J. Parallel Distrib. Comput. 79–80 (2015) 3–15

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Big Data computing and clouds: Trends and future directions

Marcos D. Assunção [a,*], Rodrigo N. Calheiros [b], Silvia Bianchi [c], Marco A.S. Netto [c], Rajkumar Buyya [b,*]

[a] INRIA, LIP, ENS de Lyon, France
[b] The University of Melbourne, Australia
[c] IBM Research, Brazil

HIGHLIGHTS

- Survey of solutions for carrying out analytics and Big Data on Clouds.
- Identification of gaps in technology for Cloud-based analytics.
- Recommendations of research directions for Cloud-based analytics and Big Data.

ARTICLE INFO

Article history:
Received 25 October 2013
Received in revised form
20 May 2014
Accepted 18 August 2014
Available online 27 August 2014

Keywords:
Big Data
Cloud computing
Analytics
Data management

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.

1. 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 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 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.

Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavour. As Yu [136] points out, Big Data offers substantial value to organisations willing to adopt it, but at the same time poses a considerable number of challenges for the realisation of such added value. An organisation willing to use analytics technology frequently acquires expensive software licences; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organisation to better understand its business, organise its data, and integrate it for analytics [120]. This joint effort of organisation and analysts often aims to help the organisation understand its customers' needs, behaviours, and future demands for new products or marketing strategies. Such effort, however, is generally costly and often lacks flexibility. Nevertheless, research and application of Big Data are being extensively explored by governments, as evidenced by initiatives from USA [20] and UK [106]; by academics, such as the bigdata@csail initiative from MIT [19]; and by companies such as Intel [122].

Cloud computing has been revolutionising the IT industry by adding flexibility to the way IT is consumed, enabling organisations to pay only for the resources and services they use. In an effort to

* Corresponding authors.
E-mail addresses: assuncao@acm.org (M.D. Assunção), rbuyya@unimelb.edu.au (R. Buyya).

http://dx.doi.org/10.1016/j.jpdc.2014.08.003
0743-7315/

reduce IT capital and operational expenditures, organisations of all sizes are using Clouds to provide the resources required to run their applications. Clouds vary significantly in their specific technologies and implementation, but often provide infrastructure, platform, and software resources as services [25,13].

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 organisations 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 organisations 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.

2. Background and methodology

Organisations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity [14,127], 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 [99,55,17]; 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 behaviour, segment customer base, offer customised 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 [43], 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) [50] aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining [133,84], 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 visualisation to drive decisions and actions [43,42,63].

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.

Analytics solutions can be classified as descriptive, predictive, or prescriptive as illustrated in Fig. 2. Descriptive analytics uses historical data to identify patterns and create management reports; it is concerned with modelling past behaviour. Predictive analytics attempts to predict the future by analysing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints.

Despite the hype about it, using analytics is still a labour intensive endeavour. This is because current solutions for analytics are often based on proprietary appliances or software systems built for general purposes. Thus, significant effort is needed to tailor such solutions to the specific needs of the organisation, which includes integrating different data sources and deploying the software on the company's hardware (or, in the case of appliances, integrating the appliance hardware with the rest of the company's systems) [120]. Such solutions are usually developed and hosted on the customer's premises, are generally complex, and their operations can take hours to execute. Cloud computing provides an interesting model for analytics, where solutions can be hosted on the Cloud and consumed by customers in a pay-as-you-go fashion. For this delivery model to become reality, however, several technical issues must be addressed, such as data management, tuning of models, privacy, data quality, and data currency.

This work highlights technical issues and surveys existing work on solutions to provide analytics capabilities for Big Data on the Cloud. Considering the traditional analytics workflow presented in Fig. 1, we focus on key issues in the phases of an analytics solution. With Big Data it is evident that many of the challenges of Cloud analytics concern data management, integration, and processing. Previous work has focused on issues such as data formats, data representation, storage, access, privacy, and data quality. Section 3 presents existing work addressing these challenges on Cloud environments. In Section 4, we elaborate on existing models to provide and evaluate data models on the Cloud. Section 5 describes solutions for data visualisation and customer interaction with analytics solutions provided by a Cloud. We also highlight some of the business challenges posed by this delivery model when we discuss service structures, service level agreements, and business models. Security is certainly a key challenge for hosting analytics solutions on public Clouds. We consider, however, that security is an extensive topic and would hence deserve a study of its own. Therefore, security and evaluation of data correctness [130] are out of scope of this survey.

3. 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 [1,113,82], 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 organisation 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