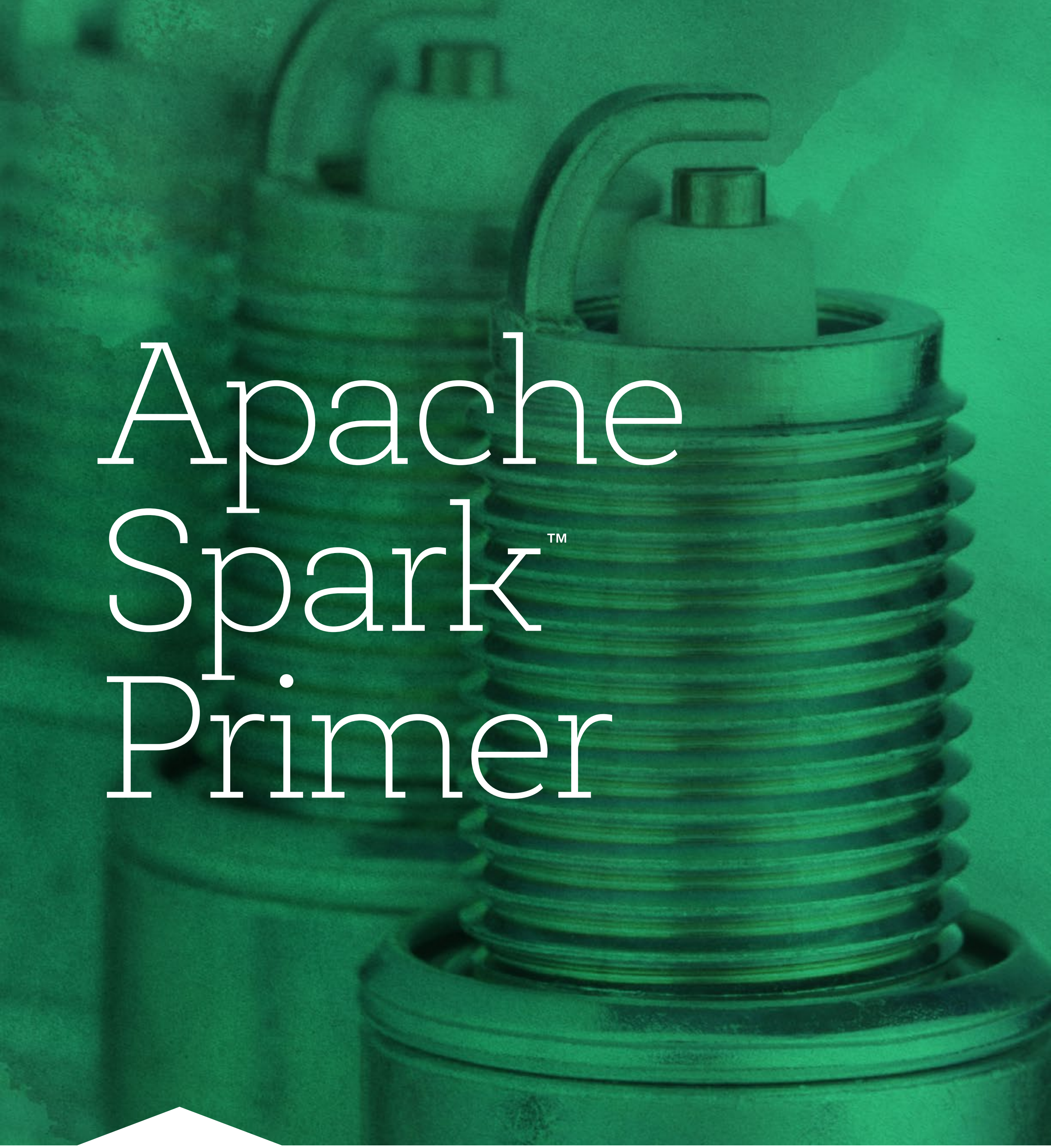


Apache Spark™ Primer

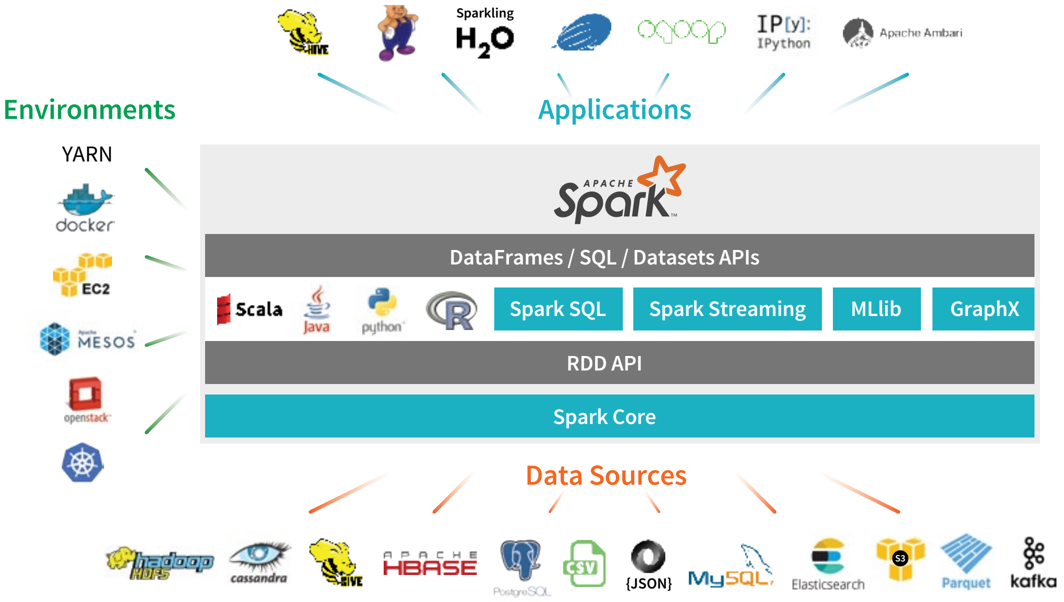
databricks™

What is Apache Spark™?

Apache Spark is an open source data processing engine built for speed, ease of use, and sophisticated analytics. Since its release, Spark has seen rapid adoption by enterprises across a wide range of industries. Internet powerhouses such as Netflix, Yahoo, Baidu, and eBay have eagerly deployed Spark at massive scale, collectively processing multiple petabytes of data on clusters of over 8,000 nodes. Meanwhile, it has become the largest open source community in big data, with over 1,000 contributors from 250+ organizations. Together with the Spark community, Databricks continues to contribute heavily to the Apache Spark project, through both development and community evangelism.

What is Apache Spark used for?

As a general purpose compute engine designed for distributed processing, Spark is used for many types of data processing. It supports ETL, interactive queries (SQL), advanced analytics (e.g. machine learning) and structured streaming over large datasets. For loading and storing data, Spark integrates with many storage systems (e.g. HDFS, Cassandra, MySQL, HBase, MongoDB, S3). Spark is also pluggable, with dozens of applications, data sources, and environments, forming an extensible open-source ecosystem. Additionally, Spark supports a variety of popular development languages including R, Java, Python and Scala.



Apache Spark

A unified engine across data sources, applications, and environments.

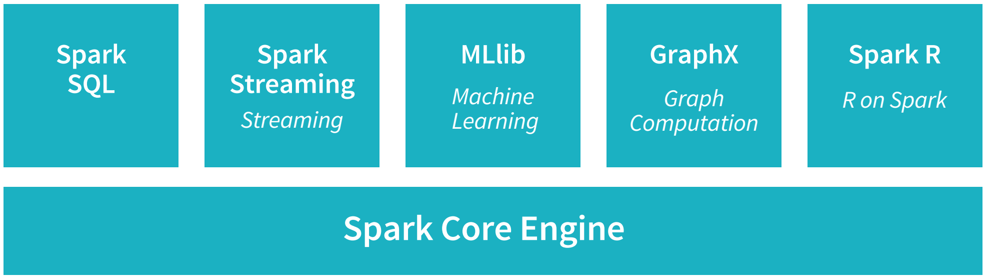
> “At Databricks, we’re working hard to make Spark easier to use and run than ever, through our efforts on both the Spark codebase and support materials around it. All of our work on Spark is open source and goes directly to Apache.”
>
> *—Matei Zaharia, VP, Apache Spark, Founder & CTO, Databricks*

How does Apache Spark work?

Spark takes programs written in a high-level concise language and distributes the execution of its tasks onto many machines. It achieves this through an API such as DataFrames and Datasets built atop Resilient Distributed Datasets (RDDs) — a distributed dataset abstraction that performs calculations on large clusters in a fault-tolerant manner.

Spark’s architecture differs from earlier approaches in several ways that improves its performance significantly. First, Spark allows users to take advantage of memory-centric computing architectures by persisting DataFrames, Datasets, and RDDs in-memory, enabling fast iterative processing use cases such as interactive querying or machine learning. Second, Spark’s high level DataFrames and Datasets APIs also enable further intelligent optimization of user programs. Third, Project Tungsten and Catalyst Optimizer as part of the Spark SQL engine significantly boost Spark’s execution speed in many cases by 5-10X.

However, sheer performance is not the only distinctive feature of Spark. Its true power lies in unity and versatility. Spark unifies previously disparate functionalities including batch processing, advanced analytics, interactive exploration, and real-time stream processing into a single unified data processing framework.



Apache Spark Components

What are the benefits of Apache Spark?

Spark was initially designed for interactive queries and iterative algorithmic computation, as these were two major use cases not well served by batch frameworks like MapReduce. Consequently, Spark excels in scenarios that require fast performance, such as iterative processing, interactive querying, batch and real-time streaming data processing, and graph computations. Developers and enterprises deploy Spark because of its inherent benefits:

Simple

Easy-to-use, high-level declarative and unified APIs for operating on large datasets. This includes a collection of over 100 operators for transforming data, familiar DataFrame/Dataset domain-specific APIs for manipulating structured or semi-structured data, and a single point of entry for Spark applications to interact with Spark.

Speed

Engineered from the bottom-up for performance, running 100x faster than Apache® Hadoop™ by exploiting in memory computing and Tungsten's and Catalyst's code optimizations. Spark is also fast when data is stored on disk, and in 2014 Spark set the world record for large-scale on-disk sorting.

Simplified and Unified Engine

Apache Spark is packaged with unified higher-level API libraries for DataFrames/Datasets, including support for SQL queries, structured streaming, machine learning and graph processing. These standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.

Through unified DataFrames/Datasets built atop SQL Engine and extended to Spark streaming and Machine Learning MLlib, developers can write end-to-end continuous applications, where they can perform advanced analytics on both static and continuous data (including real-time).

Integrate Broadly

Built-in support for many data sources, such as HDFS, Kafka, RDBMS, S3, Cassandra, and MongoDB, and data formats, such as Parquet, JSON, CSV, TXT, and ORC.

What is the relationship between Apache Spark and Apache Hadoop?

Spark is bigger than Hadoop in adoption and widely used outside of Hadoop environments, since the Spark engine has no required dependency on the Hadoop stack. Around half of Spark users don't use Hadoop but run directly against key-value store or cloud storage. For instance, companies use Spark to crunch data in "NoSQL" data stores such as Cassandra and MongoDB, cloud storage offerings like Amazon S3, or traditional RDBMS data warehouses.

In the broader context of the Hadoop ecosystem, Spark can interoperate seamlessly with the Hadoop stack. It can read from any input source that MapReduce supports, ingest data directly from Apache Hive warehouses, and runs on top of the Apache Hadoop YARN resource manager.

In the narrower context of the Hadoop MapReduce processing engine, Spark represents a modern alternative to MapReduce, based on a more performance oriented and feature rich design. In many organizations, Spark has succeeded MapReduce as the engine of choice for new projects, especially for projects involving multiple processing models and workloads or where performance is mission critical. Spark is also evolving much more rapidly than MapReduce, with significant feature additions occurring on a regular basis.

> "MapReduce is an implementation of a design that was created more than 15 years ago. Apache Spark is a from-scratch reimagined or re-architecting of what you want out of an execution engine given today's hardware."
>
> *—Patrick Wendell, Founding Committer, Apache Spark & Co-founder, VP of Engineering, Databricks*

What are some common Apache Spark use cases?

Because of its unique combination of performance and versatility, over a 1000 organizations, in many industries across a wide range of use cases, have deployed Spark. While innovators are constantly deploying Spark in creative and disruptive ways, common use cases include:

Data integration and ETL

Cleansing and combining data from diverse sources for visualization or processing or analyzing in the future. Examples include Edmund's use of data integrity for improved customer experience and Netflix's productionizing ETL at petascale.

Interactive analytics or business intelligence

Gaining insight from massive data sets to inform product or business decisions in ad hoc investigations or regularly planned dashboards. Examples include DNV GL's predictive analytics in the energy sector, Goldman Sachs' analytics platform in the financial sector, and Huawei's query platform in the telecom sector.

High performance computation

Reducing time to run complex algorithms against large scale data. Examples include MyFitnessPal / UnderArmour's food database in the health & wellness sector and Novartis' genomic research in the pharma sector.

Machine learning and advanced analytics

The application of sophisticated algorithms to predict outcomes, detecting fraud, inferring hidden information, or making decisions based on input data. Examples include Riot Games' massive gaming data for advanced analytics in the gaming sector, Alibaba's analysis of its marketplace in the retail sector, and Spotify's music recommendation engine in the media sector.

Real-time data processing

Capturing and processing data continuously with low latency and high reliability. Examples include Automatic's real-time analytics for smarter cars, Convivas' near real-time and offline analysis for the online video business of their customers, Netflix's streaming recommendation engine in the media sector, and British Gas' connected homes in the energy sector.

Who is Databricks?

Databricks' vision is to empower anyone to easily build and deploy advanced analytics solutions. The company was founded by the team who created Apache Spark™, a powerful open source data processing engine built for sophisticated analytics, ease of use, and speed. Databricks is the largest contributor to the open source Apache Spark project providing 10x more code than any other company. The company has also trained over 40,000 users on Apache Spark, and has the largest number of customers deploying Spark to date. Databricks provides a virtual analytics data platform, to simplify data integration, real-time experimentation, and robust deployment of production applications.

For more information on Databricks, download the Databricks Primer.

Try Apache Spark on Databricks for free or contact us for a personalized demo.

databricks®

The Best Place to Run Apache® Spark™

Try Apache Spark on 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.

170623