International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 04, 2020
ISSN: 1475-7192

Usage of HIVE Tool in Hadoop ECO System with Loading Data and User Defined Functions

Dr. K. Uma Pavan Kumar[1], Dr. Lakshma Reddy Bhavanam[2]

***ABSTRACT**--The general usage of Hadoop is to store the bulk data with Hadoop Distributed File System and to process the data with Map Reduce. Apart from this the eco system provides extensive functionalities like usage of query-based logics to import the data from local path and Hadoop distributed path. This article presents the usage of Hive in the context of loading the bulk data and some simple analytics applicability. The Hive User Defined functions (UDF) creation and running with eclipse is the additional context of the paper. The work explains the parameters involved in the processing of the data loading and working with UDF's so as to simplify the Map Reduce (MR) process with HIVE commands.The context of Map Reduce requires the complex coding skills, and the problem is only HDFS path is known to the MR, there is no approach of working with local file system. The basic advantage of Hive is to work with local path files and as well as HDFS path files. Similarly processing wise Hive simplifies coding and functions usage with the implementation of the simple commands.The case study taken in this article deals with various parameters like page views data, system_IP, View_time, user_id and page_url. The other case study we have taken is loading of the bulk data in the less time.The outcome of the work is loading of the data in the context of local path and Hadoop Distributed Path. Loading of the bulk data within seconds and recording of the time taken is the other outcome. The creation of the UDF and running of the tasks in HIVE is the resultant of the work. Apart from these considerations the research issues and possible extension works can be observed in the article.*

***Keywords**-- Hive, Import, UDF, Map Reduce, Data Loading.*

I. INTRODUCTION

Hive initially started at Facebook, earlier to hive FB follows a different procedure to process their huge data. Data collected in Oracle DB, and then applicability of Extraction, Transformation and Loading through hand coded Python. Once the data is loaded will be scribed and send to HDFS. To process this stored data applicability of the MR this is bit complex.The issues with the above kind of the processing is

- There is no command line interface for the end users.
- Mandatory requirement of Ad-hoc querying, but the processing is through Map reduce Jobs.

Figure1: Sample Representation of Facebook data processing. (Earlier to Hive)

[1] *Associate Professor,Malla Reddy Institute of Technology, Hyderabad, India*
[2] *Professor and Principal, SJES College of Management Studies, Bangalore, India.*

DOI: 10.37200/IJPR/V24I4/PR201080
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

The other problem is there is no schema support, which is a must while processing the huge amounts of the linked data.The flow of the work is the section II deals with Hive commands and related results while loading the data in the context of local in path and HDFS path. In Section III provides the specification of the conclusion and possible future scope of the work.

II. HIVE ARCHITECTURE AND LOADING THE DATA

The Hive tool in Hadoop eco system is having many advantages; few can be observed as follows.

- Summarization of the daily data populated in FB and daily/weekly aggregations of shares, comments and clicks.
- Ad-hoc analysis of the huge data and broken down the data across country/state/region
- Any Kind of Spam Detection
- Corporate advertisement optimization
- Assembling of the training data so as to apply various data mining techniques.

There are certain points need to observe while working with Hive tool. Basically, Hive is suitable for offline kind of the data processing. The other point is even is not suitable to Online Analytical Processing. It is not a real RDBMS. It depends on the batch processing.

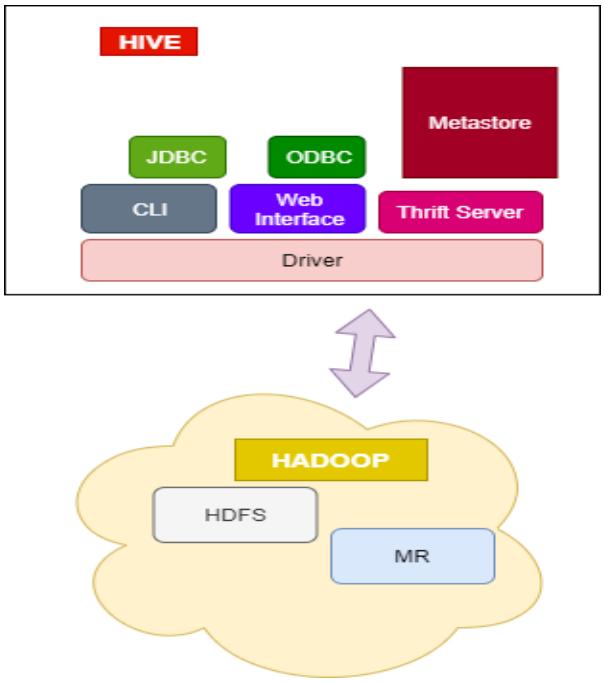

Figure 2: Hive Architecture

III. RESULTS AND DISCUSSION

Here we have created a db with name Research, and we are using the same to store the tables.

hive> desc page_view; OK

view_time int user_idbigint page_url_01String

ip String IP Address of the User dt string

country string

Time taken: 0.26 seconds

DOI: 10.37200/IJPR/V24I4/PR201080
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

hive> LOAD DATA [LOCAL] INPATH '/uma/hdp/input_hive1/page_view_20140415_IND.csv' INTO TABLE page_viewPARTITION(dt='2014-04-15', country='IN');

Copying data from file:/home/hdp/input_hive1/page_view_20140415_IND.csv Copying file: file:/home/hdp/input_hive1/page_view_20140415_IND.csv Loading data to table research.page_view partition (dt=2014-04-15, country=IN)

Processing Time taken: 0.984 seconds

hive> LOAD DATA[LOCAl] INPATH '/uma/hdp/input_hive1/page_view_20140415_US.csv' INTO TABLE page_viewPARTITION(dt='2014-04-15', country='US');

Copying data from file:/home/hdp/input_hive1/page_view_20140415_US.csv Copying file: file:/home/hdp/input_hive1/page_view_20140415_US.csv

Loading data to table research.page_view partition (dt=2014-04-15, country=US) OK

Time: 0.294 seconds

Selection of all the data from page_view_01Time taken: 0.402 seconds

2230	8	www.google.com	201.12.34.63	2019-04-15	IN
2246	10	www.twitter.com	201.12.34.79	2019-04-15	IN
2247	11	www.rediff.com	201.12.34.80	2019-04-15	IN
2248	12	www.msn.com	201.12.34.81	2019-04-15	IN
2230	13	www.nytimes.com	201.12.34.82	2019-04-15	IN
2231	14	www.guardian.com	201.12.34.83	2019-04-15	IN
2230	1	www.google.com	10.12.34.56	2019-04-15	US
2248	5	www.msn.com	10.12.34.60	2019-04-15	US
2230	6	www.nytimes.com	10.12.34.61	2019-04-15	US
2231	7	www.guardian.com	10.12.34.62	2019-04-15	US

Table 1: bulk data loading

We can observe bulk data loading in less time, here we are considering transaction data with more records and loading the data.

file:/uma/hadoop/input_hive_1/txns_1

Copying file: file: /uma/hadoop/input_hive_1/txns_1

Loading data to table retail_research.txnrecords_01

Processing Time : 0.569 seconds

hive> Select Count(*) from txnrecords_01;

Initial_Job = job_201909212044_0011,

Locator= //localhost:500130/details.jsp=Job_201909212044_0011

Terminating=/uma/hdp/Hadoop_1.0.3/ /../ /Hadoop

job - Map_Red.job_tracker=localhost:1234 -Kill job_201909212044_0011

DOI: 10.37200/IJPR/V24I4/PR201080
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

Hadoop_cluster Jobrelated info number of mappers: 1; number of reducers: 1 2019-09-21 22:26:10,131 Stage_1 Map = 0%, Reduce = 0%

2019-09-21 22:26:27,307 Stage_1 map = 100%, reduce = 0%, Cumulative_ CPU 1.55 sec

2019-09-21 22:26:33,432 Stage_1 map = 100%, reduce = 100%, Cumulative_ CPU 2.83 sec

MapReduce Total cumulative CPU time: 2 seconds 830 msec

Ended_Job =job_201909212044_0011 MapReduce Jobs_Launched:

Job_0: Map: 1 Reduce: 1 Cumulative CPU: 2.83 sec HDFS_Read: 8472303 HDFS_Write: 6 SUCCESS

Total process Time Spent: 2.0 Seconds 830 msec

95904

Process_taken: 32.476 seconds

9590004-14-2011	4007608	33.94	Exercise &Fitness	Cardio	Machine		
Accessories	Denver	Colorado	credit				
95901	01-02-2011	4007334	138.36	Outdoor Play Equipment		Outdoor	Playsets
	Huntsville	Alabama	credit				
95902	01-03-2011	4009230	32.84	TeamSports	Hockey	Everett	
	Washington	credit					
95903	09-05-2011	4005514	52.82	Jumping	PogoSticks	Scottsdale	
	Arizona	credit					

Time taken: 59.968 seconds

IV. CONCLUSION AND FUTURE SCOPE

The current work focused on the loading process of the data, and we have shown the results in terms of time taken to load the data. While loading huge data also Hive took less amount of time which is very encouraging to handle the huge data in the scenarios like FB, twitter and other job portals data. The future scope of the work is to deal with user defined functions and usage of the tables and coding so as to achieve the process of implementing the UDF in the simplest manner.

REFERENCES

1. U. P. K. Kethavarapu,S.Saraswathi, “Ontology based job recommendation system with dynamic source updates by slowly changing source detection”, International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 5, no. 3/4, pp. 164-173, 2016.
2. K.UmapavanKumar, S. V. N. Srinivasu, A. Ramaswamy Reddy, “Hadoop Cluster Performance with MR and Pig Latin in the Big Data”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 7, pp. 83-87, May 2019.
3. K.UmapavanKumar, “DWH security encapsulation with Bitmap Indexing Mechanisms”,IJETCSE, vol. 11, no. 2, pp. 10-15, Nov. 2014.
4. K.Umapavan Kumar, “Various Issues in Hadoop Distributed File System and MR future research

DOI: 10.37200/IJPR/V24I4/PR201080
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020

directions", International Journal of Pure and Applied Mathematics, vol. 120, no. 6, pp. 4441-4451, 2018.

5. U. P. K.Kethavarapu, "The ten ingredients of data base systems for improving performance and their review leading to research problems", IFRSA International Journal of Computing, vol. 2, no. 2, pp. 409-415, Apr. 2012.
6. U. P. K. Kethavarapu, "Various Computing models in Hadoop eco system along withthe perspective of analytics using R and Machine learning", International Journal of Computer Science and Information Security, vol. 14, pp. 17-23.
7. C. P. Chen and C.-Y. Zhang, "Data Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data", Information Science, vol. 275, pp. 314-347,2014.
8. J. Y. Monteith, J. D. McGregorand J. E.Ingram, "Hadoop and its EvolvingEcosystem", 5th International Workshop on Software Eco System ,pp. 57-68, 2013.
9. Mausam j. Naik (2019) mapksignalling pathway: role in cancer pathogenesis. Journal of Critical Reviews, 6 (3), 1-6. doi:10.22159/jcr.2019v6i3.31778
10. Craddock, T.J.A., Tuszynski, J.A. On the role of the microtubules in cognitive brain functions (2007) NeuroQuantology, 5 (1), pp. 32-57.
11. Georgiev, D.D., Papaioanou, S.N., Glazebrook, J.F. Solitonic effects of the local electromagnetic field on neuronal microtubules (2007) NeuroQuantology, 5 (3), pp. 276-291.

DOI: 10.37200/IJPR/V24I4/PR201080
Received: 02 Oct 2019 | Revised: 30 Sep 2019 | Accepted: 15 Jan 2020