Hala El-Ali
CSE 8331 – Data Mining
Tutorial
March 30, 2015

Abstract

As data is produced, captured, and stored at ever increasing rates, while the rate of hardware growth is slowing down, IT companies have turned to software to find solutions for processing scalability issues. One such solution is Hadoop, a framework for processing Big Data using clusters of 1000s of nodes of commodity hardware. Hadoop core components offer distributed file system (HDFS) and distributed compute model (MapReduce). Because Hadoop core components provide only low-level services, an ever increasing number of projects have been started to provide additional higher-level services and to bridge between Hadoop and existing legacy systems. These projects, together with the main components, are referred as to Hadoop ecosystem and they can be classified as: persist, run, managed, security, transfer, data interaction, data analytics, discovery and visualization. Hadoop is expected to continue be the de-facto platform for Big Data in the near future and basic knowledge of this framework and its extensions is a must for anyone involved in data processing.

Keyword: Hadoop, Big Data, HDFS, MapReduce.