API Data Source

FUTURE STATE STRATEGY

LAST UPDATE: October 11, 2024

2024

Contents

[1 Future State Vision 2](#_Toc179710231)

[1.1 Key Considerations based on Current State – Gaps and Opportunities 2](#_Toc179710232)

[1.1.1 Key Technical Capabilities 2](#_Toc179710233)

[1.1.2 Key Business Capabilities 2](#_Toc179710234)

[2 Future State Design 3](#_Toc179710235)

[2.1 End to End Future State Architecture 3](#_Toc179710236)

[2.2 Technology Stack 3](#_Toc179710237)

[2.3 Mapping of Current State Capabilities to the Future State Architecture 4](#_Toc179710238)

[2.3.1 Key Capabilities which need to be maintained 4](#_Toc179710239)

[2.3.2 Key Capabilities which can be improved 5](#_Toc179710240)

[2.3.3 Best Practices 5](#_Toc179710241)

[3 Recommendations and Roadmap 6](#_Toc179710242)

[3.1 Phases of Future State Adoption 6](#_Toc179710243)

[3.2 Cloud Data Migration Approach 6](#_Toc179710244)

[3.3 Cloud Data Migration Governance Framework 6](#_Toc179710245)

[3.4 Cloud Data Operating Blueprint 6](#_Toc179710246)

# Future State Vision

<

Describe an overview of the future state vision of this project. Provides a clear and compelling view of what it aims to achieve to ensure all stakeholders understand and support the desired outcome(s).

>

## Key Considerations based on Current State – Gaps and Opportunities

* Simplicity
* Improved performance
* Implement proper medallion architecture
* Empower end-consumers to subscribe and use their desired datasets

### Key Technical Capabilities

<

* System Integration
* Cloud Computing
* Security and Data Protection
* Network Infrastructure
* User Experience

>

### Key Business Capabilities

<

* Operational Efficiency, visibility and Analytics
* Product Development and Innovation
* Risk Management and Compliance
* Collaboration and Productivity
* Data Security and Privacy

>

# Future State Design

## End to End Future State Architecture

A diagram of a computer

Description automatically generated

## Technology Stack

* **Azure Logic Apps** – Azure orchestration capability is used to make API calls to API data sources in-scope.
* **Azure Blob Storage** – Azure storage capability to store landing zone data and Interim Storage data to upload data into Bronze Layer of RAD Data Store.
* **Azure Data Factory** - A fully managed, serverless data integration service to integrate data sources using built-in ETL services. It is used to ETL data into the Bronze layer of the Medallion architecture.
* **Azure Databricks** – Databricks helps to connect many data sources to one platform to process, store, share, and analyze data. It is used mainly for the Data Science use cases.
* **Azure Active Directory** – A cloud-based identity and access management solution that provides authentication and authorization services. It is used to authenticate users attempting access RAD Data Modernization platform and validate authorization to requested resources.
* **Azure Key Vault** - A cloud service for securely storing and accessing secrets. It supports two types of containers: vaults and managed hardware security module(HSM) pools. Credentials used in all RAD Data Modernization utilized services are stored in the Azure Key Vault.
* **Bronze Layer** – Also known as Landing zone for storing data in its original state in a medallion architecture, where data is transformed in various stages. Data pulled from API data sources is stored in this layer.
* **Silver Layer** – Also known as Refined Layer. This is where the standardization to source datasets is applied, which aligns field names across sources, applies data cleaning operations and organizes the data into well-known structures. RAD Data in-scope will be standardized in the Silver Layer.
* **Gold Layer** – Also known as Fit-for-Purpose layer. It is the final stage of data transformation and quality assurance of the Medallion architecture. It is used for reporting and uses more de-normalized and read-optimized data models.
* **Virtual Warehouse** - A virtual warehouse (VW) in Snowflake is a group of compute resources such as, CPU, memory, and temporary storage that provides the resources needed to perform tasks such as, SQL execution and DML operations. Multiple Virtual Warehouses (VW) will be used in multiple places in the RAD Data Modernization data pipelines, as well as for end-user use cases.
* **HCP Data Catalog** - A component of the healthcare platform that helps to manage and organize data assets. It provides a centralized catalog of data objects, metadata, and data. RAD Data Modernization project will leverage HCP catalog by loading it and serving end-consumers from it.
* **Data Product** - It is an intersection of business domains, organizations, and the data needed to support them. It contains significantly more technical and business metadata. A data product contains references to one or more datasets. End-consumer use cases will be served using data products in most of the cases.
* **Gold Layer Metadata** – A backend job, that runs daily, emits metadata of the data sitting in the Gold Layer. This metadata helps user create data products, which is searchable in the HCP Data Catalog.
* **Silver Layer Metadata** – same as Gold Layer metadata, except it refers to the data in the Silver Layer.
* **Compute Tenant** - It provides the ability to consume data that is available on HCP data catalog quickly and securely. Users can search for and subscribe to data products directly. A Compute tenant provides a user with a group of resources, e.g., warehouse to query the tables or views and databases and schemas to store queried information. A dedicated compute tenant is used to serve an end-consumer subscription group.

## Mapping of Current State Capabilities to the Future State Architecture

* Even though the Gold Layer data objects and structure will remain consistent with current state. However, the transformation process will change, e.g., there will be distinct layers of the medallion architecture. Transformation will be simplified and modularized. And the data platform will be migrated from on-prem to Azure, Snowflake.

### Key Capabilities which need to be maintained

* Existing APIs will remain the same, except for being called from the Azure Logic Apps
* Data serving (Gold) layer will remain the same

### Key Capabilities which can be improved

* Pseudo Medallion architecture will be redefined as formal Medallion architecture
* Complex ETL process and approach will be simplified and modularized
* Performance will be improved
* End-consumers will be able to get to the business value faster

### Best Practices

* **Ingestion**
  + **Understand the requirements**: Understand current data ingestion requirements and future scalability needs. Consider factors such as data volume, velocity, variety, and latency requirements.
  + **Design scalable architecture**: The architecture should be able to handle the expected data volume and accommodate future growth. Use scalable services, such as, configuration-based Infrastructure as Code.
  + **Ensure data security**: Implement robust security measures to protect data during ingestion. Use encryption in transit and at rest. Implement access controls and authentication mechanisms to prevent unauthorized access.
  + **Optimize data transfer**: Minimize the time and cost of data transfer by compressing and batching data before ingestion.
  + **Monitor and track data ingestion**: Use cloud-native monitoring tools or third-party solutions to identify and resolve any issues proactively.
  + **Implement data validation and transformation**: Validate and transform data during ingestion to ensure its quality and consistency.
  + **Implement data governance and compliance**: Adhere to data governance policies and compliance regulations during data ingestion.
  + **Automate data ingestion**: Automate data ingestion processes to minimize manual efforts and reduce the risk of errors.
  + **Test and validate**: Before fully transitioning, perform thorough testing and validation of the data ingestion pipeline(s).
* **Transformation**
  + **Plan and document**: Create a clear plan for the data transformation process, documenting all the steps involved.
  + **Use appropriate tools**: Choose the right tools and technologies for data transformation based on the data type, platform(s), complexity and data volume.
  + **Validate and clean data**: Before transforming the data, perform validation checks to ensure data integrity and quality. Use data cleaning techniques such as removing duplicates, standardizing formats, or addressing missing values.
  + **Use modular and reusable code**: Break down the data transformation process into modular steps. Use functions or libraries where appropriate to simplify complex transformations.
  + **Handle errors and exceptions**: Implement error handling mechanisms to capture and handle any errors or exceptions that may occur during the transformation process to minimize data loss.
  + **Optimize performance**: When dealing with large datasets, consider performance optimization techniques such as, parallel processing and optimization algorithms.
  + **Test thoroughly**: Perform comprehensive testing. use different scenarios, edge cases, and sample data to validate accuracy and consistency of the transformations.
  + **Document and maintain**: Document all the transformations, assumptions, and decisions made during the process.
* **Consumption**
  + **Clearly define requirements**: Clearly articulate and document the requirements for the data product you need.
  + **Select reliable data sources**: Choose data products from reliable and trusted sources such as HCP Data Catalog
  + **Understand the data product**: Take time to thoroughly understand the data product you are intending to consume.
  + **Foster a data-driven culture**: Encourage a data-driven culture by promoting the use of data products for decision-making and problem-solving.
  + **Comprehensive data product subscription responses**: Provide as complete and transparent responses as possible to prevent rejection of your request by the governance team(s).
  + **Secure Global Governance approval**: Get global governance review as soon as possible where needed.
  + **Data Usage Requirements**: Fill out the DUR forms and receive approvals.
  + **Ensure PHI and PII data Governance controls**: If data contains PHI or PII data, ensure that appropriate approvals are secured.

# Recommendations and Roadmap

## Phases of Future State Adoption

## Cloud Data Migration Approach

## Cloud Data Migration Governance Framework

## Cloud Data Operating Blueprint