Research on Insurance Data Analysis Platform Based on the Hadoop Framework

Mingze Xia(✉)

Heilongjiang University (HLJU), Harbin, China
hlju_xia@yeah.net

Abstract. With the development of IT technology, the traditional information technology cannot meet magnitude data analysis in GB level, let alone in TB level. So it is a perfect time for APACHE company to launch a new product, Hadoop framework, which is a JAVA based basic framework of distributed system, and the versions are now already designated as 2.X series, which means this Hadoop framework is one of the mainstream framework of massive data storage, data procession and analytical in this present.

Keywords: Hadoop framework · Insurance · HDFS · MapReduce · HBase

1 Introduction of Hadoop Framework

With the rapid development of economy of our country, we can see that people's living standard, especially the income, has been rising greatly, people begin to pay more and more attention to their health, property, pension, medical aspects and so on. In recent years, all kinds of natural disasters and the unnatural factors which caused lots of accidents, has brought immensely material and spiritual loss to the people. Just one car accident can perish more than two families. So the property insurance, life insurance and the reinsurance has quickly developed recently, and the insurance companies launched kinds of insurance products depend on different situation. According to the relevant data which announced by the CIRC (The China Insurance Regulatory Commission), the original premium in life insurance business had achieved about 174.42 billion yuan, with year-on-year growth of 31.72% in 2016. and the original premium in property insurance had achieved about 87.245 billion yuan, with year-on-year growth of 9.12%. In this internet+ age, with the information construction upgrades, the insurance industry develops rapidly. According to their own situation, every insurance company use different information platforms, which respectively designed for customers, salespeople, managers and senior leaders, for integrating, screening, excluding, combining and analyzing the insurance data to provide precise guidance, exact sale, and accurate management services.

Hadoop framework [1], which is configured under Linux platform with lower hardware environment, can provide storage and analytics capabilities for massive data. With the development of applications of big data, Hadoop framework can be applied in daily operation system, such as agriculture, finance, medical system and traffic system.

X. Gu et al. (Eds.): MLICOM 2017, Part I, LNICST 226, pp. 494–504, 2018.
https://doi.org/10.1007/978-3-319-73564-1_49

Developers and users are widely recognized this product because of its compatibility, dependability, low cost and high efficiency.Hadoop framework as shown in Fig. 1.

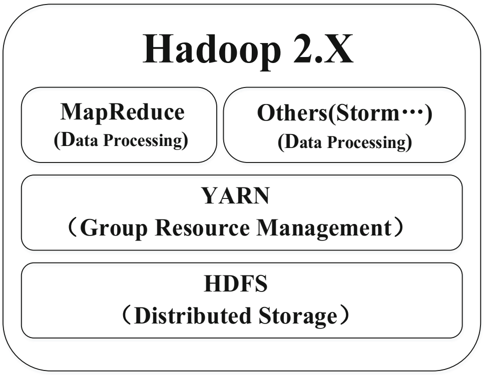

Fig. 1. Hadoop framework

1.1 Introduction of HDFS

Abbreviation in HDFS, Hadoop Distributed File System is one of the main technique in Hadoop framework, which can classify files into blocks, by default, each block set the memory limit to 64 MB, which also can be 128 MB [2]. In this way, it provides convenience for storing subsystem and backing up data. HDFS accessed though block by data stream, which fitted the design principle of HDFS well and improved the inquiry efficiency greatly [3].

HDFS includes two types of nodes, name node and data node respectively. Name node, which called manager node, is responsible for maintaining the whole file system directory, and receiving clients' requirement. At the same time, Name node can retain all information of data note, but not persistent. When the Hadoop starts, the data node will be rebuilt. So does the information.

Data note is the work node, it performs creating, replication and deletion. And it also sent the heartbeat information to name node for proving survival. Data note and name node has made TCP/TP protocol as their communication protocol. By storing data, data note can classify files into blocks, and have a backup copy of blocks in other racks. If the file got loss or damage, the system will invoke alternate block, and repair the damaged one. We can configure backup volumes in HDFS-SITE.SH, when built the Hadoop environment.

Each name node in the HDFS system is one or more corresponding date node. But when the name node got invalid, the system will be crashed. So we use standby name node in Hadoop 2.X series, when the running name node got invalid, the standby name node will take over the system in one minute for maintaining the normal operation of the system. HDFS framework as shown in Fig. 2.

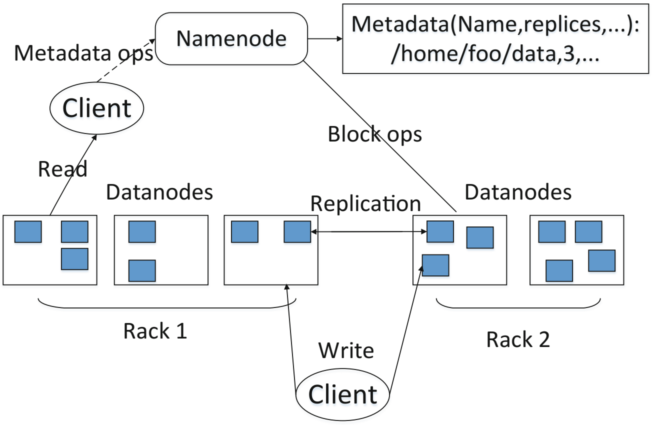

Fig. 2. HDFS framework

1.2 MapReduce

MapReduce is a distributing computing programming model with the program designing idea of “image” and “reduction” [4].

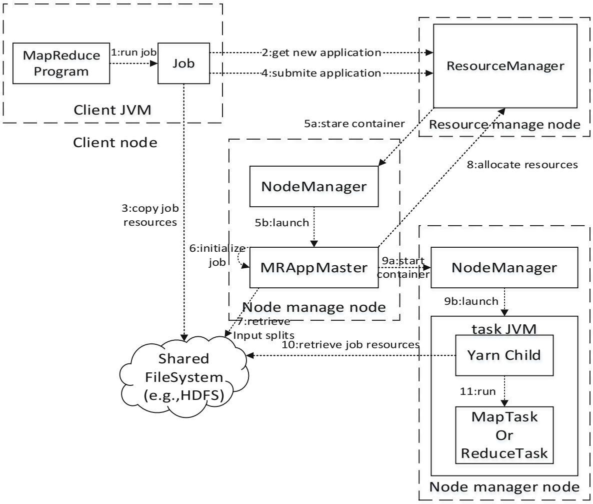

Fig. 3. MapReduce framework

There are two levels in Map Reduce task processes, map level and reduce level [5]. In map level, the system read data in data block at the first time, then map() breakdown data to result output for key-value pairs in Reduce level. Reduce begin to work after the Map finish its task. It can copy and output the received key-value pairs. Reduce has many threads, so the copy process is concurrent, and output the last result to HDFS [6].

After 2.0 version launched, Hadoop framework introduces a new mechanism named YARN, which also can be called Map Reduce2. YARN sets off Jobtrack's functions for avoiding some bottleneck questions, like development insufficient. MapReduce framework as shown in Fig. 3.

1.3 HBase Database

Hadoop Database is a distributed, column-oriented, open-source database which is JAVA language-based. HBase uses Hadoop HDFS as its file storage system, uses Hadoop Map Reduce to process the huge data in HBase, it also can use Zookeeper as collaborative service [7]. Like HDFS, Hbase use one master node to manage one or more region-server dependent computer. At this stage, HBase got widely apply because of its high-reliability, high-performance, colimn-oriented store and scalability. framework as shown in Fig. 4.

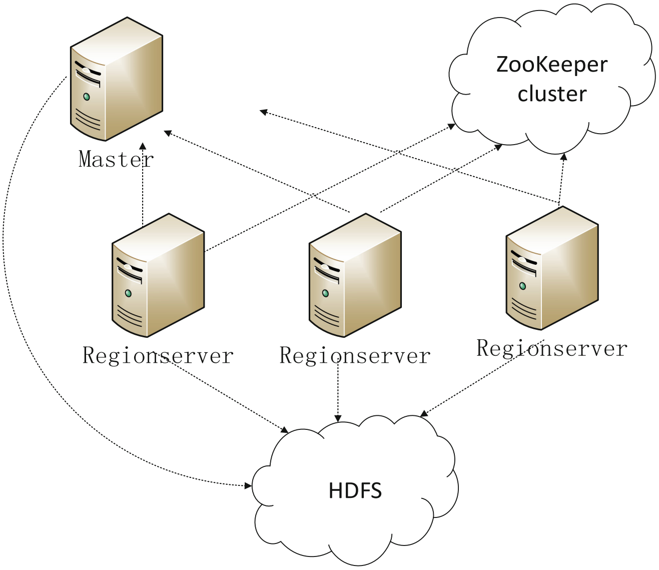

Fig. 4. HBase framework

2 Application of Hadoop Framework in Insurance Industry

At the present stage, most insurance companies, even financial enterprises use Oracle database, which is a distributed-memory relational database developed by Oracle Corporation. It is a classical row store database, but the row store database has the disadvantages.

Oracle database divides the overall data into pieces, as it implements row store, each row of data huddle together. When users access database and want to extract one row, database will put the overall data in the memory and read each row of data. Then, the database separates data for users want to get. Thus, I/O is enhanced greatly.

In insurance industry, the related data is not merely insurance policy information that we see. According to industrial standard JR/T 0048-2015 "Basic Data Model of Insurance", we can see the normal insurance data includes the following several aspects: subject of participants, contract subject, claim settlement subject, asset subject, risk evaluation subject, financial activity subject and insurance product subject. The data is very large. However, manager and decider just concern few about them, even one or several row of these data. For example, when the leader of one provincial branch makes decision, he may ask for IT personnel to acquire the overall premium and completion progress of the administrative city in a certain period, while risk control personnel only concerns if there is large sum of insurance cancellation existed among consumers. The paper, according to Hadoop frame, carries out data enquiry, index analysis, etc.

2.1 Overall Platform Achitecture

The platform mainly consists of three-layer structure, data source layer, data processing layer and data presentation layer. The Overall platform architecture is shown in Fig. 5.

1. Data source layer. It is the storage platform of data source, according to the fact of our country's insurance, it is also the Oracle database platform.
2. Data processing layer. It stores, backs up, calculates and queries the data which imported from the data source layer.
   Data importation. Data importation used open-source tools, like Sqoop, import the data from Oracle into HBase.
   (1) Data storage. Data storage import the data into HBase by Sqoop, and the HBase operated by HDFS.
   (2) Data backup. The backup of the data relies on the redundancy backup function of HDFS framework, which the number of copies is three by default.
   (3) Data calculations and inquiry. We can use the Mepraduce model of Hadoop framework when we want to query the data. And we can use Hive, the data warehouse tool, which can transform SQL into Map Reduce, to set analysis norm.

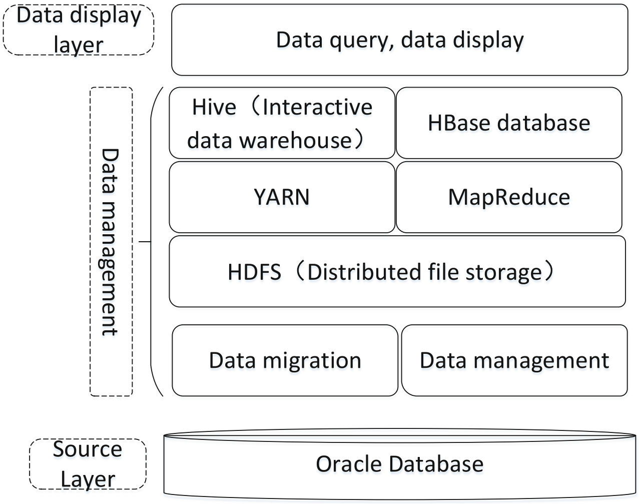

Fig. 5. Overall platform architecture

3. Data presentation layer. It can provide intuitive data for insurance industry managers and policymakers, and it also can quickly provide data querying and data index.

2.2 Related Optimization Adopted in the Implementation of Platform

1. The storage platform of data source is Oracle. During data migration, Sqoop+JDBC technology is usually adopted. According to the literatures, the system adopts open-sourcing OraOOp. In property, compared with traditional Sqoop+JDBC, OraOOp doesn't read data from the same Oracle data block, which reduces I/O. Moreover, OraOOp allocates the loading evenly according to the resource of downloading platform so as to maximize the network bandwidth and I/O.
2. HBase is non-relational database stored based on row. It has great difference with Oracle database. Owing to the difference, when importing data, it should consider the list structure of HBase. Therefore, before importing data, it associates the related lists by using internal connection method so as to form new list. At last, it imports the new list to HBase. Thus, it reduces data redundancy.
3. In job scheduling mechanism, Hadoop2.0 version above pushes two schedulers, including fair scheduler and capacity scheduler. Capacity scheduler corresponds to first in first out principle of old version, while fair scheduler supports the seizing principle. Developer makes choice according to different demands, which avoids the awkwardness that there is only one job scheduling mechanism.
4. In coding, it adopts UTF-8 coding method. Thus, it avoids messy code in the data import process.

3 Analysis and Subsequent Progress of Platform Characteristic

3.1 Characteristic Analysis

Compared with traditional relational database, data computation mode, development mode and equipment resource, the insurance data enquiry platform based on Hadoop frame has the following characteristic (Table 1):

Table 1. The characteristic of the insurance data enquiry platform

Special category	Traditional mode	Mode based on Hadoop frame
Equipment resource	High-end database, server	Common PC
Function	Single	Be expanded as per the demand
Property	Low efficient	High safety and efficiency
Capacity	Be expanded but limited	Be increased as per the demand

To sup up, the insurance data platform based on Hadoop frame has unique advantages of the storage, enquiry and analysis of mass insurance data.

1. Equipment resource is cheap. Compared with expensive advanced computer (database, server), the platform can be mounted on the cheap PC with less expense.
2. High expansibility. Owing to high expansibility of Hadoop, it is superior to traditional mode in data storage and data computation.
3. Safe and reliable. The redundant backup of HDFS enhances the data safety and motor, which avoids the economic loss caused by data loss.
4. High efficiency. The distributed storage of HDFS and parallel computation of MapReduce provide fundamental guarantee for efficient storage and computation of mass insurance data.

3.2 Data Analysis Example

For customers at the age of 22–35 and marriage bond existed between the policy holder and the insured, according to experiences, it is known such couple are not married for a long time. Such customers are potential customers of children insurance. Data mining analysis using Apriori algorithm of MapReduce can screen the potential customers. Database is implemented and computed using MapReduce model of such algorithm. The MapReduce model is described as below:

Map: (Row_ID, Transaction) list(Itemset, $V_1 = 1$)
Reduce: (Itemset, List (V_1)) (Itemset Sum(V_1))

Apriori algorithm is originality algorithm of frequent item set mined according to Boole association rules. The main thought is to find frequent k item set (fail to continue finding k + 1 item set) using iterative method of layer-by-layer searching. According to frequent k item set, produce strong association rule and compute confidence coefficient (confidence coefficient: certainty of rule occurrence). The algorithm is as below:

```
Algorithm Apriori
Input: Transaction DataBase D. Minimum support threshold min_sup .
Output:     Frequent pattern L
$L_1$=search_frequent_1_itemsets( $D$ );
for( $k = 2; L_{k-1} \neq \varnothing; k++$ ) {
 $C_k = aproiri\_gen(L_{k-1})$ ;
 for each transactions $t \in D$ {
      $C_t = subset(C_k ,t)$ ;
 for each candidates $c \in C_t$
      $c.count++$;
 }
  $L_k = \{c C_k \mid c.count \geq \min\_sup\}$
}
 Return $L = \bigcup_k L_k$ ;
Procedure apriori_gen(sssss $L_{k-1} : frequent(k-1)$ itemset )
   for each itemset $l_1 \in L_k$ sss
     for each itemset $l_2 \in L_k$
       if $(l_1[1] = l_2[1]) \wedge \ldots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-2])$
     then {
           $c = l_1 l_2$;
       if has_infrequent_subset( $c, L_{k-1}$ ) then
           delete $c$ ;
      else add $c$ to $C_k$ ;
}
  return $C_k$ ;
Procedure has_infrequent_subset($c$ , candidate $k$ itemset ; $L_{k-1}$ : frequent($k$-1) itemset)
 for each ($k$-1)subset $s$ of $c$
   if $s \notin L_{k-1}$ then
     return TURE;
 return FALSE ;
```

Combining Apriori algorithm with actual situation, the author summarizes the purchase phenomenon of customers at the age of 20–35 on Branch A. The mining result of association rule is shown in the following figure:

According to the data, it is known customers at the age of 20–35 pay more attention to accidental injury. Meanwhile, whatever the customer effect the insurance for spouse or himself, more than 50% of them buy children insurance for his children. It indicates that young customer has strong insurance sense and is capable of purchasing insurance for children (Table 2).

Table 2. The result of the association rule mining

No.	Rule description	Confidence coefficient %
1	Relation = spouse, type of insurance = accidental injury => type of insurance = children insurance	60.1
2	Relation = oneself, type of insurance = accidental injury => type of insurance = children insurance	49.7
3	Relation = spouse, type of insurance = critical illness => type of insurance = children insurance	59.4
4	Relation = oneself, type of insurance = critical illness => type of insurance = children insurance	50.5
5	Relation = oneself => type of insurance = critical illness	40.9
6	Relation = oneself => type of insurance = accidental injury	59.1
7	Relation = spouse => type of insurance = critical illness	43.3
8	Relation = spouse => type of insurance = accidental injury	56.7

Similarly, for group insurance system, according to the age structure, sex ratio and recruitment practice of the insurance application unit and loss situation at present, loss phenomenon in the future 5 years can be calculated. Thus, insurance company can forecast capital budget and human input, which provides powerful guarantee for decision in the future.

3.3 System Performance Test

When the data size is different, the author returns 30,000 items of specific data and compared time used by Oracle database and distributed cluster. In this way, the difference between distributed system performance and traditional database system performance was compared. The test result was shown in the following Fig. 6:

3.4 Subsequent Promotion of Platform

Insurance data platform based on Hadoop frame provides mass data enquiry and index analysis, which provides convenience for manager and decider of insurance industry. With social development, it is not enough for insurance data platform only providing

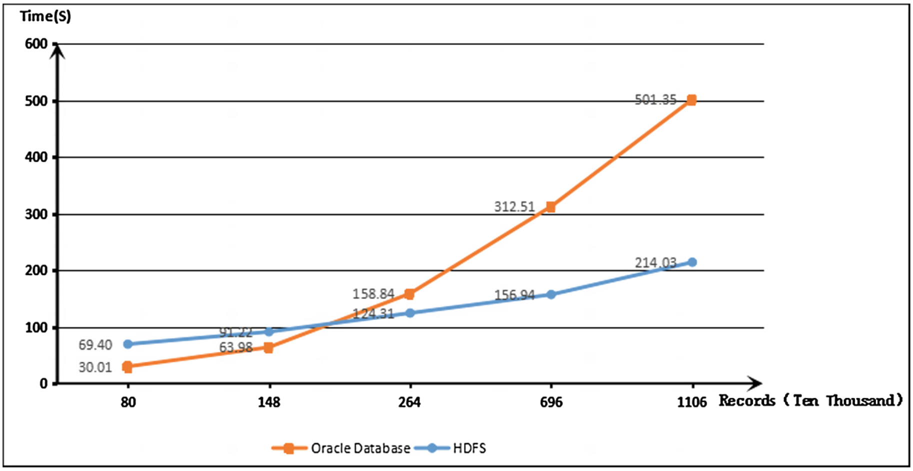

Fig. 6. The test result

data enquiry function. thus, it is necessary to carry out more research according to the characteristic of Hadoop frame.

According to the function difference of user, more data research and exploration should be done through the platform. Combined with lecturer in marketing department and the market experience of salesperson, product development is carried out through data analysis so as to push product according to the demand of different groups. Through combining with external data of medical insurance, the platform can market accurately, as well as avoiding customer loss. Model library establishment can help insurance company and supervisory agency to find illegal behavior earlier, such as money laundering so as to avoid the occurrence of delinquency.

4 Conclusion

With economic development of increasing enhancement of living level, people pay more attention to health and property safety. Meanwhile, according to the online and surrounding cases, more people purchase insurance to guarantee the health and property safety so as to reduce loss.

Based on characteristic of Hadoop frame, the paper analyzes its application in insurance industry. With the development of Internet, intelligence analysis appears very important. Finding “relation” in the mass data and making rational use of it provide great help for insurance company. Future is unpredictable, while the platform based on big data and Hadoop frame can provide relatively accurate prediction analysis for insurance company so that the insurance company can master the market quotation earlier and create more profit and benefit.

References

1. Apache: Hadoop: Open Source Implementation of MapReduce. http://hadoop.apache.org/
2. Yuanqi, C., Yi, Z., Shubbhi, T., Xiao, Q., Jianzhong, H.: aHDFS: an erasure-coded data archival system for Hadoop clusters. IEEE Trans. Parallel Distrib. Syst. **PP**(99), 1 (2017)
3. Yanfei, G., Jia, R., Dazhao, C., Xiaobo, Z.: iShuffle: improving Hadoop performance with shuffle-on-write. IEEE Trans. Parallel Distrib. Syst. **28**(6), 1649–1662 (2017)
4. Akash, H., Kiran, B.: A MapReduce based approach for classification. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1–5. IEEE Press, Coimbatore (2016)
5. Jeffrey, D., Sanjay, G.: MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation, p. 10 (2004)
6. Manjunath, R., Tejus, Channabasava, R.K., Balaji, S.: A big data MapReduce Hadoop distribution architecture for processing input splits to solve the small data problem. In: 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, pp. 480–487 (2016)
7. Frank, P., Johannes, G., David, B.: Pick your choice in HBase: security or performance. In: 2016 IEEE International Conference on Big Data (Big Data), Washington, D.C., pp. 548–554 (2016)