Nonprofit Data Warehouse Quickstart – Technical Overview

A technical overview of the architecture and resources used in the Nonprofit Data Warehouse Quickstart

Version 1.1 | 21st January 2020

Reference Information

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Type | Author | Date |
| 1.0 | Final | Microsoft | 20/01/2020 |
|  |  |  |  |
|  |  |  |  |

**Contents**

[1 Introduction 4](#_Toc30519813)

[2 Solution Architecture 5](#_Toc30519814)

[2.1 Solution Overview 5](#_Toc30519815)

[2.1.1 Logical Architecture - Quickstart 6](#_Toc30519816)

[2.1.2 Logical Architecture - “Quickstart with CDM” 7](#_Toc30519817)

[2.1.3 Security 8](#_Toc30519818)

[2.1.4 Deployment 8](#_Toc30519819)

[2.2 Orchestration 9](#_Toc30519820)

[2.2.1 Data processing and orchestration 9](#_Toc30519821)

[2.2.2 Management 10](#_Toc30519822)

[3 Azure Resources 11](#_Toc30519823)

[3.1 Storage Account 11](#_Toc30519824)

[3.1.1 Service overview 11](#_Toc30519825)

[3.1.2 Benefits 11](#_Toc30519826)

[3.1.3 How is it used in the solution? 11](#_Toc30519827)

[3.1.4 Service Details 11](#_Toc30519828)

[3.2 Azure Data Lake Storage Gen 2 12](#_Toc30519829)

[3.2.1 Service overview 12](#_Toc30519830)

[3.2.2 Benefits 12](#_Toc30519831)

[3.2.3 How is it used in the solution? 12](#_Toc30519832)

[3.2.4 Service Details 12](#_Toc30519833)

[3.3 SQL Server and Synapse Analytics 13](#_Toc30519834)

[3.3.1 Service overview 13](#_Toc30519835)

[3.3.2 Benefits 13](#_Toc30519836)

[3.3.3 How is it used in the solution? 13](#_Toc30519837)

[3.3.4 Service Details 13](#_Toc30519838)

[3.4 Azure Data Factory Gen 2 14](#_Toc30519839)

[3.4.1 Service overview 14](#_Toc30519840)

[3.4.2 Benefits 14](#_Toc30519841)

[3.4.3 How is it used in the solution? 14](#_Toc30519842)

[3.4.4 Service Details 14](#_Toc30519843)

[3.5 Key Vault 15](#_Toc30519844)

[3.5.1 Service overview 15](#_Toc30519845)

[3.5.2 Benefits 15](#_Toc30519846)

[3.5.3 How is it used in the solution? 15](#_Toc30519847)

[3.5.4 Service Details 15](#_Toc30519848)

[3.6 Power BI 16](#_Toc30519849)

[3.6.1 Service overview 16](#_Toc30519850)

[3.6.2 Benefits 16](#_Toc30519851)

[3.6.3 How is it used in the solution? 16](#_Toc30519852)

[3.6.4 Service Details 16](#_Toc30519853)

[4 Alternative Architectures 17](#_Toc30519854)

[4.1 SQL Technologies 17](#_Toc30519855)

[4.1.1 Azure SQL Database 17](#_Toc30519856)

[4.1.2 Azure SQL Managed Instance 18](#_Toc30519857)

[4.1.3 Azure Databricks 19](#_Toc30519858)

[4.1.4 PowerBI Dataflows 20](#_Toc30519859)

[5 Further Reading 0](#_Toc30519860)

[5.1 Web Links 0](#_Toc30519861)

[5.2 Books 0](#_Toc30519862)

# Introduction

The Nonprofit Data Warehouse Quickstart meets nonprofits where they’re at in their data journey, enabling aggregation and transformation of disparate data sources to connect dots and build trends that aid in decision making processes.

This document provides a technical overview of the Nonprofit Data Warehouse Quickstart, a prepackaged “one-click deployment” Azure based warehousing solution.

For a step-by-step guide on how to deploy the solution from GitHub, configure the components, and execute the solution please refer to the Deployment Guide documents.

This document is designed to give a detailed view of the solution, explaining the purpose, role, and benefits of each resource and to cover alternative architectures, highlighting the advantages and disadvantages of using different approaches.

Through this document, users will gain a strong awareness of the different technologies within the Azure platform and how they can be used within a data analytics architecture.

The remainder of this document contains the following sections:

* Solution Architecture
* Azure Resources
* Alternative Architectures
* Further Reading

# Solution Architecture

This section starts by detailing the different resources of the solution and the role each one plays in the architecture. It is then followed with a description of the implemented security measures that keep internal and external data safe and the deployment techniques applied to support the “one click” experience. The section is completed with an overview of the orchestration process, demonstrating how Azure Data Factory pipelines are used in the ingestion and transformation process.

## Solution Overview

The solution was designed to run on Microsoft Azure, using only Platform-as-a-Service (PaaS) resources. It allows an organisation to purchase and use resources on a pay-as-you-go basis and access them using a secure connection. With PaaS, the expensive and complex process of buying and managing software licences, and the underlying application infrastructure and middleware, can be avoided. This approach also means an organisation only has to manage the cloud resources, leaving Microsoft to manage the infrastructure side (e.g. updates to the operating system and patch the hardware).

The Azure services used in this solution are:

* **Orchestration:** Azure Data Factory (ADF) Gen 2
* **Storage:** Azure Blob Storage, Azure Data Lake Storage (ADLS) Gen 2 and Synapse Analytics
* **Presentation:** Power BI
* **Security:** Key Vault

This document describes two different architectures, **Quickstart** and **Quickstart with CDM**. Both use the same services, however, in the first architecture, the data stored in ADLS is structured following a standard Big Data approach and in the second it is stored in the Common Data Model (CDM) format, utilising the Nonprofit Accelerator for CDM, for more information on the Common Data Model, please see [here](https://docs.microsoft.com/en-us/common-data-model/), and for more information on the Nonprofit Accelerator for CDM, please see [here](https://docs.microsoft.com/en-us/common-data-model/nfp-accelerator).

With the Nonprofit Accelerator and Common Data Model for Nonprofits, users can develop nonprofit solutions based on entities and attributes that nonprofits commonly leverage for constituent management, fundraising, awards, program delivery and impact tracking. These entities include but are not limited to donor commitments, designations, transactions, awards, disbursements, delivery frameworks, results, indicators, benefit recipients and more.

For more information on the deployment steps of either architecture please see the Nonprofit Data Warehouse Quickstart Deployment Guide, or the Quick Start with CDM Deployment Guide.

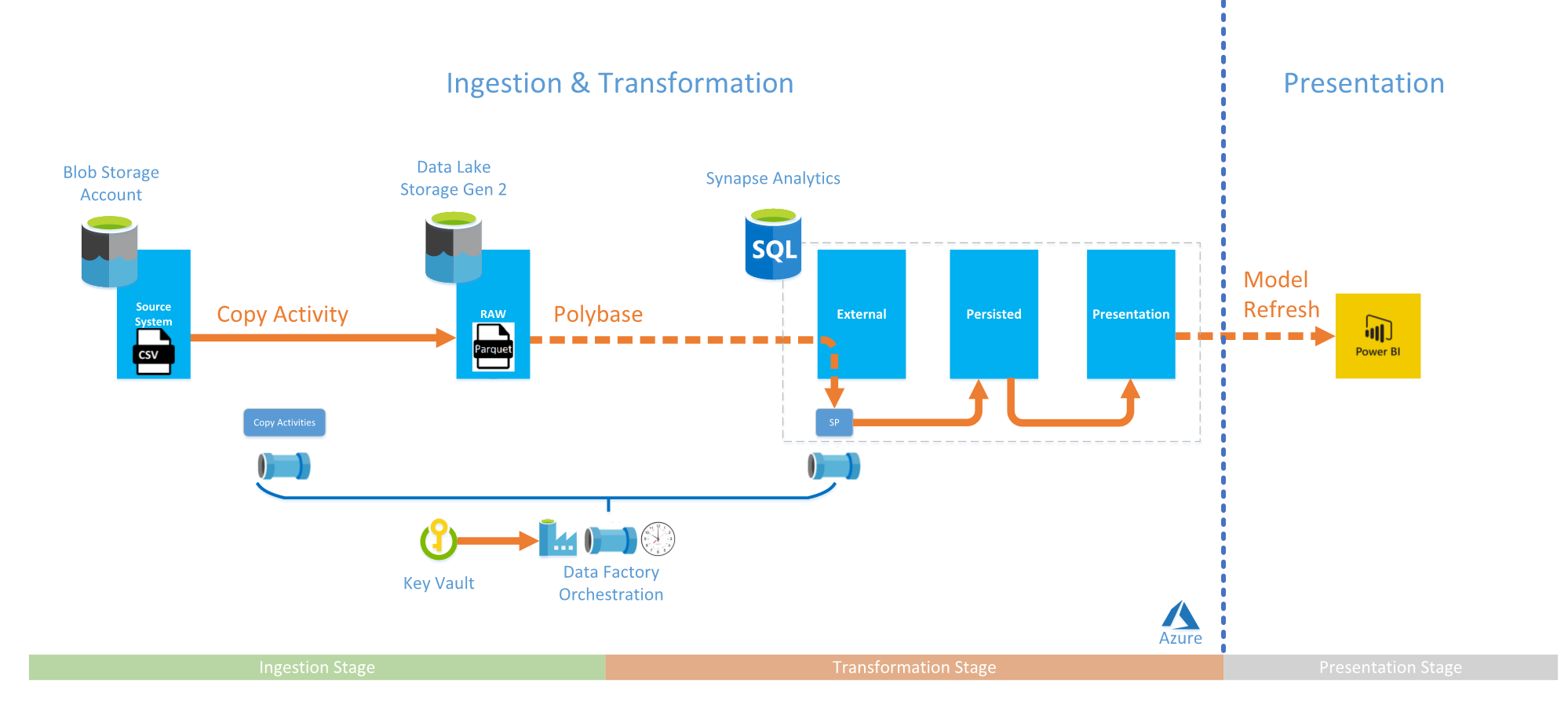
There are three staging layers in these solutions: Ingestion Transformation, and Presentation. The data is moved between Ingestion and Transformation through Data Factory pipelines and stored procedures and is presented to the users using Power BI Reports.

Sensitive information, such as login credentials, are kept in the form of secrets inside Key Vault, adding an extra layer of security to the solution.

Both solutions come with a sample source files which allow users to execute and visualise the data movement through the architecture. The sample files are based on a subset of the publicly available IATI dataset (see [here](https://www.iatiregistry.org/dataset)), and Water & Sanitation data provided by the World Health Organisation (see [here](http://apps.who.int/gho/data/node.main.WSHSANITATION?lang=en)).

The following sections present the logical architecture diagram. Later in the document, each service is described in more detail, highlighting benefits and disadvantages and describing the role they play in the architecture.

### Logical Architecture - Quickstart



#### Ingestion Layer

Ingestion is the stage where the data is loaded from the source system to Azure Data Lake Store (ADLS). The data can be extracted from a variety of source systems, such as databases, file shares or web APIs. In this scenario, the sample csv data is stored in a Blob Storage account and is landed in ADLS as Parquet Snappy files (more details about the format can be found [here](https://en.wikipedia.org/wiki/Apache_Parquet)). Azure Data Factory (ADF) pipelines extract the metadata stored in the control tables in Synapse Analytics and copies the data between the resources using a copy activity.

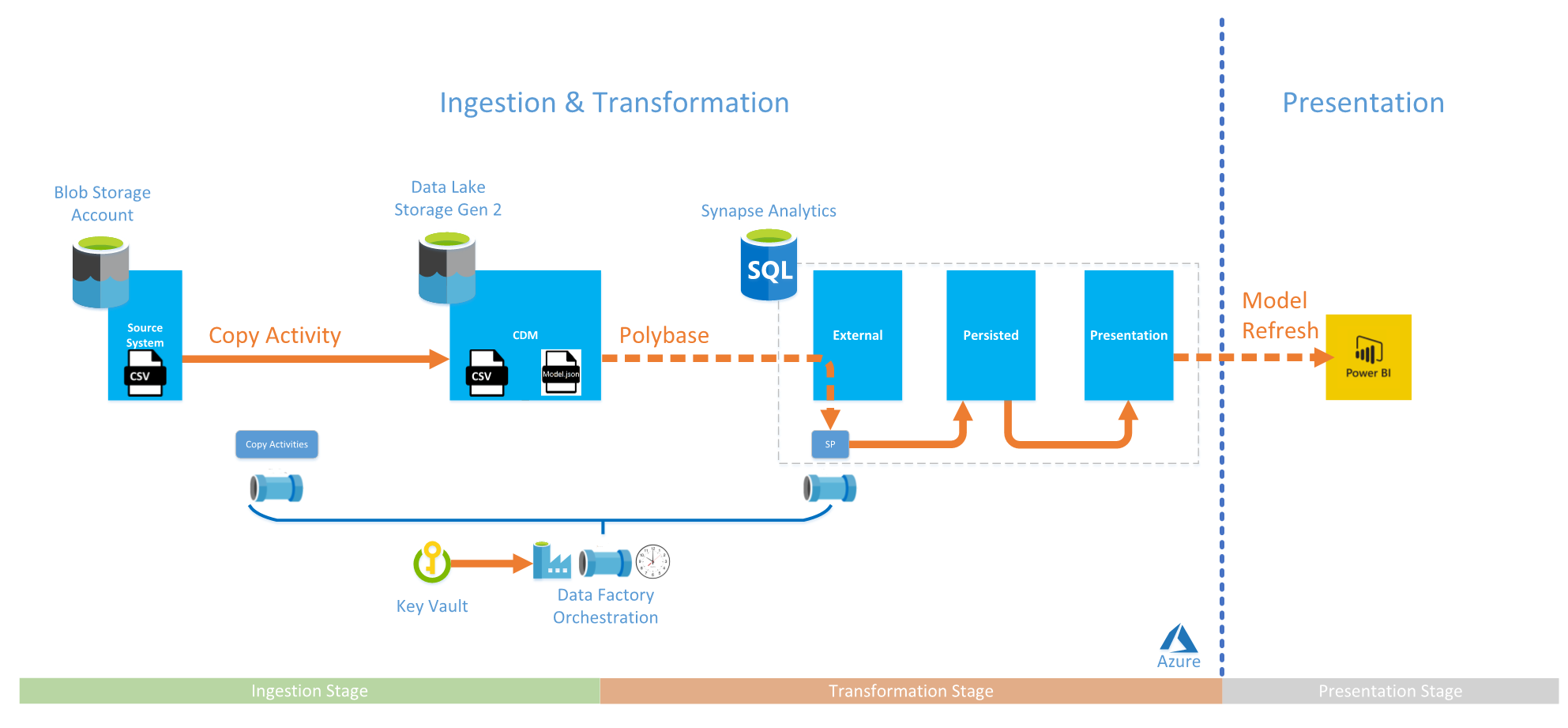
#### Transformation Layer

Transformation is the stage where the data is read and transformed according to the business logic. In this scenario, the data stored in ADLS is copied to Synapse Analytics using External tables and PolyBase (more details [here](https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-data-into-azure-sql-data-warehouse)) and transformed using stored procedures. These procedures remove duplicates and insert new records to the non-transient persisted entity tables.

#### Presentation Layer

Presentation is the stage where the data is presented to the users, most commonly using a reporting tool like Power BI. Traditionally, the facts and dimensions that form a data model would be built in the data mart/warehouse, however, for demonstration purposes, this solution used Power BI and Power Query as an alternative. Using Power BI Desktop and the import storage mode (more details [here](https://docs.microsoft.com/en-us/power-bi/desktop-storage-mode)), the data is imported from Synapse Analytics and distributed across fact and dimension tables. The data model is finally enhanced with calculated columns and measures defined in the DAX language.

### Logical Architecture - “Quickstart with CDM”



#### Ingestion Layer and Common Data Model

Ingestion is the stage where the data is loaded from the source system to ADLS. The data can be extracted from a variety of source systems, such as databases, file shares, or web APIs. In this scenario, the sample csv data is stored in a Blob Storage account and is landed in ADLS in the Common Data Model (CDM) format (more details [here](https://docs.microsoft.com/en-us/common-data-model/data-lake)). Additionally, data can also be extracted from third party applications that support CDM (e.g. Microsoft Dynamics 365) via Power BI Dataflows as detailed later in the document. An ADF pipeline extracts the metadata stored in the Control tables in Synapse Analytics and copies the data between the resources using a copy activity.

#### Transformation Layer

Transformation is the stage where the data is read and transformed according to the business logic. In this scenario, the data stored in ADLS is copied to Synapse Analytics using external tables and PolyBase (more details [here](https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-data-into-azure-sql-data-warehouse)) and transformed using stored procedures. These procedures remove duplicates and insert new records to the non-transient persisted entity tables.

#### Presentation Layer

Presentation is the stage where the data is presented to the users, most commonly using a reporting tool like Power BI. Traditionally, the facts and dimensions that form a data model are built in the data mart/warehouse, however, for demonstration purposes, this solution used Power BI and Power Query as an alternative. Using Power BI Desktop and the import storage mode (more details [here](https://docs.microsoft.com/en-us/power-bi/desktop-storage-mode)), the data is imported from Synapse Analytics and distributed across fact and dimension tables. The data model is finally enhanced with calculated columns and measures defined in the DAX language.

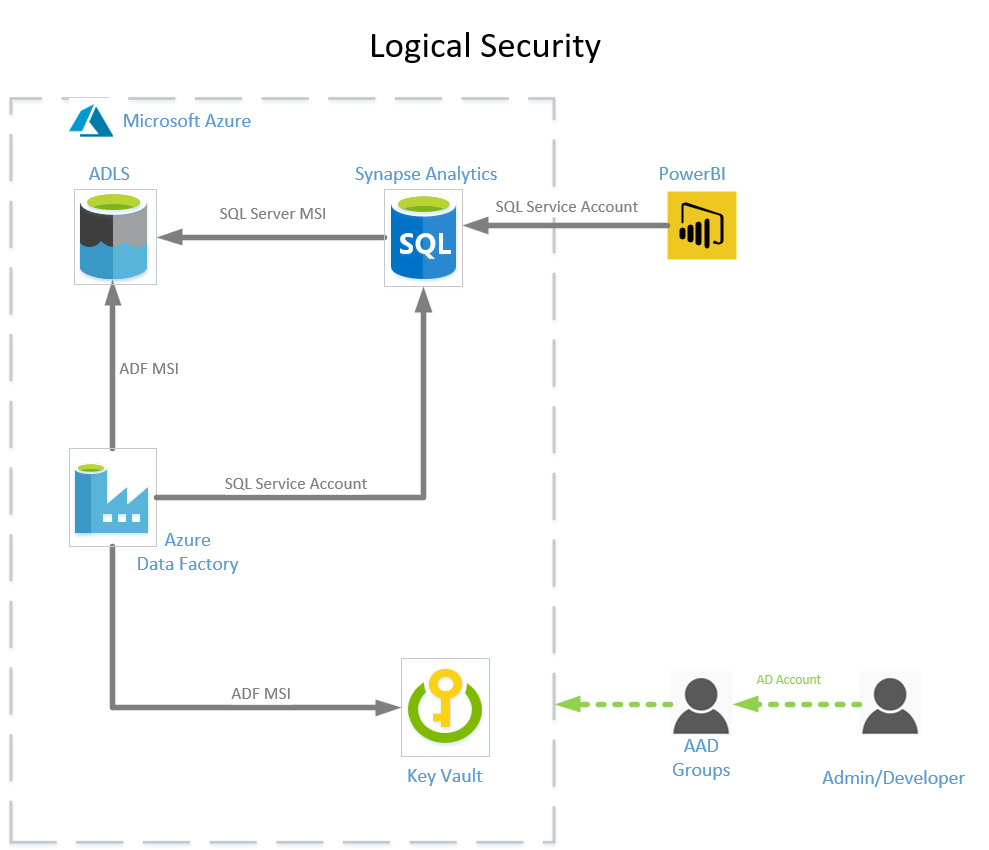
### Security

The solution is secured via Azure Active Directory (AAD) Groups, SQL Service Accounts and Managed Service Identities (MSIs) (more details [here](https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview)). MSI enables the authentication betwen compute and storage resources and reduces the need for the management of secrets, as this is handled by Azure.

SQL Users have been used only in places where MSI could not be used, such as the connection between ADF and Synapse Analytics. Users’ access to Azure resources have been configured via Azure Active Directory Groups.

There are two system-assigned MSIs used in the solution.

1. **SQL Server and Synapse Analytics** – Used to authenticate SQL Server against ADLS to PolyBase External Tables.
2. **Azure Data Factory** – Used to authenticate against all resources ADF accesses in this solution.



### Deployment

The solution is deployed using Azure Resource Manager (ARM) Templates and PowerShell Scripts. The deployment process and the required steps are described in detail in the “Deployment Guide” document.

## Orchestration

Azure Data Factory is the service responsible for the orchestration of the data processing activities. These activities can be split in two main areas:

* **Data extraction from the source systems** – copy the data from source systems to Azure Data Lake Storage (ADLS)
* **Execution of the Stored Procedures** – Read the data from ADLS, transform and stage it to Synapse Analytics Persisted tables

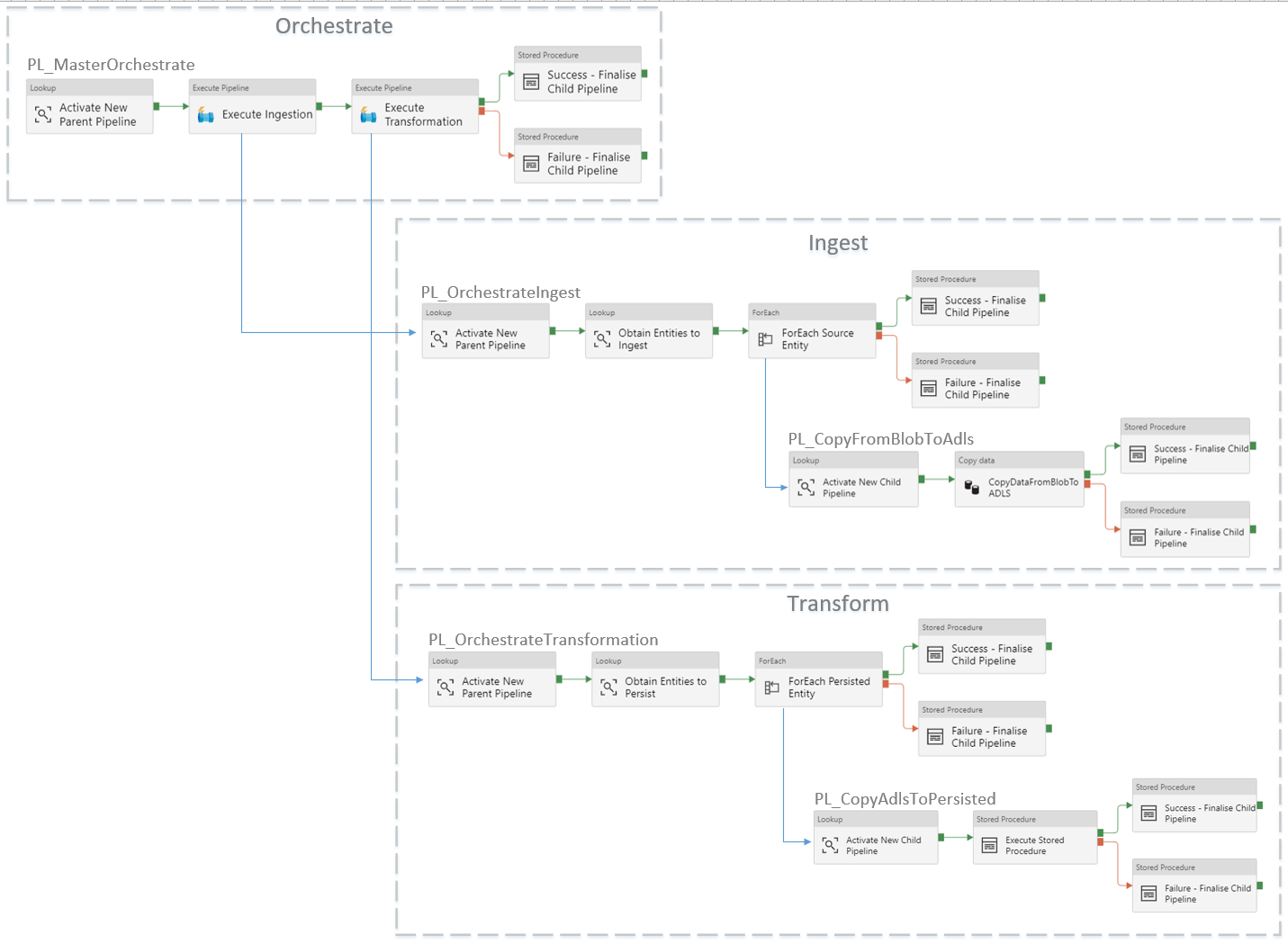
The process is structured following a parent child approach. A time triggered parent pipeline called PL\_MasterOrchestrate executes the child pipelines, PL\_OrchestrateIngest and PL\_OrchestrateTransformation, to control the ingestion and transformation of the data. Each pipeline connects to Synapse Analytics to obtain the required information about the entities that should be executed (e.g. Obtain Entities to Ingest). Additionally, all pipelines have been enhanced with ADF activities to perform custom logging to Synapse Analytics. These activities are labelled as “Activate New Parent Pipeline” and “{Result} – Finalise Child Pipeline”.

There are two types of pipelines, data processing and orchestration, and management.

### Data processing and orchestration

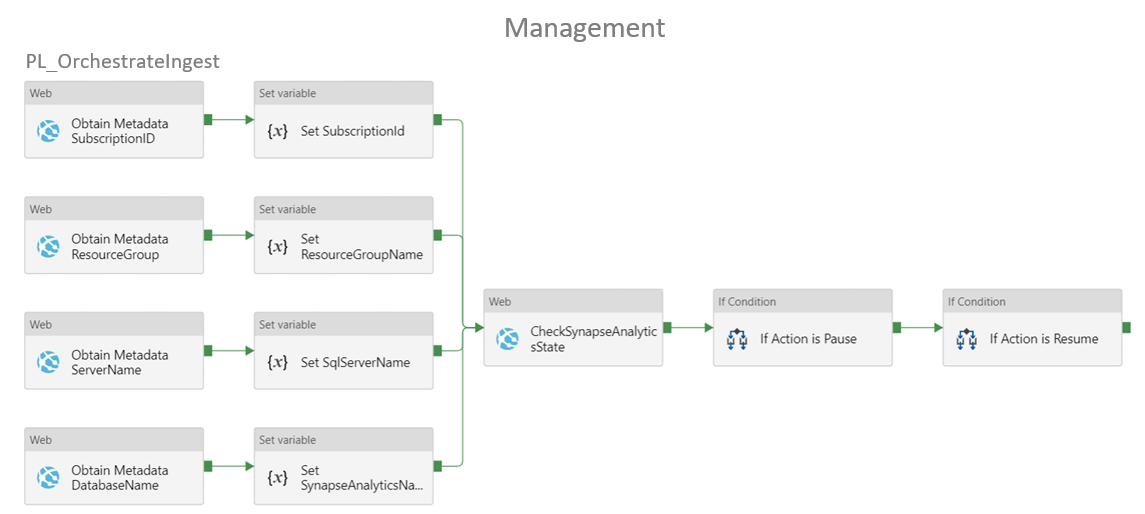
There are five pipelines involved in this process:

* **“PL\_MasterOrchestrate”** – The parent pipeline that orchestrates the entire sequence of data processing.
* **“PL\_OrchestrateIngest” –** Controls the ingestion stage.
* **“PL\_OrchestrateTransformation”** - Controls the transformation stage.
* **“PL\_CopyFromBlobToAdls”** – Moves the data from Blob Storage to ADLS.
* **“PL\_CopyFromAdlsToPersisted”** – Moves the data from ADLS to Synapse Analytics persisted tables.



### Management

The pipeline “PL\_ChangeStageOfSynapseAnalytics” manages the state of Synapse Analytics by either pausing or restarting the service. The pipeline takes a single parameter - “pause” to disable Synapse Analytics or “resume” to enable Synapse Analytics.



# Azure Resources

This section covers each key resource used in the solution. It provides a brief overview alongside the benefits of the service and describes how it is used in the solution. This should give the user the ability to understand how the solution can be improved or adapted to the organisation’s needs.

## Storage Account

### Service overview

Used as a scalable cloud storage for unstructured data, Blob storage provides an effective alternative to on-premises storage. It can be used for a wide range of applications, from storing files for distributed access through to ensuring data can be recovered in a disaster.

### Benefits

* The scalable nature means that one only has to pay for the amount of data required at a certain time, improving the cost effectiveness of Blob storage solutions.
* Different storage tiers within Blob storage, allowing the solution to be optimised for the data being stored.

### How is it used in the solution?

Used as a pseudo-source system due to its flexible nature and ability to be scaled up if further data is required to be added.

### Service Details

Further information about the Storage Accounts can be obtained in the service documentation by following this [link](https://azure.microsoft.com/en-gb/services/storage/blobs/).

## Azure Data Lake Storage Gen 2

### Service overview

Azure Data Lake Storage Gen 2 (ADLS) is highly scalable, and accessible via many different platforms, such as Hadoop and Azure Databricks. Data lakes can store information to the order of petabytes and have the ability to handle hundreds of gigabits of data throughput. ADLS also allows analytics to take place without the need for moving or transforming before the analysis.

### Benefits

* Performance – ADLS stores files in distributed form so queries can be paralleled.
* Cost effectiveness - ADLS can scale up or scale down to meet data needs, enabling customers to only pay for what they use.
* Interoperability – ADLS is a Hadoop based system that provides the variety of platforms that ADLS can be accessed by.
* Common Data Model – ADLS is the foundation for data storage and sharing when using the Common Data Model, allowing application such as Power Apps, Power BI and Dynamics 365 to utilise a single data set and providing consistency across all applications and processes.
* Manageability – ADLS can easily be managed by a variety of means, such as API REST calls, PowerShell, or SDKs from a variety of platforms.
* Security – ADLS stores data in an encrypted format. Authentication and authorisation to the services is provided by Azure Active Directory.

### How is it used in the solution?

ADLS is used to store the data after being ingested by Data Factory and to provide the ability for third-party solutions to directly consume/produce data.

### Service Details

Further information about ADLS can be obtained in the service documentation by following this [link](https://azure.microsoft.com/en-gb/services/storage/data-lake-storage/).

## Synapse Analytics

### Service overview

Synapse Analytics is a data warehouse that allows distributed solutions to be created and accessed using T-SQL queries and stored procedures at scale. This can then be integrated with big data analytics solutions provided by the wider Azure Synapse environment, allowing users from data scientists through to business analysts to be able to leverage an organisation’s data.

### Benefits

* Can easily be turned on and off based on requirements, providing a greater cost efficiency for the overall solution.
* The distributed nature of the data warehouse ensures safety for the stored data, the data is divided and replicated across 60 distributions.

### How is it used in the solution?

Used to load the data into persistent tables, which can then be accessed to run analysis, enabling organizations to discover insights about the data.

### Service Details

Further information about Synapse Analytics can be obtained in the service documentation by following this [link](https://azure.microsoft.com/en-gb/services/synapse-analytics/).

## Azure Data Factory Gen 2

### Service overview

Azure Data Factory Gen 2 is a tool for creating ETL/ELT pipelines, all within a code-free environment. Allows the data to be ingested, moved, prepared, transformed and processed comprehensively and quickly, with the code required being generated and maintained automatically. Can be used with many different Azure and on-premises services, and integrates with other cloud service providers’ solutions, such as Amazon Redshift and Google BigQuery.

### Benefits

* Integrated with Azure security measures, keeping the data safe throughout the pipeline.
* Code-free environment provides an intuitive user experience, allowing the focus to be on the overall solution rather than coding.

### How is it used in the solution?

Used to control the data ingestion and transformation processes, moving the data from the source systems to the Synapse Analytics persisted tables.

### Service Details

Further information about ADF can be obtained in the service documentation by following this [link](https://azure.microsoft.com/en-gb/services/data-factory/).

## Key Vault

### Service overview

A storage solution for the secrets associated with an overall data solution. The secrets are protected using Azure’s [industry-standard security infrastructure](https://docs.microsoft.com/en-us/azure/security/), and then requires proper authentication and authorisation before a secret can be accessed. The authentication and authorisation are implemented using Azure Active Directories and access policies, providing a straightforward method for controlling these processes. Furthermore, the secrets are accessed via URIs, allowing the secret to be used indirectly.

### Benefits

* Provides a central location for application secrets, allowing access to be easily controlled and the chances of secrets being leaked to be greatly reduced.
* Automatically scales to meet the demand of access requests to the secrets stored within the Key Vault.

### How is it used in the solution?

Used to store the sensitive details associated to different resources, keeping them secure.

### Service Details

Further information about Key Vault can be obtained in the service documentation by following this [link](https://azure.microsoft.com/en-gb/services/key-vault/).

## Power BI

### Service overview

A collection of services, apps, and connectors that take data from a variety of sources into interactive dashboards and reports. Within Power BI, data can be modelled into the required schema for a particular use case, creating visuals that display insights from data and enable sharing work with colleagues within an organisation.

### Benefits

* Use of the [DAX language](https://docs.microsoft.com/en-us/dax/dax-syntax-reference), allowing complex queries to transform the data and to gain statistical insights to be created quickly.
* Compatible with many different data services, from CSV files through to data warehouses, allowing flexibility within solutions.

### How is it used in the solution?

Used to ingest the data into analytical data models, such as Star Schema, which allow users to create performant joined up, interactive dashboards for users to gain insights from. For more information on the benefits and creating Star Schemas within Power BI, please see [here](https://docs.microsoft.com/en-us/power-bi/guidance/star-schema).

### Service Details

Further information about Power BI can be obtained in the service documentation by following this [link](https://powerbi.microsoft.com/en-us/what-is-power-bi/).

# Alternative Architectures

So far, this document has focused on the solution’s architecture and constituent technologies, however, it is important to be aware of alternatives that can either provide further capabilities to this warehousing solution or help tailor it according to the business requirements.

This section outlines three technologies that could be used within an alternative architecture. It provides an overview of the technology, the advantages and disadvantages of using it and the benefits and hindrances of each approach when compared with the current solution.

## SQL Technologies

SQL technologies can be replaced with relatively low impact. There are two alternatives to Synapse Analytics available in Azure, SQL Database and SQL Managed Instance.

### Azure SQL Database

#### Overview

Azure SQL Database is a non-distributed PaaS resource that provides a scalable cloud database and has the broadest compatibility with the SQL server engine. With hyperscale support, there could be up to 100 TB stored in the database that could be restored in minutes. It allows the scaling compute and storage resources separately and supports technologies like Always Encrypted and Transparent Data Encryption (TDE). Azure SQL Database has a very high availability level of 99.995%.

#### Advantages

Azure SQL Database is highly cost effective. If data in the database is frequently used, then it will be a cheaper option when compared with Synapse Analytics. It is much more efficient when concurrent access is required and uses similar SQL syntax to the on-premise version. Both Synapse Analytics and Azure SQL Database provide similar levels of security.

#### Disadvantages

Azure SQL Database is not distributed, meaning that the data is stored in a single database. This means it is less powerful than Synapse Analytics when dealing with high volumes of data (>60GB).

Azure SQL Database does not support PolyBase External tables on Azure Data Lake Storage (ADLS), so in this scenario data would need to be ingested into the database via a Data Factory copy activity, which is typically much slower than using PolyBase.

The database service cannot be paused, which is a key feature of Synapse Analytics. If higher service tiers are required, the cost may reach the same levels of Synapse Analytics.

### Azure SQL Managed Instance

#### Overview

The managed instance is a new deployment option for Azure SQL Database, providing near 100% compatibility with the latest SQL Server on-premises (Enterprise Edition) Database Engine. It provides a native [virtual network (VNet)](https://docs.microsoft.com/en-us/azure/virtual-network/virtual-networks-overview) implementation that addresses common security concerns and a [business model](https://azure.microsoft.com/pricing/details/sql-database/) favourable for on-premises SQL Server customers. The managed instance deployment model allows existing SQL Server customers to lift and shift their on-premises applications to the cloud with minimal application and database changes. At the same time, the managed instance deployment option preserves all PaaS capabilities (automatic patching and version updates, [automated backups](https://docs.microsoft.com/en-us/azure/sql-database/sql-database-automated-backups), [high-availability](https://docs.microsoft.com/en-us/azure/sql-database/sql-database-high-availability) ), which reduces management overhead and total cost of ownership (TCO).

#### Advantages

Managed Instance’s native VNET support provides better physical isolation of the service, which means, from a security point of view, it’s better than Synapse Analytics. Furthermore, if an organisation has “onpremises” SQL Server licences with Software Assurance, these can be used to reduce running costs. A managed instance is the best migration destination in the cloud for SQL Server instances that require high security and a rich programmability surface. The syntax is simplified compared to Synapse Analytics due to the compatibility with the “on-premises” SQL syntax.

#### Disadvantages

Despite the possible licence discount mentioned previously, a managed instance is still much more expensive than a normal SQL Database. Similarly, to Azure SQL DB, this service does not support PolyBase External tables neither can it be paused, so an organisation would need to pay for SQL Server even if it is not used at the time. Managed Instances are deployed to a VNET, which uses additional resources and adds further complexity to the solution.

### Azure Databricks

#### Overview

Azure Databricks is an Apache Spark-based analytics service, which can be used to prepare the data for analytics within a notebook-based environment. This makes use of both Spark SQL and DataFrames, allowing for those familiar with SQL and Python/R to quickly get started with it. Furthermore, the MLib library is built into Azure Databricks, providing machine learning functionality to solutions in a straightforward implementation.

#### How it can be used

The solution could use the Azure Databricks service as a transformation engine. This approach would remove the need for a high tier SQL Server layer for the data processing and allow data to be entirely modelled in Azure Databricks. The data could be directly consumed via SQL Server, via supported applications like Power BI or outputted as files and stored in ADLS. Azure Databricks has the potential to remove the SQL Database from the solution or reduce its significance by taking over all the heavy lifting associated to the data transformation process.

#### Advantages

Azure Databricks uses Spark technology and integration with Hadoop-type file systems to achieve high levels of processing performance. This is achieved by partitioning data and distributing queries among multiple files.

Azure Databricks can use ADLS as input and output storage. This means it can have input source files and output models of unlimited size. Additionally, it supports multiple file formats that can be chosen depending on the organisation’s needs. For example, files can be stored as flat delimited text files or more complex structural files like Parquet Snappy or JSON.

Azure Databricks is highly cost effective when managed properly. This is achieved due to the detachable compute clusters, that can be switched on only for the duration of the transformation. By using ADLS as the storage layer, consumers can still access the data even if the compute clusters are paused.

Azure Databricks is very secure. The resources are deployed to a VNET, which provides a good physical separation. Additionally, it is partially integrated with Azure Active Directory and supports integration with Azure Key Vault.

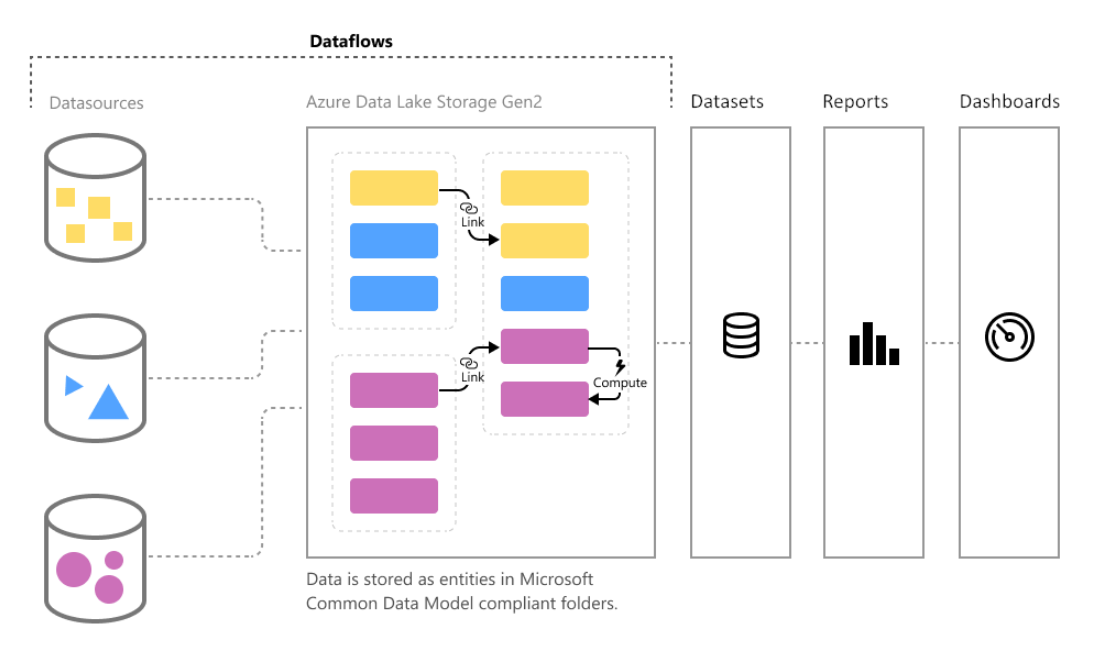
#### Disadvantages

Azure Databricks requires knowledge on one of the following languages: PySpark, Scala or Spark SQL. Because it only partially integrates with AAD, concepts like Managed Service Identities, are not available to the user.

### PowerBI Dataflows

#### Overview

Power BI has introduced Dataflows to help organizations unify data from disparate sources and prepare it for modeling. Analysts can easily create dataflows using familiar, self-service tools. Dataflows are used to ingest, transform, integrate, and enrich Big Data by defining data source connections, ETL logic, refresh schedules, and more. In addition, the new model-driven calculation engine within Dataflows makes the process of data preparation more manageable, more deterministic, and less cumbersome for data analysts and report creators alike. Similar to how spreadsheets handle recalculations for all affected formulas, dataflows manages the changes for an entity or data element on the user’s behalf, automating updates, and alleviating what used to be tedious and time consuming logic checks for even a basic data refresh. With Dataflows, tasks that once required data scientists to oversee (and many hours or days to complete) can now be handled with a few clicks by analysts and report creators.



Data is stored as entities in the [Common Data Model](https://docs.microsoft.com/powerapps/common-data-model/overview) (CDM) in Azure Data Lake Storage Gen2. Dataflows are then created and managed in workspaces by using the Power BI service. Dataflows are designed to use the CDM, a standardised, modular, extensible collection of data schemas published by Microsoft that are designed to make it easier to build, use, and analyse data. With this model, the user can go from data sources to Power BI dashboards with nearly zero friction.

#### How it can be used

Within the Quickstart with CDM solution, data is ingested and stored in ADLS in the CDM format. The model.json file contains metadata that allows Power BI Dataflows to connected to ADLS and fully understand the structure of each entity. This means SQL Server could be removed as the primary source of data to the Power BI reports.

#### Advantages

Using Dataflows as a transformation layer could reduce the costs of the platform since SQL Server could be removed from the architecture.

Dataflows enables Power BI to work as a CDM consumer and producer. As a consumer, it can attach multiple CDM models stored in ADLS as an external source in Power BI. As a producer, it can extract data from platforms like Microsoft Dynamics 365 and store it in ADLS in the CDM format.

#### Disadvantages

Dataflows uses Power Query for the transformation process, however, this has not been designed to transform large volumes of data. If an organisation uses large datasets, then Power BI could have slow performance. Additionally, Dataflows requires paid plans like Power BI Pro or Power BI Premium. For large datasets, Power BI Premium would be required, which would increase the running costs.

# Further Reading

While appropriate links for the different resources have been provided, there is a wealth of further information available to the user, often a simple web search away. This section references resources that can improve the user’s knowledge on the technologies and techniques applied in this solution, providing a further understanding of Azure and Power BI as data analytics platforms.

## Web Links

* [Certified Azure Data Engineer](https://docs.microsoft.com/en-us/learn/certifications/azure-data-engineer) - The learning paths for the Azure Data Engineer Associate credential, which contain many exercises and practical tips for implementing a data engineering solution. Can allow a user to particularly extend their knowledge of the Azure platform, and even earn a recognised certification.
* [Azure Synapse Analytics](https://docs.microsoft.com/en-us/azure/sql-data-warehouse/) - This is the documentation for Azure Synapse Analytics, formerly Azure SQL Data Warehouse. The “Concepts” section in the documentation is particularly useful in understanding key concepts when implementing a data warehouse solution.
* [Power BI Guided Learning](https://docs.microsoft.com/en-us/power-bi/guided-learning/)- Collection of resources for learning more about the complete functionality of Power BI, particularly useful in creating interactive dashboards for making use of the data.

## Books

* Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd edition),* Wiley (2013) – Useful guide to creating the data models required within a data analytics solution, providing many strategies and best practices.
* Matt Allington, *Supercharge Power BI,* Holy Macro! Books (2018) – Excellent starting point for learning more about Power BI for business intelligence solutions.