Suicides in India Big Data Analysis Using Map Reduce Environments

India's suicide statistics from year 2001 to 2012

Tejas Phopale
MSc. *In Data Analytics*
National College of Ireland
Dublin 1, Ireland
x17162301@student.ncirl.ie

***Abstract*—The Suicide in India is the major concerned subject for the country. It is necessary to analyse the reasons behind it and implement necessary prevention measures. The big data analysis report aimed at analyzing the primary factors causing this from the year 2001 to 2012. The government of India publishes suicides statistics on the official website in every five years to analyze the whys and wherefores of every region which made public for data analytics. The findings can help in making governing decisions by rewriting the current facilities and promoting education. The study focused on the most probable reasons which causing suicides. The intended findings help to reduce by organising special suicides prevention centres and psychiatric counselling. This project is focusing on implementation and design of the MapReduce paradigm to analyse the suicides in India. Moreover, few other parameters such as states, gender, education, profession, age-category are comprised for analytics. The environment using Apache Hadoop framework, JAVA environment, MySQL database system, Apache HIVE, Apache Pig. The Hadoop file system made easy in processing the suicide data of around half million rows containing all important suicide-related information.**

Keywords—Map Reduce, Hadoop, Java, PIG, HIVE Suicide Dataset

I. Introduction

Suicide is the major problem around the world. Every year, people are committing suicide astronomically. Around 80 thousand people die every year [1] which shows self-inflicted injuries. The World Health Organisation (WHO) predicted that by 2020, there would be increase in suicides around 2% worldwide and around 17% within India itself. As a result, they declared suicide as disease. The WHO's suicide intervention study imposed that by early detection of risk can reduce the suicide mortality. India is fast developing as diverse terrain and huge South Asian country. It stands world's second highest populated county but also leads to high marginal suicide rates. However, there should be an enhancement in public safety rules considering increase in number of suicides. This report can be used as guidance to implement such rules by providing the statistical evidence. Many countries such The United States, Germany has implemented the National Suicide Prevention Lifeline service. This service helps individuals whose mind started thinking of suicide. However, India does not have any national suicide prevention programmes or organisation yet. This is the major problem that country is facing but it is still neglected. This paper aims to determine the pattern and causes for number of suicides by drilling down from India's global perspective to regional perspective. Moreover, the Apache Hadoop framework is built for statistical analysis report to investigate suicides in India in detail. This huge suicide comprises of different related parameters. The different Map reduces environments are used based on their significance to process this data and get valuable report statistics. Initially, the data has been loaded on MySQL database and then into the Hadoop file system where the map reduce job is performed. From the MySQL database, JDBC connection has been made for JAVA map reduce task and store the data into a new table of MySQL. The Sqoop interface application used for transferring data from MySQL relational database to Hive for processing. The file is loaded from ubuntu environment for Apache Pig and performed analytics. The visualisation is performed using visualisation tools like Tableau and Power BI. Furthermore, this report incorporates information about suicides along with related work done with technologies and methodologies used to build this system which helps to provide the valuable outcomes.

II. Related Work

The number of suicides in India from the year 2001 to 2012 and present the number of suicides in different parts of the country. It helps the government to get better findings to focus on the regions which are affected more and reduce them by implementing some preventive measures, rules and regulations. By using the data which is provided by the Indian government on the website, it can be categories based on different factors like age categories, dead cause and states. Also, it indicates that instead of blaming to the government every time, it is individuals responsibility to support national activities by counselling and being helping hand to the social community [2]. The analysis on suicide victims in India for the year 2011 concludes that the rewriting current facilities and promoting education should be the key focus for the government of India. The control over suicide can be achieved by enabling psychiatric counselling, policy decisions [1].

The MapReduce architecture helps to separate massive high-level sequential data of complex infrastructure with the use of the Genome Analysis Toolkit (GATK). It is the next generation toolkit which maintains the sequential, robust, stable, memory efficient data and helps to analyse the required algorithms in shared memory machine as well as in distributed clusters [3]. The MapReduce programming model possesses map function with the key-value pair and reduce function is merged with the associated key of the mapper. This functionality is

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

automatically parallelised, scalable and executes on large clustered commodity machines. The input data, scheduling programme execution and machine failures maintained by the runtime system. The MapReduce is successfully used at Google for several reasons due to its features such as easy to use, fault-tolerance, load balancing scarce resource bandwidth and local optimisation [4]. The Google's map-reduce paradigm used to represent parallel speed techniques which speed up by increasing number of processors. The map-reduce utilisation helps to speed up 1.9 times on dual processor and 54 times on 64 cores [5]. The MapReduce architecture helps to separate massive high-level sequential data of complex infrastructure with the help of the Genome Analysis Toolkit (GATK). The limitation of MapReduce is that it cannot handle some of the features such as real-time, interactive, iterative and graph processing.

For extreme high-level commodity database and rapid growth in business, Hive has become an alternative solution, especially for traditional and expensive data warehousing. The Hadoop is the popular open-source map-reduce environment but it is very low level and custom programs makes hard to maintain. So, the Hive is built-up on the top of the Hadoop which supports declarative SQL as query language and compiles in map-reduce jobs. The query language is called as HiveQL supports custom map-reduce scripts, nested compositions and collections like maps and arrays. The Hive also includes metastore, catalogues and schemas which helps for query optimization. The optimizer is rule-based and its processing is scalable in nature. The limitation of Hive is that no row-level insertion, high latency, no real-time data access and does not support ACID properties [6].

The Pig Latin language is the real-life need for data processing where terabytes of data collected every day. The Teradata seems better solution but quite expensive for parallel database computing. The map-reduce paradigm emerge as proactive solution as compare to declarative SQL but it is rigid and also it involve custom user code which is hard to maintain. The Pig Latin language is the sweet spot language which lies between declarative SQL and procedural map-reduce style. It is an opensource Apache-incubator project which consist map-reduce implementation, works on incremental fashion and executed over Hadoop with sequence of map-reduce jobs. Pig reduces processing time and it consist of novel debugging environment called Pig pen which helps to gain the productivity and ability to control the executions but limitation is that it is still under upgrade status. The user can iterate and freeze the programme multiple times by examine the output.[7]. The Pig experience leads high-level data flow, low-level hacking, SQL-style data manipulation, compilation on the sequence on map-reducer jobs [8].

III. METHODOLOGY

This section represents the information about process flow of this project along with datasets information and architecture of designed system. The architecture and components of the project is as mentioned below:

1) Educational_status_of_suicide_victimes_state.csv, Profession_profile_of_suicide_victims_state.csv, Suicides in India 2001-2012.csv are the input data files.

2) Apache Hadoop Frameworks- Java MapReduce, Pig, HIVE

3) MySQL Input Database- Edu_Victims, Prof_Victims, Suicides_Stats

4) MySQL Output Database- output table (for 4 Java MapReduce operations exported as per tasks)

5) Sqoop – Data transformation from MySQL to Hive and MySQL to HBase for Java MapReduce output.

6) Java – Four different project are created for MapReduce with JDBC connection and store into table called "output" of MySQL.

7) Hive – Importing tables from HBase and generating 2 output csv files against the findings

8) Pig – Uploading data from csv and two output files are created in csv files for queries

All the input csv files are uploaded in MySQL by creating three different tables mentioned above. The Java MapReduce task performed by doing JDBC connection with MySQL and stored the output into the table called "output". The output data is stored in HBase as secondary backup storage. Later, that table in exported into csv files for all Java MapReduce tasks. The data from MySQL is imported for Hive processing using Sqoop and post-processing outputs are stored into csv. The data in Pig is directly imported from csv files and queries outputs are stored into HBase and save the output into csv. Below diagram shows the process flow of suicides big data analysis using MapReduce environments.

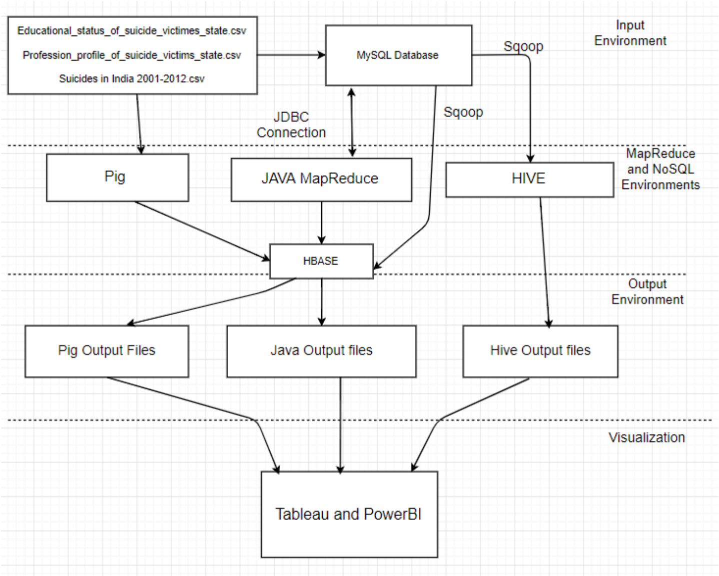
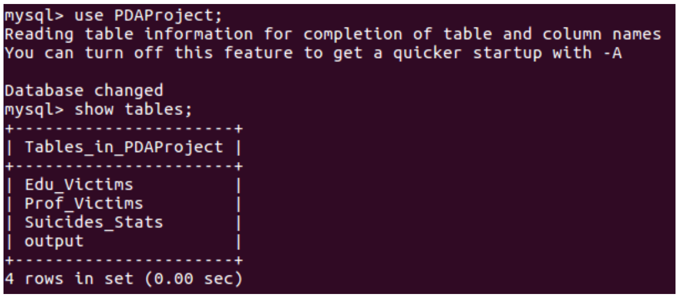


Figure 1: Process Flow

The next section describes the data source information. The suicides in India dataset has collected from Kaggle and India's official government website that is data.gov.in which gives the assurance about the data. Below figure shows the variables and the tables associated with this project after uploading it into MySQL. The database is created called “PDAProject” and tables mentioned about are created. The below figure shows MySQL table structure in PDAProject database.



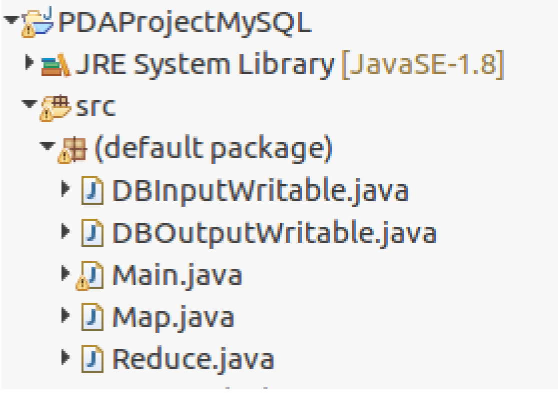
```
mysql> use PDAProject;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+----------------------+
| Tables_in_PDAProject |
+----------------------+
| Edu_Victims          |
| Prof_Victims         |
| Suicides_Stats       |
| output               |
+----------------------+
4 rows in set (0.00 sec)
```

Figure 2: MySQL database structure

The first table called “Suicides_Stats” has 237520 rows in total with 7 columns. The second table called “Prof_Victims” have 6841 rows and 16 columns. The third table called “Edu_Victims” have 4105 rows and 6 columns.

Furthermore, The Java MapReduce environment is taken for processing suicide data. The reason behind taking this approach is the methodology to drill down by focusing on highly affected states and their reasons. So, Java MapReduce framework is used by applying MapReduce patterns like Summarization, Filtering and Input-Output patterns. The input is taken by creating JDBC connection in Java and retrieved the required table with their fields. The below figure shows Java MapReduce folder structure in eclipse. The Main.java class is main java class file where MySQL JDBC connection has been performed. The DBInputWritable and DBOutputWritable java classes are used for taking input from MySQL database table’s columns and saving output into another table respectively. The Map and Reduce classes are functioned as Mapper and Reduce respectively.



```
▾ PDAProjectMySQL
  ▸ JRE System Library [JavaSE-1.8]
  ▾ src
    ▾ (default package)
      ▸ DBInputWritable.java
      ▸ DBOutputWritable.java
      ▸ Main.java
      ▸ Map.java
      ▸ Reduce.java
```

Figure 3: Java MapReduce Folder

There are total four visualisations that have been performed by digging into the data step by step. First, the analysis is performed by checking which state has a higher number of suicides rate from 2001-2012 and then checked the actual reason for that specific state. Second, the analysis performed based on what are the major reasons which lead to suicides in that state. Third, resulting which age-group has led more suicides in that state and fourth by considering that age-group, the operation performed by finding the reasons for the suicides.

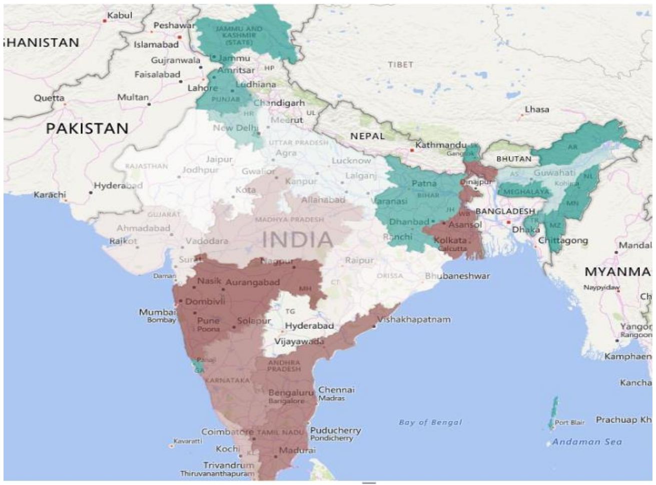
The Hive environment has been considered for the implementation of the Join patterns, for creating report scenarios and retrieving outputs with multiple parameters. First three tables are created to map against MySQL tables. By using Sqoop the data is imported into the hive data. There are total two operations are performed using Hive. First, to analyse does suicide rate increased over the years? And secondly to check that does education restricts the suicide rate in every state?

The Pig environment has been considered because of its benefits such as data flow and procedural language, easy to control the execution, less development time and lazy evaluation. There are two analyses performed using Pig. First to check which top 3 causes for suicide are affected on different age group in India and second, which profession leads to higher suicide rate.

IV. Results

I. Java MapReduce Visualisations:

1) The initial startup for the visualization of the suicides in India is performed by checking which states have highest suicide states in India. Two figures shown below illustrate that Maharashtra, West Bengal and Tamil Nadu are the states where highest number of suicides victims and are reported from 2001 to 2012. For next analyse the decision taken to drill down into Maharashtra state as Maharashtra state has highest suicide victims and finding the reasons and other factors of suicides.



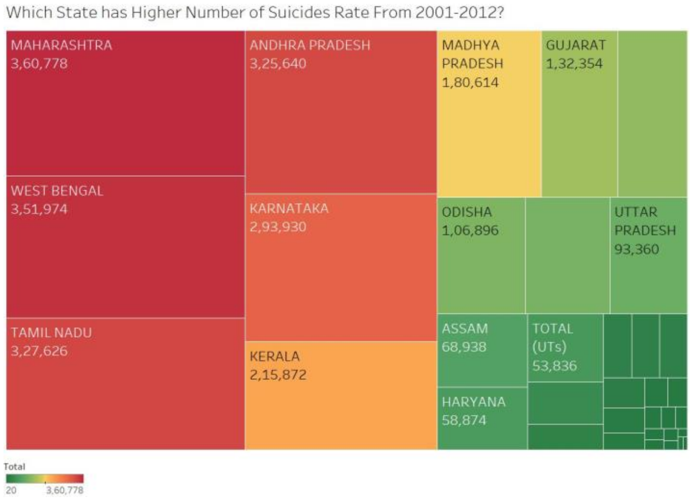


Figure 4: Map and TreeMap to Display Highest Suicide states

2) Next visualisation is performed to check what are the major reasons and factors which leads to suicide in Maharashtra. As shown in below figure, it has been observed that primary factor for the suicide in Maharashtra is because people got married forcefully. There are other major factors like family problems and many ended their lives due to daily pressure and so on. So the next visualisation needs to determine that what are the age groups who commit more suicide and check if the teenager involved in this due to depression.

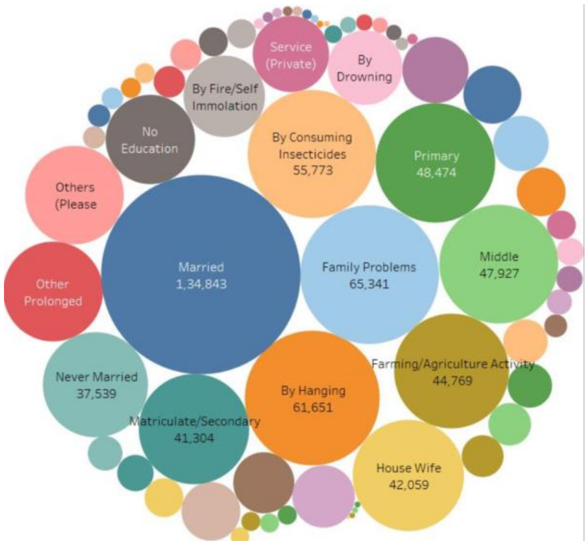


Figure 5: Reasons and Factors involve suicide in Maharashtra

3) The third visualisation helps to visualise which age group are committing more suicides in Maharashtra. The below visualisation graph shows that teenagers and young people are committing more suicides whose age ranges from 15 years to 29 years. From 2001 to 2012, around 204 thousand people have reported as suicide victim only in Maharashtra. So to understand their reason for committing suicide, the next visualisation should represent the reasons for the suicide of that age and cross verify that does it matches with regional issue which evaluated in the second visualisation or it differs.

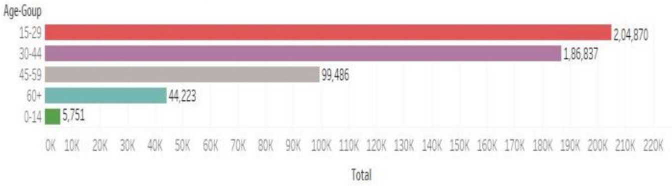


Figure 6: Victim's Age Group Statistics for Maharashtra State

4) Forth visualisation is performed for cross-validation of second visualisation and checks whether the reasons for suicide are similar or it affects by the generation gap. The below figure shows that family problems seems to be a common issue in both the visualisations. Around 28 thousand suicides occurred due to a family issue. Also, housewives and by hanging people have committed more number of suicides in common which observed during cross-validation.

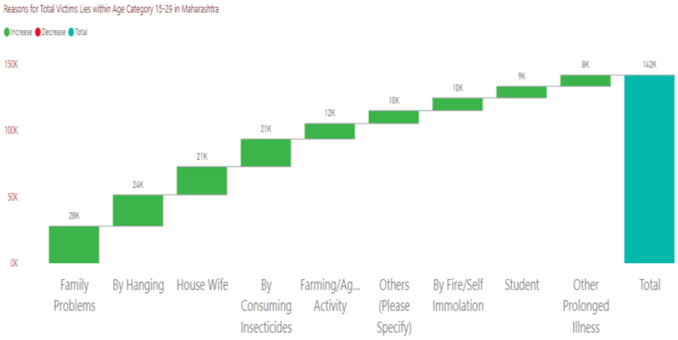


Figure 7: Suicides Reasons and Factors for 15-29 Age category in Maharashtra

II. Hive Visualisations:

1) The Hive visualization is performed to cross check against real life factors such checking the statistics to observe whether government has planned something for the serious problems and verify whether suicide rates are increasing or decreasing over the year. The below figure shows that there is no serious precaution that has been taken by Indian government nor any strict law to stop this. The observation shows increase in number of suicides every year.

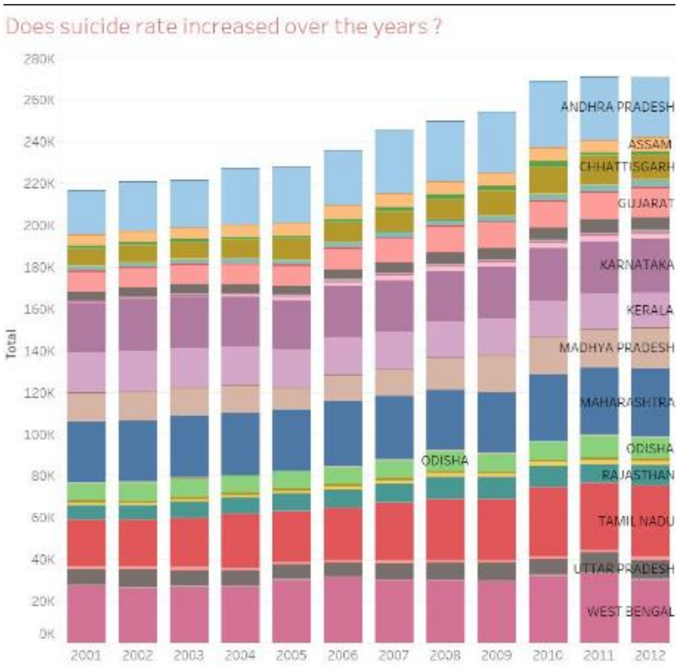


Figure 8: Increase in Suicide Victims Count Every Year

2) The second and most important visualisation of real-life scenario is that does education restricts the suicide rate in every state. The figure mentioned below shows that the people who are less literate committing more suicides. So Indian government should focus on the education and promote robust jurisdictions.

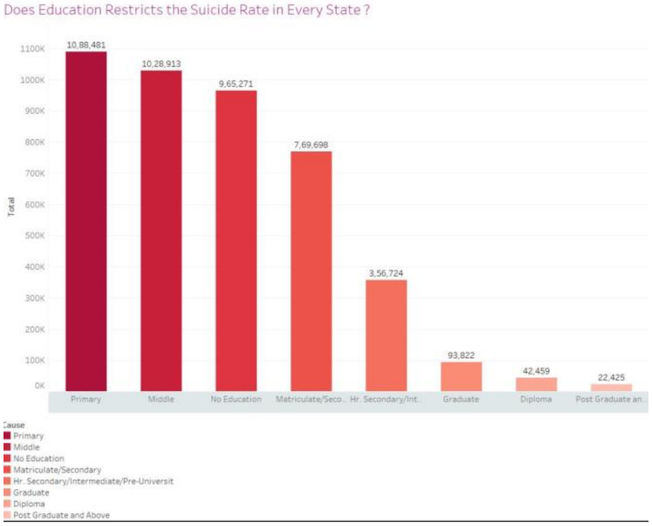


Figure 9: Suicide Statistics Categories by Education

III. Pig Visualisations:

1) Pig environment is used to write complex queries. At the end of the report, it should highlight the three important causes for every age category people which are enforcing to suicide and on which government should emphasise their focus to resolve it. The below figure shows that common problem for suicide are family problem and hanging. There are other reasons for age group 15-29 coloured in blue and 30-44 age category coloured in grey. This report could help to summarise the fact to the government and they could take quick decision on it.

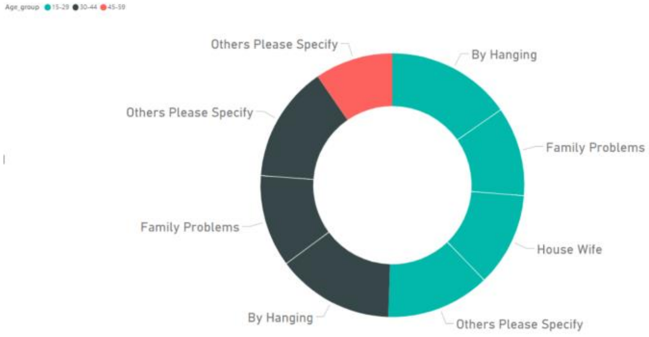


Figure 10: Three Important Causes for Suicide

2) The second and important visualisation illustrate that which profession leads to the higher suicide rate. The below figure observation shows that Self-employed, housewife and farmers are committing more suicide. The self-employed due to business loss, housewife due to family problems and farmers due to loss of food grains production or other reasons.

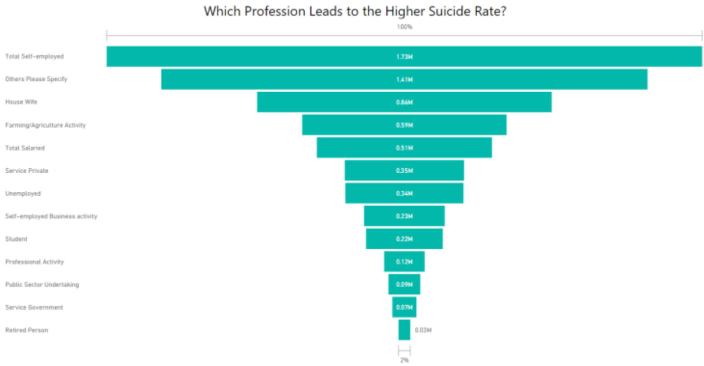


Figure 11: Profession which Leads High Suicides

Conclusions and Future Work

In this project, the use of large dataset which comprises the information related to the suicides in India. To process this large dataset, Apache Hadoop, MySQL, MapReduce environments are used. The project represents good quality of information and it can be utilise for making public safety at it's best. The data processing is performed by using MapReduce environments like Java, Pig and Hive. The output of all these environments provides information which can assist Indian government to implement new public safety rules and organise suicide prevention foundation. The statistics also helps to take proactive measure at country and region level. The education creates strong perception for making right decision at bad times.

In future, by using more data and variables can generate more depth analysis for big dataset. Also, enabling latest dataset can help to analyse the current scenarios, get more accurate and efficient results and implement the public safety law accordingly.

REFERENCES

[1] Selva Priyanka, S., Sudeep Galgali, S. Selva Priya, B. R. Shashank, and K. G. Srinivasa. "Analysis of Suicide Victim Data for the Prediction of Number of Suicides in India." (2012).

[2] Chouhan, Yuvraj Singh, Suraj Kaushik, Nagendra Sharma, Shreyansh Singh, and Shamela Rizwana. "Analysis of Suicide Database to Reduce Number of Suicides in India." Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org 8, no. 4 (2018).

[3] McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research (2010).

[4] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51, no. 1 (2008): 107-113.

[5] Chu, Cheng-Tao, Sang K. Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Kunle Olukotun, and Andrew Y. Ng. "Map-reduce for machine learning on multicore." In Advances in neural information processing systems, pp. 281-288. 2007.

[6] Thusoo, Ashish, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. "Hive: a warehousing solution over a map-reduce framework." Proceedings of the VLDB Endowment 2, no. 2 (2009): 1626-1629.

[7] Olston, Christopher, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. "Pig latin: a not-so-foreign language for data processing." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1099-1110. ACM, 2008.

[8] Gates, Alan F., Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, and Utkarsh Srivastava. "Building a high-level dataflow system on top of Map-Reduce: the Pig experience." Proceedings of the VLDB Endowment 2, no.2(2009):1414-1425.