2020 EDITION | UPDATED CONTENT

The Art of Collaborative Data Science at Scale

A unified approach that boosts data science agility and productivity

databricks

Introduction

The creation and consumption of data is accelerating at an incredible pace. This tsunami of data is fueling big investments in strategies and technologies that will empower enterprises to harness data-driven insights and drive innovation. Whether it's discovering new drugs for better patient outcomes, creating a more engaging shopping experience or uncovering ways to improve supply chain efficiencies — enterprises are turning to data science and AI to tap into the immense value embedded in the mountains of data they are generating.

Too much data. Not enough data scientists.

The pressure to extract value from data has made data scientists some of the most coveted — and hard to find — talent on the market today. That's putting pressure on every organization to both retain and get more out of their data scientists. But right now everyone's hands are tied by tools that aren't up to the task of parsing massive data sets — and methods that are painstakingly slow. Most data scientists still perform their work manually, in isolation, on laptops, grappling with data sets that are too big, too unorganized and too unreliable. And that means data scientists spend much of their time maintaining infrastructure, parsing a disparate set of tools and technologies, and organizing data instead of focusing on their core jobs.

In a world where data is siloed and data scientists are scarce, the future belongs to those who can unify all their data and all their data teams on a single, scalable platform. When this platform is open source and features collaborative notebooks, data scientists can use whatever tools and languages they prefer as they collaborate around even the most massive data sets. This maximizes productivity because it allows data scientists to focus on their core job and extract insights faster.

Just because data science is complex, doesn't mean it has to be complicated

The state of data today

The effectiveness of a data scientist is often dependent on having access to the right sets of tools, technologies and processes that they can use to do their incredibly complex tasks. Here are the longstanding challenges facing nearly every data team today.

Infrastructure Complexity

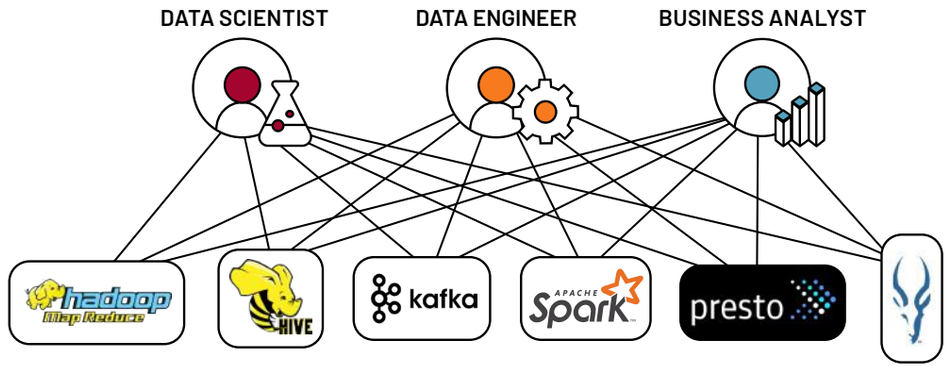
The move to the cloud is fast becoming a primary objective for businesses looking to reduce costs and create competitive differentiation. Part of the challenge associated with this shift is the complexity that surrounds setting up and maintaining big-data infrastructure. The explosion in data growth pushes organizations to move faster with infrastructure investments that can harness and derive value from this data. But doing so ultimately creates an over-reliance on DevOps teams to do the heavy lifting. For companies that don't have dedicated DevOps teams to help with these infrastructure issues, the responsibility often falls on the data scientists to fend for themselves. This creates an environment where data scientists are spending so much time configuring and setting up infrastructure that they can't focus on their core responsibilities.

Disparate technologies and analytical workflows

In their drive to become data-driven businesses, companies are deploying a patchwork of technologies. Open source projects such as Apache Spark™, Hive, Presto, Kafka, MapReduce, and Impala, offer the promise of a competitive advantage, but also come with management complexity and unexpected costs[1].

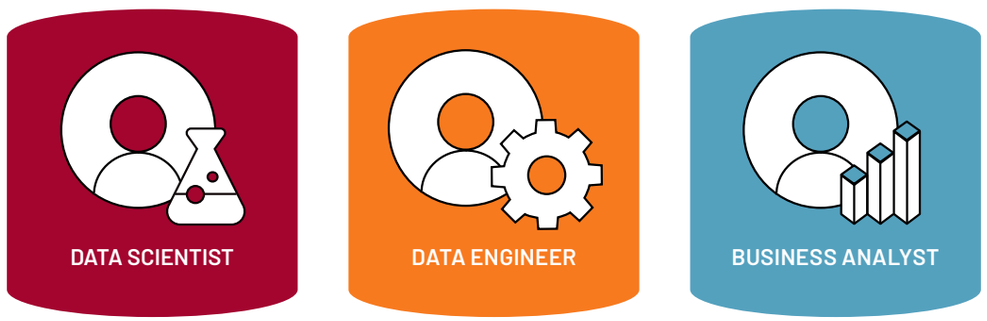
Adding to the challenge is the need to provide analysts and data scientists with support for the frameworks (scikit-learn, TensorFlow, Keras), libraries (matplotlib, pandas, numpy), scripting languages (e.g. R, Python, Scala, or SQL), tools and IDEs (JupyterLab, RStudio) they feel most comfortable using. Relying on disparate technologies to meet all these needs can be incredibly challenging as they all follow different release cycles, lack institutional support mechanisms, and have varying performance deliverables.

The net effect of deploying disparate technologies is that it throws workflows into disarray and creates bottlenecks that restrict efforts to move projects from raw data to final outcome. A lack of automation between the various steps of data ingestion, ETL, exploration, modeling and presentation of data create massive inefficiencies that can ripple through the organization[2]. This negates the potential of big data, data science and the shift to the cloud because it slows the speed of innovation to a trickle.



Siloed teams

The productivity of the team structured across a data organization can be severely impacted without a seamless and dependable big data platform. It's very difficult for functional roles of data scientist, data engineer and business user to work together because they're siloed — even within a single function — and that means people are looking at the same data through different lenses. And that complicates collaboration, erodes trust in the analytics[3], and slows the speed of innovation.



1 https://www.sungardas.com/en-GB/resources/white-papers/the-cloud-hangove

2 MSV, Janakiram (2017, February 7). Edge Computing — Redefining The Enterprise Infrastructure. Retrieved from http://forbes.com

3 https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/getting-big-impact-from-big-data

What's holding back data science today?

Data quality and exploration at scale

Data volumes continue to increase in an almost vertical trajectory, becoming highly distributed, and coming in a variety of formats. This creates new challenges to maintain quality and reliability of data sets for downstream analytics. In addition, exploring data at scale can be difficult and costly. Most organizations rely on single threaded tools to perform data exploration. The limitations of this approach are directly associated with the amount of memory on the data scientist's machine, impacting their ability to scale. As a result, they are often forced to train models against small samples of data which can result in less accurate models.

Model training is resource intensive

Model training involves using the data to incrementally improve the model's ability to solve a given problem. This process involves many steps including training the model, evaluating the results, tuning parameters to further optimize your model, and repeat. Training complex machine learning models against massive data sets can be very challenging in isolation without the ability to collaborate on models with peers. The more complex the models, the longer it will take to bring new capabilities to market.

Hard to track and reproduce results

Machine learning algorithms have dozens of configurable parameters, and whether working alone or as a team, it is difficult to track which parameters, code, and data went into each experiment to produce a model. Without detailed tracking, teams often have trouble getting the same code to work again. Whether data scientists need to pass their training code to an engineer for use in production, or need to go back to their past work to debug a problem, reproducing steps of the ML workflow is critical yet extremely challenging.

Difficult to share insights

Part of the role of a data scientist is the need to share results with team members and stakeholders for input and decision making. The trick is sharing the insights in a way that resonates with non-technical audiences. The inability to do so can hamper cross-team collaboration and slow progress.

The power of collaboration in data science

It's no secret that better collaboration often leads to improved operational efficiency and productivity. With no shortage of data to sift through and the pressures of the business to build accurate models quickly, it's more important than ever for data scientists to work effectively with other team members, colleagues across teams such as engineering, and stakeholders.

Achieving a highly collaborative environment can positively impact team efficiency, productivity, and innovation — resulting in delivering more models to production faster which can result in more revenue. As organizations continue to try to become more data-driven, creating easier access and visibility into the data, models trained against the data, and insights uncovered within the data is critical.

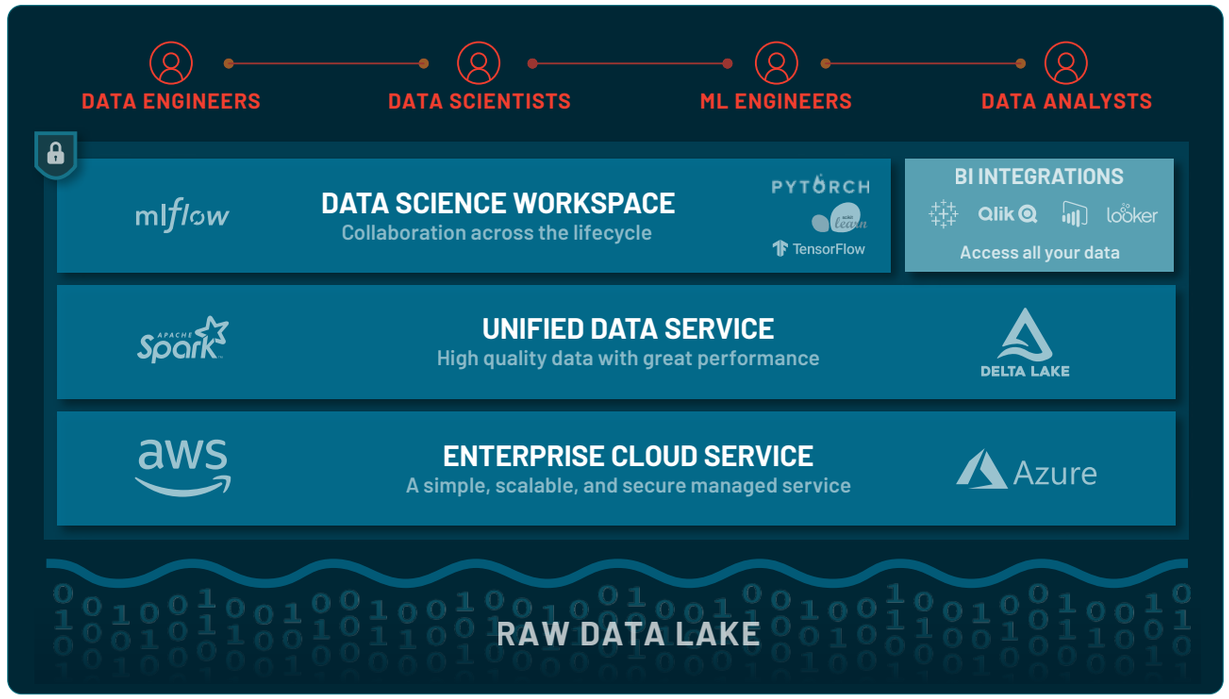
Better data. Better data science.

With so many data challenges acting as a brake on innovation — distracting data teams from their core competencies and slowing innovation and insights — a new approach is clearly needed.

By unifying data science, engineering and business, Databricks delivers a Unified Data Analytics Platform that accelerates innovation. Founded by the original creators of Apache Spark™, Delta Lake, MLflow, and Koalas, Databricks delivers a Data Science Workspace that takes traditional notebook environments to the next level.

By integrating and streamlining the individual elements that comprise the analytics lifecycle, data scientists can quickly access data, provision compute resources, and work together to build and manage ML models across their full lifecycle, creating a culture of accelerated innovation.

With Databricks, data scientists can perform all analytics in one place on reliable data, from exploratory data science to building and operationalizing state-of-the-art machine learning models as a team.



Better data. Better data science.

Focus on your data, not DevOps

Databricks' serverless and highly elastic cloud service is designed to remove operational complexity while ensuring reliability and cost efficiency at scale, so teams can focus on the data science instead of DevOps. With optimized, highly elastic clusters at their fingertips, optimized Apache Spark, Delta Lake, and ML frameworks, analysts and data scientists can. now explore petabytes of data in real-time.

- **Automated Cluster Management:** Launch expertly-tuned Spark clusters with a few clicks. Databricks Spark clusters are fully managed and automatically scale to your workload.
- **Optimized for Performance:** Built on top of Spark and native to the cloud, Databricks Runtime optimizes Spark with Delta Lake, making it 10-40x faster and more reliable.
- **Enterprise Security:** Databricks protects your data at every level with a unified security model featuring fine-grained controls, data encryption, identity management, rigorous auditing, and support for compliance standards.
- **Multi Cloud:** The Unified Data Analytics Platform is available on both Microsoft Azure and Amazon Web Services, ensuring maximum flexibility and deep ecosystems integration.

Make data pristine and ready for analytics

Data scientists need to work on a large variety of data forms and formats: small or large data sets, DataFrames, text, images, batch or streaming. All require specific pipelines and transformations. Databricks lets you ingest raw data from virtually any source, merge batch and streaming data, schedule transformations, and perform quality checks to make sure data is clean and ready for further analytics. So practitioners can now trust any data they work with, CSV files or massive data lake ingests, based on business needs.

- **Delta Lake:** Delta Lake brings enhanced reliability, performance, and lifecycle management to Data Lakes. No more incomplete jobs to rollback for clean up, suspect data added into your data lake, or difficulty deleting data for compliance changes.
- **Databricks Runtime:** The Databrick Runtime is a distributed data processing engine built on a highly optimized version of Apache Spark, for up to 50x performance gains. Build pipelines, schedule jobs, and train models with easy self-service and cost-saving performance.

Better data. Better data science.

Enable teamwork with collaborative notebooks

Increase the productivity of your data science team by 4-5x through collaboration and the democratization of data and insights. Databricks provides collaborative notebooks that eliminate 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.

- **Collaborative Notebooks:** Speed up iterative model building and tuning with interactive notebooks purpose-built to instill collaboration across teams.
- **Support for Multiple Programming Languages:** Interactively query large-scale data sets in R, Python, Scala, or SQL.
- **Built-in Visualizations:** Visualize insights through a wide assortment of point-and-click visualizations. 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.

Share insights via 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.

- **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.
- **Dashboard Widgets:** Input widgets allow you to parameterize your dashboards.

Take advantage of state-of-the-art machine learning

Move faster with 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.

- **Frameworks of choice:** Built-in TensorFlow, Keras, PyTorch, MLflow, Horovod, GraphFrames, scikit-learn, XGboost, numpy, MLeap, Pandas, and more, optimized and tested for compatibility.
- **AutoML:** 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.
- **Optimized TensorFlow:** Benefit from TensorFlow CUDA-optimized version on GPU clusters, and Intel MKL-DNN optimized TensorFlow package on Intel CPUs for maximum performance.

Complete data and ML lifecycle

Version tables, track and share experiments locally or in the cloud, package and share models across frameworks, and deploy models virtually anywhere.

- **Data Versioning:** Delta Lake Time Travel automatically keeps track of changes in your data sets, so you always stay in control.
- **Experiments Tracking:** Automatically track, share, compare, and interactively visualize experiments along with their artifacts (data, code, parameters, dependencies, etc...) with MLflow from within your notebooks.
- **Model Registry:** 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.

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 at every level with a unified security model featuring fine grained controls, data encryption at rest and in motion, identity management, rigorous auditing, and support for compliance standards like HIPAA, SOC 2 Type II, and ISO 27001.

Lower the total cost of ownership

When adopting new technologies all vendors promise to lower total cost of ownership, but often these can be empty promises. Databricks stands behind the lowered TCO claim with an enterprise cloud service that means no expensive hardware; an operationally simple platform designed to help you efficiently manage your costs; increased productivity through seamless collaboration; support for familiar languages like SQL, R, Python, and Scala; and faster performance than other analytics products — which allows you to process and analyze data, resulting in a shorter time to value.

The following pages take a deeper dive into how Databricks collaborative notebooks compare to today's common open-source notebooks.

Better data. Better data science.

FEATURE	DATABRICKS
Multi-Language Support	Enables commands across Python, Scala, SQL, or R (including markdown) in the same notebook — allowing users seamlessly to mix and match as needed
Collaboration	Designed for collaboration, Databricks' notebooks contain features such as comments, viewer log, and history
Live Sharing & Editing	Real time collaboration among team members performing data modeling or analysis
Revision History	Simplifies version control by not having to create, save and manage notebook changes made during the development of a Spark application
Run Sidebar	Automatically keep track of all your runs as experiments, including parameters, libraries, config, and output metrics. Revert to the previous version of your code in one click
Experiment Tracking UI	One click access to all previously ran experiments, sort, filter, query, and compare to find the best performing runs
GitHub and Bitbucket Integration	Enables source and version control of a notebook
SparkR Support	Allows data scientists to leverage Spark's distributed processing engine to analyze large scale datasets and interactively run jobs on them from the R shell
No Vendor Lock-in	Users can import and export notebooks in source format, allowing for seamless migration of source code (Scala, Python, R, SQL) in and out of Databricks to use across an IDE of your choice.
Attach Multiple Notebooks to a Cluster	Enables efficient cluster resource management as many users and notebooks can share a single running cluster
Debugging	Easier debugging capabilities for Spark jobs executed from notebooks
Keyboard Shortcuts	Extensive keyboard shortcuts for easy in-notebook development and navigation

Better data. Better data science.

FEATURE	DATABRICKS
REST API	Trigger or query a notebook externally via the Databricks REST API for easy programmatic access
Notebooks Workflows	Ability to invoke notebooks that can invoke each other
Autocomplete	Automatically completes function names and invocations within the notebook
Parameterized Queries	More easily reuse notebooks by utilizing parameterized queries
Extensibility	Make use of popular libraries within your notebook or job such as matplotlib, numpy, pandas, etc
One-Click Visualizations	Provide a wide range of visualizations including full big-data pivoting, histograms, scatterplots, maps, etc
Pipeline Workflows	Data Scientists and Data Engineers can use the same notebook for creating models to production jobs
One Click Publishing from Notebooks	Create shareable dashboards from notebooks with a single click
Continuous Dashboard Updates	Publish dashboards and schedule the content to be updated continuously
Parameterized Dashboards	Provide drop-downs in the dashboards to enable changing input parameters to dashboard values
Schedule Dashboards	Execute jobs for production pipelines on a specified schedule directly from a dashboard
Dashboard Widgets	Input widgets allow you to parameterize your dashboards

Case in point: How leading companies have increased data science productivity with Databricks

CONDÉ NAST

LEARN MORE

Condé Nast is one of the world's leading media companies, counting some of the most iconic magazine titles in its portfolio, including, The New Yorker, Wired, and Vogue. The company uses data to reach over 1 billion people in print, online, video, and social media.

Use case: As a leading media publisher, Conde Nast manages over 20 brands in their portfolio. On a monthly basis, their web properties garner 100+ million visits and 800+ million page views producing a tremendous amount of data. The data team is focused on improving user engagement by using machine learning to provide personalized content recommendations and targeted ads.

Solution and benefits: Databricks provides Conde Nast with a fully managed cloud platform that simplifies operations, delivers superior performance, and enables data science innovation.

- **Improved customer engagement:** With an improved data pipeline, Condé Nast can make better, faster, and more accurate content recommendations, improving the user experience.
- **Built for scale:** Datasets can no longer outgrow Condé Nast's capacity to process and glean insights.
- **More models in production:** With MLflow, Condé Nast's data science teams can innovate their products faster. They have deployed over 1,200 models in production.

> *Databricks has been an **incredibly powerful end-to-end solution** for us. It's allowed a variety of different team members from different backgrounds to quickly get in and utilize large volumes of data to **make actionable business decisions**.*
>
> — Paul Fryzel, *Principal Engineer of AI Infrastructure at Condé Nast*

Case in point: How leading companies have increased data science productivity with Databricks

LEARN MORE

SHOWTIME® is a premium television network and streaming service featuring award-winning original series and original limited series like *Shameless*, *Homeland*, *Billions*, *The Chi*, *Ray Donovan*, *SMILF*, *The Affair*, *Patrick Melrose*, *Our Cartoon President*, *Twin Peaks* and more.

Use case: The Data Strategy team at Showtime is focused on democratizing data and analytics across the organization. They collect huge volumes of subscriber data (e.g. shows watched, time of day, devices used, subscription history, etc) and use machine learning to predict subscriber behavior and improve scheduling and programming.

Solution and benefits: Databricks has helped Showtime democratize data and machine learning across the organization, creating a more data-driven culture.

- **6x faster pipelines:** Data pipelines that took over 24 hours are now run in less than 4 hours enabling teams to make decisions faster.
- **Removing infrastructure complexity:** Fully managed platform in the cloud with automated cluster management allows the data science team to focus on machine learning rather than hardware configurations, provisioning clusters, debugging, etc.
- **Innovating the subscriber experience:** Improved data science collaboration and productivity has reduced time-to-market for new models and features. Teams can experiment faster leading to a better, more personalized experience for subscribers.

> *Being on the Databricks platform has allowed a team of exclusively data scientists to **make huge strides** in setting aside all those configuration headaches that we were faced with. It's **dramatically improved our productivity**.*
>
> — Josh McNutt, *Senior Vice President of Data Strategy and Consumer Analytics at Showtime*

Case in point: How leading companies have increased data science productivity with Databricks

LEARN MORE

Shell is a recognized pioneer in oil and gas exploration and production technology and one of the world's leading oil and natural gas producers, gasoline and natural gas marketers and petrochemical manufacturers.

Use case: To maintain production, Shell stocks over 3,000 different spare parts across their global facilities. It's crucial the right parts are available at the right time to avoid outages, but equally important is not overstocking which can be cost-prohibitive.

Solution and benefits: Databricks provides Shell with a cloud-native unified analytics platform that helps with improved inventory and supply chain management.

- **Predictive modeling:** Scalable predictive model is developed and deployed across more than 3,000 types of materials at 50+ locations.
- **Historical analyses:** Each material model involves simulating 10,000 Markov Chain Monte Carlo iterations to capture historical distribution of issues.
- **Massive performance gains:** With a focus on improving performance the data science team reduced the inventory analysis and prediction time to 45 minutes from 48 hours on a 50 node Apache Spark™ cluster on Databricks — a 32X performance gain.
- **Reduced expenditures:** Cost savings equivalent to millions of dollars per year.

> *Databricks has produced an **enormous amount of value** for Shell. The inventory optimization tool [built on Databricks] was the first scaled up digital product that came out of my organization and the fact that it's deployed globally means we're now **delivering millions of dollars of savings** every year.*
>
> — Daniel Jeavons, *General Manager Advanced Analytics CoE at Shell*

Case in point: How leading companies have increased data science productivity with Databricks

LEARN MORE

Riot Games' goal is to be the world's most player-focused gaming company in the world. Founded in 2006, and based in LA, Riot Games is best known for the *League of Legends* game. Over 100 million gamers play every month.

Use case: Improving gaming experience through network performance monitoring and combating in-game abusive language.

Solution and benefits: Databricks allows Riot Games to improve the gaming experience of their players by providing scalable, fast analytics.

- **Improved in-game purchase experience:** Able to rapidly build and productionize a recommendation engine that provides unique offers based on over 500B data points. Gamers can now more easily find the content they want.
- **Reduced game lag:** Built ML model that detects network issues in real-time, enabling Riot Games to avoid outages before they adversely impact players.
- **Faster analytics:** Increased processing performance of data preparation and exploration by 50% compared to EMR, significantly speeding up analyses.

> *We wanted to free data scientists from managing clusters. Having an easy-to-use, managed Spark solution in Databricks allows us to do this. Now* ***our teams can focus on improving the gaming experience****.*
>
> — Colin Borys, *Data Scientist at Riot Games*

Case in point: How leading companies have increased data science productivity with Databricks

Quby

LEARN MORE

*Databricks, through the **power of Delta Lake and Structured Streaming**, allows us to deliver alerts and recommendations to our customers with a **very limited latency**, so they're able to react to problems or make adjustments within their home **before it affects their comfort levels**.*

— Steven Galsworthy,
Head of Data Science at Quby

Quby is the technology company behind Toon, the smart energy management device that gives people control over their energy usage, their comfort, the security of their homes, and much more. Quby's smart devices are in hundreds of thousands of homes across Europe. As such, they maintain Europe's largest energy dataset, consisting of petabytes of IoT data, collected from sensors on appliances throughout the home. With this data they are on a mission to help their customers live more comfortable lives while reducing energy consumption through personalized energy usage recommendations.

Use case: Personalized Energy Usage Recommendations – Leverage machine learning and IOT data to power their Waste Checker app which provides personalized recommendations to reduce in-home energy consumption.

Solution and benefits: Databricks provides Quby with a Unified Data Analytics Platform that has fostered a scalable and collaborative environment across data science and engineering, allowing data teams to more quickly innovate and deliver ML-powered services to Quby's customers.

- **Lowered costs:** Cost saving features provided by Databricks (such as auto-scaling clusters and Spot instances) has helped Quby significantly reduce the operational costs of managing infrastructure, while still being able to process large amounts of data.
- **Faster innovation:** With their legacy architecture, moving from proof of concept to production took over 12 months. Now with Databricks, the same process takes less than eight weeks. This enables Quby's data teams to develop new ML-powered features for their customers much faster.
- **Reduced energy consumption:** Through their Waste Checker app, Quby has identified over 67 million kilowatt hours of energy that can be saved by leveraging their personalized recommendations.

Conclusion

Founded by the original creators of Apache Spark™, Delta Lake, MLflow and Koalas, Databricks delivers an open, unified platform that accelerates innovation by unifying data science, engineering and business.

With Databricks, data teams can collaboratively perform all analytics in one place, from ETL to exploratory data science and real world machine learning for the most demanding business needs.

Get started with a free trial of Databricks and take your data science to the next level today.

START YOUR FREE TRIAL

Contact us for a personalized demo databricks.com/contact

Gartner

Databricks named a Leader in Gartner's 2020 Magic Quadrant for Data Science and Machine Learning Platforms

GET FULL REPORT

databricks