TRAINING ON BIG DATA

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

MAY 24, 2021

Submitted By:

DIWAKAR PALIWAL (171814)

PROJECT WORK COMPLETED UNDER SUPERVISED GUIDANCE AT

COGNIZANT TECHNOLOGY SOLUTIONS PVT. LTD.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS
WAKNAGHAT, H.P - 173234

TABLE OF CONTENTS

TITLE	PAGE-NO.
LIST OF FIGURES	1
PROJECT REPORT UNDERTAKING	2
DECLARATION BY THE STUDENT	3
COMPANY PROFILE	4
CHAPTER-1	5
CHAPTER-2	10
CHAPTER-3	26
CHAPTER-4	30
CHAPTER-5	35
CHAPTER-6	38
CONCLUSION	42
REFERENCES	43

LIST OF IMAGES

- Fig 1: The characteristics of Big Data
- Fig 2: An example of Structured Data
- Fig 3: An example of different formats of Unstructured Data
- Fig 4: An example of Semi-structured Data
- Fig 5: Predicted growth of data
- Fig 6: An example of a database with 3 tables to store data
- Fig 7: General Syntax of SQL Queries
- Fig 8: Data types in SQL
- Fig 9: Basic SQL commands
- Fig 10: Data Manipulation Language Queries
- Fig 11: Types of Join in SQL
- Fig 12: Aggregate commands in SQL
- Fig 13: Components of JAVA JDK
- Fig 14: Components of JAVA JVM
- Fig 15: Sample Java Program
- Fig 16: Data types in JAVA
- Fig 17: Primitive data types
- Fig 18: Example of else condition in action
- Fig 19: Example of else-if condition in action
- Fig 20: An example of Switch case
- Fig 21: Architecture of Kernel
- Fig 22: UNIX Shell
- Fig 23: Example of a general command to list files in UNIX
- Fig 24: Tree structure of File Storage
- Fig 25: Features of Python
- Fig 26: Data types in Python
- Fig 27: Various names of data warehouse

PROJECT REPORT UNDERTAKING

I Mr/Mrs DIWAKAR PALIWAL Roll-no.171814 Branch BIOTECHNOLOGY is doing my internship with COGNIZANT from 01-03-2021 to 22-06-2021.

As per procedure I have to submit my project report to the university related to my work that I have done during this internship.

I have compiled my project report. But due to COVID-19 situation my project mentor in the company is not able to sign my project report.

So, I hereby declare that the project report is fully designed/developed by me and no part of the work is borrowed or purchased from any agency. And I'll produce a certificate/document of my internship completion with the company to TnP Cell whenever COVID-19 situation gets normal.

Signature: Dpaliwal

Name: DIWAKAR PALIWAL

Roll No. 171814

Date: 24-05-2021

DECLARATION

I therefore announce that my report titled "Training on Big Data" is submitted as Project Work has been done by me at **"Cognizant Technology Solutions Pvt. Ltd."** under guided supervision. Any further augmentation, continuation or utilisation of this must be embraced with earlier express composed assent from the organization.

I further proclaim that the preparation work or any part thereof has not been recently submitted for any degree or certificate in any college.

Dpaliwal

Signature of Trainee: ……………
Name: Diwakar Paliwal
Date: 24-05-2021

Supervisor Name: Dr. Udayabanu M.
Associate Professor,
Department of Biotechnology and
Bio-informatics,
Jaypee University of Information
Technology,
Solan, H.P -173234.
Signature: ....

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 inhouse unit for providing IT-infrastructure related services for Dun & Bradstreet business, but later expanded its client base from 1996. It is headquartered in Teaneck, New Jersey, United States.

Cognizant's digital business, operations and systems and technology are the three areas which make up their business profile. To provide technological proficiency to it's 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 corporations 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-1

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 that traditional data-handling solutions find it difficult to store and process this data.

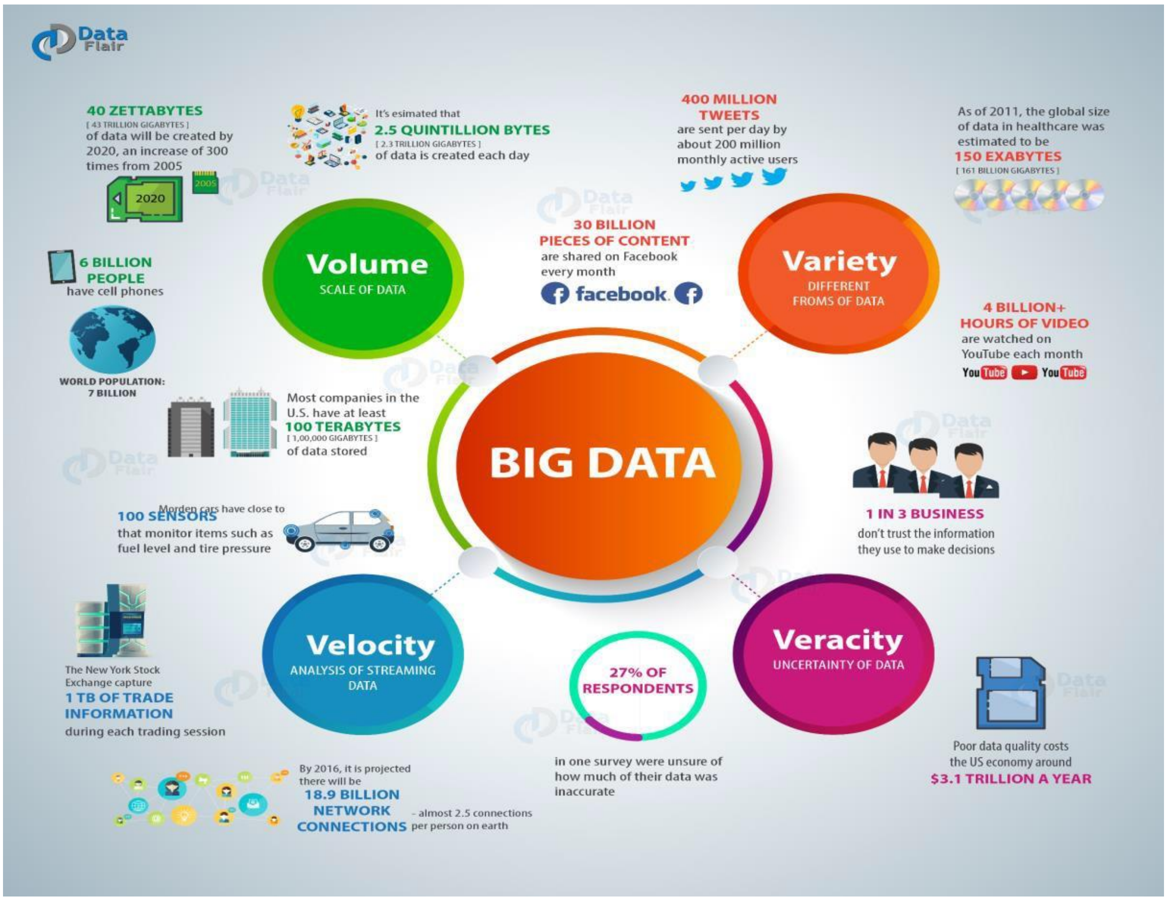
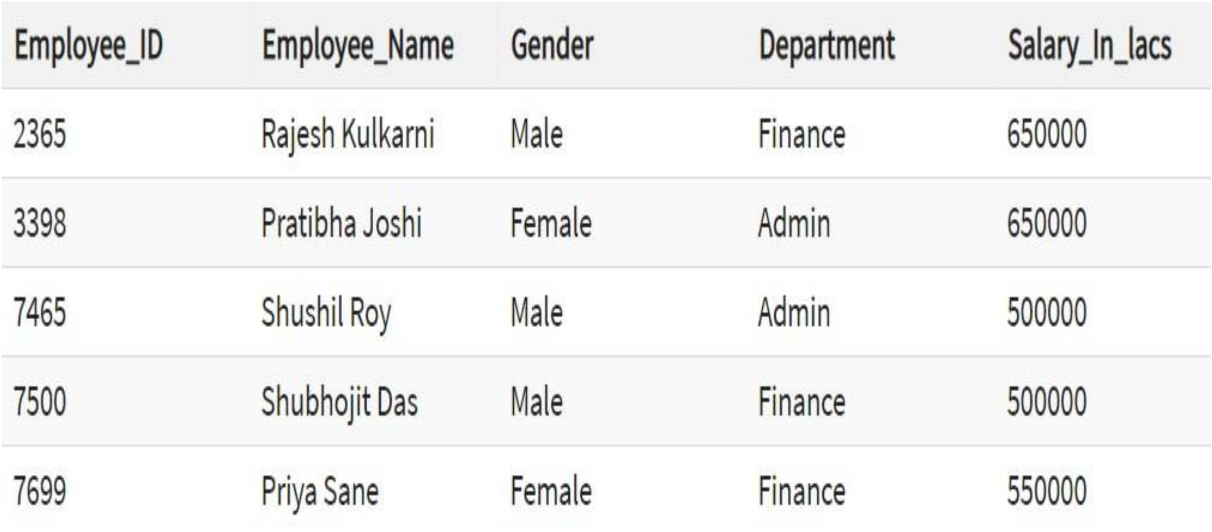

Data Flair
40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's esimated that 2.5 QUINTILLION BYTES [2.3 TRILLION GIGABYTES] of data is created each day
400 MILLION TWEETS are sent per day by about 200 million monthly active users
As of 2011, the global size of data in healthcare was estimated to be 150 EXABYTES [161 BILLION GIGABYTES]
6 BILLION PEOPLE have cell phones
WORLD POPULATION: 7 BILLION
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 YouTube
Most companies in the U.S. have at least 100 TERABYTES [1,00,000 GIGABYTES] of data stored
BIG DATA
1 IN 3 BUSINESS don't trust the information they use to make decisions
Morden cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure
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

Fig 1: The characteristics of Big Data.

TYPES OF BIG DATA:

STRUCTURED:

This type of data can be stored, assessed or processed as fixed format, hence the name Structured data. The progress 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

Fig 2: An example of Structured Data.

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 heterogenous source, the data files may include anything ranging from simple text files to mp3, mp4, jpg, flv, 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.

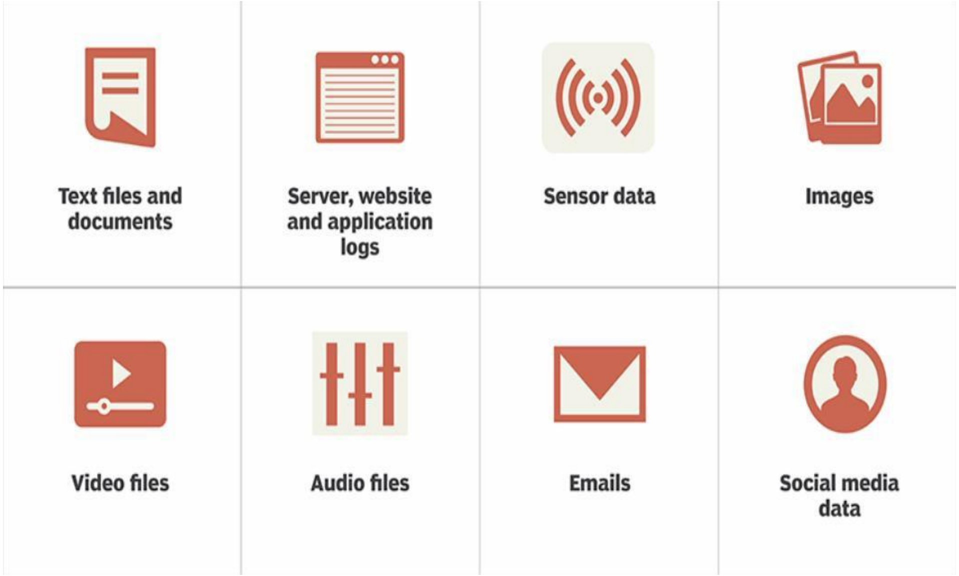
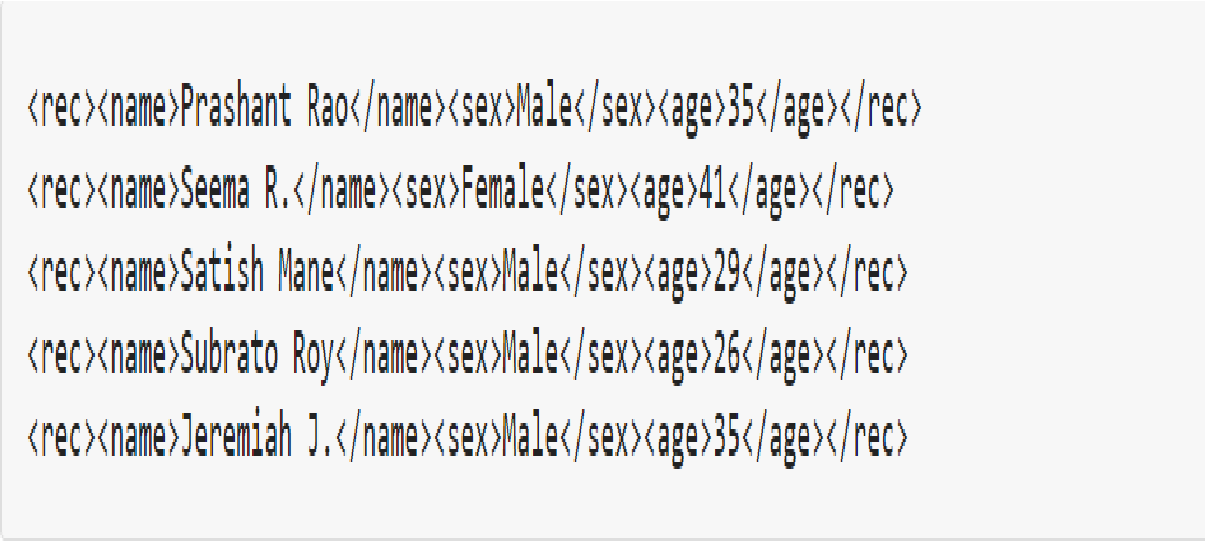


Fig 3: An example of different formats of Unstructured Data.

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>
```

Fig 4: An example of Semi-structured Data.

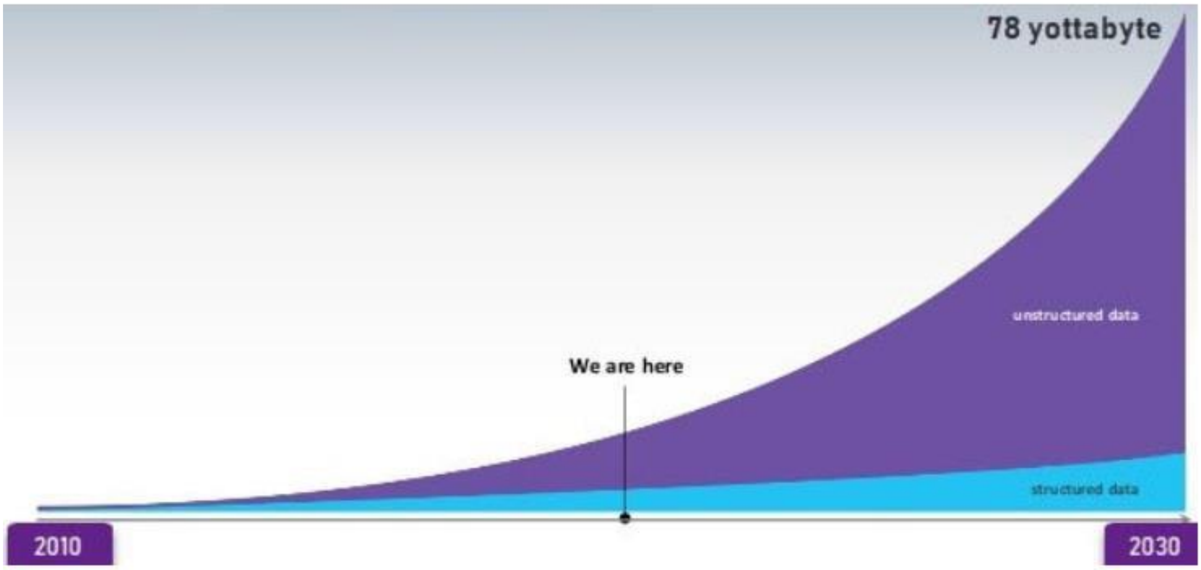


Fig 5: Predicted growth of data.

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.

CHAPTER-2

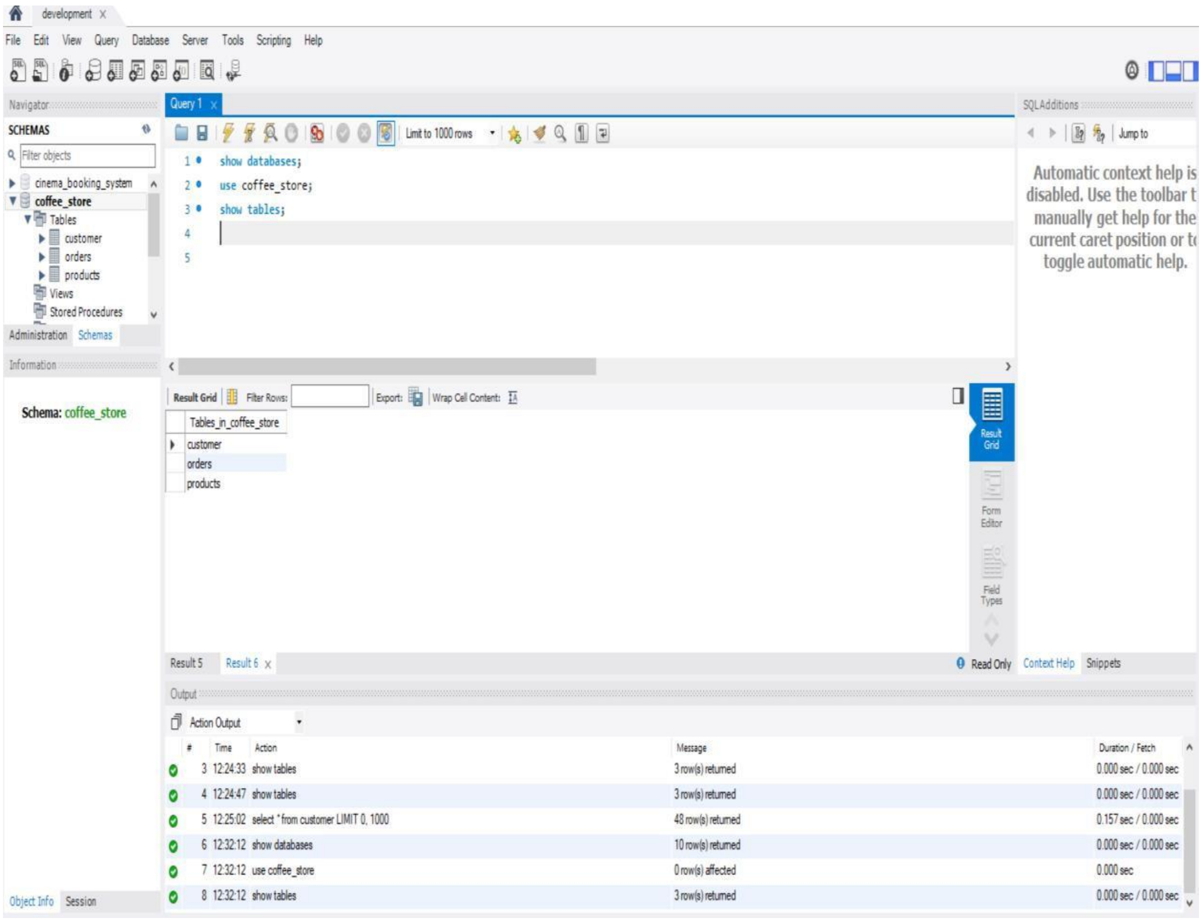
SQL AND JAVA

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[1].

The database management system is a 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.

Schema: coffee_store

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
```

Fig 6: An example of a database with 3 tables to store data.

WHAT IS SQL?

SQL stands for Structured Query Language. It is designed for use in performing management operations in relational databases. SQL is based on relational algebra and tuple relational calculus.

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[1].

In the late 1970s, relational software, 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[2]. 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[3]-

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 to generate some scalar values or even might return columns or rows or both combined, which are basically called tables.

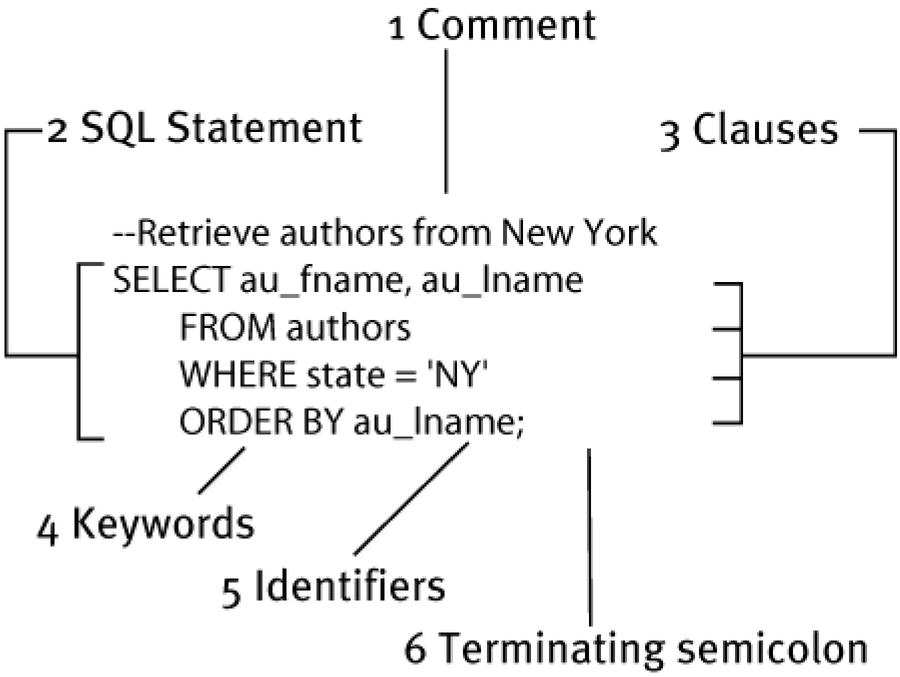


Fig 7: General Syntax of SQL Queries.

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:

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

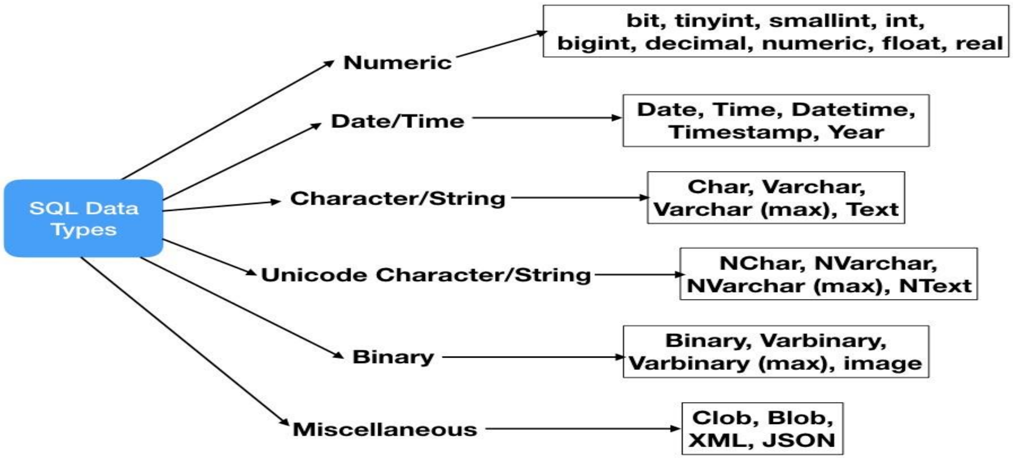

SQL Data Types
Numeric
bit, tinyint, smallint, int, bigint, decimal, numeric, float, real
Date/Time
Date, Time, Datetime, Timestamp, Year
Character/String
Char, Varchar, Varchar (max), Text
Unicode Character/String
NChar, NVarchar, NVarchar (max), NText
Binary
Binary, Varbinary, Varbinary (max), image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

Fig 8: Data types in SQL.

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>;

- **DROP-** To delete an existing database in the SQL schema
    - Drop database <database name>;

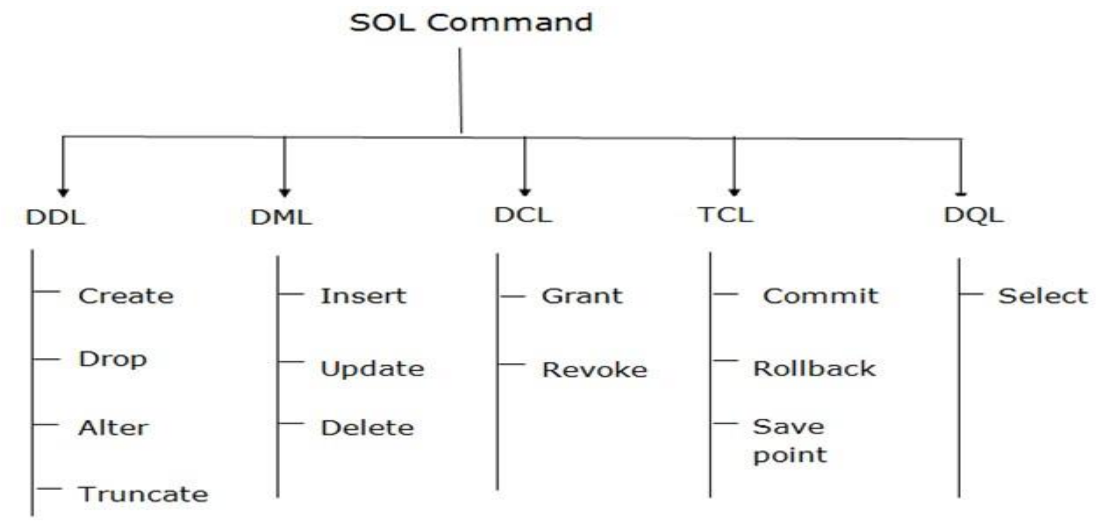
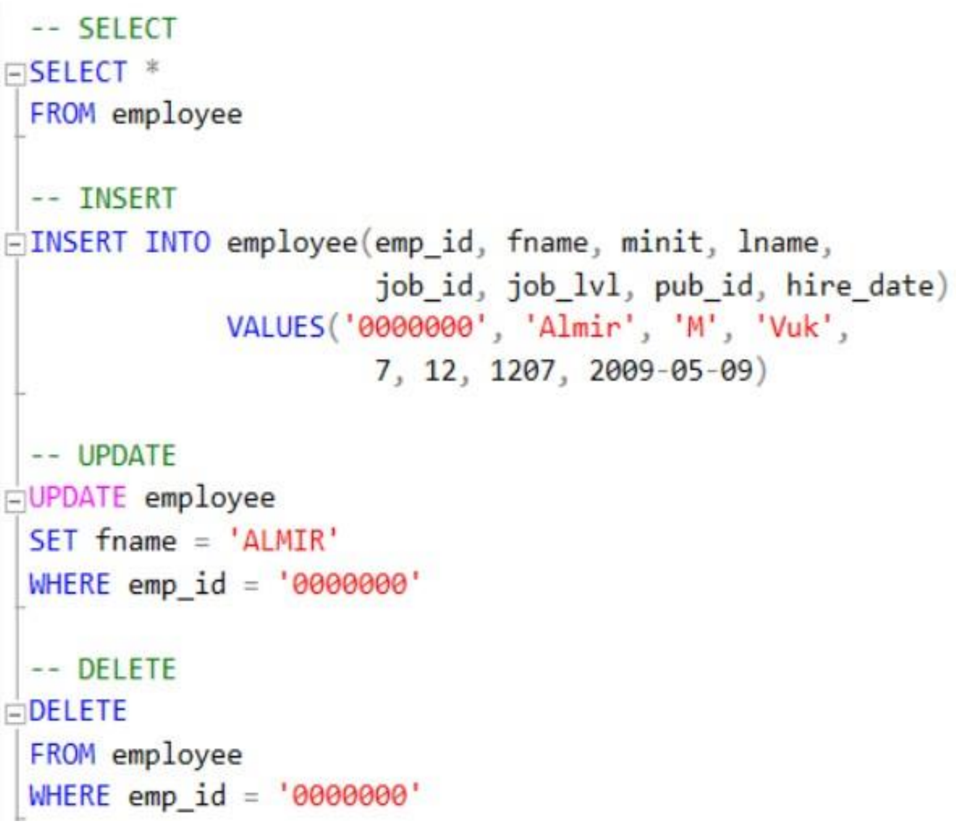


Fig 9: Basic SQL commands.

DML COMMANDS

The manipulation of data 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 Following are some examples of DML commands -



```
-- SELECT
SELECT *
FROM employee

-- INSERT
INSERT INTO employee(emp_id, fname, minit, lname,
                     job_id, job_lvl, pub_id, hire_date)
            VALUES('0000000', 'Almir', 'M', 'Vuk',
                     7, 12, 1207, 2009-05-09)

-- UPDATE
UPDATE employee
SET fname = 'ALMIR'
WHERE emp_id = '0000000'

-- DELETE
DELETE
FROM employee
WHERE emp_id = '0000000'
```

Fig 10: Data Manipulation Language Queries.

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 and 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 a 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.

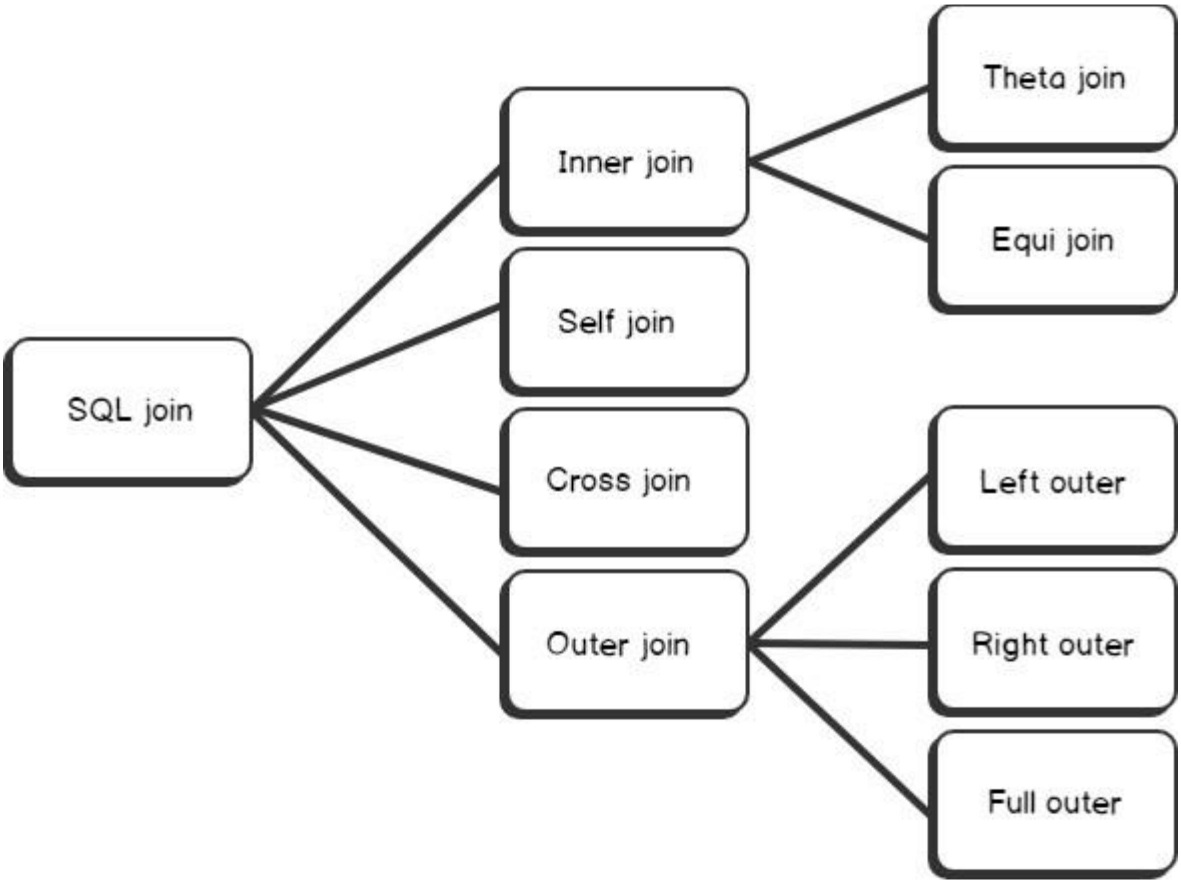


Fig 11: Types of Join in SQL.

- **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 subqueries, 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 subqueries- Here 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

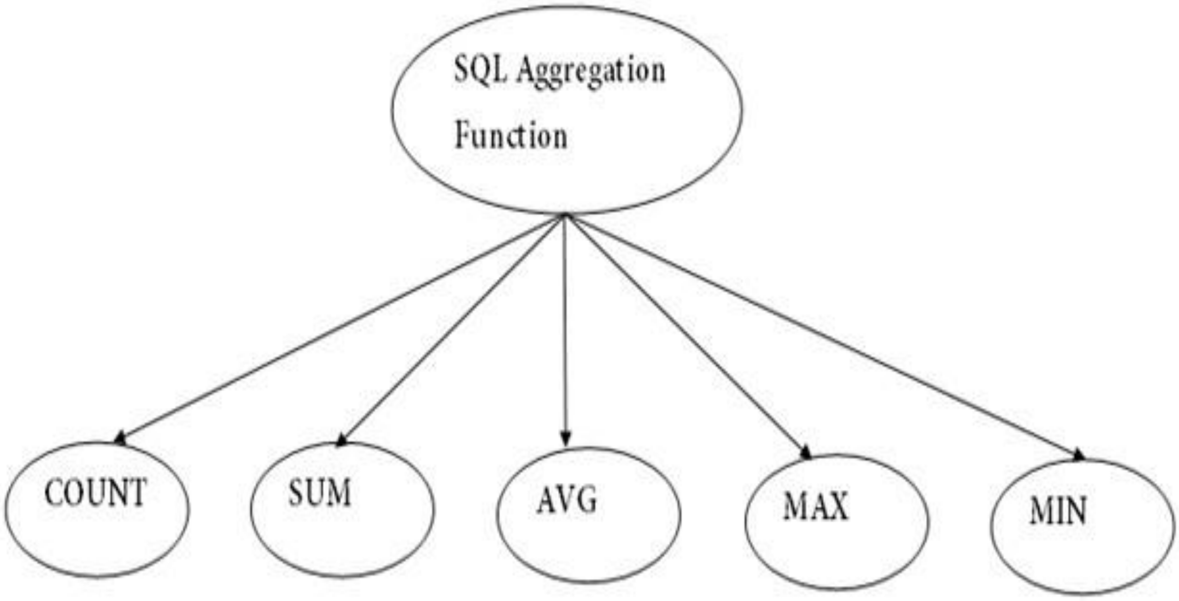


Fig 12: Aggregate commands in SQL.

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[4]. 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 OF JAVA

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 utilised 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 disassembler
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

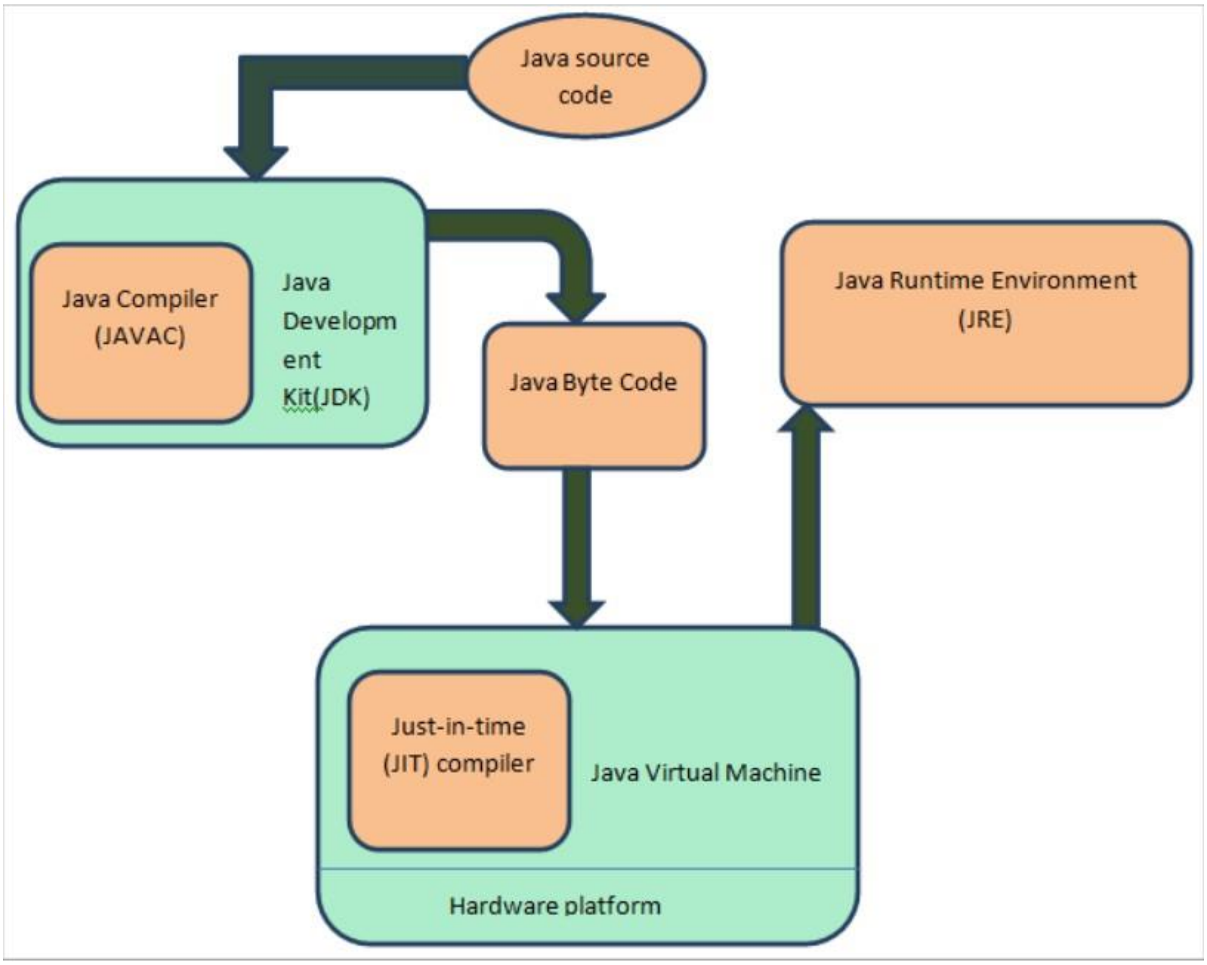


Fig 13: Components of JAVA JDK.

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 code anywhere and execute 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 bytecode, which has similar functionality at the same time. Hence, Java is both a compiled and interpreted language.

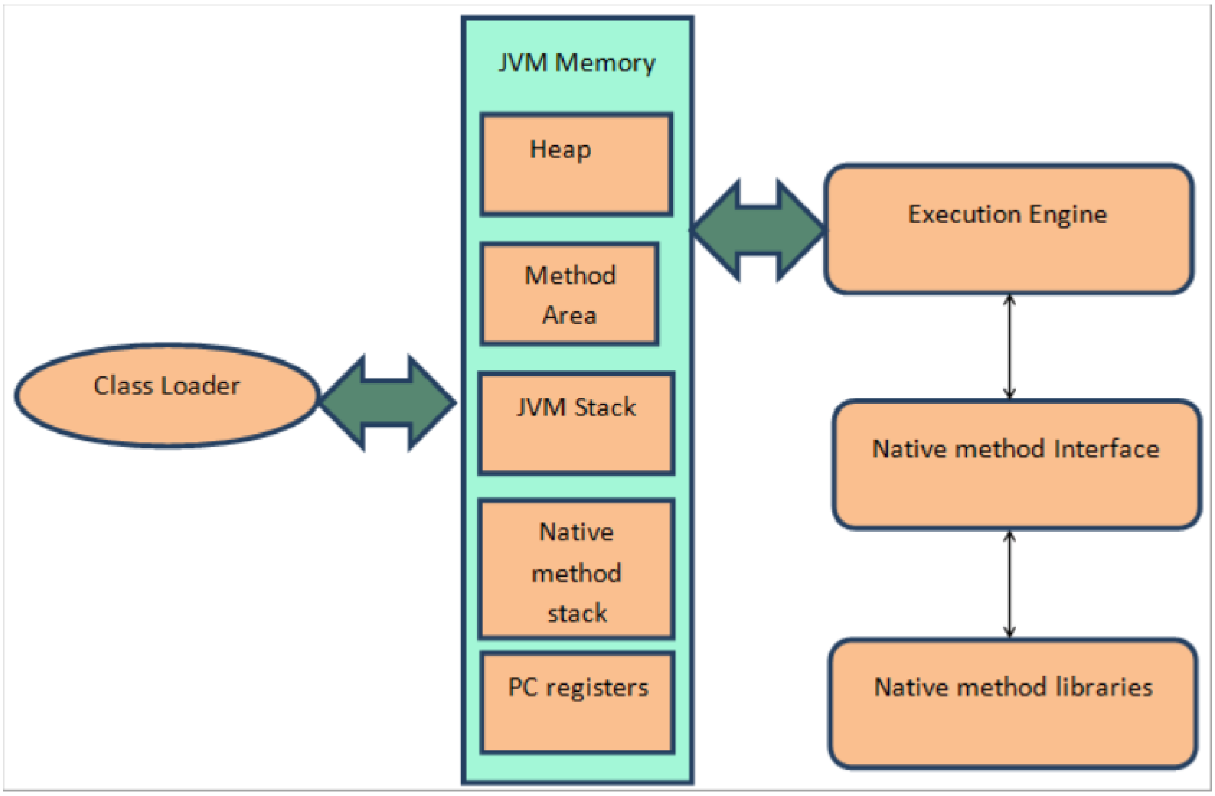


Fig 14: Components of JAVA JVM.

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- 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 Applications- 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 OOPs like regex, multi-threading etc.

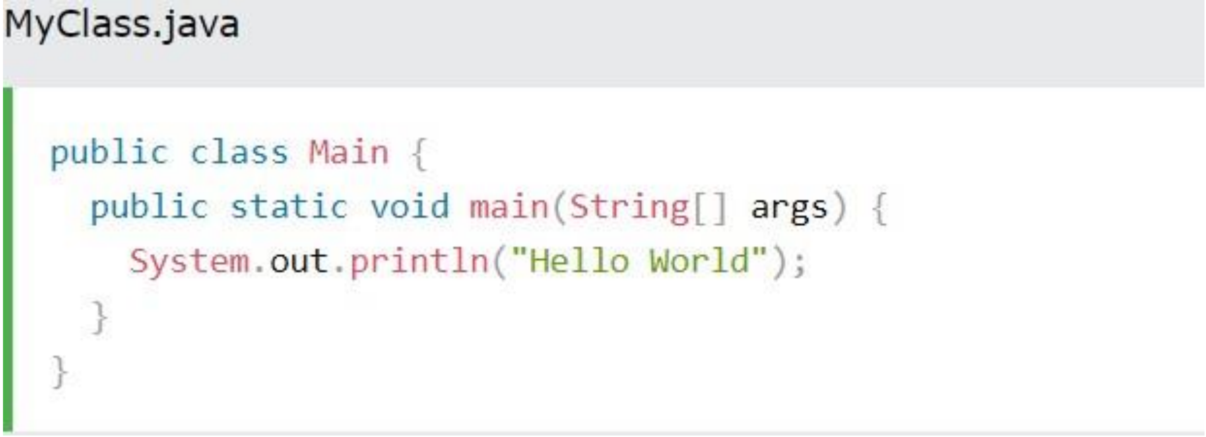
Java EE- The Enterprise Edition is used to develop applications for enterprises and web applications. This is based over the Standard Edition.

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[5]. 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");
  }
}
```

Fig 15: Sample Java Program.

JAVA DATA TYPES

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

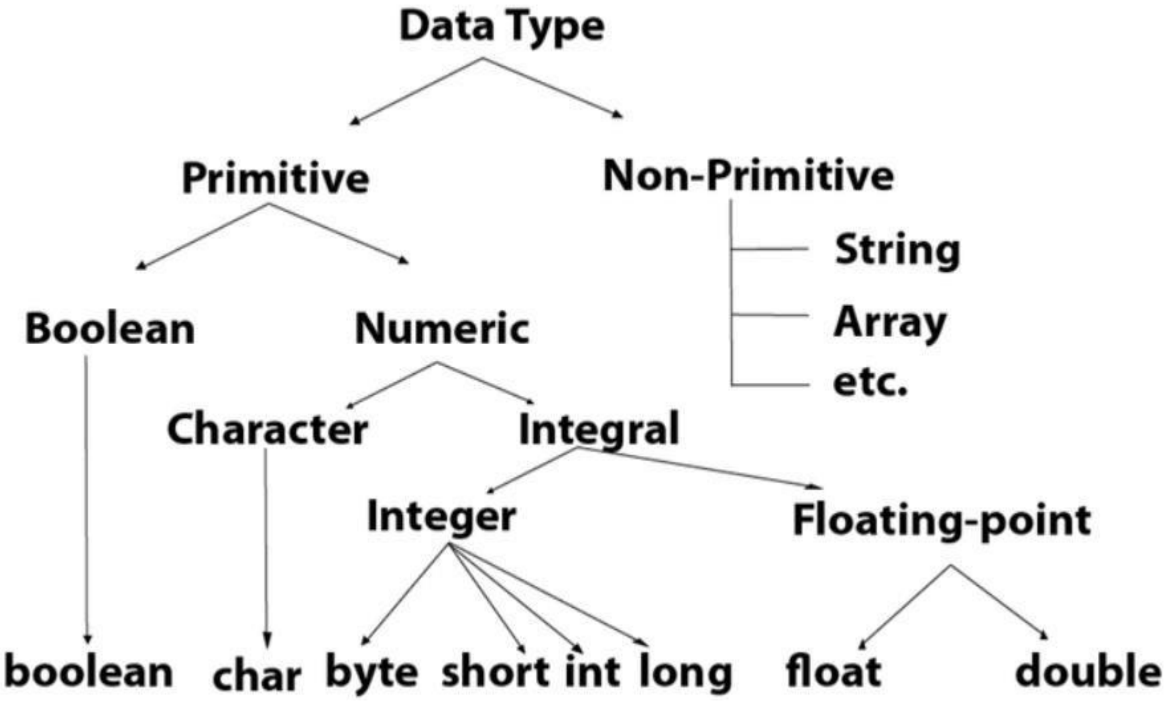
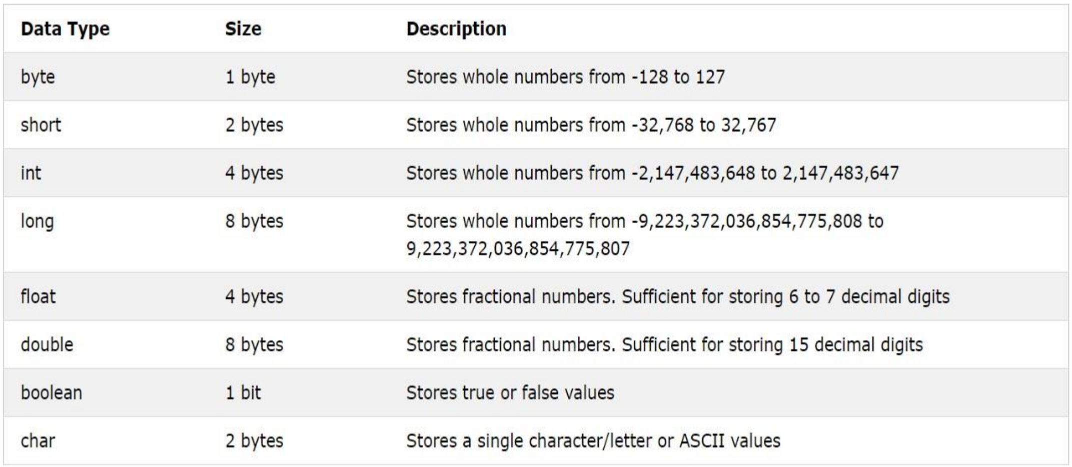


Fig 16: Data types in JAVA.

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

Fig 17: Primitive data types.

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

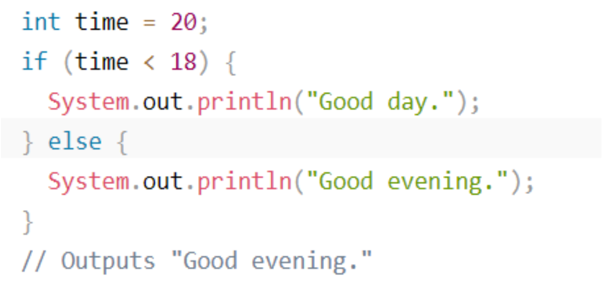
Arrays- They are utilised to store multiple values inside a single variable, instead of the need to declare multiple variables.

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.

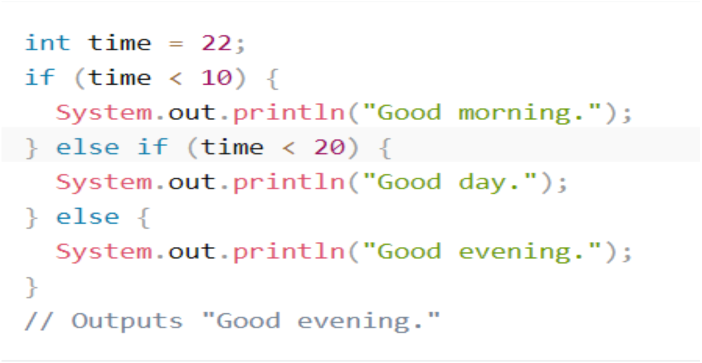
else- Specifies a code block to be executed, if the same condition evaluates to false.



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

Fig 18: Example of else condition in action.

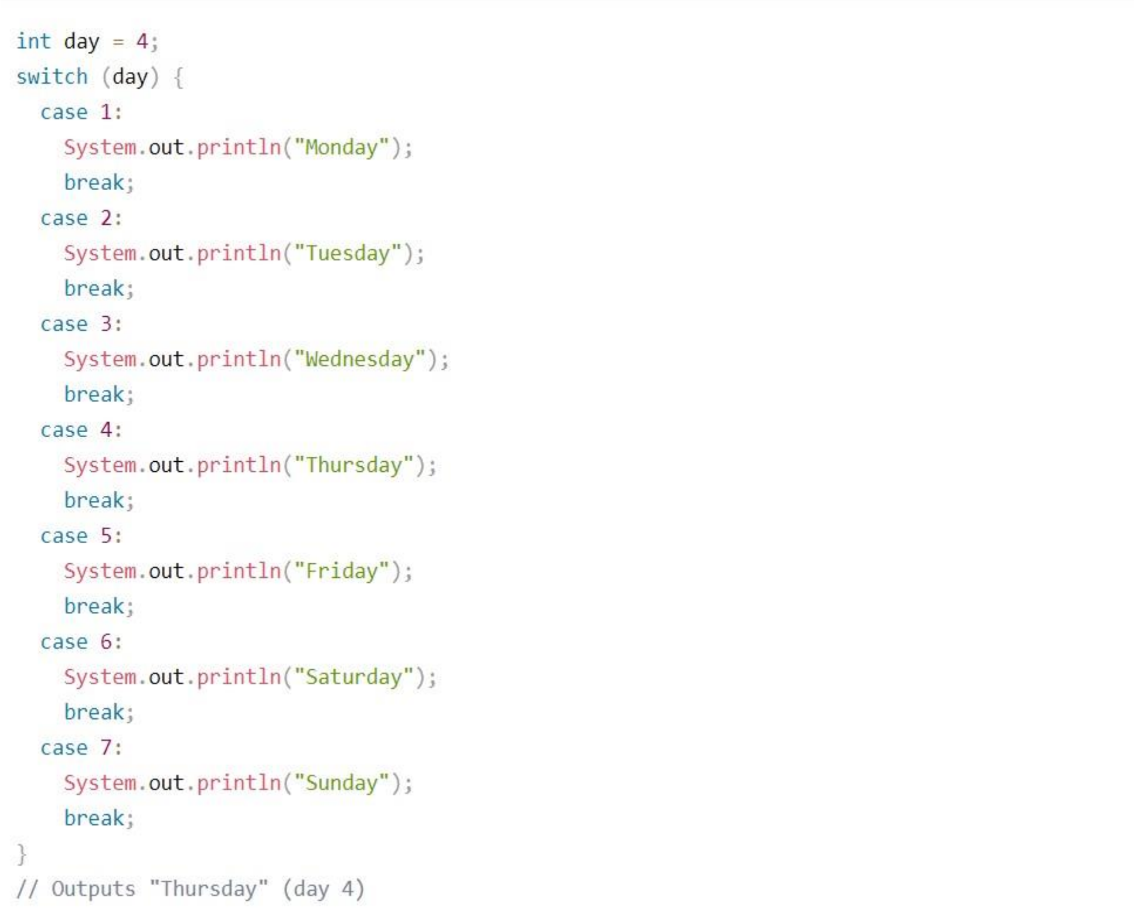
else if- Specifies a new code block to be executed if the first condition evaluates to false.



```
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."
```

Fig 19: Example of else condition in action.

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)
```

Fig 20: An example of Switch case.

CHAPTER-3
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 kickstarted in 1970’s at the AT&T’s Bell Labs research centre, 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/flavours 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.

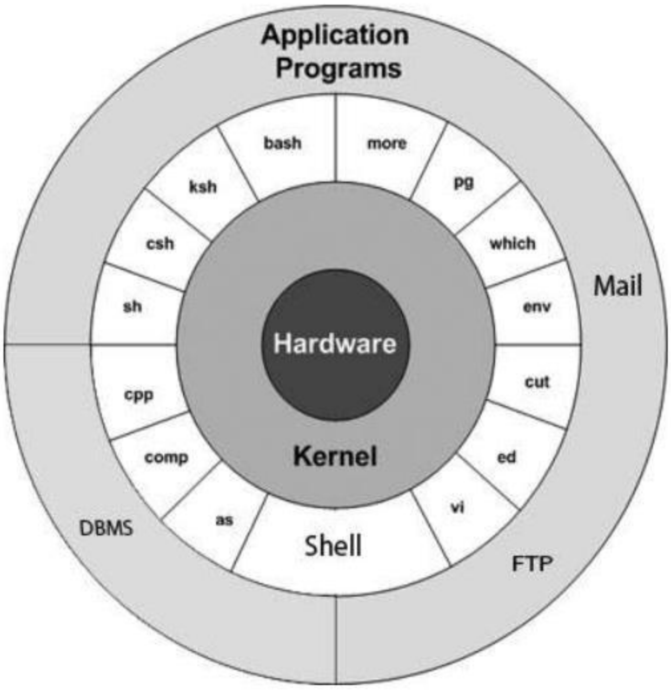
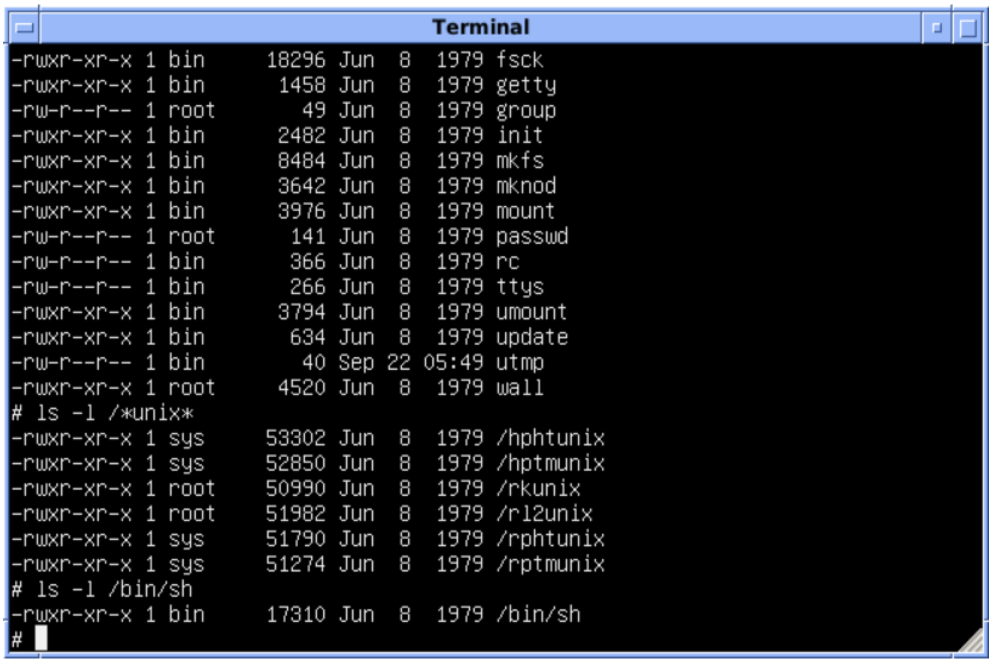


Fig 21: Architecture of Kernel.

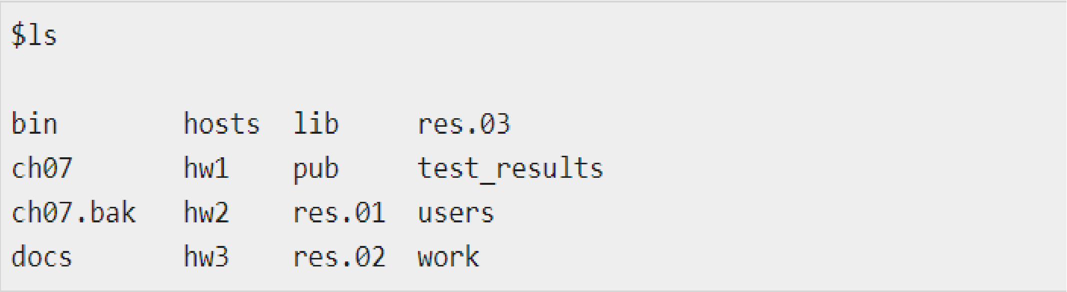
2.) Shell- A shell is a utility which processes the commands given as input in the terminal, processes them and then calls the required program to execute the task. The shell follows similar syntax for all the commands. Some types of shell available in UNIX are C shell, Korn shell, Bourne shell etc.



```
-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
# 
```

Fig 22: UNIX Shell.

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
```

Fig 23: Example of a general command to list files in UNIX.

4.) Files/Directories- Unix follows a tree-like structure for the organisation 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.

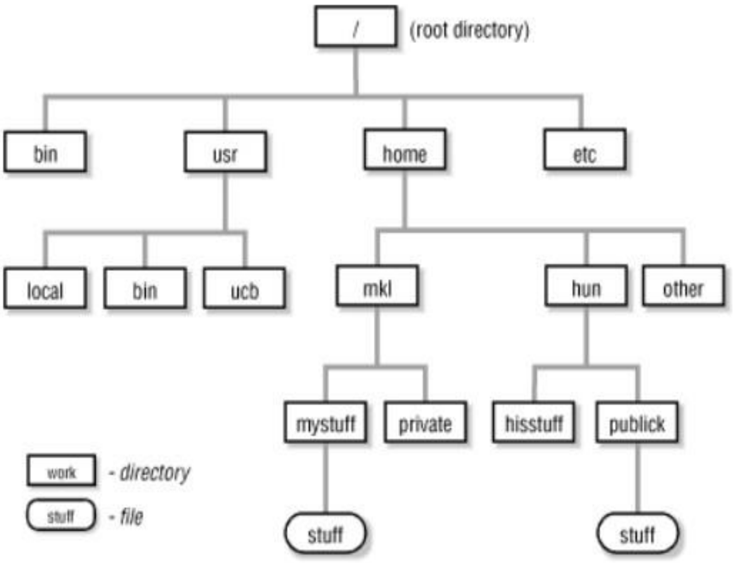


Fig 24: Tree structure of File Storage.

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 utilised for high end servers dedicated for storage and data processing. It also popularised 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.

CHAPTER-4

PYTHON

It was developed by Guido van Rossum in 1990's. It is a general-purpose high-level language. 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. Its design philosophy favours code reusability. It was a successor to ABC programming language, and its first iteration was released in 1991[6].

IDLE – Integrated Development Environment tool allows us to write and run our code easily with a simple interface.

CHARACTERISTICS OF IDLE

1) Written in python
2) Uses tkinter graphics library
3) Has an interactive python shell
4) A full featured text editor
5) A debugger

FEATURES OF PYTHON

1) **High level programming language–** Python is a high-level programming language and exhibits the features of a high-level language like code readability[6] and easy usage.

2) **Open source–** Python is free to use and can be used for personal and professional work free of cost

3) **Supports multiple programming paradigm–** It supports object-oriented programming, imperative, procedural and functional programming

4) **Extensible–** Python can easily be used combined with different languages and frameworks with simple extensions and commands

5) **GUI–** Creating buttons, text boxes, widgets is easy and achievable in python with its different technologies and tools

6) **Embeddable–** Python language is embeddable and can be used for embedded programming of certain devices, advanced technologies using its imperative programming paradigm

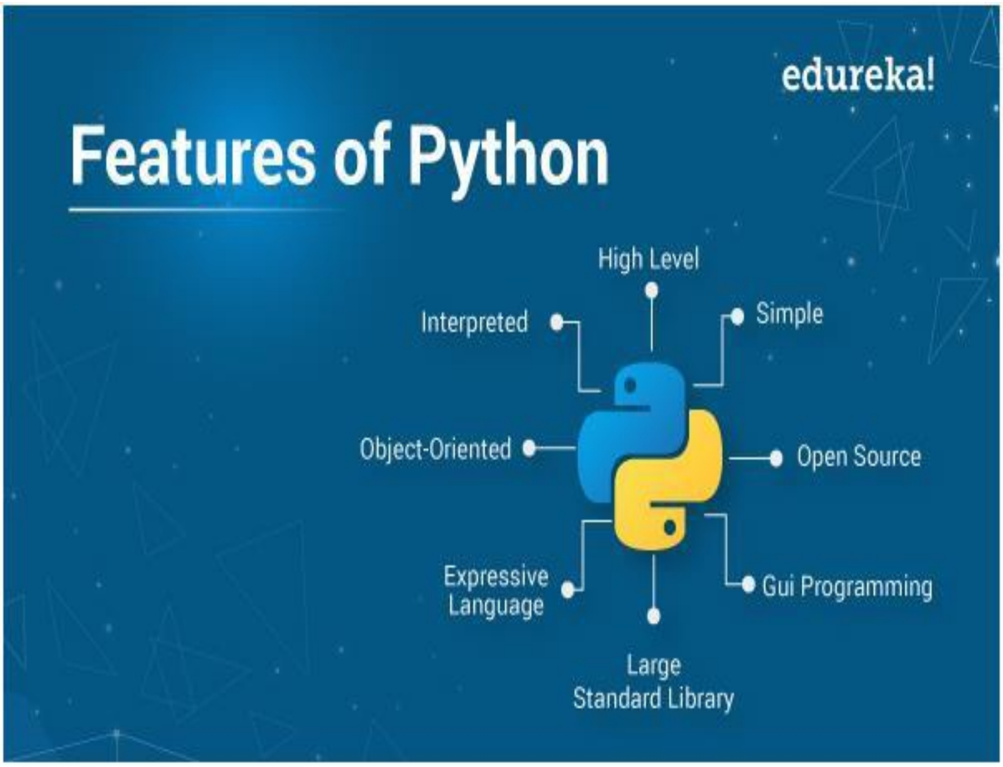


Fig 25: Features of Python.

PYTHON 2.XX vs PYTHON 3.XX- Python 2.0 came out in 2000 while the python 3.0 came out in 2008[7].

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

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

3) Stores strings as Unicode by default

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

5) 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.

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.

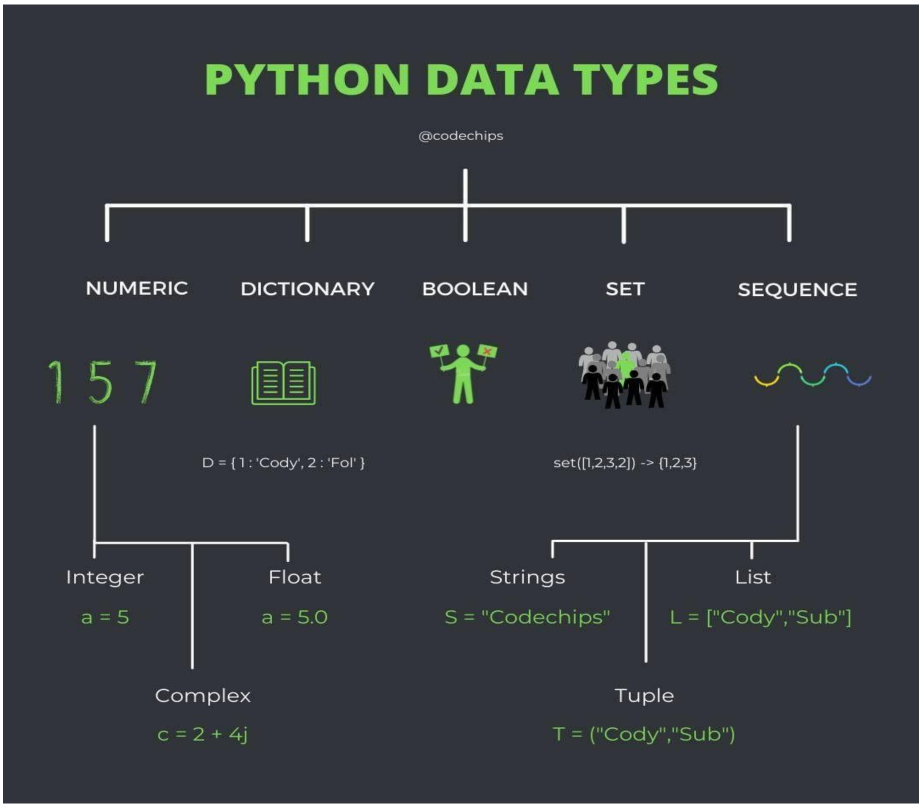


Fig 26: Data types in Python.

The four broad classifications of data are –

1) **Numeric-** Any representation of data which has a numeric value. They are of three types– **Integer** (2, -5), **Float** (1.3E, -2.8) and **Complex** (2+3i).

2) **Boolean-** Any representation of data, which has two values denoted by either true or false.

3) **Sequence-** An ordered collection of similar/different data types. Some of the built-in sequence data types are– **String**- a combination of characters e.g. ('hello'), **List**– an ordered collection of one or more data items, not necessarily of same type in square brackets represent a list e.g. [1, 'ram', 2.4, True], **Tuple**– an ordered collection of one

or more data items, not necessarily of the same type, put in parenthesis. Contents of a tuple cannot be modified e.g. (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 e.g. {1:” Adil”, 2:” Akash”, 3:” Ajay”}.

ARITHMETIC OPERATORS

It supports all mathematical operators for addition (+), subtraction (-), multiplication (*), division (/), modulus (%), exponent (**) etc. It also has a math library to utilize functions like square-root, power etc.

STRING OPERATIONS·

1) Concatenation – It appends the second string to the first.

2) Repetition – concatenates multiple copies of same string.

3) Slice – it gives the character at any given index.

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

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

```
While expression:
  {….
   }
```

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

```
For variable in sequence:
  {…
  }
```

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[7]. When we call a function, it performs the task and returns the control to the calling routine.

e.g. - def function name ():

```
  Statements
  return statement
```

CHAPTER-5

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 analysed 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 analysation 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 utilised 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.

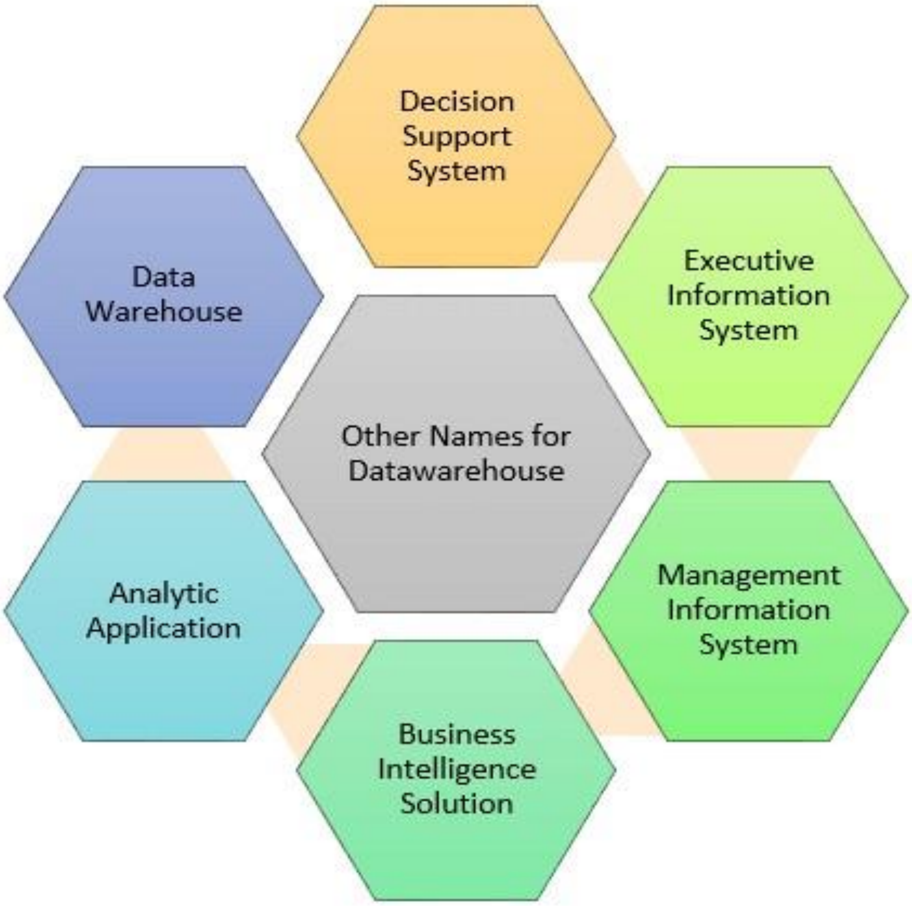

Decision Support System
Executive Information System
Data Warehouse
Other Names for Datawarehouse
Management Information System
Analytic Application
Business Intelligence Solution

Fig 27: Various names of data warehouse.

There are primarily three main types of data warehouses -

1) 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 labelled as required by the business. So, it basically provides better freedom while dealing with data.
2) 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 to store the data like employee records, which are recorded daily.
3) 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 -

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

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

3) 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.

4) 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.

5) 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 utilise the business intelligence systems.

6) Investment and finance - it is one of the most data producing, analysing 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.

7) 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-6

BIG DATA

HADOOP

MILESTONES ASSOCIATED WITH HADOOP

In **Oct 2003** Google published a research article on Google File System, which focussed on rules and regulations to store and extract value out of raw data. In **Dec 2004** Jeff Dean and Sanjay Ghemawat published a research article "MapReduce: Simplified Data Processing on Large Clusters". In **Jan 2006** Doug Cutting developed an open-source implementation of MapReduce framework[8].

These early contributions paved the way for distributed architecture approach, i.e., instead of running apps from a single system, this approach makes way to allow the same app to run parallelly on many systems at the same time, thus improving execution time. In **April 2006** Hadoop 0.1.0 was released, and in **May 2006** Yahoo successfully deployed a cluster of 300 machines.

WHAT IS HADOOP?

Hadoop is an open-source framework designed to aid in processing large amounts of data, in a distributed fashion. Hadoop has two main components-

a.) HDFS- It stands for Hadoop Distributed File System, and it takes care of the storage part for Hadoop Applications. HDFS creates multiple replicas of the files (which are stored as blocks) and store them on different nodes in a cluster. This approach confers reliability and allows for faster computations.

b.) **MapReduce-** It is the computational/processing part of Hadoop. The programs written using MapReduce are capable of processing huge amounts of data in parallel, on a large cluster having computation nodes.

HADOOP ARCHITECTURE

Hadoop deploys a master-slave kind of architecture for data storage and processing by means of HDFS and MapReduce[9]. The Hadoop master-slave architecture can be deployed locally or even on cloud.

a.) Name Node- It is a part of HDFS and represents each file present in the namespace. It regulates file access by the client. It runs on a master machine.

b.) Data Node- It is also a part of HDFS, and it allows for actual storage of business data. It runs on slave machines.

c.) Master Node- Master node confers the capability of parallel processing of the data, utilising MapReduce.

d.) Slave Node- They are the other systems which are a part of the Hadoop cluster, which actually store the data and perform complex operations. The slave nodes house a Task Tracker as well as a Data Node. This helps to establish a sync between the processes with the Name Node and Job Tracker respectively.

FEATURES OF HADOOP

1.) It is suitable for the analysis of Big Data, since it processes the logic instead of the actual data, therefore making the analysis of enormous data efficient.

2.) It is a fault tolerant system, even if a node fails the data won't be lost since it already replicates the data onto other nodes, so in case of failure the data can be retrieved from the other nodes.

3.) Hadoop clusters are highly scalable to any extent, by addition of extra cluster nodes, thus allowing to combat the growth of big data. There is even no need to modify the programming logic.

APACHE SQOOP

Sqoop (SQL to HDFS) is a tool designed for bulk import and export of data from the SQL to HDFS or from HDFS to SQL, NoSQL etc. It is a data migration tool which relies upon a connector architecture that has plugins to establish connections with external systems[9]. It is a command line utility to communicate with MapReduce.

SQOOP ARCHITECTURE

The majority of the existing DBMS are designed following the SQL as standard, but some aspects are different in each DBMS. Hence the connectors solve these problems and facilitate easy import and export of data across the systems[10]. Sqoop houses various connectors to allow working with various databases like MySQL, Oracle, SQL server etc.

THE NEED OF SQOOP

Analytical processing by Hadoop requires huge amounts of data into Hadoop clusters from various sources. So, the loading of data into Hadoop from heterogeneous sources poses a serious challenge, which is solved by using SQOOP, thus maintaining the efficiency and data consistency.

APACHE HIVE

It is a data warehousing infrastructure tool (a kind of ETL tool) to process the structured data into the Hadoop architecture. It is a database present in the HDFS. It resides on top of Hadoop to summarize the big data and make querying and analysing data easier[10]. Initially HIVE was developed by Facebook who later donated it to Apache Software Foundation, who developed it further as an open-source project under the name APACHE HIVE.

FEATURES OF HIVE

1.) Its queries are similar to that of SQL and are called as HIVE QUERY LANGUAGE (HQL) e.g., Select * from <Table Name>.

2.) First the databases and tables are created and then the data is entered.

3.) Internally HIVE translates its queries to a MapReduce job and executes the desired operation on top of the Hadoop architecture.

4.) It is extremely fast, scalable and extensible.

5.) It has a Metastore for storing information related to schemas.

6.) Since Hadoop's program work on flat files, Hive can utilise the directory structure to partition the data and improve the query performance.

HIVE ARCHITECTURE

It has 3 core components-

a.) Hive Clients- Hive provides various drivers for communicating with various applications. Like Thrift client for thrive based applications, JDBC drivers for Java based applications. They communicate with Hive server and Hive services.

b.) Hive Services- They communicate client requests for execution of the desired operations. Command Line Interface is a Hive service used for Data Definition Language operations. The Hive drivers/clients interact with the metastore and facilitate further execution.

c.) Hive Storage- Services like file storage, metastore interact with the storage component of Hive and perform the requested operations.

Hive contains data types similar to SQL, and also supports the SQL clauses like- where, from, group by etc[9]. Also, it supports mathematical and logical/conditional operators.

HBASE

HBase is another open-source tool which follows a column oriented distributed database system in the Hadoop architecture. It has a data model similar to Google's Big Table, which is a fully managed NoSQL database aiming to provide effective services for maintaining big data[10]. HBase provides quick, random access to enormous data. It runs on top of Hadoop's MapReduce. HBase happens to be a column-oriented database having tables sorted by row.

FEATURES OF HBASE

1.) Designed for low latency operations

2.) Used for random read/write operations on data

3.) Provides linear scalability

4.) It follows strict consistency when reading/writing data

HBase is highly efficient in cases where we need to search from a big set of data, because in these scenarios the traditional databases face performance failures. These performance failures are overcome by Apache HBase. In the analysis of Big Data, Hadoop confers a vital role for solving business problems dealing with huge data sets. Every component in the Hadoop architecture plays a role in Data processing, validation, storage[10]. The relational databases suffer issues of efficiency when storing enormous amounts of data either unstructured or semi-structured. Applying queries for retrieving the huge data stored in Hadoop architecture is a challenging task, so to overcome it we use NoSQL databases like HBase, MongoDB etc. HBase differs a bit from standard NoSQL databases, by having a columnar model (where all columns are grouped together as Columnar families) and storing data as key-value pairs, and provides low latency access to small sized data stored in big data sets.

PySPARK

Apache SPARK is a framework supporting blazing fast real time processing, by means of in memory computations to aid in analysis. Since MapReduce could only process in batches and lacked the ability of real time processing, to facilitate this APACHE SPARK was developed. Along with these aforementioned capabilities it also supports interactive and iterative queries and can utilise the Hadoop components, even though it also has its own cluster manager.

The APACHE SPARK was developed using Scala. To extend Python support to SPARK, the Apache Community released a tool **PySpark**, which helps in the integration of Python with the SPARK tool, by means of a library **Py4j**.

CONCLUSION

During the course of this training, I learnt the importance of managing Data in today’s world. The Data generated is increasing day by day, in fact it is increasing each minute, and the rate at which it is increasing is exponential. Apart from this fact, the generated data tends to be random like images, text files, audio/video clips etc. The organizations have enormous amounts of this data and face serious challenges to extract value out of it. Since the consumer data is increasing rapidly, we need better approaches to store data, instead of traditional databases, since they suffer from storage and performance issues. The organizations are shifting their entire database into cloud-based solutions to better manage the data storage and process it efficiently. Since, Big Data is a hot term in the industry, it surely requires specialized solutions to deal with this enormous amount of data. APACHE HADOOP is one such framework along with other sub-components like Hive, Hbase, Pig, SPARK, PySpark etc is being utilized by the organizations. The analysis of this petabyte sized data is the key area many organisations are working upon to offer insights and help businesses prosper.

During the training, I learnt a lot of new things and recent advancements in the field. I am constantly finding ways to integrate this with the Biotechnology sector.

REFERENCES

1) Molinaro, Anthony. SQL Cookbook: Query Solutions and Techniques for Database Developers. " O'Reilly Media, Inc.", 2005.

2) Beighley, Lynn. Head First SQL: Your Brain on SQL--A Learner's Guide. " O'Reilly Media, Inc.", 2007.

3) Malik, Upom, Matt Goldwasser, and Benjamin Johnston. SQL for Data Analytics: Perform fast and efficient data analysis with the power of SQL. Packt Publishing Ltd, 2019.

4) Bloch, Joshua. Effective java. Pearson Education India, 2016.

5) Sierra, Kathy, and Bert Bates. Head first java. " O'Reilly Media, Inc.", 2003.

6) Ascher, David, and Mark Lutz. Learning Python. O'Reilly, 1999. Barry, Paul.

7) Head first Python: A brain-friendly guide. " O'Reilly Media, Inc.", 2016.

8) Simon, Phil. Too big to ignore: the business case for big data. Vol. 72. John Wiley & Sons, 2013.

9) Kitchin, Rob. The data revolution: Big data, open data, data infrastructures and their consequences. Sage, 2014.

10) Davenport, Thomas. Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press, 2014.