Data Ocean Foundation Scope

Created by: Fernando Girante

Some references:

https://solvayagile.atlassian.net/browse/DAAI-294

Initiative Brief and the One page summary

As a general scope:

- The data ocean is introduced as a layer between source systems and target applications.
- All target applications source their data from the data ocean, reducing interfaces and centralizing data modeling and business logic.
- The data ocean is built in line with Solvay data platform target architecture guidelines.
- The data ocean can be built incrementally in “slices” - prioritizing where to start for MVP and how to phase is an ongoing discussion.

From the Brief

- What/Scope

Going to a data architecture that is in line with the Solvay Data Ocean Strategy.

- By providing a single model, storage and processing costs can be lowered by eliminating duplicate and redundant data, and preprocessing data up-front, resulting in a more compact and effective dataset and providing a single source of truth for analysis and decision-making
- Having a single source of truth will improve the consistency, accuracy, and integration of the data used for Predictive Modeling and in the dashboards
- Improve security and increment the Data Governance

- Why should we do this

- It will improve the quality of the data available for analysis and decision-making and will provide a consistent view of the data across different data products, simplifying the identification and correction of any inconsistencies or inaccuracies in the data, as well as ensuring compliance with data governance policies.
- It should improve the efficiency, cost-effectiveness, and accuracy of the data products by reducing storage requirements.

What are the new capabilities expected?

- There are currently a number of capabilities that need to be improved, such as the logging process, processing metadata, and data lineage.
- A Curation Engine, with a focus on Data Validation and Verification; Data Profiling; Data Standardization and Transformation; and Integration will also be included, as well as an improved Security mechanism that can be used with Reporting and Dashboarding tools.
- A Data Orchestration, Scheduling, and Monitoring, will also be evaluated.
  All of these extra features will contribute to better Data Quality procedures.

What are the deliverables planned:

- Platform Core Setup
  - Prepare & Define Setup
  - Environment Instantiation
  - Setup Schemas
  - Orchestration Definition & instantiation
  - Logging - Template Instantiation
- Domain Setup (per Domain)
  - Prepare & Define Setup
  - Create Environments
  - Environment Instantiation
- A wiki page describing the Data Ocean and exposing a collection of documents
  - Closed
    - Data Architect - Data Quality KPI_V1.0.docx
    - Data Architect - Error Log System_V1.0.docx
    - Data Architect - Mandatory Fields_V1.0.docx
    - Data Architect - Data Hub_V1.0.docx
    - Data Architect - Naming Convention_V1.0.docx
  - In progress
    - Data Architect - Scheduler _ jobs runs.docx
    - Data Architect - SCD.docx
    - Data Architect - Security System.docx
    - Data Ocean - Data ingestion reference Jobs .docx
  - Out of scope (but tentative, if not possible - to be done on Run)
    - Data Architect - Data Masking_V1.0.docx

- Data Architect - Data Curation_.docx
- Data Architect - Storage Management.docx