the average payscale of entry-level professionals are somewhere around $110,100 per annum and makes to one of the most handsome jobs in the world.

4. Apache Hadoop

When it comes to handling any huge cluster of data, Hadoop is the answer all the time. Being one of the most popular big data platforms, it's widely used for data operations that involve large-scale (unstructured) data. If you want to make your career in big data, we must understand the importance and knowledge of handling data on large scale.

Hadoop was first introduced by *Doug Cutting and Mike Cafarella* in 2005 and became public in late 2012. Ever since many implementations and development have been made. In today's time, some of the most popular components that are widely used in Hadoop are Hive, Pig, HDFS, MapReduce, etc.

5. Programming Language

This is something that creates the base of your big data career and there are certain general-purpose programming languages that enables you to work in this field smoothly. Languages like Python, R, Java, C++, SQL, Scala, Julia, etc. are some of the most widely used languages and that can also remove the learning barricade from becoming a successful data analyst expert.

Top companies prefer to hire those candidates who possess knowledge of these programming languages. You need to learn Python, Java, or R programming language (at least for your initial career) and that's where you will be able to start working on some of the most useful tools for data visualization, extraction, scraping, etc.

6. Data Visualization

Enabling the capabilities of displaying data visually is slightly more impactful than traditional methods. It helps people understand the latest trends, and patterns and help them in deciding the outcome (in many cases). That's why data visualization is among the top skill sets that you must possess to get on board in the big data field.

Companies are willing to pay much lucrative salaries to those who possess the knowledge of the best data visualization tools such as QlikView, Tableau, etc. Hence, to give your career a headstart, it is important to know what Big Data skills you need to break into analytics and start working with data.

7. Statistical Analysis

It's an important method of data analysis that helps in drawing meaningful outputs from any unstructured data. This method also helps in making fruitful business decisions based on data trends.

It can also be defined as a science of collecting and analyzing data to trace patterns and trends by involving numbers used in businesses. Being a data analyst will require you to possess this skill because it's all about data now and companies look forward to those candidates that carry such skills. Some of the most important tools for statistical analysis are MATLAB, R, SAS, etc.

Big data deals with high-volume, high-velocity and high-variety information assets,

A. True

B. False

The physical infrastructure of a big data is based on a distributed computing model.

A. True

B. False

The Big data analytics work on the unstructured data, where no specific pattern of the data is defined.

A. True

B. False

Amongst which of the following represents the Use of Hadoop,

A. Robust and Scalable

B. Affordable and Cost Effective

C. Adaptive and Flexible

D. All of the mentioned above

1. _____ is a platform for developing data flows for the extraction, transformation, and loading (ETL) of huge datasets, as well as for data analysis.

A. Spark

B. HBase

C. Hive

D. Pig

Pig is a high-level platform or tool that is used to process massive amounts of data at a high level. When processing via the MapReduce framework, it provides a high level of abstraction for the user. It includes a high-level scripting language, known as Pig Latin that is used to construct the data analysis scripts that are employed in the system.

In order to analyze all of this Big Data, Hive is a tool that has been developed.

A. True
B. False

Which of the following language is used in Data science?
A. C
B. C++
C. R
D. Ruby

Which of the following is correct skills for a Data Scientist?
A. Probability & Statistics
B. Machine Learning / Deep Learning
C. Data Wrangling
D. All of the above

Which of the following is not a part of data science process?
A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize

Which of the following are the Data Sources in data science?
A. Structured
B. Unstructured
C. Both A and B
D. None Of the above

Which of the following is not a application for data science?
A. Recommendation Systems
B. Image & Speech Recognition
C. Online Price Comparison
D. Privacy Checker

Point out the correct statement.
A. Raw data is original source of data
B. Preprocessed data is original source of data
C. Raw data is the data obtained after processing steps
D. None of the above

Which of the following is one of the key data science skills?
A. Statistics
B. Machine Learning
C. Data Visualization
D. All of the above

Raw data should be processed only one time.
A. True
B. False
C. Can be true or false
D. Can not say

Which of the following step is performed by data scientist after acquiring the data?
A. Data Cleaning
B. Data Integration
C. Data Replication
D. All of the above

What are the 3v's of Big Data?
A. Volume
B. Variety
C. Velocity
D. all the above

What was Hadoop written in ?
A. C
B. C++
C. Java
D. JSP

Which of the following platforms does Hadoop run on ?
A. Bare metal
B. Debian
C. Cross-platform
D. Unix-Like

What is Pig vs Hive vs HBase?

Hadoop is a big data eco system. HBase is a key value store (mostly), Hive is a system to execute SQL-like queries on a Hadoop system, Pig is a special query language to access big data. If you google these terms we will come up with a lot of architecture diagrams with a lot of elephants on them

Apache Flume

Apache Flume is an open-source tool for collecting, aggregating, and pushing log data from a massive number of sources into different storage systems in the Hadoop ecosystem, like HDFS and HBase. It is a highly available, distributed, and reliable service that is fault-tolerant and resilient

PIG - Program Implementation Group

What is an Oozie?

Oozie is a workflow scheduler system that manages Apache Hadoop jobs. Oozie's system operates by running the workflows of dependent jobs and permits users to create Directed Acyclic Graphs of workflows. These DAG's can be run in parallel and sequentially in Hadoop

Sqoop

Sqoop is a tool used to transfer bulk data between Hadoop and external datastores, such as relational databases (MS SQL Server, MySQL). To process data using Hadoop, the data first needs to be loaded into Hadoop clusters from several sources

Simplified Overview of Machine Learning Algorithms and Neural Networks

Introduction
Different Types of Learning
Hypothesis Space and Inductive Bias
Evaluation and Cross-Validation

- Linear Regression
- Introduction to Decision Trees
- Learning Decision Tree
- Overfitting
- Python Exercise on Decision Tree and Linear Regression

- k-Nearest Neighbour
- Feature Selection
- Feature Extraction
- Collaborative Filtering
- Python Exercise on kNN and PCA

- Bayesian Learning
- Naive Bayes
- Bayesian Network
- Python Exercise on Naive Bayes

- Logistic Regression
- Introduction Support Vector Machine
- SVM : The Dual Formulation
- SVM : Maximum Margin with Noise
- Nonlinear SVM and Kernel Function
- SVM : Solution to the Dual Problem
- Python Exercise on SVM

- Introduction
- Multilayer Neural Network
- Neural Network and Backpropagation Algorithm
- Deep Neural Network
- Python Exercise on Neural Network

- Introduction to Computational Learning Theory
- Sample Complexity : Finite Hypothesis Space
- VC Dimension
- Introduction to Ensembles
- Bagging and Boosting

- Introduction to Clustering
- Kmeans Clustering
- Agglomerative Hierarchical Clustering
- Python Exercise on kmeans clustering

What is Machine Learning?

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Features of Machine learning

Machine learning is data driven technology. Large amount of data generated by organizations on daily bases. So, by notable relationships in data, organizations makes better decisions.

Machine can learn itself from past data and automatically improve.

From the given dataset it detects various patterns on data.

For the big organizations branding is important and it will become more easy to target relatable customer base.

It is similar to data mining because it is also deals with the huge amount of data.

Types of machine learning problems

There are various ways to classify machine learning problems. Here, we discuss the most obvious ones.

1. On basis of the nature of the learning “signal” or “feedback” available to a learning system

Supervised learning: The model or algorithm is presented with example inputs and their desired outputs and then finds patterns and connections between the input and the output. The goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy on the training data. Some real-life examples are:

Image Classification: We train with images/labels. Then in the future, you give a new image expecting that the computer will recognize the new object.

Market Prediction/Regression: We train the computer with historical market data and ask the computer to predict the new price in the future.

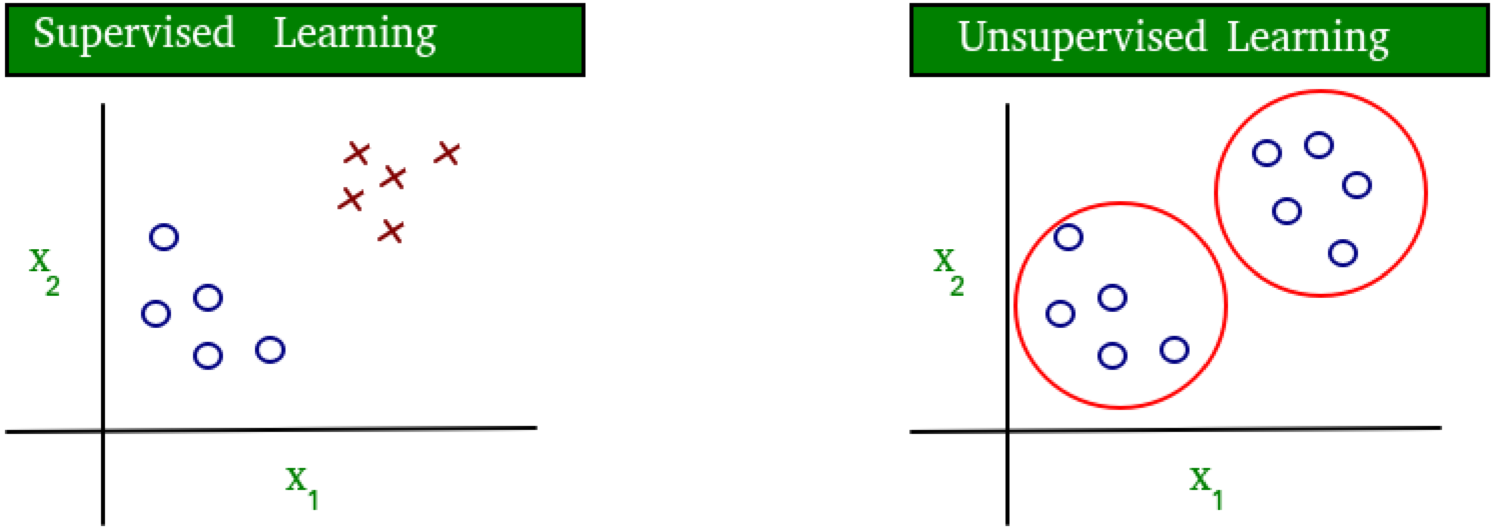
Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for clustering populations in different groups. Unsupervised learning can be a goal in itself (discovering hidden patterns in data).

Clustering: We ask the computer to separate similar data into clusters, this is essential in research and science.

High-Dimension Visualization: Use the computer to help us visualize high-dimension data.

Generative Models: After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust.

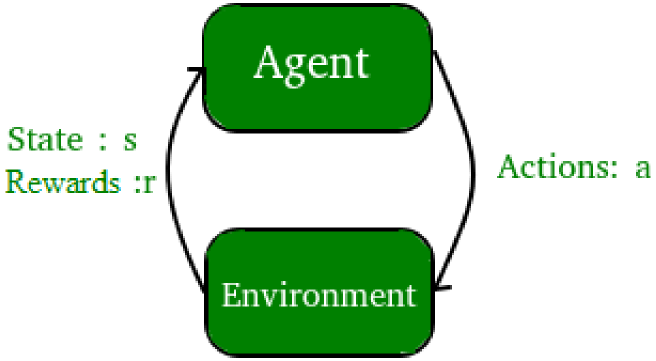
A simple diagram that clears the concept of supervised and unsupervised learning is shown below:



As you can see clearly, the data in supervised learning is labeled, whereas data in unsupervised learning is unlabelled.

- **Semi-supervised learning**: Problems where you have a large amount of input data and only some of the data is labeled, are called semi-supervised learning problems. These problems sit in between both supervised and unsupervised learning. For example, a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

- **Reinforcement learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

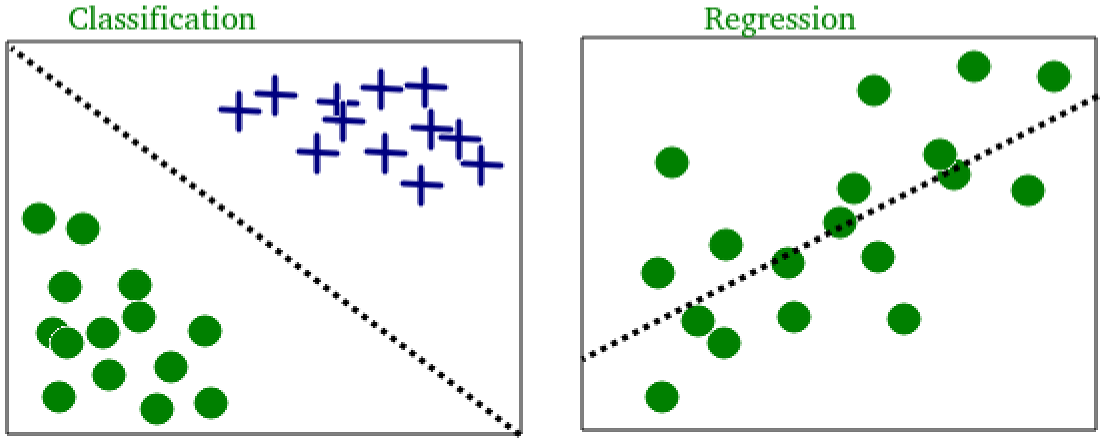


Two most common use cases of Supervised learning are:

Classification: Inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes and predicts whether or not something belongs to a particular class. This is typically tackled in a supervised way. Classification models can be categorized in two groups: Binary classification and Multiclass Classification. Spam filtering is an example of binary classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

Regression: It is also a supervised learning problem, that predicts a numeric value and outputs are continuous rather than discrete. For example, predicting stock prices using historical data.

An example of classification and regression on two different datasets is shown below:



3. Most common Unsupervised learning are:

Clustering: Here, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task. As you can see in the example below, the given dataset points have been divided into groups identifiable by the colors red, green, and blue.

Density estimation: The task is to find the distribution of inputs in some space.

Dimensionality reduction: It simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics.

On the basis of these machine learning tasks/problems, we have a number of algorithms that are used to accomplish these tasks. Some commonly used machine learning algorithms are Linear Regression, Logistic Regression, Decision Tree, SVM(Support vector machines), Naive Bayes, KNN(K nearest neighbors), K-Means, Random Forest, etc. Note: All these algorithms will be covered in upcoming articles.

Terminologies of Machine Learning

Model A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.

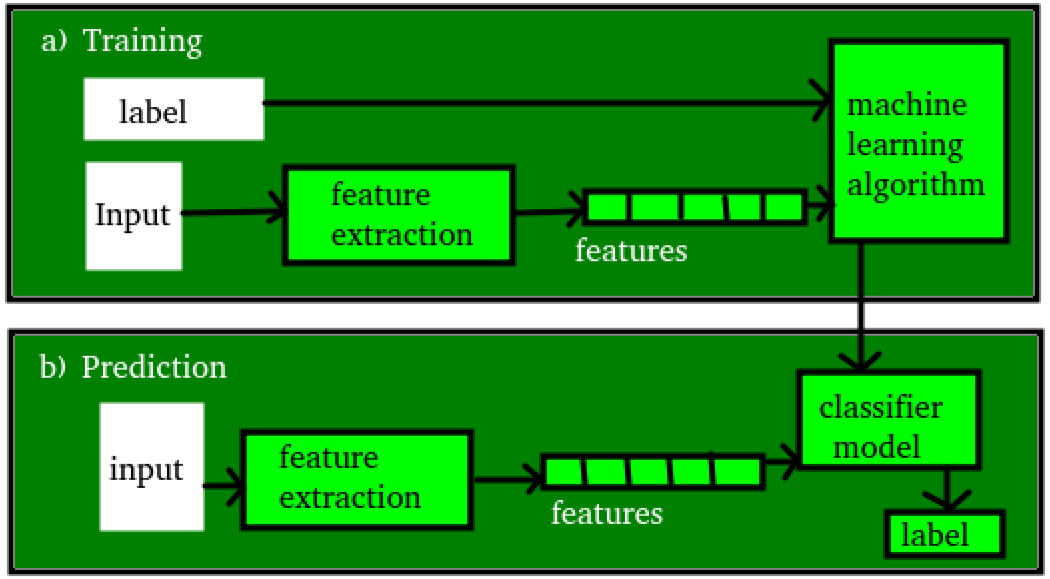
Feature A feature is an individual measurable property of our data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, etc. Note: Choosing informative, discriminating and independent features is a crucial step for effective algorithms. We generally employ a feature extractor to extract the relevant features from the raw data.

Target (Label) A target variable or label is the value to be predicted by our model. For the fruit example discussed in the features section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.

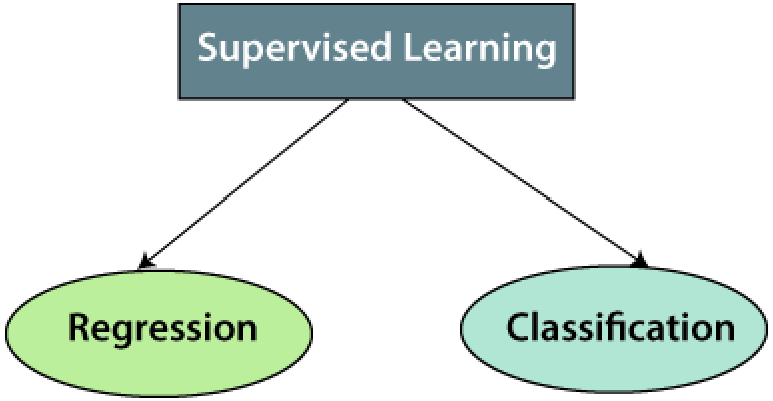
Training The idea is to give a set of inputs(features) and its expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.

Prediction Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output(label). But make sure if the machine performs well on unseen data, then only we can say the machine performs well.

The figure shown below clears the above concepts:



Types of supervised Machine learning Algorithms:



1. Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

Random Forest
Decision Trees
Logistic Regression
Support vector Machines

Applications of supervised learning

Bioinformatics

This is among the most widely used Supervised Learning applications, and we all use it regularly. Bioinformatics is the study of how individuals retain biological knowledge such as fingerprints, eye texture, earlobes, and so on. Mobile phones are now clever enough to comprehend our biological data and then verify us in order to increase system security.

The second would be speech recognition
1.It’s the type of program where you may convey your voice to the program, and it will identify you. The most well-known real-world gadgets are digital assistants such as Google Assistant or Siri, which respond to the term only with your voice.

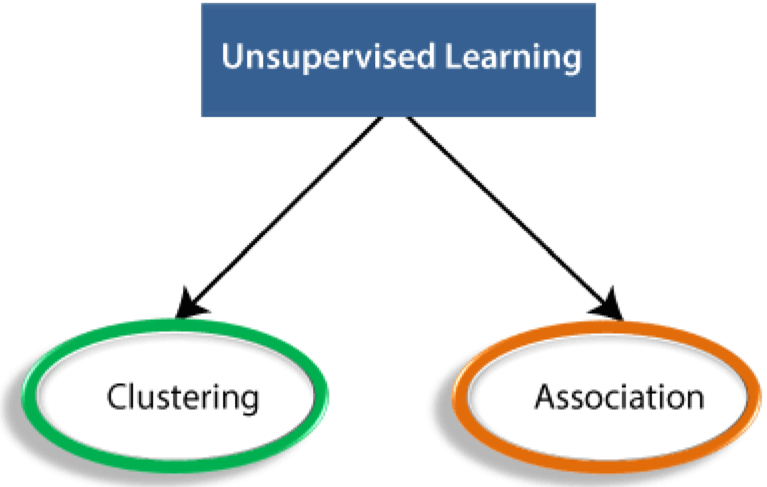
Next comes spam detection

1.This tool is used to prevent fictitious or machine-based communications from being sent. Gmail includes an algorithm that learns numerous wrong terms. The Oneplus Messages App asks the user to specify which terms should be prohibited, and the keyword will prevent such texts from the app.

There is also object recognition for the vision

This type of software is utilized when you have to define anything. You have a big dataset that you utilize to train the algorithm, and it can recognize a new object using this.

Types of Unsupervised Learning Algorithm



Clustering: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

Association: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

- **K-means clustering**
- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**
- **Singular value decomposition**

Some applications of unsupervised learning include

natural language processing,
image and video analysis,
anomaly detection,
customer segmentation, and
recommendation engines.

Advantages of Unsupervised Learning

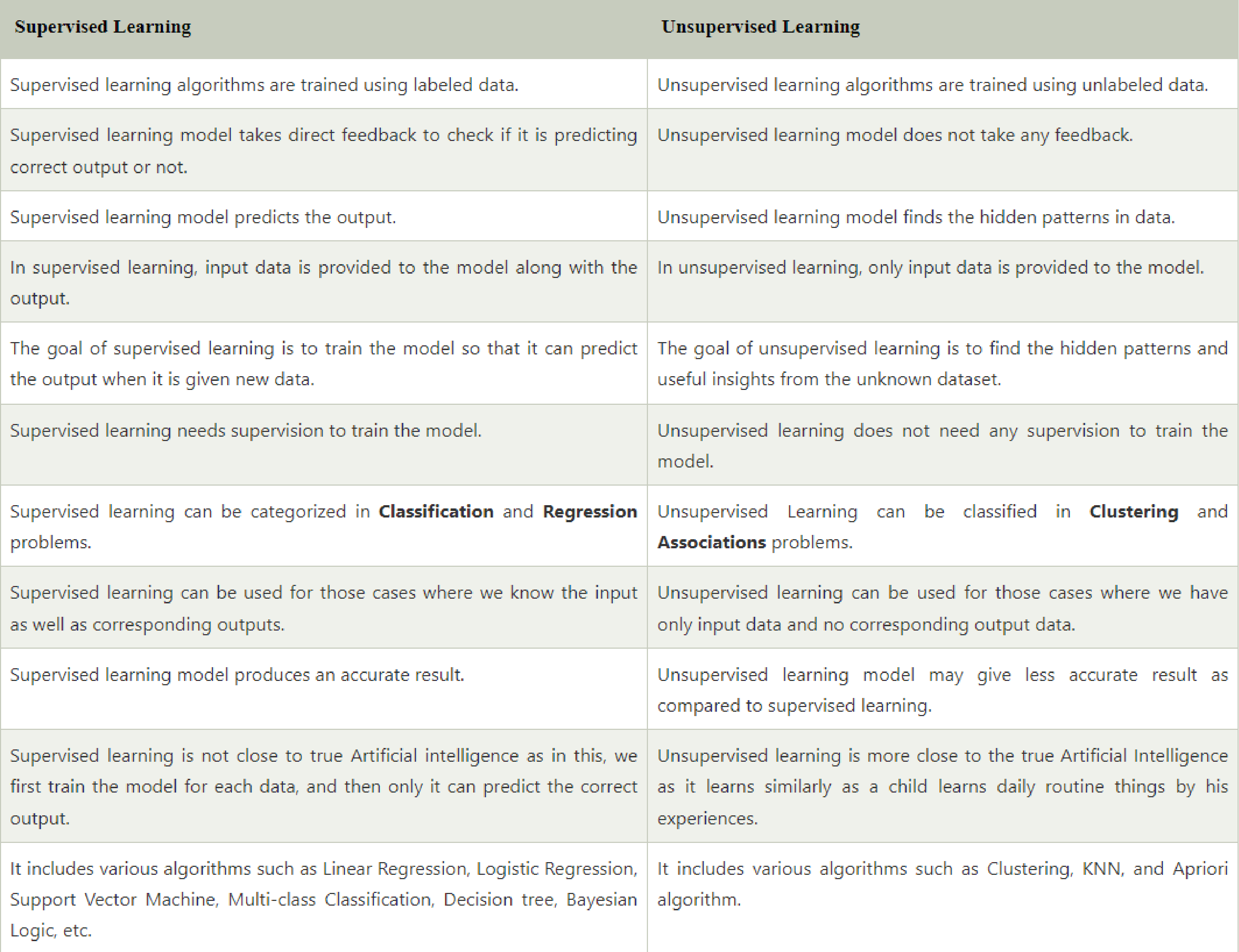
Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.

The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.



Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in **Classification** and **Regression** problems.	Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

Here are the steps to get started with machine learning:

Define the Problem: Identify the problem you want to solve and determine if machine learning can be used to solve it.

Collect Data: Gather and clean the data that you will use to train your model. The quality of your model will depend on the quality of your data.

Explore the Data: Use data visualization and statistical methods to understand the structure and relationships within your data.

Pre-process the Data: Prepare the data for modeling by normalizing, transforming, and cleaning it as necessary.

Split the Data: Divide the data into training and test datasets to validate your model.

Choose a Model: Select a machine learning model that is appropriate for your problem and the data you have collected.

Train the Model: Use the training data to train the model, adjusting its parameters to fit the data as accurately as possible.

Evaluate the Model: Use the test data to evaluate the performance of the model and determine its accuracy.

Fine-tune the Model: Based on the results of the evaluation, fine-tune the model by adjusting its parameters and repeating the training process until the desired level of accuracy is achieved.

Deploy the Model: Integrate the model into your application or system, making it available for use by others.

Monitor the Model: Continuously monitor the performance of the model to ensure that it continues to provide accurate results over time.

Which of the following are ML methods?
A. based on human supervision
B. supervised Learning
C. semi-reinforcement Learning
D. All of the above

Which of the following is not a supervised learning?
A) PCA B) Naive Bayesian C) Linear Regression D) Decision Tree

Real-Time decisions, Game AI, Learning Tasks, Skill acquisition, and Robot Navigation are applications of_________
A) Reinforcement Learning B) Supervised Learning: Classification C) Unsupervised Learning: Regression
D) None of the above

How do you handle missing or corrupted data in a dataset?
A. Drop missing rows or columns
B. Replace missing values with mean/median/mode
C. Assign a unique category to missing values
D. All of the above

ML is a field of AI consisting of learning algorithms that?
A) Improve their performance B) At executing some task
C) Over time with experience D) All of the above

A Machine Learning technique that helps in detecting the outliers in data.
A) Clustering B) Classification C) Anomaly Detection
D) All of the above

Which supervised learning technique can process both numeric and categorical input attributes?
A) Linear regression B) Bayes classifier
C) Ogistic regression D) None of the Above

Which of the following are common classes of problems in machine learning?

A) Clustering B) Regression
C) Classification D) All of the Above

The most common issue when using Machine Learning is____

A) Poor Data Quality B) Inadequate Infrastructure

C) Lack of skilled resources D) None of the Above

________algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed.

A) Deep Learning B) Artificial Intelligence

C) Machine Learning D) None of the Above

How many types of Machine Learning Techniques?

A)3 B) 5 C) 7 D) 9

Identify the type of learning in which labeled training data is used.

A) Reinforcement learning B) Unsupervised learning

C) Supervised learning D) Semi unsupervised learning

Machine learning is a subset of__________

A) Data Learning B) Deep Learning **C) Artificial Intelligence**

D) None of the above

Who is the father of machine learning?

A) Geoffrey Everest Hinton B) Geoffrey Chaucer

C) Geoffrey Hill D) None of the Above

What is true about Machine Learning?

A) Machine Learning (ML) is that field of computer science

B) ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method.

C) The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.

D) All of the above

Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

A) Random Forest B) Regression C) Classification

D) Decision Tree

Identify the successful applications of Machine Learning.

A) Learning to recognize spoken words B) Learning to drive an autonomous vehicle

C) Learning to classify new astronomical structures D) All of the Above

Big data in healthcare case study

Real-time alerting

While the concept of real-time alerts is not new, its potential in the healthcare industry is only beginning to be realized now. For example, Clinical Decision Support software analyzes medical data in hospitals, providing health practitioners with advice as they make prescriptive decisions.

However, doctors want patients to stay away from hospitals to avoid costly in-house treatments. Wearables will collect patients' health data continuously and send this data to the cloud. From there, AI-powered analytics will identify trends and anomalies, generating alerts that can proactively address potential health issues. This type of data-driven approach has the potential to transform healthcare, making it more proactive and preventative, and ultimately saving lives.

With this information, doctors can access a state-level database to see how the general public is doing regarding their health. In addition, they'll be able to compare that data against different factors like income brackets and socioeconomically disadvantaged areas, which should give them more insight into personalized care strategies.

Prevent opioid abuse

Opioid overdoses have now become the leading cause of accidental death in the United States, surpassing road accidents. The problem has gotten so severe that Canada has declared it a "national health crisis."
Care providers can use data to help solve the problem by understanding patterns of abuse and misuse to develop targeted interventions to help save lives. For example, big-data already tracks the distribution of opioids and identify "hot spots" of abuse.

Predictive analytics in healthcare

We have already recognized predictive analysis as one of the biggest business intelligence trends, but the potential applications reach far beyond business and further into the future. For example, healthcare online business intelligence aims to help doctors make data-driven decisions within seconds and improve patients' treatment. This is particularly useful in the case of patients with complex medical histories suffering from multiple conditions.

New BI solutions and tools would also be able to predict, for example, who is at risk of diabetes and thereby advise the use of additional screenings or weight management. In addition, treatment plans could be created and tailored specifically for each patient to provide the best possible outcome. With predictive analysis, we are only beginning to scratch the surface of what's possible in terms of using data to improve our lives.

Reducing fraud and enhancing security

Healthcare organizations hold a vast amount of sensitive data on their patients, which makes them a prime target for cyber-attacks. Studies have shown that 93% of healthcare organizations have experienced a data breach. The reason is simple: personal data is extremely valuable and profitable on the black markets. As a result, many organizations use analytics to help prevent security threats. They identifying changes in network traffic or any other behavior that reflects a cyber-attack. In addition, advancements in security, such as encryption technology, firewalls, anti-virus software, etc., answer the need for more security, and the benefits largely overtake the risks. Yes, there are always risks involved in using big-data. However, the potential rewards make it too valuable for healthcare organizations to ignore.

Analytical tools have helped streamline the insurance claims process. These tools result in faster payments for caregivers and better returns for patients. For example, the Centers for Medicare and Medicaid Services reported savings of over $210 million in fraud prevention within a year. These tools can also help to reduce the incidence of inaccurate claims. They also provide a more efficient and effective system for all involved. In this way, analytical tools are essential in improving the insurance claims process and protecting against fraudulent or inaccurate activity.

Big-data analytics and medical imaging

Medical imaging is vital; each year, doctors in the US prescribe about 600 million imaging procedures. However, analyzing and storing these images is expensive in terms of time and money. Radiologists need to examine each image individually, and hospitals need to store them for several years.

Big-data analytics for healthcare can help algorithms analyze hundreds of thousands of images. They can identify specific patterns in the pixels, and convert them into numbers to help the physician with the diagnosis. They even go further, saying that radiologists will no longer need to look at the images but instead analyze the outcomes of the algorithms. These algorithms will inevitably study and remember more images than they could in a lifetime. In this way, big-data analytics has the potential to revolutionize medical imaging and create significant efficiencies in the healthcare system.

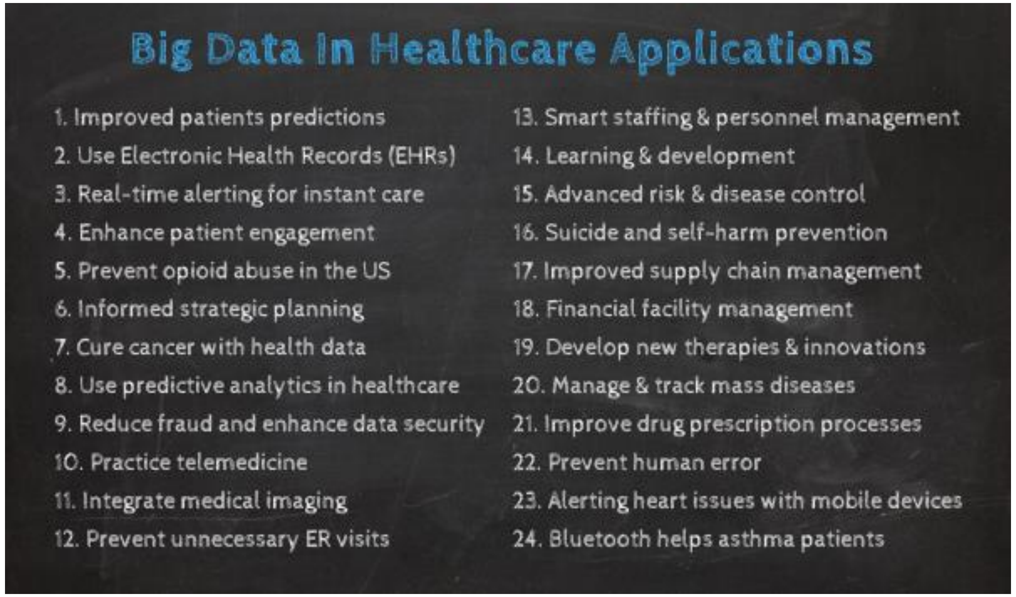
Advanced disease and risk control

Big-data has the potential to revolutionize the healthcare industry. By collecting and analyzing data, healthcare providers can better understand the hospitalization risk for patients with chronic conditions.

Care providers can then use this information to develop preventative care plans. These plans can help reduce the likelihood of deterioration and the need for hospitalization. In addition, they can also use big-data analytics to identify patterns and trends in patient behavior. This analysis empowers healthcare providers to address issues before they become problems proactively. The potential benefits of big-data are vast, and its impact on healthcare is just beginning to make impact.

These examples show how data and analytics can impact the healthcare infrastructure. It has the potential to make a real difference in the lives of patients and care providers.

Big Data Applications In Healthcare



Big Data In Healthcare Applications

1. Improved patients predictions
2. Use Electronic Health Records (EHRs)
3. Real-time alerting for instant care
4. Enhance patient engagement
5. Prevent opioid abuse in the US
6. Informed strategic planning
7. Cure cancer with health data
8. Use predictive analytics in healthcare
9. Reduce fraud and enhance data security
10. Practice telemedicine
11. Integrate medical imaging
12. Prevent unnecessary ER visits
13. Smart staffing & personnel management
14. Learning & development
15. Advanced risk & disease control
16. Suicide and self-harm prevention
17. Improved supply chain management
18. Financial facility management
19. Develop new therapies & innovations
20. Manage & track mass diseases
21. Improve drug prescription processes
22. Prevent human error
23. Alerting heart issues with mobile devices
24. Bluetooth helps asthma patients

1) Patients Predictions For Improved Staffing

For our first example of big data in healthcare, we will look at one classic problem that any shift manager faces: how many people do I put on staff at any given period? If We put on too many workers, We run the risk of having unnecessary labor costs add up. With too few workers, We can have poor customer service outcomes – which can be fatal for patients in that industry.

Big data is helping to solve this problem, at least at a few hospitals in Paris. A white paper by Intel details how four hospitals that are part of the Assistance Publique-Hôpitaux de Paris have been using data from a variety of sources to come up with daily and hourly predictions of how many patients are expected to be at each facility.

One of the key data sets is 10 years' worth of hospital admissions records, which data scientists crunched using "time series analysis" techniques. These analyses allowed the researchers to see relevant patterns in admission rates. Then, they could use machine learning to find the most accurate algorithms that predicted future admissions trends.

Summing up the product of all this work, the data science team developed a web-based user interface that forecasts patient loads and helps in planning resource allocation by utilizing online data visualization that reaches the goal of improving the overall patients' care.

2) Electronic Health Records (EHRs)

It’s the most widespread application of big data in medicine. Every person has their own digital record, which includes demographics, medical history, allergies, laboratory test results, etc. Records are shared via secure information systems and are available for providers from both the public and private sectors. Every record is comprised of one modifiable file, which means that doctors can implement changes over time with no paperwork and no danger of data replication.

EHRs can also trigger warnings and reminders when a patient should get a new lab test or track prescriptions to see if he or she has been following doctors’ orders. Although EHR is a great idea, many countries still struggle to implement them fully. U.S. has made a major leap, with 94% of hospitals adopting EHRs according to this HITECH research, but the EU still lags behind. However, an ambitious directive drafted by the European Commission is supposed to change it.
Kaiser Permanente is leading the way in the U.S. and could provide a model for the EU to follow. They’ve fully implemented a system called HealthConnect that shares data across all of their facilities and makes it easier to use EHRs. A McKinsey report on big data healthcare analytics states that “The integrated system has improved outcomes in cardiovascular disease and achieved an estimated $1 billion in savings from reduced office visits and lab tests.”

3) Real-Time Alerting

Other examples of data analytics in healthcare share one crucial functionality – real-time alerting. In hospitals, Clinical Decision Support (CDS) software analyzes medical data on the spot, providing health practitioners with advice as they make prescriptive decisions.

However, doctors want patients to stay away from hospitals to avoid costly in-house treatments. This is already trending as one of the business intelligence buzzwords in 2021 and has the potential to become part of a new strategy. Wearables will collect patients’ health data continuously and send this data to the cloud.

Additionally, this information will be accessed to the database on the state of health of the general public, which will allow doctors to compare this data in a socio-economic context and modify the delivery strategies accordingly. Institutions and care managers will use sophisticated tools to monitor this massive data stream and react every time the results will be disturbing.

For example, if a patient’s blood pressure increases alarmingly, the system will send a live alert to the doctor, who will then take action to reach the patient and administer measures to lower the pressure.

Another example is that of **Asthmapolis, which has started to use inhalers with GPS-enabled trackers** in order to identify asthma trends both on an individual level and looking at larger populations.

4) Enhancing Patient Engagement

Many consumers – and hence, potential patients – already have an interest in smart devices that record every step they take, their heart rates, sleeping habits, etc., on a permanent basis. All this vital information can be coupled with other trackable data to identify potential health risks lurking. Chronic insomnia and an elevated heart rate can signal a risk for future heart disease, for instance. Patients are directly involved in the monitoring of their own health, and incentives from health insurance can push them to lead a healthy lifestyle (e.g., giving money back to people using smartwatches).

Another way to do so comes with new wearables under development, tracking specific health trends and relaying them to the cloud where physicians can monitor them. Patients suffering from asthma or blood pressure could benefit from it, become a bit more independent and reduce unnecessary visits to the doctor.

5) Prevent Opioid Abuse In The US

Our fifth example of big data healthcare is tackling a serious problem in the US. Here's a sobering fact: as of this year, overdoses from misused opioids have caused more accidental deaths in the U.S. than road accidents, which were previously the most common cause of accidental death.

Analysis expert Bernard Marr writes about the problem in a Forbes article. The situation has gotten so dire that Canada has declared opioid abuse to be a "national health crisis," and President Obama earmarked $1.1 billion dollars for developing solutions to the issue while he was in office.

Once again, an application of big data analytics in healthcare might be the answer everyone is looking for: data scientists at Blue Cross Blue Shield have started working with analytical experts at Fuzzy Logix to tackle the problem. Using years of insurance and pharmacy data, Fuzzy Logix analysts have been able to identify 742 risk factors that predict with a high degree of accuracy whether someone is at risk for abusing opioids.

To be fair, reaching out to people identified as "high risk" and preventing them from developing a drug issue is a delicate undertaking. However, this project still offers a lot of hope for mitigating an issue that is destroying the lives of many people and costing the system a lot of money

6) Using Health Data For Informed Strategic Planning

The use of big data in healthcare allows for strategic planning thanks to better insights into people's motivations. Care managers can analyze check-up results among people in different demographic groups and identify what factors discourage people from taking up treatment.

The University of Florida made use of Google Maps and free public health data to prepare heat maps targeted at multiple issues, such as population growth and chronic diseases. Subsequently, academics compared this data with the availability of medical services in most heated areas. The insights gleaned from this allowed them to review their delivery strategy and add more care units to the most problematic areas.

7) Big Data Might Just Cure Cancer

Another interesting example of the use of big data in healthcare is the Cancer Moonshot program. Before the end of his second term, President Obama came up with this program that had the goal of accomplishing 10 years' worth of progress toward curing cancer in half that time.

Medical researchers can use large amounts of data on treatment plans and recovery rates of cancer patients in order to find trends and treatments that have the highest rates of success in the real world. For example, researchers can examine tumor samples in biobanks that are linked up with patient treatment records. Using this data, researchers can see things like how certain mutations and cancer proteins interact with different treatments and find trends that will lead to better outcomes.

8) Predictive Analytics In Healthcare

We have already recognized predictive analysis as one of the biggest business intelligence trends for two years in a row, but the potential applications reach far beyond business and much further into the future. Optum Labs, a US research collaborative, has collected EHRs of over 30 million patients to create a database for predictive analysis tools that will improve the delivery of care.

The goal of healthcare online business intelligence is to help doctors make data-driven decisions within seconds and improve patients' treatment. This is particularly useful in the case of patients with complex medical histories suffering from multiple conditions. New BI solutions and tools would also be able to predict, for example, who is at risk of diabetes and thereby be advised to make use of additional screenings or weight management.

10) Telemedicine

Telemedicine has been present on the market for over 40 years, but only today, with the arrival of online video conferences, smartphones, wireless devices, and wearables, has it been able to come into full bloom. The term refers to the delivery of remote clinical services using technology.

It is used for primary consultations and initial diagnosis, remote patient monitoring, and medical education for health professionals. Some more specific uses include telesurgery – doctors can perform operations with the use of robots and high-speed real-time data delivery without physically being in the same location as a patient.

Clinicians use telemedicine to provide personalized treatment plans and prevent hospitalization or re-admission. Such use of healthcare data analytics can be linked to the use of predictive analytics, as seen previously. It allows clinicians to predict acute medical events in advance and prevent the deterioration of patients' conditions.

By keeping patients away from hospitals, telemedicine helps to reduce costs and improve the quality of service. Patients can avoid waiting in lines, and doctors don't waste time on unnecessary consultations and paperwork. Telemedicine also improves the availability of care as patients' states can be monitored and consulted anywhere and anytime.

9) Reduce Fraud And Enhance Security

Some studies have shown that 93% of healthcare organizations have experienced a data breach. The reason is simple: personal data is extremely valuable and profitable on the black market. And any breach would have dramatic consequences. With that in mind, many organizations started to use analytics to help prevent security threats by identifying changes in network traffic or any other behavior that reflects a cyber-attack. Of course, big data has inherent security issues, and many think that using it will make organizations more vulnerable than they already are. But advances in security, such as encryption technology, firewalls, anti-virus software, etc., answer the need for more security, and the benefits brought largely overtake the risks.

Likewise, it can help prevent fraud and inaccurate claims in a systemic, repeatable way. Analytical tools help to streamline the processing of insurance claims, enabling patients to get better returns on their claims and caregivers to be paid faster. For instance, the Centers for Medicare and Medicaid Services said they saved over $210.7 million in fraud in just a year.

10) Telemedicine

Telemedicine has been present on the market for over 40 years, but only today, with the arrival of online video conferences, smartphones, wireless devices, and wearables, has it been able to come into full bloom. The term refers to the delivery of remote clinical services using technology.

It is used for primary consultations and initial diagnosis, remote patient monitoring, and medical education for health professionals. Some more specific uses include telesurgery – doctors can perform operations with the use of robots and high-speed real-time data delivery without physically being in the same location as a patient.

Clinicians use telemedicine to provide personalized treatment plans and prevent hospitalization or re-admission. Such use of healthcare data analytics can be linked to the use of predictive analytics, as seen previously. It allows clinicians to predict acute medical events in advance and prevent the deterioration of patients' conditions.

By keeping patients away from hospitals, telemedicine helps to reduce costs and improve the quality of service. Patients can avoid waiting in lines, and doctors don't waste time on unnecessary consultations and paperwork. Telemedicine also improves the availability of care as patients' states can be monitored and consulted anywhere and anytime.

11) Integrating Big-Style Data With Medical Imaging

Medical imaging is vital, and each year in the US, about 600 million imaging procedures are performed. Analyzing and storing these images manually is expensive both in terms of time and money, as radiologists need to examine each image individually, while hospitals need to store them for several years.

Medical imaging provider Carestream explains how big data analytics for healthcare could change how images are read: algorithms developed analyzing hundreds of thousands of images could identify specific patterns in the pixels and convert them into a number to help the physician with the diagnosis. They even go further, saying that it could be possible that radiologists will no longer need to look at the images but instead analyze the outcomes of the algorithms that will inevitably study and remember more images than they could in a lifetime. This would undoubtedly impact the role of radiologists, their education, and the required skillset.