International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 1, January - February 2017 **ISSN 2278-6856**

FUTURE TREND OF BIG DATA AND CLOUD TECHNOLOGY IN THE FIELD OF COMPUTATION

Prof. Amar Nath Singh[1], Er. Swapnajit Behera[2], Er. Jogindra Raut[2], Er. Sk Samillah[2]

[1,2]Gandhi Engineering College, Bhubaneswar, Odisha

ABSTRACT

As we know that, the cloud computation is now a day's became a crucial aspect for data processing as well as storage; its future prospect is very wider. Similarly the Big data is also an aspect of data processing. This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves around four important areas of analytics and Big Data, namely data management and supporting architectures model development and scoring and visualization and user interaction and business models. Through a detailed survey, we identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

Keywords:-Big Data, Cloud computing, Analytics, Data management.

1.Introduction

The modern Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Hence the data is more and the storage of data became a great challenge for us. So before the data to be processed we need to analyze the data. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web [1, 3]. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources. This paradigm is being popularly termed as Big Data.

Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavor. An organization willing to use such analytics technology frequently acquires expensive software licenses; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organization to better understand its business, organize its data, and integrate it for analytics [2, 3].

2.Background and methodology

Organizations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity [4, 7], web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network Web sites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data; a term that conveys the challenges it poses on existing infrastructure with respect to storage, management, interoperability, governance, and analysis of the data.

In today's competitive market, the following are the crucial points why we need a proper solution as;

a) Being able to explore data to understand customer behavior,
b) Segment customer base, offer customized services, and
c) Gain insights from data provided by multiple sources are key to competitive advantage.

Although decision makers would like to base their decisions and actions on insights gained from this data [4, 3], making sense of data, extracting non obvious patterns, and using these patterns to predict future behaviour are not new topics. Knowledge Discovery in Data (KDD) [5, 6] aims to extract non obvious information using careful and detailed analysis and interpretation..

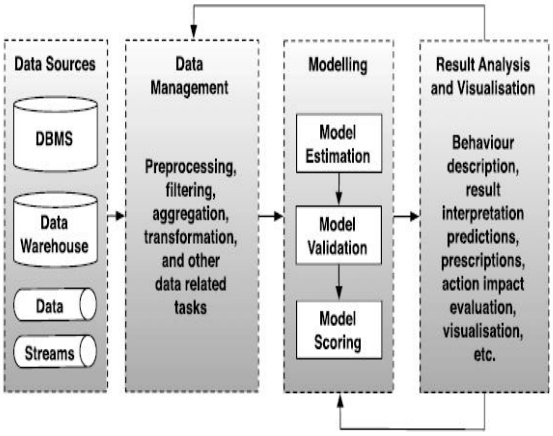

Fig. 1. Overview of the analytics workflow for Big Data.

In the above Figure -1 depicts the common phases of a traditional analytics workflow for Big Data. Data from various sources, including databases, streams, marts, and data warehouses, are used to build models. The large volume and different types of the data can demand pre-processing tasks for integrating the data, cleaning it, and filtering it. The prepared data is used to train a model and to estimate its parameters. Once the model is estimated, it should be validated before its consumption. Normally this phase requires the use of the original input data and specific methods to validate the created model.

3. Data management

It is another important area which is the most time-consuming and labour-intensive tasks of analytics are preparation of data for analysis.

Whenever a problem often exacerbated by Big Data as it stretches existing infrastructure to its limits. Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the data. Some of the challenges of deploying data management solutions on Cloud environments have been known for some time [5, 8], and solutions to perform analytics on the Cloud face similar challenges.
Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises, where Clouds can be for instance:

• **Private**: deployed on a private network, managed by the organization itself or by a third party. A private Cloud is suitable for businesses that require the highest level of control of security and data privacy [6]. In such conditions, this type of Cloud infrastructure can be used to share the services and data more efficiently across the different departments of a large enterprise.

• **Public**: deployed off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The analytics services and data management are handled by the provider and the quality of service (e.g. privacy, security, and availability) is specified in a contract [7, 8]. Organizations can leverage these Clouds to carry out analytics with a reduced cost or share insights of public analytics results.

• **Hybrid**: combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud.
Customers can develop and deploy analytics applications using a private environment, thus reaping benefits from elasticity and higher degree of security than using only a public Cloud.

3.1 Data variety and velocity

Big Data is characterized by what is often referred to as a multi-V model, which variety represents the data types, velocity refers to the rate at which the data is produced and processed, and volume defines the amount of data. Veracity refers to how much the data can be trusted given the reliability of its source [1, 3, 6], whereas value corresponds the monetary worth that a company can derive from employing Big Data computing. Although the choice of Vs used to explain Big Data is often arbitrary and varies across reports and articles on the Web – e.g. as of writing Viability is becoming a new V – variety, velocity, and volume [8, 9] are the items most commonly mentioned.

Regarding Variety, it can be observed that over the years, substantial amount of data has been made publicly available for scientific and business uses. Examples include repositories with government statistics1; historical weather information and forecasts; DNA sequencing; information on traffic conditions in large metropolitan areas; product reviews and comments; demographics. [10, 5]; comments, pictures, and videos posted on social network Web sites; information gathered using citizen-science platforms [2, 2]; and data collected by a multitude of sensors measuring various environmental conditions such as temperature, air humidity, air quality, and precipitation.

3.2 Data storage

Several solutions were proposed to store and retrieve large amounts of data demanded by Big Data, some of which are currently used in Clouds. Internet-scale file systems such as the Google File System (GFS) [5, 7] attempt to provide the robustness, scalability, and reliability that certain Internet services need. Other solutions provide object-store capabilities where files can be replicated

across multiple geographical sites to improve redundancy, scalability, and data availability.

One key aspect in providing performance for Big Data analytics applications is the data locality. This is because the volume of data involved in the analytics makes it prohibitive to transfer the data to process it. This was the preferred option in typical high performance computing systems: in such systems, that typically concern performing CPU-intensive calculations over a moderate to medium volume of data, it is feasible to transfer data to the computing units, because the ratio of data transfer to processing time is small. Nevertheless, in the context of Big Data, this approach of moving data to computation nodes would generate large ratio of data transfer time to processing time.

3.3 Data integration solutions

Forrester Research published a technical report that discusses some of the problems that traditional Business Intelligence (BI) faces [8, 5], highlighting that there is often a surplus of siloed data preparation, storage, and processing. Authors of the report envision some data processing and Big Data analytics capabilities being migrated to the EDW, hence freeing organizations from unnecessary data transfer and replication and the use of disparate data processing and analysis solutions. HANA One also offers a SAP data integrator to load data from HDFS and Hive-accessible databases. EDWs or Cloud based data warehouses, however, create certain issues with respect to data integration and the addition of new data sources. Standard formats and interfaces can be essential to achieve economies of scale and meet the needs of a large number of customers [5, 2]. Some solutions attempt to address some of these issues [10, 5, 1]. To improve the market penetration of analytics solutions in emerging markets such as India, Deepak et al. [4, 8] propose a multi-flow solution for analytics that can be deployed on the Cloud. The multi-flow approach provides a range of possible analytics operators and flows to compose analytics solutions; viewed as workflows or instantiations of a multi-flow solution. IVOCA [1, 8] is a tool aimed at Customer Relationship Management (CRM) that ingests both structured and unstructured data and provides data linking, classification, and text mining tools to facilitate analysts' tasks and reduce the time to insight.

3.4 Data processing and resource management

Map Reduce [4, 5] is one of the most popular programming models to process large amounts of data on clusters of computers.

Hadoop [10] is the most used open source Map Reduce implementation, also made available by several Cloud providers [4, 6, 7, 1]. Amazon EMR [4] enables customers to instantiate Hadoop clusters to process large amounts of data using the Amazon Elastic Compute Cloud (EC2) and other Amazon Web Services for data storage and transfer. Hadoop uses the HDFS file system to partition and replicate data sets across multiple nodes, such that when running a Map Reduce application, a mapper is likely to access data that is locally stored on the cluster node where it is executing. Although Hadoop provides a set of APIs that allows developers to implement Map Reduce applications, very often a Hadoop workflow is composed of jobs that use high-level query languages such as Hive and Pig Latin, created to facilitate search and specification of processing tasks. Lee et al. [9, 4] present a survey about the features, benefits, and limitations of Map Reduce for parallel data analytics. They also discuss extensions proposed for this programming model to overcome some of its limitations.

Hadoop provides data parallelism and its data and task replication schemes enable fault tolerance, but what is often criticized about it is the time required to load data into HDFS and the lack of reuse of data produced by mappers. Map Reduce is a model created to exploit commodity hardware, but when executed on reliable infrastructure, the mechanisms it provides to deal with failures may not be entirely essential.

3.5 Challenges in big data management

In this section, we discuss current research targeting the issue of Big Data management for analytics. There are still, however, many open challenges in this topic. The list below is not exhaustive, and as more research in this field is conducted, more challenging issues will arise.

A. **Data variety:** How to handle an always increasing volume of data? Especially when the data is unstructured, how to quickly extract meaningful content out of it? How to aggregate and correlate streaming data from multiple sources?

B. **Data storage:** How to efficiently recognize and store important information extracted from unstructured data? How to store large volumes of information in a way it can be timely retrieved? Are current file systems optimized for the volume and variety demanded by analytics applications? If not, what new capabilities are needed? How to store information in a way that it can be easily migrated/ ported between data centers/Cloud providers?

C. **Data integration:** New protocols and interfaces for integration of data that are able to manage data of different nature (structured, unstructured, semi-structured) and sources.

D. **Data Processing and Resource Management:** New programming models optimized for streaming and/or multidimensional data; new backend engines that manage optimized file systems; engines able to

combine applications from multiple programming models (e.g. Map Reduce, workflows, and bag-of-tasks) on a single solution/abstraction. How to optimize resource usage and energy consumption when executing the analytics application?

4 Open challenges

There are many research challenges in the field of Big Data visualization. First, more efficient data processing techniques are required in order to enable real-time visualization. Methods considering each of these techniques could be further researched and improved.
Cost-effective devices for large-scale visualisation are another hot topic for analytics visualization, as they enable finer resolution than simple screens. Visualization for management of computer networks and software analytics [10, 1] is also an area that is attracting attention of researchers and practitioners for its extreme relevance to management of large-scale infrastructure (such as Clouds) and software, with implications in global software development, open source software development, and software quality improvements.

5. Summary and conclusions

The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimize its operation, and reduce its costs.

Cloud computing helps to solve these problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.

Although Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud supported analytics is still in its early days.

For each of these areas, ongoing work was analyzed and key open challenges were discussed. The area of Big Data Computing using Cloud resources is moving fast, and after surveying the current solutions we identified some key lessons:
Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools;
a. It is also clear that analytics is a complex process that demands people with expertise in cleaning up data, understanding and selecting proper methods, and analyzing results.
b. Cloud computing plays a key role for Big Data; not only because it provides infrastructure and tools, but also because it is a business model that Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)).

However, AaaS/BDaaS brings several challenges because the customer and provider's staff are much more involved in the loop than in traditional Cloud providers offering infrastructure/ platform/software as a service.

Author

Prof. Amar Nath Singh is presently working at GEC, Bhubaneswar as a reader in the department of Computer Science and Engineering. His research area is Underground Mines and Surface mining using Artificial Intellegence, Fuzzy Logic. His research area includes cloud computing, WSN, Machine Learning. He has produced more than 60 M.Tech scholar till date.

References

[1]. D.J. Abadi, Data management in the cloud: Limitations and opportunities,IEEE Data Engineering Bulletin 32 (1) (2009) 3–12.
[2]. Amazon redshift, http://aws.amazon.com/redshift/.
[3]. Amazon data pipeline, http://aws.amazon.com/datapipeline/.
[4]. Amazon Elastic MapReduce (EMR),http://aws.amazon.com/elasticmapreduce/.
[5]. Amazon Kinesis, http://aws.amazon.com/kinesis/developer-resources/.
[6]. R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar, M. Shah, R. Tewari, Cloud Analytics: Do We Really Need to Reinvent the Storage Stack? in: Proceedings of the Conference on Hot Topics in Cloud Computing (HotCloud 2009), USENIX Association, Berkeley, USA, 2009.
[7]. G. Andrienko, N. Andrienko, S. Wrobel, Visual analytics tools for analysis of movement data, SIGKDD Explor. Newsl. 9 (2) (2007) 38–46.
[8]. Announcing Suro: Backbone of Netflix's Data Pipeline, http://techblog. netflix.com/2013/12/announcing-suro-backbone-of-netflixs.html.
[9]. Apache S4: distributed stream computing platform, http://incubator.apache.org/s4/.
[10]. Apache Mahout, http://mahout.apache.org.