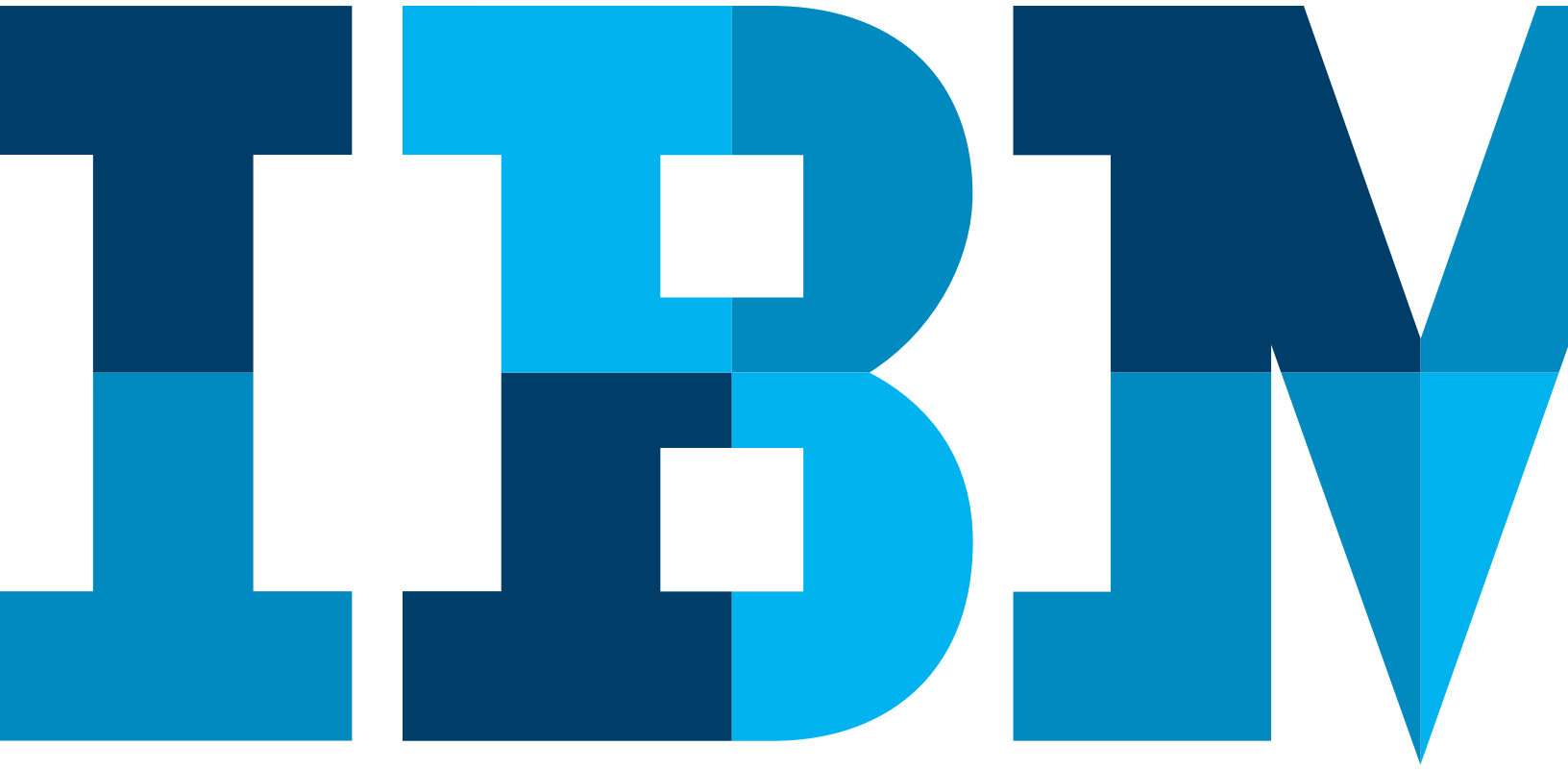
IBM Analytics
White Paper

September 2016

The democratization of machine learning

Apache Spark opens the door to the rest of us



IBM

Contents

Machine learning automates the development of analytic models that can learn and make predictions on data. It has been one of the fastest growing disciplines within the world of statistics and data science, but the barrier to entry has been high, not only in cost, but also in the need for specialized talent.

That was before IBM® Apache® Spark™.

Spark is an open-source cluster-computing framework that has all of the features required to use or develop algorithms and machine learning models. It includes a set of core libraries that enable various analytic methods, which can process data from many sources. Spark has rapidly found traction at all levels, from startups to large enterprises and industries, because it has been developed with dual objectives of ease of use and in-memory capabilities that accommodate big data.

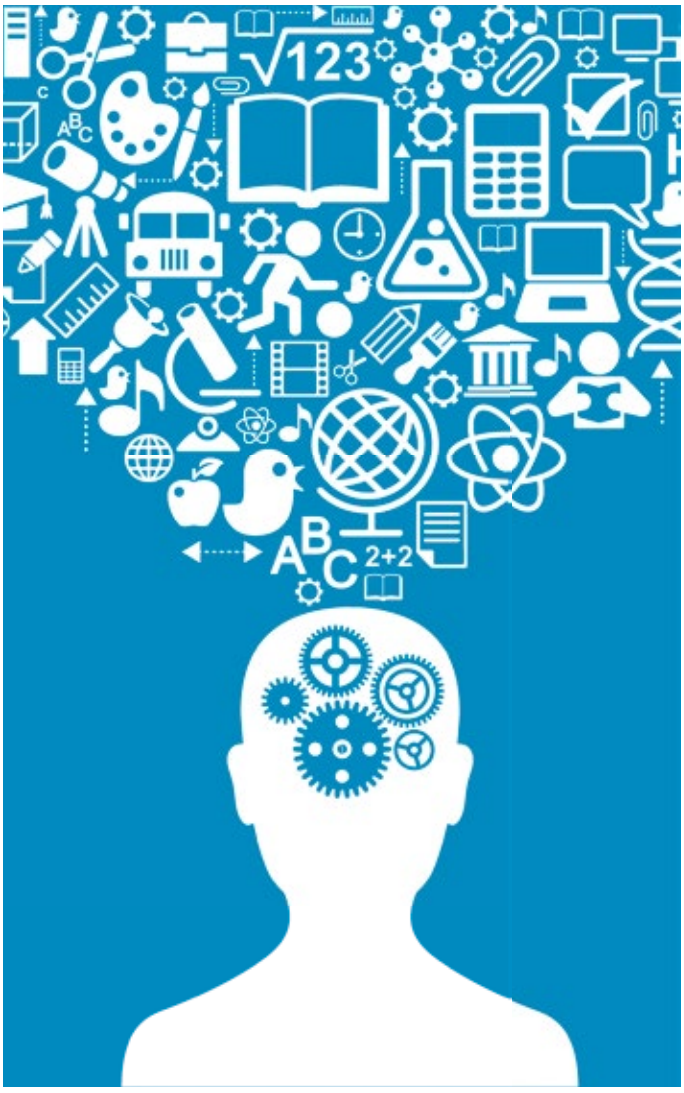
Spark has a dedicated library for common machine learning and statistical algorithms called MLlib, which is why Spark is rapidly becoming the gateway technology to machine learning. Just as web browsers opened up the internet and brought us the digital era, Spark is transforming machine learning by democratizing it to a far broader group of data professionals. Machine learning is now entering the mainstream, serving as the backing methodology for competitive advantage in a wide array of industries, from retail to security, from vacation resorts to cable companies, from competitive sports teams to space explorers.

In this white paper, we will discuss:

- How Apache Spark is broadening access to machine learning
- Various machine learning use cases and applicable scenarios
- Why Apache Spark is the ideal platform for machine learning

Apache Spark is:

- **Spark Core**
- **Spark Streaming**
  - Microbatch
  - Stateful stream processing
  - DStream
  - Socket Stream
  - File Stream
- **Spark SQL**
  - DataFrame
  - DataFrame API
  - Supported Data Formats and Sources
  - Plan Optimization and Execution
  - Rules-Based Optimization
- **MLlib**
  - Algorithms
  - Key Features
  - Pipeline
- **GraphX**
  - PropertyGraph
  - GraphViews
  - TripletView
  - Subgraph
  - Distributed Graph Representation



Broadening access to machine learning

In the past, data science professionals were intimidated by the words "algorithms" and "machine learning statistics." For many years these capabilities were delegated to people with PhDs, but Apache Spark is changing the perception of these key functions and democratizing computer learning at the same time.

The main difference between machine learning professionals and those in statistics is approach. A statistician conceivably could spend months or even years developing and testing a model, bringing its reliability and accuracy well into the 99th percentile. A machine learning professional is part of a new wave of modelers who prefer an iterative approach. The faster those iterations, the faster these modelers get their results—at far less cost and time-to-production—and Spark is ideal for that.

Spark has a distinct advantage in that it is a one-stop-shop: ETL, feature extraction, feature engineering, train algorithms, scoring, and decisioning are all contained in one platform. This empowers data science professionals to make recommendations, forecast, predict, and detect anomaly outliers—all while doing it in-memory. In combination with the libraries and APIs designed to enable a wide variety of big data and analytics uses cases, the addition of Spark MLlib opens the door for data professionals to implement machine learning algorithms into their data products.

No longer solely in the data scientist's domain, machine learning is now open and accessible to developers, data engineers, business analysts, and even business stakeholders. Using Spark, developers can build and deploy models quickly, data engineers can accelerate their respective pipelines, and business analysts can have direct access to next-gen analytics.

Spark is moving from what was the persistent and methodical realm of data science to the "quick and dirty" world of rapid iteration and, in the process, is speeding innovation and widening the practical applications of machine learning. Spark is able to do this because it provides a unified engine that all members of a data science team can work on, as well as being compatible with a variety of programming languages that breaks down accessibility and usability barriers. Spark is the breakthrough analytics operating system bridging the gap between the data scientist and the rest of the members of a data science team.

But at the same time Spark is democratizing machine learning, it also is enhancing and accelerating those already knowledgeable and experienced in the field. Spark makes testing hypotheses and running analyses iteratively easy, thanks to it's built-in machine learning library and algorithms, but it also enables scalability so that these models can be used as is or customized to fit any number of scenarios.

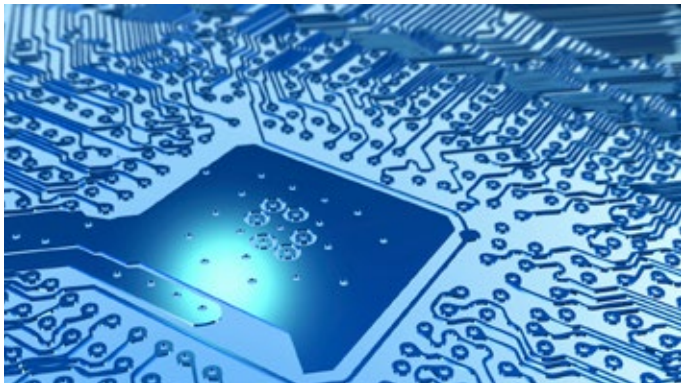
Machine learning in action

Crowdflower ran a recent survey with data scientists from a broad range of backgrounds, including those who were new to the field and those who were at a Chief Data Officer level. The survey revealed 50 percent of respondents noted machine learning had significant importance for their companies and their departments[1]. Machine learning is being used across a wide variety of industries and enterprises to drive competitive advantage. Machine learning has proven successful in cyber security and online retail. Of the former, machine learning is a strong technology for finding the data anomalies and predictions for outliers that root out fraud. Of the latter, machine learning is at the heart of recommendation engines. Many companies are finding that Spark is becoming a go-to platform to create and test these machine learning models quickly and efficiently to improve results.

Improving cyber security and fraud prevention

Spark's large data capabilities can make a key difference in the verification and approval of credit card transactions. For example, let's say a customer who lives in New York swipes his or her credit card at a coffee shop during a short, previously unexpected business trip to San Francisco. The credit card company can automatically flag the transaction and launch a workflow that—based on scoring logic that includes zip code, age, income, and other factors—will reject it.

While these models flag unusual behavior, they are not as up-to-date or detailed as they could be. Credit card companies obviously rely on their own data—customer history and status, as well as the transaction data that lists location and amount—to generate a score to determine if there is potential fraud.



But imagine that transaction within a broader data context: for example, by expanding to incorporate social media datasets, credit card companies can increase the confidence score that a transaction is legitimate. If the customer had tweeted from San Francisco earlier in the day and that social media data was available to the credit card company, it could change the score and approve the transaction.

Spark provides a platform to test these models, enabling the ability to draw from large, even unstructured, datasets to increase resolution and reliability that, once ported into a production environment, will make these decisions even better. Models can be built side-by-side, one using in-house data and one expanding to include social media data to compare the reliability of results.

Providing more accurate recommendations for online retail
The better the recommendation engine, the greater the sales and the higher the margin. Today top retailers build recommendations and test them on Spark. For example, online clothing retailers may ETL their customer data and look for specific features (feature identification). A sample feature may be a customer who has purchased ten shirts of a certain style over time. By feeding specific features into Spark, data science professionals can run them through a model to train their recommendation engine to produce a score that predicts the likelihood of a certain behavior. For example, the next time the customer makes a transaction, it will produce a score against the model to predict the likelihood that the customer will make another transaction of the same type.

Traditionally, the systems that created and operationalized these models have been separate, causing frustration for data science professionals who want to see their models in action. Spark allows you to create the model and test its scoring effectiveness all on the same platform. From a machine learning standpoint, this is a game-changer.

Spark's limitless applications and use cases
Organizations are applying machine learning to an even broader range of industries and endeavors, including:

- **Performance intelligence**. The U.S.A. Cycling Women's team employed cloud, mobile, and analytic technologies to increase performance in Team Pursuit, a four-kilometer cycling event.
- **Service optimization.** High-demand public Wi-Fi provider, SolutionInc, analyzed its massive Wi-Fi data log covering a two-year period using Spark to generate deeper and more precise business insights.
- **Data mining.** Researchers at the SETI (Search for Extraterrestrial Intelligence) Institute in Mountain View, CA, analyzed signal data from the Allen Telescope Array using limited algorithms to detect real-time signal patterns.

Why use Spark for machine learning?

Easily shift from development to production
In a production environment, Spark can be employed to test models on the fly instead of using pre-existing data feeds. Machine learning can be used to its full potential when the platform it's using enables seamless shifts from development to production environments.

An example of Spark running in a production environment is our previous e-commerce recommendation engine scenario. Online retailers want to provide a hyper-contextualized experience for their best customers. They continually optimize their recommendation engines to offer merchandise at a price point that will be the most attractive to the consumer, while at the same time generate the most profit for the retailer.

These recommendation engines rely on a variety of tactics, such as offering additional products based on past searches and purchases, suggesting alternate merchandise ("Your search found this, but were you looking for this?"), or bundling items ("You can buy this item plus another for a discounted price").

With Spark, a retailer can deploy multiple production models to test them in a live environment to perform path analysis. Developers can also swap out models in real time to speed hyper-contextualization engine development so that it doesn't take months to build and test a model. Instead, the models can be tested in real-time one after the other, in as quickly as 10 minutes, followed by another test in another 10 minutes.

A one-stop-shop for machine learning and beyond
Spark isn't solely a platform for machine learning—it includes a variety of other libraries and APIs that enable data science professionals to use Spark for a growing number of data and analytics use cases. With its growing functionality, data scientists are able to uncover new patterns and previously hidden insights. Some of the top Spark use cases include interactive querying of very large data sets, running large data

processing batch jobs, performing complex analytics and data mining across various types of data, building and deploying rich analytics models, and implementing near-real time stream event processing.

It's also built to work with a variety of data sets in different formats so that the data product you're building isn't limited by where the data lives and what form it's in. While data scientists have found increasing success building and testing models using, for example, 30 years of transaction data, with Spark they can quickly build and test other models using diverse data sources to increase reliability and confidence in scoring. Furthermore, this can all be done dynamically.

Delivering innovation through spark

The demand for machine learning keeps growing. As businesses are shifting to stay relevant in the cognitive era, machine learning will both support and drive today's data scientists and advanced analytics leaders into the future.

IBM has made significant investments in Apache Spark because of it's potential to deliver business innovation. Last year IBM announced it's mission to train one million data scientists through events, meetups, online courses and Spark Technology Center. IBM is also one of the largest contributors to Spark Core, and has a variety of partnerships with companies that are part of the Spark ecosystem aimed at showing enterprises how to get business value from Spark. Finally, IBM has infused its portfolio with Spark, meaning they have reengineered numerous products to be built on top of Spark's analytics processing engine.

IBM believes Spark is the analytics operating system for the modern enterprise.

Additional Resources:

Visit http://ibm.biz/machinelearning to learn more about Machine Learning and Apache Spark

Get started with a Machine Learning Library in Apache Spark

IBM Corporation
IBM Analytics
Route 100
Somers, NY 10589

Produced in the United States of America
September 2016

1 http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Please Recycle

CDW12360-USEN-00