The network data storage model based on the HTML

Lei Li[1], Jianping Zhou[2]
[1]Network Information Center
Chongqing University of Science and Technology
Chongqing, China
46016423@qq.com
[2]Network Information Center
Chongqing University of Science and Technology
Chongqing, China
48163779@qq.com

Runhua Wang[3], Yi Tang[4]
[3]Teaching Quality and Assessment Office
Chongqing University of Science and Technology
Chongqing, China
runhua822@sohu.com
[4]School of Electronic and Information Engineering
Chongqing University of Science and Technology
Chongqing, China
tangyi19761016@yahoo.com.cn

***Abstract*—in this paper, the mass oriented HTML data refers to those large enough for HTML data, that can no longer use traditional methods of treatment. In the past, has been the Web search engine creators have to face this problem be the first to bear the brunt. Today, various social networks, mobile applications and various sensors and scientific fields are created daily with PB for HTML data. In order to deal with the massive data processing challenges for HTML, Google created MapReduce. Google and Yahoo created Hadoop hatching out of an intact mass oriented HTML data processing tools for ecological system.**

Keywords- mass oriented HTML data for HTML; HTML

I. INTRODUCTION

With the popularity of MapReduce, a data storage layer for HTML, MapReduce and query (SMAQ) consisting of massive data processing model for HTML stack also gradually emerged. The SMAQ system is usually open, distributed, running in a general hardware.

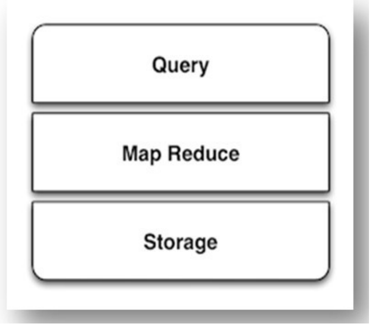

Figure 1 the structure of MapReduce

Like Linux, Apache, MySQL and PHP consisting of LAMP changed the Internet application and development fields, SMAQ will take mass oriented HTML data processing to a wider world. As LAMP becomes Web2.0 key movers, SMAQ system will support a innovation oriented HTML data driven products and services in the new era.

Although based on the architecture of the Hadoop occupies a dominant position, but the SMAQ model also contains a large number of other systems, including the popular NoSQL for HTML database. This article describes the SMAQ stack model and those who today may be included in the model under mass oriented HTML data processing tools.

II. MAPREDUCE

MapReduce is a Google for creating web webpage index created. MapReduce framework has become today most mass oriented HTML data processing plant. MapReduce is the key, will be oriented in the HTML data collection of a query division, then at multiple nodes to execute in parallel. This distributed model addresses for HTML data that is too large to store in a single machine problem.

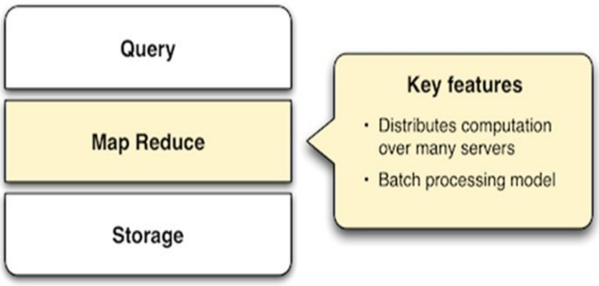

Figure 2 web webpage index

To understand how MapReduce works, we first see its name reflects the process of two. In the first stage map, HTML data is input for a processing, converted into an intermediate result sets, then the reduce stage, the intermediate results by statute to produce a we expected results of induction.

Speaking of MapReduce, usually to give an example is finding a document of different words appear a number. In stage map word is out, and then to a count value of 1, in the reduce node, the same word count value accumulation.

Published by Atlantis Press, Paris, France.

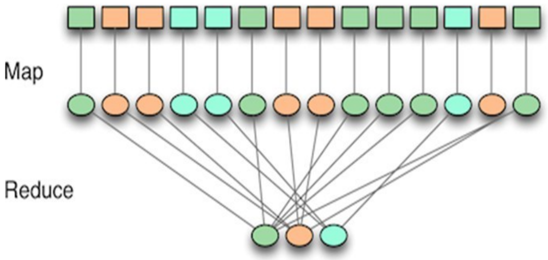

Figure 3 the MapReduce process

Is it right? Looks like a very simple work is very complex, it is MapReduce. In order for MapReduce to complete this task, map and reduce stage must comply with certain limits allows work to be parallel. The query request is converted into one or more MapReduce is an intuitive process, in order to solve this problem, some of the more advanced abstract was carried out, we will in the following query on that day were discussed.

A. Hadoop MapReduce

Hadoop is the main source of MapReduce. Funded by Yahoo, 2006 by Doug Cutting to create, in 2008 reached web scale oriented HTML data handling capacity.

The project is now managed by Apache Hadoop. With the continuous efforts, and several sub projects together constitute a complete SMAQ model.

```
    public static class Map
    extends Mapper<LongWritable, Text, Text, IntWritable>
{
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
        public void map(LongWritable key, Text value,
Context context)
        throws IOException, InterruptedException {
        String line = value.toString();
        StringTokenizer tokenizer = new
StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                context.write(word, one);
            }
        }
    }
```

The corresponding reduce function as follows：

```
    public static class Reduce
    extends Reducer<Text, IntWritable, Text, IntWritable>
{
    public void reduce(Text key, Iterable<IntWritable>
values, Context context) throws IOException,
    InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
             sum += val.get();
            }
        context.write(key, new IntWritable(sum));
          }
    }
```

Use Hadoop to run a MapReduce job includes the following steps:

1 with a java program defines the MapReduce stages

2 will face HTML data is loaded into the file system

The 3 submitted to job for implementation

4 from the file system to obtain the execution result

B. Storage

From the HTML data access to the results stored, MapReduce needs to deal with storage. With traditional HTML oriented database is different, the input of MapReduce oriented HTML data not relation type. The input for the HTML data stored in different chunk, can be divided to different nodes, then provided in the form of key-value provided to map stage. Based on the HTML data, without the need for a schema, but probably not structure. But the HTML data must be distributed, can provide different processing node.

The design and characteristics of the storage layer is important not only because it is the interface with MapReduce, but also because they directly determines the HTML oriented data loading and results of query and display of convenience.

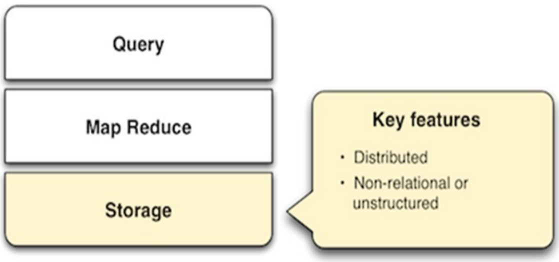

Figure 4 storage layer design

C. Hadoop distributed file system

Hadoop uses standard storage mechanism is HDFS. As the core component of Hadoop, HDFS has the following characteristics, detailed see HDFS design document:

Fault tolerant if failure is normal allow HDFS to run in the general hardware

Stream oriented HTML data access - HDFS implementation is given to the batch processing, therefore focuses on high throughput rather than for HTML data random access

Highly scalable– HDFS can be extended to PB class for HTML data, such as Facebook has such a product using

Portability - Hadoop can cross operating system transplant.

D. HBase, Hadoop for HTML database

A HDFS more usable method is the HBase. Imitation of Google BigTable for HTML database, HBase is also a design used to store massive data storage type for the HTML column for HTML database. It also belongs to the category of NoSQL for HTML database, similar to the Cassandra and Hypertable.

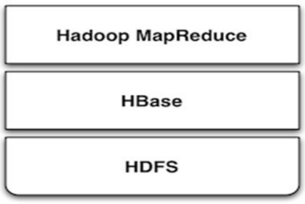

Figure 5 Hadoop for HTML database

HBase uses HDFS as the underlying storage system, it also has through a large number of fault-tolerant distributed node to store a large number of oriented HTML data capability. And other similar columns of memory for HTML database, HBase also provides access to API based on REST and Thrift.

Because the created index, HBase can provide some simple query to provide content for fast random access. For the complex operation, HBase Hadoop MapReduce for the HTML data source and a storage target. The HBase allows the system to the HTML database with MapReduce to interact, rather than through the bottom of the HDFS.

E. Hive

HTML oriented data warehouse or make reports and analysis more simple storage mode SMAQ system is an important application field. Initially developed in the Facebook Hive, is a built on top of Hadoop for HTML data warehouse framework. Similar to HBase, Hive provides a HDFS in the table based on the abstract, simplified structured oriented HTML data loading. Compared with HBase, Hive MapReduce job can only run batch oriented HTML data analysis. As the following query the part description, Hive provides a SQL query language to perform MapReduce job.

F. Cassandra and Hypertable

Cassandra and Hypertable is BigTable mode similar to the HBase column of memory for HTML database.

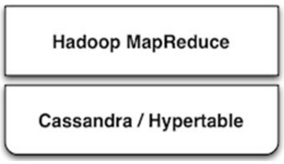

Figure 6 Hypertable mode

As a project of the Apache, Cassandra was originally produced in Facebook. Now used in many large-scale web sites, including Twitter, Facebook, Reddit and Digg. Hypertable produced by Zvents, is an open source project.

The two for the HTML database with Hadoop MapReduce interactive interface, allowing them as Hadoop MapReduce job for HTML data source and target. At a higher level, Cassandra and Pig query language (see Chapters integrated query), and Hypertable has been integrated with Hive.

G. NoSQL for HTML database MapReduce

So far we have mentioned storage solutions are dependent on the Hadoop MapReduce. There are a number of NoSQL for HTML database to store the HTML data parallel computing has built-in support for Mapreduce. Multi component SMAQ and Hadoop system structure is different, they provide a storage, MapReduce and query integrated to form a self contained system.

Riak, and the two front face is also very similar to HTML database. But more attention to high availability. You can use the JavaScript or Erlang description MapReduce.

III. RELATIONS FOR HTML DATABASE INTEGRATION

In many applications, the main source for the HTML data stored in a relational database for HTML, such as Mysql or Oracle. MapReduce usually through two ways to use these data for HTML:

Use of relational databases as a source for HTML (such as social network friends list)

The MapReduce results back into a relational database (such as for HTML based on friend interest products recommended list)

For the Hadoop system, through the use of Cascading API cascading.jdbc and cascading-dbmigrate can achieve a similar function.

A. streaming oriented HTML data source integration

Relationship oriented HTML database as well as streaming oriented HTML data source (such as a web server log, the sensor output) consists of a massive HTML oriented data system most commonly oriented HTML data sources. Cloudera 's Flume program is designed to provide streaming oriented HTML data source and Hadoop integrated convenient tool. Flume collected from clusters of machines on the oriented HTML data, will they continue into HDFS. Facebook Scribe server also provides a similar function.

B. commercial SMAQ solutions

Some of the MPP for the HTML database has a built-in MapReduce function support. MPP for HTML database has a parallel running independent nodes in distributed architecture. Their main function is for the HTML data warehouse and analysis, you can use the SQL.

Greenplum: Based on the open source PostreSQL DBMS, running in a distributed hardware cluster. MapReduce as the supplement of SQL, can be carried out in a more rapid and more extensive on the Greenplum for HTML data analysis, is reduced by several orders of magnitude of the query time. The Greenplum MapReduce allows the use of the HTML database storage and external oriented HTML data source comprising mixing based on HTML data. The MapReduce operation can use the Perl or Python function description.

IV. QUERIES

Through the above java code can be seen using a programming language to define MapReduce job map and reduce process is not so direct and convenient. In order to solve this problem, SMAQ system introduced a higher query layer to simplify the MapReduce operation and the results of inquiries.

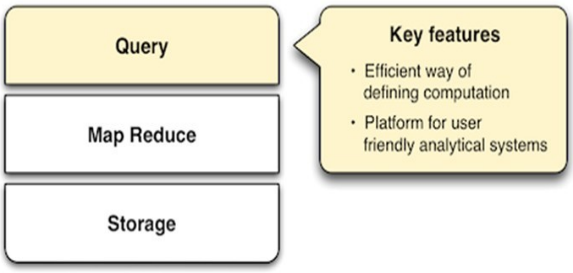
Query
Map Reduce
Storage
Key features
• Efficient way of defining computation
• Platform for user friendly analytical systems

Figure 7 the Hadoop organization

A lot of use of Hadoop tissue in order to make the operation more convenient, have to Hadoop API performed inside package. Some have become open source projects or commercial products.

Query layers are usually used to describe not only provides the calculation process, and support the HTML data access and simplified on the MapReduce cluster flow of execution.

A. Pig

Developed by Yahoo, is currently part of the Hadoop project. Pig provides a called Pig Latin advanced query language to describe and run MapReduce job. Its purpose is to make the Hadoop more accessible to those familiar with SQL developers to access, in addition to a Java API, it also provides an interactive interface. Pig has been integrated in the Cassandra and HBase for HTML database. The following is to use Pig to write the above wordcount examples, including the HTML data loading and storage process ($0 representing the recorded the first field).

```
input = LOAD 'input/sentences.txt' USING TextLoader();
words = FOREACH input GENERATE
FLATTEN(TOKENIZE($0));
grouped = GROUP words BY $0;
counts = FOREACH grouped GENERATE group,
COUNT(words);
ordered = ORDER counts BY $0;
STORE ordered INTO 'output/wordCount' USING
PigStorage().
```

B. Hive

As mentioned previously, the Hive is a built on top of Hadoop 's open source HTML oriented data warehouse. Created by Facebook, it provides a very similar to the SQL query language, and provides a simple built-in web query interface. So it is very suitable for those familiar with SQL for developers.

With Pig and Cascading need to be compiled in Hive, one of the strengths is to provide ad hoc queries. For those who have mature business intelligence systems, Hive is a natural starting point, because it provides a for non-technical more user friendly interface. Cloudera Hadoop release in Hive integration, and through the HUE program provides a more advanced user interface, so that users can submit queries and monitor execution of MapReduce job.

C. Cascading, the API Approach

The key features of Cascading is that it allows developers to MapReduce job in the form of flow through the assembly, Although very powerful, Cascalog is still only a small range in the use of language, because it is not like Hive provide a kind of SQL language, also unlike Pig is the process of. The following is the use of Cascalog wordcout Completed Example：

```
(defmapcatop split [sentence]
(seq (.split sentence "\\s+")))
(?<- (stdout) [?word ?count]
(sentence ?s) (split ?s :> ?word)
(c/count ?count))
```

D. use Solr search

Originally developed in the CENT, now as a Apache project Solr, has grown from a single text search engine for the evolution of navigation and the result of clustering support. In addition, the Solr management can also store in a distributed server on the mass oriented HTML data. This enables it to be oriented in massive HTML data on the search for the ideal solution, and build business intelligence system's important component.

V. SUMMARY

MapReduce especially Hadoop implementation provides a common server in distributed computing and powerful way. Coupled with the distributed storage and user friendly query mechanism, they form the SMAQ architecture allows mass oriented HTML data processing through a small team and individual development can achieve.

The emergence of Linux only by a table of the Linux server to the innovative developers strength. SMAQ has as much potential to improve the HTML data center efficiency, promote tissue from the periphery of the open innovation, cheap to create oriented HTML data driven business in the new era.

REFERENCES

[1] Paul Whitehead; Active Server Pages 3.0[M]; Tsinghua University press; First Edition; Beijing; 2010.8

[2] Instant ASP village Yonglong; examples of analytical ASP website programming [M]; Aerospace Press, Beijing Hope Electronic Press; First Edition; Beijing; 2010.1

[3] David Buser, John Kauffman; Asp 3 primary programming [M]; mechanical industry press; First Edition; Beijing; 2010.6

[4] Chen Yu, Meggie; software development tools [M]; Economic Science Press; First Edition; Beijing; 2006.9

[5] Active server pages Conan information; program design practical entry [M]; China Railway Publishing House Beijing 2010.4