Databricks Primer

databricks

Who is Databricks?

Databricks was founded by the team who created Apache® Spark™, the most active open source project in the big data ecosystem today, and is the largest contributor to the project. Our mission at Databricks is to empower individuals and organizations to swiftly build and deploy advanced analytics solutions. We do this through our product, a virtual analytics platform called Databricks.

Today, hundreds of organizations around the world use Databricks to build and power their production Spark applications.

To learn more about Apache Spark, read the Primer.

The Challenges We Solve

Data is spread out across the organization in disparate silos, and the use cases to create value from data are becoming more sophisticated. As the volume and complexity of data grow, the problem is only worsening — creating the need to deliver insights faster. Moreover, the ability of teams to prototype and operationalize data-driven solutions is also hindered by fragmented systems and tools, each with limited capabilities, as well as the inability to easily leverage more data science to make smarter decisions.

As a result, data professionals face many serious challenges in bridging the gap between raw data and solutions that create value for the business, including:

Providing easy and fast access to data at scale.

- Processing both structured and unstructured data.
- Ingesting from non-traditional data stores: AWS S3, others.
- Reducing the batch processing time.

Deploying production-quality machine learning and streaming applications.

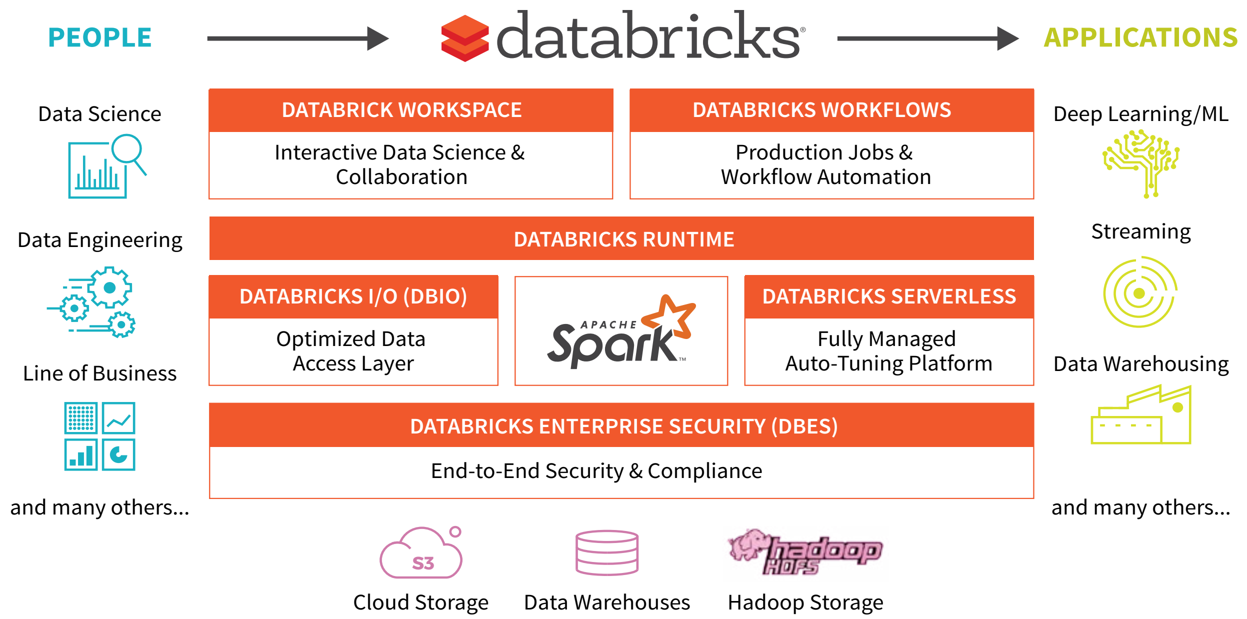
- Setting up, tuning, and scaling Apache Spark clusters for the team.
- Keeping clusters resilient and up-to-date with the latest versions.
- Scheduling, running, and debugging applications in production.

Leveraging more data science to aid decision making.

- Interactive data exploration & visualization.
- Building real-time dashboards.
- Connecting to Business Intelligence tools.

What is the Databricks Unified Analytics Platform?

Databricks, founded by the team that created Apache Spark, provides a unified analytics platform that accelerates innovation by unifying data science, engineering and business.



PEOPLE → databricks → APPLICATIONS

Data Science

Data Engineering

Line of Business

and many others...

DATABRICK WORKSPACE
Interactive Data Science & Collaboration

DATABRICKS WORKFLOWS
Production Jobs & Workflow Automation

DATABRICKS RUNTIME

DATABRICKS I/O (DBIO)
Optimized Data Access Layer

APACHE Spark

DATABRICKS SERVERLESS
Fully Managed Auto-Tuning Platform

DATABRICKS ENTERPRISE SECURITY (DBES)
End-to-End Security & Compliance

Deep Learning/ML

Streaming

Data Warehousing

and many others...

Cloud Storage

Data Warehouses

Hadoop Storage

Your Storage: By virtualizing storage, Databricks enables access to data anywhere.

- Connect directly to your data stores — no migration required.
- Separate compute from storage — scale each independently as needed.

Orchestrated Apache Spark in the Cloud: Databricks offers a highly secure and reliable production environment in the cloud, managed and supported by Spark experts.

- Powerful cluster management capabilities allow you to create new clusters in seconds, dynamically scale them up and down, and share them across teams.
- Intuitive interfaces that enable your teams to use Spark with traditional BI tools such as Tableau Software, or programmatically use the clusters via restful APIs.
- Secure data integration capabilities built on top of Spark so you can unify your data without centralization.
- Instant access to the latest Spark features as with each release.

Integrated Workspace: Through a collaborative and integrated environment, Databricks democratizes and streamlines the process of exploring data, prototyping, and operationalizing data-driven applications in Spark.

- Easy data exploration allows teams to determine what the data lets you do.
- Interactive dashboards empower teams to create dynamic reports.
- A simple and collaborative environment that enables your entire team to use Spark and interact with the data simultaneously.

Custom Spark Applications: Databricks provides a flexible job scheduler that enables a seamless transition from prototyping to production deployment without incremental work.

- Monitor progress through custom alerts for job completion and failure, and easily view historical and in-progress results.
- Enable production deployments, especially long-running applications such as streaming, to be automatically re-launched whenever a failure happens.

Databricks Enterprise Security Framework: Databricks empowers enterprises with security-enabled data democratization so that they can confidently build advanced analytics solutions when security considerations are paramount.

- *Encryption:* Provides strong encryption at-rest and in-flight with best-in-class standards such as SSL and keys stored in AWS Key Management System (KMS).
- *Integrated Identity Management:* Facilitates seamless integration with enterprise identity providers via SAML 2.0 and Active Directory.
- *Role-Based Access Control:* Enables fine-grain management access to every component of the enterprise data infrastructure, including files, clusters, code, application deployments, dashboards, and reports.
- *Data Governance:* Guarantees the ability to monitor and audit all actions taken in every aspect of the enterprise data infrastructure.
- *Compliance Standards:* Databricks has successfully completed SOC 2 Type 1 certification and can offer a HIPAA-compliant service. We also plan to achieve security compliance standards that exceed the high standards of FedRAMP as part of Databricks’ ongoing DBES strategy.

Databricks Benefits

Accelerate ETL

Make your data stores accessible to anyone in the organization and enable your teams to directly query the data through a simple-to-use interface without cumbersome ETL / ELT processes. The virtual analytics platform democratizes data access by uncoupling storage from compute and providing infinite scalability, to increase agility and better cost management. With Databricks, you can always get the resources to analyze your data by just scaling up the compute resource in a short burst.

Zero Management Apache Spark

Enable your teams to provision highly available and performance optimized Spark clusters in a self-service fashion, allowing everyone to build and deploy advanced analytics applications with no DevOps expertise. With Databricks, your team will always have access to the latest Spark features so you can leverage the latest innovation from the open source community and focus on your core mission instead of managing the infrastructure. Databricks also offers monitoring and recovery mechanisms that automatically recover clusters from failures without any manual intervention. With Databricks your infrastructure will be fast and secure without any custom work in Spark.

Agile Data Science

Databricks provides an integrated workspace that fosters collaboration through a multi-user environment that allows your team to build new machine learning and streaming applications on top of Spark. Through an interactive notebook environment, you can also create dashboards and interactive reports—allowing everyone to visualize results in real-time, train and tune machine learning models, or easily utilize any of Spark's libraries to process data. The integrated workspace helps developers and data scientists to reproduce analysis more easily, reuse more code, and simplifies the entire workflow.

Impact on Your Teams

Data Science

- Access data in all silos to form the complete picture.
- Instant access to reliable infrastructure and tools to explore data, prototype advanced analytics algorithms, and curate visual reports without building anything themselves or wait for someone else to do it for them.

Data Engineering and Application Development

- Accelerate release cycles with a more agile development and testing environment.
- Easy integration with existing tools (e.g., IntelliJ, GitHub) so the learning curve is minimum.
- Improve ETL performance with zero management Apache Spark clusters.

Architects and Administrators

- Deliver production-quality infrastructure to run advanced analytics solutions easily — higher uptime because of a managed platform built by the Spark experts.
- Satisfy the demand from their internal customers for self-service without sacrificing control and security.
- Easily integrate Spark with existing production tools (e.g., command line, Jenkins, scripts) and best practices (e.g., continuous integration).

Business Analysts

- Access data in all silos to form the complete picture with familiar BI tools.
- Run big data queries in SQL within a visual and interactive environment.
- Combine simple SQL querying with advanced programmatic approaches to perform more powerful analysis.

Why Databricks? Our Key Differentiators

Only Databricks provides a turnkey platform that leverages everything Apache Spark has to offer.

Unlike other big data providers, Databricks doesn't:

- Force you to migrate all your data into one place
- Require you to hire an army of experts to deploy and operate the platform

Databricks is the easiest and quickest way to harness the full potential of Spark:

- Access to the latest and tuned versions of Spark
- Tools to solve the most complex advanced analytics use cases
- Support from the Spark experts

In a world where Spark has become the de facto big data standard, no one better understands how to maximize the inherent power of this technology than the team who created it.

How Our Customers Use Databricks

Just-in-Time Data Warehouse

- Databricks provides a fast, simple, and scalable way to build a just-in-time data warehouse that eliminates the need to invest in costly ETL pipelines and scales on-demand, revolutionizing the way data teams analyze their data sets.
- Scale storage and compute resources independently on-demand
- Support both traditional ETL as well as directly access data to accelerate time-to-insight
- Unify a variety of data sources with powerful data sources APIs and JDBC/ODBC connectors
- Leverage advanced analytics capabilities to solve many different data problems on one platform

Machine Learning Solutions

Databricks provides an end-to-end platform designed to help data engineers and data scientists take analytics to the next level with built-in machine learning algorithms that seamlessly updates with each Spark release, interactive notebooks that support R, Python, Scala, and SQL, and automated cluster management capabilities that enable the provisioning of highly-tuned Spark clusters on-demand.

Build

- Accelerate feature data extraction at scale
- Easily support a variety of data sources and formats
- Simplify ETL and implement machine learning in a single framework

Tune

- Speed up iterative model tuning with interactive notebooks
- Interactively query large-scale data sets in R, Python, Scala, or SQL
- Visualize results with rich dashboards

Deploy

- Provision distributed clusters on-demand
- Scale storage and compute resources independently
- Ensure uninterrupted operations with seamless updates of MLlib

300+ Production Deployments Across Industries

As organizations in every industry search for faster ways to implement new and innovative use cases, the ability to perform advanced analytics to explore large data sets and extract actionable insights from this data to meet their specific needs is critical.

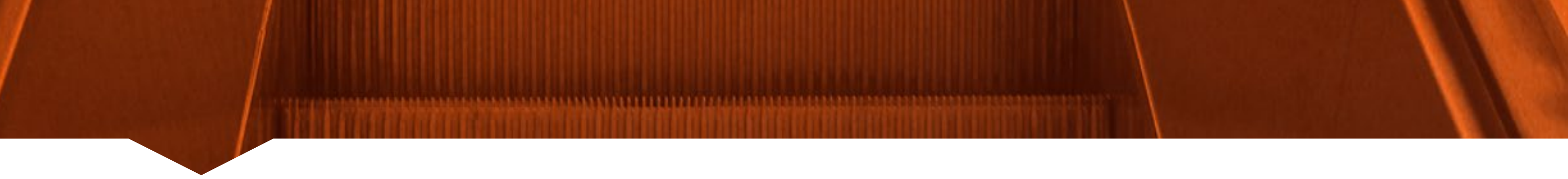
Databricks provides the performance, reliability, and ease of use to tackle advanced analytics with big data; allowing teams to focus on solving hard data problems instead of supporting infrastructure.

AIMIA
American Diabetes Association.
VIACOM
celtra

DNV·GL
edmunds.com
ELSEVIER
GoGuardian

INNERACTIVE
NBCUniversal
OpenTable
RADIUS

SAMSUNG
SAMSUNG SDS
Reltio
Two Six Capital
ZEPHYR HEALTH



How Databricks is helping organizations build data-driven applications with Apache Spark today:

SHARETHROUGH

Industry: **AdTech**
Use Case: **Just-in-Time Data Warehouse**
Learn more: **Read the Case Study**

Sharethrough uses Databricks to optimize the placement of ads based on the behavior of the visitor and to prepare customized reports for their customers by offering faster prototyping of new applications, easier debugging of complex data pipelines, and improved engineering productivity.

LendUp

Industry: **Financial Services**
Use Case: **Machine Learning**
Learn more: **Read the Case Study**

LendUp builds technology that expands access to credit. It uses Databricks to perform feature engineering and machine learning at scale, which allows them to offer credit to more people who need it and the ability to establish new products more easily.

---

myfitnesspal

Industry: **Media**
Use Case: **Machine Learning**
Learn more: **Read the Case Study**

MyFitnessPal has built one of the largest and most up-to-date food nutrition databases through crowdsourcing. They used machine learning algorithms to automatically correct inconsistencies in crowd-sourced data.

Read more case studies to learn how organizations are creating value from data with Databricks.

For more information on Spark, download the Spark Primer.

Try Databricks for free or contact us for a personalized demo.

The Best Place to Run Apache® Spark™

Try Databricks for free

databricks.com/try-databricks

Contact us for a personalized demo

databricks.com/contact-databricks

About Databricks:
Databricks' mission is to accelerate innovation for its customers by unifying Data Science, Engineering and Business. Founded by the team who created Apache Spark™, Databricks provides a Unified Analytics Platform for data science teams to collaborate with data engineering and lines of business to build data products. Users achieve faster time-to-value with Databricks by creating analytic workflows that go from ETL and interactive exploration to production. The company also makes it easier for its users to focus on their data by providing a fully managed, scalable, and secure cloud infrastructure that reduces operational complexity and total cost of ownership. Databricks, venture-backed by Andreessen Horowitz and NEA, has a global customer base that includes, Salesforce, Viacom, Amgen, Shell and HP. For more information, visit www.databricks.com