CS315: Group G10
Data Integration System

Anubhav Bimbisariye
59
11131
anubhab@iitk.ac.in
Dept. of CSE

Prashant Kumar
13
11526
prashkr@iitk.ac.in
Dept. of CSE

Indian Institute of Technology, Kanpur

Final report
11th April, 2014

Abstract

Data integration is the combining of data from multiple data sources and giving a unified view of this data. This is helpful to combine two incomplete databases and merging of databases by two organisations such as companies. It is very significant in commercial and scientific situations. It is more important with large numbers and volumes of databases. It has become the focus of extensive theoretical work, and numerous open peoblems remain unsolved. Data integration is commonly known as ”Enterprise Information Integration” (ETI).[1]

1 Introduction and Problem Statement

Given multiple heterogeneous databases, we need to build a system which shall provide a unified view of this data. Out of several approaches to do so, we have implemented Data Warehousing technique.
Data Warehousing: The warehouse system is used to extract, transform, and load selected or all data from heterogeneous sources into a single view schema so data becomes compatible with each other. This approach that the data is already physically reconciled in a single repository, so queries take less time. However, information in warehouse is not always up-to-date. Thus updating an original data source may outdate the warehouse, accordingly, the ETL process needs re-execution for synchronization. It is difficult to construct data warehouses when one has only a query interface to the data sources and access to the full data is not permitted. This is a common problem when integrating multiple commercial query services like classified advertisement or travel web applications.

In our approach, we also integrate more than one source for the same type of information for the cases where not all information is included in one source.

The problem here is being solved through manual integration because there is less amount of data as well as very few sources. For a larger scale Data Integration, the system needs to identify properly relatable data by itself, and it is not an easy task due to a variety of possible schemas and naming conventions for the same data column and different formats in which the data might be available. Morover, all sources cannot be feeded into the ETL system, so it needs to be supported by an automatic crawler for scraping the needed data from different sources accross the internet. The true magnitude of this problem is beyond scope for this course project, so we address the following issues:
1. Multiple Data sources.
2. Manual scripted integration.
3. Schema modification and interrelation.
4. Missing values integration for same data information.

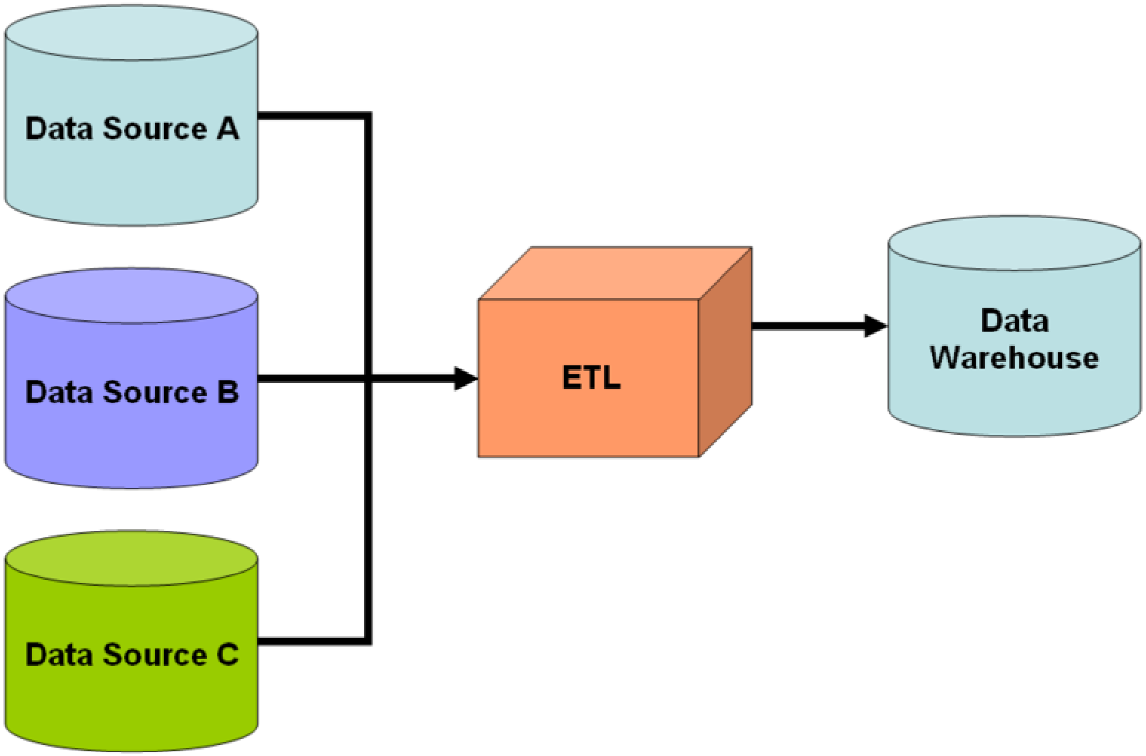

Figure 1: Simple schematic for a data warehouse. The ETL process extracts information from the source databases, transforms it and then loads it into the data warehouse..

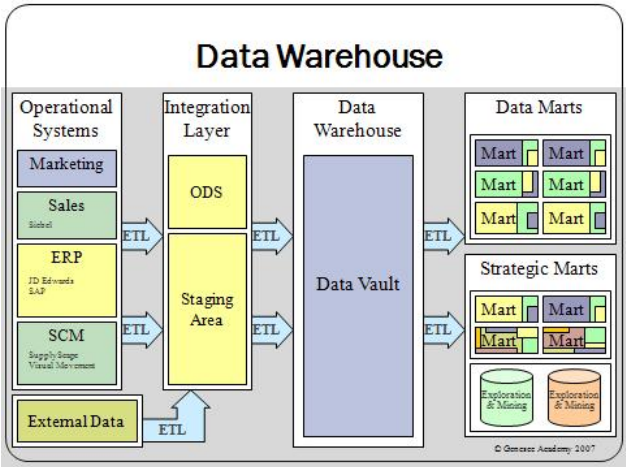

Figure 2: An example of Data Warehousing.

1.1 Related Material

Datasets from:
1. http://www.allcountries.org
2. http://apps.who.int/gho/data/node.main
3. http://data.worldbank.org/
4. Unknown.

2 Algorithm or Approach

We integrate the databases one by one into the final schema. We open the excel workbooks, and extract all the information for a tuple that we require. We do the required processing on the data like changing the format, merging the values, adding them, taking aggregate, string processing etc. Then insert the values to the relevant tables. we do this for all tuples. We use different python scripts for different modular integration. Finally we run them all using one python script. A python script to check for missing values in the warehouse and put in values from other source will be implemented.[2]
psuedo code:

```
for a database:
1. Connect to database.
2. Create database if doesn't exist(first script recreates it).
3. Create relevant tables.
4. Open Workbook.
5. Select sheet from workbook.
6. For each tuple from 1 to number of tuples in workbook
7. Extract information from relevant columns.
8. Process information, reorganise, arrange.
9. Insert into database table.

Succefully completed.
Run all scripts.
```

3 Results

Data integrates succesfully and is relatable.

Table Name	Column	Column	Column	Column	Column
ageinfo	Country	Median_Age	Life_Expectancy		
land	landarea(sq.km)	landboundary(km)	toalarea	none	
populationinfo	(0-14)years	(15-24)years	(25-54)years	(55-64)years	above-65 years
populationrateinfo	Birthrate	Deathrate	Infant mortality rate	Total literacy rate	
regiondata	country	region			

Table 1: Data warehouse schema.

4 Conclusions

Data was integrated successfully. The relations are working, data was imported successfully and is consistent with the warehouse schema.

References

[1] Data Integration, http://en.wikipedia.org/wiki/Data_integration

[2] Python Excel Support,http://www.python-excel.org/