

Master’s Degree

in

Data Science

[Title]

Supervisor Candidate

Alberto Montresor Giorgia Villa

Academic Year

2021/2022

Table of Contents

[Acknowledgments 2](#_Toc113308570)

[1. Introduction 2](#_Toc113308571)

[2. Data Overview 2](#_Toc113308572)

[3. Data Integration 3](#_Toc113308573)

[3.1 Employed technologies 4](#_Toc113308574)

[3.1.1 DBeaver – PostgreSQL 4](#_Toc113308575)

[3.1.2 Talend Open Studio for Data Integration 4](#_Toc113308576)

[3.2 Data storage 6](#_Toc113308577)

[3.2.1 Tables for pollutants concentration and weather condition data 7](#_Toc113308578)

[3.2.2 Tables for monitoring stations data 8](#_Toc113308579)

[3.2.3 Tables for auxiliary operations on monitoring stations data 9](#_Toc113308580)

[3.2.4 Table for log information 10](#_Toc113308581)

[3.2.5 Table for rejected rows 10](#_Toc113308582)

[3.3 Data ingestion 11](#_Toc113308583)

[3.3.1 Data cleaning and data quality 11](#_Toc113308584)

[3.3.2 Slowly Changing Dimension (SCD) 12](#_Toc113308585)

[3.3.3 ETL implementation in Talend Open Studio 14](#_Toc113308586)

[3.4 Data integrity 16](#_Toc113308587)

[4. Data Analysis and Visualization 17](#_Toc113308588)

[5. Conclusions and future work 17](#_Toc113308589)

[Appendix A 18](#_Toc113308590)

[Appendix B 20](#_Toc113308591)

[References 25](#_Toc113308592)

Acknowledgments

1. Introduction

2. Data Overview

3. Data Integration

The first part of this thesis was focused on the design and implementation of a data integration system to ingest, store, and ultimately update the data previously described.

Data integration usually refers to the challenge of establishing an architecture capable of providing the final user with a comprehensive view of data stored on heterogenous resources [3]. These resources constitute the first components of a data integration system, which then follows to construct a mapping between those sources and the final unified view. This final view acquires the name of global schema. The global schema can be then queried by the user to retrieve subsets a/o aggregated forms of the original view to answer specific user-intended questions [3]. Because of this, data integration has become a crucial aspect of management and coordination of enterprises, as it allows not only to gain actual information and analytics from raw data, but also to centralize and unify data from activities of different departments a/o business units. This lightens the job of higher-level management, as well as enhancing productivity and efficiency at the lower levels of the organisation [4].

Nevertheless, data integration comes with its own issues and challenges which can be both social and technical. From a social point of view, some data owners may be reluctant to allow their data to be used in integrated platforms, particularly if said platforms are directly managed by a third party. From a technical perspective, one of the main issues is the low interoperability between data sources and their semantic heterogeneity. Indeed, each data source may have its own protocols, schemas, and formats, making the establishment of the above-mentioned global schema a non-trivial task [5]. Finally, different data sources may employ a different naming system for the same kind of data, and inconsistencies or missing entries could arise once sources are combined. This is usually solved by applying data transformations and data cleaning procedures to the data; however, these procedures can be cumbersome to implement flawlessly, especially when working with considerable amounts of data which can hardly be inspected [3], as in the case of this thesis: the data employed in this work was mainly stored in large CSV files with a magnitude of millions of rows of data per CSV, reaching a maximum of 61.5 million of rows occupying 3.5GB in the largest CSV. As the size of the data source did not allow to inspect the data beforehand, data had to be first ingested and stored exploiting a first version of data integration system; once stored, it was then possible to investigate the data thanks to the use of queries to assess the necessity of data cleaning procedures. As it was indeed the case, data had to be ingested and stored a second time through an improved integration system which performed data cleaning at ingestion time. This underlines the fact that even though inarguably useful, data integration systems can be challenging to design and may need to be upgraded over time.

Considering the three main components of a data integration system (i.e., data sources, mapping, and global schema), in the concrete scenario of this work they can be resolved as follows: data sources can be described as an heterogeneous set of CSV files and APIs; the mapping between these and the global schema is implemented by an ETL pipeline developed with a dedicated software named Talend Open Studio, which is specific for data integration purposes; the global schema cab be regarded as the set to data tables stored on the Database Management System (DBMS) DBeaver, which functions as a rudimental data warehouse.

The following subsections include an exhaustive explanation of the design of the implemented data integration system. First, the employed tools and technologies will be described; then the data storage solutions will be presented; after a detailed description of the implementation of the ETL pipeline for data ingestion will be provided; finally, a brief reference to the protocols to enforce data integrity of the system will be made.

3.1 Employed technologies

In order to implement the data integration system, two main technologies were employed: Talend Open Studio for Data Integration and the database tool DBeaver. These were compulsory requirements provided by the contractor, as they are two widely used tools inside their organisation. More information on both technologies is provided below.

3.1.1 DBeaver – PostgreSQL

DBeaver is an open-source multiplatform database tool for both development and database administration purposes [6]. It is free of charge and supports many of the main DBMSs with JDBC or ODBC such as MySQL, Oracle, and PostgreSQL. It also supports some DBMSs which do not make use of xDBC drivers such as Redis and MongoDB. Among other features, it provides users with a graphical user interface and with the possibility to generate and export visualizations of database schemas and objects [7]. For this thesis, DBeaver was employed to work with PostgreSQL DBMS, which is an open-source relational DBMS based on the SQL language [8]. DBeaver and PostgreSQL were adopted in this work to implement the data storage portion of the data integration system.

3.1.2 Talend Open Studio for Data Integration

Talend Open Studio is an open-source software specifically designed for data integration, data management, and data quality purposes. It is a Java compiler providing the user with a graphical interface, where the user can design their own ETL pipelines through intuitive and code-free components. These components are usually connected forming a pipeline, which in Talend is named a job, functioning as a basic executable. Once a job is saved, Talend automatically compiles it, generating a Java class, which can then be integrated into other processes [9]. Figure 1 shows an example of Talend job. The icons represent the different components, while the arrows connecting them generate the flow of the pipeline. The light blue shaded areas correspond to the different subjobs the main job is composed by.

Diagram

Description automatically generated with low confidence

*Figure 1*: Example of Talend Job

Talend Open Studio was exploited to implement the main part of the data integration system, that is the data ingestion, data cleaning, and data transformation phases. For a better understanding of how these phases were developed, in Table 1 a list of the used Talend components and their functionalities have been provided [10].

|  |  |  |
| --- | --- | --- |
| Icon | Name | Description |
|  | tCreateTemporaryFile | Generates a temporary file in the desired location. |
|  | tDBClose | Closes an open connection to a database. |
|  | tDBCommit | Commits changes on an active database connection. Eventually closes the connection. |
|  | tDBConnection | Opens a connection to a database schema. |
|  | tDBInput | Exploits an open connection to load a database table row-by-row into the workspace. |
|  | tDBOutput | Exploits an open connection to perform row-by-row insert, delete, or update operations on a database table. |
|  | tDBOutputBulkExec | Reads and parses a delimited file, generates a modified file of the original, and exploits an open connection to perform a bulk insert operation on a database. |
|  | tDBRow | Executes any provided SQL query. |
|  | tExtractJSONFields | Reads fields of a Json file resulting from an API or HTTP request. |
|  | tFileDelete | Deletes file(s) in the specified location corresponding to the provided file mask. |
|  | tFileFetch | Retrieves a file from the provided ULR, saving it in the desired location. |
|  | tFileInputDelimited | Loads and parse row-by-row a delimited file. |
|  | tFileList | Iterates over all files a/o folders inside a specified location corresponding to the provided file mask. |
|  | tFileOutputDelimited | Creates or updates row-by-row a delimited file. |
|  | tFileUnarchive | Unzips the specified file in the desired location. |
|  | tFilterRow | Filters rows according to a personalised criterion. |
|  | tFixedFlowInput | Generates rows from the provided variables. |
|  | tFlowMeter | Saves flow information at runtime. |
|  | tFlowMeterCatcher | Catches and returns information stored in a tFlowMeter. |
|  | tFlowToIterate | Receives a list of files and parse them one at the time. |
|  | tHttpRequest | Sends an http request using the provided URL. |
|  | tJava | Executes any provided Java code. |
|  | tJavaFlex | Executes any provided Java code for each row in the flow. |
|  | tMap | Loads data from one or more sources, then allows to manipulate those data before sending the transformed data to one or more outputs. |
|  | tPostjob | Generates a subjob to be executed after the main job has finished. |
|  | tPrejob | Generates a subjob to be executed after the main job has started. |
|  | tRest | Sends a post, get, or delete request to the provided API. |
|  | tRunJob | When triggered, runs the specified job. Useful to run jobs in sequence. |
|  | tSetGlobalVar | Sets the name and value of variables accessible from any component. |
|  | tUniqRow | Filters between unique and duplicated rows. |

*Table 1*: Talend components

3.2 Data storage

As a first step to construct the required data integration system, a decision on data storage had to be made in order to find the most suitable solution considering both the type of data sources and the technologies available to the contractor.

In general, data storage refers to where and how information and data are recorded and retained on a digital system [11]. In also includes all relevant technologies capable of maintaining data stable over time and ensuring its easy accessibility, these being the two main goals of a data storage system [12]. Indeed, both human users and computers employ storage devices to save useful data for future interpolation, making it primary to these devices to ensure a certain level of protection against data loss and an intuitive and straightforward access to those data [13].

However, being able to provide such qualities has become more and more complicated as the volume and variety of data has increased exponentially over the last decades. Because of this trend, new challenges have arisen in the field of data storage, as traditional storage systems tend to easily become obsolete [1]. Data representation is one of them. As a matter of fact, data can now be collected from a considerable number of different sources, each possibly providing data with different schemas, data types, granularity, and semantics. This is true not only for sources yielding data with different nature, but also for those storing identical data. For example, given a sensor registering data on hourly basis, two different sources may collect those data using a similar, but not identical data type (e.g. float instead of double for decimal numbers), may adopt slightly different naming for the same data field (e.g. “Date” instead of “DateTime” for a field referring to a date object), or even change the granularity of data through the use of aggregations (e.g. aggregating hourly data to obtain daily information instead of maintaining the original hourly precision provided by the sensor) [1]. All these differences may cause issues when building a storage system capable of storing data from different sources, as discrepancies should be removed to provide a consistent interpretation and representation of data: failing to achieve this may in fact result in incorrect information on those data. This is directly connected to another challenge data storage systems are facing, that is data redundancy: when storing data from different sources, it can be the case that two sources, while collecting data on different phenomenon, may share a portion of data. Storing twice the data present in both sources adds complexity to the storage system without apporting any new information, and therefore should be avoided. However, recognizing which data is truly redundant may be cumbersome again for differences in naming, data types and aggregation level of the data itself [1]. One additional challenge is connected to the concept of data consistency. Data consistency is a desirable property of storage systems as it implies that when data is queried it is always up to date. However, when handling data that may change over a period of time or that requires later validation (as in the case of some data collected for this work), data consistency can only be achieved by a scheduled revision of the stored data against the one in the source, in order to decide which data should be inserted, deleted, or updated [1]. Finally, one last challenge is represented by the variety of data formats, that can hardly be stored in traditional relational DBMS, which can only accept structured table-like data and can only be scaled vertically by increasing the power of the used server at increasing costs. Hence the need of non-relational DBMS that can store also semi-structured and unstructured data and can be easily scaled horizontally [14].

In the concrete case of this work, it was decided to opt for the relational DBMS DBeaver. This, having verified that data from the different sources all exhibited a tabular structure and considered the requirements in terms of technologies stated by the contractor. Issues connected to data representation and consistency were also encountered. However, since those were handled at ingestion time, a more punctual explanation on these matters will be provided in the next dedicated section.

In the database, a total of 38 tables were created and populated, some to save the ingested data and some as support tables. More precisely:

* 17 tables were created to store pollutants concentration data,
* 15 tables were created to store weather conditions data,
* 2 tables were created to store monitoring stations data,
* 2 tables were created as auxiliary tables for the monitoring stations tables,
* 1 table was created to handle rejects during the ingestion stage,
* 1 table was created to log all relevant operations performed by the pipeline.

In Appendix A a summary table reporting the precise naming, number of rows and size for each of these tables and an overview of the database structure can be found.

For a more accurate insight on the structure of the tables, the following paragraphs provide the DDL of each kind of table, together with a description of the table fields and screenshot of the table view in the database.

3.2.1 Tables for pollutants concentration and weather condition data

Tables storing pollutant concentration and weather conditions present an analogous structure as they were all populated from similar CSV files. They are named *sens\_data\_####* and *weather\_data\_####* respectively, where #### corresponds to the year the data refer to. They all exhibit the following schema and fields [15]:

* **Idsensore** (INT8): unique identifier assigned to a sensor.
* **Dataora** (TIMESTAMP): date and time of value acquisition.
* **Valore** (FLOAT8): value registered by the sensor; if value is -9999, the record is missing or invalid.
* **Stato** (VARCHAR2): either VA (valid) or NA (not valid). As some entries may contain non-certain values, values can be validated after acquisition up until the end of March of the year following the registration date. Hence the need of this field.
* **Idoperatore** (INT4): either 1 (value in *valore* is the mean value registered in the timespan), 3 (value in *valore* is the maximum value registered in the timespan), or 4 (value in *valore* is the cumulative value registered in the timespan).

Figure 2 shows a screenshot of how the data appears in the database.

Table

Description automatically generated

*Figure 2*: Screenshot of table sens\_data\_1995

3.2.2 Tables for monitoring stations data

The two tables storing information on monitoring stations for pollutants and weather conditions are named *stations\_sens* and *stations\_weather* respectively. They were populated by ingesting data from two APIs with similar structures, but some naming differences that were handled at ingestion time. Also, since the tables store a history to address the issue of slowly changing dimension, but some records appeared to be corrupted, some additional operations had to be performed on both tables to rectified them. Their schemas and fields are as follows [16]:

* **Idsensore** (INT8): unique identifier assigned to a sensor.
* **Nometiposensore** (VARCHAR): kind of pollutant or weather condition measured by the sensor.
* **Unitamisura** (VARCHAR): unit of measure of the registered phenomenon.
* **Idstazione** (INT8): unique identifier assigned to a monitoring station.
* **Nomestazione** (VARCHAR): extended name assigned to a monitoring station.
* **Quota** (INT8): altitude at which the monitoring station is located.
* **Provincia** (VARCHAR): Italian province in which the monitoring station is located.
* **Comune** (VARCHAR): Italian municipality in which the monitoring station is located.
* **Storico** (VARCHAR2): either S (entry is active, i.e., represents the present state of the monitoring station) or N (entry is part of the history, i.e., represents a past state of the monitoring station).
* **Datastart** (TIMESTAMP): date and time corresponding to when the record became active.
* **Datastop** (TIMESTAMP): date and time corresponding to when the record became inactive. If it is set to 9999-12-31, the record is still active.
* **Utm\_nord** (INT8): value of the north coordinate in the Universal Transverse Mercator coordinate system.
* **Utm\_est** (INT8): value of the east coordinate in the Universal Transverse Mercator coordinate system.
* **Lat** (FLOAT8): latitude at which the monitoring station is located.
* **Lng** (FLOAT8): longitude at which the monitoring station is located.
* **Coordinate** (VARCHAR): compound string with latitude and longitude.

Figure 3 shows a screenshot of the data representation in the database.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

*Figure 3*: Screenshot of table stations\_sens

3.2.3 Tables for auxiliary operations on monitoring stations data

As stated above, tables storing monitoring stations data presented some inconsistencies in the implementation of the history and had to be rectified. To achieve this, two auxiliary tables were created to store intermediate data used in the cleaning operations of the pipeline. The two tables are named *stations\_check\_datastart* and *stations\_check\_datastop*, andrespectively collect information on the field *datastart* and *datastop* in the monitoring stations tables. They are populated at ingestion time to perform some rectifying operations. Their schema is as follows:

* **Idsensore** (INT8): unique identifier assigned to the sensor presenting inconsistencies i.e., a record missing either the *datastart* or *datastop* field values.
* **Prima\_acquisizione** / **ultima\_acquisizione** (TIMESTAMP): actual value of field *datastart* or *datastop* for the record, obtained through operations in the data integration pipeline.
* **Table\_name** (VARCHAR): either stations\_sens or stations\_weather, indicates which table the corrupted record comes from.

Figure 4 provides a view of how the tables appear in the database.

Text

Description automatically generated Text

Description automatically generated

*Figure 4*: Screenshot of table stations\_check\_datastart and stations\_check\_datastop

3.2.4 Table for log information

The table *log* was created to store information on the main operations performed by the ingestion pipeline. It is populated at ingestion time once a Talend job has ended. Its schema is the following:

* **Current\_date** (TIMESTAMP): date and time the operation was performed.
* **Table\_name** (VARCHAR): table on which the operation was performed.
* **Operation** (VARCHAR): kind of performed operation.
* **Number\_of\_rows** (INT8): number of rows interested in the operation.
* **Duration\_in\_secods** (INT8): duration of the operation in seconds.

In Figure 5 a snapshot of the first rows of said table can be found.

Graphical user interface

Description automatically generated with medium confidence

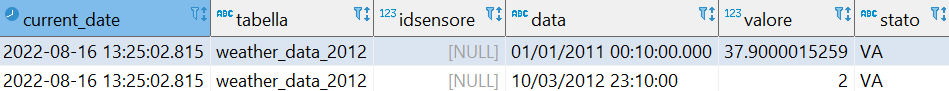
*Figure 5*: Screenshot of table logs

3.2.5 Table for rejected rows

A table *rejected* was created to handle eventual rows that could not be parsed during ingestion operations in order to monitor their fields and statutes. The schema of such table is as follows:

* **Current\_date** (TIMESTAMP): date and time the record was inserted into the table.
* **Tabella** (VARCHAR): which table the record comes from.
* **Idsensore** (INT8): *idsensore* value of the unparsable record.
* **Data** (VARCHAR): *data* value of the unparsable record.
* **Valore** (FLOAT8): *valore* value of the unparsable record.
* **Stato** (VARCHAR): *stato* value of the unparsable record.
* **Idoperatore** (INT8): *idoperatore* value of the unparsable record.
* **Errormessage** (VARCHAR): text of the error fired by Talend while parsing the record.

Figure 6 shows a screenshot of how the data appears in the database.



Graphical user interface, text

Description automatically generated

*Figure 6*: Screenshot of table rejects

3.3 Data ingestion

After having arranged the data storage, the next step of this work was to design and implement the ingestion stage of the data integration system. This was achieved by exploiting the functionalities of Talend Open Studio, to create an ETL pipeline while fulfilling the requirements of the contractor. This stage possibly corresponds to the most important part of the whole data integration system, as it is when the data storage system is populated, but also when data quality and data cleaning operations are performed.

In general, ETL processes refer to all those Extraction, Transformation, and Loading procedures aiming to: 1) identify and retrieve data from different sources; 2) clean and manipulate those data into a desired format, in order to obtain some relevant information for future analysis; 3) finally load the transformed data into a data storage system (usually a data warehouse) [17], [2]. Given the complexity and the cost of developing and maintaining ETL procedures, a proper modelling of such processes is necessary, even though no conceptual framework has been indicated as the most promising yet [17], in particular when it comes to how the ETL processes should be structured in order to grant qualities such as data integrity and data quality: the fact that ETL pipelines normally work with non-homogeneous data formats, ambiguous semantics, and data structures that may change over time, all contribute to hinder processes related to data quality and integrity. This poses some additional challenges to the already complex design of integration pipelines [18].

Issues related to data quality and integrity were indeed encountered also while designing the ETL pipeline for this work. They are briefly discussed in the first two following subsections, presenting respectively data quality concerns and the problem of Slowly Changing Dimension for data integrity. The last subsection provides instead some insights on the logic behind the implemented ETL pipeline.

3.3.1 Data cleaning and data quality

In order to summarize the different data quality issues encountered, the schematization presented in Table 1 in [18] will be taken as proxy. They delineate a categorization of quality issues according to two factors: the different stages of the ETL pipeline they can be found, and whether they relate to the schema or the instance of the data. Table 2 below, recall the structure of the proposed Table, while adapting it to the use case of this work.

|  |  |  |
| --- | --- | --- |
| **Stage** | **Problem** | **Description** |
| Extract | Embedded values | Values of different fields aggregated in one.  *Example*: lat, lng, utm\_nord and utm\_est values of monitoring stations were stored together as ‘coordinates’ in the data source. The resolution was to split the entries while ingesting. |
| Duplicate records | Repeated entries.  *Example*: duplicates were found in pollutants and weather conditions CSVs. Some of them were due to the Daylight-Saving Time change in hour and could be handled, other remained unresolved and had to be discarded. |
| Missing values | Fields appearing as NULL or empty.  *Example*: missing values in numeric fields in source CSVs are set to -9999 by the data provider. If not handled, these could be considered as valid numbers during the analysis, strongly biasing results. The resolution was to keep them in the Database, and to exclude them later from the analysis using filters. |
| Transform | Variety of data types | Fields having different data types in the data source with respect to the schema specified in the storage system.  *Example*: data ingested from APIs was saved as String type, independently of the kind of data. The resolution was to manually cast data to the correct data type during ingestion. |
| Structural conflicts | Data type (usually DateTime) has different syntax in the source and in the storage system.  *Example*: DateTime fields in source CSVs had format dd/MM/yyyy HH:mm:ss, while DBeaver default Timestamp format is yyyy-MM-dd HH:mm:ss.SSS. The resolution was to convert the Date field from the CSV to the correct DB format while ingesting. |
| Load | Wrong implementation of the SCD | History of the data can be incomplete a/o conflicted.  *Example*: in the history implemented by the monitoring stations APIs some records were missing the StartDate, while some entries were closed without EndDate. This issue is explained in more details in the next subsection. |

*Table 2*: Encountered data quality issues

3.3.2 Slowly Changing Dimension (SCD)

While for some table (usually fact tables) loading processes may only include insert operations of newly acquired data, for other tables (generally dimensions tables) records can be updated, either sporadically or in a systematic way. How to handle those updates is also part of the design decisions concerning ETL processes.

Dimensions that are expected to change over time are referred to as Slowly Changing Dimensions (SCDs). This term was first introduced by [19], who also proposed several options to deal with such matter. They all exploit the fact that only few dimensions may change, hence allowing to maintain an independent dimensional structure where only some minor modifications are needed [19]. At a higher level, two alternative solutions can be pursued: either register the change auditing for past values or replacing old values losing the history of changes [20]. At a lower level, four main types of SCDs can be delineated: Type 0, 1, 2, 3, and 4, each employing a different history storage solution [17]:

* SCD of Type 0 does not account for changes nor for history keeping: data is stored as it was first extracted and is never modified, even though source data may have changed [19].
* SCD of Type 1 accounts for changes in the data by overwriting existing values with the most recent ones [19]. This strategy has the advantage of not occupying additional storage and not needing any significant design. However, it does not allow to implement an history, as old data are merely overwritten [20].
* SCD of Type 2 accounts for both changes and history. This strategy foresees the insertion of a new row each time an update happens [20]. However, by simply adding the new row, there would be no way of knowing which one contains the updated data, as both the new and old record will share the same key. Hence, to implement this strategy three more attributes should be added to the table: StartDate, storing the date and time when the row was first inserted into the table; EndDate, storing date and time when some changes were registered for any of the attributes of the record (hence when a new row with the same key is added to the table), and a flag IsCurrent being True if the record corresponds to the most recent version of the data, or False otherwise [17]. Therefore, any current record will have a StartDate, EndDate to NULL, and IsCurrent to True. Once a new record with the same key is added to the table, first the existing record will have its EndDate set to the insertion date and its IsCurrent flag set to False, indicating that the record is now part of the history. Then the new record will be inserted into the table, having StartDate as the current date, EndDate to NULL, and flag IsCurrent to True. It should be mentioned that besides being one of the most used types of SCD, this strategy may not be scalable if dimensions are changing rapidly, as the number of rows will increase consequently [21].
* SCD of Type 3 also accounts for changes and history. However, unlike Type 2, a register is kept only for the most recent update [21]. History is implemented by adding a new column each time a dimension is changed, storing the past value for that dimension. For instance, let’s assume that Employee 1 changed their mansion from Development to Testing; then the corresponding record 1 in table Employees undergoes a change in the field *Mansion*, where the value “Development” is overwritten to “Testing”. At the same time, a new column *PreviousMansion* will be created, being NULL for any record but record 1, where it will assume value “Development”. If another employee changed their marital status, then another column *PreviousMaritalStatus* will be created, and the procedure will be repeated. Possibly a *ChangeDate* dimension could be added as well. Besides being a possible solution, it is evident how this strategy may not be scalable if several changing dimensions are present [21]: the number of columns will increase accordingly, generating a sparse table, which is quite costly to store given its several blank fields.
* SCD of Type 4 was introduced to solve the scalability issue of SCD Type 2 [21]. According to this solution, a new “Mini-Dimension” table is created for each highly changing dimension, where such dimension and its history are stored as in Type 2 [20], while records in the main table can be simply overwritten. This reduces the volume of the main table. Mini-Dimensions tables are linked to the main table through keys, which ensures the possibility to perform queries and analysis also exploiting historical data [21].

Cases of SCDs were encountered while implementing the commissioned data integration system, specifically of Type 1 and 2. A Type 1 strategy was adopted for data extracted from APIs concerning the current year data for pollutants concentration and the current month data for weather conditions. Indeed, these APIs not only store newly daily acquired data, but may also contain modified version of previous days data that have been validated (i.e., modified) by the ARPA data provider. Since data from APIs is ingested daily, then these modifications (usually on attributes *Valore* and *Stato*) were handled as a case of SCD. In particular, as data from APIs is ingested in delta and updated a/o deleted records are merely overwritten, the adopted logic was a case of SCD of Type 1.

On the other hand, the APIs storing dimensional data on monitoring stations already implements a SCD strategy of Type 2, where attribute StartDate is named Datastart, attribute EndDate is named Datastop, and flag IsCurrent is named Storico and acquires values N (True) or S (False). However, for current valid records, the EndDate field is set to the fictitious date 9999-12-31 00:00:00 instead of NULL, because of some compatibility issues in parsing NULL values between Talend Open Studio and DBeaver. Finally, even though the strategy was already implemented in the APIs, some tests were run to verify the integrity of the history and indeed some inconstancies were discovered: some records were missing the StartDate, while some others had the IsCurrent flag set to False, but EndDate still equal to NULL. Such discrepancies were handled at ingestion time, through a series of iterations and queries that will be better explained in the next section.

3.3.3 ETL implementation in Talend Open Studio

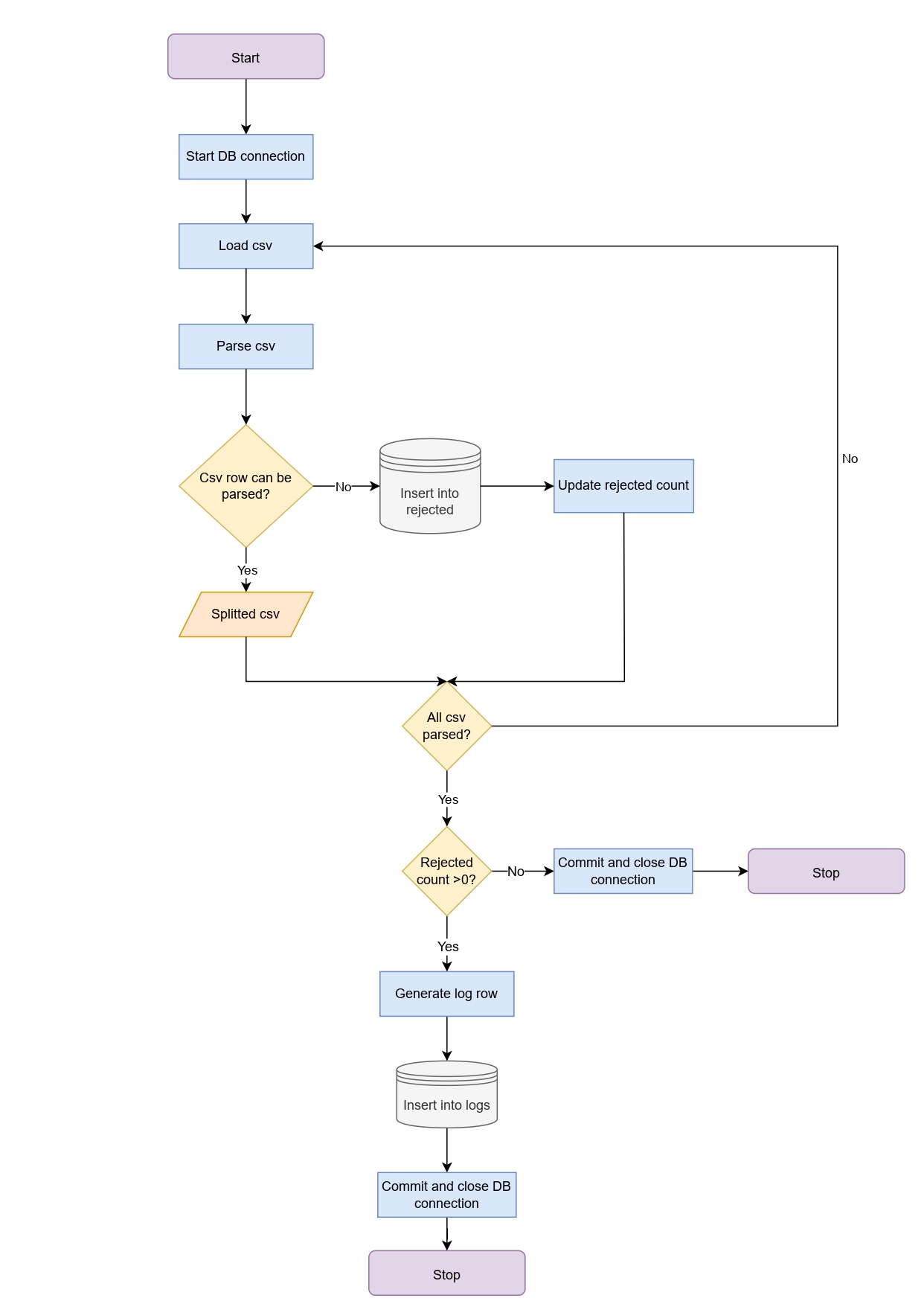
As previously stated, the ETL part of the implemented data integration system was developed using the open-source tool Talend Open Studio. The approach used can be possibly reconnected to a warehousing integration approach, as data is extracted from sources and loaded in the database, which functions as a materialized static view of the source data [5].

The pipeline is orchestrated into 10 different Talend jobs that can be categorized as:

* 3 utilities jobs: splitCSV, CombinedSplitCsv, FinalBulk;
* 2 jobs for ingesting monitoring stations data from APIs: sensorsStations\_SQLQuery, weatherStations\_SQLQuery;
* 2 jobs for ingesting historical data from CSV files for pollutants concentration and weather conditions: bulkInsertS, bulkInsertW;
* 2 jobs for ingesting current month/year weather conditions data from API: ingestion2022MonthW, ingestion2022W
* 1 job for ingesting current year pollutants data from API: ingestion2022S.

The logic and design of each of these jobs is described in detail in the following subsections.

*splitCSV*



3.4 Data integrity

What is data integrity

https://www.talend.com/resources/what-is-data-integrity/

Why do we need it here: data updates for the current year are managed by the ETL pipeline, however older CVSs could also be updated sometime later, because of how the data validation process works by the ARPA agency. Hence, there was the need to implement a simple procedure to be run yearly to check for changes in the data source and to eventually update the database tables. According to the procedure, in case some changes are detected, the interested database table is truncated, and data are re-ingested in full. Even though this may appear inconvenient as most data probably will not undergo any changes, given the specifics of Talend Open Studio and the mole of data to parse, after some tests it was decided that the most performing solution was indeed to execute an ingestion in full of the modified CSV.

Python functions descriptions

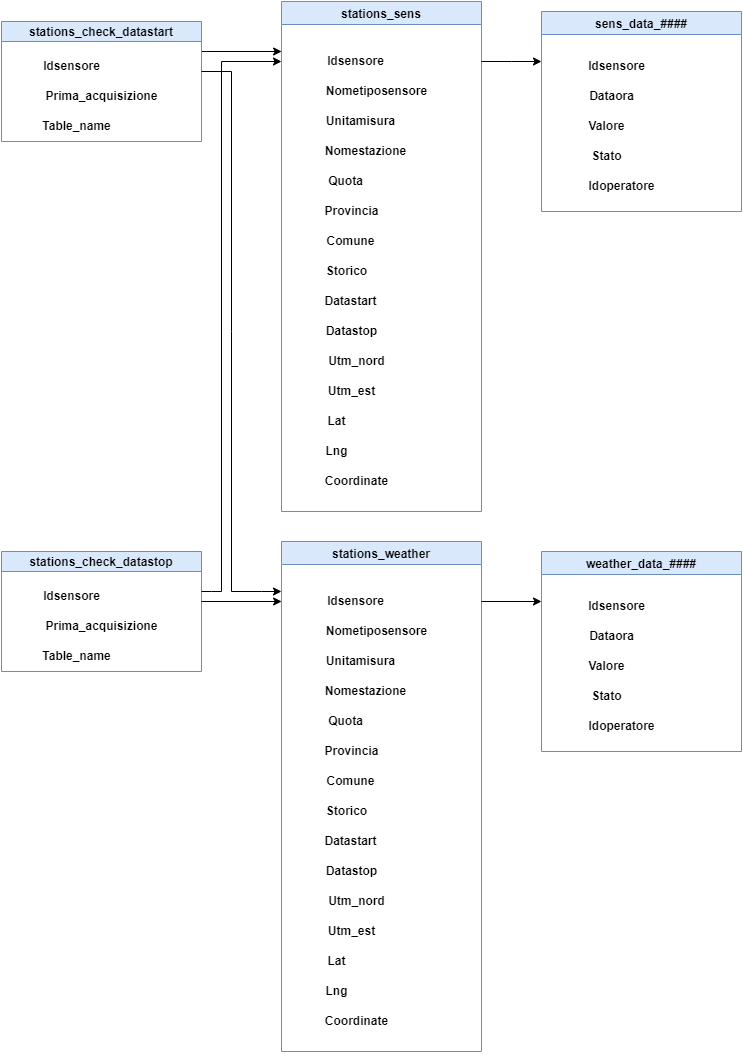
4. Data Analysis and Visualization

5. Conclusions and future work

* Use nosql db with tables bc it is more scalable
* Possible split tables to be one per year
* Put scheduled jobs on a server and not on a pc
* Connect Dashboard directly to the database for data
* Metti file check integrity su talend per automatizzare il tutto

Appendix A

**Database Overview**

****

*Figure 1*: Database tables overview and relations

|  |  |  |
| --- | --- | --- |
| Name | Number of rows | Size |
| logs\* | 63 | 48 KB |
| rejected\* | 3 | 16 KB |
| sens\_data\_1995 | 11,489,984 | 660 MB |
| sens\_data\_2000 | 14,492,734 | 832 MB |
| sens\_data\_2004 | 15,113,853 | 868 MB |
| sens\_data\_2007 | 13,277,946 | 762 MB |
| sens\_data\_2010 | 14,775,586 | 849 MB |
| sens\_data\_2011 | 4,972,158 | 285 MB |
| sens\_data\_2012 | 4,920,972 | 282 MB |
| sens\_data\_2013 | 4,913,671 | 282 MB |
| sens\_data\_2014 | 4,282,894 | 246 MB |
| sens\_data\_2015 | 4,270,563 | 245 MB |
| sens\_data\_2016 | 4,270,696 | 245 MB |
| sens\_data\_2017 | 2,595,465 | 149 MB |
| sens\_data\_2018 | 2,582,803 | 148 MB |
| sens\_data\_2019 | 2,574,002 | 147 MB |
| sens\_data\_2020 | 2,620,280 | 150 MB |
| sens\_data\_2021 | 2680340 | 154 MB |
| sens\_data\_2022\* | 1,714,296 | 112 MB |
| stations\_check\_datastart | 14 | 16 KB |
| stations\_check\_datastop | 27 | 56 KB |
| stations\_sens | 970 | 304 KB |
| stations\_weather | 1,262 | 344 KB |
| weather\_data\_2000 | 13,317,906 | 756 MB |
| weather\_data\_2005 | 39,455,657 | 2.2 GB |
| weather\_data\_2008 | 43,987,469 | 2.5 GB |
| weather\_data\_2010 | 977,107 | 56 MB |
| weather\_data\_2012 | 58,229,801 | 3.3 GB |
| weather\_data\_2013 | 42,497,020 | 2.4 GB |
| weather\_data\_2014 | 49,195,352 | 2.8 GB |
| weather\_data\_2015 | 57,877,143 | 3.2 GB |
| weather\_data\_2016 | 55,107,307 | 3.1 GB |
| weather\_data\_2017 | 56,682,542 | 3.2 GB |
| weather\_data\_2018 | 59,920,834 | 3.4 GB |
| weather\_data\_2019 | 60,126,516 | 3.4 GB |
| weather\_data\_2020 | 61,503,272 | 3.5 GB |
| weather\_data\_2021 | 59,929,021 | 3.4 GB |
| weather\_data\_2022\* | 36,922,546 | 2.1 GB |

\*Number of rows and size may increase over time as tables are updated on daily basis

*Table 1*: Information on database tables

Appendix B

**Talend Jobs**

Diagram

Description automatically generated

*Figure 1:* splitCSV job



*Figure 2:* combinedSplitCsv job

Graphical user interface, text, application, chat or text message

Description automatically generated

*Figure 3:* finalBulk job

A picture containing diagram

Description automatically generated

*Figure 4:* sensorsStations\_SQLQuery job

A picture containing chart

Description automatically generated

*Figure 5:* weatherStations\_SQLQuery job

Background pattern

Description automatically generated

*Figure 6:* bulkInsertS job

Chart

Description automatically generated with medium confidence

*Figure 7:* bulkInsertW job

Diagram

Description automatically generated with low confidence *Figure 8:* ingestion2022W job

Chart

Description automatically generated A picture containing chart

Description automatically generated

*Figure 9:* ingestion2022MonthW job

Chart

Description automatically generated with medium confidence Chart

Description automatically generated with medium confidence

*Figure 10:* ingestion2022S job

References

|  |  |
| --- | --- |
| [1] | J. RubyDinakar and S. Vagdevi, “A Study on Storage Mechanisms for Heterogeneous Sensors data on Big data Paradigm,” in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017. |
| [2] | R. Wijaya and P. B., “An overview and implementation of extraction-transformation-loading (ETL) process in data warehouse (Case study: Department of agriculture),” in *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, 2015. |
| [3] | M. Lenzerini, “Data Integration: A Theoretical Perspective,” in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002. |
| [4] | A. Halevy, A. Rajaraman and J. Ordille, “Data Integration: The Teenage Years,” in *Proceedings of the 32nd international conference on Very large databases*, 2006. |
| [5] | S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart, “Data Integration,” in *Web Data Management*, Cambridge University Press, 2011, pp. 193-222. |
| [6] | DBeaver, “About: DBeaver Community,” [Online]. Available: https://dbeaver.io/about/. |
| [7] | “What is DBeaver,” 15 March 2018. [Online]. Available: https://database.guide/what-is-dbeaver/. |
| [8] | “About: PostgreSQL,” [Online]. Available: https://www.postgresql.org/about/. |
| [9] | O. Mbaabu, “Introduction to Data Integration in Talend Open Studio,” Section, 2020. [Online]. Available: https://www.section.io/engineering-education/introduction-to-data-integration-in-tos/#:~:text=Talend%20Open%20Studio%20is%20an%20innovative%20software%20application%20that%20provides,effective%20and%20easy%20to%20use.. [Accessed 15 08 2022]. |
| [10] | Talend Inc., “Talend Components Reference Guide 7.3.1M6,” 17 01 2020. [Online]. Available: http://download-mirror1.talend.com/tosbd/user-guide-download/V731M6/TalendOpenStudio\_Components\_RG\_EN\_7.3.1M6.pdf. [Accessed 03 07 2022]. |
| [11] | R. Tiwari, “Data Storage vs Database,” Linkedin, 13 10 2021. [Online]. Available: https://www.linkedin.com/pulse/storage-vs-database-rupesh-tiwari?trk=articles\_directory. [Accessed 10 07 2022]. |
| [12] | Hwelett Packard Enterprise Development LP, “What is Data Storage?,” [Online]. Available: https://www.hpe.com/us/en/what-is/data-storage.html#:~:text=Data%20storage%20is%20the%20retention,using%20computers%20or%20other%20devices. [Accessed 10 07 2022]. |
| [13] | IBM, “What is data storage?,” [Online]. Available: https://www.ibm.com/topics/data-storage. [Accessed 10 07 2022]. |
| [14] | S. Palanisamy and P. SuvithaVani, “A survey on RDBMS and NoSQL Databases MySQL vs MongoDB,” in *2020 International Conference on Computer Communication and Informatics (ICCCI), 2020*, 2020. |
| [15] | ARPA Regione Lombardia, “Dati sensori aria,” 2022. [Online]. Available: https://www.dati.lombardia.it/Ambiente/Dati-sensori-aria/nicp-bhqi. [Accessed 11 07 2022]. |
| [16] | ARPA Regione Lombardia, “Stazioni Meteorlogiche,” 2022. [Online]. Available: https://www.dati.lombardia.it/Ambiente/Stazioni-Meteorologiche/nf78-nj6b. [Accessed 10 07 2022]. |
| [17] | J. Awiti, A. A. Vaisman and E. Zimányi, “Design and implementation of ETL processes using BPMN and relational algebra,” *Data & Knowledge Engineering,* vol. 129, no. 101837, 2020. |
| [18] | M. Souibguia, F. Atiguib, S. Zammalia, S. Cherfib and S. Ben Yahiaa, “Data quality in ETL process: A preliminary study,” in *23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2019. |
| [19] | R. Kimball and M. Ross, The Data Warehouse Toolkit Second Edition, New York: John Wiley and Sons, Inc., 2002. |
| [20] | A. U. Sepulveda, R. C. Loyola and M. W. Hernandez, “Performance comparison slowly changing dimensions using model relational and object-relational,” in *2015 34th International Conference of the Chilean Computer Science Society (SCCC)*, 2015. |
| [21] | D. Asanka, “Implementing Slowly Changing Dimensions (SCDs) in Data Warehouses,” 3 September 2021. [Online]. Available: https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/. [Accessed 25 07 2022]. |

J. Sreemathy, S. Priyadharshini, K. Radha, K. Sangeerna and G. Nivetha, "Data Validation in ETL Using TALEND," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 201hgh9, pp. 1183-1186, doi: 10.1109/ICACCS.2019.8728420.