Script-based Automation ETL Tool

Ji-zhe Li[1,a],Bei-jing Kuang[1,b],Jin-gang liu[1,c]

[1]College of Information Engineering,Capital Normal University,Beijing 100048,China
[a] yishengxiaoyao89@126.com, [b]805071937@qq.com,[c]liujg2000@163.com

Keywords: ETL;Sqoop;Hadoop;HDFS;Csv

Abstract: Due to the shortcoming of the traditional ETL tool and Sqoop, we proposed a script-based automation ETL tool. We proposed an automation ETL tool which is combined with scripting and database tools to implement the 3 ETL processes-Extraction, Transformation and Loading.At Last,we have proved that script-based automation ETL tool is faster than snoop on extracting and transforming data from Oracle to Impala.

1. Introduction

With the rapid development of science and technology, economy and society, we have stepped into the information era. We are surrounded by different kinds of information in our daily life. Thus,huge amount of data is generated with various format. The ways to store data, analyze data and find useful information become important issues. Cloud Computing technologies are developed for those issues, for example, Google File System(GFS)[1], MapReduce framework[2] and Big Table[6] proposed by Google in 2003,2004 and 2006 for distributed file system, parallel and distributed computing, and NoSQL database, respectively[3].

After GFS, MapReduce and BigTable were published, they are not open source. The Apache Hadoop[4,14] project develops open-source software for reliable, scalable, distributed computing. HDFS and MapReduce are the core part of the Hadoop project.

Most enterprises still use relational databases for business. However, as more and more data produced, relational database lacks the ability to handle such size of data[3]. ETL tools save time and money when developing a data warehouse by removing the need for "hand coding".

This rest of this paper is organized as follows. Section 2 introduces some traditional and open-source ETL tools. Section 3 introduces how script-based automation ETL tools work and its composition. Section 4 gives the performance evaluation. Finally, section 5 concludes this paper.

2. Relational Work

2.1 ETL

ETL process in data warehouse development performs data extraction from various resources, transform the data into suitable format and load it into data warehouse storage and play an important role in data migration[5].An simple ETL tools consists of Extract, Transform and Load as the Figure 1 show. The purposes of the data extract process is to collect useful data from multiple heterogeneous data sources. Data transformation is the process of transformation data from organizational forms and formats into a uniform format. The data loading process is responsible for loading the data processed by the refreshing and updating into DW.

2.2 ETL's different

Existing ETL tools are divided into two categories: homogenous mode and heterogeneous mode.

Homogenous mode is to use the ETL tool to exchange data between relational databases. For example, PowerCenter, DataStage, OWB,SSIS, DTS. Heterogeneous mode is to use the ETL tool to exchange data between relational database and NoSQL Database, such as, Kettle, Talend, Crunch, Sqoop. Differences between the distribute ETL tools(Sqoop) and traditional tools as the table below shows. Focusing on implementation and requiring high performance on the sever it the shortcoming of traditional ETL tools.

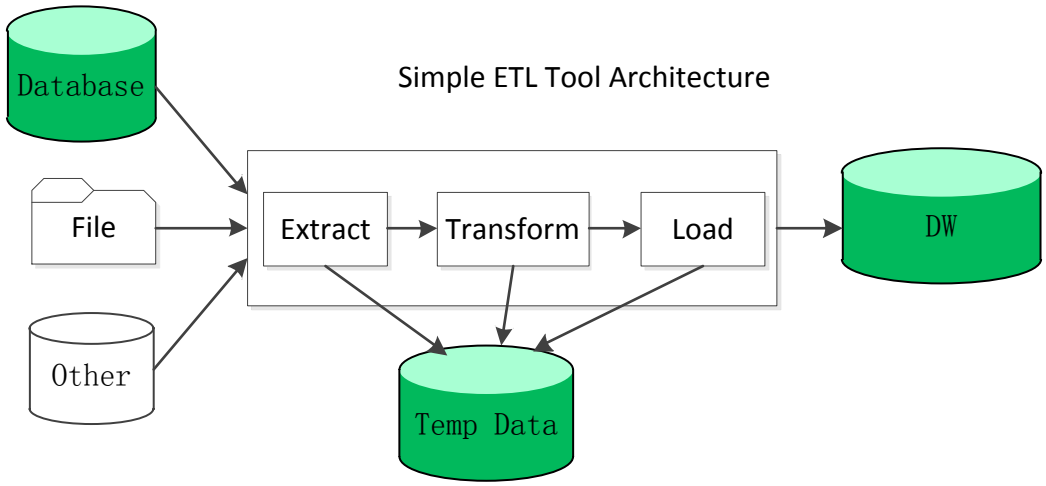

Figure 1 Simple ETL Tool Architecture

Table 1 Sqoop compare with Traditional ETL Tool

	Sqoop	Traditional ETL Tool
Introduction	Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases[7].	Data extraction from various resources, transform the data into a suitable format and load it into data warehouse storage.
Data Extract Feature	Scoop establishes connection with a relational database via JDBC.	Server for relational databases.
Integration with Hadoop	Sqoop started as a contrib module, for Apache Hadoop in May of 2009.Sqoop is compatible with Hive and HBase.	Some work to be done: programing directed to specific needs
Fault Tolerant	Hadoop gives a hand to sqoop in fault tolerant.	Friendly Web UI, Strong Fault tolerant, Location the place where the problem happened.
Mode	Distributed	Centralized(single node)
Price	Free	Every year

3.Script-Based Automation ETL tool

This section describes the framework model for ETL processes. First, we design the ETL logical framework. Then, we should decide the format which the temporary stored as. Last, we choose some proper tools to accomplish the model and further realize the optimization management of ETL processes.

3.1Scripted-Based Automation ETL Tool

A) ETL Framework Design

The Logical framework is shown in Figure 2.

During the 3 ETL processes-extraction, transformation and loading, we combine shell script with database import and export tool which depends on the database to carry out those processes.

B) Choose the intermediate files

According to [8,9,10,11], they choose xml file or excel file as intermediate files. There are some reasons why we choose the csv file as an intermediate file as the below shows.

1) The high efficiency of writing data to csv file;

2) The construction of data does not change;

3) Csv file generated by the same amount of data is smaller than other format and use little memory.

C) Choose the proper tools.

We choose import/export tool depending on the database. For example, if the data source is Oracle, we will choose the sqluldr2 for exporting data and sqlldr for importing data.

D) ETL tool implements

Script-based automation ETL tool consists of importing data from data sources to data warehouse and exporting data from data warehouse to another format. We illustrated by the example of exchanging data between oracle and HDFS.

Oracle->HDFS: Exporting data from oracle to local and generating csv file, uploading csv file to HDFS and complete impala mapping.

HDFS->Oracle: Impala quickly scan it's metastore server to find the location of data file and turn it into csv files and use sqlldr to import data to oracle database.

Exporting data and importing data as the Figure 3 below shows.

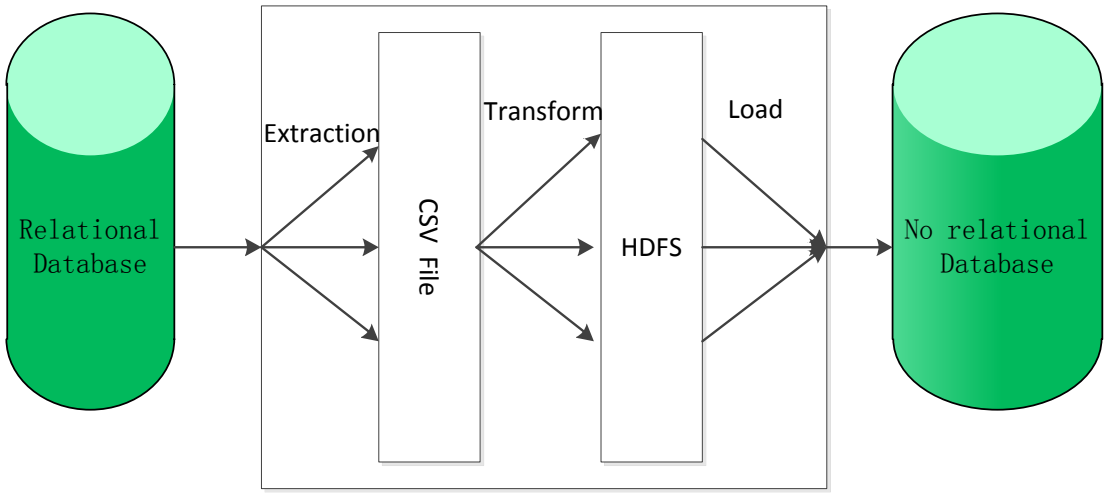

Figure 2 ETL logical Framework

4.Experimental Performance

4.1 Environment

Experimental Environment consists of 5 nodes. A Hadoop cluster consists of 4 nodes, which haves installed impala and Sqoop. The last servers is ready for installing oracle database.

4.2Dataset

The experimental data are randomly generated by a program which we wrote in Java, each record contains all of the Java basic data types. The definition of the table as the below shows:

create table test(name String,age int,sex bit, averagegrade float,address nvarchar,career varchar,birthday date,description nvarchar)

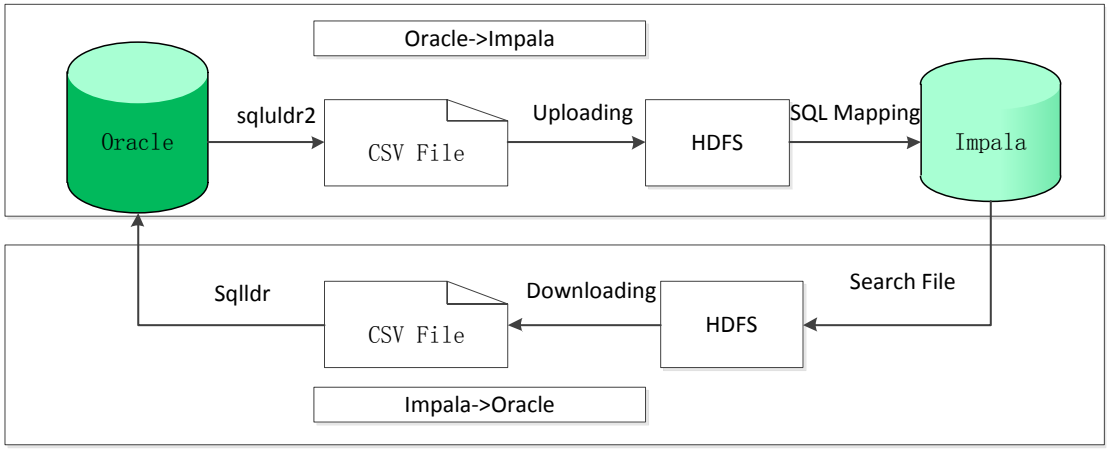

Figure 3 export and import data

4.3Experimental Step

Importing data from oracle to Impala[12,13] as the below steps:

Step 1: Exporting data from oracle to a csv file with the help of sqluldr2.

sqluldr2.bin USER=userid/keyword@db_name query=select * from test head=no FILE=/oracle/test.csv

Step 2:Uploading csv file to HDFS.

hadoop fs -put /oracle/test.csv /user/impala

Step 3:Data Mapping

impala-shell -f "show databases;use impala;show tables;create table test(name String,age int,sex bit, averagegrade float,address nvarchar,career varchar,birthday date,description nvarchar)row format delimited fields terminated by ',' STORED AS TEXTFILE location '/user/impala/';create table IF NOT EXISTS test_Parquet like test stored as parquet;insert into test_Parquet select * from test;drop table test;"

4.4Performance Analysis

The result of the experimental as the table 2 shows:

Table 2 Experimental Results

	Sqoop			Script-based automation ETL tools		
Line NO.	time(s)	size	speed	Time(s)	size	speed
10	17	1.03K	0.06K/s	4.32	1K	0.248K/s
10^2	17	12K	0.71K/s	4.32	11K	2.55K/s
10^3	17.08	115K	0.0067M/s	4.32	111.2K	0.025M/s
10^4	16	1.12M	0.069M/s	3.7	1.11M	0.03M/s
10^5	17.9	11M	0.643M/s	4.9	11M	2.24M/s
10^6	23	112M	4.7M/s	9.1	109M	11.9M/s
10^7	77	1.09G	14M/s	47	1.08G	22M/s
10^8	520	10G	19M/s	352	11G	31.24M/s

5. Conclusion

This paper introduces the concept of ETL tools as well as its key technology. Due to the

shortcoming of the traditional ETL tool with focusing on the implementation and requiring high performance on the server, the traditional ETL tool cannot match up with the requirement of the nowadays. And Sqoop assign the task unreasonable. We proposed a script-based automation ETL tool. But the script-based automation ETL tool has a lot of work to be done. For example, data verification, resume from break-point, data rollback, which is the next step we should do.

References

[1] Sven Groot, "Jumbo: Beyond MapReduce for Workload Balancing," *Fuzzy Systems and Knowledge Discovery (FSKD),2011 Eighth International Conference on Cloud Computing Technology and Science*, vol. 4, pp.2675-2678, 2011. July.

[2] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters, " *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] Hsu, J. C., Hsu, C. H., Chen, S. C., & Chung, Y. C."Correlation Aware Technique for SQL to NoSQL Transformation. *International Conference on Ubi-Media Computing and Workshops* (Vol.87, pp.395-415),2014.

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

[5] R Wijaya,B Pudjoatmodjo,"An overview and implement of extraction-transformation-loading(ETL) process in data warehouse",*3th International Conference on Information & Communication Technology,*2015.

[6] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C., Hsieh Deborah A., Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, "Bigtable: A Distributed Storage System for Structured Data," *7th UENIX Symposium on Operating Systems Design and Implementation*, pp. 205-218, 2006.

[7] http://sqoop.apache.org/

[8] Dai Li, Li Xiao-Yan, Sun Liang,"Research on the Data Transformation Technology between XML and Relational Database", *Journal of Chongqing University of Science and Technology(Natural Sciences Edition)*,vol 12(6),pp.169-173,2010.

[9] Geng Biao,Song Yuqing,Liang Quanmei,Chen Jianmei,"Research on mapping method from XML document to relational database",*Application Research of Computers*,vol 27(3),p.951-954,2010.

[10] Zhao Yanni,Gao Hualei,"On XML-Based Data Migration Technology in Information System Upgrade and Its Implementation",*Computer Application and Software*,vol 31(12),p.52-54,2014.

[11] Ge Zhongze,"Research and Implementation of MVC Pattern-Based Excel Data Migration Frame",*Computer Application and Software*,vol 31(1),p.22-25,2014.

[12] Michael," Impala:A Modern,Open-Source SQL Engine for Hadoop",*7th Biennial Conference on Innovative Data Systems Research*,2015.

[13] http://www.cloudera.com/products/apache-hadoop/impala.html

[14] Tom White,"Hadoop:the Definitive Giude, 4rd Edition",(O'Reilly Media Publish,USA,2015)

