Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.277 – 281

BIG DATA COMPUTING AND CLOUDS

M. Sravani[1], P. Laxmi[2]
Sravani886@gmail.com [1], laxmi.p16@gmail.com [2]
CSE Dept, Assistant Professor[12]
VBIT[12]

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 and supporting architectures; (ii) model development and scoring; (iii) visualisation and user interaction; and (iv) 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.

I. Introduction

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. 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 [118]. 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. The most often claimed benefits of Clouds include offering resources in a pay-as-you-go fashion, improved availability and elasticity, and cost reduction. Clouds can prevent organizations from spending money for maintaining peak-provisioned IT infrastructure that they are unlikely to use most of the time. Whilst at first glance the value proposition of Clouds as a platform to carry out analytics is strong, there are many challenges that need to be overcome to make Clouds an ideal platform for scalable analytics. In this article we survey approaches, environments, and technologies on areas that are key to Big Data analytics capabilities and discuss how they help building analytics solutions for Clouds. We focus on the most important technical issues on enabling Cloud analytics, but also highlight some of the non-technical challenges faced by organizations that want to provide analytics as a service in the Cloud. In addition, we describe a set of gaps and recommendations for the research community on future directions on Cloud supported Big Data computing.

II. Background and Methodology

Organizations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity, 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, being able to explore data to understand customer behavior, segment customer base, offer customized services, and gain insights from data provided by multiple sources is key to competitive advantage. Although decision makers would like to base their decisions and actions on insights gained from this data, making sense of data, extracting non obvious patterns, and using these patterns to predict future behavior are not new topics. Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining, more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions. Fig. 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. Finally, the model is consumed and applied to data as it arrives. This phase, called model scoring, is used to generate predictions, prescriptions, and recommendations. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data.

III. Data Management

One of the most time-consuming and labor-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:

• 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. 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.

IV. 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 utilized 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 behavior. Existing work has discussed means to offload such activities – termed here as model building and scoring – to Cloud providers and ways to parallelize certain machine learning algorithms. This section describes work on the topic. Table 1 summarizes the analyzed 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. Existing work also advocates the use of PMML as a language to exchange information about predictive models [73]. Zementis also provides technologies for data analysis and model building that can run either on a customer's premises or be allocated as SaaS using Infrastructure as a Service (IaaS) provided by solutions such as Amazon EC2 and IBM Smart Cloud Enterprise [76]. Google Prediction API allows users to create machine learning models to predict numeric values for a new item based on values of previously submitted training data or predict a category that best describes an item. The prediction API allows users to submit training data as comma separated files following certain conventions, create models, share their models or use models that others shared. With the Google Prediction API, users can develop applications to perform analytics tasks such as sentiment analysis [51], purchase prediction, provide recommendations, analyse churn, and detect spam. The Apache Mahout project [11] aims to provide tools to build scalable machine learning libraries on top of Hadoop using the Map Reduce paradigm. The provided libraries can be deployed on a Cloud and be explored to build solutions that require clustering, recommendation mining, document categorization, among others.

V. Visualization and user interaction

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

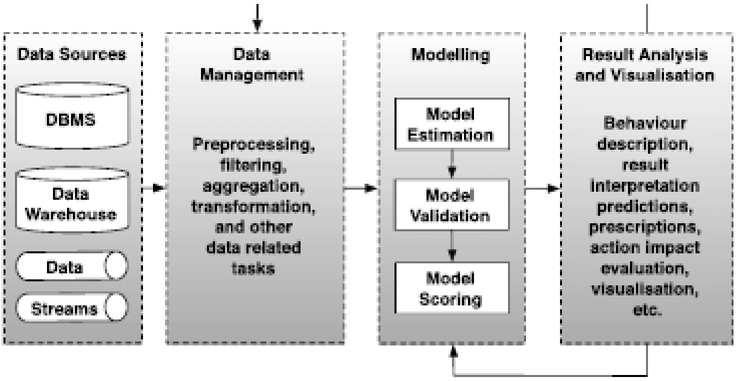
Data Sources
DBMS
Data Warehouse
Data
Streams
Data Management
Preprocessing, filtering, aggregation, transformation, and other data related tasks
Modelling
Model Estimation
Model Validation
Model Scoring
Result Analysis and Visualisation
Behaviour description, result interpretation predictions, prescriptions, action impact evaluation, visualisation, etc.

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

VI. Other challenges

In business models where high-level analytics services may be delivered by the Cloud, human expertise cannot be easily replaced by machine learning and Big Data analysis; in certain scenarios, there may be a need for human analysts to remain in the loop Management should adapt to Big Data scenarios and deal with challenges such as how to assist human analysts in gaining insights and how to explore methods that can help managers in making quicker decisions. Application profiling is often necessary to estimate the costs of running analytics on a Cloud platform. Users need to develop their applications to target Cloud platforms; an effort that should be carried out only after estimating the costs of transferring data to the Cloud, allocating virtual machines, and running the analysis. This cost estimation is not a trivial task to perform in current Cloud offerings. Although best practices for using some data processing services are available [49], there should be tools that assist customers to estimate the costs and risks of performing analytics on the Cloud. Data ingestion by Cloud solutions is often a weak point, whereas debugging and validation of developed solutions is a challenging and tedious process. As discussed earlier, the manner analytics is executed on Cloud platforms resembles the batch job scenario: users submit a job and wait until tasks are executed and then download the results. Once an analysis is complete, they download sample results that are enough to validate the analysis task and after that perform further analysis. Current Cloud environments lack this interactive process, and techniques should be developed to facilitate interactivity and to include analysts in the loop by providing means to reduce their time to insight. Systems and techniques that iteratively refine answers to queries and give users more control of processing are desired.

VII. 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. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating 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.

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/.