Hadoop Alternatives

Sources

Compiled from several sources:

- "Spark, Hydra & BigQuery: 5 enterprise alternatives to Hadoop", by James Nunns, 12th May 2016: CBR identifies five Hadoop alternatives that may better suit your business needs.
- "Three solid real-time Big Data alternatives: Spark, Storm and DataTorrent RTS", BBVA, 7th August 2015: Real time business intelligence tools: Spark, Storm or DataTorrent RTS

Introduction

Hadoop Complexity

Hadoop's progression from a large scale, batch oriented analytics tool to an ecosystem full of vendors, applications, tools and services has coincided with the rise of the big data market.

While Hadoop has become almost synonymous with the market in which it operates, it is not the only option. Hadoop is well suited to very large scale data analysis, which is one of the reasons why companies such as Barclays, Facebook, eBay and more are using it.

Although it has found success, Hadoop has had its critics as something that isn't well suited to the smaller jobs and is overly complex.

Real time streaming calculations and business intelligence tools

Data, data, data. Value, value,value. And if possible, in real time. The concept of real-time business intelligence has been on the market for some time, but until very recently only a limited number of companies used it. Today, Hadoop's stability makes it the most commonly used platform for analyzing large volumes of data, but when streaming calculations are needed, solutions such as Spark, Storm or DataTorrent RTS are a great choice.

These kinds of practices used to have no real market penetration, for two main reasons: the first, obviously, was the **lack of real-time business intelligence tools;** the second, that existing solutions **were only geared to batch data analysis and were expensive**. Spark, Storm and DataTorrent RTS provide a solution to these two problems.

Pachyderm

Pachyderm, put simply, is designed to let users store and analyse data using containers.

The company has built an open source platform to use containers for running big data analytics processing jobs. One of the benefits of using this is that users don't have to know anything about how MapReduce works, nor do they have to write any lines of Java, which is what Hadoop is mostly written in.

Pachyderm hopes that this makes itself much more accessible and easy to use than Hadoop and thus will have greater appeal to developers.

With containers growing significantly in popularity of the past couple of years, Pachyderm is in a good position to capitalise on the increased interest in the area.

The software is available on GitHub with users just having to implement an http server that fits inside a Docker container. The company says that: "if you can fit it in a Docker container, Pachyderm will distribute it over petabytes of data for you."

Apache Spark

What can be said about Apache Spark that hasn't been said already? The general compute engine for typically Hadoop data, is increasingly being looked at as the future of Hadoop given its popularity, the increased speed, and support for a wide range of applications that it offers.

However, while it may be typically associated with Hadoop implementations, it can be used with a number of different data stores and does not have to rely on Hadoop. It can for example use Apache Cassandra and Amazon S3.

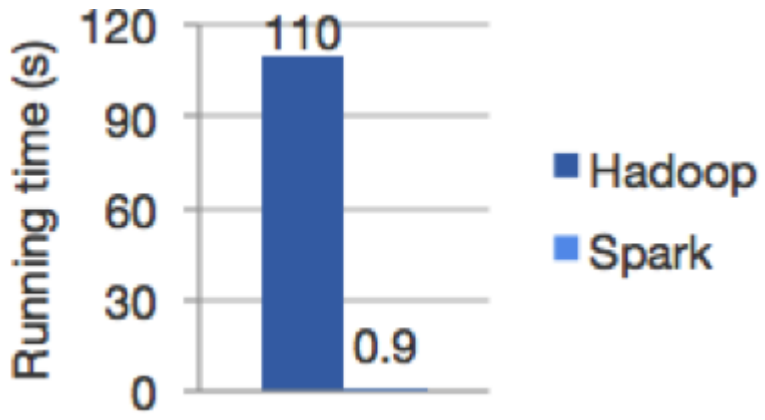
Spark is even capable of having no dependence on Hadoop at all, running as an independent analytics tool.

Spark's flexibility is what has helped make it one of the hottest topics in the world of big data and with companies like IBM aligning its analytics around it, the future is looking bright.

Apache Spark is undoubtedly the great new star of Big Data analytics. It is an open-code platform for processing data in real time, and may be executed and operated using four types of different languages: Scala, the syntax in which the platform is written; Python; R; and Java. The idea of Spark is to offer advantages in the handling of constant data entries with speeds far above those offered by Hadoop MapReduce.

Some of its key features are:

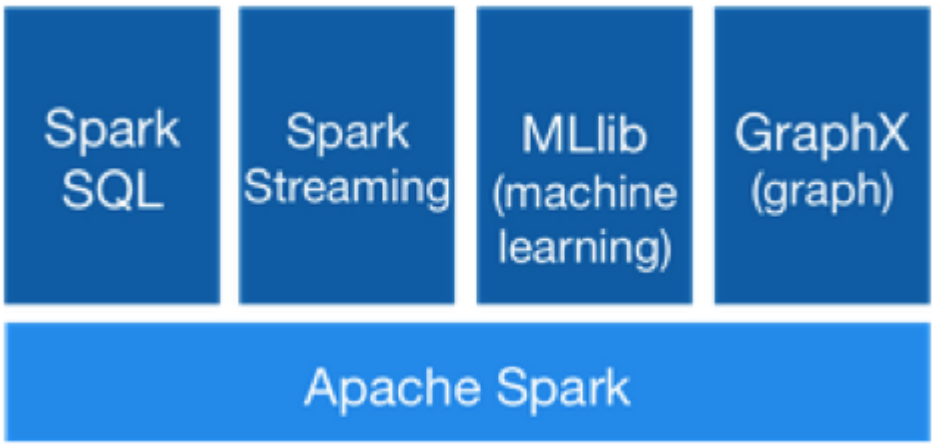
– **Speed in the calculation processes in memory and on disc**: Apache promises a calculation speed 100 times quicker than that currently offered by Hadoop MapReduce in memory and 10 times better in disc.



– **Execution on all types of platforms**: Spark can be executed on Hadoop, Apache Mesos, and EC2, in independent cluster mode or in the cloud. In addition, Spark can access numerous databases such as HDFS, Cassandra, HBase or S3, Amazon's data warehouse.

– **It incorporates a package of very useful tools for developers**: the MLlib library for implementing automated learning solutions and GraphX, Spark's API for computation services with graphs.

– **It has other interesting tools**: Spark Streaming, which allows the processing of millions of data among the clusters, and Spark SQL which makes it easier to exploit the data through the SQL language.



Apache Storm

Apache Storm is an open-source distributed real-time computation system. It allows the simple and reliable processing of large volumes of analytics data (for example, for the continuous study of information from social networks), distributed RPC, ETL processes…

While Hadoop carries out batch data processing, **Storm does it in real time**. In Hadoop the data are entered in a file system (HDFS) and then distributed through the nodes to be processed. When the task is complete, the information returns from the nodes to HDFS to be used. In Storm there is no process with an origin and an end: the system is based on the construction of Big Data topologies that are transformed and analyzed in a continuous process of information entries.

That is why Storm is something more than a system of Big Data analytics: it is a system for Complex Event Processing (CEP).

This type of solution allows companies to respond to the arrival of sudden and continuous data (**information collected in real time by sensors**, **millions of comments generated on social networks such as Twitter, WhatsApp and Facebook**, **bank transfers…**).

It is also of particular interest for developers for a number of reasons:

– It can be used in various programming languages. Storm has been developed in Clojure, a dialect of Lisp which is executed in Java Virtual Machine (JVM). Its great strength is that it offers compatibility with components and applications written in various languages such as Java, C#, Python, Scala, Perl and PHP.

– It is scalable.

– It is fault-tolerant.

– It is easy to install and operate.

DataTorrent RTS

DataTorrent RTS is an open-source solution for the batch or real-time processing and analysis of big data. It is an all-in-one tool that aims to revolutionize not only what can be done in the Hadoop MapReduce environment, but also **what is already offered in Spark and Storm in performance**. The platform is capable of processing billions of events per second and recover any node outages with no data loss and no human intervention.

Some of its key features include:

– Guaranteed event processing.

– High in-memory performance.

It is scalable.

– Fault-tolerance at platform level.

– Easy to execute.

– Applications programmed in Java.

This Big Data solution provides mechanisms for ingesting data from many different sources, directly from external databases or through their integration with native corporate applications. DataTorrent RTS provides technical teams with a group of connectors previously developed for SQL and NoSQL databases, Apache Sqoop, Apache Kafka, Apache Flume and social networks such as Twitter… Anything that generates data.

Google BigQuery

Google seemingly has its fingers in every pie and as the inspiration for the creation of Hadoop, it is no surprise that the company has an effective alternative.

The fully-managed platform for large-scale analytics allows users to work with SQL and not have to worry about managing the infrastructure or database.

The RESTful web service is designed to enable interactive analysis of huge datasets working on conjunction with Google storage

Users may be wary that it is cloud-based which could lead to latency issues when dealing with the large amounts of data, but given Google's omnipresence it is unlikely that data will ever have to travel far, meaning that latency shouldn't be a big issue.

Some key benefits include its ability to work with MapReduce and Google's proactive approach to adding new features and generally improving the offering.

Presto

Presto, an open source distributed SQL query engine that is designed for running interactive analytic queries against data of all sizes, was created by Facebook in 2012 as it looked for an interactive system that is optimised for low query latency.

Presto is capable of concurrently using a number of data stores, something that neither Spark nor Hadoop can do. This is possible through connectors that provide interfaces for metadata, data locations, and data access.

The benefit of this is that users don't have to move data around from place to place in order to analyse it.

Like Spark, Presto is capable of offering real-time analytics, something that is in increasing demand from enterprises.

Presto supports standard ANSI SQL, including complex queries,

aggregations, joins, and window functions. Lovers of Java will be happy to hear that this is what the system is implemented in.

Hydra

Developed by the social bookmarking service AddThis, which was recently acquired by Oracle, Hydra is a distributed task processing system that is available under the Apache license.

It is capable of delivering real-time analytics to its users and was developed due to a need for a scalable and distributed system.

Having decided that Hadoop wasn't a viable option at the time, AddThis created Hydra in order to handle both streaming and batch operations through its tree-based structure.

This tree-based structure means that can store and process data across clusters that may have thousands of nodes.

Hydra features a Linux-based file system in addition to a job/client management component that automatically allocates new jobs to the cluster and rebalances existing jobs, it is also capable of automatically replicating data and handling node failures.