Guidelines on Data Management for Low-enthalpy Geothermal Plants

Version 14 February 2025

nieuwe warmte nu

Guidelines on Data Management

Nieuwe Warmte Nu!

Maximize safety & efficiency

Authors:		**Reviewers:**	
Company:	Helin Data B.V.	**Company:**	TNO
From:	Nils van der Stad, Balint Takacs	**From:**	Ryvo Octaviano, Pejman Shoeibi Omrani
E-mail:	nils.vanderstad@helindata.com balint.takacs@helindata.com	**E-mail:**	ryvo.octaviano@tno.nl pejman.shoeibiomrani@tno.n
Project coordinator	Laura Precupanu, laura.precupanu@tno.nl		

Subject: Guidelines on Data Management for Low-enthalpy Geothermal Plants: Digital Twin Implementation

Date: 14 February 2025

This project is executed with funding of the National Growth Fund of the Ministry of Economic Affairs, implemented by the Netherlands Enterprise Agency. The specific funding for project *Data gedreven optimalisatie van aardwarmte systemen* (NWNI10005) concerns the funding rules 2022.

Table of Contents

1 Introduction

The integration of Digital Twin technology into a low-enthalpy geothermal plant represents a significant advancement in real-time monitoring, predictive maintenance, and overall operational efficiency. The focus of the Nieuwe Warmte Nu! (NWN) project is to develop and implement a Digital Twin for deep geothermal plants (Doublets). This will enable monitoring and simulation of plant operations, allowing for performance optimization, optimized maintenance planning, and risk mitigation.

This document summarizes the best practices for data management and integration with digital twin technologies tailored to geothermal plants and operational practices. This document was compiled by Helin Data B.V. (Helin), one of the key project partners, which is providing its expertise on creating a robust, scalable, and secure IT infrastructure to handle large volumes of data generated by geothermal facilities an reviewed by TNO which is the main partner in developing and deploying the digital twin solution. The key outcomes of the guidelines summarized in this document are to ensure compliance with industry standards, facilitate smooth project delivery, and provide recommendations on data management that can be adopted by the geothermal energy sector at large.

The rest of the report reads as follows. Section 2 gives an overview of the physical, operational infrastructure of a low-enthalpy geothermal plant. It explains what processes are typically measured at which stages of the plant's physical production workflow. Section 2 also describes the infrastructure from a data perspective and highlights the difficulties of making data -- produced by legacy industrial systems -- available in the cloud. Section 3 identifies the data necessary for harnessing the value of a digital twin. Additionally, to this data inventory, we introduce a generalized data model that we advocate should be widely adopted in the geothermal industry to achieve scalability and knowledge sharing. Section 4 provides guidance on data management and recommends tools and best practices for the back-end of the of the digital twin. Section 5 summarizes the main highlights and key messages regarding the required data management and IT infrastructures.

2 Overview of an industrial geothermal plant

2.1 Operational infrastructure

Geothermal energy is efficiently harnessed in doublet systems, which extract heat from subsurface reservoirs. In these systems, geothermal fluid (water) is drawn from production wells, passes through pumps, filters, and heat exchangers, and is then reinjected into the reservoir via injector wells. An example configuration is shown in Figure 1.

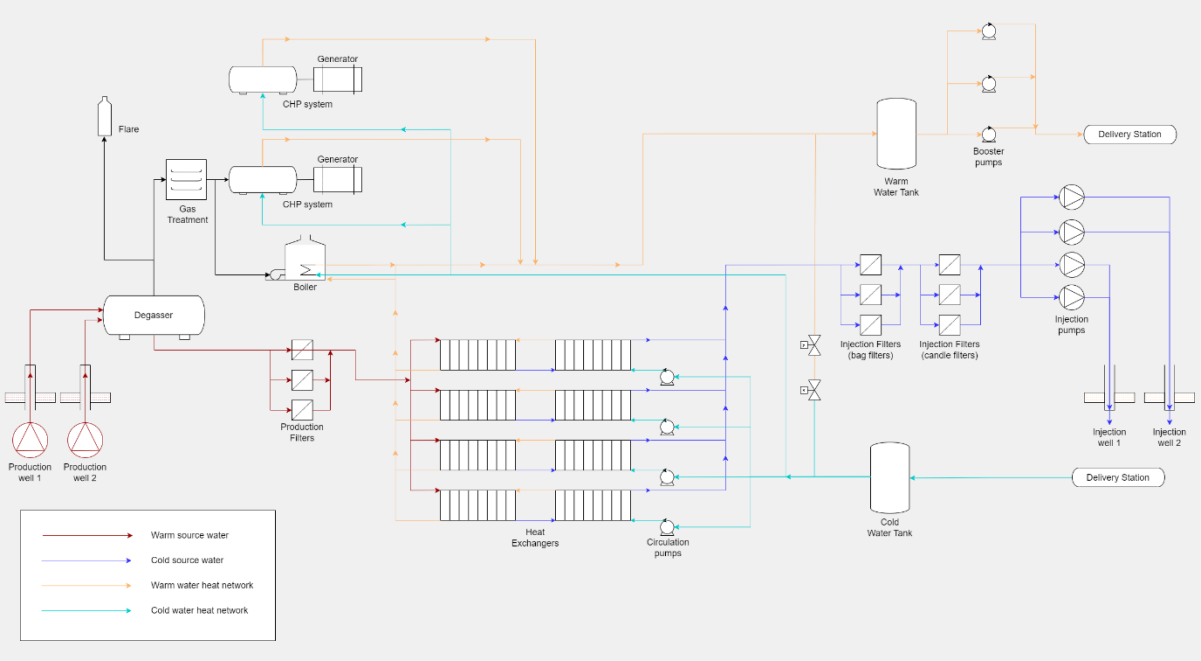
Electric Submersible Pumps (ESPs) are used in production wells, especially when reservoir depths exceed the practical limits of Line Shaft Pumps (LSPs) which is often the case for deeper reservoirs, such as those in Europe, where pump setting depths can exceed 500 meters. ESPs, with both motor and pump stages located deep within the well, generate the necessary flow rates and maintain the pressure of the extracted water to keep it in a liquid state, preventing cavitation and meeting the plant's stable pressure and flowrate requirements for optimal heat extraction.

In the schematic (Figure 1), hot geothermal water is pumped from two production wells using ESPs (red lines), ensuring a stable dynamic head above the downhole pumps. After extraction, the water passes through a degasser to remove dissolved gases, preventing gas breakout elsewhere in the system, and then through production filters to remove impurities. The filtered water transfers its thermal energy via heat exchangers. The cooled geothermal water exits the heat exchangers and flows through injection filters and pumps to the injection wells (blue lines). At the heat exchangers, heat is transferred to a secondary liquid (yellow lines), where boiler and combined heat and power (CHP) systems can provide additional heat and/or redundancy. The boiler and CHP systems are powered by gas from the degasser unit (black lines). The secondary liquid is taken to a delivery station for heating purposes and then re-enters the plant through a delivery station to be reheated by the heat exchangers (green lines).

Reinjecting the geothermal fluid is crucial for maintaining reservoir pressure and ensuring long-term sustainability. The fluid is filtered through injection filters to prevent solids from clogging the injection wells and reducing injectivity performance. Injection pumps then pressurize the cooled water and reinject it into the geothermal reservoir through dedicated injection wells.

This closed-loop system maximizes the efficiency of the geothermal plant and minimizes environmental impact while at the same time optimizing plant reliability. The setup is ensuring a long term, stable heat production from the reservoir which is essential for the heat consumers.

Figure 1: Schematic view of a low-enthalpy geothermal plant with 2 ESPs



nieuwe warmte nu

2.2 Data-infrastructure

Operational Technology (OT) data in a low-enthalpy geothermal plant involves the collection, management, and analysis of data generated by various equipment within the plant. This data is crucial for monitoring and controlling plant operations, ensuring safety, and optimizing performance. Typically, OT data at a low-enthalpy geothermal plant is managed by a SCADA (Supervisory Control and Data Acquisition) system.

A SCADA system is a computer-based system used in industrial settings to monitor, control, and optimize processes in real-time. A SCADA system collects data from various sensors, equipment, and machinery throughout an industrial plant, providing operators with a centralized view of operations. This enables real-time monitoring, protection, control, and data analysis. The SCADA system is essential to safeguard essential controls at the geothermal facility. In the future, with further automation, this will become even more essential.

- **Data Acquisition Systems (DAS)**
  - **Sensors and Actuators**: Installed throughout the plant to collect data such as temperature, pressure, flow rate, and other operational parameters.
  - **Programmable Logic Controllers (PLCs)**: Used to control and monitor the equipment based on the sensor data. They often serve as the first point of data aggregation.
  - **Remote Terminal Units (RTUs)**: Used to communicate with the central control systems, especially in remote or distributed setups.
- **Control and Monitoring system**
  - **Human-Machine Interface (HMI)**: is the user interface that allows plant operators to interact with the SCADA system. This could be a computer screen or a control panel showing graphical representations of the plant, alarms, and live data feeds.
- **Data Storage Solutions**
  - **Historian Databases**: Specialized time-series databases optimized for storing large volumes of time-stamped data. Examples include Osisoft PI, AVEVA Historian, and GE Proficy Historian.
  - **Cloud Storage**: Increasingly, OT data is often requested to be stored in cloud databases for enhanced scalability and remote access. Industry tailored solutions are available at Aveva, such as the Aveva Data Hub, but similar services can also be found at either of the large cloud providers, such as Azure, Amazon Web Services (AWS), Google Cloud Platform (GCP).
- **Data Communication Networks**
  - **OPC-UA**: An OPC UA (Open Platform Communications Unified Architecture) server acts as an intermediary for the secure and reliable exchange of data between various components of the SCADA system. While transferring data between PLCs/RTUs, SCADA computers, and Historian databases, the OPC-UA server applies conversions on the various different data source protocols, unifies the data to the OPC standard and thus enables communication between these different endpoints.

In the above-described infrastructure and in the context of this project, Helin supports by facilitating a streaming connection between edge and cloud. The **Helin-Data-Collector** (HDC) is a subscribe-publish type system, that is configurable to various data-sources and data-sinks, such as an OPC-UA server's message feed, to forward the received datapoints real-time to a cloud endpoint.

3 Data

A digital twin's scope depends on the definition of its developer and/or facilitator. Within the Nieuwe Warmte Nu project, the digital twin is developed by the project's primary contributor, TNO, who defines the digital twin as follows;

“TNO is developing the first geothermal system digital twin called GEMINI since 2021. It is a flexible web-based framework for real-time monitoring, control and forecasting a geothermal system. GEMINI acts as a decision support system (DSS) to the operators of geothermal and ATES systems:

- Enable fast, accurate and robust operational decisions
- Centralized location to access all the data and information
- Performance and environmental footprint monitoring
- Critical processes monitoring (scaling, erosion, corrosion)
- Production and operation control

- Failure and anomaly detection
- Predictive maintenance and optimum scheduling”

According to this definition, the GEMINI Digital Twin for a low-enthalpy geothermal plant with doublet system operates at its maximum potential if it has access to the ‘component’ data-points listed in Table 1.

However, a selection of data-points is only adaptable at other geothermal plants if it adheres to a standardized naming convention. In other words, a data model needs to be introduced to guarantee the scalability of the project.

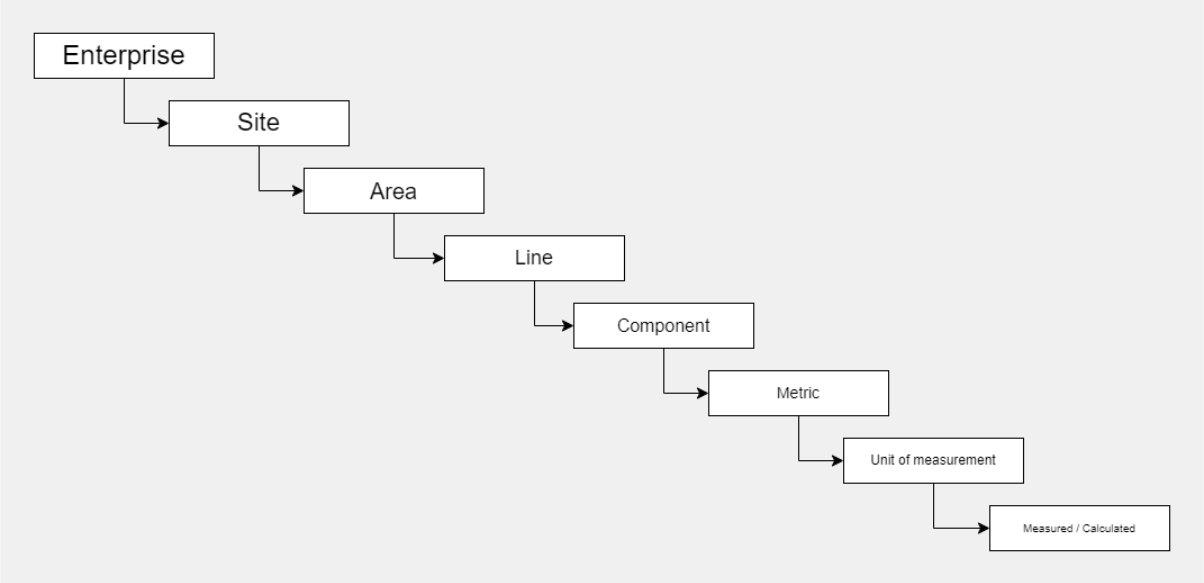
3.1 The ISA95 data model

We recommend applying the ISA95 data model in the geothermal data acquisition process. Although the International Society of Automation (ISA) developed ISA-95 already in the mid-1990s, it is an abstract model that accommodates a wide range of technologies and systems. Its scope prioritizes activities, not technologies, and its intended purpose as a tech-agnostic communication model remains relevant. (International Society of Automation, 2024)

“ISA-95, also known as ANSI/ISA-95 or IEC 62264, is an international set of standards aimed at integrating logistics systems with manufacturing control systems. It organizes technology and business processes into layers defined by activities taking place, and it outlines how an enterprise can set up an interface to communicate among these layers. The ISA-95 standards framework is widely accepted as essential to modern manufacturing. It relies on the well-known Purdue Reference Model for computer-integrated manufacturing to describe network segmentation in industrial control systems. ISA-95 establishes an architecture based on the Purdue model that enterprises can apply regardless of the technology used. This equipment hierarchy model can also be applied across discrete, continuous and logistics industries.” (International Society of Automation, 2024)

ISA95 standard consists of 8 parts. For mapping data generated during the process of geothermal heat production we propose to apply Part 1, which is often referred to as the "Enterprise->Site->Area->Line" model. Based on the example schematic, we separated 2 main ‘Areas’, Production (Prd), and Injection (Inj). Within these areas, we identified different equipment, and machinery as “Lines”. For example, ESPs, filters, heat exchangers, pumps with different purposes should each be mapped to an abbreviated equipment definition, indexed with an enumerator (e.g.: ‘_01’). Given the variety of data sources, we extended the data model with 3 more levels. First, we defined equipment ‘Components’, and mapped those to a ‘Sub-Line’ level. Second, we categorized ‘Metrics’, such as pressure, temperature, flowrate, etc. , next, we added the unit of measurement, and lastly we append one more level a suffix annotating whether the metric is measured or calculated.

Figure 2: ISA95 data model for geothermal plants



3.2 Data inventory

Advocating for this data model, each required data-point of the digital twin to the appropriate names is mapped. (See Table 1.) Enforcing these standardized tag-names allows for a modular approach when deploying GEMINI at future locations. This can also be served as an initial standardized template for the data inventory and data models in the low-enthalpy geothermal plants. Note that in different sites and projects, not all the tags and data points are required to be present but this table can serve as the baseline to start the inventory.

Table 1: Necessary input data for GEMINI Digital Twin

Item	Data-point	ISA95 mapped data-point
1	**ESP**	**'Ent'.'Site'.'Area'.'Line'.'Comp'.'Metric'.'UoM'.'MEAS/CALC'**
1.1	**Measured**	
1.1.1	esp_current	OP.Site.Prd.ESP_01.Main.Current.A.MEAS
1.1.2	esp_flow	OP.Site.Prd.ESP_01.Main.Flow.m3/hr.MEAS
1.1.3	esp_frequency	OP.Site.Prd.ESP_01.Main.Freq.Hz.MEAS
1.1.4	esp_head	OP.Site.Prd.ESP_01.Main.Head.M.MEAS
1.1.5	esp_inlet_pressure	OP.Site.Prd.ESP_01.Inlet.Pressure.Bar.MEAS
1.1.6	esp_inlet_temperature	OP.Site.Prd.ESP_01.Inlet.Temp.C.MEAS
1.1.7	esp_motor_temperature	OP.Site.Prd.ESP_01.Motor.Temp.C.MEAS
1.1.8	esp_outlet_pressure	OP.Site.Prd.ESP_01.Outlet.Pressure.Bar.MEAS
1.1.9	esp_outlet_temperature	OP.Site.Prd.ESP_01.Outlet.Temp.C.MEAS
1.1.10	esp_power_consumption	OP.Site.Prd.ESP_01.Main.PowerUsage.Mw.MEAS
1.1.11	esp_vibration_x	OP.Site.Prd.ESP_01.Main.VibrationX.m/s2.MEAS
1.1.12	esp_vibration_y	OP.Site.Prd.ESP_01.Main.VibrationY.m/s2.MEAS
1.1.13	esp_voltage	OP.Site.Prd.ESP_01.Main.Voltage.V.MEAS
1.2	**Calculated**	
1.2.1	esp_discharge_pressure	OP.Site.Prd.ESP_01.Discharge.Pressure.Bar.CALC
1.2.2	esp_efficiency	OP.Site.Prd.ESP_01.Main.Efficiency.UoM.CALC
1.2.3	esp_measured_head	OP.Site.Prd.ESP_01.Main.Head_.M.CALC
1.2.4	esp_power	OP.Site.Prd.ESP_01.Main.PowerGen.Mw.CALC
2	**Booster pump**	
2.1	**Measured**	
2.1.1	boosterpump_current	OP.Site.Prd.BPump_01.Main.Current.A.MEAS
2.1.2	boosterpump_flow	OP.Site.Prd.BPump_01.Main.Flow.m3/hr.MEAS
2.1.3	boosterpump_frequency	OP.Site.Prd.BPump_01.Main.Freq.Hz.MEAS
2.1.4	boosterpump_head	OP.Site.Prd.BPump_01.Main.Head.M.MEAS
2.1.5	boosterpump_inlet_pressure	OP.Site.Prd.BPump_01.Inlet.Pressure.Bar.MEAS
2.1.6	boosterpump_inlet_temperature	OP.Site.Prd.BPump_01.Inlet.Temp.C.MEAS
2.1.7	boosterpump_motor_temperature	OP.Site.Prd.BPump_01.Motor.Temp.C.MEAS
2.1.8	boosterpump_outlet_pressure	OP.Site.Prd.BPump_01.Outlet.Pressure.Bar.MEAS
2.1.9	boosterpump_outlet_temperature	OP.Site.Prd.BPump_01.Outlet.Temp.C.MEAS
2.1.10	boosterpump_power_consumption	OP.Site.Prd.BPump_01.Main.PowerUsage.Mw.MEAS
2.1.11	boosterpump_voltage	OP.Site.Prd.BPump_01.Main.Voltage.V.MEAS
2.2	**Calculated**	
2.2.1	boosterpump_efficiency	OP.Site.Prd.BPump_01.Main.Efficiency.UoM.CALC
2.2.2	boosterpump_head	OP.Site.Prd.BPump_01.Main.Head.M.CALC
2.2.3	boosterpump_power	OP.Site.Prd.BPump_01.Main.PowerGen.Mw.CALC
3	**Injection pump**	
3.1	**Measured**	
3.1.1	injectionpump_current	OP.Site.Inj.InjPump_01.Main.Current.A.MEAS
3.1.2	injectionpump_flow	OP.Site.Inj.InjPump_01.Main.Flow.m3/hr.MEAS
3.1.3	injectionpump_frequency	OP.Site.Inj.InjPump_01.Main.Freq.Hz.MEAS
3.1.4	injectionpump_head	OP.Site.Inj.InjPump_01.Main.Head.M.MEAS
3.1.5	injectionpump_inlet_pressure	OP.Site.Inj.InjPump_01.Inlet.Pressure.Bar.MEAS
3.1.6	injectionpump_inlet_temperature	OP.Site.Inj.InjPump_01.Inlet.Temp.C.MEAS
3.1.7	injectionpump_motor_temperature	OP.Site.Inj.InjPump_01.Motor.Temp.C.MEAS
3.1.8	injectionpump_outlet_pressure	OP.Site.Inj.InjPump_01.Outlet.Pressure.Bar.MEAS
3.1.9	injectionpump_outlet_temperature	OP.Site.Inj.InjPump_01.Outlet.Temp.C.MEAS

Item	Data-point	ISA95 mapped data-point
3.1.10	injectionpump_power_consumption	OP.Site.Inj.InjPump_01.Main.PowerUsage.Mw.MEAS
3.1.11	injectionpump_voltage	OP.Site.Inj.InjPump_01.Main.Voltage.V.MEAS
3.2	**Calculated**	
3.2.1	esp_discharge_pressure	OP.Site.Inj.InjPump_01.Discharge.Pressure.Bar.MEAS
3.2.2	injectionpump_head	OP.Site.Inj.InjPump_01.Main.Head.M.MEAS
3.2.3	injectionpump_power	OP.Site.Inj.InjPump_01.Main.PowerGen.Mw.MEAS
4	**Injection well**	
4.1	**Measured**	
4.1.1	injectionwell_bottomhole_pressure	OP.Site.Inj.InjWell_01.Bottomhole.Pressure.Bar.MEAS
4.1.2	injectionwell_bottomhole_temperature	OP.Site.Inj.InjWell_01.Bottomhole.Temp.C.MEAS
4.1.3	injectionwell_flow	OP.Site.Inj.InjWell_01.Main.Flow.m3/hr.MEAS
4.1.4	injectionwell_wellhead_pressure	OP.Site.Inj.InjWell_01.Wellhead.Pressure.Bar.MEAS
4.1.5	injectionwell_wellhead_temperature	OP.Site.Inj.InjWell_01.Wellhead.Temp.C.MEAS
4.2	**Calculated**	
4.2.1	injectionwell_bottomhole_pressure	OP.Site.Inj.InjWell_01.Bottomhole.Pressure_.Bar.CALC
4.2.2	injectionwell_injectivity_index	OP.Site.Inj.InjWell_01.Main.Inj_Index.UoM.CALC
5	**Heat exchanger**	
5.1	**Measured**	
5.1.1	hex_primary_flow	OP.Site.Prd.HeatExch_01.Main.Flow.m3/hr.MEAS
5.1.2	hex_primary_inlet_pressure	OP.Site.Prd.HeatExch_01.Inlet.Pressure.Bar.MEAS
5.1.3	hex_primary_inlet_temperature	OP.Site.Prd.HeatExch_01.Inlet.Temp.C.MEAS
5.1.4	hex_primary_outlet_pressure	OP.Site.Prd.HeatExch_01.Outlet.Pressure.Bar.MEAS
5.1.5	hex_primary_outlet_temperature	OP.Site.Prd.HeatExch_01.Outlet.Temp.C.MEAS
5.1.6	hex_secondary_flow	OP.Site.Prd.HeatExch_02.Main.Flow. m3/hr.MEAS
5.1.7	hex_secondary_inlet_pressure	OP.Site.Prd.HeatExch_02.Inlet.Pressure.Bar.MEAS
5.1.8	hex_secondary_inlet_temperature	OP.Site.Prd.HeatExch_02.Inlet.Temp.C.MEAS
5.1.9	hex_secondary_outlet_pressure	OP.Site.Prd.HeatExch_02.Outlet.Pressure.Bar.MEAS
5.1.10	hex_secondary_outlet_temperature	OP.Site.Prd.HeatExch_02.Outlet.Temp.C.MEAS
5.2	**Calculated**	
5.2.1	hex_primary_thermal_power	OP.Site.Prd.HeatExch_01.Main.ThermalPower.Mw.CALC
5.2.2	hex_secondary_thermal_power	OP.Site.Prd.HeatExch_02.Main.ThermalPower.Mw.CALC
5.2.3	hex_thermal_exchange	OP.Site.Prd.HeatExch_TT.Main.ThermalExchange.UoM.CALC
6	**Gas boiler**	
6.1	**Measured**	
6.1.1	gasboiler_heat_pipe_flow	OP.Site.Prd.Boiler_01.HeatPipeMain.Flow.m3/hr.MEAS
6.1.2	gasboiler_heat_pipe_inlet_temperature	OP.Site.Prd.Boiler_01.HeatPipeInlet.Temp.C.MEAS
6.1.3	gasboiler_heat_pipe_outlet_temperature	OP.Site.Prd.Boiler_01.HeatPipeOutlet.Temp.C.MEAS
6.1.4	gasboiler_inlet_gas_flow	OP.Site.Prd.Boiler_01.GasPipeInlet.Flow.m3/hr.MEAS
6.2	**Calculated**	
6.2.1	gasboiler_efficiency	OP.Site.Prd.Boiler_01.Main.Efficiency.UoM.CALC
6.2.2	gasboiler_thermal_power	OP.Site.Prd.Boiler_01.Main.ThermalPower.Mw.CALC
7	**Filter**	
7.1	**Measured**	
7.1.1	filter_flow	OP.Site.Inj.Filter_01.Main.Flow.m3/hr.MEAS
7.1.2	filter_inlet_pressure	OP.Site.Inj.Filter _01.Inlet.Pressure.Bar.MEAS
7.1.3	filter_inlet_temperature	OP.Site.Inj.Filter _01.Inlet.Temp.C.MEAS
7.1.4	filter_outlet_pressure	OP.Site.Inj.Filter _01.Outlet.Pressure.Bar.MEAS
7.1.5	filter_outlet_temperature	OP.Site.Inj.Filter _01.Outlet.Temp.C.MEAS
8	**Degasser**	
8.1	**Measured**	
8.1.1	degasser_inlet_flow	OP.Site.Prd.Degasser.Inlet.Flow.m3/hr.MEAS
8.1.2	degasser_level	OP.Site.Prd.Degasser.Main.Level.UoM.MEAS
8.1.3	degasser_outlet_gas_flow	OP.Site.Prd.Degasser.Outlet.GasFlow.m3/hr.MEAS
8.1.4	degasser_outlet_liquid_flow	OP.Site.Prd.Degasser.Outlet.LiquidFlow.m3/hr.MEAS
8.1.5	degasser_pressure	OP.Site.Prd.Degasser.Main.Pressure.Bar.MEAS

4 Infrastructure and data management guidelines

Setting up a robust database and IT infrastructure is critical for ensuring that the Digital Twin operates seamlessly with the high-frequency, high-volume data generated by a low-enthalpy geothermal power plant. This section outlines key considerations and recommends solutions for building an efficient, scalable, and secure infrastructure to support Digital Twin operations.

4.1 Key considerations

The following factors must be considered during the backend design of an application with extensive data dependency:

- **Multi-tier architecture:** To optimize scalability, reliability, and fault tolerance, it is recommended to avoid co-locating multiple application layers — such as the database, backend processing, and web application — on a single virtual machine or compute instance. Instead, implement a multi-tier architecture by assigning each layer to a separate compute resource. This distributed approach minimizes cross-layer dependencies, allows for independent scaling and maintenance, and reduces the risk of resource contention that could arise from co-locating containers on a single virtual machine.

- **Managing data lifecycle**: Geothermal plants generate substantial volumes of real-time operational data, which play a crucial role in supporting a responsive and insightful digital twin. However, continuous ingestion and indefinite storage of raw data alone do not ensure an efficient backend capable of delivering fast and reliable analytics. Over time, the accumulation of raw data can degrade database performance, making it essential to balance data volume with system responsiveness. For high-frequency data sources, an optimal data retention strategy often involves determining a data collection frequency that sufficiently captures relevant information for operational decision-making while minimizing redundant data. Regular aggregations to this optimal frequency help manage data growth, while aged raw data can be gradually transitioned to more cost-effective storage tiers, such as Azure's hot, cool, and cold options, and eventually to archival storage. As data reaches the end of its lifecycle and no longer serves an analytical purpose, it may be systematically phased out from the system, ensuring a streamlined and performant backend.

- **Data Latency:** Low data latency is crucial for responsive applications, particularly those requiring real-time insights. Network speed, processing time, and inter-service data transfers impact latency, but responsiveness can be improved by using low-latency storage, edge computing for localized processing, and caching. Pre-calculating and storing model results also helps reduce computation time, delivering faster insights to the application.

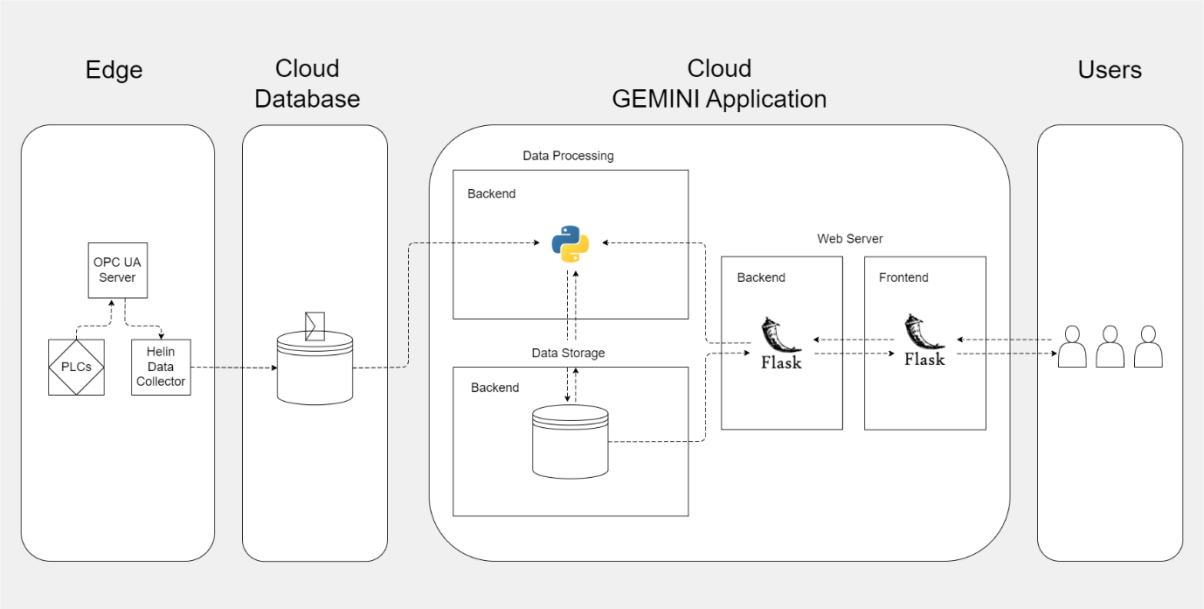
- **Cloud Integration:** Cloud integration enables scalability and smooth interoperability among application components, but there is often a trade-off between using cloud-native services and maintaining a cloud-agnostic approach. While cloud-native services like managed databases and serverless functions often reduce maintenance costs and simplify infrastructure, a cloud-agnostic setup offers flexibility but can be more complex and expensive to manage across platforms.

- **Compliance:** Regulatory compliance is essential in data-intensive applications, particularly for sensitive information. Adhering to data privacy, security, and storage regulations across all application layers ensures data integrity and legal compliance, supporting consistent standards across regions.

4.2 Design recommendations

Given a low-enthalpy geothermal plant, and the expected nature of streaming data available at edge, Helin proposes the GEMINI Digital Twin to fit into an infrastructure as described in Figure 3. That is, in an ideal setup the digital twin consists of 3 service layers. 1) A data processing layer, 2) a data storage layer, and 3) a web server layer. To establish a robust infrastructure, each layer should be supported by distinct cloud compute and storage hosts, carefully selected and dedicated to the requirements of the specific service layer.

Figure 3: Proposed IT infrastructure



4.2.1 Data processing tier

Connecting directly to industrial data-sources at edge locations is outside of the current scope of GEMINI. In future stages, the open-source version of the Helin-Data-Collector, together with stream-data-processing features can integrated into the GEMINI application. The Helin-Data-Collector is capable to send streaming data to various endpoints, data-sinks, such as TimescaleDB. However, at the current stage, the recommendation is to design the data processing service layer so that it integrates with common third-party streaming message hubs, as well as API endpoints. In other words, the data processing backend can, and should expect the same high frequency, raw data available at edge to be also available through a cloud endpoint, reachable via open-source technologies.

Building the GEMINI to be able to process stream data from a Kafka topic is one way to align with this design goal. Kafka is a data-streaming technology, capable of handling high-frequency data. Kafka is scalable, and fault-tolerant, originally built as an open-source project by the Apache Foundation. Today, Kafka is also available as a managed service by companies such Confluent. A Kafka topic, which is a key component of the Kafka infrastructure, is a publish-subscribe message hub, where multiple clients (“consumers”) can read data simultaneously from. A Kafka topic is compatible with various 3rd party data-sinks. In the case of GEMINI, through open-source Kafka-Python SDKs, the data processing backend could subscribe to this topic, to consume and process stream data becoming available to it.

In addition to Kafka, we recommend the GEMINI data processing layer to have the functionality of scheduling data retrieval requests to cloud API endpoints. It is not uncommon for geothermal plant operators to have existing cloud solutions already in place for data sinks. If this is indeed the case, like at the demonstration sites, with the scheduled, recurring requests to retrieve data for the latest time-window batches, the processing layer, similarly as in case of subscribing to a topic, receives streaming data ready to be processed further.

Handling and utilizing the raw streaming data, on the one hand, the data processing backend can become a filter on the raw data stream, if by processing we strictly mean keeping a subset of the raw data points at lower frequencies. On the other hand, capabilities in python stream data processing are endless, and the service can be tailored to reduce the incoming data set through enhanced methods, such as by calculating statistics on non-overlapping, contiguous time-windows.

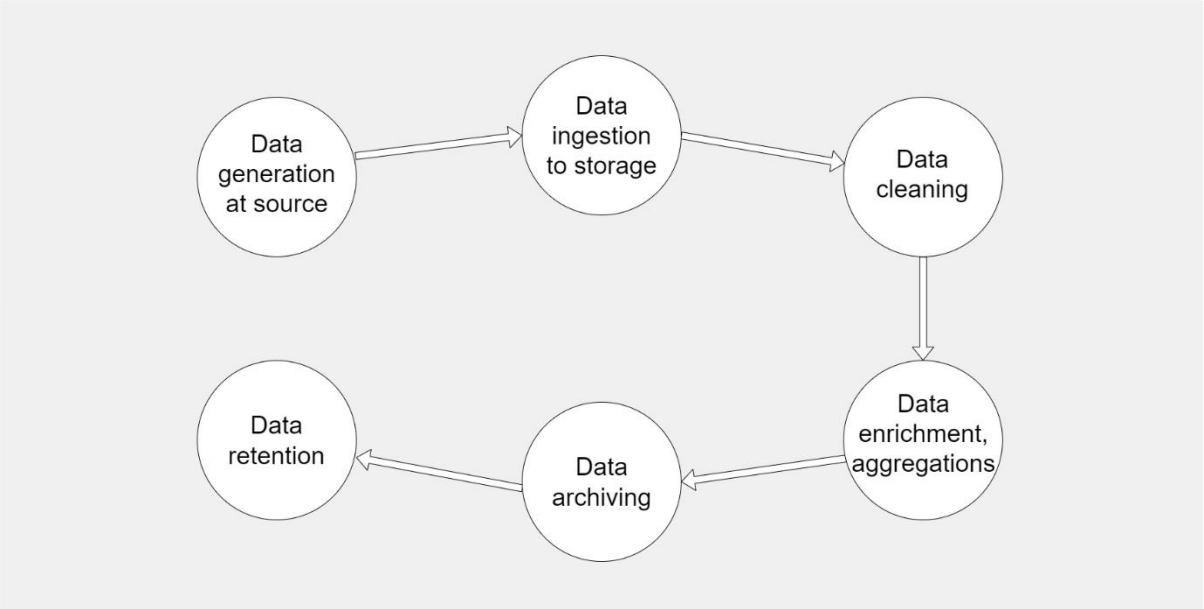
Converting raw data points into a processed dataset through sampling or time-window aggregation serves as an essential noise-reduction and data-cleaning step. This approach standardizes raw time-series observations—originating from diverse sources with varying frequencies and often overly precise timestamps—into a dataset with a unified time-series index. Achieving this matching index, with observations across all data sources, is crucial for the dataset to be viable for descriptive and predictive analysis, and input for model calculations.

4.2.2 Data storage tier

Given an efficiently setup data processing service tier, the data landing to the storage of GEMINI is cleaned, enriched with pre-calculated model metrics, and is ready to serve requests coming from the web-server layer. There are several database options available, each with its own strengths depending on the project's specific needs. The use of TimescaleDB due to its following advantages is recommended;

- **Efficient time-series data management:** Timescale ingests data into hypertables. Hypertables are PostgreSQL tables that automatically partition data by time. In other words, data is divided into smaller tables, called chunks, each of which holds data from a specific time range. Partitioning prepares data for efficient queries. Hypertables unlock several features that make it easier to manage time-series data, such as automatic partitioning, time-based queries, and continuous aggregates. (Timescale, 2024)
- **Continuous Aggregates feature** of Timescale enables efficient management and querying of recent data. Continuous aggregates function similarly to incrementally refresh materialized views in standard SQL but are specifically optimized for time-series data, providing enhanced efficiency and flexibility. Unlike traditional views that require recalculating all data, Timescale's continuous aggregates keep previous calculations cached and update only with new or changed data, minimizing computational overhead and making analytics faster and more cost-effective. Furthermore, time-based partitioning within these aggregates allows for efficient extensions with incoming data, boosting performance and scalability. (Timescale, 2024)
- Notably, Timescale's continuous aggregates can potentially take over the aggregation and down-sampling tasks from the data processing tier, streamlining the backend compute needs. In this configuration, the data processing tier would primarily handle ingesting raw data into the database, let the Timescale compute engine to manage much of the stream processing. The processing tier would access the storage tier for standardized, lower frequency aggregates to serve as input for model calculations (of which the output, again would be stored to a designated storage tier table. This setup could reduce the overall load on the processing tier by shifting certain aggregation tasks to Timescale's compute capabilities. (Timescale, 2024)
- Timescale's **automated lifecycle management** leverages tiered storage and efficient data deletion strategies to optimize performance and reduce costs for data-intensive applications, such as the GEMINI Digital Twin. As described earlier, the time-partitioning of tables enables automating the movement of older, infrequently accessed data to cost-effective storage tiers. This approach lowers storage expenses while keeping high-performance storage available for frequently accessed data, ensuring that historical data remains accessible for analysis without impacting current performance. Data retention, which stands for the complete dropping of tables from the database, capitalizes on the same mechanism of time-based partitioning. By automatically managing storage based on data age and access frequency, Timescale reduces storage costs while maintaining optimal database performance for recent, high-usage data. This automated process minimizes manual intervention, helping streamline both storage and compute resources. (Timescale, 2024)
- **Integration with Existing Tools**: As TimescaleDB is based on PostgreSQL, it integrates easily with a wide range of tools (Grafana, Python, R, etc.) commonly used in digital twin analytics and machine learning applications for geothermal systems. This makes it easier to build dashboards and machine learning models. (Timescale, 2024)

Figure 4: Data Lifecycle



4.3 Security recommendations

To safeguard each component of the distributed, multi-tier architecture, it is essential to implement robust security measures across service authorization, data encryption, and user management. Each service should authorize access to others using role-based policies, with minimal necessary privileges assigned to reduce exposure. All data should be encrypted both in transit and at rest to protect sensitive information; for example, TLS can secure data transfers, while storage-level encryption (e.g., Azure Storage encryption) ensures data protection at rest. For deployments on Azure, secrets and sensitive configuration details should be securely stored in Azure Key Vaults, where they are accessible only to authorized services. Managed identities are preferred for inter-service authentication, as they eliminate the need for hard-coded credentials. If managed identities are not feasible, service principals provide a secure alternative with granular access control. For user management, options such as Auth0 or Microsoft Entra (formerly Azure Active Directory) can be used to provide secure, scalable identity and access management with multi-factor authentication and role-based access. These measures collectively reinforce security while enabling a scalable, fault-tolerant infrastructure.

5 Conclusion and Next steps

This report provides an overview of the essential guidelines for establishing a robust data management system tailored for the GEMINI Digital Twin implementation in geothermal plants. The next steps involve implementing these recommendations at the demonstration sites, followed by ongoing refinement and optimization to ensure continued operational excellence and compliance.

6 References

International Society of Automation. (2024, 11 29). *ISA95 Standard*. Retrieved from isa.org: https://www.isa.org/standards-and-publications/isa-standards/isa-95-standard

Timescale. (2024, 11 29). *Timescale Documentation*. Retrieved from docs.timescale.com: https://docs.timescale.com

TNO. (2024, 11 29). *GEMINI: intelligent decision support system for geothermal assets*. Retrieved from TNO.nl: https://www.tno.nl/en/newsroom/insights/2023/11/intelligent-support-system-geothermal/