UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Siim Suitslepp

# Data mart for photovoltaic parks SCADA systems

Master's Thesis (15 ECTS)

Supervisor(s): Taavi Sarnet, MSc
Mozhgan Pourmoradnasseri, PhD

Tartu 2024

## Data mart for photovoltaic parks SCADA systems

**Abstract:**

Estonia and Europe have set ambitious goals to achieve climate neutrality by 2050. Achieving these goals requires a significant push toward renewable energy sources, particularly wind and solar power. This transition is crucial to ensure that future generations inherit a stable climate and environment, similar to what we experience today [Fet20]. Solar energy has made remarkable progress, with a 40% growth rate in 2023 alone. To meet these climate targets, the pace of renewable energy development in Europe and Estonia must accelerate, making it essential to attract more investment into the sector. As the renewable electricity market has rapidly evolved in recent years, traditional passive renewable energy projects have become less financially viable. The focus has shifted toward controllable renewable projects that require extensive data manipulation and analysis. This thesis addresses this need by exploring the data generated by solar parks, identifying key challenges, and proposing a new approach to managing and utilizing solar park data through the creation of a data mart. The research begins with an examination of the initial issues related to device data, followed by the development of a data pipeline designed to feed into the data mart. The thesis concludes with a discussion on the usage of data mart with a demonstration of a first feature to fill gaps in production data.

**Keywords:**

Photovoltaic (PV) parks, SCADA, Data mart, Data warehouse, Renewable energy, Solar energy, Data pipeline, Data analysis, Database

**CERCS:** P175 (Informatics, systems theory)

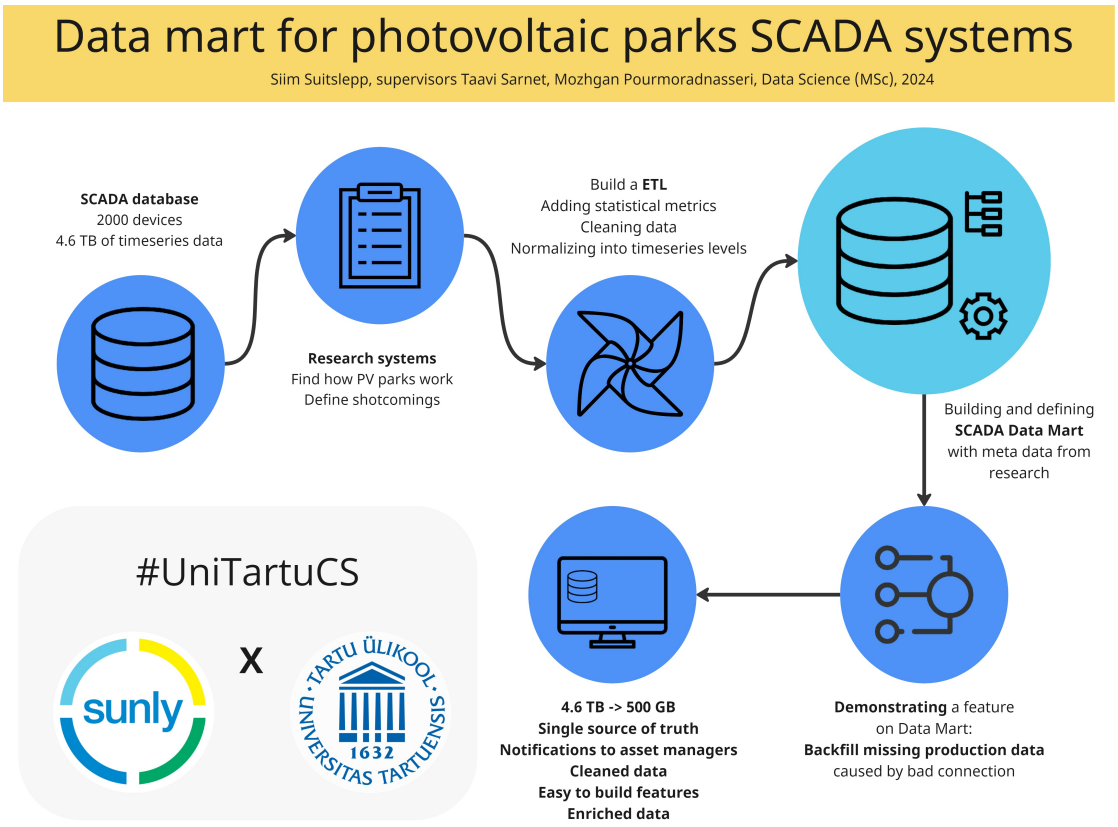



Figure 1. Visual abstract

## Andmeait päikseparkide SCADA süsteemile

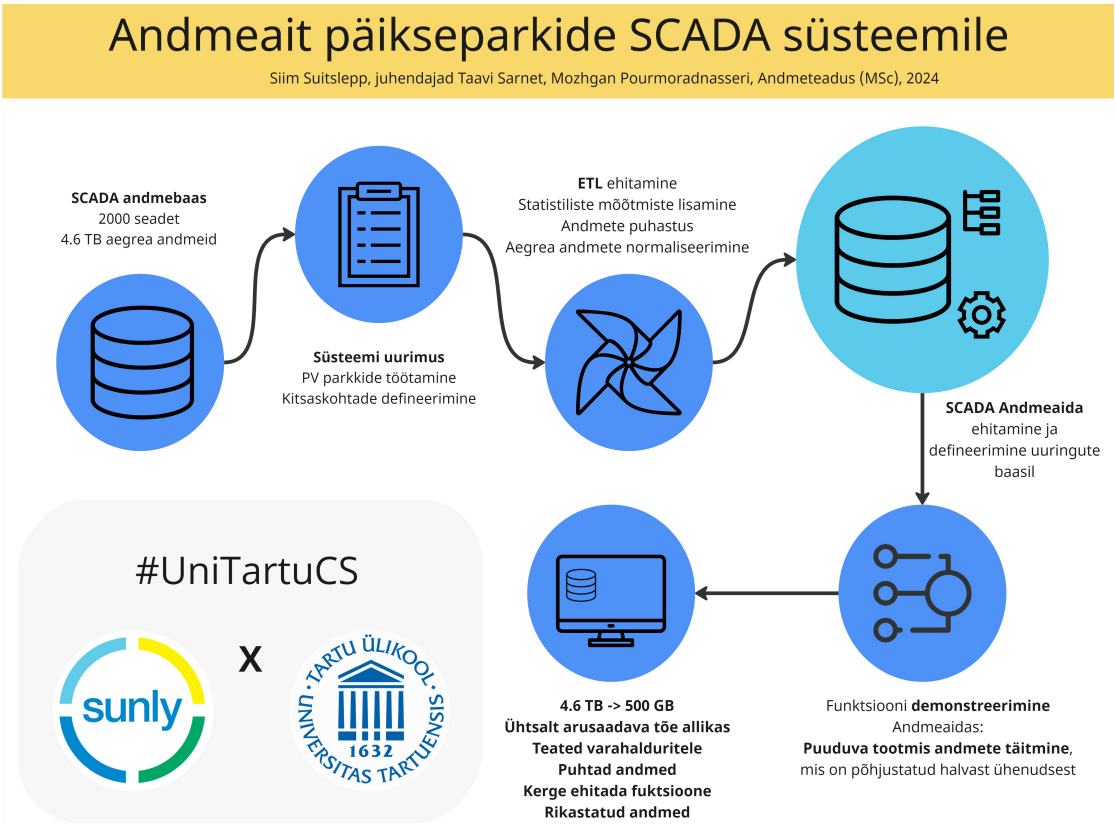
**Lühikokkuvõte:**

Eesti ja Euroopa on seadnud ambitsioonikad eesmärgid saavutada 2050. aastaks kliimaneutraalsuse. Nende eesmärkide saavutamine nõuab märkimisväärset fookust taastuvate energiaallikate, eelkõige tuule- ja päikeseenergia suunas. See üleminek on oluline, et tagada tulevasetele põlvkonnadadele stabiilne kliima ja keskkond, mis on sarnane praegusele [Fet20]. Päikeseenergia on teinud märkimisväärseid edusamme, 2023. aastal kasvas see 40% võrra. Nende kliimaeesmärkide saavutamiseks peab taastuvenergia arendamise tempo Euroopas ja Eestis kiirenema, mistõttu on oluline sellesse sektorisse rohkem investeeringuid saada. Kuna taastuvelektriturg on viimastel aastatel kiiresti arenenud, on traditsioonilised passiivse taastuvenergia projektid muutunud rahaliselt vähem tasuvaks. Fookus on nihkunud kontrollitavatele taastuvenergiaprojektidele, mis nõuavad ulatuslikku andmete töötlemist ja analüüsi. See lõputöö käsitleb seda vajadust, uurides päikeseparkide genereeritud andmeid, tuvastades peamised väljakutsed ja pakkudes välja uue lähenemisviisi päikeseparkide andmete haldamiseks ja kasutamiseks andmepargi loomise kaudu. Uuring algab seadmete andmetega seotud esialgsete probleemide uurimisega, millele järgneb andme aidale söötmiseks mõeldud andme toru väljatöötamine. Lõputöö lõppeb aruteluga andme aida kasutamise üle, demonstreerides esimest funktsiooni tootmisandmetes lünkade täitmisega.

**Võtmesõnad:**

PV pargid, SCADA, Andmeait, Andme ladu, Taastuvenergia, Päikese energia, Andme toru, Andme analüüs, Andmebaas

**CERCS:** P175 (Informaatika, süsteemiteooria)




Joonis 2. Visuaalne abstrakt

# Acknowledgements

The author acknowledges the support from the Institute of Computer Science at the University of Tartu during the writing of this thesis. Also, I am grateful to my supervisors, Mozhgan Pourmoradnasseri and Taavi Sarnet, for their patience and all the support and advice they gave me.

# Contents

# 1 Introduction

Solar energy has emerged as a crucial component in the global transition towards renewable energy sources. As the world faces the pressing challenges of climate change and environmental degradation, the need for sustainable energy solutions has never been more critical. Solar energy, being abundant, renewable, and environmentally friendly, presents a viable solution to reduce greenhouse gas emissions and dependence on fossil fuels. Estonia, with the transition of the European Green Deal [Fet20], has taken on the challenge of moving away from fossil fuel dependency and into greener and more sustainable energy with solar parks [Tat22].

However, in modern renewable projects, the market has changed; passive production alone is not sufficient for companies to meet their targets and remain profitable. To achieve the desired outcomes, there is a growing need for controllable renewable projects that can dynamically adjust production based on demand and other factors. This shift necessitates a robust data handling and manipulation platform that can provide accurate, real-time insights, long term analytical data and control mechanisms to optimize energy production and distribution [HMB18].

This is where this thesis addresses the issue of managing and utilizing the vast amounts of data generated by solar parks by transforming it into actionable insights through the creation of a comprehensive data mart with data workflows. This will provide automated data maintenance and management, which is essential for several reasons.

Firstly, it aims to enhance the reliability and performance of solar parks by providing reliable real-time monitoring and predictive maintenance capabilities. Automated systems can detect anomalies and potential issues before they escalate, ensuring uninterrupted energy production and reducing downtime, by letting asset- and project managers know of problematic assets.

Secondly, as the sheer number of solar parks will increase every year, a well-structured data mart facilitates better investment decisions in renewable energy sources. By consolidating and analyzing data from various sources, stakeholders can gain a clearer understanding of performance metrics, financial returns, and operational efficiencies. This transparency enables informed decision-making, optimizing resource allocation, and maximizing returns on investment.

Furthermore, automated data management reduces the burden of asset management. Manual data handling is not only time-consuming but also prone to errors. Automation streamlines these processes, improving accuracy and efficiency. This shift allows for more effective allocation of human resources to strategic tasks rather than routine maintenance.

The integration of automated data maintenance in solar parks is an essential step towards optimizing their control, performance, reliability, and financial viability. By addressing the challenges associated with data handling and transforming raw device data into insights, we can pave the way for more efficient and sustainable energy solutions for the modern electricity market. This thesis will explore the development and

implementation of a data mart tailored to the unique needs of solar parks, demonstrating its potential to improve the management and operation of renewable energy assets.

## 1.1 Research goals

This thesis focuses on the problems with modern renewable projects and how the need for controllable sources becomes essential, transforming large time-series databases generated by SCADA systems into a modern data mart. Achieving the best results requires a thorough understanding of the domain, which influences the design of the data mart schema, data cleaning and transformation processes, feature selection, and the overall architecture.

## 1.2 Thesis structure

The sections of thesis are organized as follow:

1. Section 2 provides an overview of the importance of PV parks and the need for data in their operation. It discusses the general structure of PV parks, including the various voltage levels and key devices used, and explores the software and data solutions currently used in PV systems.

2. Section 3 describes the initial dataset used in this research. It covers the amount and variety of data, the data-saving logic, and specific tables like sqlth_te and sqlt_data_(drvid)_(date), which are crucial to understanding the SCADA system's data architecture.

3. Section 4 introduces the SCADA data mart, explaining its purpose and the need for centralized data management. This section also outlines the solutions proposed to address issues in the existing data structures and presents the minimum viable product of the data mart, detailing its components and functionalities.

4. Section 5 details the development of the data pipeline, including the choice of technology stack, the ETL workflow, and the stages of data extraction, transformation, and loading. It also evaluates the efficiency and performance improvements brought by the new data pipeline.

5. Section 6 discusses the usage of the data mart, highlighting both internal and external applications. It demonstrates the capabilities of the data mart through a specific use case focused on filling data gaps due to communication issues, ensuring accurate performance assessments of solar parks.

6. Section 7 provides a discussion on the results of the thesis and suggests directions for future work. This includes potential optimizations of the ETL process, migration to more specialized databases like TimescaleDB, and the expansion of the data mart to support additional renewable energy sources.

7. Section 8 concludes the thesis, summarizing the main findings and contributions, and reflecting on the overall impact of the SCADA data mart in enhancing the management and control of photovoltaic parks.

# 2 Backgound

This section gives supporting information on previous works, which are related to the thesis. Subsections 2.1 provides an overview of the importance of PV parks, their problems and how data can support them. Subsection 2.2 gives a high-level overview of the structure of PV parks and their devices, which gives a quick overview of the structure and devices.

## 2.1 PV importance and need of data

Photovoltaic (PV) parks, also known as solar farms, are large-scale installations of solar panels designed to capture sunlight and convert it into electricity. The significance of PV parks is multifaceted, contributing to both environmental sustainability and economic development. PV parks harness solar energy, a sustainable and inexhaustible resource, thereby reducing dependence on fossil fuels and mitigating climate change impacts. The generation of solar energy in these parks significantly lowers greenhouse gas emissions compared to traditional power plants [OAA+21]. Additionally, the development of PV parks creates jobs and stimulates local economies through investment and infrastructure development [MA22]. They also enhance energy security and stability by reducing reliance on imported fuels [AASA+24]. The scalability and flexibility of PV technology allow installations to range from small systems to large farms, integrating easily into various energy grids. Furthermore, the growth of PV parks drives innovations in solar technology, continuously improving efficiency and feasibility.

The expansion of PV projects in Estonia is rapidly increasing, driven by directives such as the European Union's "European Green Deal" [Fet20] and Estonia's "National Energy and Climate Plan," which aim to make the EU climate-neutral by 2050 and generate 100% of energy from clean sources.

To meet these ambitious goals, it is crucial to have a comprehensive overview of asset performance, which presents new challenges in data management. One of the most cost-effective ways to achieve this is through the use of high-quality data. Regular and predictive maintenance, guided by data insights, ensures that systems operate at peak efficiency. For example, array metrics and irradiance sensors can be used to detect panel degradation, dirt accumulation, or component failures [ASMA20]. As PV parks scale, ensuring the integrity of all resources and devices becomes increasingly important. While Supervisory Control and Data Acquisition (SCADA) platforms provide a quick overview of park operations, they are not inherently designed for long-term or scalable analysis, necessitating a more comprehensive data management solution [GDC+16].

Effective data management also plays a crucial role in reducing operational costs by enabling predictive maintenance, which identifies potential issues before they impact production, thereby avoiding unplanned downtime. Additionally, compliance with industry standards and regulations is essential for the legal and safe operation of PV

parks. Proper data collection and analysis ensure adherence to these standards, preventing legal penalties.

As the renewable energy sector continues to grow, securing future investments becomes increasingly important. Data-driven insights from current and previous projects provide a solid foundation for informed decision-making. The modern trend toward controllable renewable energy sources, particularly in the Estonian and European electricity markets, further emphasizes the need for high-quality data. Reliable data enables better decision-making regarding asset management, forecasting, and market participation, ensuring that renewable energy projects remain competitive and sustainable.

## 2.2 General Structure of PV parks and devices

The PV parks are the result of different devices working together. Essential devices like inverters and PV panels, generating electricity, are being transferred to the grid by support devices like data loggers, controllers, transformers, meters and switchgear as can be seen in Figure 3. The devices are all communicating with Modbus TCP/IP protocol. In this section, we will go into each layer of PV parks and what we can take away from all of the structures to design a better data mart for PV parks.

### 2.2.1 Low Voltage

The low voltage level exists in all of the PV parks, no matter the size. Starting from the bottom, we have PV panels, which turn irradiance into DC current. From the architecture side, it is important to know what is the angle of the panels and what is the azimuth, as the popularity of east-west facing panels has been gaining more momentum, and with the difference in position to commonly south-facing PV parks, there is a difference in the power curve. Other essential devices are inverters, inverters gather the power from PV panel arrays, aggregate and convert the energy into AC output. The most amount of data is generated by inverters as this is the first place in the whole design, where we can get info to describe what is going on in the park at the lowest level.

In addition to the inverters, there are dataloggers, irradiance sensors, and weather stations that enhance the data gathered from inverters. Dataloggers play the role of data concentrators, which are connected to inverters aggreagating the info together with support devices like irradiance sensors to make it easier to manage communication with different protocols. However, as they are not essential devices, some parks don't have dataloggers nor irradiance sensors connected to them and just work on purely panels and inverters.

Irradiance sensors provide essential input to asset management to see how the park is performing comparatively to the park's production, as the physical properties at the moment of measurement are the same for irradiance sensors and PV panels, which give input into the condition of the park.

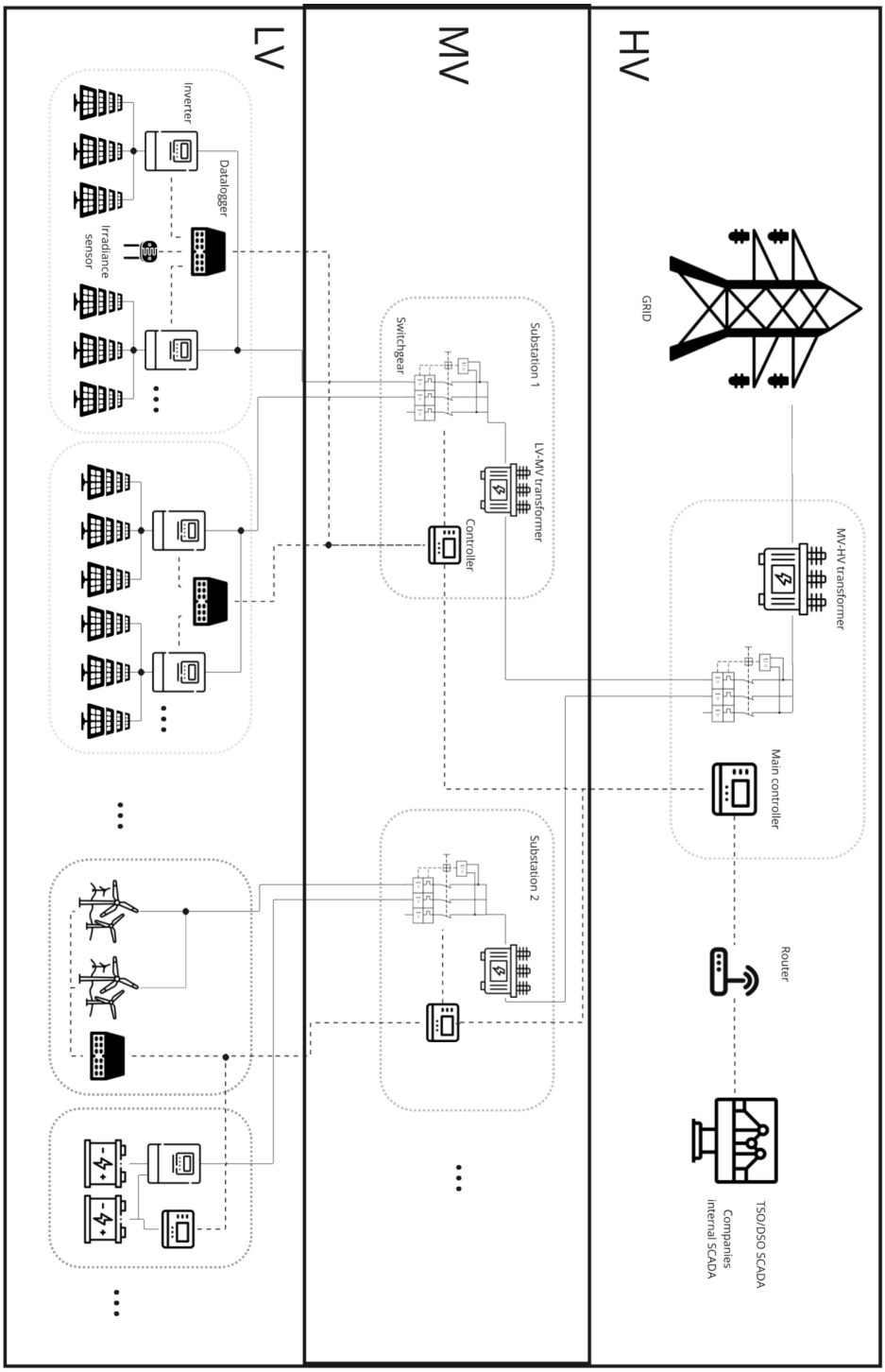



Figure 3. High level design of renewable energy parks

To add more context, PV panels and inverters could also be added together with wind turbines and/or battery packs to create hybrid projects. This could give extra info in the future of the data mart and its design, but the medium voltage and high voltage systems stay the same.

### 2.2.2 Medium Voltage

The medium voltage (MV) section acts as an intermediary between the high voltage transmission lines and the low voltage systems. Key devices in the MV system include switchgear, transformers, and protection relays. Switchgear is essential for controlling the flow of electricity and isolating parts of the system for maintenance or in case of faults, so the system can still keep on running. On smaller parks, usually, there is not any data coming from the medium voltage side, but if there is, the most important info is about input pole readings of P (active power) and Q (reactive power); each pole is connected to a section of low voltage. Additionally, park controllers are usually located here also called as remote terminal unit (RTU) which connect to grid operator, control, measure and give info about what is going on from the substation level, the location of the RTU is dependent on what is the highest voltage level in the park.

### 2.2.3 High Voltage

In the high voltage (HV) segment of a PV park, the primary components are transformers and high voltage substations. The main function at this level is to step up the voltage generated by the PV park for transmission which will be connected to the grid. The key metrics at the HV level include voltage (V), current (I), P and Q. Monitoring these metrics ensures the safe and efficient transfer of electricity to the national grid. For some parks that are connected to the lower voltage of the grid, the high voltage level is not needed as the medium voltage is already sufficient enough to be able to connect to the grid.

### 2.2.4 Takeaways

The main thing to take away from this subsection is that most of the data come from all of the inverters on the low voltage side. Additionally, the following devices are most commonly used as support devices: dataloggers, irradiance sensors, RTUs, battery units, battery support inverters, switchgear, and weather stations. Common metrics collected include P, Q, U, I, and total production at each level, alongside metadata about the park's architecture. Looking ahead, frequency data will become increasingly crucial for grid operations, especially as the Baltic countries move towards desynchronization from the Russian grid [Ele24].

## 2.3 Software and data solutions

In this section, we will be going over how an industrial and monitoring solution is used in PV parks, and what data we have as an output from these systems, with positive and negative sides brought out about the data in general.

SCADA systems are used for monitoring and controlling industrial processes and infrastructure on a large scale. These systems gather real-time data from remote locations to control equipment and conditions. SCADA is essential in industries such as utilities (electricity, water, wastewater), oil and gas, manufacturing, and transportation. A typical SCADA system includes:

1. Human-Machine Interface (HMI): Interface for operators to interact with the system.
2. Supervisory (computer) systems: Central system to gather and process data.
3. Remote Terminal Units (RTUs): Collect data from sensors, analogue signals, digital signals and states, sending it to the supervisory system.
4. Programmable Logic Controllers (PLCs): Automated controllers for process management.
5. Communication infrastructure: Networks to transfer data between RTUs, PLCs, and supervisory systems.

In this paper, SCADA is used to connect to multiple devices using Modbus TCP protocol, which is hosted on an OPC UA server. Each park has a router and a Modbus TCP connection with the OPC UA server. Ignition, by inductive automation, which is the SCADA system chosen for this thesis, collects the data in an OPC UA server and stores it in various designated databases (MSSQL, PostgreSQL) in a simple and understandable way. Tag is most commonly composed of three parts: value, quality, and timestamp.

In addition, ignition uses other means of getting data besides OPC tags which are connected to the OPC UA server, for expression tags, memory tags, query tags, derived tags, and reference tags, which enable running event base controls over a project/device, like getting the last 10 minute production from an inverter or having the average performance ratio of parks [Ind24c].

The ERD (Entity relationship diagram) of the database is the same no matter the variance in the different databases. In Figure 4, two tables are missing, called alarm events and alarm events data [Ind24a].

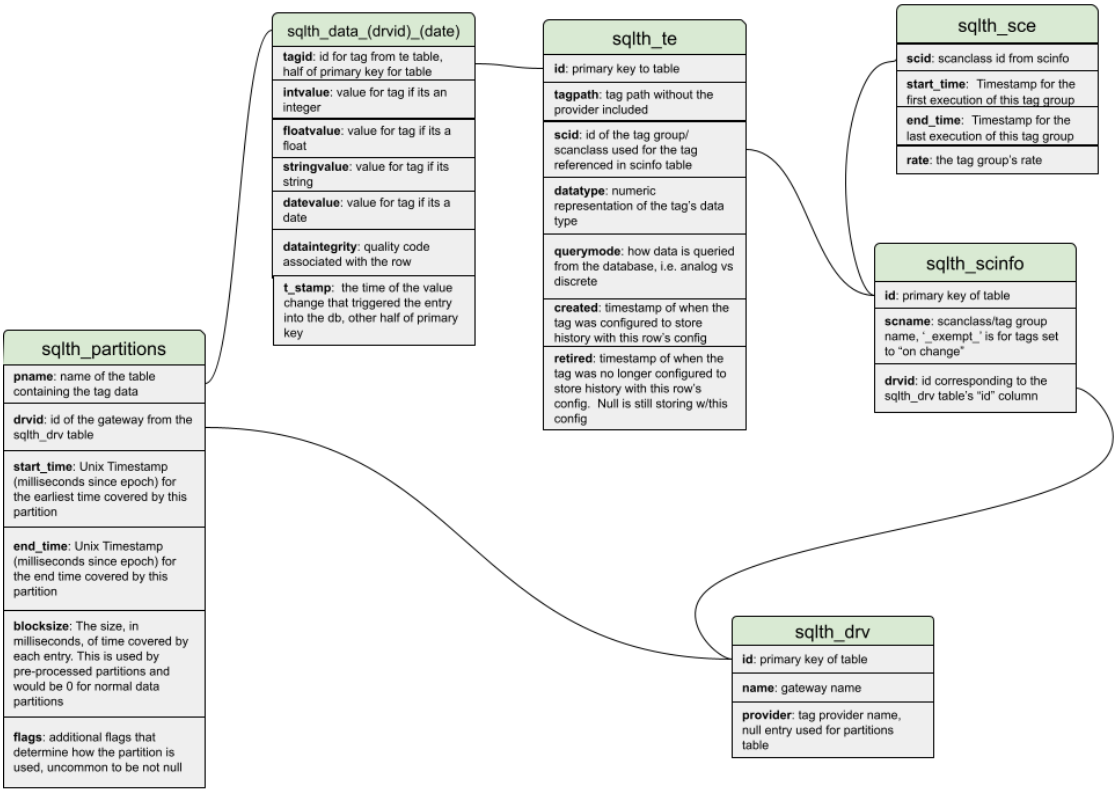



Figure 4. Base ERD without alarm journal

## 2.4 Important device data

In this section we will take a look at the most popular devices provided in the subsection and introduce the most important metrics they provide. The most common devices where data is gathered are: inverters, dataloggers and irradiance sensors

### 2.4.1 Inverters

Inverters are the most needed devices, after PV panels, which also generate most of the data. In SCADA, we are capturing 73 raw data points that are saved to the database, ranging from current data from each panel array, control settings, limitations, and fault codes all the way up to lifetime production. Additionally, in SCADA, we have expression, query, and memory tags that will already give extra info, like what has been the production of the inverter in the last 10 minutes and whether a critical alarm is active that will impact production.

However, in this thesis, we will only move the raw data points to the data mart, also called OPC tag, because the expressions, query, or memory tags could have human side input mistakes in them as the data mart needs to have a single source of truth (SSO) as raw data.

The most important metrics from this device are P, Q, F, P_DC, lifetime_production, and each phase of U and I.

### 2.4.2 Dataloggers

Dataloggers are devices that play the role of data concentrators that aggregate data together from the inverters that are connected to them. The dataloggers have 19 raw data points, where the important metrics are: P, Q, lifetime_production, rated_p, and state.

### 2.4.3 Irradiance sensors

An irradiance sensor measures the radiation level at a single point. If the park is very large, more devices are needed to ensure the measurement results are representative of the park. Each panel has a given production level it must produce at a radiation level, which is in W/m². The irradiance sensors only give us two output values: irradiance and sensor temperature.

# 3 Initial dataset

This section provides an overview of the initial dataset used to build the SCADA data mart. It covers the key aspects of data variety, structure, and the core tables, setting the stage for the detailed examination in the following subsections.

## 3.1 Amount and variety of data

In the context of this paper, we have five database servers, four of them being live databases and one being an archive:

1. MSSQL 1: 600 GB yearly, a total amount of 1.2 TB
2. PostgreSQL 1: 1.5 TB yearly, a total amount of 1.4 TB
3. PostgreSQL 2: 400 GB yearly, a total amount of 400 GB
4. PostgreSQL 3: 600 GB yearly, a total amount of 300 GB
5. PostgreSQL archive: 0 GB yearly, a total amount of 1.3 TB

The schema and the architecture of every database are the same as shown in Figure 4 with an added alarm journal. The general naming logic of the databases inside the server is also the same, with being partitioned into country based databases. In bigger database servers even a single country can be partitioned itself, like in PostgreSQL 1, which has Poland partitioned into 3.

From the structure of tables, the logged ERD is a time series approach to logging the gathered values. There are two main tables that get the most amount of action which are sqlth_te and sqlt_data_(drvid)_(date).

## 3.2 Data saving logic

In the SCADA system, you have different ways of triggering a snapshot that will take the tag value and save it to the database. The logic can be seen in Figure 5. This kind of approach is perfect to capturing the most amount of info in certain events, however, if configured incorrectly it could result in a lot unnecessary data generated.

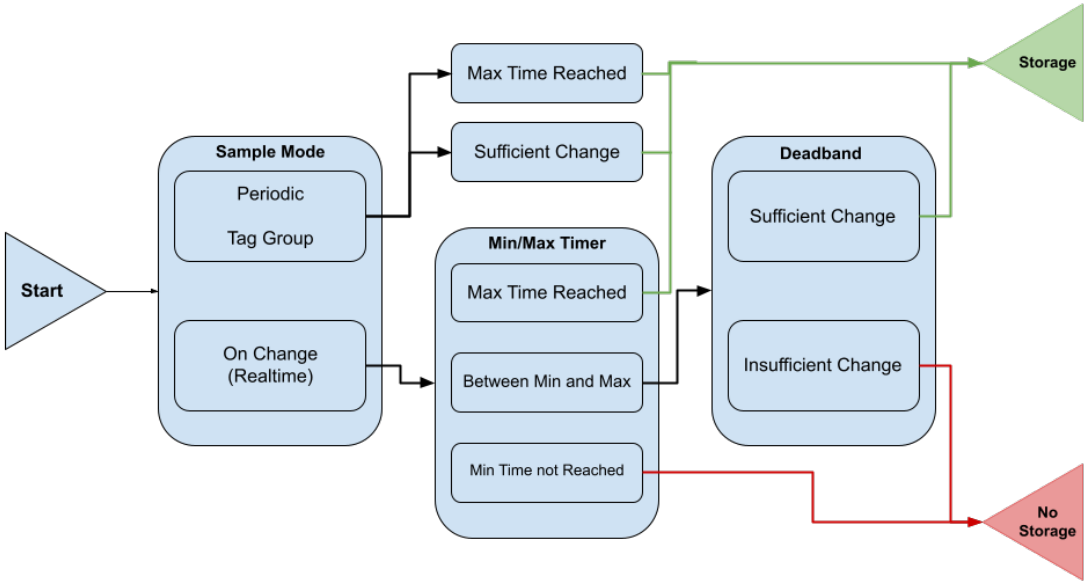



Figure 5. Ignition data storage trigger logic

## 3.3 sqlth_te table

Tables sqlth_te main purpose is to tell when the tag was created, was it retired, what is its data type, what its path is in the SCADA system, and what its tag_id is in the data table, as can be seen in Figure 6. The tag path in the SCADA system also has its logic being (country)/(park)/(device)/(metric), for example "estonia/pikkori/inverter_1/p". As we can already tell from the architecture of PV parks, then this device can consist of multiple separators, because for bigger parks, it has separators, example multiple MV substations: "poland/rzezawa_60/ST_2/inverter_1/p". The way tagpath are configured in SCADA is a semi-automatic job, each park is comprised of a template which hand fitted with the park's information, which sometimes is prone to be a hub for small mistakes, which could lead to inconsistency in the saved data:

1. Same tagpath having multiple entries: tag being retired and then being created again by getting a new row in sqlth_te_table with new tag_id.
2. Inconsistency in naming: tags are made semi-automatically which could lead to problems in naming.
3. No units/scaling: SCADA system does not log scaling or units into the database or anywhere.
4. Tags in the wrong database: As tags are country-based and are assigned a database, then some of the tags are under the wrong database.
5. Source not defined: saved tags do not say whether they came from OPC UA or any other sources. This is crucial for understanding the source of truth.

## 3.4 sqlt_data_(drvid)_(date) table

The data table provided by the SCADA system enables the saving of different data types to data tables, which are partitioned monthly. The table contains tag_id, which is defined in sqlth_te, intvalue, floatvalue, data integrity, t_stamp as can be seen in Figure 7. In the current SCADA system the data types that are being logged are only integers and floats. For a single row, there can be only one data type. The data integrity column is defined by the SCADA system, and the list of values and their definition is defined in [Ind24b]. Still, the data table has some problems, which are brought out here:

1. Inconsistent values: Some values are not correct in expression tags, which sometimes have logic changed manually.
2. Inconsistent logging times: Sometimes, the signal of the devices goes down, and they do not log the values.

| | id [PK] integer | tagpath character varying (255) | | scid integer | datatype integer | querymode integer | created bigint | retired bigint |
|---|---|---|---|---|---|---|---|---|
| 955 | 17578 | poland/ | i/datalogger/nr_grid | 2 | 0 | 0 | 1698403136297 | [null] |
| 956 | 17575 | poland/ | i/datalogger/total_energy | 2 | 1 | 3 | 1698403136297 | [null] |
| 957 | 17573 | poland/ | i/inverter_1/monthly_lost_production | 2 | 1 | 3 | 1698403130378 | [null] |
| 958 | 17572 | poland/ | i/inverter_2/monthly_lost_production | 2 | 1 | 3 | 1698403129472 | [null] |
| 959 | 17571 | poland/ | i/monthly_yield | 2 | 1 | 3 | 1698403129361 | [null] |
| 960 | 17570 | poland/ | /equivalent_power | 2 | 1 | 3 | 1698403040287 | [null] |
| 961 | 17569 | poland/ | /irradiance_sensor/irradiance | 2 | 1 | 3 | 1698403032705 | [null] |
| 962 | 17567 | poland/ | /datalogger/nr_devices | 2 | 1 | 3 | 1698403030390 | [null] |
| 963 | 17557 | poland/ | /datalogger/q | 2 | 1 | 3 | 1698403030390 | [null] |
| 964 | 17568 | poland/ | /datalogger/rated_p | 2 | 1 | 3 | 1698403030390 | [null] |
| 965 | 17558 | poland/ | /datalogger/total_energy | 2 | 1 | 3 | 1698403030390 | [null] |
| 966 | 17565 | poland/ | /datalogger/p | 2 | 1 | 3 | 1698403030390 | [null] |
| 967 | 17564 | poland/ | /datalogger/daily_energy | 2 | 1 | 3 | 1698403030390 | [null] |
| 968 | 17563 | poland/ | /datalogger/lock | 2 | 0 | 0 | 1698403030390 | [null] |
| 969 | 17560 | poland/ | /datalogger/inverter_p | 2 | 1 | 3 | 1698403030390 | [null] |
| 970 | 17561 | poland/ | /datalogger/nr_grid | 2 | 0 | 0 | 1698403030390 | [null] |
| 971 | 17562 | poland/ | /datalogger/nr_fault_devices | 2 | 1 | 3 | 1698403030390 | [null] |
| 972 | 17559 | poland/ | /datalogger/state | 2 | 0 | 0 | 1698403030390 | [null] |
| 973 | 17566 | poland/ | /datalogger/nr_off_grid | 2 | 0 | 0 | 1698403030390 | [null] |
| 974 | 17556 | poland/ | /inverter_1/monthly_lost_production | 2 | 1 | 3 | 1698403016422 | [null] |
| 975 | 17555 | poland/ | /inverter_3/daily_energy | 2 | 1 | 3 | 1698403016375 | [null] |
| 976 | 17554 | poland/ | /inverter_4/monthly_lost_production | 2 | 1 | 3 | 1698403016232 | [null] |
| 977 | 17553 | poland/ | /inverter_2/daily_energy | 2 | 1 | 3 | 1698403016086 | [null] |
| 978 | 17552 | poland/ | /inverter_3/monthly_lost_production | 2 | 1 | 3 | 1698403016044 | [null] |
| 979 | 17551 | poland/ | /inverter_2/monthly_lost_production | 2 | 1 | 3 | 1698403015311 | [null] |
| 980 | 17550 | poland/ | /average_performance_ratio | 2 | 1 | 3 | 1698403015227 | [null] |

park

Figure 6. Example of sqlth_te output

| | tagid [PK] integer | intvalue bigint | floatvalue double precision | stringvalue character varying (255) | datevalue timestamp without time zone | dataintegrity integer | t_stamp [PK] bigint |
|---|---|---|---|---|---|---|---|
| 205 | 635 | [null] | 9096 | [null] | [null] | 192 | 1722459613325 |
| 206 | 627 | [null] | 8808 | [null] | [null] | 192 | 1722459613325 |
| 207 | 16106 | 0 | [null] | [null] | [null] | 192 | 1722459615479 |
| 208 | 5089 | [null] | 15445 | [null] | [null] | 192 | 1722459614545 |
| 209 | 5214 | [null] | 8932 | [null] | [null] | 192 | 1722459610062 |
| 210 | 5088 | [null] | 15345 | [null] | [null] | 192 | 1722459614545 |
| 211 | 5212 | [null] | 9028 | [null] | [null] | 192 | 1722459610062 |
| 212 | 5202 | [null] | 15531 | [null] | [null] | 192 | 1722459610062 |
| 213 | 5229 | [null] | 8885 | [null] | [null] | 192 | 1722459613079 |
| 214 | 18512 | 6807 | [null] | [null] | [null] | 192 | 1722459617309 |
| 215 | 18531 | [null] | -2.3299999237060547 | [null] | [null] | 192 | 1722459611761 |
| 216 | 12499 | [null] | 0 | [null] | [null] | 192 | 1722459617307 |
| 217 | 18511 | 6020 | [null] | [null] | [null] | 192 | 1722459617309 |
| 218 | 18518 | 6800 | [null] | [null] | [null] | 192 | 1722459617310 |
| 219 | 5589 | [null] | 12133 | [null] | [null] | 192 | 1722459615767 |
| 220 | 5584 | [null] | 20575 | [null] | [null] | 192 | 1722459615767 |
| 221 | 5094 | [null] | 50.00899887084961 | [null] | [null] | 192 | 1722459607526 |
| 222 | 630 | [null] | 15446 | [null] | [null] | 192 | 1722459616321 |
| 223 | 15126 | [null] | -3.140000104904175 | [null] | [null] | 192 | 1722459614303 |

Figure 7. Example of the sqlt_data output

3. Problems during device startup: Values given by devices sometimes give random values upon startup.

4. Getting too big: Querying data from a month's span can take up to 3 minutes.

5. Not normalized time frames: Some log every 5-10 seconds and others from 1 to 60 minutes, as illustrated in Figure 5.

Given these shortcomings, the live database falls short of providing a reliable foundation for comprehensive data analysis tools, which must handle the variability in data effectively. To address this, a robust solution is required, one that ensures consistent data quality, uniform logging intervals, and efficient querying capabilities.

# 4 SCADA data mart

In this paper, a data mart was developed specifically for the SCADA system, forming the foundation for future tools and analysis reports. In the context of managing solar parks, the SCADA data mart is designed to handle and transform vast amounts of data generated by the SCADA systems into actionable insights. This section will explore the necessity of the data mart, its primary objectives, and its potential future applications in the management of solar parks.

## 4.1 Purpose and needs in data mart

**Centralized Data Management**: To ensure the integrity and consistency of data analysis and decision-making, the data mart must act as the single source of truth, consolidating data from various sources, including inverters, dataloggers, and sensors, into a centralized repository. This centralization is crucial as it addresses the inconsistencies and fragmentation mentioned in section 3.3 regarding the sqlth_te table, where issues like inconsistent naming and multiple entries for the same tag path. By standardizing and cleansing the data, the data mart ensures that stakeholders rely on accurate, up-to-date information essential for making informed decisions.

Additionally, the scalability of the data mart is vital, allowing it to accommodate future expansions, such as the addition of more tables for new features. The data mart must also enforce a normalized naming convention for devices, tags, and metrics to maintain uniformity across the database. This approach resolves problems identified in the sqlt_data_(drvid)_(date), section 3.4, where inconsistent logging times and data types were problematic.

**Eliminating Excel-Based Reporting**: Transitioning from Excel-based reporting to a robust data mart will significantly enhance the reliability, efficiency, and scalability of data analysis. Excel-based systems are prone to errors due to manual data entry and complex formulas, with a limit to size. The data mart automates data collection and reporting processes, reducing the risk of errors and improving efficiency. This shift not only streamlines the reporting process but also enables more complex and accurate analyses that are not feasible with Excel.

**Reducing Database Sizes and Costs**: Optimizing database sizes is another critical focus area. By lowering the volume of stored data through effective ETL processes and data archiving strategies, the data mart can reduce storage costs and improve query performance. As highlighted in section 3.4, large data volumes can lead to significant performance issues, such as lengthy query times. Implementing a data mart will help mitigate these issues, resulting in cost savings and enhanced data processing efficiency.

**Simplifying Data Analysis**: The data mart simplifies data analysis by providing normalized time-frame data and pre-calculated metrics for pure time-series data. This approach addresses the challenges identified in section 3.4, where inconsistent logging

times hindered reliable data analysis. Pre-calculating metrics ensures that analysts have immediate access to crucial data points, facilitating quicker and more accurate insights.

**Real-Time Data Insertion**: Implementing real-time data insertions from SCADA systems, once a time-frame section is fulfilled, ensures timely decision-making based on the latest data. This capability is essential for operational efficiency, allowing managers to respond promptly to changes in data trends.

## 4.2 Solutions for pointed out problems

To address the issues identified in sections 3.4 and 3.3, the following strategies will be employed:

**Inconsistent Tag Naming and Multiple Entries:** By creating a centralized data mart with a robust ETL process, we will standardize tag naming conventions and eliminate multiple entries for the same tag path. The data mart will generate unique mart tag_ids based on park, device, and metric combinations, ensuring consistency across the database.

**No Units/Scaling Information:** The data mart will include units and scaling information for all tags, ensuring that data is accurately represented and understood. This addition addresses the lack of unit/scaling data in the current system.

**Tags in the Wrong Database:** Centralizing data in the data mart will resolve the issue of tags being assigned to the wrong database, as mentioned in section 3.3. The data mart will be designed to ensure that all tags are correctly categorized and stored.

**Source Not Defined:** Including a source column in the data mart will provide clarity on the origin of each tag, whether it is from OPC UA or other sources. This transparency is crucial for understanding the data's source of truth.

**Inconsistent Logging Times and Data Types:** By normalizing time frames and ensuring consistent logging intervals, the data mart will resolve the issues of inconsistent logging times and data types highlighted in section 3.4. Pre-calculated metrics and efficient data processing will further enhance the reliability of the data.

The proposed revisions and enhancements in this section will address the critical needs of the data mart, ensuring centralized data management, eliminating reliance on Excel, optimizing database sizes, simplifying data analysis, and enabling real-time data insertion. These improvements will resolve the data issues identified, leading to a more efficient, reliable, and scalable data mart for solar park management.

## 4.3 MVP of data mart

The aim of building the MVP of the data mart was to take all of the problems with the data and needs of a data mart to build an efficient, manageable, reliable.

The SCADA data mart MVP in Figure 9 is structured to handle vast amounts of time-series data from various solar parks. The architecture ensures data integrity, consistency, and scalability. The inclusion of metadata, importance levels, and ETL versioning

enhances the ability to manage and analyze the data effectively. This centralized structure addresses the need for a single source of truth, enabling accurate and efficient decision-making and reporting. The data mart was created in PostgreSQL.

**Tag Info (tag_info):**

- id: Unique identifier for each tag.
- created_time: Timestamp when the tag was created.
- retired_time: Timestamp when the tag was retired, if applicable.
- importance: Level of importance of the tag.
- device: Device associated with the tag.
- metric: Specific metric the tag is measuring.
- project_id: Foreign key linking to the specific project (solar park).
- unit: Unit of measurement for the tag.
- source: to understand whether this is an SSOT value.

The aim of this table was to replicate a similar solution from sqlth_te table in the initial dataset in Figure 4. Some new additions were added that will deal with the problems in the sqlth_te table:

1. Same tagpath having multiple entries: now data mart tag_ids will be created from the park, device, and metric columns, which will make it consistent over different databases and solve the problem of having multiple different tag_id rows with the same info.
2. Inconsistency in naming: as described above, now the naming inside the data mart is done from park, device, and metric.
3. No units/scaling: The units are now added in the data mart with the column "unit".
4. Tags in a wrong database: Data mart solves this issue as it will be centralized, based on data that will be logged in the data tables.
5. Source not defined: This is solved from the column "source" and will provide a good understanding of what is the single source of truth when building reports.

Additionally, some additional columns were added, like column "importance" which will give a tag importance, and based on that, it will see if it will log into 15-minute data table or ignore any other interval. This is to reduce the amount of data logged into the data mart as those will not provide meaningful actions in reporting for smaller intervals.

**Park Info (park_info):**

- project_id: Unique identifier for each project (solar park).
- project_data_start_time: Start time of data collection for the project.
- project_production_start_time: Start time of energy production.
- project_end_time: End time of the project.
- latitude: Geographic latitude of the solar park.
- longitude: Geographic longitude of the solar park.
- EIC_code: Code identifying the grid connection ID.

The aim of this table was to make a central meta info for the park itself, as these parameters will only be unique for every single project. It will help get people out of Excel and make it more reliable as an SSO. The data for this was gathered from park design documents.

**Resource Metadata (resource_meta):**

- resource_meta_id: Unique identifier for each resource metadata entry.
- capacity_PV_panels: Capacity of PV panels in kilowatts.
- capacity_PV_inverters: Capacity of inverters in kilowatts.
- grid_connection_capacity: Capacity of the grid connection in kilowatts.
- PV_panel_wattage: Wattage of individual PV panels.

The aim of this data is to expand the meta of the different projects. The reason for this table is to normalize the database, as solar projects can have the same template of a park for different locations.

**Resource to Project (resource_to_project):**

- project_id: Foreign key linking to the project.

- resource_meta_id: Foreign key linking to the resource metadata.

This is a bridge table that will help to normalize the database.

**Importance Level (importance_level):**

- id_importance: Unique identifier for importance levels.
- is_15min: Boolean indicating if the tag is recorded for 15 minutes data table.
- is_1h: Boolean indicating if the tag is recorded for a 1-hour data table.
- is_1d: Boolean indicating if the tag is recorded for a 1-day data table.
- is_1w: Boolean indicating if the tag is recorded for a 1-week data table.
- is_1m: Boolean indicating if the tag is recorded for a 1-month data table.

An importance level table was created to help reduce the amount of data in the database, as not every single metric of certain devices needs to, for example, log every 15 minutes, like voltage to ground in inverters.

**ETL Versioning (etl_versioning):**

- row: Unique identifier for each versioning entry.
- project_id: Foreign key linking to the project.
- start_time: Timestamp when the versioning started.
- retired_time: Timestamp when the versioning ended.
- version_15min: Version for 15-minute intervals.
- version_1h: Version for 1-hour intervals.
- version_1d: Version for 1-day intervals.
- version_1w: Version for 1-week intervals.
- version_1m: Version for 1-month intervals.

ETL versioning tables are meant to future proof the cleaning method, as the data is being processed, any change made to the cleaning code will appear here, with time-series version control to understand if there have been any mistakes in the cleaning process, so we know what data needs to be re-processed through a new data pipeline.

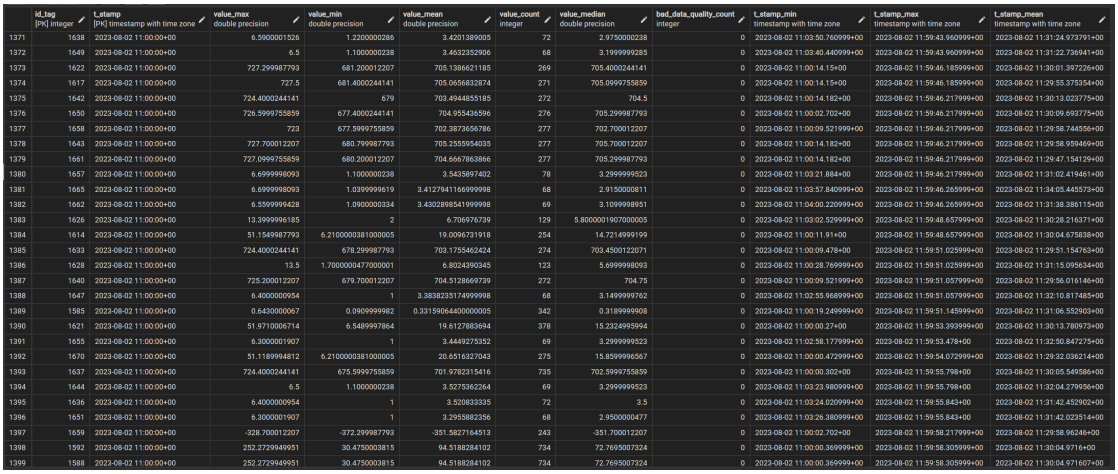
The data mart contains multiple tables to store time-series data at different intervals for a single park (15 minutes, 1 hour, 1 day, 1 week, and 1 month) and their corresponding archives. Each of these tables follows a similar structure, however, some value metrics don't appear in every data interval table, like kurtosis and skewness, and don't appear in shorter intervals.

**Data Tables:**

- id_tag: Foreign key linking to the tag.
- t_stamp: Timestamp of the recorded start_time data.
- value_max: Maximum value recorded.
- value_min: Minimum value recorded.
- value_mean: Mean value recorded.
- value_median: Median value recorded.
- bad_data_quality: Indicator of data quality issues.
- value_count: Count of values in the given interval.
- t_stamp_min: Minimum timestamp.
- t_stamp_max: Maximum timestamp.
- t_stamp_mean: Mean of the timestamps.
- value_mode: Mode value recorded.
- value_skewness: Skewness of the recorded values.
- value_kurtosis: Kurtosis of the recorded values.

The data table's main purpose was to normalize the time-series data into different interval data tables. An example is shown in Figure 8. The reasons why the different partitions were chosen are to make queries faster, prepare for the 15-minute electricity market, and enhance the development experience. Additionally, data_archive tables were introduced to make the data mart faster for querying more recent data. Data that is older than 1 year will be moved to the data archive.

In conclusion, the SCADA data mart MVP was built to address various data issues and the need for an efficient data mart. Structured in PostgreSQL, it handles vast amounts of time-series data from solar parks while ensuring data integrity, consistency,



| | id_tag [PK] integer | t_stamp [PK] timestamp with time zone | value_max double precision | value_min double precision | value_mean double precision | value_count integer | value_median double precision | bad_data_quality_count integer | t_stamp_min timestamp with time zone | t_stamp_max timestamp with time zone | t_stamp_mean timestamp with time zone |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1371 | 1638 | 2023-08-02 11:00:00+00 | 6.5900001526 | 1.2200000286 | 3.4201389005 | 72 | 2.9750000238 | 0 | 2023-08-02 11:03:50.760999+00 | 2023-08-02 11:59:43.960999+00 | 2023-08-02 11:31:24.973791+00 |
| 1372 | 1649 | 2023-08-02 11:00:00+00 | 6.5 | 1.1000000238 | 3.4632352906 | 68 | 3.1999999285 | 0 | 2023-08-02 11:03:40.440999+00 | 2023-08-02 11:59:43.960999+00 | 2023-08-02 11:31:22.736941+00 |
| 1373 | 1622 | 2023-08-02 11:00:00+00 | 727.299987793 | 681.200012207 | 705.1386621185 | 269 | 705.4000244141 | 0 | 2023-08-02 11:00:14.15+00 | 2023-08-02 11:59:46.185999+00 | 2023-08-02 11:30:01.397226+00 |
| 1374 | 1617 | 2023-08-02 11:00:00+00 | 727.5 | 681.4000244141 | 705.0656832874 | 271 | 705.0999755859 | 0 | 2023-08-02 11:00:14.15+00 | 2023-08-02 11:59:46.185999+00 | 2023-08-02 11:29:55.375354+00 |
| 1375 | 1642 | 2023-08-02 11:00:00+00 | 724.4000244141 | 679 | 703.4944855185 | 272 | 704.5 | 0 | 2023-08-02 11:00:14.182+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:30:13.023775+00 |
| 1376 | 1650 | 2023-08-02 11:00:00+00 | 726.5999755859 | 677.4000244141 | 704.955436596 | 276 | 705.299987793 | 0 | 2023-08-02 11:00:02.702+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:30:09.693775+00 |
| 1377 | 1658 | 2023-08-02 11:00:00+00 | 723 | 677.5999755859 | 702.3873656786 | 277 | 702.700012207 | 0 | 2023-08-02 11:00:09.521999+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:29:58.744556+00 |
| 1378 | 1643 | 2023-08-02 11:00:00+00 | 727.700012207 | 680.799987793 | 705.2555954035 | 277 | 705.700012207 | 0 | 2023-08-02 11:00:14.182+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:29:58.959469+00 |
| 1379 | 1661 | 2023-08-02 11:00:00+00 | 727.0999755859 | 680.200012207 | 704.6667863866 | 277 | 705.299987793 | 0 | 2023-08-02 11:00:14.182+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:29:47.154129+00 |
| 1380 | 1657 | 2023-08-02 11:00:00+00 | 6.6999998093 | 1.1000000238 | 3.5435897402 | 78 | 3.2999999523 | 0 | 2023-08-02 11:03:21.884+00 | 2023-08-02 11:59:46.217999+00 | 2023-08-02 11:31:02.419461+00 |
| 1381 | 1665 | 2023-08-02 11:00:00+00 | 6.6999998093 | 1.0399999619 | 3.4127941166999998 | 68 | 2.9150000811 | 0 | 2023-08-02 11:03:57.840999+00 | 2023-08-02 11:59:46.265999+00 | 2023-08-02 11:34:05.445573+00 |
| 1382 | 1662 | 2023-08-02 11:00:00+00 | 6.5599999428 | 1.0900000334 | 3.4302898541999998 | 69 | 3.1099998951 | 0 | 2023-08-02 11:04:00.220999+00 | 2023-08-02 11:59:46.265999+00 | 2023-08-02 11:31:38.386115+00 |
| 1383 | 1626 | 2023-08-02 11:00:00+00 | 13.3999996185 | 2 | 6.706976739 | 129 | 5.8000001907000005 | 0 | 2023-08-02 11:03:02.529999+00 | 2023-08-02 11:59:48.657999+00 | 2023-08-02 11:30:28.216371+00 |
| 1384 | 1614 | 2023-08-02 11:00:00+00 | 51.1549987793 | 6.2100000381000005 | 19.0096731918 | 254 | 14.7214999199 | 0 | 2023-08-02 11:00:11.91+00 | 2023-08-02 11:59:48.657999+00 | 2023-08-02 11:30:04.675838+00 |
| 1385 | 1633 | 2023-08-02 11:00:00+00 | 724.4000244141 | 678.299987793 | 703.1755462424 | 274 | 703.4500122071 | 0 | 2023-08-02 11:00:09.478+00 | 2023-08-02 11:59:51.025999+00 | 2023-08-02 11:29:51.154763+00 |
| 1386 | 1628 | 2023-08-02 11:00:00+00 | 13.5 | 1.7000000477000001 | 6.8024390345 | 123 | 5.6999998093 | 0 | 2023-08-02 11:00:28.769999+00 | 2023-08-02 11:59:51.025999+00 | 2023-08-02 11:31:15.095634+00 |
| 1387 | 1640 | 2023-08-02 11:00:00+00 | 725.200012207 | 679.700012207 | 704.5128669739 | 272 | 704.75 | 0 | 2023-08-02 11:00:09.521999+00 | 2023-08-02 11:59:51.057999+00 | 2023-08-02 11:29:56.016146+00 |
| 1388 | 1647 | 2023-08-02 11:00:00+00 | 6.4000000954 | 1 | 3.3838235174999998 | 68 | 3.1499999762 | 0 | 2023-08-02 11:02:55.968999+00 | 2023-08-02 11:59:51.057999+00 | 2023-08-02 11:32:10.817485+00 |
| 1389 | 1585 | 2023-08-02 11:00:00+00 | 0.6430000067 | 0.0909999982 | 0.3315906440000005 | 342 | 0.3189999908 | 0 | 2023-08-02 11:00:19.249999+00 | 2023-08-02 11:59:51.145999+00 | 2023-08-02 11:31:06.552903+00 |
| 1390 | 1621 | 2023-08-02 11:00:00+00 | 51.9710006714 | 6.5489997864 | 19.6127883694 | 378 | 15.2324995994 | 0 | 2023-08-02 11:00:00.27+00 | 2023-08-02 11:59:53.393999+00 | 2023-08-02 11:30:13.780973+00 |
| 1391 | 1655 | 2023-08-02 11:00:00+00 | 6.3000001907 | 1 | 3.4449275352 | 69 | 3.2999999523 | 0 | 2023-08-02 11:02:58.177999+00 | 2023-08-02 11:59:53.478+00 | 2023-08-02 11:32:50.847275+00 |
| 1392 | 1670 | 2023-08-02 11:00:00+00 | 51.1189994812 | 6.2100000381000005 | 20.6516327043 | 275 | 15.8599996567 | 0 | 2023-08-02 11:00:00.472999+00 | 2023-08-02 11:59:54.072999+00 | 2023-08-02 11:29:32.036214+00 |
| 1393 | 1637 | 2023-08-02 11:00:00+00 | 724.4000244141 | 675.5999755859 | 701.9782315416 | 735 | 702.5999755859 | 0 | 2023-08-02 11:00:00.302+00 | 2023-08-02 11:59:55.798+00 | 2023-08-02 11:30:05.549586+00 |
| 1394 | 1644 | 2023-08-02 11:00:00+00 | 6.5 | 1.1000000238 | 3.5275362264 | 69 | 3.2999999523 | 0 | 2023-08-02 11:03:23.980999+00 | 2023-08-02 11:59:55.798+00 | 2023-08-02 11:32:04.279956+00 |
| 1395 | 1636 | 2023-08-02 11:00:00+00 | 6.4000000954 | 1 | 3.520833335 | 72 | 3.5 | 0 | 2023-08-02 11:03:24.020999+00 | 2023-08-02 11:59:55.843+00 | 2023-08-02 11:31:42.452902+00 |
| 1396 | 1651 | 2023-08-02 11:00:00+00 | 6.3000001907 | 1 | 3.2955882356 | 68 | 2.9500000477 | 0 | 2023-08-02 11:03:26.380999+00 | 2023-08-02 11:59:55.843+00 | 2023-08-02 11:31:42.023514+00 |
| 1397 | 1659 | 2023-08-02 11:00:00+00 | -328.700012207 | -372.299987793 | -351.5827164513 | 243 | -351.700012207 | 0 | 2023-08-02 11:00:02.702+00 | 2023-08-02 11:59:58.217999+00 | 2023-08-02 11:29:58.96246+00 |
| 1398 | 1592 | 2023-08-02 11:00:00+00 | 252.2729949951 | 30.4750003815 | 94.5188284102 | 734 | 72.7695007324 | 0 | 2023-08-02 11:00:00.369999+00 | 2023-08-02 11:59:58.305999+00 | 2023-08-02 11:30:04.9716+00 |
| 1399 | 1588 | 2023-08-02 11:00:00+00 | 252.2729949951 | 30.4750003815 | 94.5188284102 | 734 | 72.7695007324 | 0 | 2023-08-02 11:00:00.369999+00 | 2023-08-02 11:59:58.305999+00 | 2023-08-02 11:30:04.971607+00 |

Figure 8. One-hour interval table output example

and scalability. By centralizing metadata, importance levels, and ETL versioning, it provides a reliable single source of truth for accurate decision-making and reporting. The data mart's architecture not only solves existing problems but also enhances future data management and analysis.

## 4.4 Evaluation

The development and implementation of the SCADA data mart MVP improved areas in managing and analyzing data generated by solar parks. The MVP effectively addresses several key challenges, including inconsistent data, the absence of unit/scaling information, and the improper categorization of tags. By centralizing and standardizing data, the data mart provides a robust platform for efficient data analysis and decision-making.

One of the critical evaluations of this project is the substantial reduction in data volume, from 4.6 TB of raw time-series data to a more manageable 500 GB in the data mart. This compression is achieved through an effective ETL (Extract, Transform, Load) process that normalizes and archives data. The normalization ensures consistency across different time frames, addressing the issue of inconsistent logging intervals and data types highlighted in the initial dataset analysis. This optimized data handling not only reduces storage costs but also significantly improves query performance, making the system more responsive and scalable.

The data mart MVP's architecture is designed to handle vast amounts of time-series data from various solar parks. It includes essential components such as tag information, park metadata, resource metadata, and data tables for different time intervals (15 minutes, 1 hour, 1 day, 1 week, and 1 month). Each of these components plays a crucial role in ensuring data integrity, consistency, and scalability. For instance, the tag information

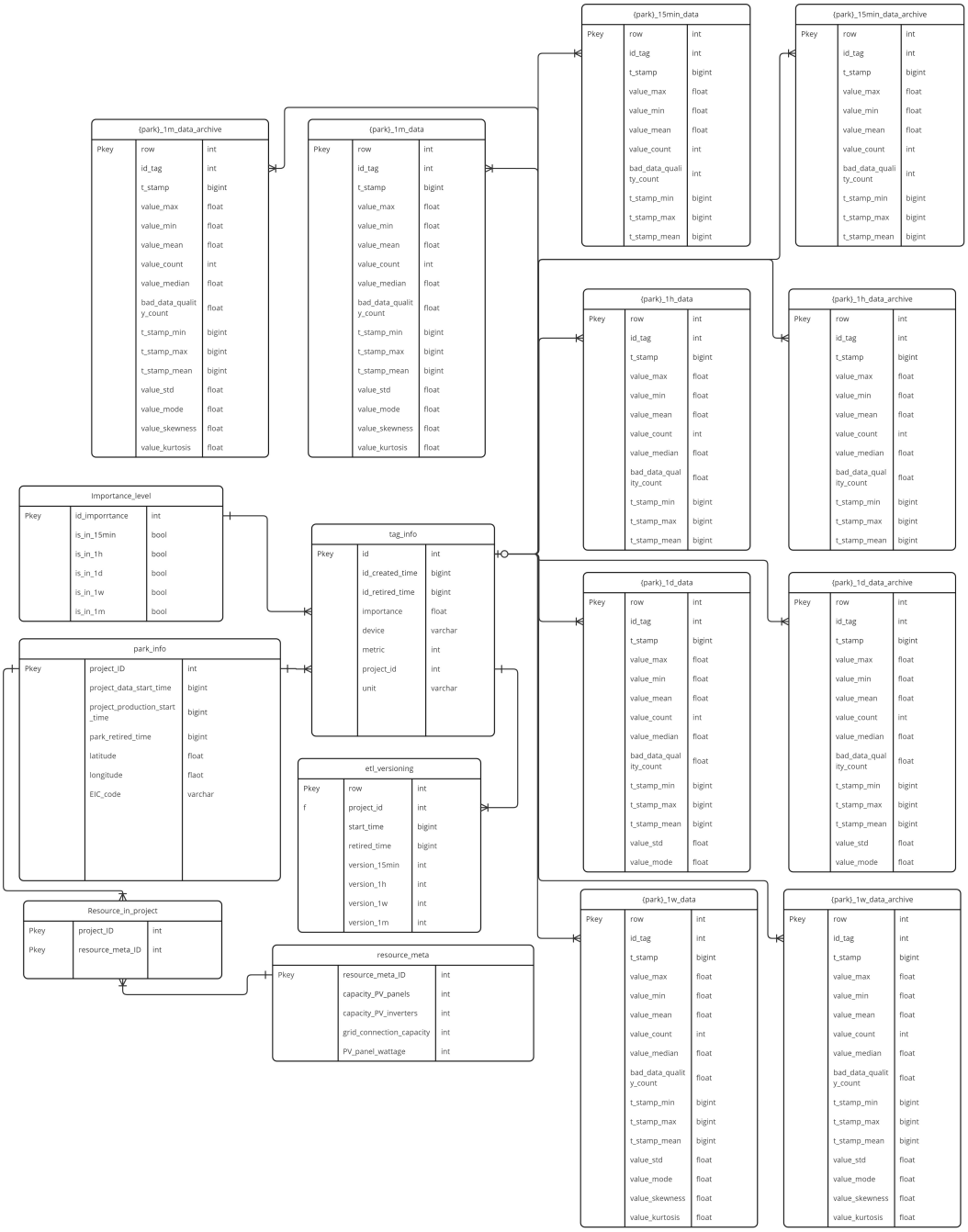



Figure 9. MVP of the data mart

table resolves issues like inconsistent naming and multiple entries for the same tag by generating unique tag_ids based on the park, device, and metric combinations. This approach ensures uniformity across the database and eliminates the problem of duplicate entries.

Moreover, the MVP includes the feature of importance levels for different metrics, allowing the system to prioritize and log data based on its significance. This feature reduces the volume of data that needs to be stored and processed, enhancing the overall efficiency of the data mart. By logging only the most critical metrics at shorter intervals, the system can focus resources on the most valuable data, ensuring that analyses and reports are both accurate and relevant.

The data mart MVP also incorporates ETL versioning to manage and track changes in data processing methods. This feature is crucial for maintaining the integrity of the data over time, as it allows for the identification and correction of any errors introduced during the ETL process. The ability to version control the ETL processes ensures that any changes can be documented and audited, providing transparency and accountability in data management.

In terms of real-time capabilities, the data mart MVP supports real-time data insertions from SCADA systems once a time-frame section is fulfilled. This capability is essential for operational efficiency, allowing managers to respond promptly to changes in data trends. The inclusion of pre-calculated metrics further enhances the reliability of the data, providing analysts with ready access to crucial insights without the need for complex calculations on the fly.

The SCADA data mart MVP demonstrates significant improvements in data management for solar parks. It addresses critical issues in data consistency, reliability, and scalability, providing a centralized platform for accurate and efficient data analysis. The successful reduction in data volume, enhanced real-time capabilities, and elimination of manual reporting errors collectively contribute to a more efficient and reliable system for managing solar park data. This project not only resolves current data management challenges but also sets the stage for future advancements in renewable energy data management.

# 5 Data pipeline

ETL processes are fundamental in building data marts, especially in contexts such as solar park management, where vast amounts of data are generated continuously. The importance of ETL in data marts can be summarized in the following points:

Data Consistency: ETL ensures that data from various sources is standardized and integrated into a single coherent format. This consistency is crucial for performing accurate analytics and generating reliable reports. Without a consistent data format, merging and comparing data from different sources would be error-prone and complex.

Data Quality: ETL processes involve cleaning and transforming raw data to remove errors, duplicates, and inconsistencies. This step is vital for maintaining high data quality, which directly impacts the insights derived from the data. Poor data quality can lead to misleading conclusions and suboptimal decisions.

Data Reliability: ETL pipelines automate the process of data collection, transformation, and loading, reducing the likelihood of human errors. Automation ensures that the data mart is regularly updated with fresh data, making the analytics and reports timely and reliable.

Accurate analytics and reporting depend heavily on the consistency, quality, and reliability of the data, which are all ensured by a well-designed ETL process.

## 5.1 Tech stack choice

Apache Airflow was chosen as the ETL tool for this project due to several key reasons:

- Open Source: Airflow is an open-source project, which makes it accessible without licensing costs. The vibrant community around Airflow ensures continuous improvements and support.
- Reliability: Airflow is a proven tool in the industry, used by many organizations to manage complex workflows. Its reliability is well-documented, and it provides robust features for error handling, retries, and monitoring.
- Ease of Prototyping: Airflow's Directed Acyclic Graph (DAG) structure allows for easy visualization and prototyping of workflows. It enables rapid development and iteration, which is crucial in a project where the ETL process may evolve over time.
- Scalability: Airflow supports parallel execution and distributed task management using Celery and Redis, which is essential for handling large volumes of data efficiently.
- Popular: Aiflow is considered one of the most popular ETL tools to build upon in the data community.

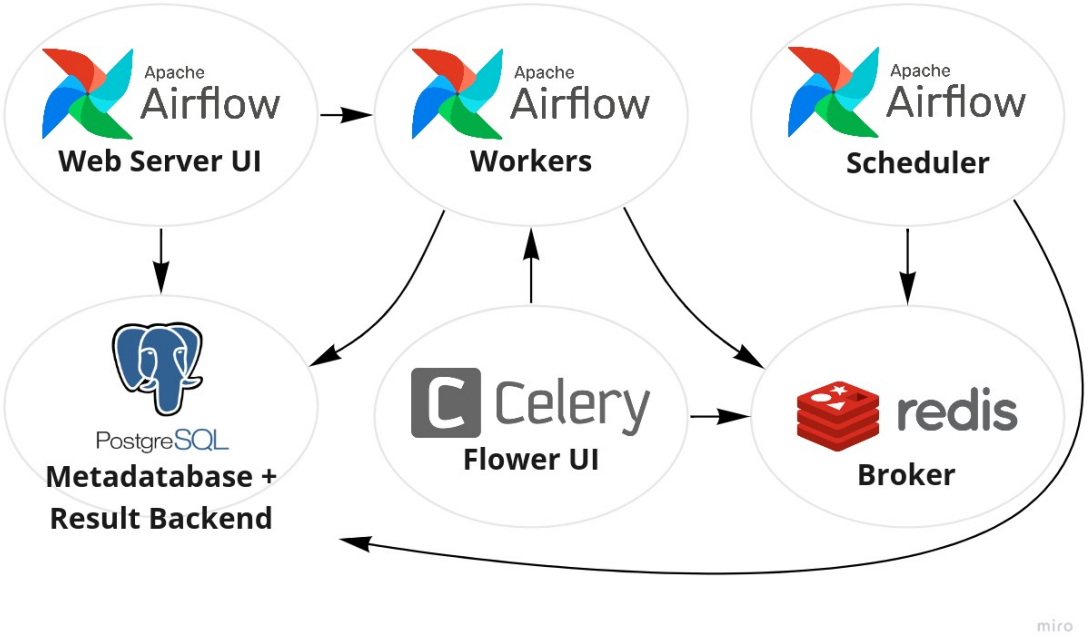



Figure 10. Chosen Airflow architecture

- Connectivity: Many tools and frameworks have made the basic connections to Airflow, which makes the ETL workflow better for faster results.

Additionally, in this paper, the Airflow architecture that was chosen for this thesis was to run Airflow in Docker with separated services with enhancement of Redis and Celery to provide scalability for this project, which documentation can be seen here with the setup, as can be seen in Figure 10. All of the data manipulation in DAGs is done in python and its libraries for its simplicity and ease of maintenance.

## 5.2 ETL workflow

ETL in this thesis consists of different stages: extraction of data, transforming, and loading. These stages were all done for different time intervals as shown in the data mart, the reasoning for using different intervals is the following:

- 15 minutes - As the electricity market is transitioning into 15-minute intervals for trading, this is the minimum time frame to understand what is happening in the parks to evaluate performances, as shown in Figure 11. There will be only the most important metrics that will be converted into 15 minute intervals.

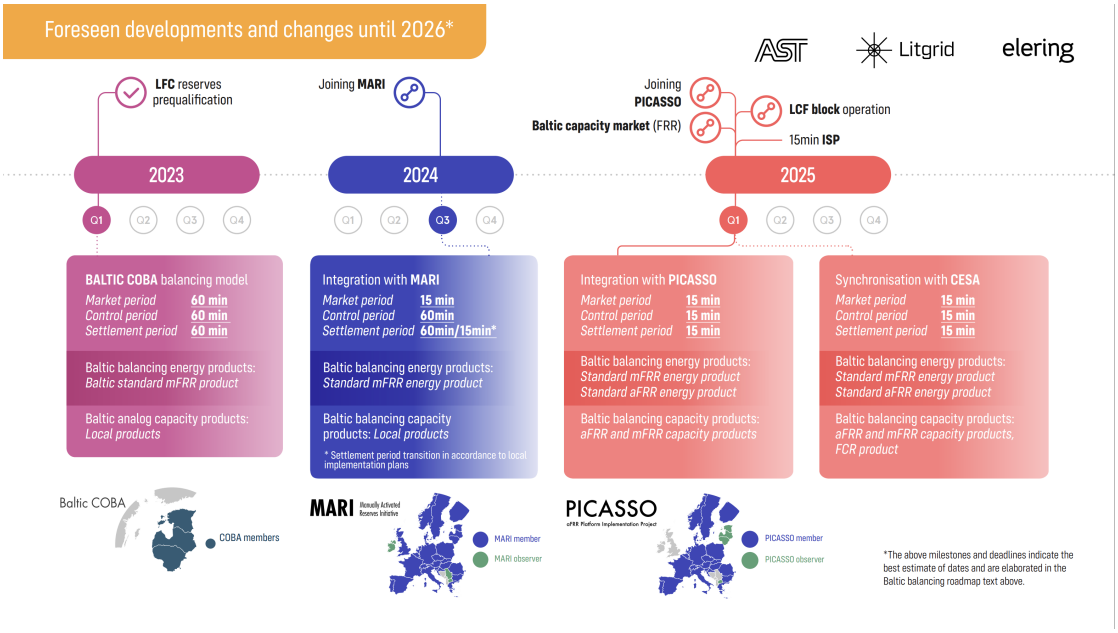



Figure 11. Development on electricity market

- 1 hour - An hour is the base of most energy-specific units, so this will serve as the core of analysis and will have the most amount of tags in this time frame.
- 1 day - The main focus is to classify how well certain parks have performed against the provided weather.
- 1 week - Give a broader overview of the performances of different parks to identify unwanted patterns.
- 1 month - Gives a good overview of how successful has the month been for accounting and other clients.

### 5.2.1 Data extraction

In the data extraction part, there are two databases that are being queried. The main ones are the data itself which will be queried for the data interval against the SCADA database table sqlt_data_drvid_year_month. Afterwards, there will be 3 other queries to prepare for the cleaning and filtering in the transform stage. One against the sqlth_te in the SCADA system to get the meanings behind the id in the data table, mainly the "tagpath" column which will be split into "country", "park", "device" and "metric", afterwards a OPC filter is run against it to get SSOT tags. With these columns, a query will be made against the data mart "tag_info" table to get the data mart id's against those the

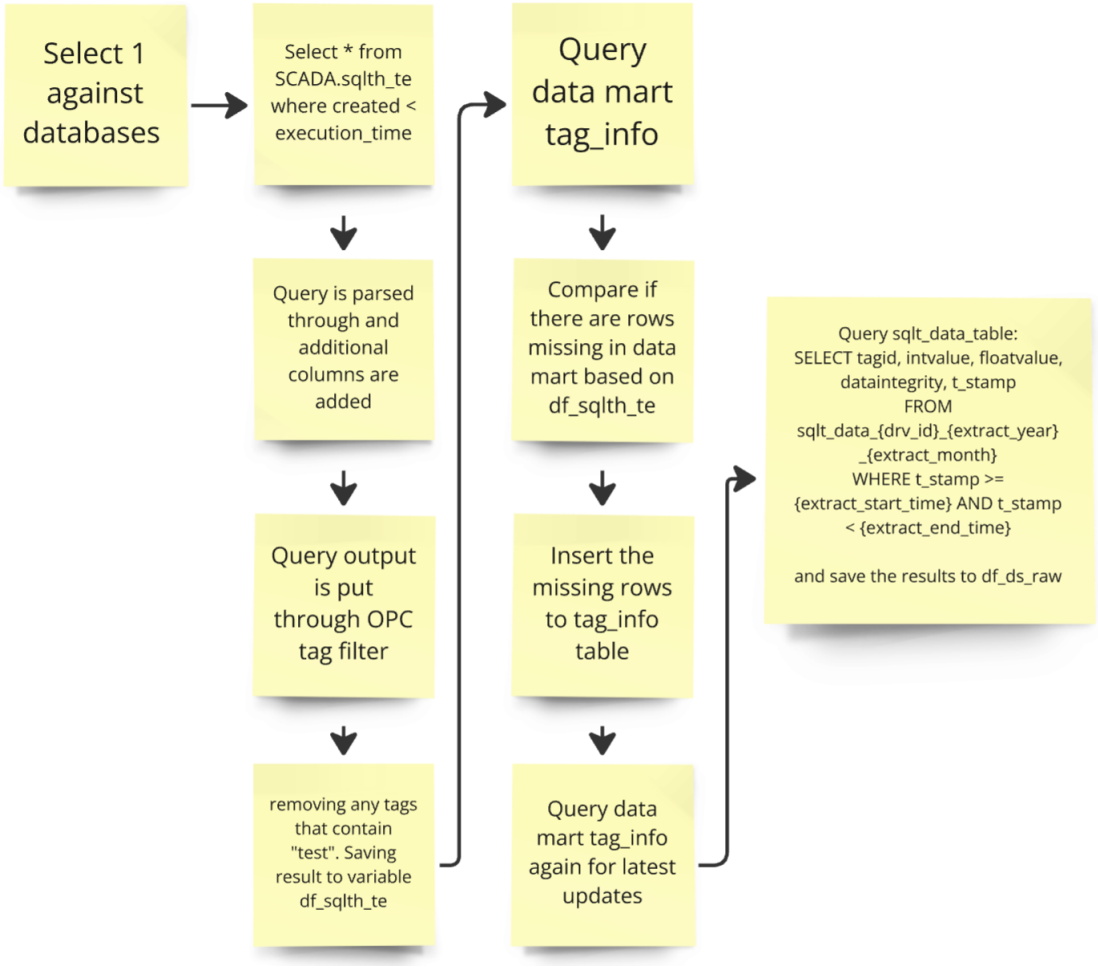



Figure 12. ETL: extraction

combination of "country", "park", "device", and "metric" if there is no combination for that an id will be created by the data mart and with the second query each of the rows in sqlth_te table will have a mart_id attached to them. This is an important process to get rid of the problem of the same tagpath having multiple entries, which was brought out in section 3.3. An example of this process is shown in Figure 12.

### 5.2.2 Data transforming

The data transforming part consists of four different stages: filtering, cleaning, and transforming. After the data gathered from the sqlt_data_drvid_year_month is being filtered for readings in the sqlth_te with mart_id and importance level. Data cleaning is divided into three distinct sections, each designed to handle different types of data.

The first category deals with data that consistently increases over time, such as the

tag total_yield. This type of data will be controlled and is always getting bigger; every point of measurement will compare that it is bigger than the last value, and when the value is not bigger, it will be dropped. However, for extra protection it will also control that the value is not too high comparing it to the last value of the given time interval, else it will be dropped.

The second type is the data that will not be cleaned. The reasoning for this is that this data is too chaotic, and any type of cleaning would mean a lot of actual data being deleted, for example, array_resistance.

The third type is the rest, which will go through a z-score filter of 3, which is calculated for each mart_id separately. With this, we can assure that measurements that are moving against a certain pattern against weather will be cleaned.

After the cleaning, the data is all added together and aggregated based on the time interval, calculating the different statistical metrics as pointed out in the data mart section. Control will be done on the same data if there hasn't been a single point of measurement for data that was gotten from the data mart according to its importance level. Then a row is added to fill the gap, with zeros and value_count set to -1. An example of this process is shown in Figure 13.

### 5.2.3 Data Loading

In this part, the final objective is to load the data obtained after the transformation phase. To achieve this, a control mechanism must be added which will go through a search for the tables in the data mart for each specific park. If there is not one, then a table will be created. Afterwards, the data will be inserted into the tables, based on the park's name. An example of this process is shown in Figure 14.

### 5.2.4 Extras

There are a couple of extras added to the process, for example, connectivity testing to database tables, filtering out unwanted sources which are not relevant parks, ETL version management and some logging info to help debug.

## 5.3 Evaluation

The data pipeline developed in this thesis provides an efficient ETL process to manage the extensive data generated by solar parks. This pipeline ensures that time-series data is collected, transformed, and loaded into the SCADA data mart in a structured format suitable for future analysis.

The ETL workflow begins with data extraction, where Apache Airflow orchestrates the collection of data from various SCADA systems. This phase addresses issues related to inconsistent logging intervals and data types, ensuring consistent data gathering.

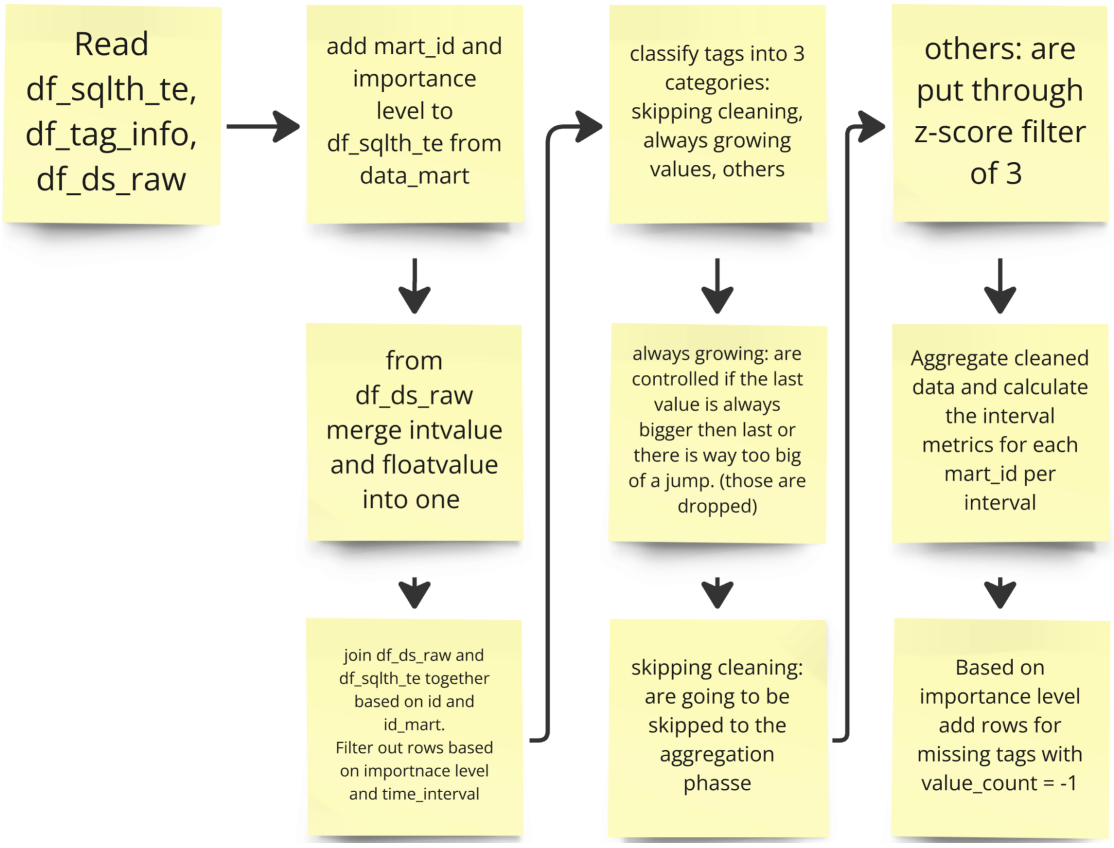



Figure 13. ETL: transform

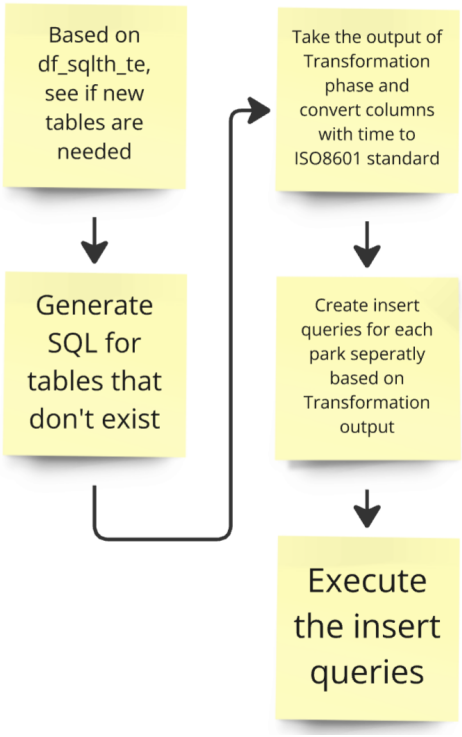



Figure 14. ETL: load

Airflow's capabilities for error handling and retries enhance the reliability and scalability of this process.

During the data transformation phase, the pipeline filters, cleans, and categorizes data based on its importance. Different techniques are used for filtering, with the most used one being such as z-score filtering for anomaly detection and outlier removal to ensure high data quality by eliminating errors and inconsistencies. This approach is crucial for maintaining the integrity of the analytical outputs from the data mart.

In the data loading phase, the cleaned and transformed data is stored in the SCADA data mart, designed to handle large volumes of time-series data across multiple intervals (15 minutes, 1 hour, 1 day, 1 week, and 1 month). This structure facilitates efficient querying and quick access to insights, aiding in informed decision-making.

A significant achievement of this ETL pipeline is the reduction of data volume from 4.6 TB of raw time-series data to 500 GB. This reduction lowers storage costs and improves query performance, making the data mart more responsive and scalable. Real-time data insertions further enhance operational efficiency, enabling timely decision-making based on the latest data.

The process was designed to handle vast amounts of data, generated continuously, starting from 2019. Despite the challenges and the extended duration required to fine-tune and debug the process, the final outcome was highly successful, resulting in the efficient reduction of 4.6 TB of data into 500 GB. The main challenge of this was debugging the data from 2019 until 2024, with each fail resulting in debugging. For example as 1 hour time interval pipeline against 1 database takes about 30 seconds to do all the tasks, it took extremely long to process all that data about 300 hours, but with airflow workers running parallelly with celery and Redis this process was cut down to 55 hours.

# 6 Usage of Data Mart

This thesis focused on building a data mart for solar assets, with the aim of improving how solar assets data is managed and controlled. In this section, we will be going into the possibilities of how the current data mart can be used and the first feature that has been built on top of it.

## 6.1 Possibilities of using data mart

The usages of data mart are separated into two categories: internal and external. Internal, which means that it can be done without relying on third-party and external, being reliant on other companies.

### 6.1.1 Internal

Internal in this context means that usage will be staying inside the data mart. This can come in various ways which will improve the overall system

- Filling essential data gaps from parks, which usually happens due to communication issues with statistics and physics. This kind of feature would help us understand more clearly that everything worked accordingly after it came back online.
- Creating an asset management model which would try to predict unwanted patterns, based on the history of the park and bring those to attention.
- Introducing asset tracking expansion, which will introduce the device tracking by serial number, metadata bout the devices and relevant documents with devices.
- Adding the possibility to link all the design documents to the parks to increase the effectiveness of analyzing parks, like performance against the PV park model analysis.
- Integrating alarm data to the data mart.
- Develop and integrate business interruption style calculations and reporting mechanisms to assess and document the financial impact of any interruptions in business operations.
- Implementing a system for verifying invoices related to accounting and trading.
- Connecting the data mart with other internal data marts like finance, project pipeline and accounting thanks to project ID in the column park_info table in Figure 9 and creating a single company data warehouse.

### 6.1.2 External

- Connecting the data with TSO (Transmission system operator), like Estonian TSO data warehouse to get extra insight from the data.
- Introducing an external API for weather info to get more location info about the weather in the parks to evaluate performance.
- Streaming data into customers' databases.

## 6.2 Demonstration of data mart capabilities

In this section, the implementation of a single usage from the previous subsection will be presented. For this thesis filling essential data gaps was chosen as a feature, with reasons brought out in the next two subsubsections.

### 6.2.1 Accurate Performance Assessment:

The reliability of solar park performance assessments is dependent on the completeness and accuracy of production data. Missing data, often a result of communication outages, can obscure true performance metrics, making it difficult to understand whether a park is underperforming or simply experiencing connectivity issues. Filling these gaps allows for:

- Normal Operation Verification: By reconstructing missing data, we can confirm if the solar park continued to perform normally despite communication issues. This verification is essential for maintaining accurate performance records.
- Benchmarking: Estimating the missing production data using irradiation API data and or nearby parks' performance enables effective benchmarking. This comparison helps identify whether a park is underperforming relative to its peers under similar conditions.

### 6.2.2 Maintenance and Troubleshooting:

Comprehensive data is critical for identifying and addressing maintenance issues promptly. Missing production data can mask underlying problems, delaying necessary interventions. Accurate data filling supports:

- Identification of Faulty Devices: Complete datasets make it easier to detect anomalies and pinpoint specific devices or components that may be underperforming.
- Proactive Maintenance: Historical data trends facilitate predictive maintenance strategies, allowing for the anticipation and prevention of potential failures before they occur, leading to minimizing downtime and repair costs.

### 6.2.3 Methods

There are multiple ways of choosing the method of filling data gaps, the most common way in data world is to take the last reading before the gap and the first new value and calculating the average of them. However, in the context of the energy domain, this is not the best way.

Instead, two different approaches are introduced and the combination of both will be the solution.

Firstly, we will introduce a method that will take the closest parks production during missing hours and then replace the missing data. For this to trigger there needs to happen 3 things.

1. There needs to be a park close to the problematic park with the same panel azimuth and angle.

2. The production data of the closest park needs to be calculated into weights based on the metadata provided by the data mart.

3. The problematic park needs to come back online to understand how much the lifetime production has changed, when being offline.

To define the closest park we will introduce a radius of 30 km, if there is not park close enough then we will choose the other option only to fill out the data gaps. The second part is to calculate the weights of the park architecture differences to converter the production of closest park to the problematic park. To calculate this, the following formula will be applied:

$$\text{missing\_hourly\_production} = \frac{\text{closest\_parks\_hourly\_production} \times \text{problematic\_parks\_panel\_MWH}}{\text{closest\_parks\_panel\_MWH}} \tag{1}$$

Additionally to this formula, an output limit is added, which is the park's maximum output, so if the calculated output is larger than the park's maximum output, it will be limited to that. After getting the theoretical missing production info of the park and the park has come back online, we can determine if there were some technical difficulties based on the big difference (more than 25%) presented by the sum of theoretical production and the lifetime production change notification will be sent out.

The second approach is more focused on a weather API that will return the irradiance info for that specific location against that park's metadata. For this we will get the hourly irradiance from the API against park's location, panels azimuth and angle, afterwards a formula is applied:

$$\text{missing\_hourly\_production} = \min\left(\left(\frac{\text{irradiance}}{1000} \times 0.98 \times 0.89\right) \times \text{panel\_MW},\ \text{max\_hourly\_MWH}\right) \tag{2}$$

From the equation, we can see that there are 2 numbers, 0.98, which represents the average efficiency of converting DC into AC and 0.89, which accounts for all the park-specific losses like shading, which comes from modeling PV parks. This number of 0.89 is usually the park average, after modeling.

Also, for this approach, a notification will be sent out if there is more than a 25% difference between the sum of missing production and lifetime production change.

After both of the methods are applied the closer one to lifetime production change is chosen and added to the data mart.

## 6.3 Evaluation

The SCADA data mart developed in this thesis provides valuable internal and external applications that enhance solar park data management and analysis.

Internally, the data mart addresses key needs such as filling data gaps caused by communication issues, ensuring a complete and accurate dataset. By using statistical methods and physics-based models, it reconstructs missing data, enabling accurate performance assessments. This capability helps verify normal operations and supports predictive maintenance by identifying potential issues early.

The data mart also enhances asset management by tracking assets, storing metadata, and linking relevant documents. Integrating alarm data improves monitoring, while features like business interruption calculations and invoice verification add financial oversight and transparency.

Additionally, the data mart connects with other internal data systems, helping create a unified company data warehouse. This integration supports comprehensive business analysis by ensuring data coherence across different domains.

Externally, the data mart links with external systems, such as the TSO data warehouse, and incorporates weather data APIs to provide more detailed performance insights. Streaming data into customer databases allows external stakeholders direct access to the data, supporting broader applications and collaboration.

A key demonstration of the data mart focused on reconstructing missing data due to communication issues. This successfully showed how the data mart ensures reliable performance assessments even during data outages, highlighting its effectiveness in improving solar park performance monitoring.

The SCADA data mart offers comprehensive solutions for managing solar park data. It enhances data accuracy, supports predictive maintenance, and integrates seamlessly with other systems, improving the overall efficiency and reliability of solar park operations.

# 7 Discussion and future work

The development of the SCADA data mart represents a significant push in the management of PV park data, several areas remain for future improvement and enhancement. These improvements could further optimize the performance, scalability, and utility of the data mart, however the foundation is laid.

**Optimization of the ETL:** The current ETL pipeline, though functional, has potential for optimization to enhance its speed and efficiency. By refining the data extraction and transformation steps, the overall process can be made faster/ optimized, reducing the time required for data processing and allowing for real-time data updates.

**Migration to TimescaleDB:** The existing SCADA database is built on standard PostgreSQL and MSSQL platforms. Transitioning to a TimescaleDB-based architecture would leverage its capabilities for time-series data, which is more aligned with the data SCADA systems provide data. TimescaleDB offers better performance for time-series data, such as faster query times and more efficient storage, making it a more suitable choice for handling the large volumes of data generated by renewable parks.

**Enhanced schema for the data mart:** To improve the robustness and flexibility of the data mart, a more refined schema could be introduced. This would include methods for handling expression tags and confirming the inclusion of relevant memory tags from the SCADA system. This would need introduction of standardized processes for validating and integrating these tags to ensure more consistent and accurate data representation in the data mart.

**Implementation of data governance:** As the data mart continues to evolve, implementing a comprehensive data governance framework will be critical. This framework should include policies for data quality management, access controls, and audit trails to ensure the integrity, security, and compliance of the data stored within the mart. Effective data governance will also facilitate better decision-making by providing stakeholders with reliable and trustworthy data.

**Support for additional renewable energy sources:** Expanding the data mart to accommodate other renewable energy sources, such as wind parks, battery energy storage systems, and hybrid renewable energy systems, can be achieved primarily through enhancements at the metadata level. By updating the metadata structure to include attributes relevant to these additional energy sources, the data mart can seamlessly integrate and manage a broader range of renewable energy data.

**Exploration of additional capabilities**: The potential uses of the data mart, as discussed in Sections 6.1.1 and 6.1.2, present numerous opportunities for expansion. These include internal applications such as filling data gaps, creating predictive maintenance models, connecting to the companies' data warehouse, and integrating asset tracking and alarm data. Additionally, external applications like connecting with (TSO) data warehouses and incorporating external APIs for weather information could greatly enhance the utility and scope of the data mart.

Future work could focus on these areas to further advance the SCADA data mart's capabilities, making it a more powerful tool for managing and optimizing the performance of solar parks. By addressing these potential improvements, the data mart could evolve into a more comprehensive, efficient, and user-friendly system that meets the growing demands of the renewable energy sector.

# 8 Conclusion

In this thesis, we developed a comprehensive data mart tailored to the specific needs of PV parks. The approach addressed the primary challenges associated with managing and analyzing the large volumes of data generated by these renewable energy sources. By transforming the raw SCADA data into an understandable, easy-to-read and structured data mart through the implementation of a data pipeline, we have improved the performance, integrity, data quality and management of solar parks.

The implementation of the SCADA data mart provided a centralized, consistent and scalable solution for data management. The data mart did not only consist of a single conversion of data, but a data pipeline that will continue to provide consistency, reliability and high quality in data by converting 4.6 TB of time-series data into 500 GB data mart. This ETL process focused on handling time-series data into normalized time intervals from 15 minutes to 1 month per tag and its importance in supporting various analytical needs.

The demonstration of the data mart feature focuses on filling the gaps in production data from communication issues, which highlighted the potential for accurate performance assessments during outages. This feature confirmed that the data mart will significantly enhance the operational efficiency of solar park management.

The results of this thesis indicate that the proposed data mart and pipeline can serve as a foundation for future advancements in renewable energy data management. Future work may include integrating additional data sources, such as device-specific metainfo and other renewable energy assets, expanding the analytical capabilities, and developing more advanced predictive models to improve the controllability of PV parks.

This thesis contributes to the field of renewable energy by providing a scalable, reliable and integratable solution for managing and analyzing PV park data leading to more sustainable, optimized and controllable energy production with data value optimization.

# References


[AASA+24] Maryam Nooman AlMallahi, Yaser Al Swailmeen, Mohammad Ali Abdelkareem, Abdul Ghani Olabi, and Mahmoud Elgendi. A path to sustainable development goals: A case study on the thirteen largest photovoltaic power plants. *Energy Conversion and Management: X*, 22:100553, 2024.

[ASMA20] Razin Ahmed, Victor Sreeram, Yateendra Mishra, and MD Arif. A review and evaluation of the state-of-the-art in pv solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, 124:109792, 2020.

[Ele24] Elering. Synchronization with continental europe, 2024. Accessed: 2024-08-06.

[Fet20] Constanze Fetting. The european green deal. *ESDN Report, December*, 2(9), 2020.

[GDC+16] Swaroop S Guggilam, Emiliano Dall'Anese, Yu Christine Chen, Sairaj V Dhople, and Georgios B Giannakis. Scalable optimization methods for distribution networks with high pv integration. *IEEE Transactions on Smart Grid*, 7(4):2061–2070, 2016.

[HMB18] Lion Hirth, Jonathan Mühlenpfordt, and Marisa Bulkeley. The entso-e transparency platform–a review of europe's most ambitious electricity data platform. *Applied energy*, 225:1054–1067, 2018.

[Ind24a] Inductive Automation. Ignition database table reference - ignition user manual 8.1, 2024. Accessed: 2024-08-04.

[Ind24b] Inductive Automation. Quality codes and overlays - ignition user manual 8.1, 2024. Accessed: 2024-08-04.

[Ind24c] Inductive Automation. Types of tags - ignition user manual 8.1, 2024. Accessed: 2024-08-04.

[MA22] Ali OM Maka and Jamal M Alabid. Solar energy technology and its roles in sustainable development. *Clean Energy*, 6(3):476–483, 2022.

[OAA+21] Khaled Obaideen, Maryam Nooman AlMallahi, Abdul Hai Alami, Mohamad Ramadan, Mohammad Ali Abdelkareem, Nabila Shehata, and AG Olabi. On the contribution of solar energy to sustainable developments goals: Case study on mohammed bin rashid al maktoum solar park. *International Journal of Thermofluids*, 12:100123, 2021.

[Tat22] Srdan Tatomir. *Estonia's climate policy: challenges and opportunities*. 2022.

# Appendix

# I. Usage of ChatGPT language model in academic writing

In the course of our research, we employed a natural language processing model, namely ChatGPT, to facilitate the academic writing process for this thesis. Specifically, ChatGPT was used to validate novel ideas and suggest alternative phrasings for sentences and paragraphs.

# II. Licence

## Non-exclusive licence to reproduce thesis and make thesis public

I, **Siim Suitslepp**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   **Data mart for photovoltaic parks SCADA systems**,

   supervised by Taavi Sarnet and Mozhgan Pourmoradnasseri.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Siim Suitslepp
***11/08/2024***