Open source Google-style large scale data analysis with Hadoop

Abstract

Our era is marked by what is referred to as the data explosion: Increasing volumes of data that need to be stored, indexed and queried for every company (such as e-mail and web logs, historical data, click streams, etc). Even Small to Medium enterprises such as start-ups with a limited budget for hardware resources and software licenses quickly come across these needs with data-intensive applications such as web indexing, data mining, log file analysis, machine learning, financial analysis, scientific simulation, bioinformatics research, etc.

Apache Hadoop is an open source Java software framework that supports data-intensive distributed applications. Hadoop enables applications to work with thousands of nodes and petabytes of data in a seamless, highly parallel and fully distributed way. Its development was mainly inspired by Google's MapReduce and Google's File System, two research papers published by Google's employees. Currently, even Amazon is offering on-demand virtual clusters with preinstalled Hadoop instances through its ElasticMapReduce service.

In this presentation, we will talk about Hadoop's architectural components and we will show how some typical data intensive problems can be easily solved through the MapReduce framework. We will discuss about some open source applications that are built on top of Hadoop and we will present how Hadoop is used by a number of companies and organizations worldwide.