ISSN:0975-9646

Gaurav Vaswani et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1255-1258

Sandbox- An Application Tool for Hadoop

Gaurav Vaswani , Ajay Chotrani , Hitesh Rajpal

Student of Computer Engineering,
VESIT, Mumbai

Abstract-The Hortonworks sandbox is a fully contained Hortonworks Data Platform (HDP) environment. The sandbox includes the core Hadoop components (HDFS and MapReduce), as well as all the tools needed for data ingestion and processing. You can access and analyze sandbox data with many Business Intelligence (BI) applications.
The Hortonworks Sandbox is a single node implementation of the Hortonworks Data Platform(HDP). It is packaged as a virtual machine to make evaluation and experimentation with HDP fast and easy. The idea is to show you how you can get started and show you how to accomplish tasks in HDP. To compare the advantage of working on Hadoop and case study to compare the files in java and Hadoop on Sandbox platform with respect to size.

Keywords: ***Sandbox, HDFS, Apache Pig, Hive,Hcat.***

INTRODUCTION :

HORTONWORKS DATA PLATFORM

The Apache Hadoop projects provide a series of tools designed to solve big data problems. The Hadoop cluster implements a parallel computing cluster using inexpensive commodity hardware. The cluster is partitioned across many servers to provide a near linear scalability. The philosophy of the cluster design is to bring the computing to the data. So each datanode will hold part of the overall data and be able to process the data that it holds. The overall framework for the processing software is called MapReduce. Here's a short video introduction to MapReduce. To instal hortonworks sandbox the requirement is to install the Virtual box and hortonworks sandbox. The basic need is the machine should have the virtualization.

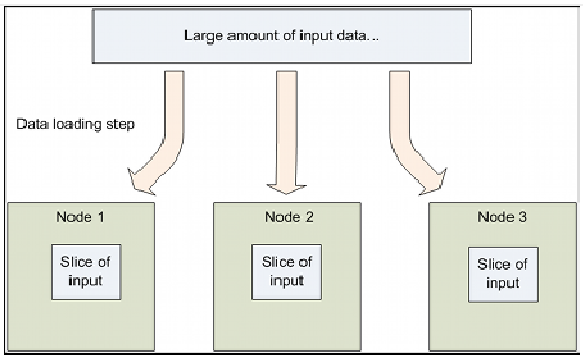

Figure 1

INTRODUCTION TO MAPREDUCE

Apache Hadoop can be useful across a range of use cases spanning virtually every vertical industry. It is becoming popular anywhere that you need to store, process, and analyze large volumes of data.

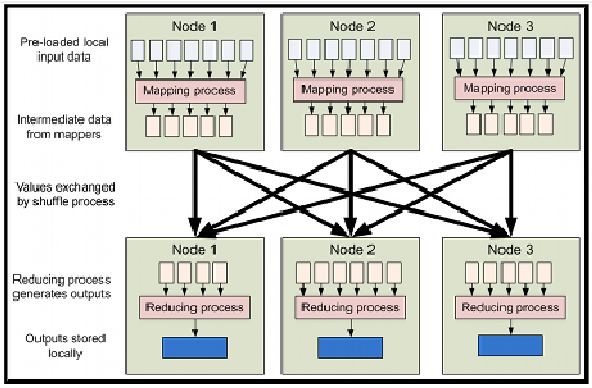

Figure 2

Examples include digital marketing automation, fraud detection and prevention, social network and relationship analysis, predictive modeling for new drugs, retail in-store behavior analysis, and mobile device location-based marketing.

HADOOP DISTRIBUTED FILE SYSTEM

Underlying all of these components is the Hadoop Distributed File System(HDFS™). This is the foundation of the Hadoop cluster.

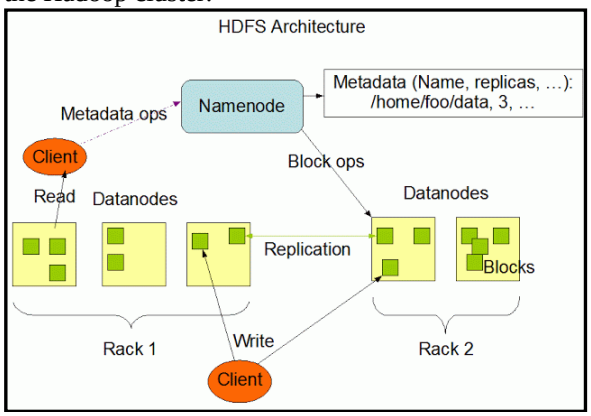

Figure 3

The HDFS file system manages how the datasets are stored in the Hadoop cluster. It is responsible for distributing the data across the datanodes, managing replication for redundancy and administrative tasks like adding, removing and recovery of datanodes.

DATA PROCESSING WITH PIG

Pig is a high level scripting language that is used with Apache Hadoop. Pig excels at describing data analysis problems as data flows. Pig is complete in that you can do all the required data manipulations in Apache Hadoop with Pig. In addition through the User Defined Functions(UDF) facility in Pig you can have Pig invoke code in many

languages like JRuby, Jython and Java. Conversely you can execute Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.
Pig can ingest data from files, streams or other sources using the User Defined Functions(UDF)

Figure 4

Pig is a language for expressing data analysis and infrastructure processes. Pig is translated into a series of MapReduce jobs that are run by the Hadoop cluster. Pig is extensible through user-defined functions that can be written in Java and other languages. Pig scripts provide a high level language to create the MapReduce jobs needed to process data in a Hadoop cluster.
That‘s all for now… let‘s get started with some examples of using these tools together to solve real problems!

DATA PROCESSING WITH HIVE

Hive is a component of Hortonworks Data Platform (HDP). Hive provides a SQL-like interface to data stored in HDP. In the previous tutorial we used Pig which is a scripting language with a focus on dataflows. Hive provides a database query interface to Apache Hadoop.
People often ask why do Pig and Hive exist when they seem to do much of the same thing. Hive because of its SQL like query language is often used as the interface to an Apache Hadoop based data warehouse. Hive is considered friendlier and more familiar to users who are used to using SQL for querying data. Pig fits in through its data flow strengths where it takes on the tasks of bringing data into Apache Hadoop and working with it to get it into the form for querying. From a technical point of view both Pig and Hive are feature complete so you can do tasks in either tool. However you will find one tool or the other will be preferred by the different groups that have to use Apache Hadoop. The good part is they have a choice and both tools work together.

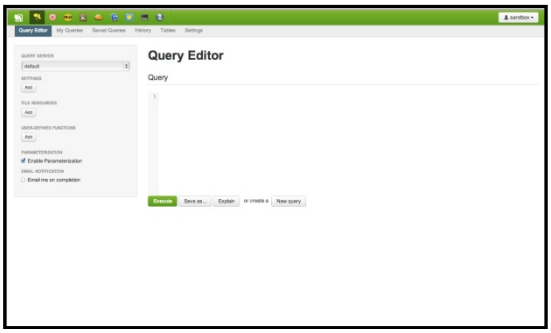

Figure 5

The Apache Hive project provides a data warehouse view of the data in HDFS. Using a SQL-like language Hive lets you create summarizations of your data, perform ad-hoc queries, and analysis of large datasets in the Hadoop cluster. The overall approach with Hive is to project a table structure on the dataset and then manipulate it with HiveQL. Since you are using data in HDFS your operations can be scaled across all the datanodes and you can manipulate huge datasets.

DATA PROCESSING WITH HCATALOG

The function of HCatalog is to hold location and metadata about the data in a Hadoop cluster. This allows scripts and MapReduce jobs to be decoupled from data location and metadata like the schema. Additionally since HCatalog supports many tools, like Hive and Pig, the location and metadata can be shared between tools. Using the open APIs of HCatalog other tools like Teradata Aster can also use the location and metadata in HCatalog. In the tutorials we will see how we can now reference data by name and we can inherit the location and metadata.

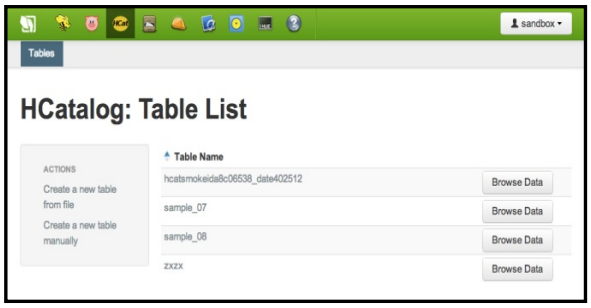

Figure 6

USING FILE BROWSER

You can reach the File Browser by clicking its icon:

Figure 7

The File Browser interface should be familiar to you as it is similar to the file manager on a Windows PC or Mac. We begin in our home directory. This is where we'll store the results of our work. File Browser also lets us upload files.

UPLOADING THE FILE

To upload the example data we create the text file to be uploaded and go to the file browser as shown in the Figure 8.

Figure 8

- Select the 'Upload' button
- Select 'Files' and a pop-up window will appear.
- Click the button which says, 'Upload a file'.
- Locate the example data file you downloaded and select it.
- A progress meter will appear. The upload may take a few moments.

When it is complete you'll see this:

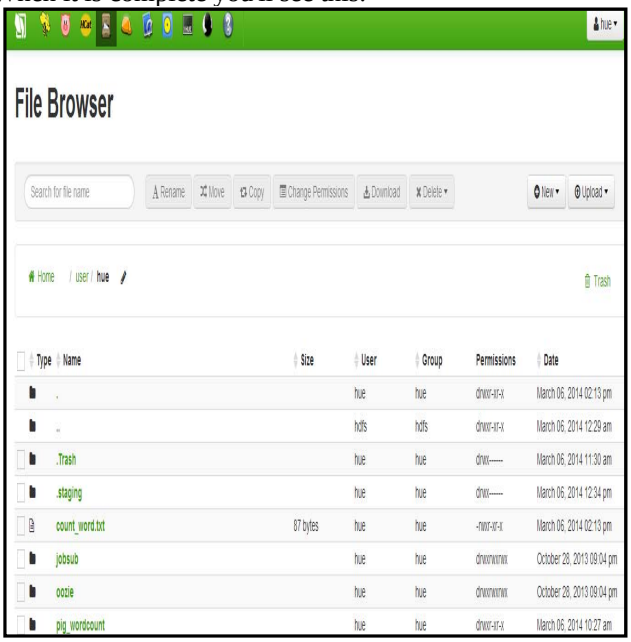

Figure 9

CASE STUDY OF WORD COUNT IN SANDBOX PIG

Word count is the simple case study to compare how the data size is reduced and the memory used when we do word count in a Java program and when we run the application in Hortonworks Sandbox.

Steps to load the word cound program in Pig.

1. Given below is the code to be executed for the word count to be put in the Query editor.

```
a = load '/user/hue/word_count_text.txt';
b = foreach a generate
flatten(TOKENIZE((chararray)$0)) as word;
c = group b by word;
d = foreach c generate COUNT(b), group;
store d into '/user/hue/pig_countword';
```

2. The size of the text file is 68 bytes.
3. The output is stored in the database.
4. The same program when executed in java takes 1.04 KB of space.
5. The query to be written for the excecution of the above code to display in the database is shown in Figure 10
6. After the query is executed the database results are as shown in the Figure 11.

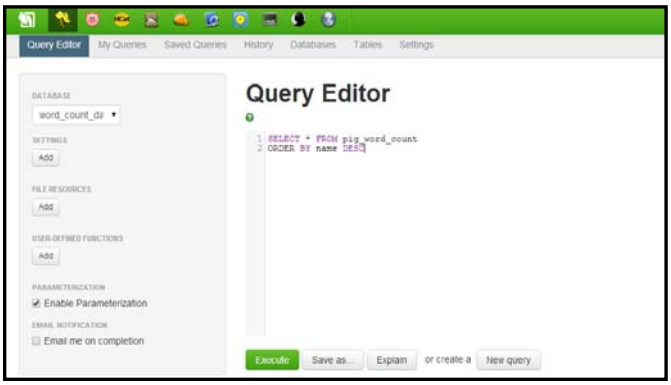

Figure 10

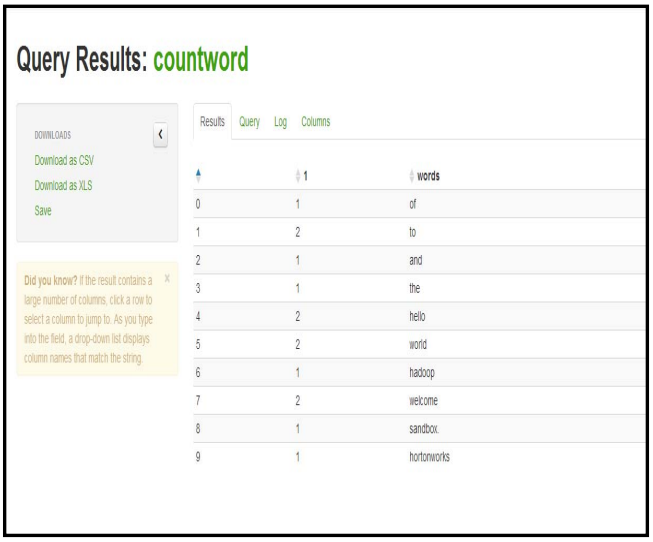

Figure 11

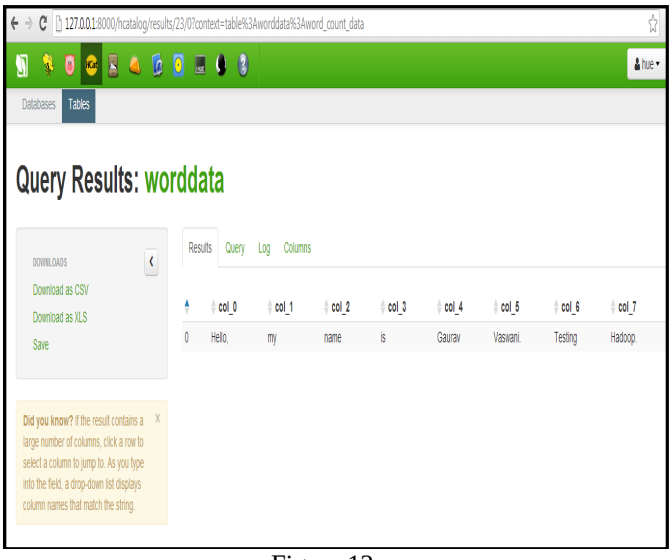

Figure 12

CONCLUSION

To working environment of the hortonworks sandbox gives ease to work with the data base and creating tables and the database. The comparitive study shows that the memory used and the data size of the output file in java is much larger as compare to that in hadoop using sandbox.

We have taken a very small case study to show the effectiveness of the implementation tool using hortonworks sandbox.

The future scope is to work on the large databases which take up large amount of space , and can be reduced wusing this tool.

REFERENCES

1. http://hortonworks.com/hadoop-tutorial/ hello- world-an-introduction-to-hadoop-hcatalog-hive-and-pig/
2. http://hortonworks.com/wp-content/uploads /unversioned/pdfs/InstallingHortonworksSandbox2onWindowsusing VB.pdf
3. http://hortonworks.com/hadoop-tutorial/ loading-data-into-the-hortonworks-sandbox/ how to load data.
4. http://www.javacodegeeks.com/2011/05/hadoop-soft-introduction.html.
5. http://hortonworks.com/hadoop-tutorial/ loading-data-into-the-hortonworks-sandbox/