Notable Internet Applications of NoSQL Cassandra Database

Jaskiranjit Kaur¹, Chitranjanjit Kaur², Swati³

# Assistant Professor CS department, KNCW, Phagwara, #Assistant Professor CS department, GNC, Phagwara

GNDU, India, respectively

Abstract

Cassandra is a distributed database designed to be highly scalable both in terms of storage volume and request throughput while not being subject to any single point of failure. This paper presents an features and notable internet applications of nosql Cassandra and discusses how its works with scalable multi-master database with no single points of failure. Additionally, a project is supplied demonstrating 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 companies that have 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.

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 actually refers only to the collection of data which stored only once and access by multiple users at same time. The system which handles Big data, transactions, meta data, problems or any other aspect of the database is the Database Management System (DBMS). Early designs and implementations were based on the use of linked lists to create relations between data and to find specific data by applying database languages. Databases were created in order to satisfy this need of storing and finding consistence, redundant free data in a good manner. After their inception in the 1960’s different types have invented, each using its own representation of data and different-different technology for handling transactions. They began with navigational databases which were based on linked-lists then moved on to relational databases with joins, afterwards object-oriented and indexes in the late 2000s and these all are based on structured format data. Unstructured data is approximately 80% of the data that organizations process daily and By 2025, predication of IDG projects that 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 web sites, digital pictures and videos, audio transmissions, sensors used to gather climate information, and unstructured content from the web. Unstructured data, with the large objects that are assigned to its keys, requires more processing and more storage, so to handle unstructured data different databases come in existence, theses databases are NoSQL(MongoDB, Cassandra, HBase/Hadoop, CouchDBetc) and has become a popular trend. To handle big data is the challenge of data management with high performance. NoSQL databases were designed to provide database solutions for large volumes of data that is unstructured. 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 data is addressed by a single key. The key can be synthetic or auto-generated while the value can be String, JSON, BLOB etc. Values are isolated and independent from another, and relationships are handled by the application logic. 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 column-oriented NoSQL database, data is stored in cells grouped in 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 a computer program designed for storing, retrieving and managing document-oriented data. XML databases are a subclass of document-oriented databases work with XML documents. It is subclass of the key-value store. Instead of columns with names and data types that are used in a relational database, a document contains a description of the data type and the value for that description. 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 as there is with a relational database.

D. Graph databases

Graph databases are part of the NoSQL databases which uses 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 labelled to be grouped. Graph databases are particularly helpful because they highlight the links and relationships between relevant data. Neo4j, orientDB, AllegroGraph, ArangoDB, Graph Engine, Grapholytic, Teradata Asterareare are 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 a tabular and key-value. Cassandra is a very scalable and resilient NOSQL database that is easy to maintain, simple to configure, fault-tolerant and handle large data with high accessibility, providing neat solutions for quite complex problems. It is a centralized storage for logs and metrics and retrieving historical information from this storage is a great task of 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 Apr 12 2010, additional feature support for integrated caching, and Apache Hadoop Map Reduce
- 0.7, released Jan 08 2011, additional feature secondary indexes and online schema changes
- 0.8, released Jun 2 2011, additional feature the Cassandra Query Language (CQL), self tuningmentables, plus support for zero downtime upgrades.
- 1.0, released Oct 17 2011, additional feature with integrated compression, leveled compaction, and improved read performance
- 1.1, released Apr 23 2012, more features self-tuning caches, row-level isolation, and support for mixed ssd/spinning disk deployments
- 1.2, released Jan 2 2013, additional feature for improve performance added clustering across virtual nodes, inter-node communication, atomic batches, and request tracing
- 2.0, released Sep 4 2013, additional feature lightweight transactions (based on the Paxos consensus protocol), triggers, improved compactions
- 2.0.4, released Dec 30 2013, more features like allowing specifying datacenters to participate in a repair, client encryption support to ss table loader, allow removing snapshots of nonlonger-existing CFs
- 2.1.0 released Sep 10 2014
- 2.1.6 released June 08, 2015
- 2.1.7 released June 22, 2015
- 2.2.0 released July 20, 2015
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 tuneable consistency.
5. No master slave issues due to peer to peer architecture, so there is 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, unstructural and semi structural.
9. Extreme resilience and fault tolerance.

In article named “Update on the relative popularity of NoSQL database skills” with URL https://blogs.the451group.com/information_management/tag/linkedin/page/2/ indicate that Canada is a hotspot 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.

<table>
<thead>
<tr>
<th>Relational Database</th>
<th>Cassandra Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>It Supports only structured data and powerful query language(sql,oracle).</td>
<td>Capability to Handle unstructured data and Supports very simple query language.</td>
</tr>
<tr>
<td>It has a fixed schema(3 level of arch.)</td>
<td>No fixed schema.</td>
</tr>
<tr>
<td>Follows ACID (Atomicity, Consistency, Isolation, and Durability) which helpful for concurrent transactions</td>
<td>It is only “eventually consistent”. Nosql based upon CAP Theorem</td>
</tr>
</tbody>
</table>
| Create table command in RDBMS using SQL syntax  
Create table tablename (colnamedatatype,colnamedatatype); | Create table command in Cassandra using cqlsh  
Create (table | Column Family)  
<Tablename> (<Col Definition> ,<Col Definition>) (with <option> and <option>) |
| Column represents the attributes of a relation       | 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 is a list of “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 key space. |

Table I: Comparison of SQL and Cassandra database
IV. CASSANDRA ARCHITECTURE

For always available and to avoid failure situation It consists of a ring type structure where its nodes are logically distributed like a ring. Theses nodes are fundamental Data storage unit of Cassandra and it has not like master 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 and sequentially written commit log on each node captures write activity to make sure data durability and consistency on each clusters. This data is also indexed and written to memtable.(MemTables is temporary memory location where Data is written in memtables after it has been written in the commit log. The data in memtables is flushed to the disk, once they are full, to form SSTables.) Once the memtable is full system write data on disk on SSTable data file. All the data is partitioned and replicated to other nodes automatically. There are two factors that are important to consider in the replication process i.e Replication Factor and 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 read/write request to any node in the cluster.

![Cassandra Architecture Diagram](image_url)

**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 overall cost of ownership. Cassandra has many key costumers like Apple, Netflix, Uber, ING,, Intuit, Fidelity, NY Times, Outbrain, BazaarVoice, Best Buy, Comcast, eBay, Pearson Education, Walmart, Microsoft, McDonalds, Macquarie Bank. According to 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 the 11th most popular database and became 1st according wide column store model.

- **Apple** uses 100,000 Cassandra nodes, as revealed at Cassandra Summit San Francisco 2015, although it has not elaborated for which products, services or features.
- **BlackRock** uses Cassandra in their Aladdin investment management platform[8][9] and Evan Chan’s presentation on FiloDB, a new OLAP database show architectures and techniques for combining Apache Cassandra and Spark to yield a 10-1000x improvement in OLAP analytical performance.
- **CERN** used Cassandra-based prototype for its ATLAS experiment to archive the online DAQ system’s monitoring information for Making data accessible, facilitates long term analysis and faster debugging.[10]
- **Cisco’s WebEx** uses Cassandra to store user feed and activity in near real time.[11] and this data get from Cassandra users survey.
- **Formspring** uses Cassandra to count responses, as well as 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 in the world, producing content such as TV series, telenovelas, TV shows, news shows, etc. exporting them worldwide. It uses Cassandra as a back-end database for their 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 systems 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 but also usable for invaluable patient insights. Babylon expand and improve its service and build highly personalized, mobile-based, 24-7 healthcare service while securing its customers’ valuable data using Cassandra. [15]

• **Mahalo.com** choose Cassandra as a critical component of their future technology architecture. With a minimal amount of effort, and build 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. CONCLUSIONS

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

REFERENCES

[10] https://cdsweb.cern.ch/record/1432912
[12] https://www.slideshare.net/martincozzi/cassandra-formspring
[17] https://www.slideshare.net/eonnen/from-100s-to-100s-of-millions