Accepted Manuscript

An experimental survey on big data frameworks

Wissem Inoubli, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri, Engelbert Mephu Nguifo

PII: S0167-739X(17)32745-0
DOI: https://doi.org/10.1016/j.future.2018.04.032
Reference: FUTURE 4110

To appear in: *Future Generation Computer Systems*

Received date : 26 November 2017
Revised date : 21 March 2018
Accepted date : 10 April 2018

Please cite this article as: W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, E.M. Nguifo, An experimental survey on big data frameworks, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.04.032

An Experimental Survey on Big Data Frameworks

Wissem Inoubli
University of Tunis El Manar,
Faculty of Sciences of Tunis,
LIPAH
Tunis, Tunisia
inoubli.wissem@gmail.com

Sabeur Aridhi
University of Lorraine, CNRS,
Inria, LORIA
F-54000 Nancy, France
sabeur.aridhi@loria.fr

Haithem Mezni
University of Jendouba,
SMART Lab
Avenue de l'Union du Maghreb
Arabe, Jendouba 8189,
Tunisia
haithem.mezni@fsjegj.rnu.tn

Mondher Maddouri
College Of Buisness,
University of Jeddah
P.O.Box 80327, Jeddah 21589
Kingdom of Saudi Arabia
maddourimondher@yahoo.fr

Engelbert Mephu Nguifo
University of Clermont
Auvergne, LIMOS
BP 10448, F-63000
Clermont-Ferrand, France
mephu@isima.fr

ABSTRACT

Recently, increasingly large amounts of data are generated from a variety of sources. Existing data processing technologies are not suitable to cope with the huge amounts of generated data. Yet, many research works focus on Big Data, a *buzzword* referring to the processing of massive volumes of (unstructured) data. Recently proposed frameworks for Big Data applications help to store, analyze and process the data. In this paper, we discuss the challenges of Big Data and we survey existing Big Data frameworks. We also present an experimental evaluation and a comparative study of the most popular Big Data frameworks with several representative batch and iterative workloads. This survey is concluded with a presentation of best practices related to the use of studied frameworks in several application domains such as machine learning, graph processing and real-world applications.

Keywords

Big Data, MapReduce, Hadoop, HDFS, Spark, Flink, Storm, Samza, batch/stream processing

1. INTRODUCTION

In recent decades, increasingly large amounts of data are generated from a variety of sources. The size of generated data per day on the Internet has already exceeded two exabytes [23]. Within one minute, 72 hours of videos are uploaded to Youtube, around 30.000 new posts are created on the Tumblr blog platform, more than 100.000 Tweets are shared on Twitter and more than 200.000 pictures are posted on Facebook [23].

Big Data problems lead to several research questions such as (1) how to design scalable environments, (2) how to provide fault tolerance and (3) how to design efficient solutions. Most existing tools for storage, processing and analysis of data are inadequate for massive volumes of heterogeneous data. Consequently, there is an urgent need for more advanced and adequate Big Data solutions.

Many definitions of Big Data have been proposed throughout the literature. Most of them agreed that Big Data problems share four main characteristics, referred to as the four V's (Volume, Variety, Veracity and Velocity) [41]. The volume refers to the size of available datasets which typically require distributed storage and processing. The variety refers to the fact that Big Data is composed of several different types of data such as text, sound, image and video. The veracity refers to the biases, noise and abnormality in data. The velocity deals with the place at which data flows in from various sources like social networks, mobile devices and Internet of Things (IoT).

In this paper, we first give an overview of most popular and widely used Big Data frameworks which are designed to cope with the above mentioned Big Data problems. We identify some key features which characterize Big Data frameworks. These key features include the programming model and the capability to allow for iterative processing of (streaming) data. We also give a categorization of existing frameworks according to the presented key features. Then, we present an experimental study on Big Data processing systems with several representative batch, stream and iterative workloads.

Extensive surveys have been conducted to discuss Big Data Frameworks [49] [33] [37]. However, our experimental survey differs from existing ones by the fact that it considers performance evaluation of popular Big Data frameworks from different aspects. In our work, we compare the studied frameworks in the case of both batch processing and stream processing which is not studied in existing surveys. We also mention that our experimental study is concluded by some best practices related to the usage of the studied frameworks