Unifying Data and AI

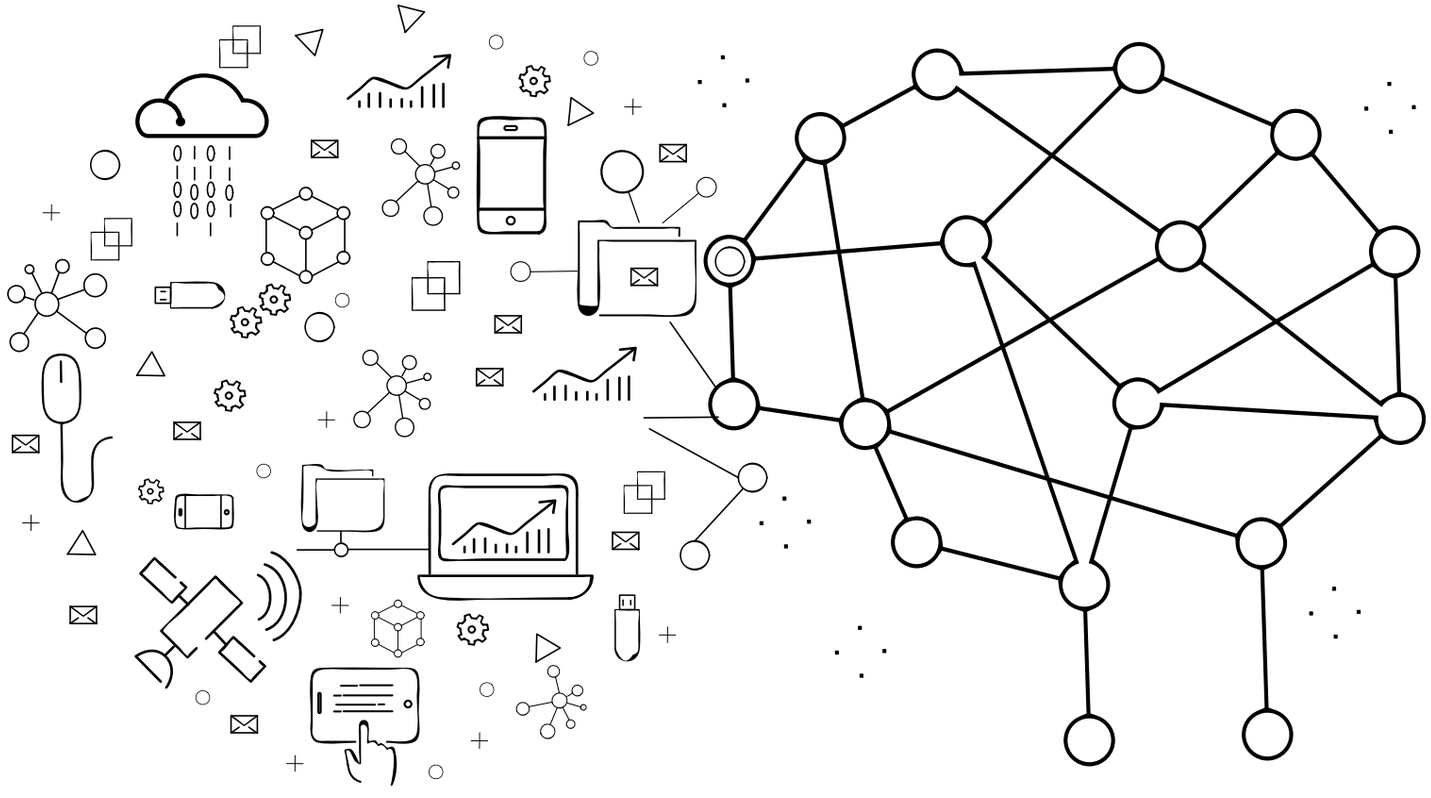
How a Modern Approach to Analytics Can Accelerate Innovation

databricks

Introduction

The world has come a long way since the early days of data analysis where a simple relational database, point-in-time data, and some internal spreadsheet expertise helped to drive business decisions. Today, enterprises are focusing massive amounts of resources to transform their business through machine learning and automation. This allows them to drive competitive advantage, improve customer experience, and more efficiently manage cost. According to a recent survey with CIO.com, nearly 90% of enterprises are investing in AI related technology.*

Data is at the core of how these modern enterprises are changing their business. With this data, enterprises are able to tap into the promise of AI to drive disruptive innovations affecting nearly every enterprises on the planet. The challenge most enterprises face is how to succeed with both their data and AI?



* https://databricks.com/cio-survey-report

Challenges

Data is Difficult to Prepare

We all know data is the key to success in today's digital age. That's why enterprises are modernizing their data architectures — moving away from legacy systems that are complex to manage and lack the flexibility for new data sources and advanced analytics. The core of this migration is the data. However, preparing data for AI is a major bottleneck. In fact, 96% of enterprises cite data related challenges as the #1 blocker to the success of AI projects. Enterprise data is often siloed across hundreds of systems such as data warehouses, data lakes, databases and file systems that are not AI-enabled. This results in an enormous amount of time is spent combining, cleaning and verifying, enriching, and featurizing the data to get it ready for the model. Furthermore, the need to manage streaming datasets (such as IoT and social) along with historical data for real-time analytics increases this complexity even more.

Underscoring this trend, 87% of enterprises are investing in technology to help with data preparation and exploration.* This work required for downstream analytics and AI is putting increasing demand on data engineering teams to enable the business with high quality datasets while keeping costs low, data secure, and complex data pipelines performant and reliable.

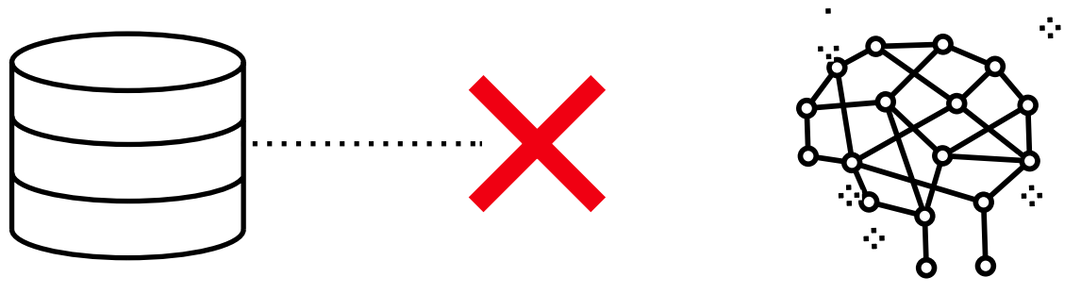
* https://databricks.com/cio-survey-report

Challenges

Siloed Teams Hinder Productivity and Time to Market

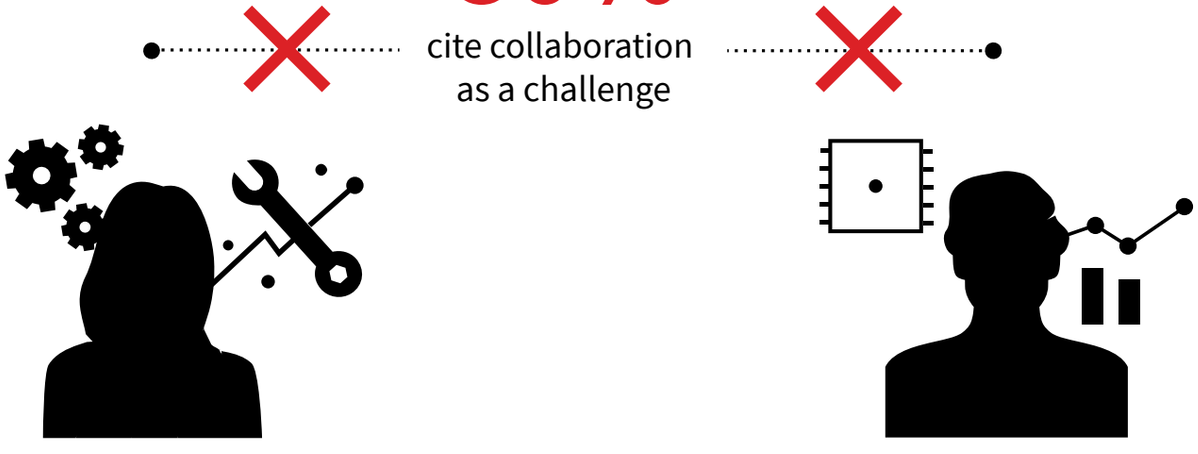
The productivity of the team structure across a data organization can be severely impacted without a seamless and dependable and unified data strategy. It is very difficult for the siloed functional roles of data scientist, machine learning engineer, data engineer, and developer to achieve synergy and work together.

Studies show that 90% of enterprises cite challenges with data engineering and data science collaboration as a reason for their inability to succeed with AI.* This organizational separation creates friction and slows projects down, becoming an impediment to the highly iterative nature of AI projects.



96%

of enterprises cite data-related challenges with AI projects



80%

cite collaboration as a challenge

Disjointed Data and AI Technologies

There has been an explosion of AI technologies like TensorFlow, PyTorch, and SciKit-Learn which are great at enabling AI capabilities but don't have the data processing capabilities necessary to bridge the gap between data engineering and data science. As a result, the capacity to feed data necessary to train a model requires multiple handoffs that open the door for errors and inefficiency. Access to data is limited without a seamless integration between data and AI technologies.

Technology and skills gaps are the largest barriers to collaboration between data engineering and data science teams.* Organizations are burdened with the limitations and complexity of setting up and maintaining distributed machine learning environments due to a multitude of point solutions and interdependencies between them. On average, large enterprises are using up to seven different tools for data engineering and data science.*

* https://databricks.com/cio-survey-report

Challenges

An Explosion of Machine Learning Frameworks Adds Complexity

TensorFlow

Caffe2

Spark MLlib

Organizations are using an average of

7

different machine learning tools and frameworks

mahout

PYTORCH

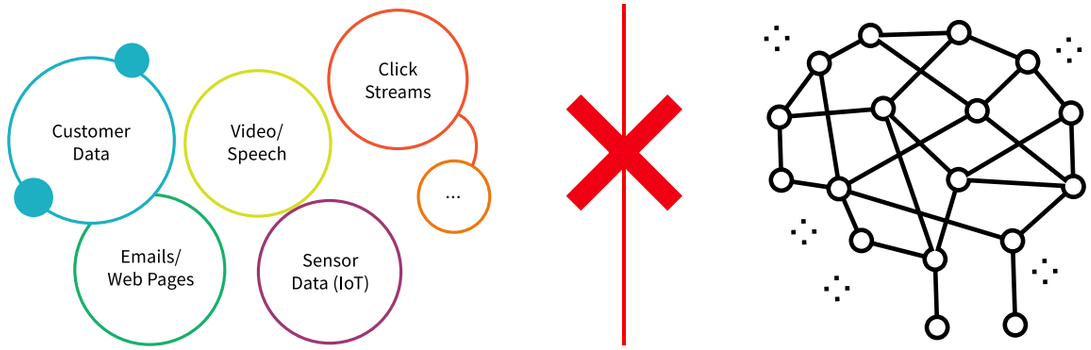
scikit learn

mxnet

theano

databricks

Beyond the myriad of tools at their disposal, data scientists and machine learning engineers are increasingly pressured to improve overall productivity to reduce the time needed to train sophisticated models. But the lack of highly specialized skills required to work with these AI technologies dramatically slows down projects and the ability to get to results.

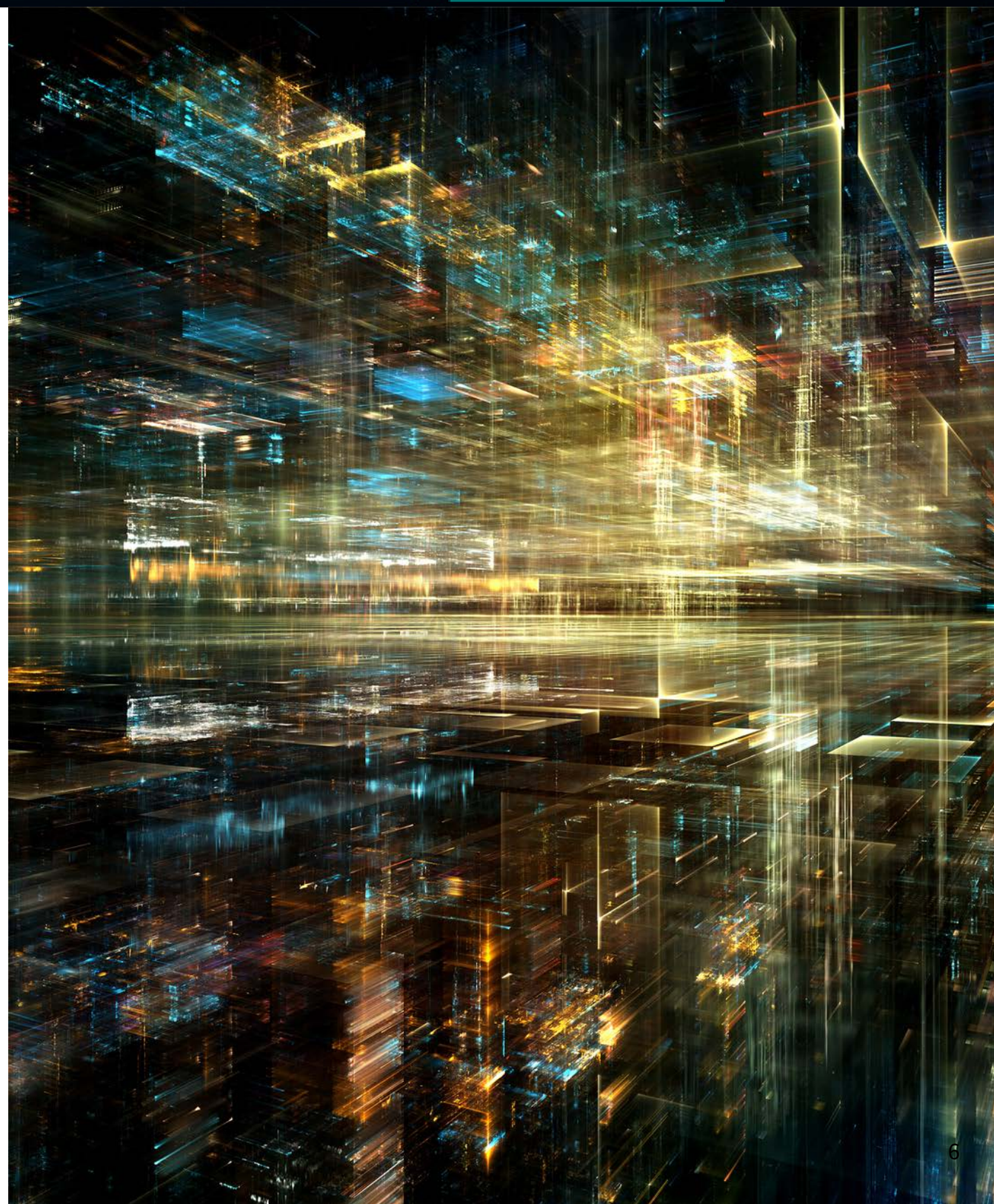


Challenges

Infrastructure Complexity, Security and Compliance Needs

As enterprises cope with rapidly growing volumes of data from various data sources, costs and operational complexity can quickly get out of control. Organizations that have not moved to the cloud and are reliant on on-premises infrastructure experience this pain tenfold as they often lack the ability to quickly and easily provision resources to meet business needs, slowing their ability to respond to demand faster, while struggling to maintain costs.

Further complicating matters, the fragmented technology set supporting the AI lifecycle and the increasing number of endpoints that needs to be secured makes it extremely hard for security stakeholders to protect one of the most valuable assets of the enterprise — its data.



The Need for a Unified Approach

> “We cannot solve our problems with the same level of thinking that created them.”
>
> — Albert Einstein

With so many data challenges facing enterprises, that act as an impediment to innovation, distracting the teams from their core competencies and increasing time to market for new products and insights, a new approach needs to be considered.

With data as the fuel for AI innovation, the modern enterprise requires a comprehensive, unified approach to analytics and AI. Over 79% of enterprises highly value the notion of a unified approach to analytics — bringing together data and AI, while enabling better collaboration and streamlining analytic workflows.*

A unified approach to analytics makes it easier for enterprises to build data pipelines across various siloed data storage systems and to prepare datasets for model building, which allows organizations to do AI on their existing data and iteratively do AI on massive data sets. Organizations also gain the benefit of integrating with a broad set of AI algorithms that can be applied to these datasets iteratively to fine-tune the models.

Lastly, unifying analytics improves collaboration across data scientists and data engineers — empowering them to work more effectively across the entire experimentation-to-production lifecycle. The organizations that succeed in unifying their data at scale and unifying that data with the best AI technologies will be the ones that succeed with AI.

Databricks, powered by Apache Spark™, provides a Unified Analytics Platform that enables organizations to accelerate innovation by unifying data and AI technologies, improving collaboration between data engineers and data scientists, and making it simpler to prepare data, train models, and deploy them into production.

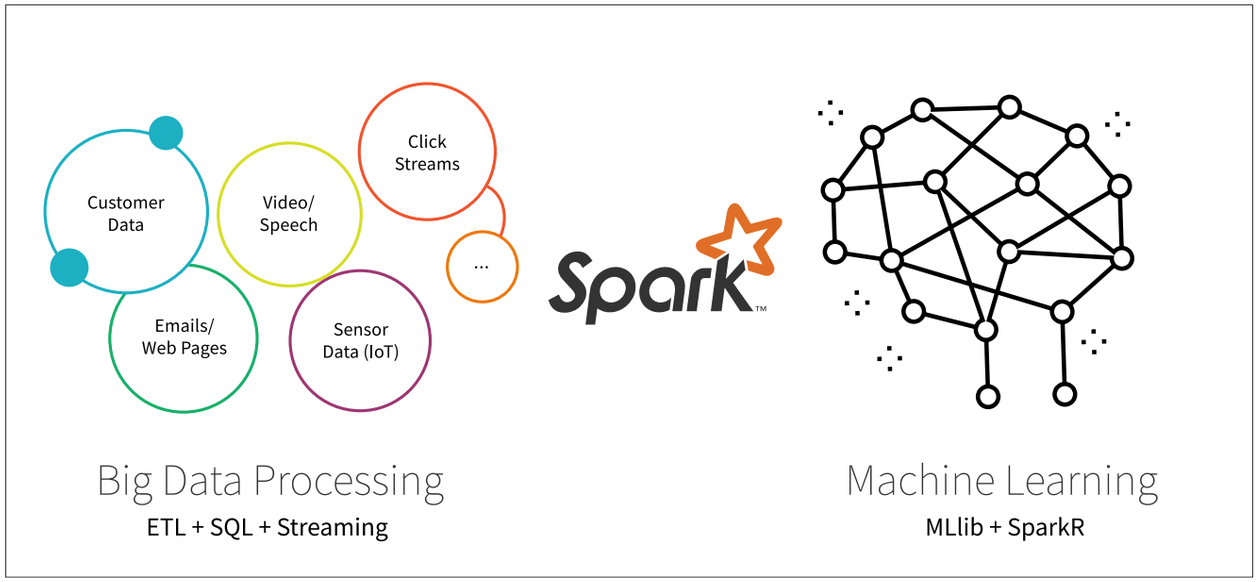
* https://databricks.com/cio-survey-report

The Databricks Advantage

Apache Spark — The Unified Analytics Engine

To avoid the problems associated with siloed data and disparate systems for handling different analytic processes, enterprises are increasingly using Apache Spark. Spark, originally created by the founders of Databricks, is the defacto standard for data processing and AI today due to its record-breaking speed, ease of use, and support for sophisticated analytics.

Spark simplifies data preparation for AI by unifying data at massive scale across various sources including cloud storage systems, distributed file systems, key-value stores, and message buses. Spark also unifies data and AI with a consistent set of APIs for simple data loading, batch/stream processing, SQL analytics, stream analytics, graph analytics, machine learning, and deep learning as well as seamless integration with popular AI frameworks and libraries such as TensorFlow, PyTorch, R and SciKit-Learn.



hadoop Parquet kafka Azure aws

The Rapid Ascension of Apache Spark

- Created at UC Berkeley in 2009 by Matei Zaharia.
- Replaced MapReduce as the de facto data processing engine for big data analytics.
- Includes libraries for SQL, streaming, machine learning and graph.
- Largest open source community in big data (1200+ contributors from 300+ orgs).
- Trusted by some of the largest enterprises (Netflix, Yahoo, Facebook, eBay, Alibaba).
- Databricks continues to drive most major efforts: Structure in Spark, DataFrames, Catalyst, Tungsten and Structured Streaming.
- Over 425,000+ meetup members around the world.

The Databricks Advantage

Unify Data Engineers and Data Scientists

With a unified approach to data and AI, data science teams can collaborate using Databricks' collaborative workspace. They can use their preferred ML frameworks and libraries to interact with the data they are modeling, and then seamlessly move those models into production with a single click.

Support for SQL, R, Python, Java, and Scala and seamless connection with popular IDEs through native integrations, or BI tools with ODBC connections allows data engineers and data scientists to use familiar languages and tools without the need to switch working environments.

By integrating and streamlining the individual elements that comprise the analytics lifecycle, these teams can create short feedback loops and work together, creating a culture of accelerated innovation. Now, thanks to Databricks it's possible to build a model and test a prototype in hours vs weeks or months with older approaches.

Databricks provides a common interface and tooling for all stakeholders (data engineers and data scientists), regardless of skill set, to foster strong collaboration. This eliminates silos and allows teams to collaborate across the AI lifecycle, from experimentation to production, which in turn benefits the organization and increases innovation.

> "We chose Databricks over Hadoop-based alternatives because it is a unified cloud-based big data processing platform that is built on top of Apache Spark, combining the fast performance and standard libraries of Spark with a user-friendly interface that fosters collaboration across our teams."
>
> — Robert Ferguson, Director of Engineering, Automatic Labs

The Databricks Advantage

Build Reliable and Performant Data Pipelines

Building best-in-class AI applications requires data, and a lot of it. Data science techniques that were actually developed years ago are only now starting to show promising results due to the sheer volume of data that can finally be used to train algorithms. And the faster you can ingest and prepare the data for analytics, the sooner you can realize the benefits of AI. Databricks has taken data processing performance to another level through Databricks Runtime. Databricks Runtime is built on top of Spark, natively for the cloud.

Through various optimizations for large-scale data processing in the cloud, we've made Spark faster and more performant. Recent benchmarks clock Databricks at a rate of 50x faster than Apache Spark on AWS — making it simpler to build highly reliable data pipelines capable of processing massive datasets at blazing speeds.

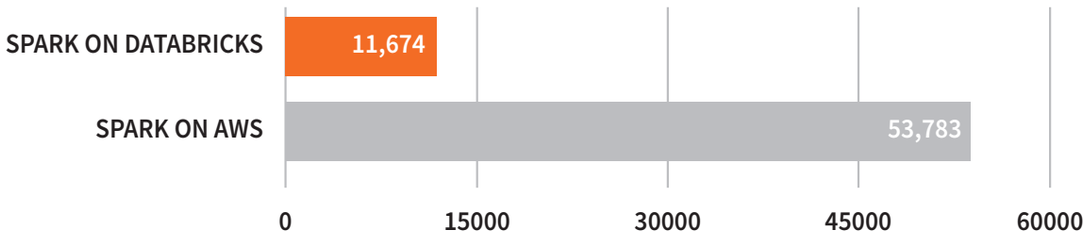
Reliability is of utmost importance when dealing with critical workloads and applications. Databricks offers a 99.9% SLA through it's fully managed cloud service, as well as transactional guarantees with the Delta technology within Databricks Runtime, making real-time data accessible quickly for downstream analytics and AI.

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.

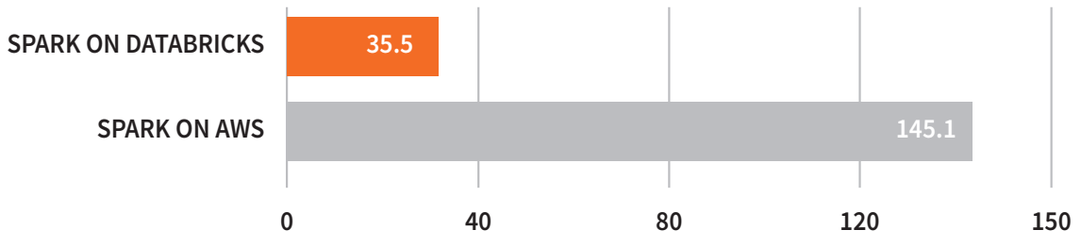
> "Databricks takes the pain out of cluster management, and puts the real power of these systems in the hands of those who need it most: developers, analyst, and data scientists are now freed up to think about business and technical problems."
>
> — Shaun Elliott
> Technical Lead of Service Engineering, Edmunds.com

The Databricks Advantage

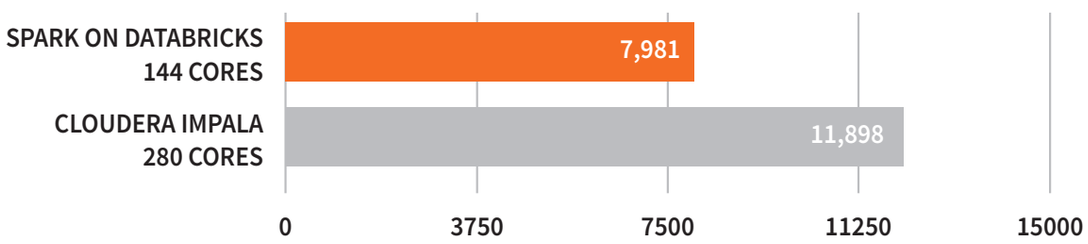
RUNTIME TOTAL ON 104 QUERIES (SECS — LOWER IS BETTER)



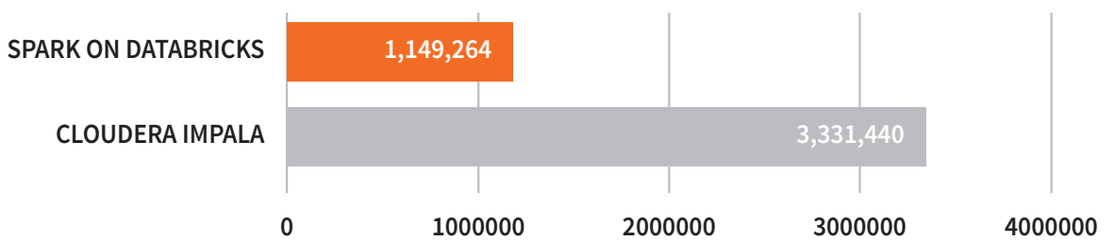
RUNTIME GEOMEAN ON 104 QUERIES (SECS — LOWER IS BETTER)



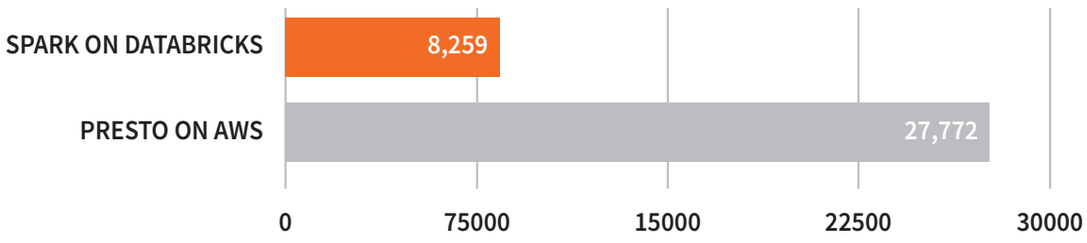
RUNTIME TOTAL ON 77 IMPALA QUERIES (SECS — LOWER IS BETTER)



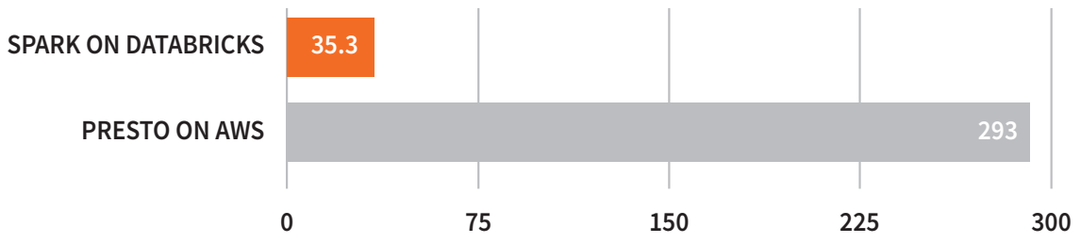
RUNTIME TOTAL ON 77 IMPALA QUERIES, NORMALIZED BY CPU CORES
(CPU TIME— LOWER IS BETTER)



RUNTIME TOTAL ON 62QUERIES (SECS — LOWER IS BETTER)



RUNTIME GEOMEAN ON 62 QUERIES (SECS — LOWER IS BETTER)



The Databricks Advantage

Build Cutting-Edge AI Models at Massive Scale

With Databricks, data science teams can leverage the Unified Analytics Platform to easily train, evaluate, and deploy more effective AI models to production. And to easily connect with data sets to perform data exploration, analysis, and transformations using SQL, R, or Python. And interactively explore data with collaborative notebooks that allow data scientists to take machine learning models from experimentation-to-production at scale. And with prepackaged AI frameworks such as TensorFlow, Horovod, Keras, XGBoost, PyTorch, SciKit-Learn, MLlib, GraphX, and sparklyr, data science teams can easily provision AI-ready Databricks clusters and notebooks in seconds on its cloud native service.

Finally, the Databricks Unified Analytics Platform significantly simplifies parallelization and distributed model training on CPU and GPU across cloud platforms via built-in optimization features and libraries (such as Horovod Estimator). It also natively decouples compute and storage, reducing the need to move data and allowing significantly faster analysis on massive amount of data at lower cost.

The Databricks Advantage

Reliability and Security in the Cloud

The proliferation of siloed-data types and point solutions for data management (data lakes, data warehouse, and streaming) is increasing costs and operational complexity. Further exacerbating the problem is the inability of on-premises infrastructure to automatically scale resources to meet changing business needs. This leads to operational costs running amok. Security also is a challenge as compliance standards such as HIPAA and GDPR are increasing pressure on the business to keep data safe and secure.

Reap the benefits of a fully managed service and remove the complexity of big data and machine learning to focus more on innovation, while keeping data safe and secure.

Databricks' elastic cloud service is designed to reduce operational complexity while ensuring reliability and cost efficiency at scale, with a unified security model featuring fine-grained controls, data encryption, identity management, rigorous auditing, and support for compliance standards.

Lowering the Total Cost of Ownership

Databricks lowers TCO with a cloud native Unified Analytics Platform that means no costly hardware, an operationally simple platform with built-in automation features designed to help you efficiently manage your costs, increased productivity through seamless collaboration, and faster performance than other analytics products which allows you to accelerate AI innovation.

Customer Proof Point: LoyaltyOne

Company

LoyaltyOne, Inc. is a global provider of loyalty marketing and programs to enterprises in retail and financial services. AIR MILES is their flagship product and Canada's largest loyalty program that serves over 11 million households.

Use Case

Their goal is to create a highly personalized experience that is optimized for conversions for their partner retailers. They call it 1:1 Conversational Marketing. Through machine learning and predictive analytics, they have built self-learning offer optimizations that help partners deliver the right offer at the right time to motivate customer behavior.

Challenge

- Their legacy Netezza data warehouse did not allow them to process both historical and real time data at scale, lacked the flexibility to easily handle different types of data, and impeded their ability to innovate and deliver machine learning capabilities.
- They struggled with vast amounts of data across different formats — millions of transactions from dozens of retailers, 100+ partners, 500 million emails/year, 1200 campaigns/year, and 11 million households served.
- They also struggled to make Spark accessible to a large and diverse analytics team that had a range of skills and needs.
- Lastly, there was pressure to accelerate speed to market to satisfy their partners and their legacy system created complexity that slowed progress.

Solution

Databricks provides LoyaltyOne with a unified analytics platform that simplifies and /accelerates ETL and empowers their data science organization to collaborate via interactive notebooks to build, train and deploy machine learning models.

Business Benefits

LoyaltyOne realized the following benefits:

- Simplified infrastructure management — They don't have to waste time provisioning clusters. Self-service cluster management with auto-scale/auto-termination of clusters helped reduce costs and saved management effort.
- Improved collaboration — Notebooks made it much easier to share work. The interactive nature of the workspace enabled them to support multiple user types across the organization.
- As a result, they were able to increase offer response rate by 2x with a 97% improvement in speed.

> "Databricks has provided us with the support and technology to modernize our architecture, enabling us to do data science at massive scale."
>
> — Bradley Kent, AVP, Program Analytics at LoyaltyOne

The Bottom Line

The goal of Databricks' Unified Analytics Platform is to accelerate innovation. It accomplishes this by uniting people around a shared objective with a common collaboration interface and self-service functionality. Additionally, Databricks unifies analytic workflows by seamlessly connecting operations and automating infrastructure — removing complexity for organizations and allowing them to innovate faster than ever before.

Get started on Databricks today with a free trial.

START YOUR FREE TRIAL

databricks