International Journals of Advanced Research in Computer Science and Software Engineering
ISSN: 2277-128X (Volume-9, Issue-4)

Research Article | **April 2019**

A Review on Big Data Cleaning and Analytical Tools

K. Dhinakaran
Department of CSE,
M. N. M. Jain Engineering College, Chennai, Tamilnadu, India
dhina79@gmail.com

G. Geetharamani
Department of Mathematics,
Anna University, BIT Campus, Tiruchirappalli, Tamilnadu, India
geeramdgl@rediffmail.com

Abstract- Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. The categories are structured, semi-structured and unstructured data. Data analysis is a collective term of gathering, organizing and analyzing data for present and future improvements. It also refers to manipulation and analysis of the large volume of data such that big data is of course a complex process. Collecting, analyzing, searching, storing and sharing of big data is a challenging task using modern big data analytic tools. In short, such data is so large and complex that none of the traditional data management tools are capable to store it or process it efficiently. This paper provides some of the cleaning, storing and analytic tools to handle big data.

Keywords- Big Data, Data Analysis, Structured Data, Unstructured Data, Challenging Tasks

I. INTRODUCTION

Big data is a data set which holds huge volume of data both structured and unstructured that exceed the range of Exabyte where the traditional methods of data processing software find inadequate to deal with them. In Business point of view, importance of big data can be well understood by how an organization utilizes the collected data instead of discussing how much data is stored by a company. The data can be taken from any source like social interaction where the people can collect the reviews of product and process. Before processing the data, it should be cleaned without noise, duplication and invalid data. Then only the desired pattern or information can be extracted from the huge data set.

II. THE 7 V'S OF BIG DATA

(i) Volume – The name Big Data itself is related to the size of which is enormous. Size of data plays a very crucial role in determining the value out of data. Whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data.

(ii) Variety – This refers to heterogeneous sources and the nature of data, both structured and unstructured. During the earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis of applications. The variety of unstructured data poses certain issues for storage, mining and analysis of data.

(iii) Velocity – The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential of the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, mobile devices, etc. The flow of data is massive and continuous.

(iv) Variability– This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

(v) Veracity – It is stated that the quality or trustworthiness of the data.

(*vi) Visualization* – It refers to the visualization of the complex data using charts and graphs.

(*vii) Value* - This refers to the ability to transform a tsunami of data into business.

III. COMPONENTS OF A BIG DATA ARCHITECTURE

The following diagram shows the logical components that fit into a big data architecture.

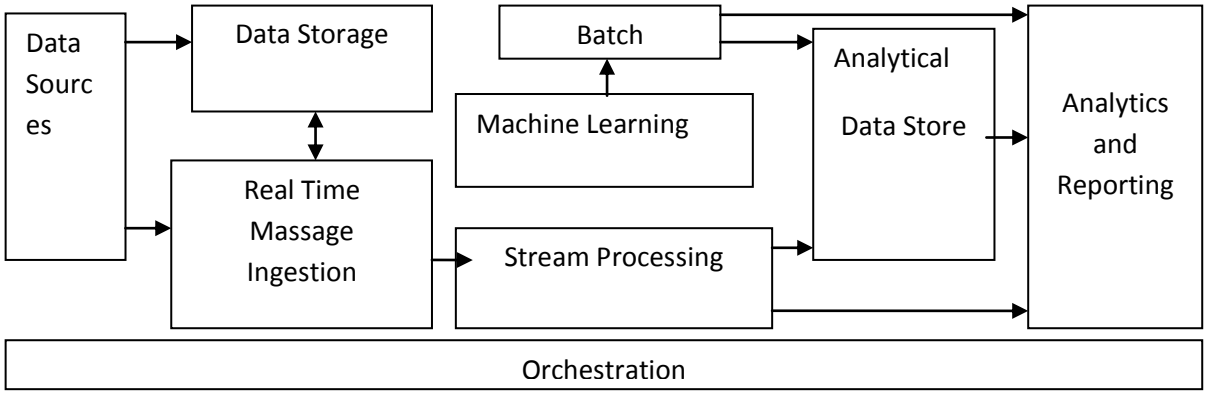

Fig. 3.1 Components of Big Data Architecture

Most big data architectures include some or all of the following components:

Data sources: All big data solutions start with one or more data sources. Examples include:

a) Application data stores, such as relational databases.
b) Static files produced by applications, such as web server log files.
c) Real-time data sources, such as IoT (Internet of Things) devices.

Data storage: Data for batch processing operations are typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a data lake. Options for implementing this storage include Azure Data Lake Store or blob containers in Azure Storage.

Batch processing: Since the data sets are so large, often a big data solution must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually, these jobs involve reading source files, processing them, and writing the output to new files.

Real-time message ingestion: If the solution includes real-time sources, the architecture must include a way to capture and store real-time messages for stream processing. This might be a simple data store, where incoming messages are dropped into a folder for processing. However, many solutions need a message ingestion store to act as a buffer for messages, and to support scale-out processing, reliable delivery, and other message queuing semantics. This portion of a streaming architecture is often referred to as stream buffering.

Stream processing: After capturing real-time messages, the solution must process them by filtering, aggregating, and otherwise preparing the data for analysis. The processed stream data is then written to an output sink. Azure Stream Analytics provides a managed stream processing service based on perpetually running SQL (Structured Query Language) queries that operate on unbounded streams.

Analytical data store: Many big data solutions prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools. The analytical data store that is used to serve these queries can be a Kimball-style relational data warehouse, as seen in most traditional Business Intelligence (BI) solutions.

Analysis and reporting: The goal of most big data solutions is to provide insights into the data through analysis and reporting. To empower the users and to analyze the data, the architecture may include a data modeling layer, such as a multidimensional OLAP (Online Analytical Processing) cube or tabular data model in Azure Analysis Services. It might also support self-service BI, using the modeling and visualization technologies in Microsoft Power BI or Microsoft Excel. Analysis and reporting can also take the form of interactive data exploration by data scientists or data analysts.

Orchestration: Most big data solutions consist of repeated data processing operations, encapsulated in workflows that transform source data, move data between multiple sources and sinks, load the processed data into an analytical data store, or push the results straight to a report or dashboard.

3.1 Data Processing

Data processing is simply the conversion of raw data to meaningful information through a process.

Stages of the Data Processing Cycle:

Collection is the first stage of the cycle, and is very crucial, since the quality of data collected will impact heavily on the output. The collection process needs to ensure that the data gathered are both defined and accurate, so that subsequent decisions based on the findings are valid. This stage provides both the baseline from which to measure, and a target on what to improve.

Preparation is the manipulation of data into a form suitable for further analysis and processing. Raw data cannot be processed and must be checked for accuracy. Preparation is about constructing a data set from one or more data sources

to be used for further exploration and processing. Analyzing data that has not been carefully screened for problems can produce highly misleading results that are heavily dependent on the quality of data prepared.

Input is the task where verified data is coded or converted into machine readable form so that it can be processed through an application. Data entry is done through the use of a keyboard, scanner, or data entry from an existing source. This time-consuming process requires speed and accuracy. Most data need to follow a formal and strict syntax since a great deal of processing power is required to breakdown the complex data at this stage. Due to the costs, many businesses are resorting to outsource services.

Processing is the stage where the data is subjected to various means and methods of powerful technical manipulations using Machine Learning and Artificial Intelligence algorithms to generate an output or interpretation about the data. The process may be made up of multiple threads of execution that simultaneously execute instructions, depending on the type of data.

***Output* and interpretation** is the stage where processed information is now transmitted and displayed to the user. Output is presented to users in various report formats like graphical reports, audio, video, or document viewers. Output need to be interpreted so that it can provide meaningful information that will guide future decisions of the company.

Storage is the last stage in the data processing cycle, where data and metadata are held for future use. The importance of this cycle is that it allows quick access and retrieval of the processed information, allowing it to be passed on to the next stage directly when needed.

3.2 Benefits of Big Data Processing

Ability to process Big Data brings in multiple benefits, such as

a) Businesses can utilize outside intelligence while taking decisions
b) Improved customer service
c) Early identification of risk to the product/services
d) Better operational efficiency

IV. BIG DATA CLEANING TOOLS

1. OpenRefine

Formerly known as Google Refine, this powerful tool comes handy for dealing with messy data, cleaning and transforming them. It's a good solution for those looking for free and open source data cleansing tools and software programs. It can also transform data from one format to another, letting the big data sets to explore with ease, reconcile and match data, clean and transform at a faster pace.

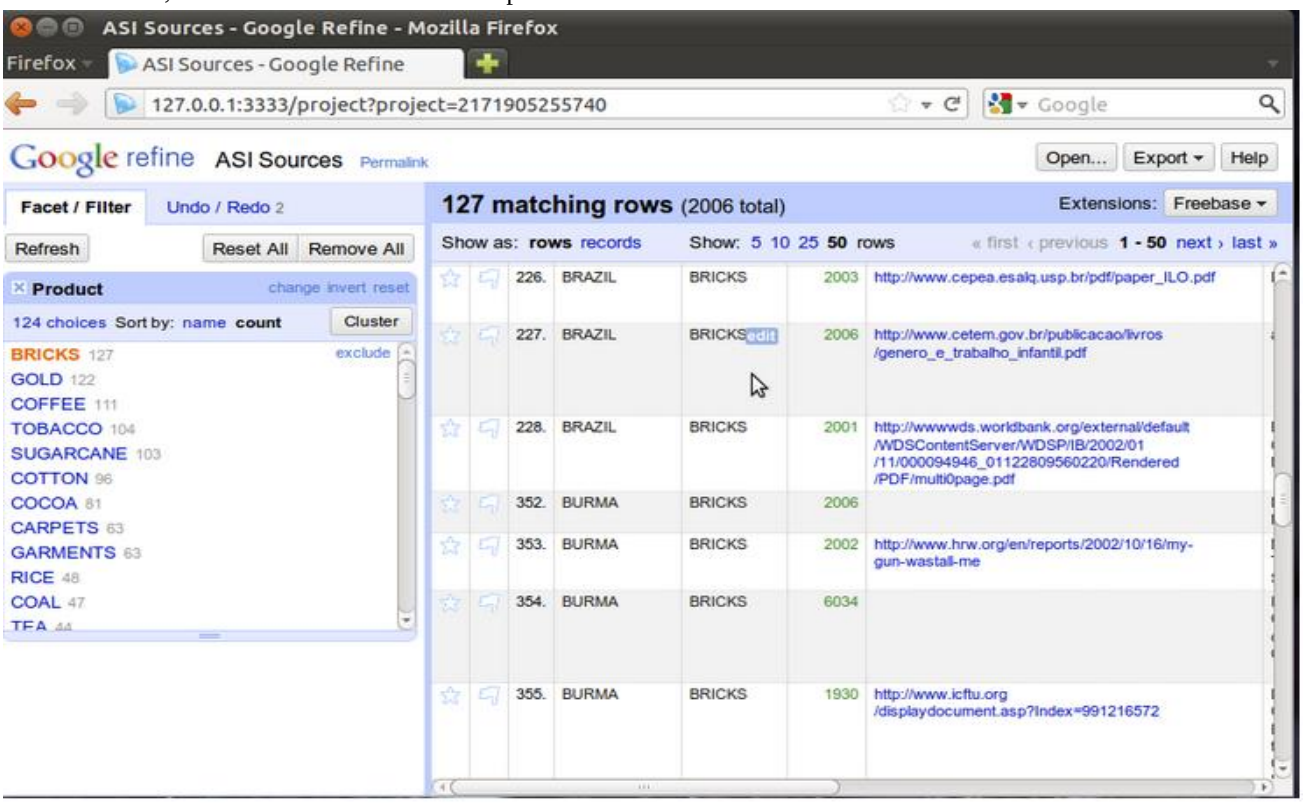

Fig. 4.1 Duplicate Record cleaning from OpenRefine Tool

2. Trifacta Wrangler

A venture started by the makers of Data Wrangler, it is an interactive tool for data cleaning and transformation. One of the best features of this tool includes less formatting time and larger focus on analyzing data. It helps data analysts in cleaning and preparing messy, diverse data more quickly and accurately. Its machine learning algorithms help in preparing data by suggesting common transformations and aggregations. This also comes at free of cost.

3. Drake

This simple to use, extensible, text based data workflow has data processing steps defined along with their inputs and output, where it can automatically resolve their dependencies and calculate the command to execute and the order that it should be executed. It is designed especially for data workflow management and organizes command execution around data and its dependencies.

4. TIBCO Clarity

This data cleaning tool offers on demand software services from the web in the form of Software-as-a-service. It allows the users to validate the data, in reduplication and cleansing addresses to help identify trends quickly and make smarter decisions. It can standardize raw data collected from disparate sources to provide good quality data for accurate analysis.

5. Winpure

It is one of the most popular and affordable data cleaning tools accomplishing the task of cleaning a large amount of data, removing duplicates, correcting and standardizing effortlessly. It can clean data from databases, spreadsheets, CRMs and more, and can be used for databases like Access, Dbase, SQL Server, and Txt files. Some of its key features include advanced data cleansing and fuzzy matching, super fast data scrubbing, multi language edition available, among others.

6. Data Ladder

It offers products Data Match, an affordable cleaning & data quality tool and Data Match Enterprise that includes advanced fuzzy matching algorithms for up to 100 million records, and has one of the highest matching accuracies and speed in the industry. These user friendly tools help businesses from any size and any industry to manage their data cleansing processes with ease.

7. Data Cleaner

Quadient Data Cleaner is a strong data profiling engine for analyzing the quality of data to drive better business decisions. The tool can find missing values, patterns, character sets and other characteristics in a data set to offer better results. With a strong profiling engine, it can detect duplicates using fuzzy logic and create single version of it. It also helps in building individual cleansing rules and composes them into several scenarios to target databases.

8. Cloudingo

This Salesforce data cleansing tool eliminates duplicates, cleans records, and maintains data quality in one place. It is suitable for business of all sizes, where the data is updated in bulk, and imported files are cleansed before accessing Salesforce. Its automation capabilities ensure that data is regularly scanned for errors. Some of its features are its simplicity, deleting unnecessary and stale records, update records in bulk, automate on a schedule, among others.

9. Reifier

With features like high accuracy, fast deployment, run time performance and others, Reifier by Nube Technologies utilizes Spark for distributed entity resolution, reduplication and record linkage. It uses machine learning algorithms to provide the best entity resolution and fuzzy data matching with a scale out distributed architecture.

10. IBM Infosphere Quality Stage

Designed to support data quality, it is one of the most popular data cleansing tools and software solutions for supporting full data quality. It allows cleansing and managing database with much ease, and build consistent views for the most important units such as customers, vendors, products, locations etc. It helps in delivering quality data for big data, business intelligence, data warehousing, master data management etc.

V. BIG DATA STORING & ANALYZING TOOLS

1. Apache Hadoop

Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster.

2. MicrosoftHDInsight

It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability at low cost.

3. NoSQL

While the traditional SQL can be effectively used to handle large amount of structured data. This NoSQL (Not Only SQL) is used to handle unstructured data. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs (Databases) available to analyze big Data.

4. Hive

This is a distributed data management for Hadoop. This supports SQL-like query option HiveSQL (HSQL) to access big data. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

```
hive> select count(*) from txnrecords;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201402270420_0005, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201402270420_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201402270420_0005
2014-02-28 20:02:41,231 Stage-1 map = 0%,  reduce = 0%
2014-02-28 20:02:48,293 Stage-1 map = 50%,  reduce = 0%
2014-02-28 20:02:49,309 Stage-1 map = 100%,  reduce = 0%
2014-02-28 20:02:55,350 Stage-1 map = 100%,  reduce = 33%
2014-02-28 20:02:56,367 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201402270420_0005
OK
50000
Time taken: 19.027 seconds
hive>
```

Fig. 5.1 select command of HiveSQL

5. Sqoop

This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

6. PolyBase

This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a data warehousing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

7. Big data in EXCEL

As many people are comfortable in doing analysis in EXCEL, Hadoop data can be connected using EXCEL 2013. Horton works, which is primarily working in providing Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. Power View feature of EXCEL 2013 can be used easily to summarize the data.

8. Presto

Facebook has recently developed the open-sourced its Query engine (SQL-on-Hadoop) named Presto to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

VI. CONCLUSIONS

In this survey, the components of big data architecture and various cleaning, storing and analyzing tools have been analyzed, discussed and presented. By using this tool, various types of data can be analyzed to enhance different types of businesses. These tools play a major role in the fields of Banking, Health Care, Products Sales, Fraud Detection, Tourism, Airline, Human Resources, Finance, Telecom, Retail, Customer response in online sales, etc. The study,

analysis and implementation of big data analytics have become a mandate in the present scenario and the above mentioned tools will be helpful for technologists and researchers dealing with big data.

REFERENCES

[1] K. Kambatlaa, G. Kollias, V. Kumar, A. Grama. Trends in big data analytics. J. Parallel Distrib. Comput. vol74, 2014:2561-2573.

[2] V. M. Schönberger, K. Cukier Big data: a revolution that will transform how we live, work and think[M]. John Murray Publishers Ltd ,2013

[3] A. P. Silva , G. R. Mateus. A Location-Based Service Application for a Mobile Computing Environment. Computer Science. 2003(79): 343-360.

[4] Mariam Adedoyin-Olowe1 et.al "A Survey of Data Mining Techniques for Social Media Analysis "

[5] Chu, Cheng, et al. "Map-reduce for machine learning on multicore." Advances in neural information processing systems 19 (2007)

[6] Groves, Peter, Basel Kayyali, David Knott, and Steve Van Kuiken. "The big data revolution in healthcare." McKinsey Quarterly 2013.

[7] Shang, Weiyi, Zhen Ming Jiang, HadiHemmati, Bram Adams, Ahmed E. Hassan, and Patrick Martin. "Assisting developers of big data analytics applications when deploying on Hadoop clouds." In Proceedings of the 2013 International Conference on Software Engineering, pp. 402-411. IEEE Press, 2013.

[8] Aggarwal, N., Liu, H.: Blogosphere: Research Issues, Tools, Applications. ACM SIGKDD Explorations. Vol. 10, issue 1, 20, 2008.

[9] Boiy, E., Hens, P., Deschacht, K., Marie-Francine, M.: Automatic Sentiment Analysis of On-line Text. In: Proceedings of the 11th International Conference on Electronic Publishing. Vienna, Austria**,** 2007.

[10] Boyd, D. M. and Ellison, N. B.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer Mediated Communication, 13: 210–230. doi: 10.1111/j.1083- 6101.2007.00393.x, 2007.

[11] Castellanos, M., Dayal, M., Hsu, M., Ghosh, R., Dekhil, M.: U LCI: A Social Channel Analysis Platform for Live Customer Intelligence. In: Proceedings of the 2011 international Conference on Management of Data. 2011.

[12] Chakrabarti, S.: Data Mining for Hypertext: A Tutorial Survey. ACM SIGKDD Explorations, 1(2):1-11. 2000.

[13] Chaomei, C., Ibekwe-SanJuan, F., SanJuan, E., Weaver, C.: Visual Analysis of Conflicting Opinions. In: 2006 IEEE Symposium On Visual Analytics And Technology: 59-66. 2006.

[14] Chaovalit, P., Zhou, L.: "Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches," In: Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2005.

[15] Chen, Z. S., Kalashnikov, D. V. and Mehrotra, S. Exploiting context analysis for combining multiple entity resolution systems. In Proceedings of the 2009 ACM International Conference on Management of Data (SIGMOD'09), 2009.

[16] Chen, Y., Lee, K.: User-Centred Sentiment Analysis on Customer Product Review. World Applied Sciences Journal 12 (special issue on computer applications & knowledge management) 32 – 38, 2011. ACM, New York, NY USA, 2011.

[17] Chi, Y., Zhu, S., Hino, K., Gong. Y., Zhang. Y.: iOLAP: A Framework for Analyzing the Internet, Social Networks, and Other Networked Data. *Multimedia. IEEE Transactions on,* 11(3):372 – 382, 2009.

[18] Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of WWW 519-528, 2003.

[19] Ding, X., B. Liu, Yu, P.: A Holistic Lexicon-based Approach to Opinion Mining. In: Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.

[20] Esuli, A., Sebastiani. F.: Determining the Semantic Orientation of Terms through Gloss Classification. In: Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2005), 2005.

[21] Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: Mining Customer Opinions from Free Text. Advances in Intelligent Data Analysis VI, pages 121–132, 2005.

[22] Godbole, N., Srinivasaiah, M., Steven, S.: Large Scale Sentiment Analysis for News and Blogs. In: Proceedings of the International Conference on Weblogs and SM (ICWSM), 2007.

[23] Daniel J. Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, Stan Zdonik, Aurora: a new model and architecture for data stream management, VLDB J. 12 (2) (2003) 120–139.

[24] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Alexander Rasin, Avi Silberschatz, HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads, in: VLDB, 2009.

[25] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, Andrew Tomkins, Pig Latin: a Not-So-Foreign language for data processing, in: SIG-MOD, ACM, 2008. ID: 1376726.

