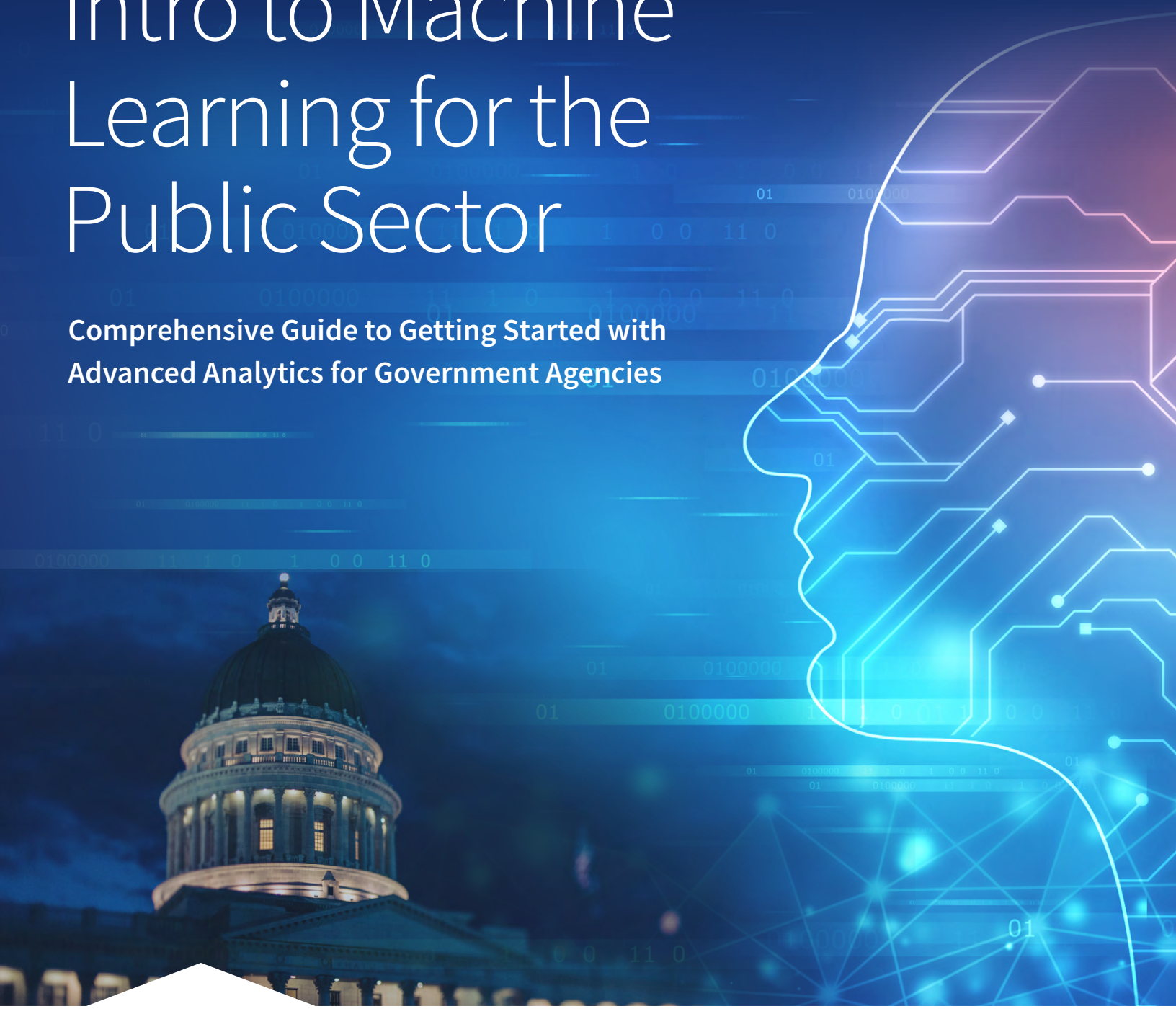
Intro to Machine Learning for the Public Sector

Comprehensive Guide to Getting Started with Advanced Analytics for Government Agencies

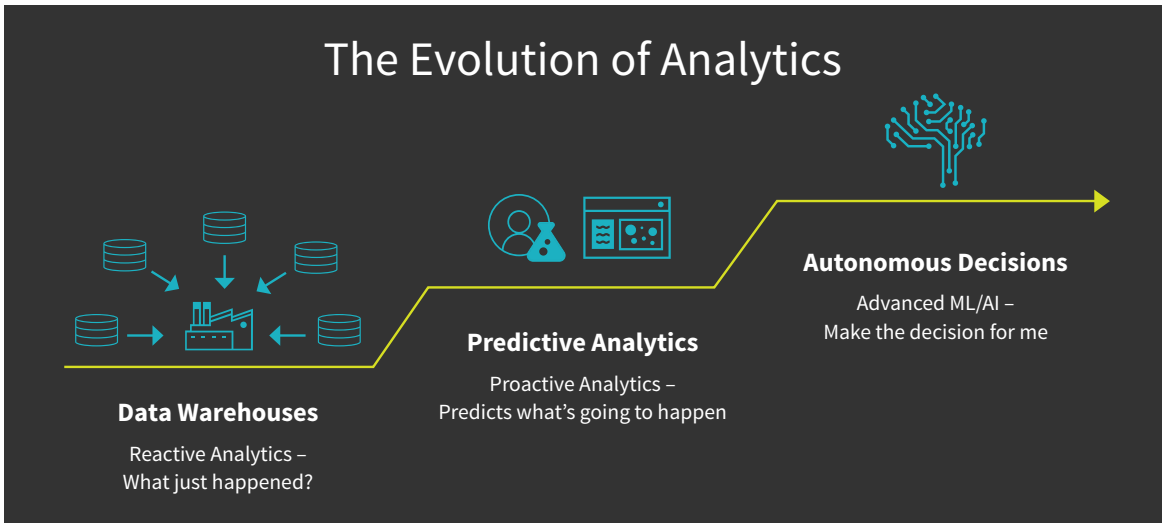


databricks™

The Evolution of Analytics: From Spreadsheets to Machine Learning

To make sound, data-driven decisions, government agencies need to take full advantage of the rich insights buried deep in their vast stores of data. Teams can no longer rely on simple relational databases, point-in-time data, and spreadsheets. Reaping the full value of your data requires the scale of the cloud and advanced analytics techniques.

For many, outdated spreadsheet-based workflows are still the rule. According to Gartner, "despite the availability of modern analytics and BI tools, spreadsheets are still deployed for significant business tasks, such as planning, forecasting, modeling, data mining, ad hoc analysis, and reporting." Moving from spreadsheets to more advanced analytics such as machine learning can alleviate the limitations of spreadsheets and accelerate innovation.



Machine Learning Adoption in the Public Sector

AI and machine learning adoption is on the rise in the public sector. Arming researchers with these powerful tools to access their data more easily and build advanced analytics and machine learning applications has the potential to improve operations, reduce waste, and even enhance national security. In fact, a report released last year by Deloitte University Press predicted that AI-driven automation could save 96.7 million worker hours every year for the federal government, translating to billions of dollars in cost savings and more time for employees to innovate and provide services.

The Evolution of Analytics: From Spreadsheets to Machine Learning

Machine Learning can help government agencies across a broad set of use cases, including:

Smarter Social Services
Predict and deliver on the needs of citizens more effectively, and reduce program waste by analyzing demographics, health statistics, claims, and other public data sets at scale.

Real-time Cyber Threat Detection
Agencies face a constant barrage of attacks, from bad actors ranging from cyber criminals to nation states. Now agencies can detect and prevent threats with predictive models that analyze network and user activity at scale, in real-time.

Fraud and Financial Crime Prevention
Protect financial markets against malicious actors with powerful predictive models that identify potential threats and help reduce false positives with real-time analytics at scale.

Stronger National Security
Analyze large volumes of geospatial, intel, and other data sources to identify risk patterns and prevent global and domestic security threats before they occur.

AI driven automation could save **96.7 million hours** each year for federal government workers alone, translating to billions in cost savings and more time for employees to innovate or provide services.

THE POWER OF AI IN GOVERNMENT

Use Case:

The Bureau of Labor Statistics (BLS) used AI to compile data on workplace injuries—a time-consuming task given the millions of data points that need to be sorted and categorized.

The Challenge:

In one year alone, the BLS identified more than 2000 job titles for a position that could generally be described as a "janitor" or "cleaner." Previously, the staff would review data manually and determine that the positions belonged in occupation code X, representing janitor and cleaner. Cleaning and prepping data required massive amounts of time before any analysis could begin.

The Impact of AI:

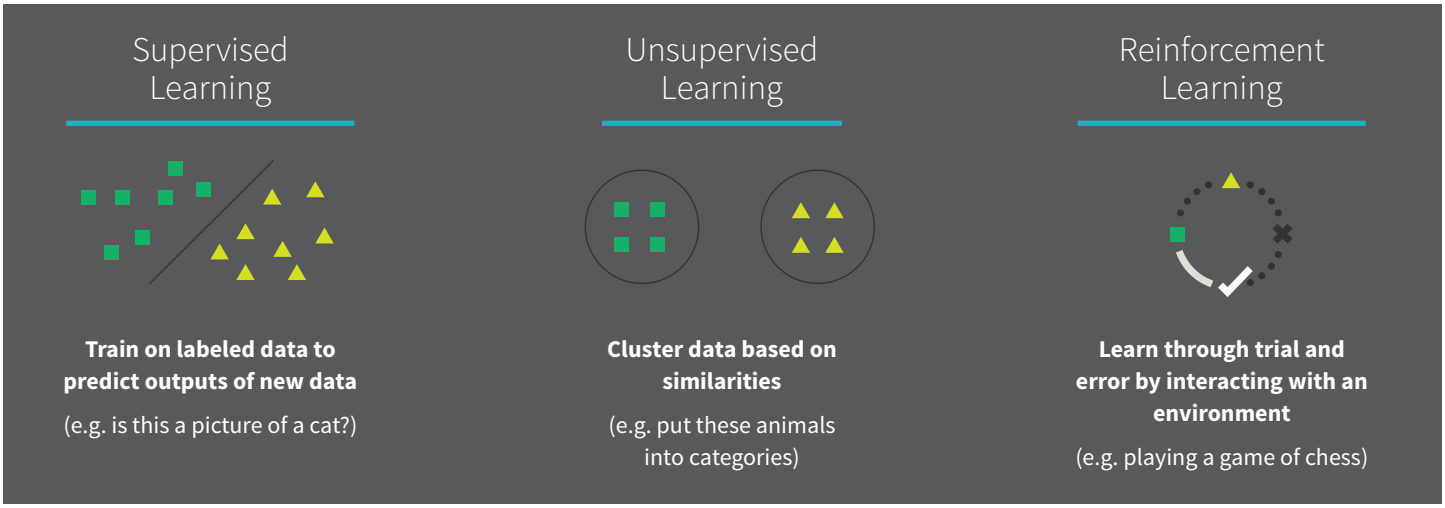
Making sense of the various titles was tedious and time-consuming for BLS employees. AI has reduced the burden by automating the effort, and teams can now focus on more important tasks.

Source:
https://fedtechmagazine.com/article/2018/01/where-ai-adoption-headed-federal-it

What is Machine Learning?

Machine learning is an application of artificial intelligence (AI) that uses statistical techniques to enable computers to solve problems without being programmed for each specific task. Machine learning focuses on developing computer programs that can access data and use it to learn and make accurate predictions on new sets of data.

Machine Learning is made possible by algorithms, which can be divided into three main categories.



1. Supervised Algorithms

Supervised machine learning is the most common form of machine learning. The goal is to predict the output of a given input. By feeding a supervised model a set of labeled data, or "training data," the model learns from that data and adjusts its variables to map inputs to the corresponding output. Once the model achieves the desired accuracy rate, the training ends.

An example is using a set of car types as a training data set. Training inputs against this data would yield outputs such as, "This car is a Toyota Camry."

Supervised learning falls into two primary categories:

- **REGRESSION:** Regression is a technique that aims to reproduce the output value. For example, regression analysis can be used to predict the price of a house in a certain city.
- **CLASSIFICATION:** Classification is a technique that aims to reproduce class assignments. The car example described above is an example of a classification analysis.

Here are some supervised learning algorithms commonly used to make predictions:

- **LINEAR REGRESSION:** an approach to modeling the relationship between a dependent variable and one or more independent variables. For example, you could use linear regression to correlate poverty-level data and geography (independent variables) with teen birth rate (dependent variable), i.e., you could determine whether the teen birth rate correlates with factors such as poverty level and geography.
- **RANDOM FOREST:** These algorithms leverage the concept of a decision tree, where questions are asked about the data based on the value of a feature. Each question has either a true or false answer. Based on the answer to the question, the data point moves down the tree. This is a very popular way to build a recommendation engine.
- **SUPPORT-VECTOR MACHINES:** Used for classification problems, a common use case for support-vector machines is image classification, which can be used to improve search accuracy for images. This approach is more accurate than conventional query-based searching techniques.

2. Unsupervised Algorithms

In this category, there is no target outcome. The algorithms cluster the data set for various groups, and the model is trained with unlabeled data. Since there is no target outcome, the model can learn and discover new and potentially interesting similarities in the data. That's why it's called "unsupervised"; it's not restricted by a desired outcome.

There are two forms of unsupervised learning:

- **CLUSTERING:** Clustering is a technique that aims to identify similar groups or cohorts. For example, this approach can help group photos of similar types of cars.
- **ASSOCIATION:** Association is a technique that enables discovery of rules that describe large portions of data; for example, "people who buy product A also tend to buy product B."

Here are some commonly used unsupervised-learning algorithms:

- **K-MEANS:** This is a very common approach for clustering problems. For example, a restaurant owner may want to open additional restaurants. How do they determine where to open them? To ensure that they select the right areas, they would need to solve a number of problems to make the most informed decisions. This is done by analyzing various groups of data based on a common goal.
- **HIERARCHICAL CLUSTERING:** This is an approach for clustering data points into "parent" and "child" clusters. For example, this algorithm can be used to split customers according to age, e.g., "old" and "young," and then split each of those groups further into more targeted age groups.
- **AUTOENCODER:** This is an example of an algorithm that extends unsupervised learning into the world of deep learning and neural networks. Financial institutions often use autoencoders to identify anomalies for fraud detection.
- **APRIORI:** These algorithms are used for association-rule learning problems. It's being used in healthcare to help detect adverse drug reactions by producing association rules to indicate combinations of medications and patient characteristics that could lead to adverse reactions.

3. Reinforcement Algorithms

Reinforcement training is a machine learning technique that trains a model to forge the best possible decision-making path based on a given goal. Based on those decisions, the algorithm will train itself depending on the success/error of output. Eventually, through experience, the algorithm will give accurate predictions. Unlike supervised training, reinforcement training has no predefined right answer; the model learns and decides what to do to perform a given task. An example is a machine trained to perform an advanced task such as playing chess.

With this machine learning technique, you can reinforce behavior based on either positive or negative outcomes. If it's a positive-reinforcement algorithm, it will strengthen the behavior as it achieves positive results. If it's a negative-reinforcement algorithm, it will strengthen the behavior if a negative result is achieved.

What is Machine Learning?

The Machine Learning Lifecycle

The machine learning lifecycle represents the various stages in driving a data science initiative forward. Delivering on a machine learning use case requires teams across data engineering, data science, and the organization to work closely throughout the process. Here are the main steps to building and deploying a machine learning model:

DATA ACCESS AND PREPARATION

The most important component of any data science project is the data. Data is often stored across myriad data sources and systems. The more data you can collect and feed into your model, the more accurate it will be.

EXPLORATION AND FEATURIZATION

In machine learning, a "feature" is a measurable data property or characteristic. Choosing clear, independent features is a crucial step for effective algorithms. Once you've collected your data, you can begin to explore the data to identify interesting features that will support your use case objectives.

MODEL BUILDING AND TRAINING

Next, you can access various algorithms via popular machine learning libraries, such as MLlib, scikit-learn, and others, and model them against the data. Once you've built your machine learning models, you train them against the data until you achieve desired levels of accuracy.

MODEL DEPLOYMENT

Deploying your models into production allows your internal and external systems to run their data against the models to get the projected outcome.

What Powers Machine Learning

Machine Learning Frameworks

A machine learning framework is an interface, library, or tool that allows developers to build machine learning models more quickly and easily, without getting into the nitty-gritty of the underlying algorithms. One of the key features of a good ML framework is optimizing for performance.

General ML Frameworks

NumPy, an extension package for scientific computing with Python

scikit-learn, an easy-to-use ML framework for numerous industries

NLTK, a Python-based human language data processing platform

Deep Learning Frameworks

Deep learning is a subset of machine learning that allows for analysis of unstructured data, such as language and images. An example of a deep-learning application is the voice assistant built into your smartphone, which analyzes voice commands and responds based on the request.

TensorFlow, a flexible framework for large-scale machine learning; and TensorBoard, a good tool for model training visualization

PyTorch, an easy to use tool for research

Keras, a lightweight, easy-to-use library for fast prototyping

Caffe2, a deep learning library with mobile deployment support

What Powers Machine Learning

Machine Learning Tools

Machine learning tools fall into three categories: languages, visualization tools, and big-data tools. We have highlighted some of these below.

Machine Learning Languages

Python, a popular language with high-quality machine learning and data analysis libraries;

C++, a middle-level language used for parallel computing on CUDA;

R, a language for statistical computing and graphics

Data Analytics and Visualization Tools

Pandas, a Python data analysis library enhancing analytics and modeling

Matplotlib, a Python machine learning library for quality visualizations

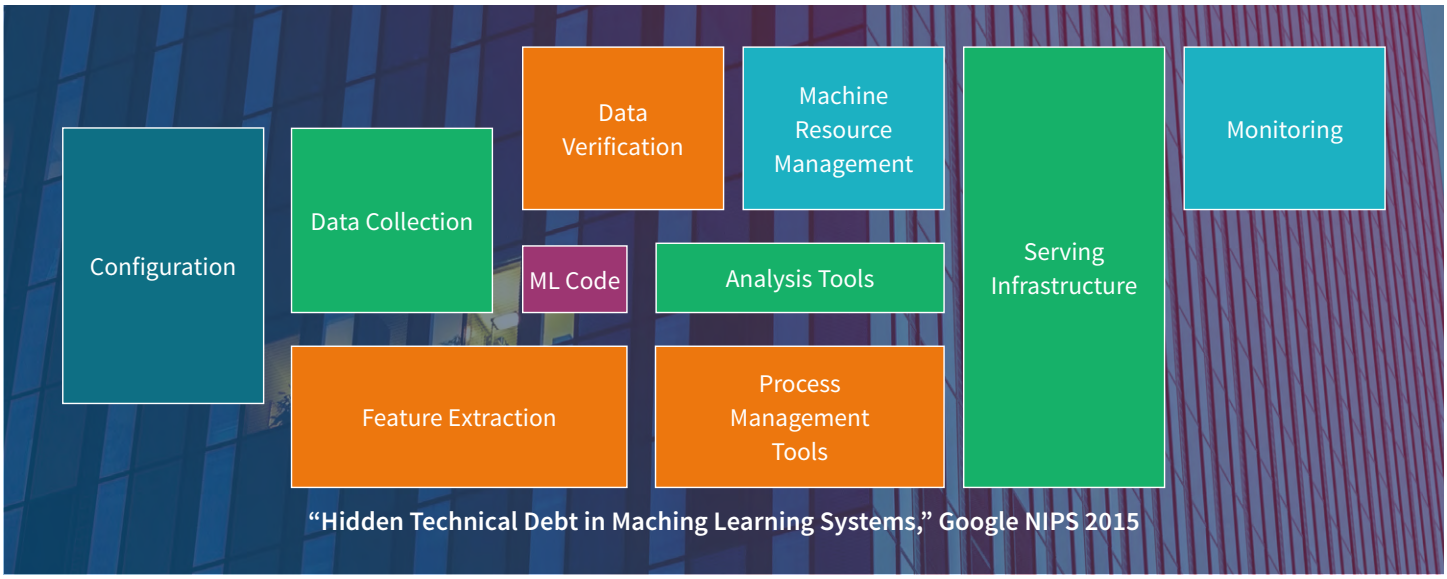
Tableau, a powerful data exploration capabilities and interactive visualization

Big Data Tools

Apache Spark, a unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing

Challenges Adopting ML in the Public Sector

There are multiple challenges with AI adoption at government agencies that's slowing data-driven innovation. Many organizations think that the biggest barrier to AI success is the complexity of machine learning. In reality, most of the work required to enable machine learning—from configuring the hardware to ingesting and preparing data for downstream analytics—is in the surrounding infrastructure.



Lack of Big Data and Data Science Skills

In order for the potential of AI to be fully realized, a fundamental shift in skill sets in the workforce needs to take place to better utilize the technology.

But competing with the private sector for engineering talent is only part of the challenge faced by the government as it drives the modernization agenda. Technology is changing so rapidly, the government will need to focus on educating the existing workforce, introducing programs and training opportunities to advance their technical skills.

Challenges Adopting ML in the Public Sector

Large Amounts of Data

The government captures massive amounts of data—everything from immigration statistics to weather data. The challenge is not in capturing the data, but in gleaning insights from it. For example, the intelligence community is eager to leverage ML and AI technologies to help make sense of its data to proactively predict threats and protect our national security. Before it can do so, however, the intelligence agencies have a big problem to solve. Their legacy IT architectures are likely comprised of siloed systems and data sources. Thus, making information available for machine learning becomes a daunting, resource-intensive task.

Stringent Security Requirements

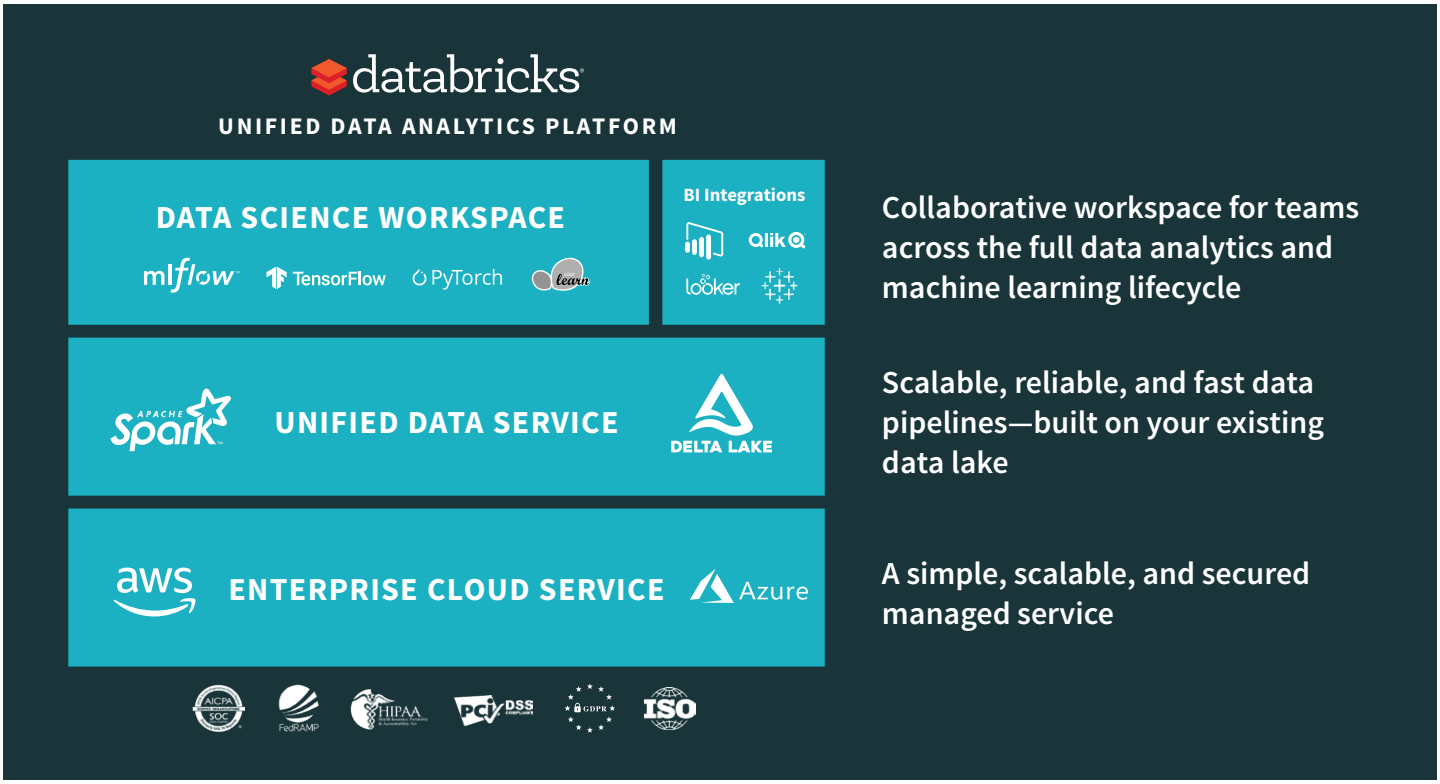
Government agencies have extremely stringent security standards. To ensure that data is protected from intruders and malicious actors, many agencies use an "air-gap" strategy, in which networks are physically isolated from unsecured networks such as the Internet. This means that the systems on these networks lack easy access to malware signatures pushed out by antivirus vendors after they identify and analyze new malware. It also makes it that much more difficult to join external data.

Legacy Technology Stack

Modernizing technology infrastructure is a top priority for government organizations today. However, fragmented technology infrastructures, many different legacy on-premise applications, and the increasing number of endpoints that need to be secured make it extremely difficult for security stakeholders to protect valuable government data.

How Databricks Makes It Easy

Data fuels AI innovation, and modern government agencies require a comprehensive, unified approach to analytics to fully harness the power of machine learning. Databricks helps agencies overcome the challenges of building machine learning applications with a platform that unifies data engineering, data science, and the organization. The Databricks Unified Data Analytics Platform, powered by Apache Spark™, helps teams become truly data-driven, to accelerate innovation and deliver transformative outcomes.

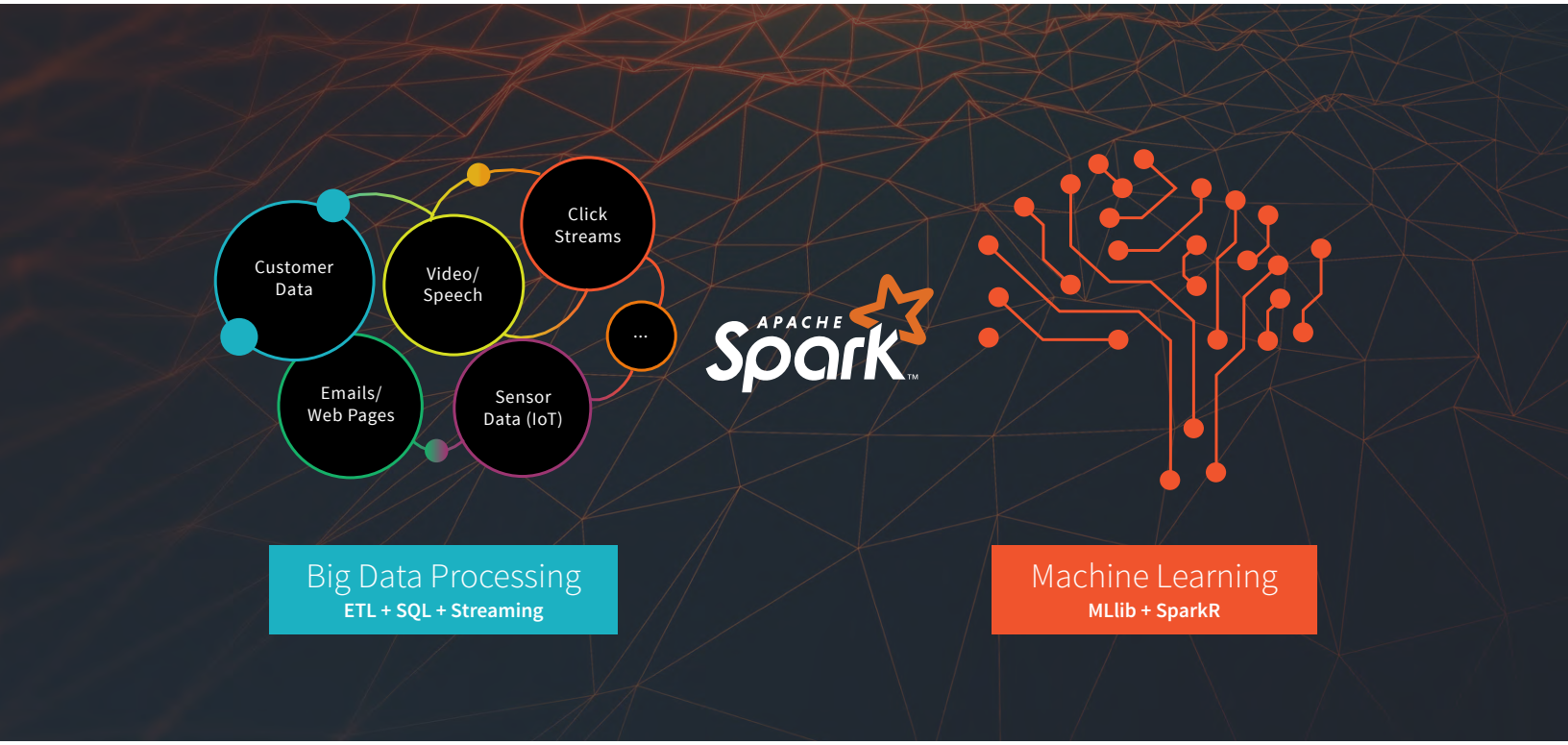


How Databricks Makes It Easy

The Databricks Advantage: Apache Spark™

A Unified Analytics Engine

To avoid the problems associated with siloed data and disparate systems for performing various analytic processes, government agencies are increasingly turning to Apache Spark. Originally created by the founders of Databricks, Apache Spark is the de facto standard for large-scale data processing and AI due to its speed, ease-of-use, and support for sophisticated analytics.



Spark is at the core of the Databricks Unified Data Analytics Platform. It simplifies data preparation for AI by bringing together data at massive scale across various sources, including cloud storage systems, distributed file systems, key-value stores, and message buses. Spark also unifies data and AI with a consistent set of APIs for simple data loading, batch and stream processing, SQL analytics, stream analytics, graph analytics, machine learning, and deep learning, along with seamless integration with popular AI frameworks and libraries, such as TensorFlow, PyTorch, R, and scikit-learn.

How Databricks Makes It Easy

The Databricks Advantage: Collaborative Workspaces

Unify Data Engineers and Data Scientists

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

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

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

Databricks ensures that teams can become heroes, by providing a common interface and tools enabling all stakeholders, regardless of skill sets, to collaborate with each other. This eliminates data silos and allows teams to collaborate across the AI lifecycle, from experimentation to production, which benefits organizations and speeds innovation.

How Databricks Makes It Easy

The Databricks Advantage: Simplified Data Management at Scale

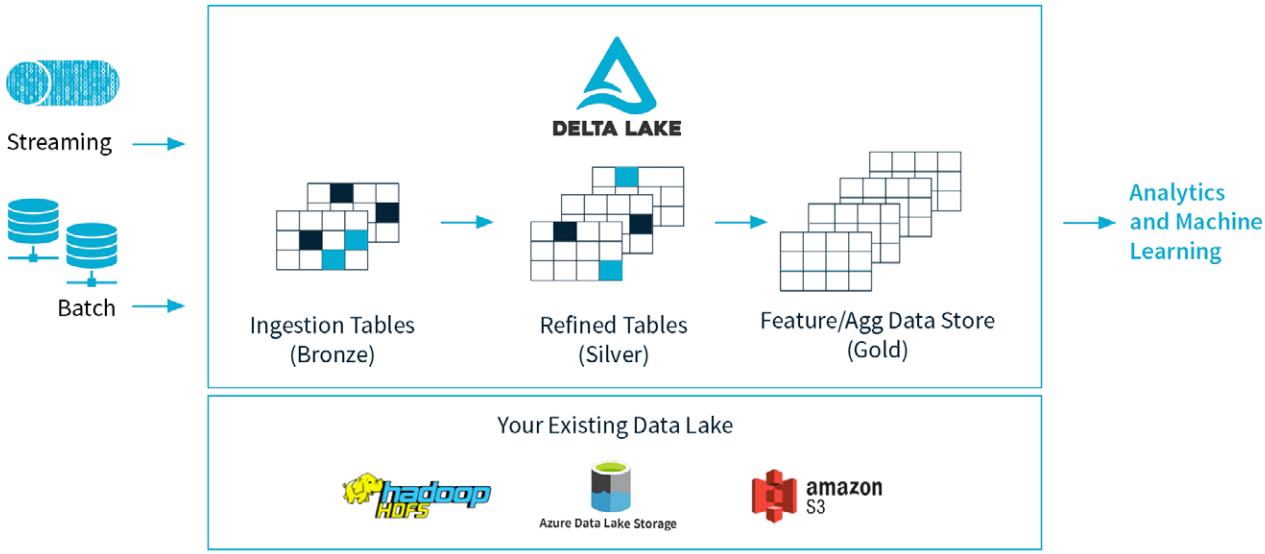
Build Reliable and Performant Data Pipelines

Building best-in-class AI applications requires large amounts of data. Data science techniques developed years ago are only now beginning to come to fruition due to the sheer volume of data available for training algorithms. The faster you can ingest and prepare the data for analytics, the sooner you can realize the benefits of AI.

Databricks has vastly improved data-processing performance. Through automated infrastructure management capabilities, such as autoscaling and various optimizations for large-scale data processing in the cloud, building data pipelines at scale is faster and more performant than alternative approaches.

Enabling Reliable Data Lakes at Scale

Many organizations have turned to data lakes to support big-data workflows. However, data lakes face challenges as a result of failed writes, schema mismatches, and data inconsistency, especially when it comes to mixing batch and streaming data.



Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads

How Databricks Makes It Easy

Delta Lake provides reliable, high-quality data that is always ready for analytics through a range of features for ingesting, managing, and cleaning data. Delta Lake provides consistent views while supporting multiple simultaneous readers and writers, even in mixed batch and streaming data environments. Delta Lake is natively integrated into the Databricks Unified Data Analytics Platform and runs on top of your existing data lake to provide the following advantages:

ACID TRANSACTIONS

Bring ACID transactions to your data lakes by providing serializability, the strongest level of isolation.

TIME TRAVEL (DATA VERSIONING)

Provide snapshots of data, enabling developers to access and revert to earlier versions for audits and rollbacks, or to reproduce experiments.

UNIFIED BATCH AND STREAM

A table in Delta Lake is both a batch table and a streaming source and sink, enabling streaming data ingestion, batch historic backfill, and interactive queries out of the box.

COMPATIBLE WITH APACHE SPARK API

Use Delta Lake with existing data pipelines with minimal change, as it is fully compatible with Spark.

SCALABLE METADATA HANDLING

Delta Lake treats metadata just like data, leveraging Spark's distributed processing power to handle all its metadata to manage petabyte-scale with ease.

OPEN FORMAT

All data in Delta Lake is stored in Apache Parquet format, enabling Delta Lake to leverage Parquet's efficient compression and encoding schemes.

SCHEMA ENFORCEMENT

Delta Lake lets you specify and enforce your schema, to ensure that data types are correct and required columns are present, thereby preventing data corruption.

SCHEMA EVOLUTION

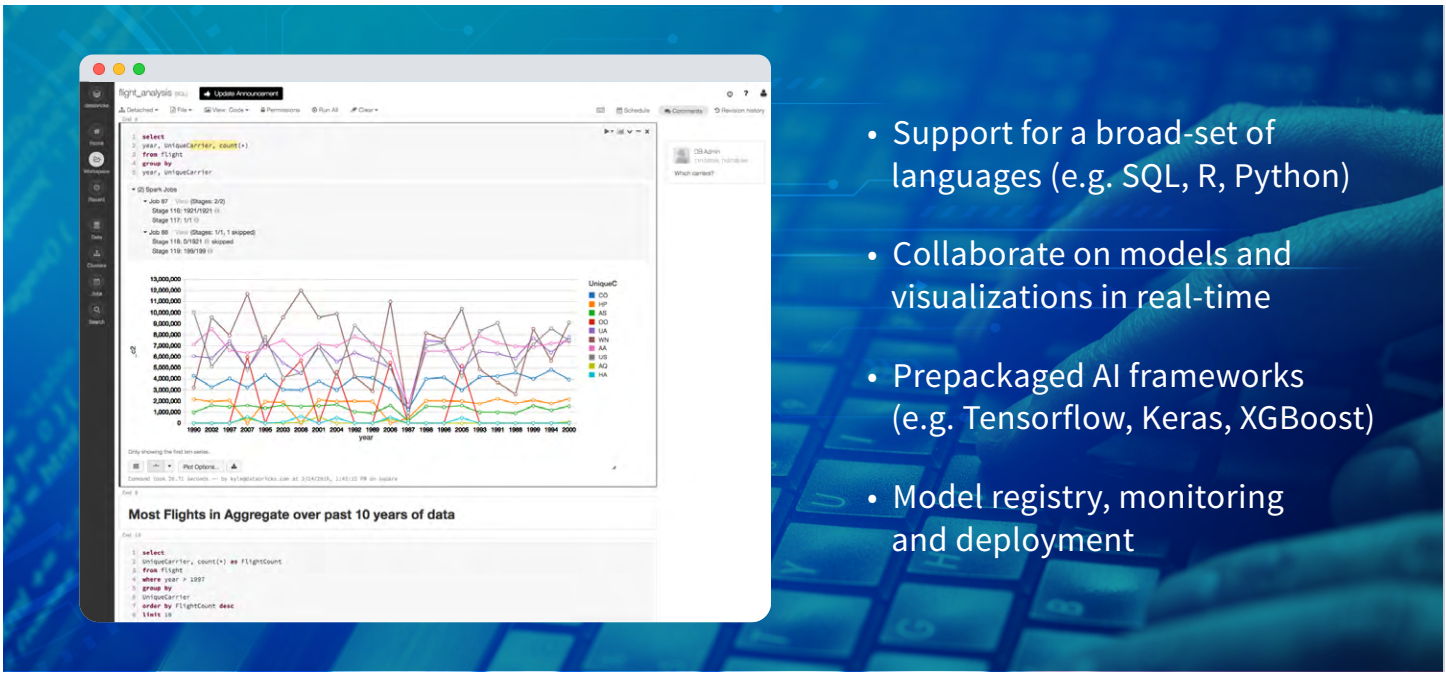
Make changes to a table schema and apply it automatically, without the need for cumbersome DDL.

How Databricks Makes It Easy

The Databricks Advantage: Machine Learning Made Easy

Build Cutting-Edge AI Models with Ease

Data science teams can leverage the Databricks Unified Data Analytics Platform to easily train, evaluate, and deploy more effective AI models to production. You can connect with data sets quickly to perform data exploration, analysis, and transformations, using SQL, R, or Python. You can also explore data interactively with collaborative notebooks that allow data scientists to take machine learning models from experimentation to production at scale.



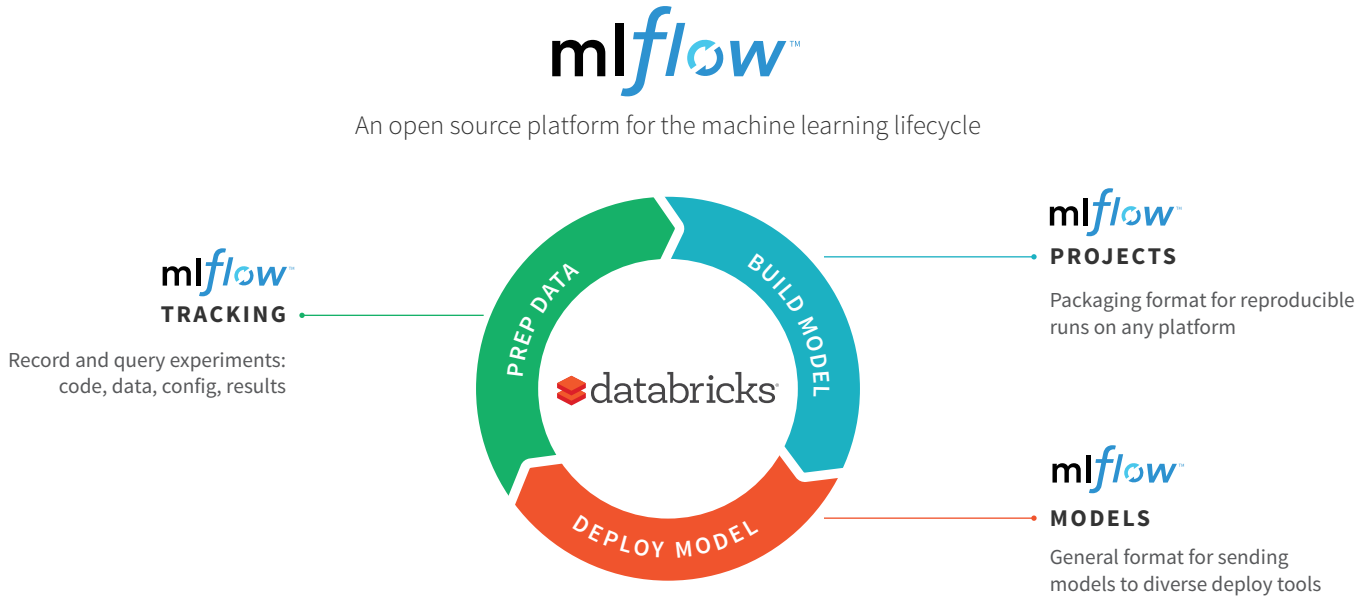
With prepackaged AI frameworks, such as TensorFlow, Horovod, Keras, XGBoost, PyTorch, scikit-learn, MLlib, GraphX, and sparklyr, data science teams can easily provision AI-ready Databricks clusters and notebooks in seconds on its cloud-native service.

How Databricks Makes It Easy

Streamlining the Machine Learning Lifecycle

Successfully building and deploying a machine learning model can be difficult to do once. Enabling other data scientists to reproduce a pipeline, compare the results of different versions, track what's running where, and redeploy and rollback updated models, is much harder.

MLflow, an open-source framework for managing the complete machine learning lifecycle, comes preinstalled on Databricks, making it easy for data scientists to track and share experiments locally or in the cloud. Teams can easily package models and frameworks and deploy virtually anywhere.



EXPERIMENTS TRACKING

Quickly record runs and keep track of models parameters, results, code, and data from each experiment, all in one place.

REPRODUCIBLE PROJECTS

Build and package composable projects, capture dependencies and code history for reproducible results, and quickly share projects with peers.

MODEL DEPLOYMENT

Quickly download or deploy any saved models to various platforms, locally or in the cloud, based on your needs, thereby eliminating code rewrites prior to deployment.

How Databricks Makes It Easy

The Databricks Advantage: Security and TCO

Lowering the Total Cost of Ownership

When adopting new technologies, all vendors promise lower total cost of ownership, but these are often empty promises. Databricks stands behind our lower TCO claim with a cloud-native unified platform that requires no expensive hardware; an operationally simple platform designed to help you manage your costs efficiently; increased productivity through seamless collaboration; support for familiar languages, such as SQL, R, Python, and Scala; and faster performance than other analytics products, resulting in lower cost and shorter time-to-value.

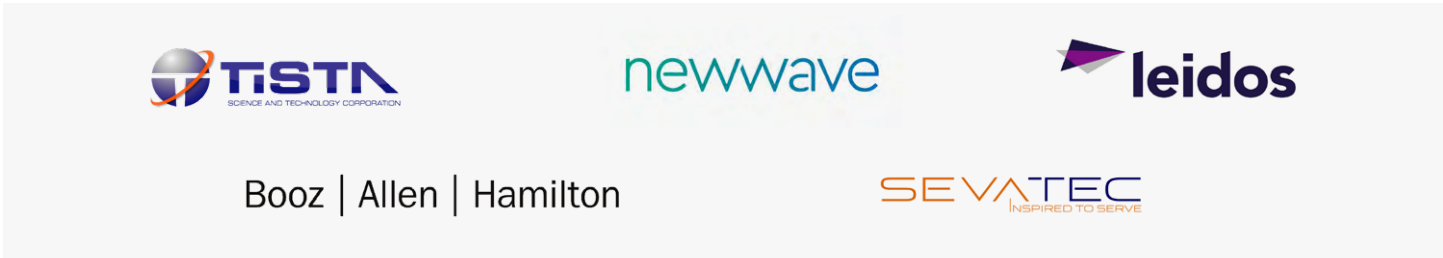
Built Secure for the Federal Government

Databricks Unified Data Analytics Platform is a secure, cloud-based platform—hosted on AWS and Azure—that makes it easy to build, scale, and deploy AI and ML applications in minutes.

- Hosted on AWS and Azure, with a federally compliant security model
- AWS Public Sector Partner, with support for AWS GovCloud
- Authority to Operate within all AWS regions, including secret and SCI clouds for the DoD and U.S. Intelligence Community
- HIPAA compliant; working toward FedRamp
- Robust encryption, user management, and data governance
- Trusted by more than 500 companies and government entities

Government Partners You Can Trust

Databricks works with federal system integrators and consulting partners that provide the expertise, technology skills, and solutions you need to better enable project success.



Customer Spotlight

The Challenges

Sevatec was struggling to provide a holistic analytics and AI solution to an agency with legacy architecture that needed to scale and meet its current needs.

- Difficult to ingest and prepare data across more than 30 disparate systems.
- Support 2000-plus siloed users who have varied skill sets (BI users, statisticians, data engineers, data scientists, business).
- Lack of ability to prepare a single view into the data for data science
- Inability to scale data science efforts as the agency was using RStudio on a single node.
- Variety of tools being used to perform large data extractions, creating substantial DevOps complexity.
- System was heavily taxed by IO-intensive queries, impacting the agency's ability to meet SLAs and the demands of the rest of its user community.

The Solution

Sevatec leveraged the Databricks Unified Data Analytics Platform to lead a digital transformation at the agency. Sevatec greatly simplified data engineering through a fully managed cloud platform and accelerated data science innovation by fostering a culture of collaboration and transparency.

- Democratized access to data across various data sources through APIs and data source connectors. Reduced access and ingest times from hours to minutes.
- Enabled building machine learning models at scale against the entire data set.
- Simplified infrastructure management and eliminated unnecessary DevOps work through automated and secured cluster management.
- Interactive workspace allows various users to collaborate on the data and run data science experiments that lead to innovative machine learning models.

The Impact

Databricks empowered Sevatec to change the agency's culture and roll out new a new platform that unifies all data, teams, and technology, and democratizes data across the agency to accelerate machine learning innovation.

- With Databricks, extracting data for data science experiments now takes minutes instead of hours, resulting in **faster data analytics at scale**.
- People were sampling data for statistical operations, and now they no longer have to sample the data, resulting in **simple, more efficient data engineering**.

SEVATEC
INSPIRED TO SERVE

Use Case

Sevatec is a high-technology services firm that engages trusted talent and leverages technology innovation to overcome the federal government's most pressing challenges. The company is focused on delivering end-to-end analytics and machine learning applications to the federal government, including agencies such as Homeland Security, Citizenship and Immigration Services, Department of Defense, Department of Transportation, Department of State, and other federal government entities.

"With Databricks, we get secure, easy access to data without worrying about the engineering behind it. Data has been held hostage, and Databricks allows us to liberate that data without compromising security."

RAKESH POL
BI Architect, Sevatec

The Bottom Line

The goal of Databricks Unified Data Analytics Platform is to accelerate data-driven innovation. It accomplishes this by uniting people around a shared objective with a common collaborative 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.

Contact us to learn more or start a free trial today:
databricks.com/trial.