Abstract

A Tutorial on Apache Spark

Harold Mitchell

Ten years ago, there was much hype around Big Data and Hadoop. Moreover, there was much excitement around Map-Reduce. Back then this was the primary tool for batch processing of big data. If we fast forward to today, big data continues to receive much attention. However, today many organizations need the ability to process big data in real-time. This requirement extends beyond the capabilities and intended usages of Map-Reduce. So, now we have Apache Spark. Apache Spark is a tool like Map-Reduce. But, it provides a vast array of features from a unified API. Most notable is this tool's primary data sharing abstraction known as Resilient Distributed Datasets or RDDs. Additionally, Spark addresses a variety of big data processing loads with a single processing engines. This engine includes: 1) Spark Streaming 2) SparkSQL 3) MLlib for machine learning 4) GraphX for graph networks. Spark's ability to process a variety of big data processing loads has made the tool worthy of adoption in many organizations across a wide range of business sectors. In this tutorial, we will take a practical look at Apache Spark. Also, a considerable amount of time will be devoted to RDDs. RDDs are probably the most widely used component of Spark. Finally, we will discuss the use case argument for Apache Spark and explain why organizations should invest in this technology.