Simplifying Data Engineering to Accelerate Innovation

databricks

The Rise of Data Engineering

With the continued growth in data generated and captured by companies across industries, the market for big data analytics capabilities is becoming more mainstream. Further amplifying this trend is the rapid ascension of modern technologies designed to help organizations harness, manage, and ultimately derive value from this data.

As a result, data engineering has quickly become one of the fastest growing functions within data driven organizations. As companies set their sights on making data-driven decisions or automating business processes with intelligent algorithms, mastering data engineering is an essential step.

Primary Data Engineering Challenges

SILOED DATA

Data often exist in disparate silos, making it difficult to access and ETL the data in a format that the data science and analyst teams can leverage — resulting in the inability to extract holistic insights that can lead to inaccurate machine learning models and misinformed decisions.

COMPLEX INFRASTRUCTURE

A primary element of any big data project involves having to build and operate the supporting data infrastructure to operationalize your deployment. Business-critical data pipelines that go down, whether ETL or feature engineering, has the potential to cost millions of dollars in lost revenue.

PERFORMANCE AT ANY SCALE

As the volume and variety of data grows to support more sophisticated use cases, the data infrastructure must be able to adapt to workload changes and handle fast growing data volumes — both structured and unstructured — to ensure efficient compute usage and infrastructure costs at any scale.

DEMANDS OF CONTINUOUS APPLICATIONS

As organizations collect massive amounts of data on a continual basis, the ability to extract actionable insights in a timely fashion becomes critical to success. Developers must be armed with the means to perform complex mission-critical data cleansing, transformations, and manipulations on data from various sources and complex formats, all while ensuring fault-tolerance across the entire pipeline.

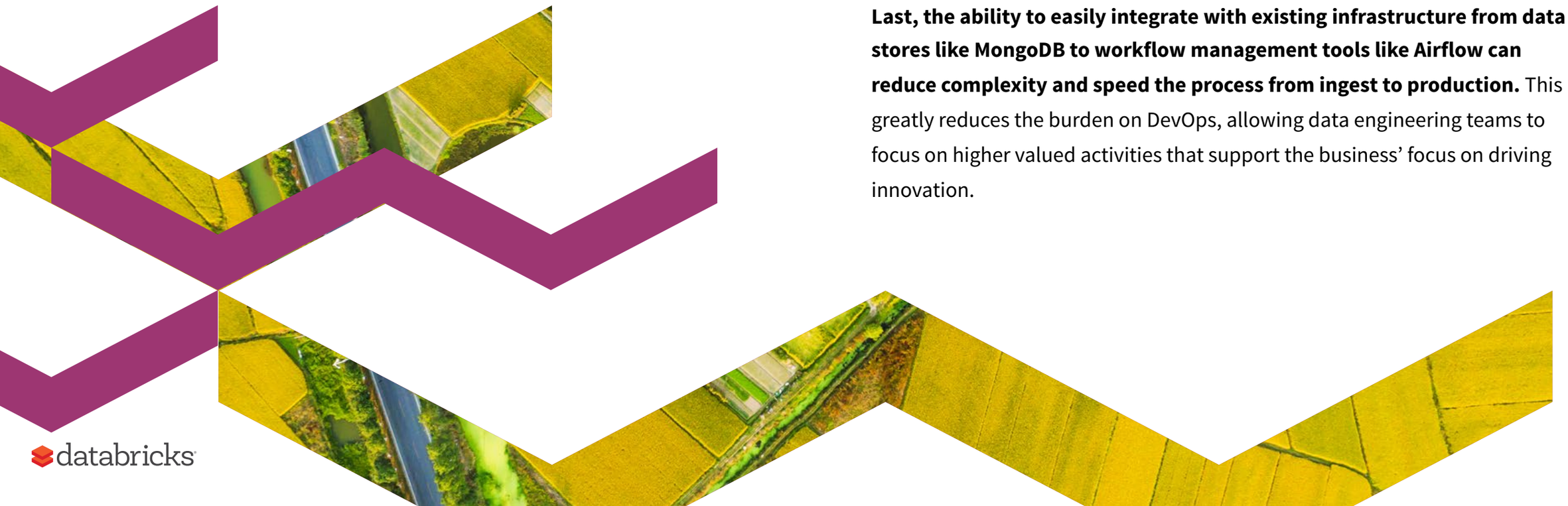
Keys to Better Data Engineering

There are **three primary keys data engineers require** to ensure they can effectively support their data science colleagues and the overall business.

First, the need for the system to be production ready is important in terms of stability and security. It's critical to build a data pipeline that is reliable and secure. Data engineering teams must be able to not only prevent outages through troubleshooting, but also ensure the necessary data protection to meet security and compliance standards.

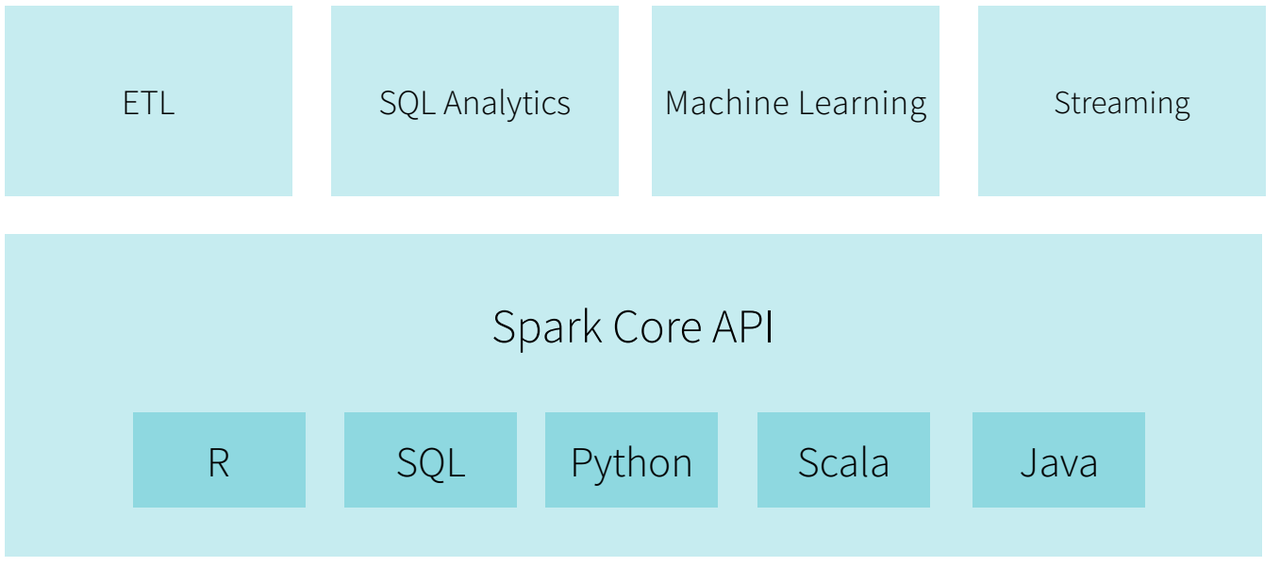
Second, the ability to process big data at breakneck speeds can help a business innovate and drive favorable business outcomes faster. Optimizing the various steps of data engineering is essential to improve process efficiency that leads to faster delivery of impactful business outcomes.

Last, the ability to easily integrate with existing infrastructure from data stores like MongoDB to workflow management tools like Airflow can reduce complexity and speed the process from ingest to production. This greatly reduces the burden on DevOps, allowing data engineering teams to focus on higher valued activities that support the business' focus on driving innovation.



The Fastest Data Processing Engine Around

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 Facebook, Netflix, Yahoo, Baidu, and eBay have eagerly deployed Spark at massive scale.



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**.

How Enterprises Deploy Apache Spark

Facebook Chose Spark for Performance and Flexibility

Facebook recently transitioned off of Hive to Spark for large-scale language model training.

Trained a large language model on

15x

more data vs. Hive

Spark was

2.5x

faster than Hive

Read the Facebook blog

to learn more >

NETFLIX

Netflix uses Apache Spark for **real-time stream processing** to provide online recommendations to its customers.

Alibaba.com

One of the world's largest e-commerce platform Alibaba Taobao runs some of the largest Apache Spark jobs in the world, **analyzing hundreds of petabytes of data** on its ecommerce platform.

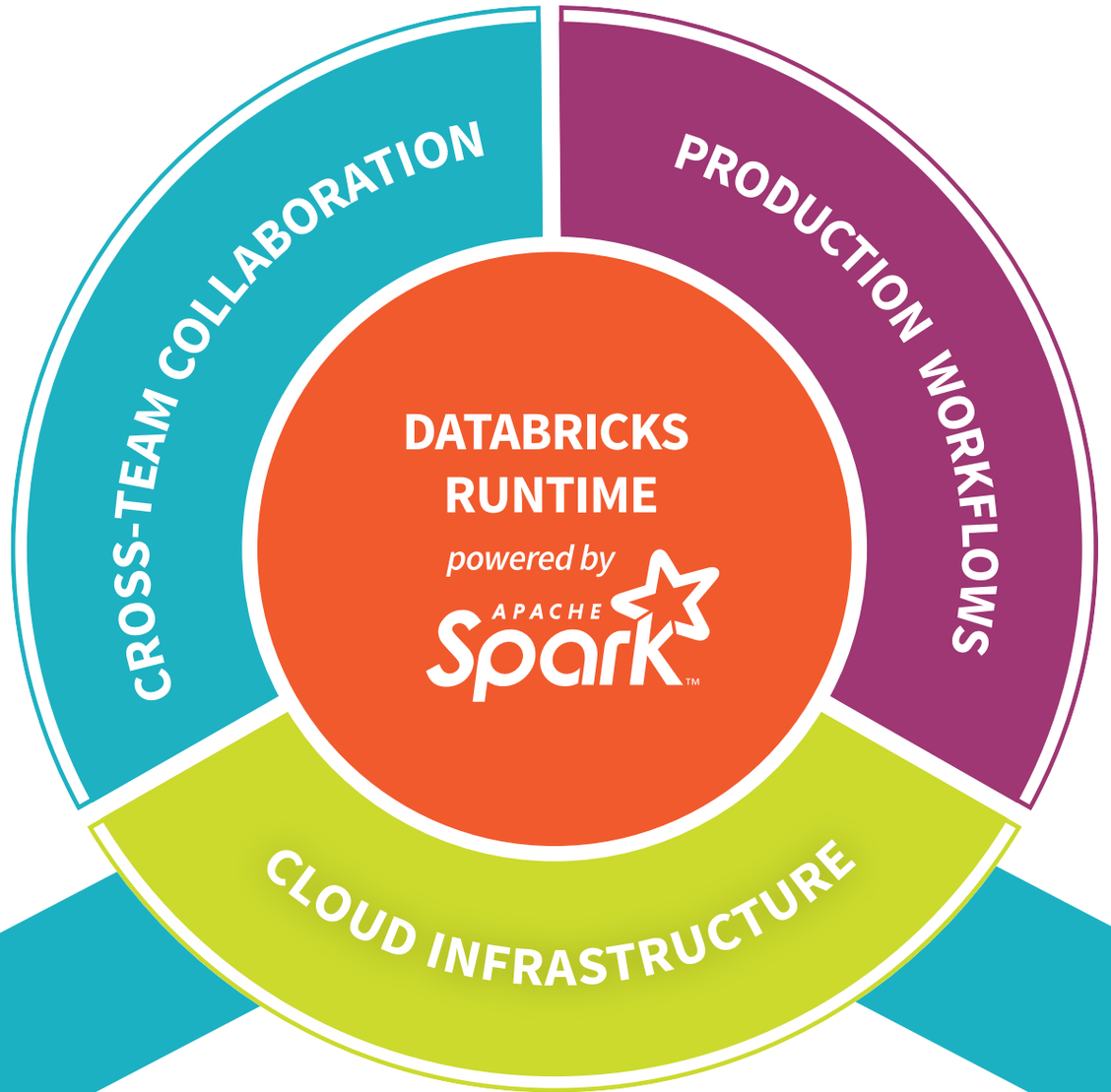
ebay

eBay uses Apache Spark to **provide targeted offers**, enhance customer experience, and to **optimize the overall performance**.

Databricks for Data Engineering: The best place to run Apache Spark

Founded by the team that created Apache Spark, Databricks' Unified Analytics Platform accelerates innovation across data science, data engineering, and the business.

Databricks provides a fully managed, scalable, and secure cloud infrastructure that helps data engineers **build and run faster and more reliable production-ready data pipelines — reducing operational complexity and total cost of ownership.**



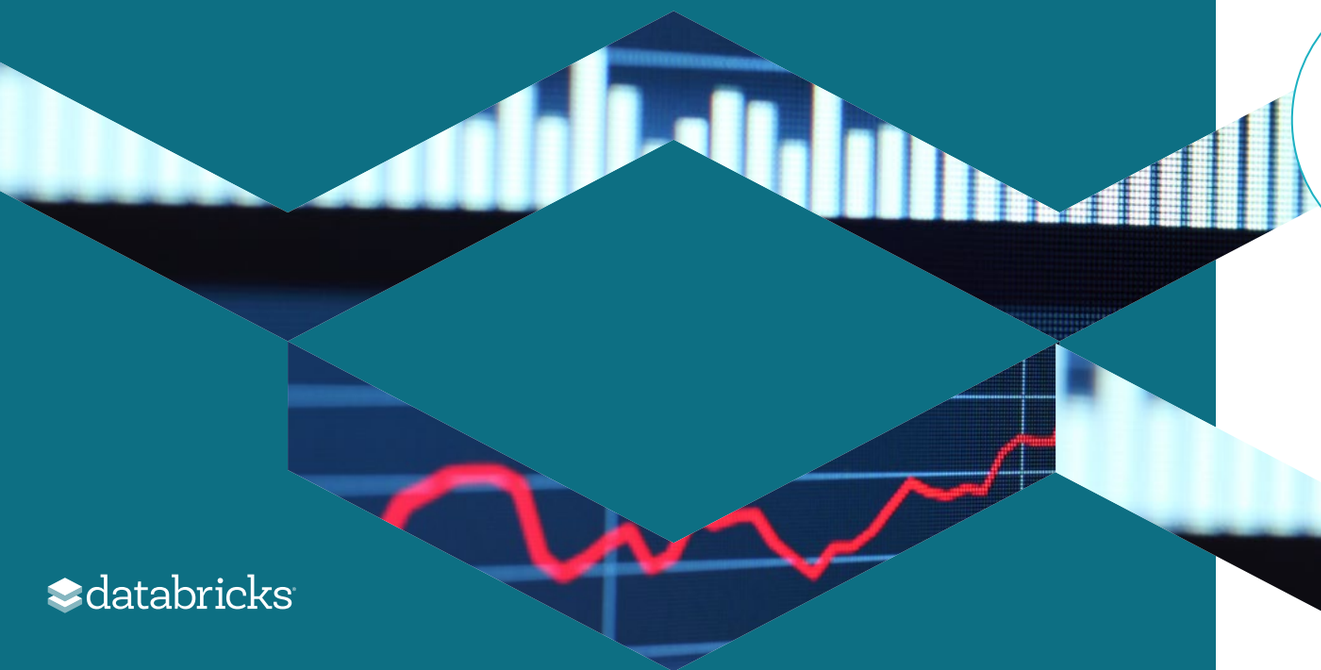
DATABRICKS' UNIFIED ANALYTICS PLATFORM

Alleviate Infrastructure Complexity Headaches

Infrastructure teams can **stop fighting complexity** and **start focusing on customer-facing applications** by getting out of the business of maintaining big data infrastructure. Databricks' serverless, fully-managed, and highly elastic cloud service **completely abstracts the infrastructure complexity and the need for specialized expertise to setup and configure your data infrastructure.**

Databricks offers ultimate flexibility by supporting all versions of Spark and the ability to run different Spark clusters to meet your workload needs — **ensuring your data engineering workloads don't get in the way of interactive queries run by your data science colleagues.**

When outages and performance degradations occur, data engineers can easily monitor the health of Spark jobs and debug issues with **easily accessible end-to-end logs in AWS S3 via the Spark UI**. And because **Databricks has the industry's leading Spark experts**, the service is fine-tuned to ensure ultra-reliable speed and reliability at scale.



SUPPORTS ALL SPARK VERSIONS

MULTIPLE SPARK CLUSTERS TO SUIT WORKLOADS

EASILY ACCESSIBLE END-TO-END LOGS

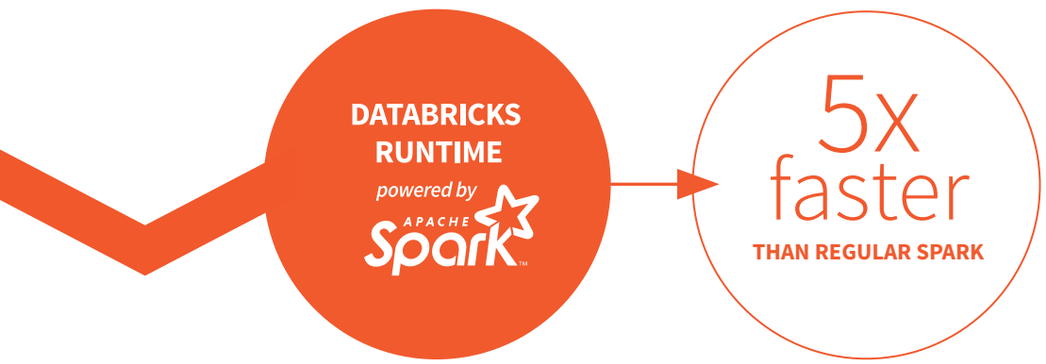
FINELY-TUNED SPARK FOR SPEED AND RELIABILITY

Faster Performance:
Databricks Runtime Powered by Spark

For Data Engineers, it's critical to process data no matter the scale as quickly as possible. Apache Spark is the proven processing engine faster than any other big data processing technology available.

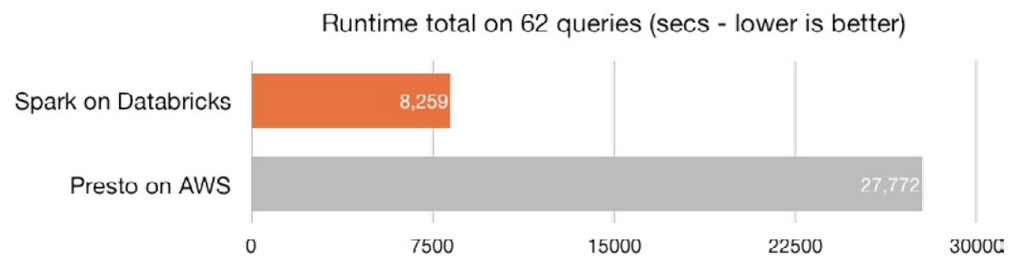
Databricks has taken Spark performance to another level through Databricks Runtime.

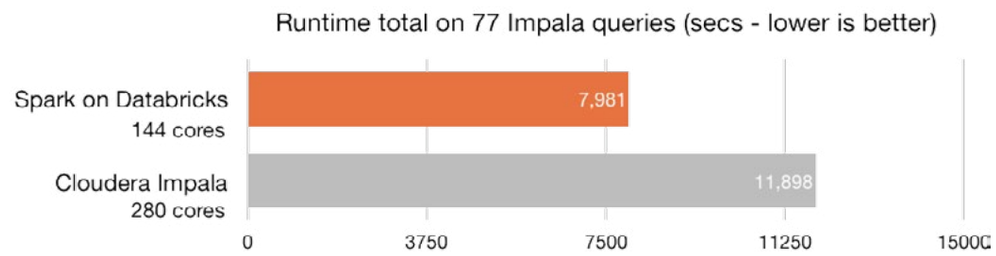
Databricks Runtime is built on top of Spark and natively built for the cloud. Through various optimizations at the I/O layer and processing layer (Databricks I/O), we've made Spark faster and more performant. Recent benchmarks clock Databricks at a rate of 5x faster than vanilla Spark on AWS. Our Spark expertise is a huge differentiator in ensuring superior performance and very high reliability.



These value-added capabilities will increase your performance and reduce your TCO for managing Spark.

In fact, in a recent performance comparison using the TPC-DS industry standard benchmark, Databricks outperformed other leading big data SQL platforms, demonstrating superior performance across the board.





FASTER DATA PROCESSING + THE CLOUD = LOWER COMPUTE AND STORAGE COSTS

Keep Data Safe and Secure

They say all press is good press, but a headline stating the company has lost valuable data is never good press. When a breach happens the enterprise grinds to a halt, and innovation and time-to-market is out the window. Databricks takes security very seriously, and by providing a **common user interface** as well as **integrated technology set**, data is protected thanks to a **unified security model with fine grained access controls across the entire stack** (such as data, clusters, and jobs) and **automatically encrypt and scale local storage**.

Lower Costs

Databricks' performance-tuned Apache Spark clusters allow you to complete jobs in a shorter time, reducing cloud compute costs.

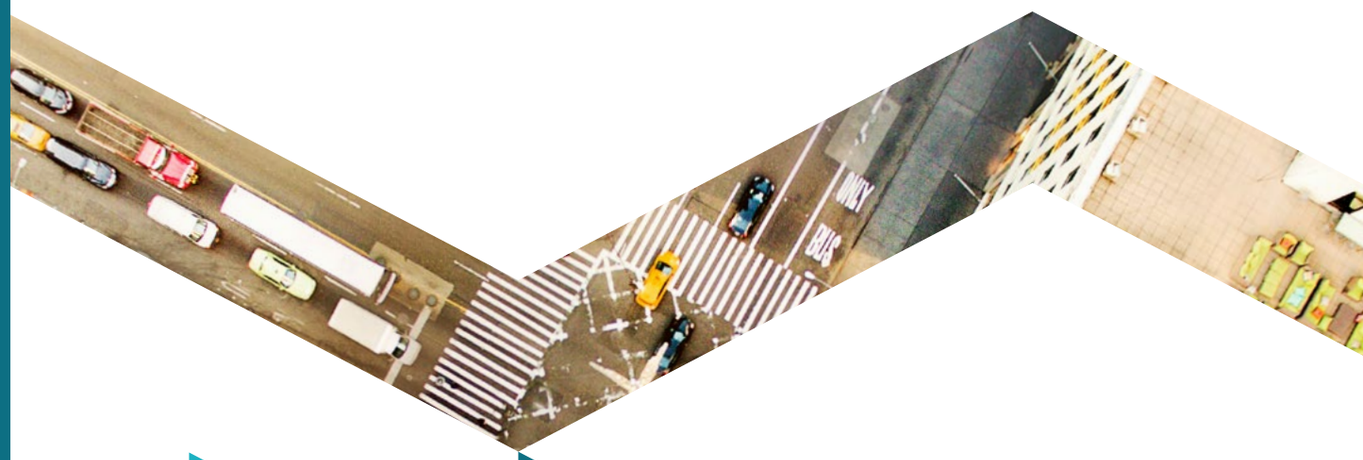
The fully-managed Databricks Spark clusters enable you to further reduce costs by **avoiding time-consuming tasks to build, configure, and maintain complex Spark infrastructure**.

In addition to being able to use **spot instances**, Databricks clusters can also **automatically scale to dynamic workloads**. Moreover, Databricks bills your usage at the minute-level, ensuring you **only pay for the resources you use**.

5 Customer Case Studies: Productionizing Data Pipelines Effortlessly

Many of our customers faced the aforementioned challenges when it came to data engineering tasks that impacted process efficiency and slowed the ability for the business and data science teams to glean insights from all the data.

The following case studies highlight how some of our customers — across all verticals — have leveraged Databricks to simplify data engineering and accelerate their ability to **build reliable and highly performant data pipelines**, allowing the business to **leverage data-driven insights to fuel innovation** and **reduce overall costs**.



Case Study: Advertising Technology

Eyeview is a video advertising technology company that provides brands with a higher return-on-investment on their video advertising spend.

USE CASE

Data is an integral part of Eyeview's platform, enabling the planning and optimization of video advertising campaigns for Eyeview's customers.

Eyeview extracts consumer knowledge and business intelligence data from first and third party sources to create thousands of ad permutations for different audiences, personalizing the ads based on factors such as location, shopping habits, and browsing history.

Due to the scale and complexity of the processing necessary to achieve this level of personalization, Eyeview needed to ensure that the technology foundation of its platform was capable of efficiently scaling to support massive volumes of data and incorporating predictive analytics through high-performing machine learning models.

CHALLENGE

Eyeview's legacy data platform struggled to scale their infrastructure to meet business growth because:

- Surging data volumes caused ETL jobs and query performance to slow down beyond acceptable performance requirements.
- Cost and labor resources to operationalize the infrastructure in support of increased demand became prohibitive.
- Lack of native support for machine learning critical for competitive differentiation.

DATABRICKS SOLUTION

Eyeview selected the Databricks Unified Analytics Platform for just-in-time data warehousing and to deploy machine learning models into production.

- Simplify provisioning of Spark clusters to automatically scale based on usage.
- Further reduce infrastructure costs through the use of auto-scaling and spot instances.
- Scale its compute and storage resources independently, providing high performance at a much lower cost.
- Effortlessly perform real-time ad hoc analysis and implement machine learning models.

BENEFITS

- Reduced query times on large data sets by a factor of 10, allowing data analysts to regain 20 percent of their workday from waiting for results.
- Sped up data processing by fourfold without incurring additional operational costs.
- Doubled the pace of product feature development, from prototyping to deployment, by increasing the productivity of the engineering team with faster and easier management of Apache Spark clusters.

> “Databricks is our go-to-system for anything requiring deep data processing and analysis. In just a short amount of time, we have been able to increase our data processing speeds by a factor of four without any added operational costs.”
>
> — Gal Barnea, CTO, Eyeview

Case Study: Travel and Hospitality

HomeAway

HomeAway allows travelers to search for vacation rentals in desired destinations. To facilitate a match between traveler and vacation rental, HomeAway must show search results that are relevant to the traveler's specific interests.

USE CASE

HomeAway, a subsidiary of Expedia, is one of the world's leading online marketplaces for the vacation rental industry, with websites representing over one million paid listings of vacation rental homes in 190 countries. Travelers use their websites and mobile applications to search for vacation rentals in desired destinations.

To achieve ideal results that enhance the user experience and drive conversions, HomeAway leverages machine learning to first comb through various data to deliver accurate search results, then they leverage context classification techniques to associated the right images based on search term.

CHALLENGE

Dealing with large volumes of structured and unstructured data, HomeAway spent too much time on DevOps work building and maintaining infrastructure with open source Apache Spark and Zeppelin notebooks.

DATABRICKS SOLUTION

HomeAway replaced its homegrown environment with Databricks to simplify the management of their Spark infrastructure through its native access to S3, interactive notebooks, and cluster management capabilities.

BENEFITS

- Reduced query time of over one million documents from over one week to 24 hours.
- Reduced over-reliance on DevOps team, increasing data science productivity by 4x.
- Automated the execution of microservices via Databricks' REST APIs.

> "Databricks takes the pain of cluster management away so we can focus on the data and not DevOps."
>
> — Brent Schneeman, Principal Data Scientist, HomeAway

Case Study: Health and Fitness / IoT

myfitnesspal

MyFitnessPal aims to build the largest health and fitness community online by helping people achieve healthier lifestyles through better diet and increased exercise.

USE CASE

MyFitnessPal, part of Under Armour, aims to build the largest health and fitness community online by helping people achieve healthier lifestyles through better diet and increased exercise.

One of the most critical data products used by the MyFitnessPal community is the food database which helps people to quickly find and log everything they eat. To support this product, they created a new feature called “Verified Foods”.

CHALLENGE

- The development of new features within their application demanded a faster data pipeline to process streams of unstructured data and to execute a number of highly sophisticated machine learning algorithms.
- Their legacy non-distributed Java-based data pipeline was slow, did not scale, and lacked flexibility.

DATABRICKS SOLUTION

MyFitnessPal chose Databricks to harness the power of Apache Spark and to build the data pipeline for “Verified Foods” to successfully deliver the feature to their users while gaining many additional benefits.

BENEFITS

- Ten-fold speed improvement over previous data pipeline implementation.
- Four times more projects completed in the past quarter resulting from an increase in team productivity.
- Improved team efficiency achieved through accessible advanced analytics and better code re-use.

> “Databricks helped us deliver a new feature to market while improving the performance of the data pipeline ten-fold. We would not have been able to fully harness the power of Apache Spark to deliver the feature to market without Databricks.”
>
> — Chul Lee, Director of Data Engineering & Science, MyFitnessPal

Case Study: Advertising Technology

SHARETHROUGH

Sharethrough builds software for delivering ads into the natural flow of content sites and apps (also known as native advertising).

USE CASE

Since Sharethrough serves ads on some of the most popular digital properties such as Forbes and People, the need for a high-performance big data scale processing platform permeates every aspect of their business.

Sharethrough offers a robust advertising platform; discovering hidden patterns in data is critical to measuring the effectiveness of their products and in making improvements to the overall product suite.

CHALLENGE

- Initial attempt to establish a self-hosted Apache Hadoop cluster with Apache Hive as the ad hoc query tool required two full-time engineers to manage the infrastructure.
- Their homegrown system was also not an effective interactive query platform, creating additional demands on data engineering to build and maintain a high performing data pipeline.

DATABRICKS SOLUTION

Databricks offered significant data engineering benefits for Sharethrough, including:

- faster prototyping of new applications
- easier debugging of complex pipelines
- improved engineering productivity

BENEFITS

- Faster prototyping of new applications.
- Easier debugging of complex pipelines.
- Improved overall engineering team productivity by freeing two full time engineers from infrastructure work.

> “Thanks to Databricks, our engineers have gone from being burdened with operations, to having the ability to easily dive right into analytics. As a result, our team is more productive and collaborative with big data than ever.”
>
> — Robert Slifka, Vice President of Engineering, Sharethrough

Case Study: Enterprise Software

Yesware enables sales teams to be more effective by providing detailed analytics about their daily interactions with potential customers.

USE CASE

As sales organizations continue to rely more and more on the data-driven decision-making approach to improving their sales forecasting and ability to close deals, they are also demanding higher-accuracy data during their decision-making process.

Yesware enhances a sales team's sales cycle by integrating with the team's e-mail application to track key metrics. Important data such as the open rate of e-mail templates, the download rate of attached collateral, and CTA click-through rates are analyzed to generate custom reports for the entire team — allowing sales teams to connect with prospects more effectively, more easily track customer engagement, and close more deals.

CHALLENGE

- Yesware needed a high-performance production data pipeline to power its main product, which provides customized intelligence to improve the performance of sales teams.
- The data pipeline built with Apache Pig was too slow, difficult to maintain, and not scalable enough for Yesware's needs.

DATABRICKS SOLUTION

Databricks provided an easier to deploy, faster, more reliable, and more efficient data pipeline; enabling Yesware to gain time improvements for deployment, processing speed, and infrastructure efficiency.

- Yesware took advantage of features including:
- Spark cluster manager simplified the provisioning of highly optimized Spark clusters simple clicks without DevOps.
- Interactive workspace enabled Yesware to prototype Scala code in small quick iterations to get the logic right, and then migrate over to a production JAR.
- Job scheduler allowed Yesware to instantly deploy code and automatically monitors the execution of production data pipelines.
- Integrations with a wide variety of data stores to set up a simple but powerful data pipeline with AWS S3 and Postgres.

BENEFITS

- Reduced time to deploy production data pipeline from six months to three weeks.
- Substantially sped up compute time, processing twice the amount of data in one-sixth the amount of time.
- Improved efficiency of data processing infrastructure by reducing Amazon Web Services (AWS) costs by 90%.

> “Databricks proved to be the easiest way to deploy Apache Spark for Yesware, reducing the time to deploy a production pipeline from six months to three weeks while enabling us to shorten the time to prototype new product features from days to mere hours.”
>
> — Justin Mills, Data Team Lead, Yesware

Data Engineering, Simplified.

Databricks' Unified Analytics Platform removes the complexity of data engineering while accelerating performance of data engineering tasks from data access to ETL to production, allowing engineers to build fast and reliable data pipelines more easily to support the business.

Get started with Databricks for data engineering today.

START YOUR FREE TRIAL

databricks