MVP Key Design Decisions

Last updated by | LIANG Zhao | Sep 26, 2023 at 12:45 PM CDT

Draft

Background

***Note: build, buy or hybrid decision is excluded from this KDDs and shall be discussed separately.**

The purpose of this document is to analyze the MVP key design points of Sustainability Data Foundation (SDF) so that the project team (product, architecture and data management) are aligned on what key components are and how we make the choice about them for MVP; and shall be read along with other detail design documents, such as data model design and etc.

Design Principles

- data foundation shall sit on the BP own infrastructure either via IaaS or PaaS, ideally platform agnostic
- all ingested data shall be immutable
- all data operation shall be traceable
- all data in foundation shall follow BP data retention and legal hold standards
- all BCP related design and processes are dictated by the product CIA rating
- zero trust security shall be implemented for all components in the foundation
- all compute and storage shall be independently scalable
- all APIs shall be managed by an API management system
- all layer 7 internet facing interfaces shall be protected by WAF

KDD - 1: Landing zone of Sustainability Data Foundation

Currently in BP environment there are two constructs which are commonly used for data analytics and lake:

- Azure Data Lake
- Data Hub (Azure or AWS)

There are many nuances between these two options; however, here are the key comparisons (note: cost analysis is not included and will be more based on the details usage patterns).

Also MSFT is providing a PaaS implementation of ESG data lake, aka, Project ESG Lake , which is currently in preview; considering its' complete coverage of ESG data management from ingestion, storage, data model and discovery it is also added into the comparison:

	ADL	Data Hub Azure	MSFT ESG lake
CE	Resource group	Subscription/Account	Subscription
Key 'provisionible' components	ADF, Databricks and SQL Servers	much more components in the DHA	tailor-fitted data lake solution from ingestion, storage, data model to discovery and analytics for ESG domain
Core common services	Data catalog, pipelines, data governance and etc.	self managed	PaaS
Storage	Provided by Data Engineering	Controlled by the Team who owned DHA	PaaS
Compute	Databricks	Databricks and others	PaaS
Analytics	Shared DW/SQL/metadata store	separated individual stack	PaaS
Data discovery and governance	shared components, such as catalog	individual components	PaaS
Security & access control	limited by resource group limitations	more flexible to accommodate the variety of scenarios	PaaS

In summary ADL is simpler and 'PaaS' version of DHA with the less control and flexibility but considerably cheaper and quicker setup.

For MVP both ADL with the design pattern 4 and MSFT ESG Lake should be considered to test identified use cases.

KDD - 2: What data shall be ingested into the SDF

Functionally SDF could and should provide the support for all aspects of the sustainability activities of the group whether it is the quarterly ESG reporting or P&O/refineries' operational carbon emission monitoring.

That said, the way to implement such vision need a careful consideration and design based on the decomposition of ISA-95 information architecture where there is a clear boundary between operational system and business (financial and non-financial) activities for many valid reasons, such as performance, proximity, ownership, security and etc.

The objectives of the SDF shall:

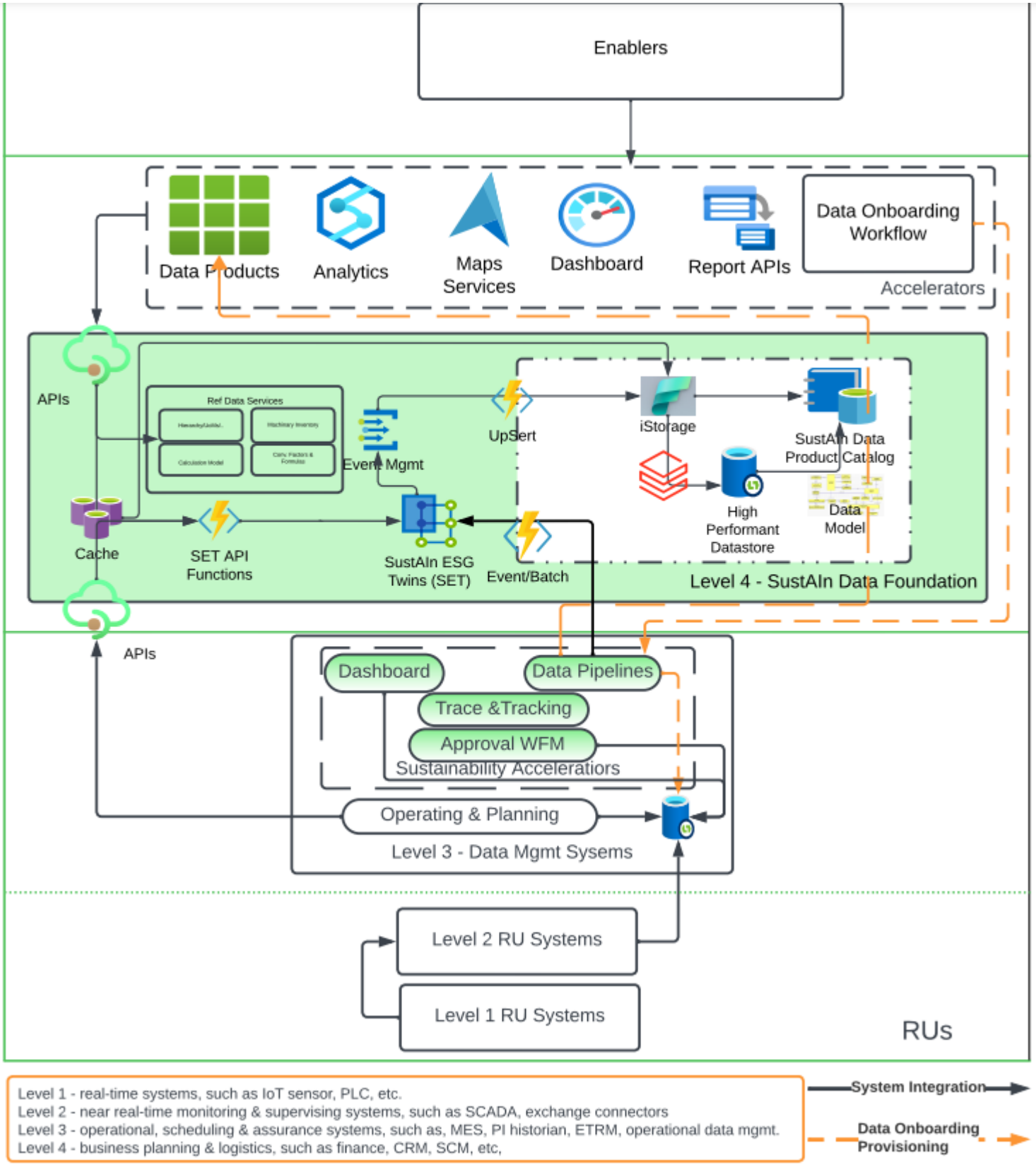
- be the single source of the truth for the corporate sustainability data at the proper information architecture level
- facilitate the accurate data sharing across functional silos

That said, the implication to the foundation implementation could be:

- any operational related data (higher frequency than monthly) shall stay at the level 3 data management systems (if there weren't any for the RUs a new ADL or similar construct shall be created to host the operational data and integrated with the existing operational, scheduling and assurance systems at RU level)
- SDF shall gather and publish such data as the data product for any potential consumption outside the owner RUs
- SDF will provide a pipeline to support such consumption pattern with either or both batch/streaming accordingly (and a 'caching' concept could be implemented for the better performance)

(This design is one of several options we could consider; as the draft nature of this document we shall consider and discuss other options the team might have.)

The high-level solutions architecture of this implementation is provided below:



What reference data should SDF be holding

There are two types of reference data within the scope of sustainability:

- records belongs to the sustainability
  - emission factors, GWPs, oxidation factors, reporting UoMs and etc.
  - BP sustainability methodologies, calculation models/formulas

- other enterprise data, such as org hierarchies, business nomenclatures, equipment inventory and etc.

SDF shall be the system of records for the sustainability data types; recommend and standardize another reference data via data quality check and transformation accelerators.

KDD - 3: How should data be stored in SDF

Conforming with the best practices two storage constructs shall be in SDF:

- source immutable dataset in delta/parquet with a scalable partition storage
- enriched high-performance data based on the industry standard or open model(s)