International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 3 Ver. 4 || March 2018 || PP01-04

Cassandra Tool

Aasif Ansari
Lecturer, Computer Technology Department, Bgit

Abstract— *With Increasing Volume Of Data And Its Importance Increasing Day By Day Along With Its Variety. It's Critical To Analyse The Variety Of Data With Relational Database With Static Schema Of Tables. The Variety That Flows As An Information Through Various Social Media Tools As Text, Images, Videos, Emoticons Etc. Cannot Be Analysed With The Similar Schema Using Sequential Query Language. Nosql Works For The Variety Of Data That We Receive From The Social Media And Cassandra Db Was Designed As A Cross Platform With Big Tables By Google And Dynamo By Amazon*

Keywords— *Nosql, Cassandra, Big Tables, Dynamo, Keyspace, Replication Strategy.*

Date of Submission: 27-02-2018 Date of acceptance 15-03-2018

I. INTRODUCTION

Cassandra Is An Open Source Distributed Database Management System For Analysinglarge Amounts Of Data Across Variouscommodities.Cassandra Is Defined As "Nosql" Or "Non-Relational" Database And It Is A Combination Between "Key-Value Store" And A "Column-Orientated" Database. Cassandra Was Initially Created At Facebook With Combination Of Google Big Table And Amazon Dynamo. It Was Created To Power The "Inbox Search" Feature At Google. Cassandra Was Released As Open Source In July Of 2008. It Became An Apache Incubator Project In February Of 2009 And It Became A Full Level Project A Year After That.Cassandra Was Modeled After Google's "Big Table" And Amazon's "Dynamo". Cassandra Even Has Similarities To A "Relational Database", But More Flexible.

Apache Cassandra Is An Open Source Largelyscalable Nosql Database. Cassandra Excels Atbeing The Underlying Database For Modern Onlineapplications That Need Extremely Fast Read And Writeoperations. It Can Analyse Thedistribution Of Data Across Multiple Data Centres Andcloud Availability Zones, And Offers Online Additionsof Capacity Via New Nodes While Providingcontinuous Availability For The System As A Whole.Columns – A "Column Family" Is Similar To A "Table" In A Relational Database Except Much More Flexible

Cassandra Encourages "Denormalization" As Oppose To Relational "Normalization".A "Column Family" Is Similar To A "Table" In A Rdbms Because It Has Columns And Rows. Relational Database Tables Use A Predefined, Fixed Schema. Column Families Do Not Which Makes Them Very Flexible. Cassandra's Data Model Promotes "Denormalization" Which Is The Complete Opposite Of The Relational Database. Columns Can Be Created On The Fly. Cassandra Is Sort Of In Its Own Data Model Class But Can Be Described As A Hybrid Of A "Key-Value Store" And A "Column-Orientated" Database.

I. CASSANDRA ARCHITECTURE

Peer To Peer Architecturewhich Is Built With The Understanding That Hardware & Software Failures Can Happen. The Architectures Are Designed For

- All Nodes Are The Same.
- Read/Write Can Be Anywhere.
- Gossip Protocol Implementation.
- Commit Log Captures For All Activities.

Key Technical Differentiators That Make Cassandra Awinning Choice In A Cloud-Computing Environmentinclude The Following:

- A Built-For-Scale Architecture That Can Handleterabytes Of Information And Thousands Ofconcurrent Users/Operations Per Second Aseasily As It Can Manage Much Smalleramounts Of Data And User Traffic
- Masterless Design That Offers No Single Pointof Failure For Any Database Process Orfunction; Every Node Is The Same, So There Isno Concept Of A Master Node Or Anythingsimilar
- Online Capacity Additions That Deliver Linearperformance Gains For Both Read And Writeoperations
- Read/Write Anywhere Capabilities Thatequate To A True Network-Independentmethod Of Storing And Accessing Data

- Guaranteed Data Safety That Ensures No Lossof Data, No Matter What Node Is Written To Ina Cluster
- Tunable Data Consistency That Allow Cassandra To Offer The Data Durability Andprotection Like An Rdbms, But With Theflexible Choice Of Relaxing Data Consistencywhen Application Use Cases Allow
- Flexible/Dynamic Schema Design Thataccommodates All Formats Of Big Dataapplications, Including Structured, Semi Structured, And Unstructured Data; Data Isrepresented In Cassandra Via Columnfamilies That Are Dynamic In Nature Andaccommodate All Modifications Online. Well Suited For Cloud Deploymentssimplified Replication That Provides Dataredundancy And Is Capable Of Being Multidatacenter And Cloud In Nature.
- Security That Includes Authorization Andauthentication Control
- Data Compression That Reduces The Footprintof Raw Big Data By Over 80 Percent In Someuse Cases
- A Sql-Like Language (Cql – Cassandraquery Language) That Lessens The Learningcurve For Developers And Administratorscoming From The Rdbms World
- It Supports Key Developer Languages (E.G.,Java, Python) And Operating Systems.
- No Requirement For Any Special Equipment;Cassandra Runs On Commodity Hardware
- Very Easy Installations In Cloudenvironments Including Amazon Machineimages (Amis) That Enable A User To Be Upand Running With A Multiple-Node Cluster Inminutes.

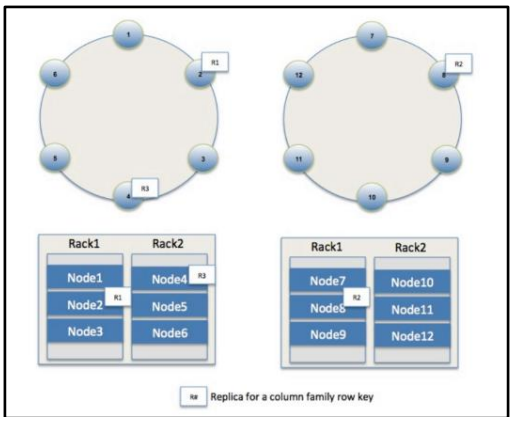

Figure 1: Ring Structure Node Representation

II. FEATURES OF CASSANDRA

The Basic Features Of Cassandra Which Makes It As An Easy Implementing Cloud Deployment Tools Are:

- Decentralized – No Master & No Single Point Of Failure. Data Is Distributed Across The Cluster
- Replication – Tailored For Multiple-Data Center Deployment
- Scalability – New Machines Can Easily Be Added With No Downtime Or Interruption
- Fault Tolerance – Failed Nodes Can Be Replaced With No Downtime
- Cassandra Query Language (Cql) – An Sql-Like Alternative
- Always On Architecture – Continuous Availability With No Downtime
- Faster Linear-Scale Performance
- Operational Simplicity – Administration Is Simplified
- Transaction Support
- No New Equipment Required - Very Economical

III. CQL (CASSANDRA QUERY LANGUAGE)

Cql Is Very Similar To Sql (Structured Query Language) In Terms Of Syntax And Commandsstatements Directly Change Data And/Or Change The Way Data Is Stored

- Maximum Flexibility When Distributing Data.
- Easy To Replicate Across Multiple Datacenters.
- Easy To Replicate In The Cloud.
- Read And Write To Any Node And All Changes Will Sync Automatically.
- All Statements End With A Semi-Colon.

Select * From Users;

- **Keyspaces**– A Keyspace Is Similar To A "Schema". A Keyspace Is A "Container" For Your Datasimilar To A Schema In An Rdbms. Used To Group Column Families Together, A Cluster Has One Keyspace Per Application. Serves As A Container For Database Objects Such As Tables And Indexes, And

Iswhere The Level Of Replication Is Set. It Is Analogous To A Microsoft Sql Server Or Mysqldatabase. The Figure 2 Shows The Creation Of Keyspace Videodb With The Columns.

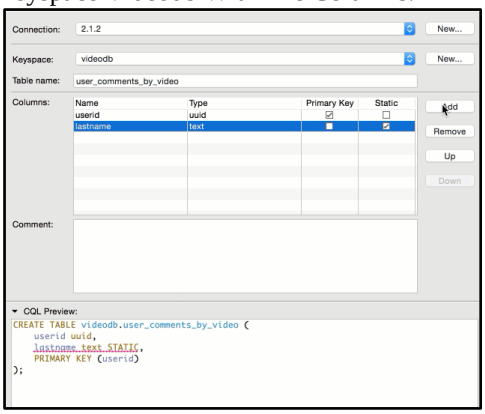

Figure 2: Keyspace And Column Creation

- **Replication** Is Controlled On A Per-Keyspace Basis. Simple Strategy – Use This For A Single Data Center. It Places The First Replica On A Node Determined By The Partitioner. Does Not Consider Topology.
- **Networktopologystrategy** - If You Plan To Have Your Cluster Span Across Multiple Data Centers. Specifies How Many Replicas You Want In Each Data Center.
- **Table** – Sometimes Referred To In Cassandra Literature As A Column Family, It Is The Primary Objectused To Store Data. A Cassandra Table Looks A Lot Like An Rdbms Table On The Surface, But Actuallyit Is A Sparse Data Object That Provides Much More Flexibility.
- **Index** –An Index In An Rdbms, It Is A Mechanism Used To Improve The Performance Ofsome Queries

II. DATASTAX

A Key Benefit Of Datastax Enterprise Is The Tightfeedback Loop It Has Between Real-Time Applicationsand The Analytics And Search Operations Thatnaturally Follow. Traditionally, Users Would Be Forcedto Move Data Between Systems Via Complex Etlprocesses, Or Perform Both Functions On The Samesystem With The Risk Of One Impacting The Other. Inbig Data Environments, This Process Can Be Timeconsuming And Burdensome.With Datastax Enterprise, Real-Time, Analytic, Andsearch Big Data Operations Take Place In The Samedistributed System, But Users Have The Ability Todedicate Certain Nodes Solely For Analytics Or Searchso Their Workloads Don't Slow Down Real-Timeprocessing. Users Simply Define One Or More Replicagroups, And Configure The Role Of Each – One Or Morecassandra, Analytics, And Search Nodes. Writes Areinstantly Replicated Between All Nodes.With Datastax Enterprise, Users Truly Have The Bestof All Worlds For Their Online Database Applications.They Have All The Power Of Cassandra Serving Theirhighest-Volume And High-Velocity, Real-Timeapplications; The Power Of Running Analytics On Theircassandra Data; And Enterprise Search On That Samedata In One Distributed Database. The Result Is Smartworkload Isolation For Big Data/Online Applicationsthat Is Much Simpler To Manage And More Reliablethan Any Alternative.

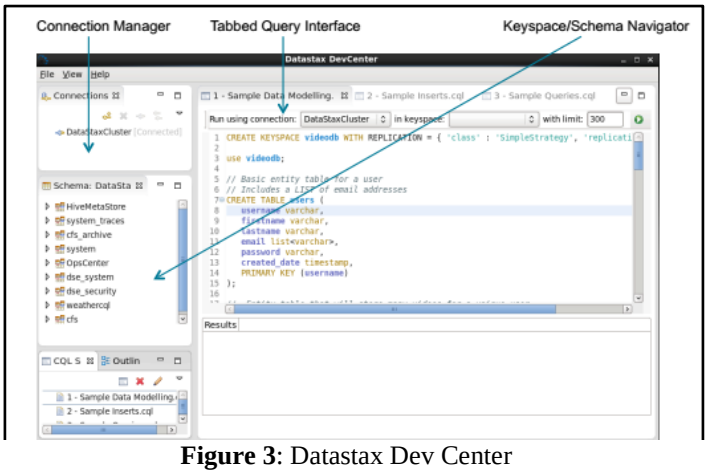

Figure 3: Datastax Dev Center

The Cql Helps To Design And Implement The Queries In Datastax. Before Designing The Entity Relation Diagram Is Designed Tfoe Development Which Give The Overall Structural View. The Entities And Their Relationships Are Considered During Table Design. Queries Are Best Designed To Access A Single Table, So All Entities Involved In A Relationship That A Query Encompasses Must Be In The Table. Datastax Enterprise Uses Single Table-Single Query Approach, Queries Can Perform Faster. Names And Ids Are Identified As 'K' For Primary Key And 'C' For Clustering Key In The Table.

III. CONCLUSIONS

The Datastax Dev Center And Opscenter Enhances The Working Functionality With Comprehensive Security Feature Of Any Nosqlfor The Cloud. Datastax Isdeployed With Confidence In Cloud Environmentswhere Data Security Is A Top Priority Because Itcontains The Types Of Security Capabilities Thatmodern Enterprises Need For Data Protection,Including Strong Authentication, Authorization,Encryption, And Data Auditing Capabilities.Moving To A Cloud-Based Infrastructure Necessitates Choosing A Database That Is Capable Of Fully Utilizing All Thebenefits The Cloud Provides. The Implementation Of Cassandra Helps To Understand The Cluster Creation And Implementation Using Cloud Infrastructure.

Creation Of Tables And Keyspace Helps To Organize The Structure With Ease. As Single Table Is Used For Query Processing, It Results In Faster Processing For Large Data. The Cluster Helps To Maintain The Data Consistency Without The Loss Of Data During The Time Os Failure.

References

[1]. Extracting Value From Chaos, By John Gantz And David Reinsel, Idc, June 2011, Http://Idcdocserv.Com/1142.
[2]. Ibid.
[3]. For Individual References For Each Statistic, Go To Http://Blog.Equinix.Com/2011/03/Optimizing-Internet-Application-Performance/.
[4]. For More Information On Big Data Management, See Big Data: Beyond The Hype, Why Big Data Matters To You, Datastax, October 2011,
[5]. Http://Www.Datastax.Com/Wp-Content/Uploads/2011/10/Wp-Datastax-Bigdata.Pdf.
[6]. "How Much Does Downtime Really Cost?," By Henry Martinez, Infomanagement Direct, August 6, 2009, Http://Www.Informationmanagement.Com/Infodirect/2009_133/Downtime_Cost-10015855-1.Html.
[7]. "Benchmarking Cassandra Scalability On Aws – Over A Million Writes Per Second," By Adrian Cockcroft And Denis Sheahan, The Netflix"Tech" Blog, November 2, 2011, Http://Techblog.Netflix.Com/2011/11/Benchmarking-Cassandra-Scalability-On.Html.
[8]. "Benchmarking High Performance I/O With Ssd For Cassandra On Aws," By Adrian Cockcroft, The Netflix "Tech" Blog, July 19, 2012,
[9]. Http://Techblog.Netflix.Com/2012/07/Benchmarking-High

Aasif Ansari" Cassandra Tool"International Journal of Engineering Science Invention (IJESI), vol. 07, no. 03, 2018, pp01-04