**ASSIGNMENT NO. 8**

**Aim**

Design database schemas and implement min 10 queries using Cassandra

**Objective**

Understand, design and implement queries in Cassandra (Column Oriented Database) .

**Theory**

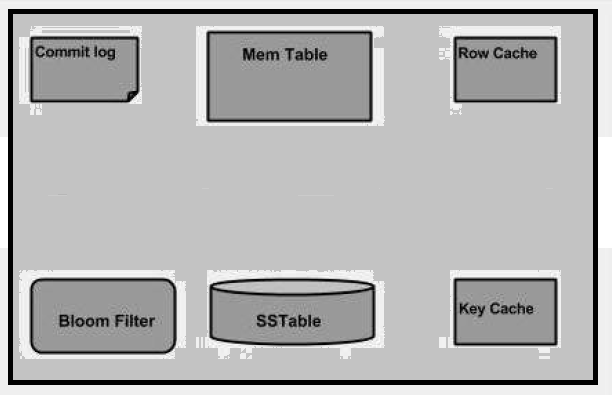
Apache Cassandra is a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. It is a type of NoSQL database.

* Cassandra was developed at Facebook for inbox search.
* It was open-sourced by Facebook in July 2008.
* Cassandra was accepted into Apache Incubator in March 2009.
* It was made an Apache top-level project since February 2010.

**Key Features**

* It is scalable, fault-tolerant, and consistent.
* It is a column-oriented database.
* Its distribution design is based on Amazon’s Dynamo and its data model on Google’s
* Bigtable.
* Created at Facebook, it differs sharply from relational database management systems.
* Cassandra implements a Dynamo-style replication model with no single point of failure, but adds a more powerful “column family” data model.
* Cassandra is being used by some of the biggest companies such as Facebook, Twitter, Cisco, Rackspace, ebay, Twitter, Netflix, and more.

**Architecture**

****

* Node - It is the place where data is stored.
* Data center - It is a collection of related nodes.
* Commit log - The commit log is a crash-recovery mechanism in Cassandra. Every write operation is written to the commit log.
* Cluster - A cluster is a component that contains one or more data centers.
* Mem-table - A mem-table is a memory-resident data structure. After commit log, the data will be written to the mem-table. Sometimes, for a single-column family, there will be multiple mem-tables.
* SSTable - It is a disk file, to which the data is flushed to, from mem-table, when its contents reach a threshold value.
* Bloom filter - These are nothing but quick, nondeterministic, algorithms for testing whether an element is a member of a set. It is a special kind of cache. Bloom filters are accessed after every query.
* Compaction - The process of freeing up space by merging large accumulated data files is called compaction. During compaction, the data is merged, indexed, sorted,
* Keyspace

Keyspace is the outermost container for data in Cassandra. The basic attributes of Keyspace in Cassandra are:

* Replication factor - It is the number of machines in the cluster that will receive copies of the same data.
* Replica placement strategy - It is nothing but the strategy to place replicas in the ring. We have strategies such as simple strategy *rackawarestrategy, old network topology strategy rack − awarestrategy, and network topology strategy datacenter − sharedstrategy.*
* Column families - Keyspace is a container for a list of one or more column families. A column family, in turn, is a container of a collection of rows. Each row contains ordered columns. Column families represent the structure of your data. Each keyspace has at least one and often many column families
* Syntax

CREATE KEYSPACE Keyspace name

WITH replication = {'class': 'SimpleStrategy', 'replication\_factor' : 3};

* **Column Family**

A column family is a container for an ordered collection of rows. Each row, in turn, is an ordered collection of columns. The following table lists the points that differentiate a column family from a table of relational databases.