**Mohammad R Islam 24/09/2020**

**Types of Cloud Service (Computing Infrastructures):**

* **IAAS: Infrastructure as a Service.**
* **PAAS: Platform as a Service.**
* **SAAS: Software as a Service.**
* **DAAS: Database as a Service (New)**

**SAAS (Google Dock): all ready!**

Hosted Application/Apps

Development tools, database management, business analytics

Operational Systems

Servers and storage

Networking firewalls/security

Data center physical facility/building

**PAAS**

Development tools, database management, business analytics

Operational Systems

Servers and storage

Networking firewalls/security

Data center physical facility/building

**IAAS**

Servers and storage

Networking firewalls/security

Data center physical facility/building

**IAAS** needs extra services to compare **SAAS.**

**Google Cloud offers**:

**AWS** (UK, widely used)

Google Cloud Platform**-GCP** (Medium range)

Azure (mostly Canada)

1. **Data Type**

* **OLTP:** Online Transitional Processing enables real-time execution of large numbers of database transactions by large numbers of people, typically over the Internet. The common defining characteristic of any database transaction is its atomicity (or indivisibility) – a transfer either succeeds as a whole or fails (or canceled). It cannot remain in a pending or intermediate state.

**Example:** ATM, Credit card payment process, Order entry, online booking, record keeping, etc.

* **OLAP:** Online Analytical Processing is software for performing multidimensional analysis at high speeds on large volumes of data from data warehouse, data mart, or some other unified, centralized data store.

**RDBMS:** Oracle, MySQL, PostgreSQL, and SQL Server. These are schema and not scalable based.

**NOSQL:** Cassandra, MongoDB. These are No-Schema and Scalable based. Example: JSON, key-value, etc.

1. **BIG DATA**

High velocity, volume and variety (3V) data.

**HADOOP:** Distributor Computing Framework (DCF). This fault tolerance and high availability based. IN this process, cluster is made up of nodes. The default replication factor is 3 nodes. There are two types:

**Processing:** Map Reduce, YARN (Yet Another Reduced Network)

**Storage:** HDFS

**File Systems**

Control how data is stored and retrieved

Metadata about the files and folders

Permissions and security

Manage storage space efficiently.

**Windows:**

**NTFS** (16 EB file/volume)

**Fat32** (4GB volume/file limit)

**Linux:**

ext3 (2 TB file, 32 TB Volume)

ext4 (16 TB File limit, 1 EB Volume limit

XFS (8 EB File limit 8 EB Volume limit)

**MacBook (Hierarchical File System (HFS)**

HFS: 2GB file 2 TB Volume limit

HFS+: 8EB file 8 EB volume limit

File default maximum block size of ext4 is 4KB

Hadoop2.X: HDFS default block size is 128MB

Hadoop1.X: HDFS default block size is 64MB

**Pros & cons**: Hadoop file format is continuous because it is easy to fetch data because the data is placed closed to each other where as in ext4 it in not continuous because the data is present in different blocks and difficult to read and write.

**Hadoop Components (nodes):**

**Name Node (NN):** Metadata (data about data)

**Secondary Name Node (SNN):** Check pointing, send metadata information to NN if NN goes wrong/lost. It edit and merge the FS image file (in the memory) and Disc file, both are copy files, into one file and end to lost NN. SNN executes every 5 minutes!

**Data Node (DN):** Physical data

**Resource Manager:** resource allocation.

**Node Manager:** Task executor. With the help of Mapper programs, sorting/shuffling and Reduce, TM provide information to the NN.

**Hadoop Architecture:**

**Cluster**

**Master Node:** NN, SNN

**Resource manager (Cluster manager)** liesin the middle.

**Slave Node:** Data Node, Node Manager

Example: 130MB data (replication factor=3):

130MB data needs 2 blocks (128MB and 2MB).

N1 N2 N3

**YARN**: In Hadoop 1.X

Resources Manager was called Job Tracker

Node Manager was called Task tracker

Hadoop: NN, DN, SNN

**Hadoop-High availability:**

Quorum Journal Node (QJN)

Meta data

N1

N1

Lock

DN2

DN3

DN1

Zookeeper

(ZK)

N1

Coordinating

System

Stand by

NN2

NN1

Active

When NN1 goes down then everything might lose. To prevent this, Hadoop2.X has a stand by Name Node (NN2), which also store Metadata. There is a Lock, which is a ***Failover Controller***. Zookeeper (ZK) component is a coordinating system; which keeps the Metadata associated with the different Data Nodes (DN1, DN2, DN3..). Failover controller regularly checks which Name Node is active. If any Name node fails then the lock shifts to the standby Name Node and QJN keeps the backup for NN1 or NN2. This way we can prevent from losing any metadata. In the mean time, Hadoop Admin replaces the old NN within a Millisecond time scale.

**Hadoop-Security:**

**KERBEROS:** Greek word means “Dog with three heads”

**OR**

Ticket Generating

Server

Access Hadoop

With encrypted

New ticket

Authentication

Server

Client or

User Name

**Hadoop Tools:**

1. **SQOOP:** Injecting RDBMS, HDFS
2. **Hive:** Data warehousing (DW) in Hadoop (OLAP)
3. **FLUME:** Near real time data injecting tool
4. **PIG:** Transformation and cleaning
5. **HBASE:** OLTP (acid transformation)

**Hive Example:**

**1TB: 128 MB (Mappers) –800 mappers**

**External table (SQL**)

IT.csv

iTB

**IT**

**csv files:**

HR.csv

iTB

HR

Finance.csv

iTB

**Finance**

**Data types:**

1. Structured: Schema (RDBMS (2D))
2. Semi-Structured: JSON, key-value, xml,csv
3. Unstructured: mp3, audio,txt

Finally

**HDFS in blocks**