Spark My MapReduce? A Shift in Processing Technology

There's a major change going on at the heart of big data. The data center industry is shifting from Hadoop processing frameworks and toward Spark data processing engines. The difference is important, and it's not just techies who need to understand what it means: business leaders, too, should learn the ramifications and understand how the shift will affect their companies.

Why consider Spark?

Virtually all enterprises generate data. But how that data gets used varies from basic, retrospective analysis in spreadsheets to dynamic, real-time analysis, which may include predictive modeling and advanced visualization.

For organizations that are ambitious about the use of their data, the quest to achieve speed and flexibility has been hampered by the limitations of physical hard drives. That situation started to change when in-memory solutions entered the mainstream market, virtually eliminating past limitations.

Before we say any more, following are some important definitions.

Hadoop is a framework for distributed and scalable processing of large data sets (big data) across computer clusters. A number of tools have been developed to run on top of or alongside Hadoop, commonly referred to as the ***Hadoop ecosystem***.

MapReduce is the original processing engine in Hadoop that brought about faster and easier implementation of existing parallel processing approaches, particularly for text processing.

Spark is a fast, general engine for large-scale data processing that can run on top of Hadoop and replace MapReduce as its computing engine. MapReduce works only with (slower) storage disk operations, whereas Spark makes heavy use of in-memory techniques to achieve speeds up to 100 times faster than its predecessor.

The most important difference between Spark and MapReduce that business users have to consider is that the latter relies solely on hard disk physical memory, which limits its utility as an interactive analytics platform. It is similar to preparing a meal and having to walk to the drawer at each step to get an ingredient. The cooking process goes much faster when all the ingredients have been laid out on the worktop, which is what Spark's in-memory computing technology does by eliminating unnecessary and time-consuming tasks. That's why Spark is much faster, claiming to speed up processing by 100 times the MapReduce rate. But even though—in the same way as you can speed up cooking prep time—you can preprocess data sets for MapReduce (a

number of custom applications do so) and you can justify the processing upgrade, it is important to stress that Spark complements (but does not replace) the Hadoop ecosystem for the same reason. We still require physical drawers to store ingredients.

What you can do with your data: Spark use cases

The journey from data to insight and then to action enables companies to focus on what actually matters: using existing and new data sources to deliver cost savings, raise productivity, and increase revenue. Your business may already be using Hadoop to crunch your latest customer or product data, or you may be considering such a transition. Either way, significant advantages can be gained by incorporating Spark into the solution.

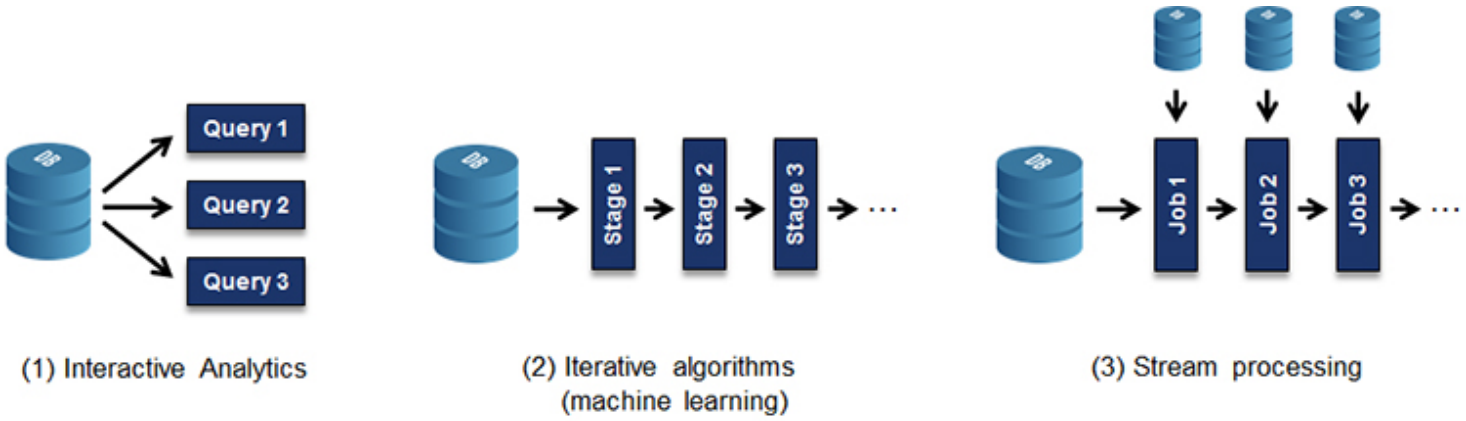
From a technical perspective, Spark's major improvement over MapReduce leads to the performance of iterative steps and calculations with much greater efficiency. For instance, a typical job that uses MapReduce to train a model complex enough to analyze a billion records so as to help predict customer response to a new promotion could take about 58 hours, assuming it uses one-thread processing. Spark could accomplish the same task in 15 minutes, thereby saving 57 hours and 45 minutes. That single change saves a significant amount of time and could result in certain other tangible benefits to your business (figure 1) as follows.

(1) Interactive big-data analytics: You can now query your large data sets in seconds rather than minutes or hours.

(2) Large-scale machine learning: You are no longer restricted to building predictive models based on only a sample of your data, and you can now more quickly deploy insights from these models into actions.

(3) Stream processing: This enables you to process and respond more quickly to live streams of data, facilitating real-time decisions.

Figure 1: Use cases in which Spark outperforms MapReduce



Source: AlixPartners

How you do it, people you'll need, and things to watch out for

Spark is not a full ecosystem like Hadoop, but it can enhance Hadoop when the two are deployed jointly. By using other Hadoop components (HDFS, HBase, Hive), you can have access to a platform in which both real-time and batch data processing coexist. The decision to deploy such a system from the outset or have it available in the cloud depends on your particular business but will also influence how much you'll have to do to make it happen.

Because part of the burden relies on administering the new systems and the licensing costs are low—because it's largely open source, 24-7 support is recommended. However, you still require a specific information technology skill set, with a larger investment required if you deploy it on your own hardware. The end user, a data scientist, or a business analyst who sits on the other side of the know-how spectrum might also affect the degree to which Spark is embraced. Here Spark shines again with support for widely used languages in data analysis such as Python, R, and SQL, thereby reducing the need for new staff hires.

In any case, enterprises should watch out for the drawbacks of being early adopters and should manage their implementation phases carefully by starting to deploy Spark to noncritical systems.

Building the business case

Adopting Spark brings business benefits from its speed and power. That's often the case at companies whose business units sell different product groups to the same customers, whereby a single, real-time data-led viewpoint facilitates collaboration and opens cross-selling opportunities. A business that makes this change may require help in identifying opportunities and may require support with its investment decisions. Carefully planned implementation helps make sure that migration projects yield the most benefit across the board.

Taking the guesswork out of business data processing choices is a bit like treating every customer to a cost-effectively-cooked meal created by a professional chef who is well-informed about customer preferences and needs.

For comments or additional information, contact:

Drew Carter
Managing Director
dcarter@alixpartners.com
+1 (646) 469-6758

Mark Giles
Managing Director
mgiles@alixpartners.com
+44 20 7098 7594

David Branch
Director
dbranch@alixpartners.com
+44 7876 344 859