Recap: Spark+AI Summit 2018

Colby Ford

From June 4th - 6th, 2018, Mike Cornell, Jared Zagelbaum, Matt Mace, and I attended the Spark+AI Summit 2018. This event offered a day of training sessions on June 4th and then two days of conference activities on the 5th and 6th. Hosted in beautiful San Francisco, the conference was packed with technical deep dives and demos from experts around the country. Couldn't attend? Don't worry, I've got you covered!

About Databricks

If you frequent the BlueGranite blog, you're probably very familiar with Azure Databricks, the hottest new data and analytics platform from Microsoft. However, did you know that Databricks is actually a company founded by some of the original creators of Spark? Databricks' founders started the Spark research project at UC Berkeley, which later became Apache Spark™. They've been working for the past 10 years on cutting-edge systems to extract value from Big Data.

databricks

In addition to making proprietary enhancements to the Spark system to create their own Unified Analytics Platform, the company has a commitment to open-source development and effective training. Databricks is also the host and main sponsor of the Spark+AI Summit.

What We Learned

Mike, Jared, and I attended a variety of sessions, from technical deep dives to Python demonstrations to talks on deep learning techniques. Personally, I enjoyed the sessions around research, which described the work that's being done (in and out of academia) in the world of Spark and distributed computing. I loved seeing what everyone else has accomplished. It certainly gave us some cool project ideas.

All of the sessions were recorded and are available here.

Each day began with a series of keynotes who provided great case studies and use cases from companies such as Regeneron and Bechtel, as well as universities and research institutions. These keynotes provided great demonstrations of state-of-the-art, game-changing solutions that have been implemented in their respective industries.

BIG Announcements

Databricks Delta

Databricks Delta is a new data management tool that provides a single interface for combining the scale of a data lake, the reliability and performance of a data warehouse, and the low latency of streaming together in a single system. Using Delta along with the rest of the Databricks Unified Analytics Platform makes it much easier to build, manage, and put Big Data applications into production.

A big talking point at the keynote for this technology was around data quality. Delta helps to make sure the data flowing through your workflows remains consistent. Plus, Delta will give you insights around changes in your data. In addition to this, Delta helps to capture and track some summary metrics around your data, as well as relevant metadata.

Read more about Databricks Delta from the blog, here.

MLflow

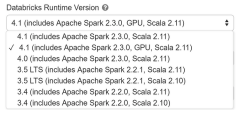
As more and more deep learning or machine learning frameworks are created, the use of state-of-the-art algorithms, which are extremely desirable, becomes difficult from an integration/compatibility front. Many data science teams struggle with using these myriad tools together, tracking experiments, reproducing results, and deploying models. Databricks has announced a new open-source solution to help combat these issues.

MLflow is designed to work with any ML library, algorithm, deployment tool or language. It's built around REST APIs and simple data formats (e.g., a model can be viewed as a lambda function) that can be used from a variety of tools, instead of only providing a small set of built-in functionality.

Check out MLflow.org to get started.

Databricks Runtime for Machine Learning

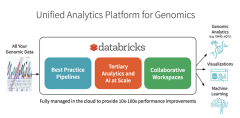
Databricks also announced a new runtime for machine learning. This includes ready-to-use machine learning frameworks, simplified distributed training, and GPU support. For many deep learning toolkits (Microsoft Cognitive Toolkit, TensorFlow, XGBoost, etc.) or machine learning packages (scikit-learn, etc.), the environment configuration can be a tricky process to getting these libraries working optimally. With the new Databricks Runtime for Machine Learning, you will now have access to pre-configured machine learning frameworks that have been optimized for Spark.



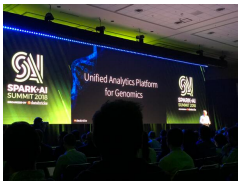
In addition to the frameworks, the folks at Databricks have simplified the task of training models in a distributed fashion using the Horovod framework. This framework facilitates distributed training across multiple GPUs. With Databricks, you will now be able to unleash the power of GPUs in Spark!

Unified Analytics Platform for Genomics

For those of you who work in the fields of healthcare or life sciences, you already know the pains of dealing with huge files with crazy file types and using command line-based tools like SnpEff, BWA, or GATK4. Traditionally, these tools only work on MPI-based computing clusters. Now, resulting from a dedicated effort from Databricks, they have ported over many of the common bioinformatics pipeline functions into Spark. Plus, since Databricks is cloud-based, genomics data pipelines and analyses are now fully scalable.



Stay tuned for an upcoming blog where I'll take an in-depth look at the new genomics-based features coming to Databricks.



All the Swag...

As with any conference, the Expos are quite interesting. Some booths were giving away fidget spinners, shirts, or stuffed animals. Others were giving away hot sauce (yes, seriously) or raffling off drones and books. In my opinion the best swag is, of course, stickers!

If you have any questions or want to learn more about BlueGranite's AI capabilities, contact us, or check out BlueGranite's free collection of Azure Databricks resources here.