Lesson 9
Cassandra Databases

Apache Cassandra DBMS

- Contains a set of programs
- They create and manage databases
- Functions (commands) for querying the data and accessing the required information

Apache Cassandra

- Has the distributed design of Dynamo
- Written in Java
- Big organizations, such as Facebook, IBM, Twitter, Cisco, Rackspace, eBay, Twitter and Netflix have adopted Cassandra

Cassandra

- Basically a column family database that stores
- Handles massive data of any format including structured, semi-structured and unstructured data
-

Cassandra Functions

- For viewing, querying and changing (update, insert or append or delete), visualizing and perform transactions on the DB.

Cassandra Characteristics

- open source
- scalable
- non-relational
- NoSQL
- Distributed (peer-to-peer distribution in the system across its nodes)

Cassandra Characteristics

- column based
- Decentralized
- Replication and thus fault tolerant
- tunable consistency

Cassandra

1. Maximizes the number of writes –
2. Maximizes data duplication
3. Does not support Joins, group by, OR clause and aggregations
4. Uses Classes consisting of ordered keys and semi-structured data storage systems

Cassandra

5. Is fast and easily scalable with write operations spread across the cluster

The cluster does not have a master-node, so any read and write can be handled by any node in the cluster

6. Is a distributed DBMS designed for handling a high volume of structured data across multiple cloud servers

Cassandra

- Components at Cassandra (Table 3.13)
- Scalability
- Transactions support
- Data Types (Table 3.14)

Cassandra Data Model Components

(i) Cluster: Made up of multiple nodes and keyspaces

(ii) Keyspace: a namespace to group multiple column families, especially one per partition

Cassandra Data Model Components

(iii) Column: consists of a column name, value and timestamp and

(iv) Column-family: multiple columns with row key reference. Cassandra does keyspace management using partitioning of keys into ranges and assigning different key-ranges to specific nodes.

Cassandra Hadoop Support

- The nodes in the Cassandra cluster can read data from the data in the Data Node in HDFS as well as from Cassandra
- A client application sends the MapReduce input to Job Tracker/Resource Manager

Summary

We learnt :

- Cassandra Column Family Data Model
- Has the distributed design of Dynamo
- Written in Java

Summary

We learnt PIG Latin:

- Decentralized
- Replication and thus fault tolerant
- Tunable consistency
- Hadoop support

End of Lesson 9 on **Cassandra Databases**