PROJECT REPORT

DATA WAREHOUSING AND BUSINESS INTELLIGENCE

DATA WAREHOUSE FOR H1B VISA PETITIONS FOR EMPLOYERS ACROSS THE UNITED STATES OF AMERICA

Project Guide
Dr. Pravin Metkewar

Submitted by
Sarvesh Prajapati
PRN 17030141070, MBA-IT, SEM-3
sap1741070@sicsr.ac.in
SICSR

TABLE OF CONTENTS

INTRODUCTION

Aim: To build a 'Data warehouse' for finding the human being diseases with respective to distribution of medicine in different region.

Objectives:

1. To identify which states across the U.S. have shown maximum increase in filing of H1B visa petitions.
2. To identify the job position that sees the most number of visa petition filing.

1. DATA COLLECTION

The data collected for this project is regarding the filing of H1B visa petitions for different job profiles offered by employers across the USA.

Data collected contains:

- Employer details consists of employer name, location
- Job profiles
- Case status of visa petitions filed over the years 2011 through 2016.

The motivation behind this project is to find out the fluctuations in filing of visa petitions across different job profiles.

2. DATA PREPROCESSING

Data pre-processing is a process of transforming raw data into meaningful data. Data collected may be of wrong format, may have errors, may have missing values, repeated values, may have incompatible values etc. Preprocessing helps resolve all these problem.

Data pre-processing consists of 4 steps:

2.1 Data Cleansing

2.2 Data Staging

2.3 Data Integration

2.4 Noise Removal

2.1 Data cleansing

Data cleansing is the process of identifying and correcting or removing the inappropriate records from the datasets. Data cleansing can be done by adding related information, to make datasets more complete. In this project data cleansing is done for some records. Data cleansing may be of different types. For instance, it may just be correction of misspellings of resolution of conflicts between state codes and zip codes in source data.

2.1.1 Filtering

Instead of some integer values, some columns had “NA”, a string, as its value which is not allowed. These values were deleted.

2.1.2 Spell checking

Spell Checker is used to find misspelled words and also to find out words which are spelled differently at different places such as names of country and products. Misspelled words are corrected using spell checker and inconsistently spelled words are added to a custom dictionary. Thus issues with spelling are resolved.

2.1.3 Removing duplicate rows

Duplicate rows are those which are repeated in the datasets. Due to duplication of rows, uniqueness of values where lost. To filter unique values “Sort and Filter” option is used. To remove duplicate values, “Remove duplicates” command in data tools is used.

2.2 Data Staging

The external data coming from several disparate sources needs to be converted and made ready in a format that is suitable to be stored for querying and analysis. The staging area involves major functions of extraction, transformation and preparation for loading. Data staging provides a place and an area with a set of functions to:

- Clean
- Change
- Combine
- Convert
- Deduplicate

2.3 Data Integration

Data integration is the process in which heterogeneous data is collected and combined in the required form. In this project data has been collected in different formats like CSV. After data integration it is stored as xlsx file.

2.4 Noise Removal

There were discrepancies in dimension ‘SOC_Name’ values (somewhere it was COMPUTER SYSTEMS ANALYST and in other records it was COMPUTER SYSTEMS ANALYSTS. It was corrected to COMPUTER SYSTEMS ANALYST).

3. DATA FORMATS

Data formats defined for each dataset defines the procedure to work over it. All data were initially available in CSV format which was converted to xlsx format. Data formats used are:

- CSV
- XLS

CSV

CSV is comma separated values. CSV files allows the user to store data in tabular format. It is more like excel files. Each value is separated by comma. In this project document containing import details and export details are in CSV format, also the output got after data integration is in csv format. Tables containing export details and import details are in CSV format.

XLS

XLS is a file extension for a spreadsheet file format created by Microsoft for use with Microsoft Excel. XLS stands for eXceL Spreadsheet. Microsoft Excel files use a proprietary format for storing Microsoft Excel documents.

4. EXTRACT TRANSFORM LOAD (ETL) TOOL

TALEND OPEN STUDIO

Talend Open Studio is an open source software which provides data agility for modern business by making use of cloud technologies to give insight. This helps to transform the business to an unexpected level.

Application of Talend are:

- Mapping
- Data Transformation
- ETL
- Business Intelligence
- Data Integration

According to facts, mapping and data integration are required.

This project has made use of data integration functionality of Talend. With the help of talend data integration and management, can be done 10 times faster than manually. Along with time saving it makes the cost $1/5^{th}$ the cost of competitors. Talend provides both ETL for data analytics and ETL for operational integration needs. Data storage formats provided by Talend - data integration are XML files, positional flat files, delimited flat files, multi-valued files and so on. Data integration is done in Talend tool by creating standard jobs. After creating job, data integration is done with different components in Talend.

Components of Talend tool made use in this project are:

- tMap - used to transform and route data from single or multiple sources to single or multiple destinations
- tfileInputExcel - opens a file and reads it row by row to split data up into fields using regular expressions. Then sends fields as defined in the schema to the next component in the job via a row link.
- tfileOutputExcel - writes an MS Excel file with separated data value according to a defined schema.

.

5. Data Modularity:

I. In data mapping, two files have been taken; both the files are connected with SERIAL_NO column. Mapping is done by inner join of both the files.

II. In data transformation, precision for decimal values has been specified and PREVAILING_WAGE column has been restricted to non-zero and non-negative values.

6. ER MODELING:

It is the way to represent the logical and physical entities in database in graphical form

6.1. Logical Design:

The process of logical design involves arranging data into a series of logical relationships called entities and attributes. This model describes the logic to find best player for each position.

6.2 Physical Design:

It is the method of transforming a description into the physical layout, which describes the position of cells and routes for the interconnections between them. It contain tables, Integration constraints (Primary key, foreign key, not null) and columns

7. Data Mapping:

Data mapping is the process of combining different datasets containing distinct data to create a data model. Data model describes the structure of data and the link between data sets. Each data sets are linked with each other with unique ids. For combining the data set, these unique ids are identified first. Mapping is done using the unique ids.

In this project two data models are created, first is to find out the fact regarding states having maximum number of visa petitions filing, and second is to find out the job profile for which there is maximum number of application for visa petition.

- Dataset containing Employer details consist of:

  - Employer key,
  - Employer name,
  - Worksite,

  Employer key is the unique id in this table.

- Dataset containing Region details consists of:

  - Region key
  - Region name
  - Worksite
  - Latitude
  - Longitude

  Region key is the unique id in this table.

- Dataset containing job profile details consists of:

  - Job id
  - Job profile name
  - Worksite

  Job id, Worksite make up the unique id in this table.

In this project, the datasets taken to be mapped are:

- Datasets containing Employer details.

  This table consists of:

  - Unique id of each Employer,
  - Employer name and Worksite

- Datasets containing details of Region.

  This table consists of:

  - Unique ids of region, its name, Worksite.
  - Geographical coordinates of worksite.

tMap

Mapping is done using Talend Data Integration Tool. Component used for mapping is tMap. Operations allowed by tMap are as follows:

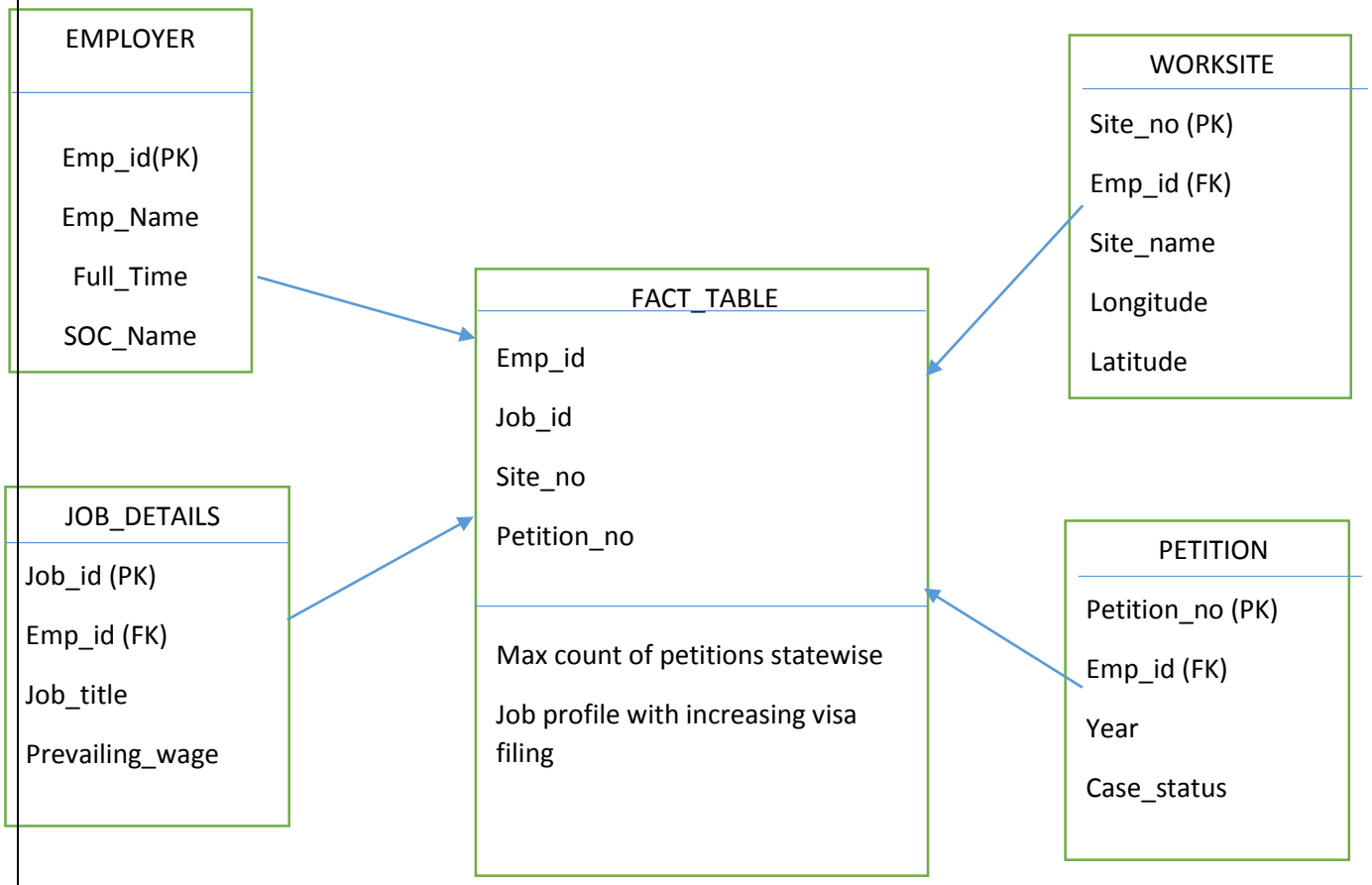
1) Data multiplexing and demultiplexing
2) Data transformation and type of fields
3) Field concatenation and interchange
4) Field filtering using constraints
5) Data rejecting

This project makes use of data multiplexing operation of tMap. Data multiplexing is the process of converting multiple inputs into one output. In tMap there are two schemas defined, therefore multiplexing is done twice in each data model.

8. Identification of facts and dimension:

a) Star schema

Star schema is created to organize the tables such that we can retrieve the results from the database easily and fast. It consists of one fact table and its dimensions.



b) Dimension reduction:

- Unnecessary dimension have been removed.

c) Factor analysis:

- All details are necessary for fulfilling the requirement of the facts.

10. METADATA REPOSITORY

Metadata repository is created to store metadata. Metadata is the information about structure of any data in any format. Metadata Repository contains data far beyond simple definitions of the data structure. Typical repository stores dozen to hundreds of separate pieces of information about each data structure.

10.1 BUSINESS METADATA

Business metadata provides information created by business people or used by business people. It can reduce the communication barriers between human and human, as well as human and computer. With business metadata, data conveyed from reports, information system or business intelligence application can be easily comprehended by non-technical business people and helps them in business decision-making.

It explains the information obtained from each column of each table.

10.2 TECHNICAL METADATA

Technical metadata is contrast to business metadata, it explains the physical character of a database such as table name, column name, data types, relationships between tables, constraints, abbreviations etc and is used by technical people.

Entity Name	Attribute	Column Null	Column data type	Primary key	Foreign key
Employer	Employer key	Not Null	VARCHAR	YES	NO
	Employer name,	Not Null	VARCHAR	NO	NO
	Worksite	Not Null	VARCHAR	NO	NO
Region	Region key	Not Null	VARCHAR	YES	NO
	Region name	Not Null	VARCHAR	NO	NO

11. DASHBOARD CREATION:

A dashboard is a collection of several Worksheet and supporting information shown in a single place. It helps to compare and monitor a variety of data simultaneously. Business tool used for creating dashboard is Tableau. Tableau is used for data visualization, data can be represented as charts, graphs, maps etc.

Steps to create dashboard:-

- Once the input files are imported to the tool, all dimensions and measures are automatically identified by Tableau
- On drag and drop of dimensions and measures at proper axes, visualization can be done. Different types of visualizations are provided by Tableau such as pie chart, stacked bars, side by side bars, vertical bars etc.
- For making visualization more clear, one may add more measures and filtering is done on rank to get increase/decline in the filing of visa petitions.

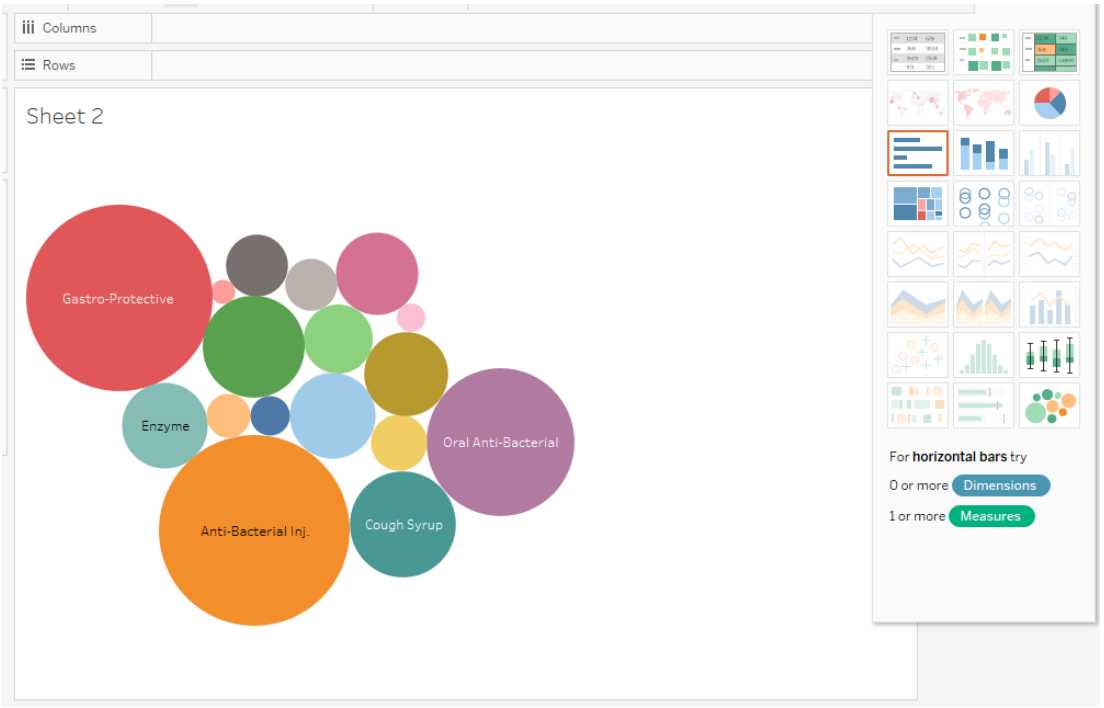
TABLEAU

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique. It is used by businesses, academic researchers, and many government organizations for visual data analysis. It is also positioned as a leader Business Intelligence and Analytics Platform in Gartner Magic Quadrant.

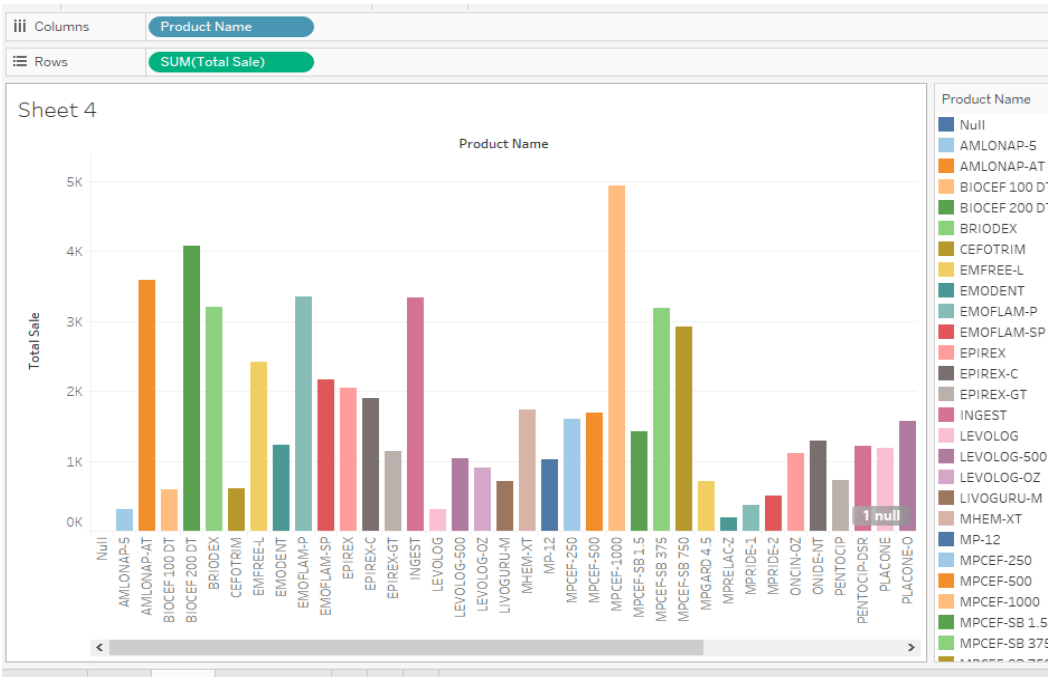
a) Data Visualization

Data visualization is the manifestation of the business analytics in order to work with data. Tableau analytics is one of the easiest and most powerful analytics tools today. It helps business managers to make quick decisions at all levels. Tableau can help to make sense of all types of

data coming in from disparate sources and blend it into one unified platform so that everybody can make sense of it.



Sample Output on Tableau



Bar Graph in Tableau

Conclusion:

The most number of visa petitions were filed for the job position of chief executive during the year 2011 – 2016 mostly for those employers located at the western coast of the USA.

12. Current status of Data Warehouse & Business Intelligence

I. Current Status of DW&BI: Today, data warehouses are not moving at the speed of the business. It takes forever to integrate a new data source into your data warehouse. We have to figure out what reports we want so you can pre-define data dimensions for aggregation. We have to figure out a schema that can accommodate all the data we are going to include. We have to set up ETL to translate our operational data into that analytic schema, and we have to maintain separate technology stacks at the operational, analytic, and archive tiers. This kind of traditional data warehouse is resistant to change.

There are three trends driving the move to a more agile model. First is the trend towards wanting to move faster and accommodate more data quickly. Waiting months to develop a schema and build the required ETL is no longer acceptable. Second is the trend towards discovery-based analytics, driven by the consumer experience with search technologies. Business analysts today want a search-based paradigm that allows them to formulate new questions to ask the data based on the results of the question they just asked a few seconds ago, and they want the results in real-time so they can figure out the question they want to ask next.

Third is the trend towards operationalizing the data from the data warehouse. This means building data services that can combine data from multiple sources and provide that data securely and performed to an operational process so that process can complete in real time. Fraud detection, eligibility for benefits, and customer onboarding are all examples of use cases that used to be performed offline but now need to be performed online in real-time.

Data Warehousing has never really been about warehousing your data. It's always been about getting value out of it. Enterprises want more agility, and they're finding that new technologies like NoSQL can deliver more value on a greater variety of data faster than ever before.

An industry peer, Ronald van Loon, writes about the future of data warehousing. Specifically, he talks about deploying a new kind of data warehousing that needs to support newer BI deployments to keep up with customer demand. The main factors that drive development and deployment of new data warehouses are being agile, leveraging the cloud and the next generation of data (as it relates to real-time data, streaming data and data from IoT devices).

A new kind of data warehousing is essential to this new BI deployment, as much of the inefficiency in older BI deployments lies in the time and energy wasted in data movement and duplication. A few factors are driving the development and future of data warehousing, including:

- **Agility** – To succeed today, businesses must use collaboration more than ever. Instead of having separate departments, teams, and implementations for things like data mining and analysis, IT, BI, business, etc., the new model involves cross-functional teams that engage in adaptive planning for continuous evolution and improvement. This kind of model cannot function with old forms of data warehousing, with just a single server (or set of servers) where data is stored and retrieved.
- **The Cloud** – More and more, people and businesses are storing data on the cloud. Cloud-based computing offers the ability to access more data from different sources without the need for massive amounts of data movement and duplication. Thus, the cloud is a major factor in the future of data warehousing.
- **The Next Generation of Data** – We are already seeing significant changes in data storage, data mining, and all things relate to big data, thanks to the Internet of Things. The next generation of data will (and already does) include even more evolution, including real-time data and streaming data.

II. How New Data Warehousing Solves Problems for Businesses

So how do new data warehouses change the face of BI and big data? These new data warehousing solutions offer businesses a more powerful and simpler means to achieve streaming, real-time data by connecting live data with previously stored historical data.

Before, business intelligence was an entirely different section of a company than the business section, and data analytics took place in an isolated bubble. Analysis was also restricted to only looking at and analyzing historical data – data from the past. Today, if businesses only look at historical data, they will be behind the curve before they even begin. Some of the solutions to this, which new data warehousing techniques and software provide, include:

- **Data lakes** – Instead of storing data in hierarchical files and folders, as traditional data warehouses do, data lakes have a flat architecture that allows raw data to be stored in its natural form until it is needed.
- **Data fragmented across organizations** – New data warehousing allows for faster data collection and analysis across organizations and departments. This is in keeping with the agility model and promotes more collaboration and faster results.
- **IoT streaming data** – Again, the Internet of Things, is a major game changer, as customers, businesses, departments, etc. share and store data across multiple devices.

List of Data Warehouse Tools and Testing Techniques:

1) Amazon Redshift

Availability: Licensed

Amazon Redshift is an excellent data warehouse product which is a very critical part of Amazon Web Services-a very famous cloud computing platform. Redshift is a fast, well-managed data warehouse that analyses the data using existing standard SQL and BI tools. It is simple and cost effective tool that allows running complex analytical queries using smart features of query optimization.

It handles analytics workload pertaining to big data sets by utilizing columnar storage on high-performance disks and massively parallel processing concepts.

A very powerful feature is **Redshift spectrum,** that allows the user to run queries against unstructured data directly in Amazon S3. It eliminates the need for loading and transformation. It automatically scales query computing capacity depending on data. Hence queries run fast.

2) Teradata

Availability: Licensed

Teradata is another market leader when it comes to database services and products. It is an internationally renowned company with its headquarters in Ohio. Most competitive enterprise organizations use Teradata DWH for insights, analytics & decision making.

Teradata DWH is a relational database management system marketed by Teradata organization. It has two divisions, namely, data analytics & marketing applications. Teradata DWH works on the concept of parallel processing. It allows users to analyze data in a simple yet efficient manner.

An interesting feature of this data warehouse is data segregation into **hot** & **cold** data. Here cold data refers to less frequently used data. It is a tool in the market these days.

3) Oracle 12c

Availability: Licensed

Oracle is a well-established3 name in data warehousing platform built for providing business insights and analytics to the users. Oracle 12c is a standard when it comes to a scalability, high performance, and optimization in data warehousing. It targets at increasing operational efficiency and optimizing end user experience. Its key features can be tabulated as:

Advanced analytics and enhanced data sets

Increased innovation and industry-specific insights

Maximum big data value

Profitability

Extreme Performance & consolidation

Additionally, Oracle 12c comes with advanced features like Flash storage and HCC (Hybrid Columnar Compression) that enables high-level data compression.

4) IBM Infosphere

Availability: Licensed

IBM Infosphere is an excellent ETL tool which uses graphical notations to execute data integration activities. It provides all the major building blocks of data integration & data warehousing along with data management and governance.The building foundation of this warehousing architecture is Hybrid Data Warehouse (HDW) and Logical Data Warehouse (LDW).

Multiple data warehousing technologies are comprised in a hybrid data warehouse to ensure that right workload is handled on the right platform. It helps in proactive decision making and streamlining the processes. It reduces the cost and is a very effective tool in terms of business agility.

This tool helps in delivering intensive projects by providing reliability, scalability, improved performance. It ensures the delivery of trusted information to end users.

5) SAP Business Warehouse

SAP business warehouse provides automated support in managing stocks in the warehouse. It is a flexible system and supports scheduled logistic processing within the data warehouse. This warehouse environment is completely integrated into to SAP environment.

7) Talend

Talend is an open source tool for data warehousing owned by Talend organization. It is a very powerful data integration and ETL tool. Its advanced features make it easy to use that has attracted many users. Talend provides progressive business solutions while having comparatively lower cost.

Conclusion:

The options in data warehouse tools that are available to companies are many. This lays stress over the importance of proper analysis of the organizational requirements and needs before picking any tool. It is always better to be prepared with a clear picture of the current requirements and future patterns beforehand.

Being the central repository, the data warehouse is extremely important to any organization in any sector and hence the choice of correct tool is a must.