IBM Analytics
White Paper

The next wave of intelligent applications, powered by Apache Spark

Brandon MacKenzie and Joel Horwitz
IBM Corporation

Life revolves around prediction—for example, the route you take to get to work, whether to go on a second date, or whether or not to keep reading this sentence are all forms of prediction. Predicating our future is very much tied to progress. We use it to help us plan our lives so we can increase our likelihood of success.

However, human judgment is intrinsically fallible. There is so much data out there today that no one can possibly process it all. For example, many companies have the data that can tell them how their customers actually feel, and when and why those customers might switch to a competitor. The problem is that most companies do not know what they don't know.

There is hope from a field called *machine learning*. Machine learning is changing not only how we interact with machines, but how we relate to the world around us. During the past decade, machine learning has given us self-driving cars, speech recognition, effective web search and a vastly improved understanding of the human genome.

Machine learning is defined as *systems that can learn from data*. Data is the teacher. Rather than explicitly programming a computer to do something, you provide a machine learning algorithm with examples from which a machine can learn a particular model. These examples are commonly referred to as *training samples*. The learning algorithm runs through these training samples to build model coefficients, much like a person builds muscle memory of what to do in certain situations.

Different machine learning algorithms learn at different rates and, like humans, can sometimes benefit from extra effort. However, many machine learning algorithms consume a lot of compute resources and require a long time to learn—they are *computationally intensive*. When machine learning algorithms are given more resources and opportunities to learn on larger, more comprehensive data sets, they become better at making predictions.

So far, the most innovative applications of machine learning have been owned by a select few. The barrier to entry for developing and productizing machine learning has been too high for most corporations. Most companies simply do not have the correct skill set or the necessary technology. However, the next big wave in this field is all about democratizing machine learning from a few to many. It's about enabling everyone to build smarter applications that can serve and interact with our world.

IBM is committed to making Apache Spark™ the engine that will power this next wave of machine learning. Apache Spark is an open source project—it's not a product. Simply put, it is an application framework for doing highly iterative analysis that scales to large volumes of data. Apache Spark provides a platform to bring application developers, data scientists and data engineers together in a unified environment that is easy to use. It is an open source in-memory compute engine powering a stack of high-level tools including Spark SQL, MLlib for machine learning, GraphX and Spark Streaming. You can combine these libraries seamlessly in the same application (see Figure 1).

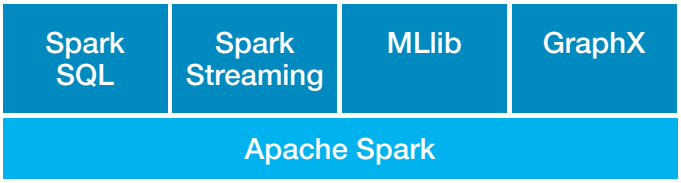

Figure 1. Apache Spark stack

The Apache Spark core engine and application programming interface (API) are major improvements over earlier processing frameworks for distributed computing, such as MPI and MapReduce. The Spark high-level API is much easier to use than those previous lower-level options and its in-memory compute engine is built from the ground up for lightning-fast distributed computation. The engine is well-suited to iterative algorithms such as machine learning. With Spark, these algorithms can execute up to 100 times faster than on MapReduce.[1]

Through its powerful engine and tooling, Apache Spark significantly lowers the barrier to entry for building analytics applications. It reduces the time and complexity around developing analytic workflows. As applications get smarter and more customized through interactions with data, devices and people, previously untapped opportunities become available. We can take on what might have been seen as unsolvable problems by using all of the information that surrounds us and bringing insight to our fingertips when it's needed most.

Over the next five years, machine learning applications will lead to new breakthroughs that will amplify human abilities, assist us in making good choices, look out for us, and help us navigate our world in interesting new ways. Here are some examples of how you can get started with Apache Spark right now to build your own intelligent analytics applications.

Natural language processing: The most expressive and insightful interactions you have with your customers are captured in unstructured form. During conversations, customers often leak the information you need to provide them with a personalized, interactive experience. Far too often, companies capture important information about what their customers think and feel, only to let that information remain unused.

Natural language processing techniques, such as Spark MLlib *term frequency-inverse document frequency (TF-IDF)*, can turn an unstructured body of text into information you can use to teach a machine learning algorithm. TF-IDF is the type of technique you often find in search engines. With Spark MLlib, you can bake natural language processing directly into your applications so that you can proactively manage customer interactions.

Prescriptive analytics: You can go further with prescriptive analytics, which predicts not only that something will happen, but the reason why it is going to happen and what you should do about it. For example, you can use machine learning to determine which attributes have the most predictive power in forecasting customer actions (that is, *attribute importance*). When you know why customers act the way they do, you can intervene in a personalized way through systems of engagement. In short, machine learning can enable you to offer a tailored next best action when it's needed most.

Intelligence of an army: As smart and as powerful as a single person may be, a group of specialists can more effectively work together to win a battle. Machine learning is no different. Spark MLlib has support for machine learning techniques called *ensembles*. With ensembles, many different models collaborate to make better predictions. This technique is well-suited for the massively parallel horsepower of Apache Spark.

Real-time machine learning: With Apache Spark, you can develop and deploy applications that can actually learn in real time. Spark Streaming and MLlib can work together to make your applications more adaptive on the fly. For example, the MLlib *streaming K-means* implementation is a technique that learns dynamically, which is useful when patterns in the data change over time. This method enables your applications to focus on what's important in the moment.

Automating automation: Machine learning applications need automation and optimization. Automating machine learning is an area where Apache Spark really shines. For example, with Spark you can automatically determine the best way to train your learning algorithm, a technique commonly referred to as *hyperparameter tuning*. The Spark community is leading the way in this area—and IBM is excited to help accelerate the charge by contributing its expertise in automation and optimization to Apache Spark.

The problem is that the human mind cannot possibly process all of the insight flowing from big data. Machine learning is the answer to this problem, through its capacity to augment our decision making in the moment to deliver transformative business outcomes. We are already seeing machine learning powered by Apache Spark changing the face of innovation at IBM. We want to bring the rest of the world along with us.

For more information

Learn how to create machine learning models yourself with our free training materials at BigDataUniversity.com, or head over to the Spark Technology Center at http://www.spark.tc to find out more about this technology.

About the authors

Brandon MacKenzie is the Data Science on Hadoop leader on IBM's Worldwide Technical Sales team for Analytics Platform. Brandon is an expert on statistical processing in Hadoop and HPC environments. He earned his master's degree from The University of Edinburgh.

Joel Horwitz is the Worldwide Director of Portfolio Marketing for the IBM Analytics Platform. He graduated from the University of Washington in Seattle with a Masters in Nanotechnology (focus in Molecular Electronics). He also earned an International MBA in Product Marketing and Financial Management from the University of Pittsburgh.

IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
June 2015

[1] https://spark.apache.org/

IMW14825-USEN-00