International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/2231280453/IJCTT-V67I3P126
Volume 67 Issue 3,137-141, March 2019

Original Article

Notable Internet Applications of NoSQL Cassandra Database

Jaskiranjit Kaur[1], Chitranjanjit Kaur[2] , Swati[3]

[1] *Assistant Professor CS department, KNCW, Phagwara,*
[2]*Assistant Professor CS department, GNC, Phagwara GNDU, India,*

Abstract *- Cassandra is a distributed database designed to be highly scalable in terms of storage volume and request throughput while not being subject to any single point of failure. This paper presents the features and notable internet applications of nosql Cassandra and discusses how its works with a scalable multi-master database with no single points of failure. Additionally, a project demonstrates how Cassandra can be leveraged to store and query high-volume, consumer-oriented data. Cassandra is in use at Digg, Facebook, Twitter, Reddit, Rackspace, Cloudkick, Cisco, SimpleGeo, Ooyala, OpenX, Netflix, and more large, active data sets. The largest production cluster has over 100 TB of data in over 150 machines. Data is automatically replicated to multiple nodes for fault tolerance. Replication across multiple data centers is supported. Failed nodes can be replaced with no downtime. Every node in the cluster is identical. There are no network bottlenecks. There are no single points of failure.*

Keywords *- column family,Cassandra,nosql database.*

I. INTRODUCTION

Databases are defined as collections of related data. Although when using the term database, we refer to the complete database system, the term refers only to the collection of data stored only once and accessed by multiple users simultaneously. The system which handles Big data, transactions, metadata, problems, or any other aspect of the database is the Database Management System (DBMS). Early designs and implementations were based on linked lists to create relations between data and find specific data by applying database languages. Databases were created to satisfy this need of storing and finding consistent, redundant free data in a good manner. After their inception in the 1960s, different types were invented, each using its representation of data and different technology for handling transactions. They began with navigational databases based on linked lists, then moved on to relational databases with joins, afterward object-oriented and indexes in the late 2000s. These all are based on structured format data. Unstructured data is approximately 80% of the data that organizations process daily, and By 2025, the prediction of IDG projects.There will be approximately 163 zettabytes of data in the world, and estimates indicate that 80% of this data is unstructured. Unstructured data comes from documents, social media websites, digital pictures and videos, audio transmissions, sensors used to gather climate information, and unstructured content from the web. With the large objects assigned to its keys, unstructured data requires more processing and more storage. So to handle unstructured data, different databases come into existence; these databases are NoSQL(MongoDB, Cassandra, Hypertable, Hbase/Hadoop, CouchDBetc) and have become a popular trend. Handling big data is the challenge of data management with high performance. NoSQL databases were designed to provide database solutions for large volumes of unstructured data. Many NoSQL databases organize the data into key-value pairs,column-based,document-based,graph based.

II. OVERVIEW OF NOSQL DATABASES

There are four main categories of NoSQL databases:

A. Key-value stores

Data is stored as unique key-pairs values. Here systems are similar to dictionaries, where a single key addresses data. The key can be synthetic or auto-generated, while the value can be String, JSON, BLOB, etc. Values are isolated and independent, and the application logic handles relationships. Riak and Amazon's Dynamo are the most popular key-value store NoSQL databases.

B. Column family database

it defines the data structure as a predefined set of columns. In a column-oriented NoSQL database, data is stored in cells grouped in the column, not in rows like

RDBMS. Columns are logically grouped into column families. These Column families can be created at runtime or the definition of the schema. Read and write is done using columns rather than rows, and the main benefit of storing data in columns is fast search/ access and data aggregation. The best-known examples are Google's BigTable and HBase& Cassandra.

C. Document-based storage

It is also called semi-structured data. A document-oriented database, or document store, is nosql, a computer program designed for storing, retrieving, and managing document-oriented data. XML databases are a subclass of document-oriented databases that work with XML documents. It is a subclass of the key-value store. Instead of columns with names and data types used in a relational database, a document contains a description of the data type and the value. Each document may have the same or different structure. To add additional types of data to a document database, there is no need to modify the entire database schema with a relational database.

D. Graph databases

Graph databases are part of the NoSQL databases, which use graph structures to represent data and schemas. A graph database works with three abstractions: node, relationships between nodes, and key-value pairs that can attach to nodes and relationships. Nodes can be labeled to be grouped. Graph databases are particularly helpful because they highlight the links and relationships between relevant data Neo4j, oriented, AllegroGraph, ArangoDB, Graph Engine, Grapholytic, Teradata Asterareare graph based databases.

III. CASSANDRA DATABASE SYSTEM

Facebook released Cassandra in July 2008 as an open-source project, and Apache Software Foundation maintains the Cassandra documentation. Cassandra uses wide column stores which utilize rows and columns but allows the name and format of those columns to change. It uses a blend of tabular and key-value. Cassandra is a very scalable and resilient NOSQL database that is easy to maintain, configurable, fault-tolerant, and handles large data with high accessibility, providing neat solutions for complex problems. It is centralized storage for logs and metrics, and retrieving historical information from this storage is a great task for Cassandra.

A. History of Cassandra

Apache Cassandra was initially developed at Facebook to power their Inbox Search feature by AvinashLakshman (one of the authors of Amazon Dynamo) and Prashant Malik. It was released as an open-source project on Google code in July 2008. In March 2009, it became an Apache Incubator project. On February 17, 2010, it graduated to a top-level project. It was named after the Greek mythological prophet Cassandra.

Releases after graduation include

- 0.6, released April 12, 2010, additional feature support for integrated caching and Apache Hadoop Map Reduce
- 0.7, released January 8, 2011, additional feature secondary indexes and online schema changes
- 0.8, released June 2, June 2, 2011, additional features the Cassandra Query Language (CQL), self-tuning memtables, and zero-downtime upgrades.
- 1.0, released October 17, 2011, additional features with integrated compression, leveled compaction, and improved read performance
- 1.1, released April 23, April 23, 2012, features self-tuning caches, row-level isolation, and support for mixed SSD/spinning disk deployments
- 1.2, released January 2, 2013, additional features for improving performance added clustering across virtual nodes, inter-node communication, atomic batches, and request tracing
- 2.0, released September 4, September 4, 2013, additional features lightweight transactions (based on the Paxos consensus protocol), triggers, improved compactions
- 2.0.4, released December 30, December 30, 2013, more features like allowing specifying data centers to participate in a repair, client encryption support toss table loader, allows removing snapshots of no-longer-existing CFS
- 2.1.0 released September 10, 2014
- 2.1.6 released June 8, 2015
- 2.1.7 released June 22, 2015
- 2.2.0 released July 20, 2015

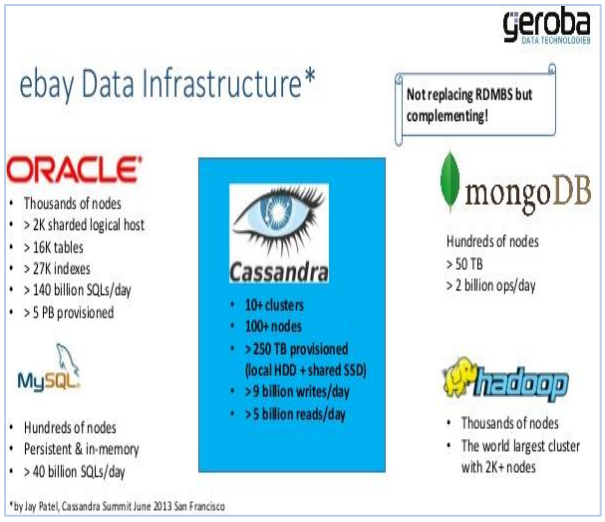

Fig. 1 comparison of Nosql databases

Following are the important advantages of using Cassandra:-

1. Free and open-source software, linearly scalable
2. Helps solve complicated tasks with ease
3. Has a short learning curve
4. Integrated caching and tunable consistency.
5. No master-slave issues due to peer-to-peer architecture. So there are no downtime problems.
6. Lowers admin overhead and costs for a DevOps engineer
7. Rapid writing and lightning-fast reading
8. It accommodates all possible data formats, either structural, unstructured, or semi-structural.
9. Extreme resilience and fault tolerance.

Fig.2 Features of Cassandra database

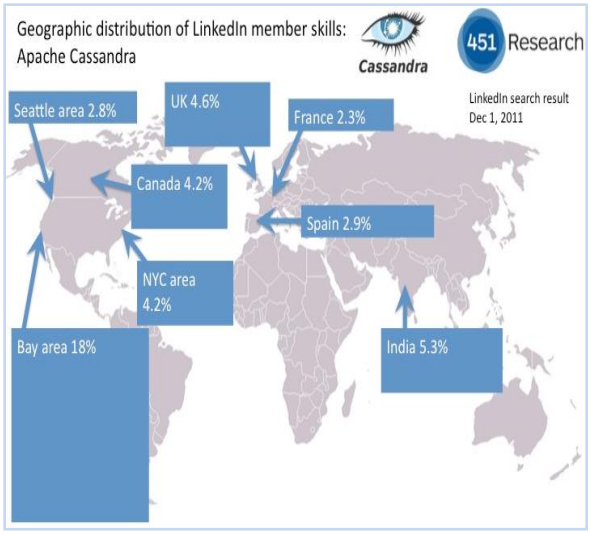

Fig. 3 Geographic distribution of Cassandra database

An article named " Update on the relative popularity of NoSQL database skills" with the URL https://blogs.the451group.com/information_management/tag/linkedin/page/2/ indicates that Canada is a hot spot for Apache Cassandra skills, with 4.1%, while Apache Cassandra is also making in-roads into Europe via France and Spain.

The following table lists the points that differentiate a relational database from a NoSQL database.

Relational Database	**Cassandra Database**
It Supports only structured data and powerful query language(SQL, oracle).	Capability to Handle unstructured data and Supports very simple query language.
It has a fixed schema(3 levels of the arch.)	No fixed schema.
Follows ACID (Atomicity, Consistency, Isolation, and Durability), which helpful for concurrent transactions	It is only "eventually consistent ."Nosql based upon CAP Theorem

Create table command in RDBMS using SQL syntax Create table name (col name data type, col name data type);	Create table command in Cassandra using cqlsh Create (table |Column Family) <Tablename> (‘<Col Definition>’) ,’<col defination>’) (with <option> and <option>)
The column represents the attributes of a relation	A column is a unit of storage in Cassandra.
RDBMS supports the concepts of foreign keys, joins	Relationships are represented using collections.
In RDBMS, a table is an array of arrays. (ROW, COLUMN)	In Cassandra, a table lists "nested key-value pairs ."(ROW, COLUMN key, COLUMN value)
Tables in rdbms are also called entities of a database	Tables or column families are the entity of a keyspace.

Table I. Comparison of SQL and Cassandra database

IV. CASSANDRA ARCHITECTURE

To always be available and avoid failure situations, It consists of a ring-type structure where its nodes are logically distributed like a ring. These nodes are fundamental Data storage units of Cassandra, and it has not like enslaver or slave nodes. Data is replicas by exchanges, and Each information among several homogenous nodes of the cluster (collection of many data centers). After this, Every write operation is written to the commit log. It is also called a crash-recovery mechanism in Cassandra. A sequentially written commit log on each node captures write activity to ensure data durability and consistency on each cluster. This data is also indexed and written to be memorable. (MemTables is a temporary memory location where data is written in memtables after being written in the commit log. The data in memtables are flushed to the disk, once they are full, to form SSTables.) Once the memtable is full system writes data on disk on the SSTable data file. All the data is partitioned and replicated to other nodes automatically. Two factors are important to consider in the replication process, i.e., the Replication Factor and the Replication Strategy. This ensures fault tolerance and reliability by using a process known as compaction. Cassandra periodically updates SSTables and removes outdated data and tombstones. A client can make a read/write request to any node in the cluster.

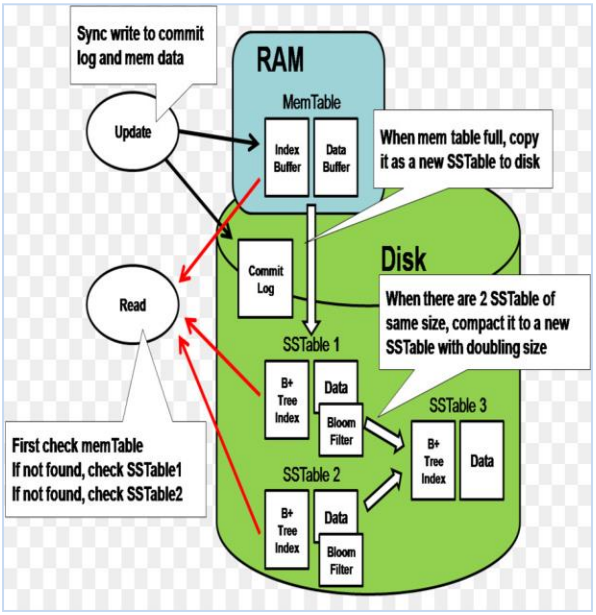

Fig. 4 Cassandra Architecture

V. NOTABLE APPLICATION OF CASSANDRA

Cassandra provides many of today's modern business applications by offering continuous availability, high scalability and performance, strong security, and operational simplicity while lowering the overall cost of ownership. Cassandra has many key customers like Apple, Netflix, Uber, ING, Intuit, Fidelity, NY Times, Outbrain, BazaarVoice, Best Buy, Comcast, eBay, Pearson Education, Walmart, Microsoft, McDonald's, Macquarie Bank.
According to the DB-Engines ranking(The DB-Engines Ranking ranks database management systems according to their popularity. The ranking is updated monthly.), Cassandra is the most popular wide column store, and in November 2018 became the11th most popular database and became 1st according to the wide column store model.

- **Apple** uses 100,000 Cassandra nodes, as revealed at Cassandra Summit San Francisco 2015, although it has not elaborated on which products, services, or features.
- **BlackRock** uses Cassandra in their Aladdin investment management platform[8][9]. Evan Chan's

presentation on FiloDB, a new OLAP database, shows architectures and techniques for combining Apache Cassandra and Spark to yield a 10-1000x improvement in OLAP analytical performance.

- **CERN** used a Cassandra-based prototype for its ATLAS experiment to archive the online DAQ system's monitoring information for Making data accessible, facilitating long-term analysis, and faster debugging. [10]
- **Cisco's WebEx** uses Cassandra to store user feed and activity near real-time.[11]and this data get from Cassandra users survey.
- **Formspring** uses Cassandra to count responses and store Social Graph data (followers, following, blocking) for 26 Million accounts with 10 million responses a day[12].
- **Globo.com** is the internet branch for GrupoGlobo, one of the 5 largest media conglomerates globally, producing content such as TV series, telenovelas, TV shows, news shows, etc., exporting them worldwide. It uses Cassandra as a back-end database for its streaming services[13]
- **Constant Contact**, Inc. is an online marketing company headquartered in Waltham, Massachusetts, with additional offices in San Francisco; Loveland, Colorado; New York, New York; Delray Beach, and London, United Kingdom. Stefan Piesche, Constant Contact CTO, spoke at the Data @Scale conference in Boston, hosted by Facebook, that We have around 350 Cassandra nodes spanning 2 data centers. That system provides 10x the performance of the old RDBMS and 1/10th of the cost.[14].
- **Babylon** Health handles an incredibly high volume of sensitive patient data that it needs to keep secure and usable for invaluable patient insights. Babylon expands and improves its service and builds highly personalized, mobile-based, 24-7 healthcare service while securing its customers' valuable data using Cassandra.[15]
- **Mahalo.com** chooses Cassandra as a critical component of their future technology architecture with minimal effort and building a data infrastructure that will support future throughput and scalability requirements while helping to control costs.[16]
- **Urban Airship** uses Cassandra with the mobile service hosting for over 160 million application installs across 80 million unique devices[17]

VI. CONCLUSION

In this paper, a detailed study is made to understand the features and working of the nosql Cassandra database and comparison with the SQL database. We also explain the working of Cassandra with the help of its architecture. Cassandra is popular among nosql databases, and Cassandra is used for various internet applications. Cassandra is the best choice for businesses due to its great features such as high availability, consistency, low downtime, fault tolerance, and high scalability in terms of both users and data, so Cassandra has many key customers. Thus we presented the features of the Cassandra distributed database management system and the benefits of its use in real-world enterprise applications.

REFERENCES

[1] https://www.tutorialspoint.com/cassandra/cassandra_data_model.htm
[2] https://blog.panoply.io/cassandra-vs-mongodb
[3] https://data-flair.training/blogs/cassandra-features
[4] http://www.datastax.com/products/datastax-enterprise-production-certified-cassandra
[5] http://www.ijettjournal.org/2015/volume-26/number-5/IJETT-V26P245.pdf
[6] https://db-engines.com/en/ranking
[7] https://en.wikipedia.org/wiki/Apache_Cassandra
[8] https://www.datastax.com/2015/09/top-cassandra-sessions-for-advanced-cassandra-users
[9] https://vimeopro.com/user35188327/cassandra-summit-2015/video/143826965
[10] https://cdsweb.cern.ch/record/1432912
[11] https://www.mail-archive.com/cassandra-dev@incubator.apache.org/msg01163.html
[12] https://www.slideshare.net/martincozzi/cassandra-formspring
[13] https://www.javacodegeeks.com/2016/06/cassandra-heart-globos-live-streaming-platform.html
[14] https://techblog.constantcontact.com/tech-talk/cassandra-sharded-mysql-scaling/
[15] https://www.datastax.com/resources/casestudies
[16] http://www.datastax.com/wp-content/uploads/2011/06/DataStax-CaseStudy-Mahalo.pdf
[17] https://www.slideshare.net/eonnen/from-100s-to-100s-of-millions.