IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 2, April 2014.
www.ijiset.com
ISSN 2348 - 7968

Framework for Data Cleaning on Weaknesses and To Eradicate Biases in Their Interpretation of the Big Data

Sowmiya Muthukumar
PG scholar,
Department of CSE,
R.M.K Engineering college,
Kavaraipettai, Chennai

Dr R.Jagadeesh Kannan,B.E,M.E Ph.D
Professor,
Department of CSE,
R.M.K Engineering college
Kavaraipettai,Chennai

Abstract: The data becomes the powerhouse of Information for major analysis in public and private entities. More number of unstructured data is getting generated every day, and these peta bytes of information are getting stored as Big Data storage and it poses a biggest challenge for analyzers to retrieve and effectively uses the Information for Statistical claims, as the major issue is of Data errors. Unreliable data which are bound to outages and losses are getting stored and there is no mechanism or framework to clean the weaknesses of the data based on the attributes identified. This Research brings the effective framework for Data cleaning and ensures the data used in the analysis are Proper reliable data

1.INTRODUCTION

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 quintillion (2.5×10^{18}) bytes of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead massively parallel software running on tens, hundreds, or even thousands of servers. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

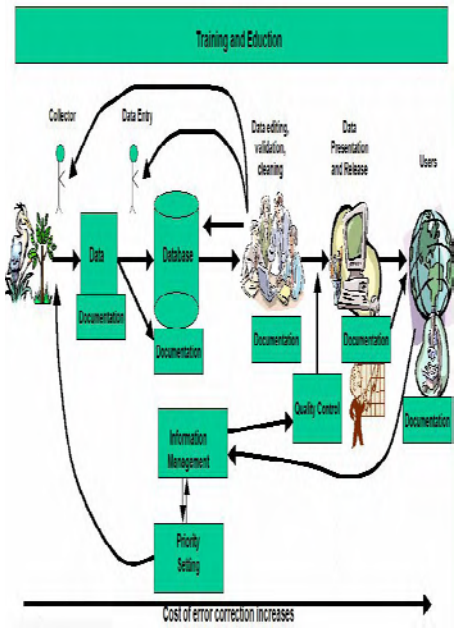
1.1 DATA CLEANING

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions.

The general framework for data cleaning is:

- Define and determine error types
- Correct the errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

ISSN 2348 - 7968



2.RELATED WORK

Antorweep chakravorty et al., (2013) discuss about Privacy Preserving Data Analytics for Smart Homes that Demonstrate a solution for reliably concealing privacy and ensuring security for analytics of data. A framework for maintaining security & preserving privacy for analysis of sensor data from smart homes, without compromising on data utility was presented.

Taoxin (2010) discussed about Framework for Data Cleaning in Data Warehouse. It is a persistent challenge to achieve a high quality of data in data warehouses. Data cleaning is a crucial task for such a challenge. To deal with this challenge, a set of methods and tools has been developed.

Anurada Bhatia and Gaurav Vaswani (2013) presented the analysis of large amount of data for making advance in many scientific disciplines. As more data becomes available from an abundance of sources both within and outside, organizations are seeking to use those abundant resources to increase innovation, retain customers, and increase operational efficiency.

Michal Shmueli et al., (2010) discussed the details of a large scale user profiling framework that we developed here in IBM on top of Apache Hadoop. The author addressed the problem of extracting and maintaining a very large number of user profiles from large scale data.

Chansup Byun et al., (2012) discus about Driving Big Data with Big Compute. Big Data (as embodied by Hadoop clusters) and Big Compute (as embodied by MPI clusters)provide unique capabilities for storing and processing large volumes of data.

3.PROPOSED SYSTEM

A.SYSTEM MODEL

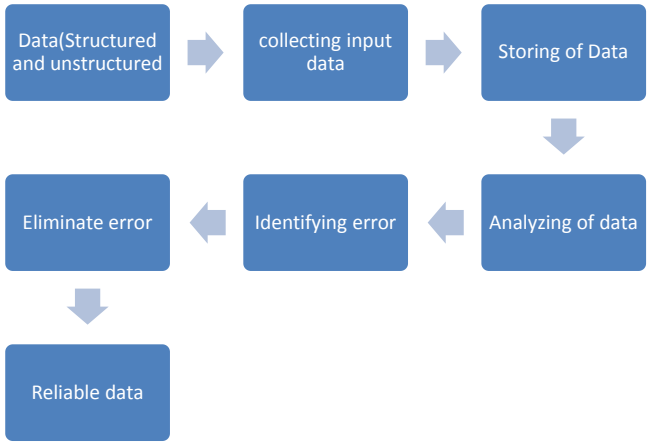


Figure describes the framework for data cleaning and finding error in generating big data.

The input of the block diagram is Structured and Unstructured data. The structured data generally refers to data that has a defined length and format for big data.

The Unstructured data usually refers to information that doesn't reside in a traditional row-column database Unstructured data files often include text and multimedia content.

These data are collecting as input for the big data database that can be used for processing in big data. The data that are collected was used to store the data using hadoop and eclipse. Once data has been stored in database has to send for data cleaning. Data cleaning is used to identify and eliminate the error. In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required. The error data are eliminating

using ETL tool and this will provide reliable data that will be stored in big data storage.

3.2 WORKING METHODOLOGY

3.2.1 Collecting and Storing Data

Collecting data means putting your design for collecting information into operation. Data collection process is necessary as it ensures that data gathered are both defined and accurate. Types of collection include census, sample survey, and administrative by-product. The way you collect your data should relate to how you're planning to analyze and use it. Regardless of what method you decide to use, recording should be done concurrent with data collection if possible, or soon afterwards, so that nothing gets lost and memory doesn't fade. Hadoop is developing in the machine in order to store large amount of data. Hadoop launches a Map Reduce job by first splitting (logically) the input dataset into data splits. Each data split is then scheduled to one Task Tracker node and is processed by a map task.

3.2.2 Analyzing Data

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Qualitative data are collected as descriptions, anecdotes, opinions, quotes, interpretations, etc., and are generally either not able to be reduced to numbers, or are considered more valuable or informative if left as narratives. Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

3.2.3 Identifying Error

The data cleansing process takes two inputs

1. Data required to be cleaned
2. Rules of cleansing from rules configuration database

This is the area where actual data cleansing processing done based on rules from rules configuration repository and output of this process provides error-free and consistent data that is ready to load into data warehouse. This output data is standardized, uniform, accurate and complete with accordance to business. The cleaned data not only provides data quality but expedite the processing speed and performance of overall ETL process.

3.2.4 Eliminating Error

Data cleaning require insight into the sources and types of errors at all stages of the study, during as well as after measurement. Once errors are detected, it is important to know how to handle them appropriately so the data can be analyzed without losing their integrity or robustness. This makes data cleansing an iterative process involving significant exploration and interaction, which may require a framework in the form of a collection of methods for error detection and elimination in addition to data auditing.

This paper mainly focuses on data extraction and data cleaning algorithms. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data, discussed a preprocessing algorithm for data cleaning, user identification and session identification.

.

4.CONCLUSION

Data cleaning is major issue of Big data. In this paper present a secure solution for providing data quality for the data that stored in big data. This project also brings out the platform for conducting further Big Data Research also acts as lessons learnt for future research to progress. Big data statistical analysis plays a vital role in Social related hubs like Judicial, Medical and Education related analysis. Big Data reports are used for Government to take appropriate actions by using the mined data. Decision making on Statistical Peta Bytes of Big Data will be proper and closest accurate results are derived by using this Framework. Processing time over Peta Bytes of data will be reduced by Replication and Parallelization strategy. High Performance computing provides a platform for further Researchers/Academicians/Students to continue the Research on Big Data without investing their time on Configuration Roadblocks.

5.REFERNCES

[1] Chansup Byun, William Arcand, David Bestor, Bill Bergeron , Driving big data with big compute , In the 2012 IEEE Conference on High

ISSN 2348 - 7968

Performance Extreme Computing (HPEC), Pages: 1 – 6, 2012.

[2] N.Laptev, Kai Zeng and C. Zaniolo., Very fast estimation for result and accuracy of big data analytics: The EARL system, In the IEEE 29th International Conference on, Data Engineering (ICDE), pages: 1296 - 1299, 2013.

[3] Nikolay Laptev, Kai Zeng, Carlo Zaniolo, Early Accurate Results For Advanced Analytics On Mapreduce, In the 38th International Conference on Very Large Data Bases, Proceedings of the VLDB Endowment, Vol. 5, No. 10 2012.

[4] S.Sagiroglu and D.Sinanc, Big data: A review, In the 2013 International Conference on Collaboration Technologies and Systems (CTS), pages: 42 – 47, 2013.

[5] Taoxin Peng, A Framework For Data Cleaning In Data Warehouse, In the 10th International Conference on Enterprise Information Systems (ICEIS), pages:473-478 ,2008.

[6] A.Chakravorty , T. Wlodarczyk and Chunming Rong , Privacy Preserving Data Analytics for Smart Homes , 2013 IEEE Security and Privacy Workshops (SPW), Page(s): 23 – 27, 2013.

[7] Holzinger, Andreas; Pasi, Gabriella (Eds.), Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Proceedings of Third International Workshop, pages: 136-442, 2013.

[8] Yaxiong Zhao ; Jie Wu , Dache: A data aware caching for big-data applications using the MapReduce framework , Proceedings of IEEE on INFOCOM, Page(s): 35 – 39, 2013.

[9] Michal Shmueli, Haggai roitman, David carmel, Extracting User Profiles From Large Scale Data, Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud, 2010.

[10] Hadley Wickham, Bin Summarise Smooth: A Framework for Visualising Large Data, An Extended Repository Approach. Information Systems, 24(3),2013.