Big Data Storage Method in Wireless Communication Environment

Liang Ye
Department of Computer Science
Beijing Foreign Studies University
Beijing, China
E-mail: liang_ye@sohu.com

***Abstract*—Big data phenomenon refers to the practice of collection and processing of very large data sets and associated systems and algorithms used to analyze these massive data sets. Big data service is very attractive in the field of wireless communication environment, especially when we face the spatial applications, which are typical applications of big data. Because of the complexity to ingest, store and analyze geographical information data, this paper reflects on a few of the technical problems presented by the exploration of big data, and puts forward an effective storage method in wireless communication environment, which is based on the measurement of moving regularity through proposing three key techniques: partition technique, index technique and prefetch technique. Experimental results show that the performance of big data storage method using these new techniques is better than the other storage methods on managing a great capacity of big data in wireless communication environment.**

Keywords- big data; storage service; partition and indexing; cache prefetch; wireless communication environment

I. INTRODUCTION

One buzzword that has been popular in the last couple of years is Big Data. [1] Big data phenomenon refers to the practice of collection and processing of very large data sets and associated systems and algorithms used to analyze these massive datasets. Architectures for big data usually range across multiple machines and clusters, and they commonly consist of multiple special purpose sub-systems. In simplest terms, Big Data symbolizes the aspiration to build platforms and tools to collect, store and analyze data that can be voluminous, diverse, and possibly fast changing. As a specific kind of application for big data, spatial application such as Mobile-based Geographical Information System (Mobile-GIS) has become more and more important in both scientific research and industry. With the development of earth observation technologies, the spatial data are growing exponentially year by year, and their categories are becoming more diverse including multi-dimensional geographic data, multi-spectrum remote sensing imageries, and so on. Besides that, as spatial applications become more popular, concurrent accesses to spatial applications are becoming highly intensive. [2]

The spatial data objects are generally nested and more complex than basic data types. [3] They are stored as multi-dimensional geometry objects, including points, lines and polygons. [4] Moreover, the spatial query predicates are complex. Typical spatial queries are based not only on the value of alphanumeric attributes but also on the spatial location, extent and measurements of spatial objects in a reference system. Therefore, spatial query processing over big spatial data requires intensive disk I/O accesses and spatial computation. [5]

Driven by the above problems, this paper will try to reflect on a few of the technical problems presented by the exploration of big data. The principles described in this paper are derived from the experiences and outcomes of various real world projects. This paper puts forward and implements a novel scheme to provide efficient storage and redrawing method over big spatial data in the wireless communication environment. Firstly and most importantly, we propose a three-tier structure composed of mobile units. The big spatial data are partitioned into blocks according to the geographic space and block size threshold, and these blocks are uniformly distributed on cluster nodes which are stored in the Mobile-GIS server. In practical spatial applications, most users only focus on a relatively small area, so the partition technique and index technique, which will be introduced in part III, could preserve the geographic proximity of spatial objects on mobile-GIS server, which are fit for transmission in the wireless communication environment and indexing the blocks in serialization order. The prefetch technique, which will be introduced at part IV, is used on fixed host to turn big data storage service into reality.

II. KEY TECHNIQUES IN BIG DATA STORAGE SERVICE

In big data storage service system, the inquiry result will be shown with map instead of with words to meet users' requirement. [6] But the problems coming with the new method must be solved. The big spatial data is difficult to be stored and analyzed on the mobile unit, so the inquiry result must be processed on the Mobile-GIS servers and then downloaded to mobile units. Sometimes the inquiry result map is enough to user to obtain the information they need. However, after people get the inquiry result map, they may adjust it to show something about it in next step, so a new result map will be made and be downloaded to mobile unit. The previous result map has to be deleted from cache because there are no connection between the old result map and the new result map, and it is a huge waste. [7, 8]

Published by Atlantis Press, Paris, France.

In order to use the result maps which are cached in the storages of mobile units effectively, improve processing efficiency of mobile units and reduce the cost of mobile communication, it is very important to design a good means to organize and analyze spatial data. Thereby, two kinds of key techniques for big data storage service are put forward, one is partition technique and index technique, and the other is prefetch technique.

III. PARTITION TECHNIQUE AND INDEX TECHNIQUE

Layer is a container for geographical spatial object in geographical spatial data model, and it is the primary unit for the structure of geographical spatial data. A layer consists of geographical spatial objects with the same kind of attribute in a certain spatial. So there are a lot of geographical spatial objects in a layer, and the quantity of data is too huge to save in the memory of mobile unit at one time. So the partition technique and index technique based on layer are put forward.

A. Partition technique based on layer

The partition technique based on layer is that the geographical spatial layer is cut into a series of the same size blocks by grid, and a geographical spatial object, which belongs to a block wholly, is still in the block, otherwise, it is cut into parts by grid, and each part belongs to the block which includes them wholly and becomes a whole geographical spatial object.

The advantage of using the partition technique based on layer is when a user wants to find out some geographical spatial object in a layer, they need not to download the whole information of the layer to the mobile unit but only to download the block which includes the object or to download several blocks if there are a lot of blocks used to save the object.

The control rectangle method is widely used method in traditional GIS. When a geographical spatial object is searched on mobile unit, the feedback will be a rectangular spatial which just include the object. Compared with the traditional control rectangle method, the partition technique is more efficient.

For example, in Figure 1, if we want to find out geographical spatial object A with our mobile unit, we could only download the block b2. If we want to search for geographical spatial object BCD, we need download blocks c2, d2, b3, c3, d3, e3, b4, c4, d4, and the blocks b2, e2, e4 which are also downloaded in control rectangle method need not be downloaded. So the efficiency is improved 25%. Otherwise, the area decided by the control rectangle method is far larger than the area of the blocks including the parts of the searched object, so the total efficiency is higher than 25%.

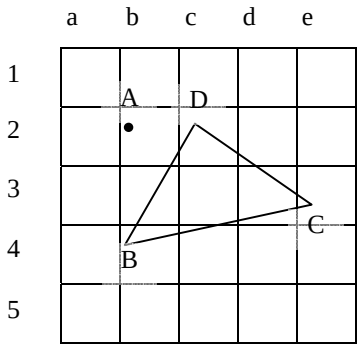

Figure 1. The objects on a layer

B. Index technique based on layer

After the geographical spatial layer is partitionized, we must prevent them from being out of order. So the index technique is needed to index the partitionized geographical spatial blocks. There are indexed files at the mobile GIS server, and each indexed file points to a geographical spatial layer. If a user needs an object on some layer, the indexed file of the layer will be downloaded to the mobile unit of the user with the result map.

The table below explains the format of binary indexed file.

TABLE I. THE FORMAT OF BINARY INDEXED FILE

The abscissa of	
the beginning block (4 bytes)	
The ordinate of	
the beginning block (4 bytes)	
The width of each block (2 bytes)	
The height of each block	
The pointer of the first block	
The version of the first block	The pointer of
the second block	The version of the second block
The pointer of the third block	
The version of the third block	…… …… ……

- The beginning block: It is used to record the first block of a layer. The geographical spatial layer is usually a rectangle, but the geographical spatial objects are pockety in the layer. So there are no objects in many blocks. But the geographical spatial information is changing all the time, so we cannot drop the blocks without object from the layer so as to add object to the block in the future. Thereby, we choose the block in the first line of the first row as the beginning block.

- The width of each block: A geographical spatial layer is cut into rows by a group of longitude lines with the same space, so the width of each block could be indicated by any two neighboring longitude.
- The height of each block: A geographical spatial layer is cut into lines by a group of latitude lines with the same space, so the height of each block could be indicated by any two neighboring latitude.
- The pointer of block: If a geographical spatial layer is cut into m lines and n rows, the number of block in this layer is m•n. The pointer and version of the m•n blocks are saved in the indexed file in serialization order. The pointer of block points to where the content of the block could be get in the mobile GIS server. According to the abscissa and ordinate of beginning block and the width and height of each block, any object in this layer could be calculated to get which block or blocks it lies in, and the pointers of the block or blocks also could be gotten.
- When the user searches for another geographical spatial object or moves the inquiry result map about, the mobile unit could calculate which blocks need to be downloaded from the mobile GIS server via the pointer of block in binary indexed file.
- The version of block: Because the city is developing quickly and the earth changes all the time, the geographical information data in the mobile GIS server must be updated with them. In the mobile GIS server, the version of block records the updating number of a block. When the block is updated again, the number adds 1. In order to reduce the communication cost, the mobile unit saves the blocks used lately or frequently in the cache. When user uses a block in the memory again, the version of the block will be sent to the mobile GIS server to confirm whether the corresponding block in the mobile GIS server is updated. If the version in the mobile GIS server is newer than the version in the mobile unit, the block is downloaded again and the version of the block is updated in the mobile unit. Otherwise, the mobile unit uses the block in the cache. In the indexed file, there is a separate version for each block, because this could add the effectiveness of cache, and void all blocks in the cache to be useless since only a block in the mobile GIS server is updated.

C. *Principle of partition and index*

If a layer is cut into more blocks, we need download blocks of query result less. That is, if the size of block is smaller, the edges of block are near to the edges of the geographical spatial object inside of the block, and the useless data downloaded with geographical spatial object is less. As if we should have cut a layer in this way to be more efficient. But in fact, we could not do in this way because of two reasons:

First, as the block is smaller, a geographical spatial object might be cut into more parts and saved into more blocks. After that, we need using the edges and vertexes of block to rebuild a new object, which is a part of the original object, in each block. Thus, the space used to save an original object is larger. At the same time, the attributes of the original object need to be saved to each block, which causes a lot of redundancy. So, the block is smaller, the efficiency might be not higher.

Second, as the block is smaller, the size of indexed file is larger. Limited by the memory of mobile unit, the block could not be in large quantities.

Through the experience we achieve in our work, if the height of screen of mobile unit is H; the width is W; the length of city from north to south is h; the length of city from east to west is w; the height of each block is h'; the width of each block is w'; the maximal scale for shown is MAX, then:

$$\eta 2 = \frac{H}{MAX} \bullet n_1$$

$$n_1 \in 2 \times N - 1 \ \alpha\nu\delta \ 1 \le n_1 \le 9 \qquad (1)$$

$$\omega 2 = \frac{W}{MAX} \bullet n_2$$

$$n_2 \in 2 \times N - 1 \ \alpha\nu\delta \ 1 \le n_2 \le 9 \qquad (1)$$

When $1 \le \left\lceil \frac{h}{h'} \right\rceil \le 200$ and $1 \le \left\lceil \frac{w}{w'} \right\rceil \le 200$ are satisfied, the effectiveness is highest.

To a certain layer, the geographical information data, which need be shown on the screen, is just in a block, because n1 and n2 are odd numbers and they are bigger than 1 or equal to 1. At the same time, we can control the quantity of blocks under 40000, and the size of indexed file could be controlled fewer than 120K, which could make sure to execute the inquiry requirement on the mobile unit.

IV. REALIZATION OF BIG DATA STORAGE SERVICE

In order to realize the big data storage service, the mobile unit, the fixed host and the mobile GIS server must work together, and the partition technique, the index technique and the prefetch which would be introduced below must be used in the system in which big data storage service is provided.

A. *Measurement of moving regularity*

The user who wants to use the big data storage service must setup the software, which supports the big data storage service, on its mobile unit. After he sends out the inquiry requirement to the fixed host, the fixed host begins to

measure the parameters of user, such as position, movement state, direction, speed, track, etc., and saves them into the database on fixed host.

B. Download the blocks

According to the obtained parameters from the user, the location where the user is could be gotten, and the fixed host could require the result map from the mobile GIS server. The mobile GIS server uses the index technique to get the blocks partitionized by the partition technique, and sends them to the fixed host. When the blocks are ready, the mobile unit could download the blocks from the fixed host, and the result map is drawn on the screen of mobile unit.

C. Prefetch technique

As time went by, the fixed host measures the movement regularity of mobile unit again and forecast the movement tendency of it, and the geographical information blocks, which might be used by the user at a certain time in the future, will be prefetched from the mobile GIS server and be sent to the mobile unit when the network connection is reliable. After the mobile unit receives the blocks, it will put them into cache. When the blocks need to be used or there is no signal with the mobile unit, the blocks will be taken from cache directly to draw result map on the screen.

V. CONCLUSION

The big data storage service has two advantages over the other storage services. On the one hand, the prefetching technique could do its utmost to guarantee that the users' requirements could be dealt with fully and duly; on the other hand, for an querying result with a large quantity of data such as a location map, the delay of network transmission could be reduced to a great extent, and the disadvantage of narrow bandwidth of wireless communication network can be removed. In a word, the technique proposed by this paper is based on mobile database technique, and it calculates and predicts the moving tendency by checking the rule of movements, so as to provide big data storage service. A simulation experiment has been done on Android platform, and the result shows that the performance of big data storage method using these new techniques is better than the other storage methods on managing a great capacity of big data in wireless communication environment. The expected goal has been achieved.

ACKNOWLEDGMENT

This paper is supported by the Fundamental Research Funds for the Central Universities (No.2012XJ031). Without this help, this work would never have been completed.

REFERENCES

[1] G. Jung, N. Gnanasambandam, and T. Mukherjee, "Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds" 2012 IEEE 5th International Conference on Cloud Computing (CLOUD 2012), 2012, pp. 811-818, doi: 10.1109/CLOUD.2012.108

[2] Edmon Begoli, and James Horey, "Design Principles for Effective Knowledge Discovery from Big Data" 2012 Joint Working IEEE/IFIP Conference on Software Architecture (WICSA) and European Conference on Software Architecture (ECSA 2012), 2012, pp. 215 - 218, doi: 10.1109/WICSA-ECSA.212.32

[3] M. Diaz, G. Juan, O. Lucas, and A. Ryuga, "Big Data on the Internet of Things: An Example for the E-health", 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2012), 2012, pp. 898-900, doi: 10.1109/IMIS.2012.198.

[4] E. Kohlwey, A. Sussman, J. Trost, and A. Maurer, "Leveraging the Cloud for Big Data Biometrics: Meeting the Performance Requirements of the Next Generation Biometric Systems", 2011 IEEE World Congress on Services (SERVICES 2011), 2011, pp.597-601,doi: 10.1109/SERVICES.2011.95.

[5] Shao Xiao-Dong, and Li Qiang, "A strategy for continuously big capacity data transmit/receive process handling in real-time", 2010 2nd International Conference on Advanced Computer Control (ICACC 2010), 2010, pp.37-40, doi: 10.1109/ICACC.2010.5487001.

[6] Rukui Tao, Baochen Jiang, and Chengyou Wang, "Sampling rate conversion and data synchronization in big merging unit" 2011 4th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT 2011), 2011, pp.531-534, doi: 10.1109/DRPT.2011.5993949.

[7] Yunqin Zhong, Jizhong Han, Tieying Zhang, Zhenhua Li, Jinyun Fang, and Guihai Chen ,"Towards Parallel Spatial Query Processing for Big Spatial Data", 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW 2012), 2012, pp.2085-2094, doi: 10.1109/IPDPSW.2012.245.

[8] B. Chandramouli, J. Goldstein, and S. Duan, "Temporal Analytics on Big Data for Web Advertising", 2012 IEEE 28th International Conference on Data Engineering (ICDE 2012), 2012, pp.90-101, doi: 10.1109/ICDE.2012.55.