**CORE SERVICE DESIGN:**

**Azure Data Factory**

atabricks

|  |  |
| --- | --- |
| IT Owner Details | |
| **Department** | DTS |
| **Contact Name** | Dominic Panzera |
| **Address** | 375 Manningham Road, Doncaster, Victoria 3108 |

|  |  |
| --- | --- |
| Document Control | |
| Title | Ambulance Victoria – Azure Data Factory Core Service Design |
| **File Name** | Ambulance Victoria – Azure Data Factory Core Service Design v2.0.docx |
| **Version** | 2.0 |
| **Status** | Released |
| **Release Date** | 7/03/2024 |

|  |  |  |  |
| --- | --- | --- | --- |
| Preparation |  |  |  |
| **Prepared** | Daniela Nikolic |  |  |
| **Authorised** | Dileep Pradeep |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version Tracking | | | |  |
| Version | Remarks | Change Requested | Pages Affected | Release Date |
| 1.0 | Initial Release | N/A | All | 29/02/2024 |
| 2.0 | Updated | Minor feedback | 13 | 7/03/2024 |

Table of Contents

[1. Overview 5](#_Toc160618544)

[1.1 Purpose and Audience 5](#_Toc160618545)

[1.2 Scope and Key Deliverables 5](#_Toc160618546)

[1.3 Glossary and Definitions 6](#_Toc160618547)

[2. Executive Summary 7](#_Toc160618548)

[3. Resource Cost 8](#_Toc160618549)

[4. WAF and Security Control Alignment 9](#_Toc160618550)

[4.1 Reliability 9](#_Toc160618551)

[4.1.1 Overview 9](#_Toc160618552)

[4.1.2 Azure Data Factory Reliability Checklist 9](#_Toc160618553)

[4.2 Cost Optimisation 10](#_Toc160618554)

[4.2.1 Overview 10](#_Toc160618555)

[4.2.2 Azure Data Factory Cost Optimisation Checklist 10](#_Toc160618556)

[4.3 Operational Excellence 10](#_Toc160618557)

[4.3.1 Overview 10](#_Toc160618558)

[4.3.2 Azure Data Factory Operational Excellence Checklist 10](#_Toc160618559)

[4.4 Performance Efficiency 11](#_Toc160618560)

[4.4.1 Overview 11](#_Toc160618561)

[4.4.2 Azure Data Factory Performance Efficiency Checklist 11](#_Toc160618562)

[4.5 Security 12](#_Toc160618563)

[4.5.1 Overview 12](#_Toc160618564)

[5. Architecture Summary 13](#_Toc160618565)

[5.1 Resource Overview 13](#_Toc160618566)

[5.1.1 Integration Runtime 14](#_Toc160618567)

[5.2 RBAC 15](#_Toc160618568)

[5.3 Design Decisions and Justifications 15](#_Toc160618569)

[5.3.1 Version 15](#_Toc160618570)

[5.3.2 Integration Runtime 15](#_Toc160618571)

[5.3.3 Connectivity 15](#_Toc160618572)

[5.3.4 Managed Identity 16](#_Toc160618573)

[5.3.5 Authentication 16](#_Toc160618574)

[5.3.6 Encryption 16](#_Toc160618575)

[5.3.7 Logging 16](#_Toc160618576)

[5.3.8 Backup 17](#_Toc160618577)

[6. Azure Policies 17](#_Toc160618578)

[7. Configuration Templates 18](#_Toc160618579)

[7.1 Primary Region Azure Data Factory 18](#_Toc160618580)

[7.2 Secondary Region Azure Data Factory 19](#_Toc160618581)

[8. Acceptance 20](#_Toc160618582)

# Overview

This document covers the baseline design for the Azure Data Factory core service. The intention of this document is to define the overall resource design in isolation from a specific application. It is aimed to highlight the general process and requirements for building a Azure Data Factory in a repeatable fashion with consistent configurations. Design decisions and justifications have been included in the Architecture section, and this document can be used as a reference for new builds that require a Azure Data Factory.

This design caters to a Level 2 design which covers both Microsoft’s WAF (Well Architected Framework)[[1]](#footnote-2) and the Department of Health Control list.

Any deviations required to the standards defined in this document will require separate exemption and approval from the Cloud Governance Forum if they are required for any reason for a specific build.

## Purpose and Audience

This document will outline the standard design and configuration of this Azure service in Ambulance Victoria’s Azure tenancy as a baseline for any application infrastructure deployments.

This design is intended to:

* Meet Microsoft WAF standards.
* Meet the controls stipulated by the Department of Health.
* Define the baseline required for the deployment of the resource.

The audience for this document is those involved in the planning, designing, and implementing of the Application/Data infrastructure. This includes:

* + Ambulance Victoria IT staff

It is assumed that the reader knows and is familiar with Azure Cloud concepts and related topics.

## Scope and Key Deliverables

The scope of this core service design is to define the baseline deployment requirements and standards for the Azure Data Factory core service.

The key deliverables for this are:

* This design to outline the service definition Level 2 baseline standards.
* A technical configuration document that defines the deployment of this resource for each of the Service Tiers, or for any other logical standard such as size
* IaC templates for repeatable deployment of this core service

## Glossary and Definitions

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **AV** | Ambulance Victoria |
| **WAF** | Well Architected Framework |
| **CAF** | Cloud Adoption Framework |
| **IR** | Integration Runtime |
| **SSIS** | SQL Server Integration Service |
| **Level 1** | Refers to a resource that has been designed to a CAF standard |
| **Level 2** | Refers to a resource that has been designed to a WAF standard with Department of Health controls overlayed |
| **AZ 2** | Refers to Ambulance Victoria’s legacy Azure Landing Zone still in use in some regards |
| **AZ 3** | Refers to Ambulance Victoria’s current Azure Landing Zone, also referred to as the Enterprise landing zone. This is the target state for migrations. |
| **SLA** | Service Level Agreement as defined by Microsoft |
| **DH** | Department of Health |
| **IaC** | Infrastructure as Code |
| **NSG** | Network Security Groups |

Table : Glossary and definitions

# Executive Summary

This design covers the baseline standards for the Azure Data Factory Core Service. This service has been assessed against the five pillars of WAF as well as the Department of Health Security Controls.

This section contains a summary of the major design decisions that have been made for defining the baseline of this resource as an outcome of the WAF and Security analysis detailed throughout this document.

Of the five WAF Pillars, it was found that Security was relevant.

For this service the main baseline configurations for the infrastructure layer include:

* Version 2 (or any latest version at the time of deployment) will be used.
* Private endpoints will be used, and public connectivity disabled.
* Azure Integration runtime with managed virtual network is the chosen Integration Runtime option.
* Source code control will take place through Azure DevOps or GitHub (with Azure DevOps the architectural standard for Ambulance Victoria).

There are no differences in the infrastructure layer configuration of Azure Data Factory across the service catalog tiers.

# Resource Cost

The pricing construct for Data Factory is based on[[2]](#footnote-3):

* Pipeline orchestration and execution
* Data flow execution and debugging
* Number of Data Factory operations such as create pipelines and pipeline monitoring

**Pipeline orchestration and execution**

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Azure Integration Runtime Price | Azure Managed VNET Integration Runtime Price | Self-Hosted Integration Runtime Price |
| Orchestration1 | **$1.517** per 1,000 runs | **$1.517** per 1,000 runs | **$2.276** per 1,000 runs |
| Data movement Activity2 | **$0.380**/DIU-hour | **$0.380**/DIU-hour | **$0.152**/hour |
| Pipeline Activity3 | **$0.008**/hour | **$1.517**/hour  (Up to 50 concurrent pipeline activities) | **$0.003034**/hour |
| External Pipeline Activity4 | **$0.000380**/hour | **$1.517**/hour  (Up to 800 concurrent pipeline activities) | **$0.000152**/hour |

Table : Pricing construct for Azure Data Factory orchestration and execution

**Data flow and execution and debugging**

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Price | One Year Reserved (% Savings) | Three Year Reserved (% Savings) |
| General Purpose | $0.422 per vCore-hour | $0.317 per vCore-hour ~25% savings | $0.274 per vCore-hour ~35% savings |
| Memory Optimized | $0.531 per vCore-hour | $0.398 per vCore-hour ~25% savings | $0.345 per vCore-hour ~35% savings |

Table : Pricing construct for Azure Data Factory flow execution and debugging

|  |  |  |
| --- | --- | --- |
| Type | Price | Examples |
| Read/Write\* | **$0.759** per 50,000 modified/referenced entities | Read/write of entities in Azure Data Factory\* |
| Monitoring | **$0.380** per 50,000 run records retrieved | Monitoring of pipeline, activity, trigger, and debug runs\*\* |

Table : Pricing construct for Azure Data Factory operations

# WAF and Security Control Alignment

The following are the five pillars of the Microsoft Well Architected Framework:

* [Reliability](https://learn.microsoft.com/en-us/azure/well-architected/#reliability)
* [Cost optimization](https://learn.microsoft.com/en-us/azure/well-architected/#cost-optimization)
* [Operational excellence](https://learn.microsoft.com/en-us/azure/well-architected/#operational-excellence)
* [Performance efficiency](https://learn.microsoft.com/en-us/azure/well-architected/#performance-efficiency)
* [Security](https://learn.microsoft.com/en-us/azure/well-architected/#security)

For this design, the security section will also cover the Department of Health Controls in addition with any Microsoft Security Best Practices. Each of these sections will detail relevant controls or baseline requirements for this core service that will be put in place.

## Reliability

### Overview

The term reliability refers to the availability of the system and its ability to recover from failure[[3]](#footnote-4). Resiliency strategies must be built into each element of the architecture. The pillars of reliability include:

* Design for business requirements
* Design for failure
* Observe application health
* Drive Automation

### Azure Data Factory Reliability Checklist

There is no guidance under WAF for Azure Data Factory under the Reliability Pillar.

## Cost Optimisation

### Overview

The cost optimisation pillar is structured to support creating cost-effective workloads in the cloud[[4]](#footnote-5). It looks at removal of unnecessary spend and improving operational efficiency. The principles of cost optimisation revolve around:

* Choosing the correct resources
* Setting up budgets and maintaining cost constraints
* Dynamically allocate and deallocate resources
* Optimising workloads whilst aiming for scalable costs
* Continuously monitoring and cost managing

### Azure Data Factory Cost Optimisation Checklist

There is no guidance under WAF for Azure Data Factory under the Cost Optimisation Pillar.

## Operational Excellence

### Overview

Operational Excellence aims to ensure that once the architecture is built, the ongoing operations are flawless. This includes repeatable and reliable deployments, automating to eliminate human error. To do this the following must be considered:

* Optimise the build and release process (including CI/CD and IaC)
* Understand Operational Health
* Test recovery and failure
* Focus on continuous improvement
* Use loosely coupled architecture

### Azure Data Factory Operational Excellence Checklist

There is no guidance under WAF for Azure Data Factory under the Operational Excellence Pillar.

## Performance Efficiency

### Overview

Performance Efficiency refers to the ability of your systems and applications to meet user demands without breaking or creating a negative user experience[[5]](#footnote-6). This covers capacity and scalability:

* Design for horizontal scaling
* Run stress and performance tests
* Continuously monitor performances, particularly in Production systems

### Azure Data Factory Performance Efficiency Checklist

There is no guidance under WAF for Azure Data Factory under the Performance Efficiency Pillar.

## Security

### Overview

Security refers to the ability of the environment to resist and manage threats.

This section covers both Microsoft Best Practices as well as relevant security controls provided by the Department of Health. With respect to the Microsoft WAF, Security is underpinned by the following[[6]](#footnote-7):

* Plan resources and how to harden them
* Automate and use least privilege
* Classify and encrypt data
* Monitor system security, plan incident response
* Identify and protect endpoints
* Protect against code-level vulnerabilities
* Model and test against potential threats

In addition to the Microsoft controls, the Department of Health has mandated security posture to Ambulance Victoria. Note there may be duplication between the Microsoft Security Best Practices and the Department of Health controls.

The following Microsoft Security Benchmark controls are applicable:

* NS-1: Establish network segmentation boundaries
* NS-2: Secure cloud services with network controls
* IM-1: Use centralized identity and authentication system
* IM-3: Manage application identities securely and automatically
* DP-3: Encrypt sensitive data in transit
* DP-4: Enable data at rest encryption by default
* LT-4: Enable logging for security investigation
* BR-1: Ensure regular automated backups

# Architecture Summary

## Resource Overview

Azure Data Factory is built to orchestrate and operationalise large and complex data functions such as hybrid extract-transform-load (ETL), extract-load-transform (ELT), and other data integration[[7]](#footnote-8). It contains a series of subservices that provide an end-to-end data platform:

A diagram of data flow

Description automatically generated

Figure : Capabilities of Azure Data Factory7

The below image illustrates a generic Azure Data Factory scenario in which Data Factory acts as an orchestration service:

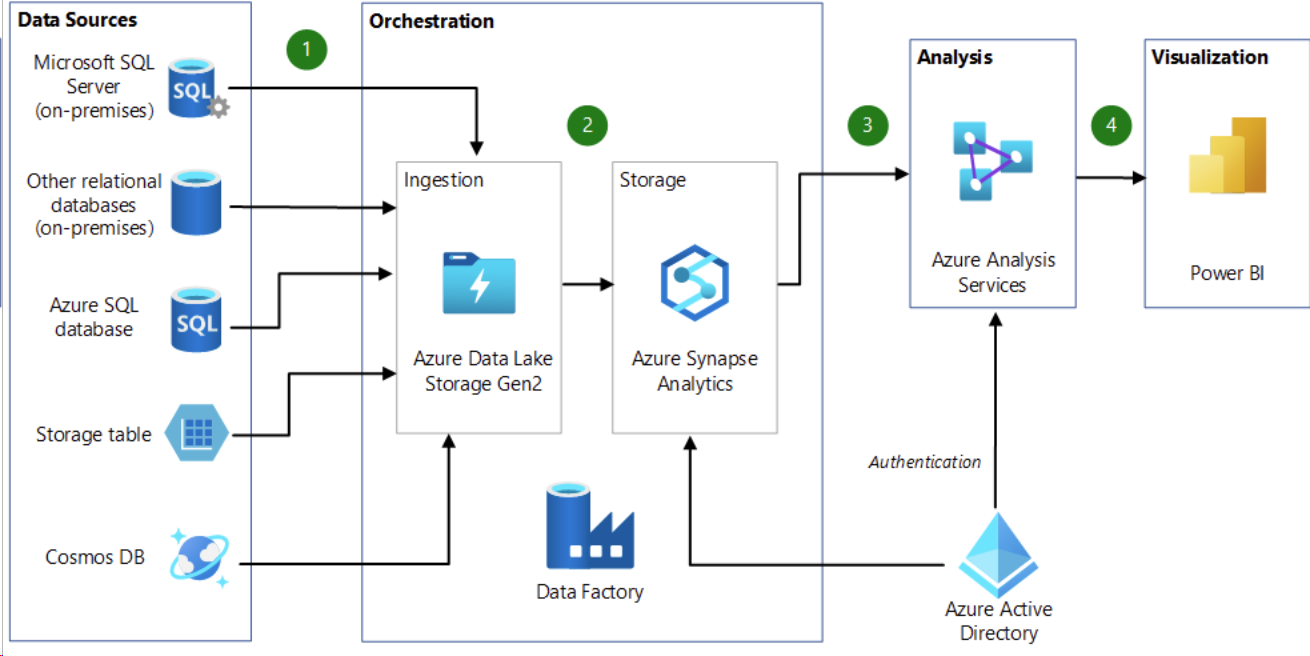


Figure : Data Factory orchestration example

### Integration Runtime

Data Factory has three types of Integration Runtime:

* Azure
* Self-Hosted
* Azure-SSIS (SQL Server Integration Services)

|  |  |  |
| --- | --- | --- |
| IR type | Public Network Support | Private Link Support |
| Azure | Data Flow Data movement Activity dispatch | Data Flow Data movement Activity dispatch |
| Self-hosted | Data movement Activity dispatch | Data movement Activity dispatch |
| Azure-SSIS | SSIS package execution | SSIS package execution |

Table : Integration Runtime supported scenarios comparison

To further assist in the selection of an appropriate integration runtime, the following features should be considered:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Azure integration runtime | Azure integration runtime with managed virtual network | Self-hosted integration runtime |
| Managed compute | Y | Y | N |
| Autoscale | Y | Y\* | N |
| Dataflow | Y | Y | N |
| On-premises data access | N | Y\*\* | Y |
| Private Link/Private Endpoint | N | Y\*\*\* | Y |
| Custom component/driver | N | N | Y |

Table : Integration Runtime selection criteria

*\* When time-to-live (TTL) is enabled, the compute size of integration runtime is reserved according to the configuration and can’t be autoscaled.*

*\*\* On-premises environments must be connected to Azure via Express Route or VPN. Custom components and drivers aren't supported.*

*\*\*\* The private endpoints are managed by the Azure Data Factory service.*

## RBAC

The following roles are applicable to this service:

|  |  |
| --- | --- |
| Role Name | Description |
| Data Factory Contributor | Create and manage data factories, as well as child resources within them. |

Table : RBAC roles relevant for this core service

## Design Decisions and Justifications

This section covers the design decisions and justifications that reflect the findings of the WAF and Security alignment. This will form the baseline requirements for the Azure Data Factory core service and will be captured in the accompanying Configuration Template with a set of pre-approved deployment settings for this resource. Any changes, modifications or removals to the pre-approved deployments must have specific approval from the Cloud Governance Forum prior to deployment.

### Version

**Design Reference:** N/A

**Design Decision:** Version 2 (currently the latest version) will be used.

**Design Justification:** The latest version of the service should be used at the time of deployment. Version 1 is no longer supported by Microsoft so any existing instances should be migrated to Version 2 as soon as possible. Version 2 has more capabilities including scale out and additional flow patterns for ETL requirements.

### Integration Runtime

**Design Reference: N/A**

**Design Decision:** Azure Integration Runtime with Managed Virtual Network is the preferred choice and should be used unless there is a specific incompatibility with application architecture.

**Design Justification:** Azure Integration runtime on its own does not support Private Endpoints which is the chosen method for connectivity, so this IR has been discounted. The Self-Hosted Integration Runtime has also been discarded as it does not have the benefits of managed compute, autoscaling, or data flow capabilities. Azure Integration Runtime with managed Virtual network has all the same capabilities of the other two IRs combined, and only lacks the custom component/driver support available in the Self-hosted integration runtime.

### Connectivity

**Design Reference:** Microsoft Security Benchmark [NS-1, NS-2](#_Overview)

**Design Decision**: Public connectivity will be disabled. Private endpoints will be used.

**Design Justification**: Private endpoints are the most secure form of connectivity so will be used preferentially to public access.

### Managed Identity

**Design Reference:** Microsoft Security Benchmark [IM-3](#_Overview)

**Design Decision:** Managed Identities will be used over Service Principals for authentication.

**Design Justification:** A managed identity will be used preferentially to a service principal for Azure authentications required by the Azure Data Factory.

### Authentication

**Design Reference:** Microsoft Security Benchmark [IM-1](#_Overview)

**Design Decision:** Azure AD will be used for authentication for Data Plane access.

**Design Justification:** Azure AD is the local authentication method that can be used for Data Factory access. It uses the central Azure AD authentication to MFA and perform other checks and is far less likely to be compromised as compared to situations where local authentication methods are possible.

### Encryption

**Design Reference:** Microsoft Security Benchmark [DP-3, DP-4](#_Overview)

**Design Decision:** Data is encrypted by default at rest with Azure Data Factory. For encryption in Transit the security is handled by the service which transfers data to the Data Factory and should be secured by HTTPS or TLS.  **Design Justification:** Data Factory is a tool that transforms data so is more a transitive service than a storage service. As such the encryption of data is determined more so the by data stores transferring the data. If they support HTTPS or TLS, which cloud native stores do, then the data is secured.

### Logging

**Design Reference:** Microsoft Security Benchmark [LT-4](#_Overview)

**Design Decision:** Logging will be enabled for the resource, and allLogs and AllMetrics will be sent to the central log analytics workspace in that region.

**Design Justification:** Sending resource logs to log analytics allows for additional troubleshooting and operational monitoring to take place. It will allow users to analyse in more details any issues or disruptions with the service in a central location.

### Backup

**Design Reference:** Microsoft Security Benchmark[BR-1](#_Overview)

**Design Decision:** Source code control will be used.  **Design Justification:** Data Factory itself cannot be backed up, however the code that is used to run pipelines and other mechanisms through Data Factory can be controlled using Azure DevOps or GitHub as the source code repositories. Using these allows you to have source-code control and versioning so that deployments are inherently backed up.

# Azure Policies

There are no additional Azure Policies required for this service.

# Configuration Templates

## Primary Region Azure Data Factory

|  |  |
| --- | --- |
| Configuration Item | Configuration Value |
| Name | adf-[env]-ause-[appname]-01 |
| Subscription | AV ALZ [Subscription Name] |
| Region | Australia Southeast |
| Version | V2 |
| ***Git Configuration Settings*** |  |
| Repository Type | Azure DevOps/GitHub |
| **Azure DevOps Settings** | Use the following if Repo is Azure DevOps |
| Azure DevOps Account | [Account where Project is created] |
| Project Name | [Project Name] |
| Repo Name | [Repo Name] |
| Branch Name | [Branch Name] |
| Root Folder | /[root folder] |
| **GitHub Settings** | Use the following if Repo is GitHub |
| GitHub Account | [GitHub account name] |
| Repo Name | [Repo Name] |
| Branch Name | [Branch Name] |
| Root Folder | /[root folder] |
| ***Networking Settings*** |  |
| Managed Virtual Network | Enabled |
| Connection Type | Private Endpoint |
| Private Endpoint Name | pep-adf-[env]-ause-[appname]-01 |
| Virtual Network | vnet-[env]-ause-[appname]-01 |
| Subnet | snet-[env]-ause-[appname]-[workload]-01 |
| Private DNS Zone | privatelink.datafactory.azure.net |

## Secondary Region Azure Data Factory

|  |  |
| --- | --- |
| Configuration Item | Configuration Value |
| Name | adf-[env]-auea-[appname]-01 |
| Subscription | AV ALZ [Subscription Name] |
| Region | Australia East |
| Version | V2 |
| ***Git Configuration Settings*** |  |
| Repository Type | Azure DevOps/GitHub |
| **Azure DevOps Settings** | Use the following if Repo is Azure DevOps |
| Azure DevOps Account | [Account where Project is created] |
| Project Name | [Project Name] |
| Repo Name | [Repo Name] |
| Branch Name | [Branch Name] |
| Root Folder | /[root folder] |
| **GitHub Settings** | Use the following if Repo is GitHub |
| GitHub Account | [GitHub account name] |
| Repo Name | [Repo Name] |
| Branch Name | [Branch Name] |
| Root Folder | /[root folder] |
| ***Networking Settings*** |  |
| Managed Virtual Network | Enabled |
| Connection Type | Private Endpoint |
| Private Endpoint Name | pep-adf-[env]-auea-[appname]-01 |
| Virtual Network | vnet-[env]-auea-[appname]-01 |
| Subnet | snet-[env]-auea-[appname]-[workload]-01 |
| Private DNS Zone | privatelink.datafactory.azure.net |

# Acceptance

Signature of this page by appropriately delegated representatives of ​Ambulance Victoria​ signifies acceptance of this design document.

Logicalis will commence build and implementation work once it receives a signed copy of this design document.

Signature of this page by appropriately delegated representatives of Ambulance Victoria signifies acceptance of this design document.

Logicalis will commence build and implementation work once it receives a signed copy of this design document.

|  |  |
| --- | --- |
| Project | Core Services |
| Document Version | 1.0 |

**Signed on behalf of Ambulance Victoria**

|  |  |
| --- | --- |
| Name | Dan Howarth |
| Position |  |
| Signature |  |
| Date signed |  |

**Signed on behalf of Logicalis Australia**

|  |  |
| --- | --- |
| Name | Daniela Nikolic |
| Position | Senior Cloud Engineer |
| Signature |  |
| Date signed |  |

1. https://learn.microsoft.com/en-us/azure/well-architected/ [↑](#footnote-ref-2)
2. https://azure.microsoft.com/en-us/pricing/details/data-factory/data-pipeline/ [↑](#footnote-ref-3)
3. https://learn.microsoft.com/en-us/azure/well-architected/resiliency/overview [↑](#footnote-ref-4)
4. https://learn.microsoft.com/en-us/azure/well-architected/cost/overview [↑](#footnote-ref-5)
5. https://learn.microsoft.com/en-us/azure/well-architected/scalability/overview [↑](#footnote-ref-6)
6. https://learn.microsoft.com/en-us/azure/well-architected/security/security-principles [↑](#footnote-ref-7)
7. https://learn.microsoft.com/en-us/azure/data-factory/introduction [↑](#footnote-ref-8)