ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Big Data Analysis: Comparision of Hadoop MapReduce, Pig and Hive

Dr. Urmila R. Pol

Assistant Professor, Department of Computer Science, Shivaji University, Kolhapur, India

ABSTRACT: Big data has fast developed into a hot topic that attracts extensive attention from academia, industry, and governments around the globe. Hadoop is a framework that allows for the distributed processing of big data sets across clusters of computers using simple programming models. Hadoop created as a faster alternative to RDBMS. To process large amount of data at a very fast rate, which earlier took a lot of time in RDBMS. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and eventually our society itself. The traditional data analytics may not be able to handle such large quantities of data. In this paper, we present a study of Big Data and its analysis using Hadoop mapreduce, pig and hive.

KEYWORDS: Big Data, Hadoop, mapreduce,pig,hive.

I. INTRODUCTION TO BIG DATA

Big Data is not just about lots of data, it is actually a concept providing an opportunity to find new insight into your existing data as well guidelines to capture and analysis your future data. It makes any business more agile and robust so it can adapt and overcome business challenges. There are four main modules in Hadoop.

Hadoop Common: The common utilities that support the other Hadoop modules.

1.Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.

2.Hadoop YARN: A framework for job scheduling and cluster resource management.

Big data' could be found in three forms:

- •Structured: Structured data has strong schema and schema will be checked during write & read operation. e.g. Data in RDBMS systems like Oracle, MySQL Server etc.
- •Unstructured: Data does not have any structure and it can be any form - Web server logs, E-Mail, Images etc.
- •Semi-structured: Data is not strictly structured, but have some structure. e.g. XML files.

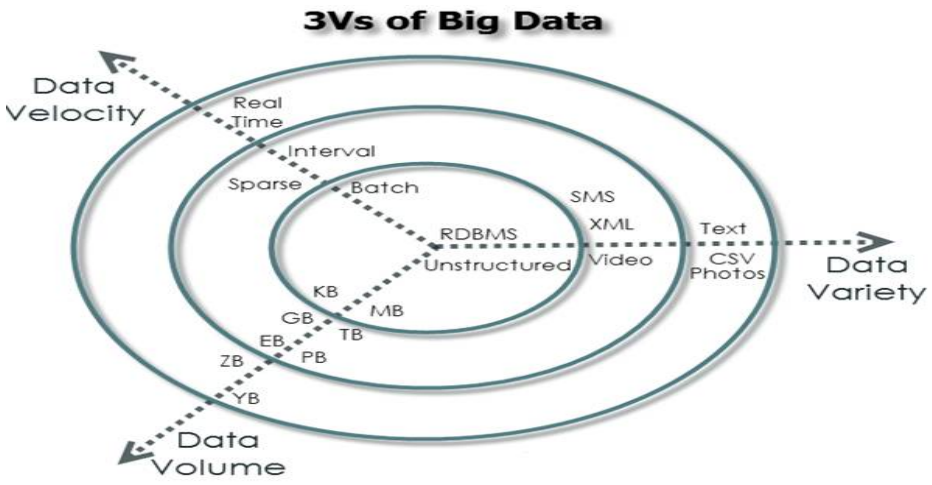
This computational logic is nothing but a compiled version of a program written in a high level language such as Java. Such a program, processes the data stored in Hadoop HDFS. HADOOP is an open source software framework. Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost. A computer cluster consists of a set of multiple processing units (storage disk + processor) which are connected to each other and act as a single system. Much of the tech industry follows Gartner's '3Vs' model to define Big Data.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016



Volume

We currently see the exponential growth in the data storage as the data is now more than text data. We can find data in the format of videos, musics and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises. As the database grows the applications and architecture built to support the data needs to be reevaluated quite often. Sometimes the same data is re-evaluated with multiple angles and even though the original data is the same the new found intelligence creates explosion of the data. The big volume indeed represents Big Data.

Velocity

The data growth and social media explosion have changed how we look at the data. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast we receive the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data.

Variety

Data can be stored in multiple format. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. Sometimes the data is not even in the traditional format as we assume, it may be in the form of video, SMS, pdf or something we might have not thought about it. It is the need of the organization to arrange it and make it meaningful. It will be easy to do so if we have data in the same format, however, it is not the case most of the time. The real world have data in many different formats and that is the challenge we need to overcome with the Big Data. This variety of the data represent represent Big Data.

II. HADOOP MAPREDUCE

Hadoop MapReduce is a software framework for easy writing applications which process huge amounts of data (multi-terabyte datasets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. A MapReduce *job* usually splits the input dataset into independent chunks which are processed by the *map tasks* in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

the *reduce tasks*. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. Typically the compute nodes and the storage nodes are the same, that is, the MapReduce framework and the Hadoop Distributed File System are running on the same set of nodes configuration allows the framework to effectively schedule tasks on the nodes where the data have been already present, resulting in very high total bandwidth across the cluster. The MapReduce framework consists of a single master `ResourceManager`, one slave `Node Manager` per cluster-node, and `MRAppMaster` per application.Minimally, applications specify the input/output locations and supply *map* and *reduce* functions via implementations of appropriate interfaces and/or abstract-classes. These, and other job parameters,include the *job configuration*.The Hadoop *job client* then submits the job (jar/executable etc.) and configuration to the `ResourceManager` which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

- Each Map task outputs the data in Key and Value pair.
- The output is stored in a CIRCULAR BUFFER instead of writing to disk.
- The size of the circular buffer is around 100 MB. If the circular buffer is 80% full by default, then the data will be spilled to disk, which are called shuffle spill files.
- On a particular node, many map tasks are run as a result many spill files are created. Hadoop merges all the spill files, on a particular node, into one big file which is SORTED and PARTITIONED based on number of reducers.

Although the Hadoop framework is implemented in Java™, MapReduce applications need not be written in Java.

- Hadoop Streaming is a utility which allows users to create and run jobs with any executables (e.g. shell utilities) as the mapper and/or the reducer.
- Hadoop Pipes is a SWIG-compatible C++ API to implement MapReduce applications (non JNI™ based).

Hadoop is typically used for storing(hdfs) and summarizing(map-reduce) large batches of data. MapReduce permits you to filter and aggregate data from HDFS so that you can gain insights from the big data. However, writing MapReduce code with basic Java may require you to write many lines of code laboriously, with additional time needed for code review and QA. So instead of writing plain Java code to use MapReduce, you now have the options of using either the Pig Latin or Hive SQL languages to construct .The benefit is that you only need to write much fewer lines of code, thus reducing overall development and testing time. The rule of thumb is that writing Pig scripts takes 5% of the time compared to writing MapReduce programs in Java, while reducing runtime performance by only 50%. Although Pig and Hive scripts generally don't run as fast as native Java MapReduce programs, they are vastly superior in boosting productivity for data engineers and analysts.

MapReduce Example

WordCount is a simple application that counts the number of occurrences of each word in a given input set. WordCount programme is available on https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Source_Code.[1]

Copy code into java file. Execute following commands.

SET ENVIRONMENT VARIABLES

export JAVA_HOME=/usr/java/default
export PATH=urp@localhost${JAVA_HOME}/bin:urp@localhost${PATH}
export HADOOP_CLASSPATH=urp@localhost${JAVA_HOME}/lib/tools.jar
Compile `WordCount.java` and create a jar:
urpol@localhost$ bin/hadoop com.sun.tools.javac.Main WordCount.java
urpol@localhost$ jar cf wc.jar WordCount*.class
Assuming that:

- /user/urpol/wordcount/input - input directory in HDFS

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

- /user/urpol/wordcount/output - output directory in HDFS

Sample text-files as input:
urpol@localhost$ bin/hadoop fs -ls /user/urpol/wordcount/input/ /user/urpol/wordcount/input/file01
/user/urpol/wordcount/input/file02

urpol@localhost$ bin/hadoop fs -cat /user/joe/wordcount/input/file01

Hello Spark Welcome Spark
urpol@localhost$ bin/hadoop fs -cat /user/joe/wordcount/input/file02
Hello Spark Goodbye Hadoop
Run the application:
urp@localhost$ bin/hadoop jar wc.jar WordCount /user/urp/wordcount/input /user/urp/wordcount/output
Output:urp@localhost$ bin/hadoop fs -cat /user/urp/wordcount/output/part-r-00000`
Goodbye 1
Spark 3
Hadoop 1
Hello 2
Welcome 1

III. INTRODUCTION TO PIG

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient propertyof Pig programs is that their structure is amenable to substantial parallelization .An application that creates map-reduce jobs based on a language called Pig Latin, which is workflow driven. See It was originally created at Yahoo! (company).Apache Pig is good for structured data too, but its advantage is the ability to work with BAGs of data (all rows that are grouped on a key), it is simpler to implement things like:[1]

1. Get top N elements for each group;
2. Calculate total per each group and than put that total against each row in the group;
3. Use Bloom filters for JOIN optimisations;
4. Multiquery support (it is when PIG tries to minimise the number on MapReduce Jobs by doing more stuff in a single Job)
5. Hive is better suited for ad-hoc queries, but its main advantage is that it has engine that stores and partitions data. But its tables can be read from Pig or Standard MapReduce.
6. One more thing, Hive and Pig are not well suited to work with hierarchical data.

Pig integration with streaming also makes it easy for researchers to take a Perl or Python script they have already debugged on a small data set and run it against a huge data set. **PIG** is best for ***semi structured data,*** for ***programming,*** used as ***procedural language,*** does not support partitions, don't have dedicated metadata of database.[1]

Pig Installation
Linux users need the following:

1. **Hadoop 0.20.2** - http://hadoop.apache.org/common/releases.html
2. **Java 1.6** - http://java.sun.com/javase/downloads/index.jsp (set JAVA_HOME to the root of your Java installation)
3. **Ant 1.7** - http://ant.apache.org/ (optional, for builds)
4. **JUnit 4.5** - http://junit.sourceforge.net/ (optional, for unit tests)

To get a Pig distribution, download a recent stable release from one of the Apache Download Mirrors (http://pig.apache.org/releases.html).
Unpack the downloaded Pig distribution. The Pig script is located in the bin directory .
Add /pig-n.n.n/bin to your path. Use export (bash,sh,ksh) or setenv (tcsh,csh). For example:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

urp@localhost$ export PATH=/usr/local/pig-0.14.0/bin:$PATH
Try the following command, to get a list of Pig commands:
urp@localhost$ pig -help
Try the following command, to start the Grunt shell:
urp@localhost$ pig

1) Run Modes

Pig has two run modes or exectypes:

- Local Mode - To run Pig in local mode, you need access to a single machine.
- Mapreduce Mode - To run Pig in mapreduce mode, you need access to a Hadoop cluster and HDFS installation. Pig will automatically allocate and deallocate a 15-node cluster.

You can run the Grunt shell, Pig scripts, or embedded programs using either mode.

2) Grunt Shell

Local Mode

urp@localhost$ pig -x local

Mapreduce Mode

urp@localhost$ pig
or
urp@localhost$ pig -x mapreduce

For either mode, the Grunt shell is invoked and you can enter commands at the prompt. The results are displayed to your terminal screen (if DUMP is used) or to a file (if STORE is used).

Wordcount example for pig

```
grunt>a= LOAD '/piginput/pigexample.txt' USING PigStorage as (word:chararray);
grunt> dump a;
grunt>words= FOREACH a GENERATE FLATTEN(TOKENIZE(word));
grunt>dump word;
grunts> grouped= GROUP words by $0;
grunt>dump grouped;
grunt>word_counts= FOREACH grouped GENERATE group, COUNT(words);
grunt>store word_counts into '/pigoutput/word_count1' USING PigStorage;
```

III.INTRODUCTION TO HIVE

Although Pig can be quite a powerful and simple language to use, the downside is that it's something new to learn and master. Some folks at Facebook developed a runtime Hadoop® support structure that allows anyone who is already fluent with SQL (which is commonplace for relational data-base developers) to leverage the Hadoop platform right out of the gate. Hive is a platform used to develop SQL type scripts to do MapReduce operations .Facebook promotes the Hive language. Hive because of its SQL like query language is often used as the interface to an Apache Hadoop based data warehouse. **Hive**™, allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL statements; now you should be aware that HQL is limited in the commands it understands, but it is still pretty useful. HQL statements are broken down by the Hive service into MapReduce jobs and executed across a Hadoop cluster. It stores schema in a database and processed data into HDFS. It is designed for OLAP. It provides SQL type language for querying called HiveQL or HQL. It is familiar, fast, scalable, and extensible.

Intsllation Of HIVE

Hadoop must be installed on your system before installing Hive. Let us verify the Hadoop installation using the following command:
$ hadoop version // It will display hadoop version information otherwise first you have to install hadoop.
Download Hive and rename its folder to /usr/local/hive
urp@localhost$ cd /home/user/Download
urp@localhsot$ mv apache-hive-0.14.0-bin /usr/local/hive
exit

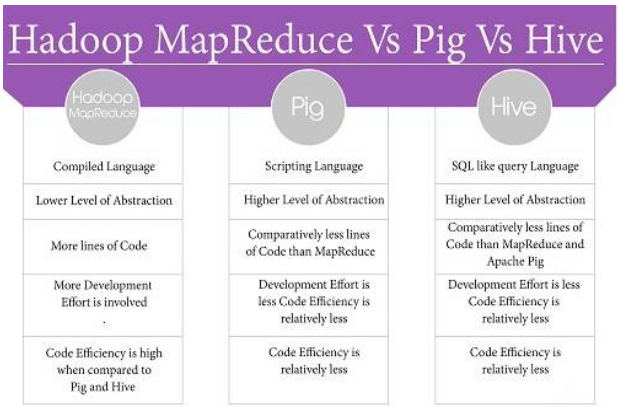
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

You can set up the Hive environment by appending the following lines to **~/.bashrc** file:
export HADOOP_HOME=/usr/local/hadoop
export HIVE_HOME=/usr/local/hive
export PATH=$PATH:$HIVE_HOME/bin
export PATH=$PATH:$HADOOP_HOME/bin
After installing Hive execute
$ cd $HIVE_HOME
$ bin/hive
Wordcount example for hive
Hive> select word,count(1) as count from (SELECT explode(split(sentence, ' ')) AS word FROM texttable)tempTable group by word
For Hadoop, there are many ways to run the jobs. The 3 programming approaches considered above i.e. MapReduce, Pig and Hive have their own pros and cons. Hadoop MapReduce is a compiled language whereas Apache Pig is a scripting language and Hive is a SQL like query language.



IV.CONCLUSION

Pig and Hive provide higher level of abstraction whereas Hadoop MapReduce provides low level of abstraction. Hadoop MapReduce requires more lines of code when compared to the Pig and Hive. Hive requires very few lines of code when compared to the Pig and Hadoop MapReduce because of its SQL like resemblance. Hadoop MapReduce requires more development effort than Pig and Hive. Pig and Hive coding approaches are slower than a fully tuned Hadoop MapReduce program. When using Pig and Hive for executing jobs, Hadoop developers need not worry about any version mismatch. There is very limited possibility for the developer to write java level bugs when coding in Pig or Hive. Pig has problems in dealing with unstructured data like images, videos, audio, text that is ambiguously delimited, log data, etc. Pig cannot deal with poor design of XML or JSON and flexible schemas. Mapreduce programming require knowledge of java. Pig Latin code can be extended through various user defined functions that are written in Python, Java, Groovy, JavaScript, and Ruby. Pig has tools for data storage, data execution and data manipulation. Pig Latin is highly promoted by Yahoo as all the data engineers at Yahoo use Pig for processing data on the biggest hadoop clusters in the world. Hive was started by Facebook to provide hadoop developers with more of a traditional data warehouse interface for MapReduce programming. For hadoop developers who know SQL, Hive is like a companion to them that makes them feel at home with all familiar and know SQL clauses like FROM, WHERE, GROUP BY ORDER BY. The drawback to using Hive is that hadoop developers have to compromise on optimizing the queries as it depends on the Hive optimizer and hadoop developers need to train the Hive optimizer on efficient optimization of queries. Apache Pig offers more optimization and control on the data flow than Hive.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

REFERENCES

[1] Ahmed Eldawy, Mohamed F. Mokbel " A Demonstration of SpatialHadoop:An Efficient MapReduce Framework for Spatial Data"Proceedings of the VLDB Endowment, Vol. 6, No. 12 Copyright 2013 VLDB Endowment 215080
[2] Mrigank Mridul , Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27
[3] Kyong Ha Lee Hyunsik Choi "Parallel Data Processing with MapReduce: A Survey"SIGMOD Record, December 2011 (Vol. 40, No. 4)
[4] Chen He Ying Lu David Swanson "Matchmaking: A New MapReduce Scheduling " in 10th IEEE International Conference on Computer and Information Technology (CIT'10), pp. 2736 –2743, 2010
[5] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.
[6] Jonathan Stuart Ward and Adam Barker " Undefined By Data: A Survey of Big Data Definitions"Stamford, CT: Gartner, 2012.
[7] http://blog.matthewrathbone.com/2013/04/17/what-is-hadoop.html
[8] https://hadoop.apache.org/docs/
[9] https://www.infoq.com/articles/apache-spark-introduction
[10] http://www.sas.com/content/dam/SAS/en_us/doc/conclusionpaper1/big-data-analytics-hadoop-107049.pdf