ISAUnited
International Security Architects

DataBricks
Threat Analysis and Security Controls

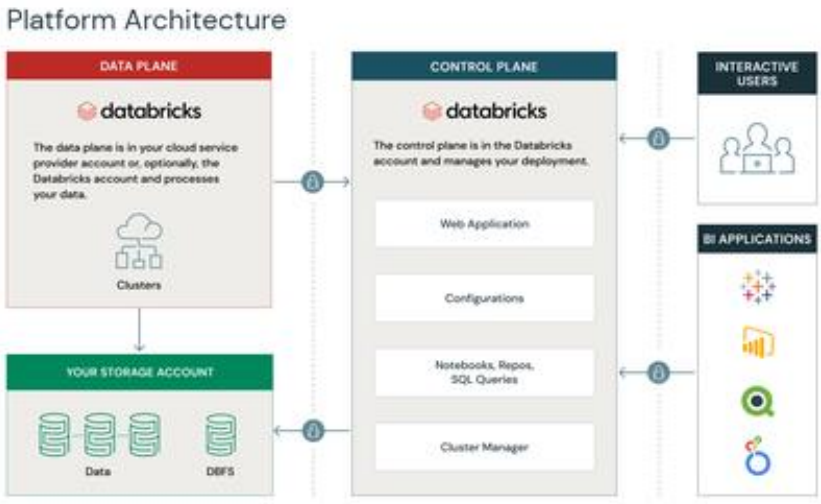
DataBricks
Threat Analysis and Security Controls

This technical security control document will cover:

- Technology Introduction and Usage
- Technology Components and or Systems
- Technology integrations with Azure and AWS
- Databricks Security Recommendations
- Mitre Attack Threat Analysis
- Microsoft STRIDE Threat Model

Technology Introduction and Usage

Databricks is a data analytics platform optimized for the Microsoft Azure cloud services platform. Databricks offers three environments for developing data intensive applications: Databricks SQL, Databricks Data Science & Engineering, and Databricks Machine Learning.



Databricks Data Science & Engineering provides an interactive workspace that enables collaboration between data engineers, data scientists, and machine learning engineers. For a big data pipeline, the data (raw or structured) is ingested into Azure through Azure Data Factory in batches, or streamed near real-time using Apache Kafka, Event Hub, or IoT Hub. This data lands

in a data lake for long term persisted storage, in Azure Blob Storage or Azure Data Lake Storage. As part of your analytics workflow, use Azure Databricks to read data from multiple data sources and turn it into breakthrough insights using Spark.

Databricks Machine Learning is an integrated end-to-end machine learning environment incorporating managed services for experiment tracking, model training, feature development and management, and feature and model serving.

Technology Components and or Systems (hardware and software used)

Databricks SQL provides an easy-to-use platform for analysts who want to run SQL queries on their data lake, create multiple visualization types to explore query results from different perspectives, and build and share dashboards.

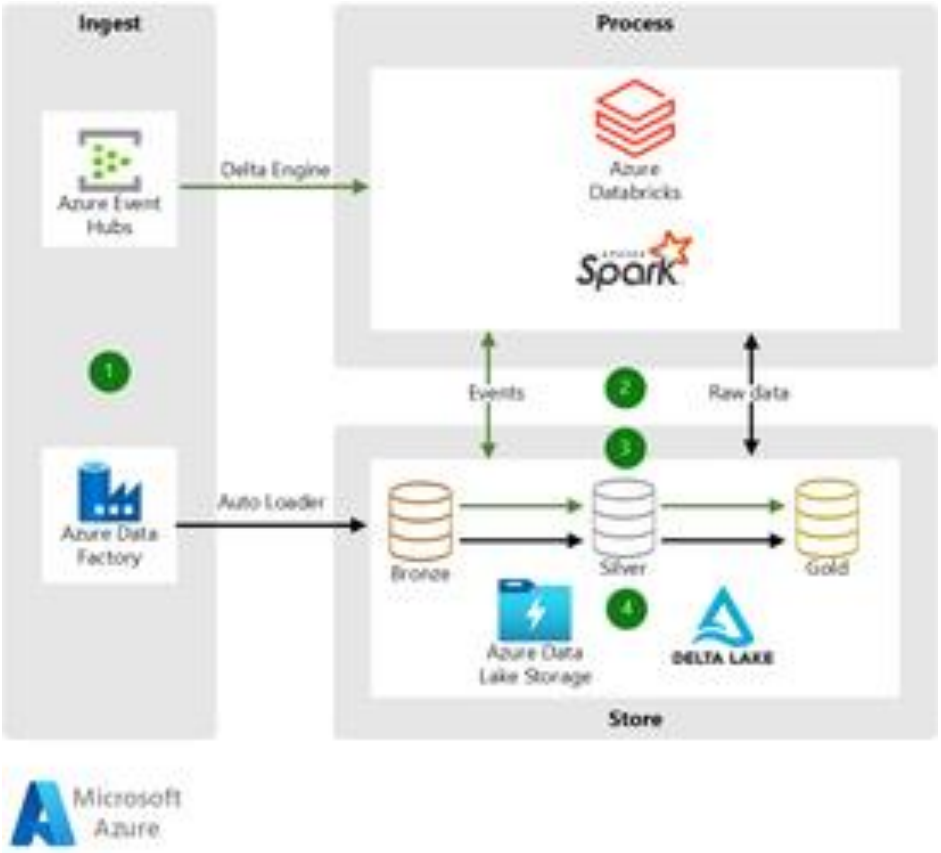
Apache Spark is what Databricks is built on-top of, is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size**.**

Technology integrations with Azure and AWS

Integrations with Microsoft Azure®

Azure Databricks Connect is a client library for Databricks Runtime. It allows you to write jobs using Spark APIs and run them remotely on an the “Azure Databricks cluster” instead of in the local Spark session.

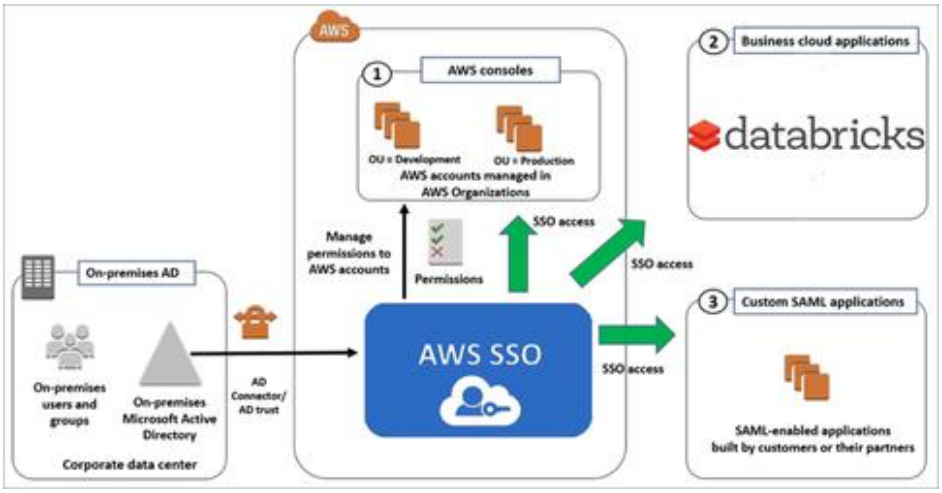
Full setup steps: https://docs.microsoft.com/en-us/azure/databricks/scenarios/connect-databricks-excel-python-r



Integrations with AWS

As an AWS Partner, integrating Databricks is simple, Databricks on AWS allows you to store and manage all your data on a simple, open lakehouse platform that combines the best of data warehouses and data lakes to unify all your analytics and AI workloads.

Full Integration/Setup: https://docs.databricks.com/administration-guide/account-settings/aws-accounts.html



Databricks Security Recommendations

Security that unblocks the true potential of your data lake

Learn how Azure Databricks helps address the challenges that come with deploying, operating and securing a cloud-native data analytics platform at scale.

Bring your own network.

What does the Azure Databricks platform architecture look like, and how you can set it up in your own enterprise-managed virtual network, in order to perform necessary customizations as required by your network security team.

Enable secure cluster connectivity.

Deploy your Azure Databricks workspace in private subnets without any inbound access to your network. Clusters will utilize a secure connectivity mechanism to communicate with the Azure Databricks infrastructure, without requiring public IP addresses for the nodes.

Trust but verify with Azure Databricks

Get visibility into relevant platform activity in terms of who's doing what and when, by configuring Azure Databricks Diagnostic Logs and other related audit logs in the Azure Cloud.

Securely accessing Azure Data sources from Azure Databricks

Understand the different ways of connecting Azure Databricks clusters in your private virtual network to your Azure Data Sources in a cloud-native secure manner.

Data exfiltration protection with Azure Databricks

Learn how to utilize cloud-native security constructs to create a battle-tested secure architecture for your Azure Databricks environment, that helps you prevent Data Exfiltration. Most relevant for organizations working with personally identifiable information (PII), protected health information (PHI) and other types of sensitive data.

Enable customer-managed key for managed services.

Azure Databricks notebooks are stored in the scalable management layer powered by Microsoft and are by default encrypted with a Microsoft-managed key. You can also bring your own-managed per-workspace key to encrypt the notebooks.

Enable customer-managed key for DBFS.

Azure Databricks creates a root storage account, DBFS, (Databricks file system), per workspace in customer's subscription. By default, the storage account is encrypted with a Microsoft-managed key. You can also bring your own-managed key to encrypt the DBFS storage account.

Simplify data lake access with Azure AD Credential Passthrough

Control who has access to what data by using seamless identity federation with Azure AD (Azure Active Directory) under the hood and get cloud-native visibility into who is processing the data and when. Please feel free to refer to cloud-native access control for ADLS (Azure Data Lake Storage) Gen 2 and how to configure it using Azure Storage Explorer. Such access management controls, including role-based access controls, are seamlessly utilized by Azure Databricks.

Authenticate using Azure Active Directory tokens.

Wherever possible, use Azure Active Directory (AAD) tokens to utilize the non-UI capabilities of your Azure Databricks workspace, including REST API, Power BI connectivity and Databricks Connect. For running jobs workloads with REST API, we recommend using Azure Service Principals with AAD Tokens.

Token management for Personal Access Tokens

For use cases where you have to use the Azure Databricks Personal Access Tokens (PAT), we recommend allowing only the required users to be able to configure those tokens. If you cannot use AAD tokens for your job's workloads, we recommend creating PAT tokens for service principals rather than individual users.

Azure Databricks is HITRUST CSF Certified

Azure Databricks is HITRUST CSF Certified to meet the required level of security and risk controls to support the regulatory requirements of our customers. It is in addition to the HIPAA compliance that is applicable through Microsoft Azure BAA.

Control which networks are allowed to access a workspace

Configure allow-lists and block-lists to control the networks that are allowed to access your Azure Databricks workspace.

©Mitre Attack Threat Analysis

Threat Type	Techniques
Cloud Storage Object Discovery	**ID:** T1619 Adversaries may enumerate objects in cloud storage infrastructure. Adversaries may use this information during automated discovery to shape follow-on behaviors, including requesting all or specific objects from cloud storage. Similar to File and Directory Discovery on a local host, after identifying available storage services (i.e. Cloud Infrastructure Discovery) adversaries may access the contents/objects stored in cloud infrastructure.
Data from Cloud Storage Object	**ID:** T1530 Adversaries may access data objects from improperly secured cloud storage. Many cloud service providers offer solutions for online data storage such as Amazon S3, Azure Storage, and Google Cloud Storage. These solutions differ from other storage solutions (such as SQL or Elasticsearch) in that there is no overarching application. Data from these solutions can be retrieved directly using the cloud provider's APIs. Solution providers typically offer security guides to help end users configure systems.[1][2][3] Misconfiguration by end users is a common problem. There have been numerous incidents where cloud storage has been improperly secured (typically by unintentionally allowing public access by unauthenticated users or overly-broad access by all users), allowing open access to credit cards, personally identifiable information, medical records, and other sensitive information.[4][5][6] Adversaries may also obtain leaked credentials in source repositories, logs, or other means as a way to gain access to cloud storage objects that have access permission controls.
Cloud Infrastructure Discovery	**ID:** T1580 An adversary may attempt to discover infrastructure and resources that are available within an infrastructure-as-a-service (IaaS) environment. This includes compute service resources such as instances, virtual machines, and snapshots as well as resources of other services including the storage and database services.

Transfer Data to Cloud Account	**ID:** T1537 Adversaries may exfiltrate data by transferring the data, including backups of cloud environments, to another cloud account they control on the same service to avoid typical file transfers/downloads and network-based exfiltration detection. A defender who is monitoring for large transfers to outside the cloud environment through normal file transfers or over command and control channels may not be watching for data transfers to another account within the same cloud provider. Such transfers may utilize existing cloud provider APIs and the internal address space of the cloud provider to blend into normal traffic or avoid data transfers over external network interfaces. Incidents have been observed where adversaries have created backups of cloud instances and transferred them to separate accounts.

Microsoft STRIDE® Threat Model

Threat	What Can Happen?	Technical Security Controls
Spoofing	Brute force and dictionary attacks can occur. Weak credentials, session token invalidation, Multiple sessions.	• Implement multiple authentication factor MFA, Limit failed login attempts, Passwords with a strong structure (uppercase, lowercase, numbers and special characters) and with a minimum of 14 characters. • Also expire session tokens after a certain time. • Role-based Access control, control Access to Azure Data Lake Storage using Azure AD credentials. • Customer managed keys to encrypt your root Azure Blob storage (root DBFS and workspace system data)
Tampering	Bypassing input data validation, which can favor code injection	• The API must validate the input data according to the expected data type.

Repudiation	Audit weaknesses.	• Any action performed by a user within the system must be recorded in logs.
Disclosure of information	Data transmission in clear text. Improper use of cryptographic keys. Lack of access control to the application	• Sensitive data must be encrypted, the channel must be encrypted with TLS 1.2 or higher, implementing digital certificates with verified entities. • For authentication and authorization with Oauth2.0 and JWT token. • Deploy an Azure Databricks workspace in your own virtual network that you manage in your Azure subscription. • Implement IP whitelist access control in the WAF. • Encrypt notebook and secret data in the control plane with an Azure Key Vault key that you manage. • Deploy your Azure Databricks workspace in private subnets without any inbound access to your network. Clusters will use a secure connectivity mechanism to communicate with the Azure Databricks infrastructure, without requiring public IP addresses for the nodes.
Denial of Service	Service unavailability.	• Through a WAF AZURE or AWS capabilities, implement tools to prevent a denial of service attack from being successful. As for example blocking of IP's with unusual behavior.
Elevation of Privilege	Access control failures, that a user has more permissions than required, outdated systems, unnecessary active ports or services, default accounts and passwords.	• Constantly review and update recommendations for patches and updates, always abide by the principle of least privilege, implementing role modeling for users who access the system, and eliminate accounts by default.

Revision Date	Revised by
September 2022	Security Architect