Data Ingestion and Key Vault

Robin Lester

Microsoft

Who wants your data?

Competitors

- For financial gain

Hackers

- For financial gain
  - Selling customer information
    - Fraudulent purposes
    - Competitors stealing customers

Internal divisions or employees

- Conflict of interest in marketing etc
- Influence the allocation of funds
- Eroding customer rights by using data in non agreed upon activities

Public interest

- Is the company doing ethical things or negative newsworthy activities

Investors

- Insider trading

Political organizations

- Foreign and domestic (political leanings etc)

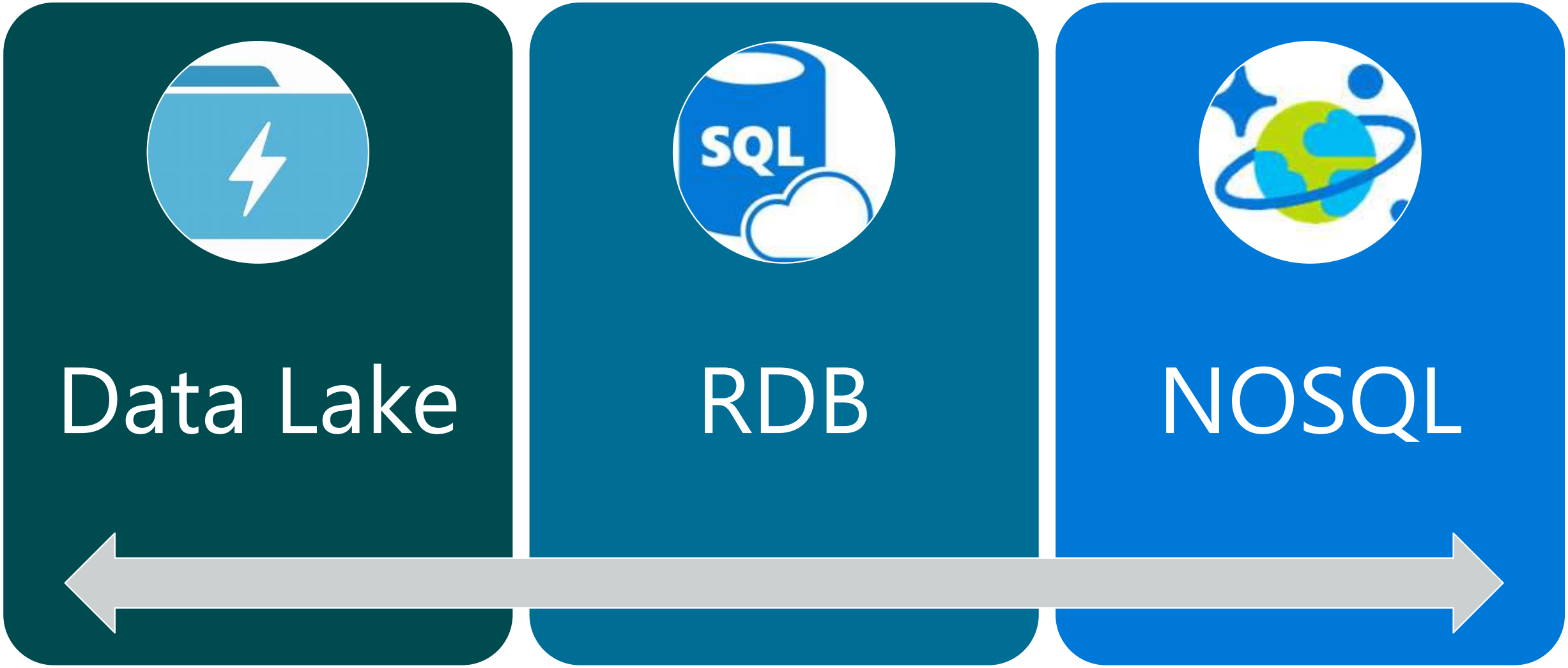
Security on data is also about ethics

- Incorrect data handling
  - GDPR
- Inflicting harm
  - Harms modelling

https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

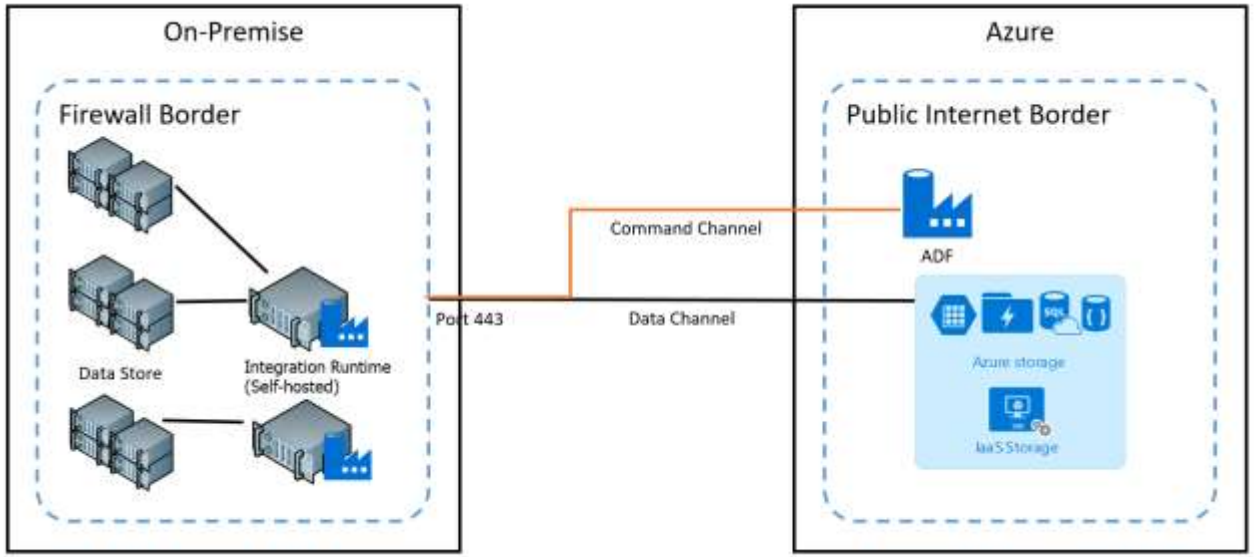
- Should your data scientists be trusted with sensitive data?
  - It puts them in danger
  - Temptation for project not fully thought out from an ethical perspective

Typical Data Repositories



Data Factory

- Cloud Credentials
  - Data store credentials
    - Default = encrypted with certificates
      - Rotated every 2 years
    - Store in own key vault
- On Prem Credentials
  - Stored on self-hosted integration runtime
  - Key Vault
- Data Movement
  - HTTPS or TLS(v1.2)



Azure Data Lake Storage Gen2

A "no-compromises" Data Lake: secure, performant, massively-scalable Data Lake storage that brings the cost and scale profile of object storage together with the performance and analytics feature set of data lake storage

SECURE

- ✓ Support for fine-grained ACLs, protecting data at the file and folder level
- ✓ Multi-layered protection via at-rest Storage Service encryption and Azure Active Directory integration

MANAGEABLE

- ✓ Automated Lifecycle Policy Management
- ✓ Object Level tiering

FAST

- ✓ Atomic file operations means jobs complete faster

SCALABLE

- ✓ No limits on data store size
- ✓ Global footprint (50 regions)

COST EFFECTIVE

- ✓ Object store pricing levels
- ✓ File system operations minimize transactions required for job completion

INTEGRATION READY

- ✓ Optimized for Spark and Hadoop Analytic Engines
- ✓ Tightly integrated with Azure end to end analytics solutions

Data Lake Gen 2 Security

- Role-based access control
  - Apply sets of permissions to security principals
- Shared Key and Shared Access Signature (SAS) authentication
  - Shared key = super-user
  - SAS
    - Limited access
    - Time bound
- Access control lists on files and directories
  - Associate a security principal with an access level for files and directories

	Directory	
File		
Read (R)	Can read the contents of a file	Requires Read and Execute to list the contents of the directory
Write (W)	Can write or append to a file	Requires Write and Execute to create child items in a directory
Execute (X)	Does not mean anything in the context of Data Lake Storage Gen2	Required to traverse the child items of a directory

Data Lake Considerations

- Data lakes are immutable
- Data is copied multiple time in multiple formats
- Watch out for data leaks
- Data can be reconstituted from multiple sources

Advanced Threat Protection for Azure Storage

- Identify suspicious user and device activity with both known-technique detection and behavioral analytics

Microsoft Azure

MEDIUM SEVERITY

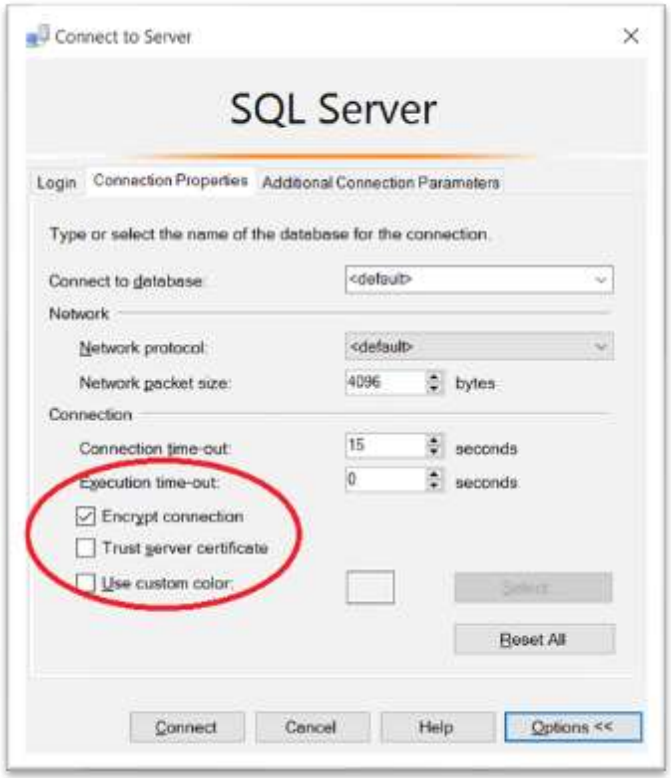
Someone has accessed your Storage account 'mystorageaccount' from an unusual location.

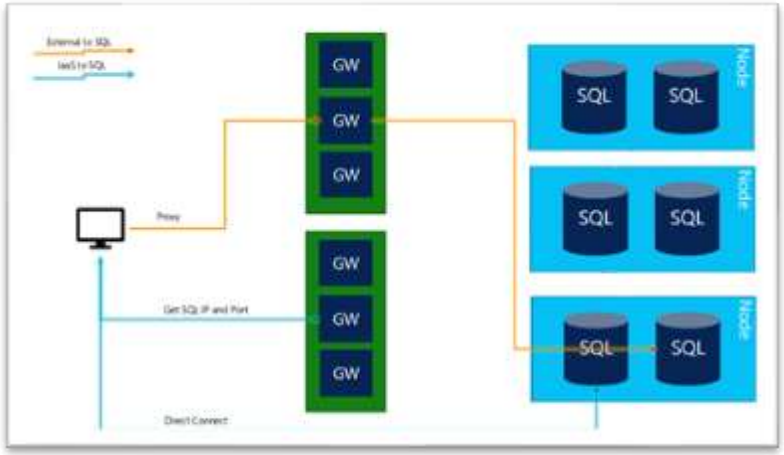
Activity details

Subscription ID	[redacted]
Storage account	mystorageacount
Storage type	Blob
Container	mycontainer
Application	myTestApplication
IP address	[redacted]
Location	Washington, United States
Data center	scus
Date	May 17, 2018 7:50 UTC
Potential causes	Unauthorized access that exploits an opening in the firewall. Legitimate access from a new location
Investigation steps	For a full investigation, configure diagnostics logs for read, write, and delete
Remediation steps	Be sure to follow the principle of "least privilege" and limit access to your data

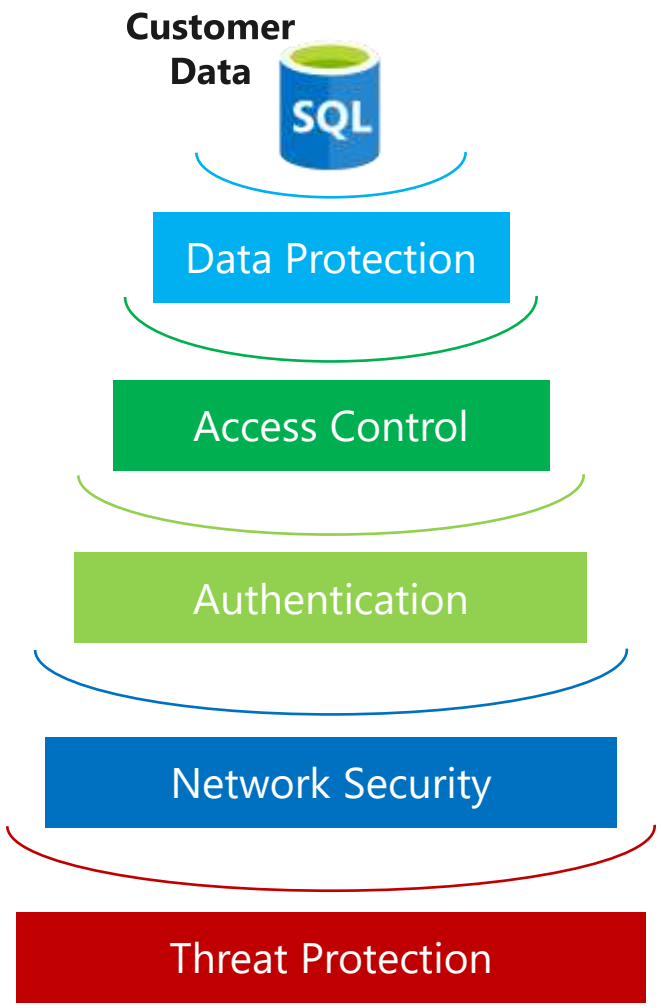
Data in SQL

- Data security
  - Always Encrypted
  - RLS/CLS
  - Dynamic Data Masking
  - Cell level encryption
  - Transparent Data Encryption
- Connection security
  - MFA
  - SQL Authentication
  - AAD authentication
  - Server certificate (MITM attacks)
- Firewall and network security
  - Private link
  - SQL endpoints
  - Database and Server Firewalls
  - Database port abstraction





Enterprise Grade Security that is Easy-to Use



Encryption-in-flight (Transport Layer Security TLS)
Encryption-at-rest (Transparent Data Encryption TDE)
Service- or User-managed keys, Backup encryption
Encryption-in-use (Always Encrypted)
Data Masking (dynamic, static)
Data Discovery and Classification

SQL Permissions
Row-level security
Column-level security

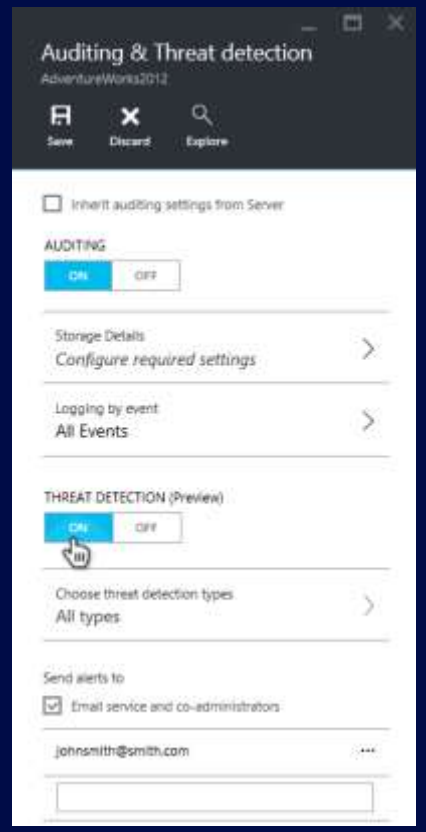
SQL Authentication
Azure Active Directory Authentication (w/ MFA)

Virtual Networks
SQL Firewall (server- and database-level)

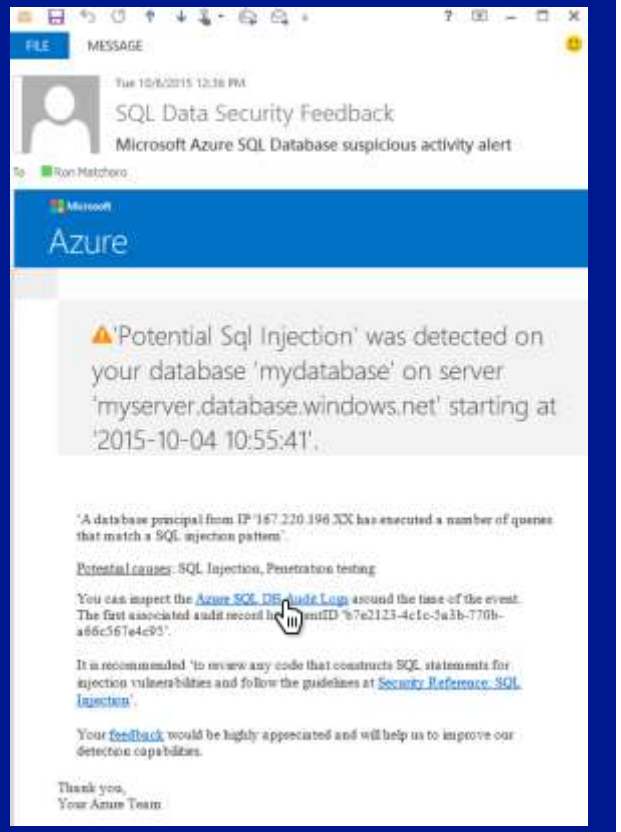
Data Classification
SQL Threat Detection
SQL Auditing
Vulnerability Assessment

Threat Detection

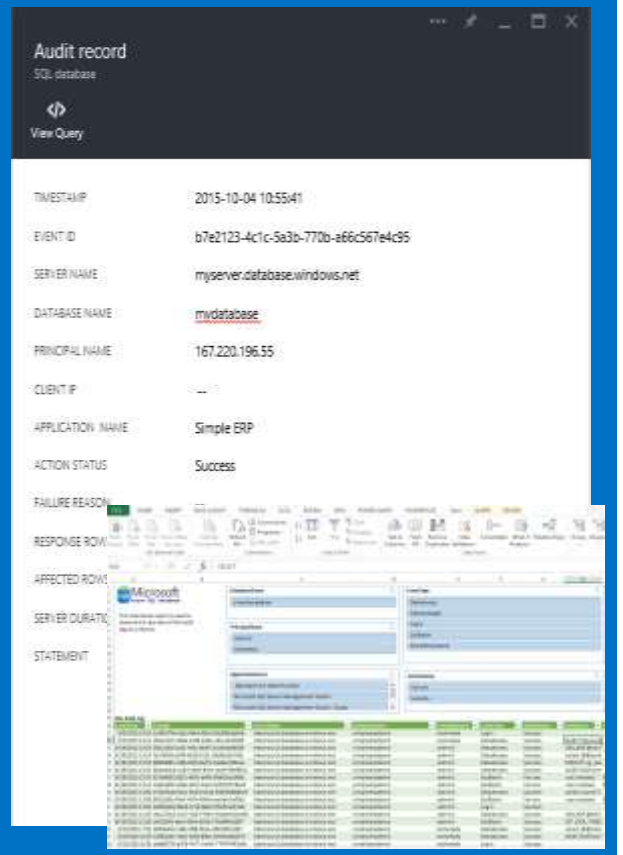
Set up



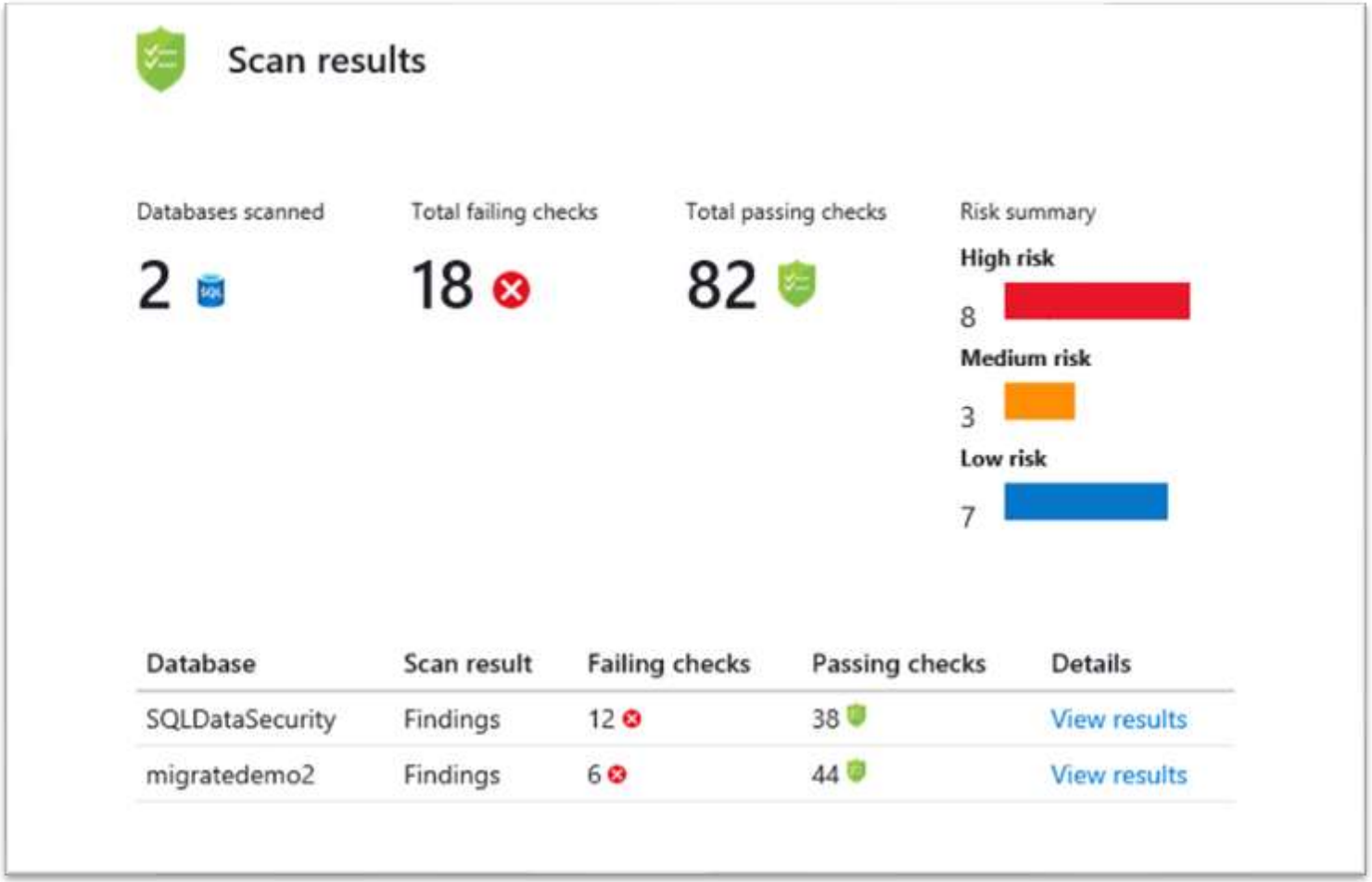
Alert



Explore



Vulnerability Assessment



Demo

Azure Networking Services

Protecting your ingestion data

Virtual Network

Virtual WAN

ExpressRoute

VPN

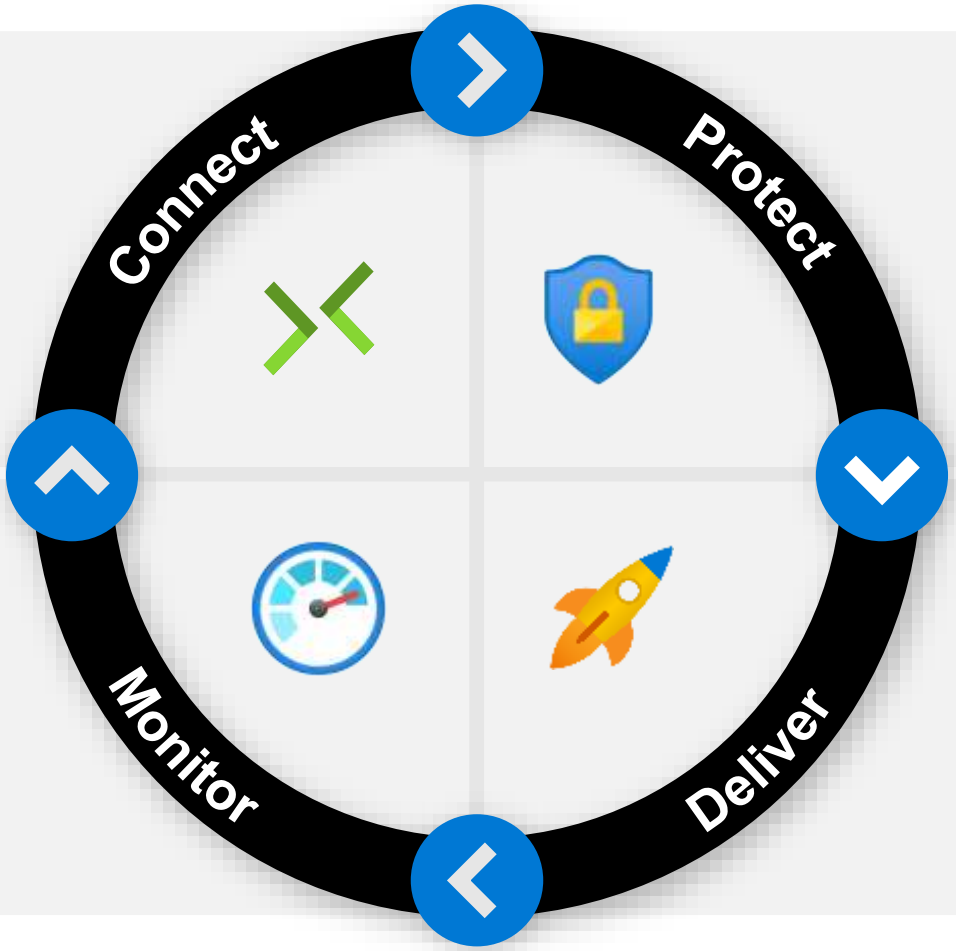
DNS

Network Watcher

ExpressRoute Monitor

Azure Monitor

Virtual Network TAP



DDoS Protection

Azure Firewall

Network Security Groups

Web Application Firewall

Service Endpoints

Azure Private Link

Azure Bastion

CDN

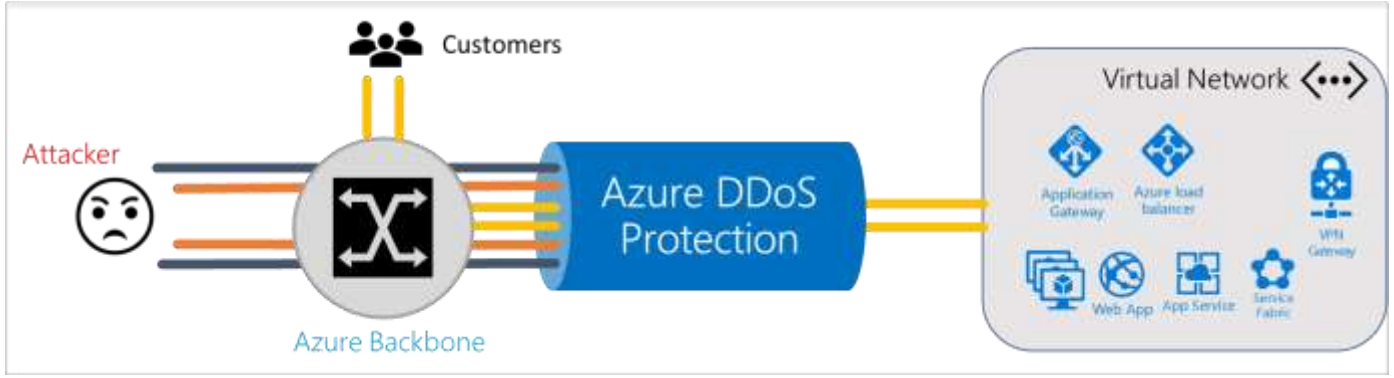
Front Door

Traffic Manager

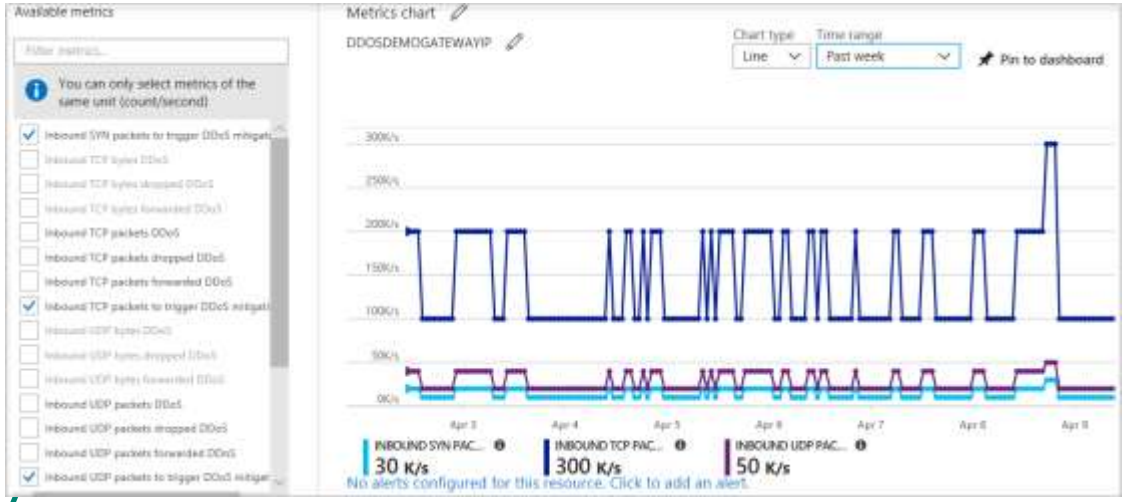
Application Gateway

Load Balancer

DDOS

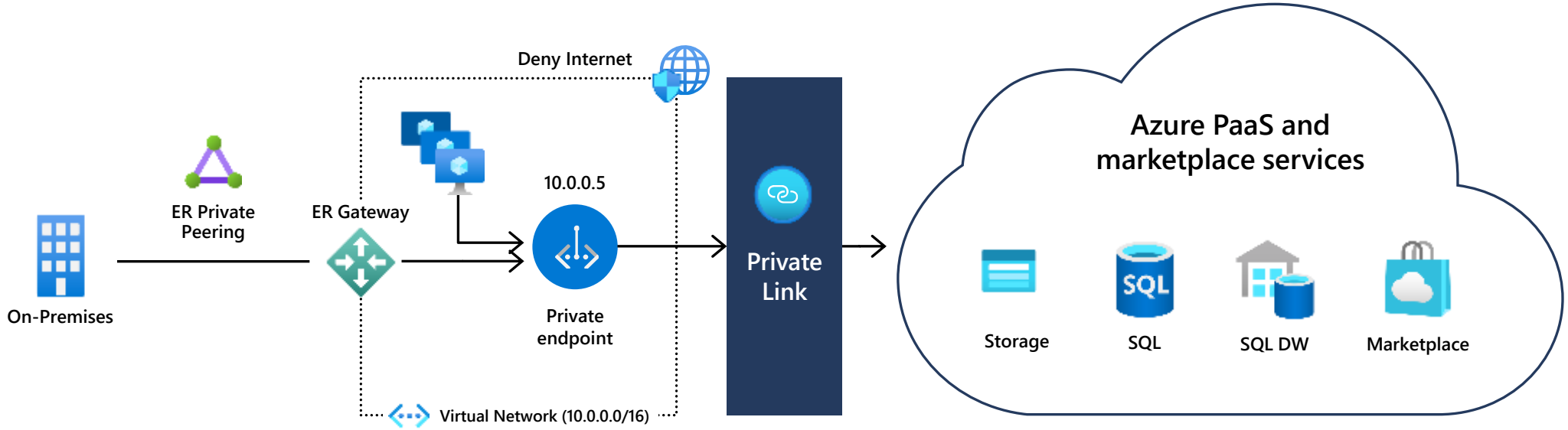


- DDoS Protection Basic
  - Free service
    - Always-on traffic monitoring and real-time mitigation
- DDoS Protection Standard
  - Automatically tuned to help protect your specific Azure resources in a virtual network
    - Logging
    - Alerting
    - Telemetry
  - Automatic learning per IP
  - DDoS mitigation policies



https://www.digitalattackmap.com/

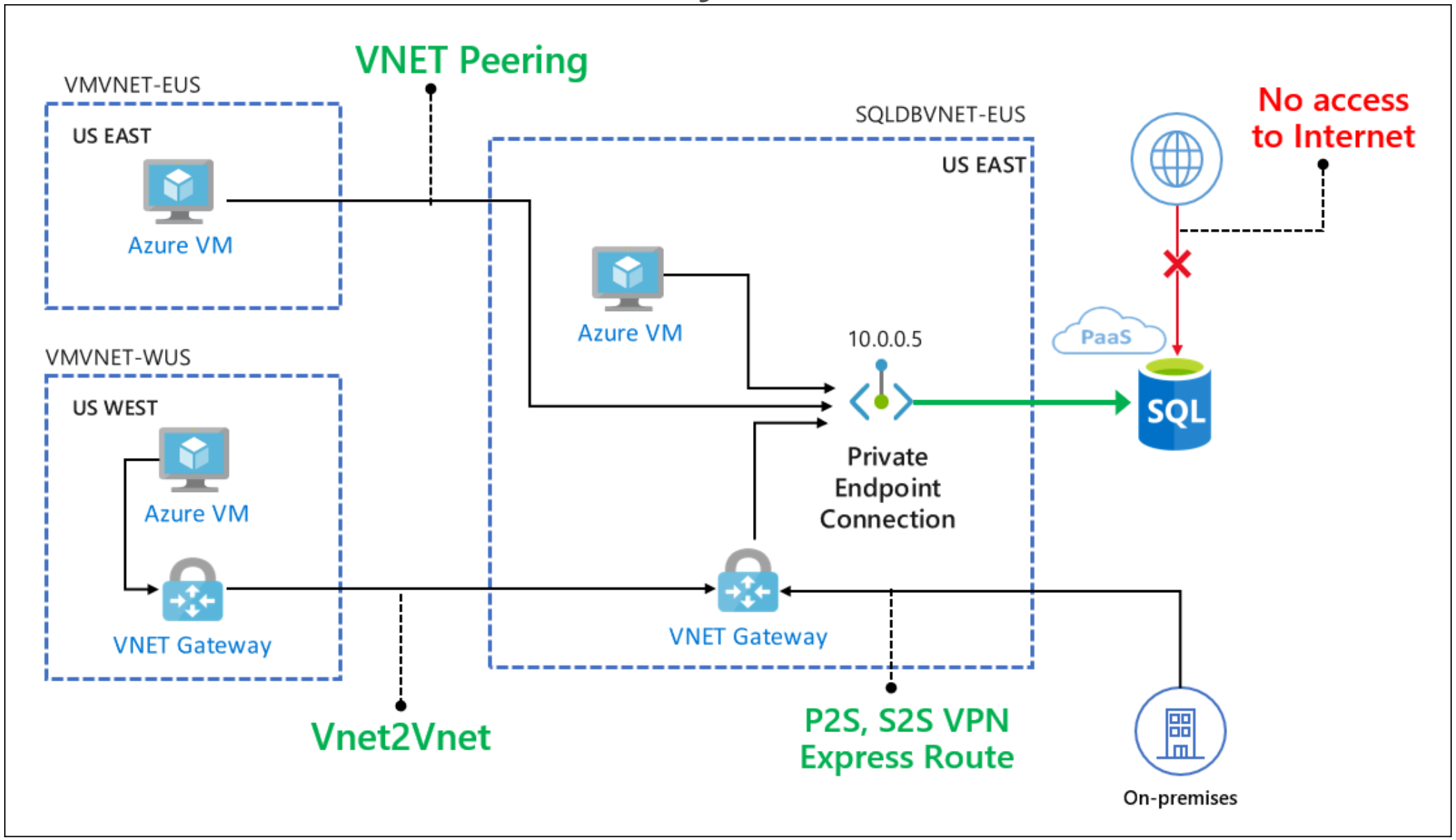
Azure Private Link



Private Link for Azure Storage, SQL DB and customer own service

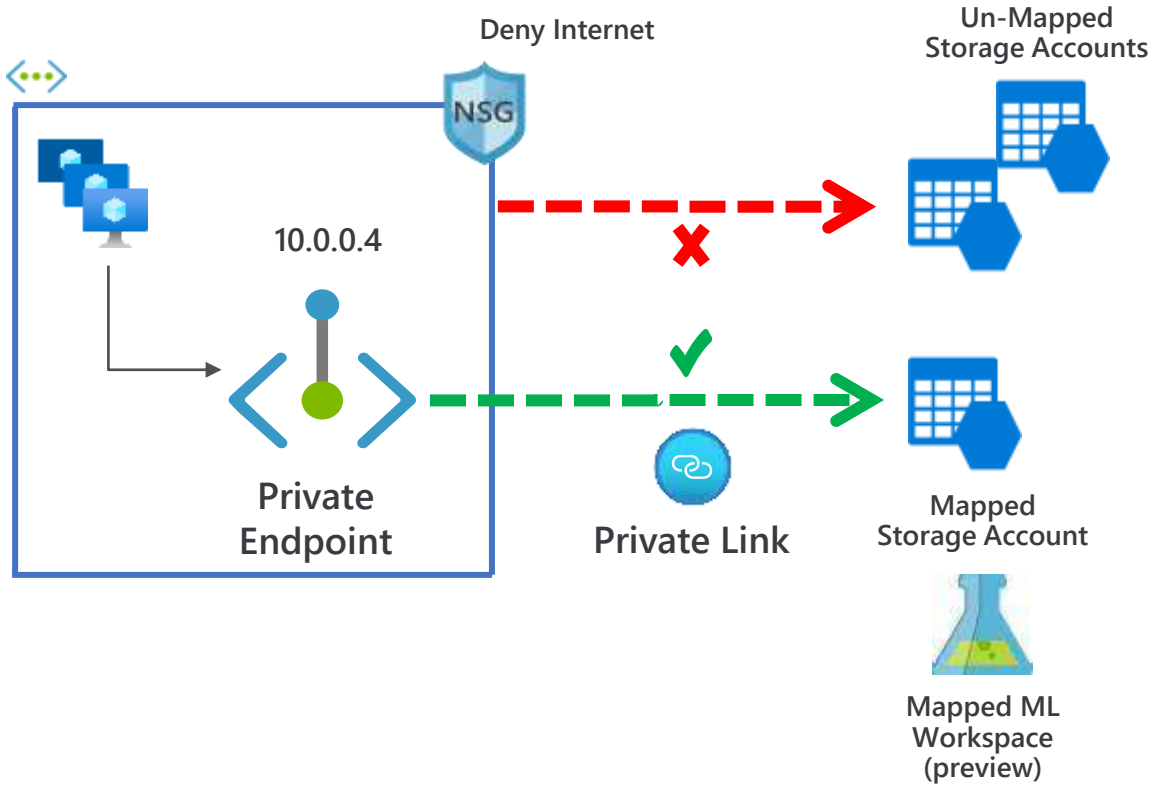
Private access from Virtual Network resources, peered networks and on-premise networks	In-built Data Exfiltration Protection	Predictable private IP addresses for PaaS resources	Unified experience across PaaS, Customer Owned and marketplace Services

Private Link – Connectivity scenarios



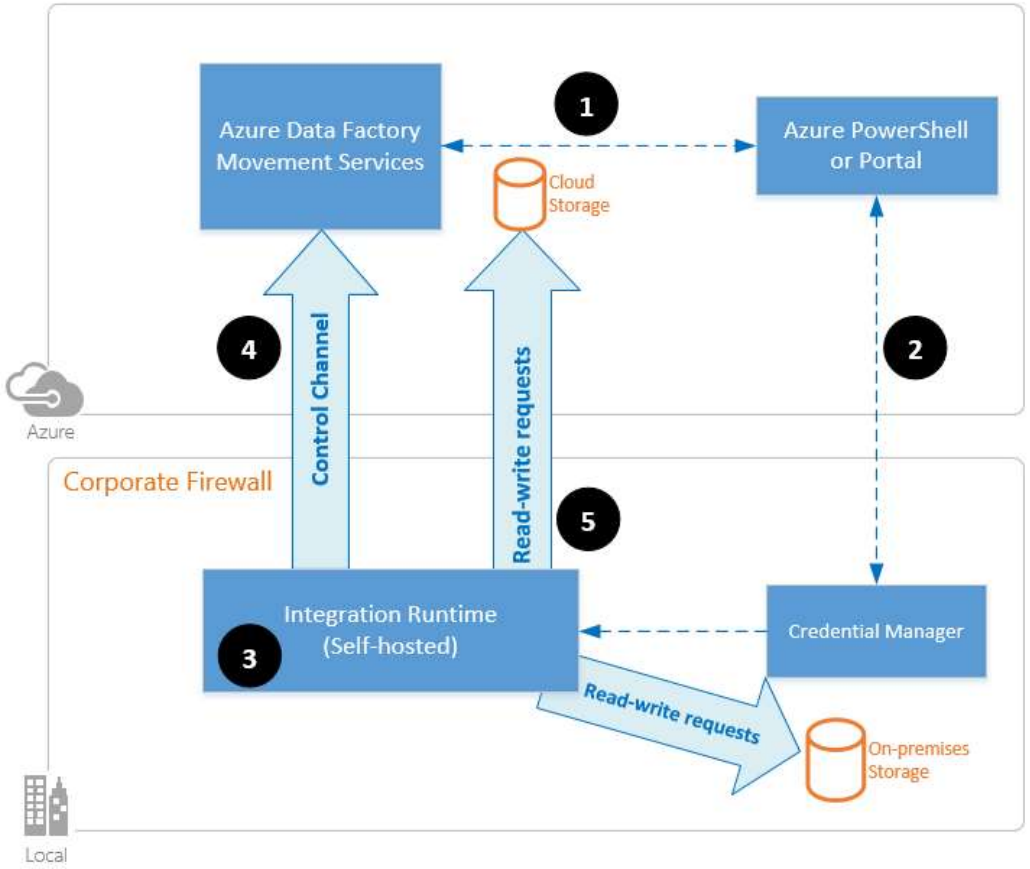
Data Exfiltration Protection

- Private Endpoint maps specific PaaS resource to an IP address, not the entire service
- Access only to mapped PaaS resource
- Data exfiltration protection is in-built



Other network conidiations

- Express Route
- Point/Site to site VPN
- Gateways and integration runtimes (Data Factory)
  - 3 - credentials are stored locally



Data in Cosmos DB

- Encryption at rest
- HTTPS/SSL/TLS encryption
- Auditing
- AAD integration
  - RBAC
    - Cosmos account
    - Database
    - Container
    - Offers (throughput)

Data Security

- When I am developing / training an ML model where do I put my blob connection keys, database connection strings?
  - In a Python function?
  - In a notebook?
  - In a secure vault?

Azure Key Vault

Organizations need to safeguard certificates deployed into their VMs.

1

Developers need to safeguard config secrets of their Azure cloud services.

e.g. Storage account key
SQL connection string

2

Organizations need to control encryption keys used by their OWN apps.

3

Organizations need to control encryption keys used by SaaS services.

4

A vault needs to have:

- Secrets and Keys encrypted at rest
- Choice of deployment country
- Choice of encryption method (Software vs Hardware)
- Security module separation
  - Create a many vaults as you like
- Easy access and rights control
  - Azure AD / RBAC / Firewall

What is Azure Key Vault?

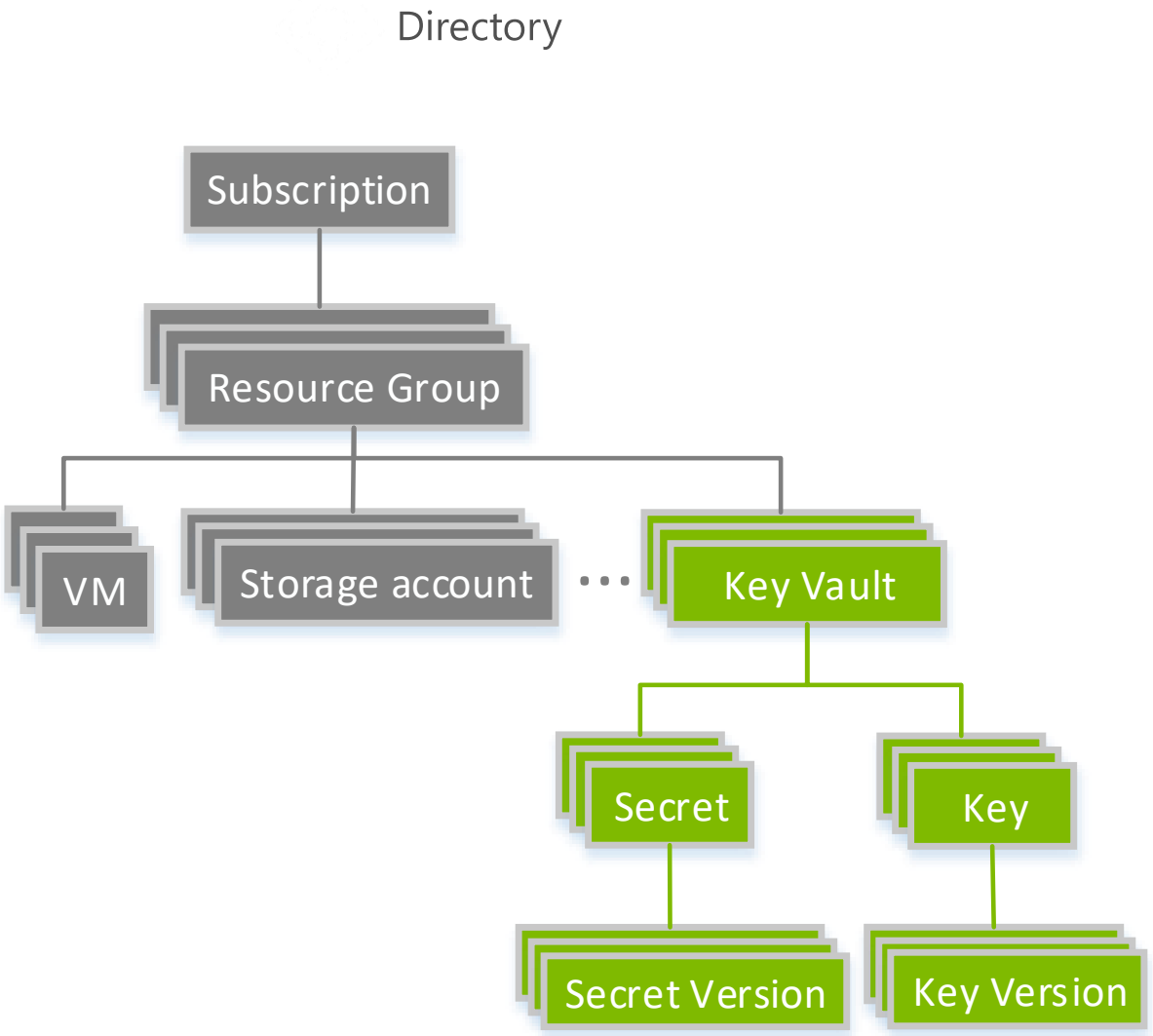
- An Azure resource provider that lets you
  - Store & manage SECRETS, and release them at runtime to authorized apps & users.
  - Store & manage KEYS, and perform cryptographic operations on behalf of authorized apps & users.

- Backed by Hardware Security Modules
  - All secrets and keys are protected at rest with key chain terminating in HSMs.
  - Keys marked as ‘HSM-protected’ are protected even at runtime with HSMs.

Terminology

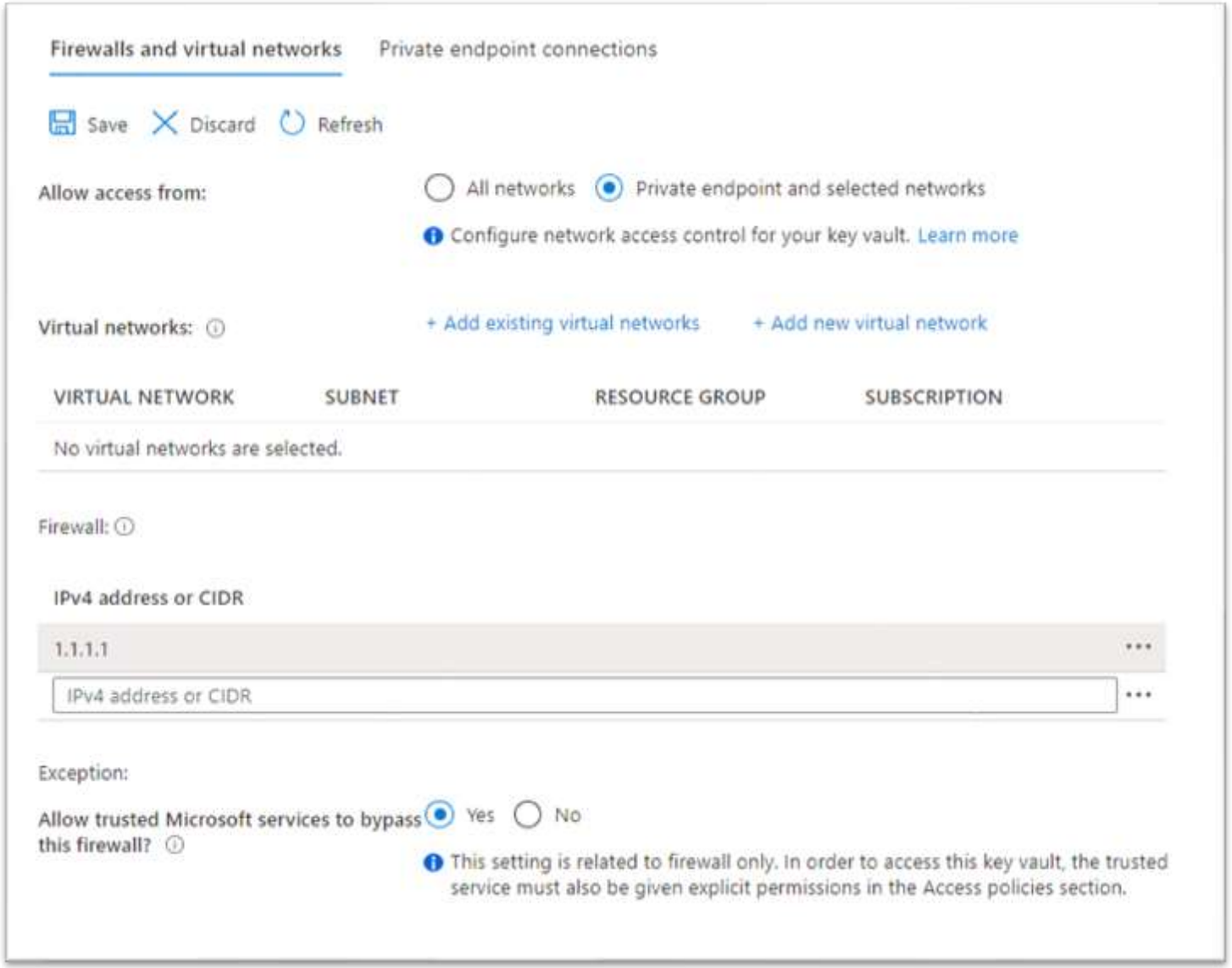
- Key Vault
  - Container for related keys and secrets that are managed together.
  - Unit of access control, unit of billing.
  - An Azure resource, like a storage account.
- Secret
  - What: Any sequence of bytes under 25KB. E.g. SQL connection string, Storage account key.
  - How used: Authorized users/apps write and read back the secret value.
- Key
  - What: A cryptographic key. RSA 2048.
  - How used: A key cannot be read back. Caller must ask the service to decrypt / sign with the key.

Key Vault within Azure object model



Additional Security Features

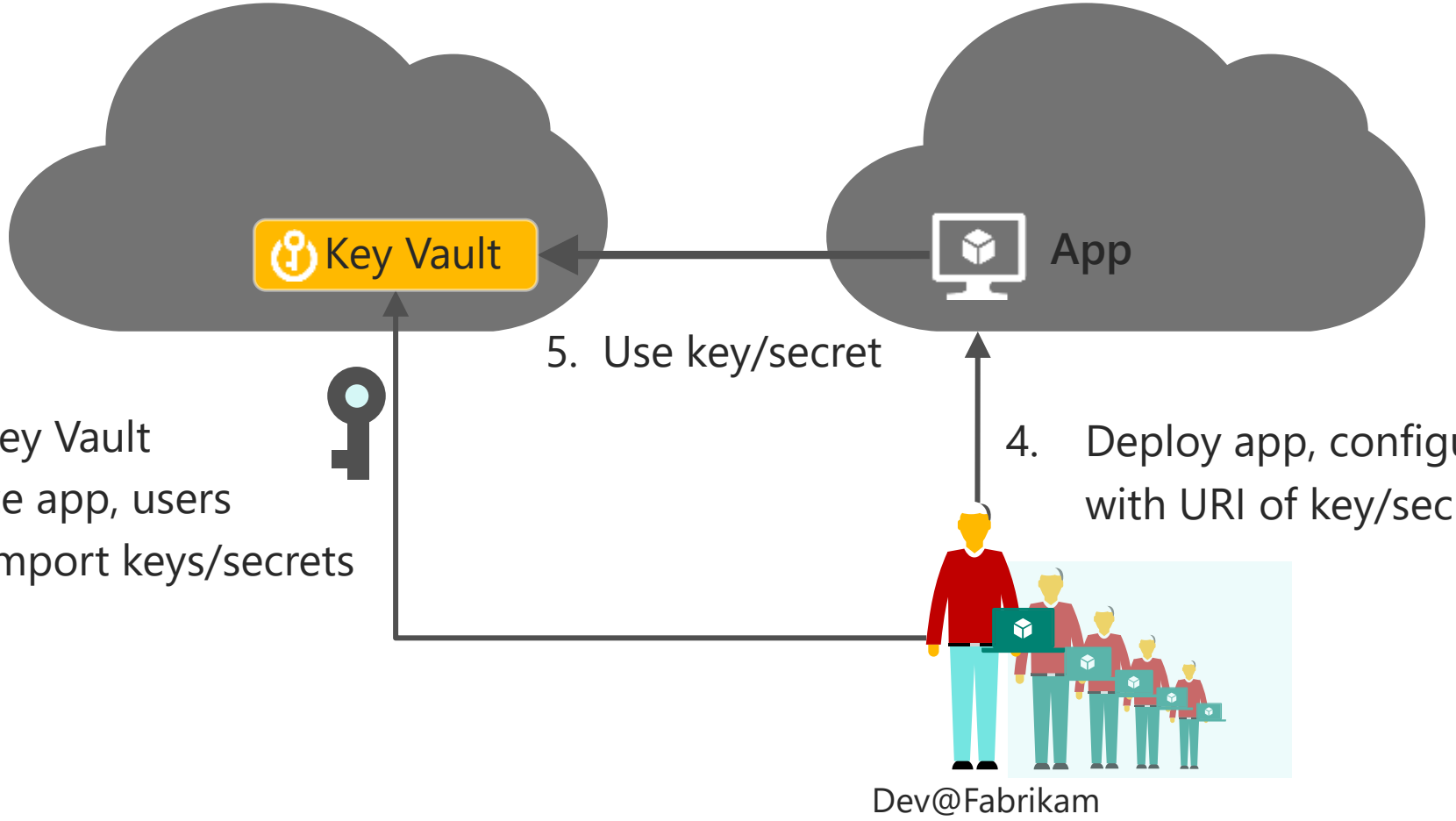
- Firewall



Additional Security Features

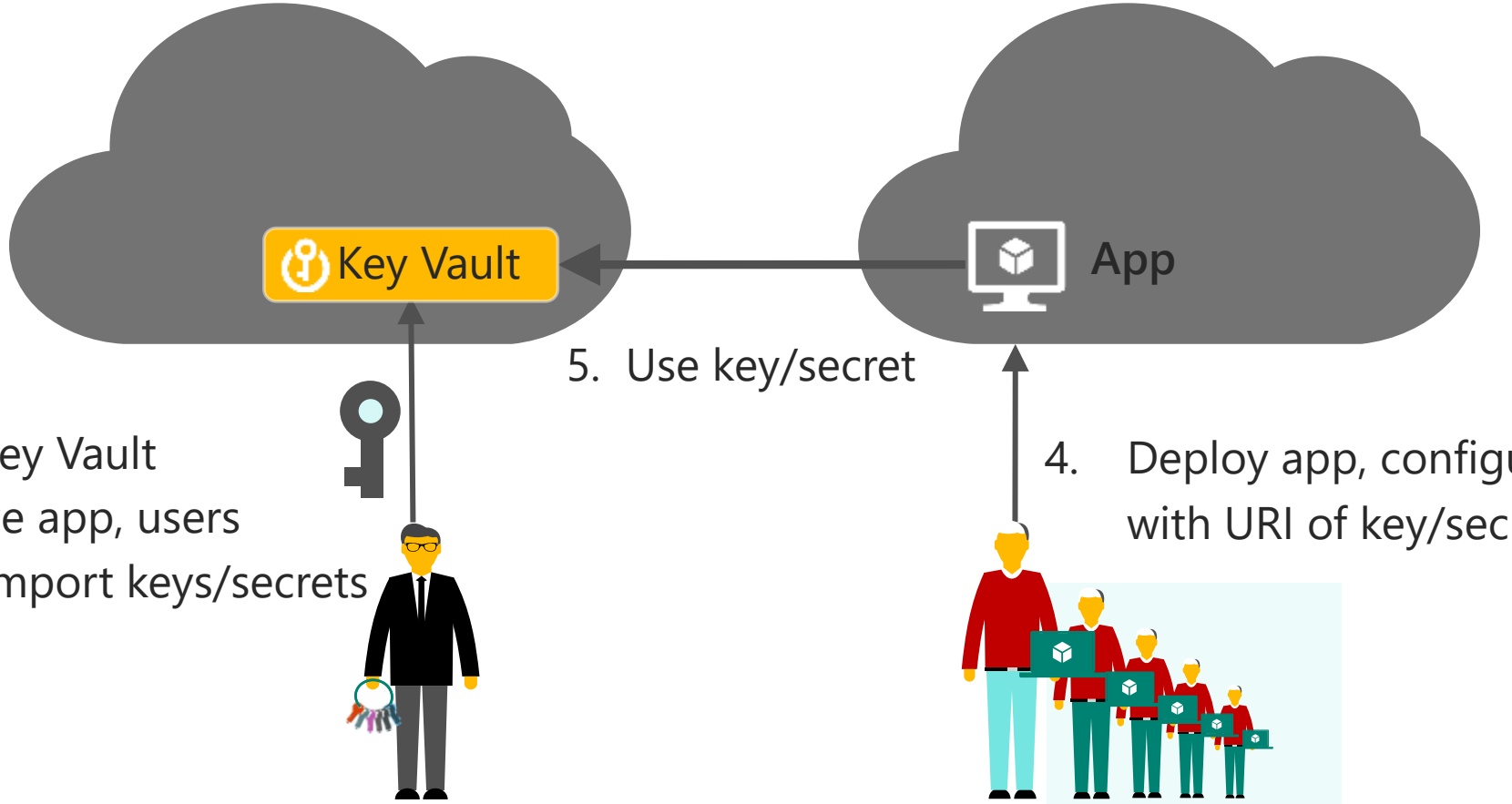
- Auditing of access
- AAD integration
- Storage account key rotation
- Store multiple key versions with start and stop date

Phase 1: Developer builds/tests application



1. Create Key Vault
2. Authorize app, users
3. Create/import keys/secrets

Phase 2: App moves into pilot / pre-prod



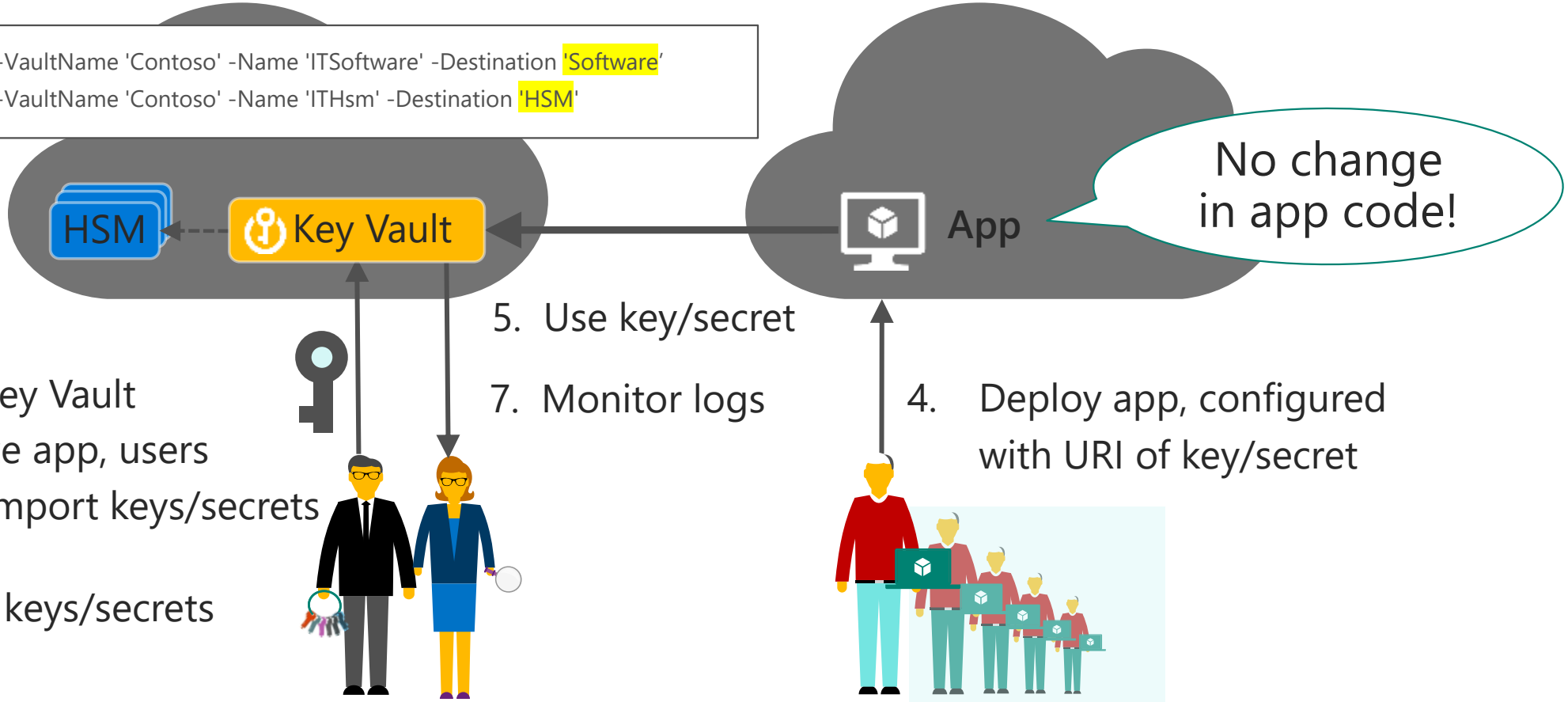
1. Create Key Vault
2. Authorize app, users
3. Create/import keys/secrets

Privileged@Fabrikam

Dev@Fabrikam

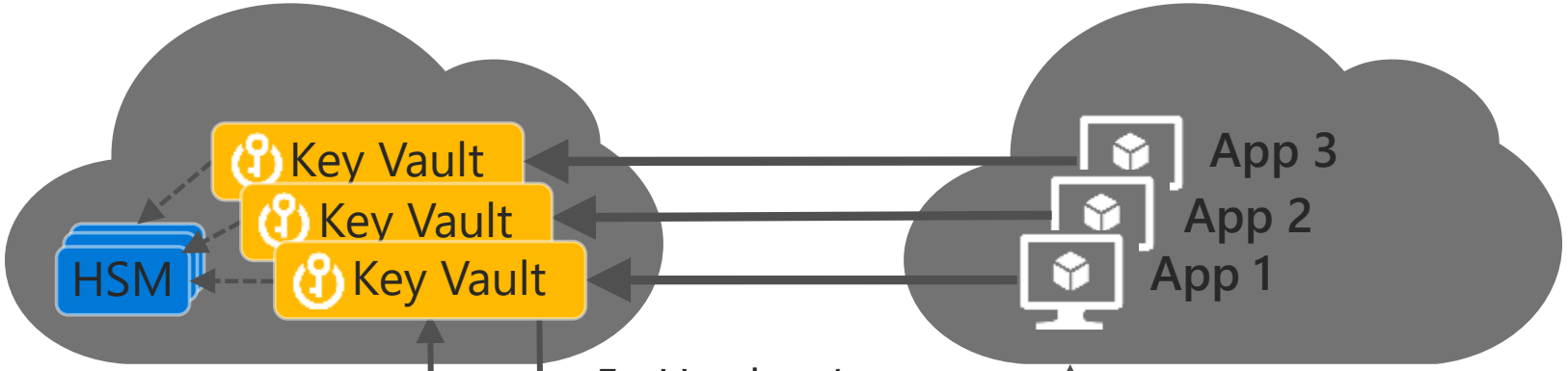
Phase 3: App moves into production

```
Add-AzureKeyVaultKey -VaultName 'Contoso' -Name 'ITSoftware' -Destination 'Software'
Add-AzureKeyVaultKey -VaultName 'Contoso' -Name 'ITHsm' -Destination 'HSM'
```



1. Create Key Vault
2. Authorize app, users
3. Create/import keys/secrets
6. Manage keys/secrets

Phase 4: Scale, deploy more apps in minutes



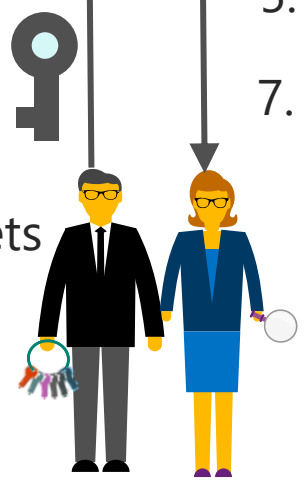
5. Use key/secret

7. Monitor logs

1. Create Key Vault
2. Authorize app, users
3. Create/import keys/secrets

6. Manage keys/secrets

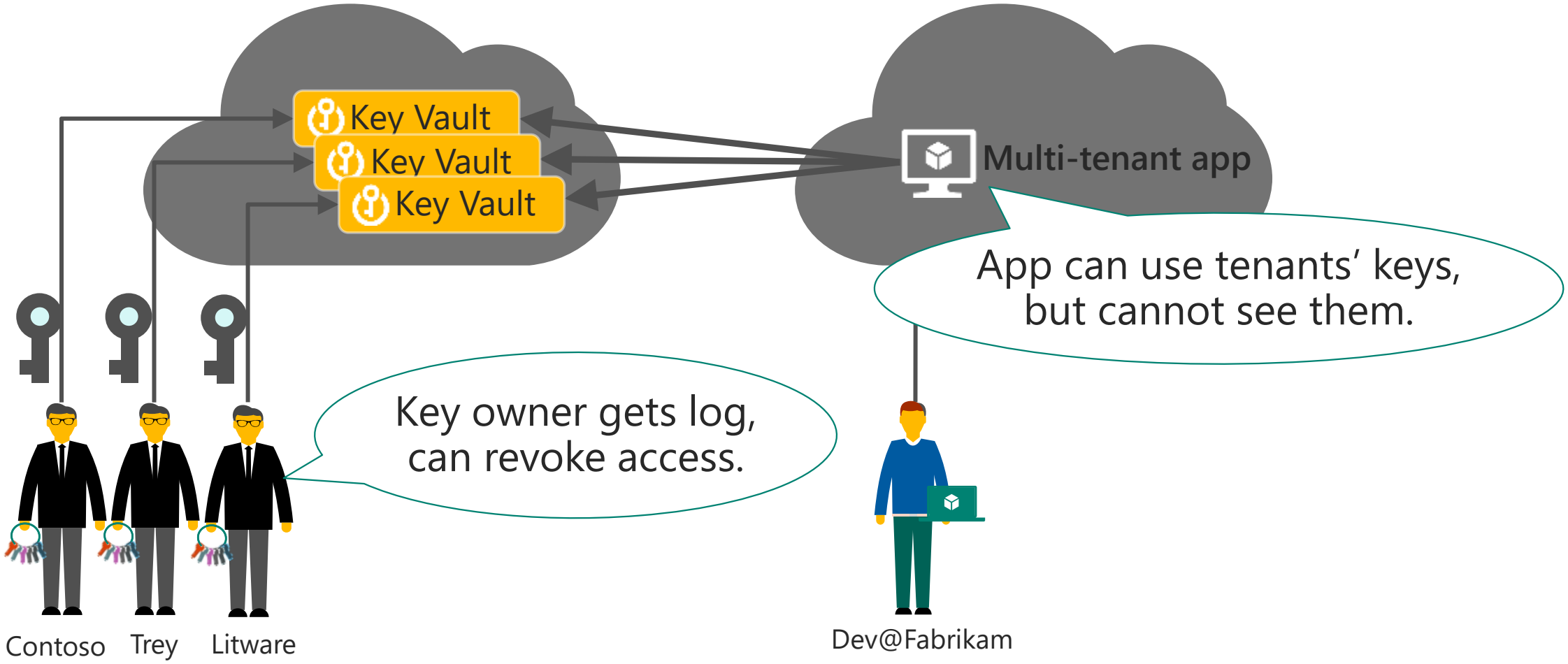
4. Deploy app, configured with URI of key/secret



CISO@Fabrikam

Dev@Fabrikam

Multi-tenant app offers customer-managed keys



Key Vault roles

- Admin
  - Allows Access to Vault
- Key Owner
  - Adds / updates Keys and Secrets
- App Owner
  - Configures app with Service Pinnacle and Secret URI
- Application
  - Holds AAD identity
  - Retrieves the keys
- Auditor
  - Audits and allows access

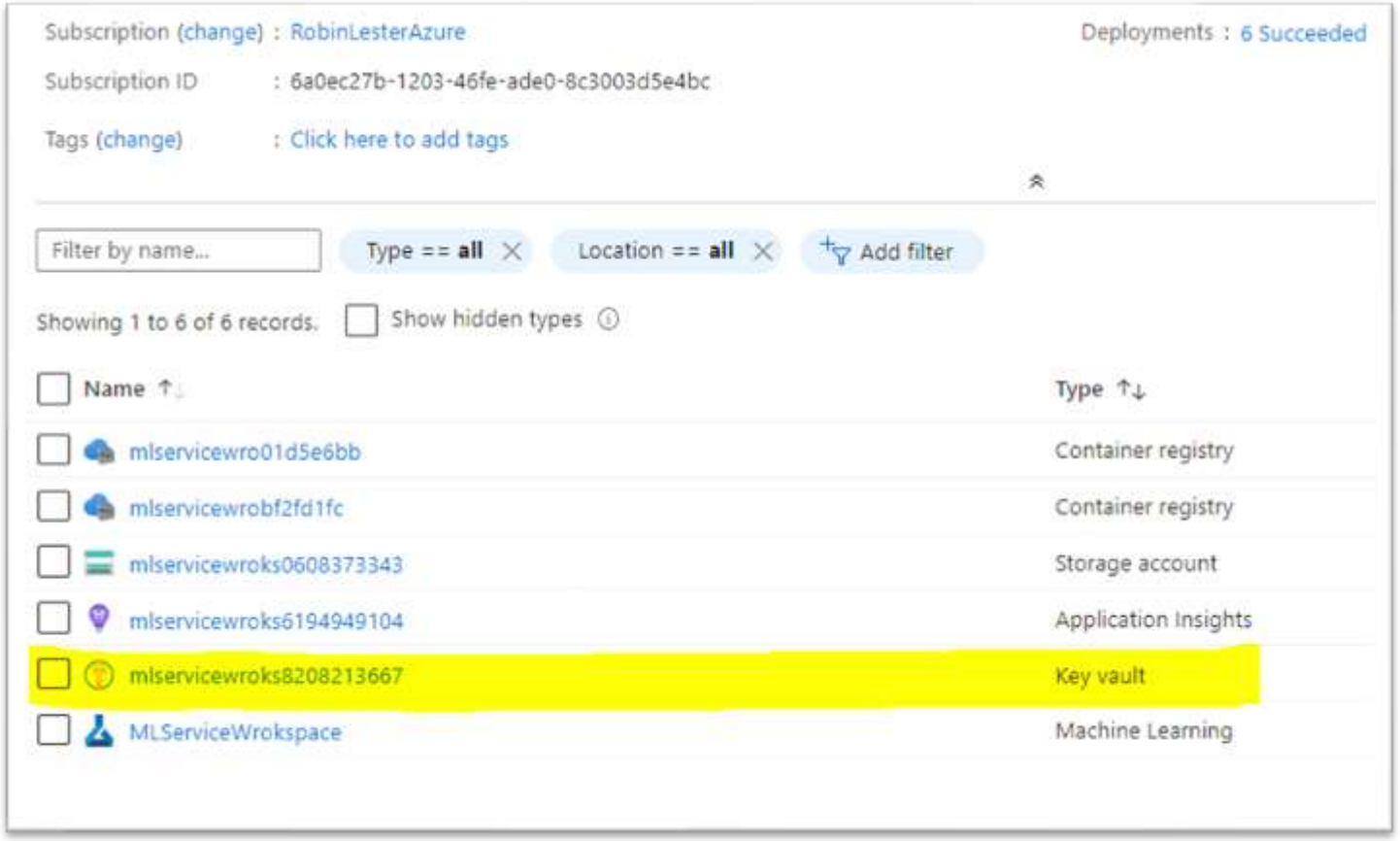
Key vault instance in ML resource group

- The key vault instance that is associated with the workspace is used by Azure Machine Learning to store
  - The associated storage account connection string
  - Passwords to Azure Container Repository instances
  - Connection strings to data stores

b77f06f9-80dd-4231-9153-6a77c4	2-x8TmUs5zJ2egwKLB...	✓ Enabled
b77f06f9-80dd-4231-9153-6	Jd72-zAKFuszYVIptiKw-Wr...	✓ Enabled
b77f06f9-80dd-423	7edd72-zeja6lxllc QrNWdq60T...	✓ Enabled
PythonSecret		✓ Enabled
workspace-secrets-b77f06f9-80dd-4231-	ɜdd72	✓ Enabled

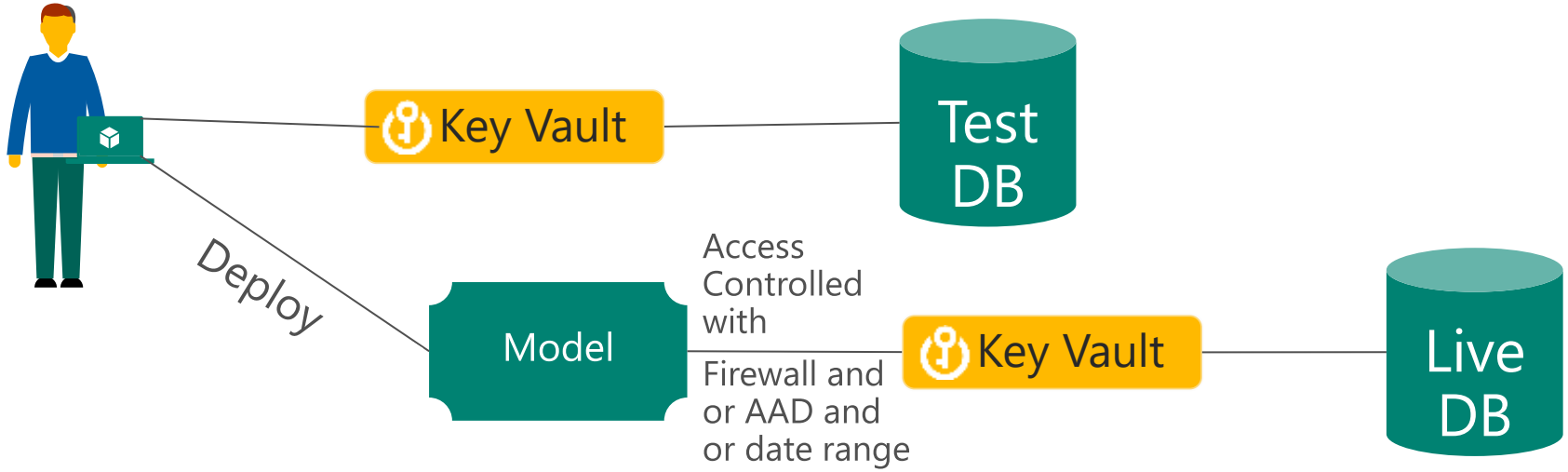
Using Key vault in ML

- R
  - https://cran.r-project.org/web/packages/AzureKeyVault/index.html
- Python
  - https://github.com/Azure/MachineLearningNotebooks/blob/master/how-to-use-azureml/manage-azureml-service/authentication-in-azureml/authentication-in-azureml.ipynb
- Demo



Usage Cases

- Using secrets in training runs
  - Keep your training data secure
  - EG: Securely Connect to a SQL database or storage system
  - https://docs.microsoft.com/en-us/azure/machine-learning/how-to-use-secrets-in-runs
- Using secrets in remote runs / pipelines
  - Allow deployments to access secure data



Data security in ML Service

- Connections
  - Owners and contributors can use all compute targets and data stores that are attached to the workspace
  - ML Service workspace owners and contributors can access attached storage through manged identity
  - Managed identity name is the same as the workspace name

Resource	Permissions
Workspace	Contributor
Storage account	Storage Blob Data Contributor
Key vault	Access to all keys, secrets, certificates
Azure Container Registry	Contributor
Resource group that contains the workspace	Contributor
Resource group that contains the key vault (if different from the one that contains the workspace)	Contributor

Microsoft