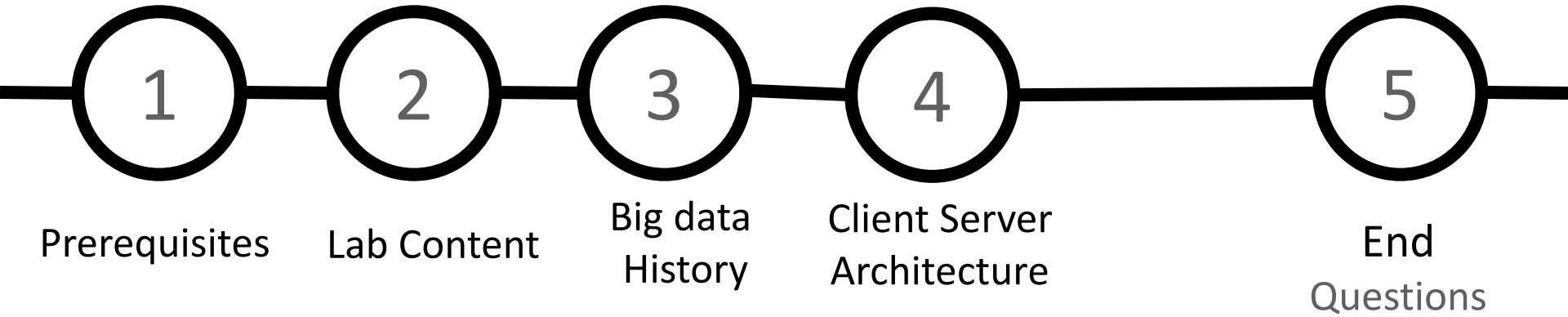
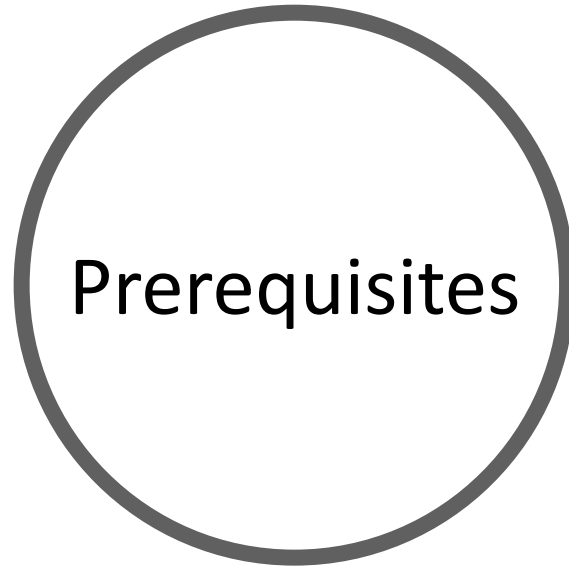


Pig Picture

# AGENDA



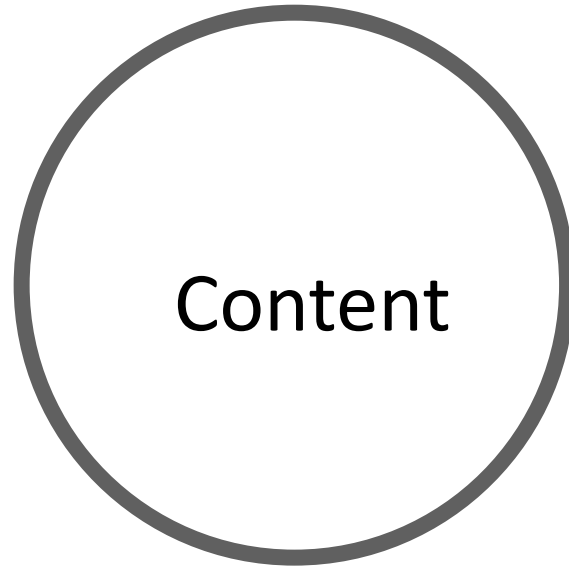


# Prerequisites

Concepts & Tools

# Prerequisites

- ✓ Any programming language (Basics , OOP, Multi Threads ,Socket).
- ✓ Database ( Basic SQL : DML , DDL).
- ✓ OS ( H/w => kernel => shell => apps) & **File System**.
- ✓ Network ( Basic Configuration e.g.: ipconfig , ping , ...etc. ).
- ✓ VMware ( Host only , Bridge , Nat )
- ✓ IDE (Main Java IDE like NetBeans , Eclipse , IntelliJ )

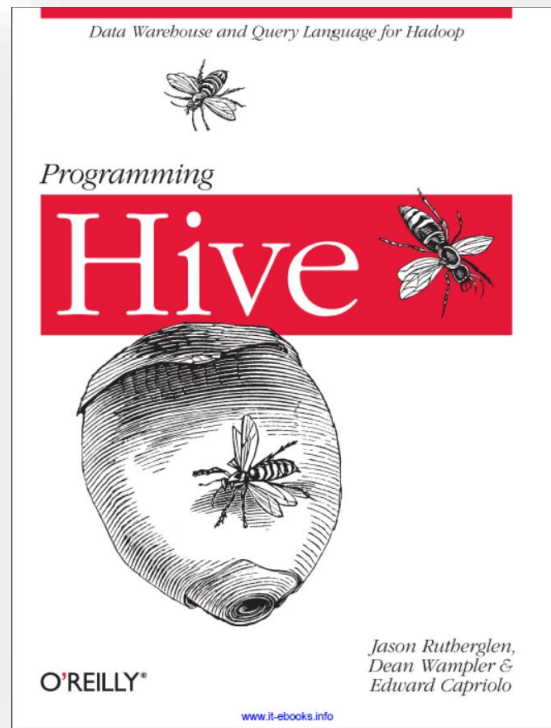
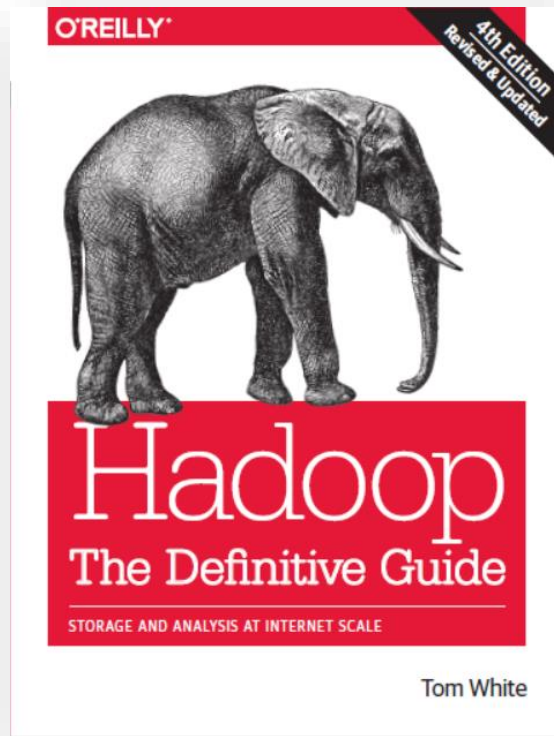


Lab Syllabus

# Lab Syllabus

- ✓ Pig Picture ( Client – Server , sockets , Hadoop History ) (1 Lab)
- ✓ Hadoop HDFS. ( 2 Lab)
- ✓ Map – Reduce Model. (2 ~ 3 Lab).
- ✓ YARN. (1 Lab)
- ✓ Hadoop Cluster (Multi Node). (1 Lab)
- ✓ Hive. ( 2 Lab)
- ✓ Spark . (1 ~ 2 Lab)
- ✓ Cloudera Navigator as a Hadoop Commercial Tool.

# Text Books & Tutorials



cloudera®

Cloudera Essentials  
for Apache Hadoop



databricks





# History

- Human need Numbers (Supermarket Case)
  - Files => DB => Server Problem (scalability).
  - Types of Scalability (Vertical vs Horizontal).
  - Main Problem data( Storing & Computation).
- **Google Search Engine** : (index for all web pages on internet )
  - 2004 [Google] (processing on intensive data) using map reduce.
  - **2003** [Google] ( Store Data in file system ) using GFS.
- **Terminologies** : Big Data , Hadoop , Map Reduce , DS , ML , DM , DS , .. etc.
- **3V - 7 V** : Data changed based on volume , variety , velocity , ... etc.

# History

- **2006 Yahoo:** Doug Cutting implemented an open source s/w called Hadoop solve data storage & computation problem.



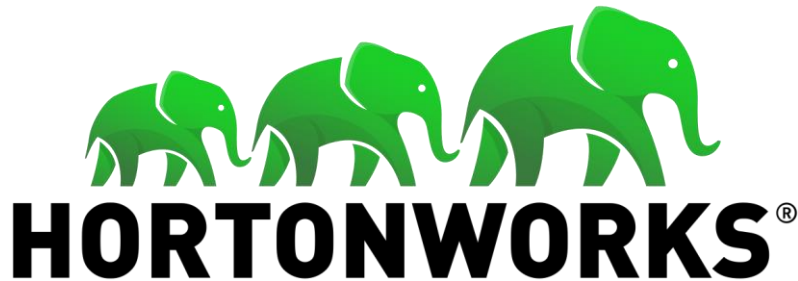
- **2008 Cloudera:** Dr.Amr et al founded his company and he provide Hadoop services ,support , training ,and certifications.



# Hadoop Vendors

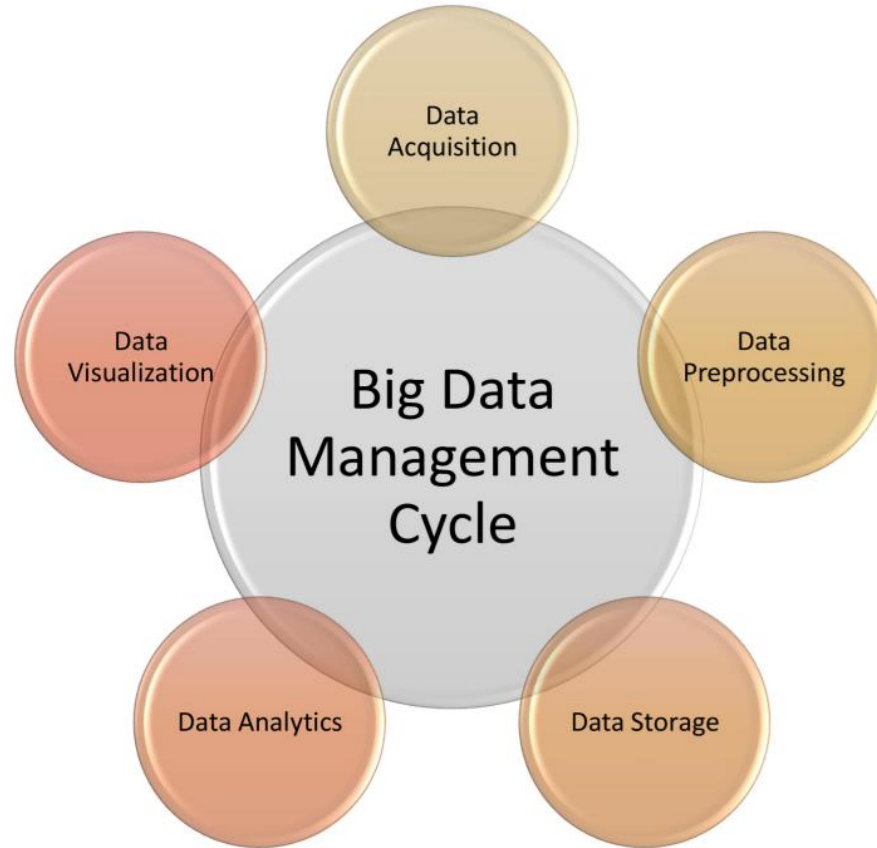
MAPR<sup>®</sup>

cloudera<sup>®</sup>



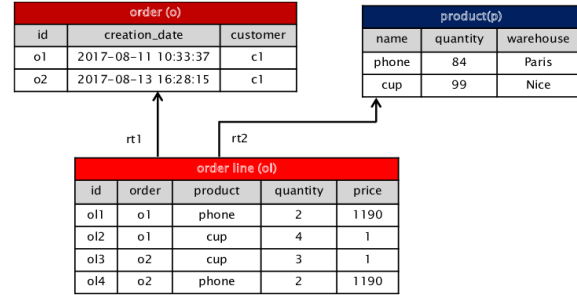
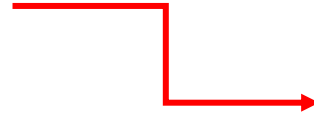
IBM  
InfoSphere DataStage

# BDMLC



# Data Types

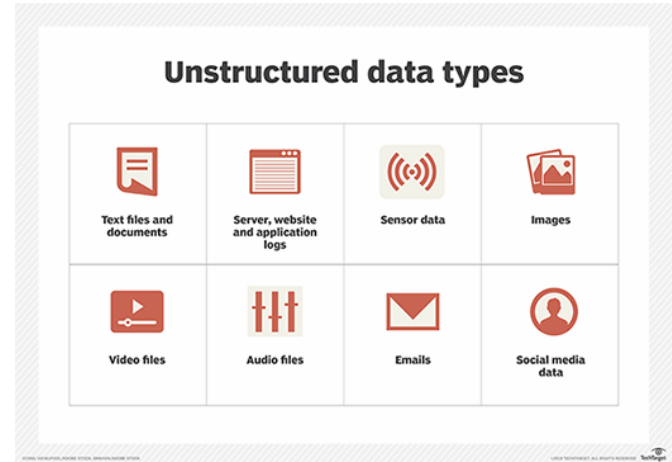
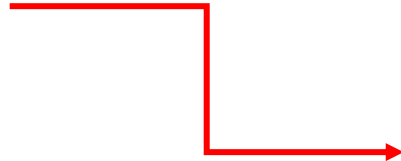
- Structure (Database)



- semi-Structure



- Unstructure (o.w)



# XML Vs JSON Example

```
<?xml version="1.0" encoding="UTF-8"?>
<DatabaseInventory>
  <DatabaseName>

  <GlobalDatabaseName>production.cubicrace.com</
GlobalDatabaseName>
    <OracleSID>production</OracleSID>
    <Administrator EmailAlias="piyush"
Extension="6007">Piyush
Chordia</Administrator>
    <DatabaseAttributes Type="Production"
Version="9i" />
    <Comments>All new accounts need to be
approved.</Comments>
  </DatabaseName>
  <DatabaseName>

  <GlobalDatabaseName>development.cubicrace.com<
/GlobalDatabaseName>
    <OracleSID>development</OracleSID>
    <Administrator EmailAlias="kalpana"
Extension="6008">Kalpana
Pagariya</Administrator>
    <DatabaseAttributes Type="Development"
Version="9i" />
  </DatabaseName>
</DatabaseInventory>
```

```
{
  DatabaseInventory: {
    DatabaseName: [
      {
        GlobalDatabaseName: "production.cubicrace.com",
        OracleSID: "production",
        Administrator: [
          {
            EmailAlias: "piyush",
            Extension: "6007",
            value: "Piyush Chordia"
          }
        ],
        DatabaseAttributes: {
          Type: "Production",
          Version: "9i"
        },
        Comments: "All new accounts need to be approved."
      },
      {
        GlobalDatabaseName: "development.cubicrace.com",
        OracleSID: "development",
        Administrator: [
          {
            EmailAlias: "kalpana",
            Extension: "6008",
            value: "Kalpana Pagariya"
          }
        ],
        DatabaseAttributes: {
          Type: "Development",
          Version: "9i"
        }
      }
    ]
  }
}
```

## Structured Data

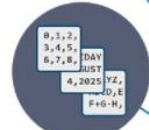
vs

## Unstructured Data

Can be displayed  
in rows, columns and  
relational databases



Numbers, dates  
and strings



Estimated 20% of  
enterprise data (Gartner)



Requires less storage



Easier to manage  
and protect with  
legacy solutions



Cannot be displayed  
in rows, columns and  
relational databases



Images, audio, video,  
word processing files,  
e-mails, spreadsheets



Estimated 80% of  
enterprise data (Gartner)

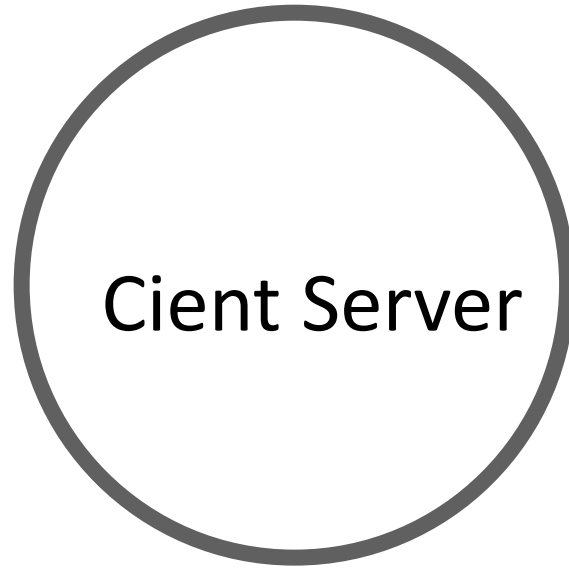


Requires more storage



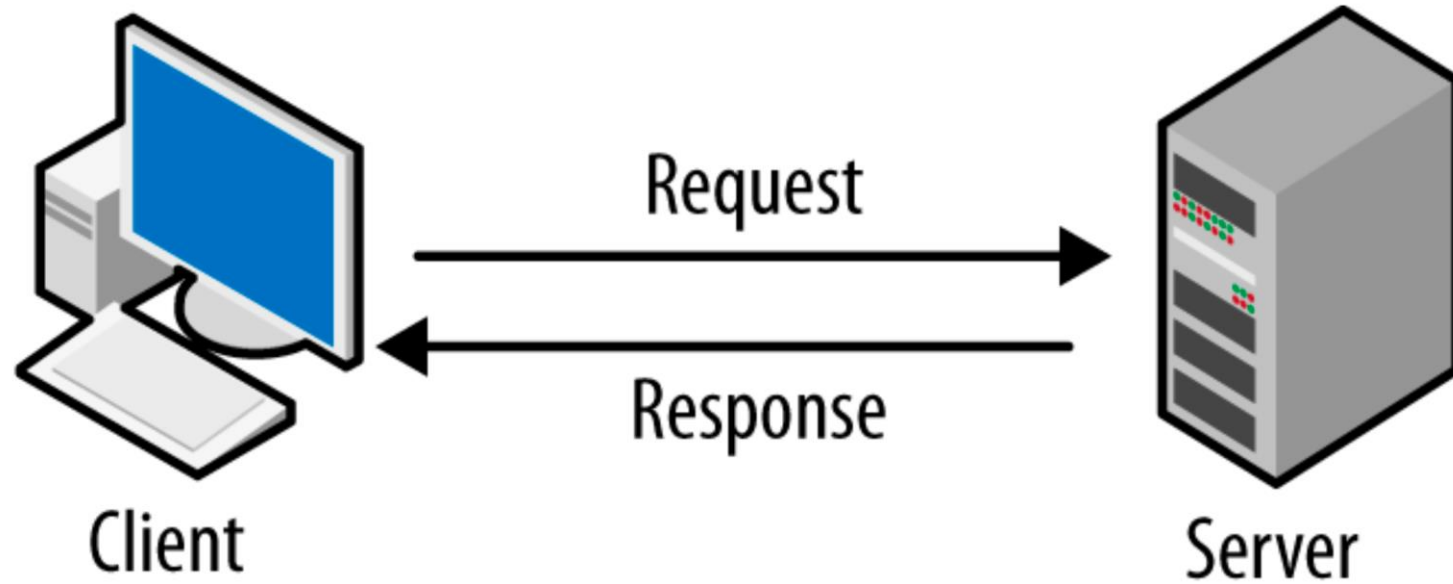
More difficult to  
manage and protect  
with legacy solutions



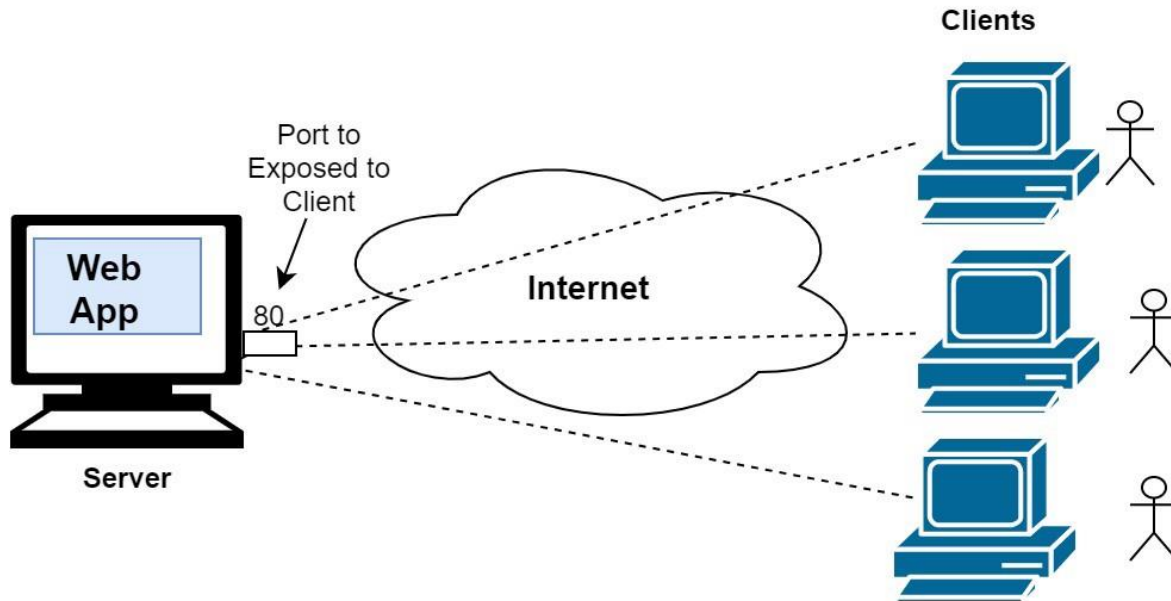


Architecture

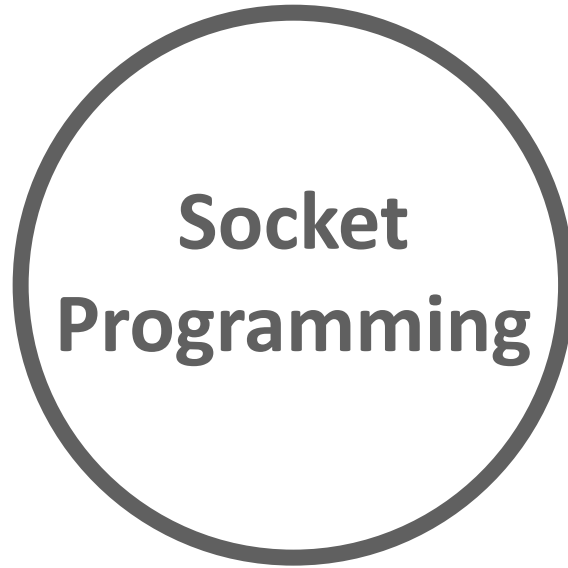




[https://madooei.github.io/cs421\\_sp20\\_homepage/client-server-app/](https://madooei.github.io/cs421_sp20_homepage/client-server-app/)



<https://nimesha-dilini.medium.com/simple-introduction-to-client-server-architecture-concept-7d2979bed31d>



Main topic

# Socket Programming

Sockets provide the communication mechanism between two computers using TCP.

A client program creates a socket on its end of the communication and attempts to connect that socket to a server.

When the connection is made, the server creates a socket object on its end of the communication. The client and server can now communicate by writing to and reading from the socket.

# Establishing a TCP connection between two computers using sockets

The server instantiates a ServerSocket object, denoting which port number communication is to occur on.

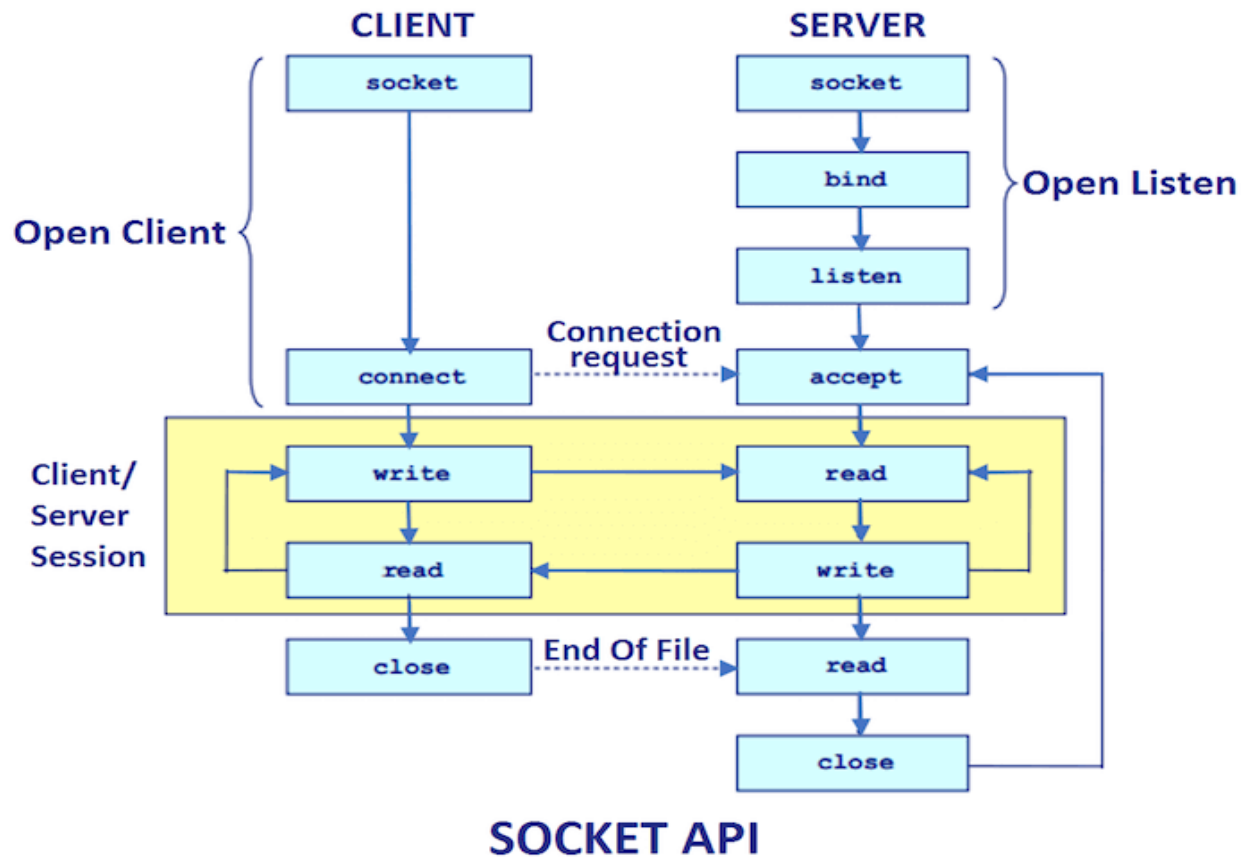
The server invokes the accept() method of the ServerSocket class. This method waits until a client connects to the server on the given port.

After the server is waiting, a client instantiates a Socket object, specifying the server name and port number to connect to.

# Establishing a TCP connection between two computers using sockets cont.

The constructor of the Socket class attempts to connect the client to the specified server and port number. If communication is established, the client now has a Socket object capable of communicating with the server.

On the server side, the accept() method returns a reference to a new socket on the server that is connected to the client's socket.



Source : <https://static.javatpoint.com/core/images/socket-programming.png>

# Assignment 1 : DeadLine Next Lab

- 1- Install Vmware & Ubunut Linux (any version).
- 2- Install Java (jdk)
  - **sudo apt-get update**
  - **sudo apt-get upgrade**
  - **sudo apt-get install openjdk-8-jdk**
  - **java -version => output java1.8**
- 3- Install JAVA IDE. (**NetBeans or IntelliJ**)
- 4- Write matrix multiplication distributed model using client/server. (Bonus)





QUESTIONS

# THANK YOU!

