Census Analysis Using Big Data Hadoop Framework & Map Reduce Technique

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Gopal Krishan Airon (141216)

Under the supervision of

Dr. Satya Prakash Ghrera

to

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology
Waknaghat, Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **" census Analysis using Big Data Hadoop Framework & Map Reduce Technique "** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2016 to December 2016 under the supervision of **Dr. Satya Prakash Ghrera** (Head of Department, Computer science & Information Technology).
The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Gopal Krishan Airon (141216)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Prof. Dr. S.P. Ghrera
Head of Department & Professor
Department of Computer Science & Engineering (CSE) and IT
Jaypee University of Information Technology Waknaghat, HP, India, 173234

Dated:

Acknowledgement

I take this opportunity to express my sincere thanks and deep gratitude to all those people who extended their wholehearted co-operation and are helping me to complete this project successfully.

First & foremost, I would like to express my sincere acknowledgements to **Dr. Satya Prakash Ghrera, who** is being so helpful to us to complete this project successfully. Special thanks to him for all the help and guidance extended to us by him in every stage during my training. His inspiring suggestions and timely guidance enabled us to perceive the various aspects of the project in a new light and bringing good out of me.

Date: **Gopal Krishan Airon (141216)**

Table of Content

List of Figures

Abstract

The project engenders statistical analysis of humungous amount of census data and strives to find every detail of that data, the project does this by using Map Reduce technique. It is intended to use the Hadoop Distributed File Systems (HDFS) idea to Distribute the plethora of Data on to distinctive nodes and then analyzing for better information. The project various topics like aging, population growth, poverty scales. The proposed design will include Apache Hadoop which takes on HDFS & uses Map Reduce techniques. The Project includes the use Hive, Flume, Apache Zeppelin.

Chapter 1: Introduction

1.1 Introduction

We are living in an era where info. is proliferated from Machines, Individuals and Institutions at an escalating rate. Basically all these info is baffled in kind and it is extremely enigmatic to make conclusion using these assorted data. They take specific structures, for instead, semi dealt and unstructured. This information accounts for “Enormous Data" considering of its large Volume, Veracity, variety and Velocity. It is very challenging for the foundations to manage such Sizably Voluminous information. Since “Census Data” falls under this category of voluminous unstructured or semi-structured information, this gigantic convoluted Data is arranged and is examine to propose distinctive insight. The accumulated info. can be open to various focuses. A skilled and flexible strategy for getting to the information is usage of algorithms known as Map reduce

The proposed work is based upon using different algorithms. Map reduce works on different criteria. Hadoop File System was build using a file system model which is distributed in nature. It works on fundamental hardware system. Dissimilar to different systems that are distributed, HDFS is fault tolerant and build with less-expensive hardware.

HDFS has an immense quantity of info. and gives simple access. To stock such huge information, reports are gathered crosswise over numerous machines. These records are put away in unnecessary mold to discharge the framework from likely information misfortunes if there is an incidence of disappointment. HDFS likewise permits applications accessible to parallel processing. HDFS works on different nodes. Name node is the item equipment that contains the GNU/Linux working framework and the name node programming. It is programming that keeps running on basic equipment. The framework with the Namenode goes about as the principle server and plays out the escorting assignments:

Runs the file system namespace.

- Controls client's entrance to files.
- It performs file system processes such as retitling, concluding, and starting files and directories.

The data node is a basic equipment that has the GNU/Linux working framework and data node software. For every node (equipment/item framework) in a cluster, there will be an data node. These node deals with the capacity of framework information.

• Data node perform read and compose tasks on document frameworks, under customer ask.

• They also perform undertaking, for example, delivering, erasing and copying hinders as indicated by the Namenode requests.

HDFS goals

• Fault location and recuperation: HDFS incorporates a humungous measure of essential equipment, component is common. Consequently, HDFS must have systems for recognizing and reestablishing quick and programmed flaws.

• Huge informational collections: HDFS must have many nodes per cluster to handle with applications with colossal data collections.

• Hardware in the information: the requested action can be executed proficiently, when the computation is performed near the information. Particularly with regards to expansive informational indexes, arrange activity is decreased and execution increments.

1.2 Big data

By the coming of newest technologies, devices and media such as social networking sites, the amount of data yielded by the humanity is escalating rapidly every single year. The volume of data yielded by us from the beginning of 2003 to 2003 was 5 billion gigabytes. If you accumulate data in the shape of disks, you could fill out a complete soccer field. The

same figure was created in every two days in 2011 and in every ten minutes in 2013. This rate continues to grow enormously. Although all this info. gathered is important and can provide insights when managed, it is neglected. Around 90% of the worldwide information has been produced lately. Big data truly implies enormous data, it is an accumulation of extensive data that can not be handled utilizing traditional computation procedures. Big Data isn't simply data, yet it has turned into an aggregate subject, which incorporates a several devices, procedures and frameworks. Big data refers to data produced by distinctive application and devices. The following are some of the fields that are included in the Big Data umbrella.

1.2.1 Benefits of Big Data

Big data is really fundamental to our lives and is emerging as one of the most significant technologies in the modern world. The following are just some of the benefits that we all know very well:

- Using information from social networks like Facebook, marketing agencies are learning the answer for their campaigns, promotions and other advertising media.
- By using information from social networks, such as consumer likings and product perception, production are planned on this basis by retail organization and product companies.
- Using data from previous patients' medical records, hospitals provide a better and faster service.

1.2.2 **Big Data Technologies**

Big Data technologies are significant in providing more precise analysis, which can lead to more tangible decisions, resulting in large operational efficiency, reduced costs and reduced risk for the company.

To control the power of big data, an infrastructure is needed that can handle and process large amount of unstructured and structured data in real time and shield data security and privacy.

There are diverse innovations available by various providers, for example, Amazon, IBM, Microsoft, and so on., To oversee huge information. In inspecting the innovations that handle enormous information, how about we take a gander at the escorting two kinds of innovation:

1.2.3 Operational Big Data

This incorporates frameworks, for example, Mongo DB that provides operational capacities to intelligent workloads progressively where information is received and stored basically.

Big Data frameworks by NoSQL intends to utilize the newest distributed computing designs that have developed amid the most current decade to empower huge estimations to be performed in a cautious and in efficient way. This makes Big Data operational loads substantially less challenging to oversee, more affordable and fast to realize.

Some NoSQL frameworks can give data on models and patterns in view of continuous information with negligible coding and without the requirement for extra information researchers and foundations.

1.2.4 Analytical Big Data

This incorporates frameworks, for example, Massively Parallel Processing (MPMP) and Map Reduce, database frameworks that give scientific abilities to review and complex breaks down that can influence most or all information.

Guide Reduce gives another technique for information examination that is complimentary to the features gave by SQL and a Map Reduce-based framework that can be resized by novel servers

These two classes of innovation are corresponding and as often as possible sent together.

1.2.5 Operational vs. Analytical Systems

	Operational	Analytical
Latency	1 MS - 100 MS	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	Map Reduce, MPP Database

Table 1

1.2.6. Big Data Challenges

Big data is a vast field, so, there are some challenges that are associated:

- Analysis
- Capturing data
- Storage
- Presentation
- Searching
- Sharing
- Transfer

1.3 CENSUS ANALYSIS USING BIG DATA HADOOP

The objective of the census is to acquire an impartial information about the fluctuations that took place in the socio-economic life and building of the country then the previous census.

To Build and then Analyze the info base of demographic and socio-economic data about the its distribution by age, population, family composition, gender, education, occupation, migratory activity, sources of livelihood, living conditions, language signs, citizenship.

Census holds a plethora of data and analyzing such a data large amount of data in a faster and efficient way is a big challenge and rather cumbersome.

Hadoop provides us with a functionality like parallel retrieval and also provide high reliability which not only help us extract the information in a faster and efficient way but help us to keep the data safe

1.4 Objectives

The goal of the making this project is that we can extrapolate distinctive information from a data that was not so useful before by using Map Reduce algorithms. Extrapolation may include aging in urban or rural areas, Marital status, Population Growth.

It is a general engineering of interconnection of various Name node and Data node of HDFS. The enormous measure of information gathered by using Internet (if online) or from the Database can be set, explored, and put away using the computational information using Hadoop. In this engineering, the info. can be competently shared by various clients and applications under flexible use situations.

1.5 Methodology

1.5.1 Hadoop

Hadoop programming by apache is a structure that allows the distributed administration of large informational indexes in collection of PCs that use simple programming models. It is designed to move from individual servers to a large number of machines, each with capabilities and local computing. Instead of relying on the equipment to transmit a high level of accessibility, the library is designed to recognize and handle faults at the application level, thus transmitting a very accessible administration on a lot of PCs, all of which could be prone to failures .

Some Hadoop modules are the following:

- Hadoop Common: these regular utilities are java libraries that will be used to start Hadoop and also for other Hadoop modules

- Hadoop Distributed File System(HDFS): A file which is distributed provides reliable data storage. Files are divided into blocks and stored at nodes.

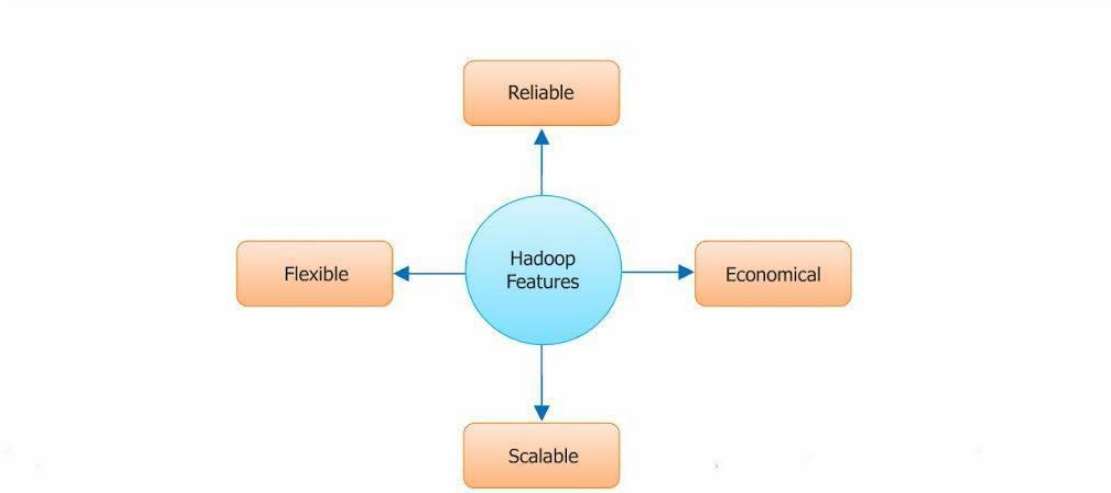

Fig. 1 Hadoop Characteristics

- Hadoop YARN: A system for job scheduling & managing the cluster.

    - Hadoop Map Reduce: It is a framework that handles data parallelly and uses key value pairs to handle information.

1.5.1 Map Reduce Algorithms

Map Reduce is a model that executes in the environment of Hadoop to enhance scalability and simple data handling solutions. This tutorial elucidates the significance of Map Reduce and how it helps in analyzing the Big Data.

Map Reduce is a programming prototype for inscribing applications that can handle large Data in parallel on many nodes. Map Reduce delivers analytical skills for analyzing large

quantity of intricate data.

1.5.1.1 What is Big Data?

It is gathering of huge data that can not be handled utilizing conventional estimation strategies. For instance, the volume of information asked by Facebook or You-tube requires that it be gathered and overseen day by day, it can fall into the class of Big Data. In any case, Big Data isn't just about scale and volume, yet it likewise includes at least one of the accompanying perspectives: speed, assortment, volume and unpredictability.

1.5.1.2 Why Map Reduce?

Traditional commercial frameworks for the most part have a brought together server to store and process information. The accompanying exhibition demonstrates a schematic perspective of an established system . The conventional prototype is definitely not fit for handling high amount of data and can not be overseen by standard database servers. Furthermore, the centralized framework makes an excessive number of bottlenecks when preparing different records in the same time.

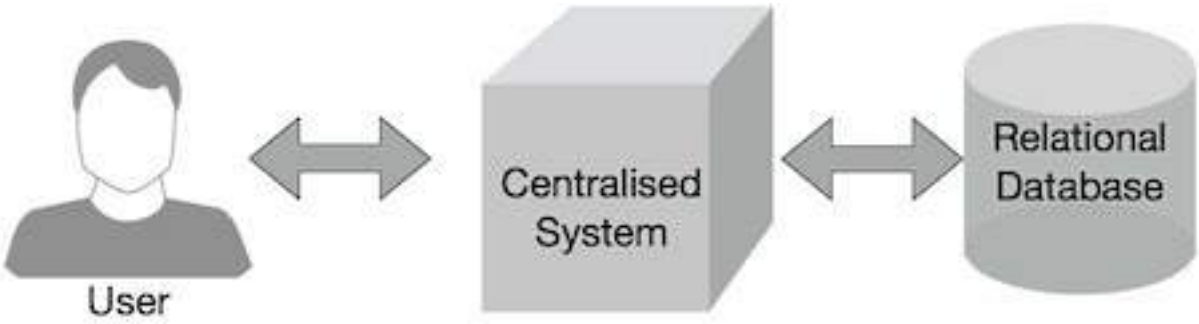

Fig 2 Hadoop Map reduce

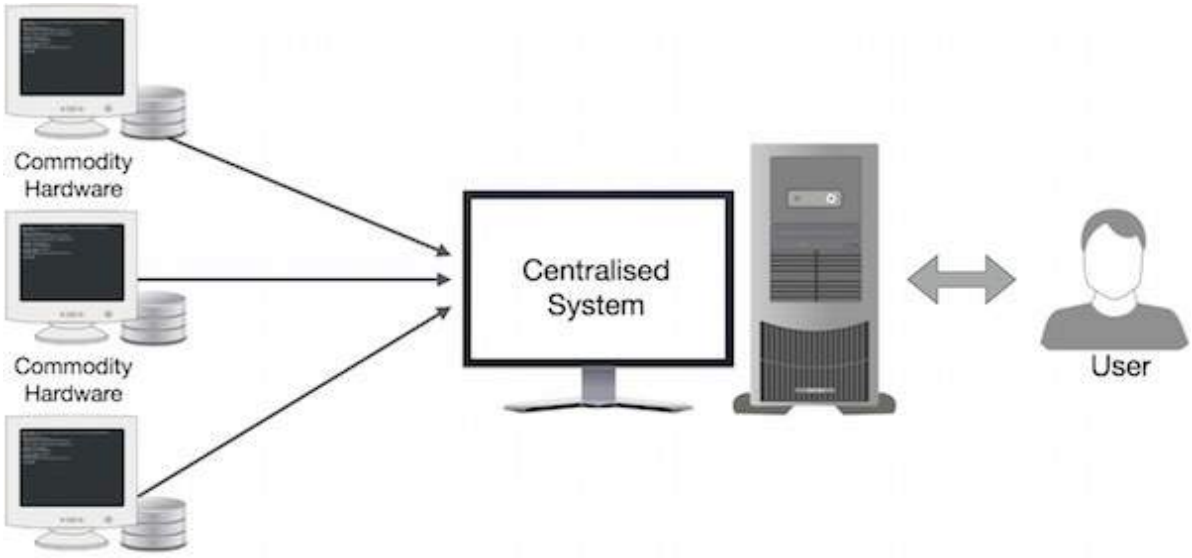

Fig 3 Map Reduce Configuration

1.5.1.3 How Map Reduce Works?

The Minimize Map calculation incorporates two imperative exercises, particularly Map and Minimize.

- The Map undertaking takes a progression of information and change it to another arrangement of information, where the single components are partitioned into tuples (key-value sets).
- The Minimize action takes the Map yield as input and joins those tuples of data (key-value sets) into a moment set of tuples.

The reduction action is always performed after the map is working.

How about we investigate every one of the stages and attempt to comprehend its importance.

Let us try to understand the two tasks Map & Reduce with the help of a small diagram

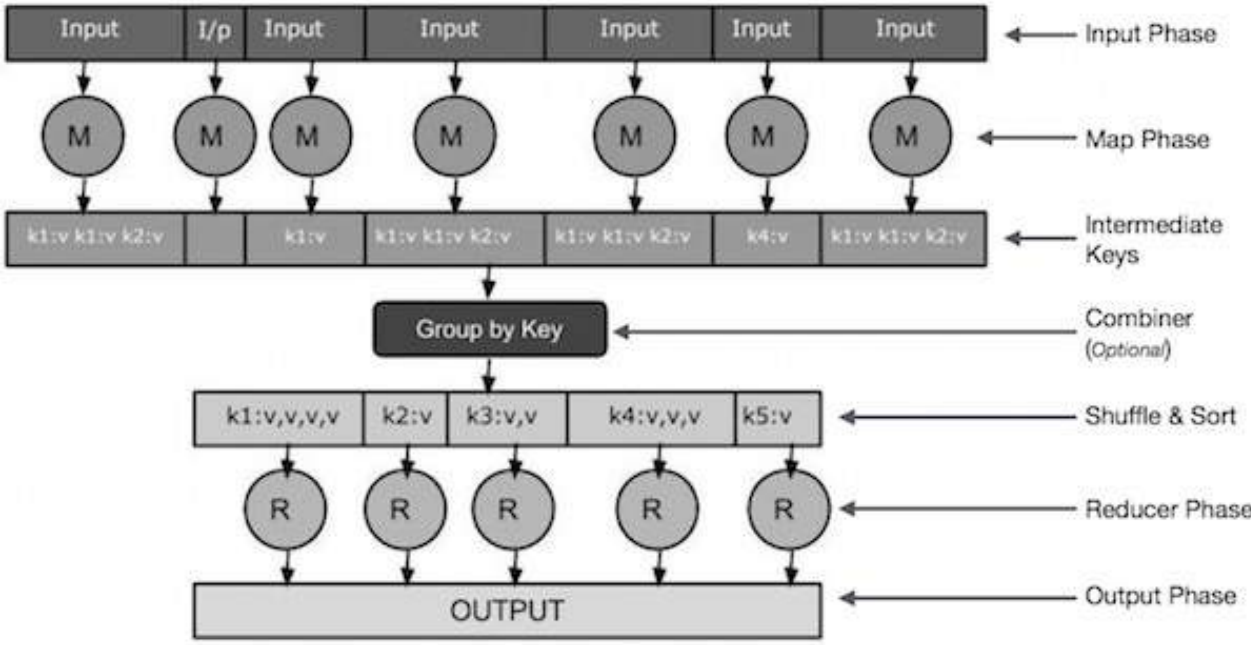

Fig 4 Map Reduce Working

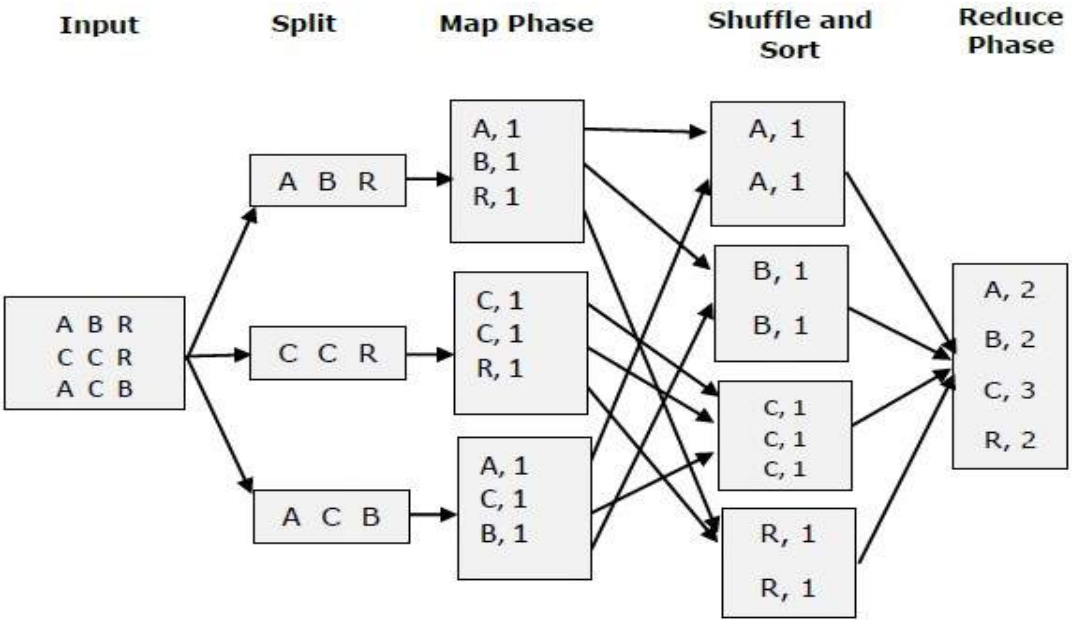

Fig 5 Key & Value Pairs

Above Images shows how Map reduce works and below is few terminology that are use while performing Map reduce algorithms.

- Input stage: Input stage is a record reader that alternates each record into an input document and guides the examined information to the mapper as key-value sets.

- Map - Map is a client characterized work that takes a succession of key-value pairs and process each to create a list of zeroes or more key-value sets.

- Intermediate keys: the key-pairs sets created by the allocator are known as intermediary keys.

- Combiner: A combiner is a sort of native reducer that gathers together same data from the period of the identifiable series map.

- Shuffle and Sort - The development of the reducer starts with Shuffle and Sort. Download the key-esteem sets assembled in the native PC, where Reducer is running. The individual key-value sets are arranged by key in a bigger data list. The information list bunches the same keys together so that there values come frequently.

- Gearbox: the gearbox takes the coupled information of the key-value sets gathered as information and plays out a reduction methods in every one of them. Here the information can be collected, separated and joined in different ways and require an extensive variety of handling. Once the execution is done, it gives zero key-value or more key value pairs for the final step

- Output stage: in the yield stage, we have a yield formatter that deciphers the last key-value sets of the Reducer function and writes them to a document using a record writer.

1.5.1.4 Map Reduce-Example

Let's take an illustration from the real world to understand Map Reduce power. Twitter gets around five hundred million tweets in a day, that is almost Three Thousand tweets every second. Below figure depicts how Tweeter is able to handle tweets with the support of Map Reduce.

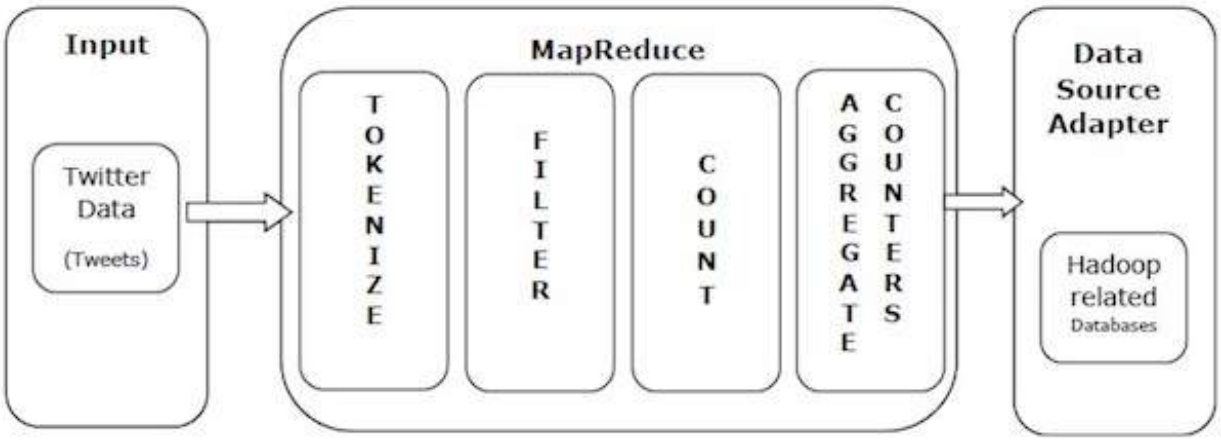

Fig 6 Twitter Working

As appeared in the figure, the Map Reduce algorithm performs out the following activities:

Tokenize: assembles the tweets in symbolic maps and writes them as key-esteem sets.

- Filter: channels the unneeded words from the token maps and composes the filtered maps as key-value sets.
- Count: produces a chip counter for each word.

Total counters: Prepare a total of similar counter values in little reasonable units.

1.5.2 Hive

Hive is an data warehouse center for handling organized data in Hadoop. Dwells in Hadoop to condense Big Data and encourages the analysis and investigation.

At first Hive was launched by Facebook, and later adjusted by Foundation of Apache software and created as an open source under the name of Apache Hive. It is utilized by numerous associations. For instance, Amazon utilizes it in Amazon Elastic Map Reduce.. The hive is not the following

- A database which is relational
- A language for real-time queries and line-level updates

1.5.2.1 Characteristics of the hive

- Stores the schema in a database and processes the data in HDFS.
- It is designed for OLAP.
- Provide the SQL type language for the query called Hive QL or HQL.
- It is familiar, fast, scalable and extensible

1.5.2.2 Hive architecture

The following component diagram describes the architecture of Hive:

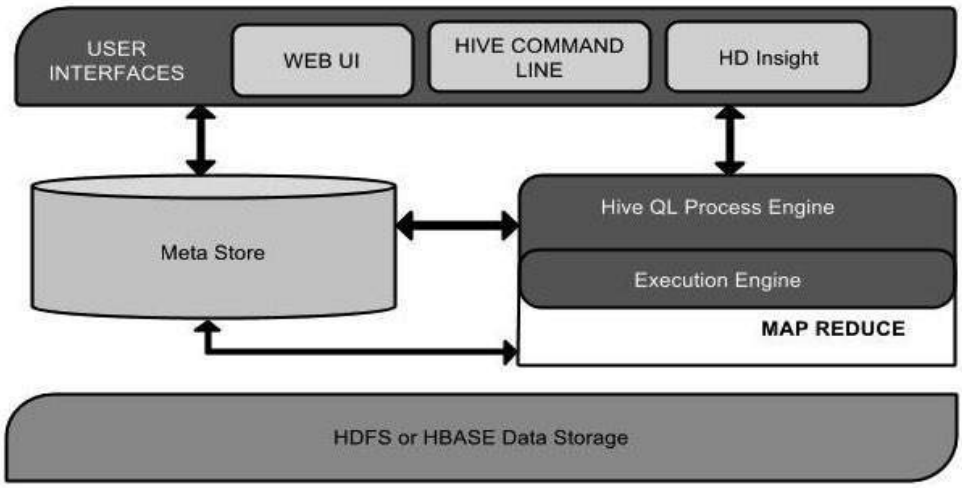

Fig 7 Architecture of Hive

The following diagram depicts the workflow between Hive and Hadoop.

1.5.2.3 Hive Partitioning

Table partitioning implies partitioning table information into a few sections in view of the estimations of specific segments like date or nation, isolate the information records into various documents/catalogs in light of date or nation.

Partitioning should be possible in light of more than segment which will force multi-dimensional structure on catalog stockpiling. For Example, notwithstanding apportioning log records by date segment, we can likewise sup isolate the single day records into nation shrewd separate documents by including nation section into dividing.

1.5.2.4 Hive Bucketing

Typically Partitioning in Hive offers a method for isolating hive table information into numerous records/indexes. Be that as it may, parceling gives successful outcomes when, There are predetermined number of Partition . Similarly, equivalent size partition.
In any case, this may be unrealistic in all situations, similar to when are dividing our tables based geographic areas like nation, some greater nations will have vast segments (ex: 4-5 nations itself contributing 70-80% of aggregate information) where as little nations information will make little Partition (staying all nations on the planet may add to only 20-30 % of aggregate information). In this way, In these cases Partitioning won't be perfect.

To conquer the issue of over apportioning, Hive gives Bucketing idea, another procedure for breaking down table informational indexes into more sensible parts.

Features

- Bucketing idea depends on (hashing capacity on the bucketed column) mod (by add up to number of buckets). The hash function relies upon the kind of the bucketing column. the bucketing column.

- Records with the same bucketed column will always be put away in a similar bucket.

- We utilize CLUSTERED BY proviso to isolate the table into buckets.
- Physically, each bucket is only a record in the table catalog, and Bucket numbering is 1-based.
- Bucketing should be possible alongside Partitioning on Hive tables and even without apportioning.
- Bucketed tables will make similarly disseminated information record parts

1.5.2.5 Comma separated file convers

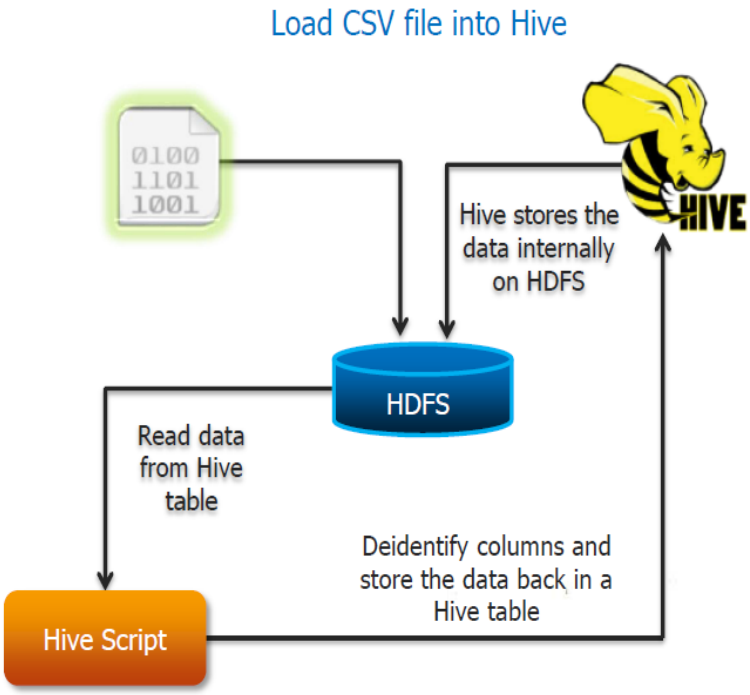

Fig 8 Loading csv file into hdfs

1.5.2.4 Apache zeppelin

Apache Zeppelin is a web-based application that allows interactive data analytics. With Zeppelin, you can make attractive data-driven, interactive and combined documents with a rich set of pre-built language back ends (or interpreters) such as Scala (with Apache Spark), Python (with Apache Spark), Spark SQL, Hive, Markdown, Angular, and Shell.

With a focus on Enterprise, Zeppelin has the following important features:

- Zeppelin verification with LDAP
- Livy integration (REST interface for interacting with Spark)
- Execute jobs as genuine use
- Security

Chapter 2: Literature Survey

I. Topic: Digitizing Indian Census Data for Analytics, Using Big Data Technology (International Journal of Advanced Research in Science, Engineering and Technology Vol. 3, Issue 3, March 2016)

The paper displays a calculated model enumeration information investigation apparatus which can be use by arrangement creator and govt. authorities to outline and casing strategies which is appropriate for the majority. This paper depicts history of statistics information gathering, how they are really digitized, how evaluation info. can be aggregated, issues of collecting enumeration information and a short portrayal of formation of an archive for these information, proposed framework for the model and openings which can be made by actualizing this model.

The Indian Census has a convention of gathering country's information and considered as the best on the planet. India is the biggest vote based systems on the planet.

So Indian census data has a remarkable importance. From more than 130 years, is used as a statistical tool for measuring the country's growth and development. From 1872 when the first census was accompanied in India by the British but it was conducted in some parts of the country. Researchers for numerous fields alike demography, economics, anthropology, sociology, statistics and many other disciplines uses the Indian Census as a basis of data for analysis and fact discovery. The rich diversity of the populaces of India i.e. the data about the population, male female ratio, literacy, income, personal and social status, etc. is gathered by the decennial census which has become one of the tools to apprehend and study India Researchers for various fields like sociology, economics, anthropology, statistics, demography and many additional disciplines practices the Indian Census as a basis of data for analysis and fact discovery. The gorgeous diversity of the populaces of India i.e. the data about the population, male female ratio, literacy, income, personal and social status, etc. is collected by the decennial census which has become one of the tools to understand and study India.

The Census Act was endorsed in 1948 to deliver for the system of conducting population census with duties, responsibilities and liabilities of census officers. The methodical collection of statistics of the country was set up in 1949. We have tons of data collected in the due course from 1872 to 2011, which will give us untouched and unexplored facts and trends of our nation. Since because of the size of this data we can refer it as a Big Data. This paper shows a wording that this information that is gathered can be broke down by utilizing refined investigation instruments that are accessible today.

Huge information alludes to enormous informational indexes described by bigger volumes (by requests of greatness) and more prominent assortment and many-sided quality, created at a higher speed. Big Data refers to data which has larger sizes more than billion zeta bytes. Big Data is about turning imperfect, complex, often unstructured data into implementable information. The data can be text, images, RFID codes, satellite images, search engine hits, tweets, face book tweets etc. Since these data are in wide variety of formats and sizes. It is referred as Big Data. Indian census data is also large sized data which will contain all historical data of 130 years along with other survey data and social networking sites‟ data. So we can refer it as Big Data. The proposed tool will gather all those data and build a ware house which will contain Big Data repository which can be made used for Analytics. The Central store will also contain social networking data since now most of the citizens of the country are active member of internet, so most of the present status, problem and condition of society can also be predicted from these online data. Online data along with census data and various government survey data is gathered under a Big Data infrastructure for analytics.

Usage of Digital Technologies in census data gathering and managing began from 1961, before this data were gathered and treated manually. The major and most significant part in structuring up this prototypical is to gather those data and stock it. Then we are gathering census data which is the secure and subtle data of a nation. Before the buildup is done we will review how those data are stored and processed and what the confronts in collecting and storing it.

-CHALLENGES FACED IN ACCUMULATING CENSUS DATA

- DIFFERENT DATA STORAGE PLATFORMS AND FORMATS

As from the above history we can see that data is recorded and stored using different technique and formats. The problem here will be bringing all those data under one platform and same schemas.

- VARIATIONS IN INFORMATION COLLECTED

In every census data gathered for the country differs accordingly, like name, age, marital status, religion etc. this factors vicissitudes in every census. The table below shows the dissimilarity in the qualities of the data that were gathered during census.

- LIMITATION OF ACCESS TO ENTIRE CENSUS INFORMATION (RAW DATA).

The raw data contains each and every record of every citizen of the nation. Most of the data available is not sufficient enough as they are summarized data. For example, we can get the per capita income of a particular village but we will not get the information about the per capita income of a particular person for the data exposed so far.

Open sources technologies can be merged to make a census data analysis tool. Fig. 1 displays the block diagram of the planned model. Hadoop Infrastructure will be finest suited to make up such an application. Map Reduce Frameworks are frequently in light of Hadoop and Hadoop-like innovations. They work by giving parallel processing capacities that move subsets of the data to distributed servers. The essential utilize is preparing huge measures of data in scalable way.

Data is logically organized into tables, rows and columns in HBASE. We can store census data as a cluster in data nodes and whose links will be managed by the name node. The grouping of collected data will be maintained in a distributed environment, using HDFS.

The clusters will also contain the data collected from the social networking sites and internet which will be refined using MAPREDUCE [5], NoSQL and HIVE will be used for retrieving data from the store.

II. Topic: A comparative analysis of population ageing in urban and rural areas of England and Wales, and Poland over the last three census intervals (Population, space and Place Volume14, Issue 5, Version of Record online:17 SEP 2008)

Populace maturing has turned into a component of numerous European and other Western nations in the course of the most recent two decades as more seasoned individuals turn out to be more prevailing in their statistic structures. This pattern has the potential for major financial and social effects, which can be placed in both the national and nearby point of view.

This record decides the plausibility of associating populace data at a provincial level over an historical period and following change in age structures in two nations, one in Western Europe (England and Wales in the United Kingdom) and the other in Central and Eastern Europe (Poland).

The correlations are expanding on the examination of the measurements of the last four populace censuses in these nations and, regardless, the outcomes are shown both at the national level and for an unmistakable arrangement of little spatial units separated into urban and rustic classifications.

The results affirm the populace development already found in the country territories of England and Wales and in the urban regions in Poland. The maturing of the populace happened first in the regions of England and Wales, in spite of the fact that the development in the quantity of elderly individuals in Polish germy quickened amid the latest time of internee. Late statistic floats in the two nations at the national and little scale reflect current financial modifications, as well as the statistic and political occasions of the past.

The outdated recovery of the northern and western territories of Poland by an energetic versatile people after the Second World War, after its movement from Germany, is reflected in the meantime in the maturing of the populace in these regions since the landing of the main entries. more established gatherings.

III Topic: SOCIOECONOMIC STATUS, MARITAL STATUS AND CHILDLESSNESS IN MEN AND WOMEN: AN ANALYSIS OF CENSUS DATA

FROM SIX COUNTRIES (J. Biosoc. Sci., (2011) 43, 619–635, Cambridge University Press, 2011 doi:10.1017/S002193201100023X First published online 11 May 2011)

This investigation thinks about the impacts of two distinct types of human capital - pay and education - on marital status and the nonattendance of kids independently by sexual orientation in six unique nations. The censuses of Brazil, Mexico, Panama, South Africa, the United States and Venezuela, which go back to at least 2000, utilized very nearly 10 million individual records of individuals in the vicinity of 16 and 50 years of age, to examine the connection between training, pay and conjugal status and childlessness in people. Concerning, the findings for both result factors are firmly steady over each of the six nations. Most elevated wage guys and bring down pay females have the most elevated extent of ever-hitched and the least extent of childlessness (utilizing an intermediary for childlessness: possess youngsters in the family or not).

There is no reliable consistency of results as to instruction between the genders or between nations. For agree, a lower perceptual purpose of low-pay men is chosen by ladies, in light of the fact that for ladies the male status and assets gave by men are vital criteria in the determination of the accomplice.

Accordingly, a higher extent of low-salary men stay unmarried and childless.

Accordingly choice appears to assume a part in present day social orders. Enumeration information from six nations are utilized as given by IPUMS International (Minnesota Population Center, Integrated Public Use Micro Data Series, International, Version 6.0 [Machine-comprehensible database], University of Minnesota, Minneapolis, 2010): Brazil, Mexico, Panama, South Africa, USA and Venezuela. Notwithstanding articulated differences in monetary improvement every one of the six nations experienced a significant richness drop from 1960 to 2010, i.e. demonstrating a cutting edge design concerning fruitfulness (Table 1). These are all the IPUMS-appropriated censuses from the present century that offer data on conjugal status, age, childlessness, instruction and salary. Conjugal status is coded in the crude IPUMS records as single/never-wedded, wedded/in association, isolated/separated/life partner truant, widowed, dwelling together. This

variable was recorded into a straightforward twofold complexity: at any point wedded versus never-wedded. Age is in years at the season of the registration. The gauge of childlessness was, operationally, a dichotomization of the quantity of one's own kids in the family: zero versus at least one. A similar variable must be utilized for the two people in the examples, as this check was not independently by conjugal accomplice. Instructive fulfillment is classified as takes after: E1, not as much as essential finished; E2, essential finished; E3, auxiliary finished; E4, college finished. As the likelihood of marriage, having youngsters, and pay are emphatically age subordinate, plots look at rates of marriage and childlessness age by age. Relapses are pooled over ages yet constantly join a term for a (direct) age effect.

Salary was implicit as takes after: South Africa 2007, USA 2005 and Venezuela 2001, Brazil 2000 for the aggregate pay of a man from all sources in the earlier year was utilized. For Panama 2001, month to month wage and pay, the main wage data IPUMS offers from this registration, was utilized. With the end goal of a portion of the nation specific investigations underneath, pay for each enumeration, each sex, and every time of age was partitioned into four quartiles named Q1 to Q4, aside from in Mexico where, attributable to especially extreme truncation of the crude information, there are just three 'quartiles'. In the relapses, for distributional reasons wage was entered as the square root (see likewise Huber et al., 2010).

There is a hazard in utilizing the variable 'no possess kids in the family unit' as a gauge of childlessness, especially in men. For example, after separation or division, kids regularly stay in their mom's as opposed to the father's family. Nonetheless, attributable to the way that the quantity of kids is now and then not inspected for ladies and never examined for men by normal censuses, this variable is the main plausibility if any gauge of childlessness for both genders crosswise over nations and societies is to be broke down.

The latest enlightening records offered by IPUMS were utilized: Brazil 2000, Mexico 2000, Panama 2001, USA 2005, South Africa 2007, USA 2005 and Venezuela 2001. Basically

people in 'their conceptive years', from the age of 16 to the age of 50 years, were joined. For Brazil 2000, USA 2005 and South Africa 2007, the information of the ethnic bigger part were bankrupt down (Brazil 2005: Whites; USA 2005: Whites; South Africa 2007: Black Africans). For Mexico 2000 and Panama 2001, which offer no additional data on ethnicity, all people of indigenous begin were restricted. For Venezuela 2001, IPUMS gives no data on ethnicity, so all people inside the right age go were combined. The outlines totaled 4,837,325 men and 5,218,288 ladies (Table 1).

For every whole number age in the vicinity of 16 and 50 the example part of people who had never been hitched and the example division who live with no of their own kids in the family unit were re-plotted. These plots are separate by sex and by wage quartile and for every training classification too. Likewise, independently by sex and nation, strategic relapses of parallel conjugal status were ascertained (0=never hitched, single; 1=married eventually) on pay, ordered instructive fulfillment and age. Independently by sex and nation, strategic relapses were likewise ascertained of the parallel intermediary for childlessness (0=no tyke in family unit; 1=child(Ren) in family) on salary, classified instructive achievement and age. As the connection amongst age and conjugal status and in addition childlessness isn't straight, a quadratic term for age was incorporated into every single strategic model for the never-wedded variable.

For the two people, the extent of never-wedded people and in addition those without possess youngsters in the family unit quickly diminishes with age through the middle30s and stays low. The extent of people without kids in the family unit correspondingly diminishes through the center 30s yet in a few examples the pattern switches at about age 40, maybe on the grounds that youngsters begin leaving the parental family unit.

out age 40, maybe on the grounds that youngsters begin leaving the parental family unit. Figure 1 exhibits the measurements of conjugal status and childlessness by age and wage quartile. Inside the men of every national example, at relatively every age the level of never-wedded people is most noteworthy in the least salary quartile, and when the most astounding three quartiles differ in rate never-wedded, the differences are almost constantly concordant with the pay sort. In like manner, at relatively every age, the level of childless

people is most astounding in the least pay quartile, yet the differences are less articulated among the quartiles over that. In ladies, the example is the turn around (Fig. 2). The most elevated level of never wedded ladies is found in the most astounding pay quartile: Q4 in Brazil, Panama, Venezuela and US and Q3 in Mexico.

In the US, differences among pay quartiles over the most minimal are little, and in South Africa there is by all accounts little effect of salary on conjugal status by any means. The level of childless ladies by wage quartile looks like the example for extent never-wedded. For this result, the effect of salary quartile seems, by all accounts, to be more articulated, even in South Africa and the US. The relationship between instructive accomplishment and conjugal status in men demonstrates a subtler picture (Fig. 3).

In Brazil and Mexico, the most astounding extents of never-wedded men are found in the most noteworthy instructive class and the least extents in the least two instructive classifications. Be that as it may, in South Africa, aside from at the most youthful ages, the extent of never-wedded men is least in the most elevated instructive classification and most noteworthy in the most minimal instructive class. In both Panama and Venezuela, a higher extent of young fellows are unmarried among the most elevated instructive classification, while among the more established men it is those of the least instructive class who all the more frequently have never hitched. These inversions recommend either a different dispersion of ages at first marriage or truly unforeseen accomplice effects (see Discussion). In the US, with expanding age the level of unmarried men increments in the least instructive class. The relationship between instructive fulfillment and the intermediary for childlessness is like that between conjugal status and instructive achievement with the exception of the US, where the bends are all the more generally spread at more youthful ages.

IV. Topic: Cluster analysis of census data using the symbolic data approach (Advances in Data Analysis and Classification
October 2008, Volume 2, Issue 2, pp 163–176)

Emblematic information investigation (SDA) is an expansion of standard information examination where representative information tables are utilized as info and emblematic items are made yield therefore. The information units are called representative since they are more intricate than standard ones, as they contain qualities or classifications, as well as incorporate interior variety and structure. SDA depends on four spaces: the space of people, the space of ideas, the space of portrayals, and the space of representative articles. The space of portrayals models people, while the space of representative items models ideas

The point of this paper is to examine the financial specialization of the Italian nearby work frameworks (sets of bordering districts with a high level of self-regulation of day by day worker travel) by utilizing the Symbolic Data approach, based on information got from the Census of Industrial and Service Activities.

Emblematic information investigation is an effective device to speak to classes of measurable units that might be inferred from the earlier through a definition, or through an accumulation calculation.

The principal precise commitment on the utilization of representative information investigation in measurements is the book distributed by Bock and D-day (2000). Since this work, much work has been done in the field of multivariate examination, perception of complex information, and so forth. All the more as of late, an audit of the ideas and techniques created under the header of emblematic information examination is the work by Ballard and D-day (2003).

All commitments on emblematic information investigation have been given to methodological issues while exact examinations on genuine information are generally few. This paper tests the utilization of the representative information investigation strategies and apparatuses to arrange, through a bunching technique, neighborhoods Italy based on their financial specialization.

The neighborhoods be breaking down are spoken to by nearby work frameworks (LLS) distinguished by the National Statistical Institute (ISTAT 2006) based on the 2001 Population Census information. A LLS is an arrangement of coterminous districts that, based on the Population Census information, display a high level of self-control of every day worker travel. Subsequently, the information are progressively organized: the LLSs are the objects of the investigation while essential information are alluded to regions.

Simply this progressively information structure recommends to utilize the emblematic information examination approach. The financial specialization of each LLS (second request protest) is depicted by interim factors whose limits are the quartiles of the district (first request question) values.

Chapter 3: System Design

3.1 Cloudera Setup

Cloudera demo is setup on oracle virtual box which is connected to servers like winscp and putty.

3.1.2 Multi Node cluster setup

One master node and two slave node are structured to analyze the data more efficiently

IP address (Master): 192.168.56.102
IP address (slave 1): 192.168.56.103
IP address (slave 2: 192.168.56.104

3.1.2 Configuration setup

Operating System Name: nn1, dn1, dn2
Type: Linux
Version: Ubuntu (64 bit)
Memory allocated: 2048 Mb
Network: Adapter1: NAT & Adapter2: Host-only Network (VirtualBox Host Only Network)
Hadoop Version: Hadoop-2.7.1
Java version: jdk1.8.0_171

3.2 Ubuntu Setup

Ubuntu 14.04 is setup on oracle virtual box which is connected to openssh servers like Winscp and Putty.

3.2.1 Single Node cluster

A single node is established on Ubuntu14.04

IP address:192.168.56.101

3.2.2 Configuration setup

Operating System Name: dn1
Type: Linux
Version: Ubuntu (64 bit)
Memory allocated: 2048 Mb
Network: Adapter1: NAT & Adapter2: Host-only Network(VirtualBox Host only Network)
Hadoop Version: Hadoop-2.7.1
Java version: jdk1.8.0_171

3.3 Winscp

WinSCP (Windows Secure Copy) is a open, uncluttered FTP, S3 SCP client, source SFTP and Amazon for Windows. Its main function is to transfer protected files amid the computer nearby & remotely. In adding, WinSCP delivers basic file management and file management functionality. For safe transmissions, it uses Secure Shell (SSH) and maintenances the SCP protocol as well as SFTP

3.4 Putty

Putty is an unrestricted and open-source terminal emulator and net file transmission app. It maintenance numerous network protocols, encompassing of SCP, SSH, Telnet, and raw socket connection. It can also associate with a sequential port.

3.5 Oracle virtual box

Oracle VirtualBox (Sun xVM VirtualBox) a free, unrestricted and open-source hypervisor for x86 computers currently being established by Oracle Corporation.

3.6 Analysis

Analysis of big data in Hadoop comprises taking the enormous data sets and managing them. In thickly network which is distributed, data impending from number of nodes entails of both structured as well as non-structured data. The prevailing database systems were intended to talk about only structured data which is in minute quantity. So, the heterogeneity with that kind of databases methods is becoming puzzling for storage and analyzing large data.

Following are some usual issues observed at the time of analysis.

i. Computation

ii. Data Management

To analyze the Big Data generated, we are using a software name Cloud Era and Ubuntu. Compendium of Map Reduce and Hive queries are executed to compute the wanted value from the data in the software.

3.7 Data Storage model-HDFS

The Apache Hadoop makes open-source programming for strong, flexible, appropriated figure. The Hadoop programming bids Hadoop Distributed File System (HDFS), an appropriated documentation structure that gives high-throughput entrance to application information. It gives Big table-like capacity over Hadoop and HDFS, and easy to use Java API for client get to (scrutinizes and creates). It is used for facilitating of immense tables on gatherings of thing gear and to support irregular, persistent read/constitute access to your Big Data.

3.8 Data processing Framework-Hadoop Map Reduce:

Map Reduce is a parallel programming perspective successfully used by broad Internet authority associations to achieve calculation on tremendous quantity of data. In the wake of being unambiguously exceptional by Google, it has in like manner been realized by the open source bunch through the Hadoop expand. The key quality of the Map Reduce show is its inherently abnormal state of conceivable parallelism. In Hadoop Map Reduce framework, the considering is divided along with two stages: Map and Reduce. Framework a key/regard join to create a game plan of mostly key/regard matches, and Reduce mixes each center regard to shape the last yield.

Nodes	Daemons	Properties
Master	Name node, Resource Manager, Secondary Name node	Main server for parallel distribution of data and its storage
Slave1	Data node, Node manager	A Data node stores data in Hadoop distributed file system
Slave2	Data node, Node manager	A Data node stores data in Hadoop distributed file system

Table2 : Hadoop Cluster

3.9 Hadoop Installation

- Hadoop is going to be installed on ubuntu12.04 and cloud era demo.
- Hadoop-2.7.1 version of is going to run with jdk-version 8.
- Java is necessary framework needed for the Map-reduce program to run
- All steps in installation should be carefully done in order for Hadoop to run.

3.9.1 Installation steps

NOTE : This installation are done on ubuntu

- Sudo-apt-get install update (In order to check for updates)
- Sudo-apt-get install open-ssh server (In order to connect ssh servers like putty and winscp)
- Move Hadoop and Java tar files into download folder
- Tar –xzvf Hadoop-2.7.1-tar.gz
- Tar –xzvf jdk1.8.0.tar.gz
- Sudo-mkdir –P usr/local/java
- Sudo-mv jdk1.8.0 usr/local/java
- Sudo-mv Hadoop /usr/local/Hadoop
- update alternatives for java:
  - Sudo-update-alternatives –install “/usr/bin/java” “java” “/usr/local/java/bin/java”

- Sudo update-alternatives –install "/usr/bin/javac" "javac" "/usr/local/java/bin/javac"
- Sudo update-alternatives –install "/usr/bin/javaws" "javaws" "/usr/local/java/bin/javaws"

➢ Set java alternatives path:

- Sudo update-alternatives –setjava "/usr/bin/java" "java" "/usr/local/java/bin/java"
- Sudo update-alternatives –setjava "/usr/bin/javac" "javac" "/usr/local/java/bin/javac"
- Sudo update-alternatives –setjava "/usr/bin/javaws" "javaws" "/usr/local/java/bin/javaws"

➢ export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/ export

➢ Set variables for Hadoop

- Export HADOOPHOME=/home/$USER/work/hadoop-2.7.1/ export
- Export HADOOPMAPREDHOME=$HADOOPHOME
- Export HADOOPCOMMONHOME=$HADOOPHOME
- Export HADOOPHDFSHOME=$HADOOPHOME
- Export YARNHOME=$HADOOPHOME
- Export HADOOP_COMMON_LIB_NATIVE_DIR="$HADOOP_HOME/lib/native"
- Export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib" export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$ JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64/ mapred-env.sh export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64

- Configure files for Hadoop

  - core-site.xml:

```
<configuration>
    <property>
      < name> fs.defaultFS </name>
       <value> hdfs://localhost:9000 </value>
     </property>
</configuration>
```

  - hdfs-site.xml:

```
<configuration>
  <property>
    <name> dfs.-Replication </name >
    < value> 1 </value>
  < /property>
  <property>
     <name> dfs.namenode. name. dir </name>
      <value> /home/$USER/work/hadoop26data/Dfs/name </value>
  </property>
  < property >
     <name > dfs.datanode . data.dir </name>
      <value> /home/$USER/work/hadoop26data/Dfs/data</value>
   </property>
  </configuration>
```

  - mapred-site.xml:

```
< configuration >
```

```
<property >
<name>mapreduce.framework.name>/name>
<value>yarn</value>
</property>
</configuration>
```

- yarn-site.xml:

```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle>/value>
</property>
</configuration>
```

- Hadoop name node -format
- start-dfs.sh
- start-yarn.sh
- jps

Now the Hadoop machine is ready to execute commands of Map Reduce

Chapter 4: Performance Analysis

4.1 Analysis on Hive

A Sample on Analysis of marital census done on Hive with Bucketing and partitioning algorithm used.

Winscp is a file transfer tool which is used here to configure Hadoop

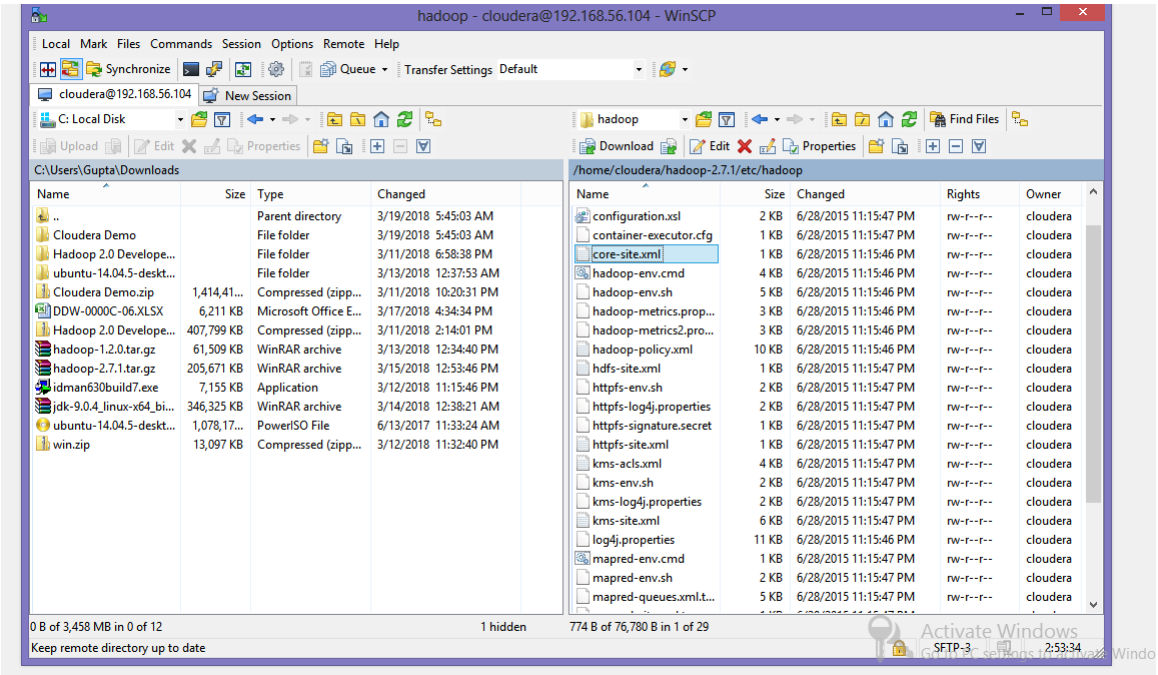

Fig 9 Working with winscp

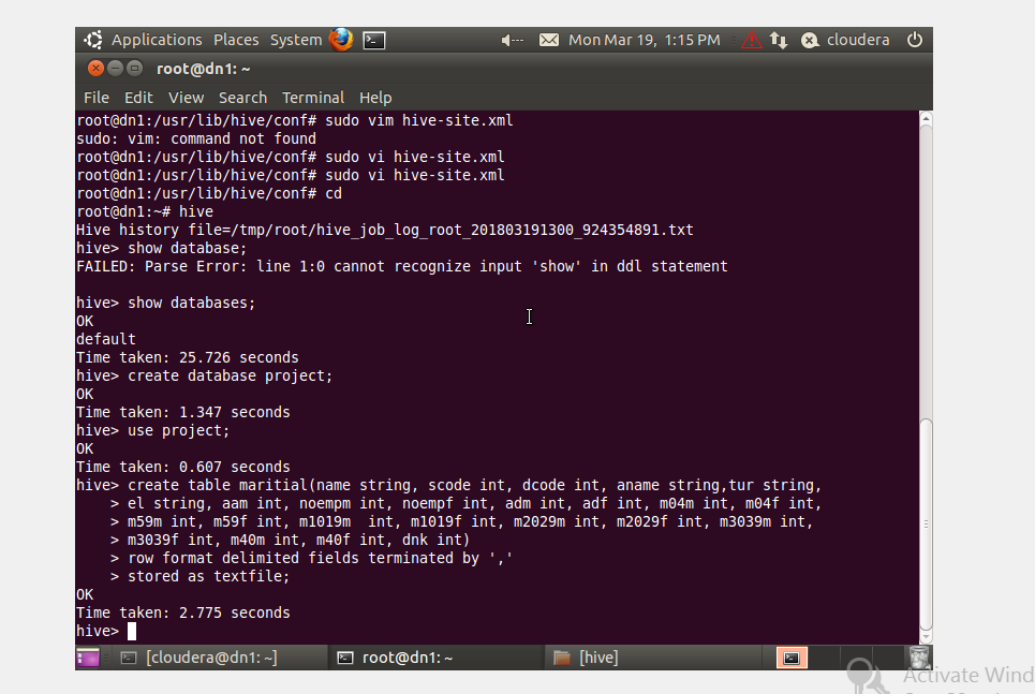

Fig 10 Creating table in hive

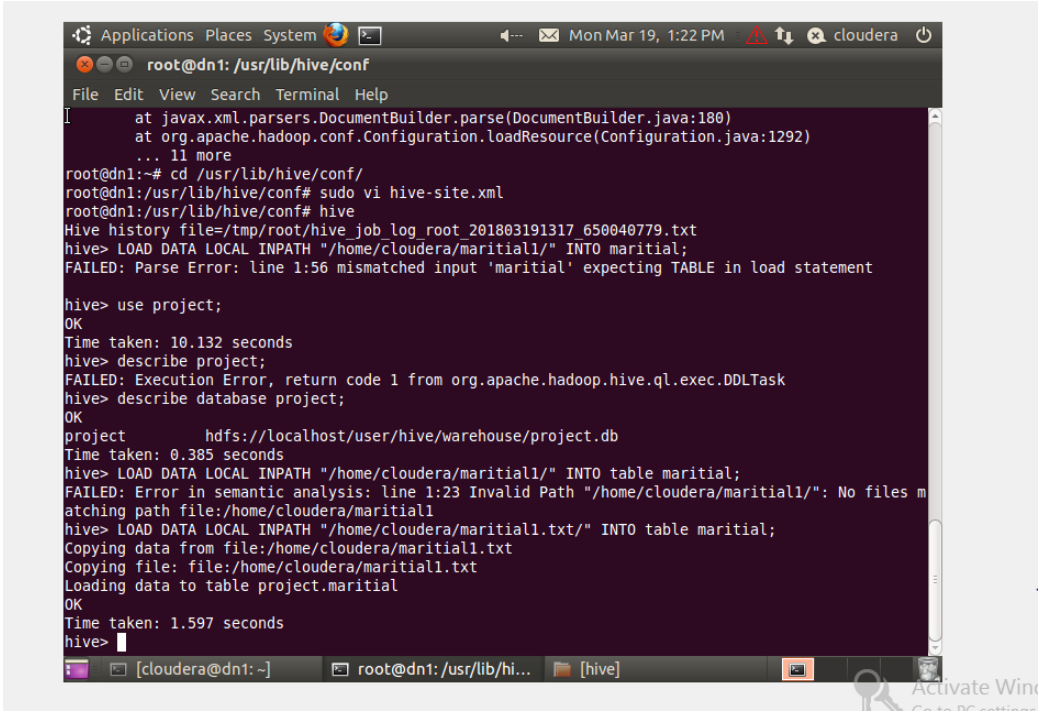

Fig 11 Loading data in hive

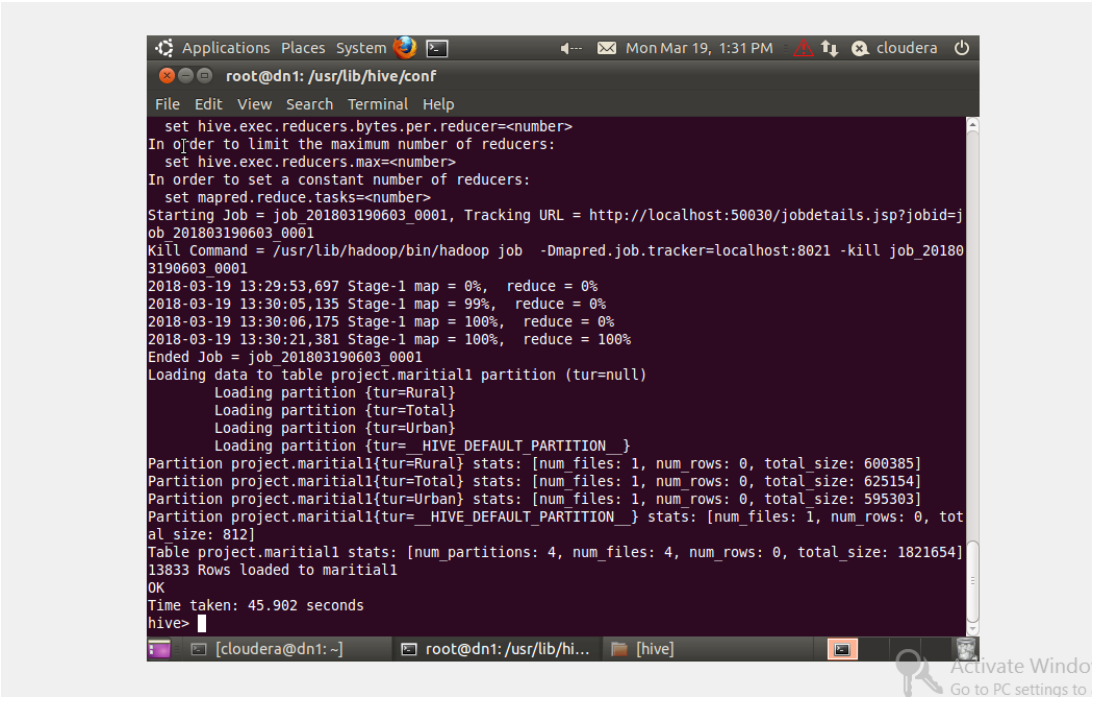

Fig 12 Applying Partition algorithim

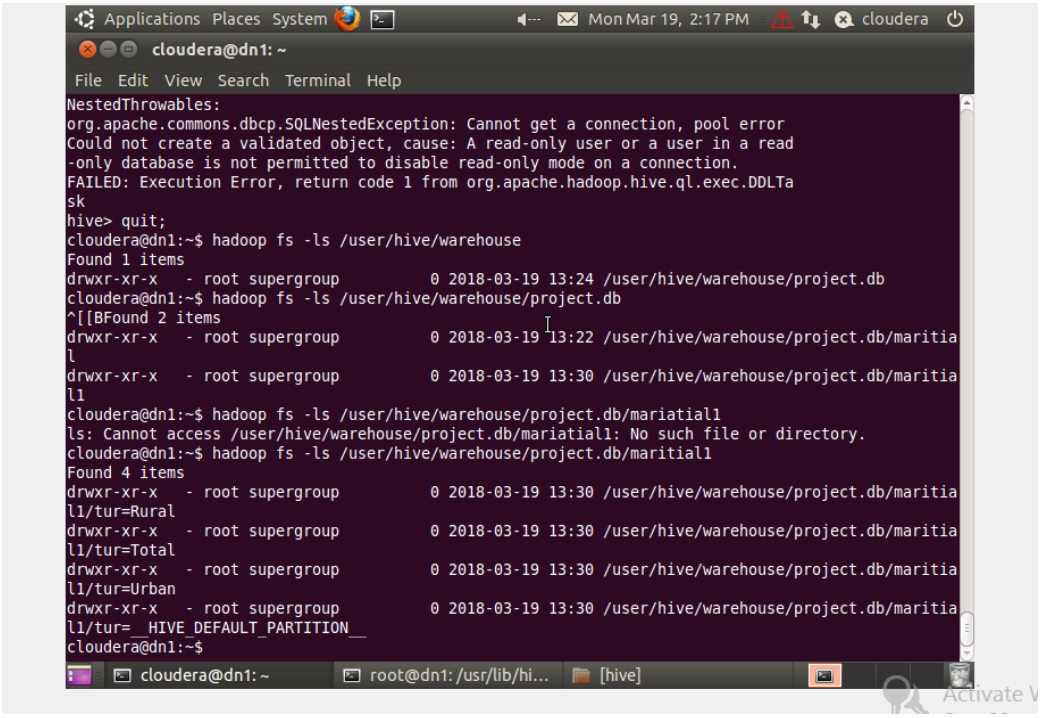

Fig 13 Checking the warehouse of hive part1

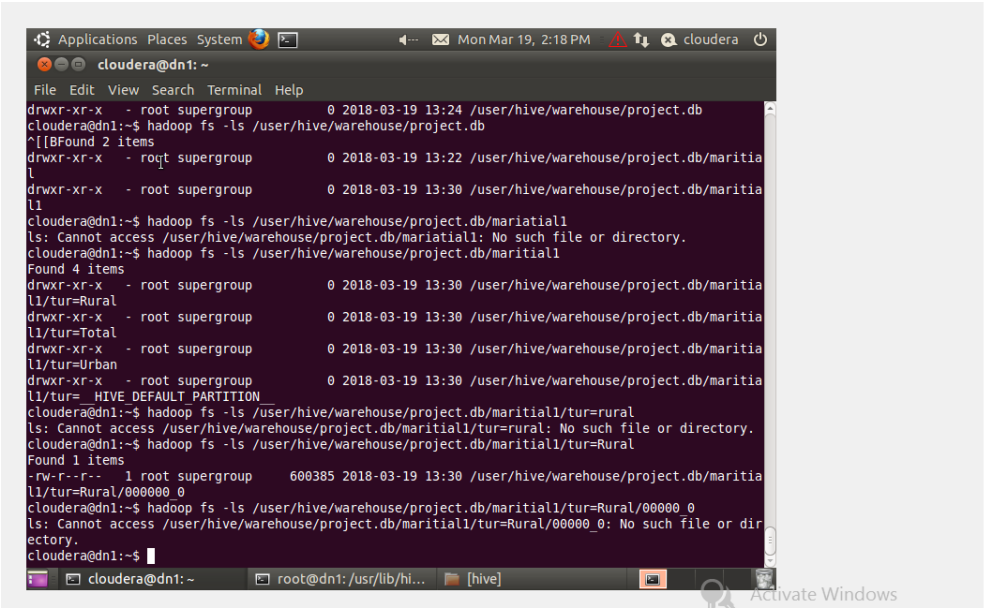

Fig 14 Checking the warehouse of hive part 2

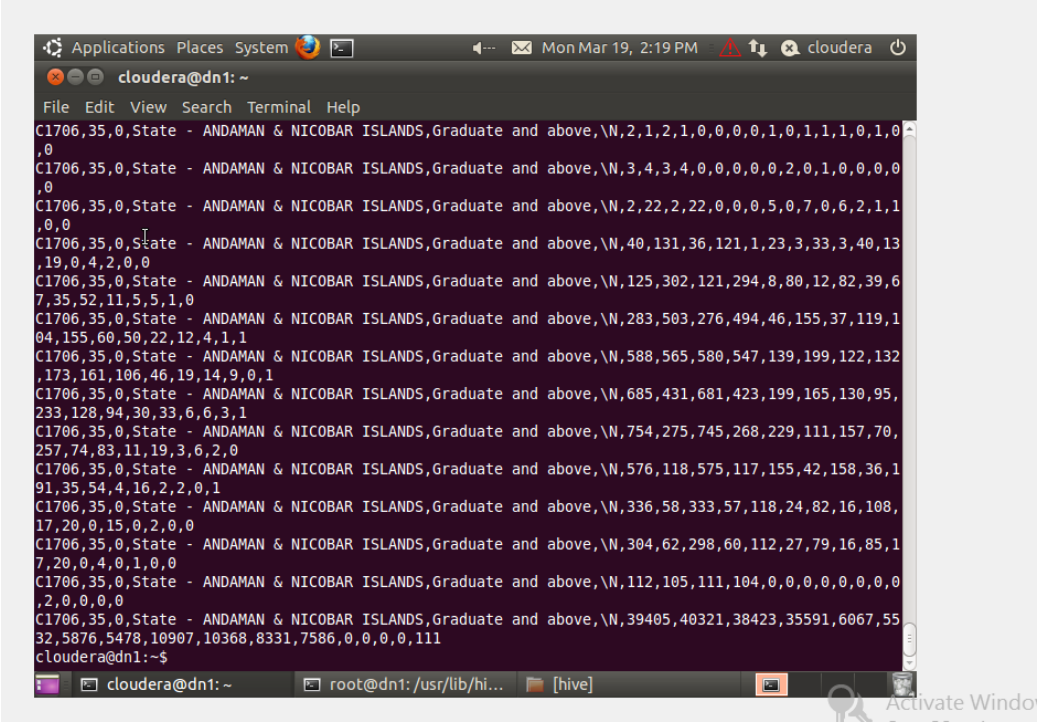

Fig 15 Data is loading into the table

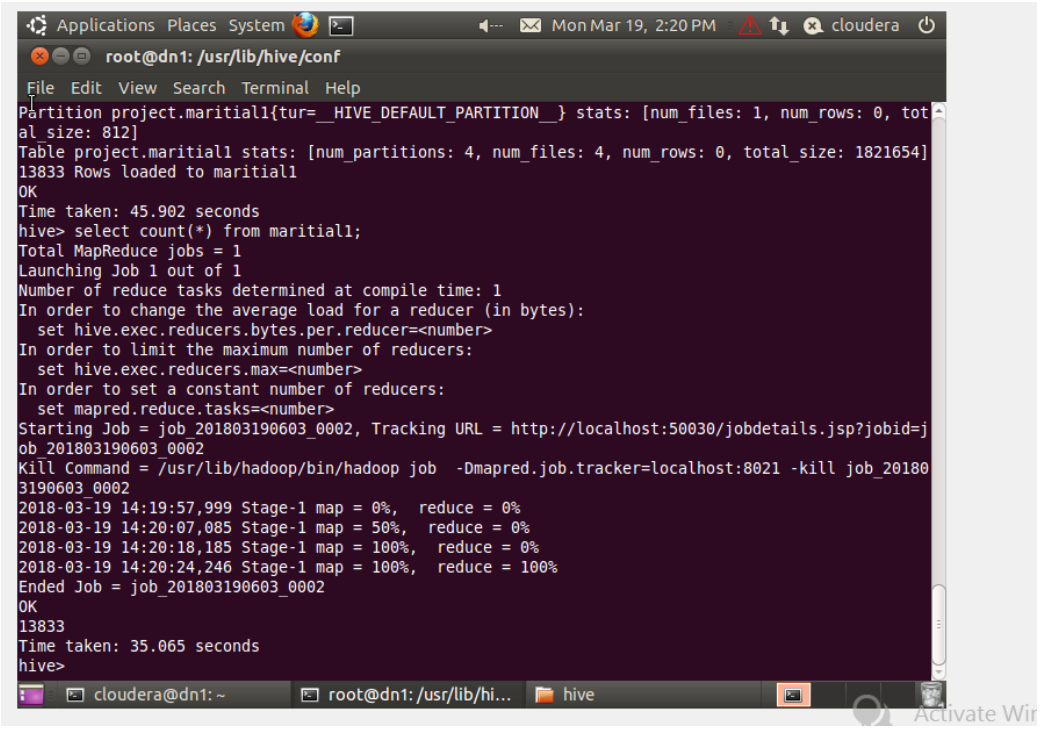

Fig 16 Running the command on Hive

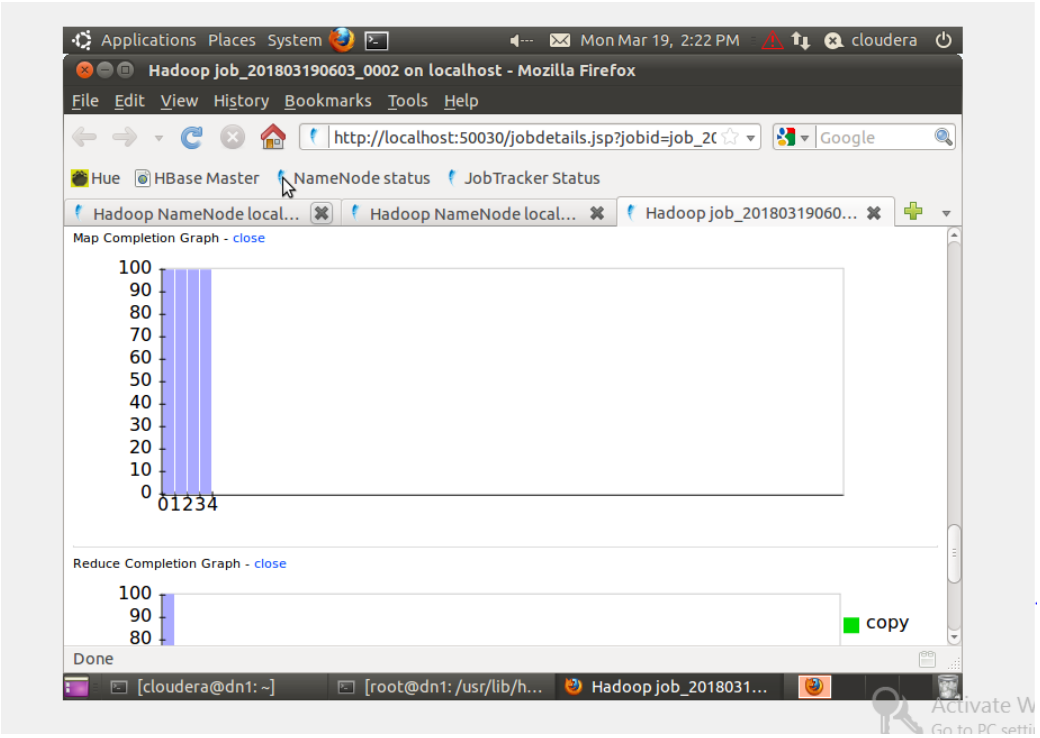

Fig 17 Checking the GUI for Hive

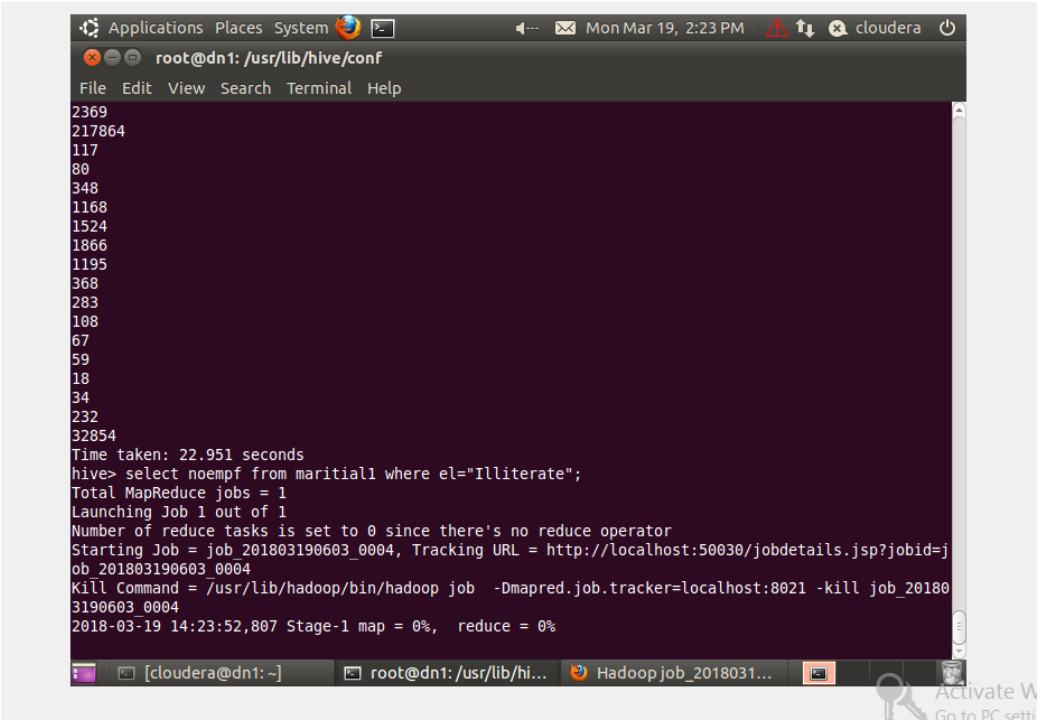

Fig 18 Running another job on the Hive

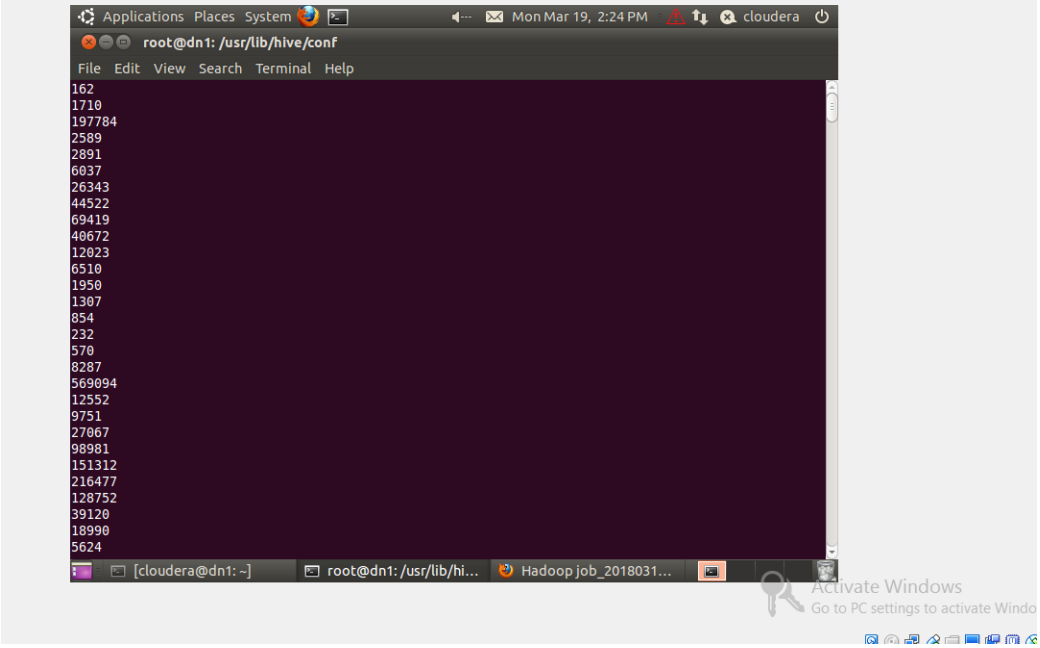

Fig 19 Outcome of the command

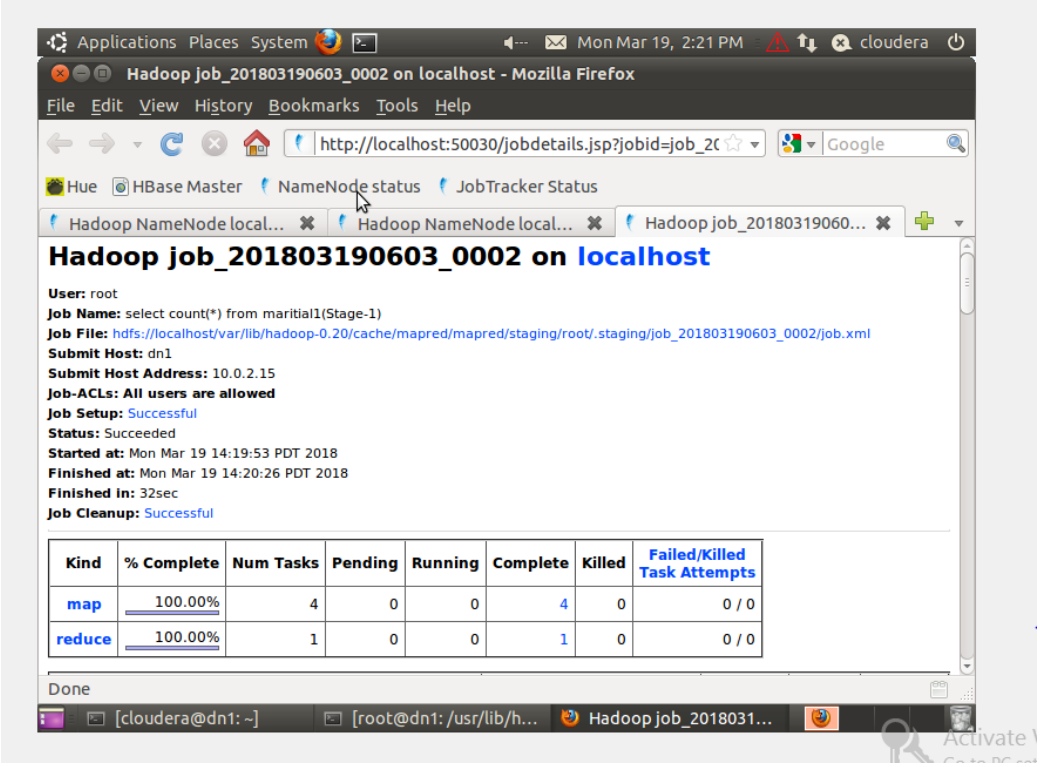

Fig 20 GUI for the command run on the Hive

Chapter 5: Conclusion

5.1 Future Scope

- Use of Apache Zeppelin to visualize the analysis results.
- Convert all the data form into a repository which is uniform based on some predefined data form.
- Try to grasp more knowledge on the topic, and if possible making it my research area
- Use upcoming new technology of Big Data in the project

5.2 Conclusion

Big data is problem set that is becoming a vast field by the gathering of huge amount of data, these data is able to provide info. that can work to help organization to make right decision and in case of my project, the census data can provide huge info. amount our countries Development. During My project, I get to knew the power of data. I learnt new technologies like Map reduce, flume, hive, java programs, Hadoop and was able incorporate all the new technology in my project. In my project I was able to built a three node cluster on cloud era, Hive analysis, Hadoop installation, work on the Hadoop with java programs, interface for human interaction, Flume and much more.

References

[1] https://www.planet-data.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf

[2] ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Big_Data_Analytics.pdf

[3] https://www.tutorialspoint.com/hive/

[4] https://www.tutorialspoint.com/map_reduce/

[5] http://www.fujitsu.com/hr/Images/WhiteBookofBigData.pdf

[6] https://www.ijarset.com/upload/2016/march/30_IJARSET_remyapanicker.pdf

[7] http://onlinelibrary.wiley.com/doi/10.1002/psp.488/pdf

[8] https://link.springer.com/article/10.1007/s11634-008-0024-5

[9]https://www.researchgate.net/profile/Susanne_Huber/publication/51467453_Socioeconomic_status_marital_status_and_childlessness_in_men_and_women_An_analysis_of_census_data_from_six_countries/links/0a85e5392d3f03227c000000/Socioeconomic-status-marital-status-and-childlessness-in-men-and-women-An-analysis-of-census-data-from-six-countries.pdf

[10] Remya Panicker "Topic: Digitizing Indian Census Data for Analytics, Using Big Data Technology", (International Journal of Advanced Research in Science, Engineering and Technology Vol. 3, Issue 3, March 2016)

[11] N. S. walford, S. kurek “Topic: A comparative analysis of population ageing in urban and rural areas of England and Wales, and Poland over the last three census intervals”, (Population, space and Place Volume14, Issue 5, Version of Record online:17 SEP 2008)

[12] MARTIN FIEDER , SUSANNE HUBER and FRED L. BOOKSTEIN “Topic: SOCIOECONOMIC STATUS, MARITAL STATUS AND CHILDLESSNESS IN MEN AND WOMEN: AN ANALYSIS OF CENSUS DATA FROM SIX COUNTRIES”, (J. Biosoc. Sci., (2011) 43, 619–635, Cambridge University Press, 2011 doi:10.1017/S002193201100023X First published online 11 May 2011)

[13] Antonio giusti, laura grassini “Topic: Cluster analysis of census data using the symbolic data approach”, (Advances in Data Analysis and Classification October 2008, Volume 2, Issue 2, pp 163–176)