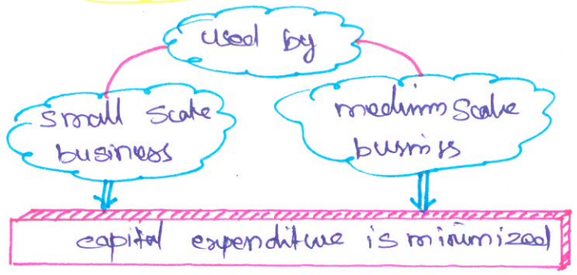
INDEX

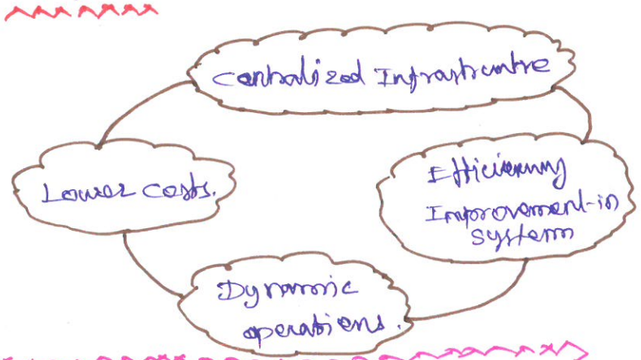
Introduction :- to cloud computing :-

* Services offering to specific -
  - Consumers

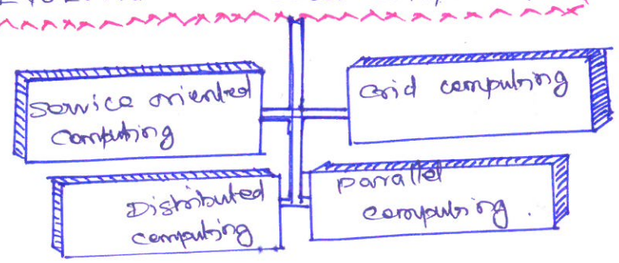


* Customer access services using

Benefits :-



EVOLUTION OF CLOUD COMPUTING :-

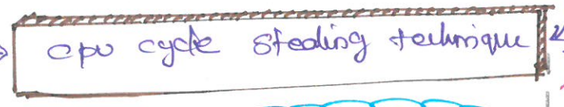


Service oriented computing :-

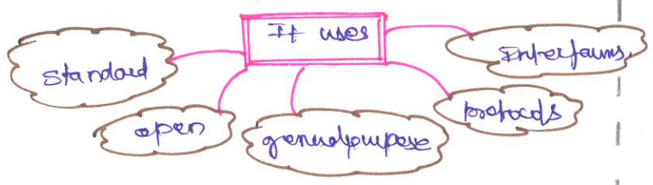
* Paradiam for distributed computing
* built using Services as fundamental

Grid computing :-

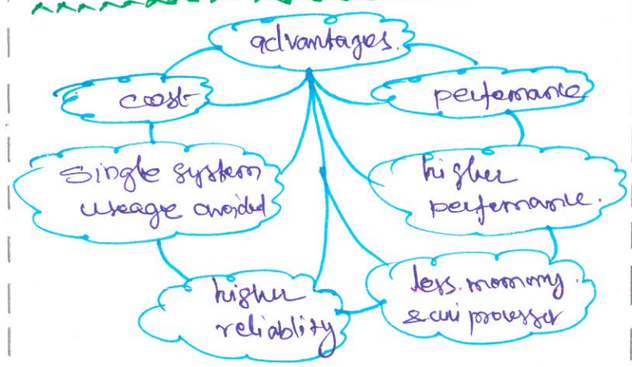
* form of distributed computing
* Working on the following technique



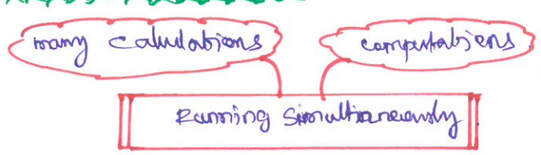
* composed of clusters of network, loosly coupled computers to perform large tasks.
* Resources are not Centralized control



Distributed computing :-



parallel computing :-

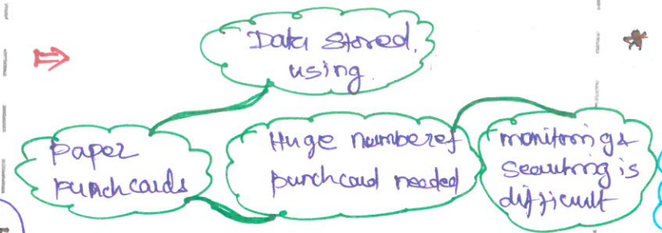


HARDWARE EVOLUTION :-

First generation computers :-

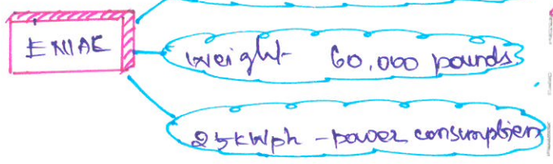
* year 1943
* built using
  - hard wired circuits
  - vacuum tubes

⇨



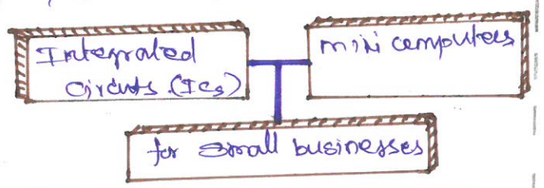
Second generation computers :-

* year 1946
* ENIAC - Electronic Numerical Integrator & computer
* It is a First digital computer
  - transistors
  - printer circuits



Third generation computers :-

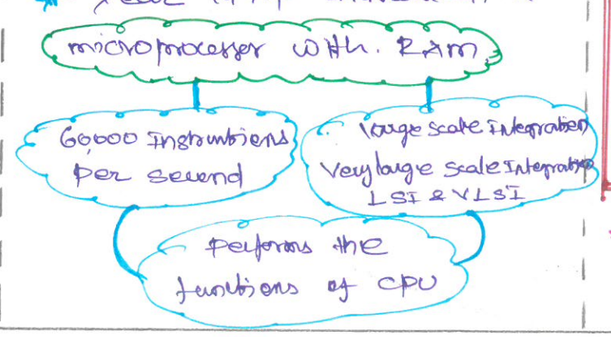
* Year 1963.



1971 - Intel released first microprocessor - Intel 4004

Fourth generation computers :-

* Year 1974 Introduced a microprocessor with Ram



* Server consolidation

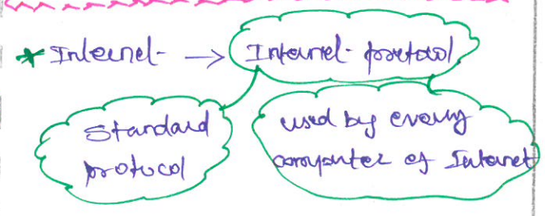
migrating network services and applications from multiple computers to single

change the world to cloud computing and shape the Business world.

Growth of mobile technology.

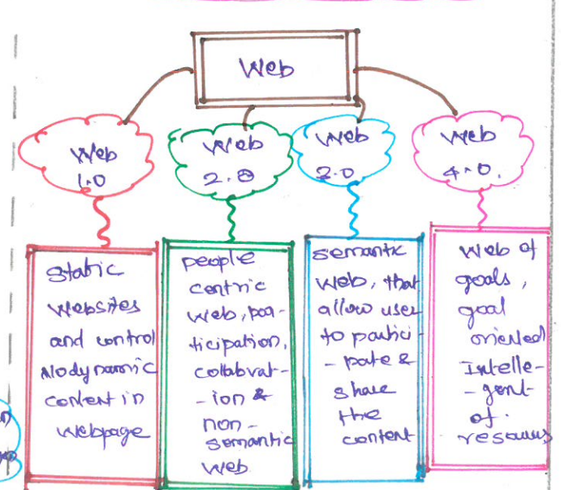
INTERNET SOFTWARE EVOLUTION

* Internet -> Internet protocol

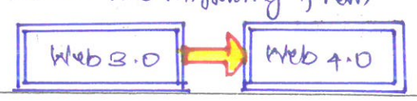


* year 1983 first computer connected with Internet (ARPANET)

uses Network control program



* We are now migrating from



Establishing Common Protocol for Internet:

* Lower layer protocol layers are provided by Interface Message Processor (IMP) host interface
* NCP provided by transport layer.
  - ARPANET Host to Host Protocol
  - Initial Connection Protocol
* File transfers: FTP
* For Email Services: SMTP
* ARPANET changed from NCP to TCP/IP protocol suite.
* TCP/IP - Start of Internet.
* Growth of Internet in 1990s,
  - Reduction in IP address (IPV4)
* IETF Released IPV6 in 1995

Building Common Interface for the Internet:

* Mosaic - first web browser - available to general Public
* Mosaic - support graphics, sound, and video clips.
* 1994 - Netscape Released its web browser.
* 1994 - Mozilla 1.0 Released
* 1995 - Microsoft Released Internet Explorer as Graphic Web browser.

SERVER VIRTUALIZATION

* Virtualization - Running multiple independent virtual OS on a single physical Computer

Parallel Processing:

* Performed by simultaneous execution of program instructions
* allocated across multiple processors
* Running a program in less time.
* It's multiprogramming

Vector Processing:

* operating in multitasking manner.
* allow single instruction to manipulate two arrays of numbers
* applications with less well-formed data, Vector Processing was less valuable.

Symmetric Multiprocessing System (SMP)

* address the problem of resource management in master/slave models.
* Same as single processor, multiprogramming platform.

Massively Parallel Processing System (MPP)

* System with many independent arithmetic units or entire microprocessor
  - Run in parallel massive numbers of Computers used to solve single problem.

* Virtual Server can run its own operating system.
* Virtual Machine (VM) is a digital version of physical Computer
* Types of Virtualization:
  - Full Virtualization
  - Para Virtualization
  - OS level Virtualization
  - Network level Virtualization

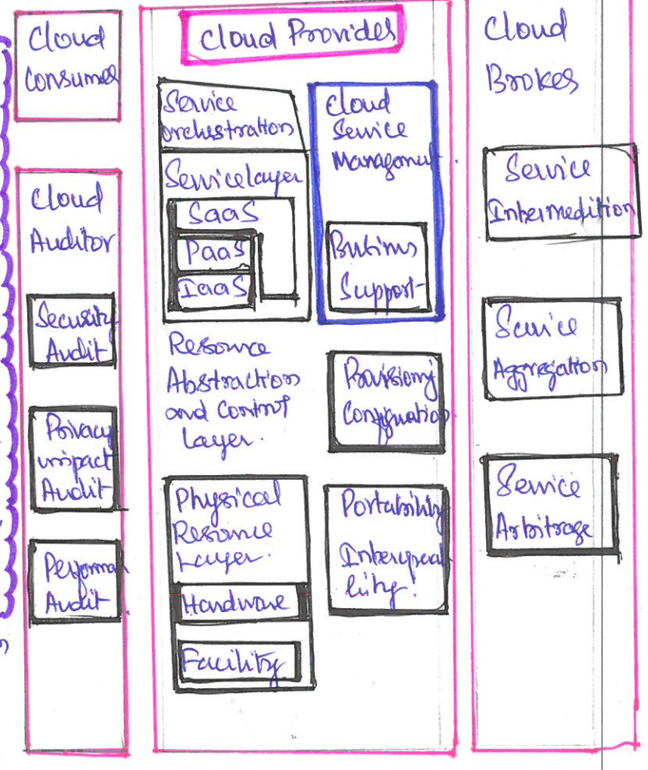
Web Services:

* WSFL - Service flow
* UDDI - Static → Service Discovery
* UDDI - Direct → Service Publication.
* WSDL - Service Description.
* SOAP - XML based Messaging
* Network - HTTP, FTP, Email, MQ, IIOP, etc.

* SOAP - exchange of information in distributed environment.
* HTTP - Application layer protocol for transmitting hypermedia documents
* UDDI - Universal Description, Discovery, and Integration is an XML based registry for businesses worldwide to list themselves on the internet.

NIST CLOUD ARCHITECTURE.

* Cloud Computing is a model.
  - enable ubiquitous
  - Convenient
  - On-demand network access to shared pool of configurable Computing resources.
  - Eg: Networks, Servers, Storage, applications, & Services.



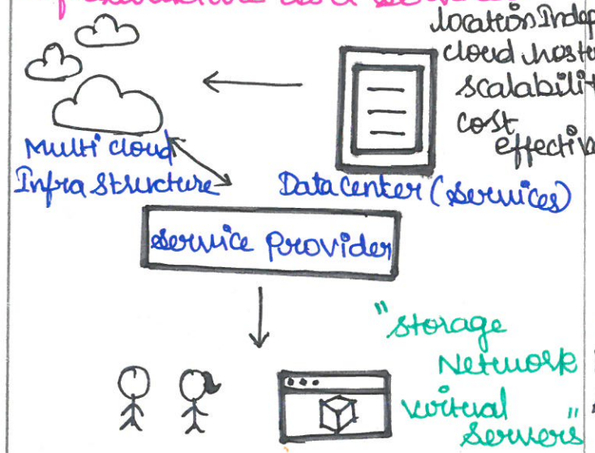
* Features of cloud:
  - Three types of cloud Service Models.
    * IaaS (Infrastructure as a Service)
    * PaaS (Platform as a Service)
    * SaaS (Software as a Service)

WEB SERVICE OVERVIEW

↳ Hardware and Software uses HTTP
↳ Some protocols respond to request from clients
↳ Displays the content of the website

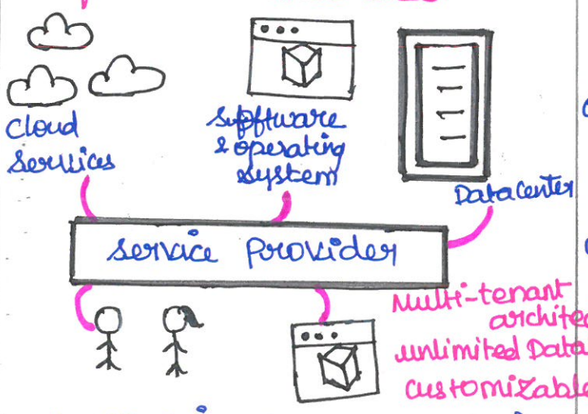
Features of Clouds:
- Infrastructure as a Services
- Platform as a Services
- Software as a Services
- Anything as a Services

Infrastructure as a Services



Application client (End user) & Application

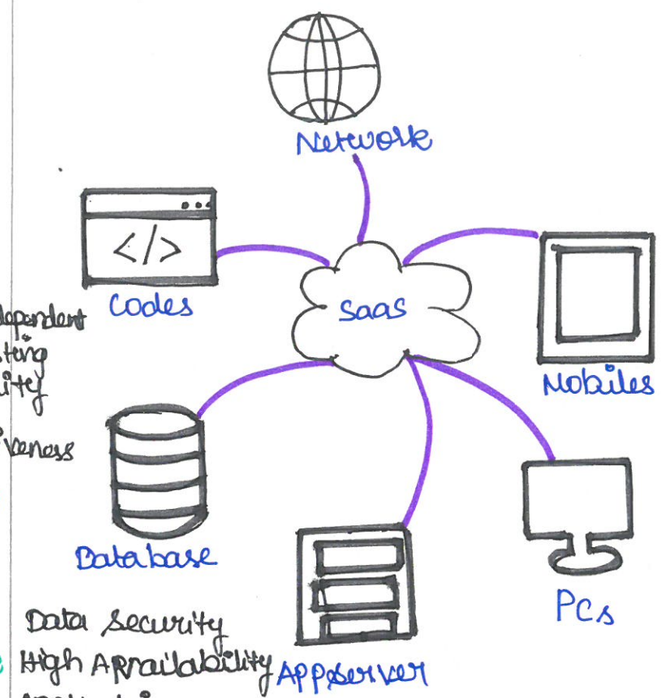
Platform - as - a Services



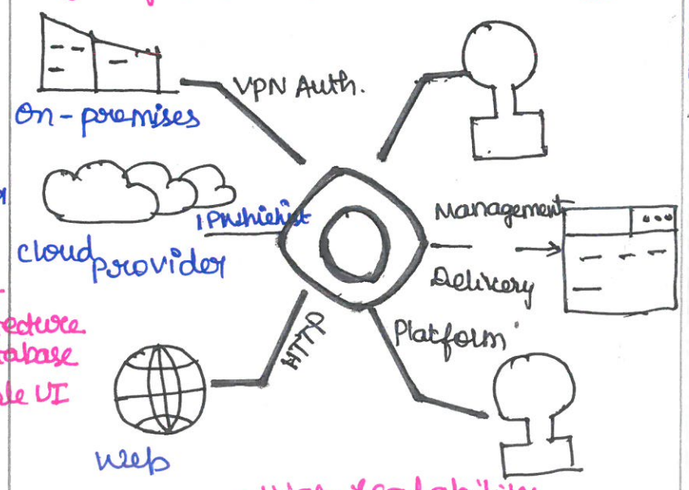
" Run time Environment
Data processing "

Software as a Service (SaaS)

" Scientific application "
Social network
End user application



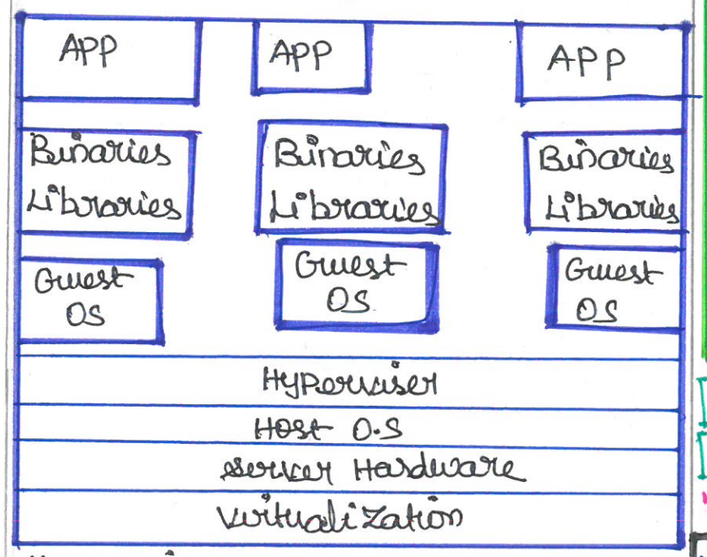
Anything as a Services (xaas)



- High Scalability
- Multi-tenancy
- Location independence
- Team based billing

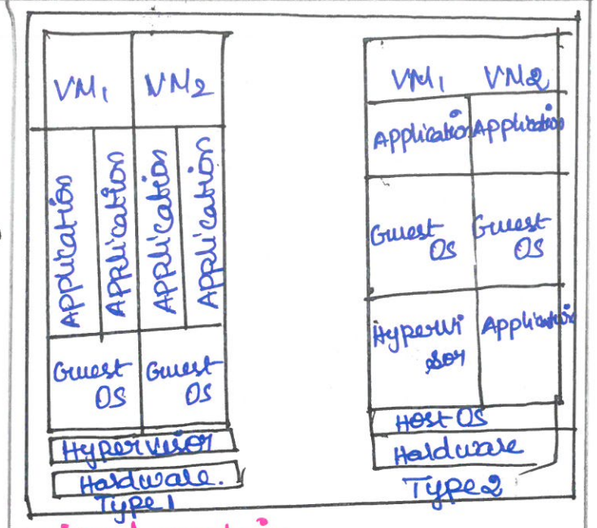
Communicating as a Services

Virtualization Architecture.

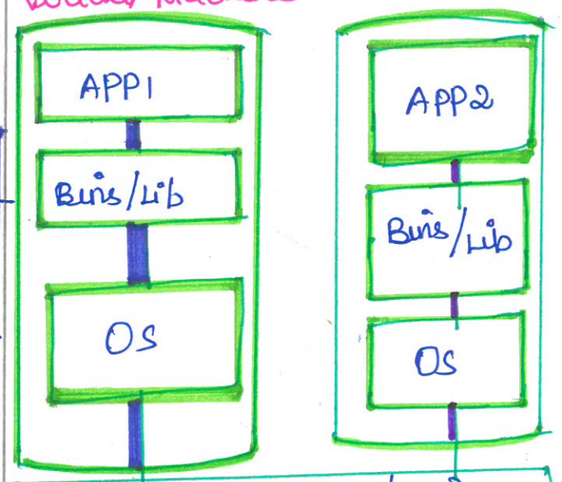


Hypervisors:
→ Computer Software & Hardware
→ Enables you to host Multiple Virtual Machine
→ Runs its own program.

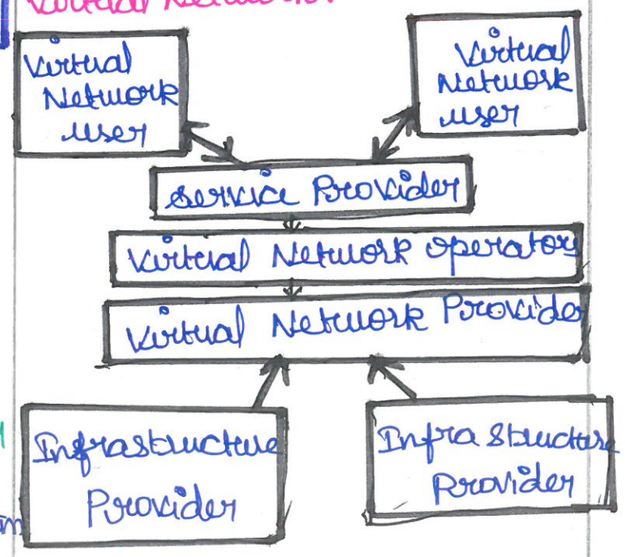
Types of Hypervisors:
* Hypervisors (Type I)
  Run on the host's hardware to control
  eg: Microsoft HyperV, Oracle VM, KVM
* Hypervisors (Type II)
  Software using an Operating System
  eg: Hosted hypervisors, Oracle VM Virtual Box



Virtual Machine



Virtual Network.

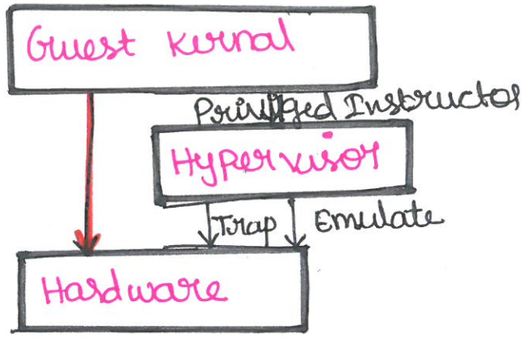


Virtualization

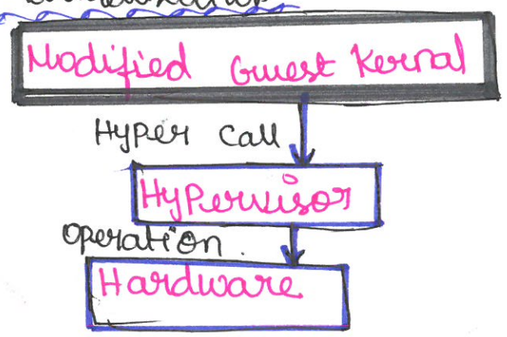
↳ Software simulate hardware function and create virtual computer system

1. Full Virtualization
2. Para Virtualization

Full Virtualization



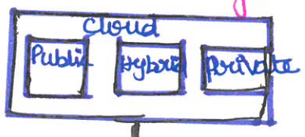
Para Virtualization



Security Problems

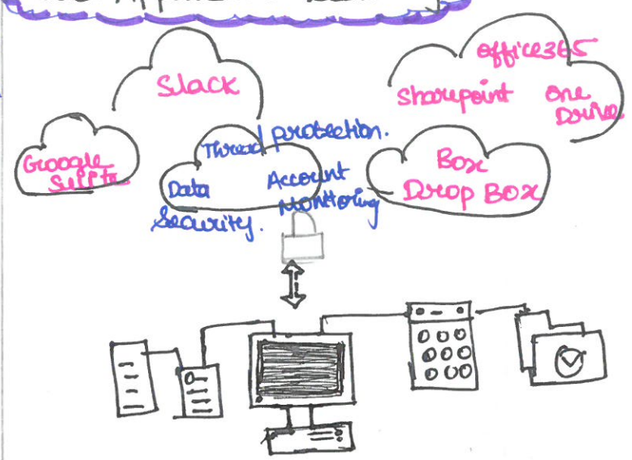
↳ insufficient identity, credential, access and key management.

Data Security:



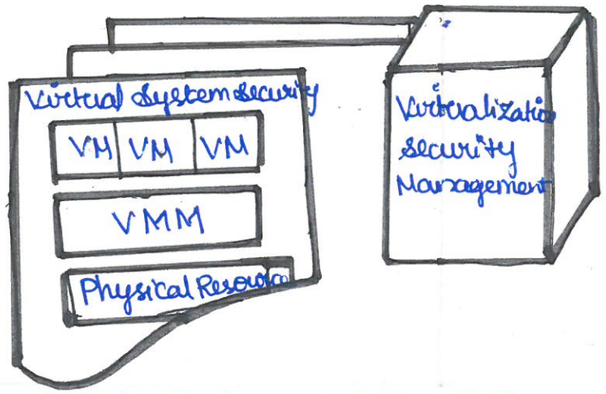
Data security & privacy
- Data integrity — S/W
- Data confidentiality — S/W
- Data availability — H/W
- Data privacy — H/W

Cloud Application Security

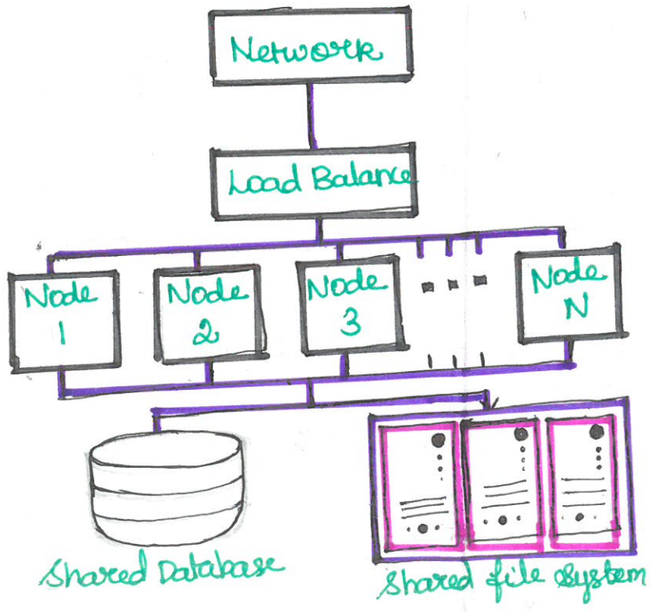


Virtual Machine Security

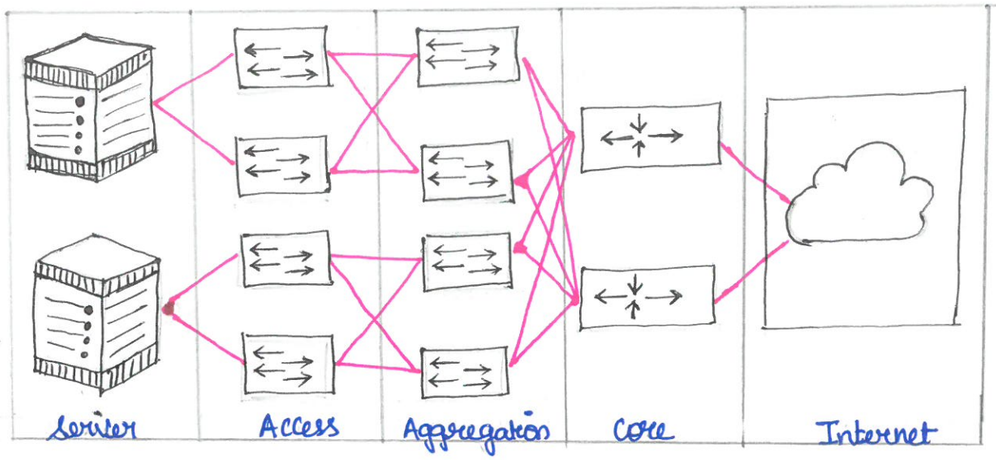
- Securing virtual machines from threats



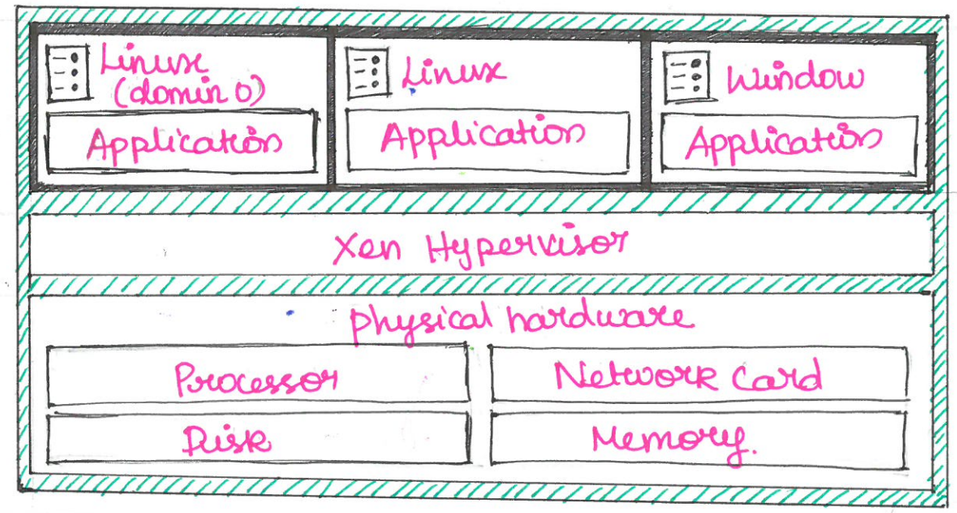
Cloud Data Center Architecture



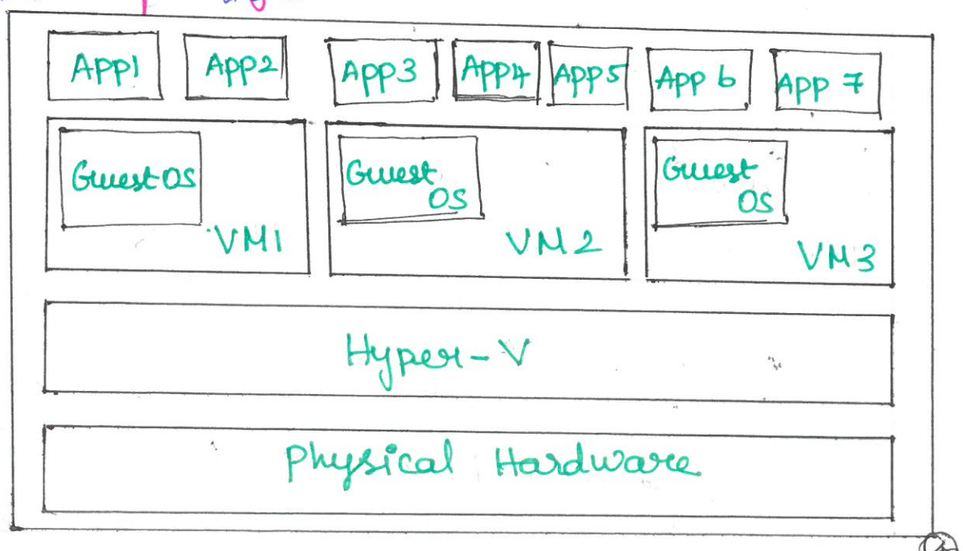
Data Center Network (DCN)



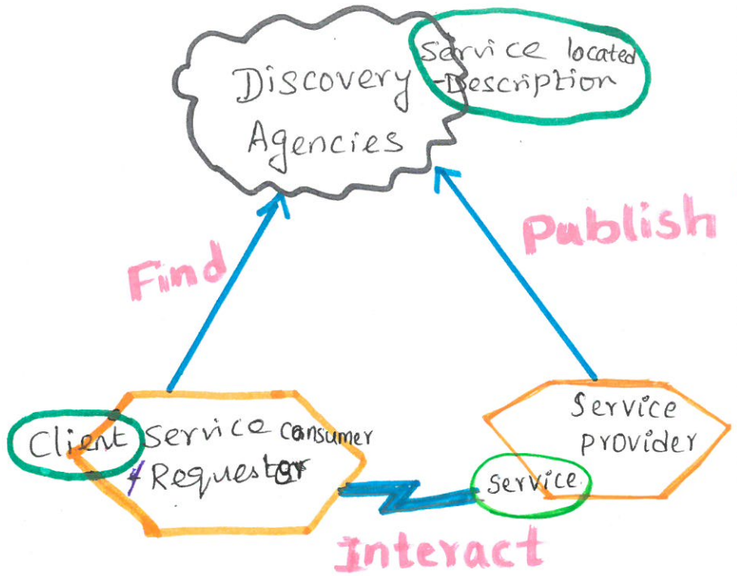
Xen Architecture - Software Defined DataCenter



Microsoft Hyper-V



Service oriented Architecture (SOA) / web service triangle

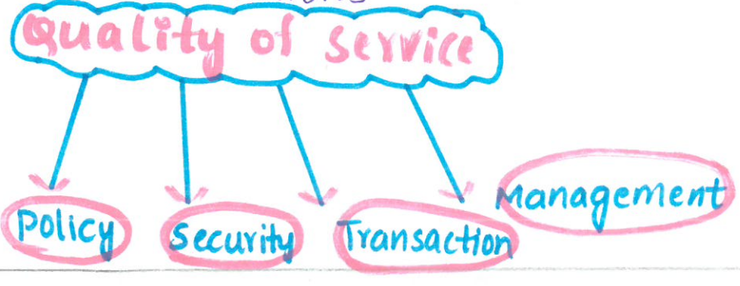


⇒ Services are the logical entities defined by one (or) more published interfaces

Service provider - software entity implements service specification.

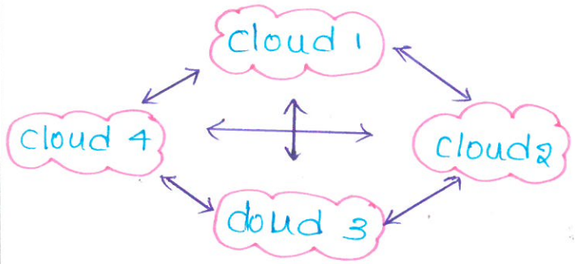
Service consumer - calls service provider can be another service (or) End user application

Service locater - Act as a Registry Examining Service provider interfaces and Service locations



Cloud federation :

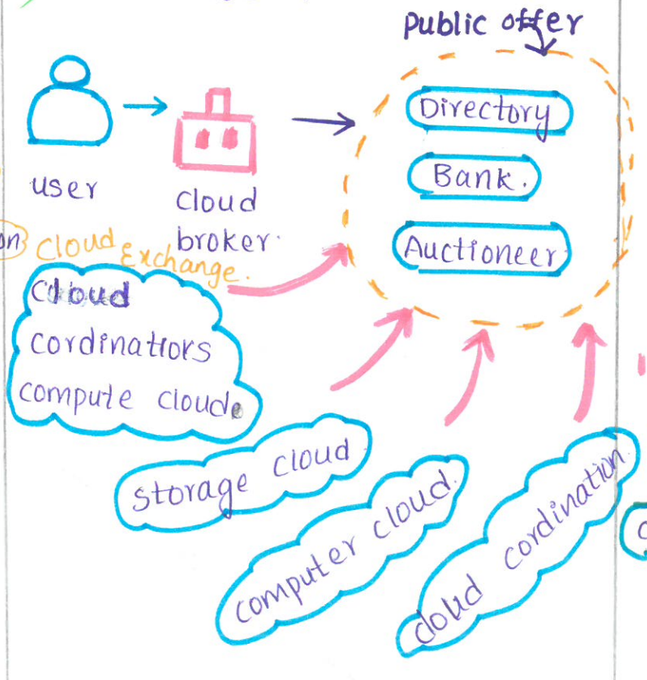
* Cloud federation also known as federation cloud
* Deployment and management ⇓ External & internal cloud
* Multi-National cloud system, ⇓ integrates that private community and public clouds



Cloud federation

⇒ federated cloud is created by connecting the cloud environment of different cloud.

⇒ It can use 2 or more clouds



properties of cloud federation

⇒ users can interact with the architecture either centrally or decentralizer manner user to Securely authenti-cate with multiple applica-

⇒ It can be Practiced with various niches like Commerical and Non commerical

⇒ marketing objects ← infrastructure, software, platform.

Characteristics

* Authentication
* Integratiy
* Monitoring
* object
* contracts
* Provisioning
* Service Management
* inter operability
* Commercia-lization

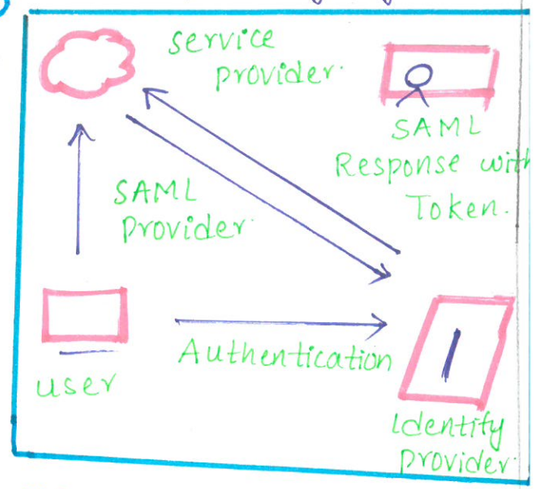
usage of cloud federation properties

1) Interaction of architecture
2) Expansion of federation
3) four centric federated cloud
   1) Business 2) provider 3) Service, customer
4) practice Niche
   - commercial
   - Non commercial
5) Volunteer
6) Visibility

SSO - Single sign on

SSO - It is an authentica-tion method that enables user to Securely authenti-cate with multiple applica-

SAML - Security Assertion Mark up language.



Presence protocol

* It Works ON SSO and

Steps involved in SSO are as follows :

Step 1 : The service provider sends SAML Request

Step 2 : If it is not authenticated it will prompted for the authenticated credentia

Step 3 : Identify Providers will send an Authentication Token in SAML Response.

Identity

* Trust between two parties
* Authenticating users and to convey information needed
* Authorize accessing resource

Federated Identity Management

Two Operational roles

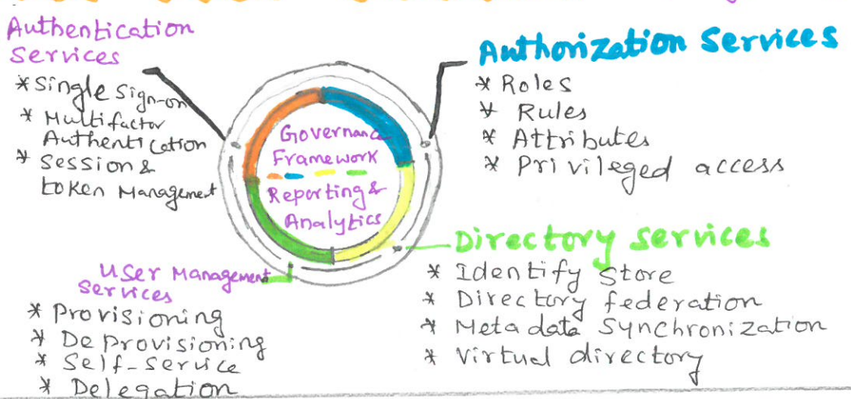
- **Idp** (Identity Provider)
- **Sp** (Service Provider)

Access Management

* Process of controlling and tracking access
* Different Privileges within a System on the Individual needs of users

Ex: Payroll System

IAM Service Components [Identity Access Management]



Identity-as-a-Service (Iaas)

Saas Model < Solves the identity Problem
Provides SSO for web application

Provide Service for Elements

- Governance
- Risk Management
- Compliance

* Iaas is a prerequistie, of Cloud computing for managing the identities and its access.

Languages used: SAML, XACML

Framework: IGF [Identity Governence Framework]

Examples: Ping Identity, Symplified Tricipher & Arcot Systems

PRIVACY

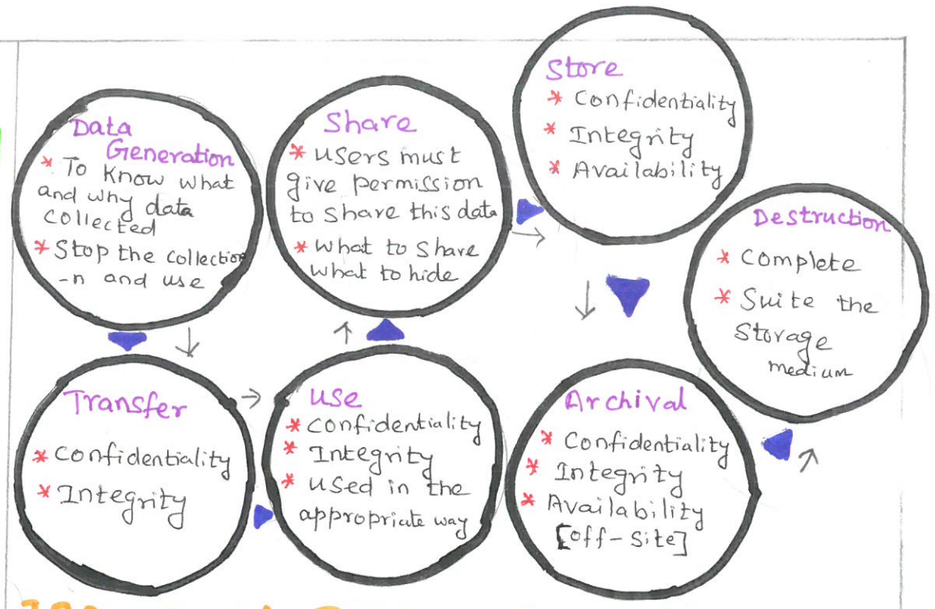
Privacy risk and Cloud

A user's Privacy and Confidentiality risks vary with terms of service and privacy policy established by Cloud Provider

* It is difficult to assess the status of information with legal uncertainties

Protecting - Privacy Information

Federal Trade Commission [FTC] is educating Consumers & businesses, importance of personal information privacy.



Issues of Privacy Protection

* Lack of Physical control
* Protection of Data copies
* Legal Problems due to Privacy Law's all over the world

Steps to provide privacy

Attention to privacy of personal Information should taken in SaaS and managed services

* Transferring personally identifiable information to and from a consumer's system.
* Storing personal Information on consumer's System,
* Transferring Anonymous data from Consumer's System
* Installing Software on a Consumer's System
* Storing & processing user data at Company
* Deploying Servers.

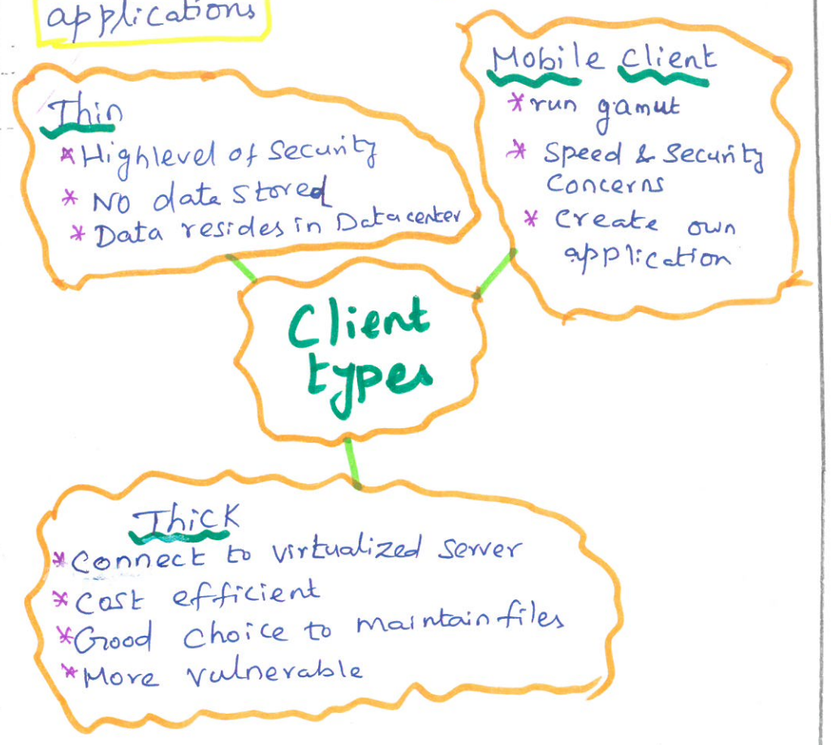
The Future of privacy in cloud

* Include better policies and practices by Cloud Providers
* Establish Standards
  - To analyze difference between cloud providers
  - Assess the risks faced by users
* Cloud Providers would benifit users with greater transparency about risks.

Hardware and Infrastructure

Client

different types of Clients Offer different ways to interact with your data and applications



Security

*Issue in Cloud computing
* Third party Stores data

Data Leakage

* Benifit is the centralization of data
* Thick Clients: download files, Maintain

* Thin clients better for Centralized data storage

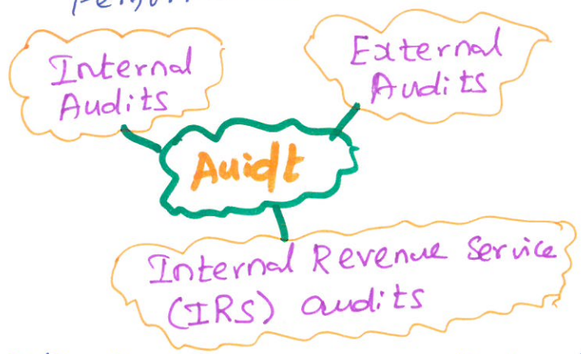
Forensics

Breach, Cloud provider respond with less downtime than investigation.

VM Cloned for offline analysis

Auditing

* Periodic examination
* assess and document vendor's Performance



* Performed by certified public Accounting [CPA]

Network

* Access cloud via Internet

Basic Public Internet

* Basic choice for Cloud Connectivity

Example: Internet Service Provider, [ISP] via Broad band, Dial-up

Accelerated Internet

* Advanced application delivery features

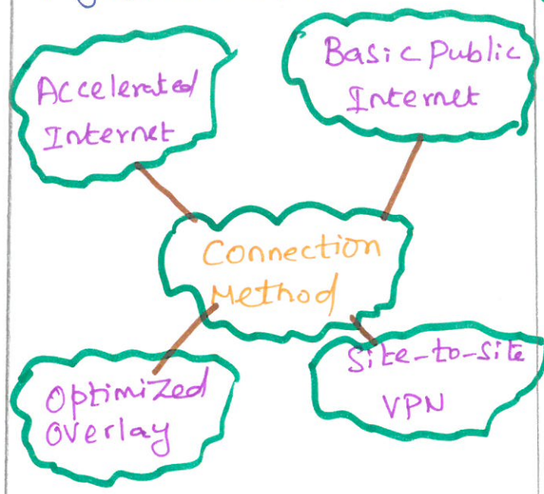
Optimized Internet overlay

Customers to access Cloud via Public Internet

Site-to-Site VPN

Connect to Service provider directly using private wide Area Network [PWAN]

* Confidentiality
* guaranteed bandwidth



Cloud-Providers

Third party Company Offers: Cloud based platform, Infrastructure, Application, Storage-Services

Example: Amazon web Services [AWS]

* Performance Improved
* Bandwidth charge reduced

Cloud Consumers

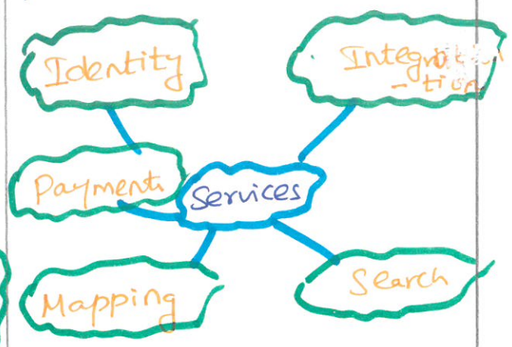
* Include: Subscribers to Cloud services; Users of Cloud Services

Example: Drop Box, Salesforce

* File hosting and Storage, Synchronization will performed on remote Servers.

Services

Depends on Cloud Provider



Identity

* users
* Digital Identity
* Single Sign-on Standard
* OpenID Authentication

Example: Google, IBM

Integration

* one permises infrastructure

Example Amazon's Simple Queue Service

* use Queue Service

Example BizTalk Services

* relay Services

Mapping

Services, Google Map, Microsoft virtual Earth, provide Cloud based functions.

Payments

* online Payments
* Signup → Accept Credit Cards → Send money

Search

* Rich feature
* Microsoft Live search is used in cloud.
* Minimal for Small organization
* Searchability is limited to the organizations

Accessing The Cloud

Platforms

How Cloud Computing Environment is delivered to you

1. Web Application Framework

Support the development → dynamic Web Sites; Web applications; Web Services

* Web Application framework Provide a Standard Way to → build, Deploy → Web Application → World Wide Web

Types

* Client Side FrameWork
  - → Used to Enhance the User Interface
  - → Front End
* Server Side frameWork
  - → Rules You've Set for the Website
  - → Back End

Examples: Ruby On Rails, Django, Angular Js, ASP.NET, METEOR

2. Web Hosting Service

That allow you to → Store your data and application

Examples:

* Amazon Elastic Compute Cloud
* Mosso

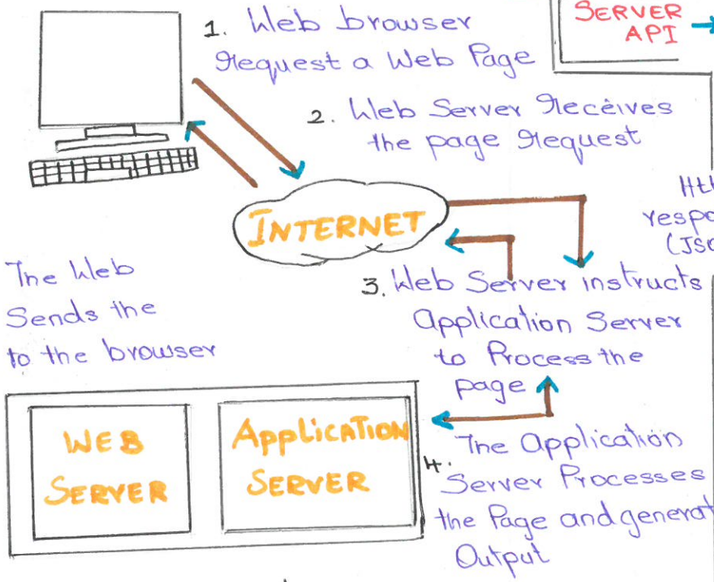
3. Proprietary Methods

↳ Connecting to the Cloud

Examples Of Own Infrastructure
↳ Microsoft, force.com

Web Applications

* Software or Mobile Application Which Catered Over the internet And Used in browser



Sample Web Applications

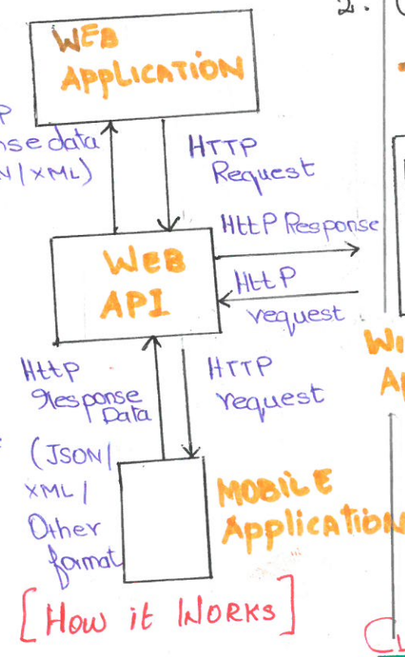
* Gmail Webmail Services
* Google Calendar Shared Calendering
* Google Talk Instant Messaging and Voice OVER IP
* Start Page for Creating a Customizable home page On a Specific domain

Web APIs [Application Programming Interface]

1. What is an API ?

— API is a Software Intermediary that allows two applications to talk to Each Other.

WEB API → Application Programming interface → WEB
BROWSER API → Extend the functionality → WEB browser
SERVER API → Extend the functionality → Web Server



[How it Works]

2. API Creators

* Google Gadgets
* Google Data Api's
* Partnership
* GoGrid
* Apex

* Some providers are huge And fill an Entire Warehouse

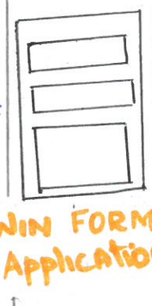
1. The Basics

* Niche-Oriented: Store Just Email Or digital Pictures
* While, Other store any type Of data
* Some providers Are Small

Web Browers

Takes you Anywhere on the Internet

1. Firefox
   * Smaller, faster, More Secure
   * Protects from —
   - Cross-Site tracking Cookies
   - Fringerprinters
   - Cryptominers
2. Chrome
   - Open Source Program
   - Accessing the World Wide Web
   - Running Web-based Applications
3. Internet Explorer
   - free graphical browser
   - Maintained by Microsoft for legacy Enterprise Uses.

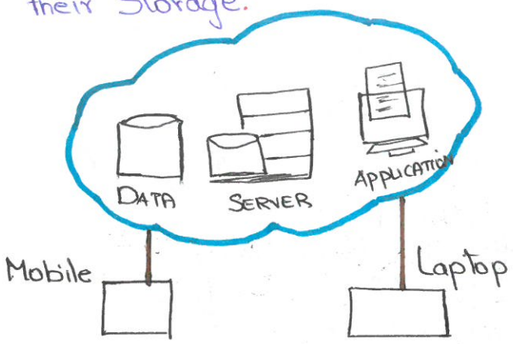


Cloud Storage

Store your Data → With Internet Access → get it from Any Location.

2. Storage As a Service

* Another name "Software as a Service"
* Third party Provider Rents Space for their Storage.



Advantage	Disadvantage
Cheaper upfront	Difficulty Chaning Provider
Easy to implement	Data Stored offsite
Intergrated With your Stack	Dependence on the Service infrastructure
minimal Commitment	Compromise Security

3. Provider As a Service

* Third Party Provider delivers hardward And Software tools Over internet

[Encryption and authentication are two Security Measures you can used to keep your data Safe on a cloud storage provider]

4. Advantages And Cautions

1. Cost 2. Availability 3. Scalability 4. Productivity 5. Upgradeability 6. Plateform Support	1. Security 2. Control 3. Reliability 4. Lock in features 5. Integration 6. Compatibility

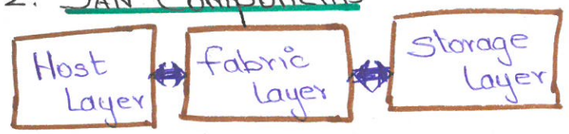
Storage Area Network (SAN)

- Network Of Storage device
- Accessed by multiple Server
- Provide a shared pool Of storage Space.

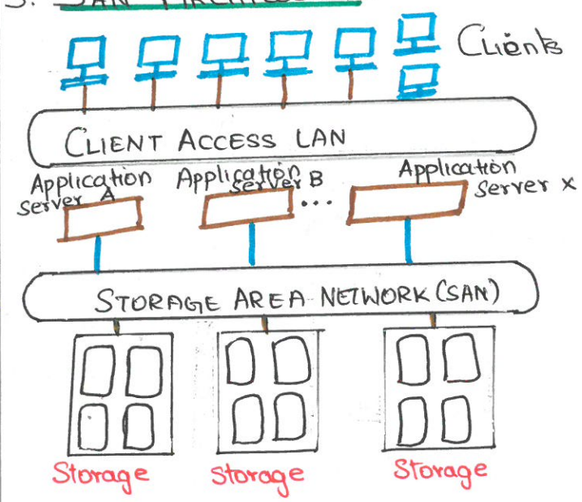
1. Types Of San

- Fiber Channel Protocol (FCP)
- Internet Small Computer System Interface (iSCSI)
- Fiber Channel Over Ethernet (FCoE)
- Non-Volatile Memory Express Over Fiber Channel (FC-NVMo)

2. San Components



3. San Architecture



Collaboration Within the Cloud

1. Benifits Of Cloud Collaboration

- Reduced investment, Scalability
- Improved Organization
- Higher levels Of Participations
- Large files Are Easy to Access
- Updates in Real time
- Improved Brainstoom

2. Top 5 Cloud Collaboration Providers/Features

- Access files Anytime through the Internet
- Real time Communication
- Setting Custom Permission levels
- Version Control
- Centralized file Storage

STANDARDS

1. Application

* The Protocols that Are Used to Manage Connection between both Parties.

A. Communication:

* Computers need a Common Way to Speak With One Another

(i) **HTTP** (HyperText Transfer Protocol)

* Stateless protocol
* transfer data between the Cloud and your Organisation

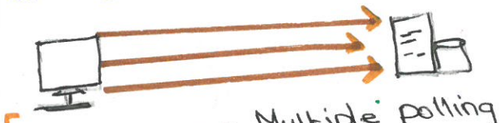
(ii) **HTTP 1.1**

* Eight Methods to describe how the desired Action to be Performed On the Server.

Request	Description
HEAD	Ask for the Response
Get	Request information from Server
Post	Submits data to be Processed to the Server
Put	Upload the Resource
Delete	Deletes the Specified Resource
TRACE	Adding or changing the Request.
Options	Returns HTTP Methods that the Server Supports for the given URL
Connect	Convert Request Connection to transparent Tcp/IP tunnel.

(iii) **XMPP** (Extensible Messaging and Presence Protocol)

* Allows for two Way Communication And Eliminates Polling.



[HTTP requires Multiple polling Events to Update Status from the Web browser]

[XMPP maintains a Connection between the Client and the Web Server]

B. Security:

* Securing your Cloud Sessions Can be Accomplished Via Encryption And Authentication

Encryption → Secure Sockets Layer (SSL)

Authentication → OpenID

(i) **SSL**: Encryption Link between Web Server and browser.

(ii) OpenID:

→ OpenID is free

→ lower Cost for Password and Account Management

→ Used to identify Web sites

2. Clients:

→ it means to Store and display information

(i) HTML:

Code in the form Of a hyperLink Which takes you to Another page.

How it Works:

Browser → Reads a file → translate to text → form the Client Wants to See

(iii) Dynamic HTML:

* Controlling the Standard HTML Codes

There Are four Parts to DHTML:

* Document Object Model (DOM)
* Scripts, * HTML
* Cascading Style Sheets (CSS)

(iii) JavaScript

* "Client-Side Web development"

Example:

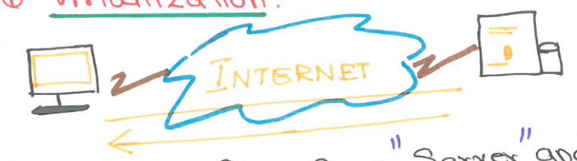
Opening or popping up new window And having Control Of the Size and Attributes Of the Window.

3. Infrastructure:

* The internet: (machine Running on a remote Server and displayed at your Organisation)

* Locally: (having your "Clients" Session Runs on a local Server and displayed at their desktops)

(i) Virtualization:



(Application Runs On a "Server" and Are "displayed" On the Client.)

(The Server Can be Local or on the Other Side Of the Cloud)

Benifits End Users by:

* Expanding Virtualization Solutions
* Expanding interoperability And Supportability
* Accelerated Availability Of new Virtualization Aware technologies

4. Service:

* Supports Machine to Machine Interaction Over a network

(i) DATA: Stored and Served up With Number Of Mechanism.

→ JSON (JavaScript Object Notation)

* Light Weight Computer data interchange formate

→ JSON Basic (Subset Of Javascript)

Driving forces

1. Popularity:
   * Software Vendors Love it
   * Enterprises Love it
2. Virtualization Benefits:
   * Easier for independent Software Vendors to adopt
3. Economic impact + Benifits:
   * Subscription based payment Model Of "Saas" Makes it more Appealing in these tough times

SOFTWARE AND SERVICES

(i) VENDORS:

* Microsoft
* Apple
* Adobe
* Google
* Salesforce.Com
* WeatherBug
* Dicentral's DiIntegrator EDI Solution

(ii) Providers:

* Creating your Own Software Plus Service deployments

(i) Adobe AIR:

* Adobe intergrated Runtime (AIR)
* Cross Operation System Application Runtime that allows developers to Use

*HTML/CSS, *AJAX, *Adobe flash
* Adobe FIEX to Extend (RIA)

(ii) Apple IPhone SDK [MOBILE DEVICE INTERGRATION]

* IPhone SDK Providers developers With a Rich Set Of (APIs)
* IPhone SDK Can download "free"
* "Third party developers are able to build native applications With rich set Of APIs.

Developing Applications

(i) Google:

* Getting Started guide
* Python Library documentation
* Examples Showing python Code for Accessing "force.com"
* "Testing" harness for the Provided Library.

* Wiki FAQ Page On developer."force.com" With best Practices And Latest tips And tricks

(ii) Microsoft:

* Tool Provided for developers Who Want to Write Application
* Runs Partially or Entirely In a Remote data Center.

(i) Live Service:

* Set Of building blocks Within the Azure Service Plateform
* Used to handle User data And application Resource.

(i) Microsoft SQL Service:

* Cloud as a Web based, distributed Relational database.
* Store and Retrieve "Structured", "Unstructured", "Semi-structured data".

(iii) Microsoft .Net Services

* Tool for developing loosely Coupled cloud-based application
* Access to Control to help "Secure application"
* On-Premises Environments to the Cloud

(iv) Microsoft Sharepoint Service And Dynamic CRM Service:

* Allow developers to Collaborate And build Strong Customer Relationships
* Tools like "Visual Studio", "Sharepoint", "CRM Capabilities"

Fundamentals of BIG DATA

Introduction :-

- Collection of data that is huge in volume
- Large size and Complexity
- Big data is also a data but with huge size.

Example :

- New York Stock Exchange
- Social Media

Importance of Big data

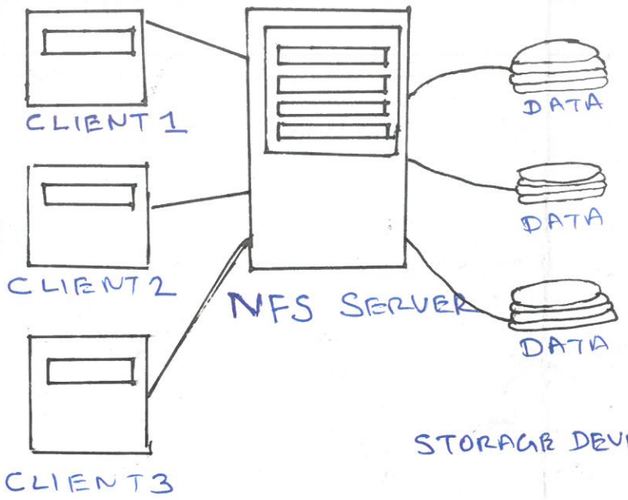
- COST SAVING
- CUSTOMER ACQUISITION
- TIME SAVING
- MARKETING INSIGHTS
- SOCIAL MEDIA LISTENING
- INNOVATION

Distributed file System :

- Distributed on Various file Servers and Locations.
- Access and Store isolated data
- Access files from any system.
- To share information and files
- Provides users access Control

Two Components of DFS

- Local Transparency
- Redundancy



Features of DFS

- Data sharing of multiple users.
- User mobility
- Location transparency
- Location independence
- Backups and System Monitoring.
- Replication
- Whole-file serving
- Whole-file Caching

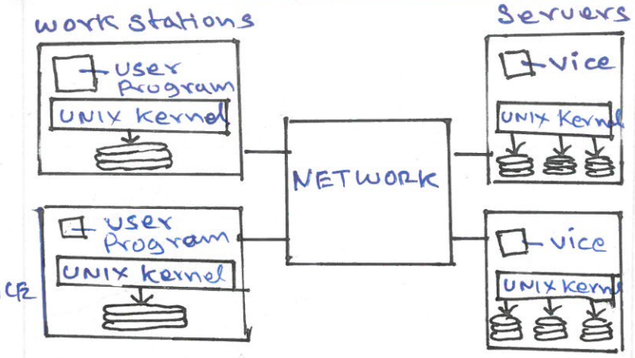
Applications of DFS

- Hadoop
- NFS (Network file System)
- SMB (Server message Block)
- Netware
- CIFS (Common Internet file system)

AFS → (Andrew File System)

- To support information Sharing on a large scale
- Transferring whole files between Server and client with caching.
- Uniform namespace
- Location Independent file Sharing
- Client-side Caching
- Secure authentication.

AFS Distribution of Processes

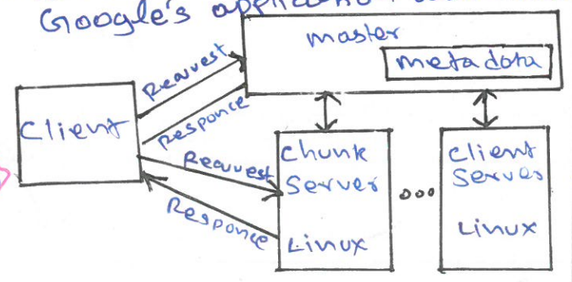


Advantages :

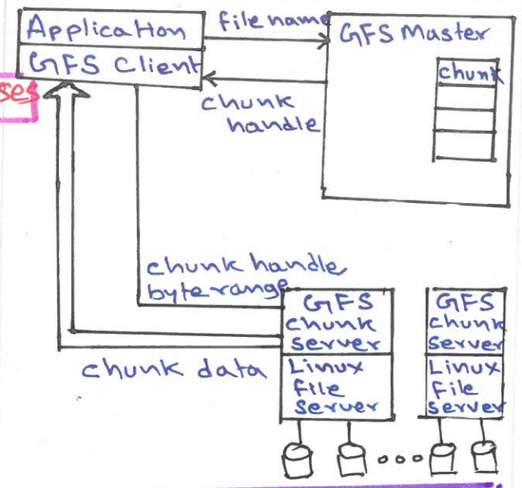
- Allows the User to access and store the data
- To improve the access time, Network efficiency and availability of files.
- provides the transparency of data.
- Permits the data to be Shared remotely.
- Helps to change the amount of data and exchange data.

THE GOOGLE FILE SYSTEM (GFS)

- Designed for System-to-System interaction, and not for user to system interaction.
- The chunk Servers replicate the data automatically.
- It can be used by Google's service.
- Driven by Observations of Google's application workloads.



GFS Architecture



Structured and unStructured

Predefinal Formatted data

Examples :-

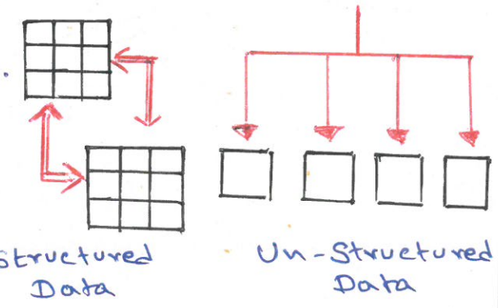
- Names
- Dare
- Address
- Geo Location
- Credit Card Number

Unstructured data

Example :-

- E-Mail
- social media posts
- IOT sensor data
- Satellete imagery

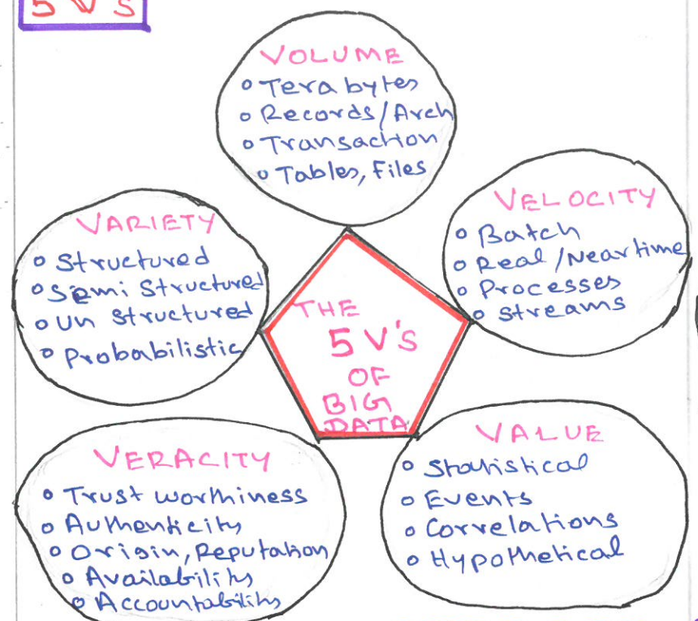
SQL Vs NoSQL



Big data and its importance

- Cost Savings
- Understand the Market Condition
- Time-Saving
- Social Media Listening

5 V's



- Volume (Huge amount of Data)
- Velocity (High speed of accumulation of data)
- Variety (Heterogeneous Sources) data)
- Veracity (Inconsistencies and uncertainty in data)
- Value (Worth of data)

Drivers of Big data

Big data has quickly become one of the most desired topic in industry.

- The Digitization of society
- The Drop in technology costs
- Connectivity through cloud computing
- Increased knowledge for Data Science
- Social media Application
- The rise of Internet of Things

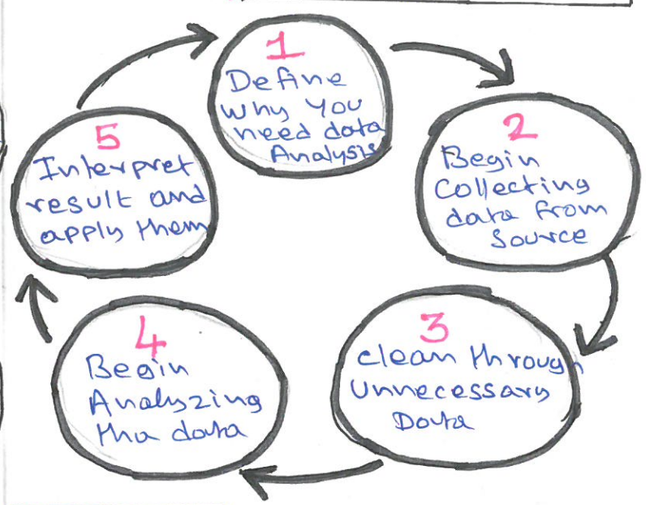
Big data analytics

Introduction

Analysis of large volums of diverse data sets, using advanced analytic techniques.

Five different steps of the big data analytics process.

- Define the question
- Collect the Data
- Clean the Data
- Analyze the Data
- Visualize/share result



Hadoop

- Open Source Framework
- Store and Process Big Data
- Distributed Environment
- cluster of computers using Programming models
- Scalable

Cassandra

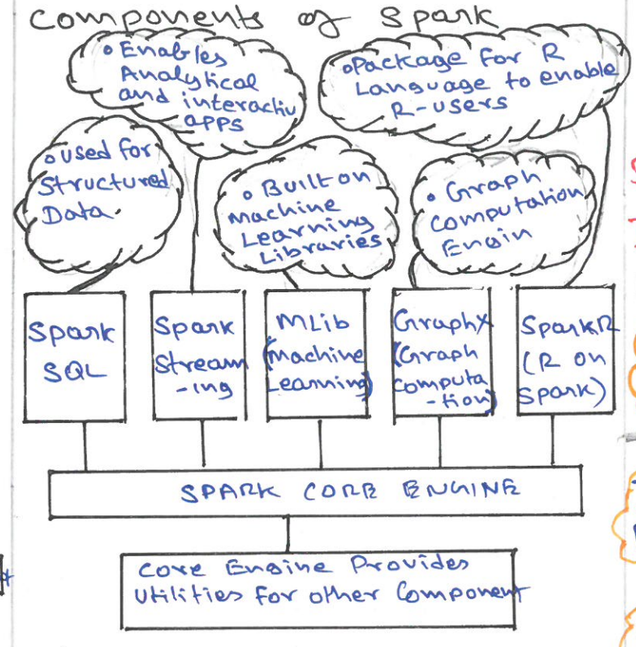
- Open Source
- Distributed and Decentralized
- Distributed Storage System
- Managing large amounts of Data

⇒ Open-Source availability
⇒ Distributed Foot Print
⇒ Scalability.
⇒ Cassandra Query Language
⇒ Fault tolerance
⇒ Schema free

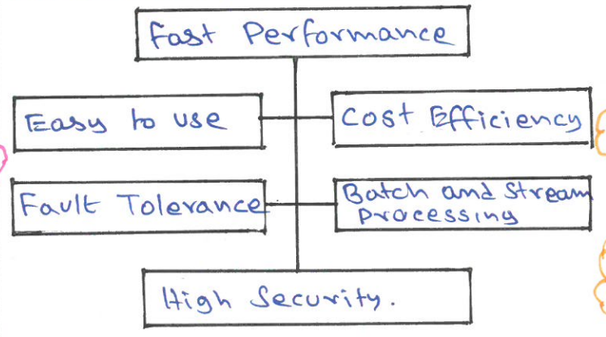
Spark

- Spark is a real time data analytics in distributed Computing
- It executes in memory to increase speed of data Processing.
- Faster for large scale data.
- It requires high processing power
- Resilient Distributed Data (RDD) is data Structure of spark

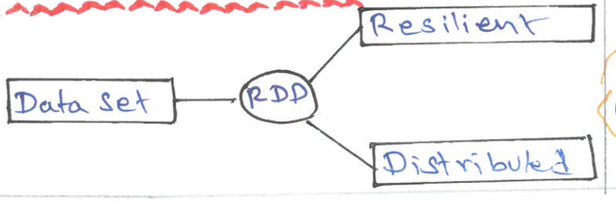
components of Spark



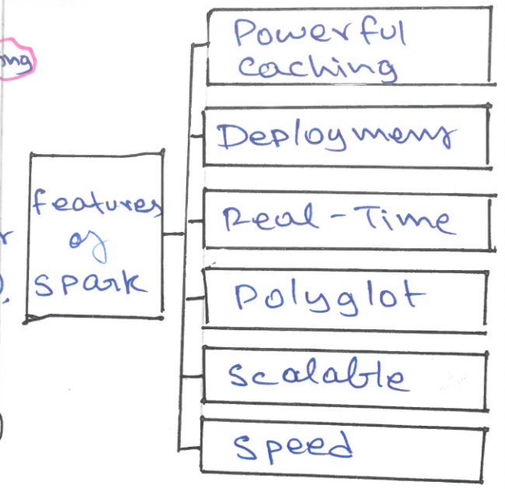
Advantages of Spark



RDD in Spark



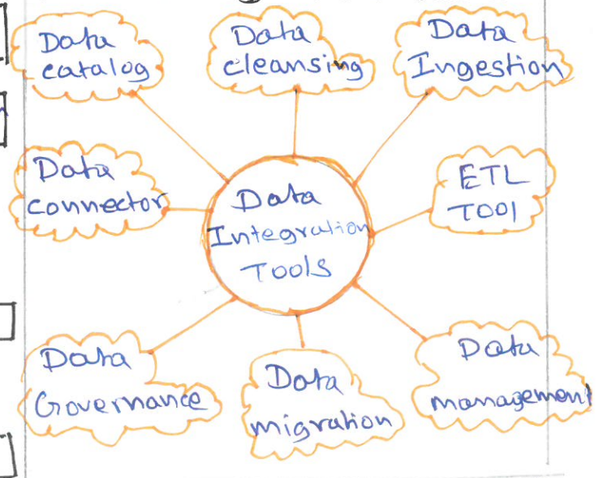
Features of Spark



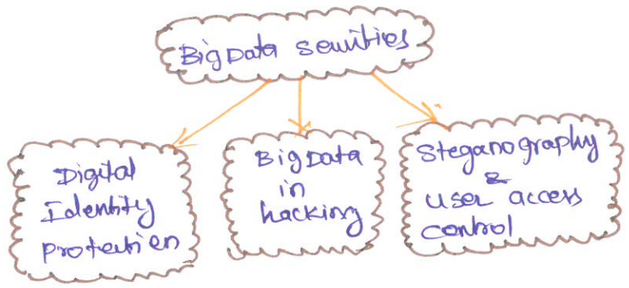
Data Appliances and Integration Tools

Benefits of data Integration

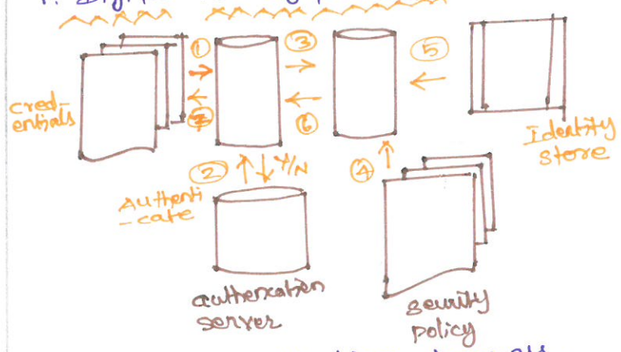
Data Integration Tools



BIGDATA SECURITIES :-

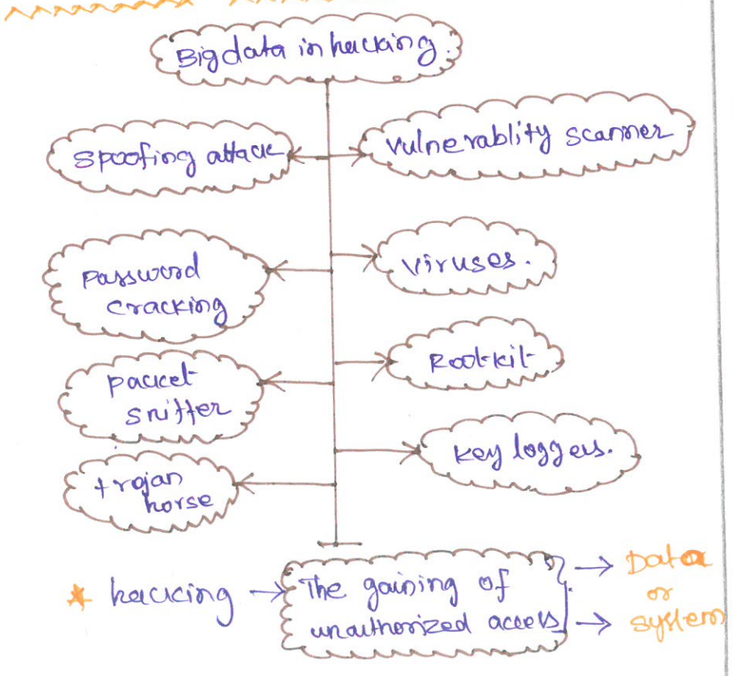


i. Digital identity protection



Authentication process.

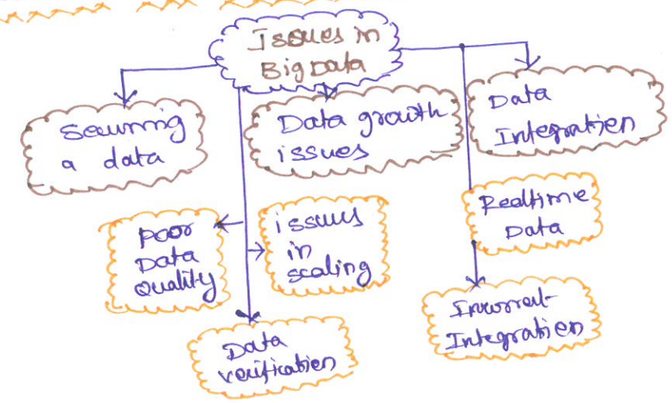
ii. Big Data in Hacking



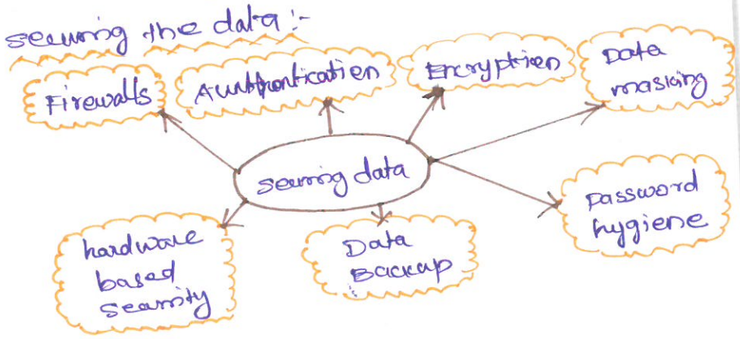
iii. steganography.

Technique of hiding secret data -

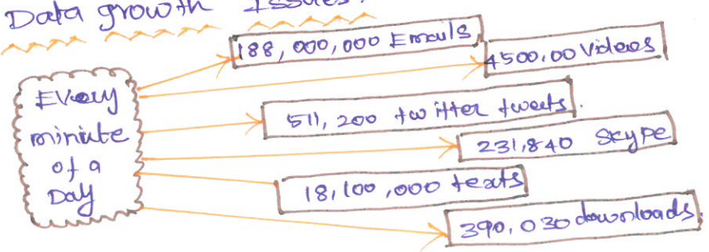
ISSUES IN BIG DATA :-



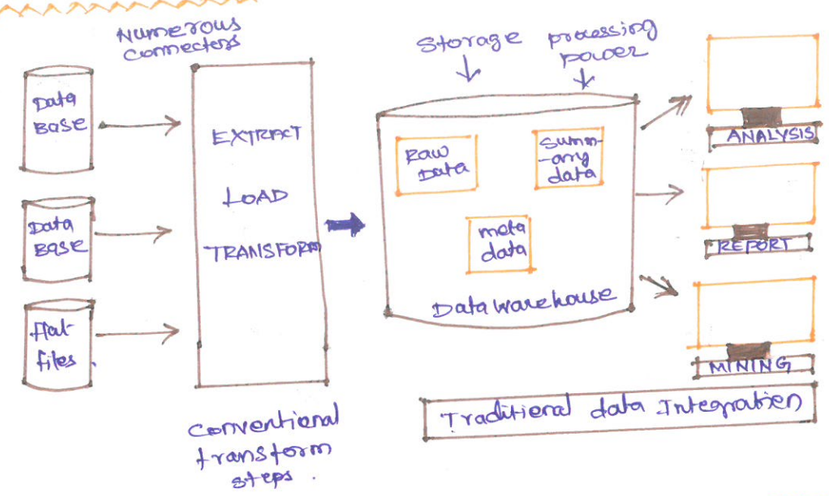
i. securing the data :-



ii. Data growth Issues :-



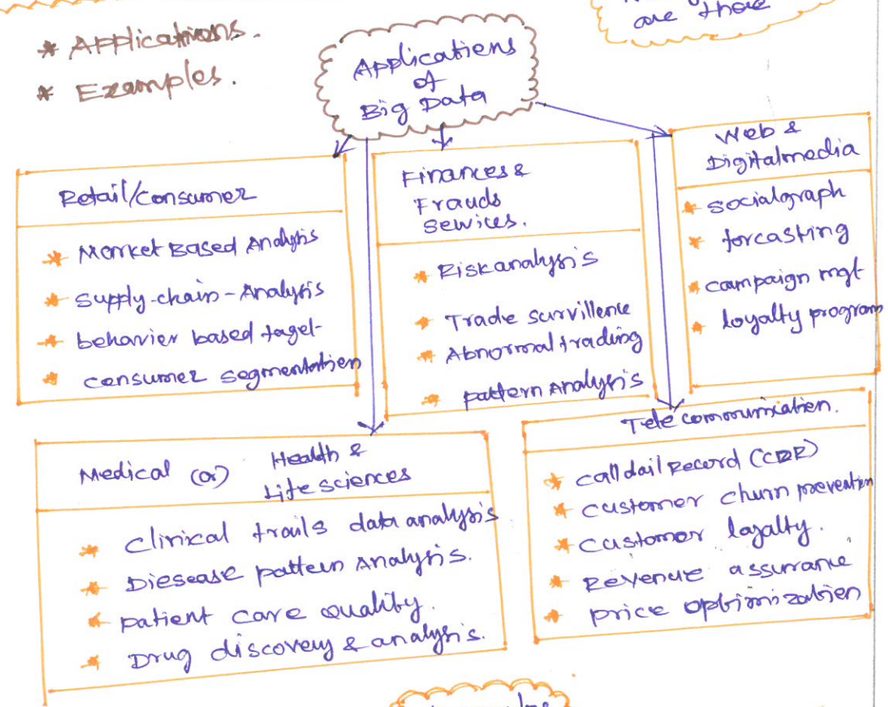
iii. Data Integration :-

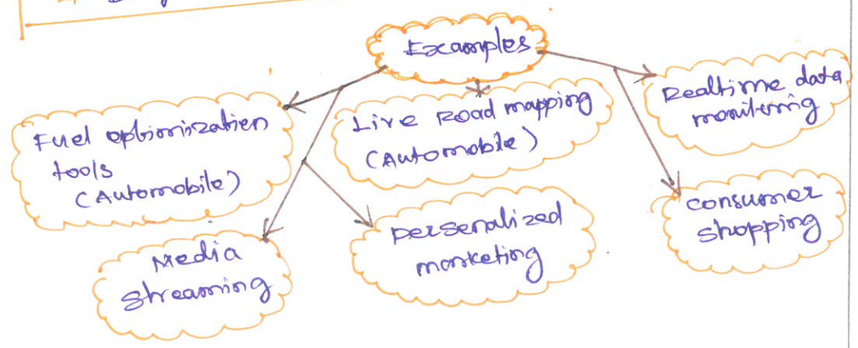


BIGDATA APPLICATIONS :-

* Applications.
* Examples.

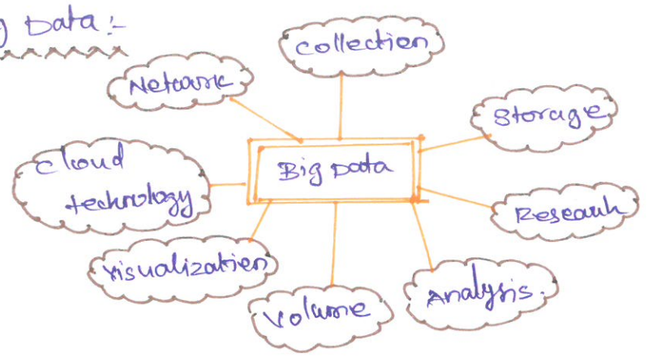
many applications are there





HADOOP AND MAPREDUCE ARCHITECTURE

Big Data :-

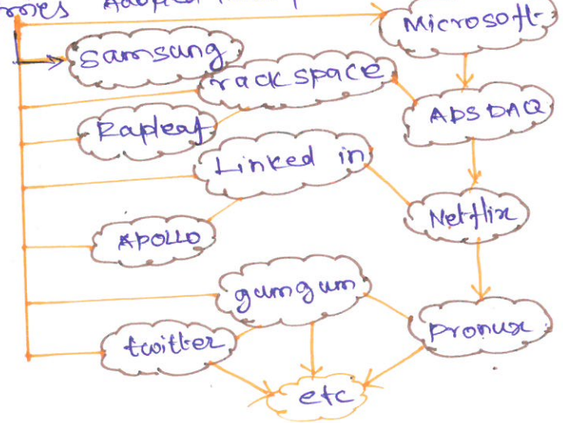


Apache Hadoop & Hadoop Echo system :-

Hadoop :- scalable fault-tolerance distributed system for → Data storage, → processing

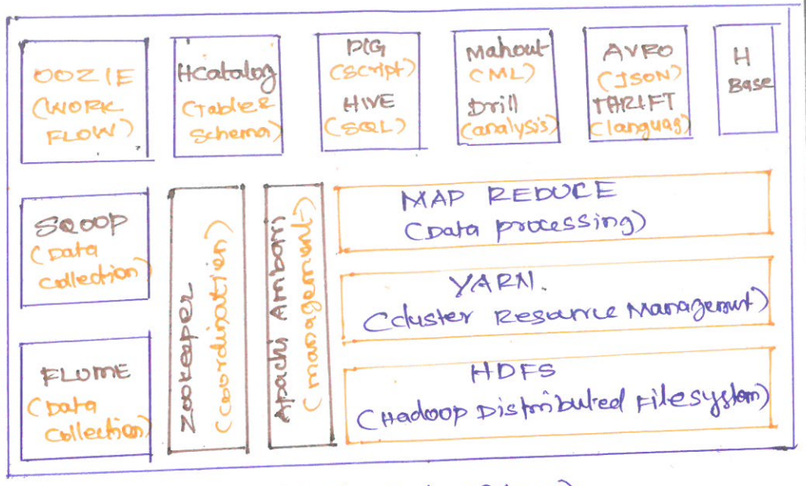
* It operates → Structured data, → unstructured data

* Industries Adopted Hadoop

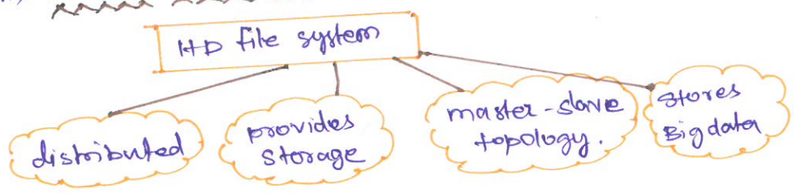


i. Hadoop ECO System :-

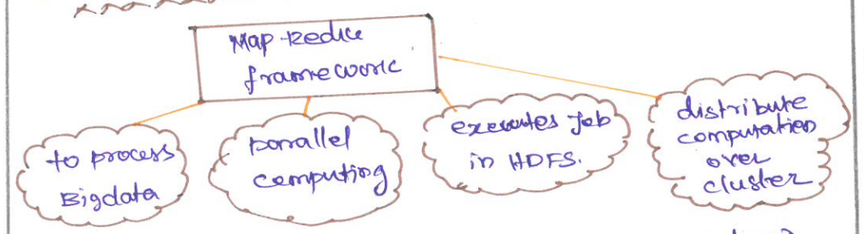
* Hadoop Distributed File System (HDFS)
* Map - Reduce (processing layer)
* YARN - (Resource management layer)
* others (spark, pic, zookeeper ..etc)



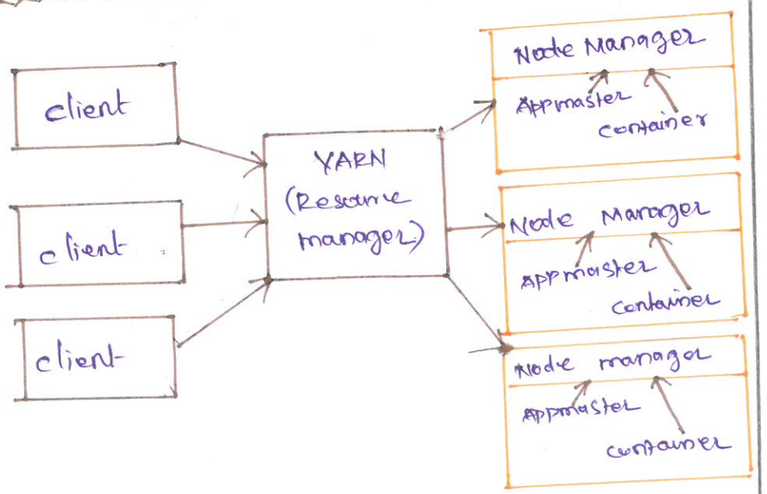
ii) Hadoop distributed File System (HDFS)



(ii) Map - Reduce :-

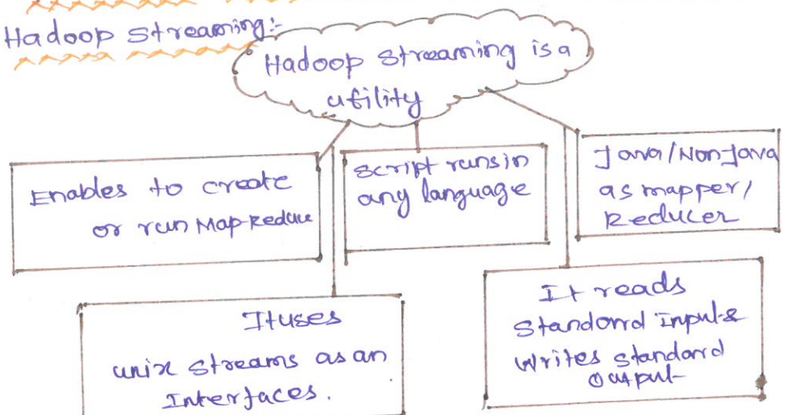


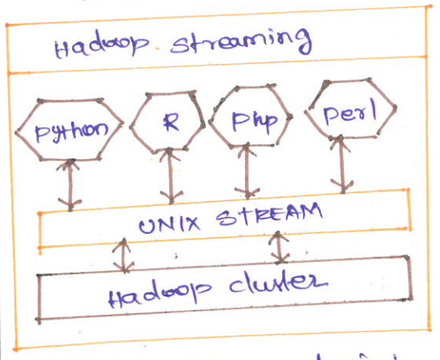
iv) YARN (Yet Another Resource Negotiator)



ANALYZING DATA WITH HADOOP STREAMING.

Hadoop Streaming :-

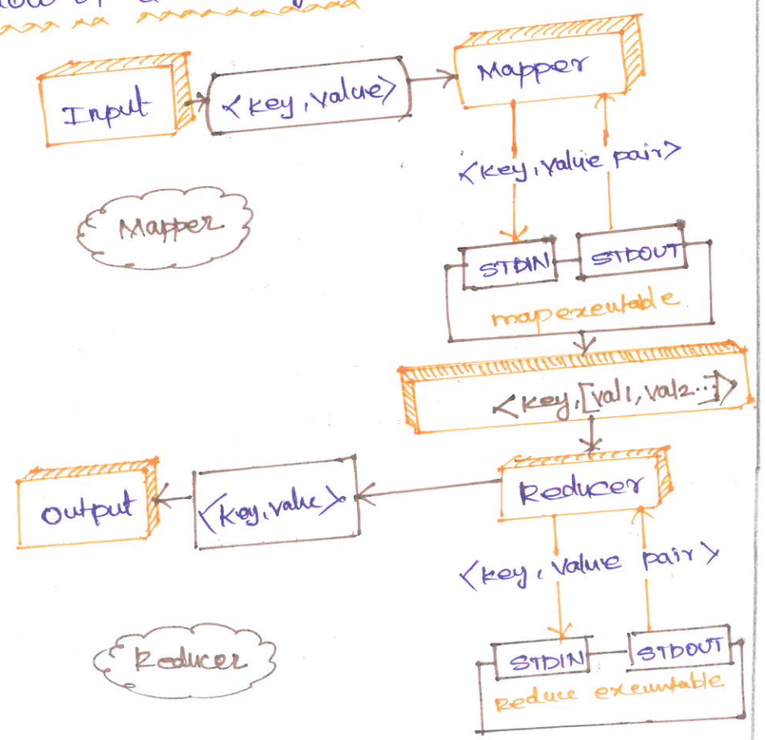




Advantages :-

* Availability
* Reduce development time
* faster conversion.
* Performance
* Learning.

Flow of data Analysis :-



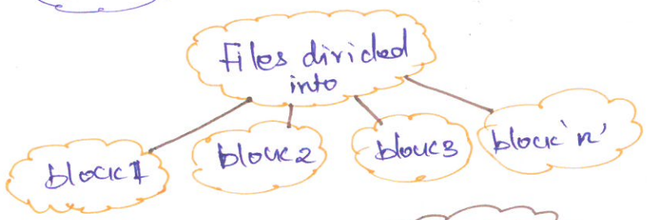
HDFS Concepts (Hadoop distributed filesystem)

Introduction:-

- provides storage for Hadoop.
- Master-slave topology.

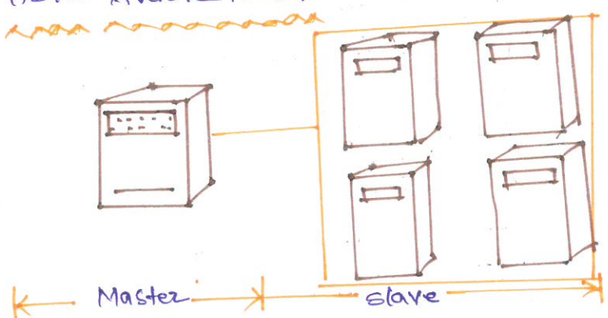
Master → High end machine

slave → Inexpensive computers

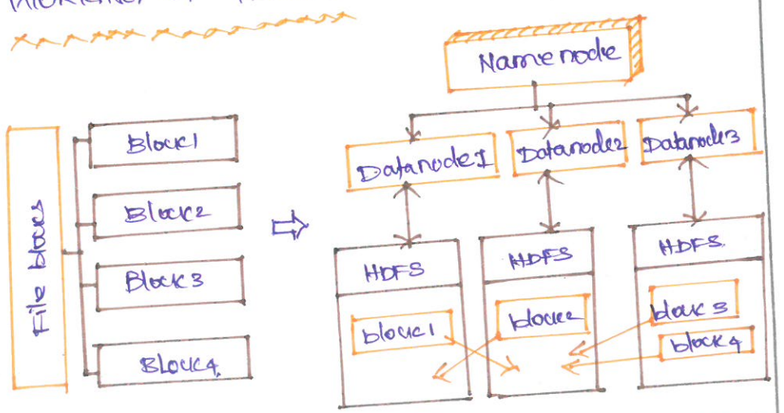


- Master Stores Metadata

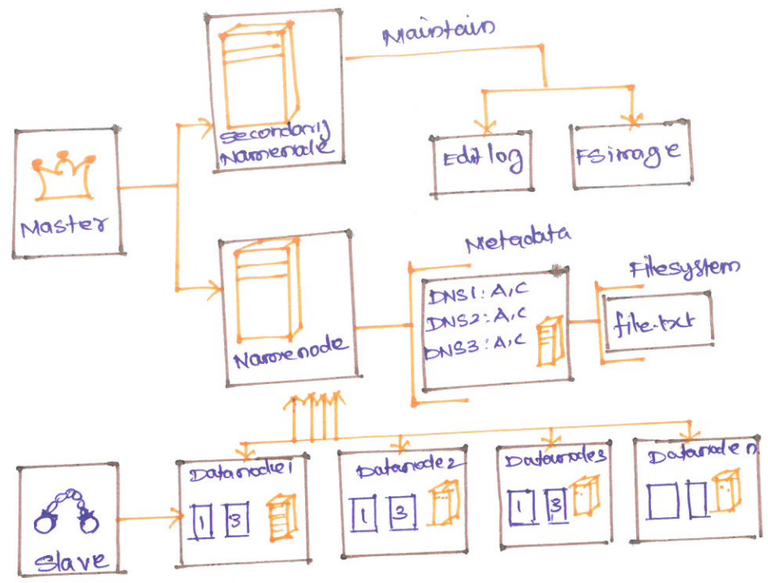
HDFS Architecture:-



WORKING OF HDFS:-

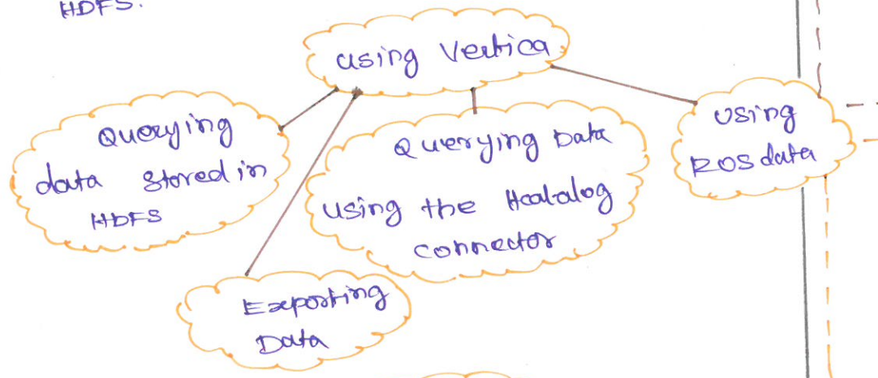


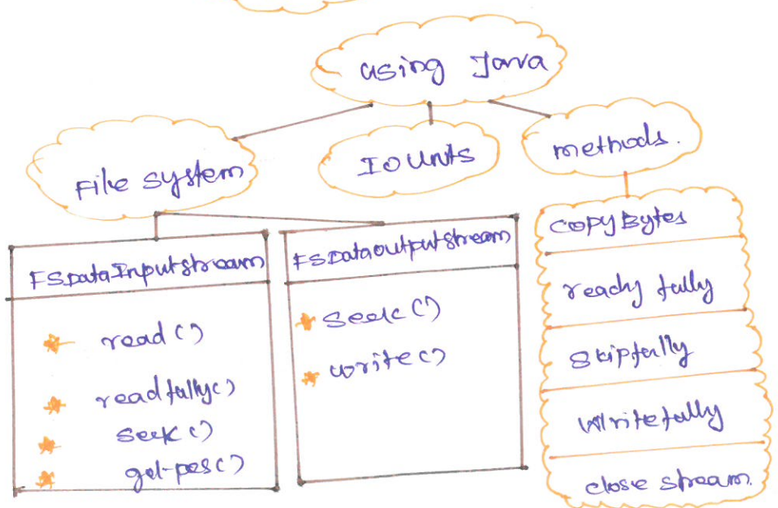
HDFS – Storage and Replication of blocks.



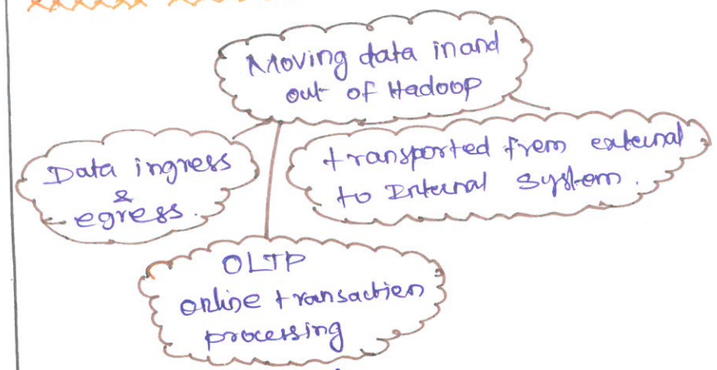
Interface to HDFS:-

- several ways to interact with data stored in HDFS.

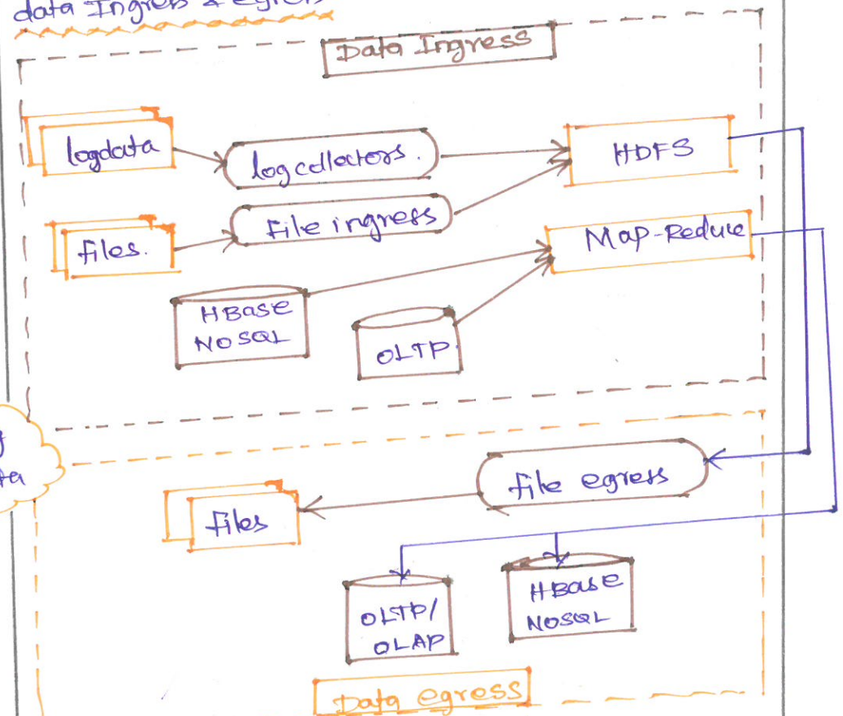




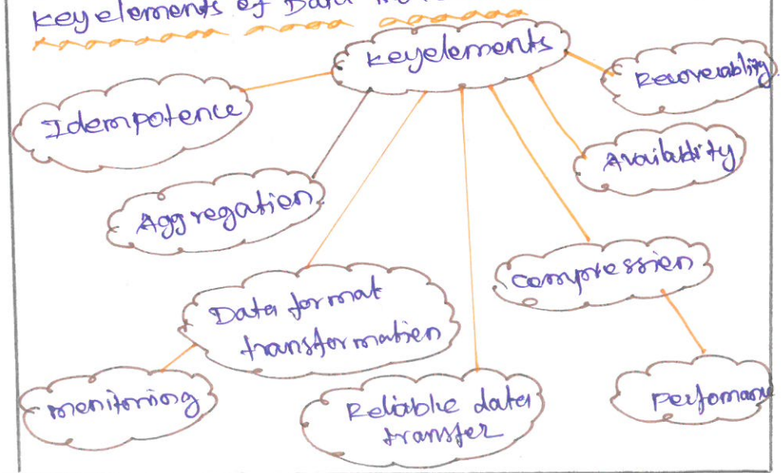
Moving Data in and out of Hadoop:-



data Ingress & egress



Key elements of Data movement:-



Introduction to MapReduce

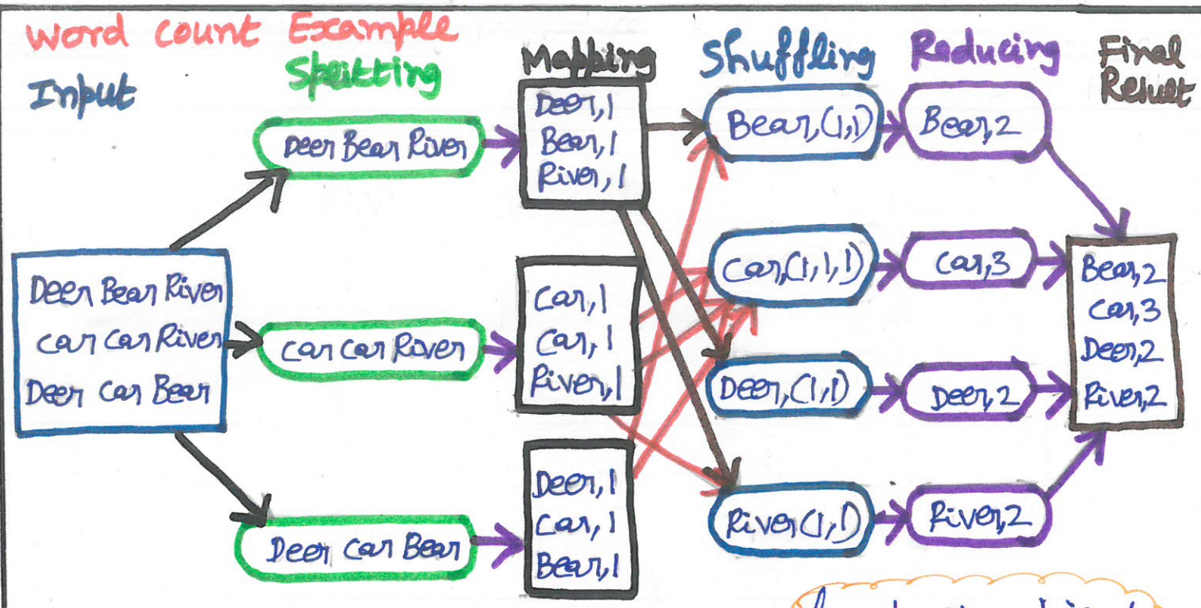
- Framework
- Process huge amounts of data
- MapReduce
  - Executes Parallely on large cluster of hardware
  - Executes Jobs on files in HDFS
  - Process Engine of the Apache Hadoop
- Application written basically in Java

MapReduce Framework (or) System

composed of 3 operations

- **Map**
  1) Worker Node applies the map function to the local data
  2) writes the output to a temporary storage
- **Shuffle**
  1) Worker node redistribute data based on the o/p keys
  Worker nodes process each group of output data, per key, in parallel
- **Reduce**
  3) Master node ensures that only one copy of the redundant input data is processed

Word count Example



Input Splits – Job is divided into fixed-size pieces.

Mapping – Data in each split is passed to a mapping function to produce output values.
→ Example – count a no. of occurrences of each word from I/p Splits. Prepare list form <word, frequency>

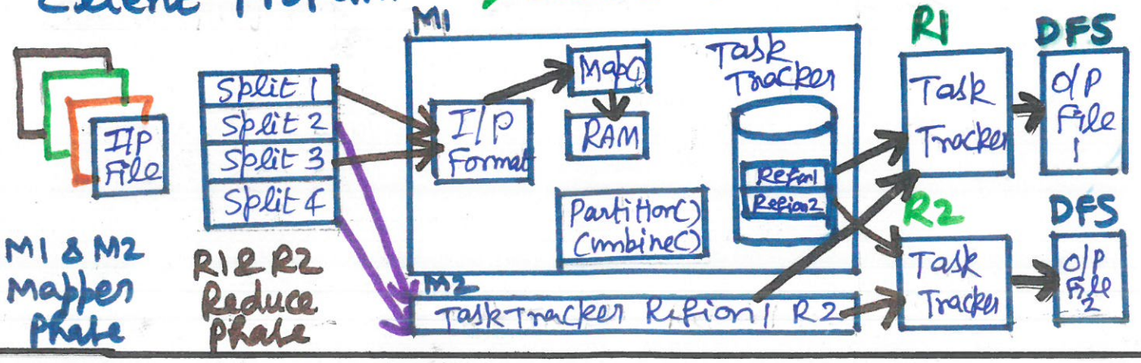
Shuffling – output of Mapping phase. Consolidate it.
→ Example – Same words clubed together with its frequency.

Reducing – o/p values from Shuffling phase are aggregated
→ Example – Summarizes the complete dataset.

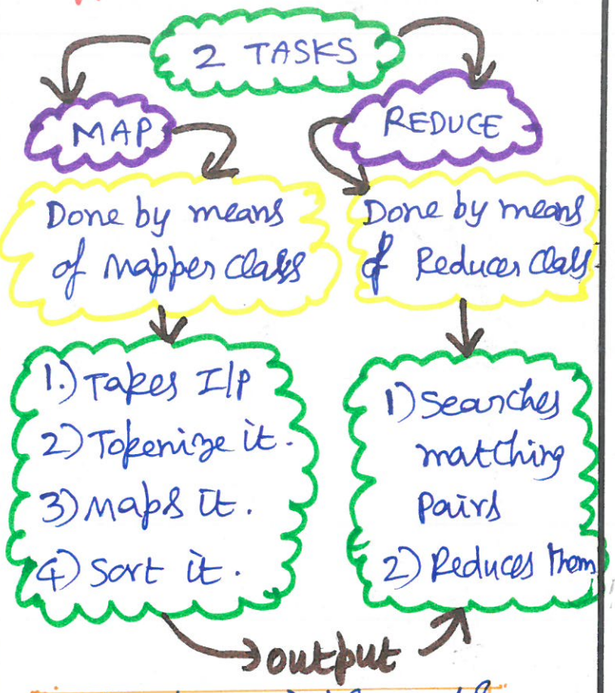
MapReduce Algorithm and Architecture

MapReduce Architecture

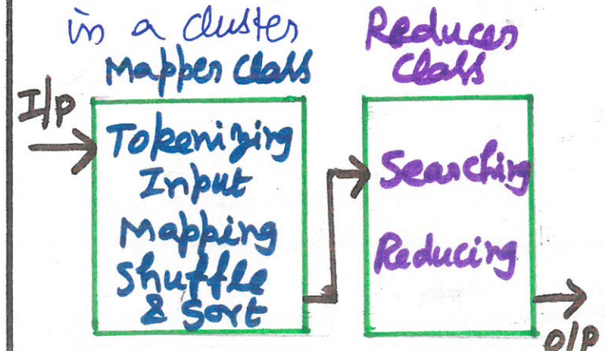
Client Program → Submit Job → Job Tracker



MapReduce Algorithm



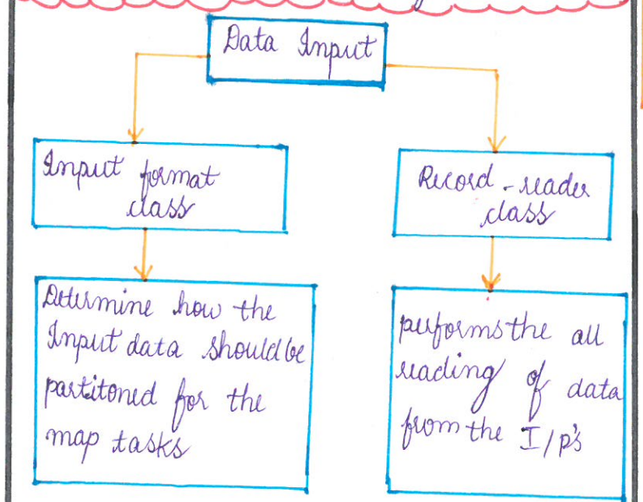
* MapReduce implements various mathematical algorithms [Sorting, Searching, Indexing]
* Divides a Task into Small parts.
* Assign them to multiple systems.
* It sends the Map & Reduce tasks to appropriate Servers in a cluster



Understanding Inputs and outputs of map Reduce :-

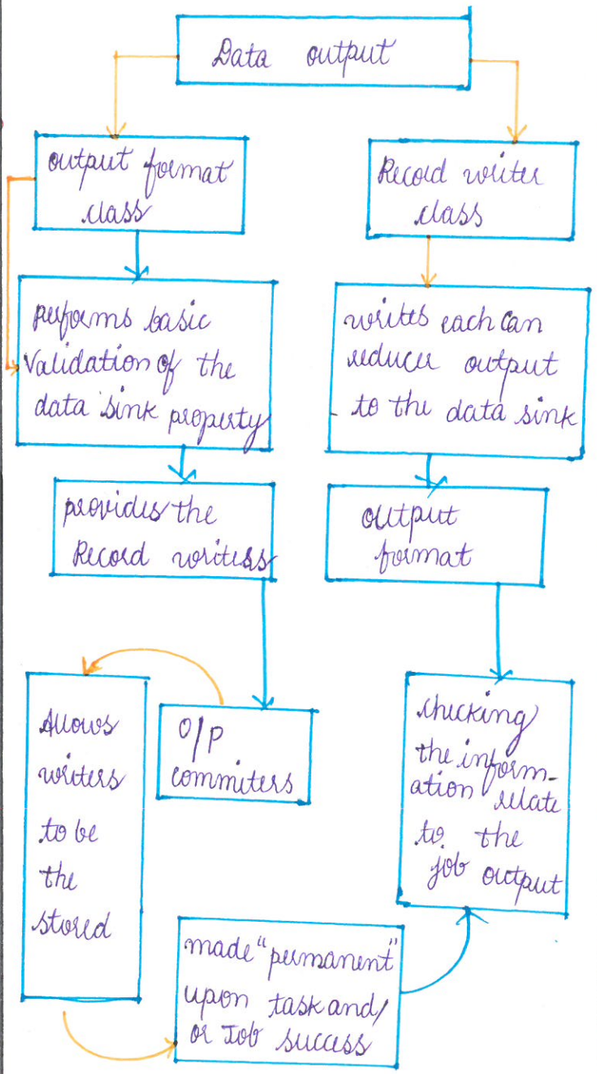
Inputs and outputs :-

* Your Data might be (XML files)
* XML (eXtensible Markup Language)
* Sits behind a no. of FTP Servers
* Text log lines of files → central web servers
* Lucene Indexes → HDFS
  Hadoop Distributed File System



Contract 2:- specify how the I/p data should be partitioned

Contract 3:- Indicate the record reader instance that should read the data from source



Anatomy of Map Reduce Job Run :-

Mapper, Shuffler & Reducer

Mapper

1) perform Map side operation (by you)
2) Write to in-memory buffer (by framework)

Shuffler

1) partitioner figure out which map output key goes to which Reducer (by framework)
2) Several unique key in one partition.
3) Specify partitioner class which implement partitioner ZD extraction process
4) sort first by partitioner ZD then by key value within partition (by framework)
5) call single combiner for each key of every partition. (by framework)
6) Spill to disk (also group by key and merge) if limit exceed (Default 100MB).

Reducer

→ Start read map O/P's from disk and from in-memory — (1)
→ Merge outputs — (2)

* Sort by partition ID and then by all the key — (3)
* Group by key — (4)
* call Reduce operation defined by you for each unique key — (5)

How Map Reduce Job is Running?

1) Map Reduce app submit Job to Hadoop client
2) client ask Resource manager to get app ID
3) copy job resources to HDFS :-
   a) checks the O/P specification of the job
   b) computes the I/P splits for the jobs
   c) copy Jar and throw error if needed.
4) submit Application
5) Resource Manager :-
   a) Allocate container on same node.
   b) Run application masker on that node.
6) Application master initialize Job.
7) Retrieve Input splits.
8) Allocate resources and start of container
9) step 10 Retrieve resource for map (or) reduce task.
10) step 11 Run map (or) reduce task.

Failures in classical mapreduce and YARN (Yet Another Resource Negotiator)

Types of Failures

Task Failure

Case 1: Child Task Failing

Map/Reduce Task throws a Runtime exception

Solution: The tasktracker marks the task attempt as Failed, freeing up a slot to run another task.

Case 2: Sudden exit of the Child JVM

Solution: The tasktracker notices that the process has exited & marks attempt as failed.

Task Tracker Failure:

- Failure of a Task Tracker is another failure mode
- Task Tracker fails by crashing, or running very slowly, it will stop sending heartbeats to the JobTracker.

JobTrackers notices Task Tracker that has stopped sending heartbeats remove it from its pool of TaskTrackers to schedule tasks on.

Task Trackers black listed by the Job Tracker

Jobtracker Failure

- Most serious failure mode
- Hadloop has no mechanism to deal it. Single point of Failure so, Job Fails.
- Low chance of occuring.
- YARN eliminates single point of failure.

Failures in YARN

Task Failure

Due to Runtime exceptions and sudden exits of the JVM

Application master Failure

Due to hardware Failures or Network Failures.

Node Manager Failure

- It will stop sending heartbeats to the resource manager
- Node manager will be removed from the resource manager's pool of available nodes

Resource manager Failure

- Neither Jobs nor task container can be launched
- Able to recover from crashes by using checkpointing mechanism to save it's state to persistent storage.
- New resource manager instance is brought up and it is recovered.

Job scheduling - Data serialization

Types of Scheduling

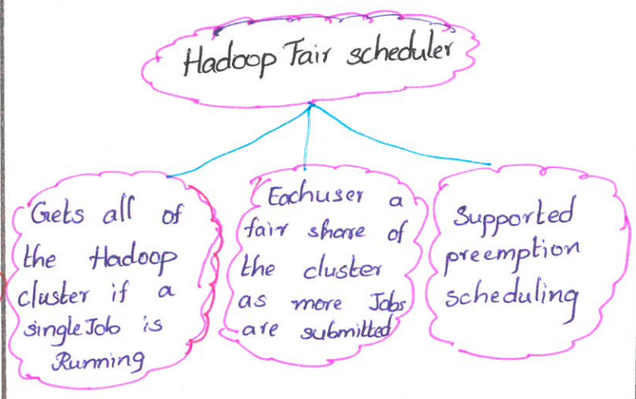
1. Hadoop FIFO Scheduler

Hadoop Job scheduling Algorithm + Job Tracker is the FIFO pulled

Jobtracker ⟹ Work queue ⟶ Job

- No concept of the priority or size of Job

2. Hadoop Fair scheduler



3. Hadoop capacity scheduler

- Multiuser scheduling simulates a separate mapReduce cluster along with FIFO scheduling

Data Serialization

A way of representing data in memory of as a series of bytes

Example : AVRO

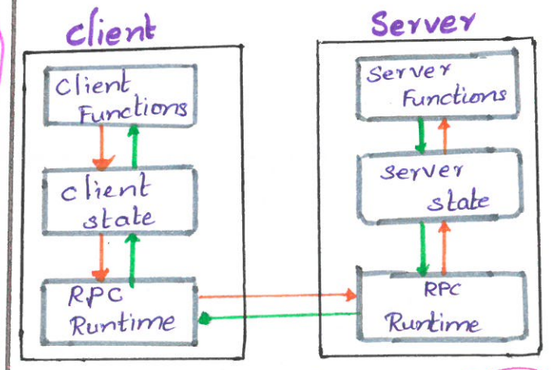
- Efficient data serialization Framework.
- Widely supported throughout Hadoop and its ecosystem.
- Using Sqoop, data can be imported to HDFS in AVRO and parquet File Format.

Remote procedure call (RPC)

- IPC communication

Used for client - server based application

A client has a request message that RPC translates and sends to the server



- Client stub called by the client
  - → Makes a system call to send messages to the server and puts parameters in the message
  - → Message sent from C → S by its OS
  - → Message is passed to the server stub by the server OS
  - → Parameters removed by server stub

Server procedure called by server stub.