A Study on Informatica Tool for Extract, Transform and Load processes of Data Warehouse

Nirali Shah, Vinaya Sawant

Abstract*—The relational database enabled access to the valuable information contained deep within data. Still improvements were needed for optimized complex reporting or analytical needs. This lead to the revolution of Data warehouse which would store current and historical data that could be used for creating analytical reports and analyzing hidden trends. As data is growing, companies will be required to adopt data warehouses so as to manage these huge stacks of data and hence Informatica can help them in building data warehouse. This paper focuses on ETL process of data warehouse and how Informatica is used for these processes using a case study on TAB Taxi to understand the working of this tool methodically. Informatica is a powerful tool for extracting the source data and loading it into the target after applying the required transformation.

***Index Terms*—Data warehouse, ETL, Informatica, Extract, Transform, Load etc.**

I. INTRODUCTION

Firstly, what is data warehouse? Data warehouses are special types of databases that are specifically built for the purpose of getting information out rather than putting data in. The data warehouse exists to provide answers to strategic questions and assist managers of the organizations in planning for future.

A. Features of a Data warehouse

W. H. Inmon, the father of data warehousing, defines data warehouse as 'subject-oriented, integrated, non-volatile and time-variant collection of data in support of management's decision' [1]. The following are some of the features of data warehouse:

Subject-oriented Data: The operational applications focus on the day to day transactions whereas the data warehouse is concerned with the things in the business processes that are relevant to those transactions. Every business stores data in relational databases to support particular operational systems. However, data warehouse stores data by subjects and not by applications.

Integrated Data: The main purpose of data warehouse is to store relevant data that can be used for decision making. The input to data warehouse is the operational databases which is cleansed and transformed to form a cohesive, readable environment. The tasks of Data cleaning and data transformation constitute the integration process. Data cleansing is removing of errors from the operational databases that form the input to this process. Data transformation deals with data from various sources and works towards transforming the data into a consistent format.

Non-volatile: The data present in operational databases is frequent data that varies from day to day, week to week or even once in two weeks. This means that operational environment is volatile, that is, it changes. Whereas, data warehouse is non-volatile, that is, the data remains unchanged once it is written into them. Moreover, the operations that can be performed on operational databases are read, write, update and delete. However, the only operation that is performed on data warehouse is read.

Time-variant: As a result of non-volatility, data warehouse have another *dimension*, that is, the time dimension. Managers and decision makers can view the data across the time dimension at granular levels to make decisions.

A major problem with databases is scalability, that is, that it becomes difficult to enlarge the database in terms of the size a database or it is troublesome to handle the load of concurrent users. As a result, companies have vested huge resources to incorporate data warehouses that can store millions of records and enable parallel usage by multiple users [5]. So, ETL is used widely before storing data into data warehouse as the main intension is to discover knowledgeable patterns and trends whilst decision making. In this paper, I will discuss the ETL process in detail succeeding towards Informatica tool and how it is used to perform ETL.

II. BACKGROUND

The brief insights of Extract, Transform and Load processes will be discussed in this section along with the Informatica tool. The sections is divided to cover the concepts of Dimension modelling (section A), ETL (section B) followed by introduction to Informatica tool (section B).

A. Dimensional Modelling

Just the way ER modelling is used to design a database; dimension modelling is required to design the dimensions that are nothing but subjects of a data warehouse. Dimension modelling describes the following:

1. Subject areas that are involved in building a warehouse.
2. The level of detail of data which is termed granularity.

Manuscript received November 2014.

Nirali Shah, Department of Information Technology, Dwarkadas J. Sanghvi college of Engineering, Mumbai, India.

Vinaya Sawant, Department of Information Technology, Dwarkadas J. Sanghvi college of Engineering, Mumbai, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

3. The time span of database. This is calculated by determining how much of archived data needs to be stored in a warehouse [1].

Data warehouse models can be built using three different schemas:

- Star Schema: Here, the fact table, which consists of measure, and facts, is arranged surrounded by dimensions which resemble a star.
- Snowflake Schema: This schema is very similar to star schema except that the dimensions are normalized.
- Fact constellation Schema: This schema is not used as it contains multiple fact and dimension tables that are shared amongst each other [1].

Fact tables can be classified based on the level of data that is stored:

- Detailed fact table: This store detail information about the facts.
- Summarized fact table: This are also called as aggregated fact table as they contain aggregated data.

B. ETL process

(I) why is ETL required? ETL is performed in the data staging phase of data warehouse. Data staging is an intermediate yet an important task in forming a data warehouse [1]. It is comparable to a construction site where the files are extracted from various sources, rules are examined, transformations are applied, and finally the data is cleansed.

ETL is generally performed in a separate server called staging server. Although, this adds an additional cost and complexity to building a data warehouse, it has various advantages:

- Security: As the staging area is not accessed by data warehouse users, it offers security and quality.
- This path helps in sharing load as 'data preparation' and data querying tasks are isolated and handled separately.

(II) What is ETL? ETL stands for Extract, Transform and Load functions that are used by data warehouse to populate data.

Data Extraction is responsible for gathering data from various homogenous, heterogeneous and external sources.

Data Transformation uses the data extracted and coverts this data into warehouse format.

Load just fills the target with a collection of data that is cleaned, standardized, and summarized [2], [3].

Fig. 1 summarizes the data staging phase while building data warehouse.

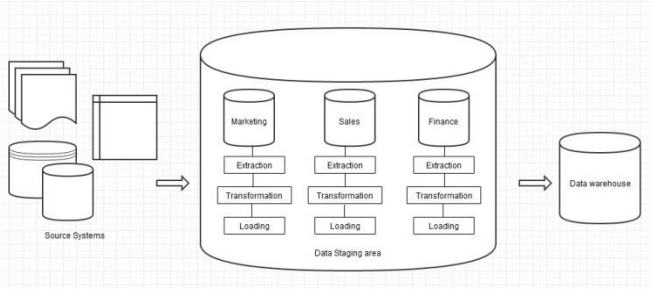

Fig.1Data moves from source to staging and finally to data warehouse

C. Informatica Interface

Informatica is a powerful tool and a widely used ETL tool for extracting the source data and loading it into target after applying the required transformation [4]. It is a successful ETL tool because easy training and tool availability has made easy resource availability for software industry; where else other ETL tools are way behind in this aspect.

As shown in Fig. 2 [8] the startup page of Informatica has repositories listed on the left side which is connected by username and password. As the repository is connected, folders could be seen. In these folders, various options are available namely Sources, Targets, Transformations, Mappings. For performing ETL, the source table should have data while the target table should be empty and should have same structure as that of source.

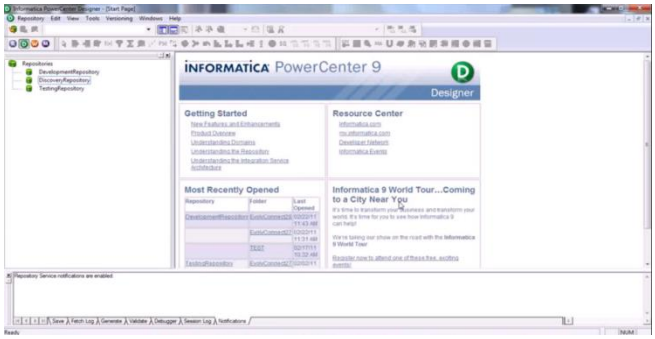

Fig.2Informatica Startup page

Steps in performing ETL using Informatica:

1. Extract: In Informatica, data can be extracted from both structured as well as unstructured sources. It can access data from the following:

- Relational Databases tables created in Microsoft SQL server, Oracle, IBM DB2, and Teradata.
- Fixed and delimited flat files, COBOL files and XML.
- Microsoft Access and Excel can also be used.

2. The source is transformed with the help of various Transformations like:

- *Expression* is used to calculate values in a single row. Example: to calculate profit on each product or to replace short forms like TCS to 'Tata Consultancy Services' or to concatenate first and last names or to convert date to a string field [7].
- *Filter* keeps the rows that meet the specified filter condition and purges the rows that do not meet the condition. For example, to find all the employees who are working in TCS.
- *Joiner* is used to join data from two related heterogeneous sources residing in different locations or to join data from the same source. Types of Joins that can be performed include Inner (Normal), Left and Right Outer join (Master Outer and Detail Outer) and Full Outer join.
- *Rank* is used to select the rank of data. Example: to find top five items manufactured by "Johnson & Johnson"
- *Aggregator* is used to summarize data with help of aggregate functions like average, sum, count etc. on

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

multiple rows or groups.

- *Sorter* is used sort data either in ascending or descending order according to a specified sort key.
- *Source Qualifier* is used to select values from the source and to create a conventional query to issue a special SELECT statement. It can also be used as a joiner. It also converts the source data types to the Informatica native data types.
- *Union* is used to merge data from multiple tables. It merges data from multiple sources similar to the UNION ALL SQL statement to combine the results from two or more SQL statements.
- *Router* is similar to filter transformation because both allow you to apply a condition to extracted data. The only difference is filter transformation drops the data that do not meet the condition whereas router has an option to capture the data that do not meet the condition.

3. Load: After transformation is complete, the final step is to load the targets. There are two types of loads that are available in Informatica:

- *Normal Load:* This type is comparatively slow as it loads the target database record by record. Also, this load writes databases logs so that the target database can recover from an incomplete session.
- *Bulk Load:* This load improves the performance as it inserts large amount of data to target database. While bulk loading, the database logs are bypassed which increases the performance [9].

As the target is loaded, let's have a look on the target types:

- Relational databases like Oracle, Sybase, IBM DB2, Microsoft SQL Server, and Teradata.
- Fixed and delimited flat file and XML.
- Microsoft access

III. CASE STUDY

In this section, we will use a case study and will follow the ETL process using Informatica.

"TAB Taxi" is a Taxi service provider offering on-hire taxi services to its customers. Taxi has a fleet of cars which includes SUV, Coupe, Sedan and Hatchback. A driver is assigned to each car and the service is charged for the Kilometers run.

Source tables:

Table1 Vehicle table in database

Attribute	*Data type*
Vehicle ID	Number
Vehicle_Type_Id	Varchar(10)
Driver_First_Name	Varchar(25)
Driver_Last_Name	Varchar(25)
Vehicle_Make	Varchar(25)
Reg_No	Varchar(25)

Table2 Vehicle_Type table in database

Attribute	*Data type*
Vehicle_Type_ID	Varchar(5)
Vehicle_Type_Desc	Varchar(20)

Table3Trip_details table in database

Attribute	*Data type*
Trip_Id	Number
Vehicle_Id	Number
Starting KM	Number
Ending KM	Number
Passenger_Name	Varchar(52)
Passanger_age	Number
Trip_cost	Number
Trip_Dt	Date

Required Dimensions:

Table4 Vehicle_Dimension in data warehouse

Attribute	*Data type*
Vehicle ID	Number
Vehicle_Type_Id	Varchar(10)
Vehicle_Type_Desc	Varchar(20)
Driver_Name	Varchar(51)
Vehicle_Make	Varchar(25)
Reg_No	Varchar(25)

Table5 Trip Dimension in data warehouse

Attribute	*Data type*
Trip_Id	Number
Vehicle_Id	Number
Starting KM	Number
Ending KM	Number
Passenger_Name	Varchar(52)
Passanger_age	Number
Trip_cost	Number

Table 6 Time_Dimension in data warehouse

Attribute	*Data type*
Time_Key	Date
Day_of_week	Varchar (10)
Month	Varchar (10)
Quarter	Varchar (2)
Year	Number

Table 7 Trip_Luxuary_Fact in data warehouse

Attribute	*Data type*
Trip_Id	Number
Vehicle_Id	Number
Trip KM	Number
Passenger_Name	Varchar(52)
Passanger_age_Group	Number
Trip_cost	Number
Trip_Dt	Date

Passanger_age_Group is calculated as bands 20 -30, 30-40, 40-50, 50-60 and senior citizens.

Trip_Dimension should have data starting from Trip_Dt 01-01-2005.

A. *Dimensional modelling*

The star schema for the above case study is as shown in Fig. 3 [6]:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

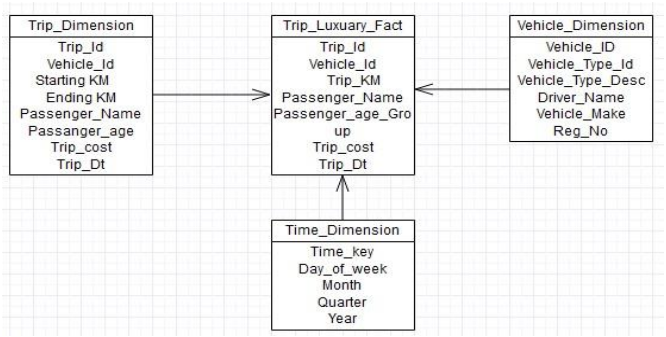

Fig.3Star schema for TAB Taxi

B. *ETL processing in Informatica*

The mappings in Informatica would be:

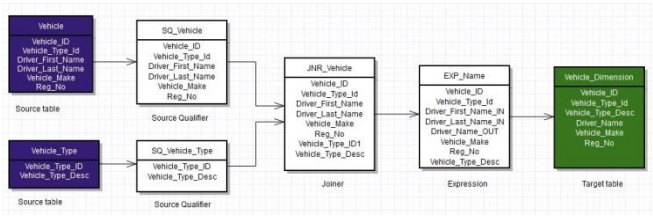

Fig.4Loading Vehicle Dimension

The prototype of transformations used in TAB Taxi is as shown in Fig. 4 [6] is:

- Joiner: To inner join tables Vehicle and Vehicle_Type based on attribute Vehicle_Type_Id.
- Expression: To concatenate Driver_First_Name and Driver_Last_Name into Driver_Name. This is done using *"Driver_First_Name_IN || " " || Driver_Last_Name_IN"*.
- Source Qualifier: To convert the source data types like Varchar (25) to Informatica data types like String.

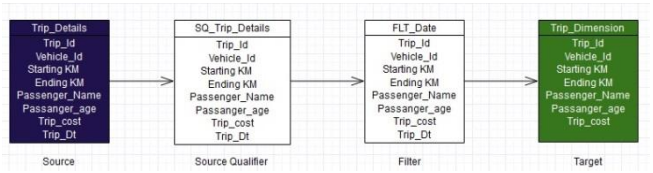

Fig.5Loading Trip Dimension

The prototype of transformations used in TAB Taxi is as shown in Fig. 5[6] are:

- Source qualifier: To convert the source data types like Varchar (25) to Informatica data types like String.
- Filter: To filter out data that is before Trip_Dt 01-01-2005.

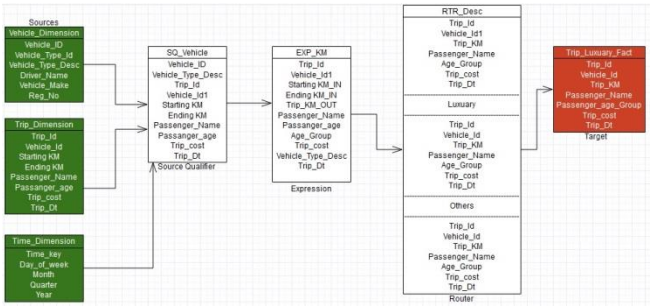

Fig.6Loading Fact table

The prototype of transformations used in TAB Taxi is as shown in Fig. 6[6] are:

- Source qualifier: Used as joiner to inner join tables Vehicle_Dimension with Trip_Dimension and Trip_Dimension with Time Dimension.
- Expression: To calculate Trip_KM using formula "Ending KM – Starting KM"
  Also used to form age bands using formula
  IIF(Passenger_age>= 20 AND Passenger_age< 30, "20-30",
  IIF(Passenger_age>= 30 AND Passenger_age< 40, "30-40",
  IIF(Passenger_age>= 40 AND Passenger_age< 50, "40-50",
  IIF(Passenger_age>= 50 AND Passenger_age< 60, "50-60", "Senior citizen")))).
- Router: To filter rows whose "Vehicle_Type_Desc = 'Luxury'".

The fact table "Trip_Luxury_Fact" is detail fact table as it contains details information about the trip taken by a passenger.

IV. Conclusion

In this paper, concepts of Data warehousing were discussed along with the ETL process. Thereafter, the Informatica tool was introduced followed by the steps involved in ETL. As discussed, the advantages of using this tool are many fold. For one, Informatica is user friendly because of which it becomes easy to understand and use. Also, Informatica has its capability of enabling Lean Integration[1] so that no resource is wasted.

Acknowledgment

We would like to acknowledge the department of Information Technology of Dwarkadas J. Sanghvi College of Engineering for their encouragement and support in writing this paper.

References

[1] ReemaThareja's book on "*Data Warehouse*" published by Oxford Higher Education.
[2] RamezElmasri, Shamkant B. Navathe, RajshekharSunderraman book on "*Fundamentals of Database Systems*" 4th Edition, published by Addison Wesley Longman
[3] Jiawei Han, MichelineKamber and Jian Pei book on "*Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann*

[1] Lean Integration is to underscore continuous improvements and it stresses on avoiding waste.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Series in Data Management Systems)" published by Morgan Kaufmann.
[4] Web page on "Informatica" by Wikepedia.
[5] SwastiSinghal, Monika Jena paper on "*A Study of WEKA tool for Data Preprocessing, Classification and Clustering*" published by IJITEE, Volume-2, Issue-6.
[6] Website named "Gliffy: Online Diagram Software and Flow Chart Software". Available: *www.gliffy.com*
[7] Web page on Informatica Tutorial. Available: *http://www.techtiks.com/informatica/beginners-guide/transformations/transformation-types/*
[8] Muhammad Abbas video on "Informatica Tutorial For Beginners". Available: http://www.youtube.com/watch?v=ufH_n5exxQw.
[9] A tutorial on Target load types. Available: *http://www.javaorator.com/informatica/interview/Target-Load-Type-in-informatica-42.code*

Nirali Shah received a degree in B.E. in Information Technology from university of Mumbai and currently working at TCS as a Developer. Her research interest includes Data warehouse and Data mining.

Vinaya Sawant received a degree in M.Tech (Computer Engineering) and currently working at Dwarkadas J. Sanghvi college of Engineering as an Assistant Professor. Her research interest includes Data warehouse and Data mining.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.