Benchmark Testing for Transwarp Inceptor—A big data analysis system based on in-memory computing

Mingang Chen[1,2,a], Zhenqiang Chen[3, b] , Wanggen Liu[3, c] , Zhengyu Liu[1,2,d]

[1]Shanghai Key Laboratory of Computer Software Testing and Evaluating, Shanghai, 201112,China

[2]Shanghai Development Center of Computer Software Technology, Shanghai, 201112, China

[3]Transwarp Technologies (Shanghai) Co., Ltd., Shanghai, 200233, China

[a]cmg@ssc.stn.sh.cn, [b]zhenqiang.chen@transwarp.io, [c]wayne.liu@transwarp.io, [d]lzy@ssc.stn.sh.cn

Keywords: Benchmark testing; Transwarp Inceptor; Big data

Abstract. The high demand for data analysis and rapid development of big data technology and application has led to a variety of commercial and open source big data processing systems launched by industry and academia. So how to test and evaluate these systems objectively has become an important research topic. In this paper, we propose and develop an automated benchmark testing solution based on TPC-DS for Transwarp Inceptor, a big data analysis system. Test includes generating and loading a 500 GB data set and creating 24 tables and testing the performance of the system and SQL compatibility by executing 99 standard SQL queries. The test results can provide reference for enterprises to compare and choose big data analysis systems.

1. Introduction

With the rapid development of the traditional Internet, the mobile Internet, the Internet of Things, the cloud computing technology, and with the continuous improvement of the IT infrastructure, our world has entered big data era. Big data brings challenges in the aspects of data storage, the method and efficiency of data processing, meanwhile several big data management and analysis systems have appeared to tackle these challenges. In recent years, the industry and academia have launched a variety of commercial and open source versions of big data management and analysis systems, such as Apache Hadoop, Spark, Storm, Oracle Exadata, IBM big SQL and Transwarp Data Hub and so on.

Big data benchmark test is a practice for the evaluation of big data management and analysis systems. Generally a benchmark first abstracts representative workload from concrete applications, then generates a scalable test data set based on the characteristics and distributions of the data, and finally analyses the system's performance according to evaluation metrics. How to use benchmark test to objectively evaluate big data management and analysis systems has become an important research topic [1-2].

HiBench developed by Intel evaluates running time of MapReduce jobs, HDFS Bandwidth and throughput [3]. TPC-DS proposed by Transaction Processing Performance Council is a decision support system oriented benchmark. TPC-DS is widely used in evaluating the SQL query throughput of structured data, such as data warehouse [4-5]. BigBench is an end-to-end big data benchmark. The underlying business model of the BigBench is a product retailer. It enriches with semi-structured and unstructured data models, and the structured part of data model is adopted from the TPC-DS[6]. BigFrame is a benchmark generator, and can be used to generate user's own benchmarks fit to their special needs [7].

In this paper, we propose a benchmark testing solution for Transwarp Inceptor v4.0, a SQL on Hadoop system for big data analysis based on in-memory computing [8], and develop automated test scripts for the solution. To verify the scalability and big-data process capability, TPC-DS scale factor 500 is adopted as the benchmark. The benchmark testing solution includes generating and loading a 500 Gigabyte test data, creating 24 tables, and testing the performance of the system and SQL compatibility by executing 99 standard SQL queries.

2. The Architecture of Transwarp Inceptor

Transwarp Inceptor is a big data analysis system based on Apache Spark, which provides capabilities of high-speed SQL analysis. It can help users to build scalable decision supporting system (such as data warehouse), and perform interactive analysis and real-time reporting. Transwarp Inceptor has a three-tier structure from bottom to top: the storage layer, the distributed computing engine layer and the interface layer, as shown in Fig.1.

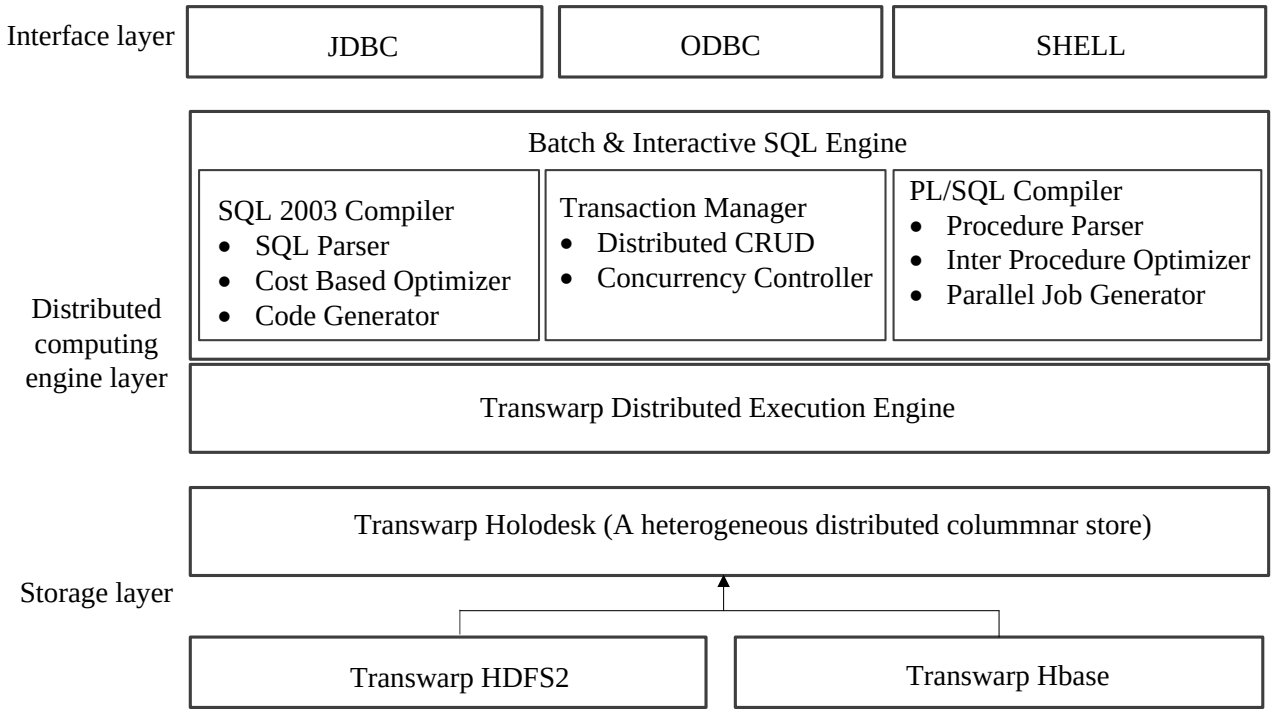

Fig.1. Architecture of Transwarp Inceptor

Transwarp Inceptor can handle data stored in HDFS, HBase or Transwarp Holodesk distributed cache (designed by Transwarp), and the amount of data that can be processed is from GB to PB. Transwarp Holodesk is a heterogeneous distributed columnar store used to cache data for Spark accesses, which is compatible with RAM, SSD and HDD. Transwarp Inceptor distributed computing engine is an in-memory execution engine based on Spark that contains the SQL 2003 compiler, PL / SQL compiler and the transaction manager to support data warehouse applications of complex analysis. The top layer of Transwarp Inceptor is JDBC, ODBC and Shell access interfaces to facilitate the migration from the existing traditional database systems to big data platform.

3. Testing design for Transwarp Inceptor

The overall process of Transwarp Inceptor system testing consists mainly of five stages, the generation of testing data and SQLs for query, the data loading, the creation of tables and data partition, the SQL queries execution and the analysis of testing, as shown in Fig.2. The test process is automatic executed by test scripts.

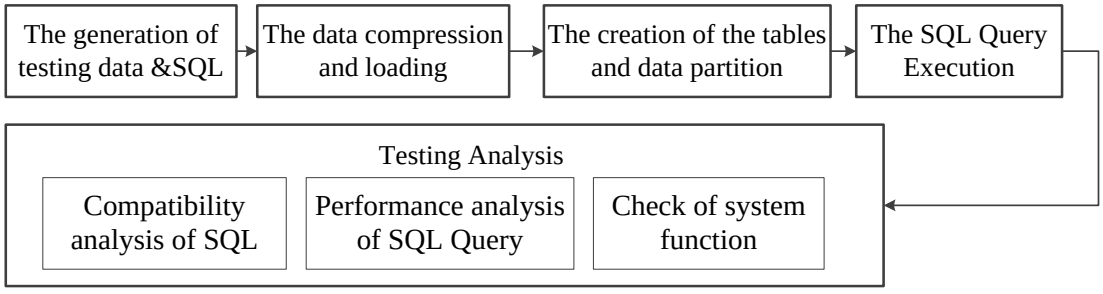

Fig.2. Testing process of Transwarp Inceptor

3.1 The generation of testing data and SQL

TPC-DS is a decision support benchmark that models several generally applicable aspects of a decision supporting system. The TPC-DS schema contains vital business information, such as customer, order, and product data. The benchmark models user's queries that are the most important component of any mature decision supporting system. User's queries convert operational facts into business intelligence. In this paper, DSTools v1.3.0 provided by TPC-DS benchmark is used to generate 500GB test data, 99 query SQLs, and the script fragments are as follows.

```
#Generate 500GB data in HDFS
1: hadoop dfs -mkdir -p LOCATION_HDFS
2: dbgen2 -scale 500 -dir LOCATION_HDFS

#Generate 99 SQL for query
3: qgen2 –query99.tpl –directory QUERY_TEMPLATE –dialect oracle –scale 500
```

The 500GB test data composes the business model of TPC-DS benchmark, which is stored in 7 fact tables and 17 dimension tables. The tables are organized with a star and snowflake mixed model in order to build a data warehouse for decision supporting. The business model of TPC-DS benchmark simulates sales, distributions and products returning in three primary channels: stores、catalogs and the web. As a whole the TPC-DS benchmark has the following characteristics:

1) A large amount of business data and test cases (SQL queries) can answer real business problems.

2) A total of 99 SQL queries follow the SQL 99 and SQL 2003 core syntax standard, and SQL queries are complex.

3) The test cases include a variety of business model, such as interactive query, statistical analysis, iterative OLAP and data mining.

4) Almost all of the test cases need high I/O loading and CPU computing.

3.2 The data compression and loading

For optimization, the testing for Transwarp Inceptor adopts ORC (Optimized Row Columnar) file format to store test data. The original 500GB text format test data is compressed to 150GB with ORC file format.

3.3 The creation of table and data partition

Next, we create 7 fact tables and 17 dimensional tables with the Transwarp Inceptor. The schemas of the tables are provided by the TPC-DS benchmark. We build a data warehouse for testing by loading the test data into the tables. Take 'store' table as an example, the following scripts illustrate how to create table and to load data into it.

```
#Create table store and load data into the table
1: create database if not exists ${DB};
2: use ${DB};
3: drop table if exists store;
4:
5: create table store
6: row format serde '${SERDE}'
7: stored as ${FILE}
8: as select * from ${SOURCE}.store;
```

Since the time span of the test data is more than ten years, Transwarp Inceptor partitions all fact tables with range partition according to date related columns to accelerate the queries. All dimension tables are free of any partition.

3.4 The SQL queries execution

The key point of the testing for Transwarp Inceptor is the execution of 99 standard queries provided by TPC-DS benchmark. We test the functional correctness and the query efficiency by executing SQL queries. According to the structure and semantics of SQL statements, 99 standard queries can be classified four classes: 9 interactive queries, 69 statistical analysis, 10 iterative

OLAPs and 11 data mining queries.

To minimize the impact of system cache on the SQL queries performance, 99 SQL are executed sequentially three rounds through automated scripts as shown in the following. The average execution time of each SQL query in three rounds is taken as the result bases on which we calculate the average execution time of each class of SQL according to the above classification.

```
# Execute all 99 SQL queries sequentially
1: for($i = 1; $i<=99; $i++ ){
2:      $sql = "query".$i.".sql";
3:      $cmd = "transwarp -t -h localhost    -f ./sql/$sql > ./logs/$sql.log 2";
4:}
5: Record the execution time of each SQL query through parsing logs of the SQL execution
```

4. Analysis of test results

Test environment is composed of a cluster of four same configured servers. The network topology of the cluster is shown in Figure 3 and the configuration of the server is shown in Table 1.

Table 1. The configuration of the servers

Model	Dell PowerEdge R720
CPUs	Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz　2 CPU x 6 core
Memory	256GB
Hard Disks	6×1T HDD hard drive，7200 rpm
Operating System	Red Hat Enterprise Linux 6.5

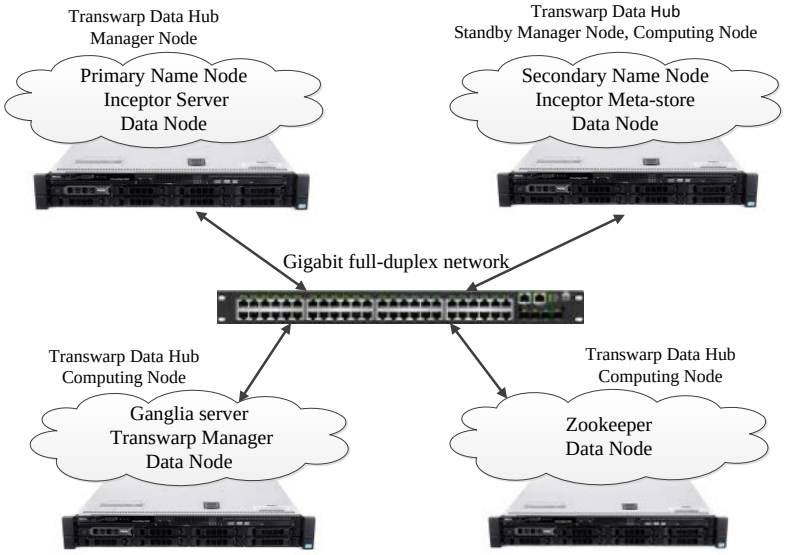

Fig.3. Network topology of Transwarp Inceptor cluster under test

In the aspect of functional testing, we focus on testing the application scenarios of Transwarp Inceptor. The test results show that Transwarp Inceptor can correctly implement data loading, table creation and partition, and complex SQL queries.

In the aspect of performance testing, we mainly test the execution time for Inceptor Transwarp to perform 99 standard SQL queries. In the case of the 500GB data and the fact table size of 18-1440 million records, the four types of SQL execution time are shown in Table 2.

Table 2. The four types of SQL execution time

The number of SQL	SQL Categories	The total execution time (seconds)	The average execution time (seconds)
9	Interactive query	197	21.9
69	Statistical analysis	7705	111.7
10	Iterative OLAP	4232	423.2
11	Data mining	3502	318.4

As compatibility test result, only 3 SQLs require minor modification and other 96 SQLs are in accordance with SQL99 and SQL 2003 core SQL standards. But according to TPC-DS benchmark specification [5], such minor modification is acceptable.

5. Conclusion

In this paper, an automated benchmark testing solution is designed and developed based on TPC-DS for Transwarp Inceptor, a SQL on Hadoop big data analysis system. This solution aims to test the functionality, performance and compatibility of complex SQL queries of Transwarp Inceptor. This benchmark testing solution can also be applied to other system of SQL on Hadoop, such as Apache Hive and Cloudera Impala. In further studies, we plan to test multiple SQL on Hadoop systems, and compare the system's functionality and performance fairly. These test results will provide the reference for enterprises to choose an appropriate big data analysis system.

Acknowledgement

This work was funded by Science and Technology Commission of Shanghai Municipality Program (15511107003, 14511106804).

References

[1] Han R, Lu X, Xu J. On Big Data Benchmarking. Lecture Notes in Computer Science, 2014, 8807:3-18.

[2] Barata M, Bernardino J, Furtado P. Survey on Big Data and Decision Support Benchmarks. Database and Expert Systems Applications. Springer International Publishing, 2014:174-182.

[3] Huang S, Huang J, Dai J, et al. The hibench benchmark suite: Characterization of the mapreduce-based data analysis. ICDE Workshops, 2010, 74:41 - 51.

[4] Poess M, Nambiar R O, Walrath D. Why You Should Run TPC-DS: A Workload Analysis. Proceedings of the 3rd international conference on Very large data bases. VLDB Endowment, 2007:1138-1149.

[5] TPC-DS specification 2015. http://www.tpc.org/tpcds/

[6] Ghazal A, Rabl T, Hu M, et al. BigBench: towards an industry standard benchmark for big data analytics. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013:1197-1208.

[7] BigFrame User Guide. https://github.com/bigframeteam/BigFrame/wiki/

[8] Transwarp Inceptor. http://www.transwarp.cn/product/inceptor?lang=en