

# Integrating Energy Data with ETL

Luís Luciano and Paulo Carreira

Instituto Superior Técnico, IST Taguspark, Av. Prof. Cavaco Silva, Tagus Park,  
2780-990, Porto-Salvo, Portugal,  
{luis.luciano,paulo.carreira}@ist.utl.pt

**Abstract.** In spite of the huge amount of energy information that is shared by cross-functional areas of companies, they aren't taking better decisions towards energy saving. Based on the existing literature, energy management systems and data warehouse architectures for energy management, a research model identifies several problems that affect intelligent building data integration. The aim of this article is to point several complexity factors that affect energy and building data integration and present a solution that could ease the integration process and applied in different environments to conform with different business requirements. To achieve it, is defined a generic prototype, which can help to define Extract-Transform-Load processes in building and energy management contexts. The prototype is intended to ease the process of extracting and transforming building and energy data by integrating specific ETL modules to an existent ETL tool.

**Keywords:** Data Warehouse, Energy Management Systems, Extract-Transform-Load, Intelligent Buildings.

## 1 Introduction

Business data analytics is becoming a key tool for managing nearly any kind of business enabling for instance analysing customer profitability, asset optimization and operation analysis to identify cost-reduction opportunities [1].

Data integration assumes an important role in business data analytics because it is responsible for combining multiple and heterogeneous sources of data (which may reside in different locations) and store them under a global schema, giving a unified view over that data. The main goal of this process is to help extracting knowledge that is scattered in different data sources [2]. Nowadays, data integration has been successfully applied in several domains, helping health professionals to extract valuable information from different medical records, managing personal information or even easing the data migration process in telecom providers [3,4,5].

Given the exponential growth of Intelligent Building we think that business analytics should be applied to building management as well. Therefore, building data should also be integrated, to ease the process of decision making and identify optimization opportunities among which, energy saving is chief. To achieve

that, building data must be collected and consolidated, analysed and aggregated in proper formats (such as reports) allowing the drilling and mining of building data. Combining data in this way is crucial for example to understand the energetic behaviour of a building and to clarify the relation of each device and appliance to energy consumption.

This paper addresses the urgent need of integrating Intelligent Building data to identify possible energy savings and improvements in the energetic behaviour of a building. Monitoring and analysis of building performance is a key mechanism to profile consumption patterns, to detect abnormal energy use and to reduce energy consumption. This document is organized into five sections. After the introductory section, Section 2 explores the different types of Intelligent Building data and the underlying complexity of integration this type of data. Section 3 details the Extract, Transform and Load (ETL) process which is responsible for integrating data in a global unified schema and also the pros and cons of applying ETL methodologies in a energy and building domain. Section 4 presents a prototype that integrates an ETL tool with protocols and standards of Intelligent Buildings and the advantages of this synergy, Section 5 presents some conclusions about this work.

## 2 Intelligent Buildings Data

The concept of Intelligent Buildings (IB) is related with the usage of Information Technology in building operations to face the progressive demand of comfort environment, the requirements for occupant control of the environment and the reduction of energy usage [6]. Furthermore, IB are also concerned with preserving the surroundings of the facilities and enhancing building operations to reduce energy consumption and environmental impact [7].

### 2.1 Sources of Intelligent Buildings Data

An IB encompasses several site-specific systems that control well defined areas and aspects of a building. An Energy Management Systems (EMS) aims at identifying energy-savings opportunities through continuous monitoring of energy consumption and equipments. Building Management System (BMS) which has to control and monitor mechanical and electrical equipments of a building [7]. The integration of multiples data sources is crucial for energy and building management, which give accurate information about the location and time of energy usage [8,9]. It is possible to categorize the types of information that are needed to perform an energy analysis, a detailed list of each data source is depicted in Table 1.

- Building Structure, refers to the type of building, the internal infrastructure and the physical layout details. To understand the energy consumption in the building we have to break up different areas physically (i.e., space breakdown and functionality) to profile the energy for each location.

External Sources	Type of information	Data Source Format
Energy Management System (EMS)	Continuous monitoring of energy consumption.	EMS specific software
Automated Meter Reading (AMR), Advanced Metering Infrastructure (AMI)	Read, transport and store meter energy data.	AMR, AMI specific software
Building Management Systems (BMS), Building Automation System (BAS)	Control and monitor building equipments(e.g. HVAC, Lighting System, Fire protection system).	Standard Protocol (e.g. BacNet, Lon-Works). Proprietary Software (e.g. TAC Vista, Metasis).
Computer-aided Facility Management (CAFM)	Supports operational management and activities related with Facility Management (FM). Provides functional space information.	CAFM specific software
Building Information Model (BIM)	Provides information about building envelope.	gbXML, ifcXML
Organizational Information	Information about the organization and the way it is structured to perform the core activities.	Database Management System (DBMS)
Energy Pricing Data	Information about tariffs from different energy service providers.	Paper form (non-structured data)
Billing Information	Keep record of information about energy expenditures (i.e. energy cost, taxes, billing date)	Paper form, Billing Information System (BIS)
Weather Data	Provides information about climatic conditions in the facilities and surroundings.	Provided by weather stations connect to the BMS or from external sources.

**Table 1.** Set of External Sources which include several Intelligent Building Systems that must be aggregated into a global schema. The provided information will ease the process of Intelligent Building data analysis by exploring data according to different dimensions.

- Operational Data, refers to data that has a direct relation with the business process, with the space, the occupant and how they perform their core activities. This is important so that energy use can be traced to activities.
- Commissioning data provides details about the operation of IB data systems, showing temperature, pressure levels and setpoints, helping to determine the cause off peak-demands and abnormal situations.
- Sensor data captures environmental and occupancy data related with the information that can be measured using sensors (luminance sensors, occupancy sensors).
- Equipment status data is essential to fully understand the energetic behaviour of a building. Equipments may be grouped as a stand-alone device

or operate as a system (e.g. HVAC). Each device as a specific electric-load, an associate activity (e.g. air conditioning, lighting) and a working period.

Accordingly IB data integration is an activity that gives support to Energy Management (EM) and Facility Management (FM). A building should be understood as a whole system and the interaction of each system to energy consumption and building maintenance must be defined. Then, building data must be correlated, aggregated according to different dimensions and hierarchies in order to process and analyse information.

## 2.2 Complexity of Intelligent Buildings Data

There are intricate factors that affect the integration of IB data making it harder to provide an homogeneous and unified view. The data integration in this context is hindered mainly due to:

- Heterogeneity of data sources. As explained before, the information required in IB comes from different sources. The buildings, devices, occupants, external factors such as weather conditions have to be brought together. This variety of sources requires the communication with several external sources in order to extract useful knowledge. IB data sources present data in different data structures (structured, semi-structured, non structured), with different sampling periods (data that is constantly being updated such as energy information provided by meters and data that is most likely not to change like building information).
- Data storage. IB data is stored in databases of proprietary systems known as “informational silos”, which makes it harder to extract and access information.
- The amount of data produced by external sources systems is often very large, since energy meters are constantly performing readings of energy consumption.
- Mapping problems. Due to the heterogeneity of data sources more efforts are need for schema matching (to identify that different schemas share similar semantics) and for schema mapping (performing transformations to integrate different schemas).
- Data Quality. In the IB context the quality of the information provided relies heavily in the accuracy of systems and devices (e.g. accuracy of data acquisition meters, sensors) and also to his fault tolerance capacity (e.g. communication losses with meters causes the storage of incorrect energy values).
- Large Data Models. The source data models are frequently very large due to the number of entities and instances requiring an additional effort to integrate these sources. Moreover, the documentation is often poor or absent.
- Organization-dependent data. In this context some information is hard to infer because is not explicitly stored in any system. This knowledge is

stored and shared by the people that compose that organization and is not stored in any physical format.

Nevertheless, collecting, archiving and analysing energy and building data requires significant computational resources with the ability of processing and analysis, making this a costly and cumbersome task and usually the information provided to the end-user is very difficult to interpret [10]. For that reason the IB data integration and data analysis is currently an handcraft process carried out by energy and building managers. Moreover, energy and building management systems are populated with inaccurate and outdated information leading to a distorted perception of the system's performance and to incorrect decisions [11].

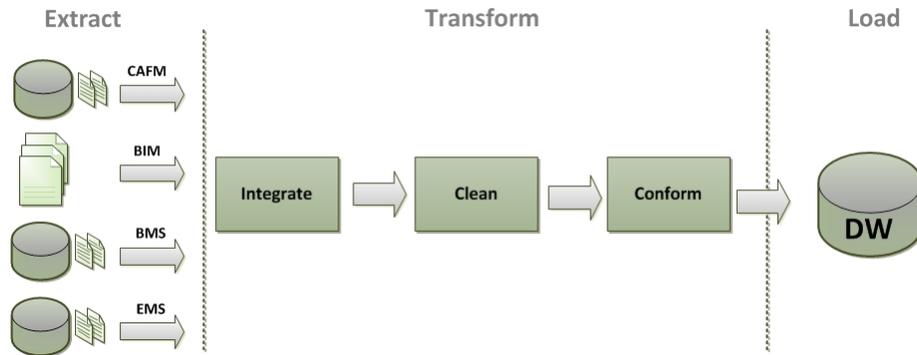
### **3 Extract, Transform and Load**

Building and energy data must be stored and integrated with a global unified schema enabling an easy access to information to specific users. This data can be integrated into a Data Warehouse (DW) which is a repository of information collected from multiple sources and integrated [12]. The main goal of a DW is supporting data analysis and decision-making, making information easily accessible and present information in a way that is consistent with the business requirements while being flexible and resilient to changes [13,14]. A DW is populated by integrating data from different sources which has to be cleaned, converted and conformed to fit in the DW schema [12]. Extraction, Transformation and Loading (ETL) processes play an important role, populating and assuring the data quality of the DW. ETL processes are responsible for: extracting data from operational source systems, transforming data which includes integrating it, checking for inconsistencies and assuring the data accuracy to meet business requirements and finally for the loading stage, which is accountable for delivering information to the data presentation area [15,16], as depicted in Fig. 1. They control the loading and refreshment of the new data [15].

Nevertheless, ETL is more than loading data from the operational source systems to the data presentation area. ETL processes are responsible for enforcing data quality and consistency, conforming and for mapping data from different data sources.

#### **3.1 Challenges of ETL**

To design ETL processes it is important to define the system requirements, to identify the data sources and the DW schema and determine the set of transformations that are needed for the project. The complexity factors of developing ETL process for building operations are due to difficult access to data sources, data source characteristics and lack of a reference model. Difficult access to data sources is related with the lack of drivers for IB protocols and with the existence of undocumented energy data models. Heterogeneous data sources have different characteristics, there are real-time data sources (that require real-time ETL)



**Fig. 1.** Overview of an ETL system to populate a DW. The extraction from structured, semi-structured and unstructured data sources, the transformation and the loading step to a centralized model (adapted from [16]).

and data quality issues due to intermittent data sources (meters and equipments that disconnect). The lack of a reference integrated model, means that data has to be integrated against a model that no one knows how to build it (no such model has been proposed until now). Moreover, autonomous and heterogeneous data should be integrated in an uniform way and consolidated through the data cleaning activity to assure data quality. Poor data quality as a strong impact in enterprise strategies, because the foundations of his success and of the process of decision-making relies heavily on this data [17]. Therefore, several authors consider that energy and building data quality must be evaluated in terms of validity, accuracy, completeness and timeliness [18,17].

The ETL activity is crucial for designing a DW for Energy Management. This activity replaces the manual task of analysing and correlating energy data, which is very difficult and error prone since the amount of information is large and heterogeneous.

### 3.2 Advantages of using ETL to integrate intelligent building data

There are several reasons for using an ETL tool to integrate heterogeneous energy data sources.

**Data Source abstraction**, eases the process of creating an ETL transformation because data can be handled and accessed in a uniform way. There are several factors that depict the advantages of using an ETL tool to effectively integrate energy data:

- Logic access to data is location-independent and implementation independent. The abstraction level hides the complexity of extracting data and the location of the data source, the developer just needs to focus in the transformation process.

- Since the architecture is source-independent, raw data can be handled in a generic way because it is independent of the extraction process. Provides a uniform access to data sources in a common representation.
- Connectivity, provides connection with a wide range of source systems (from relational databases, XML files among other formats). This feature is important because it transforms heterogeneous data into primary data, which is easier to manipulate.

**Declarative transformations**, with this paradigm it is only required to express the logic of the transformation without describing the entire execution flow.

- Algorithms are chosen based in the context conditions that can be modified, allowing the solution to evolve to face new demands.
- The strategy to implement a transformation is defined through the specific context. For each transformation it is necessary to evaluate the best algorithm to face the context characteristics.
- Improve scalability, it is possible to use computing techniques to improve the transformation process, for instance using parallelism, partitioning or clustering.

**Re-usability**, since ETL process can be disaggregated into loosely-coupled components it is possible to modify and reuse ETL components to fit in new solutions.

- Domain specificity is low, which means that ETL transformations can evolve to solve other problems in different domains. Eases the process of reproducing a new ETL solution.
- Modularity, allows the separation and recombination of ETL components. Reduces complexity and increases the flexibility of creating an ETL transformation.
- Extensibility, allows the extension and creation of new functionalities. Since ETL is a wide explored area and several ETL tools are open-source solutions it is natural to take into consideration future growth.

**Explicit knowledge**, fruitful information about ETL process is easily stored and transmitted. It is focused on the “essential” data.

- Easy to understand and control ETL transformations, the user doesn’t need to concern with implementation tasks.
- Business-oriented, it is possible to evaluate only the business logic, to identify transformation rules and constraints, determine the execution flow and necessary steps to complete a transformation.

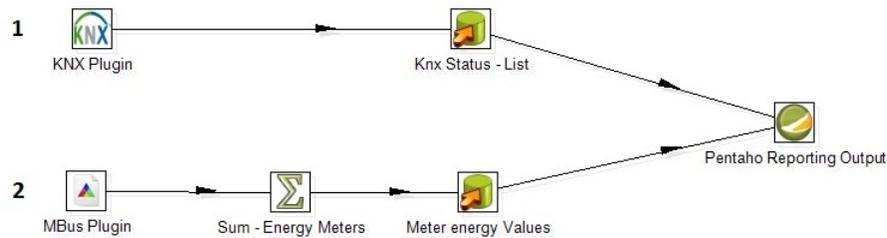
## 4 Prototype

To validate our ideas we implemented an ETL data flow using an open-source ETL tool (Pentaho Data Integration - Kettle<sup>1</sup>) to create a set of extra-features

<sup>1</sup> <http://kettle.pentaho.com/>

that allows the communication with standards and protocols of intelligent buildings. Keettle consists of an ETL engine and GUI applications that allows the developer to define data integration process using jobs and transformations.

To extract data from building management and energy management systems there are two possible solutions: (1) build a specific data extraction software that must interact with each data source, (2) create data source adapters for an ETL tool that share some features (such as similar API, similar metadata format, among others) allowing the developer to focus only in the extraction process. The first option present several drawbacks, which makes this solutions infeasible to integrate building data. With specific extraction software there is no abstraction level between the transformation and the data source in that way the extraction driver must be embedded in the software, on the other hand the missing extensibility affects the flexibility since is not possible to add or modify a component to address a specific business need [19].



**Fig. 2.** Example of a Kettle transformation step. (1) Presents the input step: KNX Plugin which is responsible for connecting with KNX devices and read their status. After this step data is saved in a Database that stores the status of each KNX device. (2) Depicts a transformation that reads the values of several energy meters (running Mbus protocol) and performs the sum of all energy meters.

We propose to implement a module with drivers for extracting data from energy and building data sources. The prototype already has a connection with an Automated Meter Reading (AMR) system which is responsible for collecting, transporting and storing energy meter data from different types of meters (electricity, gas, heat) and with building control standards (e.g. KNX<sup>2</sup>) to verify the status of sensors and actuators to evaluate the operation of building management equipments, such as lighting, shutters, HVAC systems [20]. Using this two drivers it is possible to correlate energy consumption in a specific period with the weather conditions, which can reveal abnormal energy consumption (e.g. to study how energy consumption behaves with high temperatures and high illuminance values). The transformation scenario depicted in Fig.2 can be disaggregate

<sup>2</sup> <http://www.knx.org>

into two transformations, the first connects a KNX weather station to read the lux, temperature and humidity levels and stores this information in a database with a timestamp. The second transformation connects a Mbus<sup>3</sup> data concentrator to read the values of each energy meter, later the energy values are combined to calculate the energy consumption in that period (using an aggregator step) and finally stored in a database. This information will be provided in reports so that end-users could understand how weather influence energy behaviours.

By integrating this components of building management we are able to directly extracted data from devices, sensors, meters and actuators located in any point of the building. This direct extraction decreases latency time, since energy and building data can be faster modified and loaded to the final schema and increases the flexibility because data can be cleaned and conformed using the available steps of the Kettle framework.

## 5 Conclusion

Energy management is a fast growing area along with smart grids, smart metering infrastructures and intelligent buildings. Energy managers and building owners have realized the potential of this area but they are not totally aware of the fact they are not taking the most of these systems: they have a lot of information but little knowledge about their energy consumption. Thus, integration of energy data is of utmost importance.

Nowadays data warehousing and ETL are very popular tools in business analytics to effectively integrate data residing in different data sources. However these tools have not yet been adopted to energy management and more specifically to integrate energy-related data.

The aim of this article has been to discuss the superiority of using ETL tools to integrate energy data. We presented a prototype ETL implementation that extracts data directly from building management and energy management systems. The solution is highly modular and isolates the data transformation logic from the process of communicating with meters, sensors and actuators since it reduces the number of steps in the communication and allows the interaction of other steps in the extracted data because the steps are loosely-coupled.

The advantages of this solutions compared to build up a specific software from scratch are: flexibility to create different steps based in minor modifications with minor effort, re-usability to reuse existing components to create new components and connectivity to interact with a multiplicity of source systems. We expect the results of this work to contribute to streamline the engineering practice concerning data warehouse projects in energy contexts and integrating energy-related data sources.

---

<sup>3</sup> <http://www.m-bus.com/>

## References

1. R. Kohavi, N. Rothleder, and E. Simoudis, "Emerging trends in business analytics," *ACM*, vol. 45, pp. 45–48, Aug. 2002.
2. M. Lenzerini, "Data integration: a theoretical perspective," in *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, (New York, NY, USA), ACM, 2002.
3. Y. Cai, X. Dong, A. Halevy, J. Liu, and J. Madhavan, "Personal information management with semex," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, (New York, NY, USA), ACM, 2005.
4. H. Agrawal, G. Chafle, S. Goyal, S. Mittal, and S. Mukherjea, "An enhanced extract-transform-load system for migrating data in telecom billing," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, (Washington, DC, USA), IEEE Computer Society, 2008.
5. L. Rokach, O. Maimon, and M. Averbuch, "Information retrieval system for medical narrative reports: Flexible query answering systems," 2004.
6. J. Wong, H. Li, and S. Wang, "Intelligent building research: a review," *Automation in Construction*, vol. 14, pp. 143–159, Jan. 2005.
7. SMART-ACCELERATE, "Intelligent building technology," 2008.
8. H. Gokce, D. Browne, K. Gokce, and K. Menzel, "Improving energy efficient operation of buildings with wireless IT systems," 2009.
9. D. Fong and A. Schurr, *Information Technology for Energy Managers*, ch. Relational Database Choices and Design, pp. 255 – 263. Fairmont Press, 2004.
10. X. Li, "Classification of energy consumption in buildings with outlier detection," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3639–3644, 2010.
11. D. Silva, "A data mining framework for electricity consumption analysis from meter data," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 399–407, 2011.
12. W. Inmon, *Building the Data Warehouse*. John Wiley & Sons, 2005.
13. J. Han and M. Kamber, *Data Mining: concepts and techniques*. Morgan Kaufmann, 2006.
14. R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, 2002.
15. P. Vassiliadis, "A survey of extract-transform-load technology," *International Journal of Data Warehousing and Mining*, vol. 5, no. 3, pp. 1–27, 2009.
16. R. Kimball and J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. John Wiley & Sons, 2004.
17. G. Thompson, J. Yeo, and T. Tobin, *Web Based Energy Information and Control Systems*, ch. Data Quality Issues and Solutions for Enterprise Energy Management Applications, pp. 435–446. 2005.
18. C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006.
19. M. Awad, M. Abdullah, and A. Ali, "Extending etl framework using service oriented architecture," *Procedia Computer Science*, vol. 3, no. 0, pp. 110–114, 2011.
20. D. Shu, S. Ma, and C. Jing, "Study of the automatic reading of watt meter based on image processing technology," in *Industrial Electronics and Applications. 2nd IEEE Conference*, 2007.