Research of Hive Data Warehouse

Yezehao Huai

yhuai001@fiu.edu

Abstract

[1]With the e-commerce platform and portal visits become more and more, log information is also generated more and more, traditional log file processing methods can't meet demand. By analyzing the characteristics of Hadoop Distributed File System, parallel computing framework MapReduce and Hive data warehouse technology, build a weblog data warehouse is very feasible. Aiming at the bottleneck of the traditional processing and computing of massive data, a massive web log analysis mechanism based on Hive is proposed. Hive is a data warehouse software project based on Apache Hadoop and HDFS for data query, summarization, and analysis. Hive QL is a SQL-like declarative language to compile queries expression into Map-Reduce jobs executed on Hadoop. Through the Hadoop distributed system architecture and Hive data warehouse, massive web logs are analyzed and processed, and user browsing behavior is analyzed and studied. This paper research the features of Hive and provide a design for the Weblogs data warehouse based on Hive. The results of the analysis page views, bounce rates, independent IP numbers, and new register members of user browsing are significantly guiding the website construction.

Keywords Hive, Data warehouse, Weblogs, Hadoop

ACM Reference format:
Yezehao Huai. 2017. Research of Hive Data Warehouse. SIG Proceedings Paper in word Format. In *Proceedings of ACM Woodstock conference, El Paso, Texas USA, July 1997 (WOODSTOCK'97)*, 4 pages.
DOI: 10.1145/123_4

1 Introduction

With the popularization of Internet technology, the amount of information on the network increases exponentially. The web is already the largest information system in the world. As one of the important components of this system, the web log records all the information the user is browsing the web. By processing and analyzing these log messages, we can understand the user's behavioral characteristics, so as to transform the layout of the web page and improve the website's traffic, thereby bringing higher profits to the enterprise.

Web access log contains a large number of user behavior information, such as access information, browsing information, purchasing information, preferences, and so on. By processing the log and modeling user behavior, valuable information can be obtained, such as user attribute information, user tags, user interests, purchase intention, and then the information of these users are clustered, the user set is divided, Services for web personalization. Therefore, the log file is stored and processed a very important meaning. Through data analysis technology and data mining technology, to obtain the user's behavior characteristics from web logs has become the focus of the business community.

However, with the dramatic increase in the number of users, the amount of information recorded by the web log is also getting larger and larger. Traditional methods are generally handled by the idea of divide-and-conquer or multithreaded multitasking when dealing with massive datasets. If only by improving the computer's storage and performance obviously can't essentially solve this problem. In this paper, Hive-based web log analysis system is designed and implemented by using Hadoop platform.

2 Related Technologies

2.1 Hadoop Distributed system

Hadoop is an open-source distributed framework under the Apache Software Foundation, is widely used in many large enterprises. Hadoop Distributed File System and Map / Reduce parallel programming model is two cores of Hadoop.

Hadoop mainly through HDFS to achieve the underlying support for distributed storage, and through the Map / Reduce to achieve distributed parallel computing processing support. So users can develop distributed programs without knowing the underlying details of the distribution.

HDFS uses master/slave master-slave architecture model, a management node (Name Node) as the main server, the management of the entire file system namespace and client-side access to files. A number of data nodes (Data Node) management of stored data. HDFS to store data in the form of files, it is divided into several pieces of data files (Block) stored in different Data Node dispersed, and can be configured by the Block backup. The number of backup to other Data Node, so as to achieve the purpose of disaster recovery. HDFS architecture as shown in Figure 1.

[*] Produces the permission block, and copyright information

[†] The full version of the author's guide is available as acmart.pdf document

[1] It is a datatype.

WOODSTOCK'97, July 2016, El Paso, Texas USA

DOI: 10.1145/123_4

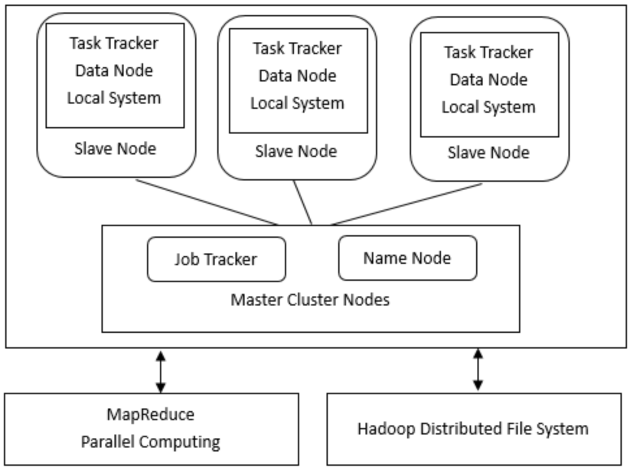


Figure 1. Hadoop distributed storage and compute framework

2.2 MapReduce Parallel Computing Model

Map / Reduce programming mode is mainly done through two sets of operations Map and Reduce, the two processes were abstracted to the corresponding map () function and reduce () function. Among them, map () is mainly to cut the task into multiple sub-tasks, to facilitate each node to work independently; reduce () is the summary of the results of the various nodes together. Map / Reduce process of processing the data set shown in Figure 2.

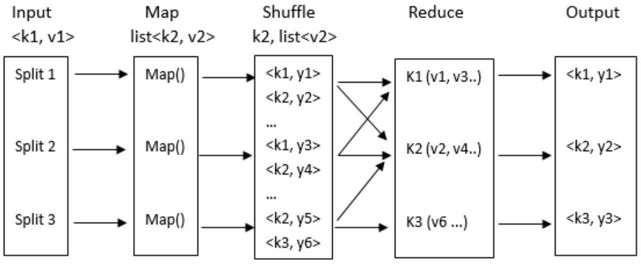


Figure 2. MapReduce Parallel Computing Model

3 Hive Data Warehouse

3.1 Data Warehouse Concept

Data warehousing is a structured data environment for Decision Support Systems (DSS) and online analytical application data sources. Data warehouse research and solve the problem of access to information from the database. Data warehouses are characterized by subject-oriented, integration, stability and time-variability.

Data warehousing, pioneered by Bill Inmon, the father of data warehousing, was introduced in 1990 with the primary function of still organizing large amounts of data that organizations have accumulated over the years through the online transaction processing (OLTP) of information systems, through data warehouse theory The unique data storage structure is systematically analyzed and collated to facilitate the analysis of various analytical methods such as online analytical processing (OLAP) and data mining, and further support such as Decision Support System (DSS), Supervisory Information System (EIS) to help decision makers analyze valuable information quickly and efficiently from a vast amount of data to facilitate decision-making and rapid response to external environmental changes to help build business intelligence (BI).

Bill Inmon's definition of the data warehouse is widely accepted - Data Warehouse is a subject-oriented, integrated, relatively stable non-volatile data collection reflecting time variant to support decision making. Figure 3 is the data warehouse architecture.

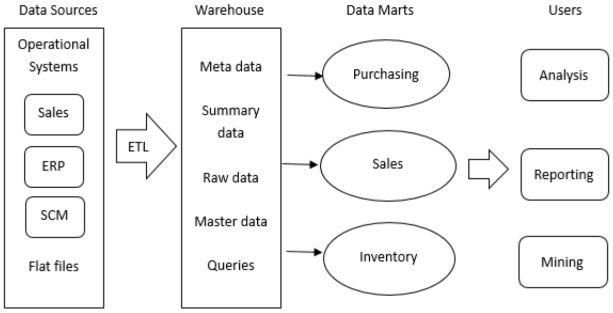


Figure 3. Data warehouse architecture

3.2 Hive Architecture

Hive is a data warehouse infrastructure built on top of Hadoop's file system and provides many capabilities for managing data warehouses: ETL (Extract, Transform, and Load) processing of data, data storage management, querying large data sets and analysis. Hive also defines a SQL-like language - Hive QL (HQL). Hive QL is a SQL-like declarative language to compile queries expression into Map-Reduce jobs executed on Hadoop. Users can manipulate HQL based on SQL syntax, parse and compile, and ultimately generate Map and Reduce tasks to process the data. To take advantage from HDFS, the tables saved in the Hive are actually the HDFS files. There will be an external table to locate the address of original table on HDFS. The Hive architecture is shown in Figure 4.

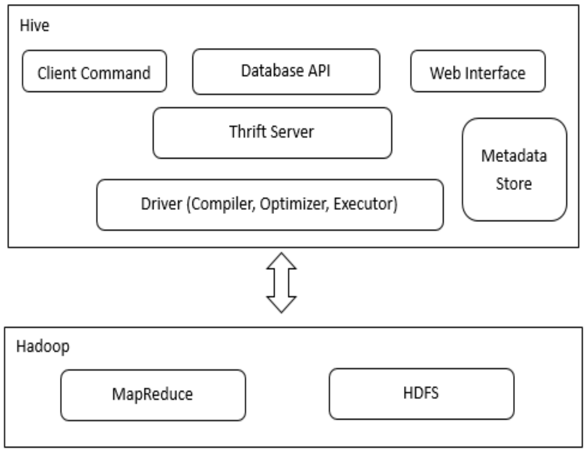


Figure 4. Hive architecture

User Interface includes the main client, database API, and web interface. Metadata storage is generally stored in RDBMS, such as MySQL, Derby and so on. Mainly include the table name, table attributes, table partitions, and directories. Driver including compiler, optimizer and so on. Mainly to complete the HQL query to the query plan generation. Hadoop Platform provide HDFS storage and MapReduce for calculating

3.3 Hive vs RDBMS

The difference between the Hive and Relational Database Management System are as follows: Query language. Because SQL is widely used in data warehouses, a SQL-like query language HQL has been designed specifically for Hive's features. Developers familiar with SQL development can easily develop with Hive.

Data storage location. Hive is built on top of Hadoop and all Hive data is stored in HDFS. The database can save the data in the block device or the local file system.

Data format. There is no specific data format defined in Hive. The data format can be specified by the user. The user-defined data format needs to specify three attributes: column delimiter (usually "\ t", "\ x001"), line delimiter (" \ n ") as well as the method of reading the file data (there are three file formats TextFile, SequenceFile and RCFile by default in Hive). Because no data format conversion from user data format to hive definition is required during data loading, Hive does not make any changes to the data itself during loading, but simply copies or moves the data contents to the corresponding HDFS directory. In the database, different databases have different storage engines, define their own data format. All data will be stored according to a certain organization, therefore, the process of loading data into the database will be time-consuming.

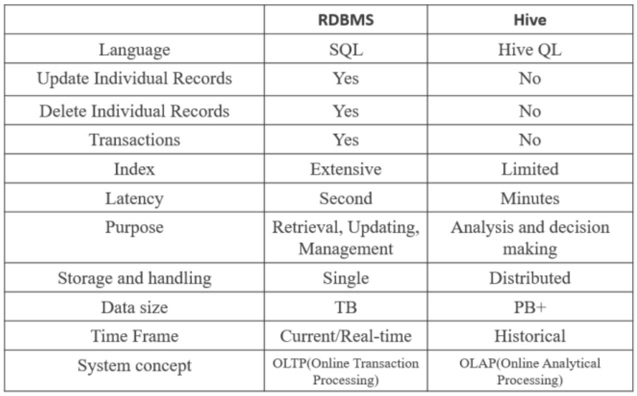
Data update. Because Hive is designed for data warehouse applications, data warehouse content is less to read and write. Therefore, Hive does not support the rewrite and add data, all the data is determined at the time of loading. The data in the database is usually need to be modified often, so you can use INSERT INTO ... VALUES to add data, use UPDATE ... SET to modify the data.

Index. As above said, Hive does not process data and does not even scan the data as it loads, so it does not index some of the keys in the data. When Hive accesses a particular value in the data that meets the condition, it needs to scan the entire data violently, so the access delay is high. Due to the introduction of MapReduce, Hive can access data in parallel, so even without an index, Hive can still benefit from the large amount of data accessed. Database, usually for one or several columns indexing, so for a small amount of specific conditions of data access, the database can have a high efficiency, lower latency. Due to the high latency of data access, Hive is not suitable for online data query.

Implementation. Most queries in Hive are executed via MapReduce provided by Hadoop (queries like *select * from table* do not require MapReduce). The database usually has its own execution engine.

Execution delay. Mentioned earlier, hive in the query data, because there is no index, you need to scan the entire table, so the delay is higher. Another factor that leads to high latency for Hive execution is the MapReduce framework. Because MapReduce itself has high latency, there is also a high latency when performing Hive queries with MapReduce. In contrast, the execution delay of the database is low. Of course, this low is conditional, that is, the data size is small, and when the data is large enough to exceed the processing power of the database, Hive's parallel computing is obviously able to demonstrate its advantages.

Scalability. Because Hive is built on top of Hadoop, Hive's scalability is consistent with Hadoop's scalability (the largest Hadoop cluster in the world at Yahoo! 2009 was around 4,000 nodes). The database due to strict restrictions on the semantics of ACID, extended line is very limited. At present, Oracle, the most advanced parallel database, has only about 100 theoretical extensions. Data size. Because Hive is clustered and can use MapReduce for parallel computing, it can support very large-scale data; correspondingly, the data that the database can support is smaller. Table 1 is the table to show the difference between Hive and Relational Database Management System.



	RDBMS	Hive
Language	SQL	Hive QL
Update Individual Records	Yes	No
Delete Individual Records	Yes	No
Transactions	Yes	No
Index	Extensive	Limited
Latency	Second	Minutes
Purpose	Retrieval, Updating, Management	Analysis and decision making
Storage and handling	Single	Distributed
Data size	TB	PB+
Time Frame	Current/Real-time	Historical
System concept	OLTP(Online Transaction Processing)	OLAP(Online Analytical Processing)

Table 1. Difference between Hive and RDBMS

4 Weblogs Warehouse Based on Hive

4.1 Weblogs Warehouse Architecture Design

The popularization of the Internet has made the web the largest information system in today's big data era. Weblogs which contains a large number of user access to information, through the web log mining, we can get a lot of valuable information. Weblogs file is a text file saved on the web server, usually stored in .txt format.

As data center websites are getting more and more traffic, more and more web log information is generated. The traditional computing power relying on a single node can’t meet the demand. The use of cloud computing technology will consume a large number of computational resources of complex computing distributed to the network through the multi-node computing has become a new and effective solution. Hadoop is a popular open source framework for storing and processing large-scale datasets on commodity hardware. However, Hadoop's Map / Reduce program model is at a low level, and developers need to develop client programs that are often harder to maintain and reuse. In this part, I use Hive to design a system for massive Weblogs processing. The Weblogs can be collected on HDFS and HBase, then use hive to do the data cleaning, statistical analysis, and get results saving into MySQL and HBase. The manager can read the summary table in MySQL and read the details information in HBase.

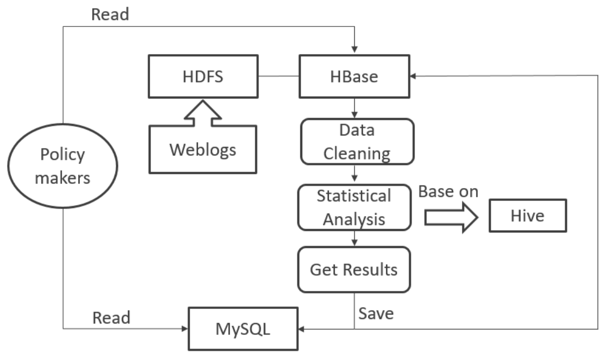


Figure 5. Weblog data warehouse architecture

4.2 Key Indicators for Weblog Analysis

Based on the research and analysis of the Web log dataset, there are many nonsense data in the Weblog record. For example, “[01/Aug/2017:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt HTTP/1.0" 200 1839”. This Weblog includes the client IP address, access date/time, access resources, HTTP status code, and network traffic. However, to get the useful information, we have to set up several key indicators. After clean the data, we can create an external table and use date as a partition indicator. Here is the example:

```
hive>CREATE EXTERNAL TABLE website_name(ip string,
time string, url string) PARTITIONED BY (logdate string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY '\t'
LOCATION '/fiuproject/weblog/cleaned';
```

The first indicator is the Page View, which refers to the sum of all the user's browsing pages. It is an independent user’s recorded once for each page opened. For the total page views, we can assess the user's interest in the site, just as the same as ratings for the TV series. But for website operators, more importantly, is the number of page views per section. To the get the number of Page View, we use the Hive QL as:

```
hive>CREATE TABLE website_name_pv_2017_08_01
     >AS SELECT COUNT(1)
     >AS PV
     >FROM website_name
     >WHERE Logdate='2017_08_01';
```

The newly registered user during a period is also an important indicator to reflect the marketing and operations results. We have to use string search function here in the Hive QL:

```
hive>CREATE TABLE fiuforum_mem_2017_08_01
     >AS SELECT COUNT(1)
     >AS MEM
     >FROM website_name
     >WHERE logdate='2017_08_01’AND INSTR(url,
registration requests)
```

Within one day, the total number of different independent IPs accessing the website. One of the same IP regardless of access to several pages, the number of independent IP is one. No matter how many computers or other users on the same IP, the number of independent IPs is the most direct measurement of website promotion activities.

```
hive>CREATE TABLE fiuforum_ip_2017_08_01
    >AS SELECT COUNT(DISTINCT ip)
   >AS IP
   >FROM website_name
   >WHERE logdate='2017_08_01';
```

Bounce Rate is the percentage of visitors to a particular website who navigate away from the site after viewing only one page. The bounce rate is a very important stickiness index of visitors. It shows the visitor's interest in the website: the lower the bounce rate is, the more interested the visitor is in the content of the website, shows the visitors are more likely to be effective and loyal users of the website. Use Hive QL to get total bounces at first and then calculate the Bounce rate = Total bounces/Total PV.

```
hive>CREATE TABLE fiuforum_bounce_2017_08_01
    >AS SELECT COUNT(1)
    >AS BOUNCES
    >FROM (SELECT COUNT(ip) AS times FROM
website_name WHERE logdate='2017_08_01' GROUP BY ip
HAVING times=1) ;
```

5 Future Work

In the initial Hive architecture, its accesses to files directly in the HDFS. Nowadays, Hive can also access files in other data storage systems like HBase. This method provides random write, random read, little range scan for the Hive users. Furthermore, now we have big data technique such as Spark and Tez whose performance is better than MapReduce in most of the cases. Hive can also translate the Hive QL query language into Spark and Tez jobs. It is really important to focus the Hive applications on the new big data process technique in the future.

6 Conclusions

Hive is a data warehouse software simplifies reading, writing and managing large data sets exist in the distributed system environment and using SQL syntax for querying. Built based on the Hadoop, Hive provides many useful feature such as empowering data warehouse tasks, reporting, and data mining. Queries of Hive QL can be executed on MapReduce, Spark, and Tez. This paper research the principle of Hive data warehouse from basic features to the use cases, provide a design for the Weblogs data warehouse based on Hive. The Website operations management can make use of Hive for the analysis of weblogs by focusing on the page view, independent IP numbers, new registered users, and bounce rate.

A Headings in Appendices

References

[1] http://hadoop.apache.org/

[2] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, Raghotham Murthy. Hive- a warehousing solution over a Map-reduce framework. August 2009

[3] Ashish Thusoo, Data warehousing and analytics infrastructure at facebook, 2010

[4] Yezheng Wang, Shufang Li, A Big Data Storage Optimization Method Based on Hive Log Analysis. Computer engineering & Software P94-100, 2014

[5] Youwei Wang, Weiping Wang, Dan Meng, Hive Data Warehouse Query Optimization Based on Statistical Methods, 2015

[6] A,Cuzzocrea, L.Bellatreche, Data warehousing and OLAP over big data: current challenges and future research directions, 2013.

[7] Long Zhao, Rongan Jiang, Research on Mass Search Log Analysis System Based on Hive, Application Research of Computers P3343-3345 2013

[8] Rui Zhang, Research and Design of Logistics Big Data Platform Based on Hive Data Warehouse, 2017

[9] N.Glance. Extracting structured data from weblogs. 2015

[10] Xiaoliang Ma, Feng Tian, Hadoop native Hbase architecture, Hive, Lealone, Phoenix and other operating components of the comparison, 2017

[11] Dewen Wang, Kai Xiao, Lei Xiao, Power Equipment Status Information Data Warehouse Based on Hive. Sep 2013

[12] T.Dokeroglu, S.Ozal, Improving the performance of Hadoop Hive by sharing scan and computation tasks. July 2014

[13] Tao Li, Jianhui Wang, Design and Research of Hive - based Payment SDK Log Analysis System, Computer Applications and Software P51-54, 2017

[14] Jinbiao Gao, Lili He, Yunyang Zou, Research on Hive and Hbase Based on Distributed Storage System

[16] Tk Das, A.Mohapatro, A study on big data integration with data warehouse. Mar 2014

[17] Mengxin Song, Hive-based mail log analysis, China CIO New, P115-117, 2016

[18] Huanliang Jiang, Hive- based log repository construction, Computer Era P21-24, 2015

[19] Ashutosh Chauhan, Alan Gates, Gunther Hagleitner, Eric N.Hanson, Jitendra Pandey, Yuan Yuan, Rubao Lee. Major technical advancements in apache hive, June 2014.

[20] C.Chiu, NH Chiu, RJ Sung, PY Hsieh, Opinion mining of hotel customer-generated contents in Chinese weblogs. 2015

[21] A. Vaisman, Data warehouse systems: design and Implementation. 2014