Available online at www.sciencedirect.com

ScienceDirect

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

Survey

Security and privacy aspects in MapReduce on clouds: A survey

Philip Derbeko[a], *Shlomi Dolev*[b], *Ehud Gudes*[b], *Shantanu Sharma*[b,*]

[a] *EMC, Beer-Sheva, Israel*

[b] *Department of Computer Science, Ben-Gurion University of the Negev, Israel*

ARTICLE INFO

Article history:
Received 1 July 2015
Received in revised form
1 May 2016
Accepted 2 May 2016
Published online 25 May 2016

Keywords:
Cloud computing
Distributed computing
Hadoop
HDFS
Hybrid cloud
Private cloud
Public cloud
MapReduce algorithms
Privacy
Security

ABSTRACT

MapReduce is a programming system for distributed processing of large-scale data in an efficient and fault tolerant manner on a private, public, or hybrid cloud. MapReduce is extensively used daily around the world as an efficient distributed computation tool for a large class of problems, *e.g.*, search, clustering, log analysis, different types of join operations, matrix multiplication, pattern matching, and analysis of social networks. Security and privacy of data and MapReduce computations are essential concerns when a MapReduce computation is executed in public or hybrid clouds. In order to execute a MapReduce job in public and hybrid clouds, authentication of mappers–reducers, confidentiality of data-computations, integrity of data-computations, and correctness–freshness of the outputs are required. Satisfying these requirements shields the operation from several types of attacks on data and MapReduce computations. In this paper, we investigate and discuss security and privacy challenges and requirements, considering a variety of adversarial capabilities, and characteristics in the scope of MapReduce. We also provide a review of existing security and privacy protocols for MapReduce and discuss their overhead issues.

Contents

[*] *Corresponding author.*
E-mail addresses: philip.derbeko@emc.com (P. Derbeko), dolev@cs.bgu.ac.il (S. Dolev), ehud@cs.bgu.ac.il (E. Gudes), sharmas@cs.bgu.ac.il (S. Sharma).

http://dx.doi.org/10.1016/j.cosrev.2016.05.001
1574-0137/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Cloud computing [1] infrastructure provides on-demand, easy, and scalable access to a shared pool of configurable resources, without worrying about managing those resources. Details about cloud computing can be found in [2,3]. Clouds provide three types of services, as follows: (i) *infrastructure-as-a-service*, IaaS, provides infrastructure in terms of virtual machines, storage, and networks, (ii) *platform-as-a-service*, PaaS, provides a scalable software platform allowing the development of custom applications, and (iii) *software-as-a-service*, SaaS, provides software running in clouds as a service, for example, emails and databases. Clouds can be classified into three types, as follows: (i) *public cloud*: a cloud that provides services to many users and is not under the control of a single exclusive user, (ii) *private cloud*: a cloud that has its proprietary resources and is under the control of a single exclusive user, and (iii) *hybrid cloud*: a combination of public and private clouds.

One of the most common *platform-as-a-service* computational paradigms is MapReduce [4], introduced by Google in 2004. MapReduce provides an efficient and fault tolerant parallel processing of large-scale data without any costly and dedicated computing node like a supercomputer. At the beginning, MapReduce was designed to be deployed on-premises under mistaken assumption that local environment can be completely trusted. Thus, security and privacy aspects were overlooked in the initial design. As MapReduce gained popularity the lack of security and privacy in on-premises deployment become severe shortcoming. In addition, MapReduce is being deployed on both hybrid and public clouds, which are prone to many attacks and security threats. In the current days, several public clouds, *e.g.*, Amazon Elastic MapReduce, Google App Engine, IBM's Blue Cloud, and Microsoft Azure, enable users to perform MapReduce cloud computations without considering physical infrastructures and software installations. Thus, the deployment of MapReduce in public clouds enables users to process large-scale data in a cost-effective manner and establishes a relationship between two independent entities, *i.e.*, clouds and MapReduce. As a downside, the deployment of MapReduce in hybrid and public cloud needs to deal with many attacks on the communication networks and (the three service layers of) the cloud.

Data processing in the cloud highlights a tradeoff between the ease of processing and security–privacy of data and computations. Specifically, on one hand, the deployment of MapReduce in a well-managed public cloud provides economical and carefree resource management. On the other hand, public clouds do not guarantee the rigorous security and privacy of computations as well as stored data. Private clouds provide security and privacy of data as well as computations, due to users' ability to physically and electronically constrain data access and execution of computations. However, the user of the private cloud manages the nodes, updates software, and replaces the failed nodes. Such management is time consuming and incurs huge monetary cost.

Our focus is on the security and privacy issues of MapReduce environment in public or hybrid clouds. Private cloud environments are more secure due to a physical security of the cloud. Many of the reviewed below results are applicable to both public and hybrid clouds, unless stated otherwise (for instance, see hybrid cloud specific research in [5,6] and Section 4.4.1). Even though there is a plethora of additional projects and frameworks that add functionality on top of MapReduce (see Apache Hive [7], Cloudera Impala,[1] HBase [8], Apache Zookeeper,[2] Thrift,[3] and Apache Solr[4]), this paper only reviews security related projects in Section 2 (readers interested in security and privacy issues of other projects may refer to [9]).

Security aspects in the context of MapReduce are crucial in order to authenticate and authorize users, auditing–confidentiality–integrity of data and computation, availability

[1] http://impala.io/.
[2] https://zookeeper.apache.org/.
[3] https://thrift.apache.org/.
[4] http://lucene.apache.org/solr/.