Research on display system for agricultural science and technology support data based on Microsoft data warehouse

Liyun Wang, *Tingting* Liu* and *Dingfeng* Wu

Agricultural Information Institute, Chinese Academy of Agricultural Sciences/Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing, 100081, China

Keywords: Microsoft data warehouse, Data dimensional modeling, Display system.

Abstract. Based on the introduction of Microsoft data warehouse service and related software architecture, in order to solve the problems of slow analysis queries caused by the great amount of the original data, complex classification and wide scope, the display system for agricultural science and technology support data was presented in this paper. In proposed system, agricultural science and technology support data were showed clearly and intuitively.

1 Introduction

Agricultural science and technology support data is an important part of agricultural science and technology information data. It is a kind of important strategic resources, also is an important foundation for the development of agricultural science and technology. It need be effectively managed in order to be better exploited and utilized, so as to promote the standardized management and efficient utilization of agricultural scientific data resources. It also promotes implementation of data resources being effectively saved, explored deeply and shared. It provides strong data support for the development of China's agricultural science and technology.

According to the data type and collection methods, science and technology support databases systems are separated and fragmented. The traditional database is one designed to make transactional systems run efficiently. Typically, this type of database is an OLTP (online transaction processing) database. In fact, an OLTP database is typically constrained to a single application, but not suitable for information processing, and its performance is low for analysis queries. In addition, the agricultural science and technology support data has the feature of large amount, complex classification and wide scopes. In order to comprehensively analyze agricultural support data quickly and conveniently, accurately grasp the status of agricultural development and the distribution of agricultural scientific and technological resources, strengthen the integrated management of resources and

* Corresponding author: liutingting@caas.cn

scientific research projects, and provide data support for agricultural development plan, it is necessary to sort out the agricultural science and technology support data. Then scientific data from different sources are fused into data warehouse. Finally, it will provide data support for data-intensive decision-making analysis and a solid foundation for subsequent data mining.

Based on the above background, scientific and technological support data are extracted from a large number of transactional databases, and then these are cleaned, transformed and finally stored into a data warehouse according to certain rules, using Microsoft data warehouse technology. It is easy for the user to query, analyze and mine information from the data through the access tool of data warehouse. It also provides complete, timely and accurate information for decision makers.

2 Information about SQL server

2.1 Overview of data warehouse

Different people have different definitions for a data warehouse. The most popular definition came from Bill Inmon [1], who provided the following: A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. Architecture of data warehouse is showed in Fig.1.

The construction of a data warehouse will greatly reduce the time of acquiring information. It congregates data from multiple sources into a single database, so all information can be obtained directly from the data warehouse. The data warehouse restructures the data and improve data quality so that it delivers excellent query performance, even for complex analytic queries [5].

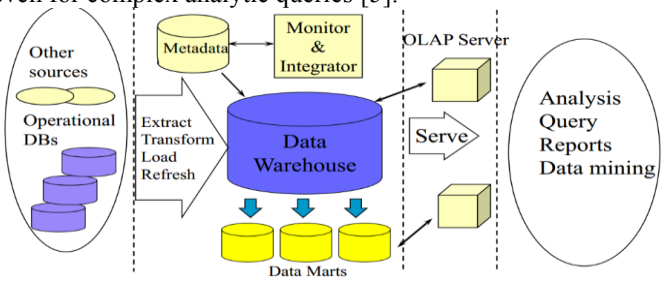


Fig. 1. Architecture of data warehouse.

2.2 Microsoft SQL server data tools

Microsoft SQL Server Data Tools (SSDT) is used to develop data analysis and Business Intelligence solutions, including SSAS (SQL Server Analysis Services) and SSIS (SQL Server Integration Services) and SSRS (SQL Server Reporting Services). While SSAS enables users to construct special databases for fast analysis of very large amounts of data, and while SSIS enables users to integrate data from many sources outside Microsoft SQL Server, SSRS enables users to quickly and easily generate reports from Microsoft SQL Server databases [3].

SSIS is a platform for building enterprise-level data integration and data transformations solutions. Integration Services can extract and transform data from a wide variety of sources such as XML data files, flat files, and relational data sources, and then load the data into one or more destinations [3].

SSAS is an online analytical processing (OLAP) and data mining tool. SSAS is used as a tool by organizations to analyze and make sense of information possibly spread out across multiple databases, or in disparate tables or files [3].

The SSRS service provides a unique interface into Microsoft Visual Studio so that developers as well as SQL administrators can connect to SQL databases and use SSRS tools to format SQL reports in many complex ways.

2.3 Key technology of data warehouse

A successful data warehouse project involves a set of concepts and methods designed to build a practical data store as a basis for the subsequent display system.

2.4 Data dimensional modeling

Dimensional Modeling technique is one of the most important techniques of Data warehouse. It is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the business [4]. Dimensional modeling has two basic concepts. Fact tables represent a business process, and contain the measurements or metrics or facts of business processes. The level of detail is called as the "grain" of the table. Fact tables contain foreign keys for the dimension tables. Dimension tables represent the who, what, where, when and how of a measurement/artifact, or real-world entities not business processes. The Dimension Attributes are the various columns in a dimension table [5].

The dimensional model is built on a star schema, with dimensions surrounding the fact table. Dimension tables are generally assigned a surrogate primary key, usually a single-column integer data type, mapped to the combination of dimension attributes that form the natural key.

Dimensional models are more unnormalized and optimized for data querying. Each dimension is an equivalent entry point into the fact table, and this symmetrical structure allows effective handling of complex queries. Dimensional models are scalable and easily accommodate unexpected new data. Existing tables can be changed in place either by simply adding new data rows into the table or executing SQL alter table commands. No queries or applications that sit on top of the data warehouse need to be reprogrammed to accommodate changes. Old queries and applications continue to run without yielding different results.

2.5 The ETL process

Data warehouse collects large volume of data from variant sources with many different data formats. The ETL (Extraction, Transformation and Loading) handles these data and transforms it into a more consistent, standard formatted data. The ETL is the core of data warehouse. Extraction is the process of extracting data from variant data sources. Transformation is the process of transforming the extracted data for storing it in the proper format or structure for the purposes of querying and analysis. Transformation occurs by using rules or lookup tables or by combining the data with other data. Loading is the process of writing the data into the final target data warehouse [6].

The ETL process typically takes the longest to develop, and this can easily take up to 50%-70% of the data warehouse implementation cycle or longer. The reason for this is that it takes time to get the source data, understand the necessary columns, understand the business rules, and understand the logical and physical data models.

2.6 OLAP cube

OLAP model comprises of building a multidimensional database after the process of ETL in a data warehouse. OLAP enables a user to easily and selectively extract and view data from different points of view. Multidimensional database technology is the fundamental approach for interactive analysis from huge amount of data.

In this work, OLAP model concentrating on building the multidimensional database where the data model is built on MOLAP architecture where all data model tables are gathered and structure a schema design of star and snowflake schema from the dimensional tables and fact tables. The multidimensional database model is also known as "cubes", which means a multidimensional view of data considers which information is stored in a multidimensional array or cube. The data cube has turned out to be a satisfactory model that provides a way to aggregate facts along multiple attributes called dimensions. The data cube is then used to access data in various methods: drill up, drill down, drill across and so on.

3 Design and implementation of the system

3.1 System design

The display system of agricultural science and technology support data is based on B/S architecture, mainly divided into four main processes. The system structure is shown in Fig. 2.

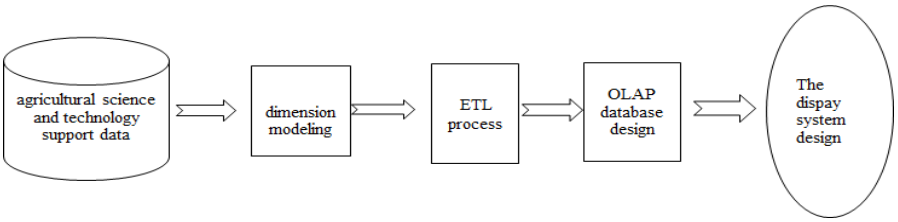


Fig. 2. The system structure.

(1) Dimensional modeling design: constructing dimensional modeling is a dynamic process with high repeatability. Based on a detailed understanding of content, meaning and business rules of original systems that dimensional modeling need, the detailed requirements documents are established, and then designers carry out data analysis and research many times. On the basis of these preparation steps, dimension model construction is generally divided into three stages: first is the advanced dimensional model design session, defining boundary of the dimension model; the second stage is the development phase of detailed model, including filling property list to each table, solving all kinds of problems and uncertainty etc.; the third stage is a series of procedures including model review, redesign and validation, the main aim is constructing model that meet the business requirements, and providing a solid starting point for the ETL process.

(2) The ETL process. So-called ETL is extracting data from variant data sources, and transforming it into a more consistent, standard formatted data that are loaded into the data warehouse. The SSIS tool is used in this stage. It has powerful data processing capabilities, and is a powerful tool to import raw data into data warehouse. The agricultural science and technology support data are imported into data warehouse with the SSIS tool according to certain rules.

(3) Design of OLAP database. Based on the relational data warehouse established in above two steps, an OLAP multidimensional database is established as a display database

for the purposes of querying and analysis. SSAS is a popular and effective tool to meet these needs. With the help of SSAS tool, a stable scalable OLAP server is built as the main query engine of the display system.

(4) Design of the presentation program. Based on the web portal platform SharePoint, SSIS is used to create and publish standard reports. The data are showed in data tables, bar charts, pie charts and other forms.

3.2 Establishment of dimension modeling

Based on the analysis of the agricultural science and technology support data source, grain and dimension of dimension modeling start to design. Taking a project of the agriculture as an example, a fact table designed is used to record the basic information of the project, including project number, project name, and project amount and so on. The grain of the fact table is each row representing project information in every year. Based on such definition of grain, the dimension tables associated with the fact table has date, unit, state, audit, etc. The fact tables and dimension tables are shown in Fig. 3. After designing the dimension model, the data is extracted, cleaned, transformed and loaded into the data warehouse through the ETL process.

3.3 Data retrieval

Existing agricultural science and technology support data scatter in a number of operating database. Data involves a wide range of class. There are some wrong data and null data. First, we collect raw data from the source systems, following by cleaning, unifying and merging. Then, consistence dimension and consistency measure are created, such as instead of null values with a default value, deleting duplicate records, renaming list, data type conversion, etc.. In addition, data quality errors are recorded into error event model. Then, after the standard dimension transformation and physical construction of data, these data are loaded to the target dimension model of the OLAP display server.

3.4 Establishment of OLAP database

In above two subsections, the data from the source system have been loaded into the data warehouse. In addition, data warehouse relation layer with surrogate keys and consistence dimension has been stablished, which results in better management of dimension change. Hence, OLAP database establishment is relatively easy. There is strong connection between the OLAP dimension and dimension tables in the relational data warehouse. The design and construction of the OLAP dimension are completed by using the dimension designer. Then, by using the OLAP cube wizard, we import the fact, create the structure of the cube and complete the construction of the cube, which can be used to implement data drill-down in a hierarchical dimension.

3.5 Data exhibition

The data of data warehouse can well match with the front desk design through the above three steps. It only need a small amount of code in system background to obtain the required data. The query performance of the system can be greatly enhanced without of repeated calculations.

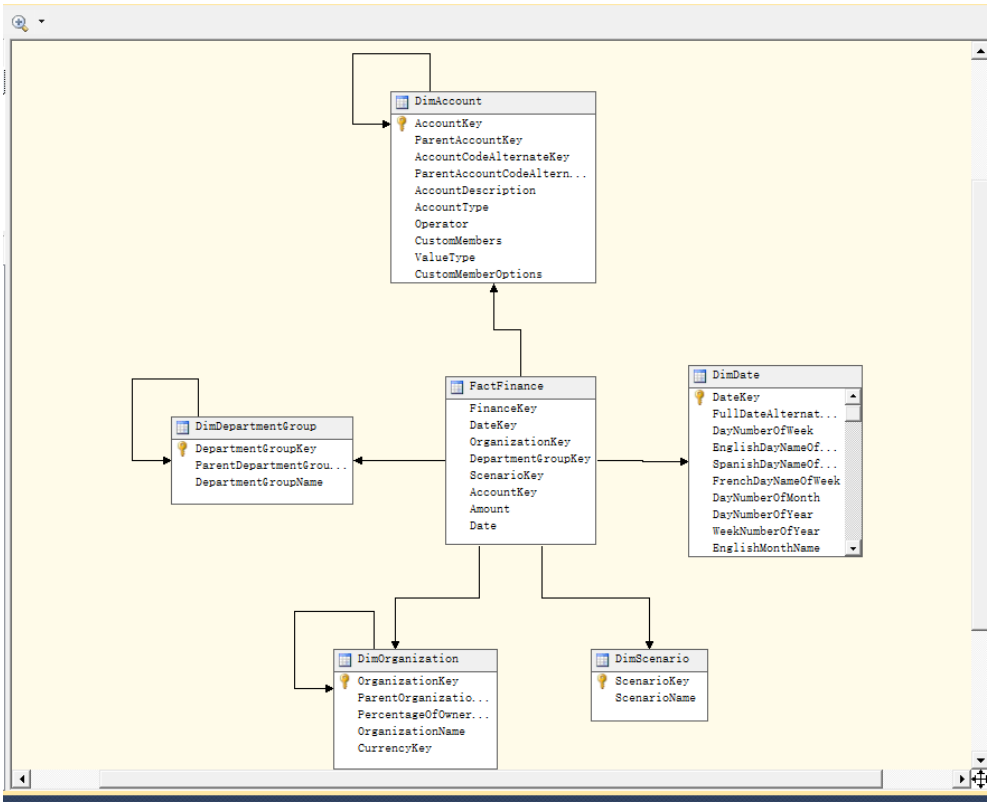
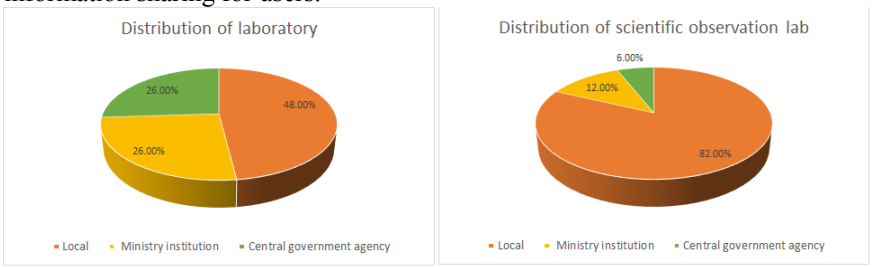


Fig. 3. Fact tables and dimension tables.

The display system for agricultural science and technology support data is developed mainly using the Reporting Services, PowerPivot, and SharePoint. Most users of this system will only use the standard reports, which created by the Reporting Services. PowerPivot tool can be used to combine various large data sets together for analysis, in order to find out new relationships and highlights. Finally, SharePoint platform is used to show users data, to provide users with the query services, and to facilitate communication and information sharing for users.



Name	Institution
Key Laboratory of Big Agricultural Genomics(Beijing)	Biotechnology Research Institute ,CAAS
Key Laboratory of Big Agricultural Genomics(Wuhan)	Huazhong Agricultural University
Key Laboratory of Crop Gene Resources and Germplasm Enhancement in ea	Nanjing Agricultural University
Key Laboratory of Crop Gene Resources and Germplasm Enhancement in Sc	Tropical Crops Genetic Resources Institute, CATAS
Key Laboratory of Crop Gene Resources and Germplasm Enhancement in Sc	Yunnan Academy of Agricultural Sciences
Key Laboratory of Crop Gene Resources and Germplasm Enhancement in nc	Jilin Academy of Agricultural Sciences

Fig. 4. Example of the data drill-down function.

The data are showed mainly through three methods: tables, bar charts and pie charts. Multidimensional data sets designed can provide the data drill-down function for dimension information with a hierarchical structure. Based on this, more detailed data can be obtained using drill-down function by clicking on a bar or pie chart. Specific implementation as shown in Fig.4.

4 Summary

In this paper, the exhibition implementation of agricultural science and technology support data based on Microsoft data warehouse is presented, including data warehouse design, dimensional modeling, data extraction, cleaning, translation and loading; the establishment of OLAP database; data exhibition. This system has laid a certain foundation for scientific management improvement of agricultural science and technology information. However, there are lots of in-depth study and improvement from data cleaning and extraction to the establishment of data warehouse and data exhibition. Such as the rationality of the data warehouse modeling, how to make the model better reflect the relationship between agricultural science and technology support data, how to improve system response efficiency and so on. These need to be further study in the future.

Reference

1. Inmon, Bill. Building the Data Warehouse. Wiley Publishing. 1992.
2. Kimball, Ralph. The Data Warehouse Toolkit. Wiley Publishing. 2011.
3. Mundy J, Thornthwaite W, Kimball R. The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset. Wiley Publishing, 2011.
4. Kimball, Ralph; Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley Publishing, 2013.
5. Herden O. Data Warehouse. Taschenbuch Datenbanken. 2015.
6. Runtuwene, J. P. A, et al. A Comparative Analysis of Extract, Transformation and Loading (ETL) Process. IOP Conference Series: Materials Science and Engineering 306(2018) 012066.