I J C T A, 9(3), 2016, pp. 1719-1726
© International Science Press

Big data Techniques and Analytics in Distributed E-commerce business

Harish Balaji*, Ujjwal Pal and Uma Pavan Kumar K.*****

Abstract: The online market potentiality is increasing day by day, with the flexible ordering, comparison of products and easy return policies that the customers are preferring purchases of the various commodities online. The advent of various online shopping sites like Flipkart, Amazon, Shopclues, Snapdeal are attracting the customers to greater extents by providing cheaper rates on the purchases when compared with the normal purchases. The other side of this activity is how to attract the customers by knowing their preferences instantly and suggesting the best outcomes as per the budget, quality and other parameters which is a challenging task. The main focus of this article is how to estimate the interest of the customers with the help of search history and pattern of buying habits. The best solution for the above mentioned scenario is usage of big data tools like hadoop along with the eco-system tools such as HDFS, HIVE, PIG and Hbase. In addition to these tools we can make use of data mining algorithms, machine learning tools such as mahout to perform intelligence analytics with artificial intelligence. The current paper will give an overview of how to use the Hadoop technology so as to achieve the solutions of big data problems and analytics on the bulk amounts of the data which provides better logic and solutions of the real time problems in electronic commerce and other areas like banking, insurance and other retail domains. The other focus area of this paper is that we will give the various areas where we can use the Hadoop techniques and solutions where the market is involved with the usage of big data analytics on problems and description and usage of hadoop eco system tools in various big data scenarios.

Keywords: Hadoop, Big data, HDFS, eco-system, E-commerce market

1. INTRODUCTION

The online marketing working process is similar in case of all the E-commerce web sites. The current discussion flow is in section II explanation about the procedure of online business, in section III discussion about the empowerment of online marketing strategies when compared with current strategies. In section IV, big data problem and the hadoop eco system solutions along with the tools specification like HDFS, Hive, Hbase, Pig, Sqoop and Flume is elaborated. In section V, the description about the integration of online marketing with big data solutions is explained.

2. ONLINE MARKET WORKING PROCESS

The online business is gearing up nowadays and the reasons behind that are simple, i.e. we can get the products with best price by comparing the available products of the same brand and same model. The other benefit is flexible payment options like cash on delivery, EMI through credit card payment, Credit card payments and Net banking through various bank interfaces. Another benefit given to customers by online business sites is easy return policy through which the customers can return their products due to various reasons. The following is the flow of the tasks that will be performed during the online purchase by the customers.

* Undergraduate Student, Department of Information Technology, Pondicherry Engineering College, Puducherry, *Email: bharish94@gmail.com*

** Assistant Professor, New Horizon College, Bangalore, *Email:palakash2008@gmail.com*

*** PhD Scholar, CSE Department, Pondicherry Engineering College, Puducherry, *Email: umapavanmtech@gmail.com*

Step 1: Customer interface with online site

Step 2: Selection of Products

Step 3: Comparisons with other sites to get best price

Step 4: Fixing the product through the specific site

Step 5: Selection of COD/Other Payment Options

Step 6: Order confirmation

Step 7: Details about Easy return policies

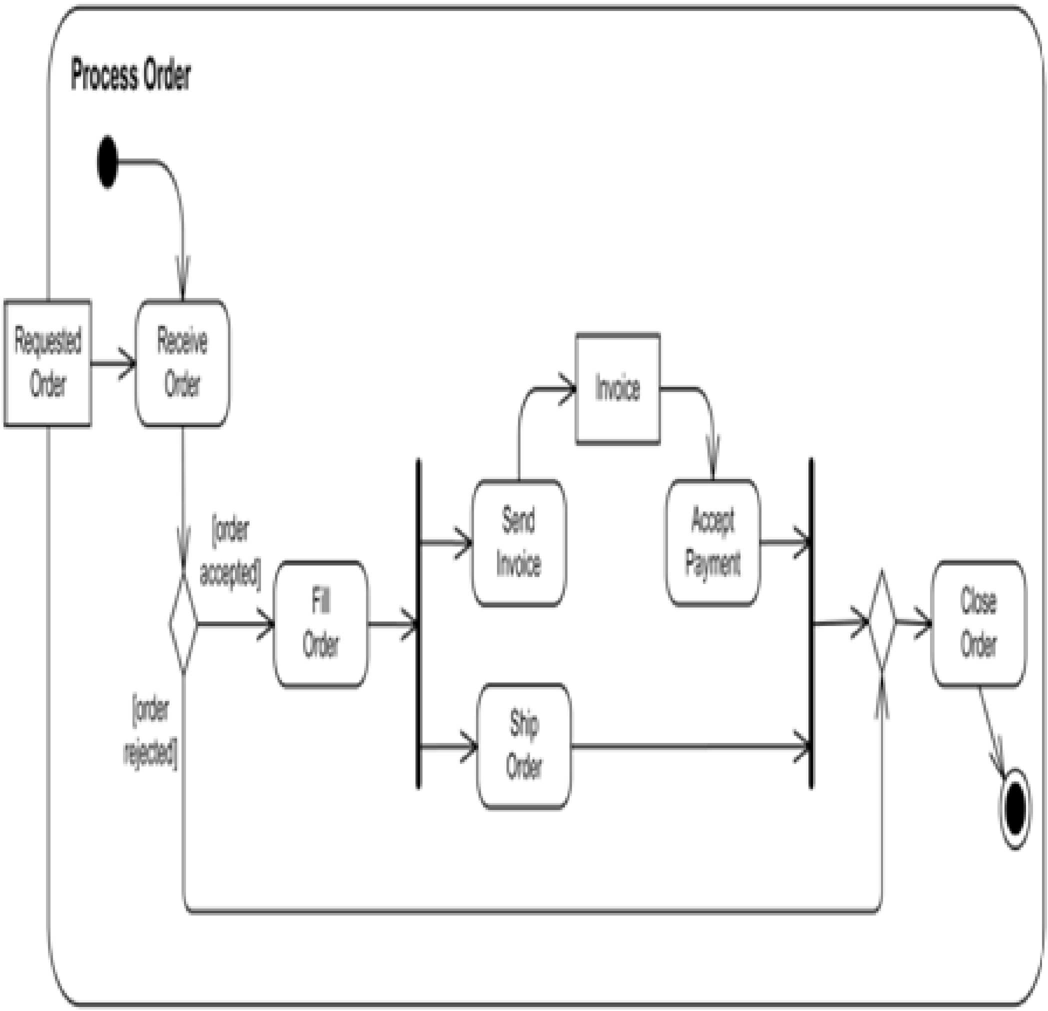


Figure 1: Online-Purchase process flow

3. EMPOWERING OF ONLINE BUSINESS

The online purchase is concentrating on order placements, payment options and order confirmation along with the return policies. The following are various possibilities through which the customers can be provided best services and various sites can make use of the following suggestions so as to get the benefit of the business.

3.1. Customer Context

1. While searching for product recommendations from various search engines like the price, warranty details along with service benefits will give the better analysis about a product, so that the customer can chose the best product.

2. The recommendation may not be of the same product but with the same parameters (Like price, Memory Capacity, Processor speed in case of laptops) which will give a global view about the item rather than company.
3. The payment benefits through credit card purchase like cash back or deductions available through the usage of particular card.
4. Analytics about the products, reviews and worldwide sales which will give the market value of the product and allowing the customers so as to get the best out of the available products.

All the above options will give the best outcomes to the customers.

3.2. Business Context

1. Identifying buying habits of the customers
2. Locating the products which are attracting the customers
3. Analytics about a product based on region, age factor and other parameters.
4. Getting information from the huge amounts of the customer profiles and products data.

The main challenge faced by the companies is huge volumes of data in unstructured format. If companies could able to get insight into the bulk data and getting the valuable information which will allows companies to have the strategic decisions based on that. In the next coming section we are going to explain the usage of big data and hadoop.

4. BIG DATA AND HADOOP

90% of the data in the world today has been created in last two years alone. Structured format has some limitations with respect to handling large quantities of data. Thus, there is a need for perfect mechanism, like Big Data, to handle these increasing quantities. Big Data is the term applied to data sets whose size is beyond the ability of the commonly used software tools to capture, manage, and process within a tolerable elapsed time. The following are various sources of big data.

- √ Web logs
- √ Sensor network
- √ Social media
- √ Internet text and documents
- √ Internet pages
- √ Search index data
    - √ Atmospheric science, astronomy, biochemical, medical records
    - √ Scientific research
    - √ Military surveillance
    - √ Photography archives

A free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment, based on Google File System (GFS). Runs applications on distributed systems with thousands of nodes involving petabytes of data with Distributed file system, provides fast data transfers among the nodes. The following are various companies that are making use of hadoop Amazon, Facebook, Twitter, Yahoo, Google. The main services of hadoop are as follows.

The core services of Hadoop are:

- NameNode
- DataNode
- JobTracker
- TaskTracker
- Secondary NameNode

The key elements of Hadoop are HDFS and MapReduce.

The key features of Hadoop MapReduce are:

- Performing distributed data processing using the MapReduce programming paradigm
- Possessing user-defined map phase, which is a parallel, share-nothing processing of input (MapReduce paradigm)
- Aggregating the output of the map phase, which is a user-defined reduce phase after a map process.

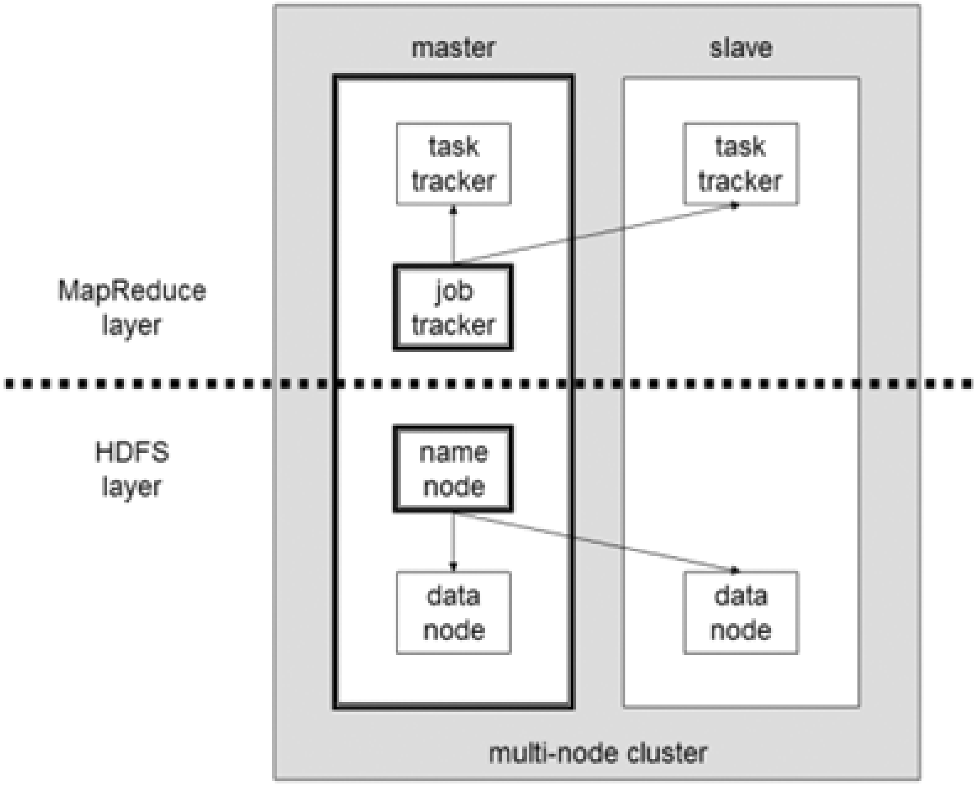


Figure 2: Hadoop Architecture

5. INTEGRATION OF ONLINE MARKET WITH BIG DATA

As mentioned in the section II the online marketing requires the uplift of the data processing with huge amounts of the data. The following are some of the ways that can be added to achieve the above mentioned challenges faced by the on line market.

1. Usage of MapReduce frame work so as to run the bulk jobs in the distributed manner.
2. Implementing MapReduce with Java requires Java professionals and needs bulk lines of code.
3. The simplest method of integrating the big data problems to MapReduce is either Hive or Pig.

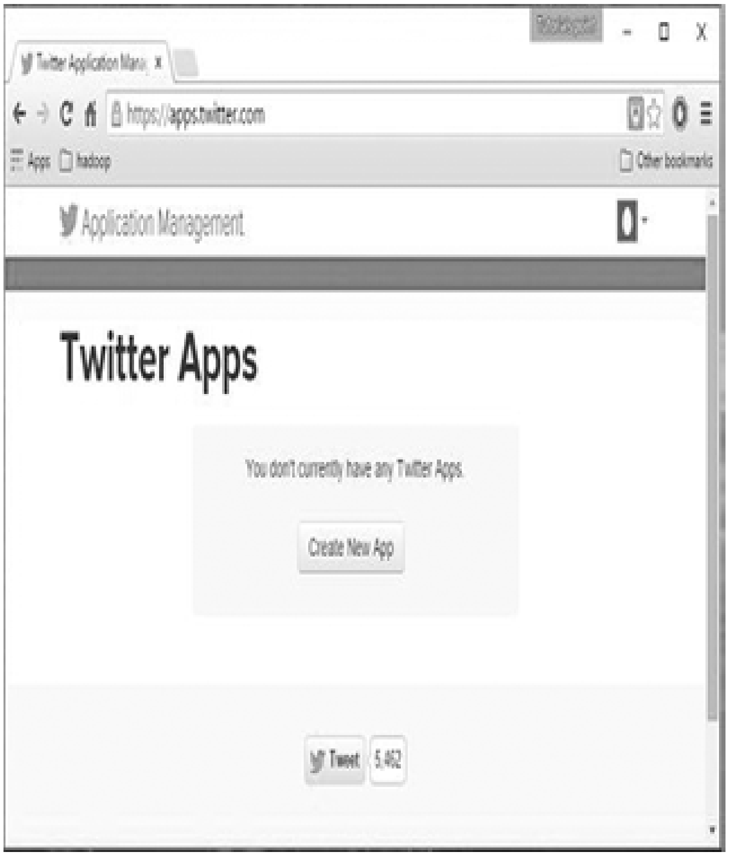
4. With Hive, which is a data warehouse without OLTP and is like SQL in the interface of HiveQL with simple coding we can achieve the MapReduce as that of Java.
5. With Pig, this is a script based language and well suited for bulk data processing to handle MapReduce aspects with simple and efficient coding.
6. Flume which is used to handle analytics of streaming data like twitter analytics.
7. Sqoop is data transfer tool through which we can change the MySQL data into Hive and other NoSQL (like HBase, Mongo DB)
8. As an ecosystem kind of interface all of the above aspects we can apply through a common file system that is Hadoop Distributed File System (HDFS) and processing can be done through MapReduce which will parallelize the jobs and proceeds in a distributed format.

6. REAL TIME IMPLEMENTATION OF ANALYTICS USING BIG DATA

Here we are giving a real time scenario where we can apply the concept of analytics. The scenario here is getting the tweets data from twitter contains “Bigdata”, “Hadoop”. Getting these words from twitter posts is requiring the following scenario. In order to get the tweets from Twitter, it is needed to create a Twitter application.

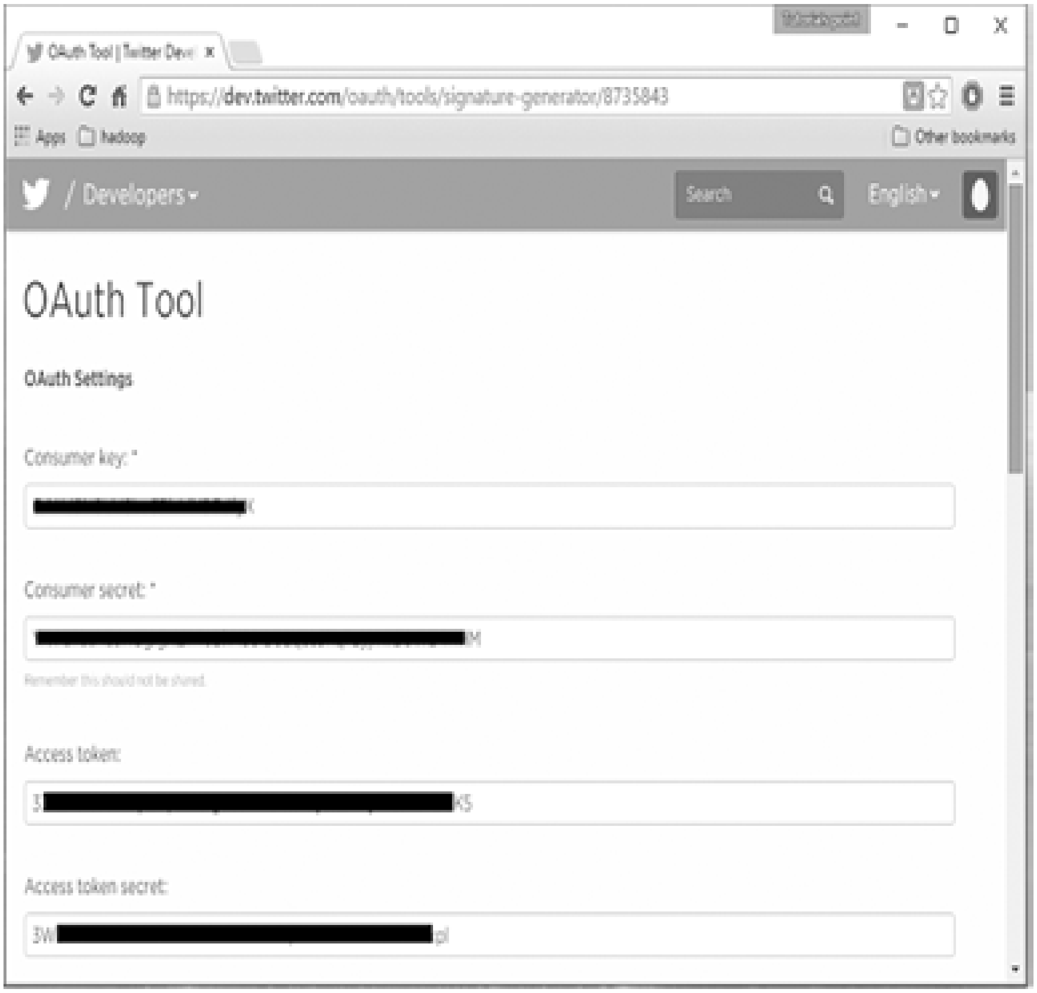
To create a Twitter application, click on the following linkhttps://apps.twitter.com/. Sign in to your Twitter account. You will have a Twitter Application Management window where you can create, delete, and manage Twitter Apps.

Fill in the details, accept the **Developer Agreement** when finished, and click on the **Create your Twitter application button** which is at the bottom of the page. If everything goes fine, an App will be



You don't currently have any Twitter Apps.

Figure 3: Twitter App Creation



OAuth Tool

OAuth Settings

Consumer key: *

[redacted]

Consumer secret: *

[redacted]

Remember this should not be shared.

Access token:

[redacted]

Access token secret:

[redacted]

Figure 4: Various keys in twitter App Creation.4

created with the given details as shown below. Finally, click on the Test OAuth button which is on the right side top of the page. This will lead to a page which displays your Consumer key, Consumer secret, Access token, and Access token secret. Copy these details. These are useful to configure the agent in Flume.

In the next step we have to get the support of flume-sources-1.0-SNAPSHOT.jar. And in the flume folder modify the flume-env.sh file as follows.

```
FLUME_CLASSPATH=”home/apache-flume-1.3.1.-bin/flume-source-1.0-SNAPSHOT.jar”
```

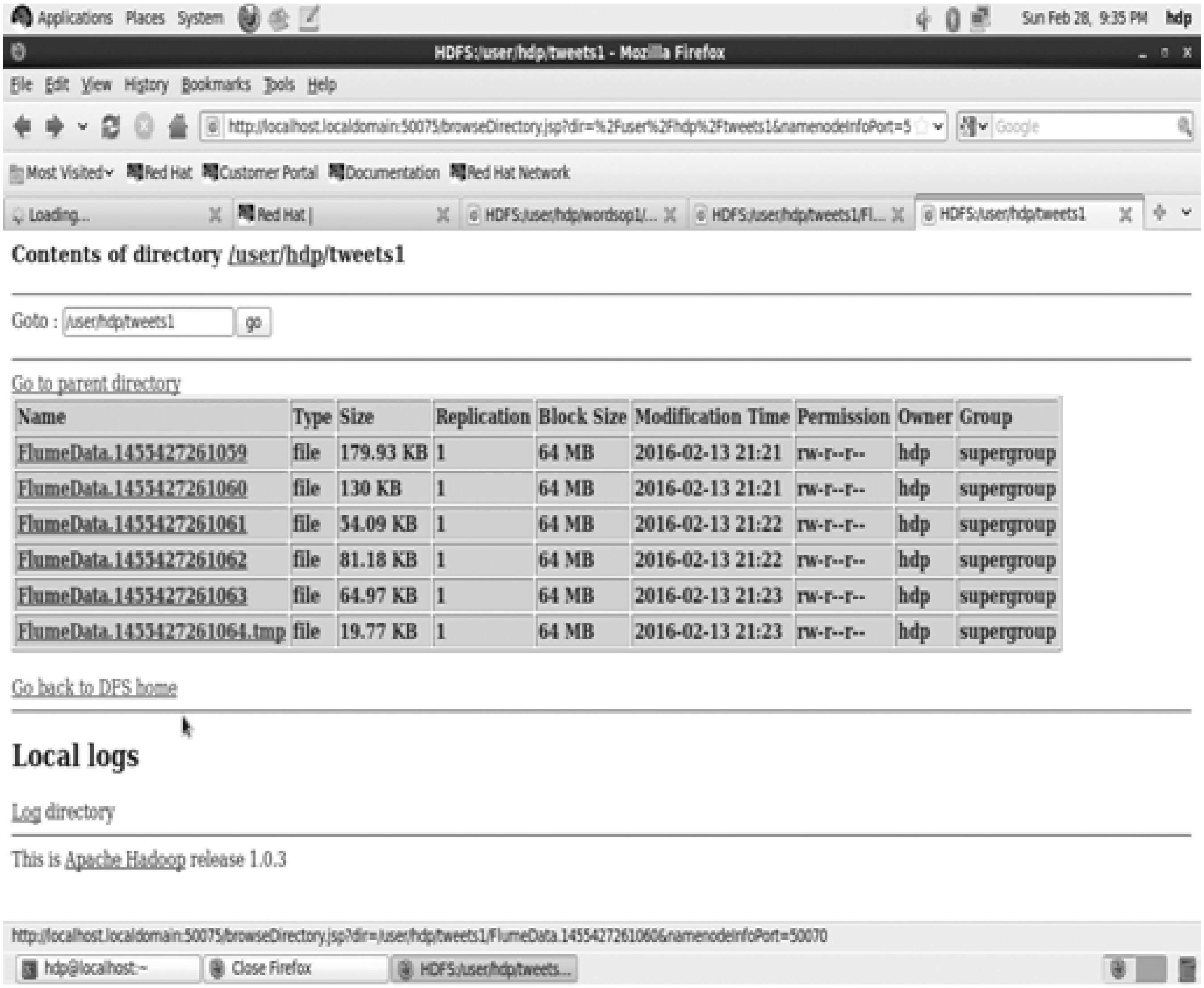
Finally run the following command so as to get the tweets data in the browser of local host hadoop machine.

```
Flume-ng agent –n TwitterAgent –c conf –f/home/hdp/apache-flume-1.6.0-bin/conf/flu, e-conf.
```

The following figure shows that how a twitter data will be saving into files with various size in KB so that we can analyse the tweets data as per the given query.

7. CONCLUSION

In this paper we have explained the concept of online marketing and the process of purchasing along with the payment options. Explanation about quality improvement of online business with some suggestions through which we can get the benefits. The description about Big Data and Hadoop with the basic concepts and HDFS, MapReduce aspects. The final conclusion is how to integrate the suggestions with common file system and common processing. The solution is HDFS and MapReduce. We have given the twitter analysis


Applications Places System
Sun Feb 28, 9:35 PM hdp
HDFS:/user/hdp/tweets1 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost.localdomain:50075/browseDirectory.jsp?dir=%2Fuser%2Fhdp%2Ftweets1&namenodeInfoPort=5
Google
Most Visited
Red Hat
Customer Portal
Documentation
Red Hat Network
Loading...
Red Hat |
HDFS:/user/hdp/wordsop1/...
HDFS:/user/hdp/tweets1/Fl...
HDFS:/user/hdp/tweets1
http://localhost.localdomain:50075/browseDirectory.jsp?dir=/user/hdp/tweets1/FlumeData.1455427261060&namenodeInfoPort=50070
hdp@localhost:~
Close Firefox
HDFS:/user/hdp/tweets...

Contents of directory /user/hdp/tweets1

Goto : /user/hdp/tweets1 go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1455427261059	file	179.93 KB	1	64 MB	2016-02-13 21:21	rw-r--r--	hdp	supergroup
FlumeData.1455427261060	file	130 KB	1	64 MB	2016-02-13 21:21	rw-r--r--	hdp	supergroup
FlumeData.1455427261061	file	54.09 KB	1	64 MB	2016-02-13 21:22	rw-r--r--	hdp	supergroup
FlumeData.1455427261062	file	81.18 KB	1	64 MB	2016-02-13 21:22	rw-r--r--	hdp	supergroup
FlumeData.1455427261063	file	64.97 KB	1	64 MB	2016-02-13 21:23	rw-r--r--	hdp	supergroup
FlumeData.1455427261064.tmp	file	19.77 KB	1	64 MB	2016-02-13 21:23	rw-r--r--	hdp	supergroup

Go back to DFS home

Local logs

Log directory

This is Apache Hadoop release 1.0.3

Figure 5: Twitter Output Stored in various files in Flume

with the help of flume through which we can analyze the trend and sentiment of the various community of the people.

References

[1] www.google.com

[2] www.yahoo.com

[3] Thilina Gunarathne, Hadoop Map Reduce cookbook, 2nd edition, Mumbai, Maharashtra: Packt.

[4] Tom White (2009), Hadoop the definitive guide, 2nd edition, Newton, Massachusetts: O'Reilly.

[5] Alexey Yakubovich, Boris Lublinsky & Kevin T. Smith (2013), Professional Hadoop Solutions, New York, NY: Wiley.

[6] Tom J. White, "The Hadoop definitive guide", Fourth edition, Orielly

[7] http://hadoopilluminated.com/hadoop_illuminated/Hadoop_Use_Cases.html

[8] Dr. B. Lakshma Reddy, (2015), Big data techniques and analytics in E-commerce business, International conference in RDA, Goa, November 14 and 15.

[9] Uma Pavan Kumar. Kethavarapu, (2015), A Model for analysis and change detection in uncertain data streamsprocessing, BIAKM, ICFAI, Hyderabad, December 15 and 16.

[10] Bhosale, "Efficient Indexing Techniques on Data Warehousing" International Journal of Scientific & Engineering Research vol 4, Issue 5, May-2013, ISSN 2229-5518.

[11] Naveen Gar, PhD scholar, SN University, Jharkhand, “Bitmap Indexing technique for data warehousing and data mining”, International Journal of Latest trends in engineering& Technology vol 2. Issue1, January 2013, ISSN: 2278-621X

[12] Zanab qays abdulhadi, school of information systems and engineering, central south university, China. “Bitmap index as effective indexing for low cardinality column in data warehouse”, International Journal of computer applications, vol 68, April 2013, ISSN: 0975-8887).

[13] Jesus Camacho-Rodriguez “Web data indexing in the cloud: Efficiency and Cost reductions”, ©ACM 2013, March 18-22.