

# A Guide to Data Science at Scale

---

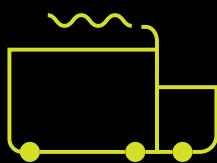
Unifying Big Data and AI



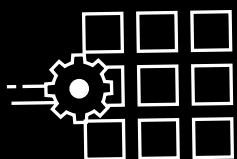
Trends driving innovation in big data and AI, the challenges this creates for enterprises, and insights on how to overcome these obstacles with a unified approach to analytics.



PRODUCT RECOMMENDATIONS



SMART LOGISTICS



NEW PRODUCT DEVELOPMENT



CUSTOMER LOYALTY

# 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 simple decision support, a modern scalable analytics platform is capable of driving monumental change in an organization. Advancements in AI and machine learning have enabled early adopters of big data to unlock new business models, grow revenue streams and deepen customer relationships.

## Leading companies across industries are using big data and AI to drive a broad range of innovative use cases:

### **Product Recommendations**

Use rich customer profiles and machine learning to recommend next best offers and products to drive higher customer conversion.

### **New Product Development**

Aggregate customer, market trend, social media and other sources of data to identify new product innovations and reduce time to market.

### **Smart Logistics**

Analyze mountains of transactions and sensor data to improve supply chain management and operational efficiency across warehouses, stores and fleets.

### **Customer Loyalty**

Holistically understand the factors that impact loyalty — such as quality of service, pricing and product features — to improve the customer experience and reduce churn.

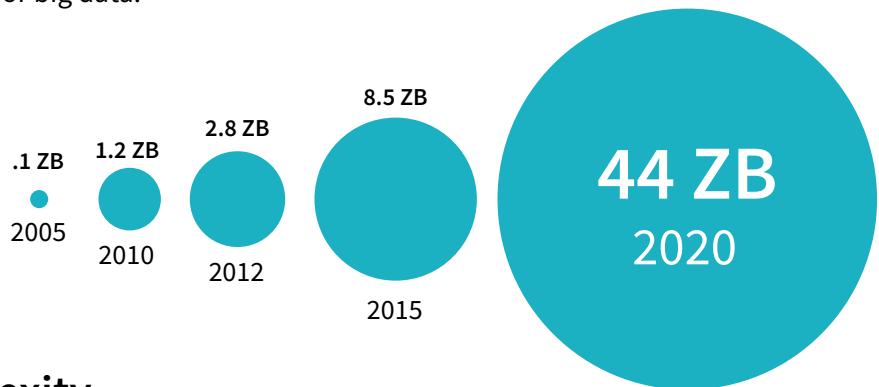
While the promise of big data and AI has never been more achievable, taking this dream and putting it into practice has never been more challenging. The success of your analytics projects hinges on knowing the challenges that lay ahead, avoiding common pitfalls and choosing the right technologies that can scale with your business.

# The Challenges

**There are six primary obstacles limiting the ability of companies to adopt and scale analytics and AI.**

## 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.<sup>1</sup> Capitalizing on the promise of big data to fuel the next phase of innovation is an incredible challenge for any organization. Exploring data at scale and building models in real-time requires on-demand compute power and elastic infrastructure that is built for big data.



## Infrastructure Complexity

The move to the cloud is fast becoming a primary objective for businesses looking to reduce costs and scale their analytics. Part of the challenge associated with this inexorable shift is the complexity that surrounds setting up and maintaining a big data infrastructure. This creates challenges for both DevOps and data scientists alike. Data infrastructure teams are tasked with connecting and managing patchwork cloud technologies while data scientists are left trying to figure out how to spin up resources and access their data across hard to navigate cloud tools.<sup>2</sup>

<sup>1</sup> Press, Gil (2017, January 20). 6 Predictions For The \$203 Billion Big Data Analytics Market. Retrieved from <http://forbes.com>

<sup>2</sup> Logicworks (2016, September 1). Why Vendor Lock-in Remains a Big Roadblock to Cloud Success. Retrieved from <http://cloudcomputing-news.net>

<sup>3</sup> The Cloud Hangover. Retrieved from <http://sungardas.com>

## Disparate Technologies

Companies are trying to use a myriad of technologies to achieve their goals of becoming a more data-driven business. 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.<sup>3</sup>

Adding to the challenge is the need to provide analysts and data scientists with support for the scripting languages (e.g. R, Python, Scala, or SQL) 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.



## **Disjointed Analytics Workflows**

One impact of 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.<sup>4</sup> This greatly reduces the speed of innovation that is the promise of big data, data science, and a move to the cloud.

## **Siloed Teams**

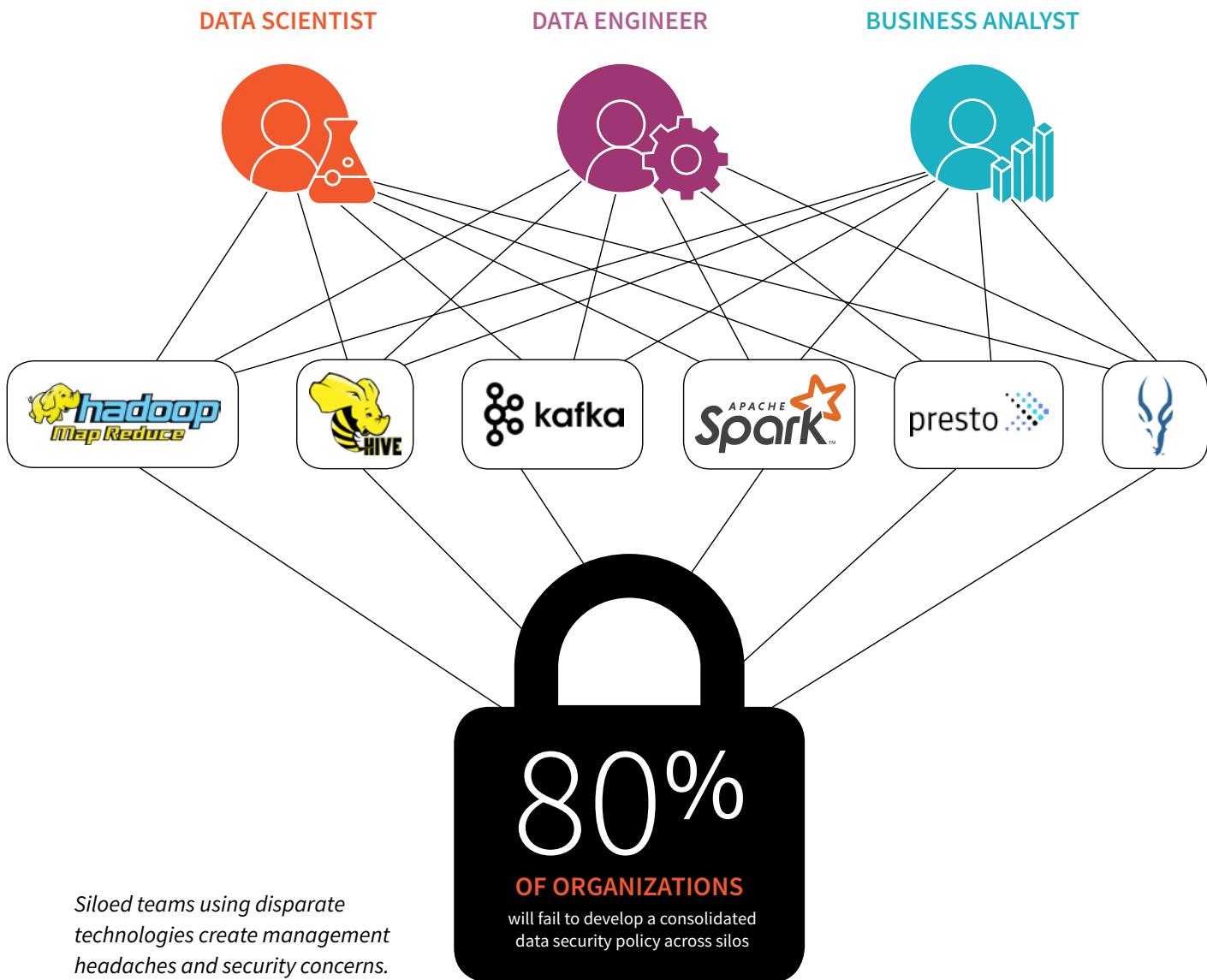
The productivity of the team structured across a data organization can be severely impacted without a seamless and dependable big data platform. It's very difficult for the traditionally siloed functional roles of data scientist, data engineer, and business user to achieve any synergy and work together both within a function and across teams. Few analytics platforms truly promote a collaborative experience. It's not uncommon for a data scientist to build and train a model in a vacuum on their local machine cut-off from their peers and data engineering. The lack of real-time feedback and collaborative capabilities can bring model development and deployment to a snail crawl.

## **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.<sup>5</sup> 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.

<sup>4</sup> MSV, Janakiram (2017, February 7). Edge Computing — Redefining The Enterprise Infrastructure. Retrieved from <http://forbes.com>

<sup>5</sup> Gartner (2014, June 4). Gartner Says Big Data Needs a Data-Centric Security Focus. Retrieved from <http://gartner.com>



# 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, lack enterprise support, 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.

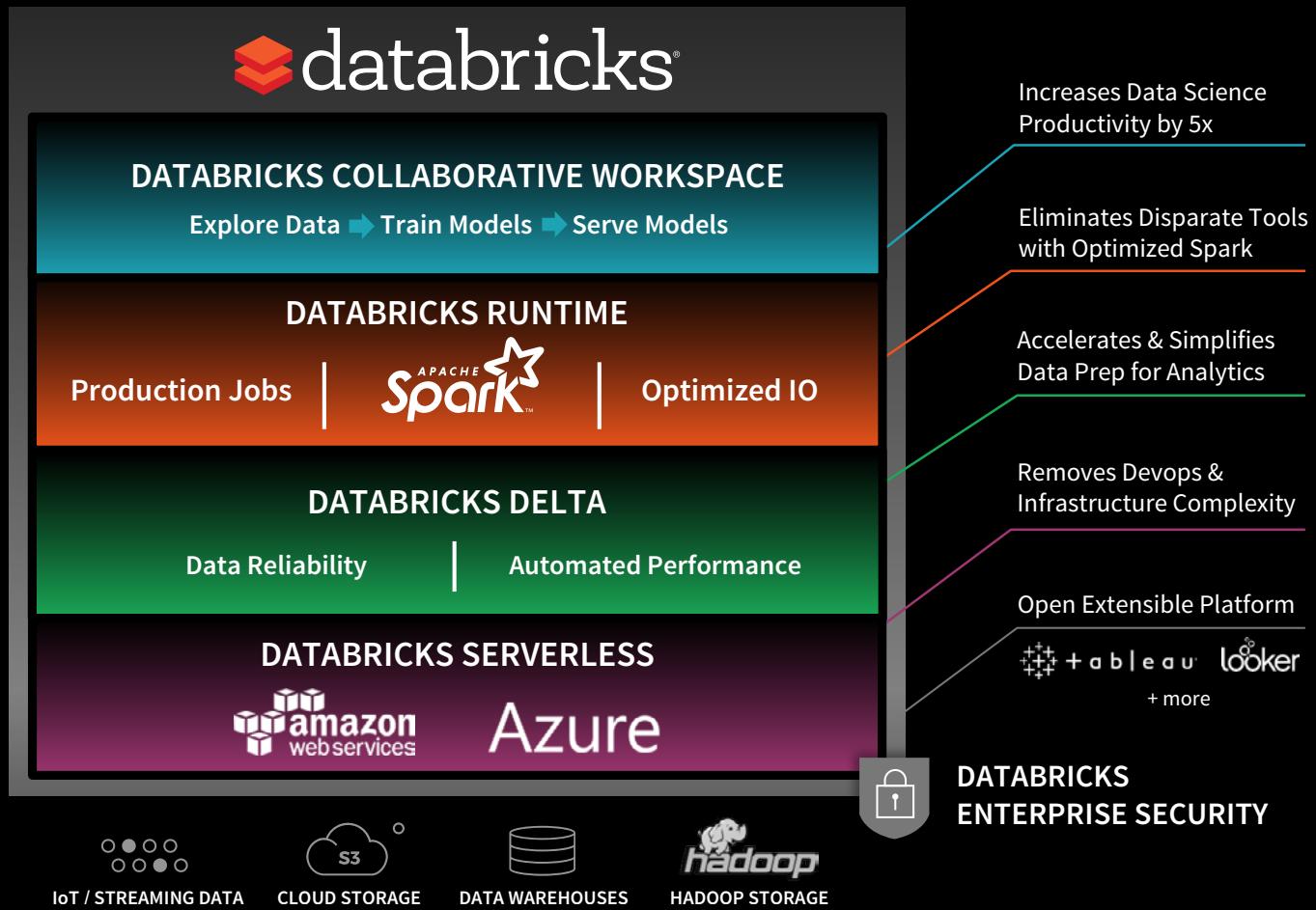
**Databricks provides the ideal solution to these challenges by providing a platform that unifies data engineering, data science, and the business. Powered by Apache Spark, the Databricks Unified Analytics Platform empowers teams to be truly data-driven to accelerate innovation and deliver transformative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

---

“ 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**

# The Databricks Unified Analytics Platform

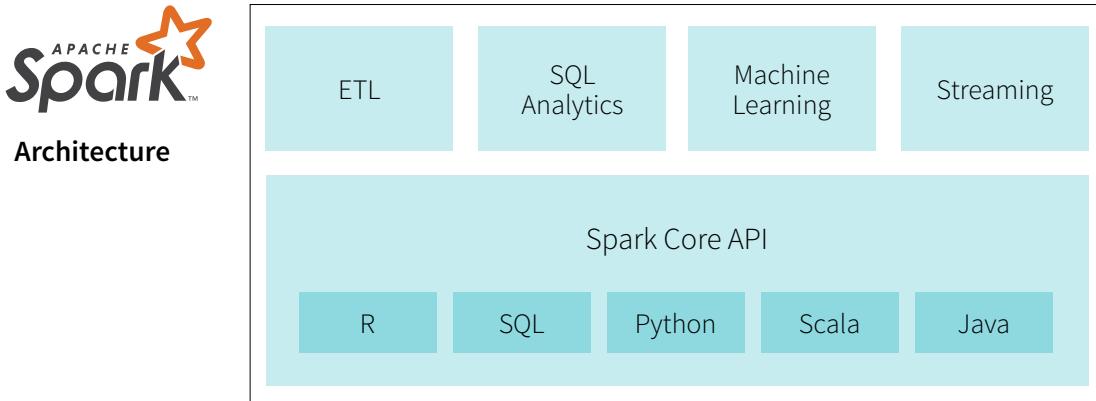


The Databricks Advantage:

# Apache Spark

## Unify Analytics with Apache Spark

To avoid the problems associated 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.



### 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: Simplified 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 cloud optimized 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 so they can focus on what's really important — building models that drive innovation. With optimized, highly elastic Spark clusters at their fingertips, analysts and data scientists can now explore petabytes of data in real-time.

+ Add Cluster

Name	Memory	Type	State	Options
Dedicated ETL cluster	1000 GB	On-demand	Running	Configure Restart Terminate
Exploration cluster - eng	100 GB	On-demand	Running	Configure Restart Terminate
ML cluster - Training	300 GB	On-demand	Running	Configure Restart Terminate

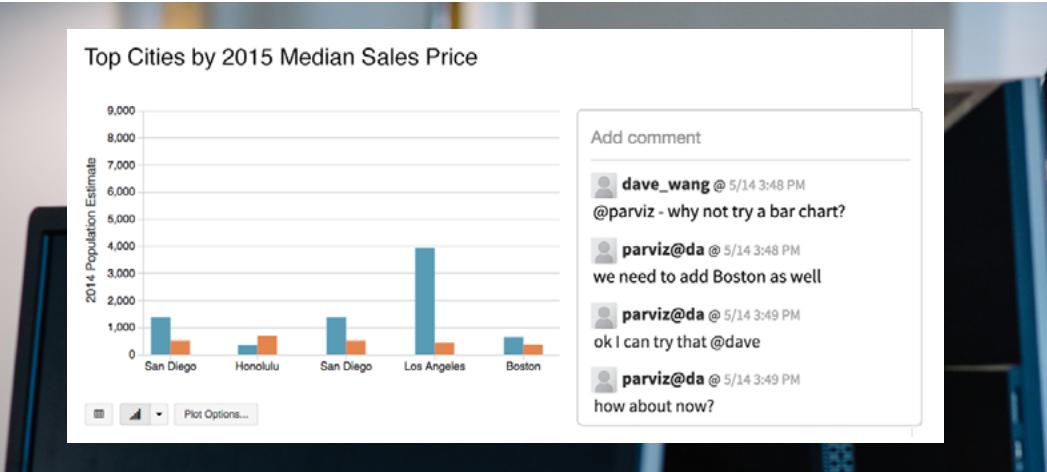
*Launch scalable Spark clusters with a few clicks of a button*

---

**“** 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

# The Databricks Advantage: Collaborative Workspaces



## Work Better Together — Become a Heroic Team

With a unified approach to analytics, data science teams can collaborate using Databricks' interactive workspace. They can use their preferred scripting languages (such as R, Python, Scala, and SQL) and libraries (such as scikit-learn, nltk ML, pandas, etc) to interact with data and build models in a shared notebook environment, and then seamlessly move those models to production with a single click. Collaborative workspaces allow teams to view and edit models in-flight and provide real-time feedback, helping to accelerate innovation.

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.

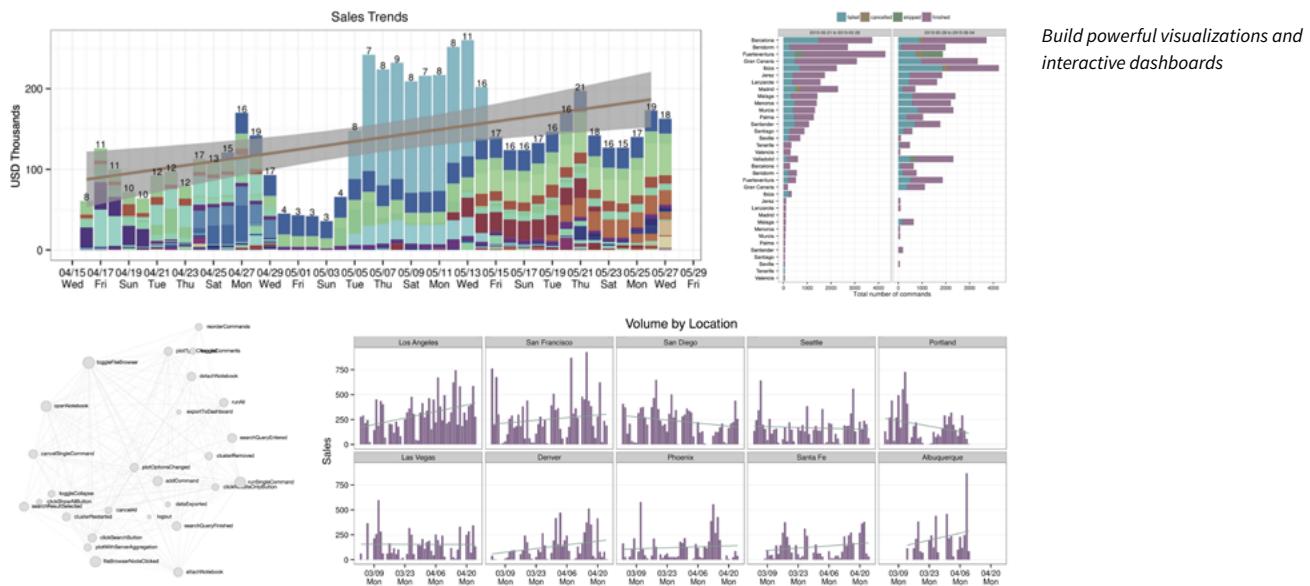
By integrating and streamlining the individual elements that comprise the analytics lifecycle, teams can now build models and test prototypes in hours, versus weeks or months with older approaches.

# The Databricks Advantage: **Unified Analytics**

# **Deliver on the Promise of Big Data and AI**

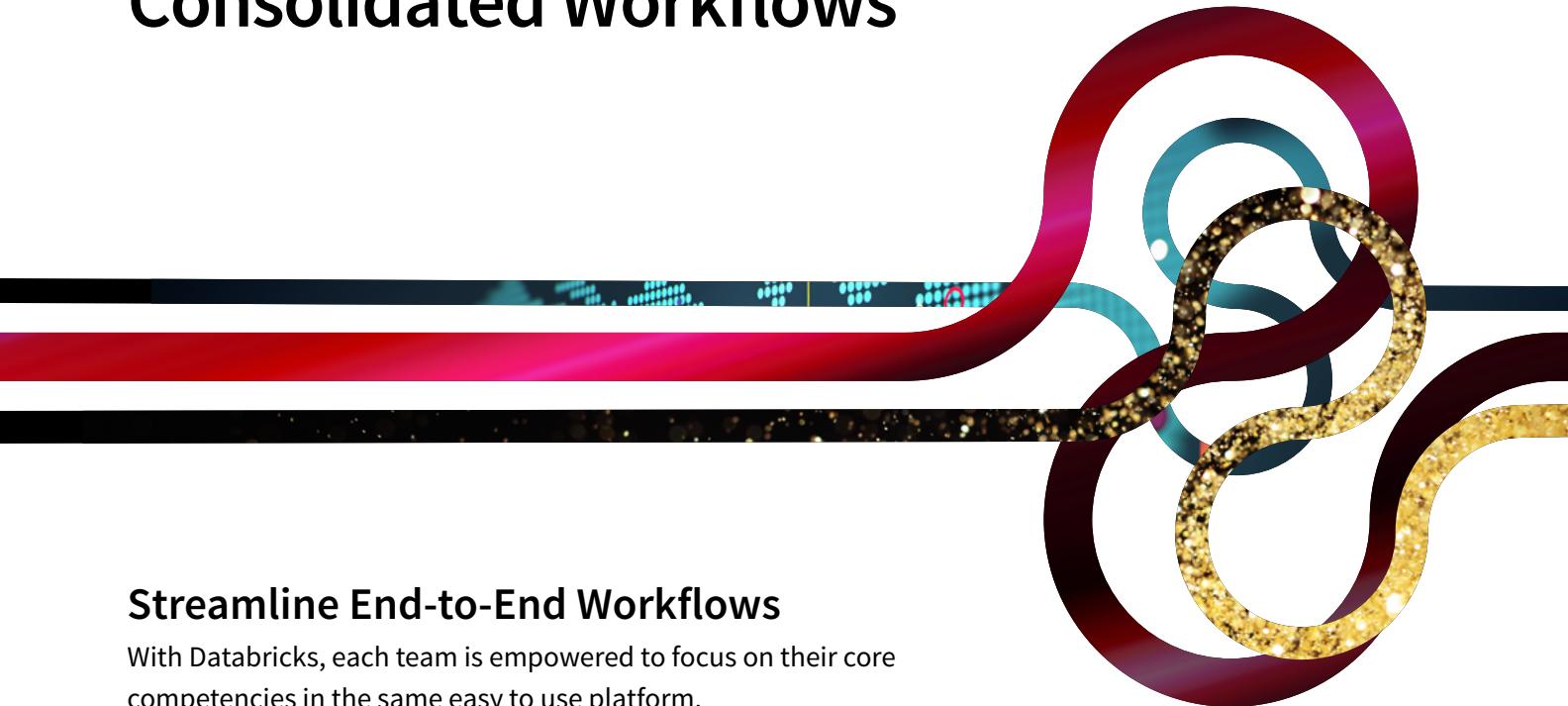
It's time to start delivering on all the analytics needs of the business. With Databricks, data scientists and analysts alike can explore data and seamlessly move across various types of analytics — including batch, ad hoc, machine learning, deep learning, stream processing, and graph — enabling them to meet the evolving needs of the business in one unified platform.

Powerful visualizations can be launched with a few clicks, including a wide range of general-purpose data visualizations (such as bar plots, maps, etc) in addition to visualizations designed for machine learning (such as Matplotlib, GGPLOT, etc) — a must for organizations serious about AI. Visualizations can be turned into interactive dashboards that are easily shared with decision makers across the business. Additionally, Databricks integrates with popular BI tools (including Looker, Tableau, Qlik and Alteryx) so existing investments can be fully leveraged.



The Databricks Advantage:

# Consolidated Workflows



## Streamline End-to-End Workflows

With Databricks, each team is empowered to focus on their core competencies in the same easy to use platform.

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. Data science teams leverage the same platform to explore and visualize data interactively, streamlining analytics workflows.

Building machine learning models is simple with collaborative notebooks that allow data scientists to work together to build and train machine learning models at scale. And interactive dashboards enable data practitioners to publish insights to business analysts and decision makers across the company. Getting from raw data to real business insight has never been easier.

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, SOC 2 Type II, and ISO 27001.

## 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.

“

*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, 



# Customer Story



A recognized leader in oil and gas exploration and production, Shell has operations around the globe. To maintain production, Shell stocks over 3,000 spare parts across their 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.

# The Challenges

- **Disjointed Inventory Distribution:** Stocking practices are often driven by a combination of vendor recommendations, prior operational experience and “gut feeling”.
- **Limited Decision Support System Availability:** There has been limited focus directed towards incorporating historical data and doing advanced analysis to come up with decisions.
- **Lost Business Agility:** This can lead to excessive or insufficient stock being held at Shell’s locations, like oil rigs which has significant business implications

# The Solution

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

- **Databricks Runtime:** The team dramatically improved the performance of the simulations.
- **Interactive Workspace:** The data science team is able to collaborate on the data and models via the interactive workspace.
- **Cluster Management:** Significant reduction in total cost of ownership by moving to the Databricks cloud solution and gains in operational efficiency.
- **Automated Workflows:** Using analytic workflow automation, Shell is easily able to build reliable and fast data pipelines that allow them to predictive when to purchase parts, how long to keep them, and where to place inventory items.

# Results

- **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 Spark cluster on Databricks — a 32X performance gain.
- **Reduced Expenditures:** Cost savings equivalent to millions of dollars per year.



# Customer Story

## Hotels.com

Hotels.com is a premier website for booking accommodations online with 85 websites in 34 languages, listing over 325,000 hotels in approximately 19,000 locations. Hotels.com required massive compute and analytics capabilities to ensure a targeted and satisfying customer experience when booking travel.

“ 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

## The Challenges

- **Leverage machine learning to drive consumer experience:** Massive volume of image files for each property listing included duplicates and lacked organization for ranking and classification. Needed to build 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.
- **Build a more robust and faster data pipeline:** On premise 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.

## 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.

# The Bottom Line

The goal of Databricks' Unified Analytics Platform is to accelerate innovation with scalable analytics and AI. 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\*\*](#) or [\*\*personalized demo\*\*](#).

