databricks

Azure Databricks:

The Best Platform to Run ML and AI

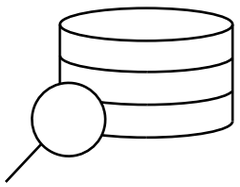
Organizations are looking to analytics to transform their businesses. With the help of concepts such as AI and machine learning, organizations see not only ways to make huge gains in terms of reducing costs, but also transformative changes through new revenue streams.

Yet only 1% of organizations today are able to take advantage of the capabilities of AI. It is the siloed nature of analytics that stifles success. Azure Databricks accelerates innovation by breaking down the silos between people, processes and infrastructure.

This whitepaper explains what makes Azure Databricks unique and how you can use it to transform your business and solve your analytics problems.

Transform your Business with Analytics

Three key elements are needed for a successful analytics program:

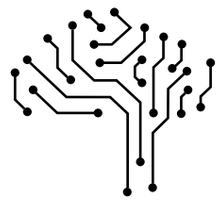


BIG DATA

First of all, the ability to ingest and analyze *all* of your relevant data in your analytics processes is key. Many organizations find they can only access certain data silos, or can only load some of the data for processing. Many organizations find that they can only process a small percentage of their data on a weekly basis, causing them to fall farther and farther behind in the effort to understand what their data is telling them. It takes the right infrastructure to enable access to your insights.

CLOUD INFRASTRUCTURE

Cloud infrastructure is required to make your processes economical. It's especially useful in big data analytics — where large analytics runs spin up and down constantly. With a cloud infrastructure, you have the ability to spin up massive analytics jobs, and then shut them down again, paying for only processing you use. It requires the right processes to scale your analytics, enabling scheduled analytics jobs to run to completion reliably.

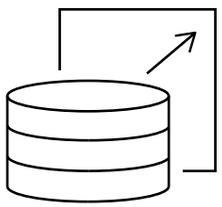


ARTIFICIAL INTELLIGENCE

Realizing the huge innovative leaps that make transformation possible requires the ability to build on the work of others. Data Engineers, Data Scientists, and Business Analysts need to able to collaborate to bring the right business problem and question, the right data set, and the right analytical model in play to answer questions. This requires the ability for people to collaborate quickly and efficiently.

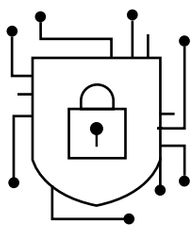
Challenges

Organizations come across fundamental challenges in achieving their analytics goals:



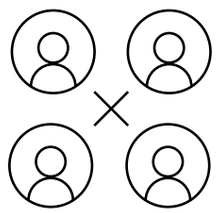
DATA VOLUME

Managing the volumes of data needed to effectively train machine and deep learning models.



SECURITY ASSURANCE

A reliable, secure, and trusted cloud to run your analytics. A lack of cohesive security features can put operations at risk, introduce vulnerabilities and jeopardize compliance.



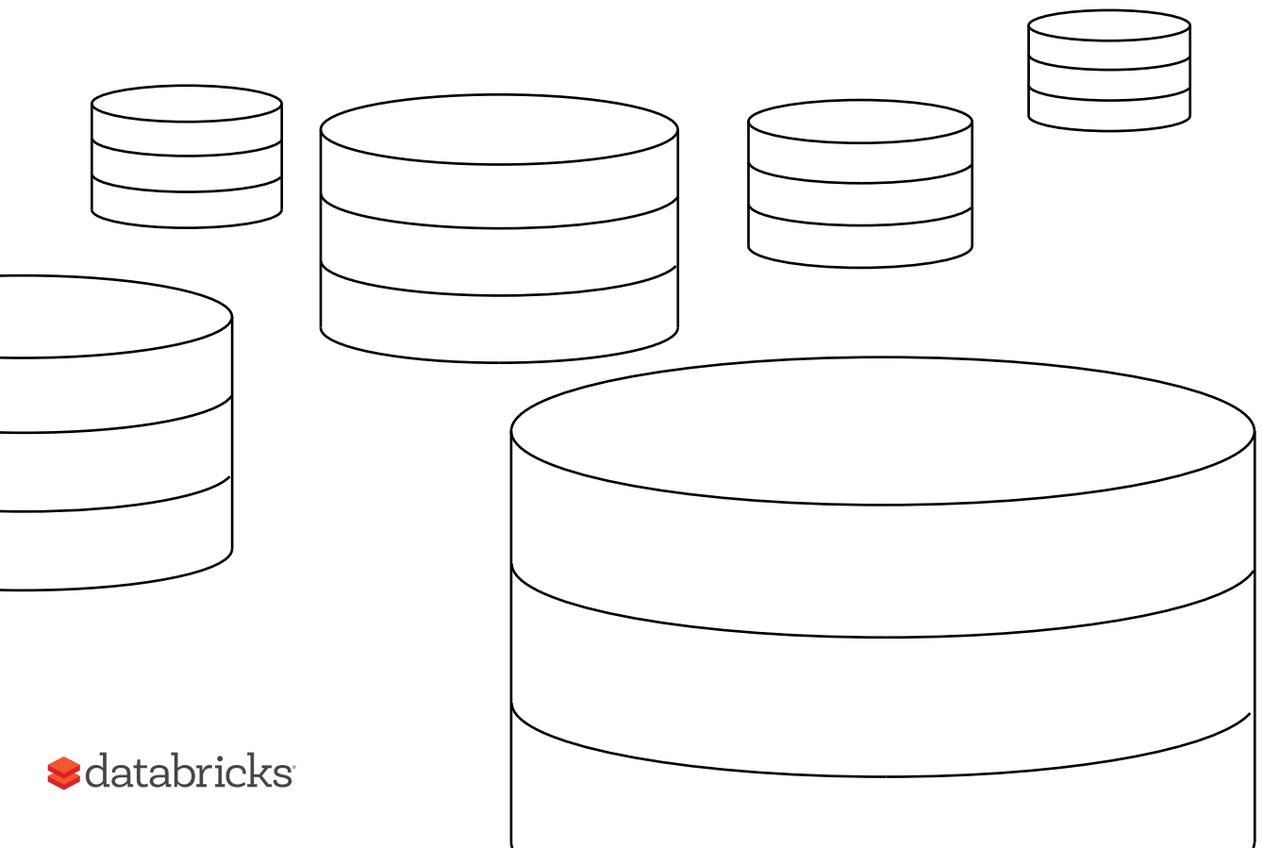
SILOED PROCESS AND PEOPLE

Technology limitations negatively impact productivity and collaboration of data science teams.

Bringing **data** and the **analytics engine** together is the key to this transformation.

Challenge: Data Volume

The flow of data seems never-ending and comes from internal and external sources alike such as line of business or CRM systems, social channels, internet bots, mobile devices and IoT sensors to name only a few.

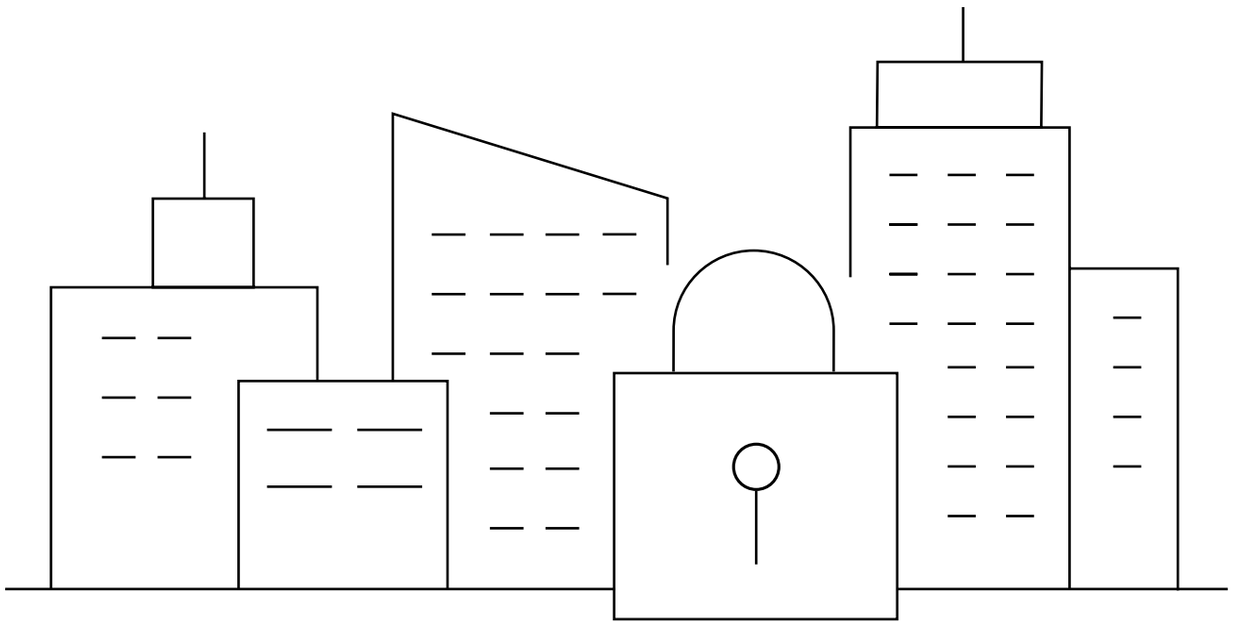


- This “any data, anywhere” can overwhelm the operational capacity of an organization.
- Organizations want to use all of the data to build better analytical models and provide the context for decisions.
- AI systems need large volumes of data to test and refine models.
- In some cases, processing this volume of data for analytics takes weeks to analyze just a fraction of the data, which puts the organization further and further behind.

Challenge: Security Assurance

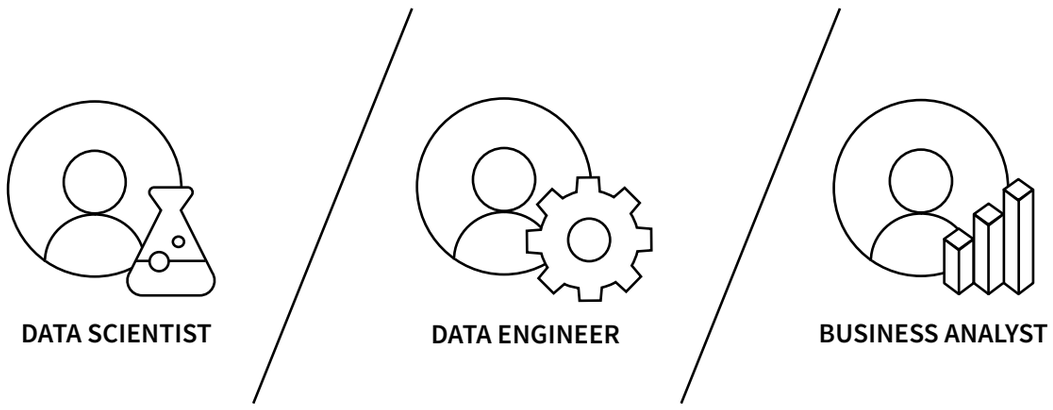
Keeping data safe is critical to an organization's reputation. Without the appropriate security infrastructure, threats and vulnerabilities proliferate and the integrity of the entire company can become compromised.

- User authentication and security policy enforcement become a performance bottleneck and stifle collaboration.
- Oversight of user activity can become resource intensive, negatively impacting productivity.
- Hindered productivity can often lead to users bypassing processes, which can introduce more threats and vulnerabilities.



Challenge: Siloed Process and People

Data Scientists use tools they are familiar with to create models and run analytics. Data Engineers use a different set of tools to blend and clean data.



- Fragmented workflows can create massive inefficiencies for big data and artificial intelligence initiatives. A lack of process automation from data ingestion to production can greatly reduce the speed of innovation
- Siloed analytics kills the productivity of data science teams. It's very difficult to explore data, train AI models and solve business problems with a disjointed analytics platform. The very speed of innovation requires the team to pivot — as it understands the data, the question that is being pursued morphs and changes, requiring fast collaboration to keep up.
- Resource inefficiencies become performance bottlenecks and cost drivers when projects must be spun up and spun down with regularity.

A technology infrastructure that is unable to meet these demands can cause productivity failures at scale.

The Azure Databricks Solution

Azure Databricks

Azure Databricks is a fast, easy and collaborative Apache Spark™-based analytics platform optimized for Azure. It was created to bring Databricks' Machine Learning, AI and Big Data technology to the trusted Azure cloud platform.

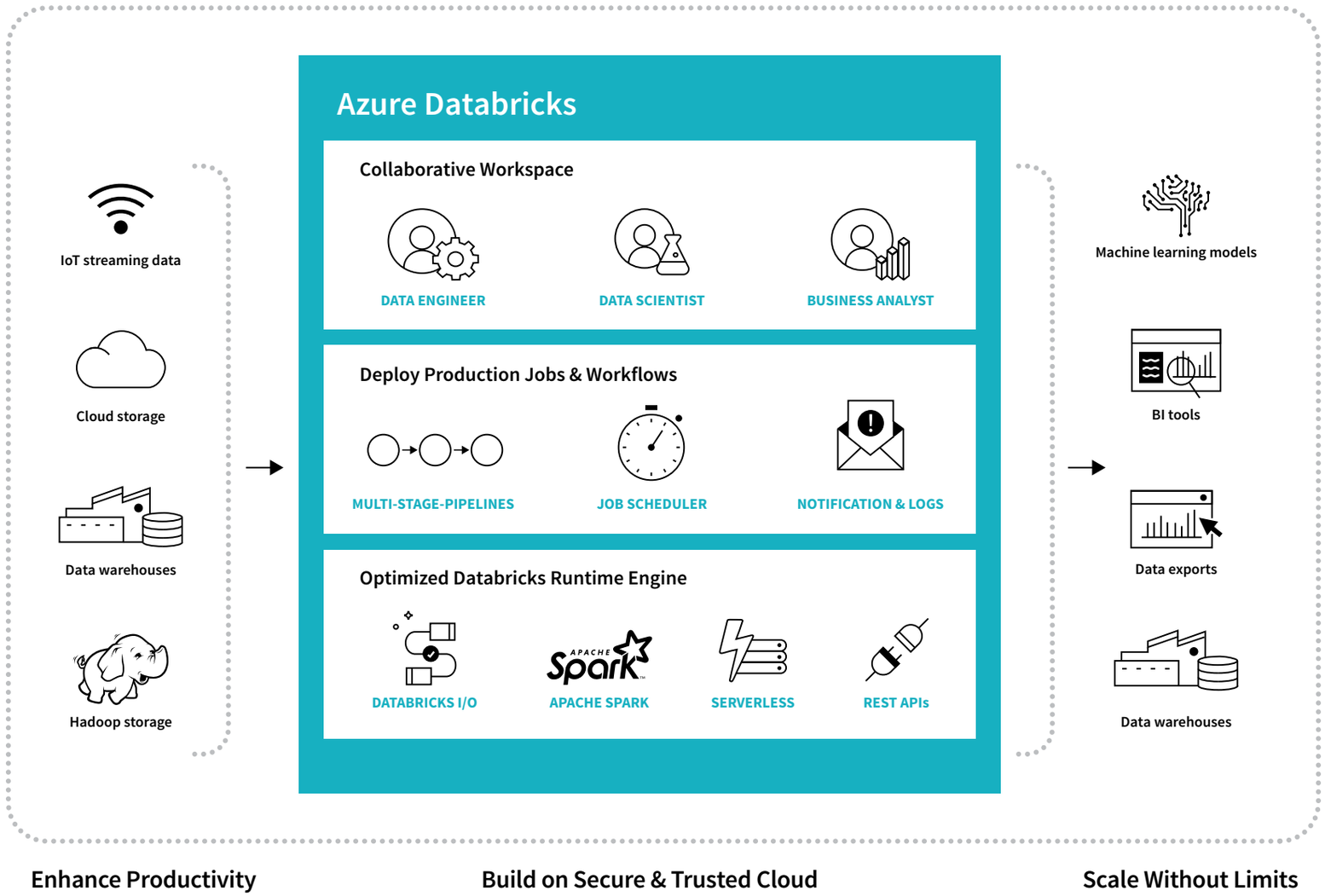
- Designed in collaboration with the team started the Spark research project at UC Berkeley — which later became Apache Spark — for optimal performance on Azure cloud.
- Data pipelines ensure analytics can be performed against growing volumes of data from multiple sources.
- Uniquely streamlined workflows and an interactive workspaces enable collaboration between data scientists, data engineers, and business analysts.
- Provides native integration with Azure services such as enterprise-grade Azure security, including Azure Active Directory integration, compliance, and enterprise-grade SLAs.

THE RAPID ASCENSION OF APACHE SPARK

- Matei Zaharia started the Spark Research project at UC Berekely in 2009.
- 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.

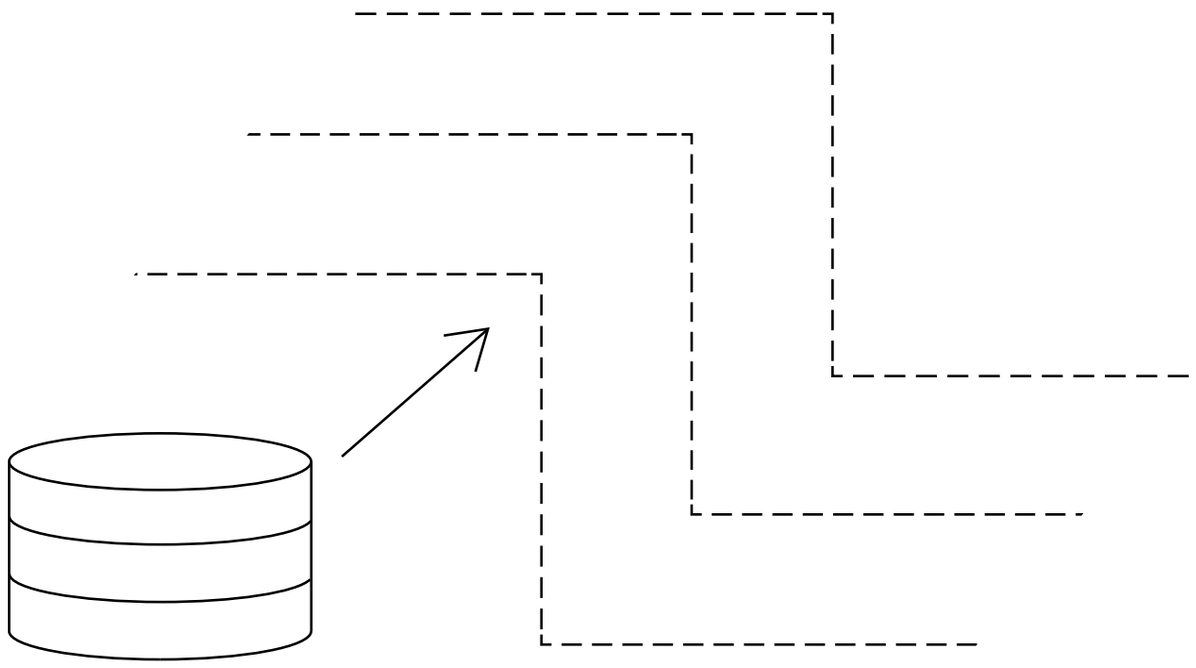
How it Works: Azure Databricks

The Azure Databricks service sits inside the Azure cloud. You can access all your Azure data sources to apply the power of the Azure Databricks analytics engine, and distribute your results by writing to visual dashboards or back to data warehouses for access.



Azure Databricks:
Scale without Limits

Azure Databricks is optimized from the ground up for performance and cost-efficiency to scale your business and handle the demands of Big Data.



- **OPERATE AT MASSIVE SCALE WITHOUT LIMITS, GLOBALLY**
  Databricks enables your analytics processes to scale up and down automatically, enabling you to process all of your data at once.

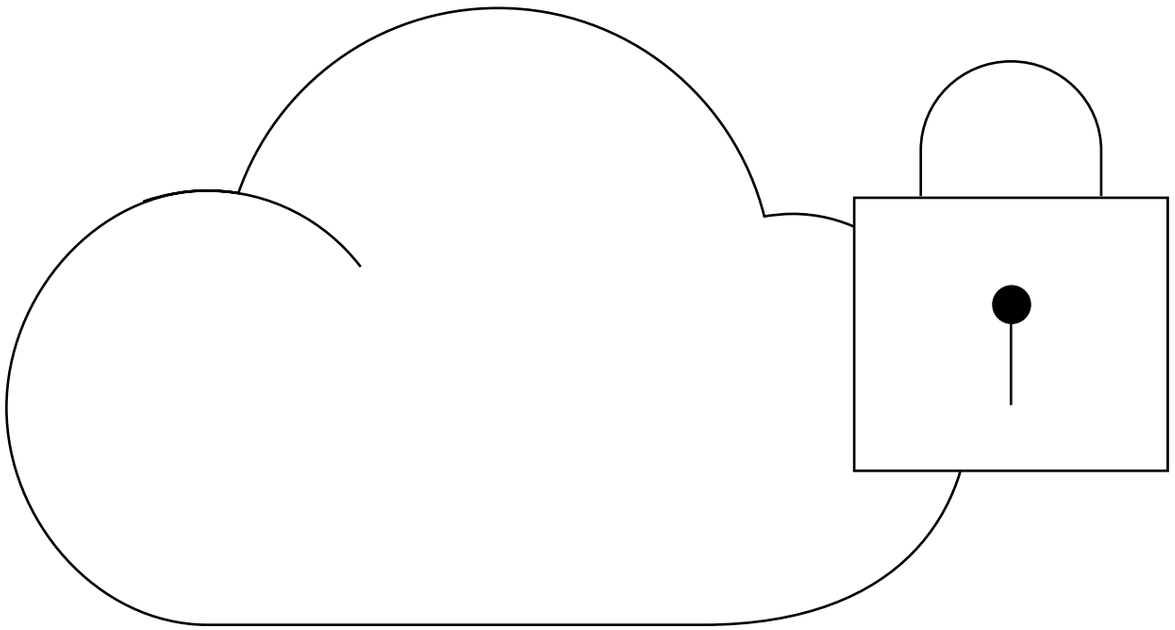
- **ACCELERATE DATA PROCESSING**
  Take analytics processes from weeks to hours or minutes with the fastest Spark engine built around speed, ease of use, and sophisticated analytics.

- **OPTIMIZED PERFORMANCE**
  Improve performance by as much as 10-100x over traditional Apache Spark deployments with performance optimizations including caching, indexing, and advanced query optimization.

Azure Databricks:

Build on a Secure, Trusted Cloud

Azure Databricks is uniquely architected to protect your data and business with enterprise-level security that aligns with any compliance requirements your organization may have.



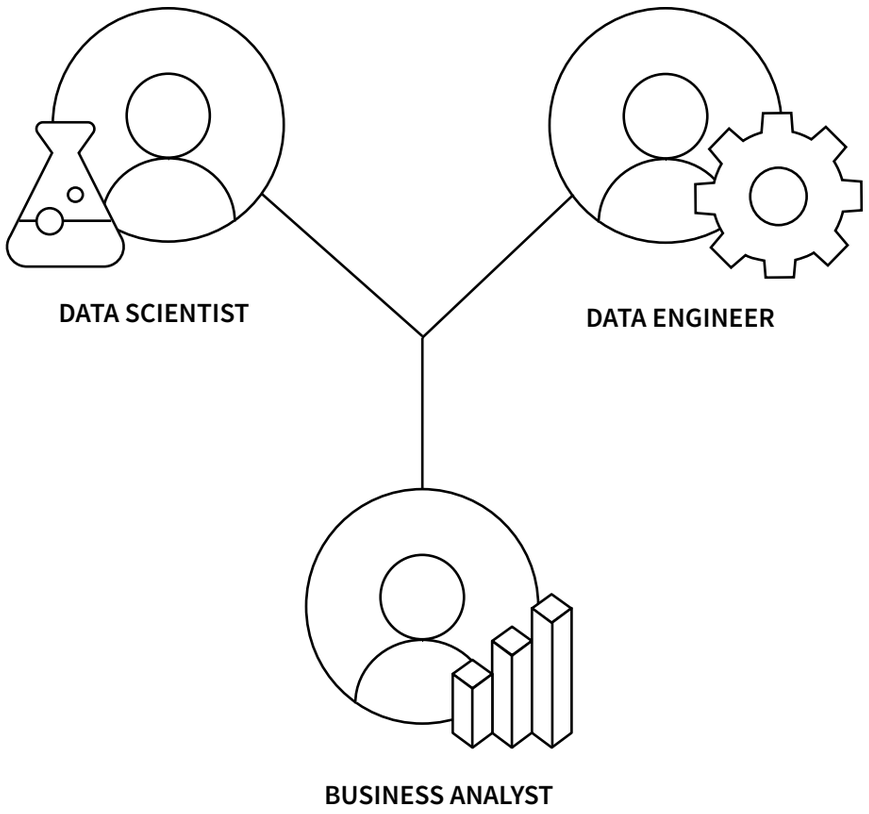
- **REGULATE ACCESS**
  Set fine-grained user permissions to Azure Databricks Notebooks, clusters, jobs, and data.

- **SIMPLIFY SECURITY AND IDENTITY CONTROL**
  Built-in integration with Azure Active Directory takes advantage of your existing roles and security settings.

- **BUILD WITH CONFIDENCE**
  Azure Databricks is backed by unmatched support, compliance and SLAs on the most-trusted cloud platform.

Azure Databricks:
Increase Productivity & Collaboration

Azure Databricks delivers the best of Azure and Apache Spark so that data science teams can be immediately productive.



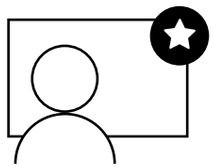
- **INSTANT PRODUCTIVITY**
  Users can launch a new Spark environment on Azure with a single click.

- **SEAMLESS COLLABORATION**
  A unified workspace provides interactive Notebooks and dashboards for real-time collaboration. Features such as seeing where each other is working in Notebooks, to the ability to add comments, enables users to work synchronously or asynchronously.

- **SHARABLE INSIGHTS**
  With rich Power BI integration, interactive visualizations can be shared across the organization, allowing for instant feedback and leading quickly to the next business question.

Sample Use Cases

These are just a few examples of the types of valuable analytics use cases you can address with Azure Databricks

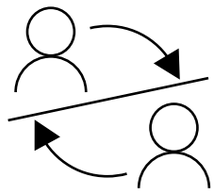


PERSONALIZED RECOMMENDATIONS

Customer Profiles, Viewing History, Online Activity, Content Sources, Channels

- Personalized Viewing and Engagement Experience
- Click-path Optimization
- Next-best Content Analysis
- Improved Real-time Ad Targeting

Faster Innovation for Customer Experience

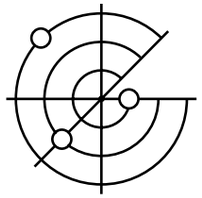


EFFECTIVE CUSTOMER RETENTION

Customer Profiles, Online Activity, Content Distribution, Services Data

- Quality of Service and Operational Efficiency
- Market Basket Analysis
- Customer Behavior Analysis
- Click-through Analysis

Improved Consumer Outcomes and Increased Revenue

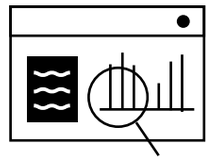


RISK AND FRAUD ANALYSIS

Transaction Data, Demographics, Purchasing History, Trends

- Real-time Anomaly Detection
- Fraud Prevention
- Customer Spend & Risk Analysis
- Data Relationship Maps

Risk Management With Machine Learning



INVENTORY ALLOCATION

Transactions, Subscriptions, Demographics, Credit Data

- Predict Audience Interests
- Network Performance & Optimization
- Pricing Predictions
- Nielsen Ratings and Projections
- Mobile Spatial Analytics

Predictive Analytics Transforms Growth

CONSUMER ENGAGEMENT ANALYSIS

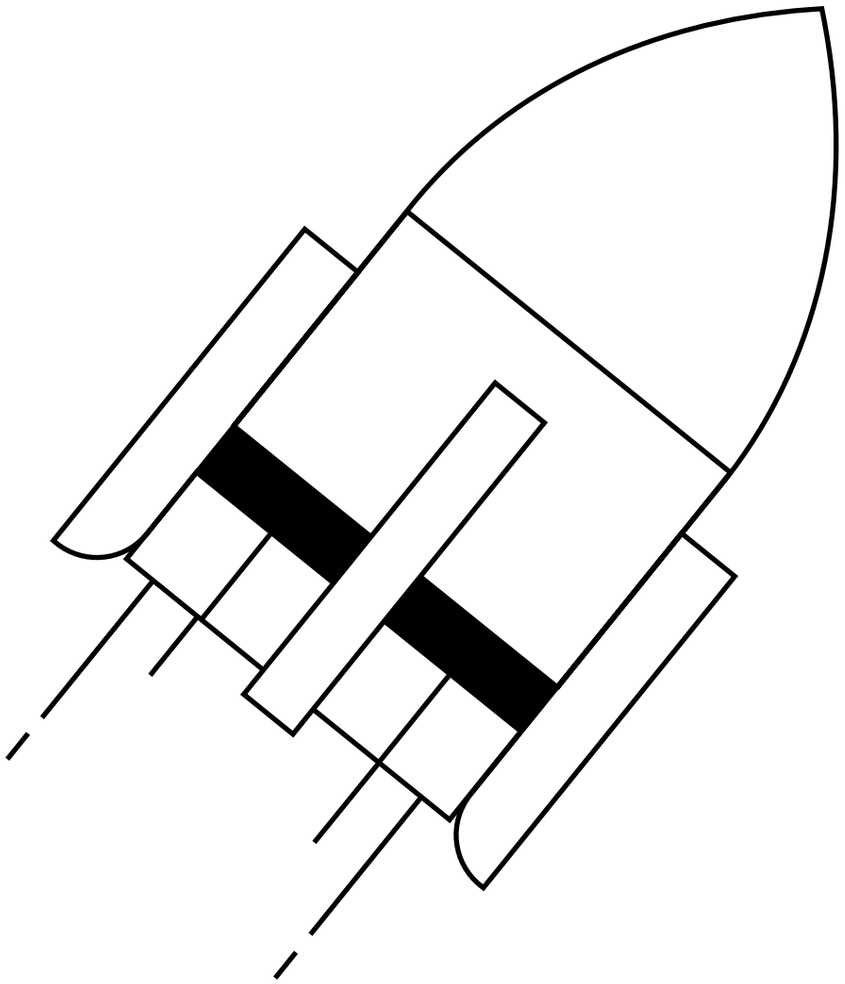
Content Metadata, Ratings, Comments, Social Media Activity

- Demand-Elasticity
- Social Network Analysis
- Promotion events Time-Series Analysis
- Multi-channel marketing Attribution

Improved Consumer Engagement With Machine Learning

Azure Databricks

By removing common analytics limitations, Azure Databricks allows organizations to innovate faster than ever before.

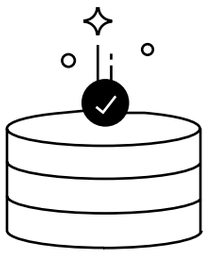


- Azure Databricks addresses the data volume issue with a highly scalable analytics engine. Processes that used to take weeks run in hours or minutes with Azure DatabricksIntegrated with Azure security, Azure Databricks provides fine-grained security control that keeps data safe while enhancing productivity.
- It enables people and processes to collaborate in a single Notebook, enabling faster iteration to get to the right answer quicker.

How to Get Started

STEP 1

Acquire and prep your data

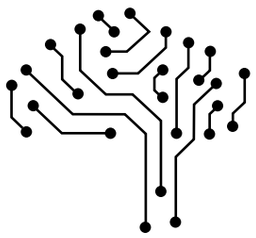


READ

Azure Databricks for Data Engineering

STEP 2

Prep and Train your ML models

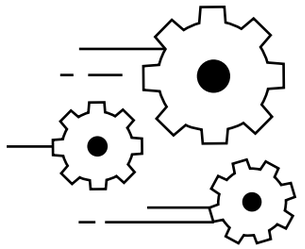


READ

Simplify Machine Learning Tech Note

STEP 3

Deploy models to production



READ

How to Productionize Your Machine Learning Models Using Apache Spark

Get started today — see tutorials and example videos at **databricks.com/azure**