DATA ARCHITECTURE

AS-IS Data Architecture

Abstract

This document describes the baseline (As Is) Data Architecture

Nozom technologies

I. Contents

I. Introduction 5
A. Data Architecture Definition: 5
B. Data Architecture Principles 5
C. Objectives of Data Architecture 10
D. List of Artifacts 10
II. Data Architecture Artifacts 12
1. Data Entity Catalog 12
1. Description: 12
2. List of Attributes 12
3. Data Entity Catalog 13
4. Observations 13
2. Data Entity to Business Function Mapping 13
1. Description: 13
2. Mapping sheet 13
3. Observations and Conclusions 13
3. Data Entity to Business Service Mapping 14
1. Description: 14
2. Mapping sheet 14
3. Observations 14
4. Data Flow Diagram 14
1. Description: 14
2. Data Flow Diagram(s) 14
3. Observations and Conclusions 14
5. Data Warehouse/Data Lake 14
1. Description: 14
2. Solution Architecture Diagram: 15
6. Database Catalog 19
1. Description: 19
2. List of Attributes 19
3. Database Catalog 20
4. Observations 20
7. Data Management 20

1. Master Data/Reference Data Management 20
2. Data Quality & Data Governance 21
3. Data Exchange and Interoperability 22
4. Data Storage and Retention 22
5. Data Security 23

III. Overall Summary 25

IV. Glossary of Terms 27

V. References 28

Document Information

Project Name:	EA for SCFHS		
Prepared By:	Ahmed Imam	**Document Version No:**	1.0
Title:	Data Architecture Baseline (AS IS)	**Document Version Date:**	12/09/2022
Reviewed By:	Amr Gaber	**Review Date:**	13/09/2022

To	Action*	Due Date	Phone/Fax/Email

* Action Types: Approve, Review, Inform, File, Action Required, Attend Meeting, Other (please specify)

Document Version History

Version Number	Version Date	Revised By	Description	Filename
0.1	10/09/2022	Amr Gaber	Initial version of the document	
0.2	11/09/2022	Amr Gaber	Updating the document and attached files	
1.0	12/09/2022	Amr Gaber	Final version	

I. Introduction

A. Data Architecture Definition:

Data architecture describes the structure of an organization's logical and physical data assets and data management resources. – TOGAF

B. Data Architecture Principles

Refer to all Data Architecture principles here. Refer to EA principles document, in particular the section for Data Architecture principles.

Name	**Data is an Asset**
Reference	DP-01
Statement	Data is an asset that has value to the enterprise and is managed accordingly.
Rationale	Data is a valuable corporate resource; it has real, measurable value. In simple terms, the purpose of data is to aid decision-making. Accurate, timely data is critical to accurate, timely decisions. Most corporate assets are carefully managed, and data is no exception. Data is the foundation of our decision-making, so we must also carefully manage data to ensure that we know where it is, can rely upon its accuracy, and can obtain it when and where we need it
Implications	• Ensure that all organizations within the enterprise understand the relationship between value of data, sharing of data, archival/backup, and accessibility to data. Stewards must have the authority and means to manage the data for which they are accountable. • The role of data steward is critical because obsolete, incorrect, or inconsistent data could be passed to enterprise personnel and adversely affect decisions across the enterprise. Part of the role of data steward, who manages the data, is to ensure data quality. Respective procedures must be developed and used to prevent and correct errors in the information and to improve those processes that produce flawed information. Data quality will need to be measured and steps taken to improve data quality - Relevant data policies and procedures will need to be developed and published across stakeholders. • A forum with comprehensive enterprise-wide representation should decide on process changes suggested by the steward. • Since data is an asset of value to the entire enterprise, data stewards accountable for properly managing the data must be assigned at the enterprise level. • Proper data archival/backup policies must be established and aligned with business continuity plans and government law and regulations.

Name	**Data Trustee is accountable for data quality**
Reference	DP-02
Statement	Each data element has a trustee accountable for data quality.
Rationale	As the degree of data sharing grows and business units rely upon common information, it becomes essential that only the data trustee makes decisions about the content of data. Since data can lose its integrity when it is entered multiple times, the data trustee will have sole responsibility for data entry which eliminates redundant human effort and data storage resources. Note: A trustee is different than a steward - a trustee is responsible for accuracy and currency of the data, while responsibilities of a steward may be broader and include data standardization and definition tasks.
Implications	• Real trusteeship dissolves the data "ownership" issues and allows the data to be available to meet all users' needs. This implies that a cultural change from data "ownership" to data "trusteeship" may be required. • The data trustee will be responsible for meeting quality requirements levied upon the data for which the trustee is accountable. • It is essential that the trustee has the ability to provide user confidence in the data based upon attributes such as "data source". • It is essential to identify the true source of the data in order that the data authority can be assigned this trustee responsibility. This does not mean that classified sources will be revealed, nor does it mean the source will be the trustee. • Information should be captured electronically once and immediately validated as close to the source as possible. Quality control measures must be implemented to ensure the integrity of the data. • As a result of sharing data across the enterprise, the trustee is accountable and responsible for the accuracy and currency of their designated data element(s) and, subsequently, must then recognize the importance of this trusteeship responsibility.

Name	**Data Access from disparate sources**
Reference	DP-03
Statement	The architecture should have clear functional layer to access the data residing in other applications. Data will be entered once, and only once, as close to its source as possible. Master & Reference Data should always be accessed, synced to/from master data management system. Transactional data if residing outside of the same database should be accessed from enterprise-wide data lake or operational datastore unless real-time data is required. Direct access to database of other application is not allowed
Rationale	The heterogeneous technology landscape makes the data access difficult and error prone. Using information must be considered from an enterprise perspective to allow access by a wide variety of applications and users. Staff time is saved, and consistency of data is improved. The layering would ensure the complexity will be encapsulated and shield the dependant layers. Improve accessibility, understanding, and usability of data for end users by providing a line of sight into data and its characteristics, and specific tools for AI development, and transforming/maintaining data sets.
Implications	• Avoid duplicates and lesser reliance on the data synchronization within the same data source. • The heterogeneous environment and lack of common master & reference data will lead to have multiple copies of data maintained across different environments. • Real-time data access requirements of other applications can be solved through publishing events or publishing APIs.

Name	Information Relevance and value
Reference	DP-04
Statement	Data must be business relevant and have value
Rationale	Relevance denotes how well a retrieved document or set of documents meets the information need of the user (Information Retrieval). Retaining data is a costly and irrelevant data has no value. Relevant data and information are more able to be identified and accessed where it is not hidden amongst irrelevant material.
Implications	• Data search-ability using effective meta-data, classifications and archiving provides for access to information that is relevant to the user. • Data should be in a format that is relevant

Name	Common Vocabulary and Data Definitions
Reference	DP-05
Statement	Data is defined consistently throughout the enterprise, and the definitions are understandable and available to all users.
Rationale	The data that will be used in the development of applications must have a common definition throughout the enterprise to enable sharing of data. A common vocabulary will facilitate communications and enable dialogue to be effective. In addition, it is required to interface systems and exchange data.
Implications	• The enterprise must establish the initial common vocabulary for the business. The definitions will be used uniformly throughout the enterprise. • Whenever a new data definition is required, the definition effort will be co-ordinated and reconciled with the corporate "glossary" of data descriptions. • Functional data administration responsibilities must be assigned.

Name	**Data Security**
Reference	DP-06
Statement	Data is protected from unauthorized use and disclosure. In addition to the traditional aspects of security classification, this includes, but is not limited to, protection of pre-decisional, sensitive, source selection-sensitive, and proprietary information.
Rationale	Open sharing of information and the release of information via relevant legislation must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information. Existing laws and regulations require the safeguarding of national security and the privacy of data, while permitting free and open access.
Implications	• Access to information based on a need-to-know policy will force regular reviews of the body of information. • In order to adequately provide access to open information while maintaining secure information, security needs must be identified and developed at the data level, not the application level. • Data at rest must be encrypted and by masked by-default and only un-masked based on proper justification. • Information security and data governance policies must be enforced on all data access.

Name	**Data Lake – Single source of truth for the enterprise**
Reference	DP-07
Statement	All the data in the organization required for analysis and research should be available in the Enterprise Data Lake.
Rationale	There is a single view for everyone and everything. This centralization view is intended to give the business whatever it needs and provide a single point where all business users can go. Data Lake is an architectural approach to information management, analytics and reporting that better matches the culture of business and better enables organizations to truly leverage the value of their information.

Implications	• Data Lake should have separate logical layers for ingestion, transformation & enrichment, handling and storing the data. The data in the data lake can be structured, semi-structured, or unstructured.
	• Data Layers in lake can be organized into:
	  ○ Raw (Data Ingestion layer),
	  ○ Curated (Processed Trusted records layer),
	  ○ Research Zone (Ephemeral data environment for time bound research and exploration purpose),
	  ○ De-identified Zone (Sandbox environment for external researchers)
	• Raw data should always be kept intact in the Raw layer
	• Architecture should be flexible to support creation of additional logical layers to handle specific purpose.
	• Existing data management policies of the enterprise should be supported by the Data Lake implementation.
	• Robust encryption and security controls are critical to meet regulatory compliance, company policy, and end-user security requirements.

C. Objectives of Data Architecture

Objective is to define the major types and sources of data necessary to support the business, in a way that is (a) understandable to stakeholders, (b) complete and consistent, (c) stable over a period of time. Having the right data at the right time is critical to an organization's operations. Data architecture creates the pathway to manage and govern the usage of data.

Key objectives:

1) Understand the conceptual/logical data entities
2) How data entities support business functions and capabilities.
3) How data is managed through its lifecycle
4) How data security is managed
5) Identify gaps in data management/data dissemination capabilities

D. List of Artifacts

List of Data Architecture artifacts is given below.

Artifact	Description	Type
Data Entity Catalog	List of conceptual/logical data entities. The catalog will have attributes for each data entity to describe it meaningfully.	Catalog
Data Entity to Business Function mapping	Which business function uses which data entity	Matrix
Data Entity to Business Service	Which business service uses	Matrix

mapping	which data entity	
Data Flow Diagram	Diagram to show all major data flows and data repositories	Diagram
Database Catalog	List of actual databases with attributes like size, expected growth, purpose, DB technology, performance qualities, any existing issues etc.	Catalog
Data Management	Critical areas of managing data that are closely related to Data Architecture.	Text, Matrix, Diagram

II. Data Architecture Artifacts

1. Data Entity Catalog

1. Description:

It is a list of all conceptual/logical data entities.

2. List of Attributes

Each Data Entity is described by a set of attributes. In the table below, we elaborate the nature and purpose of each attribute

Attribute	Attribute Definition	Purpose
Name	Name of the Data Entity	Identifies the data entity
Description	Describe the purpose and logical explanation of the entity	To understand the conceptual/logical idea behind the entity
Data Owner	Name of person/org unit who is responsible for creating/updating the data entity	Ownership is an important attribute because owners are accountable/responsible for the accuracy and timeliness of the data
Data Steward	Responsible for managing data quality, data storage etc.	
Data Area	Data Area this data entity belongs to (as per DRM)	Highest level of logical grouping of data.
Data Subject	Data Subject this data entity belongs to (as per DRM)	Logical grouping to classify data
Data Group	Data Group this data entity belongs to (as per DRM)	Logical grouping to classify data
Security Classification	Possible classifications are: Public, Sensitive, Protected, Personal	This guides with whom and how the data can be shared. It is also relevant for deciding how to store this type of data. (For example, “Protected” data may be encrypted at rest)
Master Data	Is this entity to be considered master data (yes/no)	Master data requires special handing. Changing of master

		data may require additional controls
Retention Policy	Data Owner(s) is/are responsible for retention policy	Some data entities are more important than the others. So, the data retention policy should ideally be applied at the entity level

3. Data Entity Catalog

Data Entity catalog is attached here.

Data Entity Catalog V 1.0.xlsx

4. Observations

Based on the data collected, we can see thirty-five (conceptual) data entities. The data entities represent the core business of the organization, managing training and accreditation of healthcare practitioners. Currently, the data entities are only partially mapped to data owners. There is no information on data stewards. Other attributes are also missing. Capturing complete information about the data entities should be taken up next. It is possible more data entities will be discovered/documented in the next phase of EA definition.

2. Data Entity to Business Function Mapping

1. Description:

It is a mapping of which Business Function produces/consumes which Data Entities. It allows us to understand the relationship of data with Business Functions.

2. Mapping sheet

Refer to the section on Business Functions Catalog in the Business Architecture document – [4].

3. Observations and Conclusions

This mapping is captured in Business Functions catalog. From the collected data, it appears the data entities appear in three groups across business functions. At this point, there is no observation related to this mapping. The collected data may be analyzed at a later point when more information is available about other related artifacts, for example, business processes

3. Data Entity to Business Service Mapping

1. Description:

It is a mapping of which Data Entity is consumed/produced by which Business Service. It allows us to understand relationship between data entities and business services.

2. Mapping sheet

Refer to the section on Business Services Catalog in the Business Architecture document – [4]

3. Observations

This mapping is captured in the business service catalog. At this point, there is no observation related to this mapping. The collected data may be analyzed at a later point when more information is available about other related artifacts, for example, business processes.

4. Data Flow Diagram

1. Description:

It provides a pictorial view of all the major data flows and data repositories where the data entities are hosted. It should describe the nature of the flow (synchronous, asynchronous, batch, periodic etc). Other characteristics, like volume, may also be covered if relevant.

2. Data Flow Diagram(s)

There could more than one diagrams/text illustrating different aspects of data flows. Note that we are not looking at detailed solution designs here. Just a view of major systems and high level (conceptual) data entities.

Data Flow diagrams are not available at present.

3. Observations and Conclusions

No observations as diagrams are not available.

5. Data Warehouse/Data Lake

1. Description:

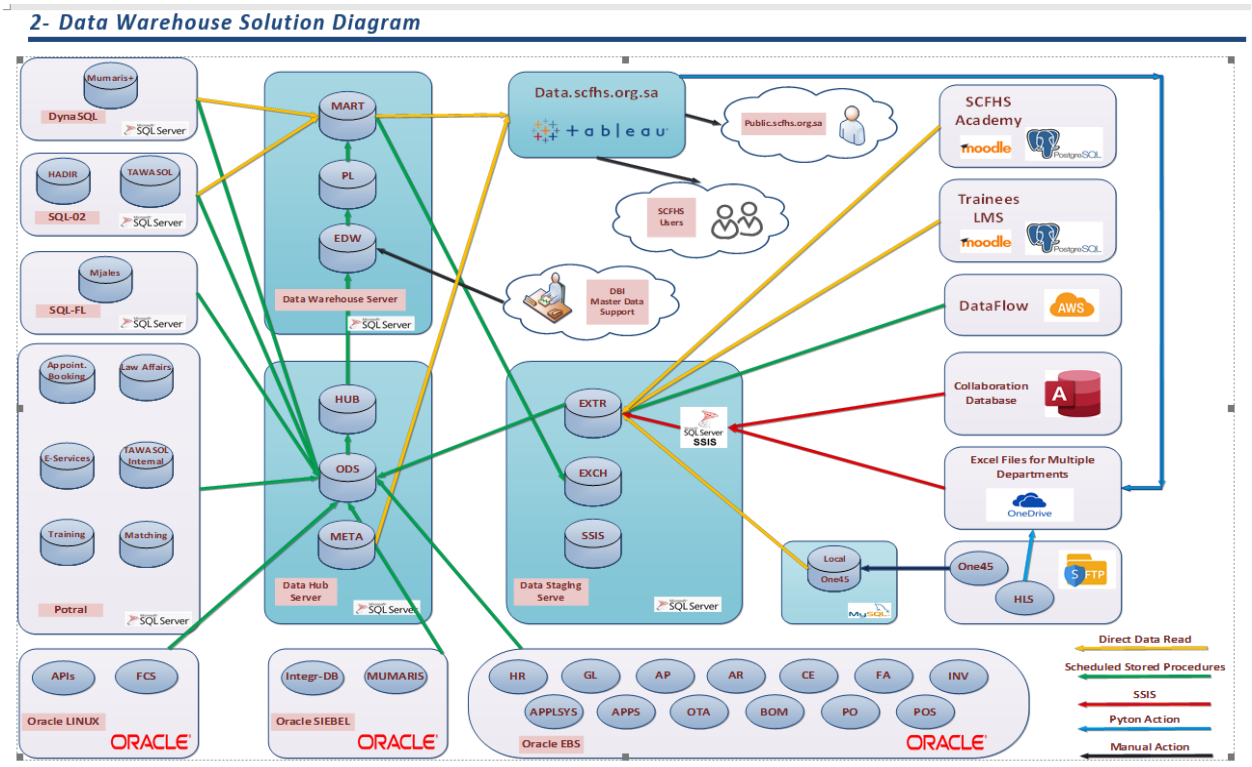
A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

A data warehouse allows an organization to consolidate data from a large number of sources into one place where they can be analyzed, cross-referenced, queried, visualized, and delivered to data consumers. By definition, it is an organized consolidation of data. Traditionally, a data warehouse has served as a repository of structured data.

A Data Lake, on the other hand, is not a structured repository and designed more for data scientists to explore and find insights.

From Data Architecture perspective, we are interested here in the current state with respect to these capabilities.

2. Solution Architecture Diagram:



e

rver is the selective technology for data warehouse' databases with 3 Production Servers, each server contains multiple databases with specified purpose per database

	Server	Database	Description
1	Data Hub Server 172.20.121.49 RUH-DHSQL-01	SCFHS-DDO-META	Data Management database handles 3 main tasks (Metadata Management – Data Quality Management – Data Warehouse Support), all these tasks are fully monitored and logged with detailed issues reporting
		SCFHS-DDO-ODS	Staging tier is created to define the data changes for all data sources (Inserts – Updates -Deletes) Filling this tier done using Merge Statements
		SCFHS-DDO-HUB	Staging tier is created to define the data changes for all data sources (Inserts – Updates -Deletes) Filling this tier done using Merge Statements

	Server	Database	Description
2	Data Warehouse Server 172.20.121.50 RUH-DWSQL-01	SCFHS-DDO-EDW	Normalized database where a single Unified database is used to consolidate, clean, integrate the data and apply Data Quality actions
		SCFHS-DDO-PL	De-Normalized database where a single Unified database is created for star schema modeling (Dimension & Facts) with all derived business values.
		SCFHS-DDO-MART	Data Marts physically located as Indexed tables and database views used extract business intelligence reports and allow business users to build their reports easily without the need of advanced technical skills.
3	Data Staging Server 172.20.121.48 RUH-SSIS-02	SCFHS-DDO-EXTR	Collect data from other external sources like Excel files and access Works as a bridge for data for other technology like MySQL and PostgreSQL
		SCFHS-DDO-EXCH	Collect data from services which require interactions and transactional operations like GSB & APIs And works as source for data sharing with other systems
		SSIS	Serves SQL Server Integration Services repository

Data Sources:

- Data Sources are essential part for Data Warehouse. These data sources differ technologies Oracle, Microsoft SQL Server, MySQL, PostgreSQL, Access and Excel Files
- Once Connection Established to data source then same connection can be used for Metadata, Data Quality and Data warehouse data load
- Databases (Oracle & MS SQL) are connected directly through database links
- Other DB technologies like (MySQL & PostgreSQL) accessed through bridged views in EXTR DB located in Data Staging Server, these views transform the data into MSSQL format

	Server	Database	Description
1	Portal Server 172.20.121.42 PORTAL-DB MS SQL Server	Training System Trainees_DB	Training Management System
		Matching SCFHSMatchingV11	Residency Enrollment Application (Matching)
		E-Services SPSCFHSServices	Database contains various services for SCFHS like License Exams, Transportation requests and Written Exams requests
		Appoint.Booking DB_AppointmentsBookingv2	Appointment Booking Service for Client Care Department

	Server	Database	Description
		TAWASOL Internal Database TawasulInternalDB	Database for Client Care department, yet to explore and identify purpose
		Law Affairs Database Law_DB	Law Affairs system holds Law Cases, this system is newly developed and yet not launched
		HealthAcademy	Health Academy data
		Wellbeing	Practitioners' issues
2	DynaSql Server 172.20.118.25 LISTN-DYNASQL MS SQL Server	Mumaris+ Mumaris_MSCRM	Mumaris+ System, CRM system implemented by Accenture with Self-Services provided to Health Practitioners
3	SQL-02 Server 172.20.121.104 RUH-SQL-02 MS SQL Server	TAWASOL servicedeskpcd2 Servicedesk	TAWASOL system is a help desk system for raising and responding to tickets by client care department
		HADIR HadirDB	HADIR system for Staff attendance
4	SQLFL Server 172.20.121.129 RUH-SQL-03 MS SQL Server	Committees (MAJALES) eMajales_SCFHS_v2	MAJALES system serving Committees department
		HealthSpecialitiesDB3	Mustamir system data
5	SEBEL Server 172.20.111.201 SBL-PRD-DB-CLUSTER-SCAN ORACLE	Integration Schema INTEGPRD	Integration Database works as middleware for multiple external API integration, mainly Exam Vendors data (PROMETRIC, PEARSON)
		Mumaris SIEBEL	OLD system for Health practitioners CRM, Now replaced by Mumaris+, Old system holds some data for CME activities
6	EBS Server 172.20.111.130 172.20.111.141 172.20.111.171	Oracle Account Payables AP	**Account Payables**
		Oracle shared APPS objects APPLSYS	**shared APPS objects**
		Oracle Application Codes APPS	**Application Codes**

	Server	Database	Description
	ERP, ERPPRDDB1, ERPPRDDB2 ORACLE	Oracle Account Receivables AR	**Account Receivables**
		Oracle Bills of Material BOM	**Bills of Material**
		Oracle Cash Management CE	**Cash Management**
		Oracle Fixed Assets FA	**Fixed Assets**
		Oracle General Ledger GL	**General Ledger**
		Oracle Human Resources HR	**Human Resources**
		Oracle Inventory INV	**Inventory**
		Oracle Training Administration OTA	**Oracle Training Administration**
		Oracle Purchasing PO	**Purchasing**
		Oracle Internet Supplier Portal POS	**Internet Supplier Portal**
7	Staging Server 172.20.121.48 RUH-SSIS-02 MS SQL Server	Excel Files dbo	Staging for Excel files received from multiple parties, these files placed on OneDrive account
		One45 One45	Bridge views for local One45 database contains extracted dump received from One45 on SFTP
		Collaboration Department Collaborate	Database Tables loaded with data from Access DB for Collaboration Department, Extract done using SSIS packages
		DataFlow Dataflow	Database tables loaded with extracted data from Amazon AWS provided by data flow, Extract into tables using stored procedure, tables used as a source for DWH
		Trainees LMS lms	Bridge views for Trainees LMS database (PostgreSQL), Trainees LMS system using open-source MOODLE
		SCFHS Academy academy	Bridge views for SCFHS Academy database for employees (PostgreSQL), SCFHS Academy system using open-source MOODLE
		Mustamir mustamir	Database Tables loaded with data from Excel Files received from ELM, Extract done using SSIS packages
		Elham	Trainees learning Experience data from API

	Server	Database	Description
		Tawasol Survey	Calls survey and feedback for tawasol tickets
		Leadership	Leadership Academy is responsible conducting programs and online webinar for leaders and particular audience
8	Linux Server 172.20.100.116 SBL-PRD-DB-CLUSTER-SCAN ORACLE	Financial Compensation FCSWS	Financial Compensation system for Assessment department
		Service Oriented Architecture WSO2AS_ANALYTICS_DB	System logs for API hits for SOA services provided by SCFHS

6. Database Catalog

1. Description:

The list of physical databases that are used/managed by the organization.

2. List of Attributes

Each Database is described by a set of attributes. In the table below, we elaborate the nature and purpose of each attribute

Attribute	Attribute Definition	Purpose
Name	Name of the Database	Identifies the database
Description	Describe the purpose of the database, along with major data entities	To understand the purpose of the database
Conceptual Owner	Name of person/org unit who is responsible changes to the schema and data. It could be the application owner (the application for which this database is attached to)	Who is responsible for changes
Physical Owner	Name of person/org unit who is responsible running the database	Who is responsible for operating it.
Type	Relational, documents, graph etc.	Different types of databases serve different purposes.
Size	Size in GB or TB	To understand how big it is.

3. Database Catalog

Catalog is attached below.

DataBase Catalog V 1.0.xlsx

4. Observations

All databases listed in the catalog are relational databases. Other types of databases – NoSQL DBs, graph DBs etc. – are not used at present. We do not have information about conceptual and physical owners of the databases.

7. Data Management

Data Management refers to treating data as an asset and managing the data assets in ways that benefits the organization. Strictly speaking, data management topics are not considered part of data architecture. Actually, it is the other way round: data architecture is considered part of data management.

However, data architecture cannot be separated from data management practices, because data architecture can be successfully realized only if it is complemented by good data management practices. In any case, some areas do overlap between the two.

In the below sections we describe the relevant data management topics. The intention is to understand the current practices in SCFHS with respect to these areas and how they impact the realization of the data architecture.

1. Master Data/Reference Data Management

Master Data represents “data about the business entities that provide context for business transactions” (DAMA-PMBOK). The most commonly found categories of master data are parties (individuals and organizations, and their roles, such as customers, suppliers, employees), products, financial structures (such as ledgers and cost centers) and locational concepts. Master data is non-transactional in nature.

Reference Data is used for classification and categorization. It is subtly different from Master Data. For example, information about a customer is master data. But customer classification is reference data. Master data can be changed during normal business operations (e.g. add a new customer), but adding reference data (e.g. a new customer segment) will likely require changes in the business process.

From the point of view of architecture, we can define master data/reference data management as a capability (architecture building block) with associated processes, technologies and roles. Defining the capability in sufficient detail is beyond the scope of this document. Typically, it requires a separate initiative to implement master data management. Solution Building Blocks (i.e. software) are available to address this capability.

To represent the current capability within SCFHS, we use the following table as a high-level view.

	Policy/Standard	Identification	Process	Tool/Implementation
Master Data	No	Yes	No	No
Reference Data	No	Yes	No	No

Policy/Standard – Does policies and standards exist with respect to master data and reference data?

Identification – Is master data/reference data identified within data objects in the organization?

Process – Are there processes/roles to manage master and reference data

Tool/Implementation – Is there a tool/software that is used to manage master data and reference data? Are the polices and processes are implemented in practice?

a) Observation

From the above table, we observe that Master Data and Reference Data have been identified, but there are no known policies/standards with respect to master data and reference data. There is no process that has been defined/followed to manage master data and reference data. Thirdly, there is no tool/implementation of automated master data/reference data management processes.

2. Data Quality & Data Governance

Data Quality refers to accuracy, timeliness, and consistency of data. Data Governance ensures quality and availability of data in conformity with policies and standards in place. In the context of Data Governance, the role of Data Steward is important, apart from the Data Owner.

From data architecture perspective, Data Quality and Data Governance capability can be addressed by existing Solution Building Blocks. Such solutions typically cover a range of capabilities, including data quality, data governance, master data and reference data management, metadata management etc.

To represent the current capability within SCFHS, we use the following table as a high-level view.

	Policy/Standard	Process	Tool/Implementation
Data Quality	No	No	Stored script services
Data Governance	Yes	Yes	No

Policy/Standard – Does policies and standards exist to ensure data quality? Are data governance roles defined? Are there standards against which governance can be applied?

Process – Are there processes/roles to manage data quality? Are there data governance processes?

Tool/Implementation – Is there a tool/software that is used for data quality and/or data governance? Are the polices and processes are implemented in practice?

a) Observation

From the above table, we observe that Data Governance policies and processes have been defined/followed. There is no automated tool/implementation of Data Governance processes. On Data Quality, all aspects are missing – no policies, processes or tools exist (although there are some scripts to ensure data quality).

3. Data Exchange and Interoperability

Data exchange and interoperability refers to policies, standards and technical capabilities to handle exchange of data between people, internal systems, and external parties.

Data standards play a key role in data exchange. Data standards are documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data.

From data architecture perspective, exchange and interoperability standards are critical to realize capabilities to ensure seamless information flow across and outside the organization.

List of Standards/formats in use/supported within SCFHS

Name of Standard/Format	Type	Description	Comment on current usage (where it is used, by whom etc.)
NA	NA	NA	NA

NA – Information is Not Available at this time

a) Observation

Data interchange formats are not known at this point. This means there is no clarity how data will be exchanged between system/services. It is assumed that decisions with respect to data exchange formats are taken on ad-hoc (case to case) basis.

4. Data Storage and Retention

Data storage and retention refers to the way data are stored, and archived. Retention policies should cover how data is to be stored, what to do if stored data is lost, how long to keep data in storage and where & how to archive data.

Storage policy might depend business usage of data and data security guidelines. For example, high usage data might be stored in faster access storage. Highly secure data may be encrypted while stored. Back-up and Disaster Recovery actions are also guided by data type/classification (business criticality, security requirements etc.). Similarly, retention policy recommendations will differ based on data type/classification. For example, financial data must be stored for minimum 5 years. Note that storage requirements pertain to current usage of the data, which may include

backups. Archival deals with long term storage of data not used currently. While a policy may cover all the aspects, for clarity of understanding, we are treating storage and archival as separate topics.

From the perspective of data architecture, we can summarize the current situation with respect to data storage and retention in the following table as a high-level-view.

	Policy/Standard	Identification	Process	Implementation
Data Storage	No	No	No	Yes
Data Archival	No	No	No	Yes
Disaster Recovery	No	No	No	No

Policy/Standard – Does policies and standards to deal with data storage and retention?

Identification – Are data objects in the organization mapped to such policies/standards?

Process – Are there processes/roles which are followed to address these areas?

Implementation – Are the polices and processes are implemented in practice?

a) Observation

From the above table, we observe that in all areas – Data Storage, Data Archival, Disaster Recovery – policies, processes are lacking. In the absence of policies/standards, there is no mechanism to identify the data that need to be treated differently, so every data is essentially treated the same way in terms of storage and archival. This is potential risk of violation of regulation as there might be specific regulatory requirements for certain type of data.

5. Data Security

Data security is the process of safeguarding digital information throughout its entire life cycle to protect it from corruption, theft, or unauthorized access. In the context of Data Architecture, we want to understand security classification of data entities and how the classification impacts the handling of the data in the systems where they reside and how they are shared across different users.

Organizations classify data based on who has access to it. According to current policy at SCFHS, there are three classifications.

Sensitive data: data whose loss, misuse, unauthorized access or modification results in serious harm or any negative impact on the activities of the Saudi Commission for Health Specialties or any party related to the data or the privacy of individuals and the protection of their rights.

Protected Data: Data that is classified as (Top Secret, Confidential, Restricted) and is not authorized to be shared.

Personal Data: Every statement, whatever its source or form, that would lead to the individual being specifically identified, to identify it directly or indirectly when merging it or make it capable of or link it with other data, including:

For example, but not limited to name, personal identification number, exact address, contact number, bank account number -credit card, still or moving images of the user, fingerprint data (such as eyes or fingers), and other Personal data.

a) Observation

The above security classifications exist, but we did not find any policy/standard on how to handle different classes of data. Ideally, there should be clear definition of the security classifications, how to identify which classification is applicable in a given situation, and once classified what are the implications (how to handle each classification).

III. Overall Summary

In this section we will summarize all the observations that have been listed throughout the document. Readers who are interested in learning the key points of this document can get an overview in this section. We also suggest opportunities corresponding to each observation.

No	Artifact	Observation	Opportunity
1	Data Entity Catalog	Currently, the data entities are only partially mapped to data owners. There is no information on data stewards. Other attributes are also missing.	Capturing complete information about the data entities should be taken up next. It is possible more data entities will be discovered/documented in the next phase of EA definition
2	Data Flow Diagrams	Data Flow Diagrams were not available, except for one solution diagram of data warehouse	Major application to application data flows should be captured such that we have an idea of how major data entities are created/consumed.
3	Data Management (Master Data/Reference Data)	Master Data and Reference Data have been identified, but there are no known policies/standards with respect to master data and reference data. There is no process that has been defined/followed to manage master data and reference data. Thirdly, there is no tool/implementation of automated master data/reference data management processes.	Policies/standards for Master Data and Reference Data should be defined. Ideally, master/reference data management will be done using a tool which will come with the capability to configure the necessary management processes.
4	Data Management (Data Governance and Data Quality)	Data Governance policies and processes have been defined/followed. There is no automated tool/implementation of Data Governance processes. On Data Quality, all aspects are missing – no policies, processes or tools exist (although there are some scripts to ensure data quality).	Policies/standards for Data Quality should be defined. Ideally, data governance/data quality management will be done using a tool which will come with the capability to configure the necessary management processes. The same tool may provide data management capabilities (for example, master data management)
5	Data Management	Data interchange formats are not known at this point. This means	Data interchange formats should be mandated using

	(Exchange and Interoperability	there is no clarity how data will be exchanged between system/services. It is assumed that decisions with respect to data exchange formats are taken on ad-hoc (case to case) basis	policy/standard. With standards, data exchange will become easier to implement and with better performance.
6	Data Management (Storage and Retention)	Policies, processes are lacking. In the absence of policies/standards, there is no mechanism to identify the data that need to be treated differently, so every data is essentially treated the same way in terms of storage and archival. This is a potential risk of violation of regulation as there might be specific regulatory requirements for certain type of data	Storage and Retention policies should be defined based the nature of data handled within/outside the organization. Policies/Standards should clearly define the types of data that should be treated differently, how to identify the correct type for a given data entity, and how to treat specific types.
7	Data Management (Security)	Security classifications exist, but we did not find any policy/standard on how to handle different classifications.	.Ideally, there should be clear definition of the security classifications, how to identify which classification is applicable in a given situation, and once classified what are the implications (how to handle each classification).
8	Database Catalog	We do not have information about conceptual and physical owners of the databases.	It would be good to identify conceptual and physical owners of the databases from the point of view of data owners and data stewards (conceptual owner is likely to be data owner, physical owner is likely to be data steward)

IV. Glossary of Terms

Term	Description
APO/APR	Application Portfolio Optimization/Rationalization
ARM	Application Reference Model
BCM	Business Continuity Management
BRM	Business Reference Model
DR	Disaster Recovery
DRM	Data Reference Model
EA	Enterprise Architecture
IT	Information Technology
KPI	Key Performance Indicator
LOB	Lines of Business
RPO	Recovery Point Objective
RTO	Recovery Time Objective
SCFHS	Saudi Council for Health Specialties
TOGAF	The Open Group Architecture Forum
TRM	Technology Reference Model

V. References

No.	Name	Description
1	Enterprise Architecture Principles	Describes the EA principles for all domains of EA
2	SCFHS Enterprise Reference Models	Describes all reference models – BRM, ARM, DRM, TRM
3	Application Assessment	Business Value assessment of applications in SCFHS
4	Business Architecture Baseline	Describes the As-is Business Architecture