

Architecture Vision

Data Platform project  
Vår Energi 2020

Table of Contents

[1 Problem Description 4](#_Toc44508709)

[2 Detailed Objectives 7](#_Toc44508710)

[3 Actors and their Roles and Responsibilities 8](#_Toc44508711)

[4 Resulting Architecture Model 9](#_Toc44508712)

[5 End Vision Statement 20](#_Toc44508713)

Document Information

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name:** | Data Platform | | |
| **Prepared By:** | Renate Endal, Christine Toft Stenersen, Nedyalka Delistoyanova, Morten Taksrud | **Document Version No:** | 1 |
| **Title:** | Architecture Vision | **Document Version Date:** | 01.07.20 |
| **Reviewed By:** |  | **Review Date:** |  |

Distribution List

| **From** | **Date** | **Phone/Fax/Email** |
| --- | --- | --- |
|  |  |  |
|  |  |  |

| **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** |
| --- | --- | --- | --- | --- |
| 1.0 | 01.07.20 |  |  |  |
| 1.1 | 14.10.20 |  | Added link to Roles and Responsibilities deliverable in chapter 3 |  |

# Problem Description

## Stakeholders and their Concerns

**Stakeholder Overview**

|  |  |  |
| --- | --- | --- |
| **Role** | **Concerns about the business** | **Concern about the Data Platform** |
| Digitalization Manager | Responsible for the Digital Portfolio and concerned about implementing digital initiatives in the organization | Data Platform is an important enabler for several digital initiatives and the digital portfolio. Thus, the digitalization manager is concerned with the progress of the project as it impacts the progress of the digital program |
| ICT Manager | Concerned about making sure that the ICT department delivers on the business unit’s needs. | Project owner for the Data platform implementation project. Concerned about implementation progression, budget, and the quality of the delivery. |
| Architecture Board |  | How the Data Platform Architecture affects and aligns with the existing architecture and the organization’s needs. |
| Infrastructure Team | Ensuring that the infrastructure support the business needs/requirements | Concerned about security and efficiency of the infrastructure in the data platform and how it integrates with the rest of the IT architecture in VE. Additionally, they are concerned about the management of the infrastructure. |
| IM Team | Enable the business to use data in an efficient manner | The Data Platform is intended to store and manage a lot of VEs data. IM team is concerned about data lifecycle management in the data platform. |
| Digitalization Team | Enable digital transformation in the organization | Data Platform important enabler for several digital initiatives and the digital portfolio, and the digitalization manager is thus concerned for the progress of the project as it impacts the progress of the digital program |
| AM Team | Ensuring that the applications support the business needs/requirements | The AM team is responsible for managing the applications in Vår Energi. They are concerned about making sure that the applications in the data platform are managed in a secure and efficient way. |

## Issues/Scenarios to be Addressed

Vår Energi is one of the largest E&P companies on the NCS and have the ambition to grow. In 2019, Vår Energi established a Digital Transformation Program, as it was expected from the owners, the competitors worked on digitalization, and employees were eager for new ways of working. When defining a digital vision and roadmap, various digital initiatives were defined to make sure Vår Energi reach their digital vision of “turning bytes to barrels”.

Vår Energis **current IT landscape lacks a common way of accessing, sharing, and collating information across units and systems**. This makes it difficult to do analyses, make reports, build dashboards that includes information from several different IT systems and business units. In addition, it is nearly impossible to collate information from the multitude of systems as there is no common way of modeling the information found in these systems. A large amount of manual work is required when collating and consolidating information, making the processes inefficient and time-consuming.

## Business Vision Statement

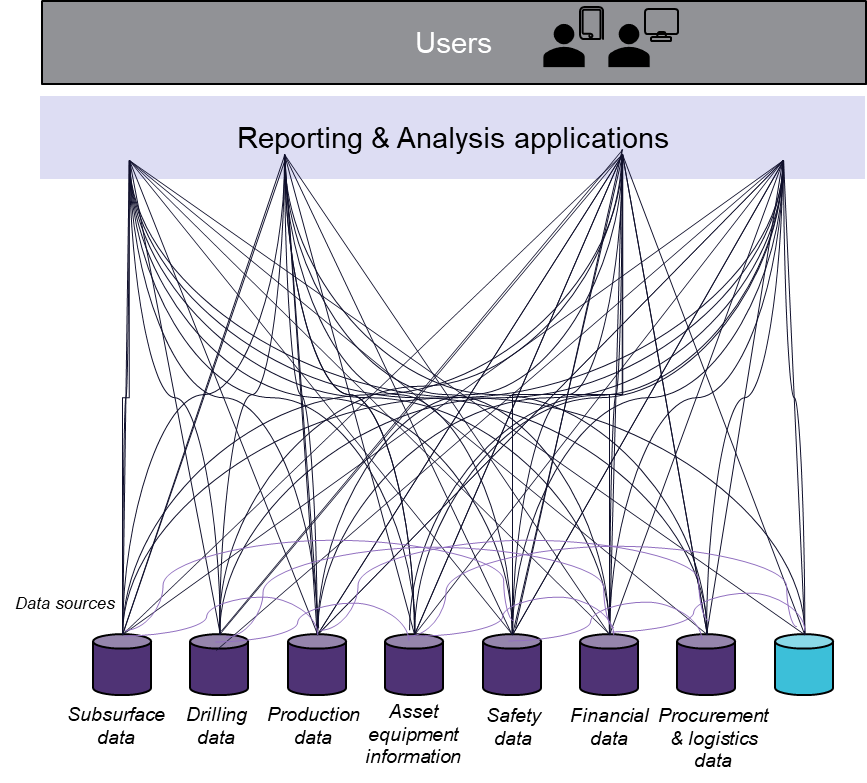
Our vision is to build a data platform that will be an essential enabler for Vår Energi’s digital strategy and roadmap and that will increase efficiency in ICT.

The current architecture makes it difficult to find and use information which leads to inefficiencies. A common data platform will be able to ingest the information found in the different IT systems across the different business units and prepare it for use. Creating a data layer will facilitate coherent data retrieval and storage, enabling a single-source-of-truth while creating a serving layer will facilitate data use and enable scalability of digital projects. It will thus be an essential enabler for many of Vår Energi’s planned digitalization initiatives and should extend the capabilities of the current underlying technology platform, information availability and governance to support Vår Energi’s digitalization roadmap and strategy. Prioritized initiatives that have the Data Platform as a prerequisite for success is the Energy Management System, Production Dashboarding, Mobility in Field Operations, and Supply Chain Control Tower. Additionally, there are several initiatives in the pipeline, such as Monitoring of Heat Exchangers, Mechanical Integrity Management, emulsion management, integrity chemicals.

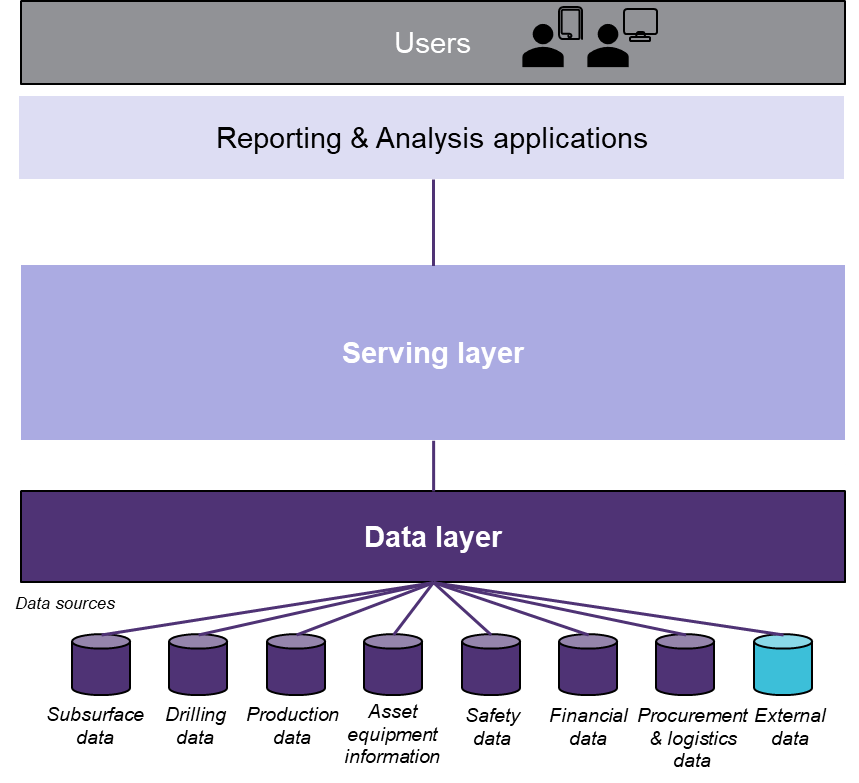
In addition to enabling digital initiatives, the data platform also enables ICT to become more efficient. ICT currently manages an IT landscape without any integration platform, which makes it difficult to ensure data flows between different systems and applications. The data platform will provide a mechanism that can make ICT more efficient and effective in their activities, by easier configuration and management of data flows.

## Business Vision Diagram

**Before**

The current architecture makes it difficult to find and use information which leads to inefficiencies

**After**

**Creating a data layer will facilitate coherent data retrieval and storage, enabling a single-source-of-truth while creating a serving layer will facilitate data use and enable scalability of digital projects.

## 

# Detailed Objectives

To resolve the identified business problem, a set of objectives is set for the architecture. Below is a list of defined objectives for the Data Platform.

**Business Objectives**

* Flexibility to start small and scale as needed and utilize services and software that can scale without license-changes according to usage
* Integration towards 3rd party vendors to get the best solution to support both current and future business needs while retaining independence and minimize vendor lock-in
* Enable the use of many different end-user tools and applications to best fit end-users’ purpose and thus increase efficiency
* Establish one single-source-of-truth
* Make it easier to find data through data and system documentation
* Increased data sharing, reuse and delivery across the business functions which will enable more efficient business operations by breaking organizational silos

**Technology Objectives**

* Replace legacy IT system with one common data platform
* An integration platform that will ensure that data flows between different systems and applications
* Enable common management of Master Data
* Make it easier to collate and consolidate information across business units
* One common way of modelling, sharing and accessing information
* Establish possibility to store historical data and for restructuring data
* Enable possibility to use advanced analytics and ML / AI
* Support integration with a wide range of data sources
* The Platform should be able to manage both structured and unstructured data, in addition to manage both secure and unsecure data

# Actors and their Roles and Responsibilities

To set up Guidelines for Role and Responsibilities is a part of the Governance Stream in the project and will be provided in the end of the project period.

The Roles and Responsibilities document can be found [here](https://varin.varenergi.no/ict_im/0091/_layouts/15/DocIdRedir.aspx?ID=8036-1144636671-16).

# Resulting Architecture Model

## Constraints

When defining the architecture, there are some constraints that need to be taken into consideration. These are 3rd party constraints, technical constraints, and business constraints.

**3rd party constraints**

The Data Platform MVP solution will use several industry products in its implementation:

* Azure (incl. Azure Data Factory and Azure storage)
* Snowflake Cloud data platform

For the whole data platform, the following products have been recommended.

* Informatica Data Cloud
* MuleSoft Anypoint API Manager
* Neo4j

They will not be part of the MVP and there needs to be taken a final decision on whether Vår Energi will implement them in the future.

**Technical constraints**

* Data Platform solution is cloud-based, hosted in Azure, and as such it should be possible to establish a secure connection between cloud and on-premise.
* For the data platform to be able to extract data from the data source system, the source system should be able to provide access to their data in a way of APIs, database access, exported files.
* Azure compatibility – every 3rd party product needs to be Azure-based or able to connect to Azure for it to be integrated in the data platform.
* Data Quality & Profiling drives the business value of the data platform, low quality data leads to wrong recommendations and decisions.
* Authentication and authorization to be achieved through Azure AD and RBAC.

**Business constraints**

* The MVP should deliver a working version of the Data Platform, even though not all components of the reference architecture will be present.
* The MVP scope does not cover the whole reference architecture and some components will be implemented in later stages.

## Principles

When defining the architecture, a set of principles were taken into consideration, hereunder business principles, architecture design principles, technology principles, and data principles.

**Business Principles**

* Primacy of Principles: These principles of information management apply to all organizations within the enterprise.
* Maximize Benefit to the Enterprise: Information management decisions are made to provide maximum benefit to the enterprise as a whole.
* Compliance with Law: Enterprise information management processes comply with all relevant laws, policies, and regulations.
* Protection of Intellectual Property: The enterprise's Intellectual Property (IP) must be protected. This protection must be reflected in the IT architecture, implementation, and governance processes.

**Architecture Design Principles**

* Ensure functional scalability: The architecture must ensure “*functional scalability*” to be able to accommodate future use cases for the data platform as these are not fully known at this point.
* Emphasize agility: Ensure flexibility and minimize vendor lock-in when selecting products to realize the components in the data platform.
* Minimize the number of Vendors: Selecting products from as few vendors as possible to realize the components in the data platform.
* API-first approach: Recognize your APIs as first-class artifacts for agility and reuse.

**Technology Principles**

* Requirements-based Change: Only in response to business needs are changes to applications and technology made.
* Responsive Change Management: Changes to the enterprise information environment are implemented in a timely manner.
* Control Technical Diversity: Technological diversity is controlled to minimize the non-trivial cost of maintaining expertise in and connectivity between multiple processing environments.
* Interoperability: Software and hardware should conform to defined standards that promote interoperability for data, applications, and technology.
* Technology Independence: Applications are independent of specific technology choices and therefore can operate on a variety of technology platforms.
* Ease-of-Use: Applications are easy to use. The underlying technology is invisible to users, so they can concentrate on tasks at hand.

**Data Principles**

* Data is an enterprise asset: Data is valued as an enterprise asset, leveraged across the data value chain to enhance competitive advantage and accelerate decision making.
* Data is shared and accessible Users have access to the data necessary to perform their duties; therefore, data is shared across enterprise functions and organizations. Data is accessible for users to perform their functions.
* Data trustee: Each data element has a trustee accountable for data quality.
* Common Vocabulary and Data Definitions: Data is defined consistently throughout the enterprise, and the definitions are understandable and available to all users.
* Data security: Data is protected from unauthorized use and disclosure. In addition to the traditional aspects of national security classification, this includes, but is not limited to, protection of pre-decisional, sensitive, source selection-sensitive, and proprietary information.

## Architecture Supporting the Process

**Overview**

The architecture will be based on utilizing **cloud computing** and selecting proven products from recognized vendors. We recommend cloud computing as it proves great flexibility to start small and scale up the Data Platform when needed. Furthermore, it is recommended to utilize software products from **vendors that have an open integration strategy** as is will provide better stability, usability, and tailored support. Lastly, it is recommended to use a **best-of-breed approach** when selecting products to realize the components in the Data Platform.

To set up the Data Platform, we will set up a **data layer** from the source systems, which manages the retrieval and storage of raw data from the source systems and makes it available for the serving layer. It is also important for delivering the future for ICT by enabling tighter data integration.

Further, we’ll set up the **serving layer**, which manages the services and frameworks needed to provide curated data. This layer, together with the data layer, will enable easier management, sharing, and reuse of data across business processes and solutions. Thus, it will facilitate data usage and enable scalability of digital projects.

Furthermore, we can **pick and choose relevant products and services for specific use cases**. Since no Data Platform Vendor in the market offers support for the entire oil and gas value chain, we proposed to enable both custom-made solutions and integration with 3rd party vendors. This will enable Vår Energi to utilize the expertise in the market to **meet current and future needs**.

**Data Platform Architecture**

Based on the Capgemini reference architecture for data platforms, a subset of platform components needed to cover the currently known business needs in Vår Energi was selected within each layer of the data platform architecture.

It is important to note that additional components can be added to the data platform at a later stage when and if they become necessary to cover a given use case not currently identified.

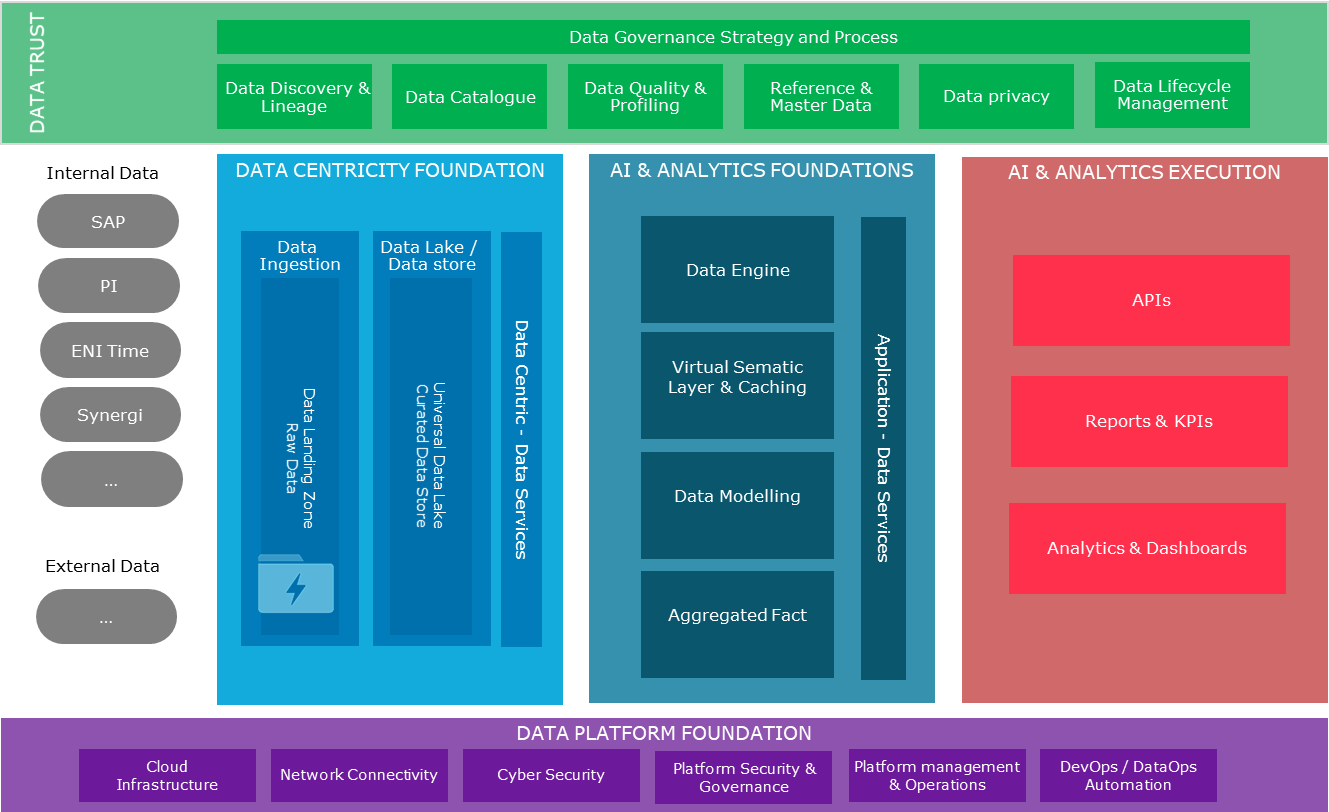


Figure 1 Data Platform reference architecture

**Data Trust**

This is where data quality, data governance and is enforced and maintained. As it concerns all stages of the data platform, it spans across all layers.

**Data Centricity Foundation**

While data scientists and certain analysts need access to raw or curated data, most users need access to data that has been prepared and transformed for analytical purposes. This is also the foundation for AI services.

**AI & Analytics Foundations**

Consumption of analytical data and AI processes happens in this layer. Consumers can be dashboards, 3rd party applications, analytical tools.

**AI & Analytics Execution**

A baseline for every data platform is security, orchestration, monitoring, and administration.

**Data Platform Foundation**

The Data Platform Foundation provides services addressing cybersecurity, network, and middleware, furthermore, DataOps, security, governance, and the operating model. Data Platform Foundation ensures that data can be securely ingested from a hybrid cloud architecture, into a secure, industrialized, optimized, automated platform with DevOps tooling.

**For the current project – the scope is as follows:**

The highlighted components in the architecture drawing are included in the MVP scope:

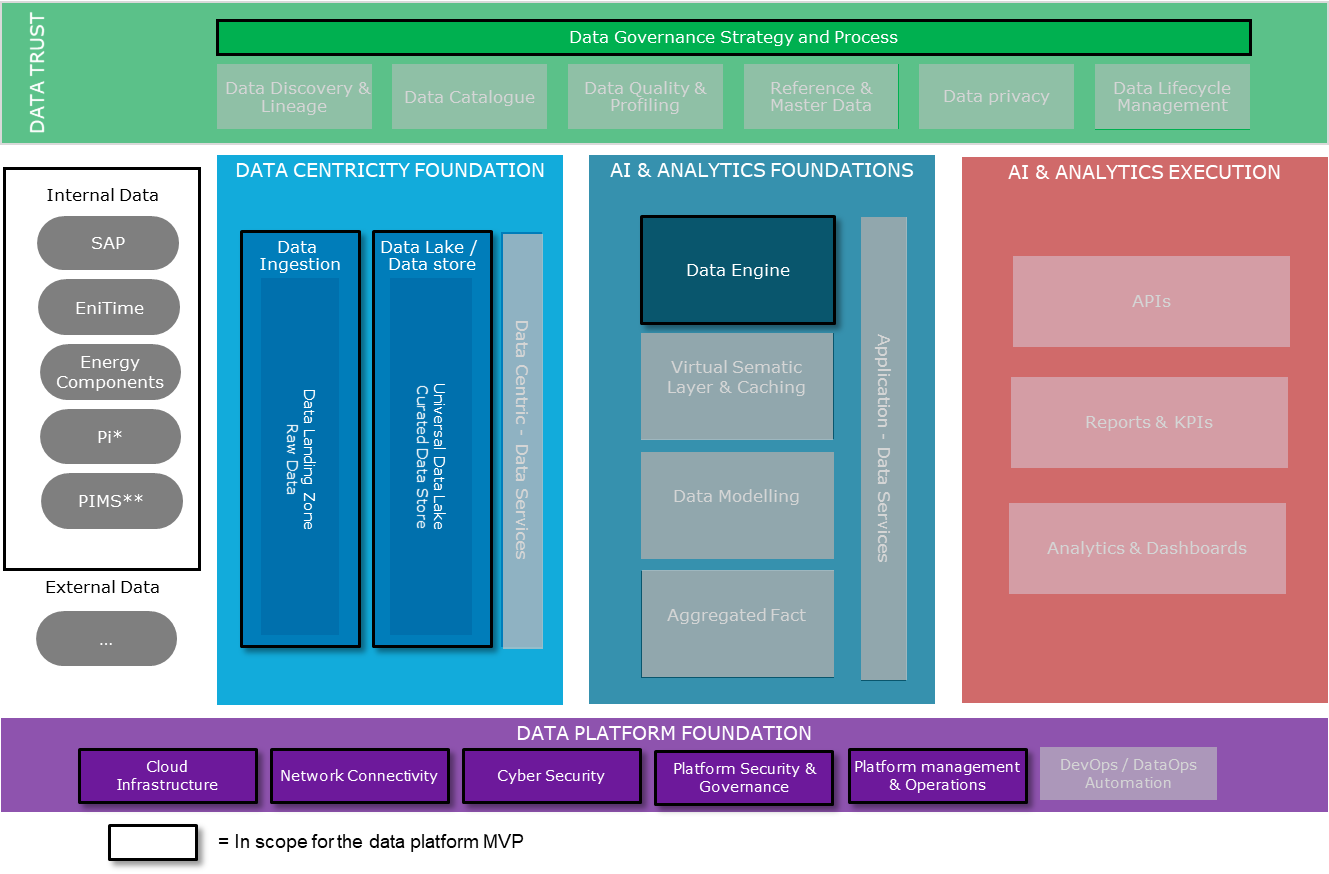


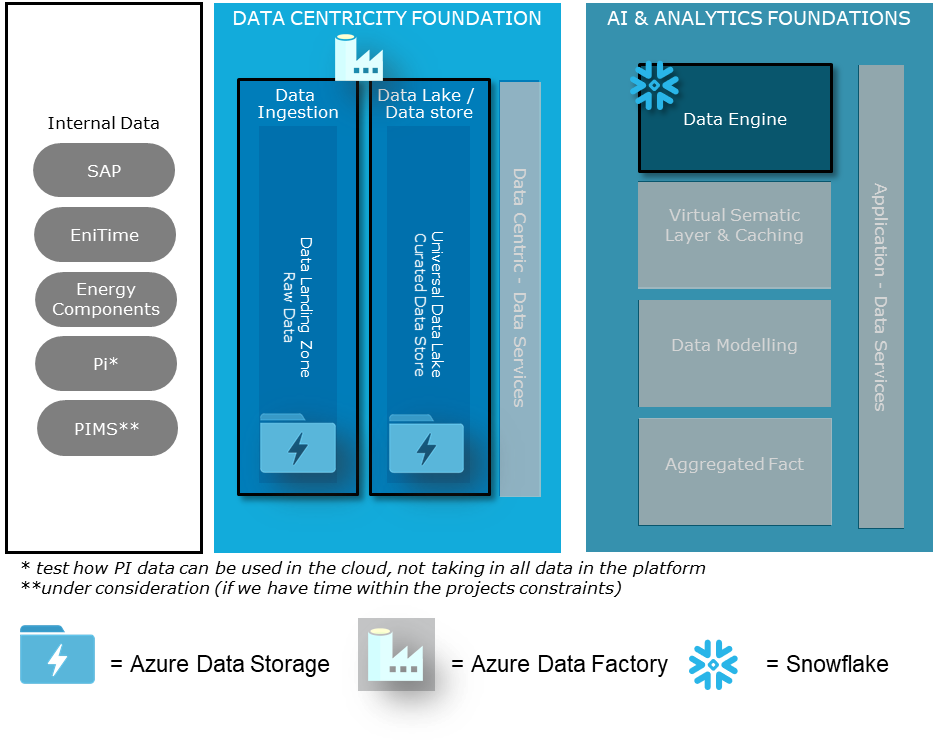
Figure 2 Scope for the Data platform MVP

**Scope for Data Platform Foundation**

* Set up an **Azure subscription** for Vår Energi
* Set up the required **Azure infrastructure** to fulfill the data platforms requirements, but also taking into account that the Azure platform will be used by other projects/initiatives/workloads in the future and prepare for that
* Establish **a minimum of governance and security measurements** such as create a naming standard and a minimum of policies to be enforced in the platform
* Do a **cyber security evaluation of the Azure platform** and get it approved. This will apply for the whole platform so that it won’t be needed to do it again for future initiatives that would like to use the platform
* **Ensure platform management** of the Azure platform. Atea will be responsible for this and the project covers the costs throughout 2020

**Scope for Data Sources, Data Centricity Foundation and AI & Analytics Foundation**

* Establish **connections between the data platform and the selected data** **sources** using Azure Data Factory (ADF) and integration runtimes (on-prem agents for ADF)
* **Build ADF pipelines** for copying data from sources to raw data lake (Scheduled batch jobs)
* We will test **how to copy PI data to cloud** 
  + PIMS will be included if we have time
* Build **transformation pipelines in Data Factory for SAP** data and load data into curated data lake
* Set up and **configure Snowflake**
* Create ADF pipelines to **copy data into Snowflake**

****

**Scope for Data Trust**

We will focus on Governance Strategy and Processes within Data Trust and have deliverables within Data Lifecycle Management and Data Platform Governance

***Data Lifecycle Management deliverables:***

* **Naming conventions:** A holistic standard for labeling data objects that helps to ensure a “common language” for data objects in the organization, which helps to keep work organized and understandable
* **Data Quality:** Establish data quality parameters from the Data Platform perspective and create rules for validation and consistency check of data coming from the source systems.
* **Structure and organization of data:** As part of the Data Platform Model we will look at how the data should be structured and organized and create rules for storage organization to increase usability and simplify security
* **Data lifecycle management:** A set of policies for managing the flow of data throughout its lifecycle and defines the steps from collection and initial storage to archiving and purging

***Data Platform Governance deliverables:***

* **Development and technical guidelines:** The guidelines apply to software developers, data engineers and architect, and represent a set of rules / standards for implementation of new features and modification of existing one
* **Maintenance and support processes:** Procedures for maintenance and support of the Data Platform aligned with the existing processes in Vår Energi
* **Request/approval/implementation for new reports and data sources:** Guidelines for request, approval and implementation process for new data sources, which will establish common routines for extension and ensure that the Data Platform is not overloaded
* **Define roles and responsibilities:** Adapt the DP to Vår Energi IT operations with defined roles and responsibilities, and define the organizational structure to support the Data Platform, incl. service and maintenance

## Requirements Mapped to Architecture

**Business Requirements**

* Make data available across different data sources and systems
  + The data platform MVP will make available data from SAP (exports), EniTime (exports) and Energy Components. It will also test how PI data can be made available, and if time we’ll also include data from PIMS
* Ensure data quality to provide correct analysis and decision-making

**Technology Requirements**

The Data Platform should be able to provide unified access to Vår Energi’s data, both raw data and analytics. While data scientists and certain analysts need access to raw or curated data, most users need access to data that has been prepared and transformed for analytical purposes. This leads to certain requirements for the data storage, the data platform should store both raw – untreated – data as well as data that has been curated with data quality and defined business rules.

**Key high-level requirements**

* Support management of Master Data
* Support integration with a wide range of data sources
* Manage both structured and unstructured data
* Manage both secure and unsecure data
* Provide capabilities for advanced analytics and ML/AI
* Support both batch data ingestion and real-time data ingestion
* Support integration towards 3rd party vendors
* Support many different end-user tools and applications
* Give flexibility to scale storage and compute according to current usage
* Utilize services and software that can scale without license-changes according to usage
* Keep infrastructure maintenance at a minimum
* Provide secure way of accessing data with Azure AD and RBAC
* Enable fast retrieval of data
* Support creating of common data model for Vår Energi

The high-level requirements and Capgemini reference architecture for data platforms served as the starting point identifying the components needed for the Vår Energi data platform. Each of the 5 different levels, Data Trust, Data Centricity Foundation, AI & Analytics Foundations, AI & Analytics Execution, and Data Platform Foundation solves known requirements for the platform.

**Data Trust**

|  |  |
| --- | --- |
| Data Governance Strategy and Process | **WHAT:** Defines & enforces the process, policies, and ownership of managing data as enterprise assets.  **WHY:** Enables control of the data platform. |
| Data Catalog | **WHAT:** Data on Data – this is the heart of unified data management and stores the “metadata” or data about the data and is critical to understand what data attributes are held and what their characteristics are. This activity is required through all the layers of the architecture. We can also talk of data catalog maintaining the Business Glossary.  **WHY:** Data Catalog will drive the adoption and usage of the Data (and eventually AI) Platform. The Catalog will in combination with Discovery & Lineage drive governance, compliance and value of data stores. |
| Data Discovery & Lineage | **WHAT:** Process of discovery of data attributes in the various sources across the enterprise and the lineage & usage of the data attributes. This process can be automated increasing the quality and trust of the information recorded. This activity is required through all the layers of the architecture.  **WHY**: Shows the flow of data between processes and systems, and who is accountable for the data. |
| Data Quality & Profiling | **WHAT:** In data quality we look to first profiling the existing data and understanding the quality. After that the data can be cleansed applying the hygiene rules by both business & IT. This process will remove duplicates, standardize and enrich the datasets.  **WHY:** Data Quality & Profiling drives business value of data platform, low quality data leads to wrong recommendations and decisions. |
| Reference &  Master Data | **WHAT:** Here we look to mastering of customer, employee, material or product data and ensuring a single view of the same. In addition, we also look at maintaining reference data tables as required.  **WHY:** Enables delivery of clean, consistent and comprehensive master data. |
| Data privacy | **WHAT:** Set of policies, processes such as Data Lineage and tools such as Data Catalog to enable compliance.  **WHY:** Given the myriad data privacy regulations worldwide, it is key that any personal data help by an organization is protect and permissioned for only those processes as agreed with the individuals. |
| Data Lifecycle Management | **WHAT:** Over time data is accumulated and needs to be managed to ensure optimum use of resources. This could include archiving of older data and also deletion of data that need not be held any longer by the organization.  **WHY:** Data Lifecycle management balances need for data and storage costs. Data Lifecycle management drives healthy data lakes, balancing need for data and too much/unwieldy data. |

**Data Centricity Foundation**

|  |  |
| --- | --- |
| Data Ingestion | **WHAT**: Ingestion design patterns & tooling to support batch, streaming data ingestion and raw data landing zone.  **WHY**: Enables data movement into the data platform. |
| Data Lake | **WHAT**: A data lake, Business hub, DW or EDW are architectural constructs that support different business demand and value propositions. This layer of the architecture could be delivered using a combination of transformation tools e.g. from traditional SQL to Open Source Algorithmics data Wrangling tools and data storage options e.g. Hadoop, RDMS, In memory, Spatial, Graph.  **WHY**: The output of this layer is the transformation and modelling of data required to development of the business product data stores and/or business products data exploitation. Creating a data centric model and architecture which becomes the constant trusted data asset of the organization, enabling and powering agility and the digital transformations. |
| Data Services | **WHAT**: Provides an ability to expose data or insights through APIs, connectors, data virtualization etc.  **WHY**: This enables a common, standard data access interface across business and analytics applications Leveraged for Search, Analytics Sandbox, Self Service Portal, BI Reporting Applications. At this layer in the architecture the data centric business model is accessed. |

**AI & AI Analytics Foundation**

|  |  |
| --- | --- |
| Data Engine | **WHAT:** Integration off Compute engines, Synapse, Snowflake, Graph engines.  **WHY:** Enables serving of persisted data to end users. |
| Virtual Sematic Layer & Caching | **WHAT:** Tooling, caching and virtual data models to interface and support performative and cost efficient traditional BI tools and business products.  **WHY:** Simplifies self-service analytics. |
| Data Modelling | **WHAT:** Design and deploy platform capabilities for data preparation & transformation.  **WHY:** Supports end users with ready-made data models. |
| Aggregated Fact | **WHAT:** Summarizing data models to support performance and cost requirements.  **WHY:** Supports end users with pre-aggregated data for consumption. |
| Application - Data Services | **WHAT:** Provides an ability to expose data or insights through APIs, connectors, data virtualization etc.  **WHY:** This enables a common, standard data access interface across business and analytics applications. Leveraged for Search, Analytics Sandbox, Self Service Portal, BI Reporting Applications. At this layer in the architecture the application business model is accessed. |

**AI & Analytics Execution**

|  |  |
| --- | --- |
| API | **WHAT**: Provides an ability to expose data or insights through APIs.  **WHY**: Gives applications a machine-readable interface for data consumption and data exchange. |
| Reports & KPIs | **WHAT**: The publishing of reports and dashboards to non-analytical users.  **WHY**: Drives business decision-making. |
| Analytics & Dashboards | **WHAT**: Building graphical reports by connecting and joining across multiple data repositories.  **WHY**: Drives business decision-making. |

**Data Platform Foundation**

|  |  |
| --- | --- |
| Cloud  Infrastructure | Provisioning of Cloud Environment and integration with existing services on-premises and/or as part of a hybrid cloud strategy. Commercial agreement with Cloud provider, Disaster recovery, SLA’s, pricing etc. |
| Network Connectivity | Middleware to support connectivity between cloud environments and on-premises. To enable cloud PaaS solutions to act as an extension to on premise data center estate, dedicated bandwidth can be a key enabler. |
| Cyber Security | Vnet and Subnet, Network Security groups, Endpoint production security, Network hardening (both ingress and egress) firewall configuration, integration into the single sign-on security (commonly active directory) and pen testing to ensure the cyber security of the cloud environment. |
| Platform Security & Governance | Account Management – Hierarchies, Account, Subscription & Resource Group, Billing and charge backs. Security - Resource Management Policies, Security Centre, Automation, Resource Locks, RBAC, Resource Groups, Tagging, Monitoring & Usage |
| Platform management & Operations | Align governance touch points with external governance mechanisms, streamlining and automating extensively. Integrate with front-line support, incident management and tracking tools and service request management to offer a seamless integration to existing IT support functions. Provide proactive, usage-based cost monitoring to minimize risk of compute cost over-runs. |
| DevOps / DataOps Automation | Continuous Integration / Continuous Delivery, Infrastructure as Code (Template based infra provisioning & automation), Test Automation, Platform (Ops) Monitoring, Cost Governance - Cost monitoring and alerts. |

# End Vision Statement

See Business Vision Statement in Chapter 2.3