SSSIHL DATA CLEANING FRAMEWORK USING HADOOP

Authors

1. J Bhanu Teja, MTech, SSSIHL*, Email: bhanu257@gmail.com
2. Phani Krishna Kandala, M.Sc., M.Tech, AVP (Actuarial) & Visiting Faculty, SSSIHL (kandala.phanikrishna@gmail.com)
3. Srivarun V, Technical Manager, SSSIHL(srivarunv@sssihl.edu.in)
4. Poorna Viswanadha Sravan Nukala, MTech, SSSIHL (poorna.sravan@gmail.com)
5. Pallav Kumar Baruah, Head of Department, Department of Mathematics and Computer Science SSSIHL. Email: pkbaruah@sssihl.edu.in
6. Satya Sai Mudigonda, M.Sc., AIAI, CPCU, PMP, Senior Tech Actuarial Consultant & Hon. Professor, SSSIHL (satyasaibabamudigonda@sssihl.edu.in)

* Sri Sathya Sai Institute of Higher Learning (www.sssihl.edu.in)

Abstract:

For any analysis, clean data is a prerequisite. This means to identify wrong data input, outliers, omissions or duplicates etc. To meet this absolute requirement, SSSIHL has developed a data cleaning framework which is cost & time efficient and maintains data quality intact. The SSSIHL DATA CLEANING FRAMEWORK is built on open source and uses distributed file computing to clean the data in parallel. This framework enables cleansing data at three different business levels and can be customized with minimum effort specific to each business need. The output can be directly used as input to other processes. This framework can also handle large sets of data and its performance is positively correlated to the size of the data.

1. Introduction

SSSIHL has created a unique data cleaning framework which works based on SSSIHL DATA CLEANING FRAMEWORK and uses mapreduce paradigm. This work is organized into 7 sections. Section 2 refers to literature review in this area of work. This framework helps clean the data at three different business levels. The SSSIHL DATA CLEANING FRAMEWORK allows for easy customization at all three levels to meet the requirements of each organization. Each level is described in section 3. Implementation and Architecture details are provided in Section 4 of this paper. Section 5, 6 refers about the various data sets used for cleaning the data and also about the performance improvements. Section 7 refers to conclusion and future work.

2. Literature Survey

Mong Li Lee [1] has examined the problem of detecting and removing duplicates records. Several different techniques to pre-process the records before sorting them so that potentially matching records will be brought to close neighborhood subsequently. Taoxin Peng [12]in his paper presented a framework for How to improve the efficiency while performing data cleaning and How to improve the degree of automation when performing data cleaning, which provides an approach to managing data cleaning in data warehouses by focusing on the use of data quality dimensions, and decoupling a cleaning process into several sub-processes Mong Li Lee et al[7] had proposed a generic knowledge based framework for effective data cleaning that implements existing cleaning strategies and more. Rahm, Hong Hai Do [2] defined the various data cleaning problems and current approaches like Single source problems and Multisource problems and Data quality problems. Nidhi Choudhary[3] had done various studies on approaches of Datacleaning. Joseph M. Hellerstein[4] discuss the quantitative cleaning of large databases, and defines the approaches to improve data quality. Heiko Müller et al[6] discussed the various data cleaning process and compare the data cleaning frameworks. Kofi Adu-Manu Sarpong et. al [8] had conceptualized the data cleansing process from data acquisition to data maintenance. Data Cleansing is an activity involving a process of detecting and correcting the errors. Rajashree Y.Patil et al [5] discussed various data cleaning algorithms for data warehouse.

3. Levels of Data Cleaning

Basic Level Data Cleaning: Data Element Level

- Each Data element type and length can be identified and checked against the definition given by the owner / data manual.
- Check whether the data elements have null/missing values.
- Check whether the data elements are in the correct format as prescribed. E.g., Date format.
- Check duplicates for primary columns.
- Check for outliers in the data.
- Produce a report of anomalies.
- Suggest potential fix for anomalies identified.

Intermediate Level Data Cleaning: Basic Rules of Business at organisational level

- Define a set of basic business rules (eg: F_1, F_2, F_3,....., F_n) at organisational level as defined by owner/data manual.
- Validate the data using rule engine and report anomalies.
- Suggest potential fix for anomalies identified.

Advanced Level Data Cleaning: Complex rules of Business at industry level

- Define a set of complex rules C_1, C_2, C_3,....,C_m at industry levels as defined by regulator, industry leaders, standard organisations etc.
- Validate the data using rule engine and report anomalies.
- Suggest potential fix for anomalies identified.

4. Implementation Details

4.1 Steps in Data Cleaning Process

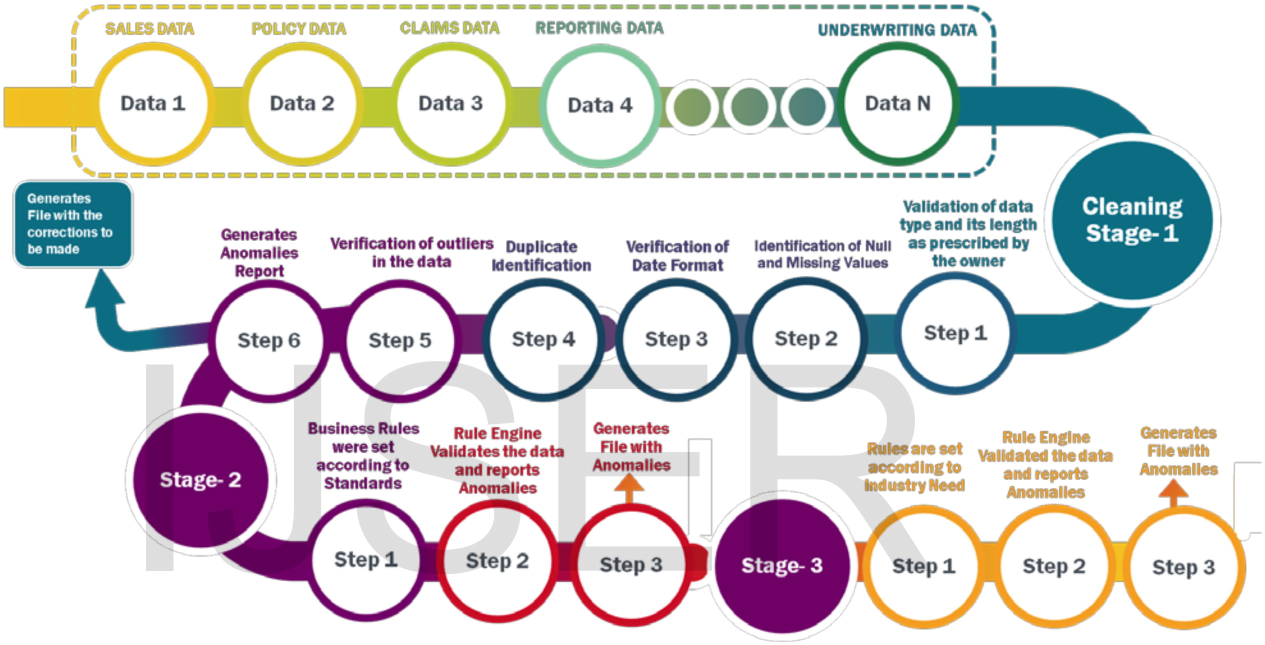

SALES DATA
POLICY DATA
CLAIMS DATA
REPORTING DATA
UNDERWRITING DATA
Data 1
Data 2
Data 3
Data 4
Data N
Cleaning Stage- 1
Validation of data type and its length as prescribed by the owner
Identification of Null and Missing Values
Verification of Date Format
Duplicate Identification
Verification of outliers in the data
Generates Anomalies Report
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Generates File with the corrections to be made
Stage- 2
Business Rules were set according to Standards
Rule Engine Validates the data and reports Anomalies
Generates File with Anomalies
Step 1
Step 2
Step 3
Stage- 3
Rules are set according to Industry Need
Rule Engine Validates the data and reports Anomalies
Generates File with Anomalies
Step 1
Step 2
Step 3

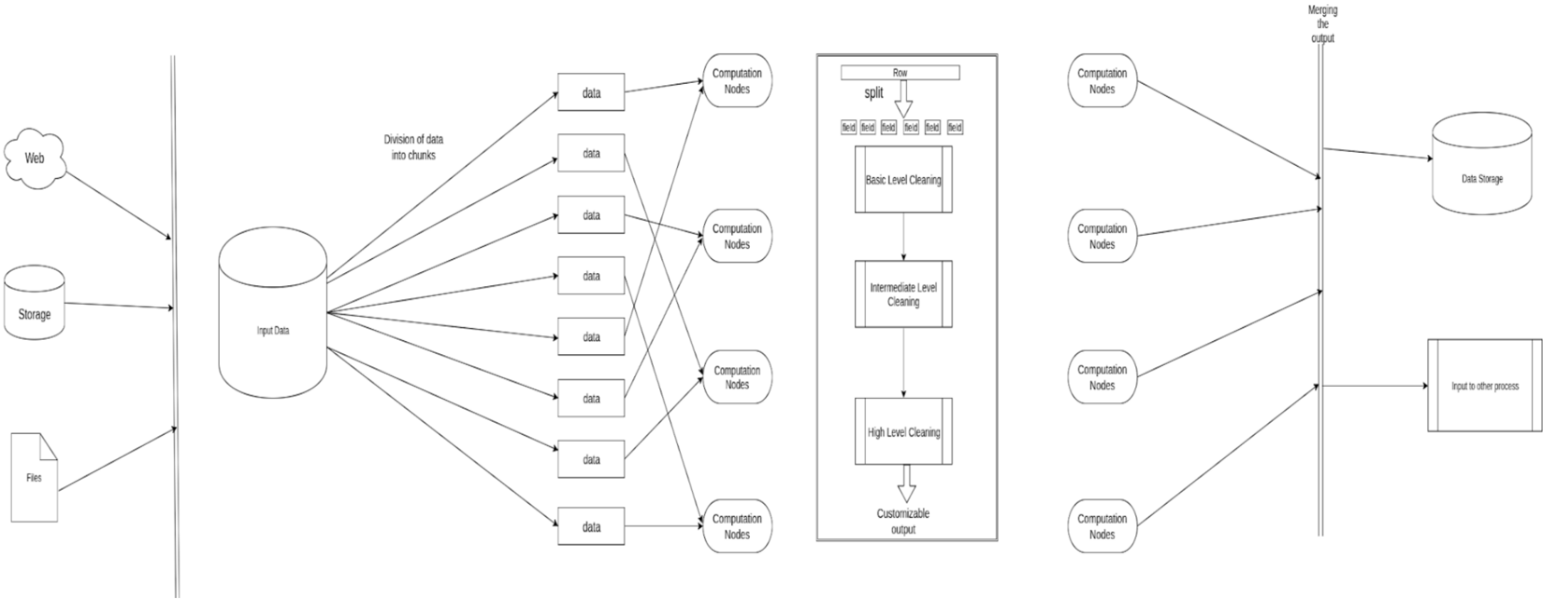
Figure 4.1 Steps in Data Cleaning Process

Figure 4.1 explains about the flow of data cleaning process. As the data arrives from different sources with different fields, data cleaning happens in three stages. Each stage is composed of different type of verification checks to produce better quality of data. At the end of each stage anomalies are identified and the report is generated to identify the potential data errors.

4.2 Algorithm:

```
For each row in the file:
    Split the row based on the specified delimiter.
    For each data element
        Check whether the data element is null value.
        Check the type of data element .
        Check whether the data elements are within the specified lengths.
        If data element is a numerical value then Check whether the numerical value is within the specified limits
        Else if data element is text value:
                If it is an Id then check whether it is in id format.
            Else check for different pattern.
        Else If boolean variable check whether the value is one of the two accepted values.
        If date variable then ensure that date is a legal date.
For each row obtained from above cleaning:
    Clean the row according to the basic rules $F_1$, $F_2$, $F_3$,....., $F_n$.
    Check the validation of  different combination of data fields.
For each row obtained from above:
    Clean the row according to the complex rules $C_1$, $C_2$, $C_3$,....,$C_n$.
```

Divide the data into chunks and give it to each node. At each node the cleaning mentioned in the above algorithm is performed. The framework is given in the figure 1.



Unique features of the SSSIHL® DATA CLEANING FRAMEWORK

1. Parallelisation using distributed file computation using HDFS.
2. Built on Open Source.
3. Cleaning using complex business rules specific to each line or class of business.
4. Framework allows easy customization to specific business needs with minimum effort.
5. User friendly output that can be directly integrated into other enterprise level sub-systems.
6. Capability to handle large sets of data input in batches.
7. Performance improvement(speed) is positively correlated to the size of the data.
8. 99.99% accuracy when compared with other validation methods.
9. Dynamic data validation while uploading data through websites.

The performance of the framework was demonstrated using two pilot projects. First one using synthetic data and the second one using real data. The details of data, performance and analysis of the framework are described below.

4.2 Experimental setup

Hadoop single system 3.4 GHz Quad Core Intel Core i5 CPU, 16GB memory with 1TB hard disk.

Hadoop Cluster: 3 systems each of them with 3.4 GHz Quad Core Intel Core i5 CPU, 16GB memory with 1TB hard disk.

A node is assigned the master status and other nodes are assigned the slave status. Compute Ratio Balancer reads block information from the NameNode and runs the distribution algorithm in order to distribute the data based on the computing capacity of the nodes.

5. Pilot project using SSSIHL DATA CLEANING FRAMEWORK on Synthetic data.

Description:

With the help of industrial experts, SSSIHL generated synthetic data which has the features of near closeness to real data. For experimental purposes, motor insurance data was created according to Indian motor market. Size of the data is around 12 million records. Each record has 20 rating factors like name, date of birth, location, car make, model, mileage, cubic capacity, premium details,mode of payment and type of damage . Coverage periods are captured according to the type of policy. Eg: 1- year policy or 3-year policy.

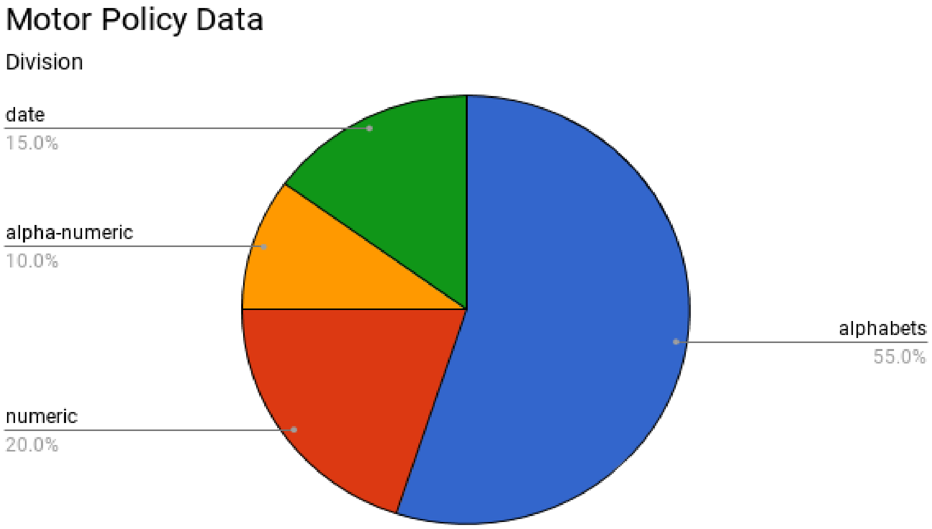


Table 5.1 specifies the different types of checks and number of checks performed on the data.

Type	count
Number of Rows	12568455
Number of Fields	20
Number of Numeric Fields per Row	4
Number of date Fields per Row	3
Number of Alphanumeric Fields per Row	2
Number of Alphabet Fields per Row	11
Number of Checks for Numeric Fields	12
Number of Checks for Date Fields	12
Number of Checks for Alphanumeric Fields	16
Number of Checks for Alphabet Fields	44
Number of Checks per Row	89
Total Number of Checks	1118592495

Table 5.1 – Types of Checks performed on Data

Then the data was cleaned using basic rules and complex rules of business. Some of the basic rules defined are, a particular model car from a specific make cannot have premium more than a certain limit. The accident date should be more than policy start date.

Some complex business rules are, a specific car from a make and model with claim being a third party damage with a premium in the given limits should not claim an amount exceeding the prescribed limit.

The framework performed accurately on data using basic and complex rules of business. The results were compared with validation tools.

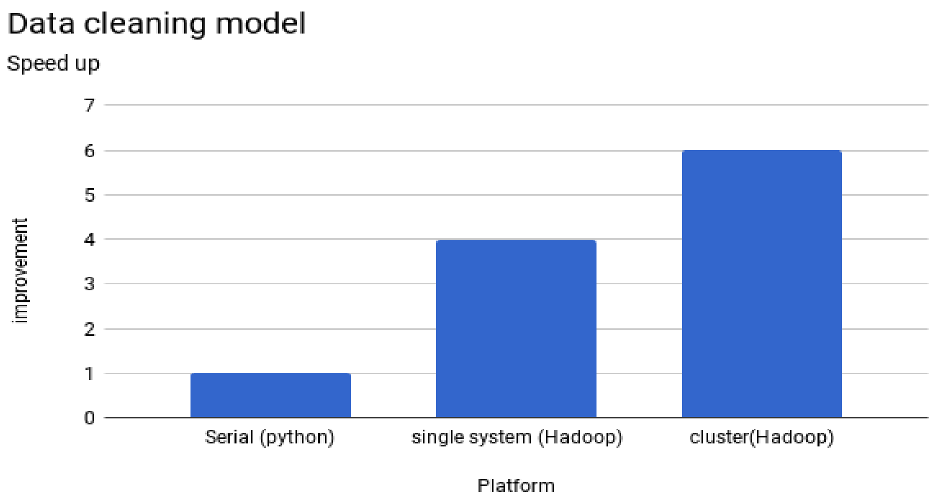


Figure 5.1 Improvement in Performance Cluster setup

Figure 5.1 shows the improvement using hadoop between single system and cluster. The single system hadoop was running almost four times faster than the serial code. Using cluster we got a maximum speed up of **6X** compared to the serial version.

6. Pilot project using SSSIHLDATA CLEANING FRAMEWORK on Real Data

Description:

There are six datasets received from IIB. First two files have the information about policy. They contain policy details and information about premium charged. The next two files contain the information about claims which are claimed because of own damage. The last two files contain the claim details which have reported as third party damage. All files had 50000 records, but each one had different number of unique fields which contain the information about policy or claims. Fields were of the format Text, Numeric, Boolean, Date and combination of Text and Numeric.

The parameters of the files are summarized in the table below

File Name	Details	Fields	Number of Records
12A_1	Policy	64	50000
12A_2	Policy	64	50000

12B_1	Claims Own Damage	58	50000
12B_2	Claims Own Damage	58	50000
12C_1	Claims Third Party Damage	101	50000
12C_2	Claims Third Party Damage	101	50000

Since the files are of very low size, the performance improvement is not being considered when discussing about the results. For the first level cleaning the following observations summarize the results which we got.

- Office Code masked -Data Element Length mismatch -0.5% records.
- Endorsement numbers Masked - Data Element Length mismatch -0.9% records.
- NCB Number Data Element Length Mismatch -0.8% records.
- Product ID Masked - Needs to be compared against master.

After doing the first level of cleaning the data, the output of this was given as input for the next level of cleaning. Using the rules of business, the data was cleaned. The below observations summarize the results obtained.

Policy Data:

- IDV of the vehicle is negative -.5% records
- NCB(Boolean and the percentage value) -10% records
  - I Boo NCB is present even when premium is 0.
  - II Num NCB is present even when Boo NCB is blank.
  - III NCB Boolean is blank.

- Premium is negative on a policy -5% records
- Policy Period is more than 365 days - 1%records
- Boolean Values fields - May need to be reviewed.

Claims Data (Own damage and third party damage):

- Date of Accident loss should be less than policy end date and greater than policy start date -0.8%records
- Claims intimation date is less than date of accident loss -0.5%records.
- Boolean values fields - may need to be reviewed.

The results obtained using the level 3 advanced data cleaning were proved to be accurate when compared with the results obtained from validation tool. There were no errors with respect to the specified complex rules.

7. Conclusions and Future Work

Results indicate that the sample data provided by IIB is almost error free. Our data cleaning framework performed as expected, not much improvement in speed up was observed because of the file sizes, which are not that large. Generated results are validated using Excel VBA, confirms the accuracy of the framework.

As the Next step, for level 3 Data Cleaning, SSSIHL framework may be configured with advanced rules to reflect complexity specific to each line of business.

References

[1] Li Lee Mong , Cleansing Datafor Mining and Datawarehousing, school of computingNational University of Singapore, 1999.

[2] Rahm E.&Hai Do Hong,Data Cleaning: Problems and current approaches,IEEE Bulletin of the Technical Committee on Data Engineering, 2000.

[3] Choudhary, Nidhi. "A study over problems and approaches of data cleansing/cleaning."*International Journal of Advanced Research in Computer Science and Software Engineering* 4.2 (2014): 774-779.

[4] Hellerstein, Joseph M. "Quantitative data cleaning for large databases."*United Nations Economic Commission for Europe (UNECE)* (2008).

[5] Y.PatilRajashree,Dr.Kulkarni R.V. ,A Review of Data Cleaning Algorithms for Data Warehouse Systems. IJCSIT ,Vol. 3 (5) , 2012.

[6] Müller, Heiko, and Johann-Christph Freytag. *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. FürInformatik, 2005.

[7] Li LeeMong, Wang Ling Tok&Lup Low Wai,IntelliClean: A knowledge-based intelligent data cleaner,Proceedings of the ACM SIGKDD, Boston, USA, 2000.

[8] Sarpong Kofi Adu-Manu, Davis Joseph George, Panford Joseph Kobina , " A Conceptual Framework for Data Cleansing –A Novel Approach to Support the Cleansing Process" International Journal of Computer Applications ,Volume 77–No.12, September 2013.

[9] Peng Taoxin ,"A FRAMEWORK FOR DATA CLEANINGS IN DATA WAREHOUSES",School of Computing,Napier University,10 Colinton Road, Edinburgh, EH10 5DT, UK.