IJRSET 2018 Volume 5, Special-Issue-2 Pages: 26-29

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

BIG DATA REFERENCE ARCHITECTURE AND CLASSIFICATION OF DATA PROCESSING IN SOCIAL MEDIA BIG DATA SYSTEMS

[1]B. Karthick, [2] V. Rajamanickam
[1] HOD, [2] Assistant Professor,
[1, 2] Department of Inforamtion Technology
[1, 2] Syed Hameedha Arts & Science College,
[1, 2] Kilakarai-623 806.

ABSTRACT: Big data is a term for data sets that are so large or complex. It is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. The rapid development data analysis on Social Networking of Big Data brings revolutionary changes to our daily lives and global business, which has been addressed by recent research. Adopting reference architecture within social media Big Data system accelerates delivery through the reuse of an effective solution and provides a basis for governance to ensure the consistency and applicability of technology use within an organization. Big Data processing techniques analyze Big Data sets at terabyte or even petabyte scale. This leads to the classification of Big Data Systems based on the processing styles. The objective of this paper is to map different architectures followed by the social media Big Data systems to the reference architecture and classify the data processing based on the processing type used by the architecture which helps in quality evaluation of Big Data as well.

Keywords: [Big Data, Reference Architecture, Stream and Batch processing]

1. INTRODUCTION

The primary goal of Big Data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modellers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things [1].Semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of Big Data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyse Big Data have turned to a newer class of technologies that includes Hadoop and related tools such as

YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases[3]. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

1.1 Reference Architecture

A reference architecture in the field of software architecture or enterprise architecture provides a template solution for architecture for a particular domain. Software reference architecture is a software architecture where the structures, respective elements and relations provide templates for concrete architectures in a particular domain or in a family of software systems.[2,9]A reference architecture often consists of a list of functions, some indication of their interfaces (or APIs), interactions with each other and with functions located outside of the scope of the reference architecture. It can be defined at different levels of abstraction.

- In high abstract one might show different pieces of equipment on a communications network, each providing different functions.
- In low level one might demonstrate the interactions of procedures (or methods) within a computer program defined to perform a very specific task [4, 11].

1.2. Processing Types

The structure of reference architecture differs slightly based on the processing type used. The data processing in the architecture can be classified in two major categories as follows:

1.2.1 Stream Processing

Data stream processing is the in-memory, record-by-record analysis of machine data in motion. The objective is to extract actionable intelligence as streaming analytics, and to react to operational exceptions through real-time alerts and automated actions in order to correct or avert the problem. Data are typically unstructured log records and sensor events, with each record including a timestamp indicating the exact time of data creation or arrival [5].

1.2.2. Batch Processing

Batch data processing is an efficient way of processing high volumes of data is where a group of transactions is collected over a period. Data is collected, entered, processed and then the batch results are produced. Hadoop focuses on batch data processing. Batch processing requires separate programs for input, process and output [6].

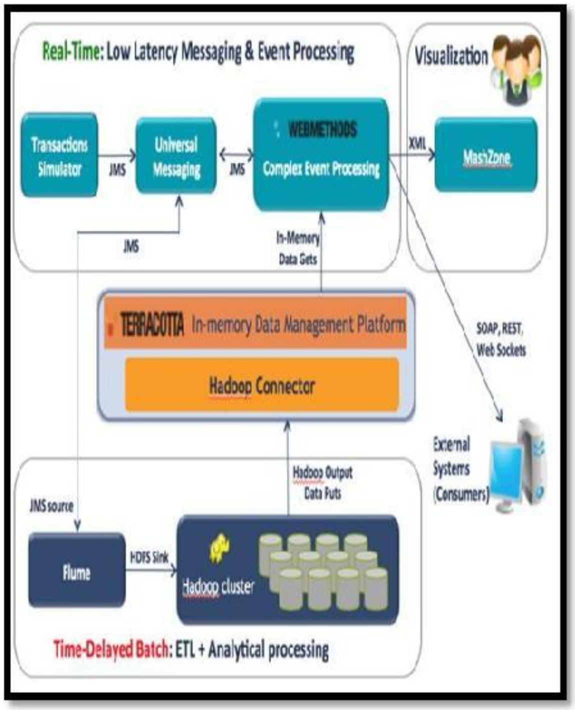

Figure 1.Real Time Processing

2. LITERATURE REVIEW

Pääkkönen et.al stated that Big Data provides a major contribution in the social networking domains like Facebook, Twitter, LinkedIn, Netflix etc**.** There is a need for independent reference architecture for the social media Big Data systems. The construction of this reference architecture is based on the published implementation of architectures by the social media Big Data use cases which is shown in fig.3This helps in classification of related implementation of technologies and products/services which helps in selecting appropriate architecture for the social media Big Data systems [8].Sumbaly et.al states that LinkedIn‟s Hadoop-based analytics stack, which allows data scientists and machine learning researchers to extract insights and build product features from massive amounts of data. It also gives an insight on how data flows into the offline system, how workflows are constructed and executed, and the mechanisms available for sending data back to the online system [10].

3. EVALUATION OF QUALITY OF BIG DATA THROUGH BIG DATA ARCHITECTURE

To improve the usability of reference architecture the quality attributes are added at the data processing stage. The quality evaluation of the Big Data is designed at the data processing stage of Big Data architecture. This quality data is used in different forms which ensure the trust bworthiness. The Information Quality Assessment Framework [12] enables information consumers to apply a wide range of policies to filter information. Fig.4 shows the general data quality process which follows five sequential steps. Different types of variation are found in the quality of Big Data:

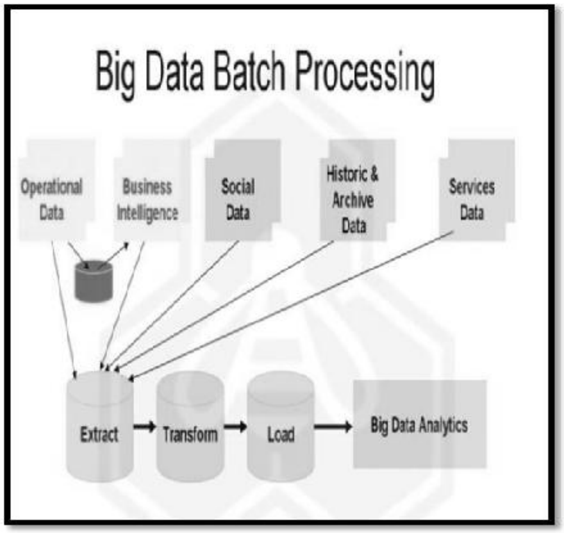

Figure 2.Batch Data Processing [6]

Target of Attribute: Certain data sources allow only specific type of quality attributes.

Applicability of Attribute: Certain quality attributes are applicable in only few stages.

Target of Metric: Data source type selection is independent of the metric type.

Applicability of Metric:Metric applicability differs from phase to phase.

Quality Policy: Variation in the data quality policy followed by the company

These variations must be properly traced to ensure the quality of the Big Data. These evaluations are used in quality data management of Big Data architectures [7].

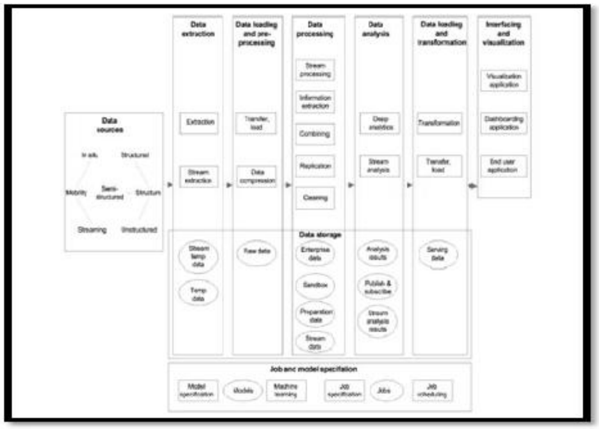

Figure 3. High-level design of the reference architecture[8]

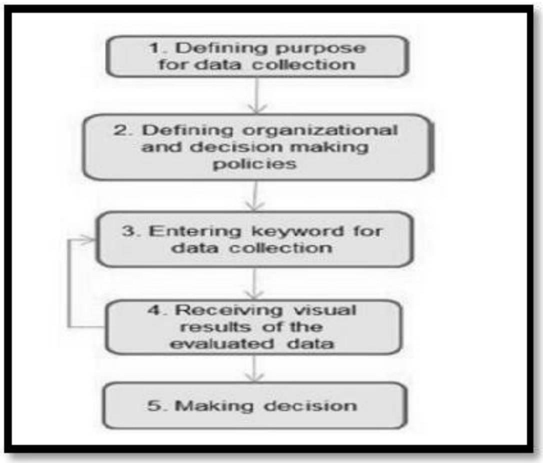
1. Defining purpose for data collection
2. Defining organizational and decision making policies
3. Entering keyword for data collection
4. Receiving visual results of the evaluated data
5. Making decision

Figure 4. The data quality evaluation process[10]

4. INFERENCES THROUGH THE SURVEY

Some common inferences which are found during the survey are as follows:

1. The various works relative to the reference architecture paves the way to construct independent reference architecture for the Big Data systems.
2. It also helps in easy understanding of architectures followed in various Big Data systems specially the social media systems.
3. It helps in characterizing the different types of architectures followed in stream and batch processing of social media Big Data systems.
4. Social media Big Data systems with fast data should have an architecture which supports instant processing of stream data.

CONCLUSION

The different architectures followed by the Social media Big Data Systems are difficult to understand. Most of them follow different forms of notations in their architecture which leads to the complexity. Constructing reference architecture helps in resolving this complexity. All the architecture contents are mapped onto the reference architecture. It provides clear view of how each processes is carried out in the Big Data systems. Based on the reference architecture, these Big Data systems are classified into two namely Stream and Batch processing systems based on their processing styles.Now-a-days all the Big Data systems uses both stream and batch processing techniques to provide better insights.

REFERENCES

[1].http://searchcloudcomputing.techtarget.com/definition/ big-data-Big-Data

[2].https://en.wikipedia.org/wiki/Reference_architecture

[3].http://searchbusinessanalytics.techtarget.com/definitio n/big-data-analytics

[4]. Office of the Assistant Secretary of Defense, Networks and Information Integration (OASD/NII), Reference Architecture Description, 2010

[5]. http://www.infoq.com/articles/stream-processing-hadoop

[6].http://www.datasciencecentral.com/profiles/blogs/batc h-vs-real-time-data-processing

[7]. Anne Immonen, PekkaPääkkönen, EilaOvaska, “Evaluating the Quality of Social MediaData in Big Data Architecture”, IEEE Access, 2016.

[8]. PekkaPääkkönen, Daniel Pakkala, “Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems”, Elsevier- Big Data Research, Volume 2, Issue 4, December-2015.

[9]. M. Meier, “Towards a Big Data reference architecture”, Master‟s thesis, October, 2013.

[10]. R. Sumbaly, J. Kreps, S. Shah, The “Big Data” Ecosystem at LinkedIn, in: 2013 ACM SIGMOD International Conference on Management of Data, New York, New York, USA, 22–27 June, 2013.

[11]. M.Galster, P. Avgeriou, “Empirically-grounded reference architectures: A proposal”, ACM SIGSOFT Conference on Quality of Software Architectures, June-2011.

[12]. C. Bizer and R. Cyganiak, “Quality-driven information filtering using the WIQA policy framework," Web Semantics, Sci., Services, Agents World Wide Web, vol. 7, no. 1, pp. 1-10, 2