Combining Web and Enterprise Data for Lightweight Data Mart Construction[⋆]

Suzanne McCarthy, Andrew McCarren, and Mark Roantree

Insight Centre for Data Analytics, School of Computing, Dublin City University,
Dublin 9, Ireland.
suzanne.mccarthy@insight-centre.org, andrew.mccarren@dcu.ie,
mark.roantree@dcu.ie

Abstract. The Agri sector has shown an exponential growth in both the requirement for and the production and availability of data. In parallel with this growth, Agri organisations often have a need to integrate their in-house data with international, web-based datasets. Generally, data is freely available from official government sources but there is very little unity between sources, often leading to significant manual overhead in the development of data integration systems and the preparation of reports. While this has led to an increased use of data warehousing technology in the Agri sector, the issues of cost in terms of both time to access data and the financial costs of generating the Extract-Transform-Load layers remain high. In this work, we examine more lightweight data marts in an infrastructure which can support on-demand queries. We focus on the construction of data marts which combine both enterprise and web data, and present an evaluation which verifies the transformation process from source to data mart.

1 Introduction

Agri companies are increasingly making use of data analytics and data warehouse technologies. Start-ups and research groups are emerging for the purpose of addressing the need for data analytics unique to the Agri sector [1, 7]. Extract-Transform-Load (ETL) as a framework for data integration gained popularity in the 1970's and quickly became the standard process for data integration. Traditional ETL grabs a 'snapshot' of the source data at given intervals, be it daily, weekly etc and updates the data warehouse or data marts [14]. This happens pro-actively at these intervals and all the source data is extracted, transformed according to relevant rules and loaded into a warehouse where it remains until it is needed for user queries. The primary issue with this approach is that it is very costly in terms of both time and resources. As the integration effort is generally significant, it is not easy to modify the ETL system, meaning that the addition of new sources is a major task.

[⋆] Research funded by Science Foundation Ireland under grant number SFI/12/RC/2289

Motivation. The Kepak Group [6] is a leading Irish meat producer and our industry partner and provided the business requirements that drive our case study for this project. Using one of these requirements, we have specified a data mart to be built and populated using our ETL framework. It is often the case that only a small subset of all the data loaded is actually required for a user's query. However, the user must still wait for the loading of the data update in its entirety [5], a burden for data scientists who require near real-time data to make predictions. Traditional ETL is typically run on a batch basis, for example on a weekly or daily schedule. All of the data from the input data source(s) are transformed and loaded to the target warehouse, with no regard for how, when or if it will be used at query-time. This is less than ideal for two main reasons: it is heavy on resources and creates redundancy; this time-consuming process also means that the warehouse is constantly out of date. In [5], researchers focus on shortening the update interval using Active Warehousing. However, doing so increases the burden on the transactional and analytical databases, as it requires performing more frequent extractions, scheduled concurrently with the frequent loads, while still not addressing the issue of redundancy in the data.

Contribution. In this paper, we present a new framework which presents a more lightweight approach to the ETL process. The framework also provides a methodology to integrate unknown data sources, from both enterprise and web environments. With the aid of an (Agri) ontology and transformation templates, the Transform and Load components are used to populate the data marts directly. As the process is largely automatic, it has a significant benefit to end users who can import new data into data marts for analysis and predictions. However, this fast transformation of unknown sources may introduce errors and thus, our evaluation contains validation routines to verify transformed data.

Paper Structure. The remainder of this paper is structured as follows: In §2, we present a review of related research in this area; Using a real world case study from a test application developed with our industry partner, §3 shows our approach to importing new data sources; §4 continues the case study with a description of data mart construction and population; in §5, we present our evaluation which validates our transformation process; and finally, in §6, we present our conclusions.

2 Related Research

As the primary focus of this paper is to build an automated ETL process to combine web and enterprise data into data marts, we focus on similar work in this area. While data integration is now a well-understood problem and ETL architectures comprise mature processes, research into systems that attempt a form of auto-integration between enterprise and web sources, or web sources alone, remains topical.

In [10], the difficulty in automating a process to assign the mappings and transformations required for ETL is addressed. While information about the semantics of the data, the necessary constraints and requirements may be hard to find or unreliable, the design of an ETL process hinges on this information. Therefore, the authors propose making use of an ontology in order to provide this domain-specific information to drive the transformation stage of the ETL process. Our approach also requires a domain-specific ontology to drive the transformation stage of the ETL process but we propose a number of metadata structures drawn from the ontology for a more lightweight ETL process which can support on-demand querying.

In [13], the authors present an ontology-based agricultural knowledge fusion method based on recent advances in the area of data fusion such as the semantic web. Their ontology defines an integrated hierarchy of agricultural knowledge: definitions and relationships. The purpose was to use an ontology to drive a knowledge fusion method to resolve disparity when integrating Agri data sources to create knowledge as well as analyse the data. We also make use of an ontology but instead combine it with transformation templates to resolve the heterogeneities between enterprise and web sources.

In [2], researchers focus on reducing latency in the data warehouse as a decision-making tool. Similar to our work, they regard the creation of a real-time data warehouse as a continuous data integration environment and end-to-end automation of the ETL process. They propose an architecture that facilitates a more streamlined approach to ETL by moving the data between the different layers of the architecture without any intermediate file storage. However, data is loaded as it becomes available and not as it is needed by the user. Our approach populates data marts from the Data Lake as they are required by the end user.

3 Data Importation Process

In traditional ETL systems, warehouse updates occur on a periodic basis through the ETL process and normally represents a significant overhead. The various components of Extract, Transform and Load are separated in our framework where the Extract process imports to the Data Lake only and involves the use of a Metabase as a management tool for the Data Lake.

3.1 Data Lake Storage

A Data Lake is a repository for large amounts of data from multiple sources, wherein the data is stored in its native format. This makes it a very lightweight and low-overhead data storage area. In our system, we adapt the concept slightly and perform a very simple transformation on each of the imported datasets, which are then stored as sets of attribute-value (A-V) pairs.

When a new data source is imported, metadata is recorded as an *Import Template*, and updated again when the data is transformed, queried and/or loaded. The Import Template then can provide the system with basic details (metadata) of the source data. An example of the Import Template can be found at [9].

3.2 Transformation

A transformation template (*Transformer*) is created for each individual data source and it stores each of the native terms used in the data and maps each of them to a standard term. In this research, we are using the Agri warehouse model described in [8]. All data marts must conform with the common model: they are cubes derived from the base schema. The base schema is a star schema comprised of 24 dimensions and 26 fact tables.

The output of the following 5-step process is a dataset which is *ready to load* into the data mart. Given a set $S = \{(A_1, V_{11}), (A_1, V_{12}), (A_{1,V_{13}})...(A_4, V_{43})\}$

1. Retrieve attribute-value pair
2. Map attribute to standard term found in Transformer
3. Map value to standard term found in Transformer
4. Create transformed attribute-value pair.
5. Update Import Template to indicate transformation status

3.3 Load

The Load process initiates a MySQL Insert statement and populates parameters before execution. Both the Import Templates and the Transformers are used again in the Loading process. To fulfil the Load process, the set of transformed attribute-value pairs are first split into batches, where each batch contains a *single item* of *each* measure and its associated dimension data. These are then used to populate a MySQL command to insert data to the data mart. The Import Template informs the process of the size of each batch and the target mart. Further details of the Transformation and Loading workflows can be found at [9].

4 Data Mart Construction

4.1 Overview

Kepak, our Agri business partner, have a requirement to analyse trade prices of beef commodities and wish to do so using their own enterprise data, combined with selected web based information. The 5 data sources are described below and, in the following section, we provide details for the Eurostat import and transformation process.

- The Eurostat [4] website is the source for EU trade data which is publicly available.
- The United States Department of Agriculture (USDA) [12] publishes Agri trade data figures which can be downloaded in bulk.
- StatCan [11] is the Canadian National Statistics agency and publishes economic, social and census data.
- Comtrade [3] is the U.N. international trade statistics database.

- Kepak Group [6] have exported from their own operational databases.

The above sources provide dimensional data to the following source attributes:

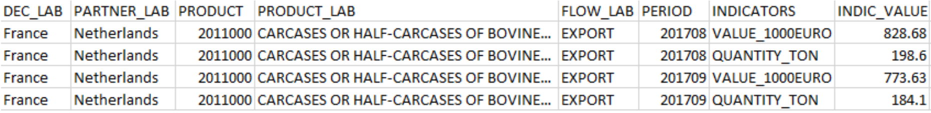
- **reporter**: the country that is reporting this data
- **partner**: the country with whom the reporter is trading
- **product**: the commodity being traded
- **flow**: the direction of the trade, i.e. imports or exports

All sources additionally have a time attribute but the **date** dimension in the warehouse is static and pre-populated with a fixed range of dates before loading time. Therefore, it is not part of the evaluation process presented in §5. All web sources use **trade weight**, measured in either tons or kg, and **trade value**, in their local currency, as their measures. The enterprise data uses both of these along with two additional measures: **offcut_value** and **yield**.

Beef trade data is extracted from each of these sources. This data will then populate a beef trade data mart with the following dimensions: **dim_geo**, **dim_trade_product**, **dim_trade_flow**, **dim_date_monthly**; and measures: **trade_weight**, **trade_value**, **offcut_value** and **yield**. Both the reporter and partner source attributes will populate the **dim_geo** dimension.

4.2 Details for Eurostat Import

The input for this process (Fig 1) is a file download from the website. Example 1 shows a single row of the Eurostat data in the attribute-value pairs which is the format for the Data Lake.



DEC_LAB	PARTNER_LAB	PRODUCT	PRODUCT_LAB	FLOW_LAB	PERIOD	INDICATORS	INDIC_VALUE
France	Netherlands	2011000	CARCASES OR HALF-CARCASES OF BOVINE...	EXPORT	201708	VALUE_1000EURO	828.68
France	Netherlands	2011000	CARCASES OR HALF-CARCASES OF BOVINE...	EXPORT	201708	QUANTITY_TON	198.6
France	Netherlands	2011000	CARCASES OR HALF-CARCASES OF BOVINE...	EXPORT	201709	VALUE_1000EURO	773.63
France	Netherlands	2011000	CARCASES OR HALF-CARCASES OF BOVINE...	EXPORT	201709	QUANTITY_TON	184.1

Fig. 1. Eurostat web output

Example 1. Eurostat data in Data Lake
S={(PERIOD,201708),
(DECLARANT_LAB,France),
(PARTNER_LAB,Netherlands),
(FLOW_LAB,EXPORT),
(PRODUCT,2011000),
(PRODUCT_LAB,CARCASES OR HALF-CARCASES OF BOVINE ANIMALS, FRESH OR CHILLED),
(INDICATORS,QUANTITY_TON),
(INDICATOR_VALUE,198.6) }

4.3 Eurostat Transformation Process

The input for this process is the Eurostat data in its Data Lake format, as seen in Example 1. For each attribute and value, the Eurostat Transformer supplies the standard term which should replace it (though it is not always the case that the source's original term is different from the standard term). The output of this process is seen in Example 2.

Example 2. Eurostat data after Transformation
S'={ (yearmonth,201708),
(reporter,FRANCE),
(partner,NETHERLANDS),
(flow,exports),
(product_code,2011000),
(product_desc,CARCASES OR HALF-CARCASES OF BOVINE ANIMALS, FRESH OR CHILLED),
(unit,ton),
(value,198.6) }

5 Evaluation

As with any automated transformation process, validation of the process itself is crucial. The goal of this paper is on the overall framework which moves data from source to data mart and the focus of the evaluation to validate transformations. This involves two sets of experiments: the first validates that all *measure* data was loaded correctly in its entirety from source to (warehouse model) data mart; the second validates that *dimensional* data underwent valid transformation. Possible states that may result during the Loading of both fact and dimension tables are:

- **Attribute error**: A mismatch between a data point from the source file and the data mart because dimensions were extracted from the Data Lake in the wrong order. This could occur with both dimensional and measure data.
- **Value error**: Missing values where a value is accidentally skipped during the Loading process means that the datasets have different cardinalities. This could occur with both dimensional and measure data.
- **Transformer error**: Errors during creation of the Transformer means that the data is incorrectly NULL. This can only occur for dimensional data.
- **True**: The validation test passes.

Measure Validation: Ordered List Test. The purpose of the Ordered List (OL) Test is to ensure that the correct number of data points were loaded to the mart (no missing or duplicated data) and that values that should remain unchanged, did so. The OL test runs against every measure in each of the data sources. The inputs for this process are an ordered list of measure values obtained from the Fact table, and an ordered list of measure values extracted from the source file. It tests for Attribute errors and Value errors, and returns True if none are found.

Frequency Pattern (FP) Test for Dimension Values. Dimensional data is treated as non-numeric and values may be required to change during the Transformation process. The purpose of this test is to ensure that term transformations are *consistent*. This test is run on every populated dimension. Two sorted lists are constructed: the frequency count of all distinct dimensional values found within the source file, and the distinct frequency count of all foreign keys found within the fact table, both sorted in ascending order. If the frequencies of the values are different, then either two source original terms have been mapped to the same standard term or a single original term has been mapped to more than one standard term, i.e. term mappings have not been consistent.

The two sorted lists are compared and tested for all of the potential errors - attribute, value and transformer, and returns True if none are found.

5.1 Results Overview

The results of all validation tests run gave a result of 6 failures out of 28 tests giving a success rate of 78.57%. On investigating the causes for the failures, all the errors listed previously were found.

- **Attribute error**: For example with USDA, the **dim_trade_product** foreign key was assigned to the **dim_trade_flow** value. This shows the importance of importing the dimension values in the order in which they are expected to populate the dimensions.
- **Value error**: For example with Comtrade trade value, a data point was skipped as a result of an error during Loading.
- **Transformer error**: For the dimensional data of our business partner, there is a high number of dimensions and not all were imported as the Agri model did not capture this level of dimensionality.

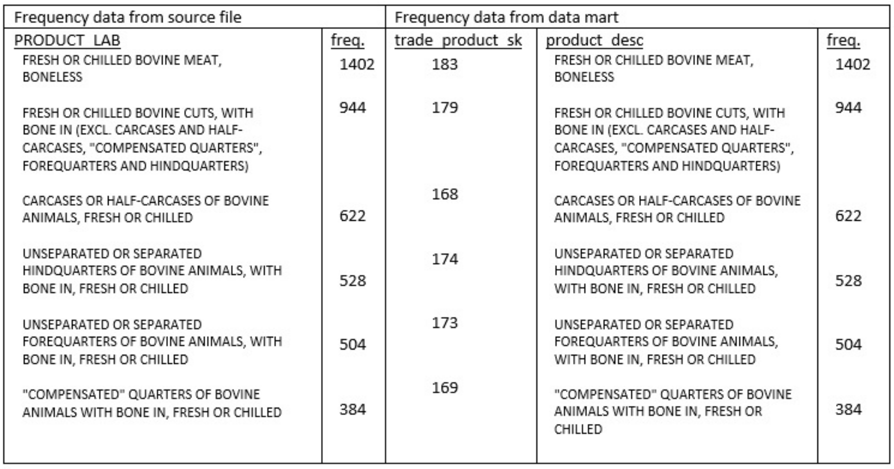
5.2 Detailed Eurostat results

The results for the validation of the Loading process for a Eurostat dimension and a measure are given.

Fig 2 shows the Frequency Pattern of the **product** source attribute as extracted from the **dim_trade_product** dimension for Eurostat. The label **trade_product_sk** refers to the ID of the fact's foreign key link with the **dim_trade_product** dimension. It can be seen that the terms in the source file and the transformed data in the data mart have the same frequency pattern and that the source original terms were mapped correctly to a standard term.

For the sake of brevity, the entire results of the Ordered List tests on the measure will not be displayed. Instead we have summed the values of **trade_weight** from the source file (TWS) and the values of **trade_weight** from the fact table (TWF).

$$\sum(tws \in TWS) - \sum(twf \in TWF) = 0$$



Frequency data from source file		Frequency data from data mart		
PRODUCT_LAB	freq.	trade_product_sk	product_desc	freq.
FRESH OR CHILLED BOVINE MEAT, BONELESS	1402	183	FRESH OR CHILLED BOVINE MEAT, BONELESS	1402
FRESH OR CHILLED BOVINE CUTS, WITH BONE IN (EXCL. CARCASES AND HALF-CARCASES, "COMPENSATED QUARTERS", FOREQUARTERS AND HINDQUARTERS)	944	179	FRESH OR CHILLED BOVINE CUTS, WITH BONE IN (EXCL. CARCASES AND HALF-CARCASES, "COMPENSATED QUARTERS", FOREQUARTERS AND HINDQUARTERS)	944
CARCASES OR HALF-CARCASES OF BOVINE ANIMALS, FRESH OR CHILLED	622	168	CARCASES OR HALF-CARCASES OF BOVINE ANIMALS, FRESH OR CHILLED	622
UNSEPARATED OR SEPARATED HINDQUARTERS OF BOVINE ANIMALS, WITH BONE IN, FRESH OR CHILLED	528	174	UNSEPARATED OR SEPARATED HINDQUARTERS OF BOVINE ANIMALS, WITH BONE IN, FRESH OR CHILLED	528
UNSEPARATED OR SEPARATED FOREQUARTERS OF BOVINE ANIMALS, WITH BONE IN, FRESH OR CHILLED	504	173	UNSEPARATED OR SEPARATED FOREQUARTERS OF BOVINE ANIMALS, WITH BONE IN, FRESH OR CHILLED	504
"COMPENSATED" QUARTERS OF BOVINE ANIMALS WITH BONE IN, FRESH OR CHILLED	384	169	"COMPENSATED" QUARTERS OF BOVINE ANIMALS WITH BONE IN, FRESH OR CHILLED	384

Fig. 2. Eurostat product dimension Frequency Pattern

The results of the Eurostat case study data shows that the data retains its integrity throughout the ETL process and the original data could, if necessary, be re-created from the data mart by reversing the transforms by the same process. This supports our vision of a lightweight ETL process that makes use of metadata structures, as opposed to a high-overhead traditional data integration process.

6 Conclusions

Data warehouses provide the basis for powerful analytical tools but are expensive to build and support, and tend to be very slow to incorporate new data sources. In this work, we looked at creating more lightweight data marts in order to accommodate new data sources that are selected by end users. We presented a framework which uses a common Agri model to which data marts must conform, and a method to transform data sources into the conforming data mart. A real world case study was specified by our Agri business partner which used their own enterprise data together with 4 selected web sources. Our evaluation used a number of validation techniques to ensure that data extracted from source and loaded into data marts were accurate. An area of work not addressed in this research is the automatic construction of the Transformer which is part of current research. Our overall aim with this platform is to provide Query on Demand data marts which extract from the data lake as required by the user.

References

1. DIT Agriculture Analytics Research Group. http://www.agrianalytics.org/. 2018.

2. Bruckner R, List B, and Schiefer J. Striving towards near real-time data integration for data warehouses. In Data Warehousing and Knowledge Discovery, 4th International Conference, DaWaK 2002, Aix-en-Provence, France, September 4-6, 2002, Proceedings, pages 317-326, 2002.
3. UN Comtrade. https://comtrade.un.org//. 2018.
4. Eurostat: Your key to European statistics. http://ec.europa.eu/eurostat/about/overview, 2018.
5. Kargin Y, Pirk H, Ivanova M, Manegold S, and Kersten M. Instant-on scientific data warehouses - lazy ETL for data-intensive research. In Enabling Real-Time Business Intelligence - 6th International Workshop, BIRTE 2012, Held at the 38th International Conference on Very Large Databases, VLDB 2012, Istanbul, Turkey, August 27, 2012, Revised Selected Papers, pages 60-75, 2012.
6. Kepak Group. https://www.kepak.com/. 2018.
7. MCR Agri Analytics. http://www.mcragrianalytics.com/. 2018.
8. McCarren A, McCarthy S, O Sullivan C, and Roantree M. Anomaly Detection in Agri Warehouse Construction. Proceedings of the ACSW, ACM press, pages 1-17, 2017.
9. McCarthy S, McCarren A and Roantree M. An Architecture and Services for Constructing Data Marts from Online Data Sources. Insight Technical Report 2018-1. Available at http://doras.dcu.ie/22386/1/DEXA-TechnicalReport-2018.pdf April 2018.
10. Skoutas D, Simitsis A, and Sellis T. Ontology-driven conceptual design of ETL processes using graph transformations. J. Data Semantics, 13:120-146, 2009.
11. Statistics Canada. https://www.statcan.gc.ca/eng/start. 2018.
12. United States Department of Agriculture. http://www.ers.usda.gov/data-products/chart-gallery/detail.aspx?chartId=40037. 2018.
13. Xie N, Wang W, Ma B, Zhang X, Sun W, and Guo F. Research on an agricultural knowledge fusion method for big data. In Data Science Journal, 2015.
14. Zhu Y, An L, and Liu S. Data updating and query in real-time data warehouse system. In International Conference on Computer Science and Software Engineering, CSSE 2008, 2008, Wuhan, China, pp 1295-1297, 2008.