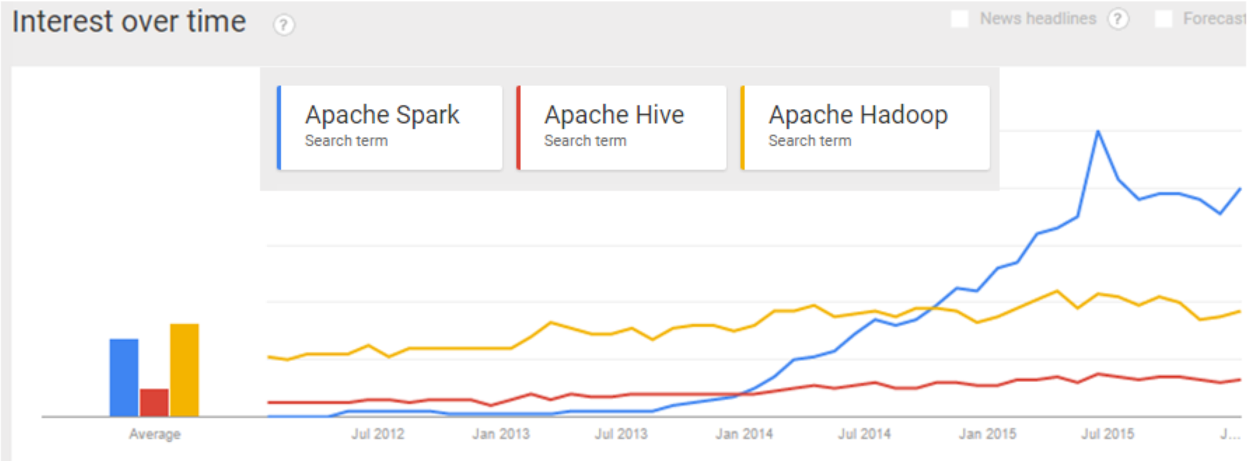
Why Spark is the new R?

As companies are moving more and more data to Hadoop, the analysts/ data scientists using them are realizing the limitations of tools they have used in past - SAS/ SPSS/ R/ KXEN. While R had promise of being able to connect to Hadoop and other tools are building connectors as well the capability they typically provide is a small fraction of what is possible on original tool. So then what is the next stage of evolution in the world of data science? There were a number of candidate earlier but for me one tool that has clear edge over others is Spark.

Before I go on about why Spark is better, I thought about checking if other felt the same. Below is a snapshot from Google Trends showing Spark in comparison to Hadoop and Hive (the most popular related Apache Project amongst Hive/ Pig/ Mahout). It can be concluded that interest in Spark has grown rapidly and still increasing.



Having looked closely at Spark (as a company we also offer Spark Training to our corporate clients), I am going to summarize the top reasons that I feel are going to make Spark a very popular tool for Data Scientists.

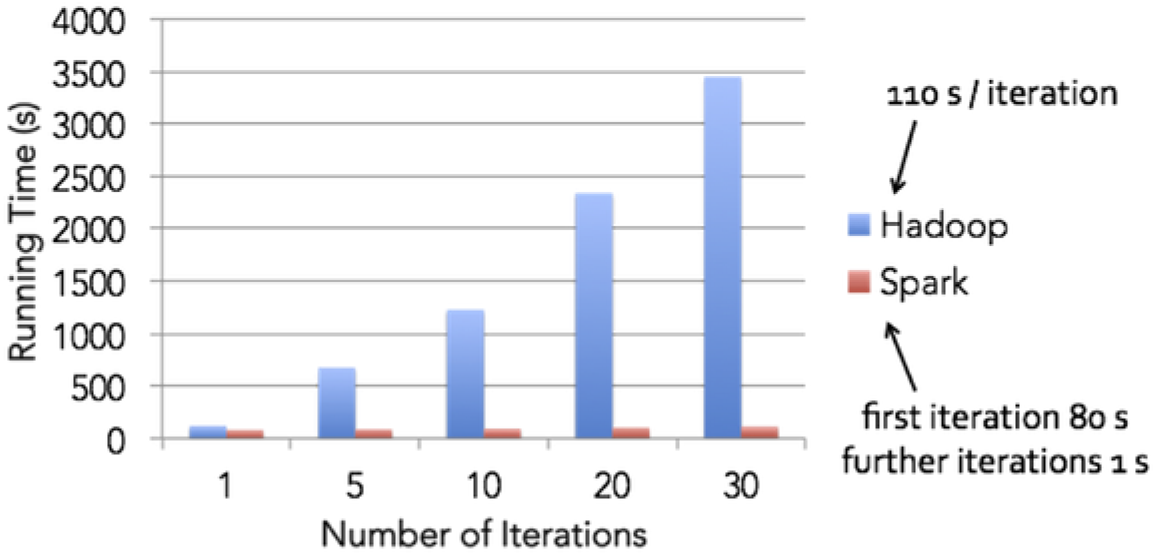
1. It is Omnidextrous: I don't know this is a proper word or not but I feel for any tool to become popular it has to support data transformation and advanced algorithms within the same environment. It is surprising how many modeling tools don't recognize this fact. Algorithm development and data transformations go hand in hand and any tool which allows data scientists to do both side by side is going to be a hit. This I feel was a major factor in success of SAS and R. Spark takes it one step further and allows to mix SQL, Machine Learning and Streaming seamlessly (at least it is working towards that). So this is the only tool a data scientist may need in time to come.

2. Hadoop is popular: Hadoop is great for data storage of both structured and unstructured data. Many companies have already moved or are in the process of moving their internal data to Hadoop. Hadoop was always a popular destination for storing the external / social media data. So now you have a scenario where all the data is going to be on one platform. Now companies are going to need tools to mine all the data. This is where Spark comes in. The fact that Spark can be installed over Yarn/ Hadoop 2.0 is important and means Spark can work directly with data that is already in HDFS.

Enterprise Hadoop is a market that is not even 10 years old, but Forrester estimates that 100% of all large enterprises will adopt it (Hadoop and related technologies such as Spark) for big data analytics within the next two years.

3. It is Open Source and free: Again zero software cost is a big plus. No doubt there would be companies, incorporating Spark in their Hadoop distributions. But the low cost is going to be factor in adoption.

4. It is fast: The in-memory processing means that Spark can perform faster than most other tools around. There is enough evidence which suggest 10-100 X improvement in many cases. In our own test, Spark SQL comprehensively outperforms same Hive query even when data is stored in Hive Metastore. I am sure there would be claims that there are specialized applications which run faster than Spark for a particular technique and they may be right. However, in general Spark would be way faster for doing a variety of operations compared to anything that is available right now.



5. Support for multiple languages: Spark allows users to write codes in Scala, Java, Python and now R. Also allows for mixing of classes/ elements with Java and Python. So it is not a Scala or nothing. Data scientists used to other languages can get started faster. However, for more complex processing it would be better in longer run to shift to Scala. I also feel I must make a disclosure that although I don't have any directly links with Spark development, as an analytics consulting company, MathLogic, has vested interest. We use and recommend Spark as appropriate to our clients and also engage in training our corporate client in Spark.