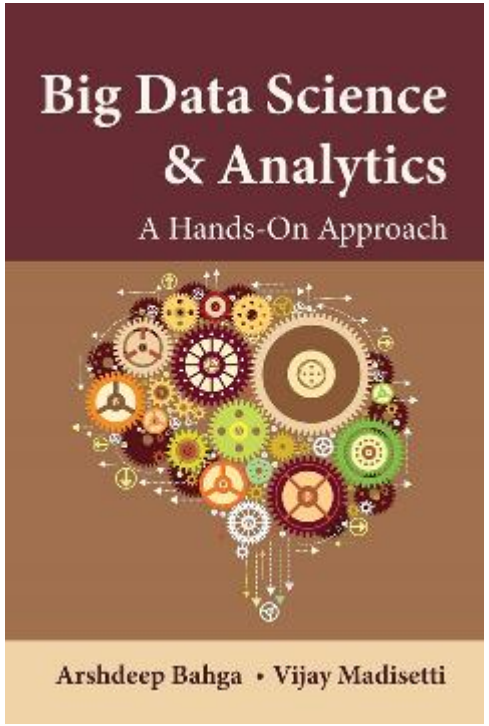
Big Data Stack Setup and Examples

Prof. Gheith Abandah

Reference



- Chapter 2: **Big Data Stack Setup and Examples**

- Arshdeep Bahga and Vijay Madisetti, **Big Data Science and Analytics: A Hands-On Approach**, 2019.
  - Web site: http://www.hands-on-books-series.com/

Outline

- Hortonworks Data Platform (HDP)
- Cloudera CDH Stack
- Amazon Elastic MapReduce (EMR)
- Azure HDInsight

Hortonworks Data Platform (HDP)

- Is an **open-source platform distribution** comprising of various big data frameworks for data access and integration, batch processing, real-time processing, interactive querying, security and operations tools.
- The book recommends using Apache Ambari to set up this platform.
- Ambari is a tool for provisioning, managing and monitoring clusters that run these frameworks.
- Ambari cluster can be setup on machines running certain OSs such as **Linux**.
- The books describe how to set up and HDP stack with Ambari on an **Amazon EC2** instance running **Ubuntu**.

Cloudera CDH Stack

- Cloudera CDH is an **open-source platform distribution** that includes various big data tools and frameworks.
- There are various methods to setup a CDH stack, the easiest one being the automated method using Cloudera Manager.
- Cloudera (the company behind **CDH**) and Hortonworks (the company behind **HDP**) **have merged**. They now are called Cloudera.
- After the merger, a new distribution was released, called the Cloudera Data Platform (CDP).
- Installing CDP on Google Cloud (YouTube)

Amazon Elastic MapReduce (EMR)

- Amazon EMR is **cloud big data platform** for processing vast amounts of data using open-source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto.

Azure HDInsight

- Azure HDInsight service allows you to setup managed clusters running the **Hadoop ecosystem of components**.
- HDInsight cluster can be created from the Azure portal, which include Hadoop, HBase, Spark, and Storm.