Assignment11

Nosql Databases:

NoSQL is an approach to database design that can accomodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.

Types of NOSQL:

There are four primary classifications for NoSQL architectures-

Key-value stores-Key-value stores, or key-value databases, implement a simple data model that pairs a unique key with an associated value

Document databases- ocument databases, also called document stores, store semi-structured data and descriptions of that data in document format. They allow developers to create and update programs without needing to reference master schema.

Wide-column stores-Wide-column stores organize data tables as columns instead of as rows. Wide-column stores can be found both in SQL and NoSQL databases. Wide-column stores can query large data volumes faster than conventional relational databases.

Graph stores-Graph data stores organize data as nodes, which are like records in a relational database, and edges, which represent connections between nodes. Because the graph system stores the relationship between nodes, it can support richer representations of data relationships.

CAP THEOREM:

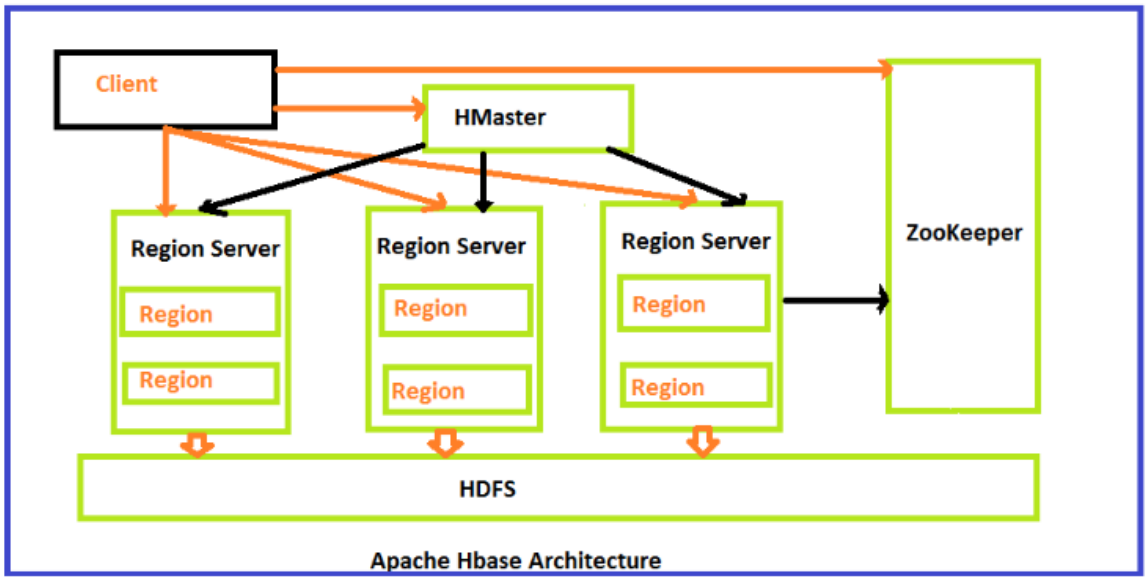
Consistency - All the servers in the system will have the same data so anyone using the system will get the same copy regardless of which server answers their request.

Availability: Even if its not the latest data or consistent across the system the system will always

respond.

Partition tolerance: System continues to operate on a whole even if individual servers fail or cannot be reached.

HBASE ARCHITECTURE:



HMaster:

HMaster is the implementation of Master server in HBase architecture. It acts like monitoring agent to monitor all Region Server instances present in the cluster and acts as an interface for all the metadata changes. In a distributed cluster environment, Master runs on NameNode. Master runs several background threads.

The following are important roles performed by HMaster in HBase.

Plays a vital role in terms of performance and maintaining nodes in the cluster.

HMaster provides admin performance and distributes services to different region servers.

HMaster assigns regions to region servers.

HMaster has the features like controlling load balancing and failover to handle the load over nodes present in the cluster.

When a client wants to change any schema and to change any Metadata operations, HMaster takes responsibility for these operations.

HRegions Servers:

When Region Server receives writes and read requests from the client, it assigns the request to a specific region, where actual column family resides. However, the client can directly contact with HRegion servers, there is no need of HMaster mandatory permission to the client regarding communication with HRegion servers. The client requires HMaster help when operations related to metadata and schema changes are required.

HMaster can get into contact with multiple HRegion servers and performs the following functions.

Hosting and managing regions

Splitting regions automatically

Handling read and writes requests

Communicating with the client directly

HRegions:

HRegions are the basic building elements of HBase cluster that consists of the distribution of tables and are comprised of Column families. It contains multiple stores, one for each column family. It consists of mainly two components, which are Memstore and Hfile.

HBASE VS RDBMS:

Hadoop and RDBMS are varying concepts of processing, retrieving and storing the data or information. While Hadoop is an open-source Apache project, RDBMS stands for Relational Database Management System. Hadoop framework has been written in Java which makes it scalable and makes it able to support applications that call for high performance standards. Hadoop framework enables the storage of large amounts of data on files systems of multiple computers. Hadoop is configured to allow scalability from a single computer node to several thousands of nodes or independent workstations in a manner that the individual nodes utilize local computer storage CPU processing power and memory.

	RDBMS
HBASE	
• Row oriented	Coloumn oriented
• Fleixible schema	Fixed

- good with sparse tables — Not optimized
- Joins using MR — optimized for joins
- tight interaction with MR — No tight interaction
- Horizontal Scalability — Hard to scale

TASK2

```
-rw-r--r--   1 acadgild supergroup         41 2018-09-19 04:34 /Hbase/bulkdata.tsv
[acadgild@localhost Hbase]$
```

```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Hbase]$ hadoop fs -cat /Hbase/bulkdata.tsv
18/09/19 04:44:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... usir
asses where applicable
1 Amit 4
2 Girija 3
3 jitin 5
4 swati 3

You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Hbase]$
```