Data Warehousing
Final Project Report

By

MOHAMED NIYAZ

Under the guidance of

Wensheng Wu, **Ph.D.**

Professor and Associate Chair

Department of Computer Science, College of Computing and Informatics

University of North Carolina at Charlotte

2013

Introduction:

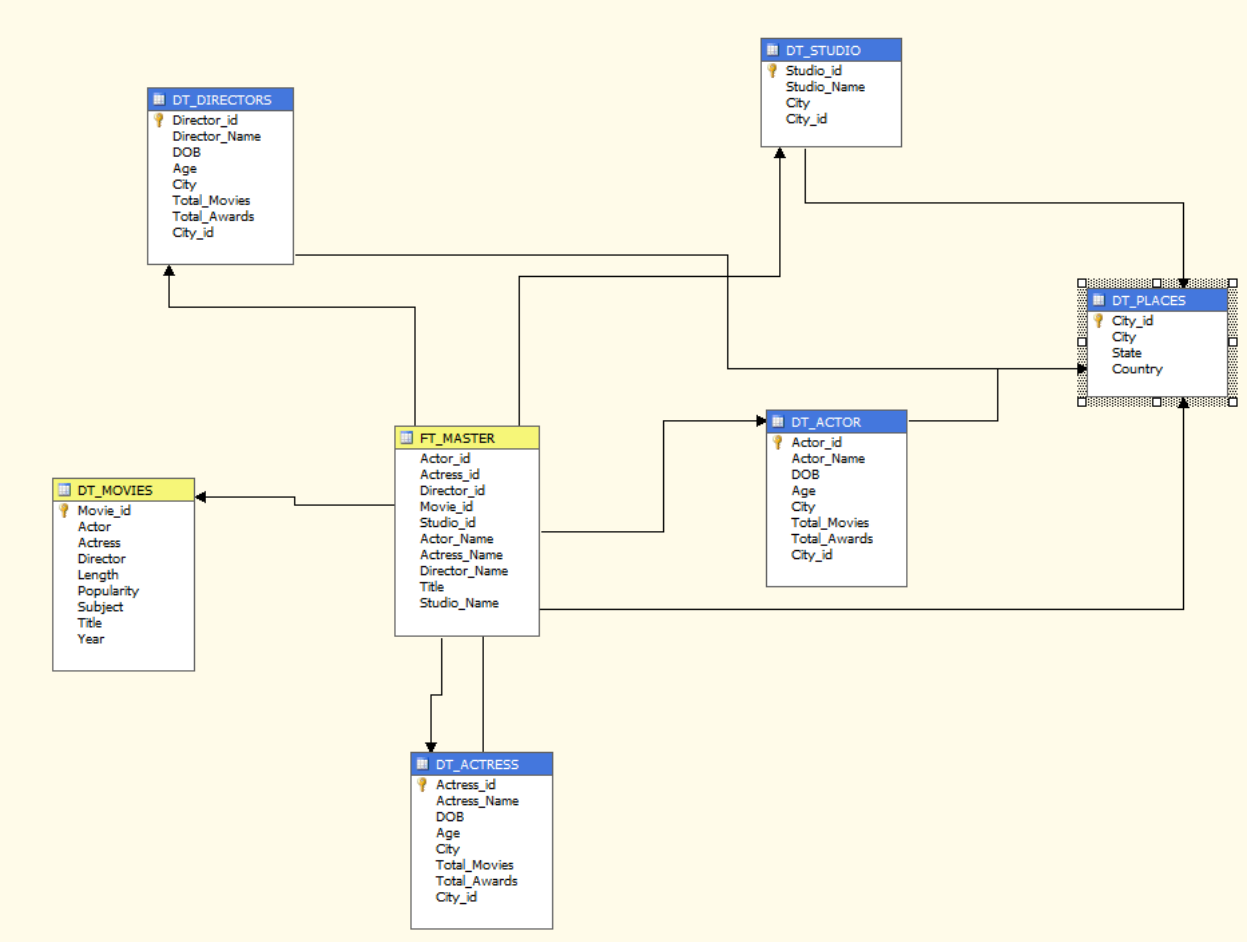
In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting.

The typical ETL-based data warehouse uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.

A data warehouse constructed from integrated data source systems does not require ETL, staging databases, or operational data store databases. The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. Unlike the ETL-based data warehouse, the integrated source data systems and the data warehouse are all integrated since there is no transformation of dimensional or reference data. This integrated data warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated source data systems.

Project Final Design and Architecture:



Schema Type: Snow Flake Schema

Staging Table

STG_Actor, STG_Actress,

STG_Director, STG_Movies,

STG_Places,STG_Studio

Fact Table

FT_Master

Dimension Table

DT_Actor, DT_Actress, DT_Director,

DT_Movies, DT_Places,DT_Studio

Project Accomplishment:

The primary responsibility of this project is to create a Data Warehouse and to implement different data warehousing strategy.

1. **ETL**

The term ETL stands for Extract, Transform and Load. ETL process involves extracting the data from the source systems. In many cases this is the most challenging aspect of ETL, since extracting data correctly sets the stage for how subsequent processes go further.

We created a movie data warehouse model which appropriately defining all the functionalities of a Data Warehouse. Design and integrate a Data Warehouse, the schema focusing on the basic functionalities like insertion, deletion and updating the respective records.

Source : Movie, Internet, Wikipedia.

File type : Comma Separated files

Extraction : Files are extracted using MS SSIS.

Transformation : Files are cleansed and parsed through a set of validation processes. On completion of the task a cleansed file is create and moved to the inbound directory.

Loading : The Files are picked from the inbound directory and loaded into the Microsoft SQL database through the built in Sql services

2. **OLAP**

Once the warehouse is created, business intelligence process is carried out, in computing online analytical processing, or OLAP, is an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture. The term OLAP was created as a slight modification of the traditional database term OLTP (Online Transaction Processing).

In order to obtain the OLAP model we have to design and develop dimensions, these dimensions are deployed together to form a cube. The dimensions in our project are Actor, Actress, Director,

Movies, Place and Studio hierarchy. These dimensions are used to create the cube master (Movies cube).

The Operations carried out in the cube are:

Roll-up **:** A roll-up involves summarizing the data along a dimension. The summarization rule might be computing totals along a hierarchy or applying a set of formulas such as movie rating = Total Rating/Number of Votes.

Drill-Down **:** Drill Down allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down).

Slicing : Slice is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension. Finding out the movie that released in the year of 2004 are "sliced" out of the data cube.

Dicing **:** The dice operation produces a sub cube by allowing the analyst to pick specific values of multiple dimensions. The new cube will show limited number of movies, the time and region dimensions cover the same range as before.

3. **Data Mining**

Data mining consists of a set of statistical techniques for analyzing observational data. The techniques are employed in the analysis step of the "Knowledge Discovery in Databases" process, or KDD. These tools discover patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Our Data Mining process was carried out by clustering the data, the scenario we followed is that the movie popularity system used by the IMDB uses a simple 0-100 scale. These values are given definite values for example 5, 10 where 0 being the minimum and 100 being the maximum. Therefore, in order to better understand this popularity system we want to cluster the movies by popularities. This will allow us to find breaking points between clusters and use these

to apply labels to certain number ranges. In other words we hope to use clustering to identify clusters of movie ratings and assign values to each cluster.

We also used the "Popularity" to find the "Titles" from the Movies table which has the maximum, minimum and the average popularities. In the same way we also used the "Popularity" to find the "Director" from the Directors table which has the maximum, minimum and the average popularities.

***Algorithm*:** Clustering

Fact Table (Nested Table) – Master.

Dimension Table (Case Table) – Director, Movies.

Predicate Value – Popularities.

Input – Actor Name, Director Name.

Key – Movie ID.

Project Plan:

All together we had 3 different phases in data warehousing. The project was planned to be used Informatica for ETL, Business Objects for OLAP and Weka for mining. The projects carried on with perception of implementing the warehouse in SQL Server and further study was done on remaining tools in the market. We finally decided to move forward with a Microsoft Business Intelligence tool. The tool was highly portable, reliable and had multi-lingual features embedded with itself. The tool had the capability to perform ETL with is in built feature named Microsoft Sql Server Integration Service(SSIS), reporting done through Microsoft Sql Server Reporting Service(SSRS) and the OLAP (Cube creation) , data mining are carried through Microsoft Sql Server Analysis Service(SSAS).

1. ETL

As planned our ETL Process is carried through Microsoft SSIS tool. We had our source to be a CSV file, these files were extracted through an inbuilt feature in the tool. Once after extraction the data's are cleansed through various transformations available in the tool. Cleansing had several phases, first we loaded the data into a temporary table named staging, after then data are

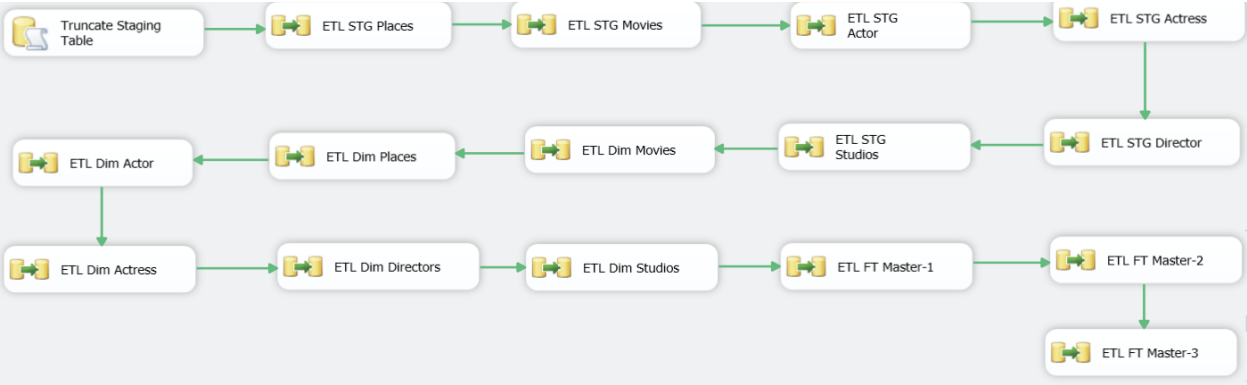
retrieved from the staging table recursively, various business validation and process are carried on the data's. The data's are finally loaded on completion of the business process.

Source : CSV File

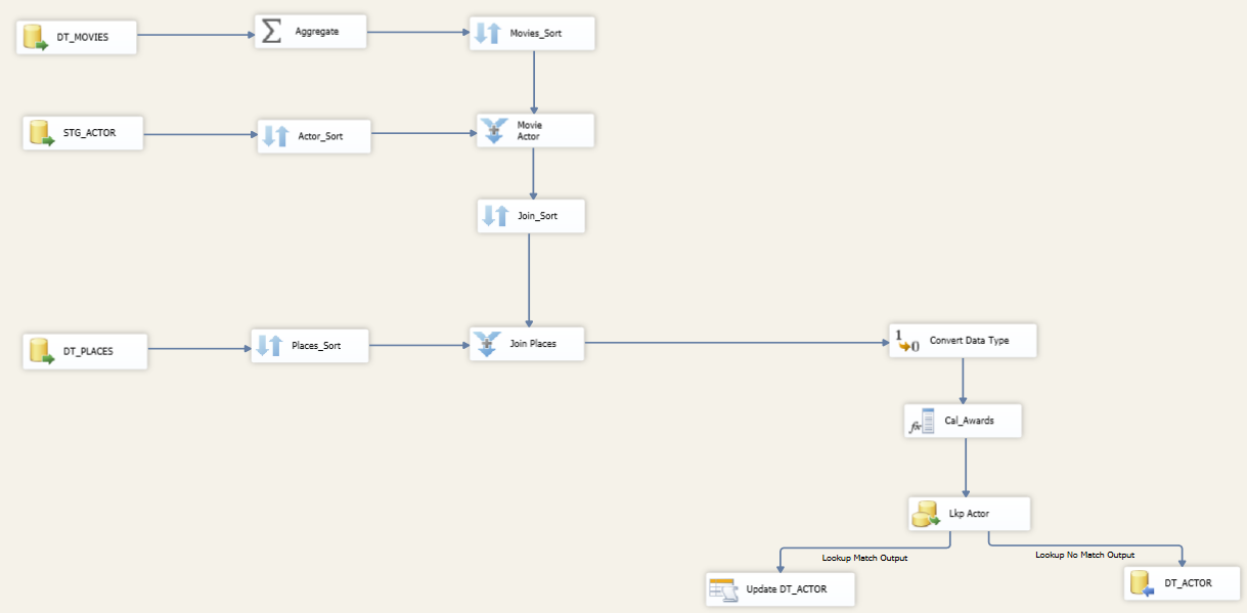
Transformations : Expression, Data Conversion, Execute Sql and many more.

Target : A table in Sql Server

High level overview of the Control tasks.

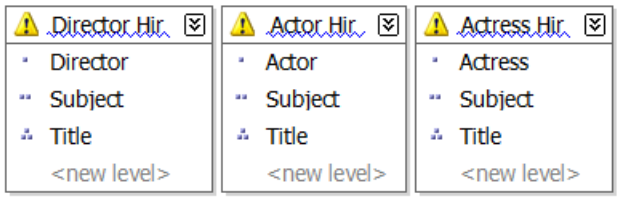


Internal view of a data task



2. OLAP

We planned to create a cube using dimensions, so we set our primary task to create a dimension. But we need to have a hierarchy to create a dimension so we planned to find the relational hierarchy between the data's. Our data set was Movies so we created 3 different movie hierarchy for a movie dimension as shown below



Similar to the above one, we created different dimension for different attributes. Having all the dimensions in place we created the cube. Below Screen shot depicts the working of the cube.

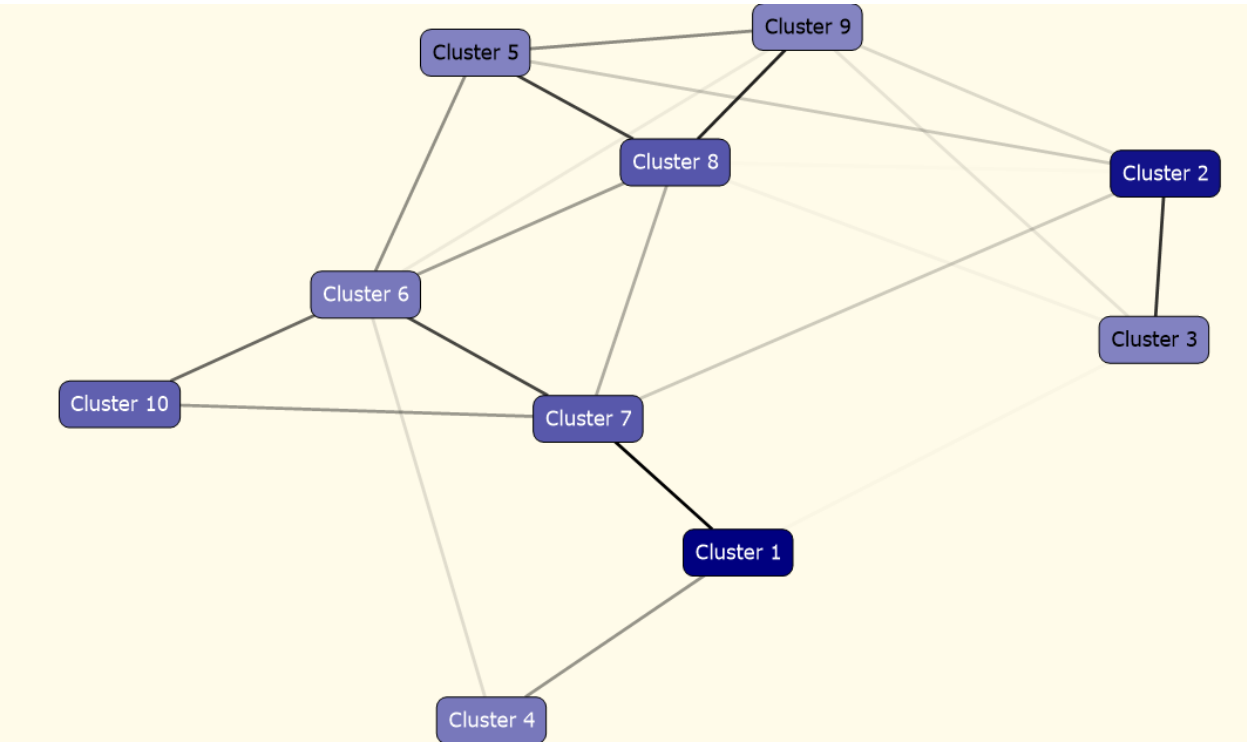
	Row Labels	FT MASTER Count
2	1920	1
3	⊞BUSTER KEATON	1
4	1924	2
5	⊞CHARLES CHAPLIN	1
6	⊞JAMES COBURN	1
7	1925	1
8	⊞LEE MARVIN	1
9	1926	3
10	⊟GENE KELLY	1
11	⊟Science Fiction	1
12	Metropolis	1
13	⊞LON CHANEY JR.	1
14	⊞MALLIKA SHERAWAT	1
15	1927	4
16	⊞FRED MACMURRAY	1
17	⊞JAMES MASON	1
18	⊞KUNAL NAYYAR	
19	⊞LOUIS DE FUNÃ¨S	
20	1928	4
21	⊞ANTONIO MORENO	1
22	⊞ARCHIE PANJABI	1
23	⊞RICHARD BURTON	1
24	⊞ROGER LIVESEY	1
25	1929	4

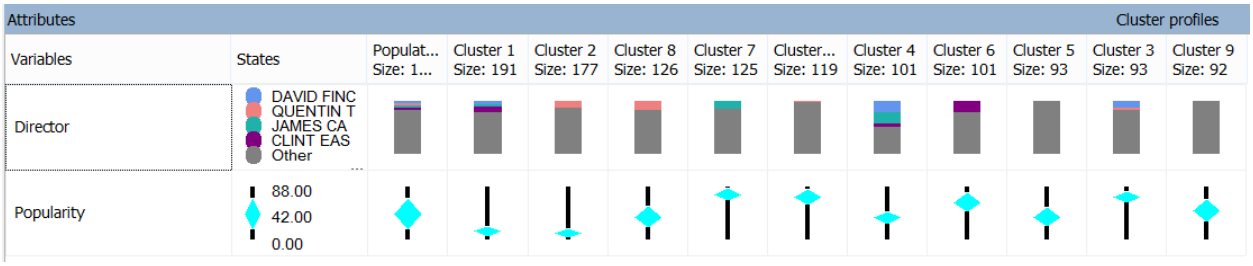
FT MASTER Count
Value: 1
Row: 1927 - JAMES MASON

3. Data Mining

We used the movie popularity system used by the IMDB uses a simple 0-100 scale. These values are given definite values for example 5, 10 where 0 being the minimum and 100 being the maximum. Therefore, in order to better understand this popularity system we want to cluster the movies by popularities. This will allow us to find breaking points between clusters and use these to apply labels to certain number ranges. In other words we hope to use clustering to identify clusters of movie ratings and assign values to each cluster. We also used director to find the correlation between the movies directed by a director which has the highest, average and minimum popularities. So ultimately we find out the director who has the highest popularity.

The planned mining results,





Project Design Specification:

1. ***Software Design Specification***

   ETL: Sql server Integration service

   OLAP, Mining: Sql Server Analysis Service

   Database: Sql Server Management Studio

2. ***Hardware Design Specification***

   Name : Asus System

   Ram : 16 GB DDR3

   Graphics : 2 GB DDR5

   Processor : i7-3630 QM

   OS : Windows 8.

Work Carried out:

Mohamed Zakriea Niyaz: Primary responsibility was to create a database design, initially created a star schema and when the requirement groomed we created a snow flake schema. In ETL , he made sure there are no redundancy in the database and also database follows the parent child relationship which is likely the primary – foreign key relationship. In OLAP the responsibility is to find the relational hierarchy among the data's and make sure the table doesn't violate any relationship. In Mining to find the Mining algorithm which best suits our data.

In ETL, wrote queries to transform data and loaded it to the database, Warehouse design, In OLAP wrote queries for slicing and dicing, prepared documents, contributed to Phase4.Contributed to all the phases

Downloaded data from IMDB website, did R & D on mining algorithms, Understood OLAP concepts and explained it to team, contributed to Phase4, Prepared documents. Exported it to Excel, Warehouse design. In OLAP wrote queries to Roll Up and Drill Down, Installed all the software's required for our project, prepared documents. Contributed to all the phases.

Issues and Concerns:

- OLAP operations can be performed only the cleanse data, so getting the cleanse data was one among the challenges.
- OLAP operations have to be performed on fact and dimension tables, forming multiple relations and hierarchy was really challenging.
- Forming the cuboid was the biggest challenge among all these, we have to make sure the entity relation is perfectly right and data are fully cleansed.
- Data Mining operations can be performed only the cleanse data, so getting the cleanse data was one among the challenges.
- Data mining operations have to be performed on fact and dimension tables, forming multiple relations and hierarchy was really challenging.
- Forming the clusters was the biggest challenge among all these, we have to make sure the entity relation is perfectly right and data are fully cleansed.
- Finding the best relation between the nested table and the case table was really hard.

Comments:

We are really privileged to say that we understood the theoretical and practical working model of warehouse. We really appreciate the efforts put forward by Wensheng Wu and Tingting to guide us in every phase and made us understand and pushed us to work on these demanding technologies and making them get well acquainted.