Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

(PROJECT REPORT - BIG DATA TRAINING)

Project report submitted in fulfillment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering/Information Technology

By

Aditya Agarwal (171479)

Under the supervision of
(Dr. Monika Bharti)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

TABLE OF CONTENT

Certificate

Candidate's Declaration

We hereby declare that the work presented in this report entitled **"Big data Training"** in fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** **"Jaypee University of Information Technology, Waknaghat**" is an authentic record of our own work carried out over a period from **January,2021** to **May, 2021** under the supervision of **Dr. Monika Bharti** (Associate Professor – CSE & IT).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)

Aditya Agarwal (171479)

d Aditya .

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Monika Bharti

(Supervisor Signature)
Dr. Monika Bharti
Associate Professor
Department of Computer Science and Engineering & Information Technology

Dated: May 20, 2021

ABSTRACT

Data is considered to be the elixir of information technology field. For every process or decision, that needs to be taken, we require data. Also, every process or transaction that takes place also generates data. Earlier, traditional methods of data storage and processing were used to deal with data. But right now, the extent of data production and the rate at which data needs to be dealt with requires advanced tools and technologies. Big data technologies are a solution, which provide faster and less costly methods to store and deal with data that is being generated and that is required. It divides the data storage with the help of distributed file systems and in order to process this data, we use the different commodity hardware to reduce and divide the computation effort for each machine. The hadoop architecture is primarily made of two components called as hadoop file distributed system and map reduce. Other tools like sqoop, hive, pig and Kafka are used to import, export, process, visualize data and generate reports which assist with the decision making processes in organizations.

LIST OF IMAGES

COMPANY PROFILE

Cognizant is a leading American multinational firm, which provides its services in business consulting, information technology, system integration, artificial intelligence, digital engineering, analytics, business intelligence, data warehousing etc. It initially began as Dun & Bradstreet Software in January 1994, established as Dun & Bradstreet's in-house unit for providing IT-infrastructure related services for Dun & Bradstreet business, but later expanded its client base from 1996.

Cognizant's digital business, operations and systems and technology are the three areas which make up their business profile. To provide technological proficiency to its clients Cognizant is organized into various verticals and horizontals. The verticals focus on specific industries like- Banking and Financial Services, Insurance, Healthcare, Manufacturing and Retail services etc.

The horizontals on the other hand focus on specific technologies and services like - Analytics, mobile computing, BPO and testing solutions. It follows a business model similar to other IT giants based on, global delivery model, which is based upon offshore software R&D and offshore outsourcing.

The first time Cognizant came in the Fortune 500 list was in 2011. In 2015, Fortune named Cognizant as the world's 4th most admired IT Services Company. It currently ranks 194 in Fortune 500 companies, 533 in Forbes Global 2000, 483 in Forbes Best Employers for Diversity in 2019.

Cognizant is among the high scientific discipline corporation that has been delivering high quality IT-infrastructure services and Business Intelligence services, extending to a list of happy clients worldwide. With various teams of highly proficient and hardworking associates working 24*7 to deliver high standard results and speedy turnarounds, it has been helping its clients in increasing their business potency.

CHAPTER - 01

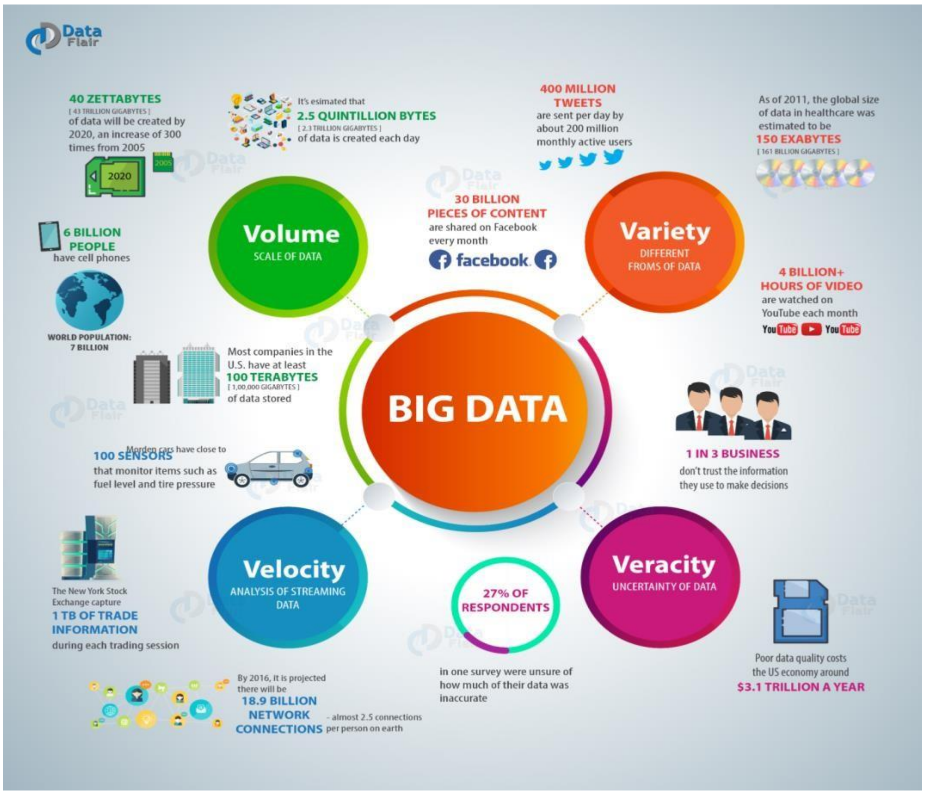
INTRODUCTION TO ASSIGNED WORK

DATA:

Data is defined as the quantities, characters or symbols upon which operations are performed by the computer, which might be stored or transmitted as electrical signals and stored as mechanical recording.

BIG DATA:

Big data is defined as the huge collection data in terms of volume, yet also growing with time, exponentially. This data being so complex in nature those traditional data-handling solutions find it difficult to store and process this data.



40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's esimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data is created each day

400 MILLION TWEETS are sent per day by about 200 million monthly active users

As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]

6 BILLION PEOPLE have cell phones

30 BILLION PIECES OF CONTENT are shared on Facebook every month

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

Most companies in the U.S. have at least 100 TERABYTES [1,00,000 GIGABYTES] of data stored

Morden cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

1 IN 3 BUSINESS don't trust the information they use to make decisions

The New York Stock Exchange capture 1 TB OF TRADE INFORMATION during each trading session

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

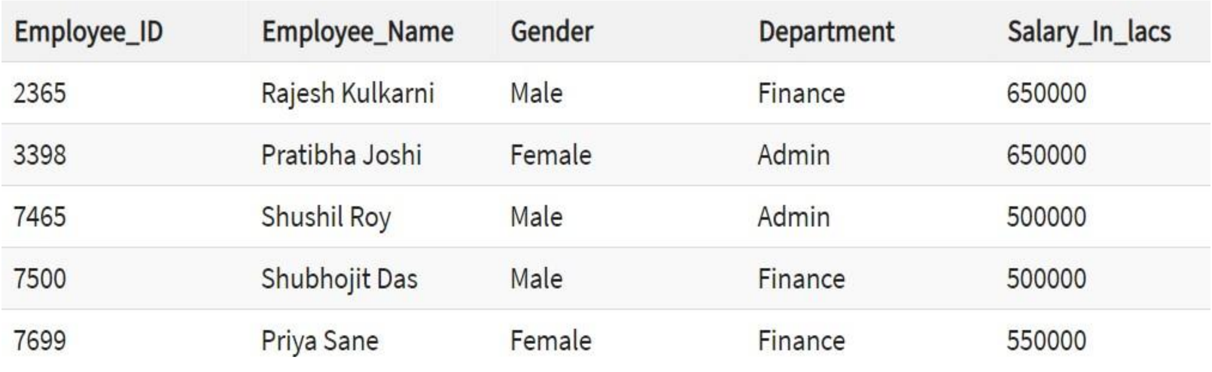
Poor data quality costs the US economy around $3.1 TRILLION A YEAR

By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS - almost 2.5 connections per person on earth

TYPES OF BIG DATA:

Structured:

This type of data can be stored, assessed or processed as fixed format, hence the name Structured data. The avent in computer science has made it easier to work with structured data as the format of the data is mostly well-known in advance,so deriving meaningful insights from it is easier. Recently, we have been facing issues to handle the huge amount of data, where a typical data size is about a zettabyte (one billion terabytes).

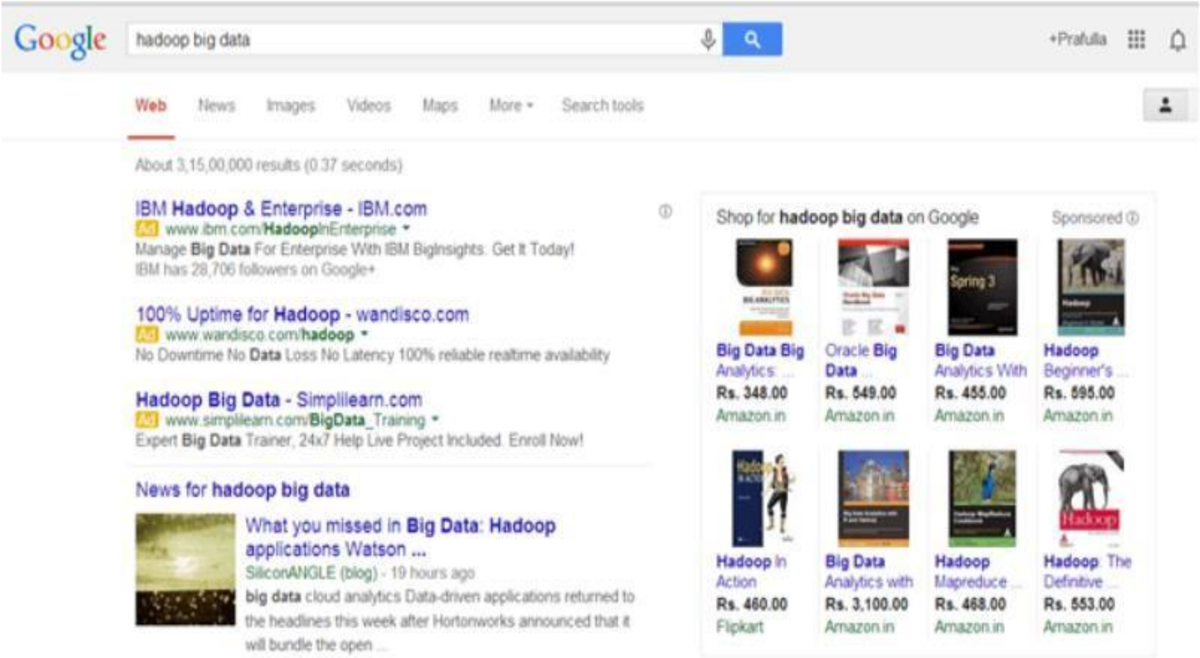


Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured:

It is a type of data having any unknown form or structure, hence the name unstructured data. In addition to its large size, what makes it difficult to process is it has various challenges, for e.g. since it is from a heterogeneous source, the data files may include anything ranging from simple text files to mp3, mp4, jpg, fly, avi etc.

Organizations might have a huge collection of unstructured data, but face difficulties in extracting meaningful insights from it, since it is in the raw form and difficult to process.



About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/HadoopInEnterprise ▾
Manage Big Data For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/hadoop ▾
No Downtime No Data Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/BigData_Training ▾
Expert Big Data Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

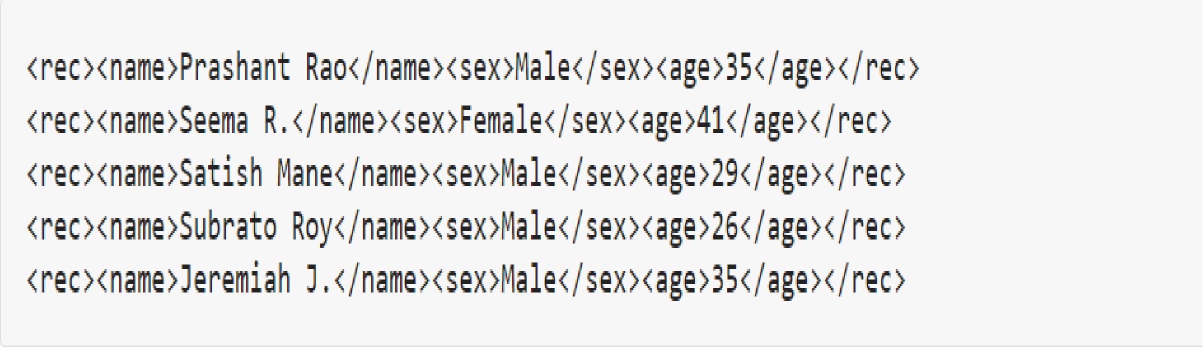
What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

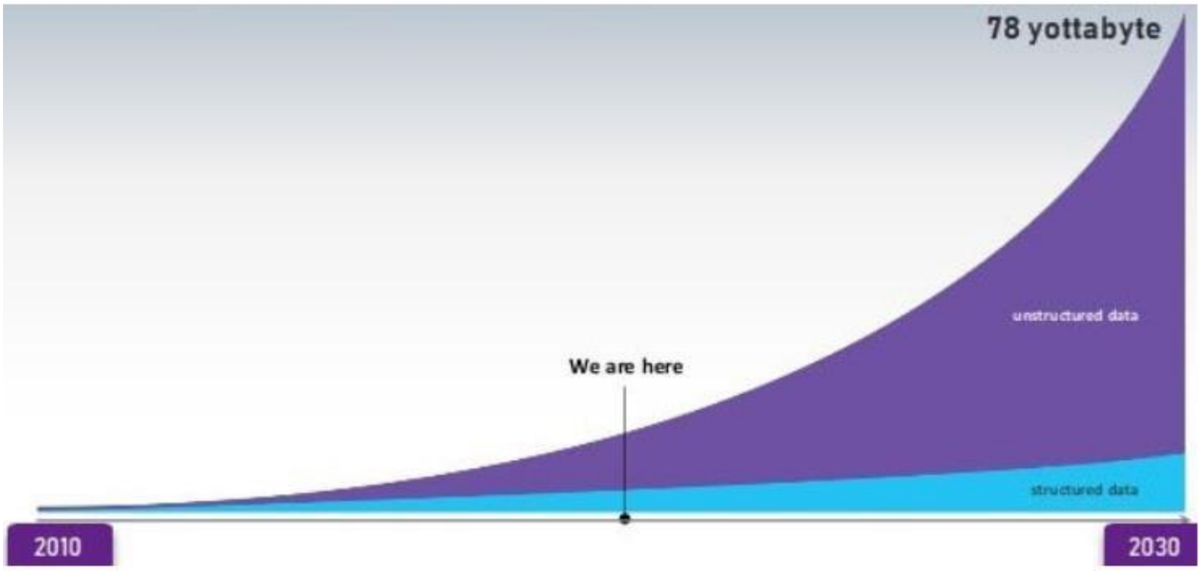
Shop for hadoop big data on Google — Sponsored

Big Data Big Analytics: ... Rs. 348.00 Amazon.in
Oracle Big Data ... Rs. 549.00 Amazon.in
Big Data Analytics With ... Rs. 455.00 Amazon.in
Hadoop Beginner's ... Rs. 595.00 Amazon.in
Hadoop In Action Rs. 460.00 Flipkart
Big Data Analytics with ... Rs. 3,100.00 Amazon.in
Hadoop Mapreduce ... Rs. 468.00 Amazon.in
Hadoop: The Definitive ... Rs. 553.00 Amazon.in

Semi-structured:

This type of data can contain both types of data that is unstructured and structured. It is usually characterized by being structured, but not defined in a tabular structure as defined by the tables in relational databases.


```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```



CHARACTERISTICS OF BIG DATA

1.) **Volume-**

It is one of the major characteristics for defining big data, since volume plays a crucial role in extracting out meaningful insights from the data. It is volume which decides if a particular set of data can be considered as big data or not.

2.) **Variety-** Variety in data arises due to heterogeneous sources of data and the nature of data. Previously spreadsheets and databases happened to be the only sources of data, considered by majority of the applications. In recent times data is obtained in various forms like images, movies, emails, monitoring-devices, audios, pdf etc. are also being included for carrying out analysis. This variety of unstructured data poses a serious challenge in storage, mining and analysis of the data.

3.) **Velocity-** Velocity refers to the rate at which the data gets generated. The real potential of the data is determined by the fact, how fast is the data generated, stored and processed for deriving value out of it. Big data velocity usually deals with the speed at which data is being generated from various sources including application logs, business processes, sensors, mobile devices, ioT devices, social media etc. The stream of data flow is continuous and humongous.

4.) **Variability-** It is defined as the inconsistency of data thus making it harder to process and derive value out of it.

ADVANTAGES OF BIG DATA:

The capability to process big data brings in certain advantages:

1.) Businesses can use outside intelligence to make decisions.

2.) Better operational frequency.

3.) Better risk management, by means of early risk identification related to products or services.

4.) Improved customer services, by means of revamped feedback systems over the traditional feedback evaluation systems.

INTRODUCTION TO SKILL TRACK REQUIRED FOR BIG DATA:

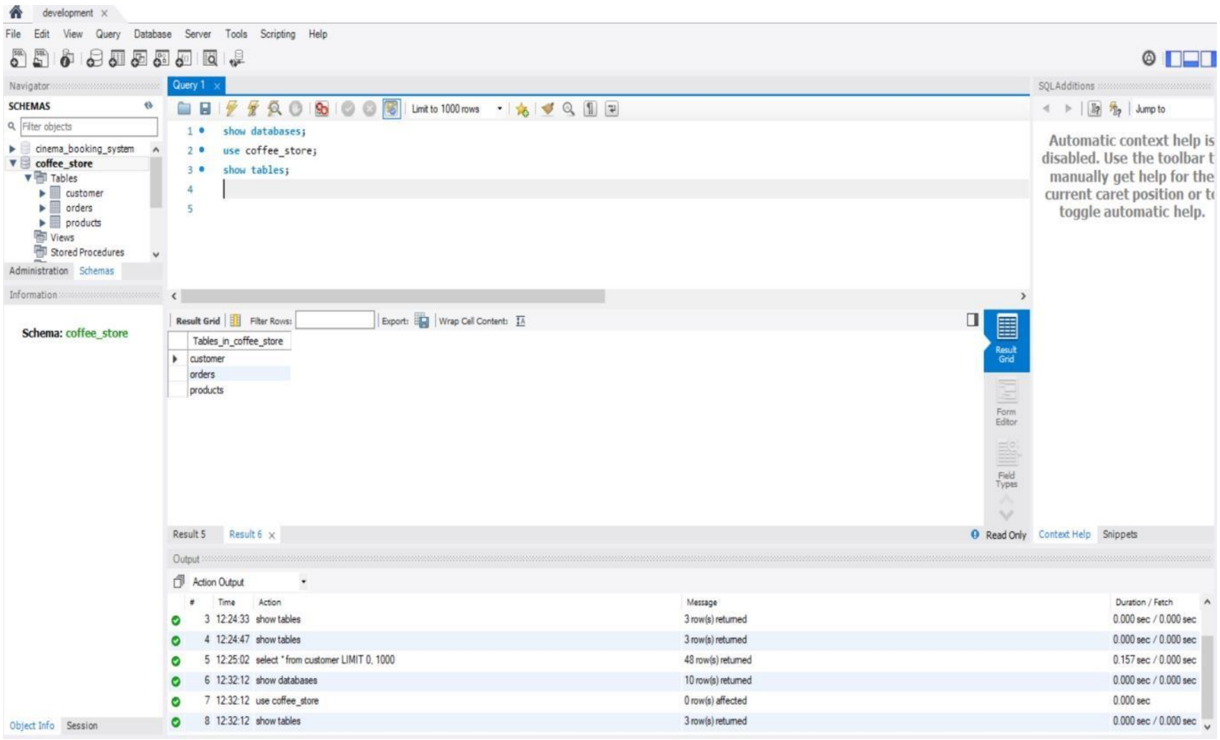
SQL-

What is a database?

A database is defined as an organized collection of data which is usually stored and accessed from a computer system. Usually relational databases are used to store and retrieve information.

A relational database is based upon a relational model of data as the name suggests. In this model the data being stored is organized into one or multiple tables consisting of various rows and columns, with each row having its own unique identifier.

The database management system is software which is used to interact with the databases.



```
show databases;
use coffee_store;
show tables;
```

Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.

Tables_in_coffee_store
customer
orders
products

#	Time	Action	Message	Duration / Fetch
3	12:24:33	show tables	3 row(s) returned	0.000 sec / 0.000 sec
4	12:24:47	show tables	3 row(s) returned	0.000 sec / 0.000 sec
5	12:25:02	select * from customer LIMIT 0, 1000	48 row(s) returned	0.157 sec / 0.000 sec
6	12:32:12	show databases	10 row(s) returned	0.000 sec / 0.000 sec
7	12:32:12	use coffee_store	0 row(s) affected	0.000 sec
8	12:32:12	show tables	3 row(s) returned	0.000 sec / 0.000 sec

What is SQL?

SQL stands for Structured Query Language. It is designed for use in performing management operations in relational databases.

Most database management systems like Mysql, Oracle, Sybase, SQL Server and Informix etc use SQL primarily.

The two main differences in the SQL databases with comparison to the traditional read write operation file systems are-

1) First, it solves the problem of accessing multiple files at the same time
2) Secondly, it eliminates the need of specification of the usage of indexes

Sql was initially developed by IBM in the early 1970s, by Donald Chamberlain and Raymond Boyce. System R, which was the original IBM database at that time was being operated and manipulated by the original version of SEQUEL (Structured English query language) which was also developed at the same San Jose laboratory in the early 1970s.

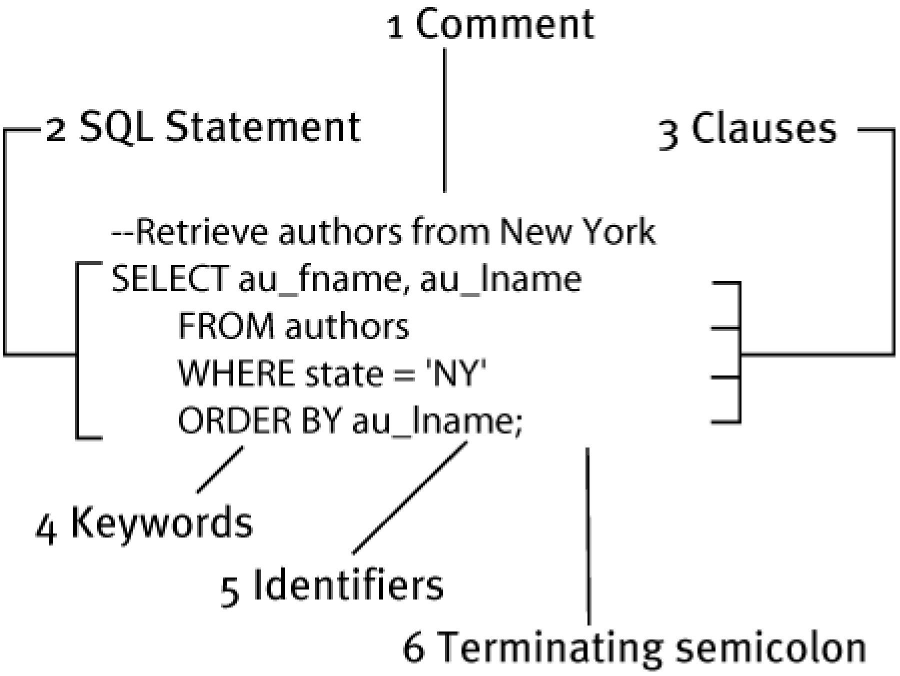
In the late 1970s, relational software, which is now known as the oracle corporation saw the potential of the system that IBM talked about and had created, so they themselves started research and development on the same and thus developed the first commercialized version of the SQL, the oracle v2 which was created to be used on the vax systems. The system was developed with the vision of selling it to the US military operations, US navy and the central intelligence agency along with many other prominent US government departments.

SYNTAX

There are multiple different constituents of the sql language, made for making it usable, readable and easier to learn and use. The various components are listed as below-

1. Clauses - clauses might or might not be used in a sql query. Some of the queries might even use multiple clauses and some might not require the use of even a single clause. Thus, these constituents of statements/expressions or queries are actually optional but very useful when required and used.

2. Expressions - expressions are the primary statements in the sql language. These are used for various basic and advanced operations which can be done by the sql language. They may be written in order t o generate some scalar values or even might return columns or rows or both combined, which are basically called tables.



3. Keywords - the various keywords or predefined words are used in sql which provide a helping hand while using sql for various purposes. Some of the keywords that are used in sql are listed as follows -
   - Order by

- Select
- Where
- From

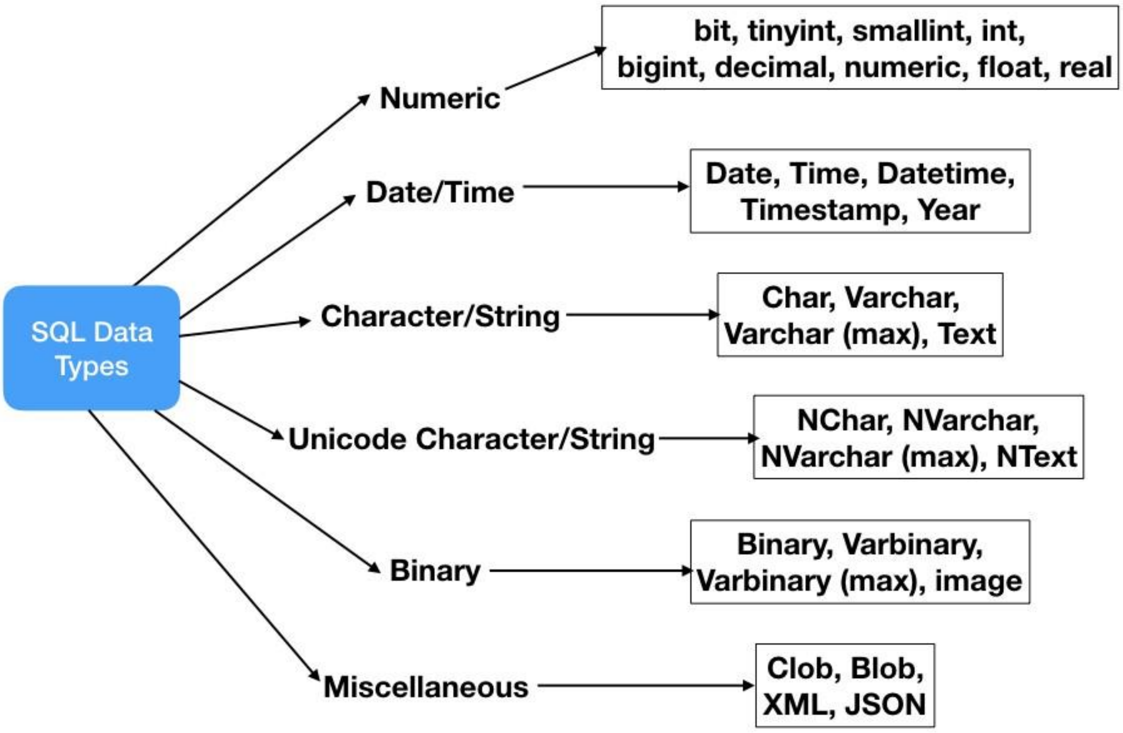
4. Statements - statements in sql may be used to control the flow, connections or diagnostics and may control the transactions etc of the program.

5. Comments - Single line comments start with --. Any text between -- and the end of the line will be ignored by the system.
Multi-line comments - they both start and end with /* and any text between /* and */ will be ignored by the system.

SQL DATA TYPES

Data types are mainly classified into three categories for every database.

- String-Data Types
- Numeric-Data types
- Date, time-Data types


SQL Data Types
Numeric
bit, tinyint, smallint, int, bigint, decimal, numeric, float, real
Date/Time
Date, Time, Datetime, Timestamp, Year
Character/String
Char, Varchar, Varchar (max), Text
Unicode Character/String
NChar, NVarchar, NVarchar (max), NText
Binary
Binary, Varbinary, Varbinary (max), image
Miscellaneous
Clob, Blob, XML, JSON

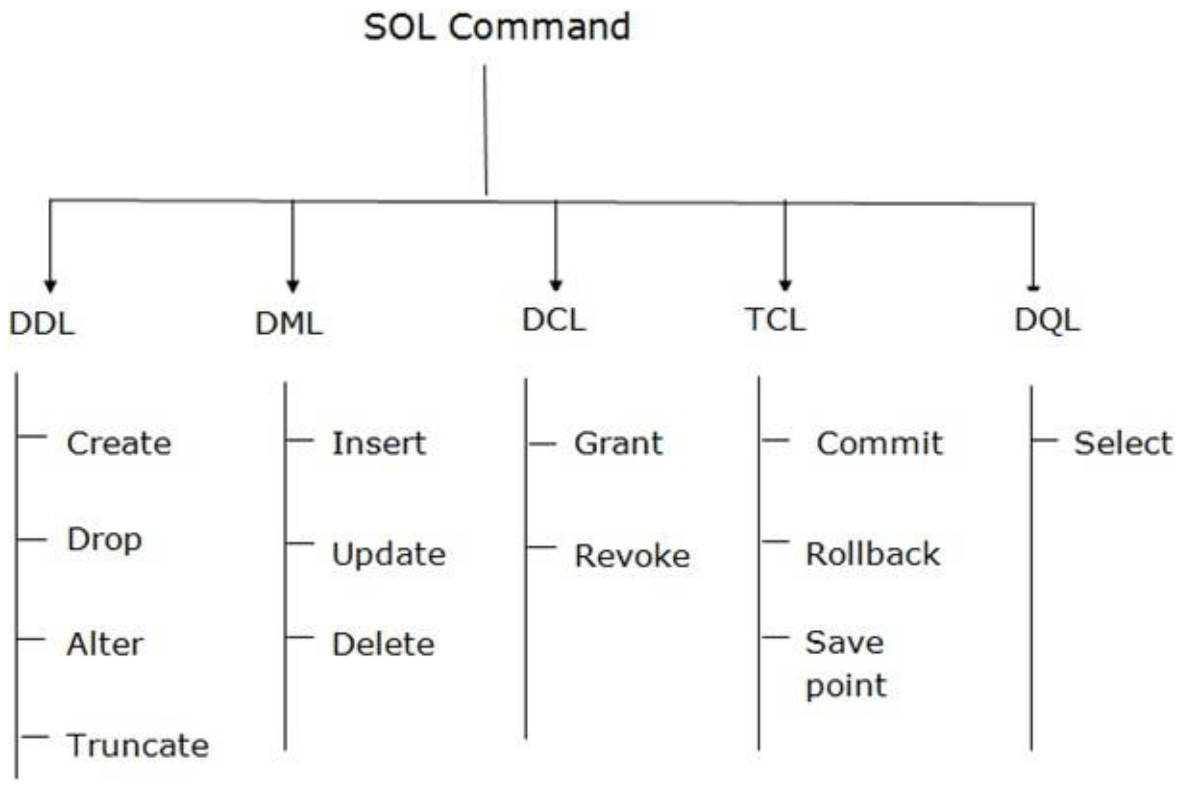
SQL DATABASE COMMANDS

- CREATE

- this is used to create a new database in the system, inside which, the multiple tables can be created and then manipulated
    - Create database database name;
    - Example - create database student_db;

- DROP - to delete an existing database in the sql schema
    - Drop database database name;
    - Example - drop database srudent_db;

Numerous commands similar to these are used in sql. A compiled summary of such commands is shown in the below picture.



DML commands

The manipulation of data that is present inside the various databases and tables inside the sql schema is done using the DML commands. There are multiple DML commands which can be used for the above stated purposes. These are known as data manipulation language commands. Following are some examples of DML commands -

- Insert - it is used to enter data into the table that already exists. The syntax for the same is as follows -
  INSERT INTO table name (col1, col2, col3) values (val1, val2, val3);

Here, the col1 etc are the columns listed and the corresponding values and val1 etc are the values that need to be entered. We can also use the same command without the column values.

- Update - it is used to modify the existing records -
  Update table name
  Set col1 = val2
  Where condition;

- Delete - it is used to delete the existing records.
  Delete from table name where condition;

DCL COMMANDS

Two commands used in DCL category are grant and revoke. They basically deal with permissions, rights etc of the users.

- Grant - provides users the access privileges to the database.
- Revoke - restricts users the access privileges to the database.

TCL COMMANDS

Deals with the transactions within the database.
Some examples are -

- Commit
- Rollback
- Set transaction
- Savepoint

JOINS

- Joins are used in SQL in order to retrieve data from multiple tables in a single select query.
- In order to access more than one table, we need to establish a single common column between the tables so that they can be connected to each other with the help of this column.
- This column having a unique value for each of the records in the table is called as the primary key in the parent table.
- The column might be an attribute of any kind and can be used as an unique value for each of the records to distinguish the records from each other.

- The column with which the primary key column matches in the other table is called as a foreign key attribute.

- Foreign key attributes also have unique values in the table and can be used to match the records throughout multiple tables.

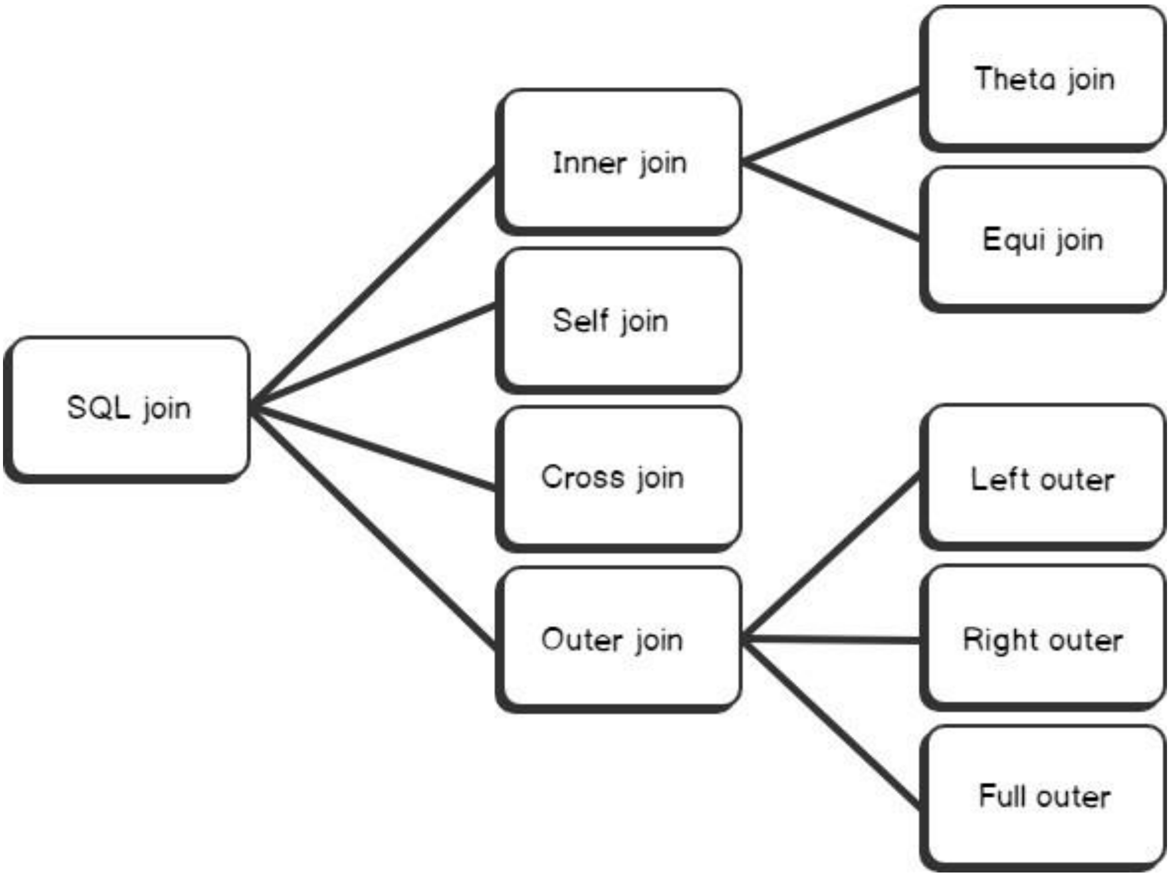
- Foreign key attribute column does not have to have the same name as the primary key and neither does it need to have the exact same values.

- But this column should have the same data type of values so that the columns can be matched with each other in order to meet the join condition

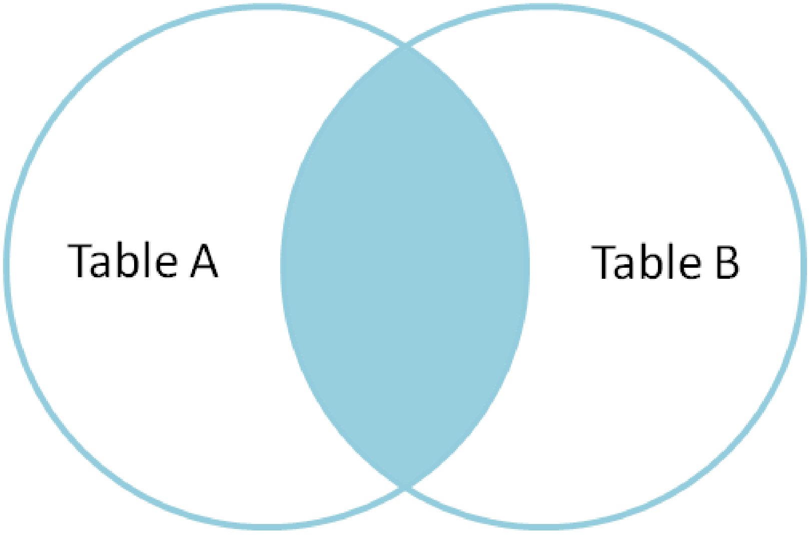
- The two columns might match all the values and for some conditions there might not even be a single match, but the condition can still be evaluated.

TYPES OF JOIN

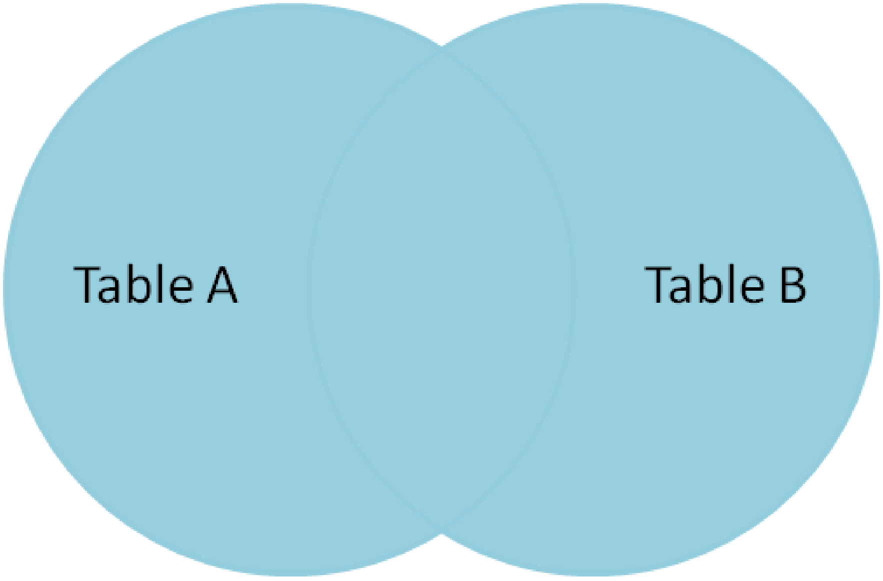
There are multiple types of joins which can be used in sql to implement the usage or retrieval of data from multiple tables at the same time. Some of the joins are demonstrated as follows -



- INNER JOIN
  It gives all the rows as long as the condition satisfies. All the respective rows from all the tables that are being queried will be returned by using the keyword inner join till the condition of join matches. For those cases, in which the condition does not match, we won't get the rows returned.

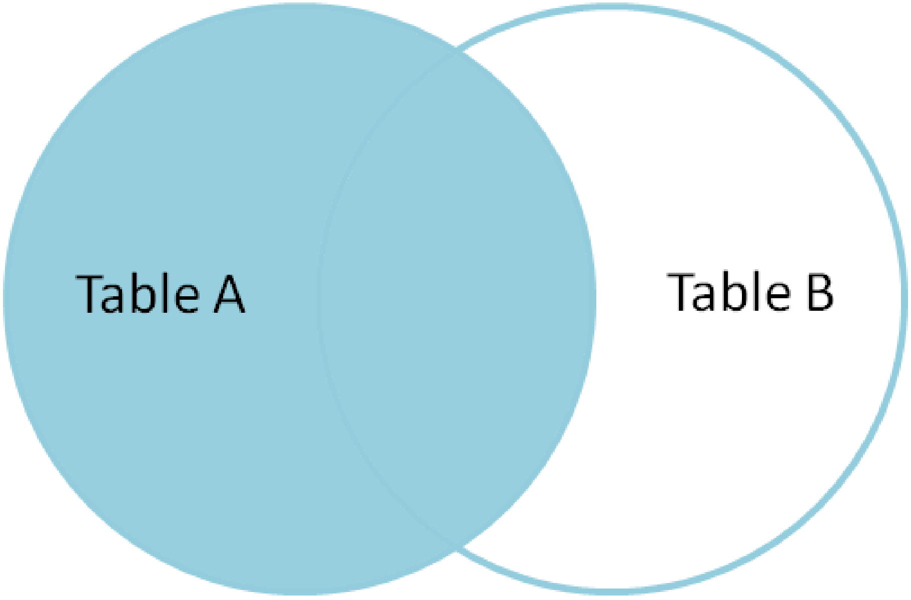


- FULL JOIN
  Returns the result of the full join query by combining the results of both the left join and right join. It contains all the rows from both the tables, irrespective of any matches or no matches, basically retrieves the combined version of both the tables



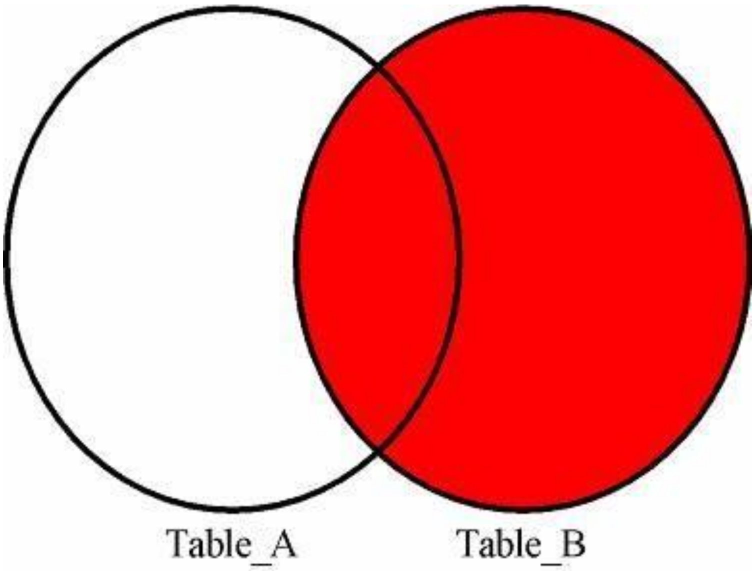
- LEFT JOIN

  This join is used to retrieve all the rows of the table which is written on the left hand side of the join while only the matching rows of the table on the right. The results which don't match have null in the columns from the table on the right. This type of join is also called the left outer join.



- RIGHT JOIN

  This join is used for retrieving the results such that all the columns and rows of the table on the right side of the join are present in the result along with the results from the left table for only the columns which match on the join condition. This join is also called the right outer join.



SUB QUERIES

- Sub queries are queries written nested inside other queries
- They can be used in select, insert, update and delete queries
- The nested part can be inside the from or where clause
- These can be of two types -
  - Non-correlated
  - correlated

Non correlated sub queries - in this type of sub queries, the inner query can run independently of the outer query.

- Inner query runs first and generates a result, which is then used by the outer query.
- It runs only once

Correlated sub queries - the inner query cannot run independently of the outer query but is dependent

- The inner query runs for every row in the outer query
- Might have columns named as a. Column

AGGREGATE FUNCTIONS

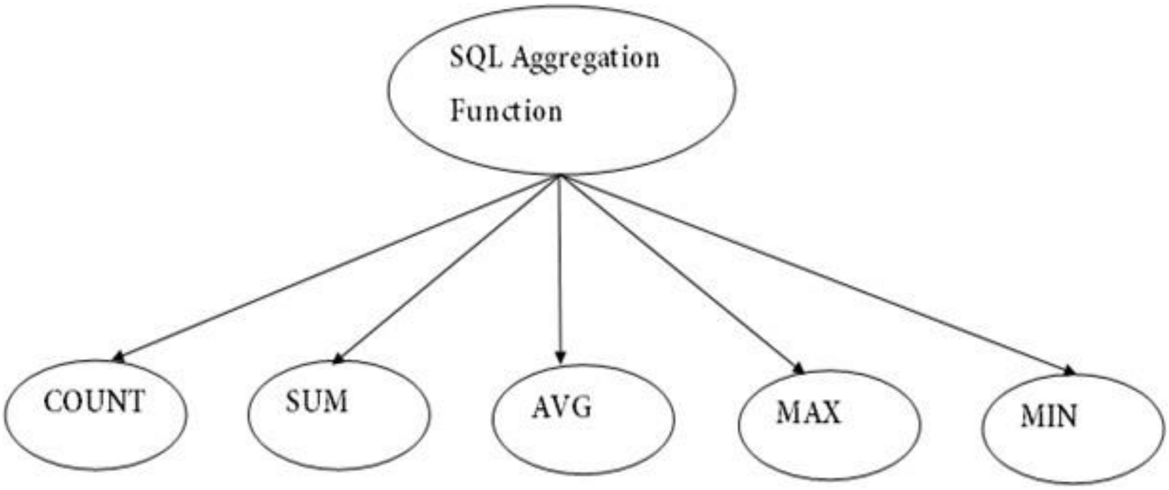
- COUNT - it counts the number of records

- Can use distinct with count to generate the result with unique values
  - Does not count the null values
  - Select count(*) from table name;

- SUM - it sums the values of a column
  - Can't do a sum on *
  - Columns should have summable data types(numeric)
  - Select sum(no seats) from rooms;

- AVERAGE - returns the numerical average of the columns
  - Can’t use where clause
  - Can only select one column to average
  - For multiple columns, use GROUP BY

- MIN and MAX - selecting minimum or maximum value of a column
  - Can use where
  - Can use other columns after the select
  - Gives a single record as result



CHAPTER – 02

JAVA

Java is a type of high level, class-dependent object-oriented programming language developed by James Gosling at Sun Microsystems. It first surfaced in May (23) 1995. Java is now owned by Oracle. Being a general purpose language it is used for application development (client-server web applications). The greatest feature of java was its architecture/platform independence, which means a java code/program written in one machine, can directly be executed in any other machine having java components, without the need of actually recompiling the program on the new machine.

James Gosling initially named the language Oak, based upon an oak tree outside his office. Following this the project was named to Green, and later renamed to Java, based upon the Java coffee from Indonesia. The first iteration of java was released for interactive televisions, but it was far ahead of its time for the digital cable television industry. It was developed with a syntax similar to C/C++ to allow familiarity for the developers.

Principles

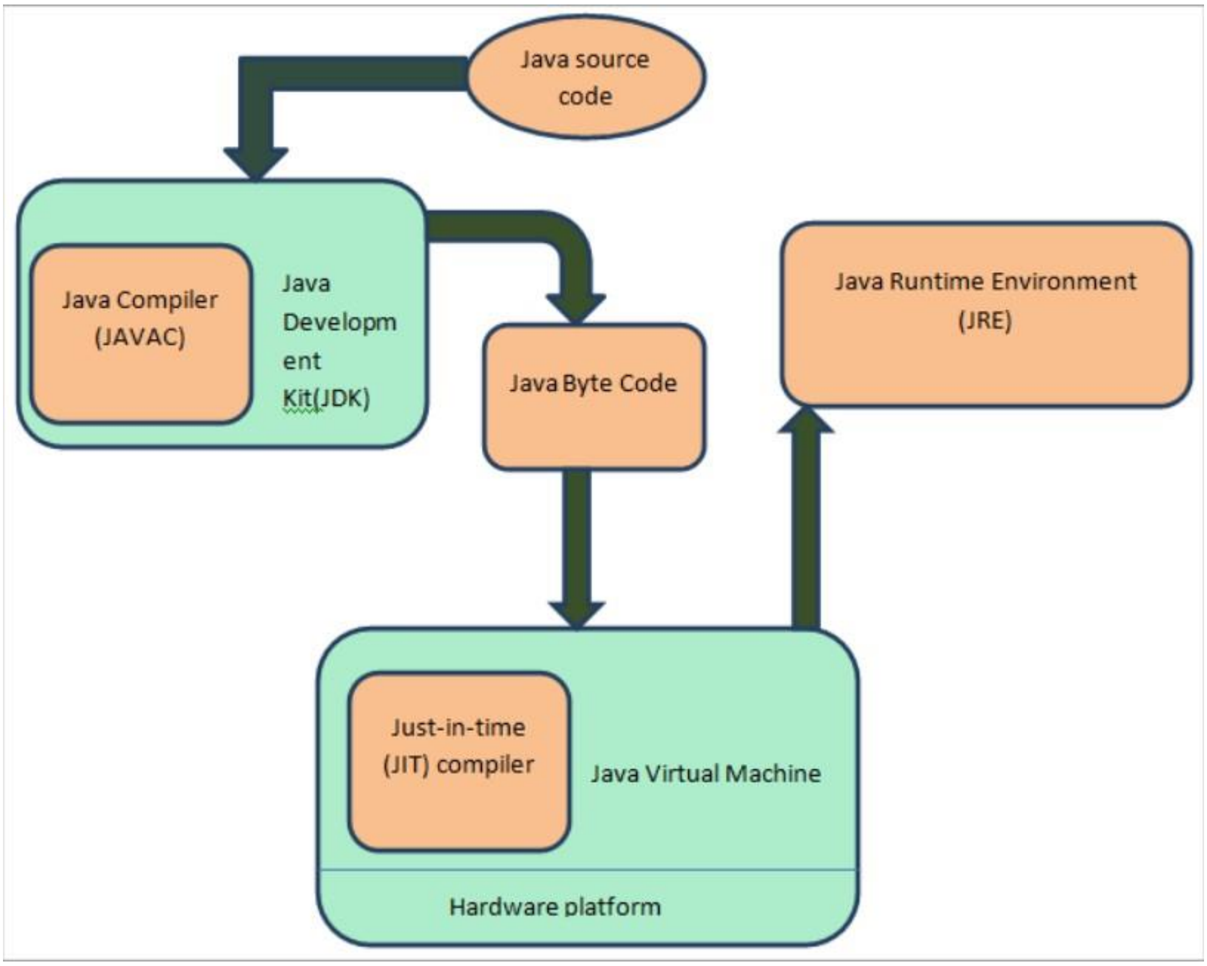
1.) The language must be simple, object-oriented and familiar
2.) It must be platform neutral and portable
3.) It must be secure and robust
4.) It should execute with high performance

5.) It must have the ability to be interpreted, threaded and being dynamic.

Components of Java Language

JAVA DEVELOPMENT KIT - It is the core component, and it contains java compiler, java runtime environment, debugger etc. It is utilized for development purposes since it provides access to all executables and binaries along with other tools required to compile, execute and debug the program. Some of its internal components are:

JConsole- The java management/monitoring console
Javap- A tool for class-files disassemble
Jar- This tool is used to archiving package related libraries into a single file
javadoc- It utilises comments from source code to generate documentation
jrunscript- It is used to help execute java queries from command-line interface



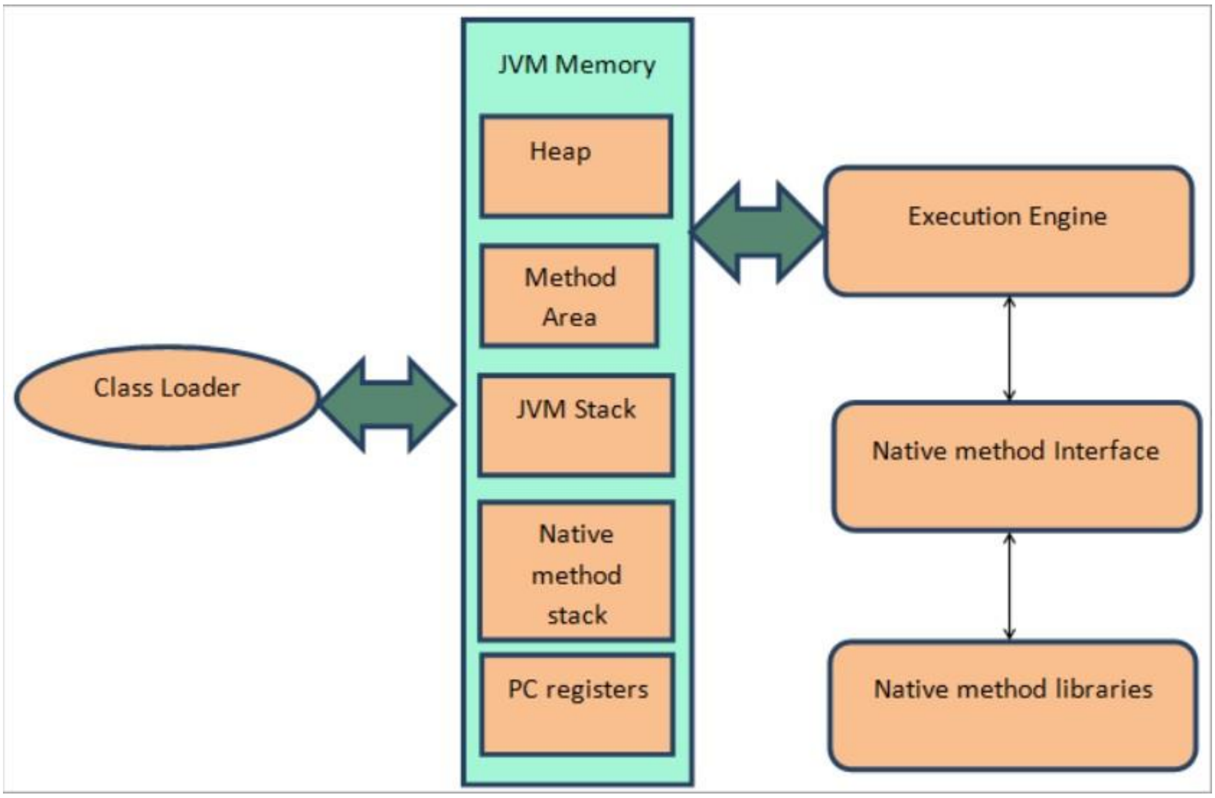
JAVA RUNTIME ENVIRONMENT- It is required for execution of java programs and applications. The JRE consists of components like Java Virtual Machine, which houses the binaries required for successful execution of any java program. Some of its components are: Files needed for management of security reasons.

DLL files
Code libraries, properties/resource files
Java extension files
Applet support files

THE JAVA VIRTUAL MACHINE- The JVM is a core component of the java language. IOT translates the byte-code into code that is understood by the machine. It also provides functionality for automatic memory management, garbage collection, security etc. It is platform independent thus allowing us the flexibility to write a java codes anywhere and executes it anywhere.

The JVM is usually present on RAM, therefore upon conversion of source file to class file, it needs to be executed. The class loader is accountable for the linking, loading and initialization of the program source code to be executed.

JVM also has the Just-In-Time compiler (JIT) which is responsible for the interpretation of a part of the byte code, which has similar functionality at the same time. Hence, Java is both a compiled and interpreted language.



JAVA COMPILER- It is the compiler for the Java programming language and its main function happens to be the conversion of java source code into java class files, following whose generation it is interpreted or compiled by the Java Virtual Machine using the Just In Time (JIT) compiler.

TYPES OF JAVA APPLICATIONS

Standalone- This type of applications is used for desktop/windows based applications, and need to be installed on every machine, e.g. Antivirus softwares.

Enterprise Applications- They are usually distributed in nature like banking applications. They have higher security, clustering, load balancing etc.

Web Applications- These applications run on the server side and create a dynamic page known as web application.

Mobile Application- These include applications created for running on mobile devices.

JAVA EDITIONS

Java SE- This is the standard edition, and contains all the java programming API's like java.sql, java.lang etc and other core stuff of OOP's like regex, multi-threading etc.

Java EE- The Enterprise Edition is used to develop applications for enterprises and web applications. This is based over the Standard Edition.

Java ME- The Micro Edition, this platform is used for developing mobile applications.

JavaFX- Used for developing richer web/internet applications.

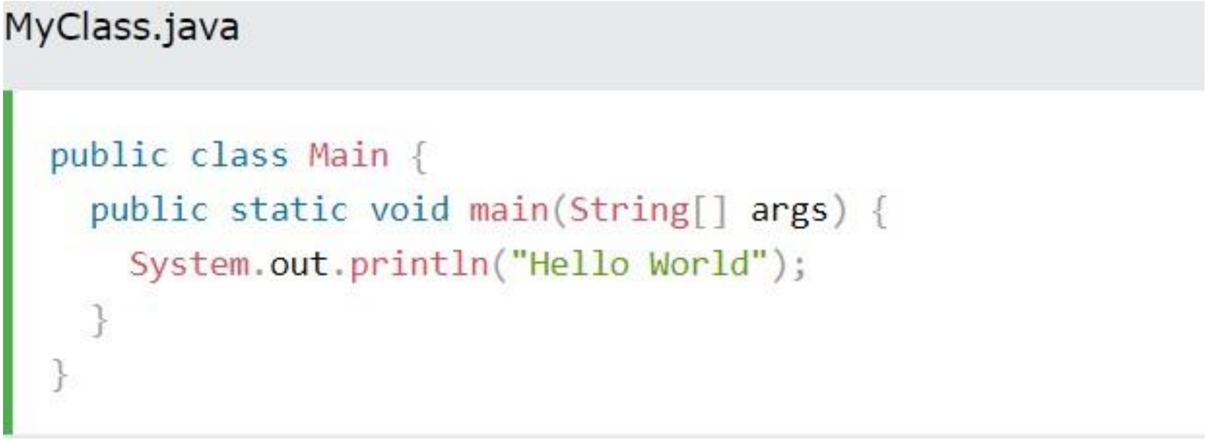
SYNTAX

Each java program must be enclosed inside a **class,** whose name should always start with an uppercase letter. Another requisite is the match between the project file name with the class name.

It is usually preceded by the **main()** method which gets executed having any code inside it. Any program needs to have a class and a main() method.

The **println ()** method is used inside the **main()** method to output information on the screen.

Curly braces {} mark the beginning and end of a block of the code. A semicolon (;) is used at the end of each sentence, to mark the end of that sentence.

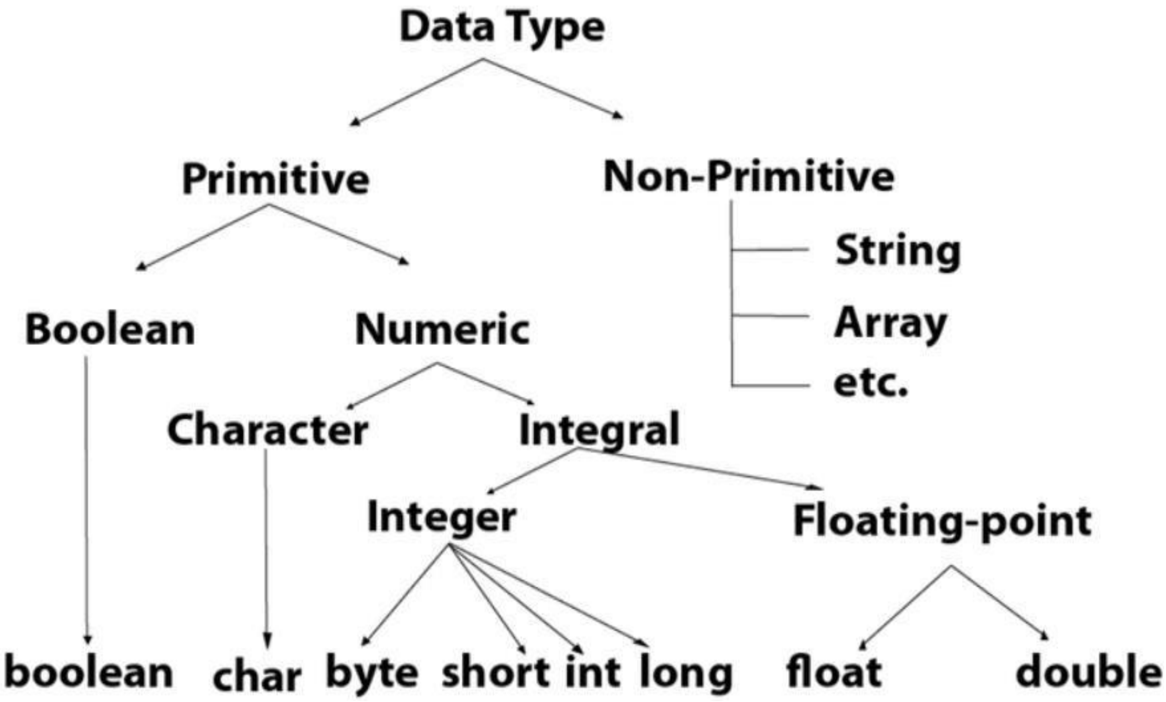


MyClass.java

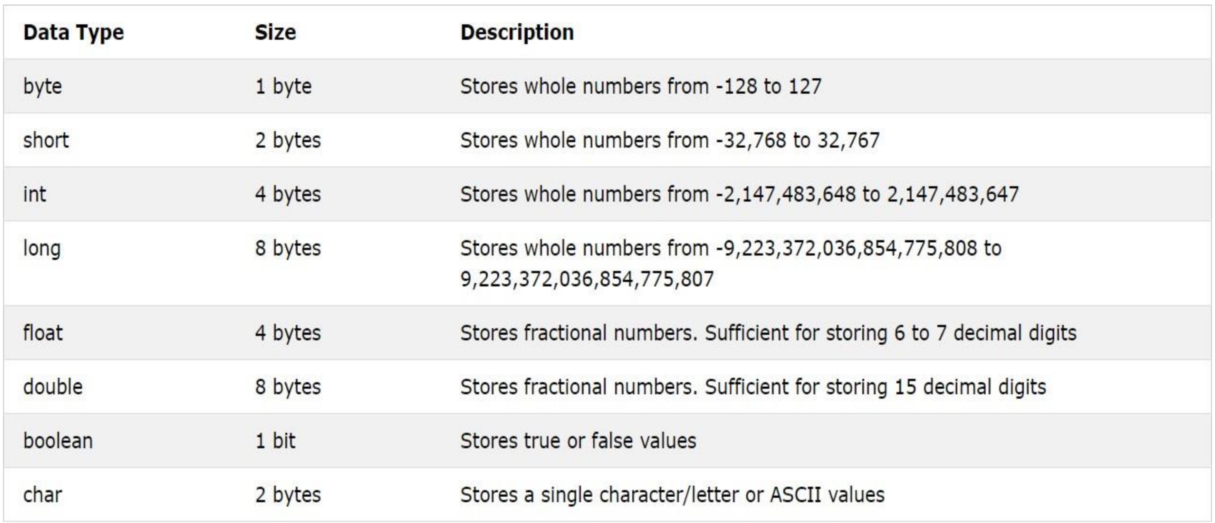
```
public class Main {
  public static void main(String[] args) {
    System.out.println("Hello World");
  }
}
```

JAVA DATA TYPES

Java has two categories of data types- Primitive and Non-Primitive data types.



Primitive Data types- They include data types like int, byte, long, float, short, double, char, Boolean.



Data Type	Size	Description
byte	1 byte	Stores whole numbers from -128 to 127
short	2 bytes	Stores whole numbers from -32,768 to 32,767
int	4 bytes	Stores whole numbers from -2,147,483,648 to 2,147,483,647
long	8 bytes	Stores whole numbers from -9,223,372,036,854,775,808 to 9,223,372,036,854,775,807
float	4 bytes	Stores fractional numbers. Sufficient for storing 6 to 7 decimal digits
double	8 bytes	Stores fractional numbers. Sufficient for storing 15 decimal digits
boolean	1 bit	Stores true or false values
char	2 bytes	Stores a single character/letter or ASCII values

Integer- It stores positive, negative or whole number values, not having decimals. The valid data-types are byte, shorting and long.

Float- It stores positive, negative or whole number values having decimal points, representing the fractional part. The valid data types are float and double.

Boolean- It is declared along with the Boolean keyword and evaluates to either true or false.

Character- Used for storage of a single character

Non-Primitive Data types- These refer to objects hence are also known as reference types. They differ from primitive data types in some aspects like- they are not predefined as in the case for primitive data types and are created during programming. The primitive data types need to have a value, while non primitive can be null. The non-primitive data types start with an uppercase alphabet while the primitive data types start with a lowercase letter. Some examples of non-primitive data types are- String, Arrays, Classes, Interfaces etc.

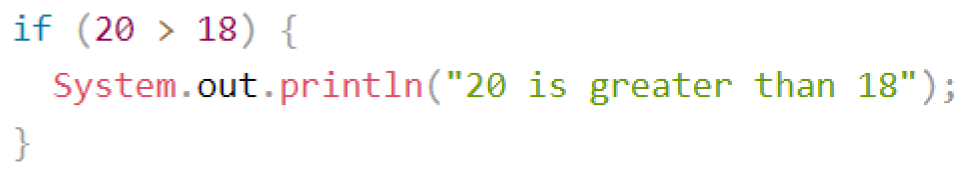
String- The String data type is generally used to store a sequence of characters. The characters must be enclosed within a pair of double quotes.

Arrays- They are utilized to store multiple values inside a single variable, instead of the need to declare multiple variables.

CONDITIONAL STATEMENTS

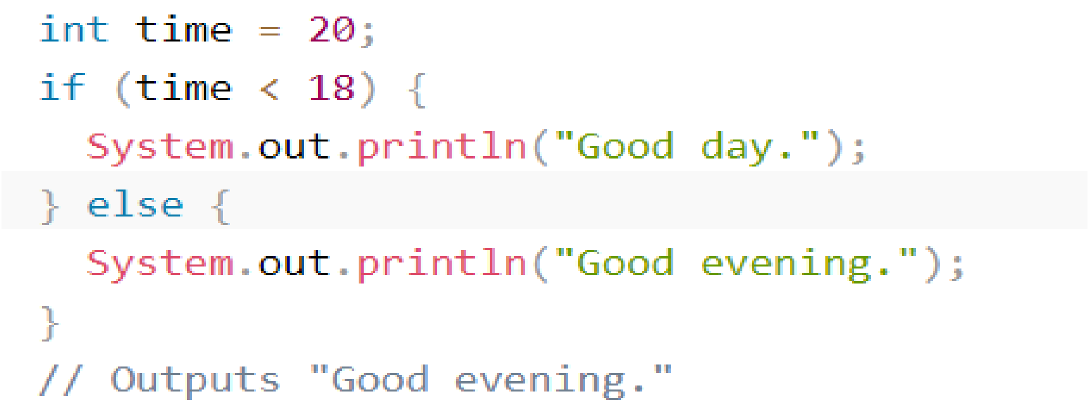
Java supports the general logic from mathematics like less than, greater than, equal to etc. which can be applied in programs. Some of the used conditional statements in java are;

If- Specifies a code block to be executed, if the condition evaluates to true.



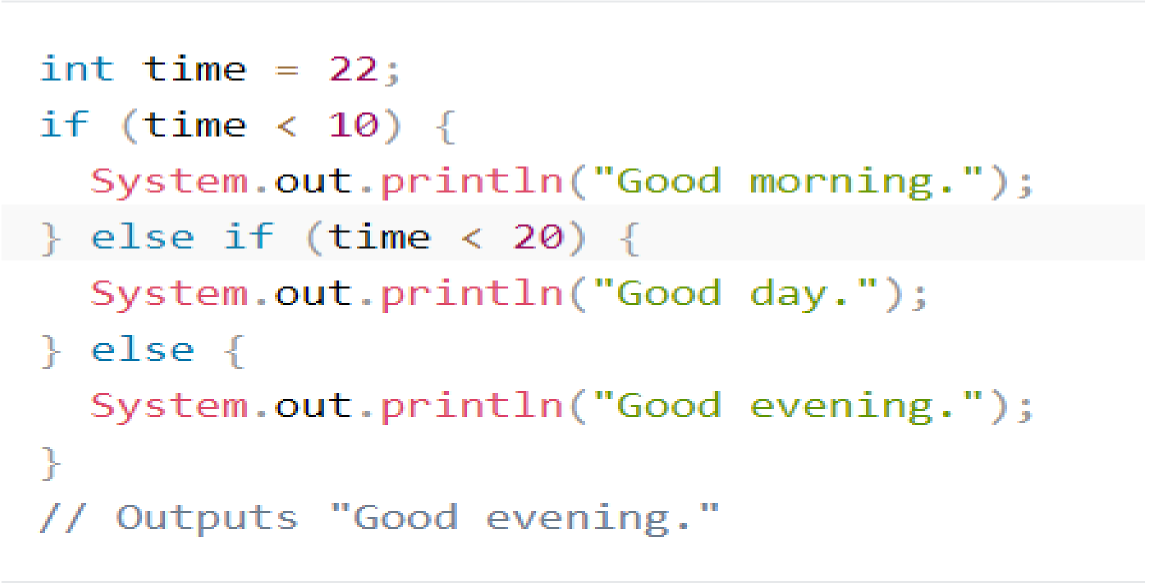
```
if (20 > 18) {
  System.out.println("20 is greater than 18");
}
```

Else- Specifies a code block to be executed, if the same condition evaluates to false.



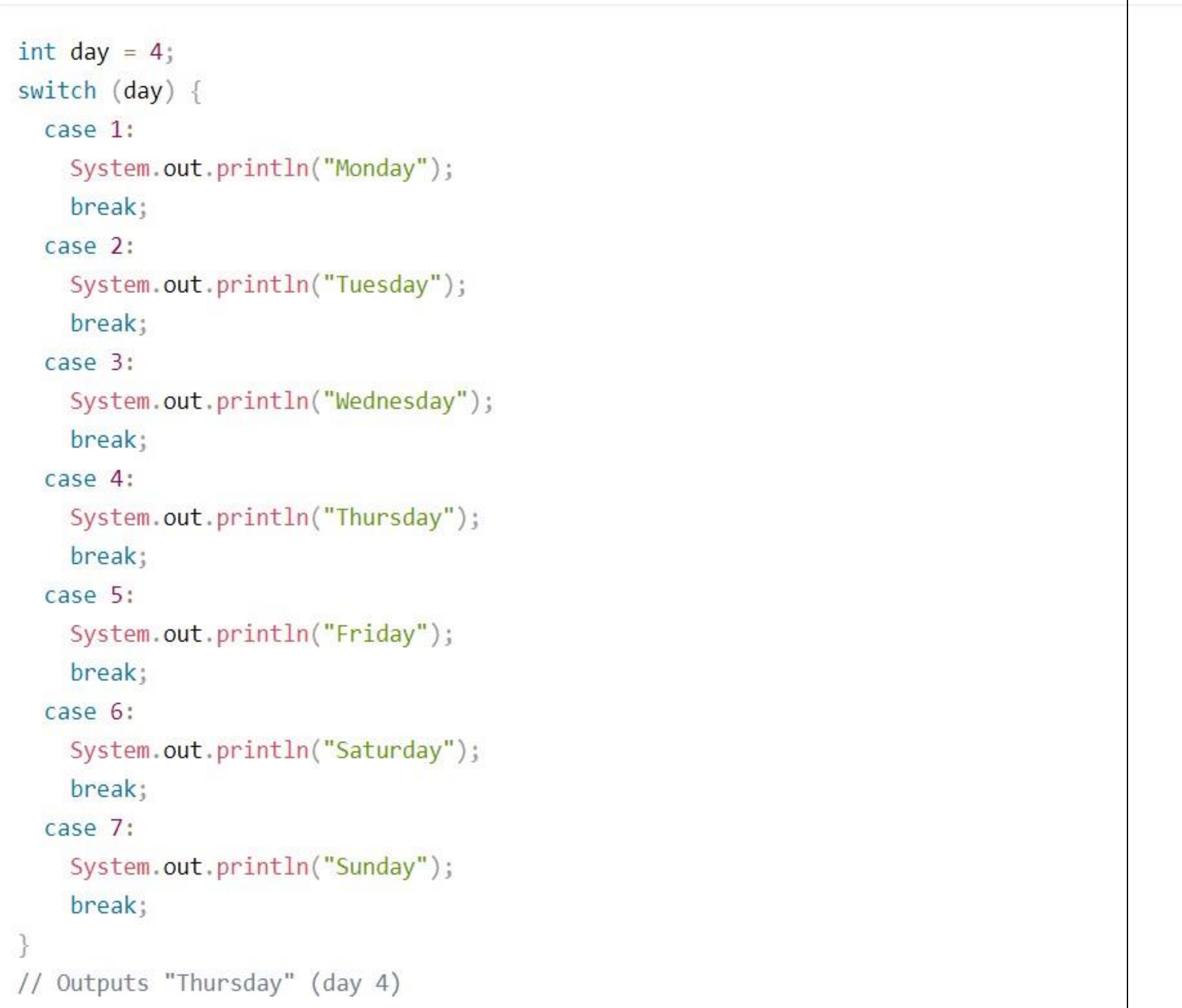
```
int time = 20;
if (time < 18) {
  System.out.println("Good day.");
} else {
  System.out.println("Good evening.");
}
// Outputs "Good evening."
```

Else if- Specifies a new code block to be executed if the first condition evaluates to false.



```
int time = 22;
if (time < 10) {
  System.out.println("Good morning.");
} else if (time < 20) {
  System.out.println("Good day.");
} else {
  System.out.println("Good evening.");
}
// Outputs "Good evening."
```

Switch- Specifies various alternative code blocks to be executed.



```java
int day = 4;
switch (day) {
  case 1:
    System.out.println("Monday");
    break;
  case 2:
    System.out.println("Tuesday");
    break;
  case 3:
    System.out.println("Wednesday");
    break;
  case 4:
    System.out.println("Thursday");
    break;
  case 5:
    System.out.println("Friday");
    break;
  case 6:
    System.out.println("Saturday");
    break;
  case 7:
    System.out.println("Sunday");
    break;
}
// Outputs "Thursday" (day 4)
```

CHAPTER – 03

UNIX AND SHELL SCRIPTING

What is UNIX?

UNIX is actually a family of operating systems, having the capability of multitasking, and multi user access at a same time. It's development kick started in 1970's at the AT&T's Bell Labs research center, by Ken Thompson, Brian Kernighan, Dennis Ritchie and others.

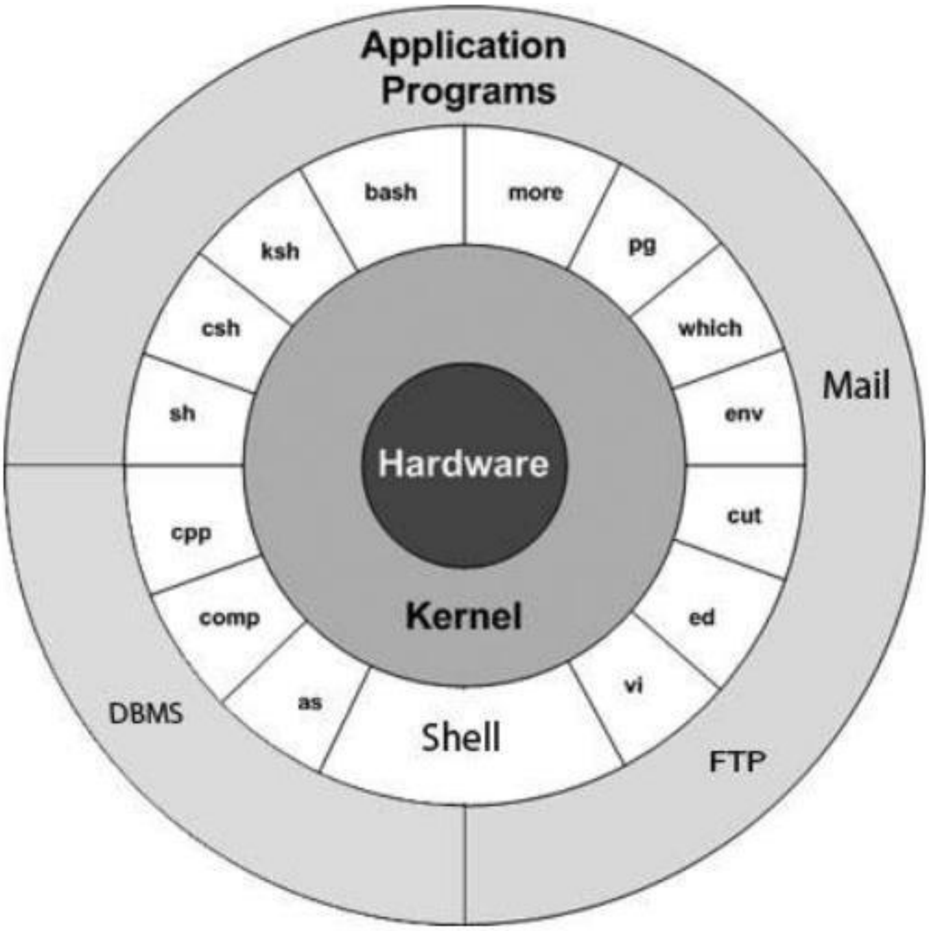
The operating system UNIX is a set of commands/programs that fuel as a link between the user and the computer system. An Operating System is a set of computer programs which allocate the system resources and further coordinate all the details of the available system internals. It is also referred to as Kernel.

The users use a **shell** to interact/communicate with the **kernel**. The **shell** is a command-line interpreter whose function is to translate the commands inputted by the user into a language which is understood by the **kernel** and thereby executing the given command. Various distributions/flavors of UNIX are available in the market like AIX, Solaris UNIX,HP UNIX etc. These are commercially licensed copies, while LINUX and its various distributions are open source and freely available. Since UNIX allows multiple programs to be executed at a single time, it is referred to as a multitasking operating system. Also since it allows multiple users to login at the same time, it is also a multiuser operating system.

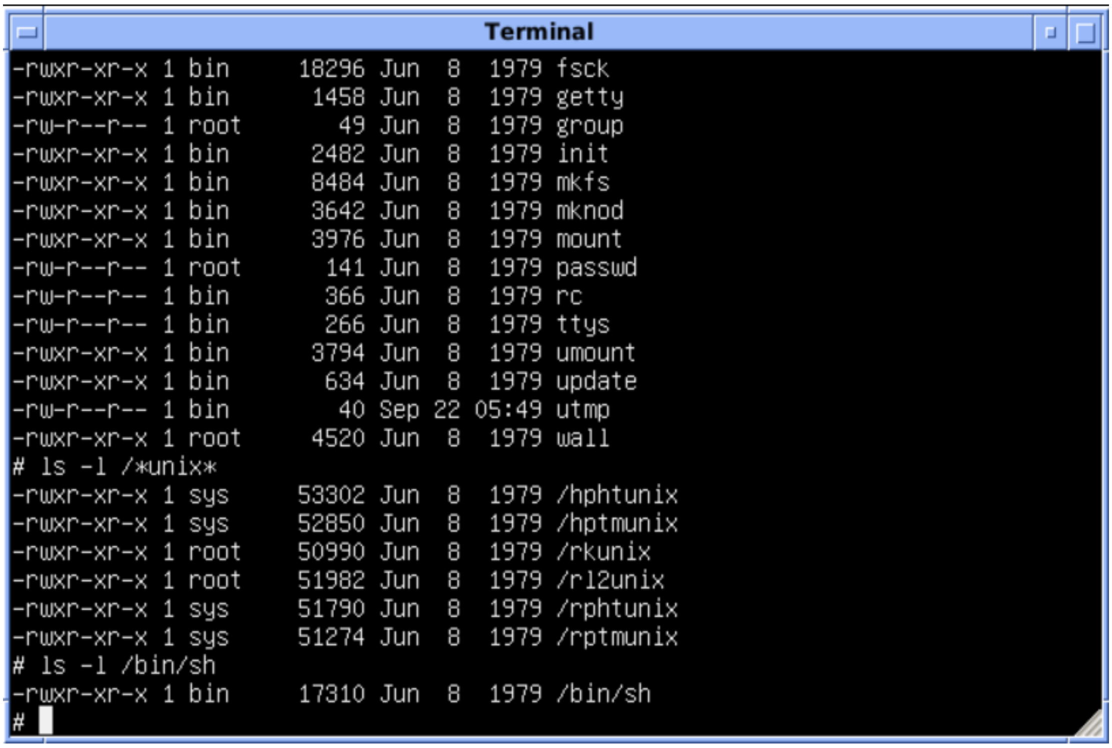
UNIX Architecture

There are four basic components of UNIX operating system-

1.) **Kernel-** It is referred to as the heart of an operating system. It is the main component which interacts with the hardware and takes care of tasks like memory and file management, task scheduling etc.

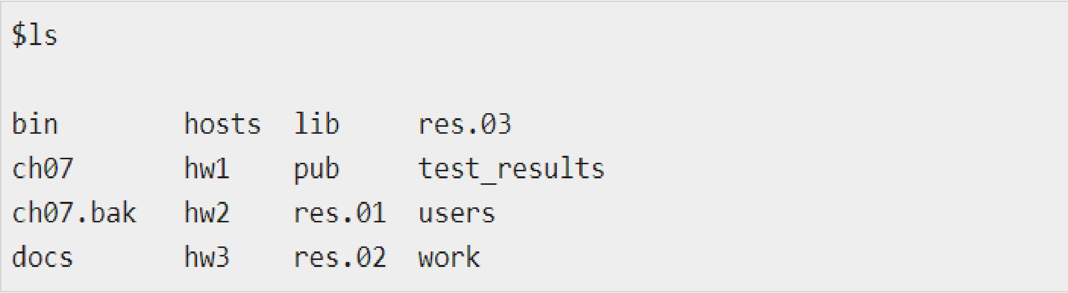


2.) **Shell-** A shell is a utility which processes the commands given as input in the terminal, processes them and then calls the required program to execute the task. The shell follows similar syntax for all the commands. Some types of shell available in UNIX are C shell, Korn shell, Bourne shell etc.



```
-rwxr-xr-x 1 bin     18296 Jun  8  1979 fsck
-rwxr-xr-x 1 bin      1458 Jun  8  1979 getty
-rw-r--r-- 1 root       49 Jun  8  1979 group
-rwxr-xr-x 1 bin      2482 Jun  8  1979 init
-rwxr-xr-x 1 bin      8484 Jun  8  1979 mkfs
-rwxr-xr-x 1 bin      3642 Jun  8  1979 mknod
-rwxr-xr-x 1 bin      3976 Jun  8  1979 mount
-rw-r--r-- 1 root      141 Jun  8  1979 passwd
-rw-r--r-- 1 bin       366 Jun  8  1979 rc
-rw-r--r-- 1 bin       266 Jun  8  1979 ttys
-rwxr-xr-x 1 bin      3794 Jun  8  1979 umount
-rwxr-xr-x 1 bin       634 Jun  8  1979 update
-rw-r--r-- 1 bin        40 Sep 22 05:49 utmp
-rwxr-xr-x 1 root     4520 Jun  8  1979 wall
# ls -l /*unix*
-rwxr-xr-x 1 sys     53302 Jun  8  1979 /hphtunix
-rwxr-xr-x 1 sys     52850 Jun  8  1979 /hptmunix
-rwxr-xr-x 1 root    50990 Jun  8  1979 /rkunix
-rwxr-xr-x 1 root    51982 Jun  8  1979 /rl2unix
-rwxr-xr-x 1 sys     51790 Jun  8  1979 /rphtunix
-rwxr-xr-x 1 sys     51274 Jun  8  1979 /rptmunix
# ls -l /bin/sh
-rwxr-xr-x 1 bin     17310 Jun  8  1979 /bin/sh
# 
```

3.) **Commands/Utilities-** Unix houses various commands to perform everyday tasks like copying files, making directories or files, adding lines to a file, counting the number of lines and words in a given file etc. Some of the commands are - ls, cp, grep, mkdir, cat etc.



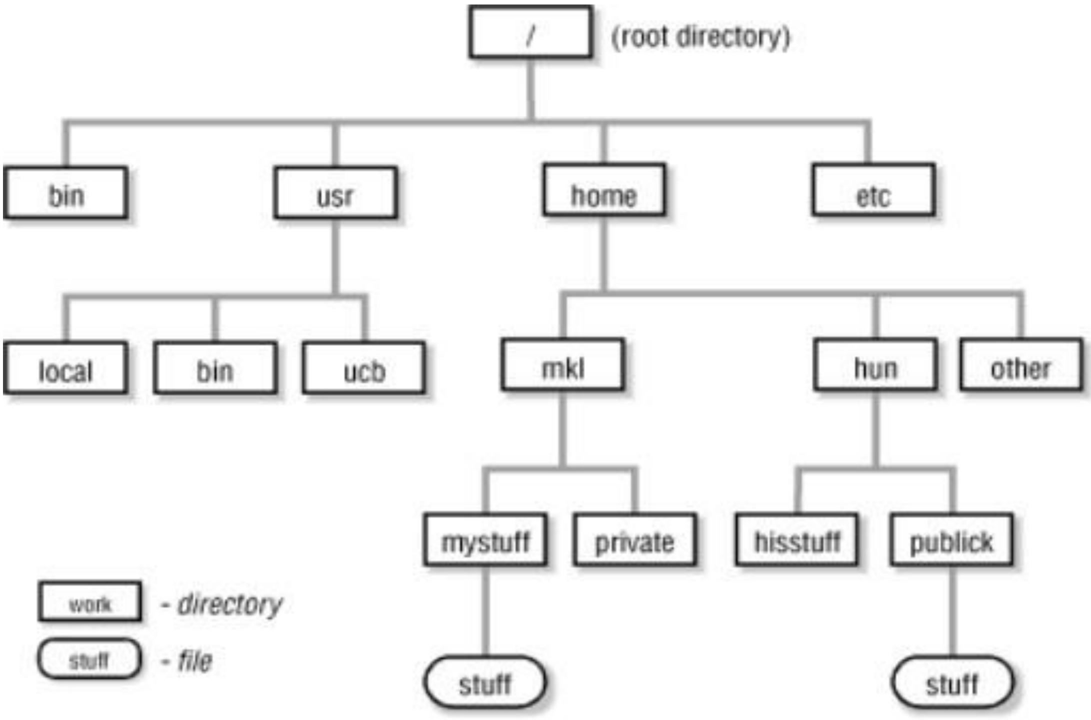
```
$ls

bin        hosts  lib     res.03
ch07       hw1    pub     test_results
ch07.bak   hw2    res.01  users
docs       hw3    res.02  work
```

4.) **Files/Directories-** UNIX follows a tree-like structure for the organization of directories. The data is organized into files and files are placed inside the directories. There are basically three types of files present in the UNIX file system-

a.) Directories- They store special as well as regular files in it. It is equivalent to folders in the Windows operating system.

b.) Ordinary files- It contains data as text files or programs.

c.) Special files- They provide access to hardware components like CD drive, network adapters etc.



IMPACT OF UNIX

Unix has a tremendous impact on operating systems, as its portable, is available at a nominal price for educational and research purposes, runs on even the hardware with lowest configurations and can be adapted easily to different systems or machines.

UNIX based LINUX is highly utilized for high end servers dedicated for storage and data processing. It also popularized its hierarchical file system with nested sub-directories. Since UNIX is majorly written in C language it makes it easier to work with and work on any kind of system. The architecture and design of UNIX is so appealing that the tech giant Apple keeps it as the core of their Mac OS operating system. Many businesses thrive upon UNIX for their regular business operations. Working knowledge of unix is recommended for establishing familiarity when working on big data, and various tools associated with it, since the associated tools and frameworks like hadoop use a similar query as found in UNIX like ls, cat, rm, rmdir, mkdir, etc.

PYTHON

- It was created by Guido van rossum
- It is used by multiple platforms like –
    - YouTube
    - Quora
    - Drop box
    - Reddit
    - Bit torrent

The various applications of python are-

- Image processing
- Graphic design
- 3D modeling
- Scientific data processing

Idle – integrated development environment tool allows us to write and run our code easily with a simple interface. Code is written in the python prompt after >>>

Characteristics of idle –

- Written in python
- Uses tkinter graphics library
- Has an interactive python shell
- A full featured text editor
- A debugger

Python is an interpreted language, which basically means that it uses an interpreter in place of a compiler. An interpreter takes one instruction at a time and executes it in real time.

Salient features of python –

· High level programming language – python is a high level programming language and exhibits the features of a high level language like code readability and easy usage.

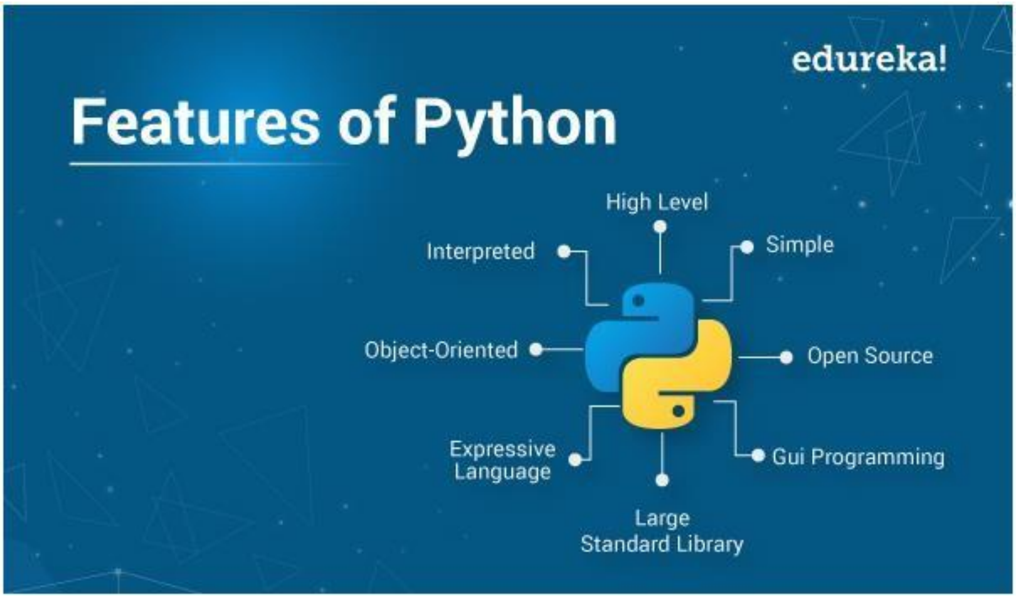
· Open source – python is free to use and can be used for personal and professional work free of cost

· Supports multiple programming paradigm – it supports object oriented programming, imperative, procedural and functional programming

· Extensible – python can easily be used combined with different languages and frameworks with simple extensions and commands

- Gui programming – creating buttons, text boxes, widgets is easy and doable in python with its different technologies and tools

- Embeddable – python language is embeddable and can be used for embedded programming of certain devices, advanced technologies using its imperative programming paradigm



Python 2.xx versus python 3.xx

- Python 2.0 came out in 2000 while the python 3.0 came out in 2008

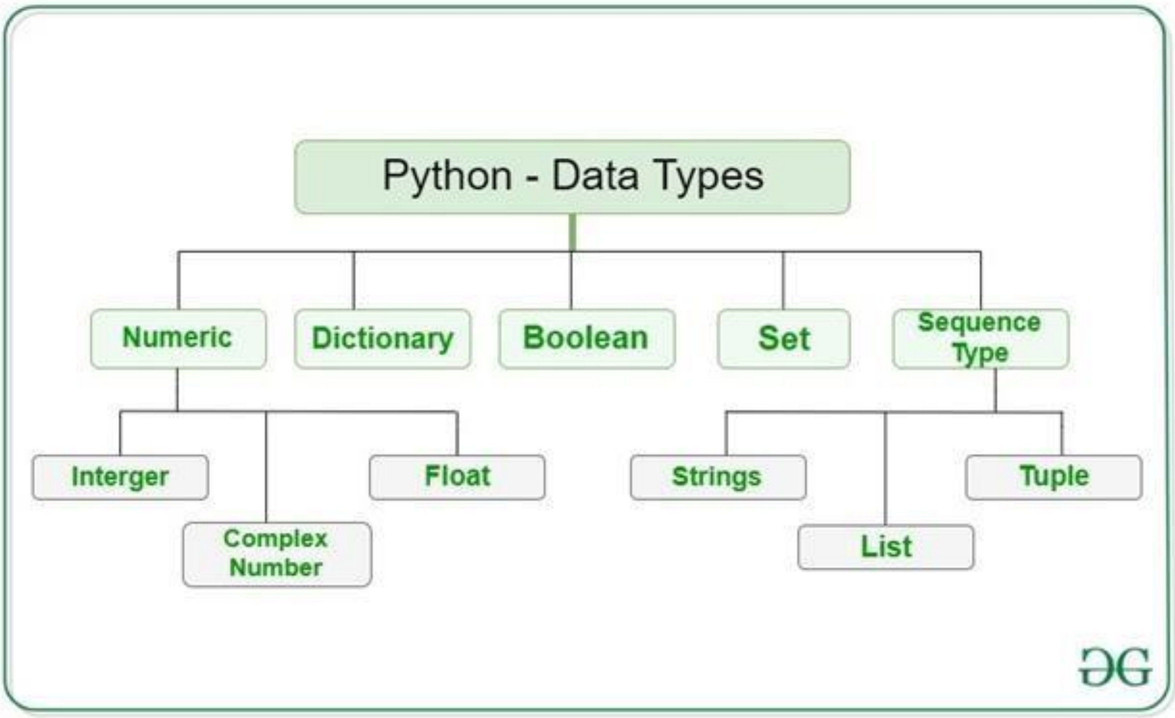
- Print is now a function and not a keyword as in the earlier version. Parenthesis are now made compulsory to use while writing the print command

- Input function is used instead of the previous raw input while doing the same job

- Results of the arithmetic division operations are now calculated as decimals only

- Stores strings as Unicode by default

- Integer objects are long by default and don’t require L as 1000L



DATA TYPES IN PYTHON

It is the classification of data items. The most common types of data types are numeric, non-numeric and Boolean. Knowing the data type helps us to understand what kind of operations and applications can be created with the usage of the available data.

The four broad classifications of data are –

- Numeric

- Boolean

- Sequence

- Dictionary

1) NUMERIC

Data values, which are numeric in nature comes under numeric data types. They are of three types –

Integer – positive or negative whole numbers. Example – 2, -5

Float – real numbers with floating point representation. Example – 1.3E, -2.8

Complex – any number with a real and an imaginary component. Example – 2+3i

2) BOOLEAN

Any representation of data, which has two values denoted by true or false is called a Boolean data type.

3) SEQUENCE

Some of the built in sequence data types are –

· String – a collection of one or more characters in single, double or triple quotes. Example – 'hello', ''abc'', ''''pqr'''

· List – an ordered collection of one or more data items, not necessarily of same type in square brackets represent a list. Example – [1, 'ram', 2.4, True]

· Tuple –multiple data items in a combined formed, when put in a parenthesis make a tuple. Contents of a tuple cannot be modified. Example – (1, 'ram', 2.4, True)

4) DICTIONARY

An unordered collection of data in key: value pair form. Collection of such pairs is enclosed in curly braces. Example – {1: "Aditya", 2: "Akash", 3: "Ajay"}

VARIABLE - a name given to an object and not to the memory location. Its value can vary throughout the execution of the program. It is not bound to a single data type but the data which is entered at the position determines the type of the variable.

That's the reason why python is known as a dynamically typed language.

ARITHMETIC OPERATORS

(PEDMAS)

- Addition - Adds the two operands on both the sides of the operator. Denoted by (+) symbol

- Subtraction – subtracts the right operand from the left operand. Denoted by the (-) symbol

- Multiplication – denoted by (*)

- Division – denoted by (/)

- Modulus – (%)

- Exponent – (**)

- Floor division – (//)

STRING DATA TYPES

- Single, double or triple quotes

- If the string has apostrophe, use double quotes, otherwise use interchangeably

- Or if it has double quotes in it, use single quotes to enclose the string

- For multi-line strings, use triple quotes

Accessing characters in a string-

A string is an ordered set of characters. Therefore, each character has an index. The first character has the index 0 and so on. To access the character, we write the name of the particular list with the character index in the [].

STRING OPERATIONS

- Concatenation - multiple strings are appended with each

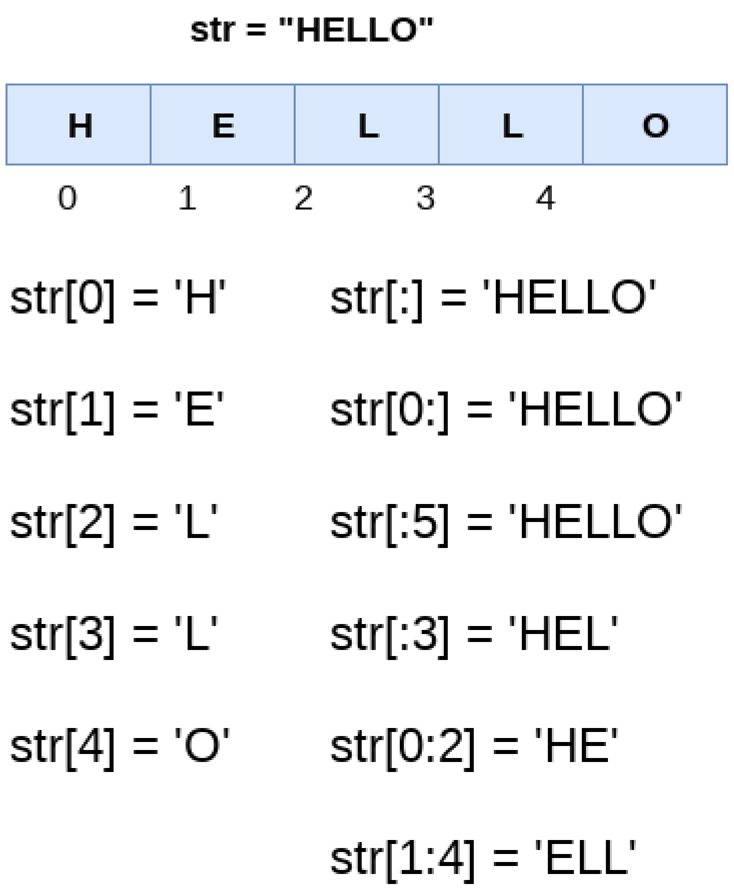
- Repetition – concatenates multiple copies of same string

- Slice – it gives the character at any given index

- Range slice – fetches characters in the range specified by two separate indexes

- In membership – returns true if the substring or character is present in a string

- Not in membership – returns true if character is not in the string



```
str[0] = 'H'     str[:] = 'HELLO'
str[1] = 'E'     str[0:] = 'HELLO'
str[2] = 'L'     str[:5] = 'HELLO'
str[3] = 'L'     str[:3] = 'HEL'
str[4] = 'O'     str[0:2] = 'HE'
                 str[1:4] = 'ELL'
```

PYTHON CONTROL STRUCTURES

Python scripts are beneficial as –

- Easily insert/delete/update statements

- Code or functions from the script can be imported

- Automate and schedule tasks

Conditional and repetitive execution –

Conditional statements are used for executing a block of code if some particular condition is true.

Repetitive statements are used for executing the blocks of code for some multiple times.

USING CONDITIONALS

- IF

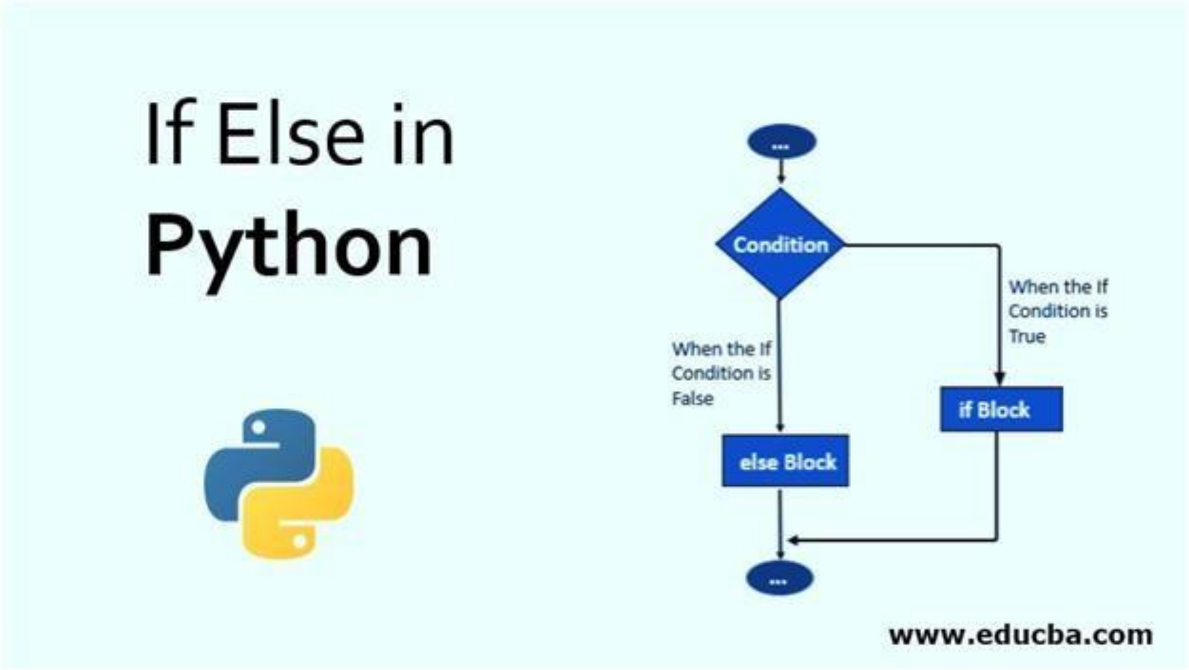
If expression value is calculated, if true, statement 1 is executed and then statement 2 is executed. If false, directly statement 2 is executed.

- ELSE

When alternate situations are required with if, we use else.

- ELIF

For multiple conditions, to reduce the else if indentation and complexity, we use elif keyword.



LOOPS

If the program flow is redirected towards any of the earlier statements, it is known as a loop. We need to specify some conditions for the loop to stop, to prevent it from going into an infinite loop.

- While loop

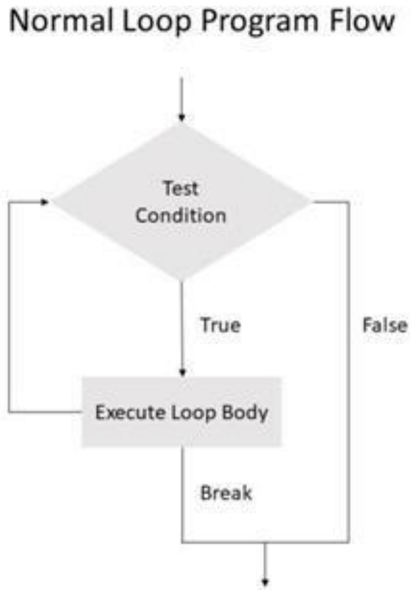
```
While expression:
  Statement 1
  Statement 2
Statement 3
```

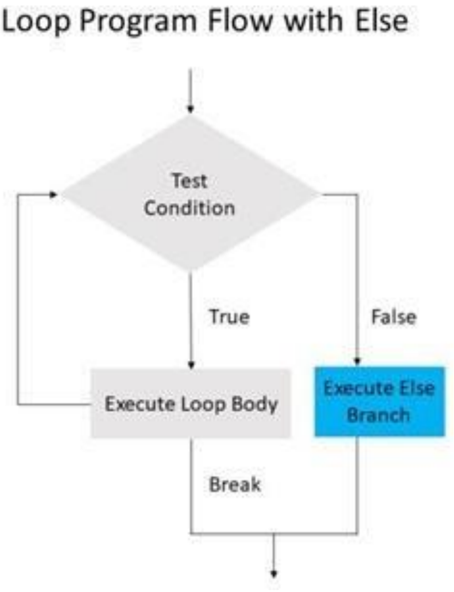
When the expression is true, the body of the loop is executed, when it becomes false, control comes out of the loop.

- For loop

```
For variable in sequence:
  Statement 1
  Statement 2
Statement 3
```

Only sequences are iterable in for loop. For numbers, use range.





USING FUNCTIONS

Independent and reusable blocks of instructions are called functions. Dividing a complex problem into functions for each program is called modular programming. Makes the code easy to develop, follow and maintain.

Modular programming also takes a top down approach towards programming. When we call a function, it performs the task and returns the control to the calling routine.

- Organizes the code
- Makes it more readable

```
Def function name ():
  Statements
  Return statement
```

Functions with arguments

Arguments are the parameters passed in the parenthesis of the function which it can use in performing a task. Two ways of specifying the arguments –

- Simply include the argument in parenthesis while calling

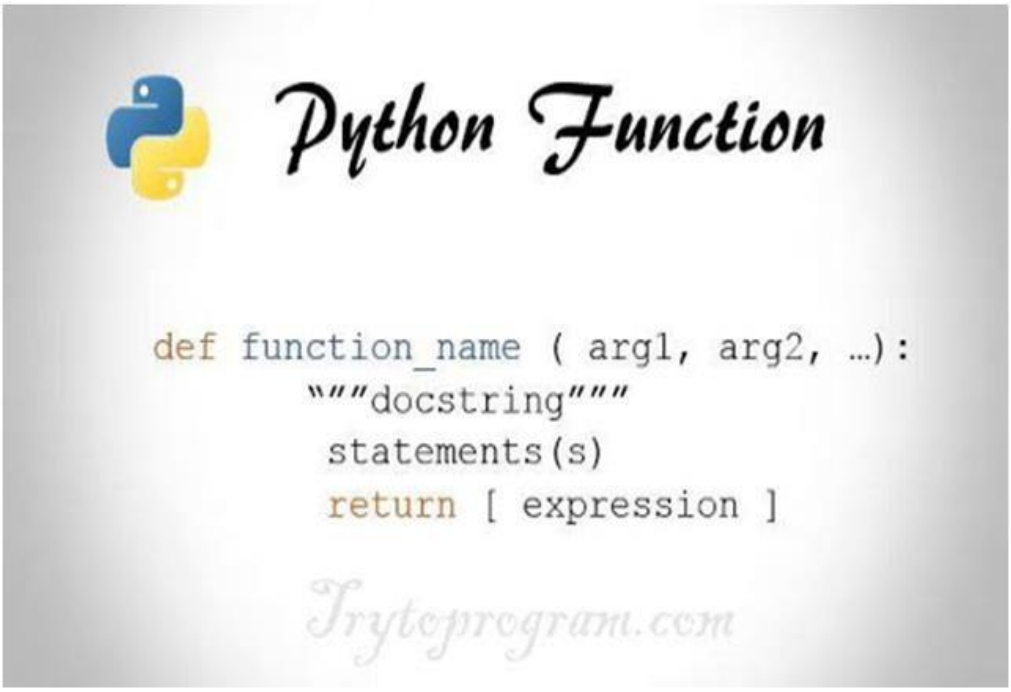
```
Def myfunc(name):

        Print(“Hello{}”.format(name))

        Return
```

To call the function,

```
Myfunc(“Aditya”):
```

- The other way is to use variables

```
Def sayhello(name):

  Print(“Hello{}”.format(name))

  Return
```



Python Function

```
def function_name ( arg1, arg2, …):
      “””docstring”””
       statements(s)
       return [ expression ]
```

Tryteprogram.com

CHAPTER - 04

DATA WAREHOUSE

The collection and management of data from various multiple sources is termed as data warehousing. This process is done keeping in mind the purpose, which is deriving meaningful insights from the data and enabling a better data driven decision making process. Heterogeneous sources might be used to draw data and the business data or the transactional data is primarily analyzed using advanced tools and technologies so that some useful information can be extracted from them.

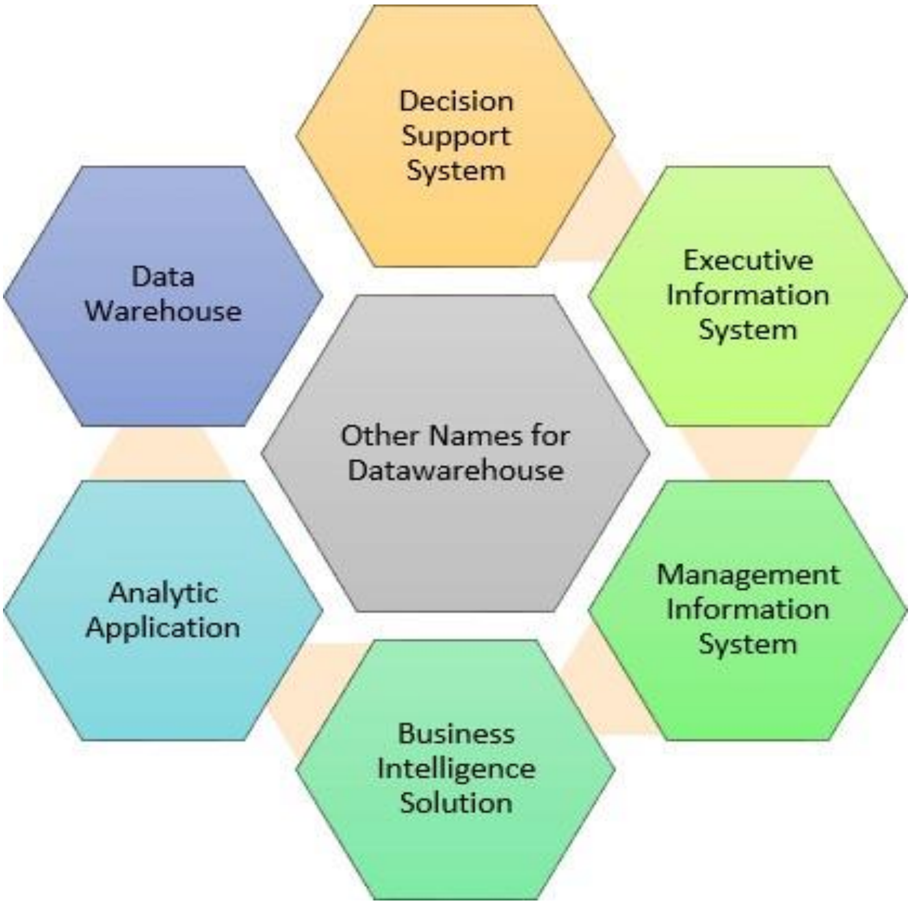
Data analysis and reporting is the main aim of the systems called business informatics systems and these systems, at their core, have the concept of data warehousing. Because, data before the canalization of data and reporting of the insights to the concerned business project, we first need to have the data required and also that data should also be in a very refined form. This refined data can then be utilized very simply by the complex and advanced systems while applying simple as well as complex algorithms on them.

A proper combination or blend of tools and technologies with all the important components help in the proper strategic use of this data. Advanced systems have been developed by companies which work in data driven industries for this very purpose and investment of billions of dollars have been made to make these systems work and help the businesses and these industries grow and generate higher revenues than ever.

After the collection and storage of this data, another main component of these systems is to transform the data. Transformation in its basic sense is conversion or change into a desired format or system. The data, to be of use while stored in these sophisticated systems also needs to be refined and in its most workable form, which means that it should be better ready for the algorithms and operations to be performed on it.

The organization must maintain this decision support data or decision support database separately from the operational database of the organization. The data warehouse, instead of being a product, is rather an environment, which is created in order for the analytics or data based intelligence operations to take place in and through it. The rise of management information systems or business information systems is the indicator of the fact that data and the tools and technologies used in relation to this data are of utmost importance to not only software or IT industry but to all other industries including the global supply chains, advertising and finance especially.

The popular and useful 3NF designed DBs are made of different tables and related to these tables might have many corresponding conditions for data to be accessed and used and hence these might take up a huge amount of time in the decision support systems of these organizations. Therefore, in order to move forward in the data usage and organization industry, we need to move forward to these types of storage and analysis of data. Data warehousing is known in many forms depending on how certain industries and academic groups see them. Following is an illustration.


Decision Support System
Data Warehouse
Executive Information System
Other Names for Datawarehouse
Analytic Application
Management Information System
Business Intelligence Solution

Data warehouses may be primarily classified into three main types -

- Enterprise data warehouse - it basically acts as a decision support system, supported by the data. The reporting and organization of data is combined through this method and industry specific data is labeled as required by the business. So it basically provides better freedom while dealing with data.
- Operational data store - it is usually called as ODS. when the need of the organization is not met by both OLTP and data warehouses, then this approach might be needed. The data is refreshed in real time in this system and thus it is usually preferred by organizations that require storing the data like employee records, which are recorded daily.
- Data mart - these are nothing but a special kind of smaller data warehouses or we could say in a sense, a subset of the complete data warehousing

system. These are usually designed for meeting the needs of the specific division of the organization such as finance, hr, sales or operations.

DATA WAREHOUSE APPLICATIONS IN DIFFERENT SECTOR

There are multiple sectors in which a data warehouse is used, some of them are mentioned below –

- Telecommunications - sales, distribution and advertisements are some of the operations that benefit from data driven decisions in the telecommunication sector

- Hospitality - promotions campaign, advertising and design for better targeting of existing and potential consumers utilizes the data warehouse systems in hospitality.

- Public sector - the present government systems are usually outdated file systems and database systems. Thus they have huge potential of improvement in their management systems and analytics systems for various government systems might be a very good idea.

- Healthcare - sharing and generation of medical reports, different personnel's files and profiles and even imaging and predicting systems are used in the healthcare systems

- Retail chains - transactional data of customers, which might vary from buying patterns, trends, new consumer engagement and retaining the customers with targeted marketing and personal discounts are some of the many fields which utilize the business intelligence systems.

- Investment and finance - it is one of the most data producing, analyzing and consuming sectors which works primarily on data. Decisions worth millions of dollars are made using the insights from share markets, studying profiles of companies and the whole investment system is very cleverly created in order to generate revenues for these organizations with the intelligent use of data.

- Airline - the whole operations of the airline systems have been working digitally from the longest time and this is one industry that smartly used data and positioning systems to create better consumer engagement, even in the Covid-era.

CHAPTER - 05

BIG DATA

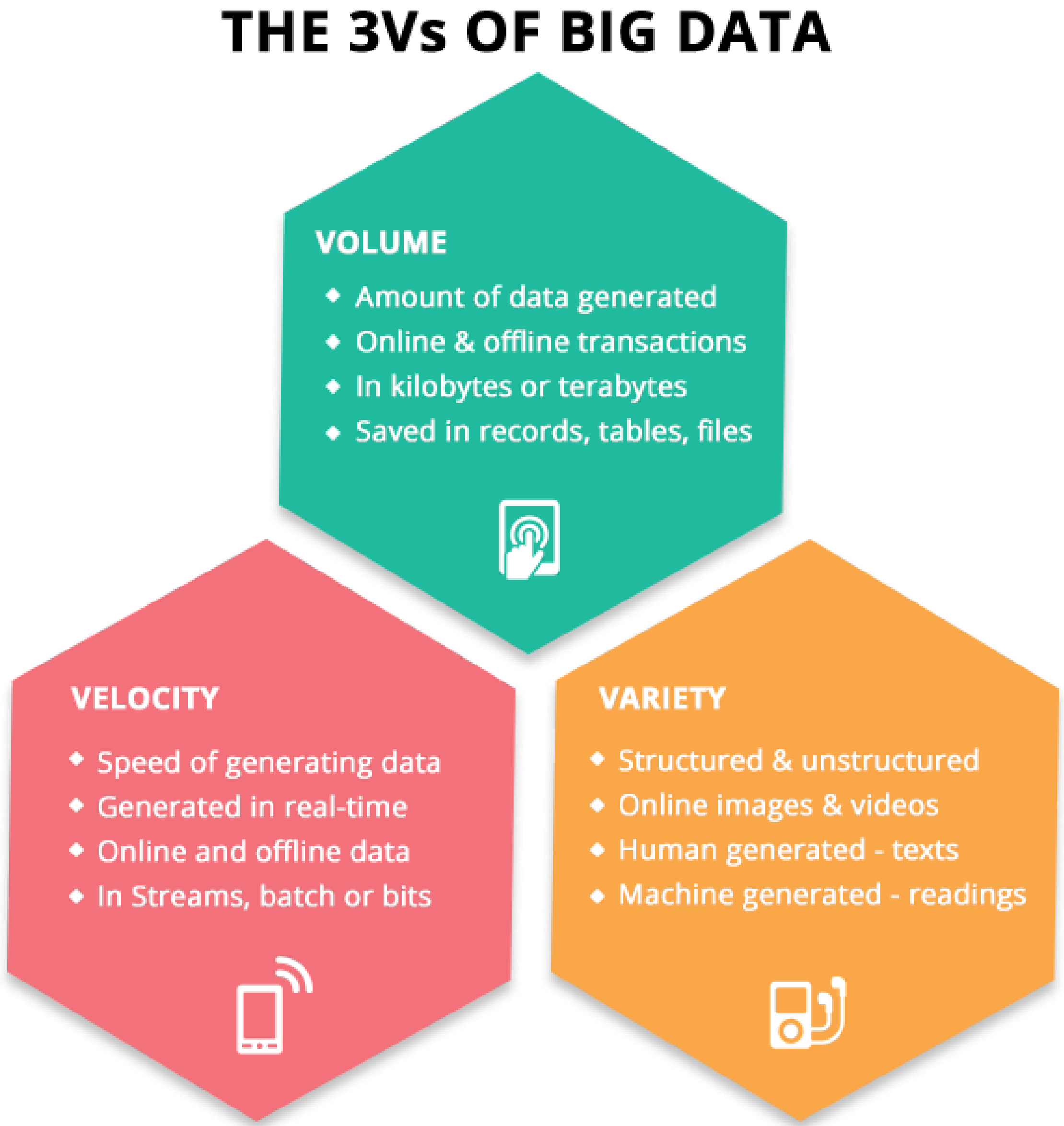
Big data is a huge collection of data. It is similar to data but very large in volume so that it cannot be managed by any traditional data management systems. The big data is primarily characterized by three parameters, the 3 Vs –

- Volume – the amount of data which is being generated is called its volume. The first main characteristic of the big data is its high volume. The data which is being generated on the lines of terabytes and petabytes regularly comes under big data.

- Velocity – the data is being generated everyday or at regular intervals. The speed at which the data is being generated is referred to as its velocity. The data explosion that is happening in the current scenario is only manageable by the use of big data and traditional methods of data storage and management are not capable of holding this data.

- Variety – the data which is being generated right now is of various different qualities and quantities as well as of various different types. The data now is not limited to records that has usual text information but even audio, video and other kinds of data is being managed and this kind of variety is acceptable and required for the data to be considered as big data.

The data which fulfills the above three conditions or is actually characterized by the above three parameters is referred to as big data.



THE 3Vs OF BIG DATA

VOLUME

- Amount of data generated
- Online & offline transactions
- In kilobytes or terabytes
- Saved in records, tables, files

VELOCITY

- Speed of generating data
- Generated in real-time
- Online and offline data
- In Streams, batch or bits

VARIETY

- Structured & unstructured
- Online images & videos
- Human generated - texts
- Machine generated - readings

www.whishworks.com

Types of big data

- Structured – when the format of the data is fixed and the data can be processed and stored in this same format, then it is referred to it as structured data. When the format of the data is known in advance, the advancement in the technology in IT allows us to handle the same with ease but the problem in dealing with this structured data arises when the size of the data that is being used is very high such that it is in multiple zettabytes.

- Unstructured – when the structure or form of the data is not known, it is classified as unstructured data. Such kind of data might contain of the data in a combined format of text, images and videos and hence its becomes very tough to store, process and derive some useful insights from such data, which in addition to this problem also has huge quantities.

- Semi- structured – the kind of data which can contain both of the above types of data in a combined as well as individual form. Semi- structured data can be in both the structured but not defined like in a database management system. An example of the semi-structured data is XML file.

Real life examples of big data

- Finance and investment – this field thrives on data and thus is bound to generate magnitudes of data. For an instance, the stock exchange at New York City only generates more than one terabytes of data per day.

- Social media – with millions and billions of users on the social media platforms right now, each of their logins, activity, likes and comments, posts, shared content are recorded and that by no surprise generates more than 500 terabytes of data on a daily basis.

- Jet engines – in only as short as 45 minutes of flight, 15+ terabytes of data can be generated by the jet engine.

HADOOP

Hadoop is an open source system, by apache, which is created to handle the storage and processing of big data with the help of simple programming model. This is done on a cluster of computers. It was written in java. Hadoop basically provides an environment for distributed storage and computation and thus the process effort is divided into multiple pieces for each of the system to store and process and hence the effort required by each system goes down multiple times in terms of both complexity and time.

Hadoop basic architecture –

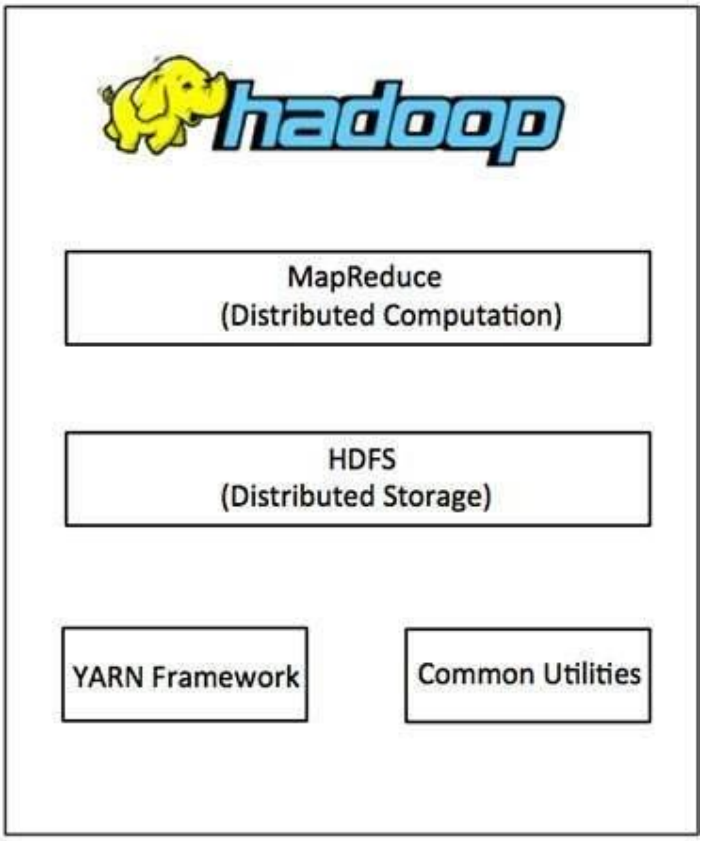
Hadoop is made up of two basic layers when it comes to the architecture –

- Processing layer called Map reduce
  It was developed by Google to support multiple programming paradigms which uses the parallel programming concepts. It uses a distributed system in order to do the processing part and makes it efficient to process large amounts of data.

- Storage layer as Hadoop distributed file system
  Based on Google file system, hadoop file system is used to avail a distributed system of storage of the big data. One of the most important benefits of this file system is its fault-tolerance and its usage of low cost hardware, which basically use commodity hardware.

  Additional modules in the hadoop framework are also present –

- Hadoop common
- Hadoop Yarn



REFERENCES –

- Molinaro, Anthony (2005), SQL Cookbook
- Beighley, Lynn (2007), Headfirst SQL
- Johnston, Benjamin (2019), SQL for data analytics
- Bloch, Joshua (2001), Effective java
- Bates, Bert & sierra, Kathy (2003), Head first java
- Ascher, David (1999), Learning Python
- Barry, Paul (2010), Head first python
- Simon, Phil (2013), Too big to ignore : Data, Analytics
- Kitchin, Rob (2014), The Data Revolution
- Davenport, T.H (2016), Big data at work

BIG_DATA_TRAINING-_Project_Report.docx

by

Submission date: 20-May-2021 06:35PM (UTC+0530)

Submission ID: 1590298870

File name: BIG_DATA_TRAINING-_Project_Report.docx (4.1M)

Word count: 7780

Character count: 38890

(PROJECT REPORT - BIG DATA TRAINING)

Project report submitted in fulfillment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering/Information Technology

By

Aditya Agarwal (171479)

Under the supervision of
(Dr. Monika Bharti)

Department of Computer Science & Engineering and Information Technology

**Jaypee University of Information Technology Waknaghat,
Solan-173234, Himachal Pradesh**

TABLE OF CONTENT

Certificate

Candidate's Declaration

We hereby declare that the work presented in this report entitled **"Big data Training"** in fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, **"Jaypee University of Information Technology, Waknaghat"** is an authentic record of our own work carried out over a period from **January,2021** to **May, 2021** under the supervision of **Dr. Monika Bharti** (Associate Professor – CSE & IT).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)

Aditya Agarwal (171479)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Dr. Monika Bharti
Associate Professor
Department of Computer Science and Engineering & Information Technology

Dated: May 20, 2021

ABSTRACT

Data is considered to be the elixir of information technology field. For every process or decision, that needs to be taken, we require data. Also, every process or transaction that takes place also generates data. Earlier, traditional methods of data storage and processing were used to deal with data. But right now, the extent of data production and the rate at which data needs to be dealt with requires advanced tools and technologies. Big data technologies are a solution, which provide faster and less costly methods to store and deal with data that is being generated and that is required. It divides the data storage with the help of distributed file systems and in order to process this data, we use the different commodity hardware to reduce and divide the computation effort for each machine. The hadoop architecture is primarily made of two components called as hadoop file distributed system and map reduce. Other tools like sqoop, hive, pig and Kafka are used to import, export, process, visualize data and generate reports which assist with the decision making processes in organizations.

LIST OF IMAGES

COMPANY PROFILE

Cognizant is a leading American multinational firm, which provides its services in business consulting, information technology, system integration, artificial intelligence, digital engineering, analytics, business intelligence, data warehousing etc. It initially began as Dun & Bradstreet Software in January 1994, established as Dun & Bradstreet's in-house unit for providing IT-infrastructure related services for Dun & Bradstreet business, but later expanded its client base from 1996.

Cognizant’s digital business, operations and systems and technology are the three areas which make up their business profile. To provide technological proficiency to its clients Cognizant is organized into various verticals and horizontals. The verticals focus on specific industries like- Banking and Financial Services, Insurance, Healthcare, Manufacturing and Retail services etc.

The horizontals on the other hand focus on specific technologies and services like - Analytics, mobile computing, BPO and testing solutions. It follows a business model similar to other IT giants based on, global delivery model, which is based upon offshore software R&D and offshore outsourcing.

The first time Cognizant came in the Fortune 500 list was in 2011. In 2015, Fortune named Cognizant as the world’s 4th most admired IT Services Company. It currently ranks 194 in Fortune 500 companies, 533 in Forbes Global 2000, 483 in Forbes Best Employers for Diversity in 2019.

Cognizant is among the high scientific discipline corporation that has been delivering high quality IT-infrastructure services and Business Intelligence services, extending to a list of happy clients worldwide. With various teams of highly proficient and hardworking associates working 24*7 to deliver high standard results and speedy turnarounds, it has been helping its clients in increasing their business potency.

CHAPTER - 01

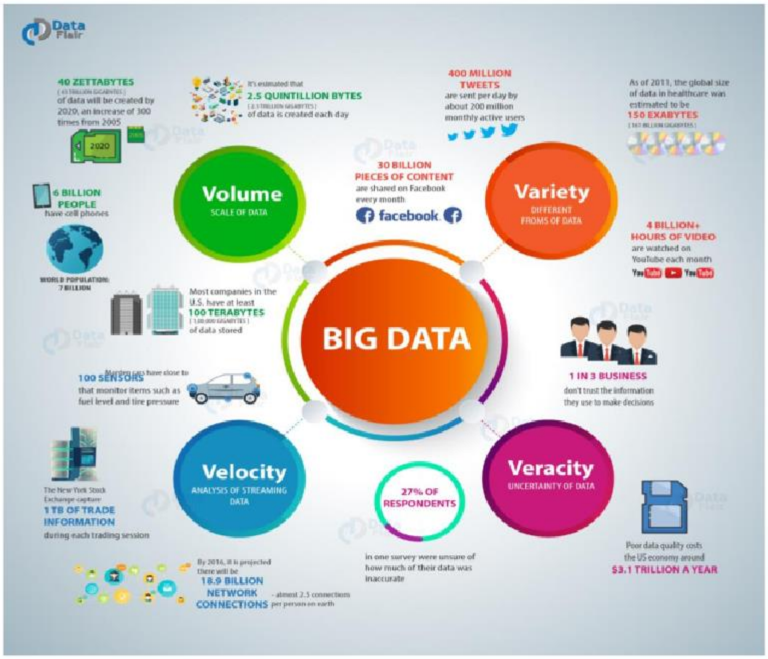
INTRODUCTION TO ASSIGNED WORK

DATA:

Data is defined as the quantities, characters or symbols upon which operations are performed by the computer, which might be stored or transmitted as electrical signals and stored as mechanical recording.

BIG DATA:

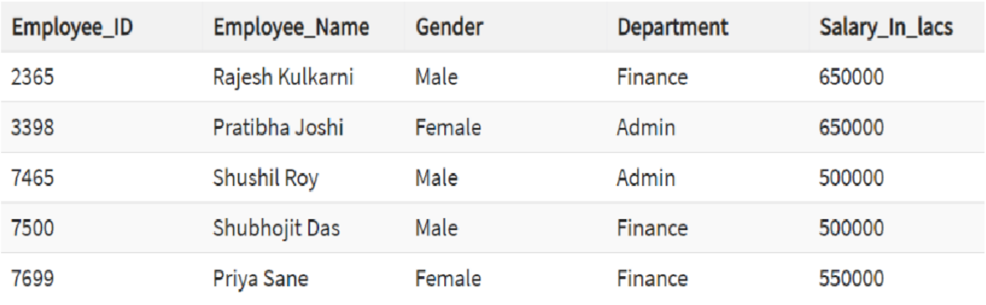
Big data is defined as the huge collection data in terms of volume, yet also growing with time, exponentially. This data being so complex in nature those traditional data-handling solutions find it difficult to store and process this data.


Data Flair
40 ZETTABYTES
(43 TRILLION GIGABYTES)
of data will be created by 2020, an increase of 300 times from 2005
2020
It's estimated that
2.5 QUINTILLION BYTES
(2.3 TRILLION GIGABYTES)
of data is created each day
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
(161 BILLION GIGABYTES)
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
facebook
Variety
DIFFERENT FROMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
YouTube
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
(100,000 GIGABYTES)
of data stored
BIG DATA
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
1 IN 3 BUSINESS
don't trust the information they use to make decisions
The New York Stock Exchange capture
1 TB OF TRADE INFORMATION
during each trading session
Velocity
ANALYSIS OF STREAMING DATA
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
- almost 2.5 connections per person on earth

TYPES OF BIG DATA:

Structured:

This type of data can be stored, assessed or processed as fixed format, hence the name Structured data. The avent in computer science has made it easier to work with structured data as the format of the data is mostly well-known in advance,so deriving meaningful insights from it is easier. Recently, we have been facing issues to handle the huge amount of data, where a typical data size is about a zettabyte (one billion terabytes).

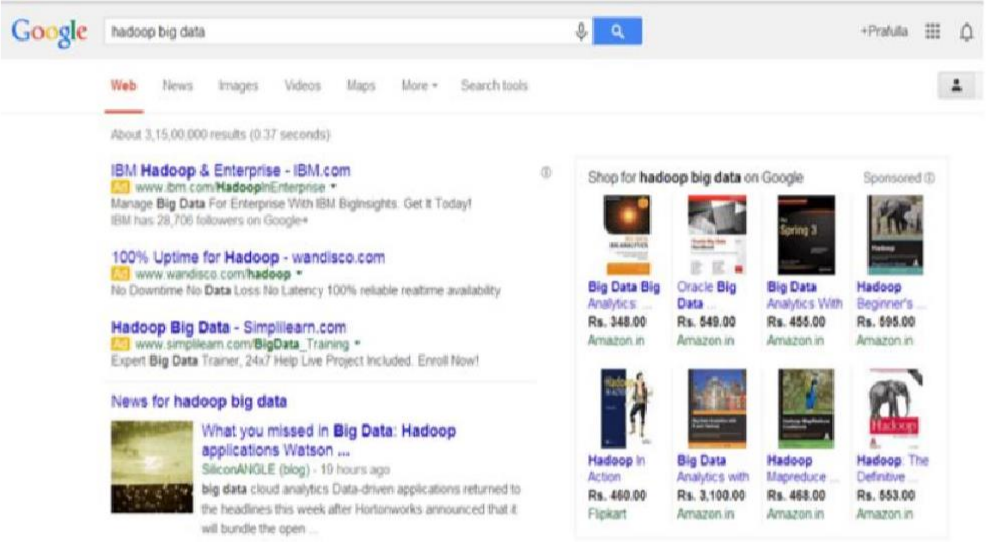


Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured:

It is a type of data having any unknown form or structure, hence the name unstructured data. In addition to its large size, what makes it difficult to process is it has various challenges, for e.g. since it is from a heterogeneous source, the data files may include anything ranging from simple text files to mp3, mp4, jpg, fly, avi etc.

Organizations might have a huge collection of unstructured data, but face difficulties in extracting meaningful insights from it, since it is in the raw form and difficult to process.



About 3,15,00,000 results (0.37 seconds)

IBM **Hadoop** & Enterprise - IBM.com
Ad www.ibm.com/**Hadoop**InEnterprise ▾
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for **Hadoop** - wandisco.com
Ad www.wandisco.com/**hadoop** ▾
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/**BigData**_Training ▾
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for **hadoop big data**

What you missed in **Big Data**: **Hadoop** applications Watson ...
SiliconANGLE (blog) - 10 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for **hadoop big data** on Google Sponsored

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle **Big Data** ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

Hadoop Beginner's ...
Rs. 595.00
Amazon.in

Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

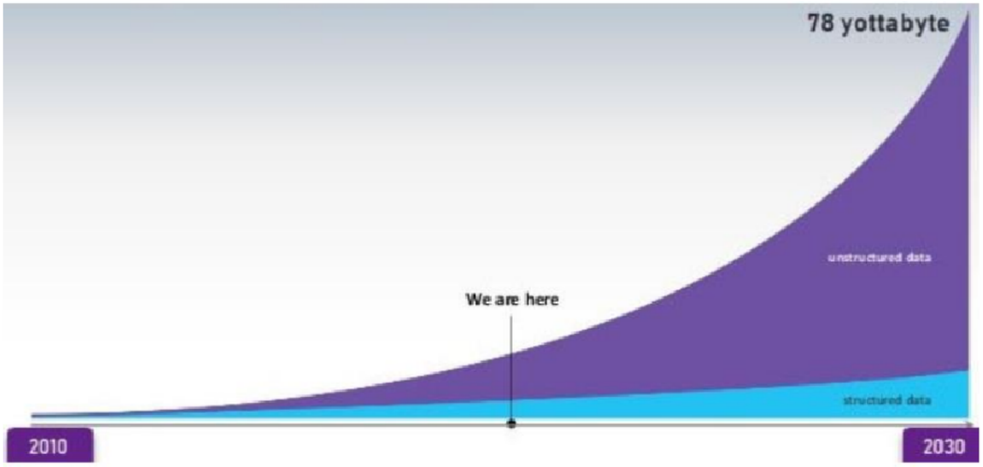
Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

Hadoop: The Definitive ...
Rs. 553.00
Amazon.in

Semi-structured:

This type of data can contain both types of data that is unstructured and structured. It is usually characterized by being structured, but not defined in a tabular structure as defined by the tables in relational databases.

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```



CHARACTERISTICS OF BIG DATA

1.) **Volume-**

It is one of the major characteristics for defining big data, since volume plays a crucial role in extracting out meaningful insights from the data. It is volume which decides if a particular set of data can be considered as big data or not.

2.) **Variety-** Variety in data arises due to heterogeneous sources of data and the nature of data. Previously spreadsheets and databases happened to be the only sources of data, considered by majority of the applications. In recent times data is obtained in various forms like images, movies, emails, monitoring-devices, audios, pdf etc. are also being included for carrying out analysis. This variety of unstructured data poses a serious challenge in storage, mining and analysis of the data.

3.) **Velocity-** Velocity refers to the rate at which the data gets generated. The real potential of the data is determined by the fact, how fast is the data generated, stored and processed for deriving value out of it. Big data velocity usually deals with the speed at which data is being generated from various sources including application logs, business processes, sensors, mobile devices, ioT devices, social media etc. The stream of data flow is continuous and humongous.

4.) **Variability-** It is defined as the inconsistency of data thus making it harder to process and derive value out of it.

ADVANTAGES OF BIG DATA:

The capability to process big data brings in certain advantages:

1.) Businesses can use outside intelligence to make decisions.

2.) Better operational frequency.

3.) Better risk management, by means of early risk identification related to products or services.

4.) Improved customer services, by means of revamped feedback systems over the traditional feedback evaluation systems.

INTRODUCTION TO SKILL TRACK REQUIRED FOR BIG DATA:

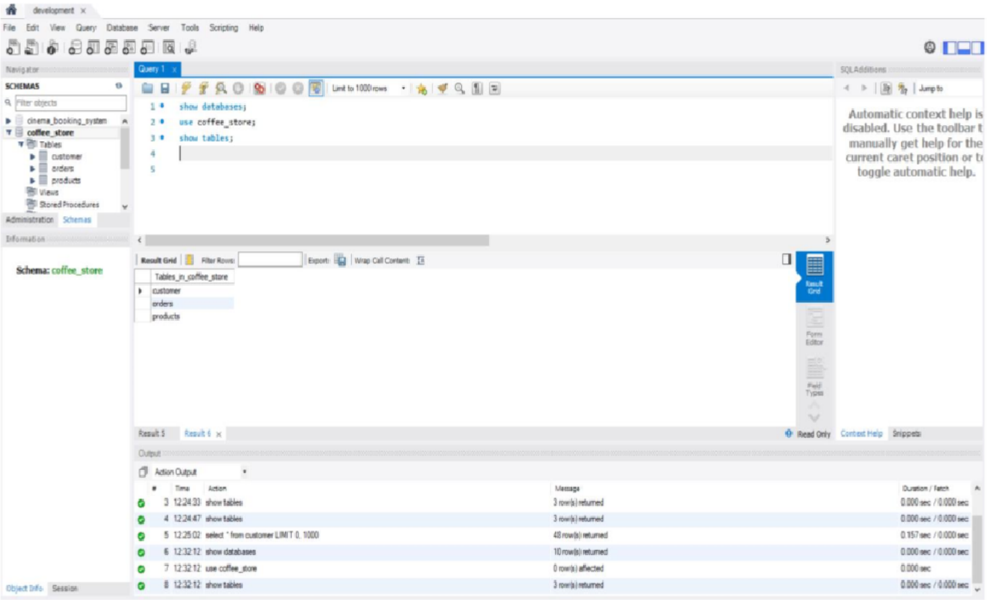
SQL-

What is a database?

A database is defined as an organized collection of data which is usually stored and accessed from a computer system. Usually relational databases are used to store and retrieve information.

A relational database is based upon a relational model of data as the name suggests. In this model the data being stored is organized into one or multiple tables consisting of various rows and columns, with each row having its own unique identifier.

The database management system is software which is used to interact with the databases.



```
show databases;
use coffee_store;
show tables;
```

Schema: coffee_store

Tables_in_coffee_store
customer
orders
products

Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.

```
 #  Time      Action                                Message              Duration / Fetch
 3  12:24:33  show tables                           3 row(s) returned    0.000 sec / 0.000 sec
 4  12:24:47  show tables                           3 row(s) returned    0.000 sec / 0.000 sec
 5  12:25:02  select * from customer LIMIT 0, 1000  48 row(s) returned   0.157 sec / 0.000 sec
 6  12:32:12  show databases                        10 row(s) returned   0.000 sec / 0.000 sec
 7  12:32:12  use coffee_store                      0 row(s) affected    0.000 sec
 8  12:32:12  show tables                           3 row(s) returned    0.000 sec / 0.000 sec
```

What is SQL?

SQL stands for Structured Query Language. It is designed for use in performing management operations in relational databases.
Most database management systems like Mysql, Oracle, Sybase, SQL Server and Informix etc use SQL primarily.

The two main differences in the SQL databases with comparison to the traditional read write operation file systems are-

1) First, it solves the problem of accessing multiple files at the same time
2) Secondly, it eliminates the need of specification of the usage of indexes

Sql was initially developed by IBM in the early 1970s, by Donald Chamberlain and Raymond Boyce. System R, which was the original IBM database at that time was being operated and manipulated by the original version of SEQUEL (Structured English query language) which was also developed at the same San Jose laboratory in the early 1970s.

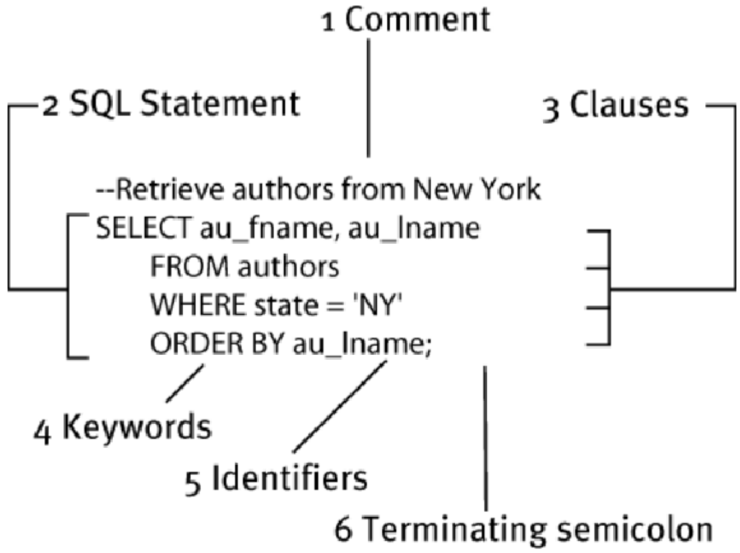
In the late 1970s, relational software, which is now known as the oracle corporation saw the potential of the system that IBM talked about and had created, so they themselves started research and development on the same and thus developed the first commercialized version of the SQL, the oracle v2 which was created to be used on the vax systems. The system was developed with the vision of selling it to the US military operations, US navy and the central intelligence agency along with many other prominent US government departments.

SYNTAX

There are multiple different constituents of the sql language, made for making it usable, readable and easier to learn and use. The various components are listed as below-

1. Clauses - clauses might or might not be used in a sql query. Some of the queries might even use multiple clauses and some might not require the use of even a single clause. Thus, these constituents of statements/expressions or queries are actually optional but very useful when required and used.

2. Expressions - expressions are the primary statements in the sql language. These are used for various basic and advanced operations which can be done by the sql language. They may be written in order t o generate some scalar values or even might return columns or rows or both combined, which are basically called tables.



```
--Retrieve authors from New York
SELECT au_fname, au_lname
     FROM authors
     WHERE state = 'NY'
     ORDER BY au_lname;
```

3. Keywords - the various keywords or predefined words are used in sql which provide a helping hand while using sql for various purposes. Some of the keywords that are used in sql are listed as follows -
   - Order by

- Select
- Where
- From

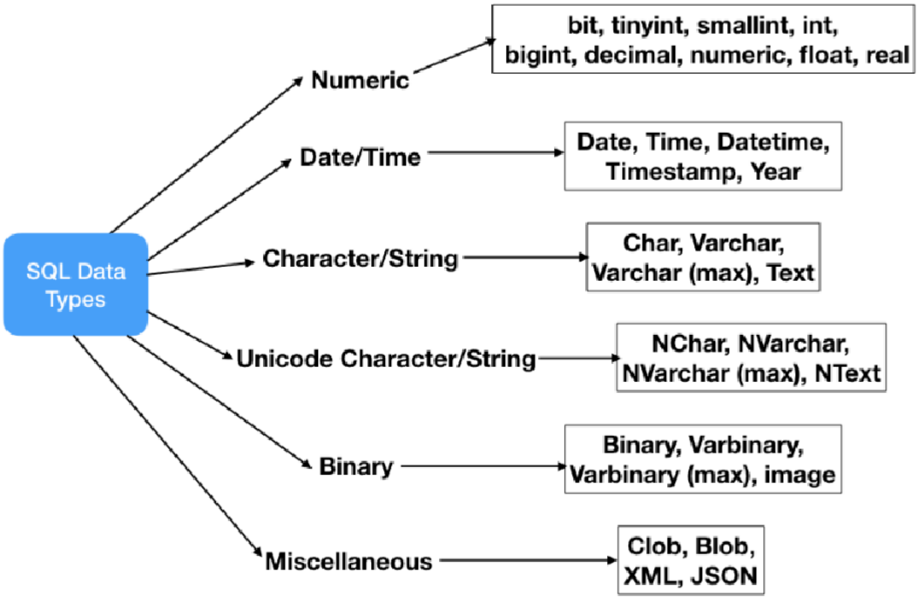
4. Statements - statements in sql may be used to control the flow, connections or diagnostics and may control the transactions etc of the program.

5. Comments - Single line comments start with --. Any text between -- and the end of the line will be ignored by the system.
Multi-line comments - they both start and end with /* and any text between /* and */ will be ignored by the system.

SQL DATA TYPES

Data types are mainly classified into three categories for every database.

- String-Data Types
- Numeric-Data types
- Date, time-Data types


SQL Data Types
Numeric
bit, tinyint, smallint, int, bigint, decimal, numeric, float, real
Date/Time
Date, Time, Datetime, Timestamp, Year
Character/String
Char, Varchar, Varchar (max), Text
Unicode Character/String
NChar, NVarchar, NVarchar (max), NText
Binary
Binary, Varbinary, Varbinary (max), image
Miscellaneous
Clob, Blob, XML, JSON

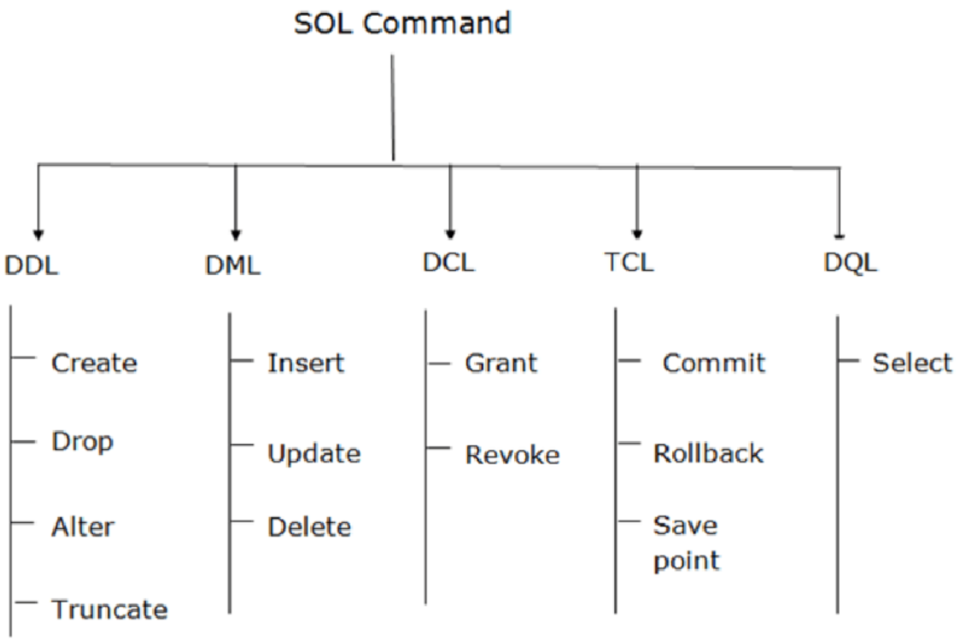
SQL DATABASE COMMANDS

- CREATE

- this is used to create a new database in the system, inside which, the multiple tables can be created and then manipulated
    - Create database database name;
    - Example - create database student_db;

- DROP - to delete an existing database in the sql schema
    - Drop database database name;
    - Example - drop database srudent_db;

Numerous commands similar to these are used in sql. A compiled summary of such commands is shown in the below picture.



DML commands

The manipulation of data that is present inside the various databases and tables inside the sql schema is done using the DML commands. There are multiple DML commands which can be used for the above stated purposes. These are known as data manipulation language commands. Following are some examples of DML commands -

- Insert - it is used to enter data into the table that already exists. The syntax for the same is as follows -
  INSERT INTO table name (col1, col2, col3) values (val1, val2, val3);

Here, the col1 etc are the columns listed and the corresponding values and val1 etc are the values that need to be entered. We can also use the same command without the column values.

- Update - it is used to modify the existing records -
  Update table name
  Set col1 = val2
  Where condition;

- Delete - it is used to delete the existing records.
  Delete from table name where condition;

DCL COMMANDS

Two commands used in DCL category are grant and revoke. They basically deal with permissions, rights etc of the users.

- Grant - provides users the access privileges to the database.
- Revoke - restricts users the access privileges to the database.

TCL COMMANDS

Deals with the transactions within the database.
Some examples are -

- Commit
- Rollback
- Set transaction
- Savepoint

JOINS

- Joins are used in SQL in order to retrieve data from multiple tables in a single select query.
- In order to access more than one table, we need to establish a single common column between the tables so that they can be connected to each other with the help of this column.
- This column having a unique value for each of the records in the table is called as the primary key in the parent table.
- The column might be an attribute of any kind and can be used as an unique value for each of the records to distinguish the records from each other.

- The column with which the primary key column matches in the other table is called as a foreign key attribute.

- Foreign key attributes also have unique values in the table and can be used to match the records throughout multiple tables.

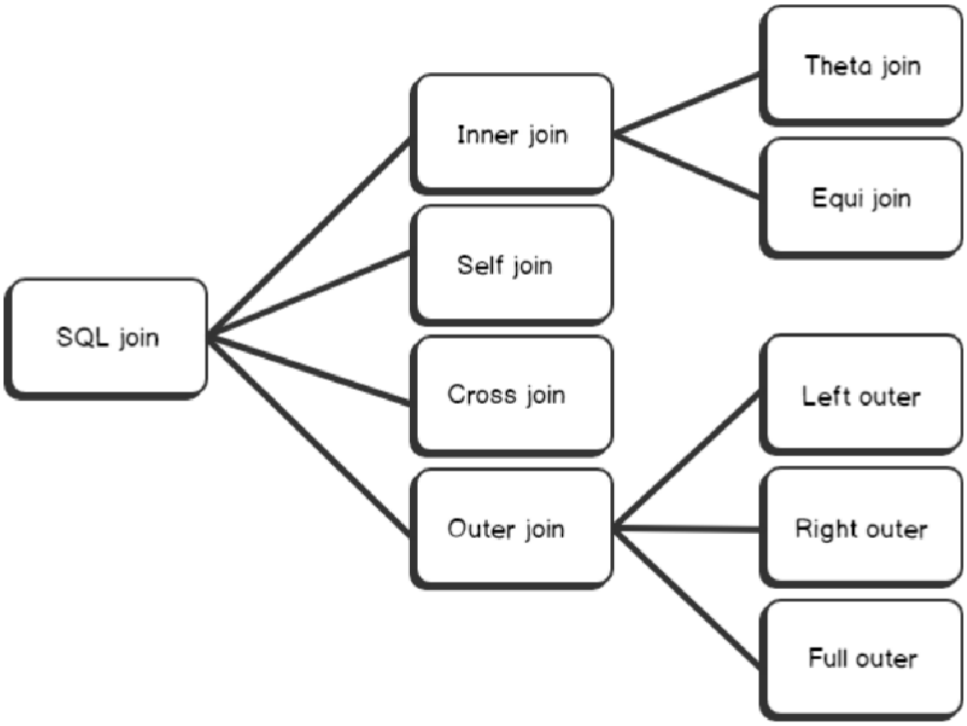
- Foreign key attribute column does not have to have the same name as the primary key and neither does it need to have the exact same values.

- But this column should have the same data type of values so that the columns can be matched with each other in order to meet the join condition

- The two columns might match all the values and for some conditions there might not even be a single match, but the condition can still be evaluated.

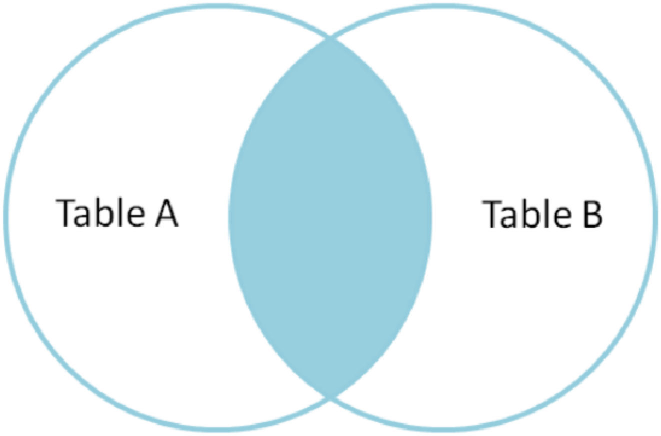
TYPES OF JOIN

There are multiple types of joins which can be used in sql to implement the usage or retrieval of data from multiple tables at the same time. Some of the joins are demonstrated as follows -



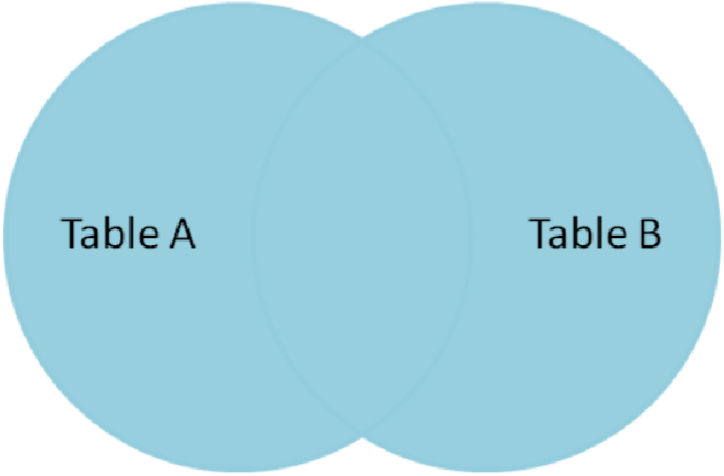
- INNER JOIN

  It gives all the rows as long as the condition satisfies. All the respective rows from all the tables that are being queried will be returned by using the keyword inner join till the condition of join matches. For those cases, in which the condition does not match, we won't get the rows returned.



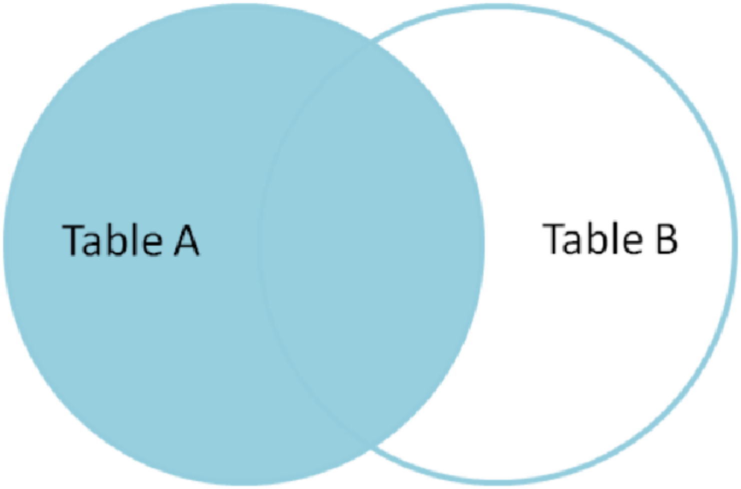
- FULL JOIN

  Returns the result of the full join query [29] by combining the results of both the left join and right join. It contains all the rows from both the tables, irrespective [6] of any matches or no matches, basically retrieves the combined version of both the tables



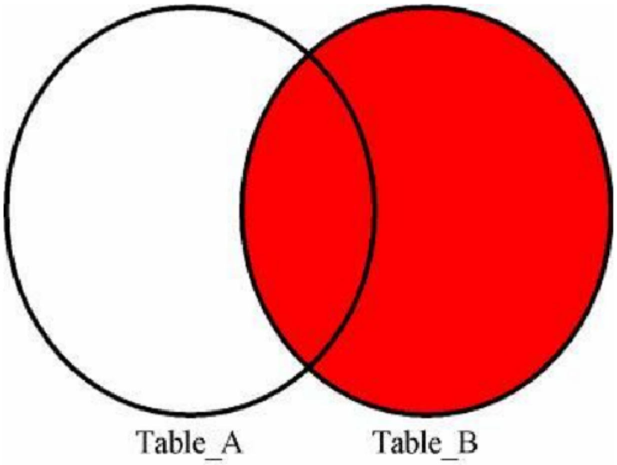
- LEFT JOIN

This join is used to retrieve all the rows of the table which is written on the left hand side of the join while only the matching rows of the table on the right. The results which don't match have null in the columns from the table on the right. This type of join is also called the left outer join.



- RIGHT JOIN

This join is used for retrieving the results such that all the columns and rows of the table on the right side of the join are present in the result along with the results from the left table for only the columns which match on the join condition. This join is also called the right outer join.



SUB QUERIES

- Sub queries are queries written nested inside other queries
- They can be used in select, insert, update and delete queries
- The nested part can be inside the from or where clause
- These can be of two types -
  - Non-correlated
  - correlated

Non correlated sub queries - in this type of sub queries, the inner query can run independently of the outer query.

- Inner query runs first and generates a result, which is then used by the outer query.
- It runs only once

Correlated sub queries - the inner query cannot run independently of the outer query but is dependent

- The inner query runs for every row in the outer query
- Might have columns named as a. Column

AGGREGATE FUNCTIONS

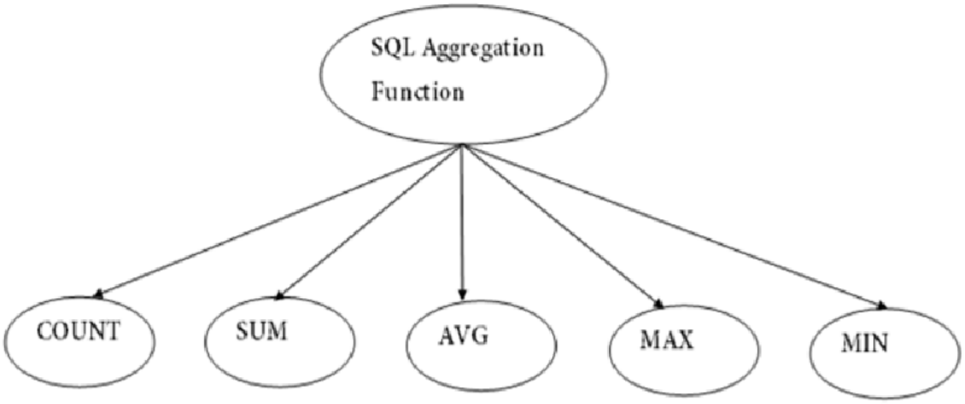
- COUNT - it counts the number of records

- Can use distinct with count to generate the result with unique values
- Does not count the null values
- Select count(*) from table name;

- SUM - it sums the values of a column
    - Can't do a sum on *
    - Columns should have summable data types(numeric)
    - Select sum(no seats) from rooms;

- AVERAGE - returns the numerical average of the columns
    - Can't use where clause
    - Can only select one column to average
    - For multiple columns, use GROUP BY

- MIN and MAX - selecting minimum or maximum value of a column
    - Can use where
    - Can use other columns after the select
    - Gives a single record as result



CHAPTER – 02

JAVA

Java is a type of high level, class-dependent [31] object-oriented programming language developed by James Gosling at Sun Microsystems. It first surfaced in May (23) 1995. Java is now owned by Oracle. Being a general purpose language it is used for application development (client-server web applications). The greatest feature of java was its architecture/platform independence, which means a java code/program written in one machine, can directly be executed in any other machine having java components, without the need of actually recompiling the program on the new machine.

James Gosling initially named the language Oak, based upon an oak tree outside his office. Following this the project was named to Green, and later renamed to Java, based upon the Java coffee from Indonesia. The first iteration of java was released for interactive televisions, but it was far ahead of its time [2] for the digital cable television industry. It was developed with a syntax similar to C/C++ to allow familiarity for the developers.

Principles

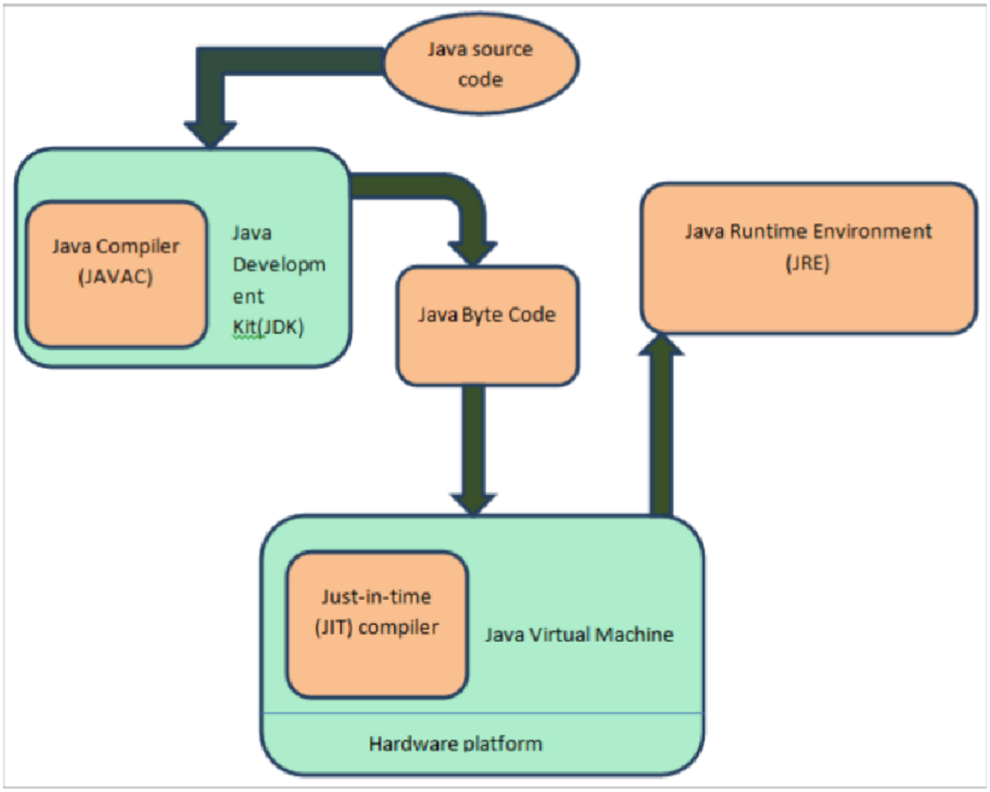
1.) The language [2] must be simple, object-oriented and familiar
2.) It must be platform neutral and portable
3.) It must be secure and robust
4.) It should execute with high performance

5.) It must have the ability to be interpreted, threaded and being dynamic.

Components of Java Language

JAVA DEVELOPMENT KIT - It is the core component, and it contains java compiler, java runtime environment, debugger etc. It is utilized for development purposes since it provides access to all executables and binaries along with other tools required to compile, execute and debug the program. Some of its internal components are:

JConsole- The java management/monitoring console
Javap- A tool for class-files disassemble
Jar- This tool is used to archiving package related libraries into a single file
javadoc- It utilises comments from source code to generate documentation
jrunscript- It is used to help execute java queries from command-line interface



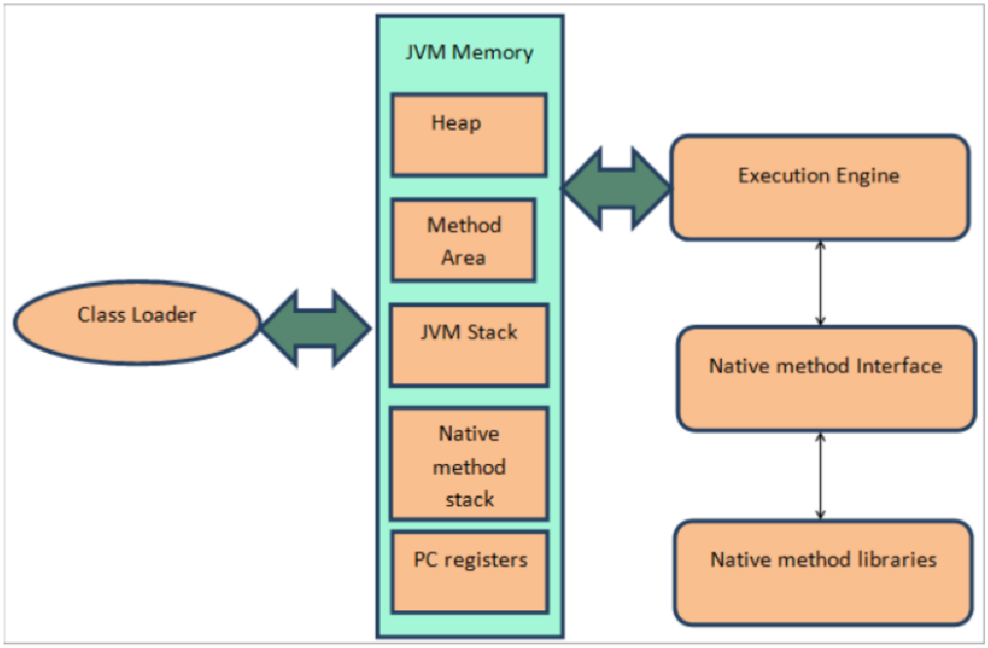
JAVA RUNTIME ENVIRONMENT- It is required for execution of java programs and applications. The JRE consists of components like Java Virtual Machine, which houses the binaries required for successful execution of any java program. Some of its components are: Files needed for management of security reasons.

DLL files
Code libraries, properties/resource files
Java extension files
Applet support files

THE JAVA VIRTUAL MACHINE- The JVM is a core component of the java language. IOT translates the byte-code into code that is understood by the machine. It also provides functionality for automatic memory management, garbage collection, security etc. It is platform independent thus allowing us the flexibility to write a java codes anywhere and executes it anywhere.

The JVM is usually present on RAM, therefore upon conversion of source file to class file, it needs to be executed. The class loader is accountable for the linking, loading and initialization of the program source code to be executed.

JVM also has the Just-In-Time compiler (JIT) which is responsible for the interpretation of a part of the byte code, which has similar functionality at the same time. Hence, Java is both a compiled and interpreted language.



JAVA COMPILER- It is the compiler for the Java programming language and its main function happens to be the conversion of java source code into java class files, following whose generation it is interpreted or compiled by the Java Virtual Machine using the Just In Time (JIT) compiler.

TYPES OF JAVA APPLICATIONS

Standalone- This type of applications is used for desktop/windows based applications, and need to be installed on every machine, e.g. Antivirus softwares.

Enterprise Applications- They are usually distributed in nature like banking applications. They have higher security, clustering, load balancing etc.

Web Applications- These applications run on the server side and create a dynamic page known as web application.

Mobile Application- These include applications created for running on mobile devices.

JAVA EDITIONS

Java SE- This is the standard edition, and contains all the java programming API's like java.sql, java.lang etc and other core stuff of OOP's like regex, multi-threading etc.

Java EE- The Enterprise Edition is used to develop applications for enterprises and web applications. This is based over the Standard Edition.

Java ME- The Micro Edition, this platform is used for developing mobile applications.

JavaFX- Used for developing richer web/internet applications.

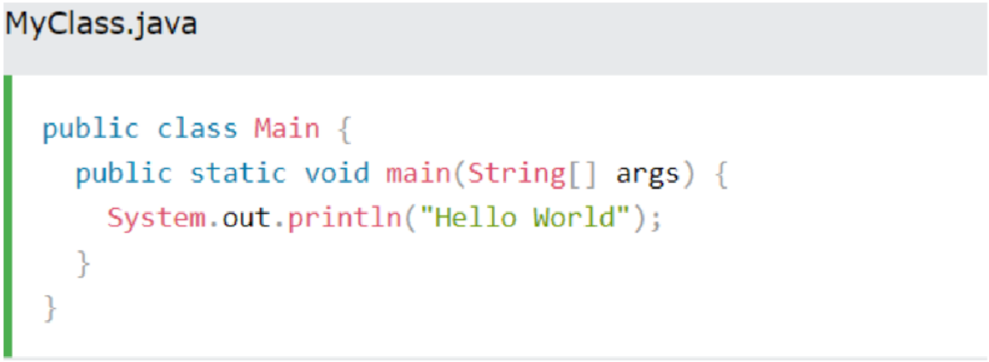
SYNTAX

Each java program must be enclosed inside a **class,** whose name should always start with an uppercase letter. Another requisite is the match between the project file name with the class name.

It is usually preceded by the **main()** method which gets executed having any code inside it. Any program needs to have a class and a main() method.

The **println ()** method is used inside the **main()** method to output information on the screen.

Curly braces {} mark the beginning and end of a block of the code. A semicolon (;) is used at the end of each sentence, to mark the end of that sentence.

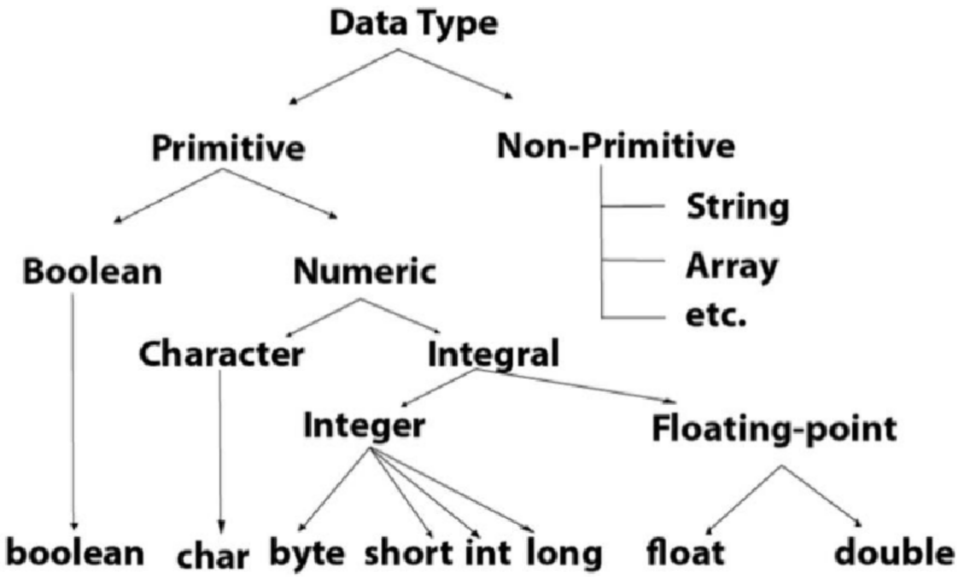


MyClass.java

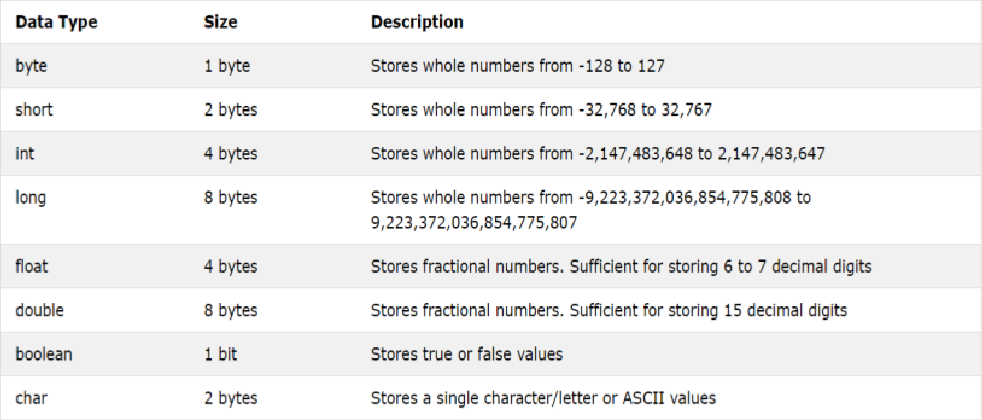
```
public class Main {
  public static void main(String[] args) {
    System.out.println("Hello World");
  }
}
```

JAVA DATA TYPES

Java has two categories of [28] data types- Primitive and Non-Primitive data types.



Primitive Data types- They include data types like int, byte, long, float, short, double, char, Boolean.



Data Type	Size	Description
byte	1 byte	Stores whole numbers from -128 to 127
short	2 bytes	Stores whole numbers from -32,768 to 32,767
int	4 bytes	Stores whole numbers from -2,147,483,648 to 2,147,483,647
long	8 bytes	Stores whole numbers from -9,223,372,036,854,775,808 to 9,223,372,036,854,775,807
float	4 bytes	Stores fractional numbers. Sufficient for storing 6 to 7 decimal digits
double	8 bytes	Stores fractional numbers. Sufficient for storing 15 decimal digits
boolean	1 bit	Stores true or false values
char	2 bytes	Stores a single character/letter or ASCII values

Integer- It stores positive, negative or whole number values, not having decimals. The valid data-types are byte, shorting and long.

Float- It stores positive, negative or whole number values having decimal points, representing the fractional part. The valid data types are float and double.

Boolean- It is declared along with the Boolean keyword and evaluates to either true or false.

Character- Used for storage of a single character

Non-Primitive Data types- These refer to objects hence are also known as reference types. They differ from primitive data types in some aspects like- they are not predefined as in the case for primitive data types and are created during programming. The primitive data types need to have a value, while non primitive can be null. The non-primitive data types start with an uppercase alphabet while the primitive data types start with a lowercase letter. Some examples of non-primitive data types are- String, Arrays, Classes, Interfaces etc.

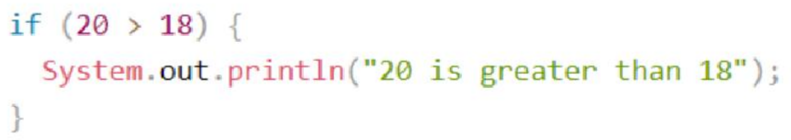
String- The String data type is generally used to store a sequence of characters. The characters must be enclosed within a pair of double quotes.

Arrays- They are utilized to store multiple values inside a single variable, instead of the need to declare multiple variables.

CONDITIONAL STATEMENTS

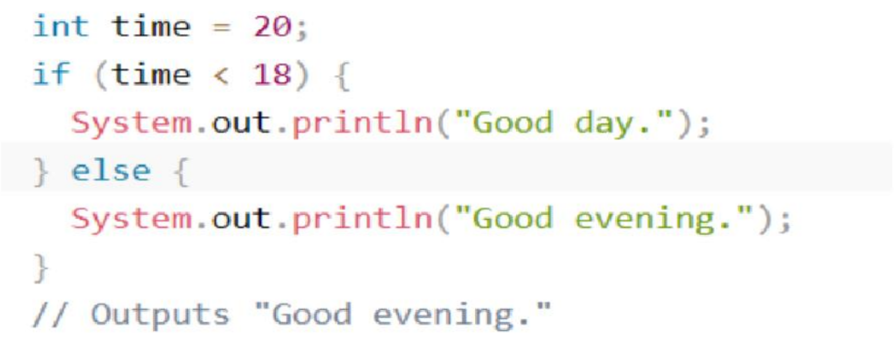
Java supports the general logic from mathematics like less than, greater than, equal to etc. which can be applied in programs. Some of the used conditional statements in java are;

If- Specifies a code block to be executed, if the condition evaluates to true.



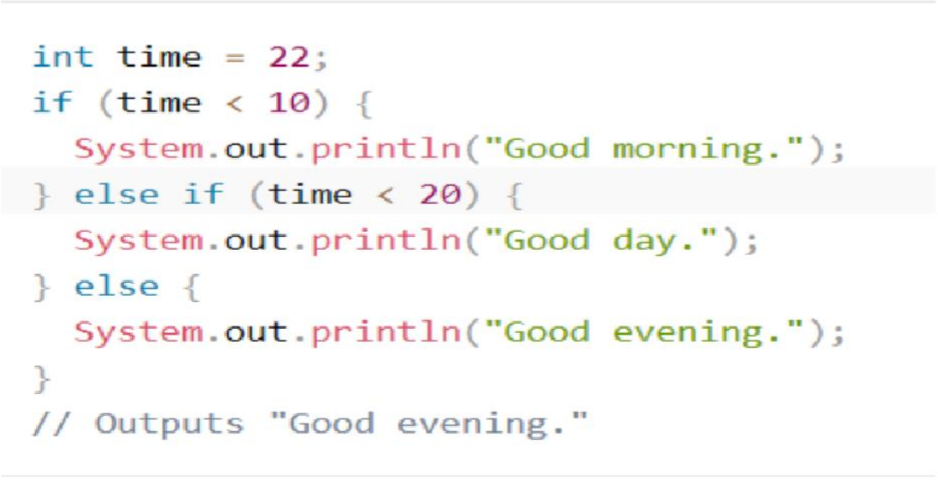
```
if (20 > 18) {
  System.out.println("20 is greater than 18");
}
```

Else- Specifies a code block to be executed, if the same condition evaluates to false.



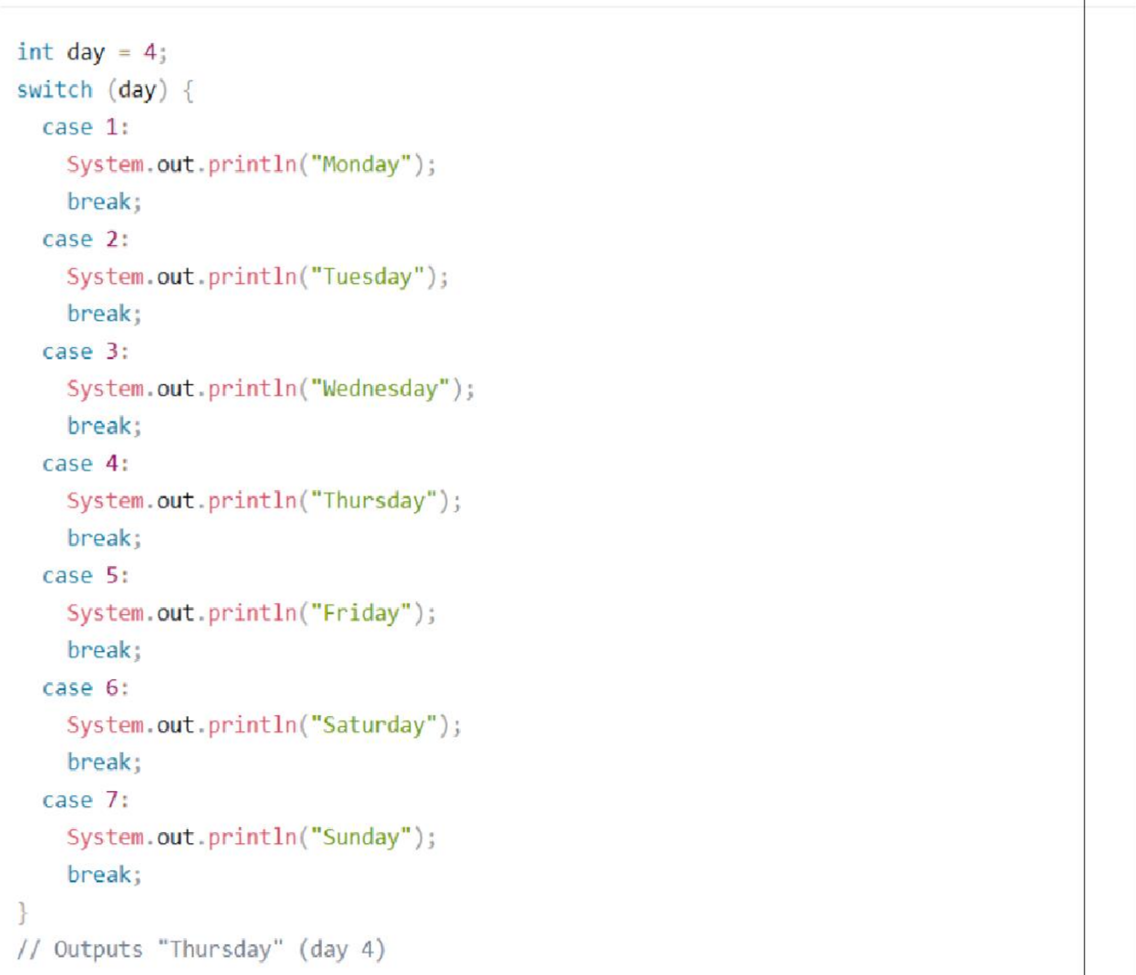
```
int time = 20;
if (time < 18) {
  System.out.println("Good day.");
} else {
  System.out.println("Good evening.");
}
// Outputs "Good evening."
```

Else if- Specifies a new code block to be executed if the first condition evaluates to false.



```
int time = 22;
if (time < 10) {
  System.out.println("Good morning.");
} else if (time < 20) {
  System.out.println("Good day.");
} else {
  System.out.println("Good evening.");
}
// Outputs "Good evening."
```

Switch- Specifies various alternative code blocks to be executed.



```
int day = 4;
switch (day) {
  case 1:
    System.out.println("Monday");
    break;
  case 2:
    System.out.println("Tuesday");
    break;
  case 3:
    System.out.println("Wednesday");
    break;
  case 4:
    System.out.println("Thursday");
    break;
  case 5:
    System.out.println("Friday");
    break;
  case 6:
    System.out.println("Saturday");
    break;
  case 7:
    System.out.println("Sunday");
    break;
}
// Outputs "Thursday" (day 4)
```

CHAPTER – 03

UNIX AND SHELL SCRIPTING

What is UNIX?

UNIX is actually a family of operating systems, having the capability of multitasking, and multi user access at a same time. It's development kick started in 1970's at the AT&T's Bell Labs research center, by Ken Thompson, Brian Kernighan, Dennis Ritchie and others.

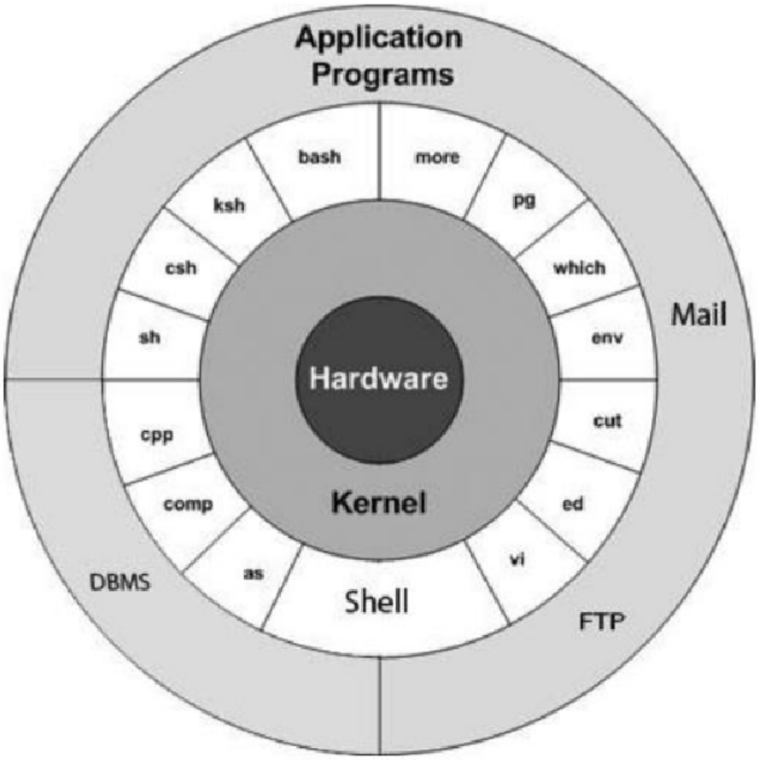
The operating system UNIX is a set of commands/programs that fuel as a link between the user and the computer system. An Operating System is a set of computer programs which allocate the system resources and further coordinate all the details of the available system internals. It is also referred to as Kernel.

The users use a **shell** to interact/communicate with the **kernel.** The **shell** is a command-line interpreter whose function is to translate the commands inputted by the user into a language which is understood by the **kernel** and thereby executing the given command. Various distributions/flavors of UNIX are available in the market like AIX, Solaris UNIX,HP UNIX etc. These are commercially licensed copies, while LINUX and its various distributions are open source and freely available. Since UNIX allows multiple programs to be executed at a single time, it is referred to as a multitasking operating system. Also since it allows multiple users to login at the same time, it is also a multiuser operating system.

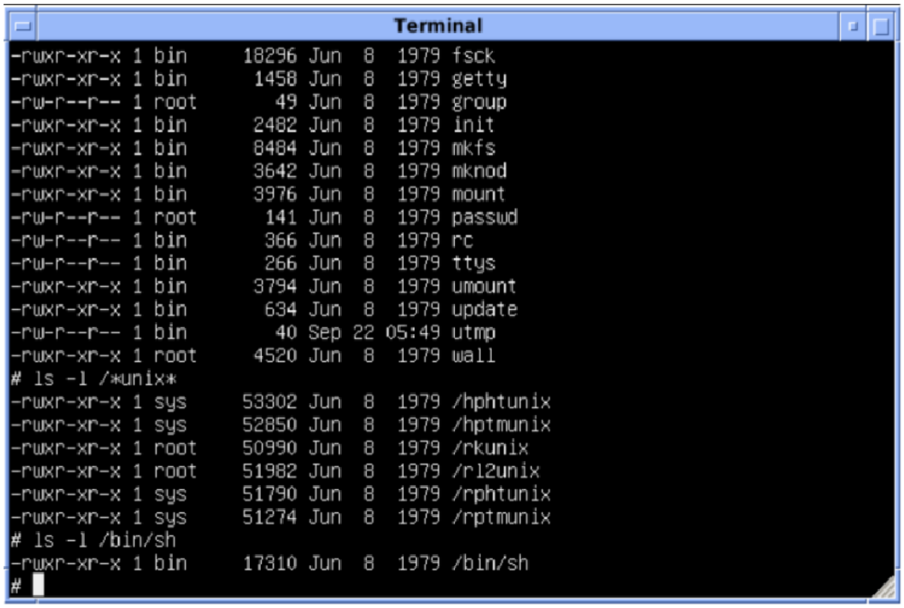
UNIX Architecture

There are four basic components of UNIX operating system-

1.) **Kernel-** It is referred to as the heart of an operating system. It is the main component which interacts with the hardware and takes care of tasks like memory and file management, task scheduling etc.

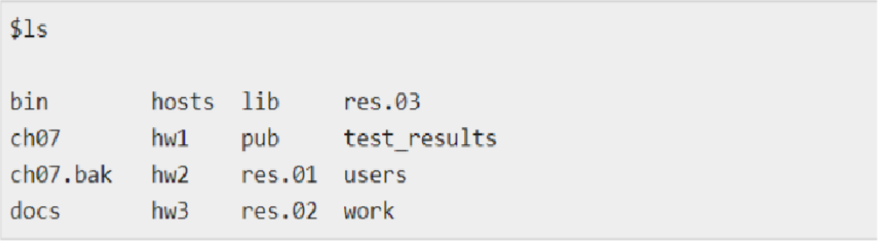


2.) **Shell-** A shell is a utility which processes the commands given as input in the terminal, processes them and then calls the required program to execute the task. The shell follows similar syntax for all the commands. Some types of shell available in UNIX are C shell, Korn shell, Bourne shell etc.



```
-rwxr-xr-x 1 bin     18296 Jun  8  1979 fsck
-rwxr-xr-x 1 bin      1458 Jun  8  1979 getty
-rw-r--r-- 1 root       49 Jun  8  1979 group
-rwxr-xr-x 1 bin      2482 Jun  8  1979 init
-rwxr-xr-x 1 bin      8484 Jun  8  1979 mkfs
-rwxr-xr-x 1 bin      3642 Jun  8  1979 mknod
-rwxr-xr-x 1 bin      3976 Jun  8  1979 mount
-rw-r--r-- 1 root      141 Jun  8  1979 passwd
-rw-r--r-- 1 bin       366 Jun  8  1979 rc
-rw-r--r-- 1 bin       266 Jun  8  1979 ttys
-rwxr-xr-x 1 bin      3794 Jun  8  1979 umount
-rwxr-xr-x 1 bin       634 Jun  8  1979 update
-rw-r--r-- 1 bin        40 Sep 22 05:49 utmp
-rwxr-xr-x 1 root     4520 Jun  8  1979 wall
# ls -l /*unix*
-rwxr-xr-x 1 sys     53302 Jun  8  1979 /hphtunix
-rwxr-xr-x 1 sys     52850 Jun  8  1979 /hptmunix
-rwxr-xr-x 1 root    50990 Jun  8  1979 /rkunix
-rwxr-xr-x 1 root    51982 Jun  8  1979 /rl2unix
-rwxr-xr-x 1 sys     51790 Jun  8  1979 /rphtunix
-rwxr-xr-x 1 sys     51274 Jun  8  1979 /rptmunix
# ls -l /bin/sh
-rwxr-xr-x 1 bin     17310 Jun  8  1979 /bin/sh
# 
```

3.) Commands/Utilities- Unix houses various commands to perform everyday tasks like copying files, making directories or files, adding lines to a file, counting the number of lines and words in a given file etc. Some of the commands are - ls, cp, grep, mkdir, cat etc.



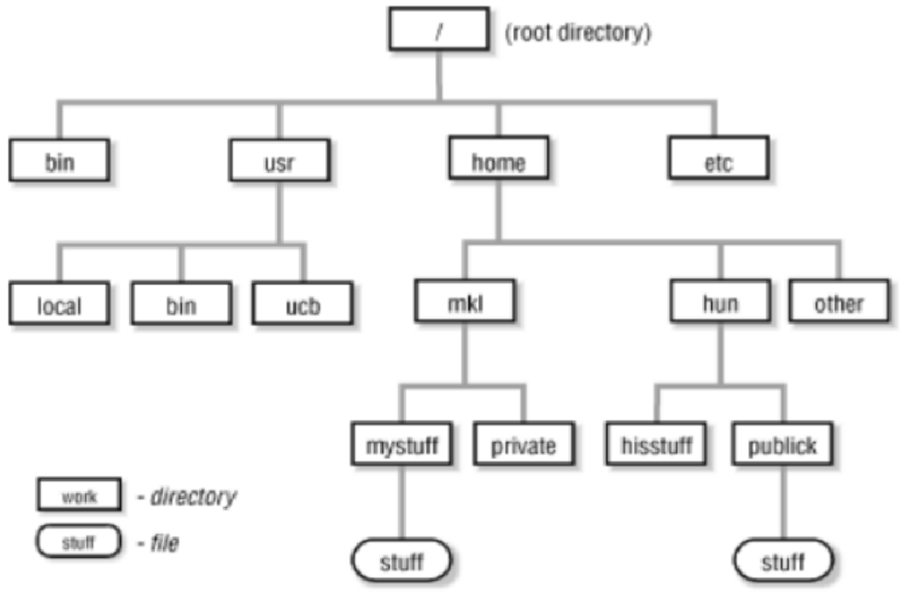
```
$ls

bin        hosts  lib     res.03
ch07       hw1    pub     test_results
ch07.bak   hw2    res.01  users
docs       hw3    res.02  work
```

4.) Files/Directories- UNIX follows a tree-like structure for the organization of directories. The data is organized into files and files are placed inside the directories. There are basically three types of files present in the UNIX file system-

a.) Directories- They store special as well as regular files in it. It is equivalent to folders in the Windows operating system.

b.) Ordinary files- It contains data as text files or programs.

c.) Special files- They provide access to hardware components like CD drive, network adapters etc.



IMPACT OF UNIX

Unix has a tremendous impact on operating systems, as its portable, is available at a nominal price for educational and research purposes, runs on even the hardware with lowest configurations and can be adapted easily to different systems or machines.

UNIX based LINUX is highly utilized for high end servers dedicated for storage and data processing. It also popularized its hierarchical file system with nested sub-directories. Since UNIX is majorly written in C language it makes it easier to work with and work on any kind of system. The architecture and design of UNIX is so appealing that the tech giant Apple keeps it as the core of their Mac OS operating system. Many businesses thrive upon UNIX for their regular business operations. Working knowledge of unix is recommended for establishing familiarity when working on big data, and various tools associated with it, since the associated tools and frameworks like hadoop use a similar query as found in UNIX like ls, cat, rm, rmdir, mkdir, etc.

PYTHON

- [39] It was created by Guido van rossum
- It is used by multiple platforms like –
  - YouTube
  - Quora
  - Drop box
  - Reddit
  - Bit torrent

The various applications of python are-

- Image processing
- Graphic design
- 3D modeling
- Scientific data processing

Idle – integrated development environment tool allows us to write and run our code easily with a simple interface. Code is written in the python prompt after >>>

Characteristics of idle –

- Written in python
- Uses tkinter graphics library
- Has an interactive python shell
- A full featured text editor
- A debugger

Python is an interpreted language, which basically means that it uses an interpreter in place of a compiler. An interpreter takes one instruction at a time and executes it in real time.

Salient features of python –

· High level programming language – python is a high level programming language and exhibits the features of a high level language like code readability and easy usage.

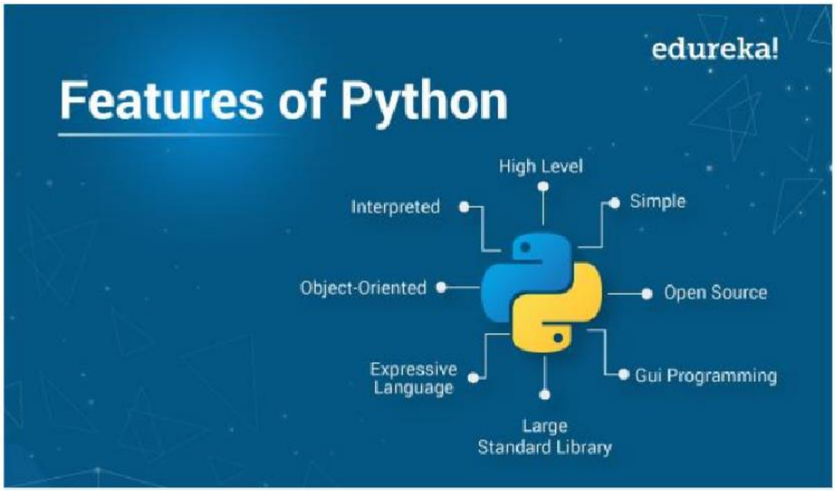
· Open source – python is free to use and can be used for personal and professional work free of cost

· Supports multiple programming paradigm – it supports object oriented programming, imperative, procedural and functional programming

· Extensible – python can easily be used combined with different languages and frameworks with simple extensions and commands

- Gui programming – creating buttons, text boxes, widgets is easy and doable in python with its different technologies and tools

- Embeddable – python language is embeddable and can be used for embedded programming of certain devices, advanced technologies using its imperative programming paradigm



Python 2.xx versus python 3.xx

- Python 2.0 came out in 2000 while the python 3.0 came out in 2008

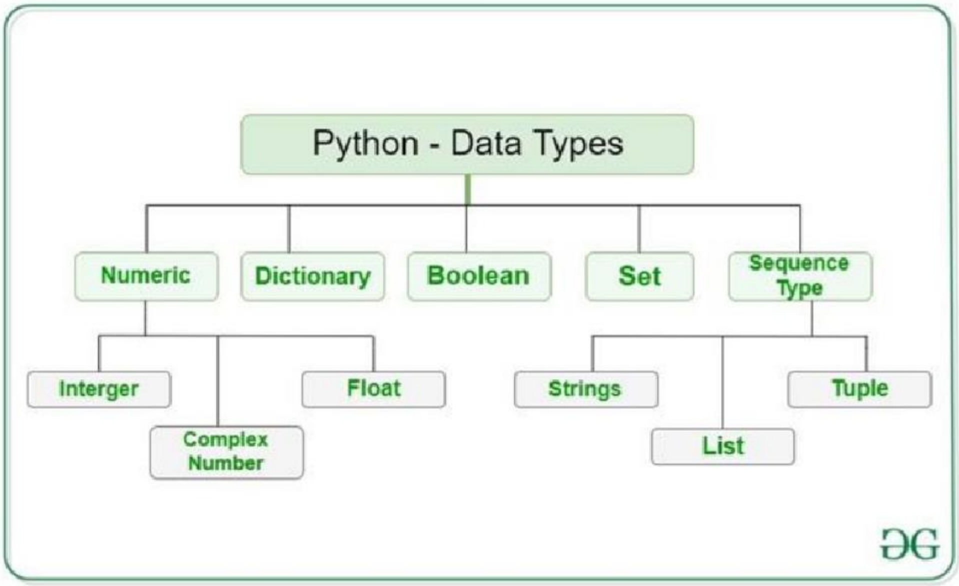
- Print is now a function and not a keyword as in the earlier version. Parenthesis are now made compulsory to use while writing the print command

- Input function is used instead of the previous raw input while doing the same job

- Results of the arithmetic division operations are now calculated as decimals only

- Stores strings as Unicode by default

- Integer objects are long by default and don't require L as 1000L



DATA TYPES IN PYTHON

It is the classification of data items. The most common types of data types are numeric, non-numeric and Boolean. Knowing the data type helps us to understand what kind of operations and applications can be created with the usage of the available data.

The four broad classifications of data are –

- Numeric
- Boolean
- Sequence
- Dictionary

1) NUMERIC

Data values, which are numeric in nature comes under numeric data types. They are of three types –

Integer – positive or negative whole numbers. Example – 2, -5

Float – real numbers with floating point representation. Example – 1.3E, -2.8

Complex – any number with a real and an imaginary component. Example – 2+3i

2) BOOLEAN

Any representation of data, which has two values denoted by true or false is called a Boolean data type.

3) SEQUENCE

Some of the built in sequence data types are –

· String – a collection of one or more characters in single, double or triple quotes. Example – ‘hello’, ‘’abc’’, ‘’’pqr’’’

· List – an ordered collection of one or more data items, not necessarily of same type in square brackets represent a list. Example – [1, ’ram’, 2.4, True]

· Tuple –multiple data items in a combined formed, when put in a parenthesis make a tuple. Contents of a tuple cannot be modified. Example – (1, ‘ram’, 2.4, True)

4) DICTIONARY

An unordered collection of data in key: value pair form. Collection of such pairs is enclosed in curly braces. Example – {1: ”Aditya”, 2: ”Akash”, 3: ”Ajay”}

VARIABLE - a name given to an object and not to the memory location. Its value can vary throughout the execution of the program. It is not bound to a single data type but the data which is entered at the position determines the type of the variable.

That’s the reason why python is known as a dynamically typed language.

ARITHMETIC OPERATORS

(PEDMAS)

- Addition - Adds the two operands on both the sides of the operator. Denoted by (+) symbol

- Subtraction – subtracts the right operand from the left operand. Denoted by the (-) symbol

- Multiplication – denoted by (*)

- Division – denoted by (/)

- Modulus – (%)

- Exponent – (**)

- Floor division – (//)

STRING DATA TYPES

- Single, double or triple quotes

- If the string has apostrophe, use double quotes, otherwise use interchangeably

- Or if it has double quotes in it, use single quotes to enclose the string

- For multi-line strings, use triple quotes

Accessing characters in a string-

A string is an ordered set of characters. Therefore, each character has an index. The first character has the index 0 and so on. To access the character, we write the name of the particular list with the character index in the [].

STRING OPERATIONS

- Concatenation - multiple strings are appended with each

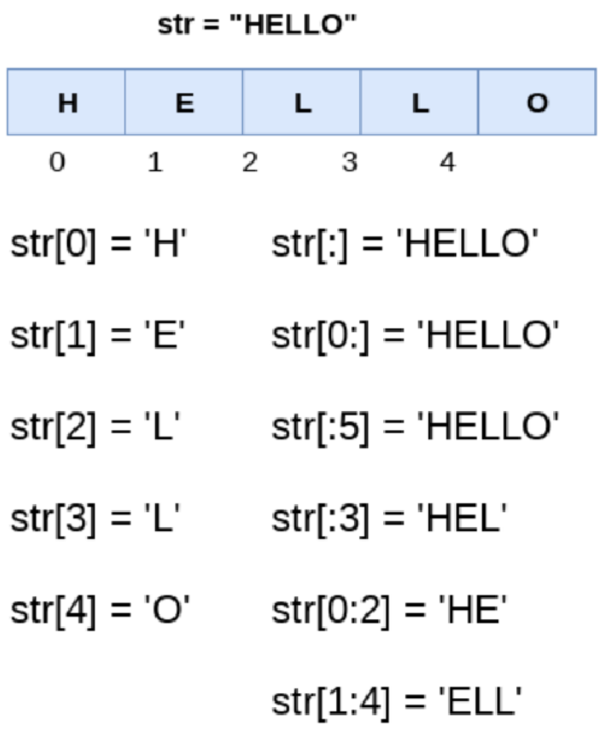
- Repetition – concatenates multiple copies of same string

- Slice – it gives the character at any given index

- Range slice – fetches characters in the range specified by two separate indexes

- In membership – returns true if the substring or character is present in a string

- Not in membership – returns true if character is not in the string



str = "HELLO"

H	E	L	L	O
0	1	2	3	4

```
str[0] = 'H'      str[:] = 'HELLO'

str[1] = 'E'      str[0:] = 'HELLO'

str[2] = 'L'      str[:5] = 'HELLO'

str[3] = 'L'      str[:3] = 'HEL'

str[4] = 'O'      str[0:2] = 'HE'

                  str[1:4] = 'ELL'
```

PYTHON CONTROL STRUCTURES

Python scripts are beneficial as –

· Easily insert/delete/update statements

· Code or functions from the script can be imported

· Automate and schedule tasks

Conditional and repetitive execution –

Conditional statements are used for executing a block of code if some particular condition is true.

Repetitive statements are used for executing the blocks of code for some multiple times.

USING CONDITIONALS

- IF

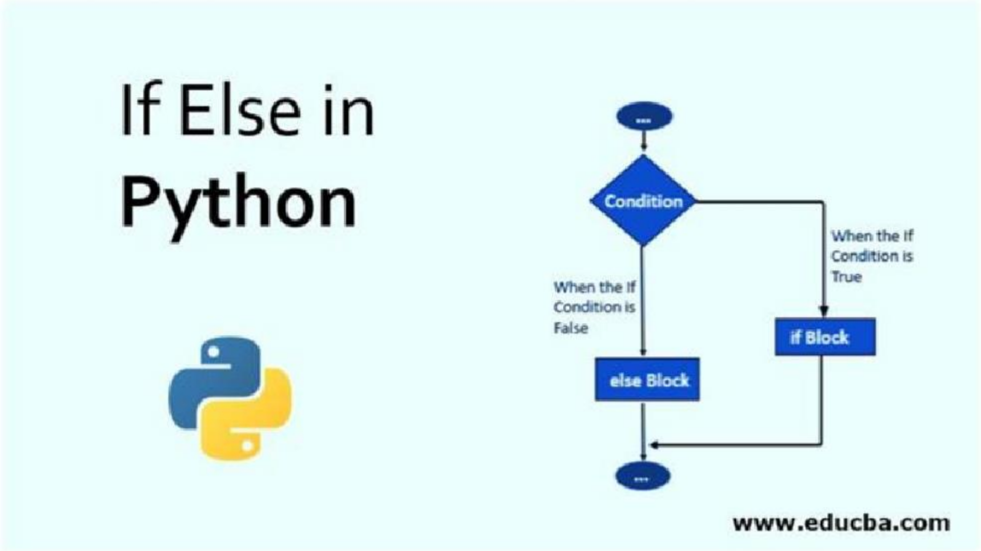
If expression value is calculated, if true, statement 1 is executed and then statement 2 is executed. If false, directly statement 2 is executed.

- ELSE

When alternate situations are required with if, we use else.

- ELIF

For multiple conditions, to reduce the else if indentation and complexity, we use elif keyword.



LOOPS

If the program flow is redirected towards any of the earlier statements, it is known as a loop. We need to specify some conditions for the loop to stop, to prevent it from going into an infinite loop.

- While loop

```
While expression:
   Statement 1
   Statement 2
Statement 3
```

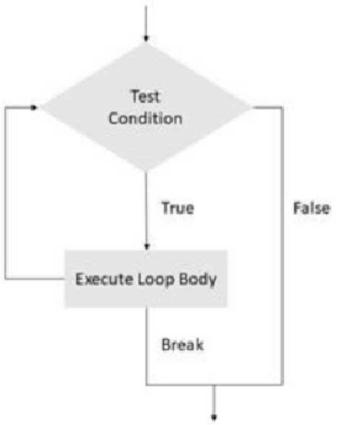
When the expression is true, the body of the loop is executed, when it becomes false, control comes out of the loop.

- For loop

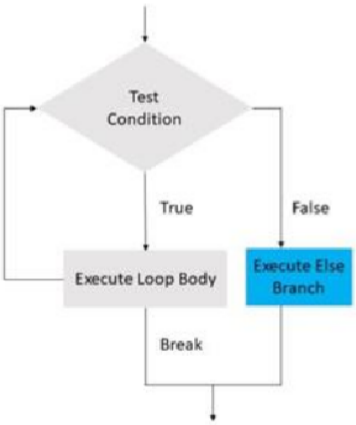
```
For variable in sequence:
   Statement 1
   Statement 2
Statement 3
```

Only sequences are iterable in for loop. For numbers, use range.

Normal Loop Program Flow



Loop Program Flow with Else



USING FUNCTIONS

Independent and reusable blocks of instructions are called functions. Dividing a complex problem into functions for each program is called modular programming. Makes the code easy to develop, follow and maintain.

Modular programming also takes a top down approach towards programming. When we call a function, it performs the task and returns the control to the calling routine.

- Organizes the code
- Makes it more readable

```
Def function name ():
  Statements
  Return statement
```

Functions with arguments

Arguments are the parameters passed in the parenthesis of the function which it can use in performing a task. Two ways of specifying the arguments –

- Simply include the argument in parenthesis while calling

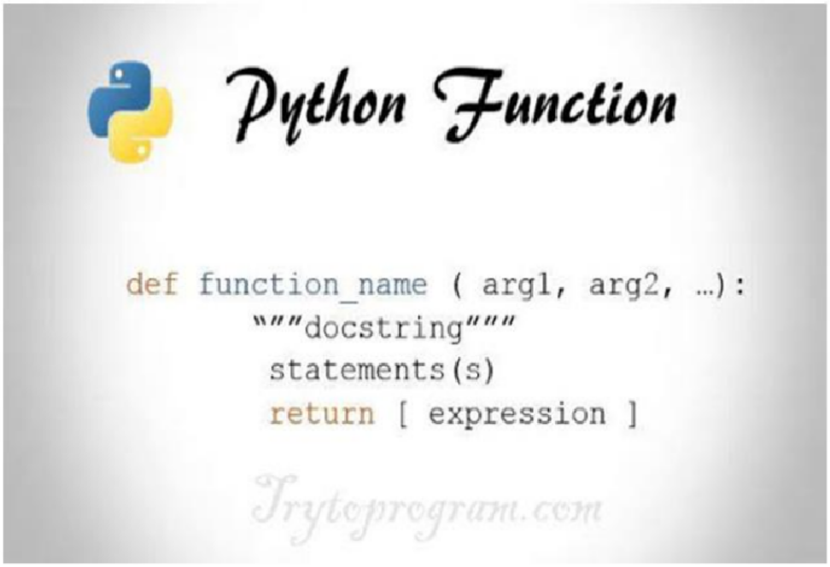
```
Def myfunc([4]name):

      Print("Hello{}".format(name))

      Return
```

To call the function,

Myfunc("Aditya"):

- The other way is to use variables

```
[4]
Def sayhello(name):

  Print("Hello{}".format(name))

  Return
```



Python Function

```
def function_name ( arg1, arg2, …):
      """docstring"""
       statements(s)
       return [ expression ]
```

CHAPTER - 04

DATA WAREHOUSE

The collection and management of data from various multiple sources is termed as data warehousing. This process is done keeping in mind the purpose, which is deriving meaningful insights from the data and enabling a better data driven decision making process. Heterogeneous sources might be used to draw data and the business data or the transactional data is primarily analyzed using advanced tools and technologies so that some useful information can be extracted from them.

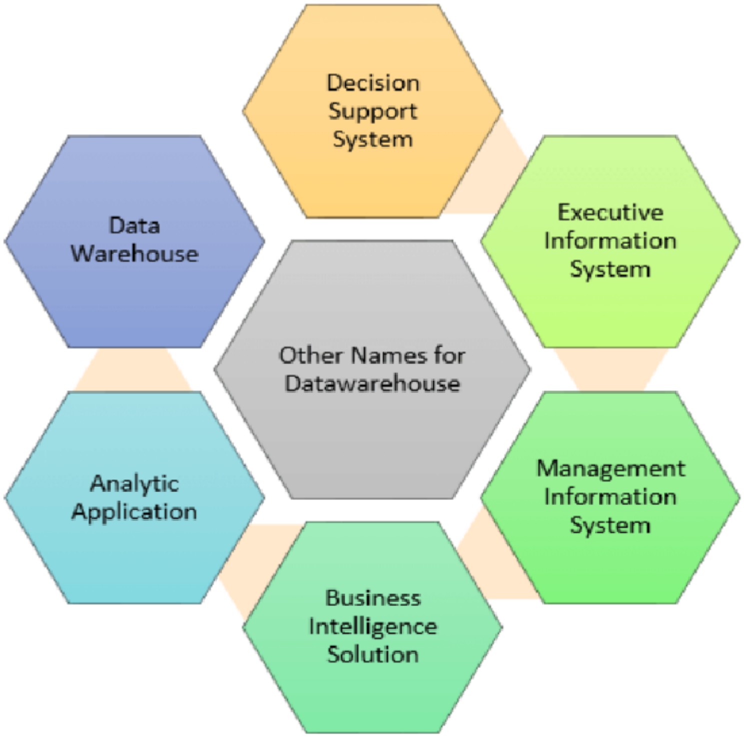
Data analysis and reporting is the main aim of the systems called business informatics systems and these systems, at their core, have the concept of data warehousing. Because, data before the canalization of data and reporting of the insights to the concerned business project, we first need to have the data required and also that data should also be in a very refined form. This refined data can then be utilized very simply by the complex and advanced systems while applying simple as well as complex algorithms on them.

A proper combination or blend of tools and technologies with all the important components help in the proper strategic use of this data. Advanced systems have been developed by companies which work in data driven industries for this very purpose and investment of billions of dollars have been made to make these systems work and help the businesses and these industries grow and generate higher revenues than ever.

After the collection and storage of this data, another main component of these systems is to transform the data. Transformation in its basic sense is conversion or change into a desired format or system. The data, to be of use while stored in these sophisticated systems also needs to be refined and in its most workable form, which means that it should be better ready for the algorithms and operations to be performed on it.

The organization must maintain this decision support data or decision support database separately from the operational database of the organization. The data warehouse, instead of being a product, is rather an environment, which is created in order for the analytics or data based intelligence operations to take place in and through it. The rise of management information systems or business information systems is the indicator of the fact that data and the tools and technologies used in relation to this data are of utmost importance to not only software or IT industry but to all other industries including the global supply chains, advertising and finance especially.

The popular and useful 3NF designed DBs are made of different tables and related to these tables might have many corresponding conditions for data to be accessed and used and hence these might take up a huge amount of time in the decision support systems of these organizations. Therefore, in order to move forward in the data usage and organization industry, we need to move forward to these types of storage and analysis of data. Data warehousing is known in many forms depending on how certain industries and academic groups see them. Following is an illustration.


Decision Support System
Data Warehouse
Executive Information System
Other Names for Datawarehouse
Analytic Application
Management Information System
Business Intelligence Solution

Data warehouses may be primarily classified into three main types -

- Enterprise data warehouse - it basically acts as a decision support system, supported by the data. The reporting and organization of data is combined through this method and industry specific data is labeled as required by the business. So it basically provides better freedom while dealing with data.
- Operational data store - it is usually called as ODS. when the need of the organization is not met by both OLTP and data warehouses, then this approach might be needed. The data is refreshed in real time in this system and thus it is usually preferred by organizations that require storing the data like employee records, which are recorded daily.
- Data mart - these are nothing but a special kind of smaller data warehouses or we could say in a sense, a subset of the complete data warehousing

system. These are usually designed for meeting the needs of the specific division of the organization such as finance, hr, sales or operations.

DATA WAREHOUSE APPLICATIONS IN DIFFERENT SECTOR

There are multiple sectors in which a data warehouse is used, some of them are mentioned below –

- Telecommunications - sales, distribution and advertisements are some of the operations that benefit from data driven decisions in the telecommunication sector

- Hospitality - promotions campaign, advertising and design for better targeting of existing and potential consumers utilizes the data warehouse systems in hospitality.

- Public sector - the present government systems are usually outdated file systems and database systems. Thus they have huge potential of improvement in their management systems and analytics systems for various government systems might be a very good idea.

- Healthcare - sharing and generation of medical reports, different personnel's files and profiles and even imaging and predicting systems are used in the healthcare systems

- Retail chains - transactional data of customers, which might vary from buying patterns, trends, new consumer engagement and retaining the customers with targeted marketing and personal discounts are some of the many fields which utilize the business intelligence systems.

- Investment and finance - it is one of the most data producing, analyzing and consuming sectors which works primarily on data. Decisions worth millions of dollars are made using the insights from share markets, studying profiles of companies and the whole investment system is very cleverly created in order to generate revenues for these organizations with the intelligent use of data.

- Airline - the whole operations of the airline systems have been working digitally from the longest time and this is one industry that smartly used data and positioning systems to create better consumer engagement, even in the Covid-era.

CHAPTER - 05

BIG DATA

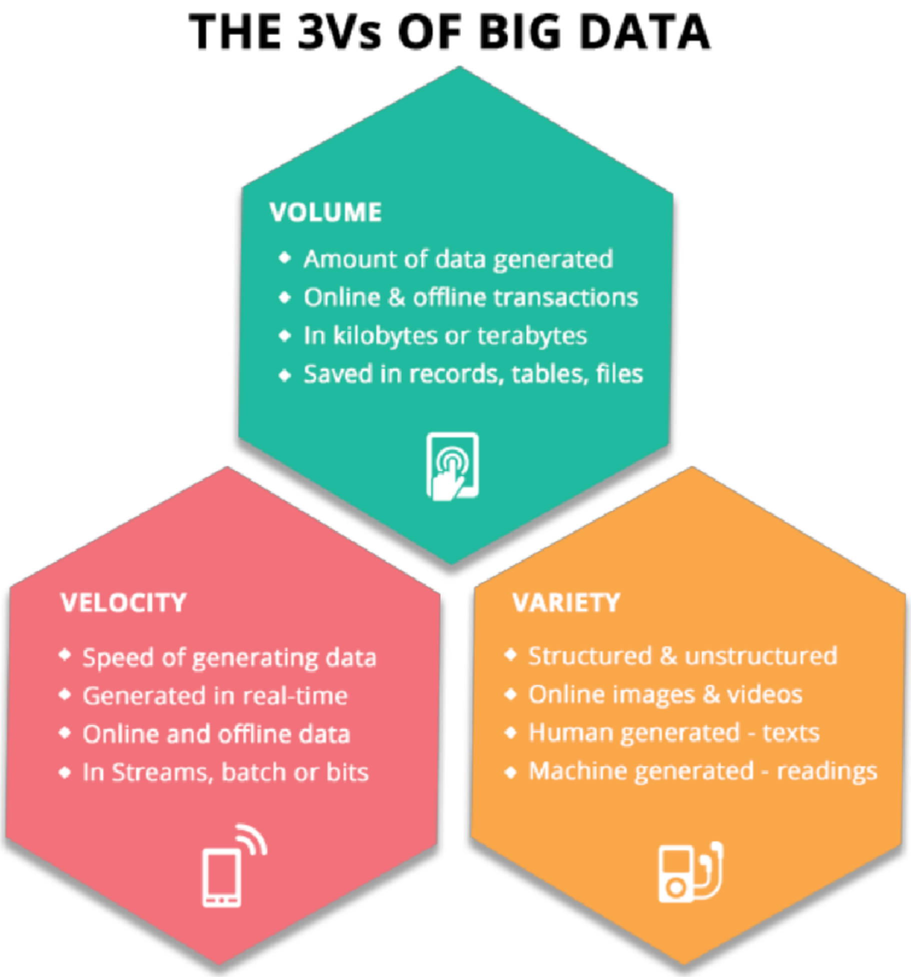
Big data is a huge collection of data. It is similar to data but very large in volume so that it cannot be managed by any traditional data management systems. The big data is primarily characterized by three parameters, the 3 Vs –

- Volume – the amount of data which is being generated is called its volume. The first main characteristic of the big data is its high volume. The data which is being generated on the lines of terabytes and petabytes regularly comes under big data.

- Velocity – the data is being generated everyday or at regular intervals. The speed at which the data is being generated is referred to as its velocity. The data explosion that is happening in the current scenario is only manageable by the use of big data and traditional methods of data storage and management are not capable of holding this data.

- Variety – the data which is being generated right now is of various different qualities and quantities as well as of various different types. The data now is not limited to records that has usual text information but even audio, video and other kinds of data is being managed and this kind of variety is acceptable and required for the data to be considered as big data.

The data which fulfills the above three conditions or is actually characterized by the above three parameters is referred to as big data.



THE 3Vs OF BIG DATA

VOLUME

- Amount of data generated
- Online & offline transactions
- In kilobytes or terabytes
- Saved in records, tables, files

VELOCITY

- Speed of generating data
- Generated in real-time
- Online and offline data
- In Streams, batch or bits

VARIETY

- Structured & unstructured
- Online images & videos
- Human generated - texts
- Machine generated - readings

www.whishworks.com

Types of big data

- Structured – when the format of the data is fixed and the data can be processed and stored in this same format, then it is referred to it as structured data. When the format of the data is known in advance, the advancement in the technology in IT allows us to handle the same with ease but the problem in dealing with this structured data arises when the size of the data that is being used is very high such that it is in multiple zettabytes.

- Unstructured – when the structure or form of the data is not known, it is classified as unstructured data. Such kind of data might contain of the data in a combined format of text, images and videos and hence its becomes very tough to store, process and derive some useful insights from such data, which in addition to this problem also has huge quantities.

- Semi- structured – the kind of data which can contain both of the above types of data in a combined as well as individual form. Semi- structured data can be in both the structured but not defined like in a database management system. An example of the semi-structured data is XML file.

Real life examples of big data

- Finance and investment – this field thrives on data and thus is bound to generate magnitudes of data. For an instance, the stock exchange at New York City only generates more than one terabytes of data per day.

- Social media – with millions and billions of users on the social media platforms right now, each of their logins, activity, likes and comments, posts, shared content are recorded and that by no surprise generates more than 500 terabytes of data on a daily basis.

- Jet engines – in only as short as 45 minutes of flight, 15+ terabytes of data can be generated by the jet engine.

HADOOP

Hadoop is an open source system, by apache, which is created to handle the storage and processing of big data with the help of simple programming model. This is done on a cluster of computers. It was written in java. Hadoop basically provides an environment for distributed storage and computation and thus the process effort is divided into multiple pieces for each of the system to store and process and hence the effort required by each system goes down multiple times in terms of both complexity and time.

Hadoop basic architecture –

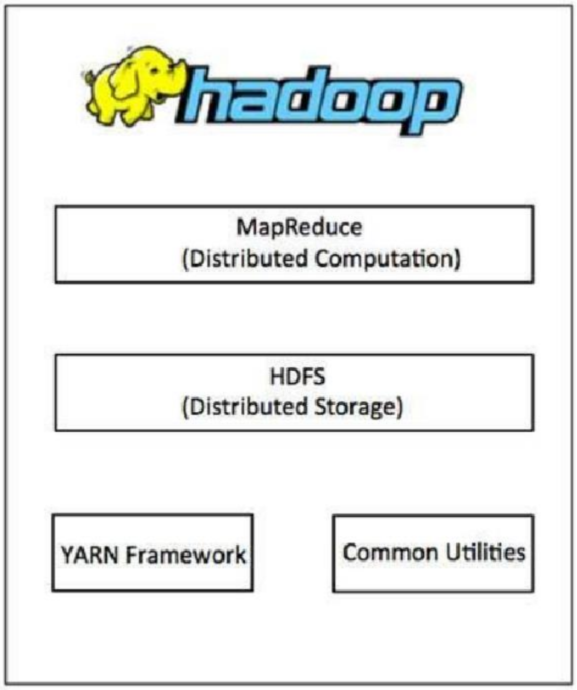
Hadoop is made up of two basic layers when it comes to the architecture –

- Processing layer called Map reduce
  It was developed by Google to support multiple programming paradigms which uses the parallel programming concepts. It uses a distributed system in order to do the processing part and makes it efficient to process large amounts of data.

- Storage layer as Hadoop distributed file system
  Based on Google file system, hadoop file system is used to avail a distributed system of storage of the big data. One of the most important benefits of this file system is its fault-tolerance and its usage of low cost hardware, which basically use commodity hardware.

  Additional modules in the hadoop framework are also present –

- Hadoop common
- Hadoop Yarn



REFERENCES –

- Molinaro, Anthony (2005), SQL Cookbook
- Beighley, Lynn (2007), Headfirst SQL
- Johnston, Benjamin (2019), SQL for data analytics
- Bloch, Joshua (2001), Effective java
- Bates, Bert & sierra, Kathy (2003), Head first java
- Ascher, David (1999), Learning Python
- Barry, Paul (2010), Head first python
- Simon, Phil (2013), Too big to ignore : Data, Analytics
- Kitchin, Rob (2014), The Data Revolution
- Davenport, T.H (2016), Big data at work

BIG_DATA_TRAINING-_Project_Report.docx

ORIGINALITY REPORT

13%	10%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Jaypee University of Information Technology Student Paper	3%
2	en.wikipedia.org Internet Source	1%
3	rolandblogsite.wordpress.com Internet Source	1%
4	www.tutorialsteacher.com Internet Source	1%
5	www.interviewbit.com Internet Source	1%
6	Submitted to International School of Management and Technology Student Paper	<1%
7	Submitted to The University of Manchester Student Paper	<1%
8	www.labautopedia.org Internet Source	<1%
9	develbyte.in	

	Internet Source	<1%
10	Submitted to University of Wolverhampton Student Paper	<1%
11	www.javatpoint.com Internet Source	<1%
12	www.knowledgehut.com Internet Source	<1%
13	dba-oracle.com Internet Source	<1%
14	Submitted to City of Westminster College, London Student Paper	<1%
15	Submitted to University of Derby Student Paper	<1%
16	cpplanguageprograms.blogspot.com Internet Source	<1%
17	Submitted to Westcliff University Student Paper	<1%
18	Submitted to Dupont Manual High School Student Paper	<1%
19	Submitted to Otago Polytechnic Student Paper	<1%
20	www.tutorialspoint.com Internet Source	

		<1%
21	Submitted to Australian Catholic University Student Paper	<1%
22	Submitted to KDU College Sdn Bhd Student Paper	<1%
23	Submitted to Roehampton University Student Paper	<1%
24	heartofcodes.com Internet Source	<1%
25	Submitted to Canterbury College, Kent Student Paper	<1%
26	Submitted to Middle East College of Information Technology Student Paper	<1%
27	Submitted to Modern College of Business and Science Student Paper	<1%
28	Submitted to RDI Distance Learning Student Paper	<1%
29	Submitted to St. Petersburg College Student Paper	<1%
30	www.allaboutweb.biz Internet Source	<1%

31	www.coursehero.com Internet Source	<1%
32	Submitted to Coventry University Student Paper	<1%
33	Sagar Ajay Rahalkar. "Chapter 2 Database Basics", Springer Science and Business Media LLC, 2016 Publication	<1%
34	Submitted to Emirates Aviation College, Aerospace & Academic Studies Student Paper	<1%
35	clean-clouds.com Internet Source	<1%
36	docplayer.net Internet Source	<1%
37	documents.mx Internet Source	<1%
38	www.btechsmartclass.com Internet Source	<1%
39	www.learnpick.in Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 16/06/2021

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | [✓ B.Tech Project Report] | Paper

Name: Aditya Agarwal Department: Computer science and engineering Enrolment No 171479

Contact No. 8384025332 E-mail. Deaditya.1999@gmail.com

Name of the Supervisor: DR. Monika Jindal

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): BIG DATA TRAINING – PROJECT REPORT

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages = 54
- Total No. of Preliminary pages = 53
- Total No. of pages accommodate bibliography/references = 1

d Aditya

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 13 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Monika Bharti

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	**Excluded**	**Similarity Index (%)**	**Generated Plagiarism Report Details** (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature **Librarian**

……