International Journal of Research in Engineering, Science and Management
Volume-2, Issue-8, August-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Study of Apache Cassandra

Omkumar Badhai[1], Shatam Bhagat[2], Rohit Mandavkar[3], Shubham Ingole[4], Nidhi Gupta[5]
[1,2,3,4]*Student, Dept. of Computer Science and Engg., Sipna College of Engg. and Technology, Amravati, India*
[5]*Professor, Dept. of Computer Science and Engg., Sipna College of Engg. and Technology, Amravati, India*

***Abstract*: Cassandra is open source and is in development at Apache. The Apache Cassandra project brings together Dynamo's fully distributed design and Bigtables Column family based data model Cassandra is adapting to recent advances in distributed algorithms like Accrual style failure detection and others. Cassandra is proven as it is in use by Digg, Facebook, Twitter, Reddit, Rack space, Cloudkick, Cisco. The largest production cluster has over 100 TB of data in over 15 machines. It is Fault tolerant, decentralizes and gives the control to developers to choose between synchronous and asynchronous data replication. It offers rich data model, to efficiently compute using key and value pairs. It is highly scalable both in terms of storage volume and request throughput while not being subject to any single point of failure. It is durable and supports third party applications. Cassandra aims to run on top of an infrastructure of hundreds of nodes (possibly spread across different data centres.)**

***Keywords*: Cassandra, apache.**

1. Introduction

A NoSQL (originally referring to "non SQL" or "non-relational") database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century, triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com. NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

Apache CassandraTM is a massively scalable NoSQL database. Cassandra's technical roots can be found at companies recognized for their ability to effectively manage big data – Google, Amazon, and Facebook – with Facebook open sourcing Cassandra to the Apache Foundation in 2009.

Used today by numerous modern businesses to manage their critical data infrastructure, Cassandra is known for being the solution technical professionals turn to when they need a NoSQL database that supplies high performance at massive scale, which never goes down. In particular, Cassandra addresses big data applications, which are exploding across nearly every industry. This paper provides a brief overview and introduction to Cassandra for those wishing to understand if Cassandra is right for them and how it is uniquely positioned to address the next phase of growth in the modern Database Marketplace.

2. Literature survey

Cassandra: A Decentralized Structured Storage System by Avinash Lakshman, Prashant Malik. Cassandra was designed to fulfil the storage needs of the Inbox Search problem. Cassandra can support a very high update throughput while delivering low latency. Future works involves adding compression, ability to support atomicity across keys and secondary index support.

Cassandra File System Over Hadoop Distributed File System by Ashish Mutha, Vaishali M. Deshmukh The main aim of design goals for the Cassandra File System were to first, it simplifies the operational overhead of Hadoop by removing the single points of failure in the Hadoop Name Node. Cassandra file system delivers the ability to run analytics on Cassandra data that comes from line-of-business applications.

3. Why Cassandra?

No Single Point of failure: The no single point of failure design principle asserts simply that no single part of a system can stop the entire from working.

Data replication: Cassandra stores replicas on multiple nodes to ensure reliability and fault tolerance. A replication strategy determines the nodes where replicas are placed. The total number of replicas across the cluster is referred to as the replication factor

CQL: The Cassandra Query Language (CQL) is the primary language for communicating with the Apache Cassandra database. The most basic way to interact with Apache Cassandra is using the CQL shell, clash. Using clash, you can create key spaces and tables, insert and query tables, plus much more.

4. The architecture of Cassandra

The architecture of Cassandra greatly contributes to its being able to scale, perform, and offer continuous availability. Cassandra was built from the ground up with the understanding that hardware and system failures can and do occur. This translates into Cassandra sporting a different way of managing and protecting data than a traditional RDBMS. Rather than using a legacy master-slave or a manual and difficult-to-maintain shared design, Cassandra has a peer-to-peer distributed architecture that is much more elegant, and easy to set up and maintain. In Cassandra, all nodes are the same; there is no concept of a master node, with all nodes communicating with each other via a gossip protocol. Cassandras built-for-scale

architecture means that it is capable of handling petabytes of information and thousands of concurrent users/operations per second (across Multiple data canters) as easily as it can manage much smaller amounts of data and user traffic. It also means that, unlike other master-slave or shared systems, Cassandra has no single point of failure and therefore is capable of offering true continuous availability.

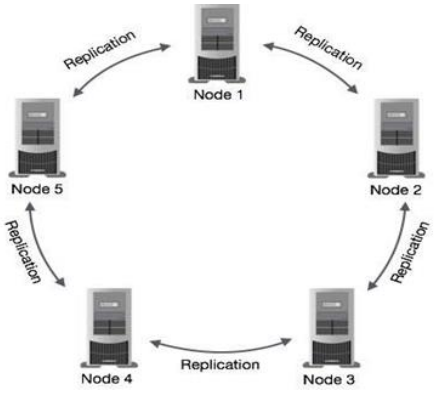

Fig. 1. Architecture Cassandra

5. Advantage of Cassandra

Open Source: Cassandra is Apaches open-source project; this means it is available for FREE! Yes, you can download the application and use the way you want.

Peer to Peer Architecture: Cassandra follows a peer-to-peer architecture, instead of master-slave architecture. Hence, there is no single point of failure in Cassandra.

Elastic Scalability: One of the biggest advantages of using Cassandra is its elastic scalability. Cassandra cluster can be easily scaled-up or scaled-down.

High Availability and Fault Tolerance: Another striking feature of Cassandra is Data replication which makes Cassandra highly available and fault-tolerant. Replication means each data is stored at more than one location.

6. Conclusion

Cassandra is fully replicated. There is no master, no slave. It's always on, its performant and these are some of the features and characteristics of Cassandra that make it a fantastic solution to the big data challenge. There is no question that many modern applications have outgrown legacy relational databases. To handle big data workloads, these systems require a massively scalable NoSQL database. While there are a number of NoSQL database providers in the market, only Cassandra is able to offer the linear scale performance and key enterprise class features that meet the expectations and requirements of big data systems.

References

[1] Avinash Lakshmana and Prashant Malik, "Cassandra - A Decentralized Structured Storage System."
[2] Ashish Mutha, Vaishali M. Deshmukh, "Cassandra File System over Hadoop Distributed File System."
[3] Bhalchandra Bhutkar, Jagannath Aghav, Sunil Dorwani, "Data Management using Apache Cassandra."
[4] http://en.wikipedia.org/wiki/Apache/Cassandra.2014
[5] http:www.datastax.com/dev/blog/apache-cassandra