Research on Quickly Search in Massive Remote Sensing Images Based on HBase

Wu Chen[1], Quan Jicheng[1], Yuan Yuwei[2], Liu Yu[1], Wang Hongwei[1], Zhao Xiuying[1], Yang Mingquan[1]
[1]Department of Specialty, Aviation University of Air Force
Changchun, Jilin, China, 130022
e-mail: wuchen4094@163.com
[2] Department of Electronic and Information Engineering, Naval Aeronautical and Astronomical University
Yantai, Shandong, China, 264001

***Abstract*—On the basis of analyzing massive remote sensing images storage and HBase, Hilbert Curve was applied in image pyramid model. A solution about quickly search in massive remote sensing images was put forward in this paper. The experiment showed that the solution not only could solve the storage problem in cluster that single computer couldn't, but also realized the quickly search in massive remote sensing images. This solution can be used in the further process of massive remote sensing images.**

Keywords—HBase; Hilbert; massive remote sensing images; image pyramid model; quickly search

I. INTRODUCTION

Remote sensing images play an important role in the geographic mapping, resource and environmental monitoring, military reconnaissance and battlefield awareness. With the rapid development of earth observation technologies, the amount of remote sensing images data that humans received and generated are geometrically increasing. But the ability of massive remote sensing images management does not keep pace with image data's growth step. This is resulting in "the more image data, the less available images" [1].

In this case, the traditional centralized storage management has been unable to meet the requirements of massive remote sensing images management. With the rise of cloud computing, the cloud computing itself has "unlimited computing power and storage capacity" [2], which naturally have been the preferred solution of massive remote sensing image storage management. Cloud computing technology is not a new invention, but the comprehensive integration of many existed technologies. Hadoop [3] is an open source cloud computing system of the Apache Software Foundation. The two key parts of Hadoop, HDFS [4] (Hadoop Distributed File System) and MapReduce, are respectively the open source implementation of GFS [5] (Google File System), MapReduce [6], Bigtable [7]. HBase is a distributed database on the base of Hadoop. HBase can randomly read and write data resulting from column-oriented storage. HBase supports fast retrieval in massive data. In order to realize the effective management of massive remote sensing images data, Hilbert curve [8] was applied in image pyramid model in this paper. An efficient management method of massive remote sensing image was proposed in this paper based on HBase.

II. HBASE DATA MODEL

HBase is an open source distributed database system in the Apache Software Foundation. HBase is built on top of HDFS. It is applicable to the random fast read and writes data on cheap computers with high reliability, high stability, scalability and column-oriented storage. HBase is completely different from MySql, SQL Server and other traditional relational database in structure. In order to get better scalability and flexibility, HBase weaken the other advantages. So that HBase has a different data model. This also led many differences in table design with traditional relational database.

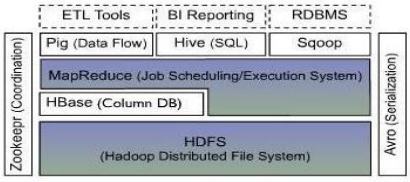

Figure 1. Hadoop Structure

HFile is the basic unit for fast retrieval in HBase that realize Bigtable. HFile is mainly responsible for the column family data storage.

A. The HBase Data Model

a) Table

HBase makes use of Table to manage data. The Table name is a string, and can be safely used in HBase.

b) Row

Table consists of many rows, and the data storages in rows. Each Row by the respective RowKey is unique. RowKey does not have the data type and is always directly believed as a byte array Byte[].

c) Column Family

Every row's data is divided into different Column Families. Column Family has a direct impact on how data is stored in HBase. Therefore, Column Family need to be defined when designing table.

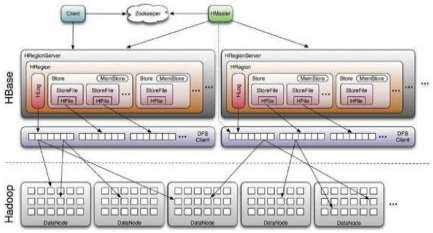

Figure 2. HBase Structure

d) Column Qualifier

The data in Column family locates through Column Qualifier. Column Qualifier does not require be predefined. Column Qualifiers between the rows do not need to be same.

e) Cell

Cell is located by RowKey、Column Family and Column Qualifier. Cell storages the data actually and does not have the data type. The data in Cell is located by timestamp from big to small which can guarantee the latest data is the most front one.

III. PYRAMID IMAGE MODEL BASED ON HILBERT CURVE

Image Pyramid is the generally acknowledged data model for massive remote sensing images management. Pyramid image model in this paper adopted the Plate Carree [9] projection. Plate Carree projection is a global range projection.

Peano curve, Hilbert curve and so on were the commonly used space filling curves, as shown in fig3. Hilbert curve comes from the classical Peano curves. Hilbert curve has the best aggregation among currently known space filling curves. 0-3 chars were adopted in Hilbert curves to express the order in a 2×2 basic unit. East and west hemispheres used "1", "0" code in the first layer. The tiles were stored in the HBase database; each tile has a corresponding Hilbert code (Hcode), such as "0323". The characteristics of Hcode organization for tile are:

- The Hcode strings have the same length with levelID;
- The adjacent tiles in Hcode are also adjacent in spatial positions. But the adjacent tiles in spatial location are generally also adjacent in Hcode.

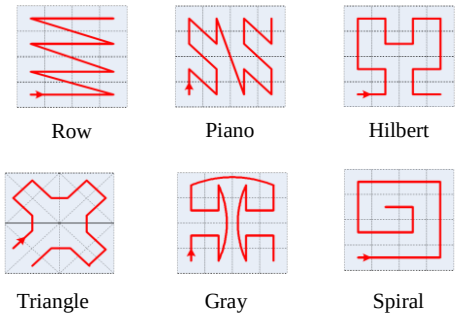

Figure 3. Space Filling Curves

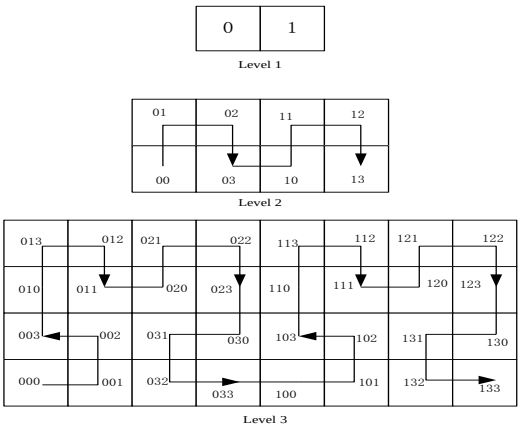

Figure 4. Image Pyramid Based on Hilbert Code

Because every search will need to convert levelID and tile location for Hcode, so we should simplify the Hilbert transformation. The time complexity of classical Hilbert encoding algorithm is $O(n^2)$, Cao Zhongsheng [10] proposed a fast encoding algorithm for Hilbert curve based on the idea of partition, the time complexity is reduced to $O(nlogn)$. This method was improved to be the fast Hilbert coding solution based on look-up table in this paper.

After observation, Hilbert curve has a strong genetic feature, and each layer is composed of four basic curves. Four basic curves are shown in Fig. 5.

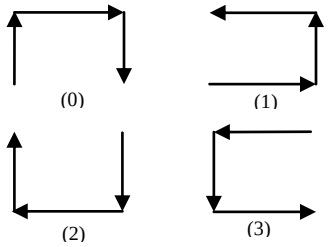

Figure 5. Four basic curves

Rule 1: We called basic curve unit in child layer as the child unit (2×2 grids in specific location); the child unit's curve type was called child type.The parent unit's curve type was called father type.

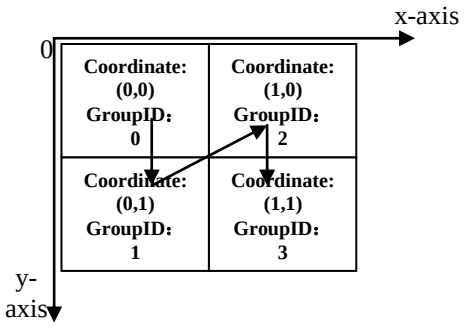

Figure 6. Coordinate and groupID

TABLE I. SERIAL TABLE

Father Type	*groupID*	*child type*	*code*
0	0	0	1
	1	1	0
	2	0	2
	3	0	1
1	0	2	3
	1	0	0
	2	1	2
	3	1	1
2	0	1	3
	1	2	2
	2	3	0
	3	2	1
3	0	3	3
	1	0	0
	2	1	2
	3	1	1

Rule 2: the coordinate diagram of basic unit was shown in Fig. 6; the origin of coordinate is the top left corner of the basic unit.

The father type and child unit's groupID decided the unique child type and code. So, we can get the Hcode of tile by looking up Table I from up to bottom. Father type and code also can only determine the child unit and child type by looking up Table II. We can get quadrant number string by cycling from top to bottom.

TABLE II. ANTI-SERIAL TABLE

Father Type	*code*	*groupID*	*child type*
0	0	0	1
	1	1	0
	2	0	0
	3	0	3
1	0	2	0
	1	0	1
	2	1	1
	3	1	2
2	0	1	3
	1	2	2
	2	3	2
	3	2	1
3	0	3	2
	1	0	3
	2	1	1
	3	1	0

IV. THE TABLE DESIGN IN HBASE

A. HilbertCodeIndex

In order to storage global image data by resolution order from low to high, we set a single tile index by Hilbert curve(HilbertCodeIndex, HcodeIndex): *levelID_Hcode*, such as *"03_021"*. HcodeIndex was the RowKey of HBase table which we could use it to search tile quickly. The rows in the HBase table were stored by the RowKey order. Therefore, this paper put level information at the beginning of the RowKey, and the level information consists of two digits, which could guarantee the tiles of same level would be stored together. The other information after level is Hcode, so that the same level's tiles stored completely according to the Hilbert sequence. The tiles that space location adjacent generally stored together which can ensure an I/O can read the required tiles in one batch search.

B. Design of Column Family in HBase Table

In order to ensure the retrieval efficiency of HBase, we designed table structure according to the HBase data storage feature. In this paper, we build an HBase table for everyone band. Image data was cut to tiles and be stored in table. The tile's size was 256×256. The RowKey of the table was *levelID_Hcode.*

In order to ensure the read and write efficiency, the number of HBase column family is generally 2 or 3. Two column families were established in the table: <ImageData:>, <ImageMeta:>. Column families ImageData and ImageMeta respectively stored tile data and metadata. In each query, the client first converted the longitude and latitude information to the row&column information, then convert row&column and level information to HcodeIndex information, then transmit HcodeIndex to the server. The server quickly searches tiles in the table because HBase uses index similar to B+ tree. At the same time, the client can cache the recent queried HRegion's address, which can accelerate the next retrieval process.

V. EXPERIMENTAL VERIFICATION

A. The Experiment Environment

There are 7 computers as the host computers which has installed the VMware virtual machine. There are one Web server, one MasterNode, four SlaveNodes and one ZooKeeper. The computers: CPU is AMD Athlon (TM) II X4 640 , Windows Server 2008 R2 Enterprise operating system , memory 8GB,virtual machine: Red Hat Enterprise Linux 5.5.

B. Image Data Retrieval Experiment

The image data is the global image of 1-17 level. And only the 1-9 level's tiles are complete. The image data capacity is about 77.6GB. The query steps: 1) computing HcodeIndex based on the spatial range; 2) read tile and attribute data corresponding HcodeIndex from the table; 3) the query data is written to the local client. The query time includes HcodeIndex computing time, searching HcodeIndex time, read and write tile and attribute data time.

Experiment one: Six space ranges query object in the ninth level were produced through a random function. The ratios of the six space ranges were 1/32, 1/16, 1/8, 1/4, 1/2. Each query time can be divided into two parts, namely the image blocks query time Ts and display time Tv on the client. The sum time is *Ts+Tv*. Experiments on the HcodeIndex storage and row&column storage were

conducted respectively. The experimental results are shown in Table III.

TABLE III. FIRST EXPERIMENT RESULT TABLE

Range Search Time Wasted	Time Wasted (S)				
	1/32	*1/16*	*1/8*	*1/4*	*1/2*
Row&Column Index	1.029602	3.276006	6.318011	10.498821	27.658851
HilbertCodeIndex	0.702001	0.998402	2.558405	5.007609	12.620421

Experiment analysis: it can be seen from the Table III that spatial query based on Hcode index was more efficient than based on row&column index. This is mainly because that the spatial aggregation Hilbert curve was far better than the row&column curve. When the query range is small (1/32), the difference was not large. This is mainly because the calculation of obtaining the Hilbert code was larger than the row&column code. So, the Hilbert code's spatial aggregation advantage was not obvious when query range is small. But with the spatial query range increased, the aggregation advantage of Hilbert curve became more obvious than row&column's computational advantage gradually.

Experiment two: In order to verify the efficiency of online browsing, we browsed a stationary location from layer first to layer thirteen. The position was set in the center of display screen when browsing layer after layer. The nodes in the figure stand for the consume time from the first level to the current level.

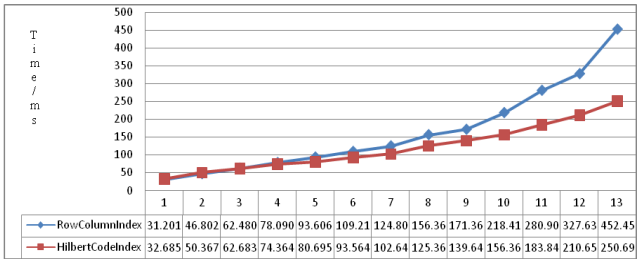

Figure 7. Experiment Figure of the second experiment

Results analysis: it can be seen from the Fig. 7 that the browse time based on Hcode index from level 1 to 4 is higher than based on row&column index. This is mainly because that the spatial aggregation advantage of level 1-4 in Hcode index of is obvious. However, with the level increased, the aggregation advantage of Hilbert curve became more obvious than row&column's computational advantage gradually. On the whole, image data retrieval task can be done more efficiently by Hilbert index than row&column index in HBase.

VI. CONCLUSION

In this paper, the characteristics of column storage model in distributed database HBase was applied with the Hilbert curve in image pyramid. A fast Hilbert coding algorithm based on look-up table was proposed a used in the designed HBase table structure and RowKey. The experiment results verified the validity and practicability of the proposed methods. The next step will be how to improve the memory efficiency of HBase and further images processing based on MapReduce.

REFERENCES

[1] Li Fei, "Research on the key Technology of Image Database Management System", Beijing:Graduate University of Chinese Academy of Sciences, 2008,pp.18–20.

[2] Lv Xuefeng, Cheng Chengqi, and Gong Jianya, "The Overview of Massive Remote Sensing Data Storage Management Technology". China Science, Science, vol. 21, Dec. 2011, pp. 1561–1579.

[3] Lu Jiaheng, Hadoop in Action. CA:Mechanical Industry Press, 2011, pp. 260–261.

[4] L.George, HBase:The Definitive Guide,CA: ,2011,pp. 5–6.

[5] GHEMAWAT S, GOBIOFF H, LEUNG ST, "The Google file system". Proceeding of 19th ACM Symposium on Operating Systems Principles, 2003, pp. 20–43.

[6] DEAN J, GHEMAWAT S, "MapReduce: Simplified data processing on large clusters". OSDI'04:6th Symposium on Operating Systems Design and Implementation, 2004, pp.137–150.

[7] CHANG F, DEAN J, GHEMAWA S, et al,"Bigtable: A distributed storage system for structured data". OSDI'06: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation, 2006, pp. 205–218.

[8] Xu Hongbo, Hao Zhongxiao, "The nearest neighbor query algorithm For space filling curve based on grid division," Journal of computer science, vol.37,2006,pp. 184–188.

[9] Huo Shumin, "Research on Key Technologies of massive remote sensing image management," Changsha: Master's thesis of National University of Defense Technology, 2011:17–18.

[10] Cao Zhongsheng, Li Chenyang, "A Hilbert curve fast coding algorithm based on partition thought," computer engineering and science, vol.21, 2006,pp. 63–65.