Simplify Machine Learning on Spark with Databricks

by Denny Lee (https://databricks.com/blog/author/denny)
Posted in **COMPANY BLOG (HTTPS://DATABRICKS.COM/BLOG/CATEGORY/COMPANY)** | June 4, 2015
(https://databricks.com/blog/2015/06/04/simplify-machine-learning-on-spark-with-databricks.html)

Join us at Spark Summit (https://spark-summit.org/2015/) to hear more about new functionalities of Apache Spark. Use the code *Databricks20* to receive a 20% discount!

As many data scientists and engineers can attest, the majority of the time is spent not on the models themselves but on the supporting infrastructure. Key issues include on the ability to easily visualize, share, deploy, and schedule jobs. More disconcerting is the need for data engineers to re-implement the models developed by data scientists for production. With Databricks, data scientists and engineers can simplify these logistical issues and spend more of their time focusing on their data problems.

Simplify Visualization

An important perspective for data scientists and engineers is the ability to quickly visualize the data and the model that is generated. For example, a common issue when working with linear regression is to determine the

model's goodness of fit. While statistical evaluations such as Mean Squared Error are fundamental, the ability to view the data scatterplot in relation to the regression model is just as important.

Training the models

Using a dataset comparing the population (x) with label data of median housing prices (y), we can build a linear regression model using Spark MLlib's Linear Regression with Stochastic Gradient Descent (LinearRegressionWithSGD). Spark MLlib is a core component of Apache Spark that allows data scientists and data engineers to quickly experiment and build data models – and bring them to production. Because we are experimenting with SGD, we will need to try out different iterations and learning rates (i.e. alpha or step size).

An easy way to start experimenting with these models is to create a Databricks notebook in your language of choice (python, scala, Spark SQL) and provide contextual information via markdown text. The screenshot below is two cells from an example DBC notebook where the top cell contains markdown comments while the bottom cell contains pyspark code to train two models.

(https://databricks.com/wp-content/uploads/2015/06/Figure-1a.png)

Figure 1: *Screenshot of Databricks Notebook training two models with Linear Regression with SGD*

Evaluating the models

Once the models are trained, with some additional pyspark code, you can quickly calculate the mean squared error of these two models:

```
valuesAndPreds = parsedData.map(lambda p: (p.label, model.predict(p.featu
MSE = valuesAndPreds.(lambda (v, p): (v - p)**2).mean()
print("Mean Squared Error = " + str(MSE))
```

The definition of the models and MSE results are in the table below.

	# of iterations	Step Size	MSE
Model A	100	0.01	1.25095190484
Model B	1500	0.1	0.205298649734

While the evaluation of statistics most likely indicates that Model B has a better goodness of fit, the ability to visually inspect the data will make it easier to validate these results.

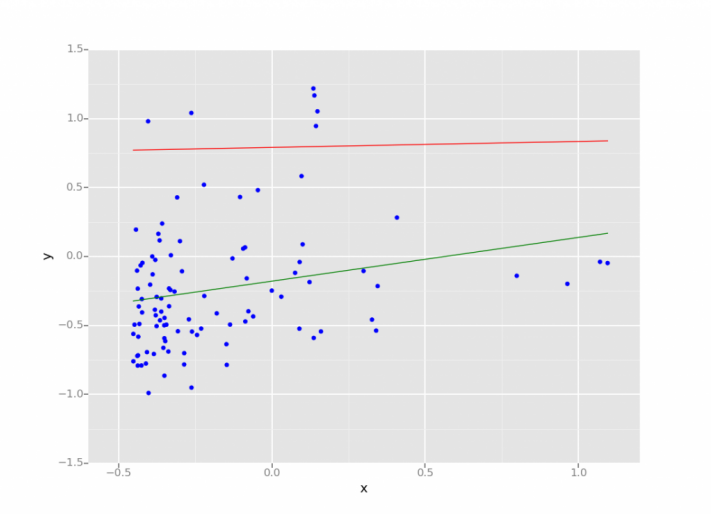
Visualizing the models

With Databricks, there are numerous visualization options that you can use with your Databricks notebooks. In addition to the default visualizations automatically available when working with Spark DataFrames, you can also use matplotlib, ggplot, and d3.js – all embedded with the same notebook.

In our example, we are using ggplot (the python code is below) so we can not only provide a scatter plot of the original dataset (in blue), but also graph line plots of the two models where Model A is in red and Model B is in green.

```
p = ggplot(pydf, aes('x','y')) + \
    geom_point(color='blue') + \
    geom_line(pydf, aes('x','y2'), color='red') + \
    geom_line(pydf, aes('x','y3'), color='green')
display(p)
```

Embedded within the same notebook is the median housing prices ggplot scatterplot figure where the x-axis is the normalized population and y-axis is the normalized median housing price; Model A is in red while Model B is in green.



(https://databricks.com/wp-content/uploads/2015/06/Figure-2a.png)

Figure 2: *Screenshot of a ggplot scatterplot embedded within a Databricks notebook*

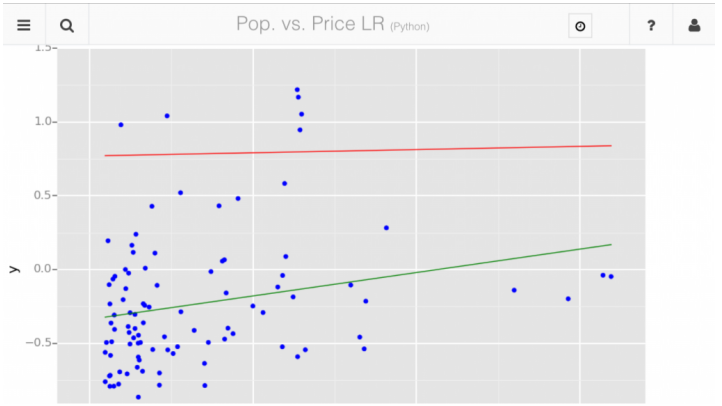
As you can see from the above figure, the green line (Model B) has a better goodness of fit compared to the red line (Model A). While the evaluation statistics pointed toward this direction, the ability to quickly visualize the data and the models within the same notebook allows the data scientist to spend more time understanding and optimizing their models.

Simplify Sharing

Another crucial aspect of data sciences is the collaborative effort needed to solve data problems. With many developers, engineers, and data scientists often working in different time zones, sched important to have an environment that is de

Portability

With Databricks, you can make it easier to collaborate with your team. You can share your Databricks notebooks by sharing its URL so that any web browser on any device can view your notebooks.



(https://databricks.com/wp-content/uploads/2015/06/IMG_1596.png)

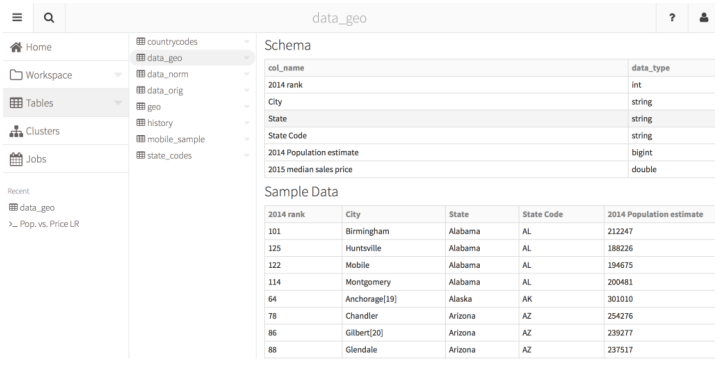
***Figure 3**: Databricks notebook view of a the same linear regression SGD model via matplotlib on an iPhone 6.*

Non-proprietary

While these notebooks are optimized for Databricks, you can export these notebooks to python, scala, and SQL files so you can use them in your own environments. A common use-case for this approach is that data scientists and engineers will collaborate and experiment in Databricks and then apply their resulting code into their on-premises environment.

Share Definitions

As a data scientist or data engineer working with many different datasets, keeping up with all of the changes in schema and locations itself can be a full time job. To help keep this under control, D table definitions. Instead of searching for in schema, go the tables tab within Databricks and you can define all of your tables in one place. This way as a data engineer updates the schema or source location for these table, these changes are immediately available to all notebooks.

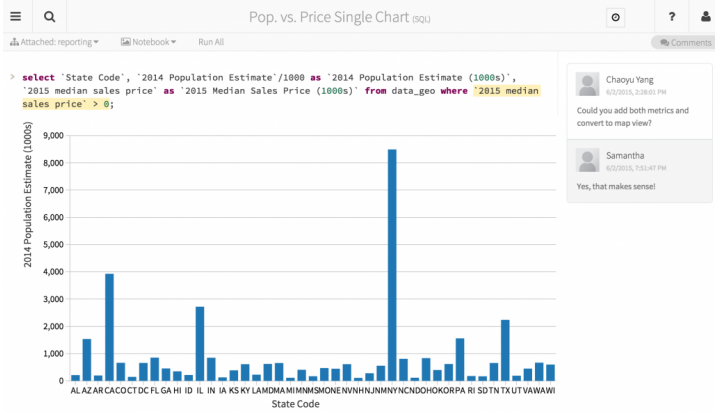


(https://databricks.com/wp-content/uploads/2015/06/Figure-41.png)

Figure 4: *View of table definitions (schema and sample data) all from one place.*

Collaborate

As notebooks are being created and shared, users can comment on the code or figures so they can provide input to the notebooks without making any changes to them. This way you can lock the notebooks to prevent accidental changes and still accept feedback.



(https://databricks.com/wp-content/uploads/2015/06/Figure-5.png)

Figure 5: *Users commenting on a Databricks notebook to more easily facilitate feedback*

Simplify Deployment

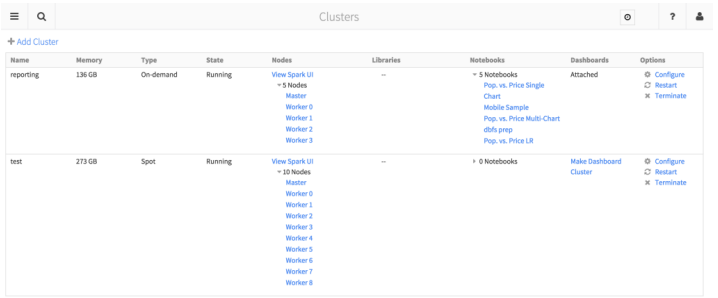
One of the key advantages of Databricks is that the model developed by data scientists can be run in production. This is a huge advantage as it reduces the development cycle and tremendously simplifies the maintenance. In contrast, today data scientists develop the model using single machine tools such as R or Python and then have data engineers re-implement the model for production.

Simplify Infrastructure

As a data engineer, there are many steps and configurations to deploy Apache Spark in production. Some examples include (but are not limited to):

- Configuring High Availability and Disaster Recovery for your Spark clusters
- Building the necessary manifests to spin up and down clusters
- Configuring Spark to utilize local SSDs for fast retrieval
- Upgrading or patching your Spark clusters to the latest version of the OS or Apache Spark

With Databricks, the management of your Spark clusters are taken care by dedicated Databricks engineers who are supported by the developers and committers of the Apache Spark open source project. These clusters are configured for optimal performance and balance the issues surrounding resource scheduling, caching, and garbage collection.
Once deployed, you can quickly view what clusters are available and their current state including the libraries and notebooks that are attached to the cluster(s). Concerns around high availability, disaster recovery, manifests to build and deploy clusters, service management, configurations, patching, and upgrades are all managed on your behalf using your own (or your company's) AWS account.

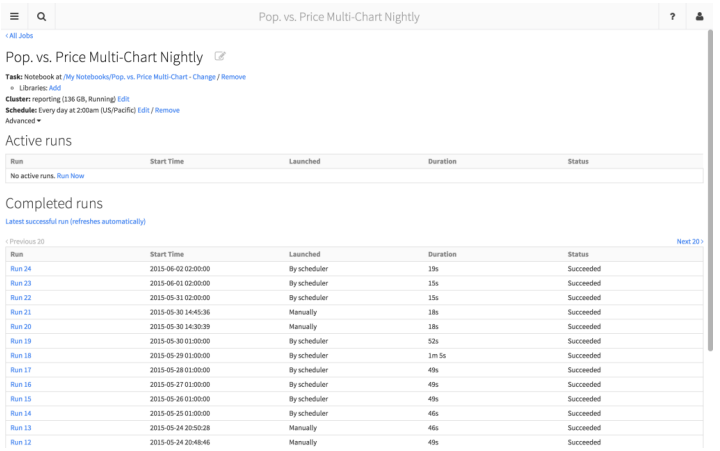


(https://databricks.com/wp-content/uploads/2015/06/Figure-6.png)

Figure 6: *Databricks Cluster view for easier management of your Databricks infrastructure*

Simplify Job Scheduling

Traditionally, transitioning from code development to production is a complicated task. It typically involves separate personnel and processes to build the code and push it into production. But Databricks has a powerful Jobs feature for running applications in production. You can take the notebook you had just created and run it as a periodic job – scheduling it minute, hourly, daily, weekly, or monthly intervals. It also has a smart cluster allocation feature that allows you to run your notebook on an existing cluster or on an on-demand cluster. You can also receive email notifications for your job as well as configure retries and timeouts.



(https://databricks.com/wp-content/uploads/2015/06/Figure-7.png)

Figure 7: View of the Population vs. Price Multi-Chart Notebook Nightly Job

As well, you can upload and execute any Spark JAR compiled against any Spark installation within the Jobs feature. Therefore any previous work can be used immediately instead of recreating and rebuilding the code-base.

Try out Databricks

We created Databricks to make it easier for data scientists and data engineers to focus on experimenting and training their models, quickly deploy and schedule jobs against those models, easily collaborate and share their learnings, and easily share the schema and definitions for their datasets. Let us manage the cluster, configure it for optimal performance, perform upgrades and patches, and ensure high availability and disaster recovery.

Machine Learning with Spark MLlib is a lot more fun when you get to spend most of your time doing Machine Learning!

READ MORE

Apache SparkR On-Demand Webinar and FAQ

Two months ago we held a live webinar – Enabling Exploratory Analysis of Large Data with Apache Spark and R – to demonstrate one of...

(https://databricks.com/blog/2016/07/12/sparkr-on-demand-webinar-and-faq.html)

New eBook Released: Lessons for Large-Scale Machine Learning Deployments on Apache Spark

We are excited to announce that the third eBook in our technical blog book series, Lessons for Large-Scale Machine Learning Deployments on Apache Spark, has...

(https://databricks.com/blog/2016/07/06/new-ebook-released-lessons-for-large-scale-machine-learning-deployments-on-apache-spark.html)

Introducing Getting Started with Apache Spark on Databricks

We are proud to introduce the Getting Started with Apache Spark on Databricks Guide. This step-by-step guide illustrates how to leverage the Databricks' platform to...

(https://databricks.com/blog/2016/06/30/introducing-getting-started-with-apache-spark-on-databricks.html)

SEE ALL COMPANY BLOG POSTS (HTTPS://DATABRICKS.COM/BLOG/CATEGORY/COMPANY)

(https://twitter.com/databricks)

(https://www.linkedin.com/company/databricks)

(https://www.facebook.com/pages/Databricks/560203607

(https://databricks.com/feed)