Research on Architecture Construction of Information System of Soil Environment under the Perspective of Big Data

Yan Gao¹, Hongbo Zhao¹, Ruoshi Hao², Zhiyong Cao¹, *, Mengran Liu³
¹Yunnan Agricultural University, Kunming, 650201, China
²Yunnan Plateau Characteristic Agricultural Industry Research Institute, Kunming, 650201, China
³Yunnan Agricultural Environmental Protection and Monitoring Center, Kunming, 650201, China

Yan Gao and Hongbo Zhao are co-first authors.
*Corresponding Author: Zhiyong Cao

Abstract: It is the demand of national big data development strategy in information construction of soil to utilize history data and monitoring data validly to deal with soil pollution prevention and promote the level of information management of soil environment. The necessity and feasibility of the construction of soil environment information system are analysed firstly. Then, relative technology in the framework construction process is introduced, especially the realization path of multi-variate data fusion, storage and services. The construction of the framework could provide data monitor and control for realizing comprehensive application of soil environment information.

1. Introduction
The Chinese government put forward the plan of soil pollution prevention action in 2016 and deeper investigated on soil environment quality. The periodic investigation regime of soil environment quality situation every ten years was carried out and soil environment quality monitoring network was established actively. At the end of 2020, we have to realize the quality monitoring of soil environment of all cities and counties to promote the information management value of soil environment. The proposition of prevention action plan provided scenery demand for the big data application of soil environment quality monitoring. Considering practical reality, this paper introduces the big data technology in the basic framework construction of soil environment quality information system.

2. Necessity of applying big data technology to promote the information level of soil environment
Since 1980s, with the development of information construction, soil, as one of the main objects of environment research, has been investigated on and mass data foundation is formed, including structured numerical data and half-structured and unstructured pictures, videos and view data. On the whole, the data of soil environment has the following characteristics: (1) Low volatility and public sensitiveness. It is hard to find the potential influence on environment and people in relatively short time. (2) Long monitoring time and demand for data accumulation. (3) Multi-dimension of monitoring data. Comprehensive influence of other environment factors should be considered under the condition of taking soil as the main environment factor.
This requires us to consider the fusion of original history data, and to set a set of administration and management of extensible and multi-variate data management and storage framework. The information construction of soil environment based on relative techniques of big data is the inevitable requirement in the future.

3. Fusion and storage of multi-variate data
Data fusion and storage is based on the characteristics of soil environment data. Under the condition of setting soil as the main environment factor, comprehensive influence of other environment factors is considered as the start of decision. Multiple influence factors and data with different sources should be extracted, pumped, processed and stored as a data set which can be analysed, which put forward challenges for the data management and service platform with the core of database management system.

There are problems of four aspects in data fusion and storage. Firstly, the source of multi-variate data. Secondly, technology fusion of multi-variate data. Thirdly the distributed-memory and lastly the establishment of data warehouse.

3.1 The source of multi-variate data
Soil environment is a developing, complex and open system. Data of various dimensions should be accumulated to monitor the quality of soil environment validly. There are mainly five types of data source.

| Data Source                        | Data Type          | Illustration of Data Source                                                                                                                                                                                                 |
|-----------------------------------|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Maps and drawings                 | Unstructured data  | Vector data and remote sensing image data of plough, administration, agriculture, land utilization, soil type, spatial distribution of soil environment quality, landform, address, drainage system, settlement place and facilities and transportation with different measuring scale. |
| Pollution source data             | Half-Structured data | Soil pollution data, including enterprise type of soil pollution key industry, waste water discharge, pollution enterprises, influence on soil environment and atmospheric bulk deposition situation of pollution enterprises and industries. |
| Monitoring data of soil environment | Structured data     | Including the background value of national soil environment, chemistry investigation of multiple goals regions, sifting of national soil pollution situation.                                                                 |
| Quality data of agricultural products | Structured data     | General investigation on heavy metal pollution of soil and soil environment monitoring in agricultural producing areas.                                                                                                       |
| Relative social-economic data     | Half-Structured data | Population situation, labor force, industrial distribution, level of income, planting regime and cultivation habits in relative regions.                                                                                                                                       |

Among the sources of multi-variate data, structured data has specific definitions and data types and mature products are managed. Since the increment is small, mature products are the main objects to be monitored. Since the collection process is slow, they are used frequently and everywhere when entering the system. With the contents of all sorts of images, texts, voice and videos, unstructured and half-structured data has large amount. It is required to manage through unstructured data management system based on traditional database system extensions, so the collection process is relatively quick.
3.2 The technology implementation of multi-variate data fusion

Multi-variate data is extracted from different sources according to the data fusion technology. The pump, processing and storage are carried out, as shown in the figure below:

![Fusion process of multi-variate data](image.png)

Sqoop open source tool is used during data extraction to complete data delivery in Hadoop and traditional databases (mysql, postgresql, etc.). It can also be used to transmit the data from relational databases (Such as MySQL, Oracle and Postgresql) to Hadoop’s HDFS or in turn. When dealing with text unstructured data processing, databases Apache POI and Apache Commons CSV are used to read and write Microsoft Office format file and plain text file of CSV. Fusion task management requires to use Oozie framework, which is an open-source framework based on workflow engine. Control nodes and action nodes are realized through Workflow, Coordinator and Bundle Job.

3.3 Distributed storage technology of multi-variate data

(1) Unstructured storage

Unstructured data includes the plough soil images, videos, PDF documents, Word documents of achieving results and relative GIS data by collecting at the scene or using unmanned aerial vehicles. The storage of unstructured data uses HDFS distributed file system in Hadoop framework.

The size of the file data of monitoring images, videos and achievement report document of plough soil differs from hundreds of KB to tens of MB. The storage is completed through the modes of file archiving, sequencefile and HBase.

File archiving is settling and preserving to make it easy to manage and preserve. The archiving in Hadoop is to establish a new abstract layer on HDFS, namely HAR (Hadoop Archives) and the format turns into har:// URL. The realization principle is shown as the figure below.
Archiving files is to complete the read of the final file by searching the index file of the second layer. So, it is relatively slower in efficiency comparing to normal HDFS. Hadoop archiving can be produced by archiving command of Hadoop archive. The production of the archiving file will not delete the original file, and the user can determine whether to preserve the original file or not.

SequenceFile is writing into the file with the format of sequence file by changing the writing pattern of the file. The unique storage format of SequenceFile determines that it can meet the demand of small file storage. The pattern of SequenceFile’s storage of data is the joint by using key-value.

**Figure 2. The way of archiving**

**Figure 3. The way of sequencefile**

HBase is used to complete data writing, rather than write directly into HDFS. The storage format is similar to the K-V format of SequenceFile.

(2) Structured storage

For other structured data such as detailed data, achievement basic archiving, farmland data, agricultural resource of agricultural plough soil, HBase distributed databases can store data through data. HBase is suitable for the storage of structured and unstructured data.

3.4 The implementation and design principles of data warehouse

On the basis of soil detailed data, sampling monitoring results and other relative data in distributed databases, the extraction, transformation, calculation and transmitting calculation results to data warehouse are completed by designing scientific and rational data modes.

(1) The design of data warehouse

In the designing process, first we should determine relative subjects and consider the technical indexes, such as the content of inorganic substances of agricultural plough soils (Heavy metal, e.g.
cadmium, mercury, plumbum, arsenic, chromium, copper, zinc and nickel). Since the extent of polymerization of measurement is different, minimum granularity principle is adopted, namely setting the granularity to minimum. If the data is subdivided into single days, the granularity is set as day. After determining the key value, gradation and hierarchy, redundancy should be reduced, such as the region dimension, time dimension and plough type dimension of agricultural plough soil. The fact table is used to record all information, including the specific elements of each event and specific issues.

(2) Distributed calculation engine

Aiming at soil data, GIS data and agricultural product data, distributed calculation engine is utilized to complete the analysis of agricultural plough soil, the rank analysis of agricultural products, network analysis of GIS.

Batching calculation is aiming at the analysis and calculation of single batch and storing the results in the data warehouse. Batching calculation mainly uses Spark, which offers a big data processing demand of data set and data source (batch data or real-time streaming data) of different properties (texts, charts and figures).

Aiming at online real-time monitoring data, such as the online monitoring data of surface runoff of the comprehensive control of non-point source pollution, Kafka is adopted to transmit to Spark Streaming to complete real-time processing, analysis and calculation.

Spark streaming is an extension of Spark’s core API, which has characteristics of extendibility, high throughout and fault tolerance. The data can not only be accessed from sources of kafka, flume, Twitter, ZeroMQ and Kinesis, it can also be calculated through complex algorithms of higher-order functions of map, reduce, join and window. At last, the data processed can be pushed to the file systems, databases and real-time instrument panel, as shown in the figure below.

![Diagram of Spark Streaming and Data Processes](image)

Figure 4. Process of stream computing

(3) Data service

After data fusion and calculation, agricultural plough data and relative statistics should provide upper applications with relative data service by using programming, command line and SQL execution mode.

The platform sets up unified regulated data category by the statistical sum and arrangement. Service is also offered, and RESTful and RPC ports are utilized to provide data exchange and sharing for upper application systems and external system. The whole data fusion and storage architecture is shown as below.
In this paper, according to the requirement of improving the information management level of the soil environment in preventive action plan, by building big data basic platform, according to the united standard, it conducts the structural and non-structural distributed storage, implements the standard and normal data storage. By establishing the united data warehouse, implementing the united management of data, it breaks out data barriers and information isolated islands, provides the services of interview, data index for resource data, reaches the exchange and share of various information, develops the values of big data to maximal extent, and offers the guarantee for the next step of realization of comprehensive application analysis and high-quality knowledge acquisition.

References

[1]Mao W, Wan Y, Sun Y, et al. Applying dredged sediment improves soil salinity environment and winter wheat production[J]. Communications in Soil Science & Plant Analysis, 2018, 49(2):1-8.
[2]Garg V K. Interaction of tree crops with a sodic soil environment: potential for rehabilitation of degraded environments[J]. Land Degradation & Development, 2015, 9(1):81-93.
[3]Kader M A, Senge M, Mojid M A, et al. Recent advances in mulching materials and methods for modifying soil environment[J]. Soil & Tillage Research, 2017, 168(5):155-166.
[4]Yao P, Li X, Liu J, et al. The role of maize plants in regulating soil profile dynamics and surface emissions of nitrous oxide in a semiarid environment[J]. Biology & Fertility of Soils, 2018, 54(1):1-17.
[5]Busari M A, Kukal S S, Kaur A, et al. Conservation tillage impacts on soil, crop and the environment[J]. International Soil & Water Conservation Research, 2015, 3(2):119-129.
[6]Wang S, Wang H. Adsorption behavior of antibiotic in soil environment:a critical review[J]. Frontiers of Environmental Science & Engineering, 2015, 9(4):565-574.
[7] Chem A. Soil Dehydrogenases as an Indicator of Contamination of the Environment with Petroleum Products[J]. Water Air & Soil Pollution, 2015, 226(11):372.

[8] Zollinger B, Alewell C, Kneisel C, et al. Soil formation and weathering in a permafrost environment of the Swiss Alps: a multi-parameter and non-steady-state approach[J]. Earth Surface Processes & Landforms, 2017, 42(5):814-835.

[9] Ginandjar P, Saraswati L D, Martini. Soil-Transmitted Helminth Infection in Elementary School Children: An Integrated Environment and Behavior Case Study in Bandungan Sub-District, Semarang District[J]. Advanced Science Letters, 2017, 23(4):3565-3568.

[10] Ji Z, Fu B, Gao G, et al. Effects of precipitation and restoration vegetation on soil erosion in a semi-arid environment in the Loess Plateau, China[J]. Catena, 2016, 137(137):1-11.

[11] Yang J, Yan N, Liu Q, et al. Laboratory test on long-term deterioration of cement soil in seawater environment[J]. Transactions of Tianjin University, 2016, 22(2):132-138.

[12] Hu W, He Y, Chen D, et al. Diversity of ammonia-oxidizing bacteria in relation to soil environment in Ebinur Lake Wetland[J]. Biotechnology & Biotechnological Equipment, 2016, 30(2):1-9.