Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

Analysis and Design of ETL in Hospital Performance Appraisal System

Fengjuan Yang
Department of Computer and Information Science, Fujian University of Technology
Fuzhou 350100, China
Tel: 86-591-2270-2070 E-mail: yangfj99@126.com

Abstract

Taking the hospital performance appraisal items as the background, the method and mechanism of ETL process, data extraction, data cleaning, data transformation and increment updating are concretely analyzed and designed in the article. The data preparation area is added in the ETL process of the system, and the monitoring and restarting mechanism is set up in the system, which can effectively enhance the efficiency of ETL process.

Keywords: ETL, Hospital performance appraisal, Data Extraction, Data Transformation

1. Introduction

ETL (Extract-Transform-Load) is the process of data extracting, data transforming and data loading, and it is the core and soul of BI/DW (Business Intelligence/Data Warehouse), and it can integrate and enhance the value of data according to uniform rules, and it is responsible for the process transforming from the data source to objective data warehouse, and it is the important approach to implement data warehouse. The hospital performance appraisal system involves many business systems which are developed by different development teams in different terms. Because the data sources differ in thousands ways in data content, data format and data quality, which brings certain difficult and larger workload to the ETL process of the system. It is one of key factors to design a high-effective ETL process for constructing the system.

2. System structure of the performance appraisal system

The performance appraisal system is the appraisal system composed by KPI (Key Performance Indicators) which are independent and associated, and can completely express the appraisal requirements. KPI are the references for performance management, objective management, organization design and strategic management, and the performance appraisal method generally emphasized in modern enterprises (Dai, 2007, P.91).

The data sources of the hospital performance appraisal system root in multiple heterogeneous data sources such as HIS, human resource management system, OAS, financial management system and material management system, and the databases of these systems include Oracle and SQL Server. And the data sources with other document characters are also included in the hospital performance appraisal system. The hospital performance appraisal system is to utilize the mass data produced by the business support system, adopt the computer technologies such as data warehouse and data digging to extract, integrate, analyze and dig data, and provide timely, exact and scientific appraisal references for the performance appraisal of the hospital. The structure of the hospital performance appraisal system is seen in Figure 1.

3. Introduction of ETL tools

The representative ETL tools include Informatica, Datastage, OWB and Microsoft DTS at present.

Informatica Power Center is the advanced ETL tool in the industry, and it can conveniently extract data from heterogeneous system and data sources for users to establish, deploy and manage the data warehouse of the enterprise, and help the enterprise to make quick and exact decisions. This product can provide extensive supports for the application and data sources such as ERP system (Oracle, PeopleSoft and SAP), CRM (Siebel), electric business data (XML, MQ Series), legacy system and host computer data. The ETL tool in the solution project of IBM DB2 is Visual Warehousing which is included in Data Warehousing Manager. The Datastage of Ascential is the manufacturer with the highest share in the market, and it is the solution to support various systems and platforms such as host computer, ERP, UNIX and NT. The basic frame of Oracle Warehouse Builder includes two parts, i.e. design environment and operating environment. OWB can automatically generate the SQL codes corresponding with database object, and these codes can be distributed into the database, and ETL is implemented by the codes which are distributed into the database by the Oracle Enterprise Manager. DTS is the data integrating tool of Microsoft which can complete data extracting, transforming and loading on the Windows platform (Webmaster, 2006).

All above tools are all-purpose graph interface tools, and they can screen complex coding tasks, enhance the speed and reduce the difficulty. For the hospital performance appraisal system, the extracted data quantity is huge and the parameters are numerous. To enhance the efficiency and the flexible expansibility of the system, the ETL in the article combines OWB tool with SQL to quickly establish the ETL project by OWB, and utilize the flexibility of the SQL method to enhance the development speed and efficiency of ETL.

4. ETL design in the hospital performance appraisal system

Figure 2 is the structure of the ETL, and in the designing process of the system, the extracting program first extracts exterior data to the data preparation area, and then the system cleans the data in the data preparation area, and transforms the data according to the data model, and finally loads the transformed load to the data warehouse.

4.1 Data preparation area

Because the data extracting, cleaning and loading of the data warehouse need long work time, and to reduce the influence to the data source system and enhance the extracting efficiency, the system sets up the data preparation area.

The data preparation area is the work platform of data preparation, and its function mainly includes three aspects. First, the data extracted from the data source in the data preparation area can enhance the extracting efficiency and reduce the data extracting time, and reduce the influence of data extraction on the business support system. Second, the data preparation area can extract multiple data sources, and enhance the reliability and coherence of the extracted data. Third, some simple data preprocessing can be made in the data preparation area, which can enhance the efficiency of cleaning and transforming (Qi, 2005).

4.2 Restarting mechanism

Set up the restarting mechanism of data extracting, cleaning and loading in the data preparation area. In the data extracting, cleaning and loading process, because of the reasons of the system or some unpredictable factors, these activities will often fail, and if the system is restarted after failing, large numbers of resources of the system will be wasted. Therefore, the monitoring mechanism of data extracting, cleaning and loading in the data preparation area can be set up to dynamically monitor these activities, once they failure, the system can restart from the position of failing. To complete this mechanism, the data extracting, cleaning and loading activity can be divided into many approaches, and when the system enters into certain approach, it can hold the present status.

4.3 Data extraction

When designing the data extraction, the system should mainly consider the extracting mode, extracting content and increment updating of data.

4.3.1 Extracting mode

The data extracting mode includes the active mode and the passive mode. The active mode means that the source system actively extracts the data according to the data format defined by two parties. The active mode will make the source systems or other development teams depend on the performance and network of the source system. The passive mode means that the ETL program directly interviews the data source to acquire the data mode, and under this mode, the ETL works independently and extracts the data by itself.

The system in the article adopts the passive data extracting mode because the extracting time can be flexible and the structure change of the business system can not influence the normal work of ETL program.

4.3.2 Extracting content

The second problem of the data extracting is “what data the system extract”. The hospital performance appraisal system involves many tables, and the extracting must fulfill two conditions, and the first one is that the extracted data should fulfill the requirements of corresponding indexes in the performance appraisal system, and the second one is that the extracting process should not influence the performance of the original business system. Therefore, the system adopts the combination of the full extraction and the increment extraction. For the tables with small data quantity, the full extraction can be adopted, and all dictionary tables in the system all belong to small tables, and the data quantity is less than thousands of records. And these tables include the section office table, the doctor table, the medicine table and the illness table, and for these tables, the full extraction should be adopted. For the tables with large data quantity, such as the charge list, the illness diagnosis record table and the patient record table, and the data quantity of these tables can achieve ten-millions-class, so relatively flexible increment extracting modes must be adopted. The increment extraction can reduce the extracting data quantity, reduce the influences on the transformed and loaded data quantity, network flux and the business system performance, and enhance the performance of the whole process.

4.3.3 Updating of increment

The increment updating is the most important problem in the ETL process design, and extracting mode of data increment directly influences the performance of the system. At present, the changeable data methods in common use include trigger, time-stamp and log comparison. The extracting mode of time-stamp is used in the fields with time-stamp, and it can distinguish whether the record belongs to the newly added record, and the comparison of the ending time of the last extracting and the time-stamp field in the table can decide the extraction of the data increment.

Taking the in-patient charge list increment of HIS as the example, according to the control flow table of the system ETL (Table 1), the extracting period is week, and the extracting time is the two o'clock in every Sunday, and the extracting SQL sentences are

```
//select * from in-patient charge list
where pricing date and time > to_date('05/10/2008 23:59:59', 'DD/MM/yyyy HH24:MI:SS')
and pricing date and time <=to_date('11/10/2008 23:59:59', 'DD/MM/yyyy HH24:MI:SS')"//
```

For the table without time-stamp fields, the log comparison increment extracting mode can be adopted to analyze the log of the database and judge the changing data. The CDC (Changed Data Capture) of Oracle is the representative technology in this aspect, and CDC can identify the changed data after last extracting. By means of CDC, when the source table implements many operations such as inserting, updating or deleting, the system can extract the data, and the changed data are stored in the changed table of the database. So the changed data can be captured, and are provided to the objective system by a kind of controllable mode through the database view.

4.4 Data cleaning and transformation

The data extracted from the business system are put into the data preparation area and cleaned in the data preparation area, and the dirty data can be filtrated. Then the system completes incomplete data and transforms the cleaned data according to the designing requirements.

4.4.1 Data cleaning

The data falling short of requirements mainly include incomplete data, false data and repetitive data.

(1) Incomplete data. These data are some information deficiencies such as the sex of patient, birth date and region. These data can be completed by the concealed information according to patients' ID number. For the data which can not be completed, some user-defined types can be used for later analysis, for example, when the patient's family address is deficient and can not be completed, fill in "undefined", and these data can be extracted in the future conveniently, and deleted according to actual situation.

(2) False data. The main reason of the false data is that the business system is not complete, and after the data are incepted, the data are directly wrote into the backstage database without being judged, for example, the numerical data follow an enter operation, or the input of the character string is false, the date format is not correct or the date is beyond the mark. The name of the doctor is "Zhang San", but the input may be "Zhang Shan" or "Zhang Sun". These data should be classified, and found out by the SQL sentence, and extracted when the business department modifies the business system.

(3) Repetitive data. For these data especially in the dimensional table, all fields recorded by repetitive data should be educed to confirm and process by the business department.

Data cleaning is a repetitive process, and it can not be completed in several days, and it can continually discover and solve problems. The business department will confirm whether the data need to be filtrated and modified, and the filtrated data are wrote into the data table by the form of Excel file, and in the initial stage of ETL development, the e-mails transmitting the filtrated data to the business department will make the department to modify the mistakes as soon as possible and regard the data as the references in the future. The data cleaning should validate each filtrated rule and be confirmed by the business department, and should not filtrate useful data.

4.4.2 Data transformation

The task of data transformation mainly includes the inconsistent data transformation, the transformation of data granularity, and the calculation and integration of some business rules.

(1) Inconsistent data transformation. This process is a process integrating data with same type in different business operations, for example, the sexes in HIS are denoted by "M and F", but they are denoted by "Male and Female" in the office system, and they are denoted by "0 and 1" in the financial system. After extracting, the sexes are denoted by "0 and 1" uniformly. If the quantity of the data needing transformation is large, the transformation comparison table can be designed to conveniently transform the data, for example, the section office codes include hundreds even thousands of records, and they can be designed as the comparison table such as Table 2 to convenient for the transformation of the section office codes.

(2) The transformation of data granularity. The business system generally stores fined data, but the data in the data warehouse are used for analysis, so generally the data in the business system will be integrated according to the data

warehouse granularity.

(3) The calculation of business rules. Different enterprises have different business rules and different data indexes, and these indexes sometimes can not depend on simple adding operations, and these data indexes calculated in ETL need to be stored in the data warehouse for analysis and use.

4.5 Data loading

Data loading is to load the data extracting, transforming and cleaning in the source business system into the data warehouse. In the system loading process, not only the data loading method should be considered from the performance angle, but also the data validating mechanism should be established, for example, validating the input and actual loading record amount, and processing and transferring the abnormal mistakes. This system adopts asynchronous and batch processing mode to load the data, and because the data loading involves many system resources, and needs the processing, interior memory and exterior memory equipments of data source and data warehouse. The data loading of data warehouse is implemented at the two o'clock, because the business system in this period is spare.

In addition, the loading and renovating of large number of data can only be implemented in the first-time data loading when the data warehouse is just established, and the sequent data loading always needs adopting increment data loading method. When implementing increment data loading, some necessary preparation works should be completed in the loading in the data preparation area.

5. Conclusions

The effective strategies and implementation projects are proposed in the article for the data extraction mode, data cleaning, data transformation and data loading mode in the ETL design process of the hospital performance appraisal system, and these strategies and projects can ensure the clarity and high-efficiency of the whole ETL process, enhance the veracity and reliability of data in the system, and provide powerful data guarantee for the data integration and data digging with high quality for the hospital performance appraisal system. The system proposed in the article has been implemented successfully and acquires good effects in certain provincial Class Ⅲ-A general hospital.

References

Dai, Huazhen, He, Liangyu & Yang, Feihong. (2007). Design and Realization of Bank Achievement Inspection System Based on ETL Technology. *Modern Computer*. No.273(12). P.91.

Qi, Yinfeng & Wang, Manshu et al. (2005). Research of Chinese Enterprise Investment and Financing Behaviors: Based on Results of Questionnaires. *Management World*. No.3.

Webmaster. (2006). White Book of Disaster Recovery. [Online] Available: http://www.ibm.com.

Table 1. ETL control flow

No.	Data source	Name	Extracting mode	Extracting condition	Operating time	Sign of success
1	HIS	Section office dictionary table	Full		2008.10.12 02:00:00	Success
2	HIS	In-hospital charge list	Increment	2008.10.5 23:59:59	2008.10.12 02:00:02	Success
3	Office system	Human resource table	Full		2008.10.12 02:10:00	Success
4	Financial system	Salary table	Increment	2008.10.5 23:59:59	2008.10.12 02:10:05	Success

Table 2. Comparison table of section office codes

HIS system code	Name	Corresponding	Code of section office	Name of section office
C210	Earthquake relief nursing unit	⇄	FF	Flexible section office
C308	ENT nursing unit	⇄	FF02	ENT nursing unit
C30801	ENT medicine-chest	⇄	FF0201	ENT medicine-chest
A101	Director of hospital	⇄	1	Director of hospital
A102	Department of medical affairs	⇄	2	Department of medical affairs
A103	Department of politics	⇄	3	Department of politics
A104	Department of hospital affairs	⇄	4	Department of hospital affairs
A105	Department of nursing	⇄	5	Department of nursing
A10208	Department of planning	⇄	6	Department of planning
A10201	Department of medical treatment	⇄	7	Department of medical treatment
A10202	Departmnet of training	⇄	8	Departmnet of training
A10203	Department of economic management	⇄	9	Department of economic management
A1020301	Office of outpatient service and charge	⇄	901	Office of outpatient service and charge
A1020302	Office of plastic surgery department charge	⇄	902	Office of plastic surgery department charge

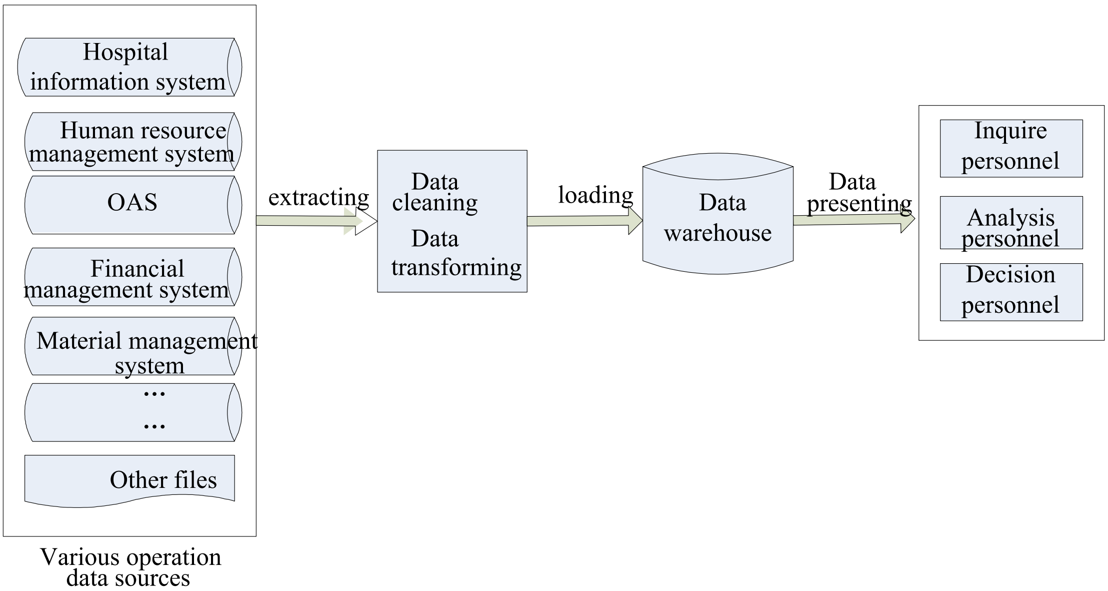

Figure 1. Structure of Hospital Performance Appraisal System

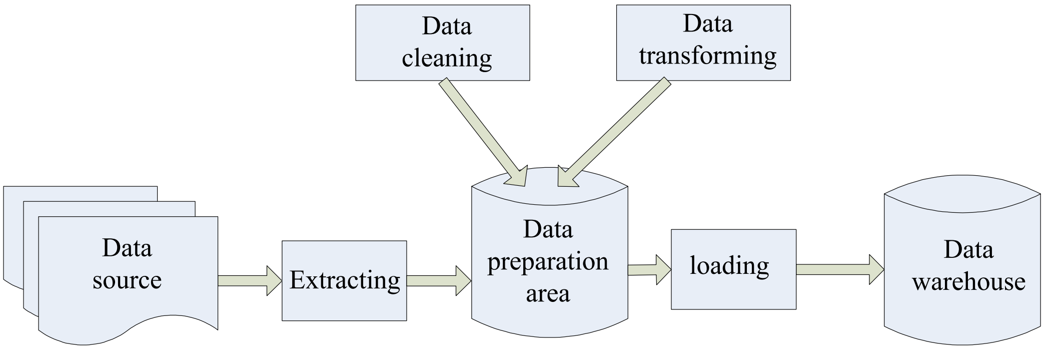

Figure 2. ETL System Structure