International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-04, July 2019

Data Preprocessing on Cassandra Data through Spark SQL

[*]Dr. K. UshaRani, [#]K. Anusha

[*]Professor, [#]Research Scholar, Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam, Tirupati, AP, India. usharanikuruba@yahoo.co.in, siri.bachina@gmail.com

Abstract: Big Data is a collection of large volumes of data generating from different sources. With this rising growth of data, now-a-days storage and processing of data are becoming very difficult. Apache Spark, an open source, general-purpose distributed computing engine used for processing and analyzing large amount of data. Similar to Hadoop Map Reduce, it also works with the system to allocate data across the cluster and process the data in parallel. The quality and representation of data is the first important one before running an analysis. If there is much irrelevant, noisy and unreliable data, then knowledge discovery is more difficult to conduct. Data preparation can take significant amount of processing time. Hence, in this paper, we presented a Big Data preprocessing system based on Spark platforms. The large amount of data is stored in Cassandra database and data cleaning, one of the data preprocessing techniques is applied on Cassandra data using Apache Spark. Performing data cleaning on raw data can reduce the processing time of application. Hence, in this study we experimented data cleaning on weather data using Spark SQL and Cassandra and made a comparison of Spark processing time with and without data cleaning.

Key Words: Big data, Apache Spark, Apache SparkSQL, Cassandra, Data Preprocessing, Data Cleaning.

I. INTRODUCTION

Big Data, as the name indicates, is a collection of huge amount of data which supports various types of data from different sources. Big data analytics is one of the important solutions for extracting useful information from large data sets. It is essentially used to discover hidden patterns, correlations, etc., from large data sets. In general, Hadoop Map Reduce is the commonly used framework for storing and processing of Big Data. But it is very difficult and time consuming task. So Apache Spark a unified processing engine and an efficient framework are used for processing of Big Data which reduces the processing time and also overcomes the limitations of Hadoop Map Reduce.

Apache Spark performs Big Data analytics which provides faster and more general data processing platform. Apache Spark runs programs up to 100 times faster than Hadoop. Apache Spark framework consists of different higher level libraries for Big Data Processing. Spark Machine Learning is one of the libraries of Spark which provides machine learning API built on the Data Frames. This can be used for developing and managing the Machine Learning Pipelines. The machine learning pipeline process involves different steps like Data Ingestion, Data Cleaning, Feature Extraction, etc,.

Pre-processing is a data mining technique that involves transferring raw data into an understandable format. World data are definitely inappropriate, inconsistent and / or lacking specific attitudes or trends, and likely there are a number of mistakes. Data processing is a methodology that addresses these issues. Preprocessing is preparing raw data for further processing.

Data pass through some of the steps during the previous treatment:

- Data Cleaning: Data is cleaned up by processes such as filling in missing values; smoothing the noisy data or resolving data discrepancies.
- Data Entry: Data with different representations is collected and data conflicts are resolved.
- Data Conversion: Data is usually collected and aggregated.
- Data Reduction: This step is intended to provide data reduction data in the database.
- Data Discretization: Includes the reduction of the attribute's attribute values, dividing the range from the attribute.

Data Cleaning, one of the data preprocessing techniques is the first and important step in the overall data analytics pipeline. It is used to detect and eliminate the errors, missing values and inconsistencies in the data and then improves data quality. The architecture of data cleaning is shown in the following figure 1[2].

 DOI : 10.35291/2454-9150.2019.0305

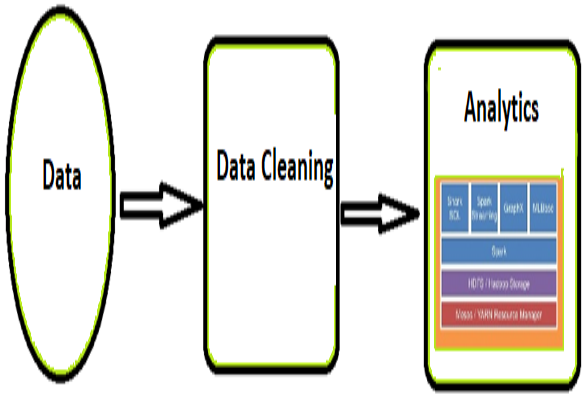

Fig 1: Data Cleaning Architecture [2]

Storage of Big Data is another major issue. Traditional relational databases are not so efficient to store Big Data. Cassandra, one of the NoSQL databases provides greater flexibility in storing Big Data.

Hence, in this study data cleaning is performed on Cassandra data using Spark SQL which enhances the processing speed.

II. LITERATURE REVIEW

Huadong Dai et.al [5] proposed a Big Data Preprocessing system based on Hadoop in which the system utilizes idle computing resources in cluster storage nodes to perform preprocessing works in parallel.

Ashish R. Jagdale et.al [6] proposed a Data Preprocessing for Big Data and executed under Apache Hadoop Framework. They presented a pre-processing algorithm to extract real time user accessed data from windows operating system environment and an approach from Apache's Hadoop Distributed File System (HDFS) framework using Map Reduce functionality to mine and analyze this large dataset.

Xindong Wu et.al [7] proposed several data mining techniques with Big Data and represents the performance evaluation.

K.Anusha et.al [8] proposed an integrated approach on weather data and made a comparison of processing time of extracting data from external sources using both Hadoop Map Reduce and Spark processing frameworks.

Changming Zhu et.al [9] explains about influence of data preprocessing and concluded that using different preprocessing methods gives different classification performances.

Sumian Peng et.al [10] conducted a detailed study of preprocessing process in Web log mining.

Ming-hua Zhu [11] analyses the exam analysis system in detail and introduced a general technique of data preprocessing.

Fatin Zulkepli et.al [12] shows data preprocessing techniques used to produce clean and quality data for University Technology Malaysia (UTM) research performance analysis.

Vivek Agarwal et.al[13] aims to highlight the data preprocessing steps required for review analysis of a newly launched smartphones in the market by collecting tweets from the Twitter data feed.

K.Anusha et.al [15] proposed an experiment about comparison of integrated approaches for Batch Processing.

Sapna Devi1et.al [16] provides an overview of data cleaning and comparison of data cleaning tools.

Yang Bao et.al [17] analyzes the types and causes of dirty data, and proposes several key steps of typical cleaning process.

Shivangi Rana et.al [18] presents a survey of sources of error in data, data quality challenges and approaches, data cleaning types and techniques and an overview of ETL Process.

III. PROPOSED WORK

The proposed system is experimented using Apache Spark Cassandra Connector environment which access the data from Cassandra database through Apache Spark shell, preprocess the data and then generates the output dataset. The block diagram of the proposed system is shown in the following figure 2.

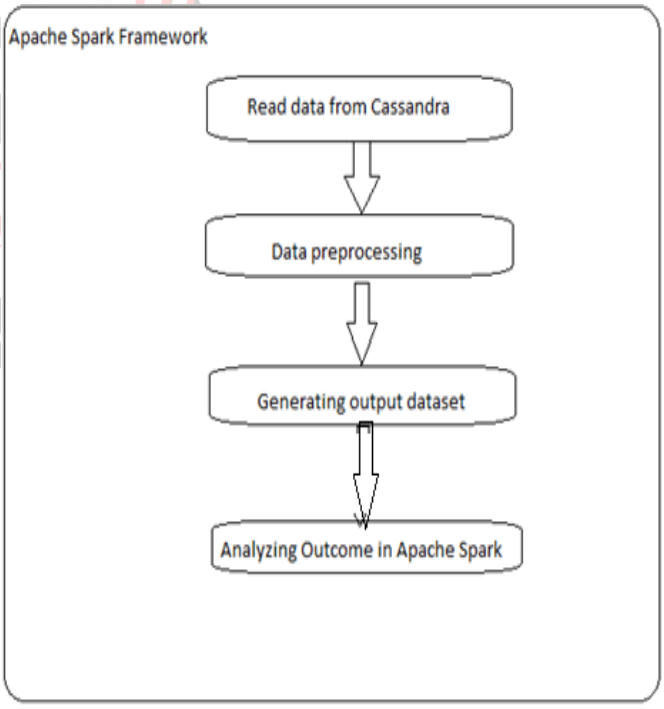

Fig 2: Block Diagram of Proposed System

The important phases of proposed system are:

- Data Collection: In this phase the batch data is collected and loaded into Cassandra NoSQL database. In this study we experimented on Weather data. Weather data is collected from National Climatic Data Centre (NCDC) and loaded into Cassandra database.

 DOI : 10.35291/2454-9150.2019.0305

- Data Preprocessing: Data cleaning, called data filtering, data cleansing or data manipulation is the process of collecting data and making it available in our favorite statistics program. Cleaning includes poor data removal, creating a valid label and code and everything. Sometimes it is inevitable to collect incompatible data. In this study we experimented Data cleaning and is applied on the data stored in Cassandra using Apache Spark.
- Generation of Output dataset: After applying data cleaning on Cassandra data, the output is generated in Apache Spark shell and the output is analyzed and then compares the processing time of reading data from Cassandra with and without data cleaning.

IV. EXPERIMENTAL SETUP AND RESULTS

The configuration we used to conduct an experiment in this proposed system is:

Ubuntu Operating system, Intel Core i7 Processor. We have experimented this proposed system on Weather data containing ten lakhs of records and the output is evaluated.

The results obtained in proposed system i.e., applying data cleaning on Cassandra data using SparkSQL are compared with the results obtained in our previous study i.e., SparkSQL-Cassandra Connector [15]. The comparison of Spark results with and without Data Cleaning is shown in the following table 1.

Metrics	SparkSQL-Cassandra without Data Cleaning	SparkSQL-Cassandra with Data Cleaning
Duration	0.7s	0.3s
Scheduler Delay	5ms	4ms
Task Deserialization Time	8ms	3ms
Result Serialization Time	1ms	0ms

Table 1: Comparison of Experiment Results

The following figure 3 shows the graphical representation of the analysis.

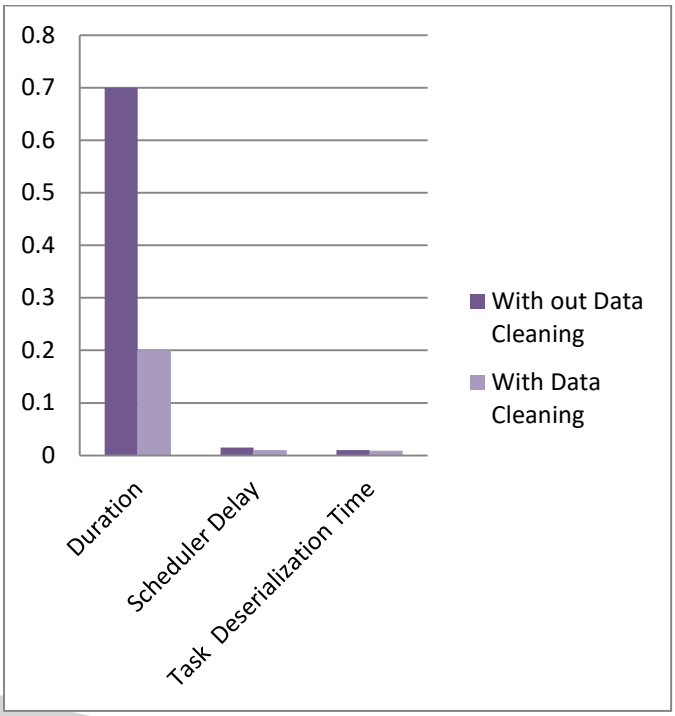

Fig 3: Comparison of Experiment Results of SparkSQL-Cassandra Without and With Data Cleaning

V. PERFORMANCE ANALYSIS

The performance analysis of this proposed system is explained based on various factors like Processing time, Scalability and flexibility of the data. Time taken for getting the desired output quickly and efficiently defines the efficiency and time complexity of the algorithm used. The reduction of scheduler delay and Task Deserialization time improves the scalability and flexibility of data. In general, the traditional data mining algorithms are very time consuming if applied on large data sets with more number of records as compared to the machine learning library in Apache Spark Framework.

It also improves the Scalability and Flexibility of data by scaling the data from thousands to millions of records without reducing the performance. Spark also can handle large amount of data in parallel and the workloads are simultaneously distributed on different executors which makes it more flexible.

VI. CONCLUSION

Analyzing Big Data is a very challenging and complicating task. Apache Spark SQL and Apache Spark Machine Learning plays very important role in handling and processing of large datasets. A big data preprocessing system based on Spark was designed and experimented in this study. Apache Spark Data Cleaning plays an important role in the proposed system in terms of processing time, Scalability and Flexibility. It also enhances the performance of SparkSQL Cassandra Connector and makes a comparison of an experiment with and without Data Cleaning.

 DOI : 10.35291/2454-9150.2019.0305

REFERENCES

[1] https://www.infoq.com/articles/apache-sparkml-data-pipelines

[2] https://www.slideshare.net/jeykottalam/sample-clean-amp-camp-demov9

[3] https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-baseddatasets/quality-controlledlocal-climatological-data-qclcd.

[4] Apache Cassandra [Online]. Available: https://www.datastax.com/wp-content/uploads/2012/09/WPDataStax HDFSvsCFS.pdf

[5] Huadong Dai, Shu Zhang, Li Wang, Yan Ding "Research and Implementation of Big Data Preprocessing System Based on Hadoop" National University of defense technology.

[6] Ashish R. Jagdale, Kavita V. Sonawane "Data Mining and Data Pre-processing for Big Data" International Journal of Scientific & Engineering Research, Volume 5, Issue 7, July-2014

[7] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", (In Press) IEEE Transactions on Knowledge and Data Engineering, 2013.

[8] K.Anusha, K.UshaRani "Big Data Techniques for Efficient Storage and Processing of Weather Data" International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; Volume 5 Issue VII, July 2017.

[9] Changming Zhu, Daqi Gao "Influence of Data Preprocessing" Journal of Computing Science and Engineering, Vol. 10, No. 2, June 2016.

[10] Sumian Peng ; Qingqing Cheng "Research on Data Preprocessing Process in the Web Log Mining" The 1st International Conference on Information Science and Engineering (ICISE2009).

[11] Ming-hua Zhu "Research on Data Preprocessing in Exam Analysis System" Communication Systems and Information Technology.

[12] Fatin Zulkepli, Faisal Saeed "Data Preprocessing Techniques for Research Performance Analysis" Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2016.

[13] Vivek Agarwal, Kashibai Navale, 'Ram Darshan "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis" International Journal of Computer Applications (0975 – 8887) Volume 131 – No.4, December2015.

[14] https://www.dataquest.io/blog/data-retrieval-and-cleaning/

[15] K.Anusha, K. Usha Rani "Performance Evaluation of Spark SQL for Batch Processing" accepted for publication in Springer series "Advances in Intelligent Systems and Computing".

[16] Sapna Devi1, Dr. Arvind Kalia "Study of Data Cleaning & Comparison of Data Cleaning Tools" International Journal of Computer Science and Mobile Computing, Vol.4 Issue.3, March- 2015.

[17] Yang Bao, Shi Wei Deng, Wang Qun Lin "Research of Data Cleaning Methods Based on Dependency Rules" World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:9, No:10, 2015.

[18] Shivangi Rana, Er. Gagan Prakesh Negi, Kapil Kapoor "A Study over Importance of Data Cleansing in Data Warehouse" International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 4, April 2016 ISSN: 2277 128X.

 DOI : 10.35291/2454-9150.2019.0305