Enterprise Data Warehouse Implementation

Solution Design Document

Prepared By

GLOSSARY

Abbreviation	Description
BI	Business Intelligence
DWU	Data warehouse Units
I/O	Input /Output
IR	Integrated Runtime
IT	Information Technology
SSL	Secure Sockets Layer
SQL	Standard Query Language
TSL	Transport layer Security

Contents

DOCUMENT CONTROL

Document History

Version	Author	Issue Date	Changes
Version 0.1	Adeola Ogunnaike	27.04.2023	Final Draft

Project Information

Project Name	Enterprise Data Warehouse Project

Approvals List/ Signatories

Name	Organization	Designation	Signature	Date
Ayobami Ekundayo	Keystone Bank	Project Manager		
Chinedu Asogwa	Keystone Bank	Technical Project Lead		
Yvonne Ojibah	Keystone Bank	Group Head - Strategy & Implementation		
Adebiyi Adeniji	Keystone Bank	Department Head - Solutions Devt. & Analytics		
Arinze Mbaeto	Keystone Bank	Ag. Chief Information Officer		
Yemi Olaniyi	Keystone Bank	Chief Information Security Officer		
Yemi Odusanya	Keystone Bank	Executive Director - South & Corporate		
Tope Ajao	Bluechip Technologies	Head, Delivery		
Olawale Alao	Bluechip Technologies	Head, Analytics		

1 INTRODUCTION

1.1. Summary

Keystone Bank today has potentially valuable data about our customers sitting in disparate systems, but do not have a single source of truth for critical insights apart from transactions as recorded on the core banking application. In order to compete successfully in the global environment, the bank must develop the capacity to operationalize data around their core business interests. The ability to aggregate data from multi variety sources, clean, ingest and transform it into meaningful information and insight is a source of sustainable competitive advantage.

The successful implementation of this project is therefore very strategic. At the end of this project, the expectation is to have a central source of truth as envisaged by the bank in initiating this Enterprise Data warehouse project. Our Project approach will focus on an iterative delivery to ensure the Bank begins to realize benefits quickly and a laser focus on capacity development for its resources in order to sustain and expand the implementation post-go-live.

1.2. Intended Audience

The target audience for this document is all architects, solution developers, data architects, technical leads, ETL developers, Power BI developers, and other roles involved in the delivery of the various components of the Data Reservoir project.

1.3. Document Purpose

The purpose of this document is to outline the standard implementation methodology, design approach, and architecture, which would be used to deliver the Enterprise Data Warehouse Project This document would also help define the following:

- A feasible, well planned technical architecture that meets the current and future requirements of the system and its users
- The solution components and technologies necessary to implement that architecture
- The hardware configuration requirements necessary to support the system through its lifecycle

2 PROJECT DETAILS

2.1. Objectives

The main objectives of this project are;

- Creation of a Single Customer View, enabling the Bank to have a unified view of all offerings availed by a customer
- Generation of insights from deep analytics in a range of domains including (but not limited to) Customer Relationship Management (CRM), Product, Risk, MIS, Performance Management, Collections and Recovery, Data Mining, Operations, Compliance, Cost Management, Credit decisions, etc.
- Generation of analytics models in the above domains as required by the business users at all levels of the Bank for their internal purposes duly ensuring speed, data integrity, and consistency.
- Understanding customer value, profitability, and interactions better to improve service levels and drive business growth.
- Ability to focus on customer interactions and satisfaction through integrated Operational Customer Relationship Management solutions supported by analytical capabilities.
- Aid strategic and operational management teams of the Bank in the decision-making process by implementing a data warehouse and business intelligence system with device-agnostic reporting tools.
- To create a single source of truth for all enterprise data in Keystone Bank.

2.2. Project Scope

The following are expected outcomes and deliverables upon successful project completion:

- Implementation and development of Microsoft Enterprise Data Warehouse & Business Intelligence
- Design and document the Data Warehouse Solution Architecture
- Implement data governance and data quality
- Create a comprehensive data architecture
- Implement the data warehouse solution with adequate data security & protection

3 SOLUTION DESIGN

3.1. Reference Architecture

As advised by Microsoft Figure 1 below shows the Referenced Architecture advised by Microsoft Corporation, which includes additional baseline components.

NB: This is a referenced architecture from Microsoft, for this project the proposed architecture in fig 2 will be implemented.

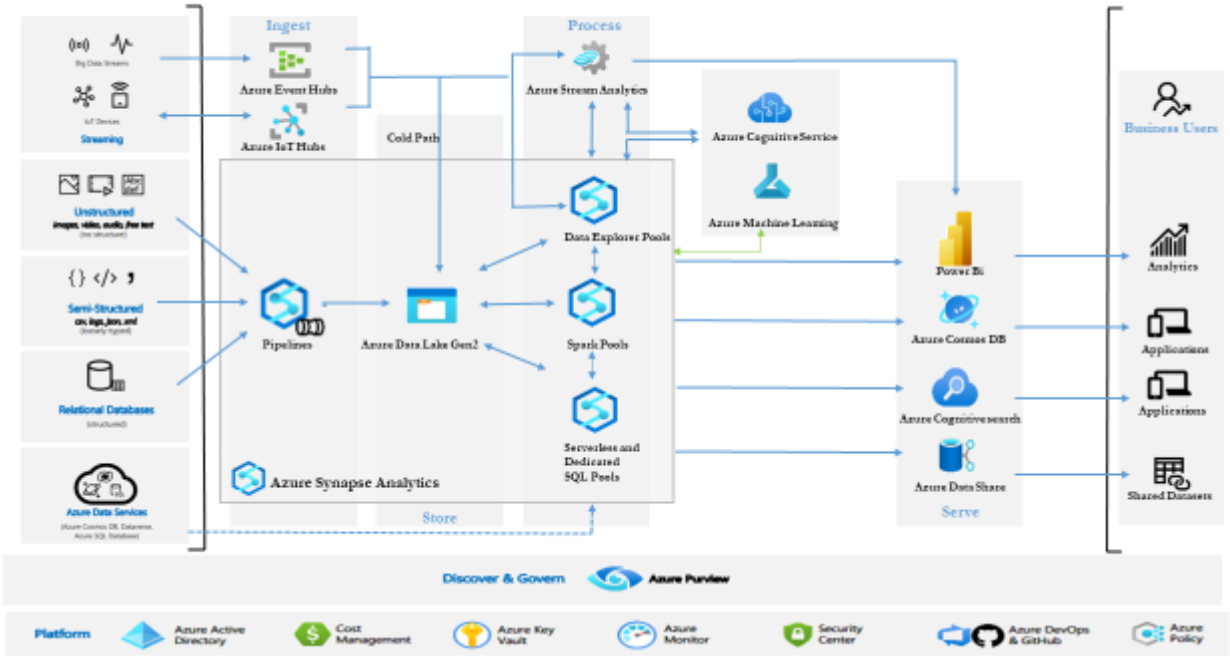


Figure 1: Modern Data Warehouse Architecture for Keystone Nigeria

3.2. Solution Architecture

Based on the activities undertaken with Keystone Business and Technical teams (Bluechip & Keystone), the Modern Data Warehouse implementation will be achieved with the logical architecture in Figure 2 below. Here the main emphasis lies in carrying out the fundamental tasks of Data Extractions from source systems and databases into ADF for Data Storage, the Azure Synapse would be used to Model the data, while Power BI would be used to serve reports and dashboards for business users

Models would be built on Azure ML components leveraging data in Azure Synapse and presented to business users

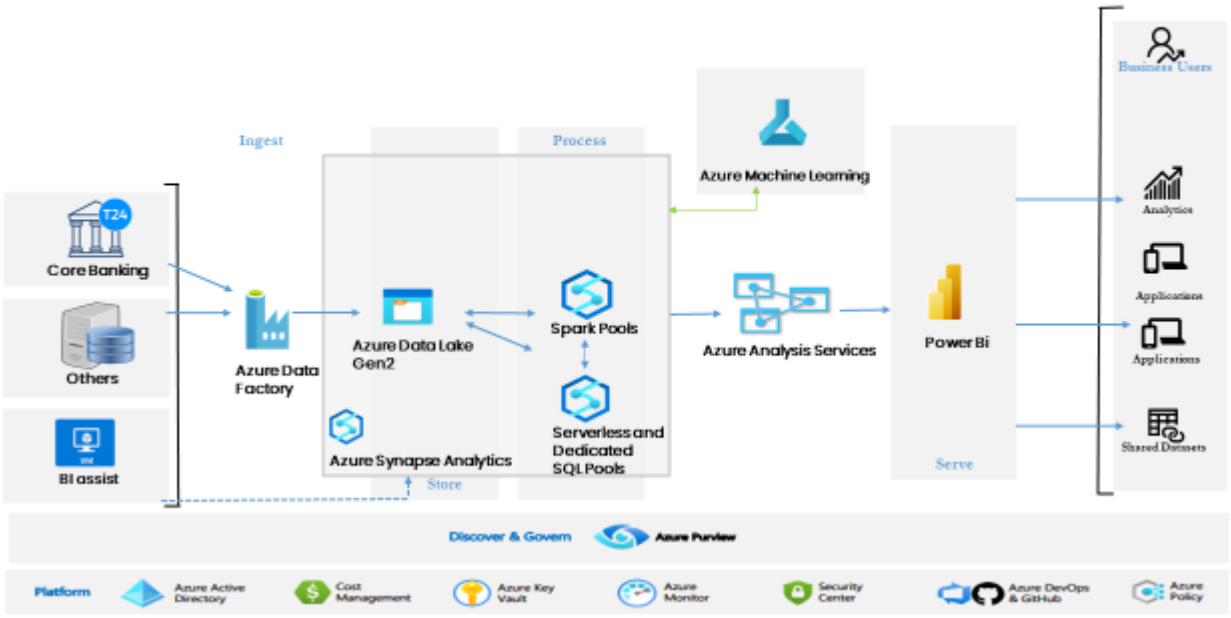


Figure 2: Modern Data Warehouse Architecture for Keystone Nigeria

3.2.1 Data Sourcing

For the purpose of this implementation, Two Major Data Sources are considered are Structured and Semi-structured Data. The Bluechip Team is responsible for extracting, loading and Transforming Data.

- Data from the Core banking application should be available in a backup/reporting server.
- The Keystone team will be responsible for moving the data to the reporting server to be used for both development and production.
- Data from reporting server are loaded into the ADLS via ADF Pipelines where further processing will take place.
- Bluechip BI Assist will be used to upload unstructured data captured in Excel Files into a preprovisioned Virtual Machine on Azure. Thereafter, the database will be integrated into the solution via structured data sources.

3.2.2 Design Considerations

The technologies in this architecture were chosen because they met Keystone's requirements for scalability and availability while helping to control cost effectively.

- Bluechip team will be responsible for data Ingestion into the Data Lake for data storage and preparation.
- Azure Data Factory will be used to Ingest structured and semi structured data into the Data Lake.

- XML Parser processing via Azure Synapse automated spark pools will periodically fetch T24 files from the ADLS, using Apache Spark spool and T-SQL procedures. The output will be stored in the Dedicated SQL Spool.
- The massively parallel processing architecture of Azure Synapse provides scalability and high performance.
- Azure Synapse has guaranteed SLAs and recommended practices for achieving high availability.
- When analysis activity is low, Keystone can scale Azure Synapse on demand, reducing or even pausing compute to lower costs.
- Keystone Bank will provide the Azure Services and Power BI licenses to be used for the project.
- The IT security protocol for this project will be in line with the standards of Keystone bank

3.3. Solution Components Description

Below are proposed items in the solution to address Project requirements:

- Azure Data Factory
- Azure Data Lake Gen2
- Azure Synapse
    - Azure Synapse Pipeline
    - Azure Spark Pool
    - Dedicated SQL pool
    - Serverless SQL pool
- Power BI
- Platform Services
- Bluechip BI Assist

3.3.1 Azure Data Factory

It is a managed cloud service that's built for complex hybrid extract-transform-load (ETL), extract-load-transform (ELT), and data integration projects. It will be used to consume the data sources in the cloud and on premise to dump data into the Azure Data Lake Storage Gen 2.The stream added to consume data for from on-premises is to be used to process batched data . There will be multiple pipelines for the respective data sources that are provided by the bank.

3.3.2 Azure Data Lake Gen2

Azure Data Lake Storage is a highly scalable and cost-effective data lake solution for big data analytics. It combines the power of a high-performance file system with massive scale and economy to help speed up time to insight. Data Lake is cost effective and has corresponding Rest APIs to be consumed by other clients such as App services. Azure Data Lake includes all the facilities required to make it easy for data scientists, developers, and analysts to store data of any shape, size, and speed. It does all types of analytics and processing across platforms and languages. It removes all the difficulties of ingesting and storing all data while making it faster to get up and running with streaming, batch, and interactive analytics.

3.3.3 Azure Synapse Analytics

Azure Synapse is a limitless analytics service that brings together enterprise data warehousing and Big Data analytics. It provides a unified environment by combining the data warehouse of SQL, the big data analytics capabilities of Spark, and data integration technologies to ease the movement of data between both, and from external data sources. Azure Synapse has four components:

- **Synapse SQL**: Complete T-SQL based analytics
    - Dedicated SQL pool (pay per DWU provisioned)
    - Serverless SQL pool (pay per TB processed)

- **Spark**: Deeply integrated Apache Spark
- **Data Integration**: Hybrid data integration
- **Synapse Pipelines**: Data Flows orchestration
- **Studio**: Unified user experience.

Synapse SQL leverages a scale-out architecture to distribute computational processing of data across multiple nodes as shown in diagram below:

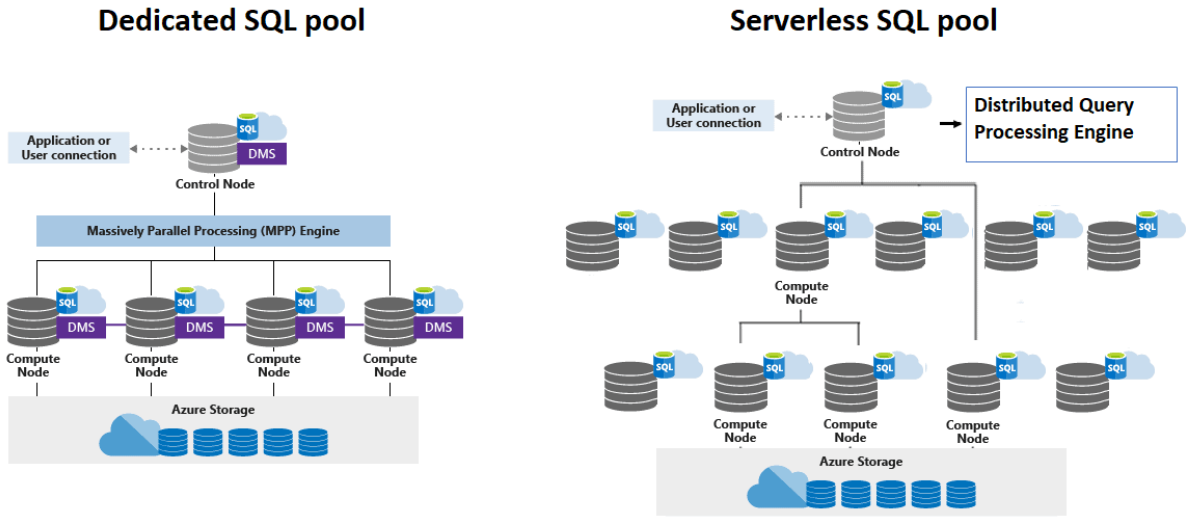


Figure 3: Azure Synapse Parallel Processing

Component	Details
Azure Storage	Synapse utilizes Azure storage to keep data safe
Control Node	The Control node is the brain of the architecture. It is the front-end that interacts with all applications and connections.
Compute Nodes	The Compute nodes provide the computational power. Distributions map to Compute nodes for processing.
Data warehouse units	A Synapse SQL pool represents a collection of analytic resources that are being provisioned. Analytic resources are defined as a combination of CPU, memory, and IO.

3.3.1 Azure Synapse Pipelines

Azure Synapse Pipelines will be used to orchestrate the Data Flow within the Synapse ecosystem such as from Transformed/Conformed to Business Curation Data. Using Azure Synapse Pipelines for orchestration, data is written to the different databases that will house the Curated data for business reports and different data products.

3.3.2. Azure Synapse Dedicated SQL Pool

Dedicated SQL Pools will expose an SQL endpoint which external tools like Power Bi will connect to. Azure Dedicated SQL Pool also makes it possible to create a Power BI dataset.
Materialized Views will be created in the Dedicated SQL Pools, which will be used to pre-calculate and store dataset ready for use in Power BI.

3.4. Power BI

This is the business analytics tool that will be used to provide interactive visualizations with business intelligence capabilities through an interface simple enough for end users to create their own reports and dashboards. For effective reports that are not limited by the licensing packages, it is recommended that Keystone bank should have Existing Power BI licenses.

3.5. Azure Machine Learning

In order to deliver analytics models – SEGMENTATION ANALYSIS & MODELLING, DELINQUENCY MODELS, CROSS-SELLING ANALYSIS MODELS, Azure Machine Learning Component will be required.
Azure Machine Learning is a secure platform designed for responsible AI Applications in machine learning. It empowers data scientists and developers to build, deploy, and manage high-quality models faster and accelerate time to value with industry-leading machine learning operations (MLOps), open-source interoperability, and integrated tools.

3.5.1 Delinquency Models

Delinquency prediction model presents a result which has been vetted, time and again, using machine learning, selecting and tuning models, and adjusting missing variables. Delinquency prediction using this method can help lenders substantially reduce their lending and refinancing risks.

Business Goal:

- Reduce non-performing loans and loan loss provisioning

Project Goal:

- Define and highlight the Definition of the default
- Develop application probability of default scoring models
- Analyze existing historical customer application data that has defaulted
- Select the development window for the models

- Proactively identify any potential hurdles such as data quality (NULLs) & availability.

Data Required: Customer loan Information and Demographic Information

3.5.2 Cross-Selling Analysis Models

The main objective of this analysis is to increase the sales revenue and profit from the already acquired/existing customer base of a company by recommending more or supplementary products to the customers. Using the Customer Segmentation model, products would be layered across segments and recommendations would be made with algorithms. If cross-sell prediction results are expected as output, then supervised classification models will be used. And if the goal is to recommend supplementary or more products to customer's recommendation system algorithm will be used.

3.5.3 Segmentation Analysis Models

The objective of this analysis is to identify segmentation approaches that can help identify customer value based on their transaction and spending patterns using state-of-the-art modeling framework i.e. RFM (Recency, frequency, and Monetary) analysis and K-means Clustering on customer Transactional data. As such knowing the customer's segment bucket and their preference of service would help in a lot of further use cases;

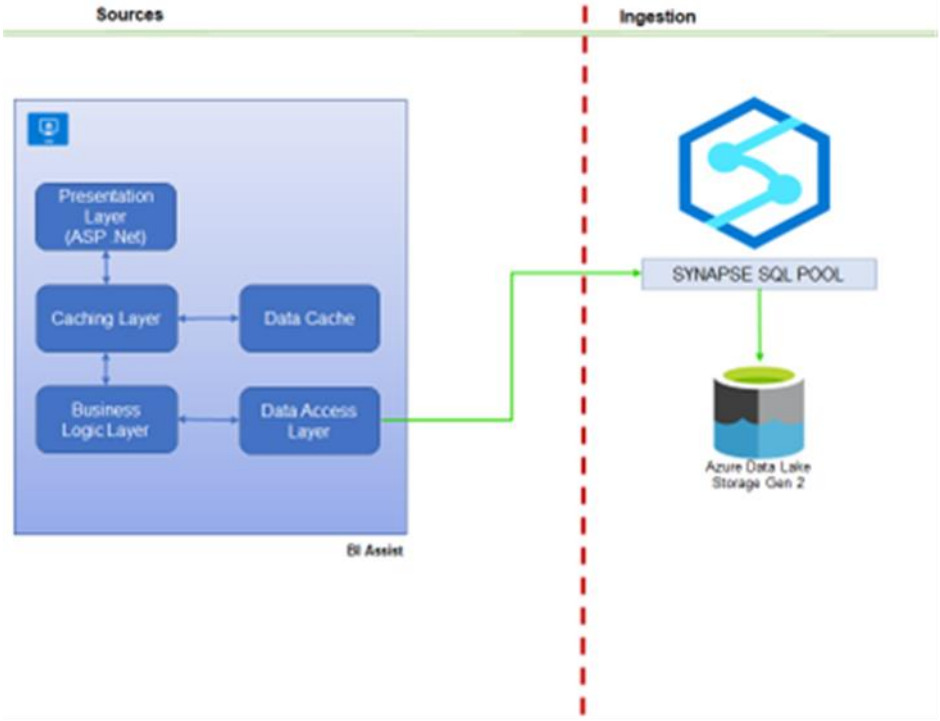
- Marketing campaigns,
- Cross-selling,
- Better customer experience,
- Customer acquisition etc.

Data Required: Demographic and Transactional Data

3.6. Bluechip BI Assist

BI Assist is a .NET application that helps data warehouse acquire data that needs to be inputted or files such as excel or CSV that are used in personal computers in a highly governed manner.
It helps consolidate all the files such as excel (CSV) that are used in personal computers, in order to centralize the processing in the Data warehouse. A walkthrough of the tool and its features will be done for the relevant stakeholders. It is an added value to the solution and at no cost. However, additional custom features may incur some cost and will impact delivery. Bluechip will provide the source code and supporting documentations such as a User manual and As-built information for the tool and therefore Keystone Application team can support it effectively.

The BI Assist portal thus becomes another data source to the Data Platform



3.7. Azure Platform Services

Azure platform services provide the foundation for building and running applications and services in the cloud. They include a wide range of infrastructure, platform, and software services that can be used to support various types of workloads, from web and mobile applications to big data and artificial intelligence.

3.7.1 Active Directory (Azure AD)

Azure Active Directory (Azure AD) is a secure your environment with multi-cloud identity and access management. It is an enterprise identity service that provides single sign-on and multi-factor authentication to help protect users from 99.9 percent of cybersecurity attacks. Azure Active Directory helps to manage user identities in one location, enable access to Azure Synapse Analytics and other Microsoft services with Azure Active Directory user identities and groups.

3.7.2 Cost Management

Azure Cost Management is used to monitor, allocate, and optimize cloud costs with transparency, accuracy, and efficiency using Microsoft Cost Management.

3.7.3 Azure Key Vault

Azure Key Vault is used to safeguard cryptographic keys and other secrets used by cloud apps and services Active Directory.

3.7.4 Azure Monitor

Azure is used to provide full observability into your applications, infrastructure, and network. It can collect, analyze, and act on telemetry data from Azure and on-premises environments. Azure Monitor helps you maximize performance and availability of your applications and proactively identify problems in seconds.

3.7.5 Azure DevOps & GitHub

Azure DevOps is an end-to-end solutions on Azure to implement DevOps practices throughout application planning, development, delivery, and operations. It allows for application of the right combination of DevOps technologies, culture, and processes to enable continual software delivery and better value for customers.

3.7.6 Azure Policy

Azure Policy helps to enforce organizational standards and to assess compliance at-scale. Through its compliance dashboard, it provides an aggregated view to evaluate the overall state of the environment, with the ability to drill down to the per-resource, per-policy granularity. It also helps to bring your resources to compliance through bulk remediation for existing resources and automatic remediation for new resources.

4 SECURITY

4.1 Synapse Dedicated SQL Pool Security

4.1.1 Network Connection Security

Firewall rules are used by both the logical SQL server and its databases to reject connection attempts from IP addresses that haven't been explicitly approved. To allow connections from your application or client machine's public IP address, you must first create a server-level firewall rule using the Azure portal, REST API, or PowerShell.

4.1.2 Authentication

Dedicated SQL pool currently supports SQL Server Authentication with a username and password, and with Azure Active Directory.

To connect to dedicated SQL pool, the following information must be provided:

- Fully qualified server name
- Specify SQL authentication
- Username
- Password

4.1.3 Azure Storage Firewalls

Azure Storage provides a layered security model. This model enables you to secure and control the level of access to your storage accounts that your applications and enterprise environments demand, based on the type and subset of networks or resources used. When network rules are configured, only applications requesting data over the specified set of networks or through the specified set of Azure resources can access a storage account. You can limit access to your storage account to requests originating from specified IP addresses, IP ranges, subnets in an Azure Virtual Network (VNet), or resource instances of some Azure services.

Storage accounts have a public endpoint that is accessible through the internet. You can also create Private Endpoints for your storage account, which assigns a private IP address from your VNet to the storage account, and secures all traffic between your VNet and the storage account over a private link.

4.1.4 Transport Layer Security

This ensures that data is encrypted in transit to and from the database and reduces susceptibility to "man-in-the-middle". TLS 1.2 will be used to ensure security for data in transit.

4.1.5 Authorization

Authorization privileges are determined by role memberships and permissions. There are ways to further limit what a user can do within the database:

- Granular Permissions controls which operations can be carried out on columns, tables, views, schemas, procedures, and other objects in the database.
- Database roles other than db_datareader and db_datawriter can be used to create more powerful application user accounts or less powerful management accounts.
- Stored procedures can be used to limit the actions that can be taken on the database.

4.1.6 Azure Active Directory

Azure Active Directory will be used to control access to the data in the Azure Data Lake. Azure AD will control Authentication and authorization of users into azure resources though Role based Access Control (RBAC).

5 STANDARDS

The following object prefixes will be adopted for the Data Reservoir Project

5.1 Azure

ADLS_ Azure Data Lake Storage Gen 2
AFIL_ Azure File System
ASQL_ Azure SQL Database
ASDW_ Azure Synapse Analytics

5.1.1 Files

FILE_ File System
FTP_ FTP
SFTP_ SFTP

5.1.2 Naming Convention

The appropriate naming conventions that will be adhered to for each of the solution development components. This should include Fact tables, Dimension tables, Temporary tables, Schemas, Report names, data tables, Packages etc.

	Database Objects	Naming Conversion	Description
1	Temporary tables	Tmp_	Tables used to store data temporarily during processing
2	Derived tables.	Drv_	Tables that store processed data
3	Facts	Fact_	Database objects used to store quantitative data for analysis
4	Procedures	Prc_	Database objects used for procedural data processing and design
5	Functions	Fn_	Database objects used for data manipulation, processing and design
6	Indexes	Ind_	Database objects used for table performance tuning
7	Dimensions	Dim_	Measures used to categories data to answer business questions

5.2 System Requirements

Resource	Resource Description
PROD – BI Assist Application	1 D4s v5 (4 vCPUs, 16 GB RAM) Windows (License Included), OS Only; 1 X 512 Gb Standard SSD managed disk
TEST – BI Assist Application	1 D2 v3 (2 vCPUs, 8 GB RAM)Windows (License Included), OS Only; 1 X 256 Standard SSD managed disk
PROD - Data Storage Gen 2	Data Lake Storage Gen2, Standard, LRS Redundancy, Hot Access Tier, Flat Namespace File Structure, 10,000 GB Capacity - Pay as you go, Write operations: 4 MB x 100,000 operations, Read operations: 4 100,000 x MB operations, 100,000 Iterative read operations, 100,000 Iterative write operations, 100,000 Other operations. 1,000 GB Data Retrieval, 1,000 GB Data Write.
PROD - Azure Synapse Analytics	Tier: Compute Optimized Gen2, Dedicated SQL Pools: DWU 100 x 1 Month, 10 TB of storage; Serverless SQL Pools: 10 TB of data queried; Apache Spark Pools: Memory optimized: (1 Instance(s), Small x 300 Hours); West Europe Region, 0 GB of data collected per day, 7 days of Hot Cache, 30 days of total retention, 7 times estimated data compression, 0 Hours of 2 x Extra Small (2 vCores) Engine Instances, 0 Hours of 2 x 1 vCore Data Management Instances.
PROD – Azure Data Factory	Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 10 Activity Run(s), 100 Data movement unit(s), 1,000 Pipeline activities, 1,000 Pipeline activities – External, Azure VNET Integration Runtime: 10 Activity Run(s), 50 Data movement unit(s), 50 Pipeline activities, 50 Pipeline activities – External, Self-hosted Integration Runtime: 10 Activity Run(s), 1,000 Data movement unit(s), 1,000 Pipeline activities, 1,000 Pipeline activities – External, Data Flow: 0 x 8 General Purpose vCores x 100 Hours, 1 x 8 Memory Optimized vCores x 100 Hours, Data Factory Operations: 10 x 50,000 Read/Write operation(s), 10 x 50,000 Monitoring operation(s).
Azure Machine Learning	1 D4ds v4 (4 Core(s), 16 GB RAM) x 586 Hours, Pay as you go
Azure Analysis Service	Basic B2 (Hours), 1 Instance(s), 1 Month
Power BI	Power BI existing Licences

REFERENCES

- Enterprise Data Warehouse (https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/enterprise-data-warehouse)
- Azure Synapse SQL Architecture (https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/overview-architecture)
- Azure Data Lake Storage Gen2 https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction
- Microsoft Power BI Pro https://powerbi.microsoft.com/en-us/power-bi-pro/
- Azure Analysis Services https://docs.microsoft.com/en-us/azure/analysis-services/analysis-services-overview
- Microsoft SLA for Cloud Services (Service Level Agreements - Home | Microsoft Azure)