# CREDIT CARD DATA PROCESSING AND E-STATEMENT GENERATION WITH USE OF HADOOP

**Ashvini A.Mali[1], N. Z. Tarapore[2]**

[1]*Research Scholar, Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra*
***ashvini63@gmail.com***
[2]*Assistant Professor, Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra*
***ntarapore@yahoo.com***

## Abstract

*The Bigdata for most of part is known as virtually wealthy storage and processing .This paper is related to basic knowledge of Bigdata and Hadoop environment. Now a day's various organizations deal with problems like data sharing ,data privacy and secure data transfer where they take lots of time for all this work , so to avoid this we are using Hadoop. Hadoop is used for large storage and parallel processing , we can say it also works in distributed fashion. There are different hadoop tools available like hive, sqoop, pig and more which we will discuss later.*

***Keywords:*** *Bid Data, Hadoop, Parallel Processing, Map-Reduce, Hdfs*

----------***----------

## I. INTRODUCTION

Big organizations like Infosys, cognizant and more on they have big data analysts ,who analyze global market and supply chains and also analyze insights of customer demands from various information collected on business transactions and market needs. This data which is used to analyze is collected from various sources like newspapers, social sites , business blogs, discussion forums etc[3]. After collection analysts check this data from various ways and makes decisions which are useful to improve supply chains and customer behavior and also effective toward enhancing business.

Every organization has rapid growing data day by day and it becomes tough for the system to process and respond for every query or function. It is difficult for the bunch of data in single drive makes reading and writing slower for a system in-charge. Hadoop is an solution of common storage and parallel processing[2] . The scope of using Hadoop is, it can handle large volumes of data and more efficient in handling data losses which is more important for all sectors and even fair enough for processing the data faster. All the sectors like financial, healthcare ,business are trying to use hadoop for faster performance and to reduce workload.

## II. RELATED WORK

The Big organizations, health care centers and more financial sectors like bank all deal with large data .so to process this data our normal processing and database does not respond well ,so hadoop comes in this situation. where it deals with big data and processing . There are different tools available in market which are useful like HBase , Hive , Pig , sqoop ,flume etc etc . The Hive and PIG both are used to work with big scale data .working at small data it will affect on time using this tools but if you are dealing with big scale data then it gives fabulous results, also other tools like sqoop and flume used for data import and export purpose or for data collection .How this all work will be described in section IV later.

## III. BIG DATA CONCEPT

Big data is similar to "small data " but it is bigger than small one. we are part of it every day, which is lengthy to process n store, but data having bigger requires different approaches to handle this large data .There are different approaches used like technologies, tools and architectures, which are used to solve new problem also old problems with better way. We all are using smart phones, tablets ,camera and more on social networking sites like facebook, twitter etc ,which generates large volume of data which needed to be stored and processed simultaneously as per our or customer demands. This large data yield in petabytes or terabytes so to store this data require large storage space so hadoop comes into this situation to solve storage and processing problem. Big data is not only related to size but also relates with other data analysis and processing norms. The enhancement of this big data is achieved through the 3 v's , which are nothing but variety, velocity and volume.

Volume :- The amount of data we collect and store like environmental data, entertainment data , event data and more on ,so data is either structured or unstructured. Most of the relational databases support structured data.
Variety :- Data is in different types for example text ,image, video, audio .This makes complex to handle this data. for e.g. Time required to retrieve text data is less than other types.

Velocity:- Data is growing faster ,which need to be handled.

Big Data is something new for startup companies but its plays a major part of big organizations .they are wrestling around this every day. one of the widely used technology among organizations is hadoop cluster , where data stored and analyzed in less time and also reduce cost.

## IV. HADOOP

Hadoop is an open source tool for big data analytics. Hadoop works in distributed fashion where it has one or more clusters. It has main two parts which plays main role in its architecture , hadoop HDFS and Map-reduce.

*A.Purpose of hadoop*

The storage and processing of data is main concern today. Every organization has rapid growing data day by day which becomes tough for system to respond for every query. Hadoop writes the data sets into chunks and stores as three data blocks with 64MB (64MB not mandatory as per Admin settings) as default. The loss of data may also happen for a data in a single drive. But, Hadoop has three data blocks for each chunk of dataset at the time of loss. The other problem is combining two data chunks is more challenging. But, Hadoop can manage two different data sets when combining through key-value pair's technique. Hadoop is a map-reduce processor which deals with huge data and process with in no time. The RDBMS databases can also handle large datasets by using more disks but yet it slower because of its seek times and also troubles the transfer rate too.

*B.Goals*

- Abstract and facilitate the storage and processing of large and/or rapidly growing data sets, also handle structured and non structured data with the available Simple programming models
- High scalability and availability of data any time when it will be rquired
- Use commodity (cheap!) hardware with little redundancy and Fault-tolerance

*C. Hadoop Architecture*

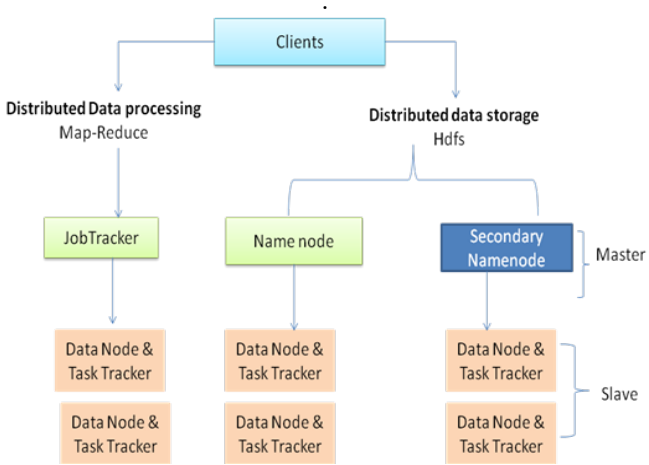


**Figure 1:** Hadoop Architecture

We have discussed the importance of Big Data, next is how to architect systems that can deliver a Just in Time Analytical Analysis Any Big Data architecture will almost always address data storage and query, time dimension considerations are an amendment to it. So it is important to understand the foundation architecture before moving into to any Just in Time Big Data Architecture. For better or worse, Hadoop is becoming the basis for big data storage and query.

From above figure (Figure 2:Hadoop Architecture) we can depict that hadoop works with both distributed data processing and distributed data storage in parallel way. We will discuss how this Map-Reduce and Hdfs works in detail in this section.

*A.*Hadoop HDFS

This is distributed storage file which is used by hadoop ecosystem. It is suitable for distributed storage and processing. hadoop has command line interface through which we can interact the HDFS system. There are servers available of namenode and datanode which provides service information or status information to user easily.

HDFS is exexutes on low-cost hardware and it is highly fault-tolerant. It takes data which is stored in blocks which replicates those data three times (by default).

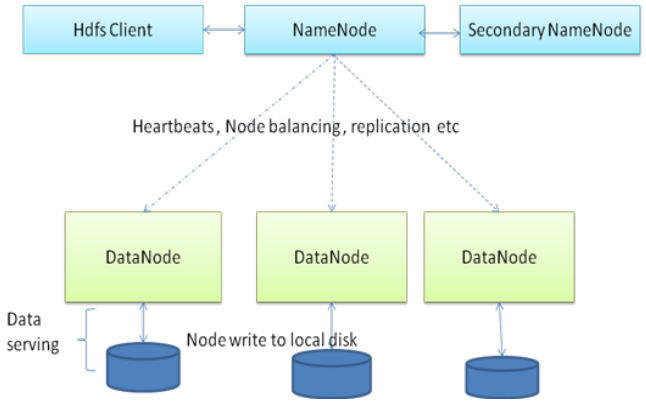


**Figure 2.** HDFS Workflow

HDFS is built to support applications with large data sets, including individual files that reach into the terabytes. It uses a master/slave architecture, with each cluster consisting of a single NameNode that manages file system operations and supporting DataNodes that manage data storage on individual compute nodes.

HDFS has some characteristics like fault detection and automatic recovery ,streaming data access, has large data sets, moving computation is cheaper than moving data ,data availability ,data replication etc etc.

*D. Hadoop Map-Reduce*

Hadoop Map reduce has two major parts one is map and another is reduce .Map takes data as input and convert it in set of outputs using some key-value pair .it is mostly java based. Mapper does mapping of input data using key and

then sort and shuffle is done on that mapped data and finally reducer collect that final data. for e.g word count program where we give input as an text file and get output as count of words appeared in that text. It is parallel and distributed algorithm or technology which works in less time as compared to other techniques. MapReduce program executes in three parts, map , shuffle, and reduce.

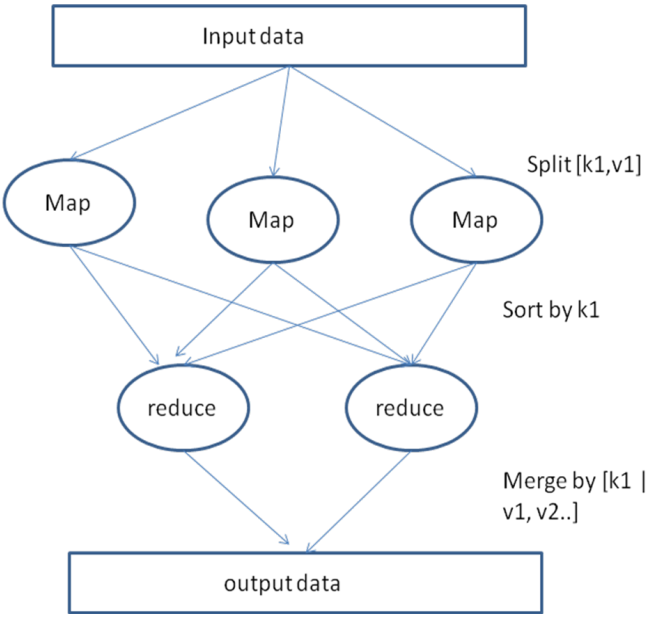


**Figure 3:** Map reduce workflow

- **Map** : The map does the work of processing input data. The input data which itself stored on HDFS system allready. this file is given as input to the mapper function line by line. The mapper processes the data and creates several small splits of data.
- **Reduce stage** : This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer does the work of to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Hadoop is a map-reduce processor which deals with huge data and process with in no time. The RDBMS databases can also handle large datasets by using more disks but yet it slower because of its seek times and also troubles the transfer rate too.

## IV. HADOOP TECHNOLOGY AND TOOLS

*A.Hadoop Environment*

Hadoop environment which is made up of Hadoop master , hadoop slaves , one sqoop ,one hive and one pig which is shown as in figure 5 below, where the import and export of data is done by sqoop and hive and pig are responsible for data processing whereas HDFS used for data storage. Below figure shows the representation of hadoop

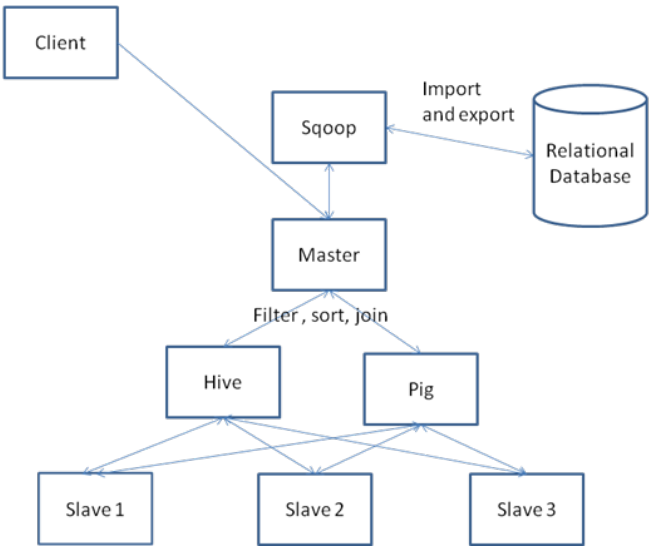


**Figure 4:** Hadoop Multi-cluster environment

multi cluster environment where it has three slaves ,one hive ,one pig and one master and finally sqoop sqoop relocate data from mysql or sql server to the HDFS file system in hadoop. Each tool of hadoop ecosystem is used for specific reasons e.g. hive is used for join and pig is scripting language

## V. CREDIT CARD STATEMENT GENERATION

This are the steps which are carried out in this system which are as follows:-

1. *Pig* generates the **confirmation file** by filtering the account numbers holding e-statements from raw data (*fig 6*) and stores it in HDFS.

2. **NIT file** is loaded into *Hive* table including account numbers and card holder details.

3. **Supplement file** is loaded into *Hive* table including account numbers, record type and legal identity details.

4. **Cycle file** is loaded into *Hive* table including account numbers, product type, language preference, state code, bill type and event time-stamp details.

5. **WCC file** is imported from Sql-db through *Sqoop import* command listing account numbers, first name and last name as fields.

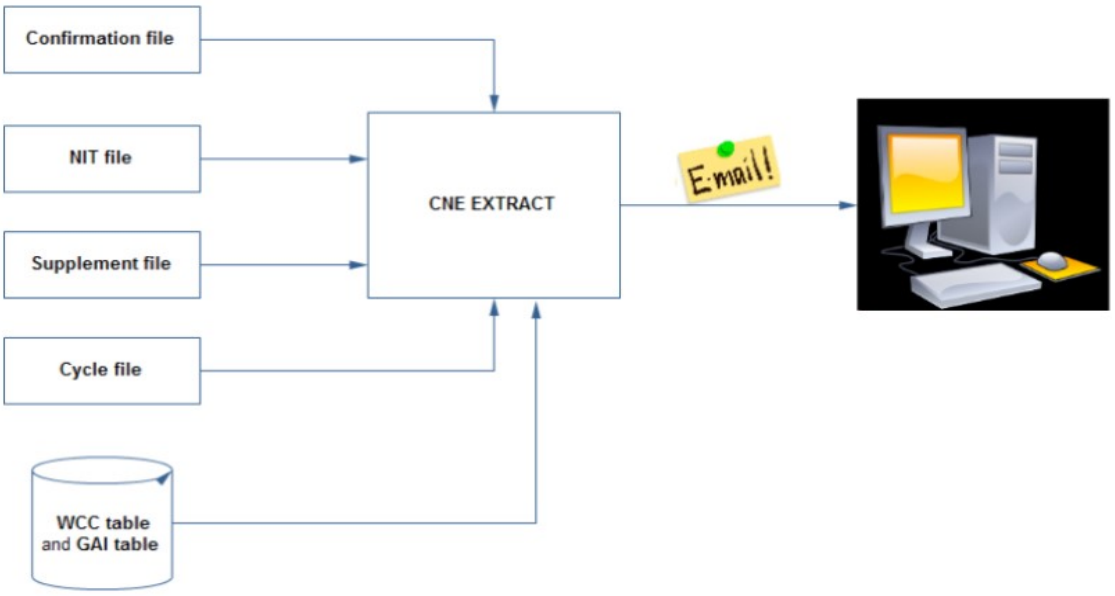


**Figure 5.** High level diagram of credit card generation

6. **GAI file** is also imported from sql-db through *Sqoop import* command listing account numbers, global ID, party ID, e-mail and last-login as fields.

7. The generated e-statement account numbers from *pig* are later loaded it into *Hive* table and stored in HDFS.

8. WCC and GAI imported files are also loaded into *Hive* table and stored in HDFS.

9. The **Default** fields containing are alert type, record type, account entity type, product type, and locale are also loaded into *Hive* table for further processing.

10. Finally, all the hive tables are joined using **Hive join** and processed through map-reduce in order to generate **CNE** file.

## VI. RESULTS AND SCREENSHOTS

The input dataset is available which have the all information abpout card holder like account number , card holder name ,address, gender and more on also the credit card payment details and product details.Pig is chosen, because of its comfortable scripting and maintenance. The current Framework makes the usage of Pig easily, by filtering the records according to the requirement.

PIG generates confirmation file by filtering and extracting account number field from dinput dataset and generates new file where all this data resides. Using hive query create and load this specific fields in each file which will be used later for hive join, so NIT, CYCLE and SUPPLEMNT files are available

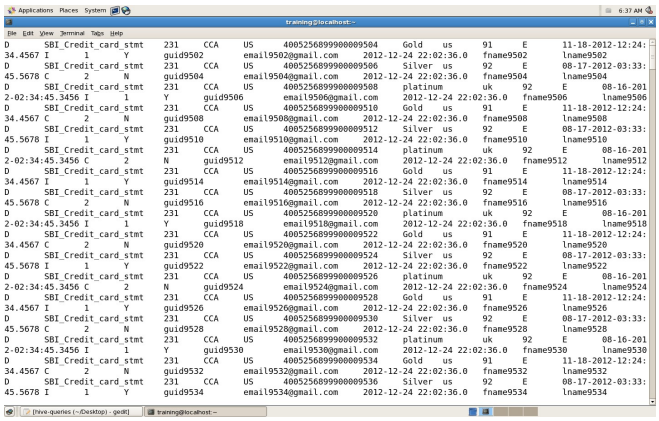

**Figure 6.1** Screenshot of raw file of credit card details from murchant or middle processor

The Nit files contains card holder details and account numbers, supplement files contains account numbers, record type and legal identity and cycle file contains account numbers, product type, language preference, state code, bill type and event time-stamp as fields. It is very easy to generate the files with different fields using HIVE.

```
hive> select * from supplement_table;
OK
4005256899900000002      N       2
4005256899900000004      Y       1
4005256899900000006      N       2
4005256899900000008      Y       1
4005256899900000010      N       2
4005256899900000012      Y       1
```

**Figure 6.2** supplement file in Hive

```
hive> select * from cycle_table;
OK
4005256899900000002     SILVER  91      E       Uk      05-26-2012-09:45:34.6789
4005256899900000004     PLATINUM        92      E       Uk      07-19-2012-07:43:34.5678
4005256899900000006     GOLD    91      E       uk      10-18-2012-11:25:34.3456
4005256899900000008     SILVER  91      E       Uk      05-26-2012-09:45:34.6789
4005256899900000010     PLATINUM        92      E       Uk      07-19-2012-07:43:34.5678
4005256899900000012     GOLD    91      E       uk      10-18-2012-11:25:34.3456
```

**Figure 6.3** Cycle file in Hive

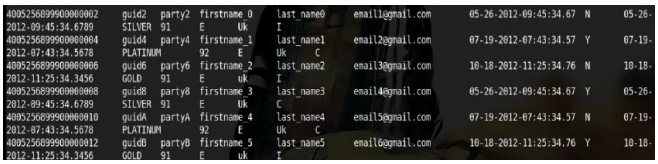


**Figure 6.4** CNE file

Later mysql is used for creating tables in hive using mysql queries this work is accomplished in less time. Here gai and wcc tables are created and data is stored in this files later again this data or sql-databases are loaded or imported using sqoop in the HDFS system.

After this all generated files are stored in HDFS and Hive joins are done on those files from which we get specific fields and e-statement gets generated which is CNE file.

*A. Experiences*

1. Make sure that the disk space is available for storing five million records
2. While joining all the Hive tables, make sure all the tables has the common field with common name.
3. Nulls may be raised, if the fields in a file are not delimited properly.
4. Once the data is ready, run the sample data first which makes the programmer easy to identify the condition and later it is easy to make the decision of increasing the reducers or stay with it.

## CONCLUSION

Pig is chosen, because of its comfortable scripting and maintenance. The current project makes the usage of Pig easily, by filtering the records according to the requirement. Hive is used to join two or more tables easily with a simple query. It is more useful for joining more tables in order to achieve the CNE table. Sqoop can import data directly to Hive / H-base or HDFS directly without reducing. Because it is known for, map only process.

The other systems are using different strategies for processing and storing of data. The storing of data is easily generated but while accessing the data the complexity raises. In case of Hadoop, it is very effective in accessing, because it knows the data locality through Namenode. The other systems cannot handle the failures in remote. But, Hadoop can handle to some extent by re-promoting / re-running the failed tasks for better results. According to Chapter-3, the discussion based on back ground theory clearly conclude that Hadoop is good when compared to the other distributed systems and other processing mechanisms through its remote handling, reducing seek times and combining two datasets with key-value pairs.

## REFERENCES

[1] Katal, A.; Wazid, M.; Goudar, R.H., "Big data: Issues, hallenges, tools and Good practices," Contemporary Computing (IC3), 2013 Sixth International Conference on, vol., no., pp.404,409, 8-10 Aug. 2013

[2] "Welcome to Hive!" December, 2012; http://hive.apache.org/

[3] Ms. Vibhavari Chavan, Prof. Rajesh N.Phursule "Survey Paper on Big Data" International Journal of Computer Science and Information Technology volume 5(6).

[4] Jeffrey Shafer, Scott Rixner, and Alan L.Cox Rice University Houston TX " The Hadoop Distributed Filesystem :Balancing Portability and Performance "

[5] Poonam S.Patil, Rajesh N.Phursule "Survey Paper on Big Data Processing and Hadoop Components" International Journal of Science and Research.

[6] Sabia and Love Arora "Technologies to Handle Big Data : A Survey" Department of Computer Science and Engineering, GuruNanak Dev College University, Regional Campus, Jalandhar, India.

[7] Apache Hadoop. Available at http://hadoop.apache.org

[8] Apache HDFS. Available at http://hadoop.apache.org/hdfs

[9] Apache Hive. Available at http://hive.apache.org

[10] Apache HBase. Available at http://hbase.apache.org

[11] A community white paper developed by leading researchers across the United States "Challenges and Opportunities with Big Data"

[12] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy "Hive – A Petabyte Scale Data Warehouse Using Hadoop" By Facebook Data Infrastructure Team.

[13] Jeffrey Dean and Sanjay Google, Inc." MapReduce: Simplified Data Processing on Large Clusters"

[14] Sanjeev Dhawan1, Sanjay Rathee2, Faculty of Computer Science & Engineering, Research Scholar " Big Data Analytics using Hadoop Components like Pig and Hive" AIJRSTEM 13- 131; © 2013, AIJRSTEM.

[15] pig is referenced from "http://pig.apache.org/docs/r0.7.0/piglatin_ref2.html" "http://blog.cloudera.com/wp-content/uploads/2010/01/IntroToPig.pdf". "http://wiki.apache.org/pig/PigLatin"

[16] Hive join referenced feom "'https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Joins"