ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 10, October 2016

NoSQL based Cassandra Cluster with Combination of Tesseract for Optical Character Recognition

M.Ramya, S.Suresh Kumar

Abstract- **Processing huge data simultaneously in a distributed platform is difficult. Our present database like RDBMs which is SQL base database cannot be done. So we need a NoSQL data organization that can perform huge processing. Parallely taken an image and perform OCR creates huge complexity in the form of computational speed and ETL. Hence we combine the processing of OCR using Tesseract, which is effective OCR but underlying database document not support its full efficiency. Hence we create a model that combine tesseract to which in a distributed database where NoSQL based Cassandra. By through tesseract with Cassandra into we achieve high efficiency and high throughput.**

Index Terms—**Cassandra, NoSQL database, Tesseract, Optical Character Recognition.**

I. Introduction

A traditional database like RDBMs, where the system support to fail large amount of data to process. To overcome this problem , we use Apache Cassandra, it is a distributed database from Apache that is highly scalable, high performance distributed database designed to manage very large amounts of structured data, it can also handle unstructured which is a NoSQL database. It provides high availability with no single point of failure

a. NoSQL Database

- A NoSQL database (sometimes called as Not Only SQL) is a database that provides a mechanism to store and retrieve data other than the tabular relations used in relational databases. . These databases are schema-free, support easy replication, have simple API, eventually consistent, and can handle huge amounts of data. Shows in Fig 1 Objectives to have simplicity of design, horizontal scaling, and finer control over availability. Shows in Fig 1.

b .Apache Cassandra

- Apache Cassandra is an open source, distributed and decentralized storage system (database), for managing very large amounts of data spread out across the world. It provides highly available service with no single point of failure. It is scalable, fault-tolerant, consistent and column-oriented database. Created at Facebook, it differs sharply from relational database management systems. Cassandra is being used by some of the biggest companies such as Facebook, Twitter, Cisco, Rackspace, eBay, Twitter, Netflix, and more.

c. Tesseract –OCR

- **Tesseract** is an optical character recognition engine for various operating systems .It is free software, released under the Apache License, Version 3.0,and development has been sponsored by Google . Tesseract is considered one of the most accurate open source OCR engines currently available. It is available for Linux, Windows and Mac OS X, however, due to limited resources only Windows and Ubuntu.

Biotech
Web
Defense
NoSQL Market $14 Billion in 2013-2018 CAGR 21%
Mobile Telecom
ERP/CRM
Image signal processing
Finance/ Banking
Geoscience weather

Fig 1:NoSQL Features

Version 3.0 Tesseract has supported output text formatting, hOCR positional information and page layout analysis. Support for a number of new image formats was added using the Leptonica library. Tesseract can detect whether text is monospaced or proportional.

Manuscript received Nov, 2016

M.Ramya, *PG Student, Department of Computer Science and Engg, Rajalakshmi Engg College, Chennai, India*

S.Suresh Kumar, *Associate professor, Department of Computer Science and Engg, Rajalakshmi Engg College, Chennai, India*

II. Related Work

A lot of research has been carried out in the field of NoSQL database. Optical Character Recognition has increase the efficiency of the Converting Image representation of character into the readable text formatting. [1] had proposes the general overview technology of the NoSQL storage .Thorough the analysis will highlight the strengths, features and limitations of six most popular NoSQL databases and thus would help the organizations to choose a NoSQL database. They evaluated the best NoSQL solutions and also discussed their architectural working and best use cases. [2] Gave guidance that would combine the application requirements to a suitable NoSQL database. They also find how many nodes in a cluster for which a NoSQL database produces the best performance. They had taken three most popular NoSQL database MongoDB (document based), HBase(column based),Cassandra(hybrid) and the performance of each database on its ability to scale with different dataset sizes, operation counts and CRUD operation is analyzed.[3] introduced a new methodology and implementing for increasing the availability of unstructured data by presenting a new distributed storage system called Mystore, based on the combination of MongoDB and some advantages from other NoSQL system. Consistent hash is used to distribute data on multiple Mongo DB nodes, NWR mode is applied to automatic backup operation and guarantee data consistency. Gossip protocol is taken for exchanging information of failure in the system. Based on above requirements, a high available and scalable system for storing unstructured data is realized, which can also provide complex query function like relation database.[4] had published a popularity of NoSQL database, as uses many companies are developing Cassandra applications, they may require new tools to monitor database performance efficiently. When problems related to performance occur and proper analysis is required, the statistical data generated by monitoring tool will be of a lot help. To optimize NoSQL applications, developers need to have an idea about how the database is behaving in different working scenarios. Cassandra is easy to configure, but for the proper performance tuning it is necessary to study the performance requirements for a particular application. This can be evaluated by monitoring tool. The design of such monitoring tool and the results generated such as statistics and graphs. The tool will be used primarily for low end machines as they are cost effective.[5] had proposes a methodology to increase I/O performance of database appliances running in the cloud environment with distributed object storage as the underlying data stores. The proposed method involves separating the distributed storage's journal and data partitions to different hard drives and also separating a few database application directories to multiple RBD images from different storage pools in order to speed up the I/O operations. Experiments with SATA, SAS, and SSD type-drives with Ceph distributed storage system have been conducted based on proposed method and the results show significant performance compared to local drives and default distributed storage setup. Ref[6]had proposed a set of training regimes for Tamil using the Tesseract engine that have enabled us to develop a robust Tamil OCR system. They describe in detail our training regime, which results in a performance improvement of 12.5 % over the default Tamil module shipped with Tesseract on a set of ancient Tamil documents, which were part of an authentic project to digitize important Tamil manuscripts of Sri Lanka. Ref [7] proposes in their work they compare OCRs for printed Tamil texts on four different types of documents: books, magazines, newspapers and pamphlets .They propose a post-processing error correction technique to the tested OCRs which reduces the overall mean error rate by nearly 10% on those four categories. There are several well-known works in Tamil such as Tolkaappiyam, KambaRaamaayanam, Aaththisoodi, Silappathigaaram and Thirukkural. Tamil has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and one special character (called as 'aayutha ezhuththu') counting to a total of 247 characters. Ref[8] proposes an introduction of Optical Character Recognition (OCR) method, History of Open Source OCR tool Tesseract, architecture of it and experiment result of OCR performed by Tesseract on different kinds images are discussed. They conclude this paper by comparative study of this tool with other commercial OCR tool Transym OCR by considering vehicle number plate as input. From vehicle number plate they tried to extract vehicle number by using Tesseract and Transym and compared these tools based on various parameters .Ref[9] published data modeling approach that ensures sound and efficient schema design. They i) proposes the first query-driven big data modeling methodology for Apache cassandra, ii) defines important data modeling principles, mapping rules, and mapping patterns to guide logical data modeling, iii) presents visual diagrams for Cassandra logical and physical data models, and iv) demonstrates a data modeling tool that automates the entire data modeling process. Ref [10] presented their efforts in supporting the data storage and processing requirements for such characteristics. To achieve efficient queries about target data subsets, they propose a general customizable and scalable indexing framework that can be built over distributed NoSQL databases. This framework allows users to define suitable customized index structures for their query patterns against social media data, and supports scalable indexing of both historical and streaming data. They implement this framework on HBase, and name it Indexed HBase. Ref [11] had proposes a new technique of optical character recognition using Hough transform and statistical method. This method proves to be very effective with the use of randomized Hough transform for feature extraction and Statistical method for developing the model. The model proposed is trained and validated for two languages- cursive English and Tamil script and the results are found to be very much encouraging. The model developed works for the entire character set in both the languages.

II. Proposed Architecture

The scanned imaged in distributed database as taken into image processing for recognize Tamil character. By using Tesseract is an optical character recognition engine for various operating systems. It is free software, released under the Apache License and development has been sponsored by Google . Tesseract is considered one of the most accurate open source OCR engines currently available. The initial versions of Tesseract could only recognize English language text. Tesseract v2 added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese, and Dutch). Version 3 extended language support significantly to include ideographic (Chinese & Japanese) and right-to-left (e.g. Arabic, Hebrew) languages as well many more scripts. Also new languages are included Tamil. The language code for Tamil is “tam”.

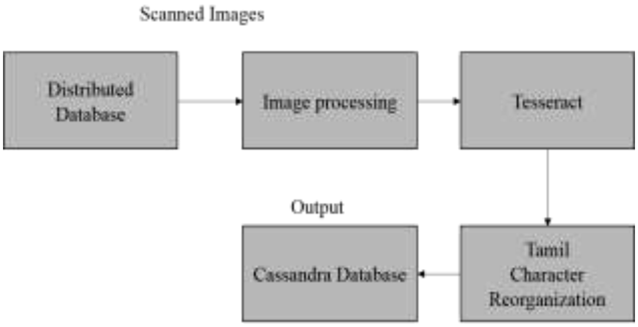

Fig 2: Proposed Architecture

Step1: The Cassandra distributed database has been created for increasing the speed and efficiency of the process. Because In Cassandra, one or more of the nodes in a cluster act as replicas for a given piece of data. If it is detected that some of the nodes responded with an out-of-date value, Cassandra will return the most recent value to the client. After returning the most recent value, Cassandra performs a **read repair** in the background to update the stale values. So it can increase the efficiency.

Step2: Further we recognize Tamil character from the scanned images by using Tesseract. This process behind performing their image processing functionality such as pattern reorganization, edge reorganization and so on.

Step3: The recognized Tamil characters are converted into readable text format and it can store in the form of text document in the system.

Step4: The output will store in the Cassandra distributed database, Because of its outstanding technical features. Given below are some of the features of Cassandra:

- **Elastic scalability** – It is highly scalable
- **Always on architecture** -It has no single point of failure.
- **Fast linear-scale performance -** It maintains a quick response time.
- **Flexible data storage** – It can accommodates all possible data formats including: structured, semi-structured, and unstructured
- **Easy data distribution -** Cassandra provides the flexibility to distribute data where we need by replicating data across multiple data centers.
- **Transaction support -** Cassandra supports properties like Atomicity, Consistency, Isolation, and Durability (ACID).
- **Fast writes -** Cassandra was designed to run on cheap commodity hardware

III. Conclusion

A RDBMs, where the system support to fail large amount of data to process. To overcome this problem , we use Apache Cassandra, it is a distributed database from Apache that is highly scalable, high performance distributed database designed to manage very large amounts of structured data, it can also handle unstructured which is a NoSQL database. It will provides speed, efficiency and more availability with no single point of failure. The system can achieve high throughput and fault tolerance.

References

[1] Pragati Prakash Srivastava, Saumya Goyal, Anil Kumar “Analysis of Various NoSql Database” 978-1-4673-7910-6/15/$31.00 c 2015 IEEE.

[2] Surya Narayanan Swaminathan, Ramez Elmasri “Quantitative Analysis of Scalable NoSQL Databases” 978-1-5090-2622-7/16 $31.00 © 2016 IEEE.

[3] Wenbin Jiang, Lei Zhang, Weizhong Qiang, Hai Jin, Yaqiong Peng “MyStore: A High Available Distributed Storage System for Unstructured Data” 978-0-7695-4749-7/12 $26.00 © 2012 IEEE.

[4] PrasannaBagade, Ashish Chandra, Aditya B.Dhende “Designing Performance Monitoring Tool for NoSQL Cassandra Distributed Database” 978-1-4673-2225-6/12/$31.00 ©2012 IEEE.

[5] Mohd Bazli Ab Karim, Luke Jing Yuan, Wong Ming-Tat, Hong Ong “Improving Performance of Database Appliances on Distributed Object Storage” 978-1-5090-0144-6/15 $31.00 © 2015 IEEE.

[6] Chamila Liyanage1, Thilini Nadungodage2, Ruvan Weerasinghe3 “Developing a commercial grade Tamil OCR for recognizing font and size independent text” 978-1-4673-8270-0/15/$31.00 ©2015 IEEE.

[7] M. Ramanan A. Ramanan E.Y.A. Charles “A Performance Comparison and Post-processing Error Correction Technique to OCRs for Printed Tamil Texts”.

[8] Chirag Patel, Atul Patel, PhD. Dharmendra Patel “Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study” Volume 55– No.10, October 2012.

[9] Artem Chebotko, Andrey Kashle, Shiyong Lu “A Big Data Modeling Methodology for Apache Cassandra” 978-1-4673-7278-7/15 $31.00 © 2015 IEEE.

[10] Xiaoming Gao, Judy Qiu “Supporting Queries and Analyses of Large-Scale Social Media Data with Customizable and Scalable Indexing Techniques over NoSQL Databases” 978-1-4799-2784-5/14 $31.00 © 2014 IEEE.