Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Big Data Computing Tools Management & Clouds

E K Radhika
HOD, Department of Computer Science Dept. Sindhi College, India

Abstract: 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 (i) data management (ii) model building and scoring (iii) smart grid (iv) Visualisation and user interaction provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

I. INTRODUCTION

Society is becoming increasingly more instrumented and as a result, organisations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organisations gain insights not only from their privately acquired data, but also from large amounts of data publicly available. 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 organisations to understand the needs of their customers, predict their wants and demands, and optimise the use of resources. This paradigm is being popularly termed as Big Data.

II. BIG DATA TOOLS

Customers data is in the order of Terabytes and in a variety of formats. So, it requires high velocity, scalability and fault tolerance in data processing, storage and visualization. Big Data implementation can be done using several tools, but the analytics tools are the most critical in business choice. Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavour.

Data management

One of the most time-consuming and labour-intensive tasks of analytics is preparation of data for analysis; 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, 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.

Data variety and velocity

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, 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 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, comments, pictures, and videos posted on social network.

Web sites; information gathered using citizen-science platforms and data collected by a multitude of sensors measuring various environmental conditions such as temperature, air humidity, air quality, and precipitation. An example illustrating the need for such a variety within a single analytics application is the Eco-Intelligence platform.

Eco-Intelligence was designed to analyse large amounts of data to support city planning and promote more sustainable development. The platform aims to efficiently discover and process data from several sources, including sensors, news, Web sites, television and radio, and exploit information to help urban stakeholders cope with the highly dynamics of urban development. In a related scenario, the Mobile Data Challenge (MDC) was created aimed at generating innovations on smart phone-based research, and to enable.

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) 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. Examples include Amazon Simple Storage Service (S3), Nirvanix Cloud Storage, OpenStack Swift and Windows Azure Binary Large Object (Blob) storage. Although these solutions provide the scalability and redundancy that many Cloud applications require, they sometimes do not meet the concurrency and performance needs of certain analytics applications.

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. Thus, a different approach is preferred, where computation is moved to where the data is. The same approach of exploring data locality was explored previously in scientific workflows and in Data Grids.

In the context of Big Data analytics, MapReduce presents an interesting model where data locality is explored to improve the performance of applications. Hadoop, an open source

MapReduce implementation, allows for the creation of clusters that use the Hadoop distributed File System (HDFS) to partition and replicate data sets to nodes where they are more likely to be consumed by mappers. In addition to exploiting concurrency of large numbers of nodes, HDFS minimises the impact of failures by replicating data sets to a configurable number of nodes. It has been used by Thusoo. to develop an analytics platform to process Facebook's large data sets. The platform uses Scribe to aggregate logs from Web servers and then exports them to HDFS files and uses a Hive–Hadoop cluster to execute analytics jobs. The platform includes replication and compression techniques and columnar compression of Hive7 to store large amounts of data. Among the drawbacks of Cloud storage techniques and Map Reduce implementations, there is the fact that they require the customer to learn a new set of APIs to build analytics solutions for the Cloud. To minimise this hurdle, previous work has also investigated POSIX-like file systems for data analytics.

By using the concept of meta-blocks, they demonstrated that IBM's General Parallel File System (GPFS) can match the read performance of HDFS. A meta-block is a consecutive set of data blocks that are allocated in the same disk, thus guaranteeing contiguity.

The proposed approach explores the trade-off between different block sizes, where meta-blocks minimise seek overhead in MapReduce applications, whereas small blocks reduce pre-fetch overhead and improves cache management for ordinary applications.

1) http://aws.amazon.com/s3/
2) http://www.nirvanix.com
3) http://swift.openstack.org

Model building and scoring

The data storage and Data as a Service (DaaS) capabilities provided by Clouds are important, but for analytics, it is equally relevant to use the data to build models that can be utilised for forecasts and prescriptions. Moreover, as models are built based on the available data, they need to be tested against new data in order to evaluate their ability to forecast future behaviour. Existing work has discussed means to offload such activities – termed here as model building and scoring – to Cloud providers and ways to parallelise certain machine learning algorithms. This section describes work on the topic. summarises the analysed work, its goals, and target infrastructures. Guazzelli et al. use Amazon EC2 as a hosting platform for the Zementis' ADAPA model scoring engine. Predictive models, expressed in Predictive Model Markup Language (PMML), are deployed in the Cloud and exposed via Web Services interfaces. Users can access the models with Web browser technologies to compose their data mining solutions.

Smart Grid

A smart grid is an intelligent electricity grid that optimizes the generation, distribution and consumption of electricity through the introduction of Information and Communication Technologies on the electricity grid. In essence, smart grids bring profound changes in the information systems that drive them: new information flows coming from the electricity grid, new players such as decentralized producers of renewable energies, new uses such as electric vehicles and connected houses and new communicating equipments such as smart meters, sensors and remote control points. All this will cause a deluge of data that the energy companies will have to face. Big Data technologies offer suitable solutions for utilities, but the decision about which Big Data technology to use is critical. In this paper, we provide an overview of data management for smart grids, summarise the added value of Big Data technologies for this kind of data, and discuss the technical requirements, the tools and the main steps to implement Big Data solutions in the smart grid context.

Big Data implementation in smart grid

In this section, the focus will be on customers data analytics, because it involves the smart consumers concept, which makes consumers as potential producers of clean energy, players in their consumption and also main actors in production and consumption balancing. Customer data analytics is a great opportunity for utilities to understand customer behaviour better, and be able to make strategic decisions. Big Data analytics of customers data become a necessity and not a choice for electrical companies. Consumers are participating in smart grids as end customers through smart meters that offer them better control of their own consumption. Demand Response (DR) programs are used by utilities to obtain real-time information of the demand curves in the various points of consumption in order to calibrate and prognosticate more precisely. Thus, the production curve can be regulated according to demand more efficiently and reduce the losses of "overproduction". This will also make it possible to make a real-time diagnosis of meters and equipment close to the consumer, sending alarms, executing "self-healing" systems, etc.. Improving customer engagement is among the motivations of DR, because it helps utilities interact with the customers energy needs even during power outage.

Visualisation and user interaction

With the increasing amounts of data with which analyses need to cope, good visualisation tools are crucial. These tools should consider the quality of data and presentation to facilitate navigation. The type of visualisation may need to be selected according to the amount of data to be displayed, to improve both displaying and performance. Visualisation can assist in the three major types of analytics: descriptive, predictive, and prescriptive. Many visualisation tools do not describe advanced aspects of analytics, but there has been an effort to explore visualisation to help on predictive and prescriptive analytics, using for instance sophisticated reports and storytelling. A key aspect to be considered on visualisation and user interaction in the Cloud is that network is still a bottleneck in several scenarios. Users ideally would like to visualise data processed in the Cloud having the same experience and feel as though data were processed locally.

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.

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? Data storage: How to efficiently recognise 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 optimised 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 centres/Cloud providers? 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. Data Processing and Resource Management: New programming models optimised for streaming and/or multidimensional data.

SUMMARY AND CONCLUSION

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, optimise its operation, and reduce its costs.. Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud supported analytics is still in its early days. In this paper, we discussed the key stages of analytics workflows, and surveyed the state-of-the-art of each stage in the context of Cloud-supported analytics. Surveyed work was classified in three key groups: Data Management (which encompasses data variety, data storage, data integration solutions, and data processing and resource management), Model Building and Scoring, and Visualisation and User Interactions.

For each of these areas, ongoing work was analysed and key open challenges were discussed. This survey concluded with an analysis of business models for Cloud-assisted data analytics and other non-technical challenges. The area of Big Data Computing using Cloud resources is moving fast, and after surveying the current solutions are analytics can be descriptive, predictive, prescriptive; Big Data can have various levels of variety, velocity, volume, and veracity. Therefore, it is important to understand the requirements in order to choose appropriate Big Data tools; • 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 analysing results. Tools are fundamental to help people perform these tasks. In addition, depending on the complexity and costs involved in carrying out these tasks, providers who offer Analytics as a Service or Big Data as a Service can be a promising alternative compared to performing these tasks in-house; • 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. Big Data is an evolving field, where much of the research is yet to be done. Big data at present, is handled by the software named Hadoop. However, the proliferating amount of data is making Hadoop insufficient. To harness the potential of Big Data completely in the future, extensive research needs to be carried out and revolutionary technologies need to be developed.

References

[1] Apache Hive. Available at http://hive.apache.org
[2] http://blogs.worldbank.org/voices/meet-winners-and-finalists-first-wbg-big-data-innovation-challenge
[3] http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/
[4] http://www.exist.com/wp-content/uploads/2014/10/3Vsbigdata.png
[5] http://www.internetlivestats.com/twitter-statistics/
[6] http://www.internetlivestats.com/google-search-statistics/
[7] Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013
[8] http://lsst.org/lsst/google
[9] http://en.wikipedia.org/wiki/Parkinson's_law
[10] http://www.economist.com/node/15557443
[11] http://www.youtube.com/t/press_statistics/?hl=en
[12] http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/