Recent Developments on Security and Reliability in Large-Scale Data Processing with MapReduce

Christian Esposito, University Federico II, Naples, Italy
Massimo Ficco, Seconda Università degli Studi di Napoli, Naples, Italy

ABSTRACT

The demand to access to a large volume of data, distributed across hundreds or thousands of machines, has opened new opportunities in commerce, science, and computing applications. MapReduce is a paradigm that offers a programming model and an associated implementation for processing massive datasets in a parallel fashion, by using non-dedicated distributed computing hardware. It has been successfully adopted in several academic and industrial projects for Big Data Analytics. However, since such analytics is increasingly demanded within the context of mission-critical applications, security and reliability in MapReduce frameworks are strongly required in order to manage sensible information, and to obtain the right answer at the right time. In this paper, the authors present the main implementation of the MapReduce programming paradigm, provided by Apache with the name of Hadoop. They illustrate the security and reliability concerns in the context of a large-scale data processing infrastructure. They review the available solutions, and their limitations to support security and reliability within the context MapReduce frameworks. The authors conclude by describing the undergoing evolution of such solutions, and the possible issues for improvements, which could be challenging research opportunities for academic researchers.

KEYWORDS

Hadoop Distributed File System, MapReduce, Parallel Processing, Reliability, Security

1. INTRODUCTION

Big data has become a buzzword for several research projects since the research in such a topic is currently encouraged by huge funding possibilities. As a practical example, big data analytics represents one of the Excellent Science priorities of Horizon 2020 on scientific infrastructures and the development of innovative high value added services for the European Commission, and a recent call “FP7 - ICT - 2013 - 11 - 4.2 Scalable data analytics” had a budget of 26 M€ on the topic of “Big Data R&D”. Big data is also attracting a lot of attention within industry, since, as shown by several surveys, such as the one conducted by the IAIS in Sankt Augustin, it is able to provide to companies strategic competitive advantages, increased sales, higher productivity and reduced costs (Brynjolfsson et al., 2011). The definition of big data within the IT dictionary of Gartner is the following one: “Big data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”. Indeed, the need of processing a large amount of data has caused a revolution in analytics, as stated by Roger Barga from Microsoft at his last keynote speech at ACM DEBS 2013. Traditionally, data...
processing is conducted according to the classic data mining approach or Batch Processing (Babcock et al., 2002): data need to be stored prior to analysis taking place. Companies used to have their own data centre to store and process their information of interest. Due to the increasing volume of data to be stored and processed, it is no more economical affordable to possess data storing and processing capabilities within small- and/or medium-sized companies. The current tendency is to outsource such capabilities by using third-party cloud platforms and to apply the so-called Data as a Service solution (Hong-Linh & Dustdar, 2009), which makes data management very cost-effective.

Not only has the ICT infrastructure for data management changed so as to accomplish the vision of big data analytics, but also the way of elaborating data. Besides the traditional batch processing, there is also the “data-in-motion” approach or Real-Time Processing (Gupta et al., 2012), where analytics is applied as soon as data are produced and while moving from their sources to their intended destinations. Such a different data processing approach imposes novel requirements so as to provide the scalability, low latency, and performance needed to handle big data projects with real-time processing demands. At the moment, IT managers are using both batch and real-time processing equally, with a 50-50 split. However, as illustrated in a survey of 200 IT professionals about big data analytics conducted by Intel4, we will experience an increase of demands for real-time processing.

Such a change in the way of looking at data management in the current industry is pushing towards a radical rethink of the ICT platforms for data management, and leaving behind the old means for analytics, i.e., descriptive analytics based on warehouses. We are currently witnessing the rapid success of novel data processing platforms based on the MapReduce programming paradigm (Dean & Ghemawat, 2008), which offers easiness of implementing programs for processing large amounts of data in a parallel fashion and the possibility of using non-dedicated distributed computing hardware. As a concrete example, an open-source implementation of MapReduce, named Apache Hadoop5, has emerged as the leading big data solution, with more than 1,200 people across 80 companies contributing to its development since its emergence in 2005. However, if we consider the Emerging Technologies Hype Cycle in 2013 by Gartner6, we can see that the area of Big Data is approaching its peak of expectation. This means that such analytics are not yet mature technologies, and Apache Hadoop and similar products are still relatively new. Thus, there is a great deal of confusion about their strengths and weaknesses. In fact, there are several possible research challenges still to be completely resolved, and several research opportunities to take advantage of. The above-mentioned survey of Intel clearly indicates the importance of having a better analytics than the one possible with the current platforms based on MapReduce as Hadoop.

The 1010Data company surveyed about 158 U.S. executives7, so as to find the specific concerns in their business projects and to identify the shortcomings of Apache Hadoop, and other similar products based on MapReduce. The obtained results, shown in Figure 1, indicate that the 45% of the responders cited Security and Reliability among their main concern when using Hadoop for business. In fact, the common belief is that Hadoop still is not reliable and secure enough to meet the business requirements in this sense. Therefore, industrial practitioners are waiting for upgrades in these directions. In addition, neither commercial alternative to Hadoop exhibits considerable improvements on these topics than Hadoop, leaving the provision of a higher reliability and security degrees in platforms based on MapReduce an open issue to be investigated in a more exhaustive manner.

This paper aims at answering to the following questions:

1. How to connote the security and reliability aspects of MapReduce frameworks? Which requirements have to be satisfied in order to have a secure and reliable MapReduce framework?
2. Which methods are currently available to achieve a secure and reliable MapReduce framework? Is there something missing in the available MapReduce implementations that needs to be handled?
3. What are the possible research opportunities within the context of security and reliability in MapReduce frameworks and in large-scale data processing infrastructures?
Related Content

A Directed Acyclic Graph (DAG) Ensemble Classification Model: An Alternative Architecture for Hierarchical Classification
[www.igi-global.com/article/a-directed-acyclic-graph-dag-ensemble-classification-model/185659?camid=4v1a](www.igi-global.com/article/a-directed-acyclic-graph-dag-ensemble-classification-model/185659?camid=4v1a)

Resource Constrained Data Stream Clustering with Concept Drifting for Processing Sensor Data
Minimizing the Minus Sides of Mining Data
www.igi-global.com/chapter/minimizing-minus-sides-mining-data/7520?camid=4v1a

Collaborative Filtering Based Data Mining for Large Data
www.igi-global.com/chapter/collaborative-filtering-based-data-mining-for-large-data/159498?camid=4v1a