What is Apache Spark?

Over the past two years, about 90% of the world’s data has been generated from mobile apps and devices, IoT devices, search queries, and social data. Many companies have recognized this explosive growth of data as opportunities — however, these opportunities require additional capabilities and resources that they may not be prepared for like big data storage systems and sophisticated analytics tooling. Add in higher demands to conduct rapid analytics in an efficient manner, and the accelerated drive towards adopting flexible, fast, and increasingly powerful computing systems makes these opportunities all that much more difficult to fulfill as the suffusion of data continues.

Apache Spark is a “lightning-fast cluster computing” solution for big data analytics. Originally developed by students at UC Berkeley, Spark has become one of the largest and most active open source communities in big data processing today. It offers the competitive advantage of processing high volumes of data at higher velocities compared to other big data and MapReduce technologies. Spark’s in-memory data crunching also allows for stronger and faster performance all while remaining well suited for machine learning algorithms.

To learn more about Apache Spark visit: ibm.com/jstart/spark

Spark’s Benefits

- **Speed**: Allows programs to run up to 100x faster than Hadoop MapReduce by manipulating in memory computing and can also run 10x faster when stored on disk.
- **Easy to Use**: Offers data scientists a variety of popular development languages like Java, Python, or Scala allowing for faster results and easy adoption.
- **Unified Platform**: Packaged with high level libraries, Spark combines SQL queries, streaming data, and complex analytics all in one application.
- **Compatible**: Spark can perform in a variety of environments, including Hadoop, Mesos, standalone, or even in the cloud.

How is IBM using Spark?

Since its release, Spark has been rapidly gaining momentum among the leading big data companies, including IBM. With significant advancements in speed that reduce the time-to-value to customers, Spark’s strong and flexible framework has helped to spike the growth in interest and support for the Spark community.

IBM has committed several notable resources to support Spark including the Spark Technology Center, enabling a Spark-As-A-Service model on its Cloud Platform, Bluemix, as well as dedicating thousands of IBM employees to building solutions and engaging in the Spark community. This interest in Spark was influenced not only by IBM executives, but also by customer and market demand for Spark’s capabilities, as well as positive feedback from the data science community.

What’s Next for Spark?

As the Spark community continues to grow, IBM will remain committed to leveraging, contributing and participating in its rapid analytics capabilities to further help solve real business needs. But where does the technology go from here?

- **Extending Spark to the Cloud as a Service**: This offering allows cloud app developers to leverage Spark’s analytic capabilities.
- **Portable Analytics**: The ability to move analytics capabilities and models seamlessly from environment to environment rapidly.
- **Evolution into an Integrated Platform**: The ability to tie together data from disparate data sources from internal and external stores while using a variety of endpoints/methods of accessing data.
- **Development of Next Gen Spark Systems**: Further development will allow customers to avoid having to build and maintain complex back end systems.

For more information about IBM’s integration with Apache Spark, contact:

info@ibmjstart.com