Contents lists available at ScienceDirect

Advances in Engineering Software

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

# Cloud agnostic Big Data platform focusing on scalability and cost-efficiency

Róbert Lovas*, Enikő Nagy, József Kovács

*Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI), Budapest, Hungary*

ARTICLE INFO

*Keywords:*
Cloud computing
Big Data
Orchestration
Hadoop
Occopus
Contextualisation
Scalability
Cost analysis

ABSTRACT

Nowadays a significant part of the cloud applications processes a large amount of data to provide the desired analytics, simulation and other results. Cloud computing is becoming a widely used IT model to address the needs of many scientific and commercial Big Data applications. In this paper, we present a Hadoop platform deployment method for various cloud infrastructures with the Occopus cloud orchestrator tool. Our automated solution provides an easy-to-use, portable and scalable way to deploy the popular Hadoop platform with the main goal to avoid vendor locking issues, i.e. there is no dependency on any cloud provider prepared and offered virtual machine image or “black-box” Platform-as-a-Service mechanism. The paper presents promising performance measurements results and cost analysis.

## 1. Introduction

Big Data [1] refers to the massive volume of both structured, semi-structured and unstructured different type of data that has the potential to be mined for information. However, the data is so large, complex, and rapidly growing in this case that it is overly difficult to process using traditional databases and software techniques. Traditional sequential data processing algorithms are not sufficient to analyze this large volume of data. This phenomenon demands new strategies, techniques and tools for processing and analysing data.

Cloud [2] infrastructures are continuously emerging over these years and they play a crucial role in solving Big Data applications. Such grand challenge applications require a significant amount of IT resources that might be accessed in a user-friendly way and on-demand from cloud providers by utilising (among others) virtualisation technologies and rapid elasticity. However, the data scientists face several problems once they start planning the use or deployment of any Big Data platform on cloud(s). On one hand, the selection of the appropriate cloud provider(s) is always a cumbersome process since the potential user community has to take into consideration several factors and trade-offs even if they need only a generic Infrastructure-as-a-Service (Iaas) provider: private institutional (e.g. SZTAKI Cloud [3]), federated cloud (e.g. MTA Cloud [4] or pan-European EGI FedCloud [5]) or public cloud (e.g. Amazon [6]) provider is the most suitable for the addressed scenario? Some of them are not enough mature or having strict access and quota policies but less costly, or providing high-standard SLA but more expensive services, etc. Moreover, due to the rapid technology changes, we are the witnesses of the fast evolution of these IT infrastructures, and new lightweight virtualisation and management technologies became available, such as Docker [8] and its surrounding ecosystem. On the other hand, the selection process can be even more complex, once users need widespread and popular platforms, e.g. a Hadoop [7] cluster for Big Data offered by several commercial cloud providers as Platform-as-a-Service (Paas) under various names (HDInsight by Azure [9] or EMR by Amazon [10]). In this case, there is a high chance to face the vendor lock-in problem, and the high fees charged by these companies based on the generated intensive network traffic to/from the cloud sites, which is unavoidable in the targeted Big Data application scenarios. Therefore, the two most distinguishing key features of any cloud orchestrator tool for Big Data platforms are being cloud agnostic, and the capability of building a PaaS platform as flexible as possible, even from scratch, i.e. using only the most basic and common IaaS functionalities that are available on commercial clouds having proprietary solutions or academic clouds based on open-source technologies such as OpenNebula [11] or OpenStack [12].

In this paper, we focus on Apache Hadoop [7], an open-source software framework for storing data in a distributed cluster environment and running applications to process this large amount of data in a fast and efficient way. In the last few years, Hadoop has become a very popular system for analysing Big Data with its MapReduce [13] paradigm introduced by Google in 2004. Many scientific applications, such as numerical weather forecasting [14], DNA sequencing [15], and molecular dynamics [16], have now been parallelised using Hadoop. However, the deployment of a fully functional Hadoop cluster is not a trivial task, it is currently not in line with the capabilities of the average data scientists, and therefore there is still a significant barrier for this

* Corresponding author.
*E-mail addresses:* robert.lovas@sztaki.mta.hu, rlovas@sztaki.hu (R. Lovas), eniko.nagy@sztaki.mta.hu (E. Nagy), jozsef.kovacs@sztaki.mta.hu (J. Kovács).

https://doi.org/10.1016/j.advengsoft.2018.05.002
Received 1 September 2017; Received in revised form 23 March 2018; Accepted 9 May 2018

technology to spread among data scientists.

Combining Hadoop, Cloud and an orchestration tool for dynamically build up Hadoop clusters would help these scientists run their Big Data applications. Complex virtual infrastructures, like Hadoop, with all of its configuration and network design, needs special planning, maintenance and skills by the end-users to have proper functioning Hadoop cluster. One of our main targeted user groups is the Hungarian academic research community and their new computing infrastructure, the MTA Cloud [4], where MTA stands for the acronym 'Hungarian Academy of Sciences'. In 2014, the MTA Wigner Data Center and the Institute for Computer Science and Control (MTA SZTAKI) initiated the MTA Cloud project together as a joint effort to establish a community Cloud for the further mostly non-IT specialised member institutes of the Hungarian Academy of Sciences including among others the Research Centre for Natural Sciences, and the Biological Research Centre. The recently opened OpenStack [12] and Docker container-based cloud infrastructure combine resources from Wigner Data Centre and MTA SZTAKI relying on the nationwide academic internet backbone and other federated services, e.g. eduGain for authentication and authorisation. The total capacity of the two deployed sites is 1160 virtualised CPU with 3.3 TB memory and 564 TB storage facility. More than 30 research teams have started utilising the MTA Cloud since 2016 with no or little experiences with advanced cloud usage scenarios such as multi-VM deployment and orchestration of Big Data tools.

Our presented work focuses on utilising a hybrid, cloud orchestration tool called Occopus [18,19]. The solution, presented in this paper, provides automatic deployment of a fully functional Hadoop cluster without the need for the low-level understanding of Hadoop architecture or cloud computing. Moreover, (1) it is portable since the solution does not depend on any cloud-specific feature, (2) it is scalable by utilising Occopus and Hadoop dynamicity, (3) it does not require any prepared image, (4) it gives the possibility to fine-tune the configuration of the Hadoop components for advanced users and finally (5) it supports short or long-term usage scenarios.

The solution described in the paper focuses on the dynamic deployment and scaling of Hadoop, but not on data locality. Regarding data locality, one of the most expensive tasks is the data transfer between the data storage and Hadoop which can be significant at the level of petabytes. The cost of data transfer can be minimised with two simple scenarios. For example, when the storage and Hadoop deployment is located on the same network (like in our lab where storage and cloud is on the same network segment) the cost can be significantly reduced. Another possible scenario is to reduce the cost (per calculation) by utilising the transferred data by performing many calculations on it. Beyond these two simple scenarios, there can be other application and environment specific solutions.

The paper is structured as follows: in Section 2 we overview some of the related works. A short summary of Hadoop and Occopus can be found in Sections 3 and 4. The details of the solution including the architecture, the explanation of operation and contextualisation will be discussed in Section 6. Performance related results with summary are provided in Section 7. The final Section 8 provides a conclusion with the future plans.

This paper is based on Lovas et al. [20], but the current paper includes the following additional research: automatic scaling of Hadoop cluster, deeper performance evaluation involving further clouds and metrics, detailed cost analysis of the elaborated solution on Microsoft Azure cloud, comparison with more strongly related projects, and many minor enhancements.

## 2. Related work

There have been several efforts to build Hadoop cluster for MapReduce jobs in the cloud.

A solution [23] from the University of Westminster, provides a generic method based on infrastructure aware workflows. A concrete implementation of this concept was presented to integrate Big Data processing based on the MapReduce paradigm and Hadoop to scientific workflow systems. By integration Hadoop with WS-PGRADE workflow tool [29], the user can configure one or more workflow nodes to execute Hadoop. This system can be used to create complex applications for large-scale scientific simulations. Compared to the solution presented by us, the destruction of the infrastructure is obligatory, as this Hadoop infrastructure implementation is based on a workflow system with some limitations (lifetime of jobs/services in the workflow, direct acyclic graph approach, etc.). It makes the building and the decommission dynamically before and after the Hadoop MapReduce jobs. Uploading the necessary data involves high costs, which will be deleted after the job is done. Therefore, this approach is not optimised for maintaining Hadoop clusters for a longer period of time and to work with a large amount of data. In the tested version, there was a missing scaling functionality and it worked on a specific cloud infrastructure (EGI's FedCloud [5]).

CloudMesh [30] is another promising academic toolkit from the Indiana University that enables the management of virtual machines in a multi-cloud environment including Amazon AWS, OpenStack and OpenNebula as well. Its core services and PaaS launcher are able to deploy several multi-VM platforms, e.g. Hadoop clusters, but the automatic scaling of such clusters is not supported by CloudMesh.

Hewlett Packard Laboratories together with University of Murcia elaborated an architecture [31] with the main aim to allow the dynamic deployment of Hadoop clusters in virtual infrastructures provided by either public or private cloud providers. Their solution has many common features with our approach: they install also the necessary Hadoop packages on-the-fly and manage the deployed Hadoop cluster with a specialised tool (SmartFrog). Their presented scalability results (concerning the infrastructure deployment time) seem better however, they used a dedicated HP private cloud with pre-installed SmartFrog and local repository in the cloud.

Amazon Elastic MapReduce (Amazon EMR) [10] is a commercial cloud PaaS that provides a managed Hadoop framework, with dynamically scalable Amazon EC2 instances. By distributing the computation work across a cluster of virtual servers running in the Amazon cloud users can analyse and process vast amounts of data. Open-source projects that run on top of the Hadoop architecture (like Hive, Pig or HBase) can also be used on it. Other commercial cloud providers also offer on-demand solutions for MapReduce applications under various names; Hadoop is a vital part of HDInsight [9] from Microsoft (available on Windows Azure cloud), and Google named its related services as Dataproc [24] including some other open-source Apache tools as well for Big Data applications. All of these solutions are vendor-specific, provided as black-box commercial services from the end-users' point of view. On the other hand, Cloudbreak from Hortonworks [25] is an available tool for provisioning and managing Apache Hadoop clusters across cloud infrastructure providers (including Amazon Web Services, Microsoft Azure, Google Cloud Platform and OpenStack) but its functionalities are limited to Hadoop clusters. Our Occopus based approach is not limited to Hadoop, other (and even more complex) platforms, such as Internet of Things (IoT) backends, can be described, deployed, and managed automatically.

## 3. Hadoop overview

Apache Hadoop is an open source software platform for distributed storage and processing of very large datasets on computer clusters. Due to the special storage method, which is based on a distributed file system (HDFS, Hadoop Distributed File System [17]) Hadoop can process efficiently terabytes of data in just minutes, and petabytes in hours. It is a highly scalable storage platform, it can store and distribute large datasets across hundreds of nodes which operate in parallel. Unlike traditional relational database systems (RDBMS), Hadoop enables to run applications on thousands of nodes. Hadoop's flexibility enables