**UNIT-1**

**Introduction to Big data**

**TOPICS:-**

Types of Digital Data, Classification of Digital Data, Characteristics of Data, Evolution of Big Data, Definition of Big Data, Challenges with Big Data, What is Big Data?, Other Characteristics of Data Which are not Definitional Traits of Big Data, Why Big Data?, analyzing Data with Unix tools, Analyzing Data with Hadoop, Hadoop Streaming, Hadoop Echo System.

Types of Digital Data:-

Big data, at its core, involves vast amounts of digital data, but in a wide variety of formats and gathered at mind-boggling speeds. Different types of digital data, including batch or streaming, can be collected and processed for consumption by machines or people, through big data integration. Digital data in big data is commonly referred to by its characteristics.

ITS TYPES: -

* Structured Data
* Semi-Structured Data
* Unstructured Data

**Structured Data:-**

**Structured data** is the data which conforms to a data model, has a well define structure, follows a consistent order and can be easily accessed and used by a person or a computer program.

Structured data is usually stored in well-defined schemas such as Databases. It is generally tabular with column and rows that clearly define its attributes.

SQL (Structured Query language) is often used to manage structured data stored in databases.

**Characteristics of Structured Data:**

* Data conforms to a data model and has easily identifiable structure
* Data is stored in the form of rows and columns  
  **Example : Database**
* Data is well organised so, Definition, Format and Meaning of data is explicitly known
* Data resides in fixed fields within a record or file
* Similar entities are grouped together to form relations or classes
* Entities in the same group have same attributes
* Easy to access and query, So data can be easily used by other programs
* Data elements are addressable, so efficient to analyse and process

**Sources of Structured Data:**

* SQL Databases
* Spreadsheets such as Excel
* OLTP Systems
* Online forms
* Sensors such as GPS or RFID tags
* Network and Web server logs
* Medical devices

**Advantages of Structured Data:**

* Structured data have a well defined structure that helps in easy storage and access of data
* Data can be indexed based on text string as well as attributes. This makes search operation hassle-free
* Data mining is easy i.e knowledge can be easily extracted from data
* Operations such as Updating and deleting is easy due to well structured form of data
* Business Intelligence operations such as Data warehousing can be easily undertaken
* Easily scalable in case there is an increment of data
* Ensuring security to data is easy
* **Semi-Structured Data:-**

**Semi-structured data** is data that does not conform to a data model but has some structure. It lacks a fixed or rigid schema. It is the data that does not reside in a rational database but that have some organizational properties that make it easier to analyze. With some processes, we can store them in the relational database.

**Characteristics of semi-structured Data:**

* Data does not conform to a data model but has some structure.
* Data can not be stored in the form of rows and columns as in Databases
* Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored
* Similar entities are grouped together and organized in a hierarchy
* Entities in the same group may or may not have the same attributes or properties
* Does not contain sufficient metadata which makes automation and management of data difficult
* Size and type of the same attributes in a group may differ
* Due to lack of a well-defined structure, it cannot used by computer programs easily

**Sources of semi-structured Data:**

* E-mails
* XML and other markup languages
* Binary executables
* TCP/IP packets
* Zipped files
* Integration of data from different sources
* Web pages

**Advantages of Semi-structured Data:**

* The data is not constrained by a fixed schema
* Flexible i.e Schema can be easily changed.
* Data is portable
* It is possible to view structured data as semi-structured data
* Its supports users who can not express their need in SQL
* It can deal easily with the heterogeneity of sources.

**Disadvantages of Semi-structured data**

* Lack of fixed, rigid schema make it difficult in storage of the data
* Interpreting the relationship between data is difficult as there is no separation of the schema and the data.
* Queries are less efficient as compared to [structured data](https://www.geeksforgeeks.org/structured-data/).

**Problems faced in storing semi-structured data**

* Data usually has an irregular and partial structure. Some sources have implicit structure of data, which makes it difficult to interpret the relationship between data.
* Schema and data are usually tightly coupled i.e they are not only linked together but are also dependent of each other. Same query may update both schema and data with the schema being updated frequently.
* Distinction between schema and data is very uncertain or unclear. This complicates the designing of structure of data
* Storage cost is high as compared to structured data
* **Unstructured Data:-**

**Unstructured data** is the data which does not conforms to a data model and has no easily identifiable structure such that it cannot be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

**Characteristics of Unstructured Data:**

* Data neither conforms to a data model nor has any structure.
* Data cannot be stored in the form of rows and columns as in Databases
* Data does not follow any semantic or rules
* Data lacks any particular format or sequence
* Data has no easily identifiable structure
* Due to lack of identifiable structure, it cannot used by computer programs easily

**Sources of Unstructured Data:**

* Web pages
* Images (JPEG, GIF, PNG, etc.)
* Videos
* Memos
* Reports
* Word documents and PowerPoint presentations
* Surveys

**Advantages of Unstructured Data:**

* It supports the data which lacks a proper format or sequence
* The data is not constrained by a fixed schema
* Very Flexible due to absence of schema.
* Data is portable
* It is very scalable
* It can deal easily with the heterogeneity of sources.
* These types of data have a variety of business intelligence and analytics applications.

**Disadvantages Of Unstructured data:**

* It is difficult to store and manage unstructured data due to lack of schema and structure
* Indexing the data is difficult and error prone due to unclear structure and not having pre-defined attributes. Due to which search results are not very accurate.
* Ensuring security to data is difficult task.

**Problems faced in storing unstructured data:**

* It requires a lot of storage space to store unstructured data.
* It is difficult to store videos, images, audios, etc.
* Due to unclear structure, operations like update, delete and search is very difficult.
* Storage cost is high as compared to structed data.
* Indexing the unstructured data is difficult.

**Classification of Digital Data**: -

we are going to discuss the classification of data in which we will cover structured, unstructured data, and semi-structured data. Also, we will cover the features of the data. Let’s discuss one by one.

**Data Classification:**

Process of classifying data in relevant categories so that it can be used or applied more efficiently. The classification of data makes it easy for the user to retrieve it. Data classification holds its importance when comes to data security and compliance and also to meet different types of business or personal objective. It is also of major requirement, as data must be easily retrievable within a specific period of time.

**Types of Data Classification :**

Data can be broadly classified into 3 types.

**1.** **Structured Data:**

Structured data is created using a fixed schema and is maintained in tabular format. The elements in structured data are addressable for effective analysis. It contains all the data which can be stored in the [SQL database](https://www.geeksforgeeks.org/sql-tutorial/) in a tabular format. Today, most of the data is developed and processed in the simplest way to manage information.

**Examples –**

[Relational data](https://www.geeksforgeeks.org/relational-model-in-dbms/), Geo-location, credit card numbers, addresses, etc.

Consider an example for Relational Data like you have to maintain a record of students for a university like the name of the student, ID of a student, address, and Email of the student. To store the record of students used the following relational schema and table for the same.

|  |  |  |  |
| --- | --- | --- | --- |
| **S\_ID** | **S\_Name** | **S\_Address** | **S\_Email** |
| 1001 | A | Delhi | A@gmail.com |
| 1002 | B | Mumbai | B@gmail.com |

**2. Unstructured Data :**

It is defined as the data in which is not follow a pre-defined standard or you can say that any does not follow any organized format. This kind of data is also not fit for the relational database because in the relational database you will see a pre-defined manner or you can say organized way of data. Unstructured data is also very important for the big data domain and To manage and store Unstructured data there are many platforms to handle it like [No-SQL Database](https://www.geeksforgeeks.org/introduction-to-nosql/).

**Examples –**

Word, PDF, text, media [logs](https://www.geeksforgeeks.org/log-based-recovery-in-dbms/), etc.

**3. Semi-Structured Data :**

Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in a [relational database](https://www.geeksforgeeks.org/relational-model-in-dbms/) but is very hard for some kind of semi-structured data, but semi-structured exist to ease space.

**Example –**

[XML data](https://www.geeksforgeeks.org/xml-basics/).

**Features of Data Classification:**

The main goal of the organization of data is to arrange the data in such a form that it becomes fairly available to the users. So it’s basic features as following.

* **Homogeneity –**The data items in a particular group should be similar to each other.
* **Clarity –**There must be no confusion in the positioning of any data item in a particular group.
* **Stability –** The data item set must be stable i.e. any investigation should not affect the same set of classification.
* **Elastic –**One should be able to change the basis of classification as the purpose of classification changes.

**Characteristics of Data(4’v)**:-

* **Volume**

Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using**distributed systems**, to store data in several locations and brought together by a software Framework like [**Hadoop**](https://www.edureka.co/blog/what-is-hadoop/).

Facebook alone can generate about **billion** messages, **4.5 billion** times that the “like” button is recorded, and over **350 million** new posts are uploaded **each day.** Such a huge amount of data can only be handled by Big Data Technologies

* **Variety**

As Discussed before, **Big Data** is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured

* **Veracity**

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

* **Value**

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights. You can get a better understanding with the

* **Velocity**

Last but never least, Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Dat is to provide data on demand and at a faster pace.

**Evolution of Big Data:-**

In the era of data and information, [big data](https://www.analyticsinsight.net/opening-alert-top-big-data-jobs-to-apply-in-december-2021/) is no longer new to businesses and society. It is a known fact that via big data solutions, organizations can generate insights and make well-informed decisions, discover new market trends, and improve productivity. As the amounts of data continue to grow, organizations are looking for new innovative ways to optimize big data. One of the major relationships of big [data analytics](https://www.analyticsinsight.net/what-are-the-best-5-big-data-applications-in-education-for-the-upcoming-2022/) with businesses is that their dependence on the internet increases, along with the amount of data generated by the rapid development and evolvement of technology. In 2022, the global big data [market](https://www.expertmarketresearch.com/reports/big-data-market) powered by big data analytics trends attained US$208 billion. It is expected that the big data market is expected to reach US$250 billion by 2026, with a CAGR of 10%. In this article, we have listed some [big data](https://www.analyticsinsight.net/application-of-big-data-in-the-fashion-industry/) analytics hidden trends to get to the core of its evolution in 2022.

**Big data analytics powering digital transformation**

Digital transformation is a global phenomenon that is driving a technological revolution all over the world. The transformation will continue to grow as IaaS providers scamper to cover the ground and build data centers. Digital transformation goes hand in hand with [big data](https://www.analyticsinsight.net/top-10-big-data-stocks-to-buy-before-it-explodes-in-2022/), AI, machine learning, and the Internet of Things (IoT). Machine learning and AI tools will continue to handle the data generated from the data analytics to operate systems, make sense of complex hidden relationships, and store and project insights beyond human understanding.

**Transformation from SaaS to IPaaS**

SaaS has been around for quite some time and has helped businesses optimize their businesses on the cloud. Earlier, the integration of SaaS has made headlines since it was relatively a new concept. But in 2022, we might not be able to see any revolutionary contributions from it. This is where iPaaS comes to play. As businesses try to avoid [data](https://www.analyticsinsight.net/top-big-data-analytics-project-ideas-for-engineering-students/) losses and disjointed information between departments and platforms, iPaaS may provide logical solutions and become the next best trend in 2022.

**Big data will help climate change research**

Solid data and proof might put the raging climate change research to rest by backing up the views and predictions by the climate change organization. The data might reveal some interesting insights about what is going on. The presence of legitimate data exempted from human biases will productively benefit the climate change debate.

**Big data might be used in local stores**

Almost 90% of the local businesses and enterprises are using data to generate productive insights from these tools. The use of data-as-a-service is becoming more commonplace and is predicted to grow by US$10.7 billion by 2023. Customers might encounter DaaS in the form of purchased music, videos, and image files from multiple sources online.

**The use of small data is on the rise**

Large enterprises can save massive amounts of time by just evaluating the most vital [data](https://www.analyticsinsight.net/heard-of-smart-data-discovery-all-you-need-to-know-about-it/) instead of entire lots of the generated data. This can be efficiently achieved if businesses shift from big data to small data. It can enable more streamlined, fast, and bandwidth-sparring innovations to take place.

**Data fabric will be the foundation**

As data becomes increasingly complex and digital business accelerates, data fabric will become the architecture that will support composable data and analytics in its various forms. Data fabric reduces the time for integration by 30%, and for development by 70% since the technology designs will draw on the ability to reuse and combine different data integration styles.

**Composable data and analytics**

The goal of composable data and analytics is to use components from multiple data, analytics, and AI solutions for a flexible, user-defined, and usable experience that will enable leaders to connect data insights to business outcomes. Composing new applications from the packaged business capabilities of each promotes productivity and agility.

**Big data to search novel medical cures**

It is a primary responsibility for businesses to invest in human welfare. So, the use of raging big data applications in innovating cures for novel diseases might increase. Many scientists hope that by consolidating all the medical records accumulated on the planet, the discovery of medical cures will become faster and sooner than ever imagined.

**The use of XOps**

The goal of XOps is to achieve efficiencies and economies of large scales using the best practices of DevOps for efficiency, reliability, and reusability while reducing the duplication of technology and processes and enabling automation.

**Planning and forecasting**

The increased use of predictive analytics has also boosted the availability of affordable applications in the market, for both BI platforms, like Qlik or Anaplan and standalone cloud services like Amazon Forest, which can help the users to easily integrate predictive analytics in the systems. These tools can be used for planning based on the generated forecast data to make intelligent and profitable decisions.

**Defination of big data**:-

## What Is Big Data?

Big Data is the data which cannot be managed by using traditional databases. Here is Gartner’s definition: The Data sets with huge volume, generated in different varieties with high velocity is termed as Big Data. These are considered as 3 Vs of Big Data.

These humongous volumes of data can be used to generate advanced patterns & address business problems you wouldn’t have been able to handle earlier.

### KEY TAKEAWAYS

* Big data is a great quantity of diverse information that arrives in increasing volumes and with ever-higher velocity.
* Big data can be structured (often numeric, easily formatted and stored) or unstructured (more free-form, less quantifiable).
* Nearly every department in a company can utilize findings from big data analysis, but handling its clutter and noise can pose problems.
* Big data can be collected from publicly shared comments on social networks and websites, voluntarily gathered from personal electronics and apps, through questionnaires, product purchases, and electronic check-ins.
* Big data is most often stored in computer databases and is analyzed using software specifically designed to handle large, complex data sets.

**Challenges in Handling Big Data**:-

**Challenge #1:** Insufficient understanding and acceptance of big data

**Solution:** To ensure big data understanding and acceptance at all levels, IT departments need to organize numerous trainings and workshops.

**Challenge #2:** Confusing variety of big data technologies

**Solution:** If you are new to the world of big data, trying to seek professional help would be the right way to go.

**Challenge #3:** Paying loads of money

**Solution:** Resorting to **data lakes** or **algorithm optimizations** (if done properly) can save money and migrating to cloud platforms.

**Challenge #4:** Complexity of managing data quality (will all the data be useful For analysis?)

**Solution:** Compare data to the single point of truth Match records and merge them, if they relate to the same entity.

**Challenge #5:** Dangerous big data security holes

**Solution:**The precaution against your possible big data security challenges is

putting security first. It is particularly important at the stage of designing your

solution’s architecture.

**Challenge #6:** Tricky process of converting big data into valuable insights

(There is a dearth of skilled professionals who possess a high level of proficiency in data science.)

**Solution:**The idea here is that you need to create a proper system of factors and data sources, whose analysis will bring the needed insights, and ensure that nothing falls out of scope.

**Challenge #7:** Troubles of up scaling

Solution: The first and foremost precaution for challenges like this is a decent

architecture of your big data solution.

**Challenge #8:** The other challenge is to decide on the period of retention of Bigdata

Solution: Proper planning for financial and support.

**Other Characteristics of Data Which are not Definitional Traits of Big Data**:-

**There are yet other characteristics of data which are not necessarily the Definitional Traits of Big Data**

* **Veracity:** It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control. Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.

***Example:***Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

* **Variability:**How fast or ,available data that extent is the structure of your data is changing ?.

How often does the meaning or shape of your data to change?

**Data flows can be highly inconsistent with periodic peaks**

* **Volatility:** How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?
* **Validity:** Similar to veracity, validity refers to how accurate and correct the data is for its intended use. According to Forbes, an estimated 60 percent of a data scientist's time is spent cleansing their data before being able to do any analysis. The benefit from big data analytics is only as good as its underlying data, so you need to adopt good data governance practices to ensure consistent data quality, common definitions, and metadata.

**Why Big Data**?:-

Big Data initiatives were rated as “extremely important” to 93% of companies. Leveraging a Big Data analytics solution helps organizations to unlock the strategic values and take full advantage of their assets.

It helps organizations:

* To understand Where, When and Why their customers buy
* Protect the company’s client base with improved loyalty programs
* Seizing cross-selling and upselling opportunities
* Provide targeted promotional information
* Optimize Workforce planning and operations
* Improve inefficiencies in the company’s supply chain
* Predict market trends
* Predict future needs
* Make companies more innovative and competitive
* It helps companies to discover new sources of revenue

Companies are using Big Data to know what their customers want, who are their best customers, why people choose different products. The more a company knows about its customers, the more competitive it becomes.

Advertisement

We can use it with Machine Learning for creating market strategies based on predictions about customers. Leveraging big data makes companies customer-centric.

Companies can use Historical and real-time data to assess evolving consumers’ preferences. This consequently enables businesses to improve and update their marketing strategies which make companies more responsive to customer needs.

**Analyzing Data with Unix tools**:-

* Data analysis inspects, cleans, transforms, and models data to extract insights and support decision-making.
* The classic tool for processing line-oriented data is *awk.*

Example:

*1) finding the maximum recorded temperature by year from NCDC weather records*

*#!/usr/bin/env bash*

**for** year in all/\*

**do**

echo -ne `basename $year .gz`"\t"

gunzip -c $year | **\**

awk '{ temp = substr($0, 88, 5) + 0;

q = substr($0, 93, 1);

if (temp !=9999 && q ~ /[01459]/ && temp > max) max = temp }

END { print max }'

**done**

* The script loops through the compressed year files, first printing the year, and then processing each file using *awk.*
* The *awk* script extracts two fields from the data: the air temperature and the quality code.
* The air temperature value is turned into an integer by adding 0.
* Next, a test is applied to see whether the temperature is valid (the value 9999 signifies a missing value in the NCDC dataset) and whether the quality code indicates that the reading is not suspect or erroneous.
* If the reading is OK, the value is compared with the maximum value seen so far, which is updated if a new maximum is found.
* The END block is executed after all the lines in the file have been processed, and it prints the maximum value.
* Here is the beginning of a run:

% **./max\_temperature.sh**

1901 317

1902 244

1903 289

1904 256

1905 283…

The temperature values in the source file are scaled by a factor of 10, so this works out as a maximum temperature of 31.7°C for 1901 (there were very few readings at the beginning of the century, so this is plausible). The complete run for the century took 42 minutes in one run on a single EC2 High-CPU Extra Large instance.

To speed up the processing, we need to run parts of the program in parallel.

we could process different years in different processes, using all the available hardware threads on a machine.

Problems:

1. Dividing the work into equal-size pieces isn’t always easy or obvious.
2. combining the results from independent processes may require further processing.
3. you are still limited by the processing capacity of a single machine.

So, although it’s feasible to parallelize the processing, in practice it’s messy. Using a

framework like Hadoop to take care of these issues is a great help.

**Analyzing Data with Hadoop:-**

**Map and Reduce**

MapReduce works by breaking the processing into two phases:

* + 1. Map Phase
    2. Reduce Phase
* Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer.
* The programmer also specifies two functions:
  + the map function.
  + the reduce function.
* The input to our map phase is the raw NCDC data.
* We choose a text input format that gives us each line in the dataset as a text value.
* The key is the offset of the beginning of the line from the beginning of the file, but as we have no need for this, we ignore it.
* We pull out the year and the air temperature, because these are the only fields we are interested in.
* The map function is just a data preparation phase, setting up the data in such a way that the reduce function can do its work on it: finding the maximum temperature for each year.
* The map function is also a good place to drop bad records: here we filter out temperatures that are missing, suspect, or erroneous.
* To visualize the way the map works, consider the following sample lines of input data (some unused columns have been dropped to fit the page, indicated by ellipses):

0067011990999991950051507004...9999999N9+00001+99999999999...

0043011990999991950051512004...9999999N9+00221+99999999999...

These lines are presented to the map function as the key-value pairs:

(0, 0067011990999991950051507004…9999999N9+00001+99999999999…)

(106, 0043011990999991950051512004…9999999N9+00221+99999999999…)

**Map Phase:**

The keys are the line offsets within the file, which we ignore in our map function. The map function merely extracts the year and the air temperature (indicated in bold text), and emits them as its output (the temperature values have been interpreted as integers):

(1950, 0)

(1950, 22)

(1950, −11)

(1949, 111)

(1949, 78)

The output from the map function is processed by the MapReduce framework before being sent to the reduce function. This processing sorts and groups the key-value pairs by key. So, continuing the example, our reduce function sees the following input:

(1949, [111, 78])

(1950, [0, 22, −11])

**Reduce Phase**

Each year appears with a list of all its air temperature readings. All the reduce function has to do now is iterate through the list and pick up the maximum reading:

(1949, 111)

(1950, 22)

This is the final output: the maximum global temperature recorded in each year.



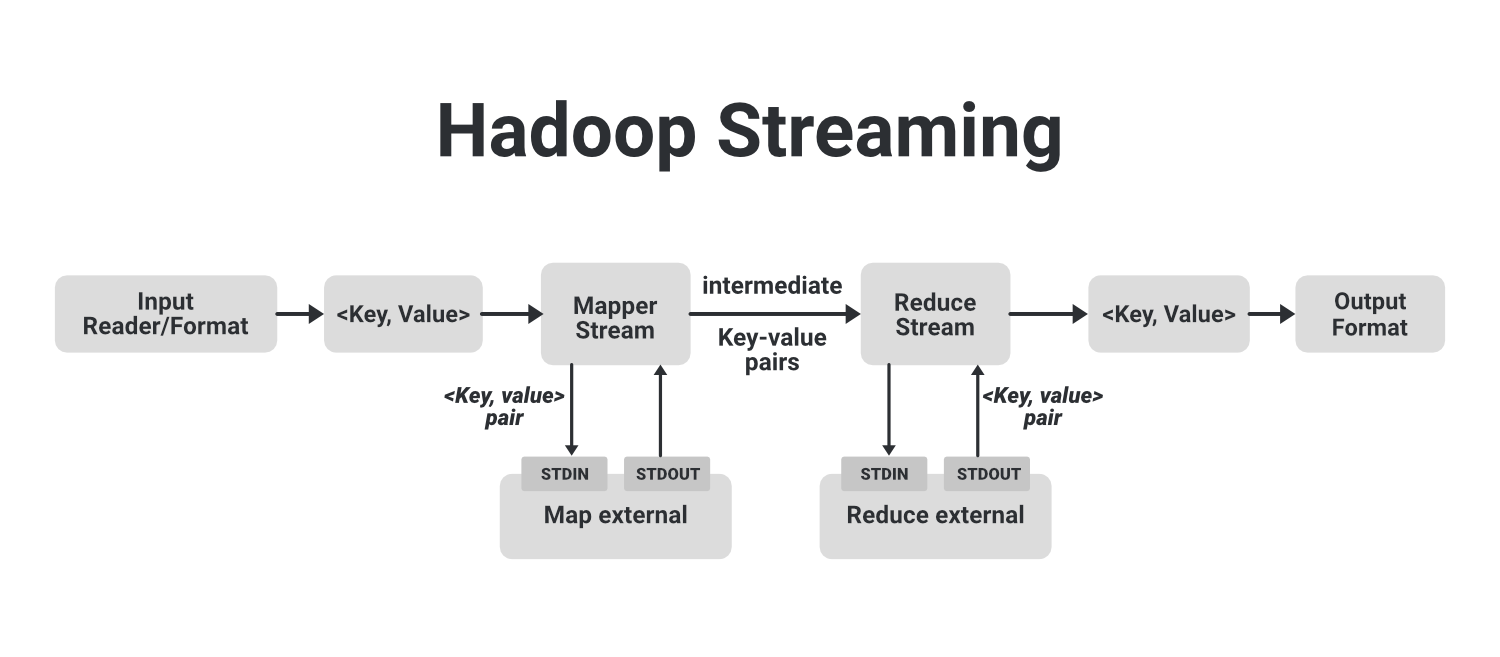
**Hadoop Streaming:-**

* *Hadoop Streaming* uses Unix standard streams as the interface between Hadoop and your program, so you can use any language that can read standard input and write to standard output to write your MapReduce program.

**Unixstd.io**

**Hadoop**

* Streaming is naturally suited for text processing.
* Map input data is passed over standard input to your map function, which processes it line by line and writes lines to standard output.
* A map output key-value pair is written as a single tab-delimited line.
* Input to the reduce function is in the same format — a tab-separated key-value pair — passed over standard input.
* The reduce function reads lines from standard input, which the framework guarantees are sorted by key, and writes its results to standard output.



* A utility that allows one to run map/reduce jobs on the cluster with any executable.
* Executables reads input from STDIN and writes output to STDOUT.
* Maps input line data from STDIN (key, value) pairs in STDOUT in our case.
* Reducer task convert I/P (k,v) pair into lines and feeds the lines to STDIN.
* Reducer gather the lines and generate O/P from STDOUT and convert each line into(k,v) pair result of reducer

Advantages of Hadoop streaming:

1. Availability – No Separate software to install.
2. Learning – doesn’t require to learn new technology.
3. Faster development time.
4. Faster conversion takes less time to convert data from one format to another format.
5. Testing – quickly tested using Unix shell scripts.
6. Requirement for business (simple filtering operations, aggregations)
7. Ability to use the existing code library.

Disadvantages of Hadoop streaming:

1. Performance
2. Suited for handling data that can be represented as text.
3. Uses excessive amount of RAM or fork excessive no of processes.
4. MR written using streaming.

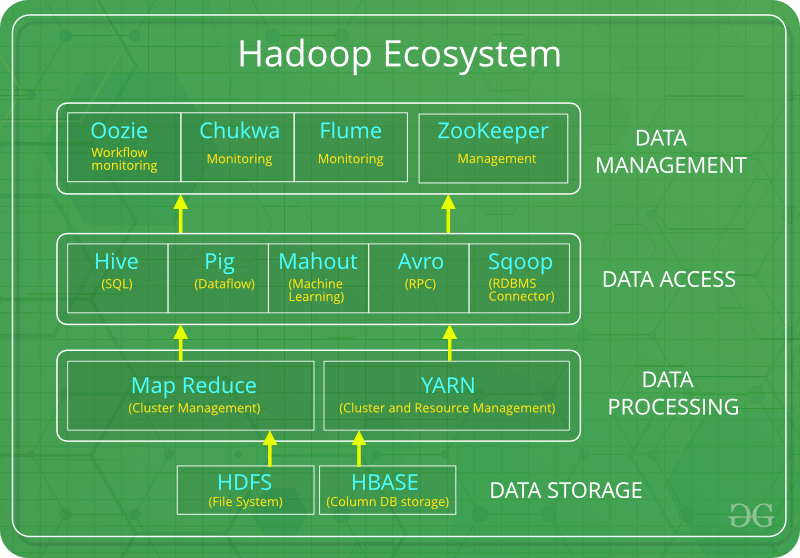
**Hadoop Echo System**:-

 Apache Hadoop is an open source framework intended to make interaction with [**big data**](https://write.geeksforgeeks.org/geek/the-world-of-big-data/) easier, However, for those who are not acquainted with this technology, one question arises that what is big data ? Big data is a term given to the data sets which can’t be processed in an efficient manner with the help of traditional methodology such as RDBMS. Hadoop has made its place in the industries and companies that need to work on large data sets which are sensitive and needs efficient handling. Hadoop is a framework that enables processing of large data sets which reside in the form of clusters. Being a framework, Hadoop is made up of several modules that are supported by a large ecosystem of technologies.

**Introduction:** Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. **HDFS, MapReduce, YARN, and Hadoop Common**. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

Following are the components that collectively form a Hadoop ecosystem: 

* **HDFS:**Hadoop Distributed File System
* **YARN:** Yet Another Resource Negotiator
* **MapReduce:** Programming based Data Processing
* **Spark:** In-Memory data processing
* **PIG, HIVE:** Query based processing of data services
* **HBase:**NoSQL Database
* **Mahout, Spark MLLib:** [Machine Learning](https://www.geeksforgeeks.org/machine-learning/)algorithm libraries
* **Solar, Lucene:** Searching and Indexing
* **Zookeeper:** Managing cluster
* **Oozie:** Job Scheduling



**Note:** Apart from the above-mentioned components, there are many other components too that are part of the Hadoop ecosystem.

All these toolkits or components revolve around one term i.e. *Data*. That’s the beauty of Hadoop that it revolves around data and hence making its synthesis easier.

* **HDFS:**
* HDFS is the primary or major component of Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.
* HDFS consists of two core components i.e.
  1. Name node
  2. Data Node
* Name Node is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that stores the actual data. These data nodes are commodity hardware in the distributed environment. Undoubtedly, making Hadoop cost effective.
* HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.
* **YARN:**
* Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.
* Consists of three major components i.e.
  1. Resource Manager
  2. Nodes Manager
  3. Application Manager
* Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager. Application manager works as an interface between the resource manager and node manager and performs negotiations as per the requirement of the two.
* **MapReduce:**
* By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing’s logic and helps to write applications which transform big data sets into a manageable one.
* MapReduce makes the use of two functions i.e. Map() and Reduce() whose task is:
  1. *Map()* performs sorting and filtering of data and thereby organizing them in the form of group. Map generates a key-value pair based result which is later on processed by the Reduce() method.
  2. *Reduce()*, as the name suggests does the summarization by aggregating the mapped data. In simple, Reduce() takes the output generated by Map() as input and combines those tuples into smaller set of tuples.
* **PIG:**

 Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.

* It is a platform for structuring the data flow, processing and analyzing huge data sets.
* Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.
* Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the [JVM](https://www.geeksforgeeks.org/jvm-works-jvm-architecture/).
* Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

**HIVE:** 

* With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).
* It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.
* Similar to the Query Processing frameworks, HIVE too comes with two components: *JDBC Drivers* and *HIVE Command Line*.
* JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.
* **Mahout:**
* Mahout, allows Machine Learnability to a system or application. [Machine Learning](https://write.geeksforgeeks.org/geek/ml-what-is-machine-learning-2/), as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
* It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.
* **Apache Spark:**
* It’s a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.
* It consumes in memory resources hence, thus being faster than the prior in terms of optimization.
* Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.
* **Apache HBase:**
* It’s a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google’s BigTable, thus able to work on Big Data sets effectively.
* At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data