DATA ENGINEERING FOR STRONG DATA WAREHOUSE

HARSHIL J JOSHI

Abstract: Data Engineering is a practice of extracting data completely or partially from source of data to analyze the data which becomes information. Data engineering uses data as the means for understanding a process. A system which collects data from different sources and provides information is called an Information System. This information is useful to take business decisions. At that time this system refers as Decision Support System. The whole process from collecting data from more than one sources, generate information and works as a decision support system is covered in Business Intelligence. Data Engineering is a part of Business Intelligence. A word 'Data Engineering' is clarified by an' Engineered Data', which is useful for business decision. Business Information is purely based on engineered dat. In Data Engineering two major processes are covered Data Mining and Data Warehousing.

Keywords: data warehouse, data mining, data engineering, business intelligence.

Introduction: The volume of data maintained by today's companies is enormous – and growing rapidly. A study conducted by a well known Information Technology magazine shows that approximately 2.5 quintillion bytes of data are generated on a daily basis. Ninety percent of the data that exists today, the report also claims, has been created in the last two years. Lack of integration among the systems that house this big data can make it difficult to manage operations, compliance, and risk across the business.

To take decisions in a same moment with a huge amount of data is itself a big task for any organization. In now a day's almost all companies have standard ERP which covers their business processes. ERP makes an organization in a well formed format, it organize a whole company with its standard functions, processes and business rules. A system which holds big amount of data in a standard way for reporting is known as Business Intelligence. Business Intelligence is a Data Warehousing and Data Mining technology which is covered in Data Engineering.

Business Trends: Until Now, the goal behind the implementation of classic data processing systems and transaction processing systems are automated in individual business areas. In this way information is very crucial upon any enterprise software. In parallel of increasing globalization, decentralization of organizations has also been increased. To gain complete information, decision-makers in modern, globally operating enterprises rely ever more frequently on the effective use of this information; unfortunately this information is widely spread across the globe in all business areas. This is precisely the challenge that modern data warehouses attempt to meet. Extensive solutions are required to cover the entire process from the retrieval of source data to its analysis. Metadata, dimension- and aggregation- data are treated differently in this process.

Data Mining & Data Warehousing: Due to continuous innovation in data processing possibilities, more and more information is stored in a more and more detailed form. As a result, there is the need to both reduce and structure this data so it can be analyzed meaningfully. A data warehouse can help to organize the data here. A data warehouse performs activity to get all operative data sources (these are mostly heterogeneous and have different degrees of detail) in order to provide this data in a scalable form to the whole organization. This data can then be used for future requirements.

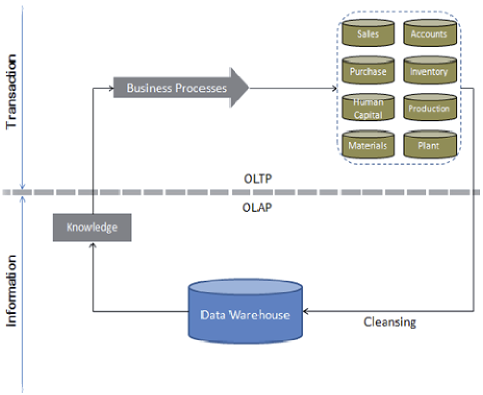

Fig 3.1

A data warehouse has the following properties:

- Read only access: End users have read-only access, meaning that the data is primarily loaded into the data warehouse via the Extraction, Transformation and Loading (ELT) process.
- Cross-organizational focus: Data sources from the entire organization (production, sales and distribution, controlling) and possibly external sources too make up the basis of the system.
- Data warehouse data is stored persistently over a particular time period.
- Historization: Data is stored by time on a long-

term basis.

- Designed for efficient query processing: The technical environment and data structures are optimized for answering business questions - not transaction processing.
- Analysis tool: Users can use comprehensive analysis tools to access data. These tools offer a user-friendly interface which simplifies query creation.

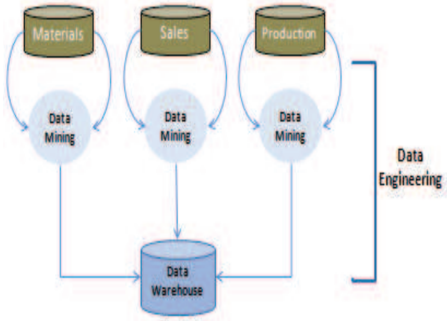

Fig 3.2

A data warehouse is a copy of transaction data, specially restructured for queries and analyses, in: R. Kimball: The Data Warehouse Toolkit, 1996, page 310.

Extract, Transform and Load (ETL) Process:

In Data Warehousing, data is to be extracted from many source systems, like from ERP, CRM, SRM, HCM, and Flat File etc. Data which are important for reporting or analysis will be extracted from source system and transformed to data warehouse system. Here data mining activity performs a major role. E.g. in ERP system there are many modules and transactions are performed in that but not all data need to be load to data warehouse, here as per the requirements or necessity data are load to data warehouse. The process to identify the necessary data and load it to data warehouse is covered in data engineering. As described in Fig 3.2, data from Materials, Sales and Production is load in Data Warehousing after Data Mining process.

Objectives of Data Engineering: The main purpose of data warehousing is standardize data structure and display of all information. Decision makers require an up-to-date and comprehensive picture of individual business area and of the business as a whole, so data architect need not pull whole data in data warehouse, only required data need to be pulled.

An important aspect here is that the data is defined uniquely across the entire organization, in order to avoid errors arising through varied definitions in different sources. When implementing the data warehouse, an influential cost factor is its integration into an OLTP system and the straight-forward loading of heterogeneous data. Alongside the Metadata Repository, Business Content also has an important role here.

Reporting and analysis require fast information access, so system should be stabilized with structured data; for structured data, data engineering is required. Data engineering also include data cleansing. Data cleansing includes identifying and removing (or updating) invalid data from the source systems.

The ETL process utilizing the staging area can be used to implement business logic to identify and handle "invalid and unwanted" data. Constraints may additionally be placed on staging area structures (such as table constraints in a relational database) to enforce data validity rules.

Data archiving can be performed in, or supported by, a single staging area or multiple. In this scenario the staging area(s) used to maintain historical records for data load process. In addition, data can be maintained within the staging area for extended periods of time to support technical troubleshooting of the ETL process.

Life Cycle – Data Engineering

Lifecycle data for engineered object (sales, purchase, materials, production etc.) – are all the data that appear and are used at any moment of the life cycle, from transaction and up to reporting.

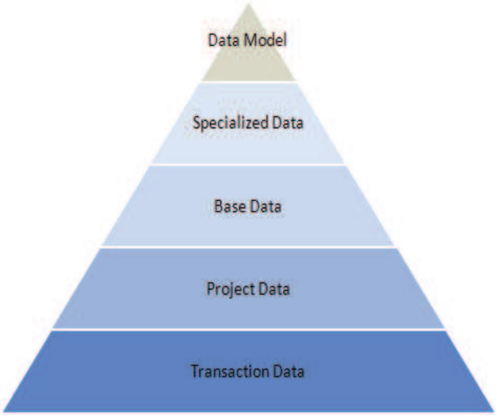

Fig 3.3

Fig 3.3 explains about the data staging in data warehousing. Transaction Data are non standardized & non structured data which is not used for any kind of reporting in organization, but based on these data Project Data is developed.

Project Data is the first step of data mining in data warehousing. All required data for reporting are stored in project data for warehousing purpose. Now based on project data Base Data is developed. Base Data is base not only for reporting but for analytics also.

Specialized Data is an aggregated data of base data which is purely for analysis. Data Model is updated dataset for reporting. Business reporting can be performed on base data, specialized data or data model. Fig 3.4 shows just example of how data are filtered from source system for business reporting.

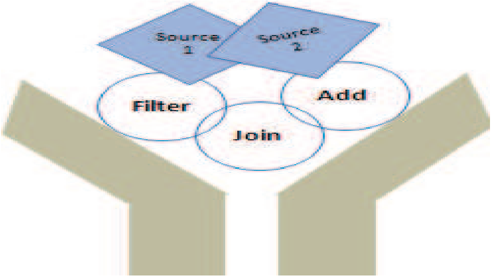

Fig 3.4

An enterprise performs well when it is having right information at right time (RIART), RIART is possible when enterprise is having strong data warehouse, and strong data warehouse is possible when there is strong data mining process. This happens if data engineer performs his job well.

Roll of Data Scientist/Engineer: Data scientists who spends time and apply brainpower on data science and analytic results to critical business issues helping an organization turn data into information, information into knowledge and insights - and valuable, actionable insights into better decision making and game changing strategies.

Data engineers/architects are the designers, builders and managers of the information or "big data" analysis. They develop the architecture that helps analyze and process data in the way the organization needs it with smooth system performance.

(A)Unstructured Data Management

- Content Management & Discovery
- Retrieval, Searching & Indexing
- Security & Protection
- Storage & Retention
- Backup & Recovery

(B) Data Architecture & Design

- Value Chain Analysis
- Enterprise Data Modeling
- Enterprise Data Integration

(C)Data Quality Management

- Data Quality Analysis
- Data Mining & Cleansing
- Data Conversion
- Continues Monitoring & Reporting

(D) Data Governance

- Data Ownership & Protection
- Master Data & Metadata Management
- Data Dictionary & Standards Maintenance
- Logical Layer Security & Audit Validation
- Data Lifecycle Management
- Compliance & Privacy Management

(E) Data Warehousing & Business Intelligence

- Relational & Dimensional Data Modeling
- Data Mining
- Extract, Transform & Load (ETL) Development
- ETL Quality Assurance
- Business Rules Development
- Online Analytical Processing (OLAP) cube development

(F) Information Presentation

- Dashboard & Interactive Reports
- Historical Reporting & Predictive Analysis
- Mobile Reporting

Data Warehousing Tools : As per the market research (data gathered from internet and assumed on latest trend) Oracle, SAP and IBM are top competitors while Microsoft and Informatica are widely used in SMEs. SAS is also focusing on Big Data analytics with statistical analytics. Open source business intelligence tools are also available in market e.g. SpagoBI, JasperSoft, and Pentaho BI etc. They are open source business intelligence suite which provides data warehousing, data mining, ETL and business analytics etc. Here are list of some ETL or Data Warehousing tools:

Informatica – Power Center
IBM Websphere Data Stage
IBM Cognos Data Manager
IBM DB2 Warehouse Edition
Microsoft SQL Server Integration Services (SSIS)
Oracle Data Integrator
Pentaho Data Integrator
SAS Data Integration Studio
SAP BusinessObjects Data Integrator (Services)
SAP Business Warehouse

References:

1. All figures are designed from real time industry experience.
2. Role of Data Scientist and Data Engineer/Architect is detailed information from real time industry experience and practical work.
3. Explained ETL process is industry specific, it defers from industry to industry as per business process.

SAP BI/BO Consultant,harshiljoshi.sap@gmail.com
B-35 Ashirvad Society Harni toVarasia Ring Road,Vadodara - 390022, Gujarat, India.