Accelerating Innovation Through a Unified Approach to Analytics

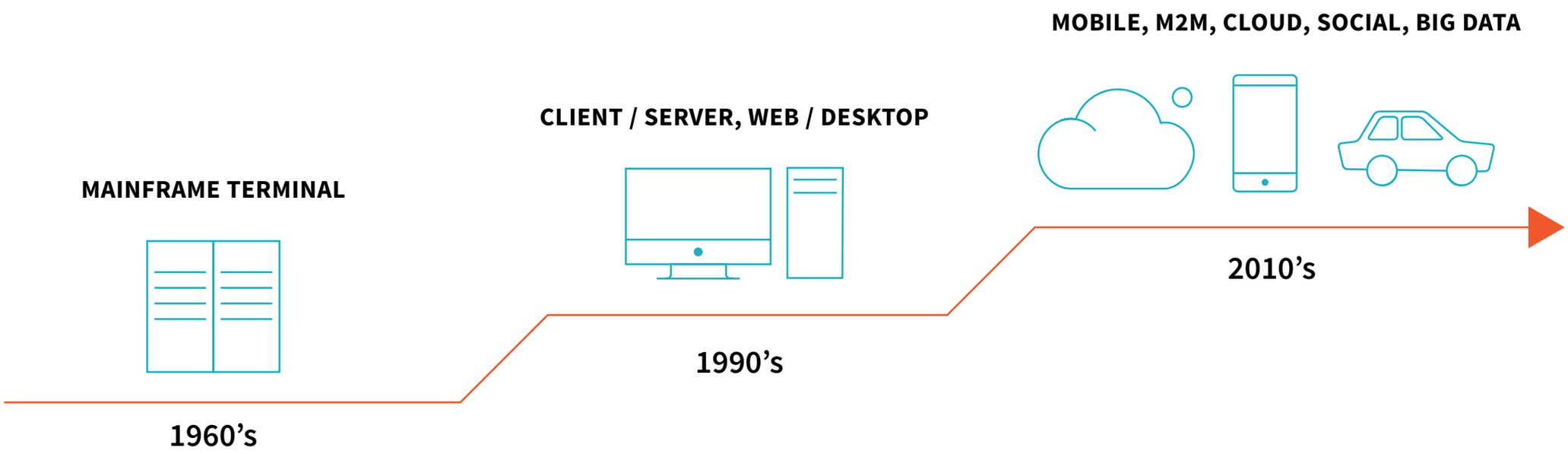
databricks

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, the challenges related to data analysis as a driver of innovation are far more complex and yet very exciting.

Beyond business decision support, a strong analytics platform is capable of supporting change across market verticals, enabling healthcare providers with new drug discovery in the fight to cure cancer, supporting financial institutions with fraud detection to protect consumers, empowering utilities with smart grid technology to improve energy consumption, and many more use cases that impact businesses and users.

Never before has the speed of innovation been more rapid and the goal of business agility more attainable — as long as the right technology is in place that helps avoid pitfalls that can derail successful analytics projects.



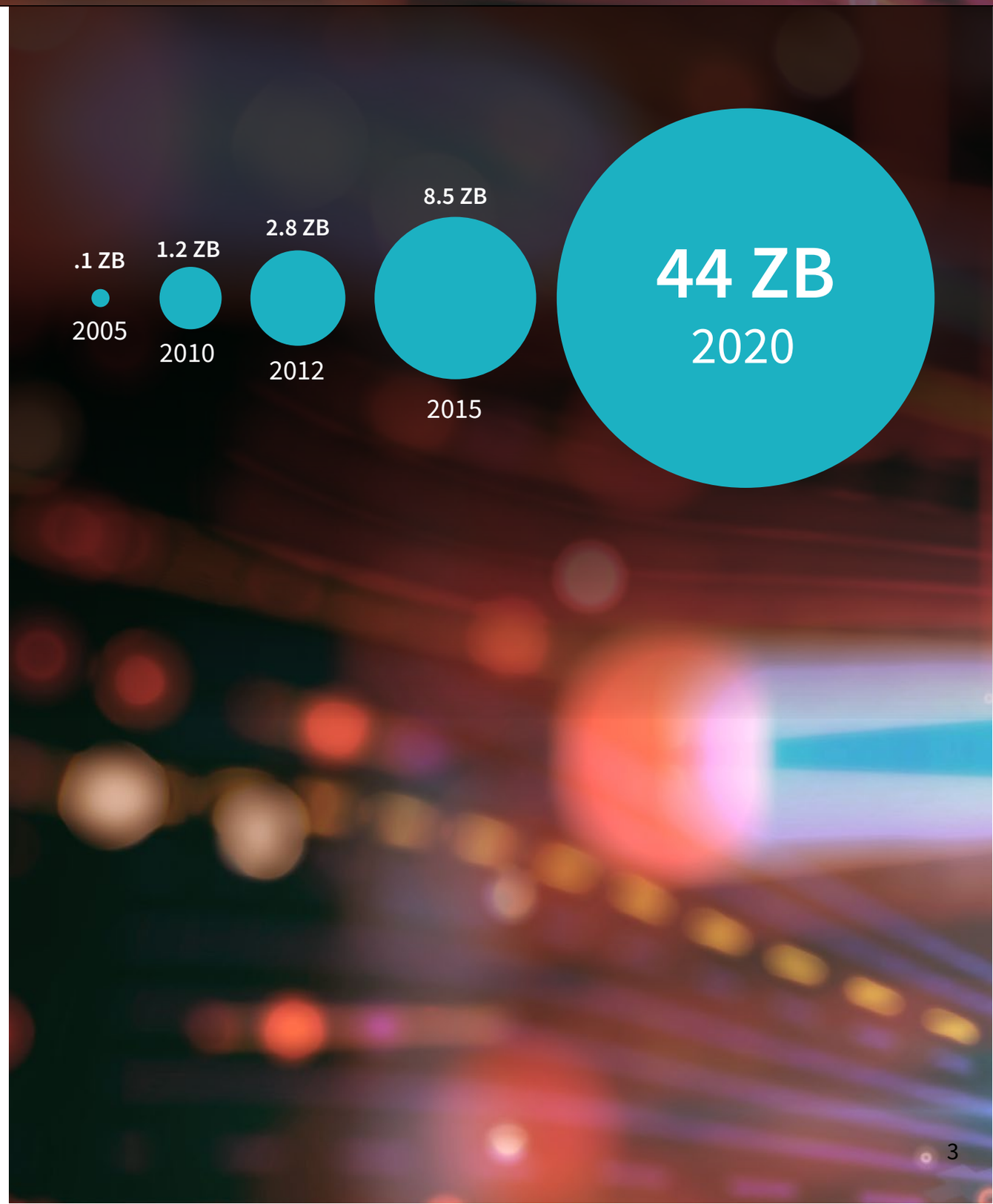
Challenges

Data Growth

Data, and how it is put to use, are key to any business success. At issue is that data volumes are increasing in an almost vertical trajectory, are becoming highly distributed, and can come in a variety of formats. According to IDC, global data generation will reach 180 zettabytes by 2025 — up from close to 10 zettabytes today.[1] Capturing and unlocking the power of this data in order to fuel the next phase of innovation is an incredible challenge for any organization and this challenge will not become any easier.

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 inexorable shift is the complexity that surrounds setting up and maintaining a big data infrastructure. The explosion in data growth pushes organizations to move faster with investments in infrastructure that can harness and derive value from this data, but with this speed comes the trap of an over-reliance on DevOps teams to do the heavy lifting and potential vendor lock-in.[2]

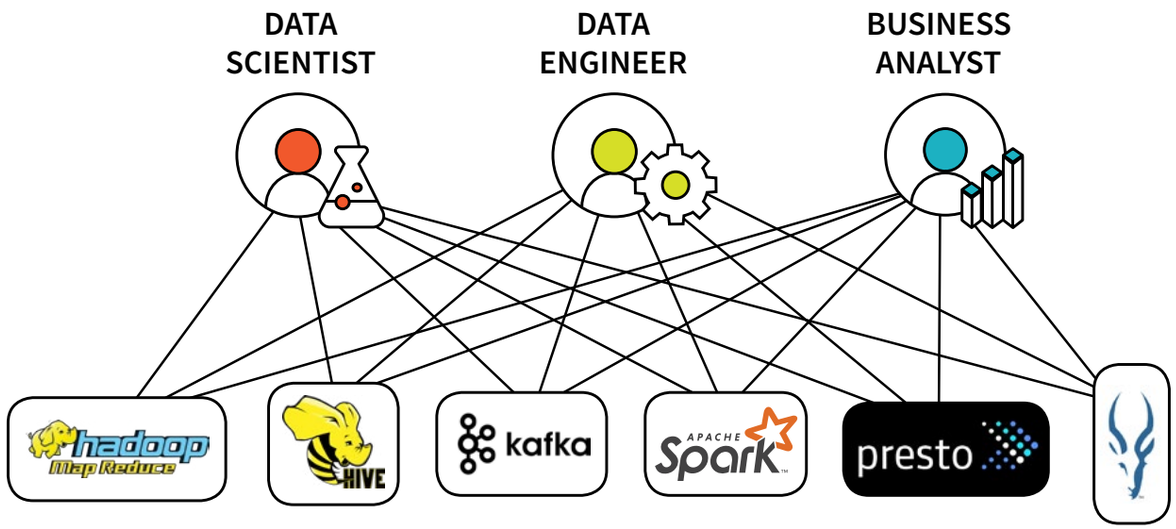


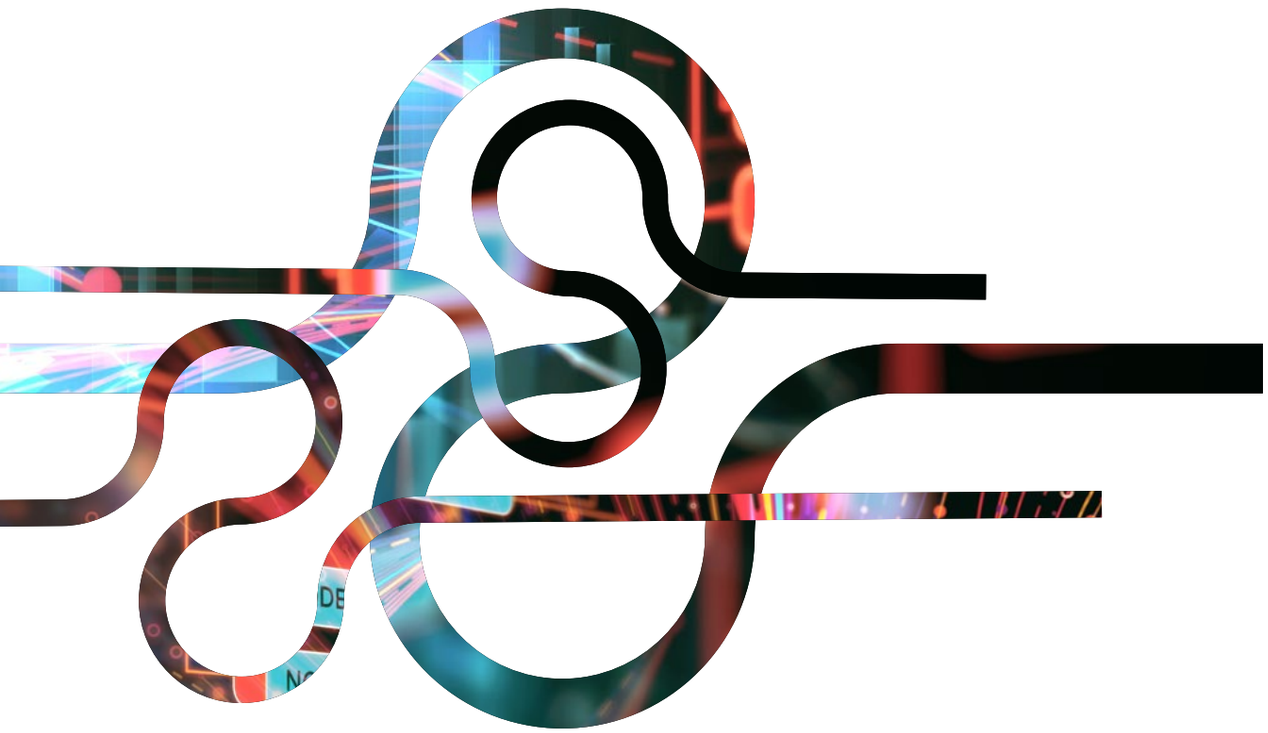
1 https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#198a132b2083
2 https://www.cloudcomputing-news.net/news/2016/sep/01/vendor-lock-in-is-big-roadblock-to-cloud-success-survey-finds/

Challenges

Disparate Technologies

Companies are trying to use a myriad of technologies to achieve their goals of a more data-driven business. Open source projects such as Hive, Presto, Kafka, MapReduce, Impala, and Apache Spark™ offer the promise of a competitive advantage, but also come with management complexity and unexpected costs.[3] Relying on disparate technologies can be incredibly challenging as they all follow different release cycles, lack institutional support mechanisms, and have varying performance deliverables. Additionally, the skills to integrate these technologies are in short supply and can jeopardize innovation in order to maintain a stitched together infrastructure.



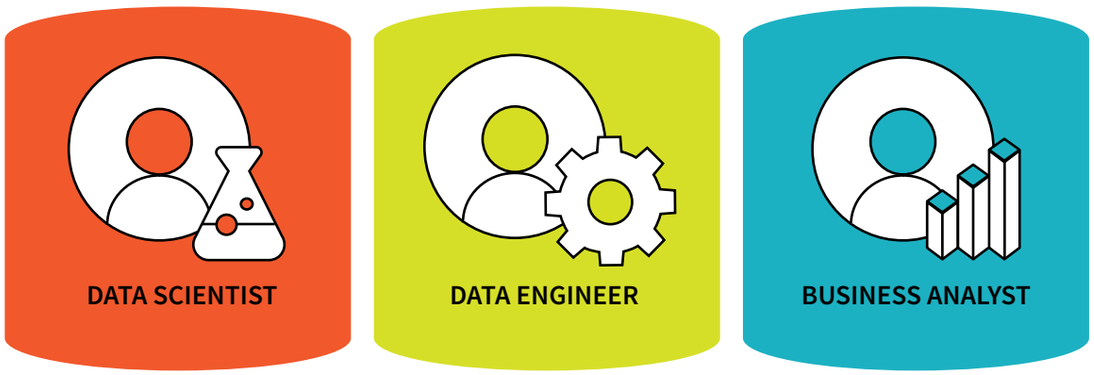


Disjointed Analytics Workflows

One impact of disjointed 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.[4] This greatly reduces the speed of innovation that is the promise of big data, data science, and a move to the cloud.

3 http://www.sungardas.com/en-GB/resources/white-papers/the-cloud-hangover/
4 https://www.forbes.com/sites/janakirammsv/2017/02/07/edge-computing-redefining-the-enterprise-infrastructure/#30119bb07549

Challenges



Siloed teams

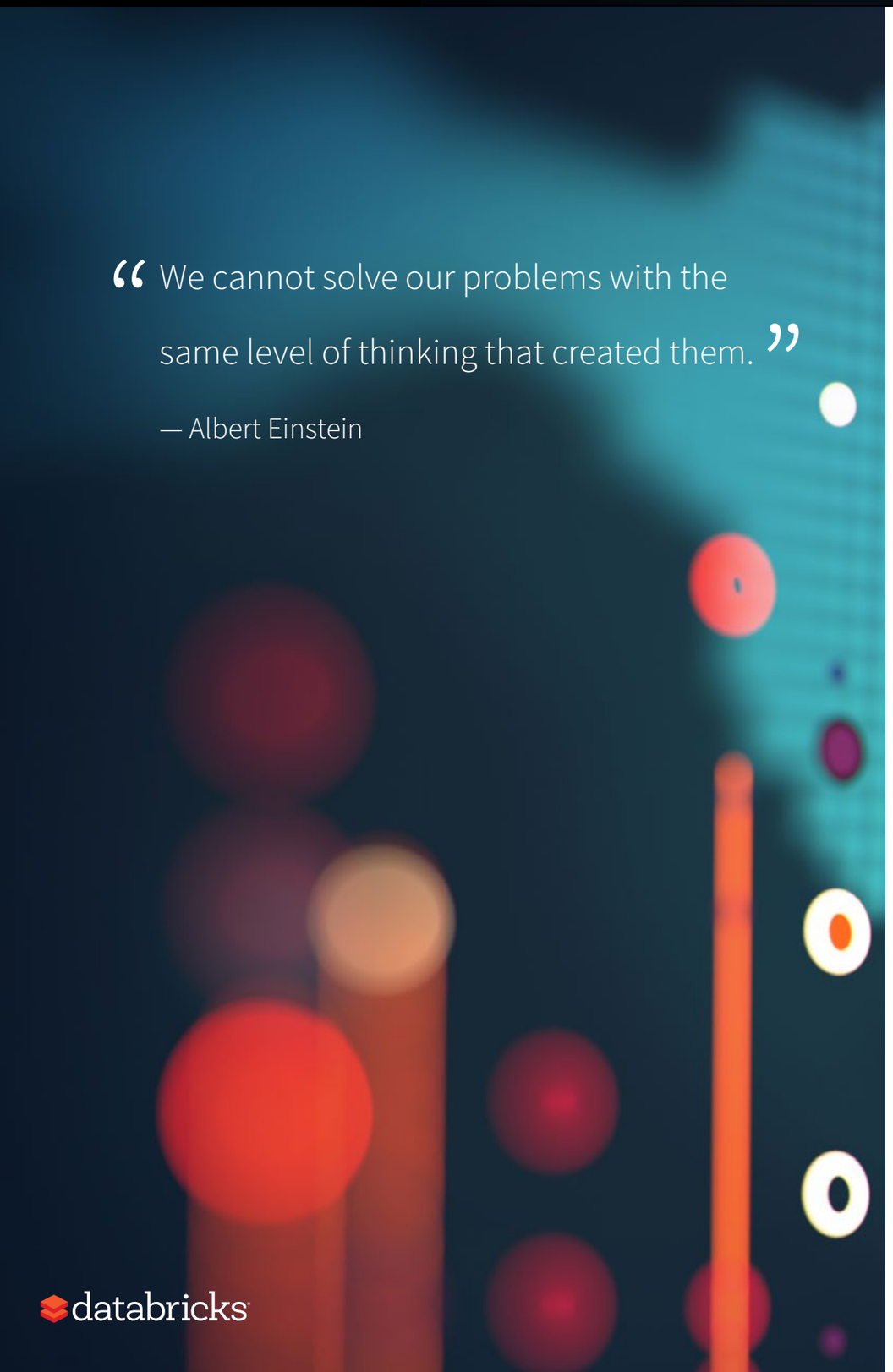
The productivity of the team structured across a data organization can be severely impacted without a seamless and dependable big data platform. It's very difficult for the traditionally siloed functional roles of data scientist, data engineer, and business user to achieve any synergy and work together — both within a function and across teams — to explore data and solve business problems. By viewing data through separate lenses, collaboration is very difficult, trust in the analytics can be misplaced,[5] and speed of innovation is slowed.

Protecting the data

Ironically, even if there is a successful implementation of fragmented technologies allowing organizations to leverage the value of their data, ensuring that the data itself is secure is called into question. Configuring individual technologies so that they comply with a cohesive security strategy can max out even the most seasoned security stakeholders. According to Gartner, 80 percent of organizations will fail to develop a consolidated data security policy across silos, leading to potential noncompliance, security breaches and financial liabilities.[6] The increased number of endpoints that need to be secured in this splintered infrastructure makes protecting the most valuable asset of the business incredibly challenging. But if achieved, a secure foundation can provide the necessary assurances necessary to unlock the possibilities within the data.

5 http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/getting-big-impact-from-big-data
6 http://www.gartner.com/newsroom/id/2758717

The Need for a Unified Approach



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

With data as the fuel for innovation, the modern era’s enterprise requires a comprehensive, unified approach to analytics. This approach should enable the goals of the organization to become an innovation hub, creating virtuous cycles where developers and data scientists can focus on the data, and collaboration, rather than fighting disparate technologies, and working in silos.

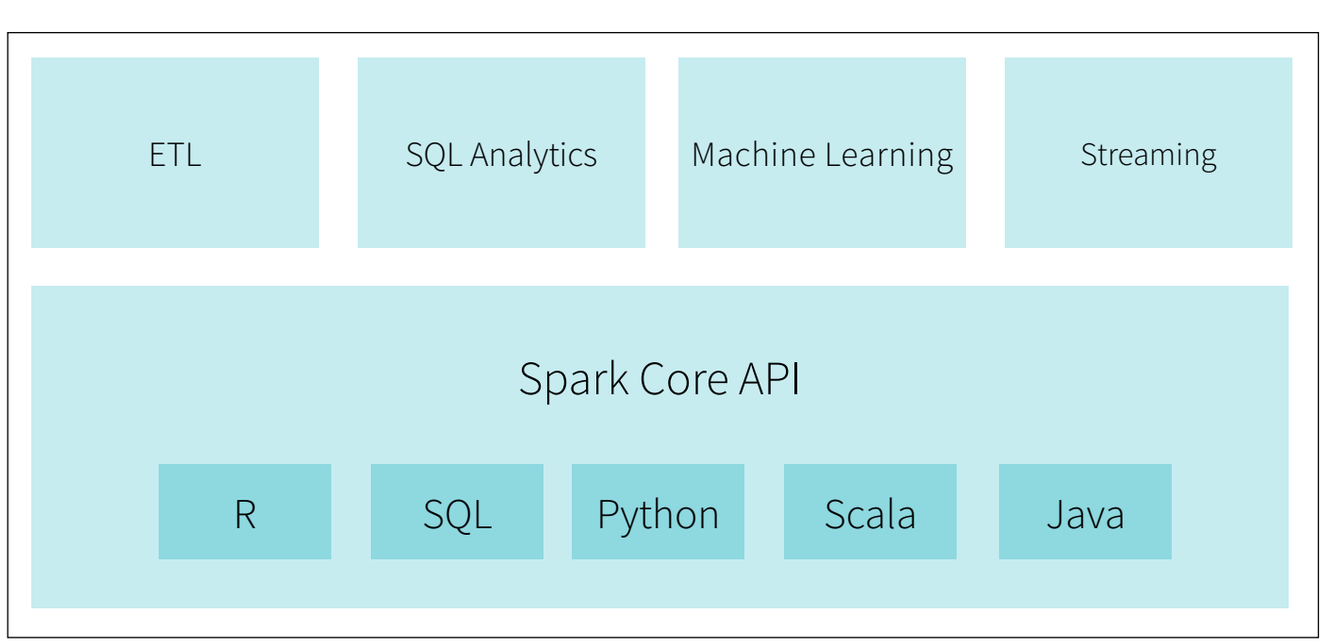
Likewise, engineering teams should be freed from the mundane tasks of maintaining different open source projects. These projects may not work well with one another, introduce unnecessary security risks, and become outdated quickly. The engineering team should instead be able to focus on the important mission of ensuring optimal performance of the customer-facing applications that drive revenue for the business.

Databricks provides the ideal solution to these challenges by providing a platform that unifies data engineering, data science, and the business. Powered by Apache Spark, the Databricks Unified Analytics Platform empowers teams to be truly data-driven to accelerate innovation and deliver transformative business outcomes.

The Databricks Advantage: Apache Spark

Unify Analytics with Apache Spark

To avoid the problems associate with siloed data and disparate systems handling different analytic processes, enterprises need Apache Spark. Spark is the de facto standard for big data processing and analytics that can handle any and all data sources, whether structured, unstructured, or streaming. Additionally, the unified system is agnostic to whether data is fed from the cloud or on-premises, enabling teams to extract valuable insights, and build performant models to fuel innovation.



Apache Spark

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 (1000+ contributors from 250+ orgs)
- Trusted by some of the largest enterprises (Netflix, Yahoo, Facebook, eBay, Alibaba)
- Databricks contributes 75% of the code, 10x more than any other company
- Over 365,000+ Meetup members around the world

The Databricks Advantage: Performance

Faster Performance — Databricks Runtime Built on Apache Spark

For Data Engineers, it's critical to process data no matter the scale as quickly as possible. Databricks has taken 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.

DATABRICKS RUNTIME = 5X **FASTER** than vanilla Spark on AWS

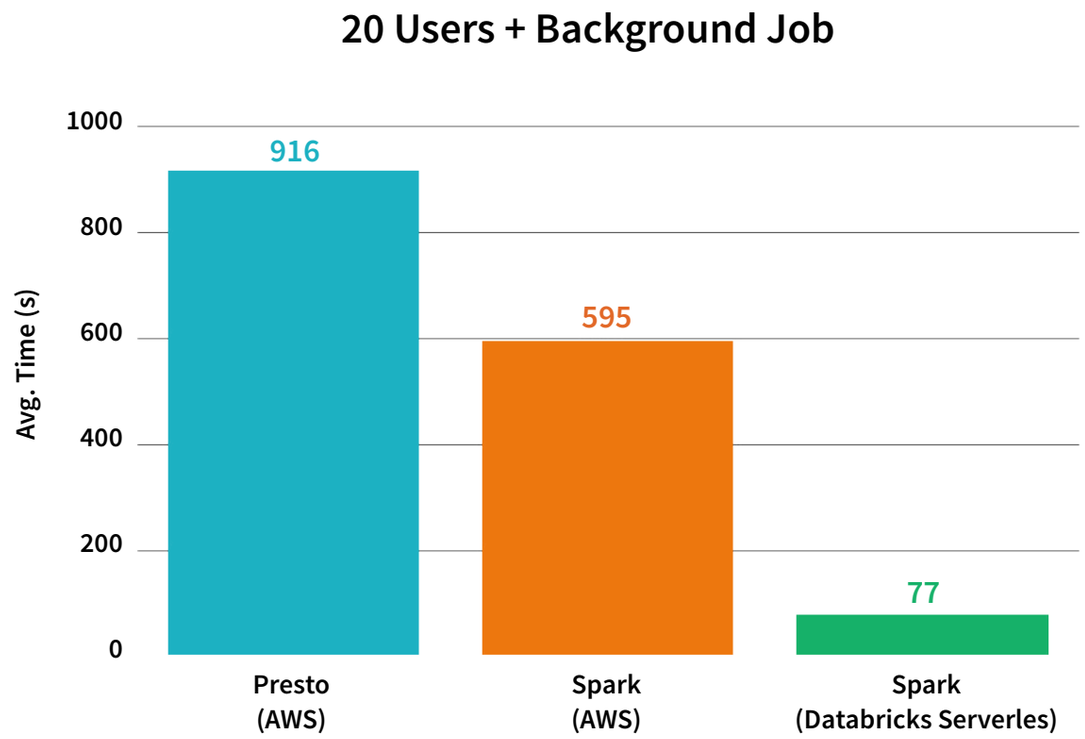
> "We chose Databricks over Hadoop-based alternatives because it is a unified cloud-based big data processing platform that is built on top of Apache Spark, combining the fast performance and standard libraries of Spark with a user-friendly interface that fosters collaboration across our teams."
>
>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

The Databricks Advantage: Infrastructure

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 complex data infrastructure. This is thanks to Databricks' serverless, fully-managed, and highly elastic cloud service. And because Databricks has the industry's leading Spark experts, the service is fine-tuned to ensure ultra-reliable speed and reliability at scale.

Data scientists no longer have to wait for an infrastructure team to provision and configure hardware for them, but instead, can be up and running in minutes, and focusing on building models and finding patterns in data that fuel innovation and accelerate time to market for transformative business outcomes.



> "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: Collaboration

Work Better Together — Become a Heroic Team

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

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, versus weeks or months with an older approach.

Technology is nothing without people to make it great, and Databricks ensures a team can become heroes, by providing a common interface and tooling for all stakeholders, regardless of skill set, to collaborate with one another. This eliminates data silos and frees teammates to focus on what they do best, which in turn benefits their organization and increases innovation.

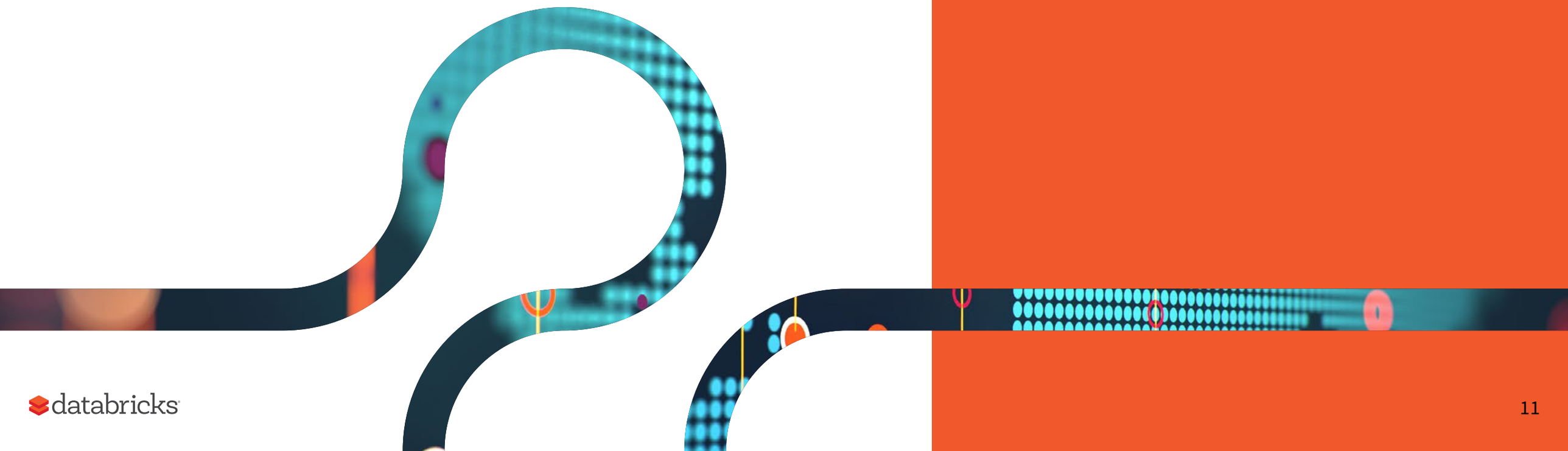
The Databricks Advantage: Workflows

Streamline End-to-End Workflows

With Databricks, each team is empowered to focus on their core competencies. No matter the analytic workload, the engineering team can focus on data preparation and productionizing the models that the data scientists build through a common, unified framework. This entails accessing and ETLing the data as well as monitoring and debugging issues to ensure the highest levels of reliability.

Data science teams can leverage the same platform to explore and visualize data interactively. Building machine learning models is simple with collaborative notebooks that allow data scientists to build and train machine learning models at scale. And interactive dashboards enable all data practitioners to publish insights to distribute across the company to business analysts and decision makers alike.

> “Databricks has allowed us to focus on data science rather than DevOps. It’s helped foster collaboration across our data science and analyst teams which has impacted innovation and productivity.”
>
> — John Landry, Distinguished Technologist, HP Inc.



The Databricks Advantage: Security & TCO

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.

Lowering 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 a cloud-native unified platform 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.

Customer Proof Points: Hotels.com

Hotels.com

The Challenges

- **Build a more robust and faster data pipeline:** On-premises Hadoop cluster using SQL and SAS to do data science at scale was slow and limiting — taking 2 hours to process the data pipeline on only 10% of the data.
- **Leverage machine learning to drive consumer experience:** Massive volume of image files corresponding to each property listing included duplicates and lacked organization for ranking and classification. Needed to build in real-time scoring and become more efficient at deploying machine learning models into production.
- **Increase customer conversions:** Being able to understand customer trends in real-time to develop strategies to drive conversion and lifetime value.

The Solution

Databricks has helped Hotels.com to realize its goal of becoming 'data science focused' so that they can anticipate customer behavior and provide a more optimized user experience.

- **Cluster Management:** Able to scale volume of data significantly without adding infrastructure complexity.
- **Interactive Workspace:** Foster a culture of collaboration among data science teams within Hotels.com as well as other business units within Expedia.
- **Databricks Runtime:** Increase processing performance of streaming data even at scale.

Results

- **Accelerate ETL at scale:** Able to increase the volume of data processed by 20x without impacting performance.
- **Optimized user experience:** Highly accurate and effective display of images within the context of property searches by customers.
- **Increased sales efficiency:** Providing the right hotel with the right images based on searches has resulted in higher conversions.

> "Agility and flexibility were critical for us to successfully support our data science and engineering goals. Moving to Databricks' Unified Analytics Platform to run 100% of our workflows has been a huge boost for our business and our customers."
>
> — Matt Fryer, VP, Chief Data Science Officer, Hotels.com

Customer Proof Points: Viacom

VIACOM

The Challenges

- **Improving user experience:** Streaming petabytes of video data across the world puts a strain on the delivery systems, resulting in videos failing to load or constantly stuttering as they rebuffer.
- **Growing the audience:** Making sense from huge troves of viewing data and determining the best actions to drive viewer retention and loyalty.
- **Targeted advertising:** With TV ad sales falling in recent years, Viacom needed to find better ways to engage with their audience via advertising.

The Solution

Viacom has built a real-time analytics platform based on Apache Spark and Databricks, which constantly monitors the quality of video feeds and reallocates resources in real-time when needed.

- **Cluster Management:** Able to scale volume of data significantly without adding infrastructure complexity.
- **Interactive Workspace:** Foster a culture of collaboration among data science, engineering, and the business.
- **Databricks Runtime:** Increase processing performance of streaming data even at scale.

Results

- **Predict trends and issues to provide superior viewing experience:** Reduced video start delay by 33%.
- **Increase customer loyalty:** Leveraged data to identify how to increase customer retention by 3.5-7x.
- **Improve ad conversions:** Targeted customers with personalized ads based on comScore ratings and viewing behavior.

> “Databricks lets us focus on business problems and makes certain processes very simple. Now it’s a question of how do we bring these benefits to others in the organization who might not be aware of what they can do with this type of platform.”
>
> — Dan Morris, Senior Director of Product Analytics, Viacom

Customer Proof Points: HP

The Challenges

- **Technical architecture:** Zero new budget for big data initiatives due to the company split; lots of money had been spent on old data centers and data warehousing technologies and tooling.
- **Data integration:** Data stored in multiple systems, in multiple places, that was hard to access and analyze.
- **Improved customer experience:** Needed to figure out a new way to program analytics that was much more efficient, and a workflow that could use internal analysis of data that could be turned into customer facing services.

The Solution

Databricks provides HP Inc. with a scalable, affordable, unified analytics platform that allows them to predict product issues and uncover new customer services:

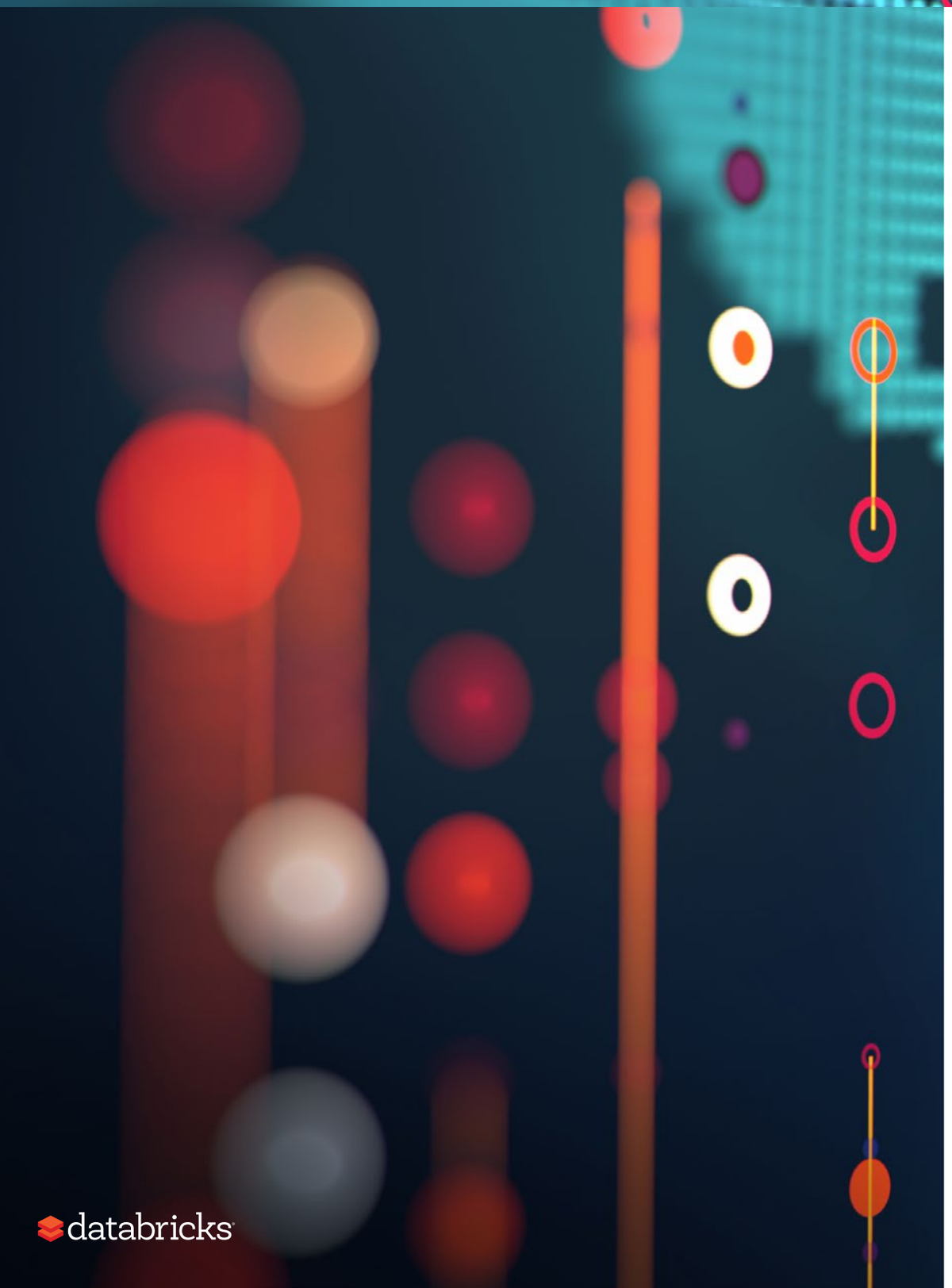
- **Fully managed, cloud-native:** Able to streamline the provisioning of infrastructure and scale to meet the growing volume of data without adding complexity.
- **Interactive Workspace:** Foster a culture of collaboration among data science teams accelerated productivity and ability to build and train predictive models.
- **Databricks Runtime:** Increase processing performance of streaming data even at scale.

Results

- **Simplified management of Apache Spark:** Turnkey unified platform removed infrastructure complexity and empowered the data scientists to focus on the data and not DevOps.
- **Analysis of high volume data in flight:** Analyze data from more than 20 million devices with a seamless workflow.
- **Improved collaboration and productivity:** a unified platform for business analysts, data scientists, and subject matter experts

> “Databricks’ Unified Analytics Platform has helped foster collaboration across our data science and engineering teams which has impacted innovation and productivity.”
>
> — John Landry, Distinguished Technologist, HP, Inc.

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