Removal of Semi-Duplicated and Fully Duplicate Shards using Hadoop Techniques for Elastic Search

Subhani Shaik, Nallamothu Naga Malleswara Rao

Abstract*: *Duplicate Records Identification is the most complex issues in information distribution center. This issue occurs when multiple databases are formed as a cluster. The duplicate records identification needs to be incorporated on both semi and completely copied records. Duplicate data identification is a technique for identifying all instances of numerous representation of some true values, client relationships, administration or information mining. Another application is Data Mining i.e. to adjust input information is important to develop helpful. In this manuscript a efficient algorithm is proposed for effective removal of the partial and fully copied data. Here the data in the database is divided into small parts called shards in which the duplicate data can be identified easily and accurately. In this paper dynamic duplication calculation is done with the assistance of Hadoop and mad reduce methods. The duplicate shards are identified and they are completely erased from the dataset. An Enhanced De-Duplicate Remover (EDDR) algorithm is proposed in this manuscript to erase the excess copied information and to effectively process the information on the final stage. To identify data repetition, the information utilize a few parameters, and afterward the recognized excess information will be erased by a few constraints as determined. The duplicate shards are removed and the memory wastage is reduced.

Index Terms*: *Duplicate Detection, Data Cleaning, Map Reduce. Information purifying, incomplete duplication, Hadoop, Map Reduce, Duplicate information Removal method,

I. INTRODUCTION

Database is the essential hotspot for each association and that can be used from various sources. Each heterogeneous source has diverse portrayal for same element, which prompts imitation in the database. In this way vast data shards are made by associations to clean the copy from the database. Information mining is the famous innovation which separates the valuable data required by the association for taking a superior choice. In an information store, a record that alludes to a similar true substance or protest is alluded as copy records. What's more, that copy record is additionally called as ―dirty data. Because of this grimy information numerous issues are happened as takes after:

1) Performance debasement—as extra futile shards request additionally handling and extra time is necessary to answer straightforward inquiries.

2) Quality trouble—the proximity of copies and different irregularities prompts twists in reports and mislead conclusions in view of the current information.

3) Increased cost — due to the extra volume of ineffective information, speculations are necessary on additional medium and additional calculation handling energy to keep the reaction time levels worthy. The issue of identifying and expelling these coped shards from an archive or database is recognized as shard de-duplication. It is additionally alluded as shard linkage [1], information cleaning [2]. Information de-duplication can be utilized to enhance information quality and honesty, which serves to re-utilization of existing information hotspots for new investigations, and to decrease expenses and endeavors in getting information. In the de-duplication procedure, interesting lumps of information, or byte designs, are distinguished and put away amid a procedure of investigation. As the assessment proceeds, different pieces are contrasted with the duplicate shards and at whatever point a match happens, the repetitive shard is replaced with a little reference that focuses to the original shard. De-duplication is a key activity in coordinating information from numerous sources. In the present condition because of expanding interest of web empowered gadgets we are getting enormous quantity of information. Presently days, information develops quickly from numerous social locales and media utilized by people groups like Facebook, twitter satellites, Airplanes, stock advertising and so forth. They all are producing bunches of information every second and all information put away and associated with cloud for sufficient memory. Examination and preparing of such kind of information is so troublesome [1]. It winds up hard to process this extensive measure of information utilizing close by Database administration devices or customary information handling application. To beat this sort of issue Hadoop is to be utilized. Hadoop gives Hadoop Distributed File System (HDFS) that can deal with unstructured information productively. In enormous information condition de-duplication and its procedures were utilized to expel copy information from bigdata [2].

II. LITERATURE SURVEY

As per the information granularity, De-duplication techniques can be sorted into two fundamental classes: document level De-duplication and shard level De-duplication, which is these days the most well-known system. In shard based De-duplication, the shard size can either be settled or varied. Another arrangement foundation is the area at which De-duplication is performed if information are de-copied at the customer, and after that it is called source-based De-duplication, generally target-based.

Revised Manuscript Received on 24 February 2019.

Subhani Shaik, Department of CSE, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Nallamothu Naga Malleswara Rao, Department of IT, RVR & JC College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India.

In source-based De-duplication, the customer first hashes every data portion he wishes to transfer and sends these outcomes to the capacity supplier to check whether such information are now put away: in this way just "not de-copied" information sections will be transferred by the client on the cloud. While De-duplication at the customer side can accomplish transfer speed funds, it tragically can make the framework helpless against side-channel assaults whereby aggressors can instantly find whether a specific information is put away or not. Then again, by de-copying information at the capacity supplier.

M. Sathiamoorthy et.al [1] Proposed a paper that characterizes what information purifying is and the components that information contains. Information cleaning process are quickly portrayed i.e. arranging, breaking down for cleaning, actualizing robotization, attaching missing information, and checking information. It likewise incorporates noteworthiness of information quality and the difficulties emerging while at the same time cleaning information. Different methodologies for cleaning information are proposed.

D. Borthakur et.al [2] Proposed a method that are ordinarily used to recognize comparable field sections, and we show a broad arrangement of copy discovery calculations that can identify roughly copy records in a database. We additionally cover numerous strategies for enhancing the productivity and adaptability of rough copy location calculations.

D. Borthakur et.al [3] Proposed a diverse sort of blunders modify information predominance from the heterogeneous spaces. As an option, assess every target portrayal by methods for a most likely composite like technique, to recognize whether the protest is true or not. This paper has given itemized study examination and foundation on copy recognition in various leveled information.

S. Yan et.al [4] Proposed a method for recognizing record duplication [6]. Character based system haggles well with the typographical mistakes. Be that as it may, once in a while typographical traditions prompt reworking of words. Character related system flops keeping in mind the end goal to think about such sort of strings. The token base system is utilized to defeat this issue.

Yang and Chen [6] proposed a versatile information component in view of similitude estimation among the partitioned data pieces in Cloud figuring. They proposed an information piece system for NoSqldatabases in light of the trait structure tree. This component at first parts the single private characteristic into a consolidate properties. What's more, viewed as the multicast resource allocation issue for video spilling in Orthogonal Frequency Division Multiple Access (OFDMA) frameworks.

III. COPY DETECTION & ELIMINATION

In the phase, just a single duplicate shard of correct copied shard are held and dispensed with other copy records [4]. The final procedure is imperative to create cleaned information. Prior to the final procedure, the comparability edge esteems are figured for every one of the shard which are accessible in the informational index of the database where every shard is compared with all the remaining shards in the database. The similarity limit esteems are essential for the data clearance procedure. In the data clearance procedure, select every conceivable match from each group and look at shards inside the bunch utilizing the chose traits. The greater part of the end forms look at shards inside the group as it were. In some cases different groups may have copy shards, same incentive as of different bunches.

The accompanying systems are utilized to distinguish or identify copies of shards and clearing or removing of duplicate shard copies.

i. Get edge an incentive from LOG table from the database.

ii. Ascertain conviction factor i,e shard ID and duplicate level.

iii. Ascertain information quality factor for each shard based on its uniqness.

iv. Recognize or distinguish shard copies utilizing assurance factor, threshold esteem and information quality factor .

v. Clear copy shards in view of information quality, threshold esteem, number of missing worth and range of each field esteem.

vi. Hold just a single copy shard which is having high information quality, high limit esteem and high certainty factor .

3.1 Copy Data Identification/Detection Rule

Copy shard location is the way toward recognizing extraordinary or numerous shards that allude to one of a kind genuine substance or question if their comparability surpasses a specific cutoff esteem. The shards comprise of various fields, making the copy recognition issue substantially more confused [13]. A constraint based approach is proposed for the copy shards recognition issue. This administer is produced with the additional confinement to get great consequence of the principles. These principles indicate the conditions and criteria for two shards to be named copies. A general if then else lead is utilized as a part of this exploration work for the copy information ID and copy information end. An administer will by and large be of the frame:

Consider Dataset DS where S is a shard having ID

(S_1,S_2,…..S_n) ε (DS)

on the off chance that <condition >

at that point <action >

The activity part of the lead is enacted or terminated when the conditions are fulfilled. The intricate predicates and outer capacity references might be contained in both the condition and activity parts of the administer [10]. In existing copy recognition and end strategy, the standards are characterized for the particular subject informational index as it were. These standards are not pertinent for another subject informational index. In copy information identification govern, edge estimations of shard sets and sureness factors are essential.

Normally copy information disposal is executed as the last advance and this progression needs to occur while coordinating two sources or performed on an officially incorporated source. The mix of credits can be utilized to recognize copy shards. In the copied shard clearance, just a single best duplicate of copy shard must be held and staying copy shards ought to be wiped out. Copied shards are recognized utilizing assurance factor and limit esteem. Copy information is disposed of in view of the quantity of missing worth, scope of each field esteem, information nature of each field esteem and portrayal of information.

Assurance factor= (Primary Key(S1(ID))=1)
Limit Esteem=Unique($S_{1,}S_2\ldots S_n$)>1
Each shard copied information or general similarity of two shards are resolved from the similarities of chosen shard fields. A case of copy information is that two shards with (i) indistinguishable field value (ii) high limit value (ii) are of a similar length, and (iii) have a place with a similar kind of information, are copied. The administer can be spoken to as: high limit esteem ^ same length of field esteem ^ same portrayal of information with slight changes → not copied shard.

The accompanying components are utilized as a part of the copy information
disposal run the show.
i. Number of missing values.
ii. Scope of qualities.
iii. Information portrayal.
iv. Absence of value.
v. Limit of value.

Copy shards are recognized by utilizing particular and high segregation control properties. As a rule, copy shards can have such a large number of missing fields. Henceforth, shards can be wiped out in light of the quantity of missing values in each copy shard. Copy shard is dispensed with if the copy shard is has more missing parameters than other copy shards. The size and scope of each field esteem is figured and contrasted with other copy information field with wipe out low quality copy information. For instance, now and again copy information can have easy route frame or condensing. In this way, the scope of each field esteem is figured to evacuate copy information which have low range than other copy shards. More often than not shard is copied due to the diverse configuration utilized for information portrayal. For instance, 'M' and 'F' are utilized for male and female however '1' and '0' are utilized for sex portrayal. In this way, there is a need to recognize correct arrangement for each field portrayal. Generally, copied shards are distinguished and wiped out in view of the edge estimation of each copy shard. Most noteworthy edge value of shard is held and least limit value copy shards are wiped out.

3.2 De-Duplication HDFS Architecture

In reality, the measure of information is developing exponentially; little and medium ventures or instructive association will experience the ill effects of inadequate space trouble. In this paper, to confront this test for capacity frameworks, a dynamic de-duplication head to enhance the utility of storage room in HDFS is proposed, which is right now a standout amongst other answers for huge data.Then, the proposed framework will be displayed as the accompanying. In this area, we will make a depiction of the framework system and the points of interest of the parts, calculations, and stream diagrams independently.

Looking through various databases, notwithstanding, brings about the recovery of various copy references. Likewise, because of the idea of the distributing cycle in the field of drug, gathering edited compositions and full-content articles detailing a similar data are frequently recovered simultaneously. Moreover, albeit numerous have gotten out against such practice, a few creators "cut, reformat, or repeat material from an examination" [4], which makes dreary, copy, and excess productions. Expelling these copy references, otherwise called de-duplication, can be a tedious procedure however is important to guarantee a substantial and solid pool of concentrates for incorporation in an efficient audit

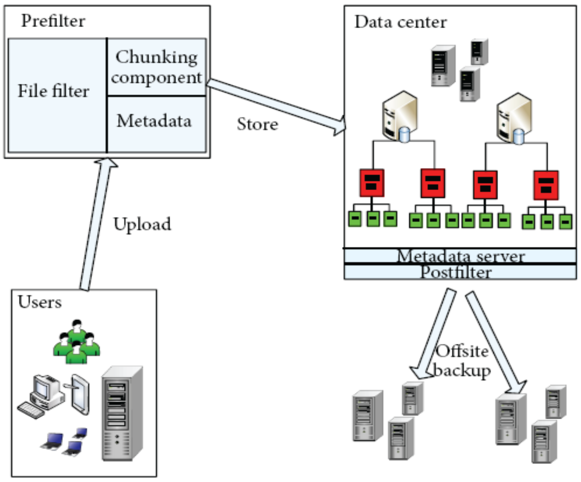


Fig-1 Proposed method framework

3.3 Algorithm to eliminate Duplication Data

```
Input: {S(ID), Table(ID),t} where S is shard, t is table name.
Output :{Unique(Table),Count(S)} displays table without duplicate values and count of
            removed shards
Algorithm EDDR(S(ID),Table(ID),t)
{
S= {S1, S2,……Sn}ε Dataset (DS)
MAX=Count(Tablerows)
for each $i in {S1, S2,……Sn}
for ( $i=1;$i<MAX;$i++)
for ( $j=1;$j<MAX;$j++)
for ( $k=$j+1;$k<=MAX;$k++)
if ((Sj==Sk)
Sk==NULL;
return Unique (Table(ID));
else
return Unique(Table(ID));
if(match(Sj)>match(Sk))
delete(Sk)
return DS
}
```

Σ[((radix) position × alphvalue) % m] (1)

where alpha value is set apart from 0-9 and aA=10, bB=11,… … ,

Input: Table with various information configuration.

Output: Uniform arrangement table with additional attached quality having unique values

Algorithm

Start

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
For credit j = 1 to keep going characteristic, n
For push I = 1 to last column, m
1. Bring quality parameters into uniform organization
2. Expel the exceptional character
3. Remove  the variety of characteristic esteems
4. Change over every one of the qualities into numeric shape
5. Put the numeric incentive into affixed characteristic isolated with comma (,)
end
```

The proposed EDDR algorithm effectively scans all the duplicate shards and elimanets them resulting in reduction of memory wastage.

3.4 Hadoop Technique for Duplicate Shards Removing

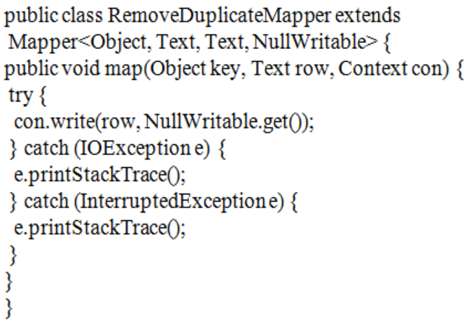
Utilize MD5 & SHA-1 hash capacities to figure the shards hash level and after that pass the incentive to HBase.

- Compare the new hash an incentive with the current parameters. On the off chance that it exists prior in HBase de-duplication table, HDFS will check the quantity of connections, and if the number isn't zero, the counter will be increased by one.
- HDFS will store source documents, which are transferred by clients, and comparing join records, which are consequently created. Connection documents record the source record's hash esteem and the coherent way of the source record.

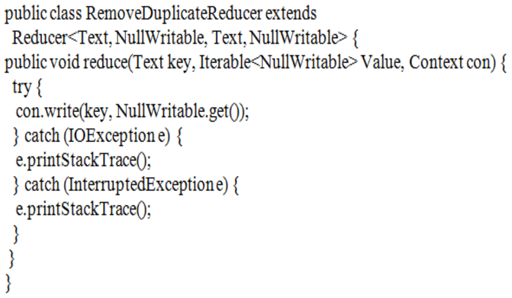
A portion of the key things to note in this approach include:

- shard level de-duplication to keep the file as little as conceivable with a specific end goal to accomplish high query effectiveness.
- MD5 & SHA-1 parameters are combined to dodge unintentional impact.

The following map and reduce methods are applied on the data set for effective duplicate data removing.



```
public class RemoveDuplicateMapper extends
 Mapper<Object, Text, Text, NullWritable> {
public void map(Object key, Text row, Context con) {
 try {
  con.write(row, NullWritable.get());
 } catch (IOException e) {
  e.printStackTrace();
 } catch (InterruptedException e) {
  e.printStackTrace();
 }
 }
 }
```



```
public class RemoveDuplicateReducer extends
 Reducer<Text, NullWritable, Text, NullWritable> {
public void reduce(Text key, Iterable<NullWritable> Value, Context con) {
 try {
  con.write(key, NullWritable.get());
 } catch (IOException e) {
  e.printStackTrace();
 } catch (InterruptedException e) {
  e.printStackTrace();
 }
 }
}
```

IV. Results

The proposed algorithm is applied on the considered dataset and the data which is identified by elastic search methods are removed in this stage. The proposed algorithm effectively removes all the partial and full copied shards in the database thus reducing the memory wastage.

Based on the dataset size the time taken for execution process to remove the duplicate is illustrated in below figure

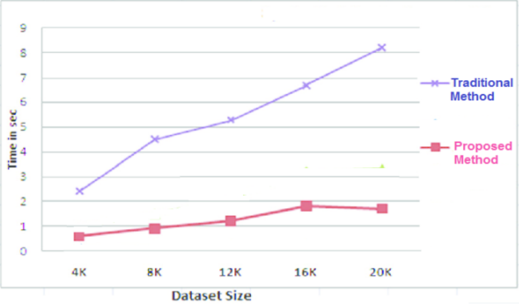


Fig-2 Execution time for removing duplicate shards

By deleting the partial and fully copied shards from the bigdata database a huge amount of memory can be reused and the wastage of memory is reduced. The memory wastage reduction levels of the existing and proposed methods are illustrated in below figure.

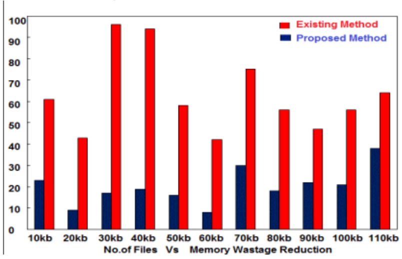
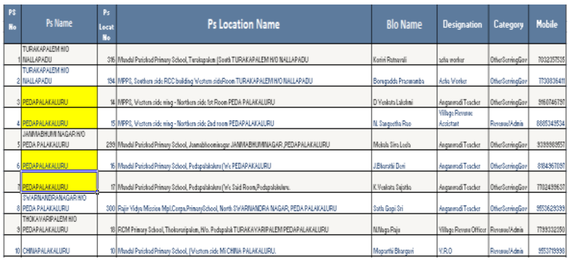


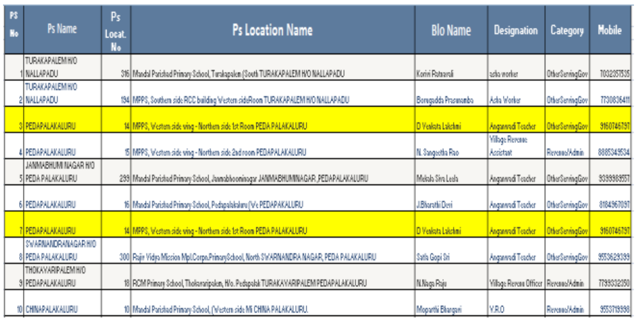
Fig-3 Memory Reduction Level Calculation.

By using Elastic search method the duplicate data is identified from the election commission. The dataset contains 286 shards where 73 shards are copied.



PS No	Ps Name	Ps Locat No	Ps Location Name	Blo Name	Designation	Category	Mobile
1	TURAKAPALEM H/O NALLAPADU	316	Mandal Parishad Primary School, Turakapalem (South TURAKAPALEM H/O NALLAPADU	Korivi Ratnavali	asha worker	OtherServingGov	7032357535
2	TURAKAPALEM H/O NALLAPADU	194	MPPS, Southern side RCC building Western sideRoom TURAKAPALEM H/O NALLAPADU	Borugadda Prasunamba	Asha Worker	OtherServingGov	7730836411
3	PEDAPALAKALURU	14	MPPS, Western side wing - Northern side 1st Room PEDA PALAKALURU	D Venkata Lakshmi	Anganwadi Teacher	OtherServingGov	9160746797
4	PEDAPALAKALURU	15	MPPS, Western side wing - Northern side 2nd room PEDAPALAKALURU	N. Sangeetha Rao	Village Revenue Assistant	Revenue/Admin	8885349534
5	JANMABHUMI NAGAR H/O PEDA PALAKALURU	299	Mandal Parishad Primary School, Janmabhoominagar JANMABHUMINAGAR ,PEDAPALAKALURU	Mekala Siva Leela	Anganwadi Teacher	OtherServingGov	9399969557
6	PEDAPALAKALURU	16	Mandal Parishad Primary School, Pedapalakaluru (W/s PEDAPAKALURU	J.Bharathi Devi	Anganwadi Teacher	OtherServingGov	8184967097
7	PEDAPALAKALURU	17	Mandal Parishad Primary School, Pedapalakaluru (W/s Said Room,Pedapalakaluru.	K.Venkata Sujatha	Anganwadi Teacher	OtherServingGov	7702499637
8	SWARNANDRANAGAR H/O PEDA PALAKALURU	300	Rajiv Vidya Mission Mpl.Corpn.PrimarySchool, North SWARNANDRA NAGAR, PEDA PALAKALURU	Satla Gopi Sri	Anganwadi Teacher	OtherServingGov	9553629399
9	THOKAVARIPALEM H/O PEDAPALAKALURU	18	RCM Primary School, Thokavaripalem, H/o. Pedapalak TURAKAVARIPALEM PEDAPALAKALURU	N.Naga Raju	Village Revenue Officer	Revenue/Admin	7799332350
10	CHINAPALAKALURU	10	Mandal Parishad Primary School, (Western side Mi CHINA PALAKALURU.	Moparthi Bhargavi	V.R.O	Revenue/Admin	9553719998

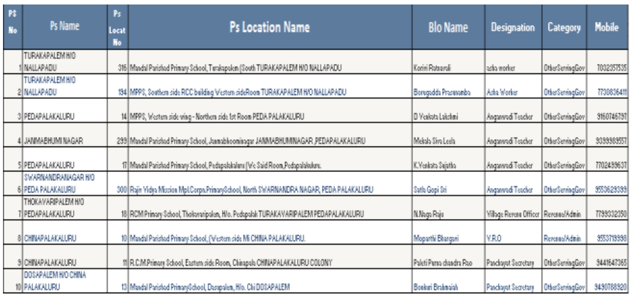
Table-1: Semi Duplicate data identified.



PS No	Ps Name	Ps Locat. No	Ps Location Name	Blo Name	Designation	Category	Mobile
1	TURAKAPALEM H/O NALLAPADU	316	Mandal Parishad Primary School, Turakapalem (South TURAKAPALEM H/O NALLAPADU	Korivi Ratnavali	asha worker	OtherServingGov	7032357535
2	TURAKAPALEM H/O NALLAPADU	194	MPPS, Southern side RCC building Western sideRoom TURAKAPALEM H/O NALLAPADU	Borugadda Prasunamba	Asha Worker	OtherServingGov	7730836411
3	PEDAPALAKALURU	14	MPPS, Western side wing - Northern side 1st Room PEDA PALAKALURU	D Venkata Lakshmi	Anganwadi Teacher	OtherServingGov	9160746797
4	PEDAPALAKALURU	15	MPPS, Western side wing - Northern side 2nd room PEDAPALAKALURU	N. Sangeetha Rao	Village Revenue Assistant	Revenue/Admin	8885349534
5	JANMABHUMI NAGAR H/O PEDA PALAKALURU	299	Mandal Parishad Primary School, Janmabhoominagar JANMABHUMINAGAR ,PEDAPALAKALURU	Mekala Siva Leela	Anganwadi Teacher	OtherServingGov	9399969557
6	PEDAPALAKALURU	16	Mandal Parishad Primary School, Pedapalakaluru (W/s PEDAPAKALURU	J.Bharathi Devi	Anganwadi Teacher	OtherServingGov	8184967097
7	PEDAPALAKALURU	14	MPPS, Western side wing - Northern side 1st Room PEDA PALAKALURU	D Venkata Lakshmi	Anganwadi Teacher	OtherServingGov	9160746797
8	SWARNANDRANAGAR H/O PEDA PALAKALURU	300	Rajiv Vidya Mission Mpl.Corpn.PrimarySchool, North SWARNANDRA NAGAR, PEDA PALAKALURU	Satla Gopi Sri	Anganwadi Teacher	OtherServingGov	9553629399
9	THOKAVARIPALEM H/O PEDAPALAKALURU	18	RCM Primary School, Thokavaripalem, H/o. Pedapalak TURAKAVARIPALEM PEDAPALAKALURU	N.Naga Raju	Village Revenue Officer	Revenue/Admin	7799332350
10	CHINAPALAKALURU	10	Mandal Parishad Primary School, (Western side Mi CHINA PALAKALURU.	Moparthi Bhargavi	V.R.O	Revenue/Admin	9553719998

Table-2: Fully Copied identified data.

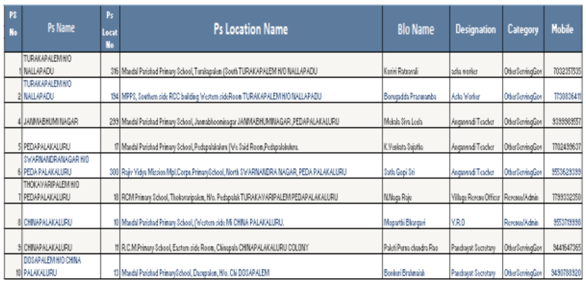
After removing the semi copied data from Table-1 the dataset is displayed in Table-3.



PS No	Ps Name	Ps Locat No	Ps Location Name	Blo Name	Designation	Category	Mobile
1	TURAKAPALEM H/O NALLAPADU	316	Mandal Parishad Primary School, Turakapalem (South TURAKAPALEM H/O NALLAPADU	Korini Ratnavali	asha worker	OtherServingGov	7032357535
2	TURAKAPALEM H/O NALLAPADU	194	MPPS, Southern side RCC building Western sideRoom TURAKAPALEM H/O NALLAPADU	Borugadda Prasunamba	Asha Worker	OtherServingGov	7730836411
3	PEDAPALAKALURU	14	MPPS, Western side wing - Northern side 1st Room PEDA PALAKALURU	D Venkata Lakshmi	Anganwadi Teacher	OtherServingGov	9160746797
4	JANMABHUMI NAGAR	293	Mandal Parishad Primary School, Janmabhoominagar JANMABHUMINAGAR ,PEDAPALAKALURU	Mekala Siva Leela	Anganwadi Teacher	OtherServingGov	9399969557
5	PEDAPALAKALURU	17	Mandal Parishad Primary School, Pedapalakaluru (W/c Said Room,Pedapalakaluru.	K.Venkata Sujatha	Anganwadi Teacher	OtherServingGov	7702499637
6	SWARNANDRANAGAR H/O PEDA PALAKALURU	300	Rajiv Vidya Mission Mpl.Corpn.PrimarySchool, North SWARNANDRA NAGAR, PEDA PALAKALURU	Satla Gopi Sri	Anganwadi Teacher	OtherServingGov	9553629399
7	THOKAVARIPALEM H/O PEDAPALAKALURU	18	RCM Primary School, Thokavaripalem, H/o. Pedapalak TURAKAVARIPALEM PEDAPALAKALURU	N.Naga Raju	Village Revenu Officer	Revenue/Admin	7799332350
8	CHINAPALAKALURU	10	Mandal Parishad Primary School, (Western side Mi CHINA PALAKALURU.	Moparthi Bhargavi	V.R.O	Revenue/Admin	9553719998
9	CHINAPALAKALURU	11	R.C.M.Primary School, Eastern side Room, Chinapala CHINAPALAKALURU COLONY	Palcti Puna chandra Rao	Panchayat Secretary	OtherServingGov	9441647365
10	DOSAPALEM H/O CHINA PALAKALURU	13	Mandal Parishad PrimarySchool, Dosapalem, H/o. Chi DOSAPALEM	Bonkuri Brahmaiah	Panchayat Secretary	OtherServingGov	9490788920

Table-3: Dataset After Removing Partially copied data.

After removing the Fully copied data from Table-2 the dataset is displayed in Table-4.



PS No	Ps Name	Ps Locat No	Ps Location Name	Blo Name	Designation	Category	Mobile
1	TURAKAPALEM H/O NALLAPADU	316	Mandal Parishad Primary School, Turakapalem (South TURAKAPALEM H/O NALLAPADU	Korini Ratnavali	asha worker	OtherServingGov	7032357535
2	TURAKAPALEM H/O NALLAPADU	194	MPPS, Southern side RCC building Western sideRoom TURAKAPALEM H/O NALLAPADU	Borugadda Prasunamba	Asha Worker	OtherServingGov	7730836411
4	JANMABHUMI NAGAR	293	Mandal Parishad Primary School, Janmabhoominagar JANMABHUMINAGAR ,PEDAPALAKALURU	Mekala Siva Leela	Anganwadi Teacher	OtherServingGov	9399969557
5	PEDAPALAKALURU	17	Mandal Parishad Primary School, Pedapalakaluru (W/c Said Room,Pedapalakaluru.	K.Venkata Sujatha	Anganwadi Teacher	OtherServingGov	7702499637
6	SWARNANDRANAGAR H/O PEDA PALAKALURU	300	Rajiv Vidya Mission Mpl.Corpn.PrimarySchool, North SWARNANDRA NAGAR, PEDA PALAKALURU	Satla Gopi Sri	Anganwadi Teacher	OtherServingGov	9553629399
7	THOKAVARIPALEM H/O PEDAPALAKALURU	18	RCM Primary School, Thokavaripalem, H/o. Pedapalak TURAKAVARIPALEM PEDAPALAKALURU	N.Naga Raju	Village Revenu Officer	Revenue/Admin	7799332350
8	CHINAPALAKALURU	10	Mandal Parishad Primary School, (Western side Mi CHINA PALAKALURU.	Moparthi Bhargavi	V.R.O	Revenue/Admin	9553719998
9	CHINAPALAKALURU	11	R.C.M.Primary School, Eastern side Room, Chinapala CHINAPALAKALURU COLONY	Palcti Puna chandra Rao	Panchayat Secretary	OtherServingGov	9441647365
10	DOSAPALEM H/O CHINA PALAKALURU	13	Mandal Parishad PrimarySchool, Dosapalem, H/o. Chi DOSAPALEM	Bonkuri Brahmaiah	Panchayat Secretary	OtherServingGov	9490788920

Table-4: Dataset After Removing Fully copied data.

V.CONCLUSION

In this paper, we have proposed copied shards removal strategy, which helps to remove partial copied data and fully copied data from the specified dataset. Since at long last the measures of information preprocessed are considerably small than the first information after the evaluation procedure, the quantity of mappers in Map Reduce is reduced and in the final stage removing duplicate data can be avoided. The proposed method uses Hadoop method which is very efficient in data cleaning process in the data base which reduces memory wastage.

REFERENCES

1. M. Sathiamoorthy,M. Asteris, D. Papailiopoulos et al., "XORing elephants: novel erasure codes for big data," Proceedings of the VLDB Endowment, vol. 6, no. 5, pp. 325–336, 2013.
2. D. Borthakur, The Hadoop Distributed File System: Architecture and Design, 2007, http://hadoop.apache.org/docs/r0.18.0/ hdfs design.pdf.
3. Thorsten Papenbrock, ArvidHeise, and Felix Naumann,‟ Progressive Duplicate Detection‟ IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 25, no. 5, 2014.
4. S. Yan, D. Lee, M. yen Kan, and C. L. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," in International Conference on Digital Libraries (ICDL), 2007.
5. M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," Data Mining and Knowledge Discovery, vol. 2, no. 1, 1998.
6. X.Dong, A.Halevy, and J.Madhavan, "Reference reconciliation in complexinformation spaces," in Proceedings of the International Conference on Management of Data (SIGMOD), 2005.
7. S.E.Whang, D.Marmaros, and H.Garcia-Molina, "Pay-as-you-go entity resolution" IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 25, no. 5, 2012.
8. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicat record detection: A survey," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 19, no. 1, 2007.
9. S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, ―Robust and Efficient Fuzzy Match for Online Data Cleaning,‖ Proc. ACM SIGMOD Int‟l Conf. Management of Data, pp. 313-324, 2003.
10. A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios, ―Duplicate Record Detection: A Survey,‖ IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16, Jan. 2007.
11. H.B.Newcombe, J.M. Kennedy, S. Axford, and A. James, ―Automatic Linkage of Vital Records,‖ Science, vol.130, no. 3381, pp. 954-959, Oct. 1959
12. I.P. Fellegi and A.B. Sunter, ―A Theory for Record Linkage,‖ J. Am. Statistical Assoc., vol. 66, no. 1, pp. 1183-1210, 1969.
13. M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, ―Adaptive Name Matching in Information Integration,‖ IEEE Intelligent Systems, vol. 18, no. 5, pp. 16-23, Sept./Oct. 2003.
14. D. Jeffrey and G. Sanjay, "MapReduce: simplified data processing on large clusters", Communications of the ACM 2008, vol. 51, no. 1, (2008), pp.107-113.
15. P. Pupunwiwat and B. Stantic, "Location Filtering and Duplication Elimination for RFID Data Streams", International Journal of Principles and Applications of Information Science and Technology, vol.1, no.1, (2007), pp 29-43.
16. H. Mahdin and J. Abawajy, "An Approach for Removing Redundant Data from RFID Data Streams", Sensors 2011, vol. 11, (2011), pp. 9863-9877.
17. W. M. Choi, J. S. Jeong, B. J. Kim and D. G. Kim, "Garlic cold storage", Korea, 1020110012155, (2011).
18. J. S. Seo, M. S. Kang, Y. G. Kim, C. B. Sim, S. C. Joo and C. S. Shin, "Implementation of Ubiquitous Green-house Management System Using Sensor Network", Journal of Korean Society for Internet Information, vol. 9, no. 3, (2008), pp. 129-139.
19. K. O. Kim, K. W. Park, J. C. Kim, M. S. Chang and E. K. Kim, "Establishment of Web-based Remote Moni-toring System for Greenhouse Environment", Journal of The Korea Institute of Electronic Communication Sciences, vol. 6, no. 1, (2011), pp. 77-83.
20. C.-I. Fan, S.-Y. Huang, andW.-C. Hsu, "Hybrid data deduplication in cloud environment," in Proceedings of the International Conference on Information Security and Intelligence Control (ISIC '12), pp. 174–177, Yunlin, Taiwan, 2012.
21. W. Xia, H. Jiang, D. Feng, L. Tian, M. Fu, and Z. Wang, "PDedupe: exploiting parallelism in data deduplication system," in Proceedings of the IEEE 7th International Conference on Networking, Architecture and Storage (NAS '12), pp. 338–347, Fujian, China, 2012.
22. B.Mao,H. Jiang, S.Wu, Y. Fu, and L. Tian, "SAR: SSDassisted restore optimization for deduplication-based storage systems in the cloud," in Proceedings of the IEEE 7th International Conference on Networking, Architecture and Storage (NAS '12), pp. 328–337, Fujian, China, 2012.
23. B. Fan, W. Tantisiriroj, L. Xiao, and G. Gibson, "DiskReduce: RAID for data-intensive scalable computing," in Proceedings of the 4th Annual Petascale Data Storage Workshop (PDSW '09), pp. 6–10, Portland, Ore, USA, November 2009.
24. Y. Zhang, Y. Wu, and G. Yang, "Droplet: a distributed solution of data deduplication," in Proceedings of the ACM/IEEE 13th International Conference on Grid Computing (GRID '12), pp. 114–121, Beijing, China