International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

# Data Analyser GUI

**Hardik Goel**[1], **Tushar Ahuja**[2]

IT Department, Maharaja Agrasen Institute of Technology (IPU), Delhi, India

**Abstract:** ***Global digital content created will increase some 30 times over the next ten years – to 35 zettabytes. Big data is a popular, but poorly defined marketing buzzword. One way of looking at big data is that it represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. Examples of this data include high-volume sensor data and social networking information from web sites such as FaceBook and Twitter. Organizations are interested in capturing and analyzing this data because it can add significant value to the decision making process. Such processing, however, may involve complex workloads that push the boundaries of what is possible using traditional data warehousing and data management techniques and technologies. This article looks the benefits analyzing big data brings to the business. It examines different types of big data and offers suggestions on how to optimize systems to handle different workloads and integrate them into a single infrastructure. Two important data management trends for processing big data are relational DBMS products optimized for analytical workloads (often called analytic RDBMSs, or ADBMSs) and non-relational systems (sometimes called NoSQL systems) for processing multi-structured data. A non-relational system can be used to produce analytics from big data, or to preprocess big data before it is consolidated into a data warehouse. Big Data is a concept that is leading the world right now and taking it by storm. We have tried to discuss on the fundamentals of Big Data and tools and techniques associated with it. We also have tried to categorize the Big Data elements into a model and tried to derive Big Data Ecosystem from it. The V Model for the Big Data has been defined and categorized into 3V, 4V or 5V dependent on the organization which uses it and under which business scenario. Catering to the aforementioned models, we have classified data into various forms and explanations have been provided on the same to gain a better insight and understanding on the same.***

**Keywords:** Big Data, RDBMs, Data Warehouse

## 1. Introduction

With changing trends in the field of information technology it has become difficult for traditional database software to manage huge datasets and make quick query processing on them and generating results out of them in a small time, with addition of a big data tools in it field it has become easy to handle and manage live and real time datasets and work upon them and processing them within a very small time, moreover big data tools and frameworks have made it very easy for the user to store very huge amount of data online and on cloud and uploading and downloading it whenever it necessary .

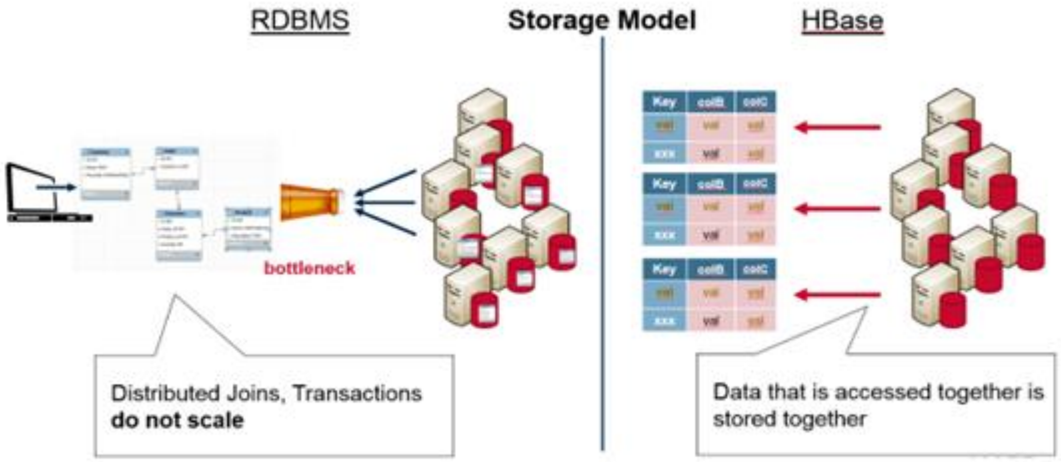


**Figure 1:** Comparison-Relational vs Non Relational DBMS's (Architecture)

Our project data analyser and viewer designed and managed using hadoop and netbeans provide the best view outputs graphically as well as tabular to each individual user which has been cropped from large datasets. these datasets are in gigabyte storages and for traditional rdbms it becomes difficult to perform quick analysis out of them. Datasets selected include records over years and specific data has been pulled out using proper querying process and fixed partitioning to handle easy analysis. a gui is the best way for a person who is unaware about of how to access an os and provides the user point to point details regarding data and graphical content he wishes to see without including any extra data cuts.

Data analyser is an java backend application that has been designed to carry several datasets analysis using data sets queried from hadoop cluster and collected from various government sources. Datasets have been secreted for whole Indian data and according to all states and from year ranging around 1950-2018.Not only relational view but it also provides graphical views with several ranges. Easy managing of large dataset and usage of map reduce algorithm in hive framework makes easy managing and handling of large datasets to carry out easy analysis.

Various pie charts or bar graphs have also been generated based on data which has been classified according to the range eg: separate pie chart according to quantity of production and year of production. View in output frame - login frame to mark entry if new user signing two sections one for graphical view and one for tabular view scroll down menu containing division based on district name. Datasets handled includes:

Analysis and Graphs automatically update with update in dataset

### 1) Climatic Evolution And Agriculture domain

#### a) Indian crop production state wise (1997-2016)

Type: csv file structured
Analysis task: On basis of

- States
- Districts within states
- Season wise - Kharif, Rabi, Whole year, Autumn
- Crop wise - Arecanut, rice, banana, cashew-nut, coconut, dry ginger, sugarcane, sweet potato, tapioca, dry chillies, dry ginger, black pepper, other kharif crops
- Crop Area wise distribution
- Production wise

#### b) Daily Retail Price of Potato (1997-2016)

Type: csv file format
Analysis task: On basis of

- Year
- Center name
- Price

#### c) India Air Quality (1990-2016)

Type: csv file format structured
Analysis task: On basis of

- State wise
- Location within states
- Sampling year
- Type of area (residential, rural, industrial, urban, other areas)
- So2 and No2 levels in air
- Spm

#### d) Rainfall in India (1901-2016)

Type: xml file semi structured
Analysis task: On basis of

- State wise
- Year wise
- Month wise

#### e) Air pollutant PM2.5 and PM10 India (2018)

Type: xml semistructured
Analysis Task: On basis of

- City
- Parameter (PM2.5 and PM10)

#### f) India Affected Water Quality Areas (2009)

Type: csv file format
Analysis task: On basis of

- State names
- District names
- Block name
- Village name
- Panchayat name
- Quality Parameter (Salinity, Fluoride)
- Year

#### g) Delhi weather dataset (1997-2017)

Type: csv file
Analysis Task: On basis of

- Year
- Condition (Smoke, Clear, Haze, Fog, Shallow Fog, Unknown)
- Pressure
- Temperature
- Wind direction (West, North, ssw.nne)

### 2) Demography domain

#### a) India States Analysis (2018):

Type: csv file
Analysis Task: On basis of

- State
- State code
- District code
- Population male
- Population female
- Literates total (male, female)
- Sex ratio (chile sex ratio)
- Total graduates (male, female)

#### b) Literacy Rate In India state wise (1951-2011)

Type: csv file structured
Analysis task: On basis of

- State wise literacy

#### c) Indian Elementary School (2005-2017)

Type: 10 csv file format datasets
Analysis Task: On basis of

- Year
- State
- Primary Divisions (middle, upper etc)

#### d) Suicides in India (2001-2011)

Type: csv file format
Analysis task: On basis of

- State wise
- Year wise
- Type (illness, bankruptcy, dowry, dispute, property dispute etc)
- Gender (male, female)
- Age group (0-60yrs+)
- Total suicide count)

#### e) Crime Rate in India (2001-2018)

Type: 15 csv file format datasets on basis of individual crime

Analysis task: On basis of
Type of crime (Property stolen, victims of rape, complaint against polices, period of trial by court, auto theft, serious fraud, murder victim age sex, human rights violation , police housing, specific purpose of action)

## 2. Literature Review

What Is Big Data: When we speak about Big Data, as we have done above, we often identify it as a jargon, catch phrase which means a exponential volume of unstructured and structured data that contains so many huge datasets which cannot be processed by traditional database management techniques and associated software techniques. With the size of the big data and simply the capacity of the data that it encompasses, it carries in itself the potential that will help companies, in making far better, intelligent and data driven decisions and help in improving operations. For most of the organizational scenarios, it can be easily identified either the data is in excess of the current storage and processing capacity, or the volume of the data is too big or it moves too fast. To give insights using the same data, that we have spoken about earlier, it has to help us in giving insight which would help us in gain competitive advantage, increasing revenues and customer retention and for that we need to capture the data, clean the data, format, manipulate, store and analyze the same. Big Data is a concept and a concept can have various interpretations, for which the same topic can have multiple definitions: Big Data is the amount of data beyond the ability of technology to store, manage and process efficiently (Manyika et.al, 2011). Big Data is a term which defines the hi-tech, high speed, high-volume, complex and multivariate data to capture, store, distribute, manage and analyze the information (TechAmerica Foundation, 2014). Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization (Gartner, 2014; Gürsakal, 2014).

Data Forms: Structured: When we talk about structured data, we often conclusively identify that, as soon as we placed our current data ware house in the relational database management system, the structure of the relational database management system was enforced on the current data ware house system, which is inclusive to understand the meaning associated with it. So we know, which columns are placed where, whom are they associated with and how the columns are associated in between tables and table spaces. The format of the data can be in text or numerical, but it is common understanding that for every person there is a unique identifier in terms of Age.

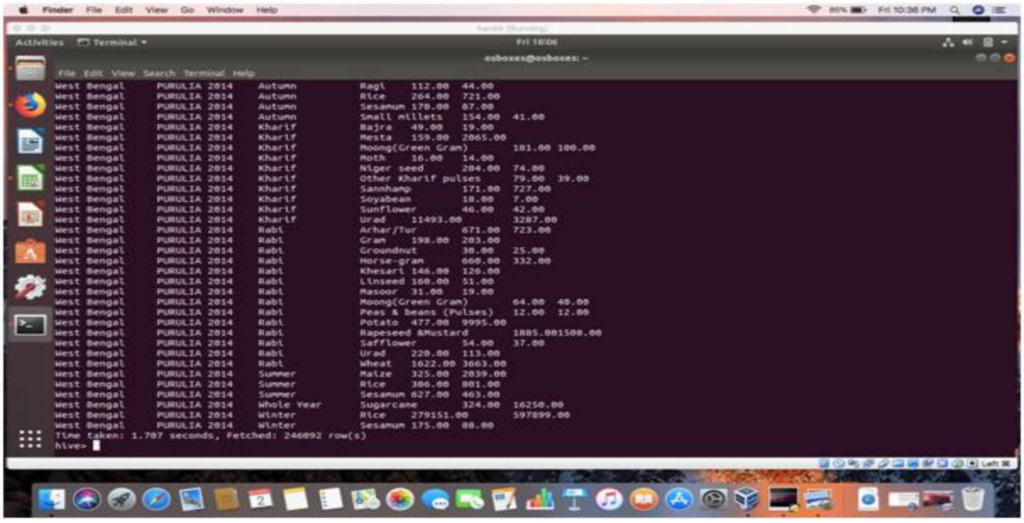
Semi Structured: As we move on from structured data to semi structured data, there is little to demarcate and often the differentiating lines goes blurry. The data format that we are describing here does not conform to an explicit and fixed schema, however the tags associated with the data, if found associated with organizational structure, then the same data would be easier to analyze and organize. The same concept described here would predate the idea of XML but not HTML

Unstructured: We have already discussed about the Structured and Semi Structured formats. Moving on to the unstructured format, this type would consists of formats that cannot be easily indexed.

## 3. Analyse Data

After organizing data, it has to be analyze to get fast and efficient results when a query is made. Mapreducer’s are mainly used to analyze data. Mapreducer in Pig, Hive are very efficient for this purpose.
Hive – A Warehousing Solution Over a MapReduce Framework

**What is Hive?**
Hive is a data warehousing infrastructure built on top of apache Hadoop. Hadoop provides massive scale-out and fault-tolerance capabilities for data storage and processing (using the MapReduce programming paradigm) on commodity hardware. Hive enables easy data summarisation, adhoc querying and analysis of large volumes of data .It is best used for batch jobs over large sets of immutable data (like web logs).It provides a simple query language called Hive QL, which is based on SQL and which enables users familiar with SQL to easily perform ad-hoc querying, summarisation and data analysis. At the same time, Hive QL also allows traditional MapReduce programmers to be able to plug in their custom mappers and reducers to do more sophisticated analysis that may not be supported by the built-in capabilities of the language.



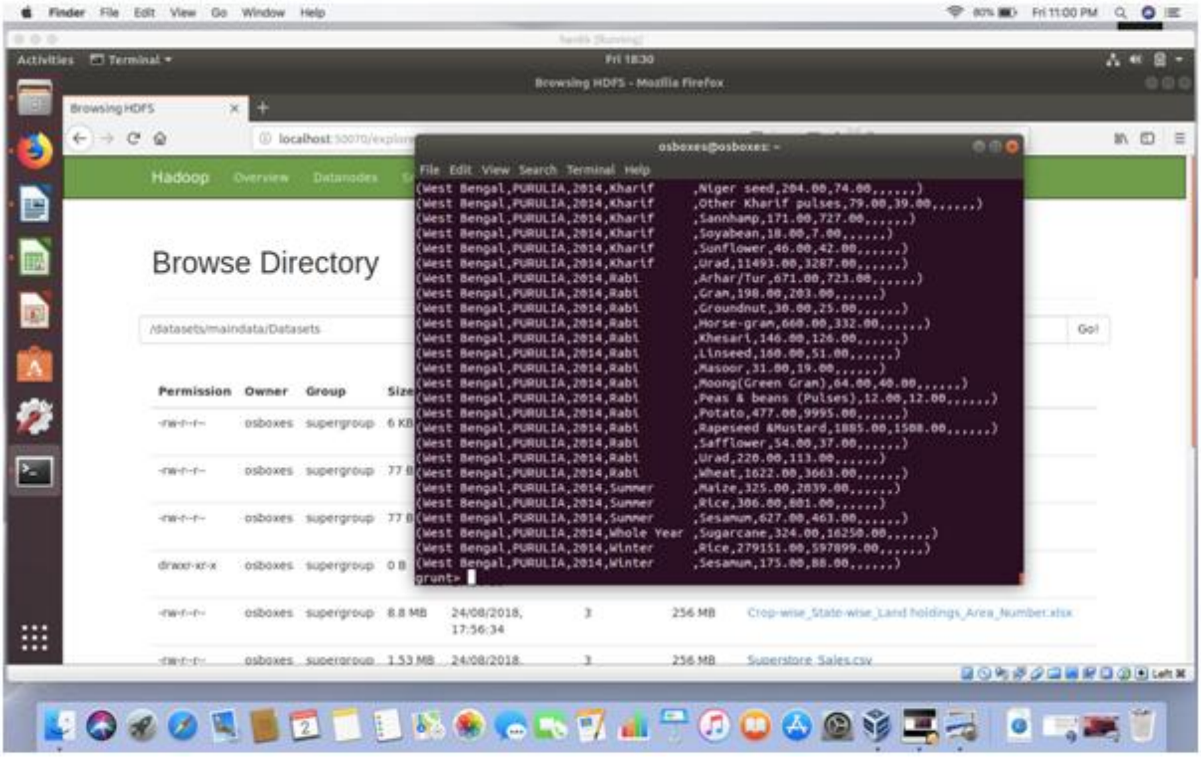
```
West Bengal     PURULIA 2014    Autumn          Ragi    112.00  44.00
West Bengal     PURULIA 2014    Autumn          Rice    264.00  721.00
West Bengal     PURULIA 2014    Autumn          Sesamum 170.00  87.00
West Bengal     PURULIA 2014    Autumn          Small millets   154.00  41.00
West Bengal     PURULIA 2014    Kharif          Bajra   49.00   19.00
West Bengal     PURULIA 2014    Kharif          Mesta   159.00  2065.00
West Bengal     PURULIA 2014    Kharif          Moong(Green Gram)       181.00 100.00
West Bengal     PURULIA 2014    Kharif          Moth    16.00   14.00
West Bengal     PURULIA 2014    Kharif          Niger seed      204.00  74.00
West Bengal     PURULIA 2014    Kharif          Other Kharif pulses     79.00  39.00
West Bengal     PURULIA 2014    Kharif          Sannhamp        171.00  727.00
West Bengal     PURULIA 2014    Kharif          Soyabean        18.00   7.00
West Bengal     PURULIA 2014    Kharif          Sunflower       46.00   42.00
West Bengal     PURULIA 2014    Kharif          Urad    11493.00        3287.00
West Bengal     PURULIA 2014    Rabi            Arhar/Tur       671.00  723.00
West Bengal     PURULIA 2014    Rabi            Gram    198.00  203.00
West Bengal     PURULIA 2014    Rabi            Groundnut       30.00   25.00
West Bengal     PURULIA 2014    Rabi            Horse-gram      660.00  332.00
West Bengal     PURULIA 2014    Rabi            Khesari 146.00  126.00
West Bengal     PURULIA 2014    Rabi            Linseed 160.00  51.00
West Bengal     PURULIA 2014    Rabi            Masoor  31.00   19.00
West Bengal     PURULIA 2014    Rabi            Moong(Green Gram)       64.00  40.00
West Bengal     PURULIA 2014    Rabi            Peas & beans (Pulses)   12.00  12.00
West Bengal     PURULIA 2014    Rabi            Potato  477.00  9995.00
West Bengal     PURULIA 2014    Rabi            Rapeseed &Mustard       1885.001508.00
West Bengal     PURULIA 2014    Rabi            Safflower       54.00   37.00
West Bengal     PURULIA 2014    Rabi            Urad    220.00  113.00
West Bengal     PURULIA 2014    Rabi            Wheat   1622.00 3663.00
West Bengal     PURULIA 2014    Summer          Maize   325.00  2039.00
West Bengal     PURULIA 2014    Summer          Rice    306.00  801.00
West Bengal     PURULIA 2014    Summer          Sesamum 627.00  463.00
West Bengal     PURULIA 2014    Whole Year      Sugarcane       324.00  16250.00
West Bengal     PURULIA 2014    Winter          Rice    279151.00       597899.00
West Bengal     PURULIA 2014    Winter          Sesamum 175.00  88.00
Time taken: 1.707 seconds, Fetched: 246092 row(s)
hive>
```

**Figure 2:** Analysis using Hive shell

## 4. What is Apache Pig?

Apache Pig is an open-source technology that offers a high-level mechanism for the parallel programming of map reduce jobs to be executed on Hadoop cluster. Pig enables developers to create query execution routines for analysing large, distributed data sets without having to do low-level work in MapReduce, much like the way the Apache Hive data warehouse software provides a SQL-like interface for Hadoop that doesn't require direct MapReduce programming,

The key parts of Pig are a compiler and a scripting language known as Pig Latin. Pig Latin is a data-flow language geared toward parallel processing. Managers of the Apache Software Foundation's Pig project position the language as being part way between declarative SQL and the procedural Java approach used in MapReduce applications. Apache Pig grew out of work at Yahoo Research and was first formally described in a paper published in 2008. Pig is intended to handle all kinds of data, including structured and unstructured information and relational and nested data. That omnivorous view of data likely had a hand in the decision to name the environment for the common barnyard animal. It also extends to Pig's take on application frameworks; while the technology is primarily associated with Hadoop, it is said to be capable of being used with other frameworks as well. The underlying Hadoopframework grew out of large-scale Web applications whose architects chose non-SQL methods to economically collect and analyse massive amounts of data. Apache Pig is just part of a long list of Hadoop ecosystem technologies that also includes Hive, H Base, ZooKeeper.


Browsing HDFS - Mozilla Firefox
Hadoop
Overview
Datanodes
Browse Directory
/datasets/maindata/Datasets
Go!
Permission
Owner
Group
Size
osboxes@osboxes: ~
File Edit View Search Terminal Help

```
(West Bengal,PURULIA,2014,Kharif      ,Niger seed,204.00,74.00,,,,,,)
(West Bengal,PURULIA,2014,Kharif      ,Other Kharif pulses,79.00,39.00,,,,,,)
(West Bengal,PURULIA,2014,Kharif      ,Sannhamp,171.00,727.00,,,,,,)
(West Bengal,PURULIA,2014,Kharif      ,Soyabean,18.00,7.00,,,,,,)
(West Bengal,PURULIA,2014,Kharif      ,Sunflower,46.00,42.00,,,,,,)
(West Bengal,PURULIA,2014,Kharif      ,Urad,11493.00,3287.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Arhar/Tur,671.00,723.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Gram,198.00,203.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Groundnut,30.00,25.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Horse-gram,660.00,332.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Khesari,146.00,126.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Linseed,160.00,51.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Masoor,31.00,19.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Moong(Green Gram),64.00,40.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Peas & beans (Pulses),12.00,12.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Potato,477.00,9995.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Rapeseed &Mustard,1885.00,1508.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Safflower,54.00,37.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Urad,220.00,113.00,,,,,,)
(West Bengal,PURULIA,2014,Rabi        ,Wheat,1622.00,3663.00,,,,,,)
(West Bengal,PURULIA,2014,Summer      ,Maize,325.00,2039.00,,,,,,)
(West Bengal,PURULIA,2014,Summer      ,Rice,306.00,801.00,,,,,,)
(West Bengal,PURULIA,2014,Summer      ,Sesamum,627.00,463.00,,,,,,)
(West Bengal,PURULIA,2014,Whole Year  ,Sugarcane,324.00,16250.00,,,,,,)
(West Bengal,PURULIA,2014,Winter      ,Rice,279151.00,597899.00,,,,,,)
(West Bengal,PURULIA,2014,Winter      ,Sesamum,175.00,88.00,,,,,,)
grunt>
```

**Figure 3:** Analysis using pig (grunt shell)

## 5. Figures

**Interfaces**

a) Relational interface
b) Graphical interface

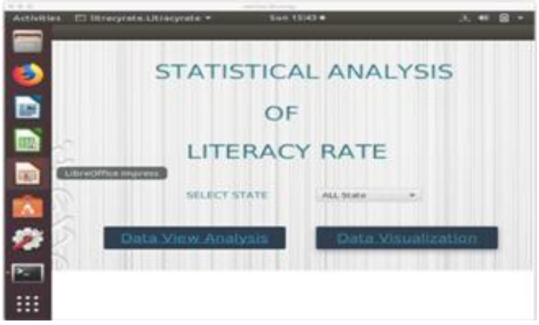
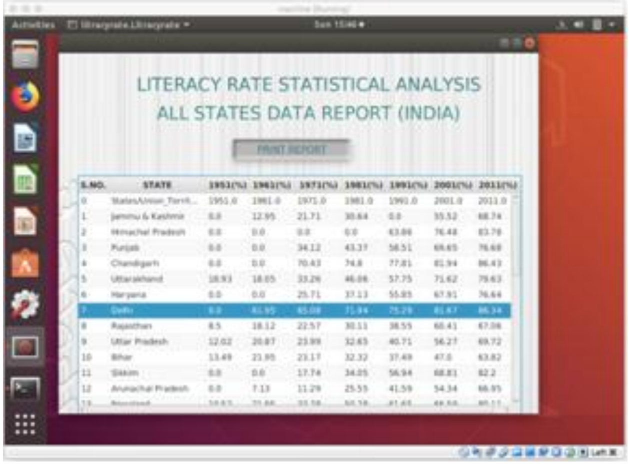


**Figure 4:** Statistical View (Literacy Rate Data Analysis)

Layout presents user with options of the kind of views of analysed data- data view analysis & data vizualisation

Year wise (1997-2016)

- Crop wise distribution.
- State + District wise
- Season wise



LITERACY RATE STATISTICAL ANALYSIS
ALL STATES DATA REPORT (INDIA)

PRINT REPORT

| S.NO. | STATE | 1951(%) | 1961(%) | 1971(%) | 1981(%) | 1991(%) | 2001(%) | 2011(%) |
|---|---|---|---|---|---|---|---|---|
| 0 | States/Union_Territ... | 1951.0 | 1961.0 | 1971.0 | 1981.0 | 1991.0 | 2001.0 | 2011.0 |
| 1 | Jammu & Kashmir | 0.0 | 12.95 | 21.71 | 30.64 | 0.0 | 55.52 | 68.74 |
| 2 | Himachal Pradesh | 0.0 | 0.0 | 0.0 | 0.0 | 63.86 | 76.48 | 83.78 |
| 3 | Punjab | 0.0 | 0.0 | 34.12 | 43.37 | 58.51 | 69.65 | 76.68 |
| 4 | Chandigarh | 0.0 | 0.0 | 70.43 | 74.8 | 77.81 | 81.94 | 86.43 |
| 5 | Uttarakhand | 18.93 | 18.05 | 33.26 | 46.06 | 57.75 | 71.62 | 79.63 |
| 6 | Haryana | 0.0 | 0.0 | 25.71 | 37.13 | 55.85 | 67.91 | 76.64 |
| 7 | Delhi | 0.0 | 61.95 | 65.08 | 71.94 | 75.29 | 81.67 | 86.34 |
| 8 | Rajasthan | 8.5 | 18.12 | 22.57 | 30.11 | 38.55 | 60.41 | 67.06 |
| 9 | Uttar Pradesh | 12.02 | 20.87 | 23.99 | 32.65 | 40.71 | 56.27 | 69.72 |
| 10 | Bihar | 13.49 | 21.95 | 23.17 | 32.32 | 37.49 | 47.0 | 63.82 |
| 11 | Sikkim | 0.0 | 0.0 | 17.74 | 34.05 | 56.94 | 68.81 | 82.2 |
| 12 | Arunachal Pradesh | 0.0 | 7.13 | 11.29 | 25.55 | 41.59 | 54.34 | 66.95 |
| 13 | Nagaland | [illegible] | [illegible] | 33.78 | 50.28 | 61.65 | 66.59 | 80.11 |

**Figure 5:** Relational Layout (Literacy Rate Data Analysis). Final data analysed represented in relational format in organised database format/

- Basic Histogram
- Line Groups
- Scatter Plots with Legends
- Box Plots
- Box Plots with errors
- Pie charts 2D/3D

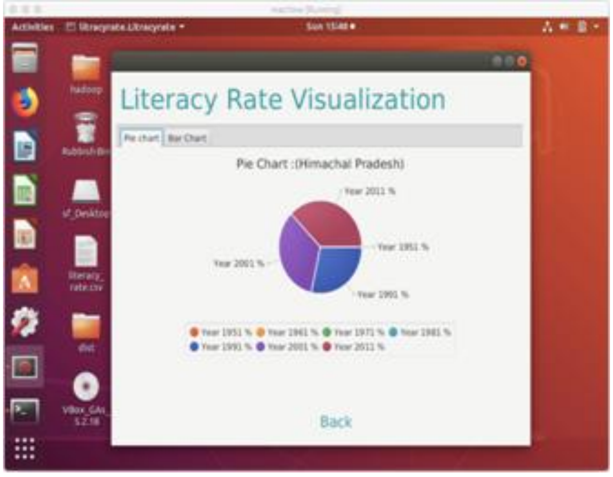


**Figure 6:** Graphical layout (Literacy Rate Data Analysis). Vizualisation representation representing pie carts-2D for each state separately based on its literacy rate .

## References

[1] Ahlawat, T., & Rambola, R. K. (2016). Literature Review On Big Data. *International Journal of Advanced in Engineering Technology Management & Applied Science*, *3*.

[2] Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business intelligence and analytics: from big data to big impact."*MIS quarterly* (2012): 1165-1188.

[3] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[4] Kline, Rex B. "Beyond significance testing: Reforming data analysis methods in behavioral research." (2004).

[5] Phillips-Wren, Gloria E., et al. "Business Analytics in the Context of Big Data: A Roadmap for Research."*CAIS* 37 (2015): 23.

[6] Mays, N., & Pope, C. (1995). Qualitative research: rigour and qualitative research. *Bmj*, *311* (6997), 109-112.

[7] Wang, Yonggang, and Sheng Wang. "Research and implementation on spatial data storage and operation based on hadoop platform."*Geoscience and Remote Sensing (IITA-GRS), 2010 Second IITA International Conference on*. Vol. 2. IEEE, 2010.[8] Walsham, G., 1995. Interpretive case studies in IS research: nature and method. *European Journal of information systems*, *4* (2), pp.74-81.

[8] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A review paper on Big Data and Hadoop."*International Journal of Scientific and Research Publications* 4.10 (2014): 1-7.

[9] Rathee, Sanjay. "Big data and Hadoop with components like Flume, Pig, Hive and Jaql."*International conference on cloud, big data and trust*. Vol. 15. 2013.

[10] Harford, Tim. "Big data: A big mistake?."*Significance* 11.5 (2014): 14-19.

[11] Singh, Ranjit, and Kawaljeet Singh. "A descriptive classification of causes of data quality problems in data warehouseing."*International Journal of Computer Science Issues (IJCSI)* 7.3 (2010): 41.