Vol. 4, Issue III, Jul-Sep, 2021 http://www.bharatpublication.com/journal-detail.php?jID=25/IJTSE

DEVELOPING A SMART INTEGRATED WAREHOUSING FOR THE SCIENCE AND TECHNOLOGY MANAGEMENT SYSTEM

Chandan Bansal

Shivaji College, University of Delhi, New Delhi, India

ABSTRACT

The science and technology management system plays a pivotal role in assimilating and leveraging vast data from distributed systems through data warehousing technology. This paper delves into this critical function. Initially, we provide an overview of science and technology management, elucidating project management business flows. Subsequently, we propose a tailored star model to delineate the structure of science and technology data, augmenting it with various dimensional tables to interlink diverse attributes with science and technology projects. Notably, we introduce a 4-D data cube to accommodate intricate query conditions in the science and technology domain. Additionally, we present an architecture for the Science and Technology Projects Data Warehouse (STPDW)—a five-layered system that aggregates raw data from science and technology management systems and processes it through a standardized workflow using Extract Transform and Load (ETL) tools with predefined rules. Finally, we outline a prototype system designed to realize the functionalities of the STPDW.

INTRODUCTION

Science and technology are playing a growing number of roles, and are the embodiment of a country's comprehensive strength in modern society, to a certain extent. Science and technology management innovation improves the scientific and technological level of a country, and usually needs to be supported by science and technology management systems [1].

Therefore, the systems have been popular in most universities, institutions, government in China as well as abroad because of its significance, such as National Science and Technology Information System (NSTIS) [2] and Internet-based Science Information System (ISIS) [3] in China, Project Information and Management System (PIMS) [4] in India, and Research Data Management (RDM) service in Japan [5]. Modern science and technology management requires data of comprehensive depth and breadth to improve our understanding of the progress and details of projects, and ultimately achieves data analysis, decision support, and supervision. Thus, data integration in the science and technology management system becomes a real business requirement. However, the data is distributed, heterogeneous, variant, voluminous and generated in a very high frequency, which complicated their extraction, integration, storage, and analysis.

Therefore, a integration method often called Data Warehouse (DW) has emerged for the management, maintenance and querying of data [6]. DW is the revolution in the world of database management systems, consolidating large amounts of data of interest from heterogeneous and distributed data sources into specialized database [7]. As a result, to face the big challenge of data integration in science and technology management systems, we adopt the data warehouse technology, which is based on effective fundamentals and a result of long work, findings and methods universally recognized [8]. This paper deals with the design of the customed data

warehouse in order to propose a solution for the integration of project data from science and technology management system.

RELATED WORKS

A. Data Warehouse

There is a large body of literature around adopting data warehouse technology in specific application areas, such as nursing [9], aeronautical service [10], clinician [11], question answering [12]. G. Garani [9] developed the conceptual and logical modelling of a semantic trajectory data warehouse, which stores data about nursing personnel shifts as trajectories of moving persons, and schedules nurses' shifts by the computation of On-Line Analytical Processing (OLAP) operations over the data. C. Tang [10] discussed the typical characteristics of Aeronautical Data Warehouse of Traffic Flow (ADWTF) as well as its relation with aeronautical database, found out the demands of characteristics of traffic flow, and designed a multi-dimensional star schema for parameters of traffic flow. N. Garcelon [11] described a clinical data warehouse named Dr. Warehouse, which is an open source data warehouse oriented toward clinical narrative reports and designed to support clinicians' day-to-day use. A. Ferrández [12] proposed an integrated method that is achieved seamlessly through the presentation of the data returned by the DW and the Question Answering (QA) systems into dashboards that allow the user to handle both types of data.

B. Science and Technology Management System

There have been some researches on science and technology management system recently. In China, the most typical systems are NSTIS [2] and ISIS [3]. The NSTIS is a complicated information system, which aims to support the management and implementation of national science and technology research and development programs and to provide all-round services for various groups of users in China [2]. ISIS is developed by the National Foundation Science of China (NSFC), which is constructed for managing the national natural science foundation of China, and serves the declaration and implementation of many kinds of research projects[3].

As in China, science and technology management systems have also been widely used in other countries. PIMS has been designed and developed in India to help in taking decisions to check duplication in research projects both at divisional, as well as inter divisional level for on-line monitoring and concurrent evaluation of the ongoing research projects and for other management requirements [4]. RDM service is a national scale system that is developed for the need in universities and research institutions to archive their research data in ten years for ensuring research integrity and the promotion of open science [5].

Nowadays, the massive data mainly generated from science and technology management systems is progressively increasing and difficult to be extracted, integrated, stored, and analyzed by conventional database methods. Combining with data warehouse technology, the huge amount of data should be integrated in a standard architecture. Objective of this paper is to design a data warehouse which meets the needs of data integration of science and technology management systems. To achieve the objective, this paper focus on the data model and system architecture of STPDW.

SCIENCE AND TECHNOLOGY PROJECT BACKGROUND

Considering the previous research on the management of projects [13], we divide project management processes into four stages which are related to three types of users. In this paper, we define the four stages as application, approval, implementation, and acceptance, and the three types of users as researchers, managers, and experts. In each stage, there is a particular management process, and different types of users do customized operations in sequence. For example, after project data is submitted by scientific researchers, managers review the data immediately, and then inform domain experts of further reviewing the data.

In general, in the four stages, business processes consist of the operations of researchers, managers, and experts. Therefore, the scheme of general project management business flows is presented. As shown in Fig.1, researchers and managers take part in the four stages. Particularly, experts are required for reviewing projects only in the first and fourth stages.

Furthermore, data transmission throughout the management processes is also illustrated by the arrows in Fig. 1. The solid lines represent the information transmission, and the dashed lines represent the message transmission. In summary, data is generated in each stage by different users, and then transferred from one stage to another in project management business flows.

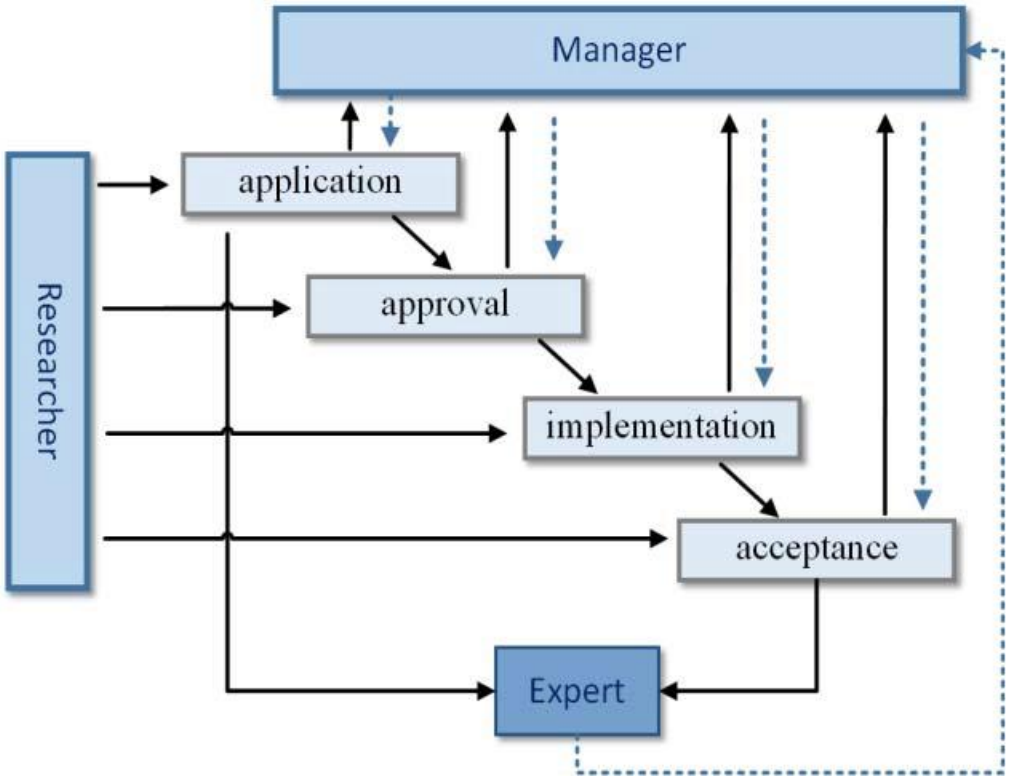

Fig. 1. Scheme of project management business flows.

DATA MODEL

A. Star Model

There are many models of data storage in existing researches, such as star model [14], snow model [15], and constellation model [16]. According to general introductions [17], the star model is a

multi-dimensional data relations consisting of a fact table and multiple dimension tables. The main information of the target object is stored in the fact table, and its different attributes are stored in the dimension table. These attributes are connected to the fact table through the primary key to form a data storage structure (star model). In the science and technology management systems, the project is the core object of data integration, including multi-dimensional attributes, such as field, stage, location, user, etc. Therefore, we adopt star model which is suitable for our study to illustrate the definition of data model.

According to the project management business flows, the fact table stores the information of the project including the project number, project name, project fund and so on. Many related attributes—such as the project filed, project stage, project user, and location—form different dimension tables of the project. These dimension tables and the fact table are connected by the primary keys, which forms the conceptual storage structure of the star model of the science and technology project, as shown in Fig. 2.

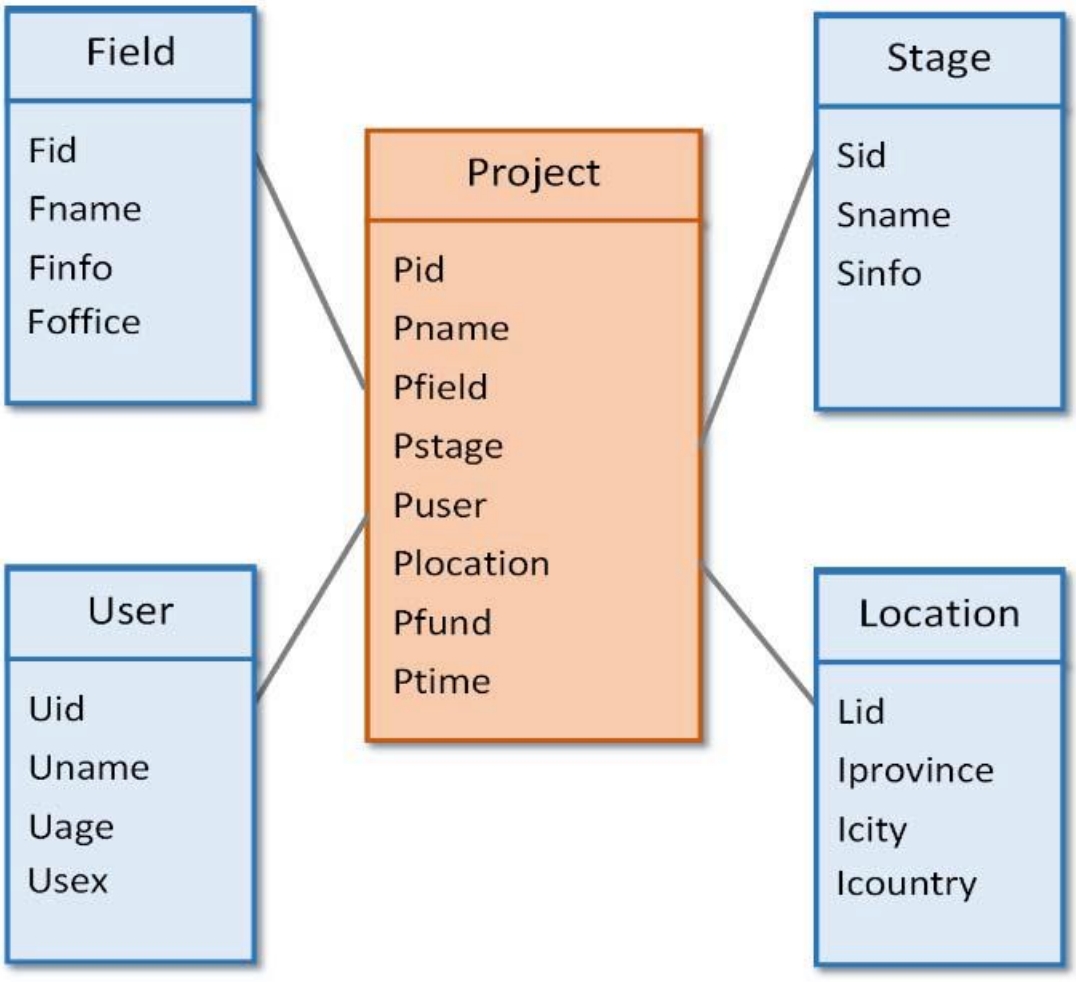

Fig. 2. Star model of science and technology project data.

B. 4-D Data Cube

Based on the basic concepts of OLAP, data cube is used to represent multi-dimensional data model [18]. In general, data cube is three-dimensional. However, it can be a N-dimensional cube, as an extension of the three-dimensional cube, achieving complicated query conditions. From the

previous section, the data of science and technology projects is multiple dimensions including stage, field, location, and user. These dimensions appear in the common query and

analysis scenarios of science and technology project management. As a result, we use the dimensions to construct a 4-D data cube, as shown in Fig. 3. S is the stage, F represents the field, L denotes the location, and U refers to the user. In the 4-D data cube, different dimensions are combined in sequence to satisfy the various query and analysis requirement. For example, if there is a requirement of calculating the number of projects in the biological field declared in Beijing area, a 2-D cuboid FL which represents the projects segmented by field and location dimensions is used for the query.

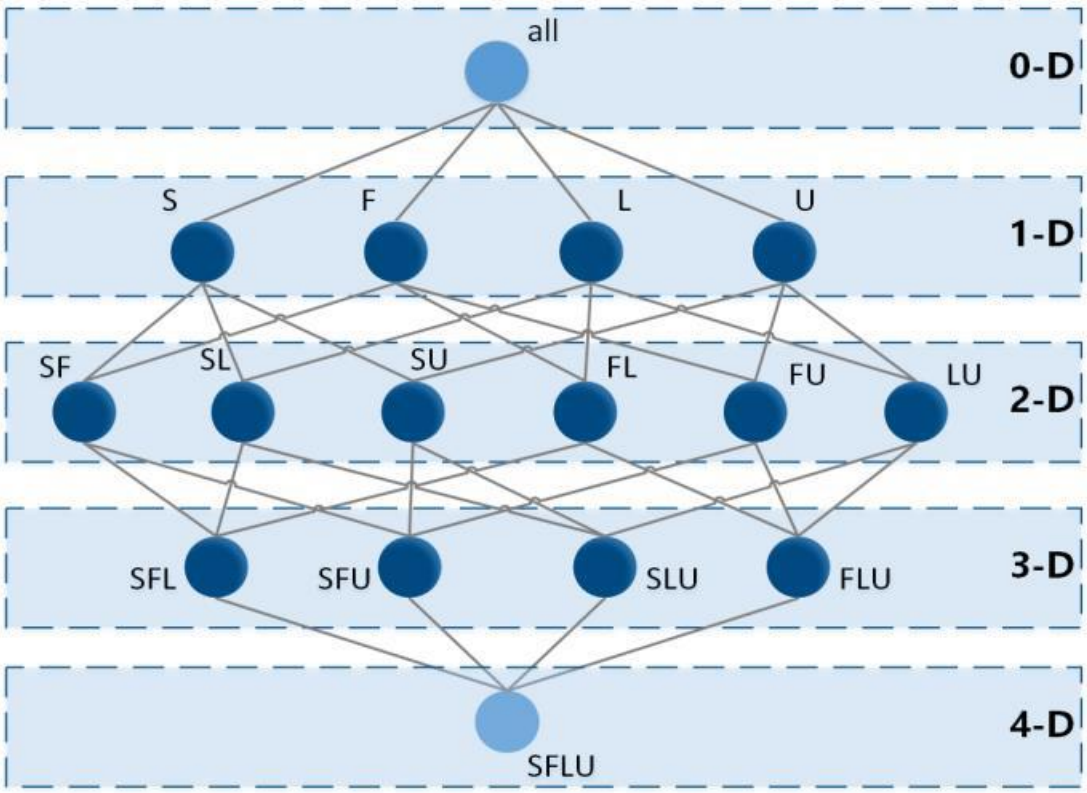

Fig. 3. 4-D data cube of project

DATA WAREHOUSE ARCHITECTURE

A. Architecture

Drawing from existing research on data warehouse architecture [19][20], our focus is on integrating data from original systems and utilizing the data model outlined in the preceding section to establish a suitable architecture for STPDW. Consequently, a five-layered architecture for STPDW is delineated in Figure 4, comprising the source layer, transform layer, quality layer, storage layer, and service layer.

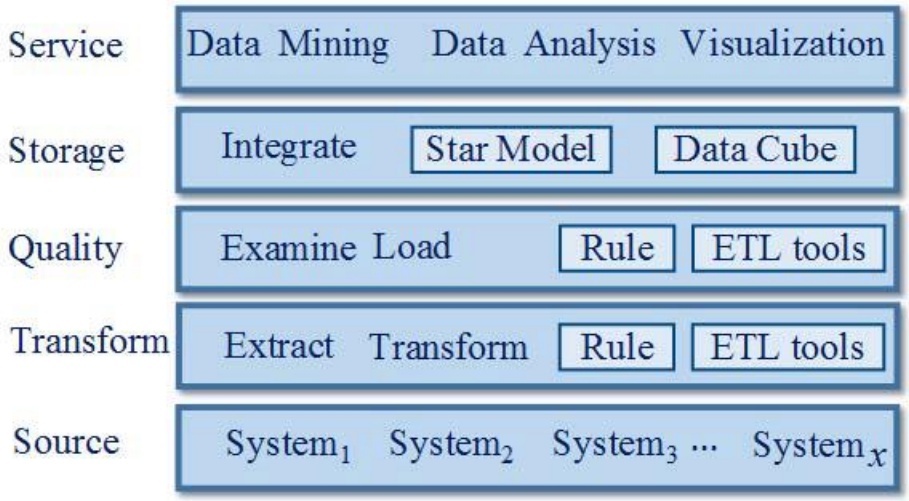

Fig. 4. Architecture of STPDW.

1) Source layer: This layer encompasses the original systems associated with the four project stages, housing both structured data (e.g., relational databases containing project data with diverse dimensions) and unstructured data (e.g., Word and PDF files). Common database software types include MySQL, SQL Server, Oracle, and HBase.

2) Transform layer: Serving as the initial stage of science and technology project data integration, this layer is tasked with extracting structured and unstructured data from these sources and transforming them via ETL (Extract, Transform, Load) tools such as Kettle.

3) Quality layer: Prior to loading data into STPDW, the quality layer plays a crucial role. Utilizing ETL tools, this layer establishes rules for validating data extracted from the transform layer, ensuring data quality during extraction, transformation, and loading processes.

4) Storage layer: Based on the star model and 4D data cube, the storage layer integrates science and technology project data, serving as the core of the architecture and facilitating unified data storage.

5) Service layer: Functioning as an interface, the service layer interacts with the storage layer to efficiently retrieve organized data, catering to various query and analysis requirements in science and technology business management.

B. Workflow

In current data integration research, data is typically replicated from source systems into warehouses using ETL processes [21]. To manage the complexity of such ETL processes, they are often implemented in environments offering libraries of connectors to different source types, transformation operators, and graphical workbenches for constructing complex workflows. A prominent solution is Pentaho Data Integration (PDI), the standard tool for data integration [22]. In this section, we outline the workflow of STPDW. Illustrated in Figure 5, the heterogeneous data process in the proposed workflow comprises extraction, filtering, transformation, examination, and loading.

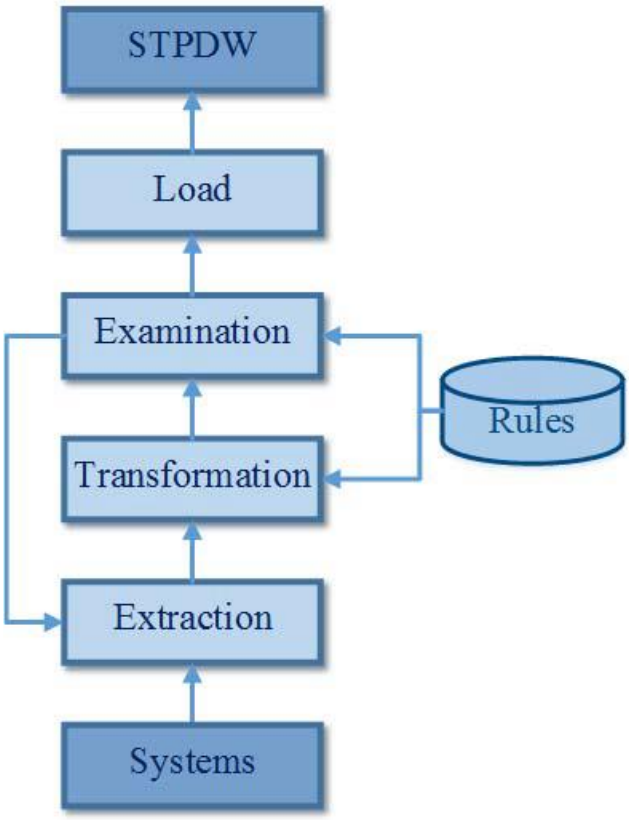

Fig. 5. Workflow of STPDW.

1) Extraction: This process involves extracting heterogeneous data from various systems deployed in the source layer, defining data extraction based on business requirements, establishing connections between ETL tools and source systems, and specifying the frequency and volume of data extraction.

2) Transformation: Transformation, achieved through data format and semantic processing, addresses inconsistencies in extracted data. A rule base is employed to define transformation rules for both format and semantic transformation, ensuring uniformity and accuracy.

3) Examination: This process examines the quality of data to be loaded into STPDW, enforcing high-quality data integration. Verification rules are defined to examine value, format, and semantics of transformed data. Issues encountered during data examination are logged for resolution.

4) Load: As the final process, loading involves establishing correspondence between original systems and the proposed data model, configuring necessary information such as path, table, and field in ETL tools, and executing data loading into STPDW. Figure 6 presents an example workflow integrating data from two sources.

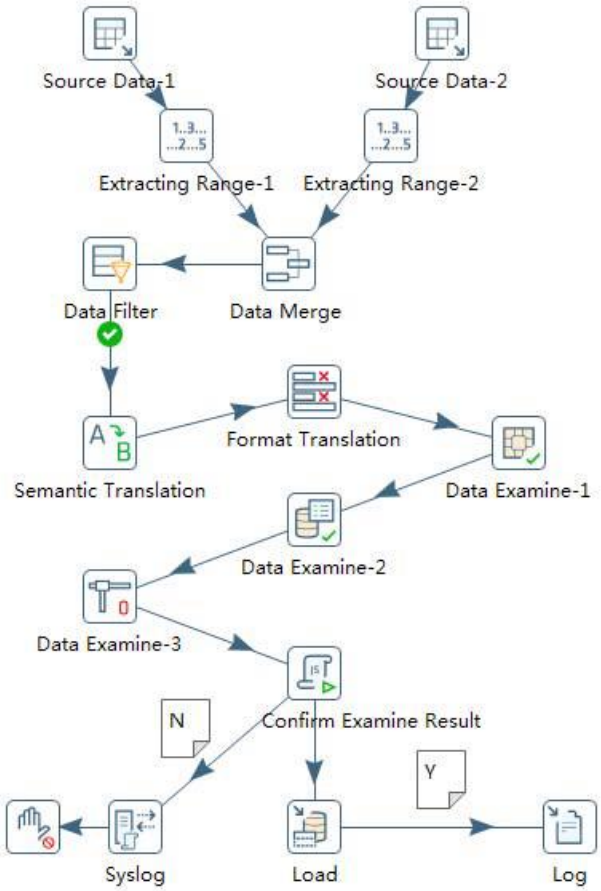

Fig. 6. An example of the workflow.

THE PROTOTYPE SYSTEM

To demonstrate the efficacy of STPDW, we have developed a prototype system supporting data statistics and analysis in science and technology management business flows. Illustrated in Figure 7, the prototype system showcases project distribution across different stages, fields, and locations. It provides data analysis and decision support for science and technology management departments, effectively integrating data from science and technology management systems.

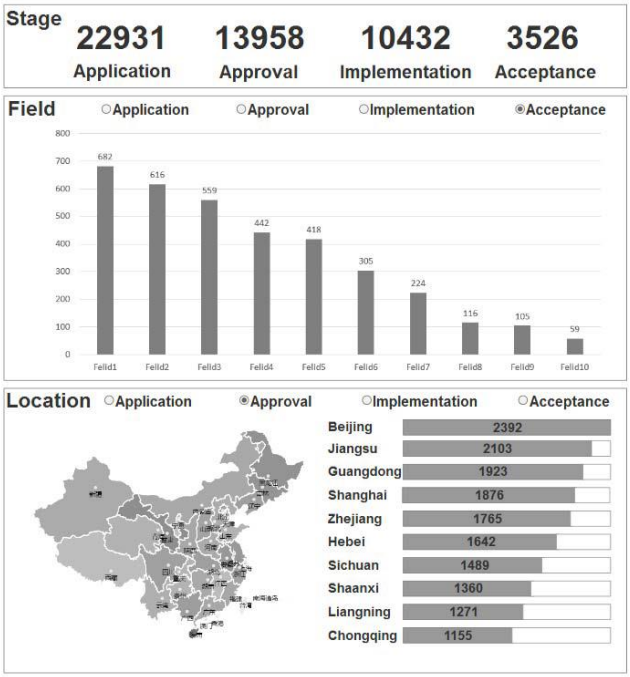

Fig. 7. Prototype system.

CONCLUSION

The essence of science and technology management lies in extracting information concealed within project data, a task that necessitates an effective data integration method for science and technology management systems. Designing a suitable architecture is thus pivotal for successful data integration. This paper has discussed the business context of science and technology management and proposed a star model and data cube to define project data structure. Subsequently, STPDW, a specialized data warehouse enabling integration of science and technology project data from distributed system environments, has been introduced. Furthermore, a five-layered architecture for STPDW has been presented, followed by an elucidation of the workflow using ETL tools. Finally, the proposed STPDW has been validated through a prototype system displaying data analysis across various project dimensions. In future endeavors, we aim to enhance STPDW with high-performance query methods integrated into the proposed architecture to achieve comprehensive data services.

REFERENCES

[1] Y. Bai, Z. Li, K. Wu, J. Yang, S. Liang, and B. Ouyang, et al., “Researchain: Union Blockchain Based Scientific Research Project Management System,” Proceedings 2018 Chinese Automation Congress (CAC), November 2018, pp. 4206-4209.

[2] S. Hu, Z. Wang, X. Song, J. Chen, H. Men, and D. Wang, et al., “Implementation and Application of National Science and Technology Information System,” China’s e-Science Blue Book 2018, 2020, pp.139-154.

[3] M. Hu, “Evaluation of the Internet-based Science Information System from the perspective of user's needs,” Bulletin of National Natural Science Foundation of China, 2016, Vol. 30, No. 4, pp.

360-364.[4] P. Singh, Sudeep, A. Arora, R. C. Goyal, and P. K. Malhotra, "Project information and management system of ICAR (PIMS-ICAR)," 2014 International Conference on Computing for Sustainable Global Development (INDIACom), March 2014, pp. 405-407.

[5] Y. Komiyama, and K. Yamaji, "Nationwide Research Data Management Service of Japan in the Open Science Era," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), July 2017, pp. 129-133.

[6] G. Garani, G. K. Adam, "A semantic trajectory data warehouse for improving nursing productivity," Health Information Science and Systems, Springer, 2020, pp:1-13.

[7] J. W. Teixeira, L. P. Annibal, J. C. Felipe, R. R. Ciferri, and C. D. A. Ciferri, "A similarity-based data warehousing environment for medical images," Computers in Biology and Medicine, Elsevier, 2015, Vol. 66, pp. 190-208.

[8] F. Jenhania, M. S. Gouidera, and L. B. Said, "Streaming Social Media Data Analysis for Events Extraction and Warehousing using Hadoop and Storm: Drug Abuse Case Study," Procedia Computer Science, 2019, Vol. 159, pp. 1459-1467.

[9] G. Garani, and G. K. Adam, "A semantic trajectory data warehouse for improving nursing productivity," Health Information Science and Systems, Vol. 8, August 2020, pp. 1-13.

[10] C. Tang, H. Wang, J. Liu, and C. Liu, "A design for AIS data warehouse of traffic flow," 2017 4th International Conference on Transportation Information and Safety (ICTIS), August 2017, pp. 895- 898.

[11] N. Garcelon, A. Neurazb, R. Salomona, H. Faoura, V. Benoita, and A. Delapalmea, et, al., "A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse," Journal of Biomedical Informatics, Vol. 66, April 2015, pp. 190-208.

[12] A. Ferrandez, A. Mate, J. Peral, J. Trujillo, E. D. Gregorio, and M. Aufaure, "A framework for enriching Data Warehouse analysis with Question Answering systems," Journal of Intelligent Information Systems, Vol. 46, December 2014, pp. 61-82.

[13] D. Wang, M. Zhou, S. Ali, P. Zhou, Y. Liu, and X. Wang, "A Novel Complex Event Processing Engine for Intelligent Data Analysis in Integrated Information Systems," International Journal of Distributed Sensor Networks, Vol. 2016, 2016, pp. 1-14.

[14] G. Garani, A. V. Chernov, I. K. Savvas and M. A. Butakova, "A Data Warehouse Approach for Business Intelligence," 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2019, pp. 70-75.

[15] A. Dahlan, and F. W. Wibowo, "Design of Library Data Warehouse Using Snowflake Scheme Method: Case Study : Library Database of Campus XYZ," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation, January 2016, pp. 318-322.

[16] T. Bhowmik, A. Sarkar and N. C. Debnath, "OLAP umbrella: Visualization model for multidimensional databases," ACS/IEEE International Conference on Computer Systems and Applications(AICCSA 2010), 2010, pp. 1-8.

[17] M. Barkhordari, and M. Niamanesh, "Chabok: a Map-Reduce based method to solve data warehouse problems," Journal of Big Data, Vol. 5, No. 1, October 2018, pp.1-25.

[18] J. Han, M. Kamber, J. Pei. "Data Warehousing and Online Analytical Processing", Data Mining, 2012, pp. 125-185.

[19] N. A. Farooqui, R. Mehra, "Design of A Data Warehouse for Medical Information System Using Data Mining Techniques," 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), December 2018, pp. 199-203.

[20] G. Blazi, P. Posi, and D. Jaksi, "Data Warehouse Architecture Classification," 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2017, pp. 1491-1495.

[21] F. Prasser, H Spengler, R. Bild, J. Eicher, and K. A. Kuhn, "Privacy enhancing ETL-processes for biomedical data," International Journal of Medical Informatics, 2019, Vol. 126, pp. 72-81.

[22] H. A. Sulistyo, T. F. Kusumasari, E. N. Alam. "Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration," 2020 3rd International Conference on Information and Communications Technology (ICOIACT), November 2020, pp. 12-16.