Chapter 42
Open Source Software (OSS) for Big Data

Richard S. Segall
Arkansas State University, USA

ABSTRACT

This chapter discusses Open Source Software and associated technologies for the processing of Big Data. This includes discussions of Hadoop-related projects, the current top open source data tools and frameworks such as SMACK that is acronym for open source technologies Spark, Mesos, Akka, Cassandra, and Kafka that together compose the ingestion, aggregation, analysis, and storage layers for Big Data processing. Tabular summaries and categories for 38 Open Source Statistical Software (OSSS) are provided that include for each listing of features and URLs for free downloads. The current challenges of Big Data and Open Source Software are also discussed.

OPEN SOURCE SOFTWARE AND TECHNOLOGY FOR BIG DATA

In the past, companies had been writing big checks to database corporations such as Oracle, Microsoft and IBM. After 2000, Google started to encounter a problem that the data they collected were so large that no single database vender will be able to store and process their data anymore, and hence the need for Big Data technology evolved as well as Big Data Analytics.

Balihausen (2019) indicated that the percentage of FOSS (Free and Open Source Software) in the average application exceeds the amount of proprietarily applications exceeds the amount of proprietarily licensed code. According to 2019 Open Source Security and Risk Analysis (OSSRA) report published by Synopsys Cybersecurity Research Center (2019), open source represented 60% of the code analyzed in 2018, up from 57% in 2017, and 64% Open Source Software (OSS) was used for financial services, Big Data, artificial intelligence, business intelligence machine learning.

An entire book on the technology of hands-on approaches to Big Data has been published by Bahga and Madisetti (2016). The following are a few of the tools and frameworks for "batch processing" of Big Data.

DOI: 10.4018/978-1-7998-9158-1.ch042

Several studies have been completed with the "best" Open Source Software for Big Data in specific categories. These include those of Freeman, Garza et al. (2018) that discussed the best Open Source Software for cloud computing, Heller et al. (2018) the best Open Source Software for data storage analytics, Heller and Pointer (2018) the best Open Source Software (OSS) for machine learning, and Freeman, Heller et al. (2018) the best open software for software development.

Riehle (2019) discussed that the needs of open source processes have led to two major tool investigations that have since become an important part of corporate software development: Software forges and distributed version control. A software forge is a website that allow the creation of new projects and provides developers with all of the tools needed for software development. According to Riehle (2019), distributed version control is version control in which one copies the original repository for Big Data and work with your copy that does not need commit rights or permission to start work. Git and Mercurial are the two best-known examples of such software for Big Data.

Harvey (2017b) provides a detailed study of the top 35 open source companies that use Big Data and play a major role in developing and maintaining the Open Source Software that powers today's businesses.

Frampton (2018) published a complete guide to open source Big Data stack that includes components for visualization, resource management, framework queueing, processing, storage monitoring, and resource management that interact with Apache CloudStack. The following Table 1 provides representative Open Source Software for Big Data for each of these components of Big Data Stack as presented by Frampton (2018) with many of which are discussed in more depth below Table 1 and elsewhere in the following part of this chapter.

Table 1. Representative Open Source Software (OSS) for Big Data stack

Big Data Stack Component	Representative Open Source Software (OSS)
1. Visualization	Zeppelin
2. Resource Management	Mesos
3. Frameworks	Akka Spring
4. Queueing	Apache Kafka
5. Processing	Apache Spark
6. Storage	Hadoop Distributed File System (HDFS) Riak Apache Cassandra
7. Monitoring	Brooklyn Mesos
8. Release Management	Brooklyn Mesos

[Derived using Frampton (2018)]

Apache Brooklyn

Apache Brooklyn is an open-source framework for modeling, deploying and managing distributed applications defined using declarative YAML blueprints. (Wikipedia, 2019a)