Databricks for Data Science

Accelerate Innovation through Unification

Businesses are generating data at a faster pace than ever: 90% of the world's data was generated within the last two years. The increased data volume is rapidly outpacing our ability to consume it. Data science allows businesses to efficiently predict future outcomes, and even preemptively take action, based on insights from terabytes of business data.

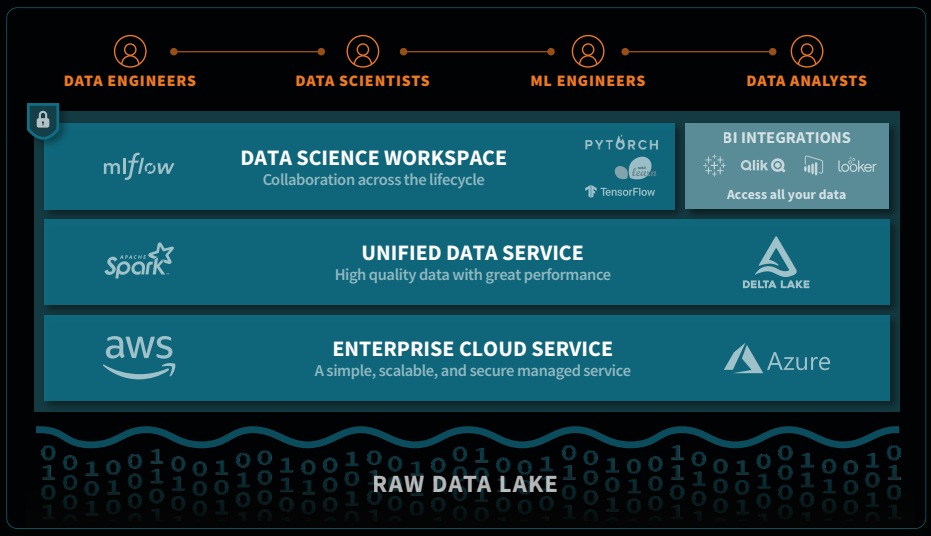
However, as the data continues to grow in volume, new challenges arise that can impede time-to-insight and innovation:

- Spending too much time maintaining infrastructure rather than the data.
- Complexity and cost to train machine learning models at scale.
- Poor collaboration among team members and across the organization.

By combining big data with data science techniques such as machine learning and deep learning, businesses can build and train scalable models that drive new and extraordinary business use cases.

Better Data Science with Databricks

Founded by the team who created Apache Spark,™ Delta Lake, and MLflow, Databricks provides a Unified Data Analytics Platform that accelerates innovation by unifying data science, engineering, and business. With Databricks, data scientists can perform all analytics in one place, from exploratory data science to building state-of-the-art machine learning models as a team.



Automated Infrastructure

Databricks' serverless and highly elastic cloud service is designed to remove operational complexity while ensuring reliability and cost efficiency at scale, so you can focus on the data science instead of DevOps. Through the first serverless API for Apache Spark, organizations can remove the barriers of infrastructure for both end-users and DevOps.

Key benefits of Databricks' serverless infrastructure are:

AUTO-CONFIGURATION

The Spark version deployed in serverless pools is automatically optimized for interactive SQL and Python workloads.

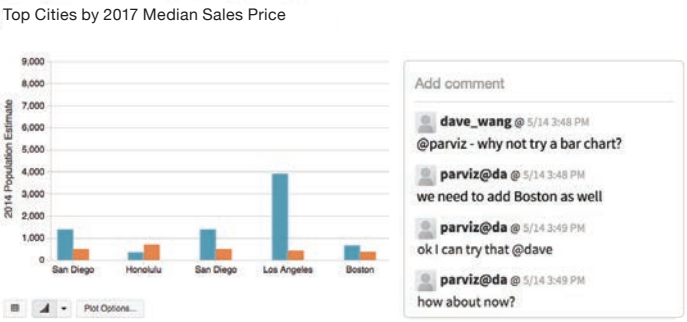
ON-DEMAND ELASTICITY

Databricks automatically scales the compute and local storage resources in the serverless pools in response to Apache Spark's changing resource requirements for user jobs.

OPTIMIZED TENSORFLOW

Benefit from TensorFlow CUDA-optimized version on GPU clusters, and Intel MKL-DNN optimized TensorFlow package on Intel CPUs for maximum performance.

Databricks for Data Science



Collaborative Notebooks

Databricks provides collaborative notebooks that eliminates the need to integrate third party tools and libraries. Support for multiple programming languages (R, Python, Scala, and SQL) ensures you use the right tool for the job. Improve team productivity by enabling team members to collaborate on the data and models in real time, while tracking usage through viewer logs and revision history.

Databricks' interactive workspace allows data science teams to leverage the following capabilities:

COLLABORATION

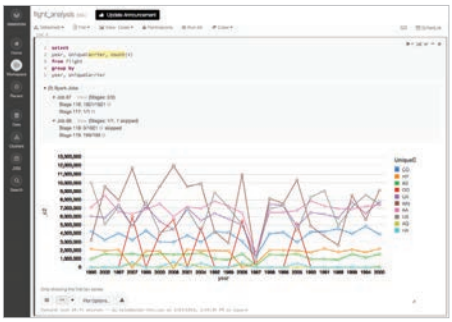
Built-in collaboration features to increase productivity across the entire data science team.

POINT-AND-CLICK VISUALIZATIONS

Data visualizations allow you to easily create and embed a wide range of point-and-click visualizations into your notebook or use powerful scriptable options like matplotlib, ggplot, and D3.

VERSION CONTROL

Revision history and Github integration for version control which is extremely useful when building notebooks for production and ad-hoc querying.



Interactive Dashboards

Turn your analysis from a notebook into a dynamic dashboard with one click. Databricks Dashboards allow you to easily share insights with your colleagues and customers, or let them run interactive queries with Spark-powered dashboards.

Key capabilities built into the Databricks Dashboards include:

ONE CLICK PUBLISHING

Create shareable dashboards from notebooks with a single click. One notebook can be tailored into multiple dashboard views.

CONTINUOUS UPDATES

Publish dashboards and schedule the content to be updated Continuously.

PARAMETERIZED DASHBOARDS

Enable non-technical users to perform scenario analysis directly from published dashboards.

> *Databricks is really an **important part of our AI strategy**. As we continue to digitize our business, and explore opportunities and challenges through AI and machine learning, we will continue to activate Databricks **throughout the rest of the business**.*
>
> — Mainak Mazumdar, Chief Research Officer nielsen

databricks

Databricks for Data Science

Machine Learning Environment

One-click access to preconfigured ML clusters powered by a scalable and reliable distribution of the most popular ML frameworks, built-in autoML capabilities, and optimizations for unmatched performance at scale.

Key capabilities built into the Databricks ML Runtime include:

FRAMEWORKS OF CHOICE

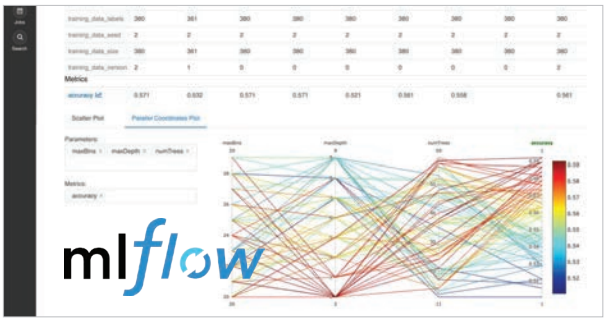
Built-in TensorFlow, Keras, PyTorch, MLflow, Horovod, GraphFrames, scikit-learn, XGboost, numpy, MLeap, Pandas, and more.

AUGMENTED MACHINE LEARNING

Accelerate machine learning from featurization to inference, including hyperparameter tuning and model search with Hyperopt.

SIMPLIFIED SCALING

Go from small to big data effortlessly with an auto managed clusters infrastructure and simplified distributed training on Horovod.



ML Lifecycle Management

MLflow is an open-source platform from Databricks for managing the complete lifecycle. With MLflow, data scientists can track and share experiments locally or in the cloud, package and share models across frameworks, and deploy models virtually anywhere.

Key capabilities built into MLflow include:

EXPERIMENTS TRACKING

Automatically track, share, compare, and interactively visualize experiments with MLflow from within your notebooks.

MODEL MANAGEMENT

One place to share ML models, collaborate on moving them from experimentation to online testing and production, integrate with approval and governance workflows, and monitor ML deployments and their performance.

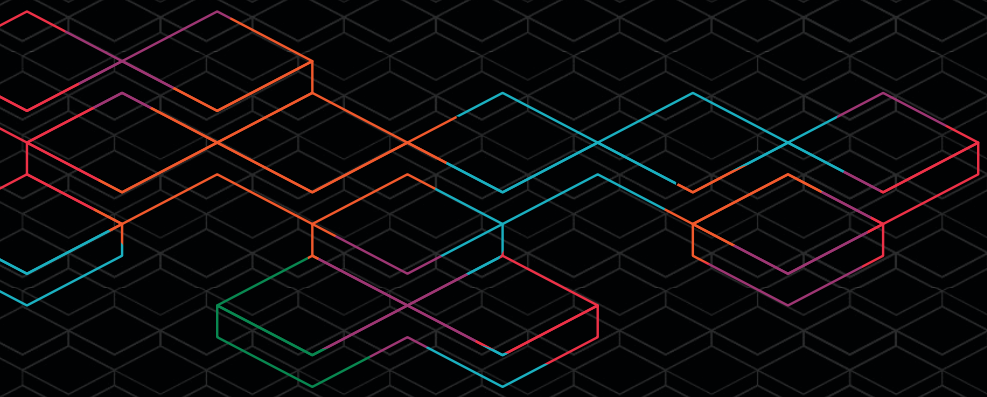
FLEXIBLE DEPLOYMENT

Quickly deploy production models for batch inference on Apache Spark™ or as REST APIs using built-in integration with Docker containers, Azure ML, or Amazon SageMaker.

TensorFlow™ Keras PyTorch HOROVOD scikit learn XGBoost

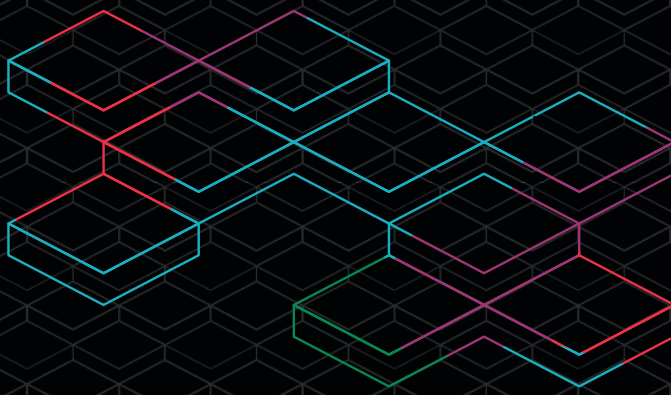
Get started with Databricks for data science today with a free trial

databricks



databricks®

Get started with Databricks for data science today with a free trial



Machine Learning and AI for the Masses

Databricks' Unified Data Analytics Platform takes the complexity out of data science at scale, allowing data scientists of all backgrounds and levels of experience to tap into the power of advanced analytic techniques such as machine learning and deep learning.

BRANDLESS™ COMCAST devon Hotels.com hp McGraw Hill Education nielsen

Quby Riot GAMES RUE GILT GROUPE SHOPRUNNER SHOWTIME viacom