

Data Warehousing Concepts Using ETL Process for Social Media Data Extraction

Rohita Yamaganti, Usha Manjari Sikharam

Abstract— The importance of using social media has increased enormously and focus of the software analyst has shifted towards analyzing the data available in these social media sites. For Example if a college wants to retrieve the alumni data from the social networking sites. Yes it can be done by using power tool called ETL Tool that is available for analyzing the data. In order to analyse the data it's important to have cleaned or preprocessed data. This preprocessed data helps us in retrieving the desired data. In this process, we need to create special type of database that is specifically built for the purpose of getting information out rather than putting data in. Data warehousing concepts using ETL process trying to build the Data warehouse. Not only alumni data can get the list of friends along with their bio-data, list of employees who are working in different industries. The data Warehouse exists to provide answers to strategic questions and assist managers of the organizations in planning for future. The beauty of creating Data Warehouse will enable the user to analyse the data. We use powerful Tool called Informatica to create this warehouse from social media.

Index Terms—Data warehouse, ETL, Informatica, Extract, Transform, Load, Preprocess. 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:

A.1 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.

A.2 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.

A.3 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.

A.4 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.
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:

1. Security: As the staging area is not accessed by data warehouse users, it offers security and quality.
2. 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 converts 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.

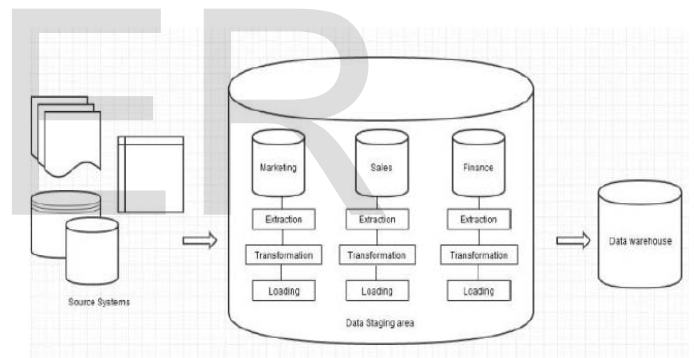


Fig. 1 Data moves from source to staging and finally to data warehouse

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.

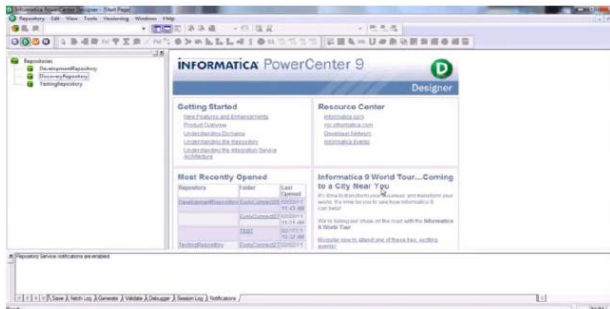


Fig.2 Informatica Startup page

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 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].

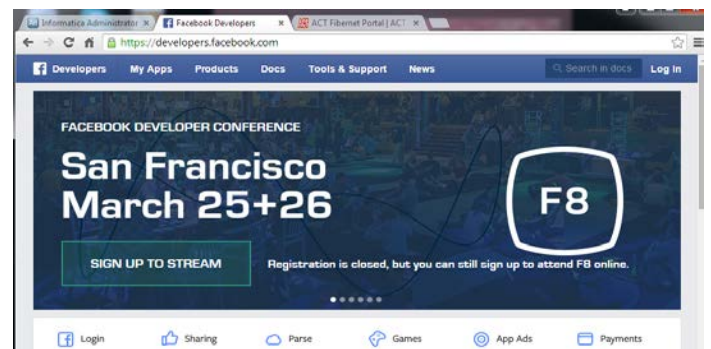
As the target is loaded, let's have a look on the target types:

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

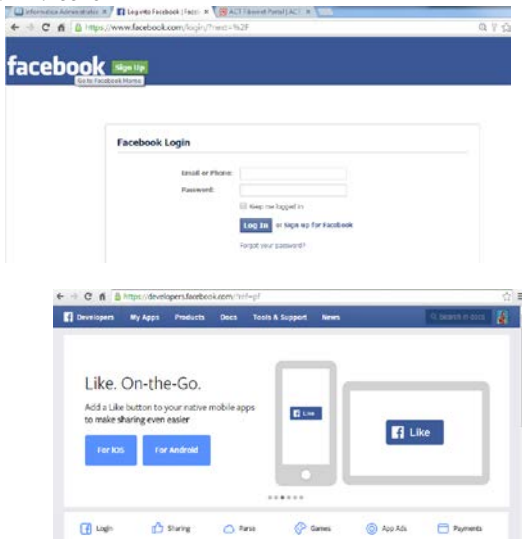
III Case Study:

We will show the implementation of using facebook.

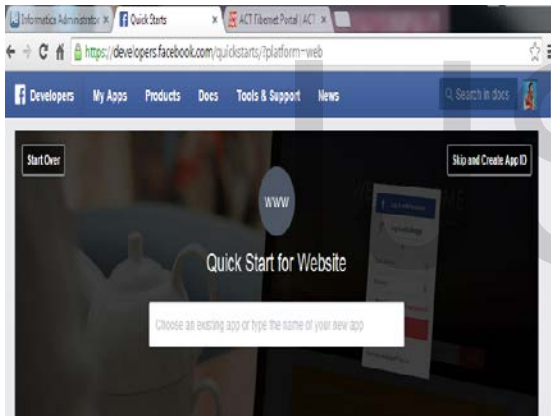
1. Firstly we have to create facebook app through <https://developers.facebook.com/>



1. facebook account



2. Create an application
- Go to my app -> add a new app
3. Select www
4. Next click the option (skip and create a app ID)



5. Next fill the details

Create a New App ID
 Get started integrating Facebook into your app or website

Display Name:

Namespace:

Category:

By proceeding, you agree to the Facebook Platform Policies



6. App id is consumer id and app secret is consumer secret. Click on show to create App Secret.

App ID: App Secret:

Display Name: Namespace:

App Domains: Contact Email:

7. Next go to settings
8. Next enter the email and save the changes.
9. Open Informatica OAuth Utility using a URL: **domain-name: informatica port number/ows/**
10. Enter the consumer key and secret.

Informatica Administrator | SNIST | ACT Fibernet Portal | ACT | Informatica OAuth Utility

usha-pc7009/ows/

INFORMATICA OAuth Utility
 OAuth Utility for Social Media Adapters

Application:

Do not have consumer key and secret? [Click here](#)

Consumer ID:

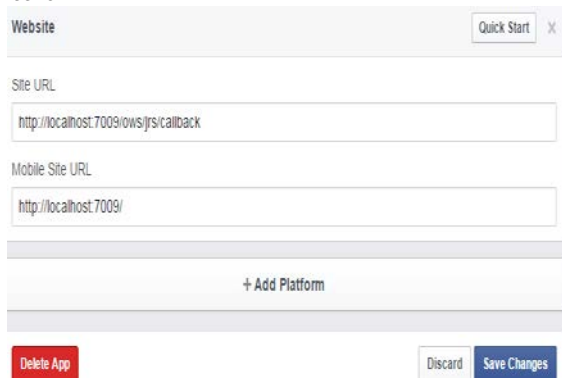
Consumer Secret:

OAuth Callback URL:

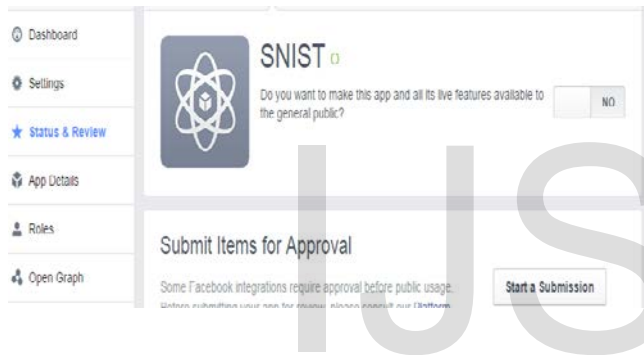
Access Token:

Access Secret:

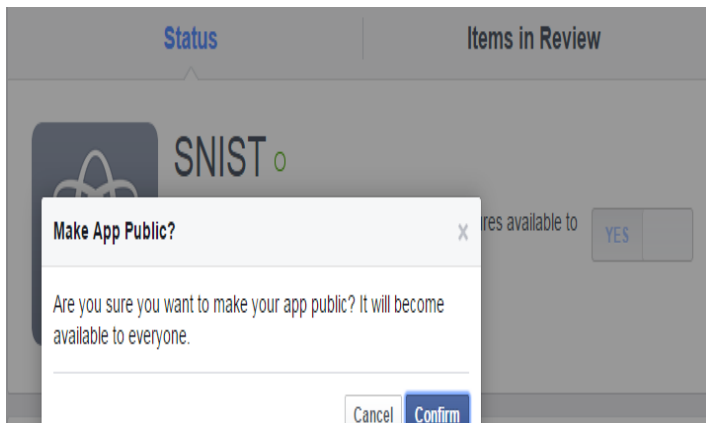
11. Copy the default callback URL .and go to your created facebook application.
12. Click Add Platform.



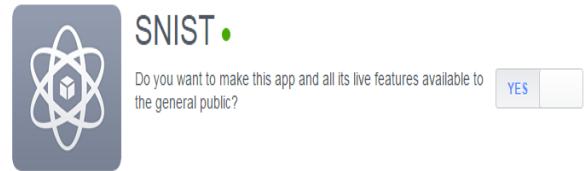
- 13. Select website. And paste the OAuth callback URL
- 14. And enter the mobile site URL.
- 15. Click save changes
- 16. Next go to status and review.



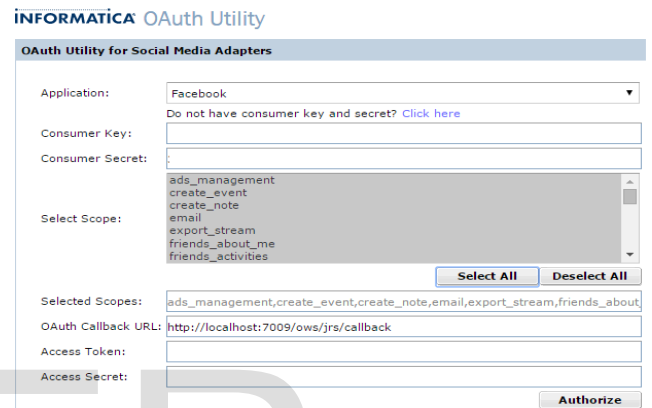
- 17. Select yes to make your app in live mode



- 18. Click confirm.
- 19. Now the application is in live mode.



- 20. Next go to OAuth and fill the required details



- 21. Click Authorize.

SNIST will receive the following info: your public profile, email address, custom friends lists, messages, News Feed, relationships, relationship interests, birthday, work history, status updates, education history, events, groups, hometown, interests, current city, photos, religious and political views, videos, website, personal description, likes and games activity.

[Edit the info you provide](#)

This does not let the app post to Facebook.

You are using a display type of 'page' in a small browser window or popup. For a better user experience, show this dialog with our JavaScript SDK without specifying an explicit display type. The SDK will choose the best display type for each environment. Alternatively, use display type 'popup' if you have special requirements precluding you from using the SDK. This message is only visible to developers of your application.

Cancel Okay

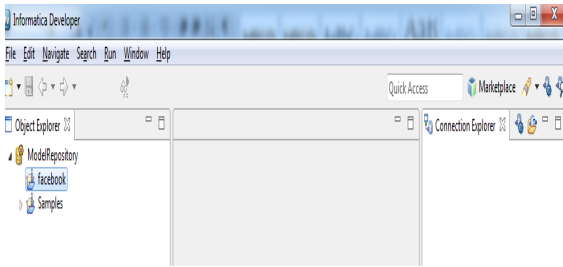
- 22. Click Okay.
- 23. We have successfully connected informatica to social media.

INFORMATICA OAuth Utility

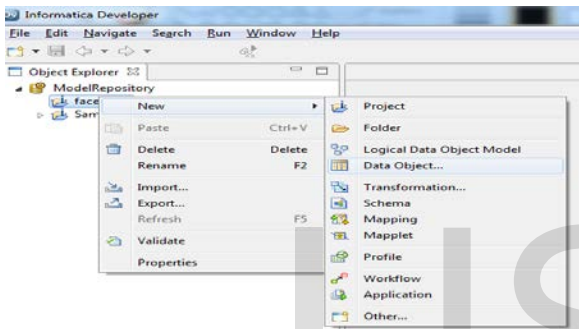
OAuth Utility for Social Media Connectors

Authentication Successful !
You may close the window.

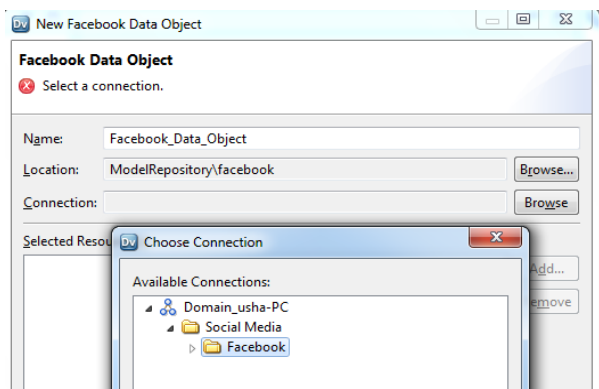
- 24. Close the window.
- 25. Next go to Informatica developer.
- 26. Create a new project.



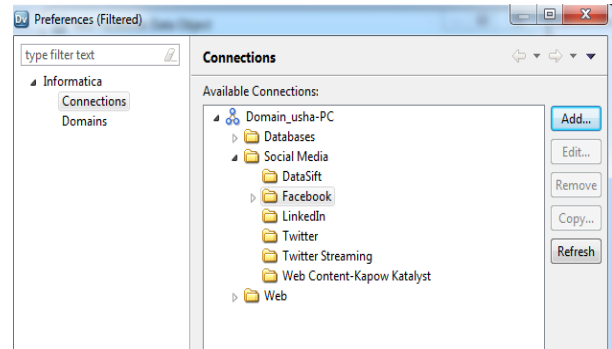
- 27. Next right click to project Name next select new and next select Data Object.



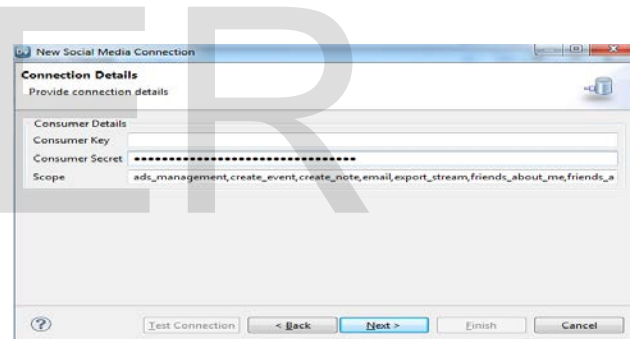
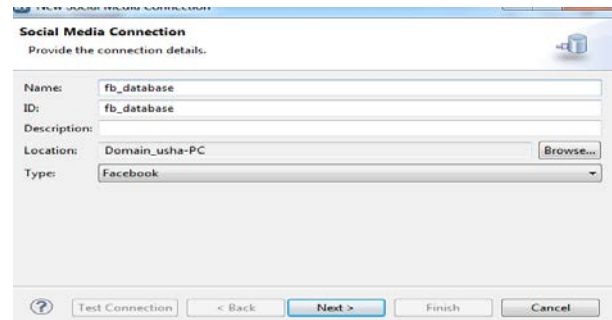
- 28. Next select facebook data object.
- 29. Next connections click browse button and select facebook next click more option.



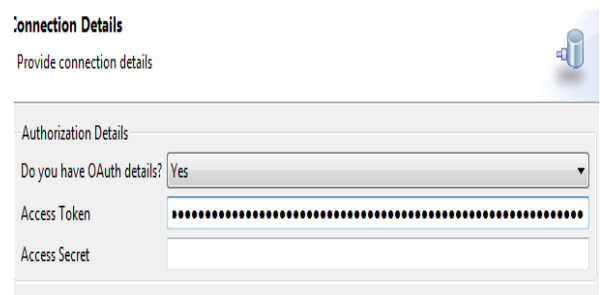
- 30. Double click Domain and select social media->facebook next click add button.



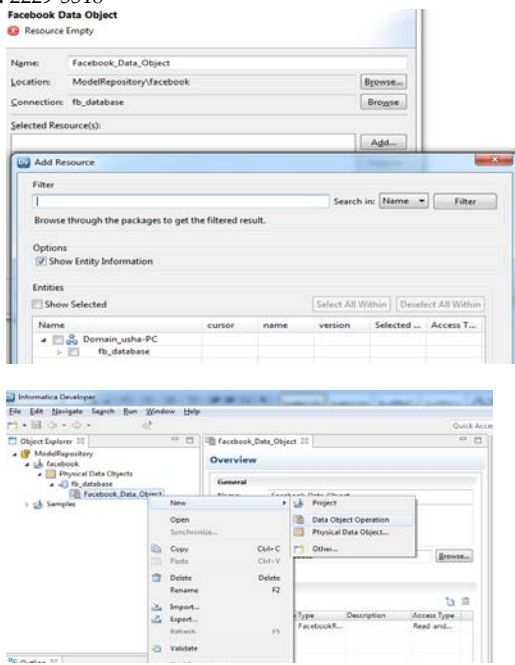
- 31. Fill the details click next button



- 32. Enter app id & secret and next click.



- 33. Next OAuth Access key copy and paste this window.
- 34. Next click test connection. And click finish.



Output

Name: Application Data Object Operation

actions	applic...	appl...	appl...	ca...	co...	created_time	description	from_category	from_id	from_name	icon
36	["link"/https://www.fac...	["u...	2015-03-07T...	Internet/softw...	9030753203	Informatica C...	https://www...				
37	["link"/https://www.fac...	["u...	2015-03-07T...	Internet/softw...	9030753203	Informatica C...	https://www...				
38	["link"/https://www.fac...	de...	2015-03-07T...	Your Sales, M...	Internet/softw...	9030753203	Informatica C...	https://www...			
39	["link"/https://www.fac...	["u...	2015-03-07T...	Internet/softw...	9030753203	Informatica C...	https://www...				
40	["link"/https://www.fac...	["u...	2015-03-06T...	Internet/softw...	9030753203	Informatica C...	https://www...				
41	["link"/https://www.fac...	["u...	2015-03-06T...	Internet/softw...	9030753203	Informatica C...	https://www...				
42	["link"/https://www.fac...	["u...	2015-03-05T...	Internet/softw...	9030753203	Informatica C...	https://www...				
43	["link"/https://www.fac...	tab...	2015-03-05T...	At Tableau we...	Internet/softw...	9030753203	Informatica C...	https://www...			
44	["link"/https://www.fac...	["u...	2015-03-04T...	Internet/softw...	9030753203	Informatica C...	https://www...				
45	["link"/https://www.fac...	["u...	2015-03-04T...	Internet/softw...	9030753203	Informatica C...	https://www...				
46	["link"/https://www.fac...	["u...	2015-03-04T...	Internet/softw...	9030753203	Informatica C...	https://www...				
47	["link"/https://www.fac...	["u...	2015-03-04T...	Internet/softw...	9030753203	Informatica C...	https://www...				
48	["link"/https://www.fac...	["u...	2015-03-03T...	Internet/softw...	9030753203	Informatica C...	https://www...				
49	["link"/https://www.fac...	["u...	2015-03-03T...	Internet/softw...	9030753203	Informatica C...	https://www...				
50	["link"/https://www.fac...	["u...	2015-03-02T...	Facebook/Infor...	9030753203	Informatica C...	https://www...				

IV. CONCLUSION

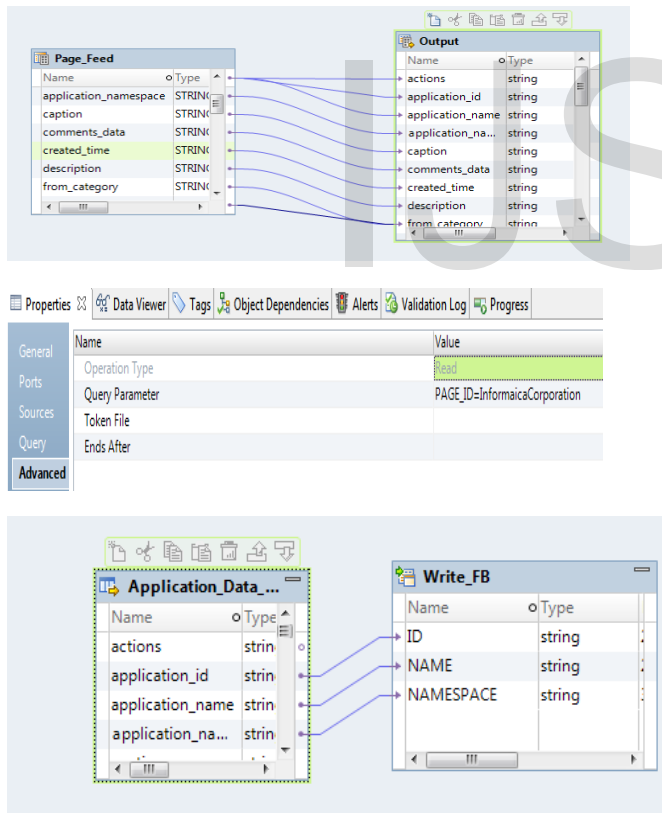
Data ware housing concepts along with the ETL processes are covered in depth in this paper. This paper gave an insight to the various aspects and steps in the ETL tools and also explains the user friendly nature of the tool. As discussed in the paper, the advantages of using Informatica are many fold. Informatica is very user friendly as it becomes easy to understand and use. Also, Informatica has its capability of enabling Lean Integration so that no resource is wasted during the process. We have implemented the complete steps in the Informatica; the results (screen shots) are attached to this paper. The paper will surely help in understanding the ETL tool and helps the researcher to create their own data warehouse.

ACKNOWLEDGMENT

We would like to acknowledge the department of Computer Science and Engineering of Sreenidhi Institute of Science and Technology 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).
- [4] Web page on "Informatica" by Wikipedia.
- [5] SwastiSinghal, Monika Jena paper on "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



Next run mapping.

Output:

- [7] Web page on Informatica Tutorial. Available:
<http://www.techtiks.com/informatica/beginners-guide/transformations/transformation-types/>
- [8] Muhammad Abbas video on “Informatica Tutorial ForBeginners” 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>

Rohita Yamaganti received a degree in M.Tech in Computer Science and Engineering from university of JNTUH and currently working as Assistant Professor in Sreenidhi Engineering College research interest includes Data warehouse and Data mining.

Usha Manjari Sikharam received a degree in M.Tech Computer Science and Engineering from university of JNTUH and currently working as Assistant Professor in Sreenidhi Engineering College. Her research interest includes Data warehouse and Data mining.

IJSER