International Journal of
Georesources and Environment
http://ijge.camdemia.ca, ijge@camdemia.ca

IJGE 2017 3(1): 9-21
Available at http://ojs.library.dal.ca/ijge

Section – Geohazards and Environment

Automated Data Processing and Integration of Large Multiple Data Sources in Geohazards Monitoring

Chaoyang He[1,2], Nengpan Ju[1,2], Qiang Xu[1,2], Jian Huang[1,2]
[1]College of Environment and Civil Engineering, Chengdu University of Technology, 1 Third Road East, Erxianqiao, Chengdu, Sichuan, China 610059
[2]State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, 1 Third Road East, Erxianqiao, Chengdu, Sichuan, China 610059

Abstract: The development of geohazard information management system has greatly promoted the wireless automation monitoring technology for geohazards. More monitoring instruments are increasingly used in geohazard monitoring. Consequently, the types of monitoring data become more and more complicated, and massive amount of monitoring data are collected, which raises new demands in data storage and retrieval. In order to meet the requirements of data processing in geohazard monitoring, this paper presents a method of geohazard monitoring data processing, realizing the heterogeneous data integration, data access optimization, and abnormal data processing. Having analyzed the wireless automation monitoring process and the features of geohazard monitoring data, we defined the data integration standards of multiple data sources. Based on this, we developed a Geohazard Monitoring Data Integration System, with optimization in both hardware and software. This system allows automatic integration of large monitoring data from multiple sources. It has important significance for geohazard monitoring and early warning. A Geohazard Monitoring Data Analyzing System based on the monitoring data integrated by this system and data mining technology is developed to fully explore the hidden values of Big Data. Through field tests in Guizhou province with 92 sets of monitoring equipment and 5 types of databases, this method is proven to meet the system requirements with satisfactory performance.

Keywords: geohazard monitoring, monitoring and early warning, big data, data integration, data mining, data analyzing

1 Introduction

Geohazards occur frequently in China, especially those secondary geological disasters triggered by earthquakes greatly threaten life and property safety (Ju et al 2010, Parker et al 2011, Xiao and Li 2012, Huang and Fan 2009, Huang et al 2013, Wei et al 2014). Technology development in geohazard monitoring, information management, and Dynamic Monitoring and Early Warning System (Liu et al 2009), have greatly promoted the wireless automatic monitoring on geohazards. Many emerging monitoring instruments have been successively applied to the real-time monitoring on geohazards, such as GPS (Global Positioning System), GPR (Ground Penetrating Radar), TDR (Time-Domain Reflectometry), photogrammetry, infrasound monitoring, and 3-D laser scanning. Thanks to the development of information technology including Cloud Technologies, Cluster Technology, IOT (Internet of Things), Mobile Devices and Mobile Internet Technology, geohazard monitoring is advancing from the traditional manual operation to real-time wireless automation monitoring (Zhang et al 2009). This change has greatly improved the real-time and reliability of monitoring data, thus providing a better data support for geohazard warning. However, the monitoring work usually needs a variety of monitoring instruments at the same time. Those instruments are often produced by different vendors who have their own data collection system, along with constant updating and improvement as an integral part of the vendors' development. Thereby, the obtained monitoring data are different in structures, and also scattered in various independent databases with their own database management system (Chen and Liu 2010). These multi-source heterogeneous data constitute a large and complex dataset. Due to the difficulty of integrating all data into a uniform platform, monitoring data can only be analyzed in each individual monitoring instrument management system. This makes the analysis and management of monitoring data extremely complex.

Following with the rapid development of geohazard monitoring technology, the accumulation of monitoring data has increased dramatically, while data types have become more complex. Nowadays how to integrate the heterogeneous data and how to store and efficiently retrieve

* Corresponding Author: Chaoyang He, Email: hechaoyang2013@mail.cdut.edu.cn, Tel: +86 135-1820-6640

DOI: 10.15273/ijge.2017.01.003

the vast amounts of diverse monitoring data (Big Data) have become a challenge. This paper presents the development and construction of the Geohazard Monitoring and Early Warning Platform to process Big Data sets, as well as applications of this system in some monitoring work we have done.

2 Features of Geohazard Monitoring Data

2.1 Multi-sources

Geohazard monitoring data are featured with multi sources when multiple monitoring instruments are used simultaneously in the field. Data stored in those instruments can be in different databases, such as Oracle, Microsoft SQL Server, MySQL, Access, even a text file. To integrate those multi-source data, it is necessary to develop a corresponding data access interface for each type of database. This interface also has to be flexible to switch and expand.

2.2 Heterogeneousness

Heterogeneousness is one of the most significant features of geohazard monitoring data, referring to non-uniform data structure. Under general circumstances, due to the lack of industry data protocol, data structures are prescribed by manufacturers themselves. As a consequence, data collected from different monitoring instruments, or even the same type of monitoring instrument produced by different manufacturers, do not have a unified structure. Countless different data dictionaries and various data qualities present great difficulty in data integration. The ultimate reason of data heterogeneousness is due to the lack of monitoring data standards and a common protocol among software developers. Meanwhile, the software developers are usually not familiar with the professional knowledge in the field of geohazard monitoring, therefore are not able to strictly apply the corresponding specification to establish a data dictionary, instead, they develop the software from the point of view of Software Engineering, resulting in the variation of data features, including column names, data types, and data structure. The biggest problem is that most software developers are only concerned with the realization of the software function, but neglecting the uniformity of the data structure. The key to solving the problem of heterogeneous data is Data Mapping, i.e., according to the original data to establish the corresponding data mapping relations to extract data of interests.

2.3 Big Data

Big Data refers to the huge amount of geohazard monitoring data. Since monitoring instruments collect data over day and night, as time goes on, data stored in the database becomes larger and larger. In our monitoring platform, there are 12 types of monitoring instruments, including rain gauge, mud level gauge, GPS, deep displacement, soil moisture, etc., a total of 193 sets of instruments, monitoring rockfall, landslides and debris flow in 32 stations across Sichuan, Guizhou, Anhui and Gansu provinces. As of May 1, 2016 at 00:00:00, the cumulative data collection is of totally 32,532,594 rows, and increasing at a daily rate of about 36,000 rows.

3 Key Techniques for Data Integration

With the development and popularization of Distributed Application, and increasing improvement of Independent Research and Development Platform, data structures become more and more complex. Therefore, data Integration System requires high scalability of the platform to meet the requirements for data Plug-and-Play, of which the traditional data integration techniques is not able to achieve. This Multi-Source Heterogeneous Data Integration System is based on Service-Oriented Architecture (SOA) System and uses C# .NET combined with multiple database technologies. It runs as a System Service program. The Middleware is built with Data Mapping, Data Conversion and other technologies based on Three-tier Architecture. Three-tier Architecture (Eckerson 1995) is a client/server (C/S) software architecture pattern in which the user interface (presentation), functional process logic (business rules), computer data storage and data access are developed and maintained as independent modules, mostly on separate platforms. It does not need to change original data forms, but simply modify the configuration file to achieve data integration. It also is compatible to custom SQL (Structured Query Language) to support different types of data sources.

3.1 Multiple Database Support

The key to integrate heterogeneous multi-source data is that the system has to support multiple databases, as aforementioned, including Microsoft SQL Server, Oracle, MySQL and other databases. Based on C#.NET, we implemented the interface provided by System.Data.dll (Fig. 1), including IDbConnection, IDbCommand, IDbDataAdapter, and IDataParameter, to develop a Universal Data Access components library (HCY.DBUtility.dll, Fig. 2).

The key Class and Method in HCY.DBUtility shown in Fig. 2 have achieved common database operations, such as connection, disconnection, retrieval, adding, deleting, updating and query optimization for massive data (DbHelper class). Also, all the database operations use certain parameters to avoid SQL injection attacks. The database connection string, including database user name and password, uses MD5 encryption to protect the security of the database. The dynamic link library (HCY.DBUtility.dll) support most common existing databases (MySQL, Microsoft SQL Server, Oracle, SQLite, etc. Fig. 2 - DatabaseType) and can easily be extended to support other databases as well.

3.2 Heterogeneous Data Processing

For heterogeneous data integration, we need to extract all data that are scattered in multiple databases and then insert them into one database (Data Center). Therefore, it is necessary to establish the corresponding data mapping relations, and create a unified data structure, in other words, to form a unified data standard. Based on the characteristics of the geohazard monitoring data, we created a monitoring data table in the data collection platform, which is mainly comprised of monitoring data encoding, monitoring time, data compositions and values.

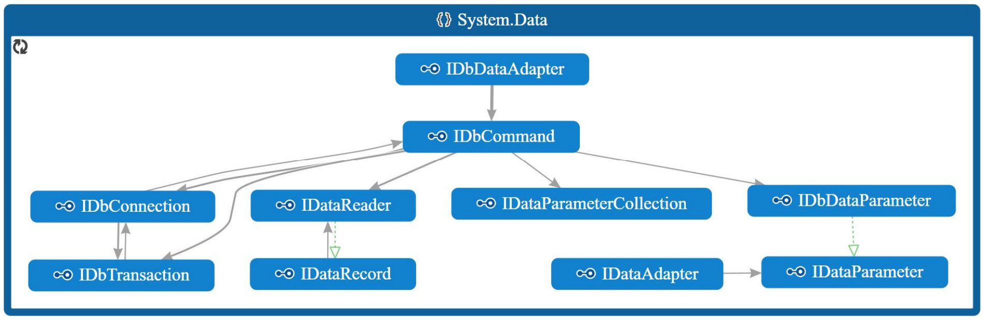


Fig. 1 Part of the database operation interface in namespace of System.Data

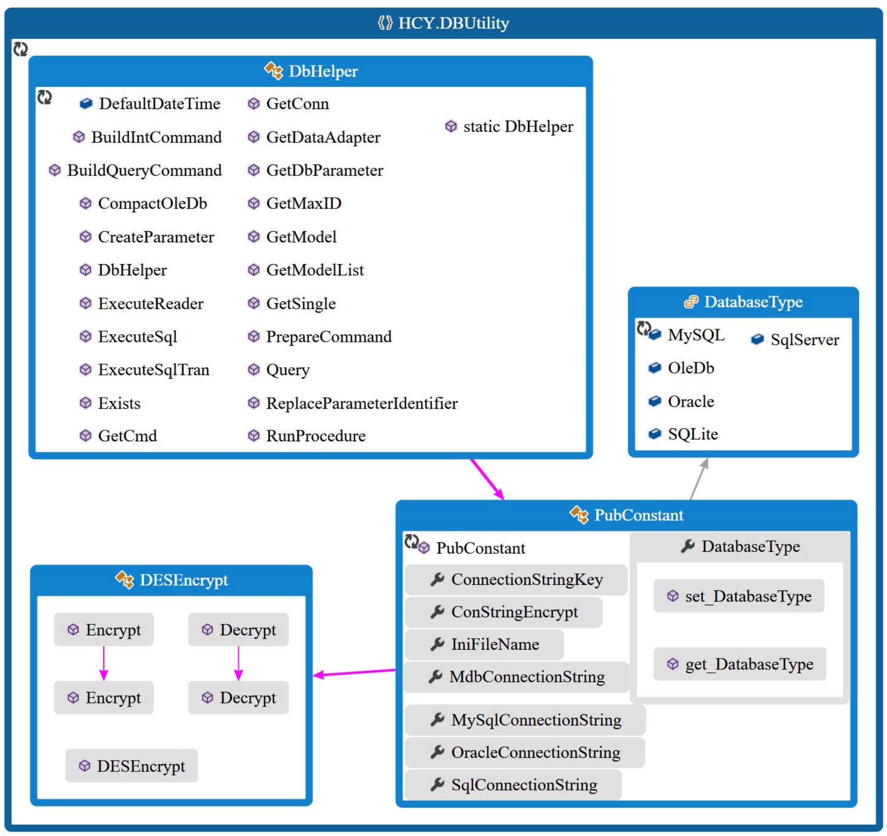


Fig. 2 Class diagram of HCY.DBUtility

Some monitoring instruments collect multiple values simultaneously, for example, a GPS collects three-dimensional data in X, Y and Z directions. So with consideration of these actual situations, we defined monitoring data coding rules. It consists of 18 characters (Fig. 3): Geohazard Number (12 characters) + Monitoring Type Code (2 characters) + Monitoring Number (2 characters) + Data Type Code (2 characters). Table 1 shows some of the "Monitor Type Code" and "Data Type Code", e.g. "520121010001YL0101" represents the rainfall of #1 rain gauge in Longjingwan landslide located in Kaiyang County, Guizhou Province, China.

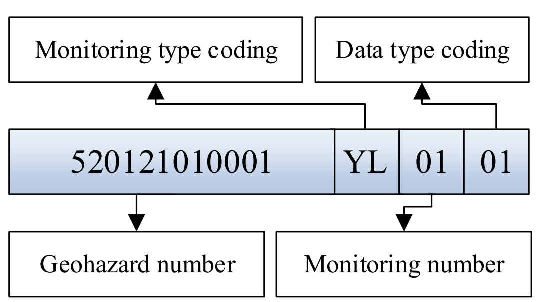


Fig. 3 Monitoring data coding rules

Table 1 Part of geohazard monitoring type coding rules

Monitoring Type Encoding	Data Type Coding	Meaning	Unit	Remark
CJ	01	settlement	mm	Settlement meter
	10	voltage	V	
CL	01	the tensile force of cable	KN	Dynamometer
	10	voltage	V	
GP	01	X-displacement	mm	GPS
	02	Y-displacement		
	03	Z-displacement		
	10	voltage	V	
HH	01	water content	%	Water cut meter
	10	voltage	V	
LF	01	the width of crack	mm	Crack meter
	10	voltage	V	
NW	01	mud meter	m	Mud Meter
	10	voltage	V	
QX	01	inclination of direction A	°	Inclinometer; direction A is perpendicular to direction B
	02	inclination of direction B		
	10	voltage	V	
SQ	01	displacement of 1# direction A	mm	Deep inclinometer; 1#, 2# and 3# are measuring points with different depths; direction A is perpendicular to direction B
	02	displacement of 1# direction B		
	03	displacement of 2# direction A		
	04	displacement of 2# direction B		
	05	displacement of 3# direction A		
	06	displacement of 3# direction B		
	10	voltage	V	
SY	01	pore-water pressure	KPa	Osmometer
	10	voltage	V	
YL	01	rainfall	mm	Rain Gauge
	10	voltage	V	

3.3 Data Access Optimization

Data storage is a very important part of the geohazard monitoring system. Because the amount of data collected from monitoring systems is very large, it requires a very high storage capacity with good security and efficient retrievals. Combining our previous experience with the actual situation of this project, we mainly focused on hardware and software to ensure the data security and retrieval efficiency.

3.3.1 Hardware Optimization

For hardware optimization, we focus on data security, stability and accessing speed. Existing facilities in our Information Center provide a good operating environment and hardware platform for data integration, among of which the RAID 5 (Redundant Arrays of Independent Disks) storage system consists of 24 pieces of 1TB hard drives and is equipped with optical fiber switches to ensure the data safety and the data accessing speed.

Disk Array (Chandy 2008, Thomasian and Xu 2011) is a combination of two or more disks, with the same types, capacity and interface, managed by a disk array card. Disk Array selectively distributes data into multiple disks, which improves not only data accessibility, but also data accessing speed and fault tolerance, thereby avoiding disastrous consequences caused by disk failure. There are 6 common methods of RAID configurations (Liu et al 2005), including RAID 0, RAID 1, RAID 2, RAID 3, RAID 4 and RAID 5.

3.3.2 Software Optimization

In the aspect of software, we focused on database optimization.

(1) Partition Storage

As described in section 2.3, monitoring data recorded in our system has achieved over 32 million rows. If all of those data were stored in one table or a single disk partition section, it would have low efficiency for data retrieval. Our system was based on the Oracle Database, according to Oracle (Oracle 2002), if a data table is larger than 2GB, or the data table contains many historical data, it is recommended to use the partition storage solutions. Table 2 shows a monthly statistics of monitoring data in our system. There are over 0.8 million rows data added to the system per month, and even more during rainy seasons (e.g. June and July in 2015 amounted to 3.23 million rows). The Oracle 10g (a version of Oracle Database) supports 1024 k-1 = 1,048,575 partitions (Oracle 2006). If our data are stored in partitions by month, it can be used for 87,381 years. So the Oracle 10g is enough to meet the actual demand, therefore we designed the partitioning scheme on a monthly base.

Table 2 Monthly statistics of monitoring data in Monitoring Data Integration System

Month	Data Rows	Month	Data Rows
May-2016	1,184,205	Jul-2015	1,591,245
Apr-2016	792,028	Jun-2015	1,639,137
Mar-2016	495,384	May-2015	842,929
Feb-2016	892,912	Apr-2015	852,069
Jan-2016	991,139	Mar-2015	679,962
Dec-2015	653,035	Feb-2015	710,902
Nov-2015	506,655	Jan-2015	871,699
Oct-2015	427,962	Dec-2014	829,768
Sep-2015	530,300	Nov-2014	702,259
Aug-2015	1,069,122	Oct-2014	892,168

(2) Arterialized View

View is a virtual table consisting of a set of columns and rows defined by a query (SQL). However, a View does not contain real data. What define the rows and columns in a View referenced in the query are produced dynamically when the query is run. The query that defines the View can be from one or more tables or from other Views in one or more databases. Distributed queries (queries that access data from multiple data sources) can also be used to define Views that pull data from multiple heterogeneous sources, such as a SQL Server database, an Oracle database, a text file or an excel file etc.

There is a need in the "Geohazard Monitoring Data Analyzing System" of comprehensively analyzing existing monitoring data to extract additional information, namely Data Mining. With a large amount of data, the time of view retrieval is generally 8~10 seconds. This long waiting time for each operation is unacceptable to users. Therefore, we created a Materialized View using the ON PREBUILD TABLE provided by Oracle. The amount of data produced by data mining (analysis result) in Materialized View is small with a response time of milliseconds. Thus, when retrieved from the materialized view, the user does not substantially feel the delay.

3.4 System Service

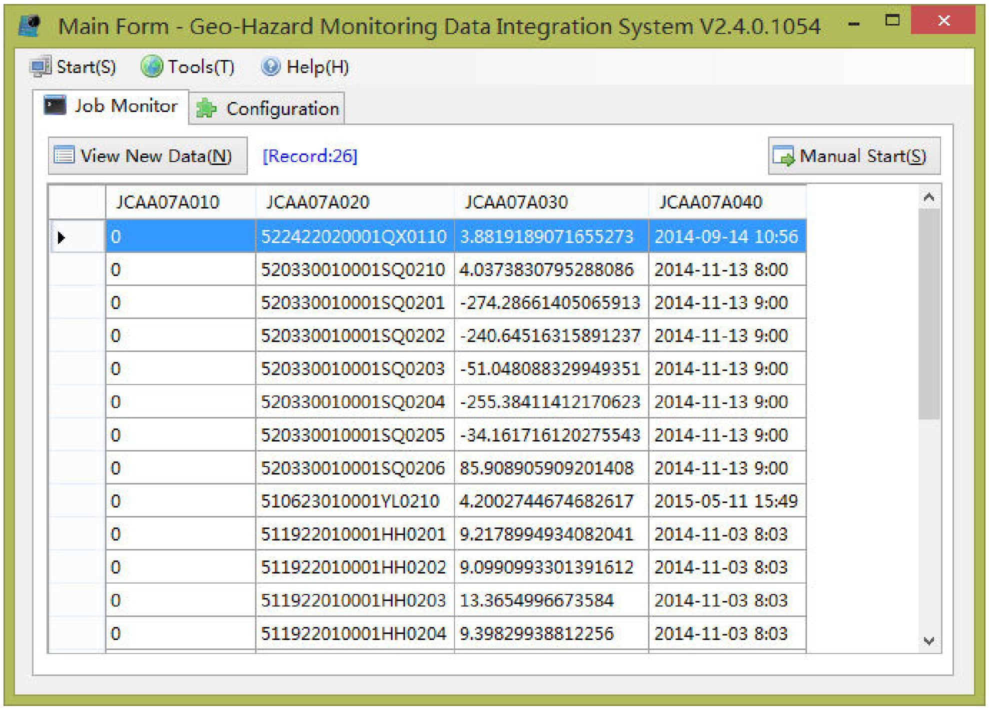
During monitoring process, data are continuously collected. It requires the system must keep stable and run smoothly all the time. Once the system has any problems or stops working, it will inevitably lead to delay in data collection, or even lose data during the entire failure period. "System Service", defined at the designing stage of the Monitoring Data Integration System, is particularly target to this problem (Fig. 4), to ensure an integrated system with the server running automatically without user's intervention.

In addition, to avoid memory overflow after the system running over long time, we have conducted some specific treatments in memory management for variables and database connection. We also developed a system service monitoring module (HCY Service Watcher) to monitor the system operation. In any exceptional cases that the system fails, the service will reboot the system immediately and also automatically send E-mails or short messages to the administrator, to ensure the monitoring data can be processed immediately.

4 Wireless Automation Monitoring of Geohazard

Based on Internet of Things and the characteristics of the geohazard monitoring, we built a Geohazard Wireless Automatic Monitoring System (Fig. 5). The system is divided into four parts: data collection, data transfer, data processing and data application. These four functions correspond to the core work of geohazard monitoring, i.e. data collection, transmission, processing, storage, analysis and application. In this paper, we focus on monitoring data processing.

(1) Data Collection
Typically, a variety of types of monitoring equipment, including rain gauge, inclinometer, GPS, are installed in a field monitoring project. Monitoring data are acquired through field monitoring instruments automatically, as shown in Fig. 5.



JCAA07A010	JCAA07A020	JCAA07A030	JCAA07A040
0	522422020001QX0110	3.8819189071655273	2014-09-14 10:56
0	520330010001SQ0210	4.0373830795288086	2014-11-13 8:00
0	520330010001SQ0201	-274.28661405065913	2014-11-13 9:00
0	520330010001SQ0202	-240.64516315891237	2014-11-13 9:00
0	520330010001SQ0203	-51.048088329949351	2014-11-13 9:00
0	520330010001SQ0204	-255.38411412170623	2014-11-13 9:00
0	520330010001SQ0205	-34.161716120275543	2014-11-13 9:00
0	520330010001SQ0206	85.908905909201408	2014-11-13 9:00
0	510623010001YL0210	4.2002744674682617	2015-05-11 15:49
0	511922010001HH0201	9.2178994934082041	2014-11-03 8:03
0	511922010001HH0202	9.0990993301391612	2014-11-03 8:03
0	511922010001HH0203	13.3654996673584	2014-11-03 8:03
0	511922010001HH0204	9.39829938812256	2014-11-03 8:03

Fig. 4 Data transfer system service

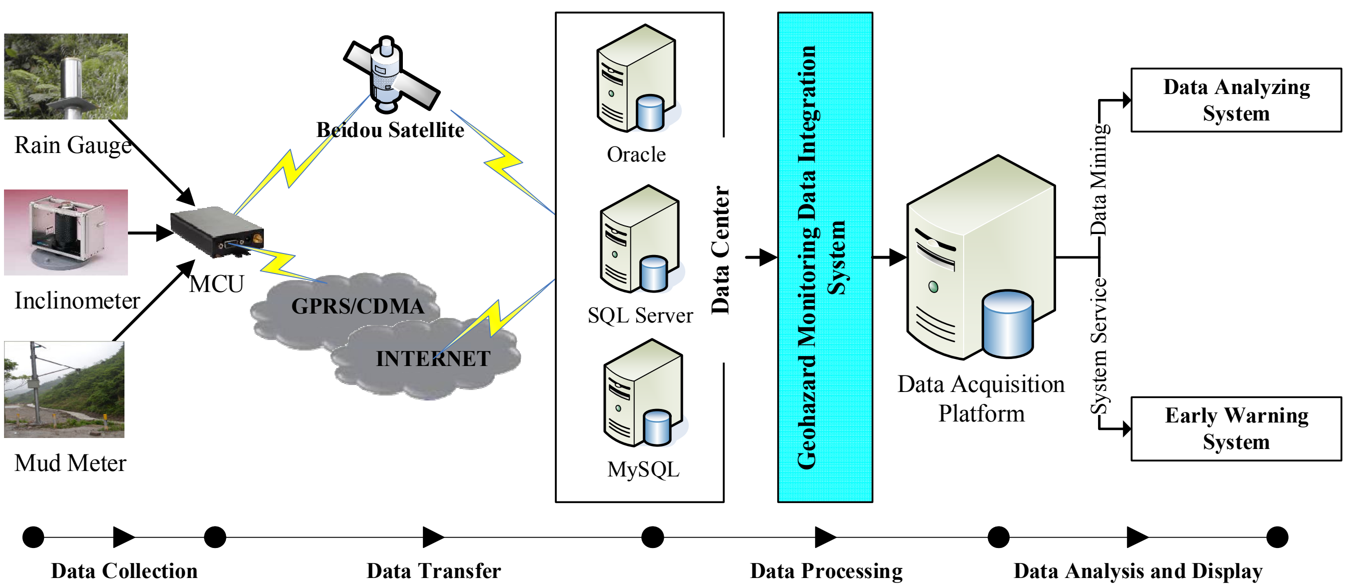


Fig. 5 Automatic monitoring system of geohazard modified from He et al (2014)

(2) Data Transfer
Monitoring data are sent to the Monitor Data Center via wireless communication modules, e.g., GPRS (General Packet Radio Service), 3G/4G, Wi-Fi, even Beidou satellite communication if monitoring equipment networks are available. As monitoring devices are from different manufacturers, monitoring data collected by these devices are usually stored in multiple databases, such as Oracle, Microsoft SQL Server, or Access, and do not have a unified data structure.

(3) Data Processing
Data Processing includes abnormal data processing and data integration. The process of data collection may be

interrupted or abnormal under some circumstances, e.g. when battery runs low, network is disconnected, the instruments fail or, a sensor is damaged. The monitoring data can be wrong or even parts of them are lost. So we have to deal with those abnormal data. To integrate data, we developed a Geohazard Monitoring Data Integration System (GMDIS) based on the corresponding data dictionary and storage rules, in which the monitoring data in different structures are integrated into the Data Acquisition Platform.

(4) Data Analysis and Display
Based on the data in the Data Acquisition Platform, we adopted data mining technology to develop the Monitoring Data Analyzing System, and a mobile client interface. Additionally, combining the Web Service (using C#) and AJAX (Asynchronous Java script and XML) technology with some drawing components, we developed a Monitoring Data Display Platform to comprehensively analyze the monitoring data. In addition, we developed a Geohazard Early Warning System based on System Service (He et al 2014).

5 System Design and Implementation

5.1 System architecture

This Data Integration System uses Oracle database (11 g R2), which is installed on a Windows Server and freely accessible through the Internet. All monitoring data are stored in this database. The Data Integration System are based on SOA, using C# language combined with multiple database technologies. All parameters in this system were defined in a configuration file, and run in the windows service mode. A Three-tier Architecture (Presentation Layer, Business Logic Layer and Data Access Layer) and Data Mapping and Transformation Technology are used to build the Middleware. Data integration can be achieved without changing original data storage and management methods, but only modifying the configuration files according to corresponding rules. This Middleware also supports custom SQL statements for a variety of data sources. Its configuration is very flexible.

5.2 System operation process

The common method for data integration is ETL (Extract, Transform and Load) (Papastefanatos et al 2012) (Fig. 6). The process contains three steps based on Schedule Table: 1) extracting data of interests from various data sources through the Data Access Components (HCY.DBUtility.dll) to get the Original Dataset (Extract); 2) using the Middleware to process the original dataset (mapping transform, formula transform, abnormal data processing, etc.) to get the Final Dataset; 3) inserting the Final Dataset into the database by the Data Access Components (Load).

The core of this system is the Middleware. Middleware is located between the heterogeneous database system (Data Layer) and the target database system (Application Layer). In the upper-stream, it provides multiple source databases with data standard and data access interface; in the down-

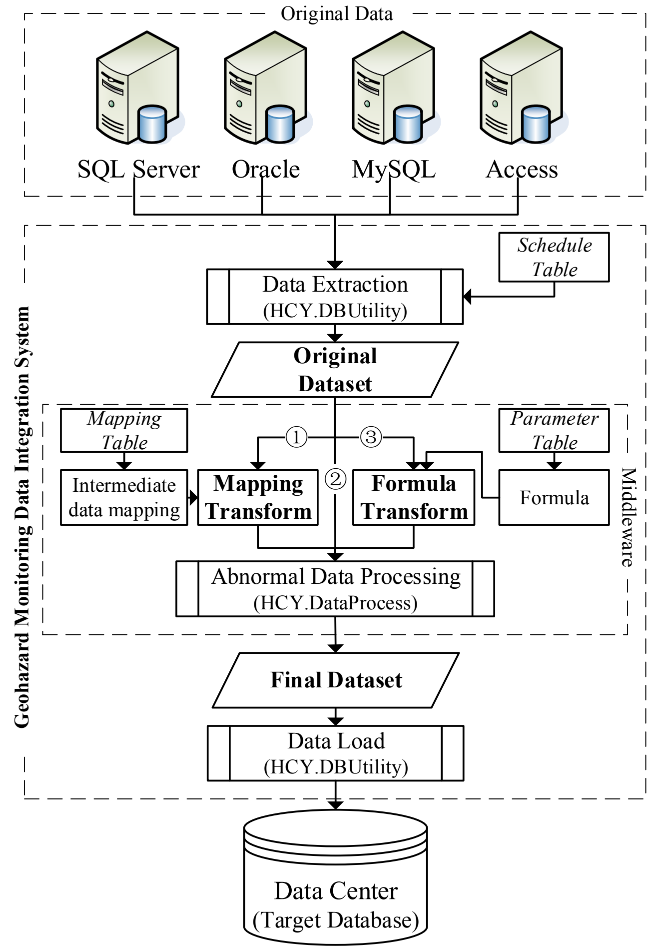


Fig. 6 A diagram of multi-source heterogeneous data integration

stream, it provides a uniform data format, and ultimately achieves the goal of the multi-source heterogeneous data integration. Each source database system runs independently without any interference. The mission of the Middleware is to support heterogeneous data integration in data Retrieval and Filtering. It has two main functions:

(1) Unifying Data Structure
In most cases, Original Dataset cannot be directly inserted into target database due to their different data structures. There are three common ways to format an Original Dataset, Mapping Transform, Direct Extraction and Formula Transform (Fig. 6 ①, ② and ③).

(i) Mapping Transform
This method applies where the source data and target data tables have differ rent field names but no other special processing of the data is needed except mapping each data field. The common approach is to use database function, i.e. restructuring the dataset using SQL. Samples of codes are shown as follows, and also in Fig. 7.

```
select top 100 ID as JCAA07A010, @pid as JCAA07A020, PadValue as JCAA07A030, PGsmTime as JCAA07A040 from TrackTable where DeviceID = @id and PGsmTime > @start_time order by PGsmTime
```

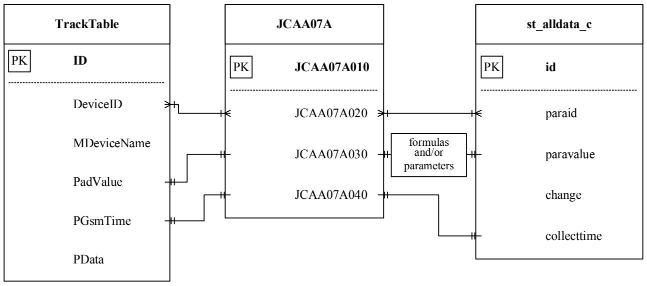


Fig. 7 Data fields mapping. L: Original Dataset (Mapping Transform); M: Final Dataset; R: Original Dataset (Formula Transform)

The original data and transformed data in this sample are shown in Tables 3 and 4.

(ii) Direct Extraction

When the structure of the source data and target data is consistent, the original dataset can be extracted directly without any special operation. This is a special case of "Mapping Transform".

(iii) Formula Transform

If the source data are of original forms recorded in monitoring instruments, data conversion will need certain formulas and/or parameters from the configuration file depending on the monitoring instruments. Based on this, the original records are converted to physical quantities, and then the converted data are inserted into the target database. This is the most common situation in data transformation. Take Osmometer as an example, it uses the following equation for data transformation:

$$P = G\,(R_1 - R_0) - k\,(T_1 - T_0) \qquad (1)$$

where, P is the Osmotic Pressure (KPa), G is the Calibration Coefficient (KPa/Digit), k is the Temperature Correction Coefficient (KPa/℃), R_0, R_1 is the Original Data (Digit) at initial time and study time, respectively, and T_0, T_1 is the Temperature (℃) at initial time and study time, respectively.

For convenience, we stored parameter values of each monitoring instrument in the configuration file. The following row is an example of a monitoring instrument's parameters stored in the configuration file:

3=02B80001000301|520121010001SY0201|0|-0.1771361,8000.5,15.3,-0.00232

Table 3 The raw data stored in the source database (Fig. 7L– TrackTable)

ID	DeviceID	MDeviceName	PadValue	PGsmTime	PData
62334	BD000024	Kualiangzi 1#	237	2016/05/19 09:01:46	<Binary>
62336	BD000024	Kualiangzi 1#	237	2016/05/19 10:01:46	<Binary>
62338	BD000024	Kualiangzi 1#	239	2016/05/19 11:01:46	<Binary>
62340	BD000024	Kualiangzi 1#	239	2016/05/19 12:01:46	<Binary>
62342	BD000024	Kualiangzi 1#	239	2016/05/19 13:01:46	<Binary>

Table 4 The transformed data stored in the final database (Fig. 7M - JCAA07A)

JCAA07A010	JCAA07A020	JCAA07A030	JCAA07A040
3345163	510623010001LF0101	237	2016/05/19 09:01:46
3345836	510623010001LF0101	237	2016/05/19 10:01:46
3346484	510623010001LF0101	239	2016/05/19 11:01:46
3347129	510623010001LF0101	239	2016/05/19 12:01:46
3347789	510623010001LF0101	239	2016/05/19 13:01:46

Table 5 The raw data stored in the source database (Fig. 7R– st_alldata_c)

id	paraid	para value	change	collecttime
3739381	02B80001000301	27596	27	2016/01/01 01:59:55
3739884	02B80001000301	27612	16	2016/01/01 08:00:03
3740454	02B80001000301	27571	-41	2016/01/01 14:00:08
3740957	02B80001000301	27603	32	2016/01/01 20:00:05
3741440	02B80001000301	27610	7	2016/01/02 02:00:14

Table 6 The processed data stored in the final database (Fig. 7M - JCAA07A)

JCAA07A010	JCAA07A020	JCAA07A030	JCAA07A040
29917604	520121010001SY0201	68.30529	2016/01/01 01:59:55
29921713	520121010001SY0201	66.74060	2016/01/01 08:00:03
29925722	520121010001SY0201	70.74831	2016/01/01 14:00:08
29929072	520121010001SY0201	67.62085	2016/01/01 20:00:05
29932429	520121010001SY0201	66.93624	2016/01/02 02:00:14

We can get the parameters (G = -0.1771361 KPa/Digit, R_0 = 8000.5 Digit, T_0 = 15.3 ℃, k = -0.00232 KPa/℃) from this row. The osmotic pressure values (Table 6) can be calculated after these parameters and raw data (Table 5) are taken into formula (1).

(2) Abnormal Data Processing

With normal data, after Middleware has unified data structure, source data can be directly inserted into the target database. However, under some circumstances data collection is interrupted, such as low battery, network disconnection, instruments failures or, a damaged sensor, this can make the monitoring data wrong or even result in data loss. We have to deal with those abnormal data, and make the final dataset be able to reflect the real situation to the largest extent. A common processing method is to integrate them into a module named HCY.DataProcess. This module provides some functions, e.g. AGO (Accumulated Generating Operation, a data processing method for de-noising), to deal with the abnormal data. With the aid of this module, the original monitoring data can also be backup automatically at the same time, which fully guarantees the authenticity and reliability of the monitoring data.

6 Research Results and Applications

To verify the effectiveness of this data integration system, a test was conducted on the Geohazard Monitoring and Early Warning Platform in Guizhou province (Fig. 8). Twenty monitoring locations (Table 7) were chosen as a demonstration pilot project of automation geohazard monitoring. It included a total of 92 sets of monitoring

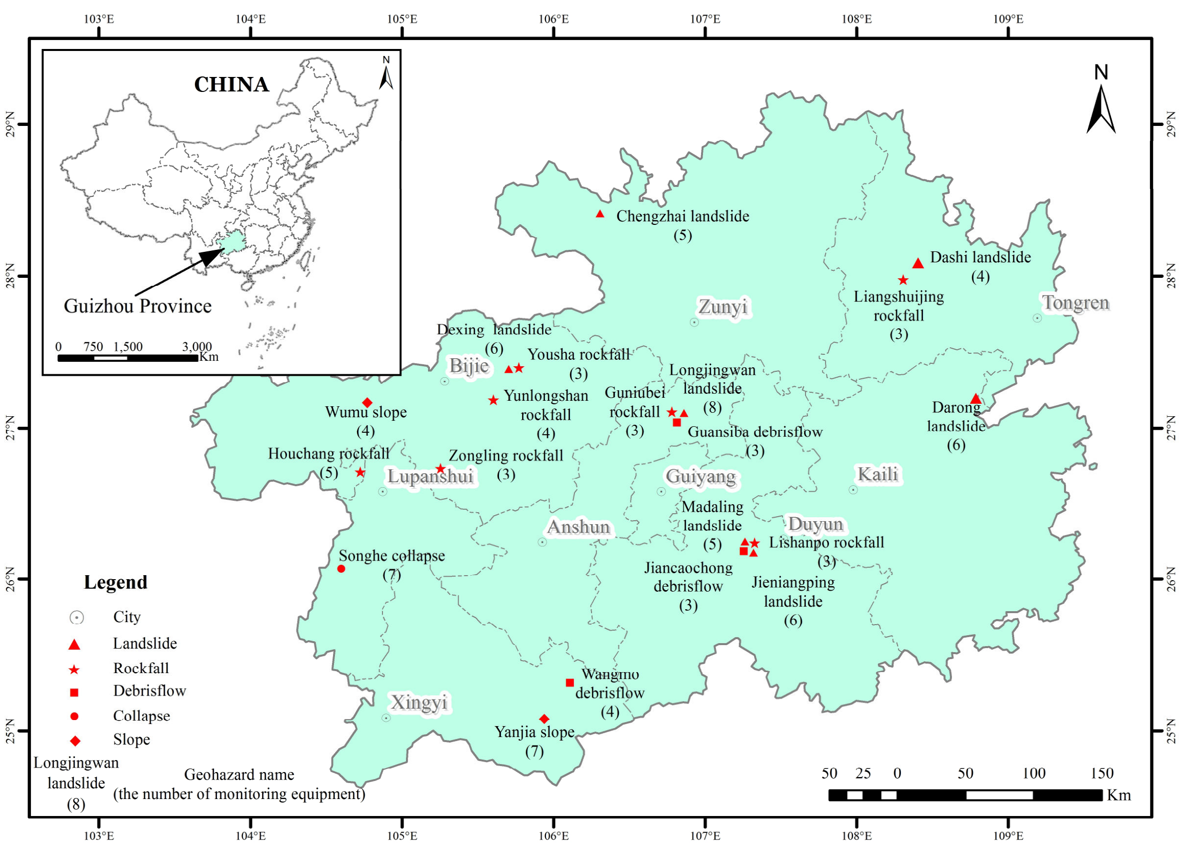


Fig. 8 A test in Guizhou Province, China. Twenty monitoring locations included 92 sets of monitoring devices and more than 14.3 million rows data have been collected so far

Table 7 Twenty monitoring locations

#	Location (City)	Geohazard Name	Monitoring Type Encoding*	Number of Equipment	Data Format
1	Anshun	Longjingwan landslide	HH	1	SQL Server
			SQ	3	SQL Server
			SY	3	SQL Server
			YL	1	SQL Server
2	Anshun	Guniubei rockfall	LF	1	SQL Server
			CL	1	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server
3	Anshun	Guansiba debrisflow	NW	1	SQL Server
			YL	2	SQL Server
4	Lupanshui	Songhe collapse	GP	2	Oracle
			LF	2	SQL Server
			YL	1	SQL Server
			CJ	2	Access
5	Zunyi	Chengzhai landslide	HH	1	SQL Server
			SQ	2	SQL Server
			SY	1	SQL Server
			YL	1	SQL Server
6	Tongren	Liangshuijing rockfall	LF	2	SQL Server
			QX	1	SQL Server
7	Tongren	Dashi landslide	HH	1	SQL Server
			SQ	2	SQL Server
			SY	1	SQL Server
8	Xingyi	Wangmo debrisflow	NW	1	SQL Server
			YL	2	SQL Server
9	Xingyi	Yanjia slope	LF	5	SQL Server
			YL	2	SQL Server
10	Bijie	Dexing landslide	GP	3	Oracle
			HH	1	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server
11	Bijie	Yousha rockfall	LF	1	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server
12	Bijie	Yunlongshan rockfall	LF	2	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server
13	Bijie	Zongling rockfall	LF	2	SQL Server
			YL	1	SQL Server
14	Bijie	Houchang rockfall	LF	3	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server

Continued:

#	Location (City)	Geohazard Name	Monitoring type encoding*	Number of equipment	Data format
15	Bijie	Wumu slope	LF	2	MySQL
			QX	1	SQL Server
			YL	1	SQL Server
16	Kaili	Darong landslide	HH	1	SQL Server
			SQ	3	SQL Server
			SY	1	SQL Server
			YL	1	SQL Server
17	Duyun	Madaling landslide	LF	3	SQL Server
			QX	1	SQL Server
			YL	1	SQL Server
18	Duyun	Jieniangping	GP	4	Oracle
			LF	1	SQL Server
			YL	1	SQL Server
19	Duyun	Lishanpo rockfall	LF	2	SQL Server
			QX	1	SQL Server
20	Duyun	Jiancaochong	NW	1	SQL Server
			SY	1	SQL Server
			YL	1	SQL Server
Total number of monitoring equipment				92	/

equipment, all of which consists of 10 types of monitoring instruments, e.g. Rain Gauge, GPS, Osmometer, Mud Meter, etc. This monitoring data has been stored in 5 databases: 1 MySQL database, 1 Access database, 2 Microsoft SQL Service databases and 1 Oracle database.

After more than three years of monitoring, a large number of monitoring data were accumulated. But these data were scattered in each individual database, and also were not accessible through remote retrieval, which caused great inconvenience to scientific research work. In order to enable data sharing, the System we designed was used in this project to solve the problems.

Through applying this system, the problem of heterogeneous data and sharing was solved with all types of monitoring data successfully integrated into the data center platform database. Up to present, GMDIS has totally processed monitoring data of more than 14.3 million rows collected by field equipment in Guizhou province. The monitoring data processed by GMDIS can be drawn as a curve and a bar chart, as shown in Fig. 9.

Such as the rainfall data, in the same chart, the cumulative rainfall can be drawn as a curve and the rainfall intensity can be drawn as a bar chart (as shown in Fig. 9a). Other different types of data also can be drawn in the same chart. This is very useful for analyzing the correlation of different types of data. For example, the rainfall data and Osmometer can be drawn in the same chart by using unified X-axis (time) and different Y-axis, as shown in Fig. 9b, the Osmometer value increases after the rainfall event, with an obvious time lag effect (usually 1-4 hours).

Moreover, based on the monitoring data, we developed a Data Analyzing System. This system is used to obtain additional information based on the analysis of monitoring data (data mining, Table 4), which provides useful scientific messages to support the operation of the monitoring instruments and geohazard early warning.

7 Conclusions

The purpose of this paper is to present a data processing and integration method of large multiple data sources in geohazard monitoring. According to the features of geohazard monitoring data, we defined the standard for data integration. Based on this standard, heterogeneous data integration is achieved and support for subsequent data analysis and geohazard early warning. Unified data structure has important significance in geohazard monitoring and early warning.

To satisfy the safe, stable and fast response retrieval requirements for massive monitoring data, we used a disk array storage system (RAID 5) as of hardware to secure data safety, while in the aspect of software, we optimized data table in Oracle Database. Finally, we developed a Geohazard Monitoring Data Integration System. To fully excavate the values of the Big Data, a Geohazard Monitoring Data Analyzing System based on the monitoring data and data mining technology was developed. After a long period of operation, these two systems show high stability and security, and can be applied in other monitoring project. The herein proposed solution can achieve the desired purpose.

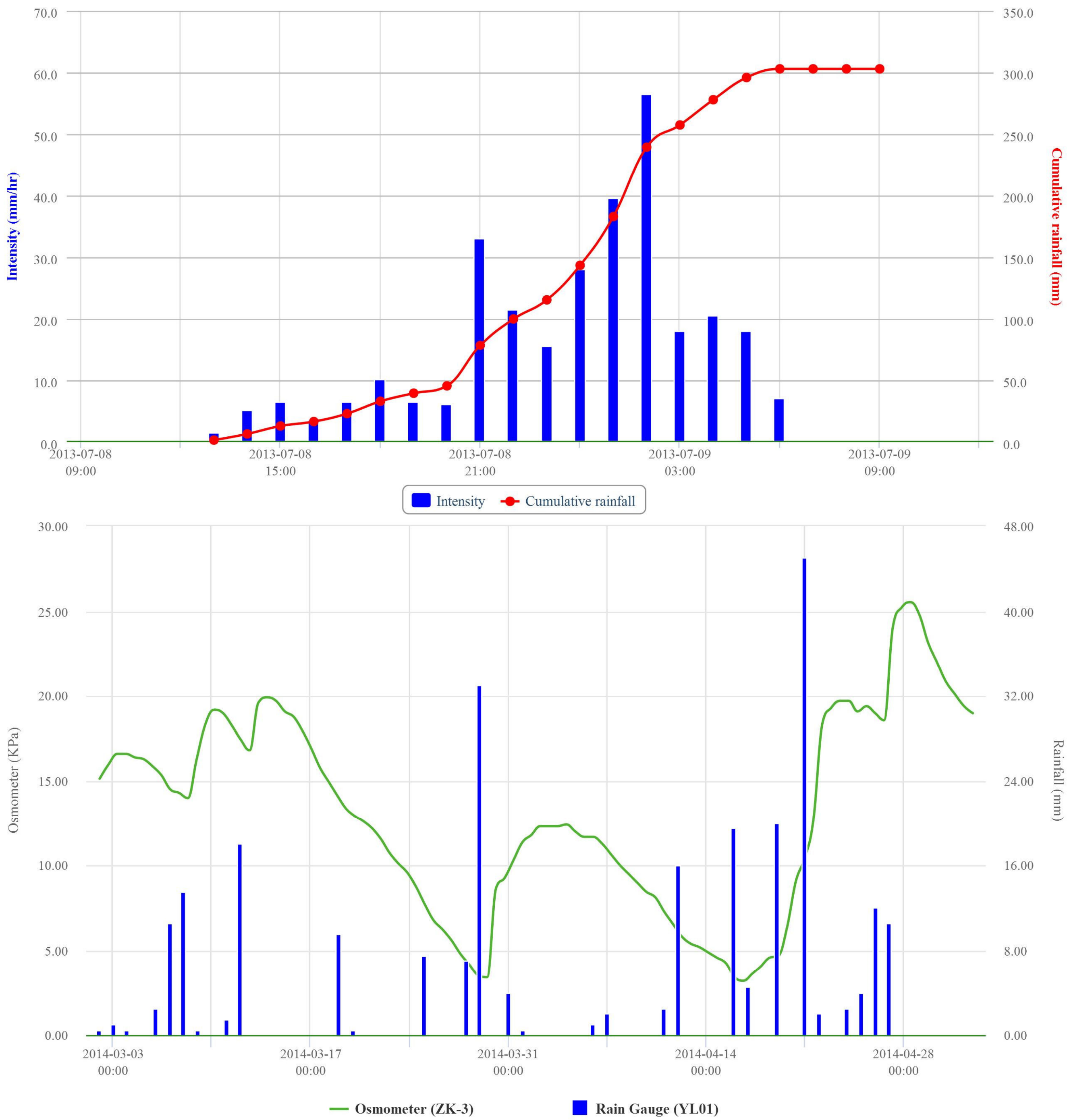


Fig. 9 Charts of monitoring data processed by GMDIS: (a) is the chart of rainfall, red curve is the rainfall intensity and the blue bar chart is the cumulative rainfall; (b) is conjoint analysis of rainfall and Osmometer

Acknowledgments

This study is financially supported by the Funds for Creative Research Groups of China (Grant No. 41521002) and Opening Fund of State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (Chengdu University of Technology) (No. SKLGP2015Z006). We are also grateful to the anonymous reviewers for their useful comments and careful review of the manuscript.

References

Chandy, J.A., 2008. RAID0.5: Design and implementation of a low cost disk array data protection method. The Journal of Supercomputing, **46**(**2**): 108 - 123. DOI: 10.1007/s11227-007-0159-8

Chen, W. and Y.W Liu, 2010. The research and implementation of test software for heterogeneous database migration process. Journal of Anhui

Polytechnic University, **25**(**4**): 35 - 39. DOI: 10.3969/j.issn.2095-0977.2010.04.011

Eckerson, W., 1995. Three tier client/server architectures: achieving scalability, performance, and efficiency in client/server applications. Open Information Systems Journal, **3**(**20**): 46 - 50.

He, C.Y., N.P. Ju and J. Huang, 2014. Automatic integration and analysis of multi-source monitoring data for geo-hazard warning. Journal of Engineering Geology, **22**(**3**): 405 - 411. DOI: 10.13544/j.cnki.jeg.2014.03.008

Huang, R.Q. and W.L. Li, 2009. Analysis of the geo-hazards triggered by the 12 May 2008 Wenchuan Earthquake. China. Bulletin of Engineering Geology and the Environment, **68**: 363 - 371. DOI: 10.1007/s10064-009-0207-0

Huang, R.Q. and X.M. Fan, 2013. The landslide story. Nature Geoscience, **6**: 325 - 326. DOI: 10.1038/ngeo1806

Ju, N.P., W.L. Hou, J.J. Zhao and L.Z. Zhao, 2010. Geohazards of Jushui River in the Wenchuan earthquak area. Journal of Mountain Science, **28**(**6**): 732 - 740. DOI: 10.3969/j.issn.1008-2786.2010.06.013

Liu, P., S. Li, CML Francis, Y. Shi and F. Huang, 2005. RAID-M: A high performance RAID matrix mass storage. Science in China Series F: Information Sciences, **48**(**4**): 409 - 420. DOI: 10.1360/04yf0060

Liu, Y.H., W.Q. Chen and G.H. Ye, 2009. Integration of geologic hazard multi-sources data. Journal of Guangxi University (Natural Science Edition), **34**(**2**): 246 - 250. DOI: 10.3969/j.issn.1001-7445.2009.02.028

Oracle Corporation, 2002. Partitioned Tables and Indexes. http://docs.oracle.com/cd/B10500_01/server.920/a96524/c12parti.htm

Oracle Corporation, 2006. Managing Partitioned Tables and Indexes. http://docs.oracle.com/cd/B19306_01/server.102/b14231/partiti.htm

Papastefanatos, G., P. Vassiliadis, A. Simitsis and Y. Vassiliou, 2012. Metrics for the prediction of evolution impact in ETL ecosystems: a case study. Journal on Data Semantics, **1**(**2**): 75 - 97. DOI: 10.1007/s13740-012-0006-9

Parker, R.N., A.L. Densmore, N.J. Rosser, M. de Michele, Y. Li, R.Q. Huang, S. Whadcoat and D.N. Petley. 2011. Mass wasting triggered by the 2008 Wenchuan earthquake is greater than orogenic growth. Nature Geoscience, **4**: 449 - 452. DOI: 10.1038/ngeo1154

Thomasian, A. and J. Xu, 2011. RAID level selection for heterogeneous disk arrays. Cluster Computing, **14**(**2**): 115 - 127. DOI: 10.1007/s10586-010-0129-4

Wei, X.L., N.S. Chen, Q.G. Cheng, N. He, M.F. Deng and J.I. Tanoli. 2014. Long-term activity of earthquake-induced landslides: a case study from Qionghai Lake basin, Southwest of China. Journal of Mountain Science, **11**(**3**): 607 - 624. DOI: 10.1007/s11629-013-2970-4

Xiao, J. and H. Li, 2012. Geology environment evolution process of Wenchuan Earthquake epicenter and disaster prevention and control measures for postreconstruction. Journal of Engineering Geology, **20**(**4**): 532 - 539. DOI: 10.3969/j.issn.1004-9665.2012.04.008

Zhang, Q.Y., X.P. Chen, D.W. Liu, J.Z. Hu, J. Li and D.W. Cai. 2009. Development of monitoring information management and monitoring data analysis network system for geotechnical engineering and its application. Rock and Soil Mechanics, **30**(**2**): 362 - 366. DOI: 10.3969/j.issn.1000-7598.2009.02.013