International Journal of Engineering & Technology, 7 (2.21) (2018) 417-419

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Effective processing of unstructured data using python in Hadoop map reduce

K. Kousalya[1*], Shaik Javed Parvez[2]

[1]*Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies (Vistas), Chennai, India.*
[2]*Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies (Vistas), Chennai, India.*
**Corresponding author E-mail: kousalyakumar80@gmail.com*

Abstract

In present scenario, the growing data are naturally unstructured. In this case to handle the wide range of data, is difficult. The proposed paper is to process the unstructured text data effectively in Hadoop map reduce using Python. Apache Hadoop is an open source platform and it widely uses Map Reduce framework. Map Reduce is popular and effective for processing the unstructured data in parallel manner. There are two stages in map reduce, namely transform and repository. Here the input splits into small blocks and worker node process individual blocks in parallel. This map reduce generally is based on java. While Hadoop Streaming allows writing mapper and reducer in other languages like Python. In this paper, we are going to show an alternative way of processing the growing unstructured content data by using python. We will also compare the performance between java based and non-java based programs.

Keywords: *Hadoop map reduce, unstructured data, streaming, performance, non-java based.*

1. Introduction

An examination work in 2011 exhibited that 92% of the world's whole information had been made inside the most recent six years [1]. In starting of 2004, the wide amount of data was maintained on single server. If any query was run on data store, it leads to data loss, reduce scalability. Mapreduce was first introduced by Google, it's architecture was widely used for parallel processing of un-organized manner data. While normal unstructured data processing leads to more time taken to maintain and process the data. To solve this issue Apache Hadoop introduced map reduce in dec-2004 [3]. By using this hadoop platform the data are process within minutes, rather than days. The excess amount of developing text information are regularly unstructured. Unstructured information does not have an organized structure. The interpersonal interaction media widely gives the services to real world with advanced features like Time Saving & an Attractive way [1]. Since the present scenario data are unstructured, data should be process effectively, and managed the data to satisfy our future needs.

Hadoop is an open source system which is used to store and process a ton of data in a parallel manner. It is a Java-based programming which is used for Processing the unstructured data. However the hadoop framework is based on Java, programs for hadoop need not be only coded in Java, but instead we also make use of various programming like c++ or Python. There are two phases in Hadoop mapreduce framework: Mapper phase & Reducer phase. It is essentially more Flexible, Economical, and Scalable [3]. Map reduce strategy is used for splits and combine of data, the two phases namely Map and Reduce. It processes the data and produce the <Key-Value > pair.

Hadoop allows us to process terabytes or petabytes of information and even more. We present the idea of utilizing non-Java programs for Hadoop Map Reduce execution for effective processing of unstructured data. MapReduce's streaming highlight enables software engineers to write coding on the mapper and reducer phase other than Java. For example, c++ or Python compose the Map Reduce programs like non-java based. The "trap" behind the Python code is that we will utilize the Hadoop Streaming API for helping us passing information between our Map and Reduce code through STDIN and STDOUT [12] [13]. We will just use Python's sys.stdin for input information and store our own result to sys.stdout.

The picture underneath demonstrate how the map reduce framework utilize the spilling highlights.

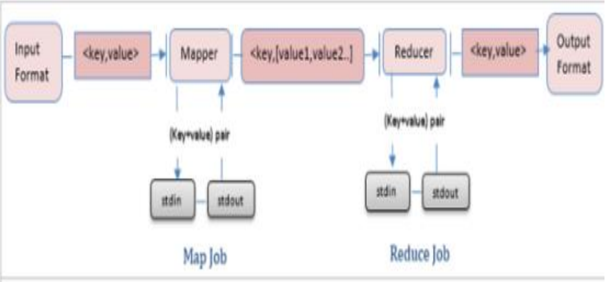


Fig.1: Processing unstructured data using python

Here we use UNIX operating system with hadoop. In map reduce framework, there are two phases namely Mapper and Reducer. These both phases executes the input, process it and finally that data are stored in the Reducer phase. This is the way hadoop system passes the input data and resulting the output through the mapreduce streaming technique. When the developer don't have much knowledge of Java to compose Mapper/Reducer, At this point when contrasted with streaming technique that extra overhead of beginning a scripting (python). This promotes a great deal between to processing java based and non-java based programme.

2. Methodology

2.1. Steps implemented in hadoop streaming

First, the mapper phase splits the input into lines and place into the stdin of the process. Then mapper collects the output of the process from stdout. It convert into <key-value> pairs. First, it converts the <key-value> pair into lines and put it into the stdin of the process. Then reducer collects the line output from the stdout of the process and prepare <key-value> pair of the final output.

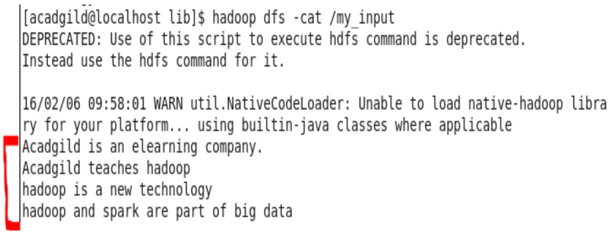
Step 1: Create one input file and store it in HDFS

Suppose the given input is- An learning company, Acadgild teaches hadoop. Being a new technology, hadoop along with spark are part of big data. We need to store this in HDFS.

First, make a directory to store the text file in desired folder. To make directory, the command is-$ *hdfs dfs –mkdir/user/hadoop/dir1*

Then, copy the data in a new text file, say file.txt. Then, we use the below command to save the file in hadoop directory. *$ Hadoop fs –copyFrom Local source-path destination-path*

Here, source-path is file.txt and destination-path is hadoop destination path in system. This command is used to create an hdfs directory.



```
[acadgild@localhost lib]$ hadoop dfs -cat /my_input
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/02/06 09:58:01 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Acadgild is an elearning company.
Acadgild teaches hadoop
hadoop is a new technology
hadoop and spark are part of big data
```

Fig.2: Input file for processing

Step 2: Write mapper class in Python and save in our local directory.

Here we use Cloud era Quick start VM to these examples. In Linux, first create one new document and type your mapper program in Python. Save the file as mapper.py. Open the command prompt execute the python script.

Step 3**:** Write reducer class in Python and save in our local directory.

Next, again create one new document and type your reducer program in Python. Save the file as reducer.py. Open the command prompt and execute the python scripts.

Step 4: Execution of map reduce using Hadoop streaming jar.

Copy the mapper.py & reducer.py scripts to the same folder where the above file exists. Open terminal and locate the directory of the file. This hadoop streaming utility allows any script executable to work as Mapper/Reducer provided they can work with stdin and stdout.

Command: ls – to list all files in the directory.

To see the content of the file: *command* is - cat file_name.

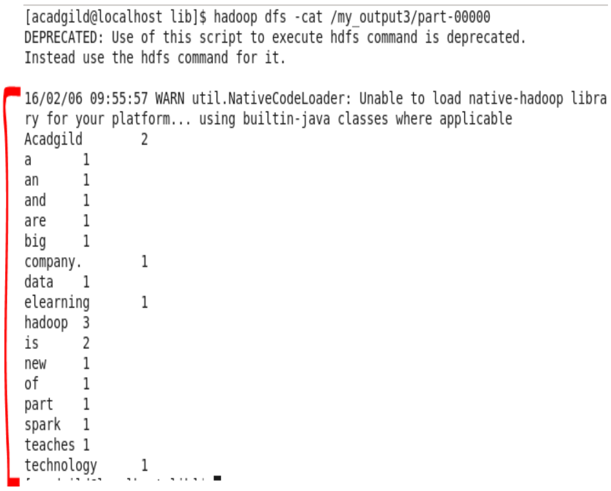
Before running the Map Reduce task on Hadoop, copy local data(word.txt) to HDFS.

Ex: hdfs dfs-put source_directory hadoop_destination_directory

Step 5: View the output from HDFS.

When you want to view the output on terminal, use this below command. This final ***output*** shows the occurrences of the word which appeared in the text file. This way we can extract the result from unstructured data.

Command: hadoop fs –cat/user/edureka/file name/part-0000.



```
[acadgild@localhost lib]$ hadoop dfs -cat /my_output3/part-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

16/02/06 09:55:57 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Acadgild        2
a       1
an      1
and     1
are     1
big     1
company.        1
data    1
elearning       1
hadoop  3
is      2
new     1
of      1
part    1
spark   1
teaches 1
technology      1
```

Fig.3: View the output

3. Map reduce with python

Map Reduce is a programming model that is widely designed for parallel processing of massive amount of information to be handled and is created for input spliting in blocks inturn executing that in parallel over a large range of machines [5]. At each stage Map Reduce program process the information twice, once in the map stage and once in the reducer stage. Python is a programming model and it has nothing to do with the Hadoop system. Using Hadoop streaming is possible by using non-java programmes like C++ or Python. Yet it is utilized for further future works like web improvement, progressed investigation, manmade brainpower, and so on [7].

3.1. Processing flow of map reduce job

There are three stages namely Map, rearrange and sort, and reduce.

3.1.1. Map stage

It is the first phase of Map Reduce framework. Under the map stage, the input is split into a small blocks with corresponding key-value. It forms a <key-value> sets [6]. The mapper successively forms each key-value match separately, delivering more yield <key-value> sets [12].

3.1.2. Rearrange and sort phase

The second stage of Map Reduce is the rearrange and sort. As the mapper being completed, the direct yield from the map organize are moved to the reducer. This method of moving yields from the mapper to the reducer is called as revising [12]. The last stage before the reducer start planning data is the masterminding technique. The center keys and characteristics for each bundle are organized by the Hadoop structure before being shown to the reducer.

3.1.3. Reducer phase

The third phase of Map Reduce is the reducer phase. Within the reducer phase, an iterator of values is provided to a function known as the reducer [13]. The iterator of values is a non-unique set of values for each unique key from the output of the map phase. The reducer aggregates the values for each unique key and produces output <key-value> pairs.

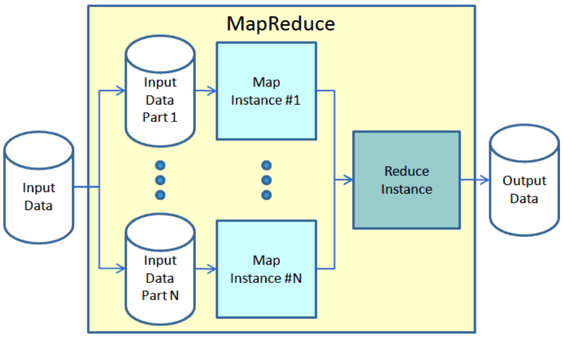


Fig.4: Processing flow of map reduce job

a). Datasets

The datasets are unstructured content of information. Along these lines the datasets is given as contribution to outline process for the refinement and organize.

b). Hadoop Streaming

Hadoop Streaming is widely used package like API. It allows Map Reduce occupations to be made with any executable as the mapper and the reducer. The Hadoop stream utility engages with c++ and Python. Hadoop Streaming that improves execution of instances: Python code having map and reduce capabilities process and executes superior than Java.

c). Running Hadoop map reduce

The datasets which is given as input to the map reduce process. The process could execute the input at any size. Hadoop is an compatible Binary operation. Hadoop utility allows that any script executable works on mapper and reducer phase.

d). HDFS

HDFS (hadoop distributed file system) is intended to store the data, regularly tera bytes and peta bytes and even more. This is refined by utilizing a piece organized file system. The design plan of HDFS is made out of two systems: it known as name node and data node. First one is name node that holds the metadata for the file system, and second one data node to stores the reports.

e). Structuring of the unstructured data

At last, the successful execution of the hadoop map reduce program, the final output is structured with the particular order.

3.2. Python Example

To demonstrate how the Hadoop Streaming utility can run Python in Map Reduce framework. It has two programs: mapper.py and reducer.py [13]. The mapper.py is the python program that first parallel process on mapper phase. It reads data from std.in with <key-value> pair, and yield each word to stdout. The reducer.py is the python program that executes on the reducer phase. It examines the results of mapper.py from std.in, whole unstructured information of each word, and puts the result to std.out.

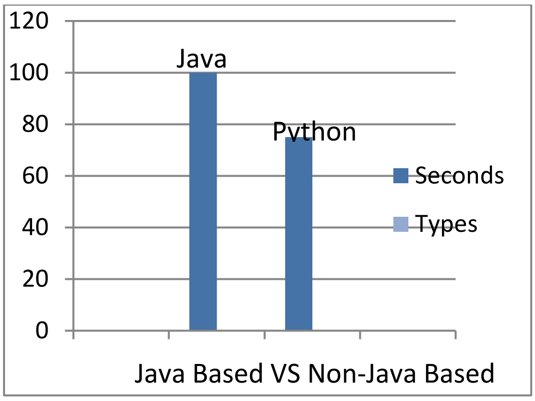


Fig.5: Performance based on time analysis

4. Conclusion

Map Reduce is one of the recent framework for processing the wide range of unstructured data in parallel manner. Actually, Hadoop Map Reduce based is on java, but it takes to much time and space for processing the Unstructured data. Mainly because of jar file creation. This survey demonstrates, the process of map reduce & how effectively it processes the unstructured text by using Python. Here the Map Reduce with Streaming Model, which provides an effective Processes to run non-java based programme within the context of a java-based Map Reduce framework.

References

[1] Subramaniya SV & Vijayakumar V, "Unstructured Data Analysis n Big data using Map Reduce", *2nd International Symposium on Bid data & cloud computing*, (2005).

[2] Leu JS, Yee YS & Chen WL, "Comparison of map-reduce and SQL on large-scale data processing", *IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, (2010), pp.244-248.

[3] Sudha P & Gunavathi R, "A Survey Paper on Map reduce in Big data", *International Journal of Science and Research (LJSR)*, Vol.5, No.9, (2016).

[4] Grolinger K, Hayes M, Higashino WA, L'Heureux A, Allison DS & Capretz MA, "Challenges for map reduce in big data. *IEEE World Congress on Services (SERVICES)*, (2014), pp.182-189.

[5] Ekanayake J, "Map Reduce Implementation for Streaming Science Application", *IEEE 8th International Conference on E-Science*, (2012).

[6] Dittrich J & Quiané-Ruiz JA, "Efficient big data processing in Hadoop MapReduce", *Proceedings of the VLDB Endowment*, Vol.5, No.12, (2012).

[7] Simone L & Gianluigi Z, "Python Map Reduce and HDFS API for Hadoop", *Proceeding of the 19th ACM International Symposium on High Performance Distributing Computing*, Chicago, USA, (2015).

[8] Lammel, "Google's Map Reduce programming model Revisited". Science Computer Program.

[9] Apache Hadoop http://hadoop.apache.org/

[10] Dinesh P, Processing Unstructured Data", Senior Architect Specialist, Virtusa Private Limited, (2015).

[11] Zaharia M, Konwinski AJ & Katz AD, "Improving Map Reduce performance in heterogeneous environments", Proceeding of the 8th USENIX conference on Operating system design and implementation, 2008.

[12] Michael, G, "Big Data & Distributed Systems", *International Journal of Science and Research*, (2015).

[13] Zachary R & Donald M, "Programming in Python", Published by O'Reilly Media, (2016).