**CHAPTER 1**

**INTRODUCTION**

**1.1 Data**

Data is not just information. It requires interpretation to become information. There must be several factors to be considered to translate data into information. Those factors are decided by the creator of data and the information. Data can be measured, collected reported and visualized using graphs, images etc. data becomes information only when it is analyzed. The term metadata is a data which gives information about other data. Data have reference to process or event can have temporal component. Meta data can be involved, specified or given. The temporal components are involved in all the cases so consider when temperature logger device receives data from temperature sensor and it will record the date, time and temperature.

Mainframe data is of binary number where the data is represented by 0’s and 1’s. It is just opposite to analog representation. Digital data is used in latest computer system. When the data move from or to it will be in sequential data manner and if data moves within a computer it will be in lateral data manner. Computer data is the information stored in computer. The information can be of text, images and audio etc. if we consider a program it contain data that contain computer operations. The address and the data storage will be present in computer memory. Mainframe data are stored in relational database like SQL db, tables. Data can be organized in array, graphs and objects etc. data structure contain information like different types, numbers, string etc.

**Characteristics of Data**

As the sequence of instructions is in the form of data, set of instructions are used to do task called program. Binary machine code will be present when computer execute the program. Data can be shared among different types of users and applications. The collection of data and storage will result in unreliable so I will reduce accuracy. Collection data information are depends on completeness of data. Data will always be same and used properly and data maintain consistent in dataset which help to determine accuracy.

**Accuracy**

Data need to be sufficiently accurate for the supposed use and need to be captured most effectively once, although it may have a couple of makes use of. Data should be captured on the point of pastime. Data is usually captured at the point of pastime. Data used for a couple of functions, together with population and range of families, is input as soon as by the machine administrator.

**Validity**

Data is verified and stored for several application requirements, consisting of the precise software of different policies or definitions. It will make certain consistency among intervals and with similar establishments, measuring what's supposed to be calculated.

**Reliability**

As Data increases it should be stored correctly with different collection methods with correct time. Progress towards performance objectives should reflect real adjustments instead of versions in data collection procedures or techniques.

**Timeliness**

Data which is captured should not take more time to load and it must be more feasible to store and do not any extra Data should be available quick and frequently sufficient to aid information desires and to steer provider or control decisions.

**Relevance**

Data captured should know about for what purpose it is used.This will require a periodic evaluate of necessities to reflect changing wishes. It’s about having a responsibility to collect and file performance records in opposition to a wide variety of statutory signs.

**Completeness**

Data necessities must be simply special based totally at the records needs of the enterprise and facts series methods matched to these necessities. Checks may be made to make certain for completeness of information. An annual assessment of this is undertaken via Internal Audit.

**1.2 Big Data**

Big data is similar to data but in large in size, which contain both structured and unstructured data. To maintain large data traditional methods and software are difficult to maintain. To describe the big data 3V’s model refers to variety,volume and velocity. Variety refer to the size of data and velocity refer to speed of the data. Nowadays data is increased rapidly because if the increasing large number of mobile devices, aerial, camera etc. by increasing accuracy it is easy to make decision confidently and that wil result in efficiency and reduce the cost and risk.Big data contain the data which is more than the capacity of old software tools to capture, manage and process the data. Big data size is increasing fast like peta byte in range.

**Characteristics of Big Data**

Data means which is always used and stored forever, Consider any application that stored the data in a decade’s means which can be stored forever and when the data required it can be retrieved from the application. There are so many problems in data due to the increasing size of data exponentially and data organization.

**Volume**

Data is not just a text as data growth increases fast. Data can be of videos, music or pictures. It is used to store everywhere. By using large database the application and architecture are built to support data. Sometimes same data is used with multiple views and even through the original data is present the new data which is similar creates the explosion of data. The big volume shows similar like big data.

**Velocity**

The increase of data and social communication changed the way how people look at the data. There was belief of thinking yesterdays data as recent . The matter of newspaper is still following the same logic but news channels and radios have changed how fast so only receives the news. People can send something or can get replay with the fraction of second. In media communication peoples are not interested in few seconds old message, they will destroy old and read new messages.

**Variety**

Data can be stored in different formats like database,excel,csv,access ar etc. data can be stored in textfile. Data is not only just text it can be of broadcast, sms, pdf or something. It should be managed correctly to make it meaningful and usefull. It is easy to manage if we have data in same format,it is not possible in all the time. In existing world different formats of data challenges need to be overcome using big data.

**Applications of big data**

Big data which is used in so many applications and it has become more important in industry, academics ,security purpose, banking and etc.

**Banking and security**

Industries have some challenges like securities thief warning, thief detection, client data transformation, social analytics etc. security exchange use bigdata to manage financial market activity.

**Communications, media and Entertainment**

Media is most used in different formats for different device. Collecting ,utilizing and analyzing the customer data and details about mobile and social media. Industry understands the customer data and its behavioural data.

**Healthcare provider**

In hospitals huge number of data is stored so by using bigdata healthcare challenges are reduced. Free public health data are been used by some hospital. If the customer or a patient wants his health information he can get his health information by cell phone application.

**Education**

Big data is used quite effectively in education to store the information of student who want to learn, which subject and how are entred etc. and also used to know about the effectiveness and performance of the teachers.

**1.4 Data mining**

Data mining is used to find the actionable information from large sets of data it uses mathematics to derive patterns and trends in data. Data mining model can be constructed using this trends and patterns. Data mining model can be constructed using thistrends and patterns. Data mining is used to produce the extracted information from large data sets.

As there is huge quantity of information available in industry, these data are not useful until. It is converted into useful information. So it is necessary to know about large amount of data and extract useful information.

Data mining not only do extraction but also perform data cleaning, data intefration and data transformation. When all the process are performed it will be used in many application.

**Data mining Application**

Data mining used in various areas like market analysis, corporate analysis, fraud detection, also in production control, customer retention, sports and internet.

**Market Management**

In market technique data mining used for various purpose. Customer Profiling-data mining helps to detect who buys what kind of product. Identify customer need for different products. It will use assumptions to find different ways to attract customers. Cross market analysis is used to do association in productions and also help to find customers clusters where the attributes are shared as income, interest etc. customer purchasing pattern will be easy to discover. Provide various multidimensional report information.

**Corporate analysis and Risk Management**

Data mining provide cash flow analysis and prediction. Resource planning is used to compare and summarize resource and spending. monitor competitors and market directions.

**Fraud Detection**

In the fraud detectionof credit card and telecommunication etc., the data mining is used. For the phone call fraud detection the call destination, call duration and time etc are used. to know about what kind of data to be mined there are two categories like descriptive and classification prediction.

**Characteristics of Data Mining**

The large quantities data can be analyzed and volume of the data is great so it is easy to analyze by the automated technique. It is used to collect all the data irrespective of noisy and imprecise data and it collects the all type of characteristics of data. It is an complex data structure conventional statically analysis connection possible. It heterogeneous data can be solved in the legacy system. Different technique in data mining like next generation rules, classical techniques, trees and networks.

Classical techniques are mainly used for the aiding in the optimization of customer relation unique system. Statistics techniques are used mainly for business purpose. It is also used in mathematical analytics it is used in every daily life. Nearest neighbour is an oldest techniques, it works on the basics of historical value which one is nearest value to the centroid that one is classified as closet nearest neighbours.Clustering is meant by the group of data, it mainly used to mean segmentation which mainly used in market segment. It is also used in K-nearest neighbour algorithm. It consist of two different clustering algorithm namely agglomeration which is used for only one record. Non hierarchical clustering is used to compact the database. It consist of two types single pass method and reallocation method.

**1.5 How data linked in Data mining**

In general state data linked is collection of data and elements to reforming a new data. It defined the different model by using the mathematical methods. It also contains the different mathematical methods namely statistical it defines the non deterministic and logical about deterministic method. It will going to compressing the data using different algorithm. The basic goal is prediction and description.

**1.6 Techniques used for data mining analysis**

The data analysis techniques which are more frequently used are frequent itemset and association rule mining. These are used to find the frequently getting data pattern and relationship of different patterns in large databases.Frequent pattern means frequently occurring data pattern. The different type of patterns are frequent itemset, frequent subsequence and substructures. Frequent itemset is a itemsets that appear together in a transactional data set for example milk and bread, these are brought together in market by customer.

Association rule mining utilized to find the regularly occurring items, correlations, association and structures from data sets from relation db, transaction db or other forms of data source. Transaction db is of the transactions, where particular transaction is set of data item and individual transaction identification numbers.

A client who buy bed will also buy pillow is the main thing in association rule mining algorithm, such as apriori, fp-growth are designed for centralized db where the central site of mining will store the raw data. Problem was privacy concerns in the setting. Vaidya and Clifton are the used to recognize and give privacy issues in vertically or horizontally divided databases. Because of the increase of data the importance of data privacy, so many privacy preserving solutions are expressed. In this view, multi data owners who want to learn association rules or frequent itemset from multiple data and also data owner are not ready to send their sensitive data to central site because of privacy issues. If client has more than one column data in the joint db, it is called as vertically partitioned here are focus on vertically partitioned for example, different business, like costumiers and a high cost clock designer, sells variety of item to a society.

**Frequent itemset mining**

If the data should be mined means it should have some knowledge about the frequent itemset mining and association rule mining. The regularly happening patterns in the dataset are called the frequent pattern. The itemsets which occur commonly are called frequent itemsets.consider if a customer buying something like pencil and sequentially if he buys rubber means it’s called sequential patterns. For association and correlation with relationship of data frequent itemset play main rule. structured pattern means the frequently generated. Mainly in data mining research the frequent itemset generation is very necessary. To find out association correlation in transaction database the frequent itemset is used.

As the amount of data increased the many organizations are concerned in mining.The identification of relationship in transaction record help in business process. The transaction database is used in many business for catalogue design, cross marketing and behaviour analysis.By considering market analysis where the items are analysed by clients buying practice. First and most important thing to owner is that which item customer want to buy is performed by retailed data. Consider an example of world where so many stores are there and there are so many items in the shop and the items presence and absence is calculated by using Boolean numbers. For buying items the Boolean vectors are analyzed and represented in association rules. Consider if a client want to buy a cur also want to buy a bike then the rule of association can be:

Tooth paste=>brush [support=1%, confidence=50%]

The main rules are support and confidence which is needful to find the necessary revealed rules. The support of 1% means the study of transaction show the number of tooth paste and brush brought jointly. The value of confidence is 50% means the tooth paste brought ratio of customer is 50% means that much percentage of people also buy brush. By using both minimum support threshold andconfidence threshold the association rule mining can become more curious. Field experts or area user used to set the threshold. To find the association correlation among items the supplementary techniques are performed.

Consider an itemset i={i1,i2,…..,in} and task related data to be considered as the D which contain transaction database[T] with non empty itemsets. In every transaction database there is each transaction id. As an example X is an item which is present in transaction. inference of a association rule can be represented in the form X=>Y where X belong to itemset i and Y belong to itemset i and the value of X and Y are non empty set. X is association with Y with the relevant data D which contain support s. the s value is used to ind the ratio of transaction. The support and confidence is calculated by using formula shown below.

Support(X=>Y) = P (XUY)

Confidence(X=>Y) = P (Y|X)

The minimum support and threshold should be calculated is called strong.the supp and confidence values lies between 0% to 100%.the number of patterns called itemset. The set of tooth paste and brush is of 2itemsets. The transaction item number can be calculated by regularly obtaining pattern this can also be called as support count.

**Finding frequent itemset mining**

There are so many frequent itemset algorithm like apriori, Eclat and fp growth. **Apriori** is used to find frequent itemsets in Boolean association rule. It uses the level wise calculation to find the frequent itemsets. First frequent itemset1 is calculated by scanning the database count for each item with minimum support. The second itemset is calculated by using the itemset1 and the third one is calculated by using the second itemset2.it will be repeated until no frequent item set is found. It is founded by using full scan of database. The non empty regular itemset should be repeated. It has some properties like if the minimum support is not satisfied then ‘i’ is not frequent.

Consider an example of apriori algorithm where the transaction database is D and there are eight transaction in this database and |D|=8 the apriori algorithm is used to find itemsets in D.

* In the first step the each item is a part of itemset1 and itemset should be scaned to calculate the occurrence of the each item.
* First the minimum support should be known and by using that support the first frequent itemset is calculated as itemset1.
* To identify the second itemset this method make itemset1 is combinationed with the itemset2 which generates the second set of frequent itemset.
* Then the transaction database is scanned and support count is checked for each itemset in itemset2.
* Then the third itemset is calculated using above procedure only with minimum support count. it will calculate the itemset level wise and the give the frequent itemsets.

**Generating association rule mining from frequent itemset**

When the frequent itemset is created and database D is generated it is full straight to produce association rule mining by using minimum support and the confidence. Confidence can be calculated by using formula:

Confidence(X=>Y)=P(B|A)= sup\_count(XUY)/sup\_count(X)

The conditional probability is expressed in terms of support count of itemset sup\_count(XUY) is used to find the number of transaction in (XUY) and sup\_count(X) is number of transaction in itemset X. For association rule mining frequent itemset generate all the non empty subset. The minimum support is automatically calculated usinf minimum support.

**Motivation of the project**

Privacy preserving mining solution provide high level privacy compare with other solutions. This solution provide more privacy, as currently existing solution need partition of data owners. this will lead to the draining of confidential information. This can be avoided by privacy preserving solutions. This solution can avoid complication. The frequent itemset solution can work same as this solution and also association mining cannot be constructed by using frequent itemset mining solution. So by using both association and frequent itemset mining solution are designed. This solution is more significant because of the use of homomorphic encryption technique.To use of semi-trusted third partyallow to safely calculate supports and differentiate support efficiently.

**1.7 Main Components of Hadoop**

**Hadoop distributed file system[HDFS]**

Hadoop is an open foundation software which is built by the software foundation. There are so many algorithms used to store the large data and to do dispensation. The name of the hadoop can be comprises as hadoop distributed file system with handing out is called mapreduce. Hdfs is a cloud storeroom which provide high access to the data applications. It is mainly used to store large data it is of tera byte to peta byte. For large scale management the mapreduce methods are used.

The efficient data are used to write only one time and it can read any times in the itemsets. The streaming of data can be accessed very well in the built in system. The dataset is copied from source and then many analysis are performed in all time. More amount of analysis can be done to reduce the time and latency of the first record. It does not require any hardware of more cost. But the problem is if any node loss occur means there will be more loss in cluster.

**Map Reduce**

input

Key,value

Key,value

…

Key value

Key value

…

Key value

Key value

…

Key value

Key value

…

output

Key,value

Key,value

…

**=**

=

**Figure 1.1:Mapreduce flow**

Figure 1.1 shows for the distributed processing on computer cluster MapReduce software framework used foe large datasets. As shown in Figure 1.1, MapReduce consists of two part map and reduce phase so it is called MapReduce. Input and output are in the form of key in each part of mapreduce. Before sending to mapreduce the input value will be divided into key pair value. Every time mapreduce produce latest key pair value. The reduce method is processed for every different key value. For each distinct key the reduce part produce one key value pair. key value pair is the final output. MapReduce will process in input file manner.

To use the mapreduce the calculation that need to be composed is calculated. The calculation of subset is easy for the data and it can be used to combine the another subset data and to calculate. One form or any different form of aggregation method is done here compare to source starting data the destibation data is smaller. The mapreduce is used to calculate the input with no dependency with external input being processed.Our dataset size should be big enough that splitting it up for independent computations will not hurt overall performance.

It is also important to see in which situation the use of mapreduce is not much suitable mainly in subsequent situation. Mainly for online processing the map reduce is not suitable. It is more suitable for group dealing out.it is not suitable to give fresh and fast result. To do user request on the mapreduce it will be more expensive and it cannot be done easily. Recursive problem which cannot be done easily on mapreduce. The Fibonacci series cannot be applied on mapreduce because it applies on previous value before calculating present one. The computation should done independently.

Mainly hadoop consist of namenode, seconadrynode, datanode , jobtracker and task tracker. In big data the centralized node is used for namenode. When the request is given to the node it is retrived by namenode. There are two setups in the hadoop where one is singlenode and another is multinode setup. Where the all first requirements of hadoop will stored in same node and second one can have different nodes. The hdfs is not good for tiny files.

**Problem Statement**

Earlier centralized databases are used by the frequent itemset and association rule techniques by the apriori and fpgrowth algorithms to store the data in central sites. But the problem was there is no privacy.Due to increased importance of data mining many privacy solutions are proposed but the data owner who are having multiple data are not ready to send their data to central site due to the privacy concern. In this project by providing cloud aided service for frequent itemsets work same as the solution and cant build the association mining using frequent itemset solutions, so by using both association and frequent itemset mining solution are designed. This solution is more useful due to the use of “homomorphic encryption” method. By using semi trusted 3rd party also allow to compute support securely and differentiate efficiently.

**Objectives**

* The privacy for association mining solution is provided using cloud aided frequent itemsets for vertical divided solution.
* It helps to give security for association mining solution, where these methods used for the application and owners who have multiple data should have high privacy.
* Privacy preserving mining solution is most effective and efficient because of the homomorphic encryption method.
* By designing efficient homomorphic encryption can get secure outsourced computation of confidence and outsourced comparison privacy technique for comparing confidence and threshold.
* Ciphertext tag approach is used to avoid fictious data’s result on mining result.

**1.8 Scope of the Project**

The scope of project is used to provide privacy for frequent itemsets using homomorphic encryption. The Risk factors considered when data owner has multiple joint data. The cloud added privacy service using homomorphic encryption provide efficient privacy for the frequent itemset.

**1.9 Organization of the Report**

The report is begins with the table of content which contains list of figure and table then chapters included in the report contain screen shots and explanation of the project. The body of the report contains the introduction to Data, Big data, Data mining and Data mining analysis and then literature survey to know about different techniques used in the different papers and then the system requirement to collect the requirement and design and the system architecture is used to explain about the plans and requirement of the project. The implementation of the project include how the project developed and how development environment is used. Next chapter explains about the test cases used to evaluate the project which is followed by conclusion.The body of the report is divided into 6 chapters.

Chapter 1: Importance of the data, characteristics of data ,big data and their properties,data mining and problem statement are discussed in this chapter.

Chapter 2: Literature Survey.

Chapter 3:Give complete view of system requirement. hardware and software requirements explained here.

Chapter 4: Design constraints, system architecture, and data flow diagrams are discussed in this chapter.

Chapter 5: Implementation details, languages chosen, supporting libraries,coding guidelines, platform selected and database chosen is discussed in this chapter. Chapter 6: Explain about the testing stage of the project which contain test cases and their solutions.

Chapter 7: The experimental results analysis, evaluation metrics, performance analysis and inference from result are presented.

Chapter 8: Conclusion and future enhancements are discussed in this chapter.

**CHAPTER 2**

**LITERATURE SURVEY**

**Zhiqiang Yang and Rebecca N[1]** explained Bayesian network for providing security and privatness for vertically partitiond data.Bayesian network it is formed by acyclic graph with random variables and their situation. There had been many statistics mining strategies designed within the centralize version wherein all information is accrued and available in one relevant web sites. As the data increases the expansion of data and statistics can be stored in touchy records. the protection and mining result is calculated in privacy preserving method with finding association mining. If clients have private database and need to examine Bayesian community on the aggregate of their database without leaking their information to each different. The accuracy of the data with performance privateneess and security is provided using MSK. Regular overhead is computed using Bayesian network.Bayesian community used most effective acylic graph but can’t do on cyclic graph.

**Jaideep Vaidya and Chris Clifton[2]** describes privacy preserving association rule mining in vertically partitioned data. here mainly considering the privacy conservation.They mainly addressing the problem of data rule algorithm.Data is distributed over all the sites,each site consisting different types of attributes and then need to consider all the attributes.Two partition algorithm is used.It shows the individual transactions. clustering algorithm is used partitioning of the data. Vertically partitioned data is analyzed by using market based techniques.Market based technique consisting mainly two parts grocery purchase and clothing purchase like it shows the relationship between the grocery purchase and clothing purchase like using the ATM cards. Examples are tires,electronic products and car equipments,etc. It gives the better privacy setting. Large data can be transmitted.By using this method there are some problems like It leads to the homogeneous population, Duplication can occur, unable to detect cross-site correlation and value for a single data can split across other site it leads to confusion.

**Fosca Giannotti,Laks and Anna[3]** proposed privacy keeping mining of affiliation regulations from transactin databases.It specially describes the 1/3 character correlated data and how to get right of entry to that records using simple example, the enterprise proprietors owns the information from the outsourced however it isn't always green so it contact the 0.33 individual for the statistics or patterns and all of the statistics may be non-public statistics. It is going to keep the information in the one-of-a-kind servers for privacy reason. It might be accomplished by way of privacy maintaining framework to protecting the non-public information and then a number of the facts may be used. Mining algorithm can provide proper protection and scalability for private information.Encryption schema is used to discover the cipher information or item from correlated facts effortlessly. It works like including one-one substitution cipher for objects and including fake transaction to wake each cipher object share the same frequency as>=okay-1 others.In this technique cryptographic notation schema for studying the privateness splitting.Easily become aware of the cipher iteams from the precise styles.Affords higher safety. Some dangers like complex networking analysis.Tough for information, because it incorporates the greater AND/OR operations and it only consider the encryption gadgets.

**Mohmoud Hussein[4]** explained about security and preservation of association rule is provided to increase the performance using stegnography. Mainly the security is provided for the vertically divided databases with realization of affiliation regulation and to reduce Boolean computation. To avoid the multi divide problem the steganography change is made. To avoid the fixing problem of scalar product so many present techniques are used.In early techniques the main principle knowledge is to increase the overall performance with providing high privacy. When the large amount of data is considered the security will get more priority where the performance will be in decreased state so many new algorithms are proposed to to provide privacy fo huge datasets. They proposed rapid approach for the good performance with high privacy in large datasets for scalar products. The smaller matrix can be eliminated or hide by decreasing calculation time of sclar products. In this paintings they suggest a alteration of steganography based multiparty protocol for computing scalar product.This amendment gives suitable solution for tradeoff between the overall performance and privateness. The proposed amendment high-quality song theperformance to be quicker in case of very massive database, with suitable degree of discount in privateness. It will save you the invention of sensible records.

**Madhuri N. Kumbhar[5]** proposed privacy retaining for affiliation rule on horizontal and vertically partitioned database. The principal seek privacy retaining records mining is to keep individual web page facts. So many techniques are proposed for privacy on association rule mining partitioning is main constraints. The privateness constraints for vertically partitioned databases, algorithm based on cryptography strategies, Homomorphic encryption, Secure Scalar product and Shamir’s secret sharing technique are used. The RSA algorithm and homomorphic is combined for the horizontal divided database that uses Paillier cryptosystem to compute international helps are used.Data miner is used to provoke the procedure with the aid of sending help threshold and public key. Data miner also used in encryption and decryption manner for frequent object sets a good way to defend person websites information.This scheme isn't always relaxed to preserve privacy of master sites. Boolean association rule mining is designed for semi honest version.

**Fang liu [6]** addressed the affiliation rule which is used to discover the exciting and beneficial relation of the unrelated statistics in big database. As the data growth extended hastily and the applications which can be available like clinical records evaluation,actual time monitoring, huge scale financial collect massive amount of data from diverse gain like performance and control however a few trouble is in when statistics proprietor doubt confidentially for the sake of facts proprietors encrypted records might be uploaded so it's going to save you the server to use statistics illegally.Fang liu designed protocol for outsourced position mining[PORM] distinctive users or owners have set of objects that are vertically partitioned. PORM need to satisfies a few layout like produce accurate frequent itemsets and consumer need to produce correct guidelines. Consumer to server have to be secured and also person to person safety need to be there with no person have plaintext of any other consumer information. They have used protocol for outsourced rule mining. Outsourced version is used by the association rule mining to perform PORM,in which the statistics is encrypted and outsourced.Mainly PORM need to perform consumer to server security and person to person protection.PROM is best used for vertically partitioned database not aid horizontal database.

**Mohammed J Zaki [7]** explains about the affiliation mining undertaking that is used to understand about the unique attributes among huge database. Main undertaking in this paper is to find out the common association in large databases. Vertical tid-listing database is used to find the each itemset in a transaction and lattice-theoretic technique is used to decompose the lattice into sub-lattice. In which they don’t use foremost reminiscence. First the all of the common itemsets ought to be searched. Main venture in this paper is to discover the common itemsets here the brand new algorithm is used for efficient enumeration of frequent itemsets to search area into small, impartial subspace used to prefix-primarily based or maximal-clique based strategies. Tid listing is used for easy be a part of operations. Frequent sequences will increase because the tid list decreases.Complicated hash tree is not used. Poor locality is gift within the structure. Long itemsets are identified correctly and that they may be capable of avoid enumerating subsets. Efficient lattice traversal method are used to identify long frequent itemsets. The impact of the use of exceptional database layout schemes combined with decomposition and traversal techniques.

**Duc H Tan[8]** explained about building a real system is one of the major demanding situations of privacy retaining information mining. In this paper, recommend CRYPPAR is a novel which produce the complete fledged framework to provide security for association rule data using cryptogtaphy.The use of scalar protection fabricate and public key cryptography effectively on vertically divide data.It also additionally produce the partial topology to lower common unique cost as tons as feasible.The experimental effects show that the framework is green in privacy preserving association regulations and may come to be a popular framework for PPDM structures. There are two methods in the system one will be of coordinator and another is used for entire method.The second is the opposite events referred to as clients. Their obligation are exceptional.The main loops also are unique. The framework could be more effective in terms of running time if the implementation techniques such as pre-encrypted data values and caching results. To improve the CRYPPAR different efficient method used that will help for future enhancement.

**Yiqun HUANG[9]** explained about the privacy security in the dispension of essential record mining technique. To solve the privacy retaining method theSecure multiparty computation technique is more beneficial in mining. The data is more viable for vertically and scalar product tto discover frequent itemset for association mining. To evaluate the insecurity the scalar product first calculated. The two scalar product is used to increase the green protocol method. For high security and evaluation of the effectiveness this method keep integrity. It will provide low communication and computation cost for high security records. In association policies the frequent itemset generation is viable to find. For analyzing and providing high privacy several protocols are used. It is a good protocol with top individual protection and coffee price of verbal exchange and computation.

**Dragos¸ Trinca and Sanguthevar Rajasekaran[10]** describes Privateness preserving method is earlier used for research where mining is more importanat mainly because o the numerous programs. By considering the distinct rule mining technique it receive lagre data in different technique with affiliation technique and received massive interest, and mainly horizontal and vertically divided data has maximum methodology to get the different information correctly. This paper mainly explain about the vertically divided data and advise multi data protocol to calculate large itemset. In this paper they have considered algebraic method rules with recursive nature data and lately proposed multi party protocol for the identical problem. It is not good for similar rule and it need extra additional similar protocols are extra comfortable and additionally gift a variant of the protocol this is immune to collusion amongst parties.

**CHAPTER 3**

**SYSTEM REQUIREMENT AND SPECIFICTION**

The software requirement is one of the most important life cycle of software development and it will describe the full characteristics of the system which need to be developed. System requirement and specification is the software development basic, SRS is announced in IEEE 830 standard, which is used to validate the final produce.

SRS is used to describe the factors which explain about product and its requirement. Goals of used and tasks are achieved by SRS it also contain the details about the functional and non functional requirements. It helps to check the software requirement.

**3.1 Functional Requirement**

Functional requirement is used to describe function of software and how the system act when the specific inputs or condition. This is used to describe the various functionality like data manipulation and processing. The following functional requirements are

* The data owner first specifies the number of multiple data in the storage.
* Data owner send their data to do association rule mining to get frequent itemset.
* Substitution cipher encryption is done to change the character of the data.
* Data mining is done to manage the cloud added management with homomorphic encryption.
* Homomorphic encryption is used to do plaintext encryption with symmetric encryption.
* The result of homomorphic encryption will be decrypted in the other hand.
* The final decrypted frequent pattern will be given to the data owner.

**3.2 Non Functional Requirement**

Non functional requirement is used to do system operation than the particular behaviours. they may be used in the property like reliability, portability, response time. Non functional problem is due to the client need.

Portability: it has an good portability result. It can be adaptable for hardware and software technique. The hadoop software run on ubantu and it is more portable.

Correctness: the procedures are well followed and it will give good performance result and also the data will be of secured and correct.

**Ease of use**: user interaction is of easy way which provide effective result in less time.

**Robustness**: this software is used to develop the system with good performance and the user result will always correct and relevancy

**Modularity**: the flexibility of the product is important and very beneficial with the well formed modules.

**Reliability**: the tools used are of more reliable and it will not produce any conflict result. It is reliable in the means of security and more compatible.

**Security**: the data which will be stored in the cloud will be more secured by using two type of encryption. The third party clients can’t use the data because of the security.

**Hardware Requirement**

System : Pentium iv 2.4 Ghz

Hard disk : 50GB

Ram : 4GB

**Software Requirement**

OS : Hadoop,cloud[AWS],Linux [ubuntu]

Language : Python, Java

Software : Hypervisor [Virtual Box]

**CHAPTER 4**

**DESIGN FOR FREQUENT PATTERNS GENERTION AND PRIVCY PROTECTION**

In software development the most important part is the design. System architecture and data flow is mainly explained in high level design, which gives the plan or blueprint of the system.

**4.1 Design content**

In this chapter mainly discussed about the design plan and needed things to resolve the addressing problems. To start with the project design is the main thing to know about the problem which occurs in the future.

**4.1.1 Main design constraints**

* Consider a linux operating system of 64bit.
* Install hadoop in the linux operating system.
* The data owner who have multiple data they send their joint data to do mining to generate frequent pattern.
* The encryption technique like substitution cipher is used for that frequent pattern.
* The encrypted pattern is send to the cloud.
* The data pattern which is in the cloud should be provided with privacy using homomorphic encryption.
* The encrypted data is decrypted and then the association result will be given to data owner.

**The advantage of cloud added privacy model**

* Easy to store large amount of data using hadoop.
* Security is provided using two different encryptions.
* Symmetric homomorphic encryption is used.

**4.1.2 Architectural ideas**

This part explain about the methods used to design the techniques in the high level strategy and this strategy help to give the abstractions of key used in architecture.

**4.2 System Architecture**

Hadoop

Report

Generates

Encryption

Cloud

Encrypt

Decrypt

A

W

S

RSA advanced algorithm

Substitution algorithm

Cloud move engine

Reports.csv

Reports.txt

Reports.axl

Association rule mining & aprior algorithm

**Figure 4.1: System Architecture**

The Figure 4.1 shows the system architecture of the project where the data owner who have multiple data send their data to hadoop to generate frequent itemsets using association rule mining.the frequent itemset is generated by using the apriori algorithm. By using this algorithm the frequent itemsets are generated. For the security purpose the frequently generated items should be encrypted the encrypted data will be sended to cloud and then again for security purpose the data is encrypted using homomorphic encryption scheme.then the encrypted data should be decrypted and sent back to data owner.

**Existing System:**

Earlier the frequent itemset and association rule mining methods like apriori and fp growth are only designed for centralized database and the data stored in central datasets. Where the problem is due to privacy concern. Firstly in horizontal and vertical database the privacy issues was identified by vaidya and Clifton. As the amount of data growth increases the more number of privacy solutions are proposed. The data owners who have multiple data they want to learn about the association rule and frequent itemset mining, but data owners not ready to send their data to centralized database because of privacy matter. Horizontal partitioned database means the data is in the form of one or more rows and vertically partitioned database which as one or more column in joint database. These type of databases are used in market analysis and etc. the main task is to identify the privacy issue in vertically partitioned database to identify the address of the privacy.

**Disadvantages:**

* As the data growth increases the privacy issues was most considered.
* Data owners not ready to send their data to central sites.
* There were no solutions for vertically partitioned databases.
* The data is stored in centralized database form.

**Proposed system**

First the frequent itemsets are generated using the associationrule miningfor the data owners easy purpose, then the frequent itemset is secured using substitution encryption. The ecncypted data is uploaded to cloud for the mining purpose where the mining process with another encryption is done here.By using this encryption security for the cloud is increased and outsourced scheme is built securely.previously used homomorphic encryption is of asymmetric but here the more efficient symmetric method is used.this technique is used to help the addition , multiplication with comparision of algorithm of homomorphic encryption and decryption.

**4.3 Detailed Design**

Data owner have multiple data and they are not ready to send their data to centralized database because of privacy issue where the data is stored in central sites.Multiple data owners have multiple joint data. The multiple joint data has random itemset it should be in a frequent manner so it will be given to hadoop where the frequent items are generated.

**Figure 4.2: Detailed Design**

Figure 4.2 shows The frequently generated items should be secured so the encryption is done there using substitution encryption. Then the encrypted data should be given to cloud for mining. Mainly in cloud data should be encrypted using homomorphic method then the data will be more secured. When data owner want a data it will be decrypted and given to data owner. The Architecture of the system shows the requirements for the project.

**4.4 Flow Chart**

Read item in

Transaction

Remove item

Sup>=minsp

Frequent itemset

Substitution encryption

Encryption 2

Move to cloud

Check credential

Enter correct credential

Data moved

Decrypt data

**Figure 4.3: Flow Chart**

Figure 4.3 Shows The flowchart diagram explains about how the projectsgoes step by step. As it start the data from different owners is given to do association rule mining to get the itemsets which occurs frequently. The transaction database data is read and then the support is calculated to find frequent itemsets. If the support value is greater than minimum support then it will be added to generate frequent itemsets otherwise remove the frequent itemsets and using that frequent itemset need to generate the association rule result confidence value is calculated an it gives the frequent itemset result which is given for encryption.

The frequent itemset result is given for encryption which is of substitution encryption and then the encrypted data is given to do another encryption. The encrypted data is given to cloud for storage the credentials of aws cloud should be checked and send the data to cloud. Whenever the data is required data owner can decrypt the data.

**4.5 Data flow diagram**

**Level** **0**

The level0 data flow diagram is a general first stage flow diagram sometimes called as context level diagram. The main work of context level diagram is it interact with system and outside environment of the system. This level0 diagram shows the ho w the data flow is modelled.

Source

Destination

File upload data

**Figure 4.4 : interaction between source** **and destination.**

Figure 4.4 shows the level0 data flow diagram. In this data flow diagram there are two entities like source and destination. The complete security protocol is shown in the particular process. The source gives the file to upload and the owner given file will be given to hadoop tool for association rule mining to generate frequent itemset generation and the generated files are given for encryption and then the file stored in cloud for security.

**Data flow diagram Level** **1**

This level1 dataflow diagram is more efficient and give more data details. Figure4.5 shows the dataflow diagram level2 where it contain several steps like first the uploaded data is given to the hadoop for the generation of frequent itemset generation.

Association Rule mining with A-prior Algorithm

Hapood map reducer

Report.txt

Encryption substitute Algorithm

AWS

**Figure 4.5: Data encryption flow**

This data flow diagram shows overview of how the data set is given to do association rule mining in hadoop and the generated file will be given to do encryption and then it will be moved to cloud bucket.

**Data flow diagram Level** **2**

Preprocessing

phase

Generates

Association rule mining

Reports

Encryption 2

Encrypted report 1

Encryption 1

Encrypted report 2

Cloud move engine

cloud

G

AWS Bucket

This flow diagram shows the detailed procedure of how the data sets are given to hadoop to do preprocessing and it will generate the association mining result. The generated result is given to do two type of encryption and then the encrypted data is moved to cloud engine and stored in AWS cloud bucket.

**CHAPTER 5**

**IMPLEMENTATION**

The most significant phase in the project is implementation where it is main part in the output generation. The main part of implementation is of full fledged which work in the final software. the designed code should be in machine understandable way. This step is performed by code generator. There is no complication for the code generation when the design is correctly executed. For generating code compiler interpreter and debugger etc tools are used. For coding high level programme languages are used like c, c++, java, python. For each application the related language is used.

**5.1 Frequent Itemset generation using association rule mining**

The main task of data mining is to use the frequent itemsets essentially and to find the itemset in datasets like clusters sequences and association rules.the main aim for searching frequent itemset is to analyze the different itemsets in the market shops. An efficient confidence and minimal support will be more than help then it is just a common set. When functioning on frequencies of sets as opposed to their help, it use the relative minimal frequency threshold minimum support,with zero≤min\_suprel≤1. Obviously min\_supabs = [min\_support] can usually use absolutely the minimal assist threshold and pass over the subscript.

When the frequent items are detected from database, it is uncomplicated to organize the correct tough association rules from them, where strong association guidelines assure mutually minimal help and minimal self declaration. This may be completed the usage of the subsequent equation: self assurance(A) = support\_count(AUB) / support\_count(A) the probability of condition support is expressed in, where support\_count(AUB) is the variety of dealings contain the itemsets AUB and support\_count(A) is the number of transactions contain the itemset A. by using this formula frequent itemset generated using association rule, create all nonempty subsets of l. For every nonempty subset s of l, output the rule “s => (l-s)” if support\_count(l)/support\_count(s)>=min\_conf, where in min\_conf is the minimm self assurance threshold.

**5.2 Substitution encryption methodology**

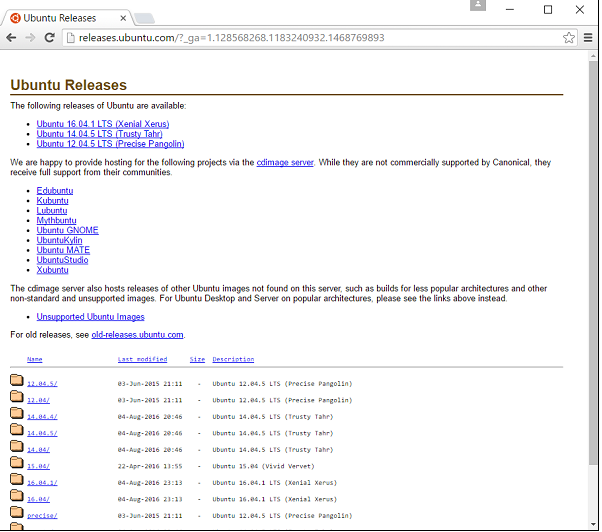
Substitution code is one the encryption method which is used to exchange the plaintext to ciphertext, invariable with unchangeable arrangement. The "devices" can be solo letters, double combination of letters, triplets of letters, combination of the above or, and so on. The decryption can be done by the reverse process of the replacement encryption. Substitution encryption may be in contrast with transposition ciphers. In a transposition cipher, the units of the plaintext are rearranged in a specific and generally quite complicated order, however the devices themselves are left unchanged. By evaluation, in a substitution , the gadgets of the plaintext are retained inside the same sequence in the ciphertext, however the gadgets themselves are changed

There are so many unique forms of substitution cipher. If the encryption is of the solo letter then I will be of easy substituition and if the data encryption with more letters called poly grphic. A mono alphabetic cipher uses unchanged substitution over the complete message, while a poly alphabetic cipher makes use of a number of substitutions at wonderful positions in the message, in which a unit from the plaintext is mapped to one among numerous opportunities in the ciphertext.

**5.3 Technologies used for project**

Linux based operating system is **ubuntu** and it is considered for network service and various applications. It is an open source system for software development. Various features of ubuntu like it support software lik firefox , chrom etc. and also support various libraries like libreoffice. To view and edit pics there are so many applications for free. To share and to manage there are so many applications. Normal user can frequently use this operating system and it will help to build the software easily. Many applications like email, multimedia and other are available in desktop. For various application like server and databases server version is used.

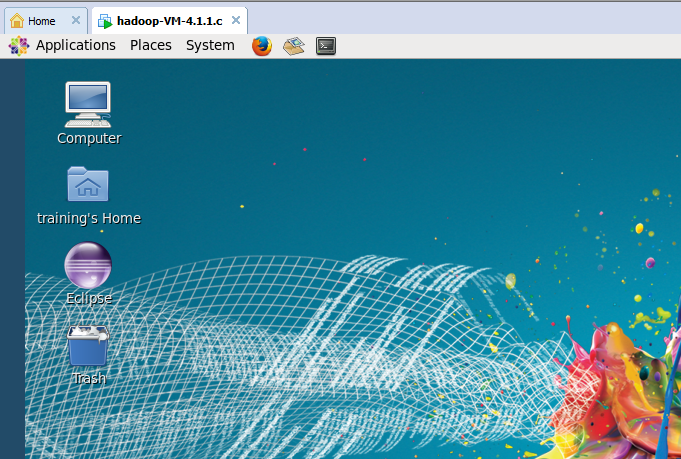
Figure 5.1 shows the Ubuntu has so many versions which is used for several application like wb application. Server can be downloaded to ubuntu and then it will be added to USB device. It is used by prompt to select language for installation.



**Figure 5.1 Downloading Ubuntu.**

**Hadoop** is used for huge network project not only for Mapreduce and HDFS which is more useful for big scale data analysis. HBase was the first module to provide online access and it uses HDFS for storage method. For the access of individual rows HBase provide online read or write contact. The YARN is used to establish processing form. When there is multiple work node and single master means it will be of small hadoop cluster. it mainly contain master node and slave node and master node contain job tracker,task tracker name node and data node. Slave node contain data node and task tracker it will have data and compute worker node. HDFS is a scalable and portable file system which will be written in java or python. Redundancy will be in the namenode which is present in hadoop cluster. Remote procedure call is used by clients for communication.

Large files are stored in HDFS as giga bytes and tera bytes form in large machines. By replicating the data among multiple hosts it will provide high reliability. Three nodes are used for data storage where the two data stored in one rack and another on different rack. The secondart namenode is present in HDFS it will lead to some problem when backup node is incorrect while primary namenode goes offline. As big data means large data it is quite difficult to build advanced server with high configuration which help to process large datasets. By using single CPU with multi computer large server can be constructed and it will lead in increase in throughput. Main aim of using hadoop is to run all the cluster using machines which is of low cost. The main task of hadoop is first partition the data into files and directories and the size will be of 128M to 64M. For some dealing out files are distributed over cluster nodes.

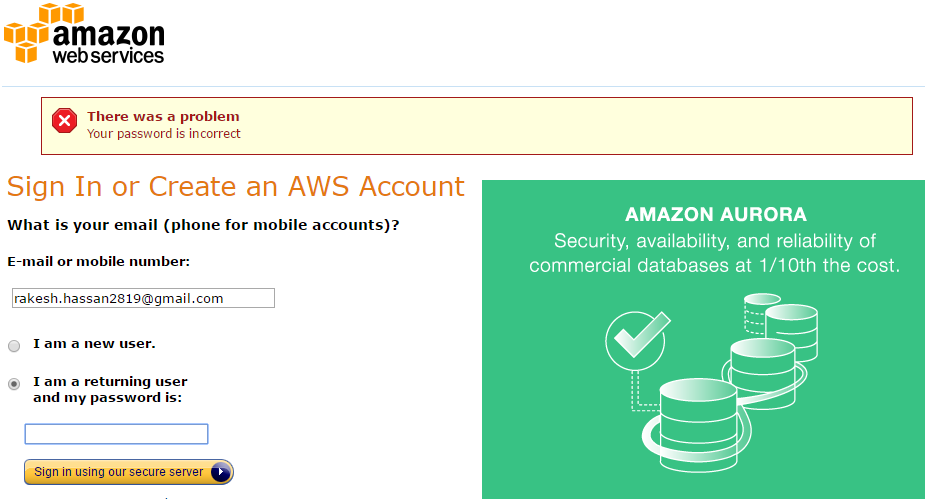


**Figure 5.2: hadoop on virtual machine**

The processing is supervised by the HDFS on local file. To manage the failure of hardware duplication of blocks are done.Check the thriving implementation of code. For each job debugging log is written. By using hadoop user can fastly write and check the correctness of distributed system. The cpu core parallelism is utilized by efficient use of distributed data. The fault tolerance and availability are not depended by hadoop and it is used to handle failure. Hadoop operate without any intermission. Main improvement of hadoop it is compatible in all platforms.

**Cloud** is a storage device which is used to store large data. It is an internet based system used to share the resource and data. Cloud computing is used to share resource and other devices. it is used to enable ubiquitous and on demand access to share the configured data. Cloud computing provide solution for user venture and computing is used to do coherence of similar utility. Amazon web service is a web service which is used to provide service for the storage. The aws virtual system contain so many real system like hardware which is of CPU and GPU, memory which is of local or ram, hard disk and any of the operating system. Mainly it provide data for storage. They operate in different fields across the world. There is different type like elastic, simple storage service by using AWS storage of important information and the database with many applications can be stored in the cloud with high growth rate. Cloud will have so many AWS products.

In virtual machine there will be some instances it work as the private server for virtual machine. The AWS service is provided by amazon.com. the most important thing is security. It will be provided by the same method as traditional method in data center. The security for any organization is efficient when the AWS is used to store the data in cloud first AWs cloud account should be created, then the virtual machine should be started to capture the files and media. The advantage of cloud like paying for the space used instead of booking large space it will lead to reduce the cost and consume less space. Amazon has higher economie than other services with less cost and good service.



**Figure 5.3 Starting AWS service**

It will avoid the guessing structure before starting the work guessing the capacity will lead to the problem. It will help to guess all the type of data from large to small. It will load the data very fastly and make resource available to who build the data easily. There are three different methods of cloud computing infrasture, platform and software as a service, AWS help to develop the flexibility of the business. In most of the time cloud is used for development and IT deparment and avoid the problem of maintaining, planning and not differentiated work. It provide different work such as management and flexibility.

Iaas is used to provide access for network. It is the important part of cloud and it will provide high security for management control with the similar content of the resources. It is used to provide the organization management and mainly focus on the deployment of the applications. Software as service is used to provide the service to run the product correctly it will be used to give refer to end user application.

**Python**

It is one of the most frequently used programming language and it will write the plain program on the small and huge scale. It is of both object oriented and structured program programming language.

**CHAPTER 6**

**SOFTWARE TESTING**

The main aim of system testing is to do different series of testing which is held in the computer based system. The system must be integrated correctly to enquire all the work and it use the different methods and it also perform correct allocation. This process technique is used to see the product work correctly and it do the same work as it required. The following steps to be followed in the testing process:

* The excellence of the project should confirm.
* Should to delete the errors of the previous steps.
* The original complexity should be evaluated by software solutions.
* The operations which are done on system should be reliable.

**6.1 Unit Testing**

The process of testing full one unit is call unit testing. It can cooperate with much functionality but combine it as one unit. Interpretation is another step of unit testing. Scaffolding is used for the individual test which is sometimes called supporting testing.

The black box testing is derived from the architecture and focus because the data interface is visible in test. Unit test are used to generate limits.

**6.1.1 Characters used for unit testing**

* Collect the data files from the data owners.
* Send the files to data hadoop for frequent item set generation.

**Table 6.1 frequent itemset generation**

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 1** |
| Test description | To test the frequent item set |
| Input | Transaction dataset |
| Expected output | By using association rule mining frequent item set generation |
| Actual output | Frequent pattern generated |
| Remarks | Pass |

Table 6.1 this test is to check if the frequent data is correctly generated by the association rule mining result. If the actual output is obtained then the test case is considered as pass.

**Table 6.2 Testing for substitution encryption**

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 2** |
| Test description | To test the generation of substitution encrypted data |
| Input | Frequent item sets |
| Expected output | Encrypted data |
| Actual output | The output as encrypted data |
| Remarks | pass |

Table 6.2 checks the frequent itemset generated items correctly encrypted. If the encryption correct means it will give correct encrypted data.

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 3** |
| Test description | To test the generation rsa algorithm result |
| Input | Substitution encryption data |
| Expected output | The rsa encrypted data |
| Actual output | The data which is encrypted by rsa algorithm |
| Remarks | pass |

**Table 6.3 Testing for RSA algorithm**

Table 6.3 checks the rsa algorithm for the encrypted data. The encrypted data will be from substitution encryption.

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 4** |
| Test description | Check if the file uploaded to cloud |
| Input | Encrypted data file |
| Expected output | File is secured in the cloud |
| Actual output | The secured and encrypted data in cloud |
| Remarks | pass |

**Table 6.4 To check if the file uploaded cloud**

Table 6.4 checks whether the encrypted data uploaded to the cloud. If the data uploaded correctly then test case is remarked as pass.

**Table 6.5 decrypted the data**

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 5** |
| Test description | To generate decryption for encrypted data |
| Input | Secured cloud data |
| Expected output | The decrypted data ready. |
| Actual output | The decrypted correct data is ready for data owner |
| Remarks | pass |

Table 6.5 Tests the data which is taken out from the cloud that will decrypted and given to data owner.

**6.2 System testing**

The main aim of system testing is to check the interface of each and every module of the system. The main thing is to find the input and output correlation between the modules and they are not different with others in a transaction module. the system is reliable and it will not execute the unexpected things in the system.

**Table 6.6 To Generate association rule result**

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 5** |
| Test description | To generate frequent itemset using hadoop |
| Input | Transaction database |
| Expected output | Frequently generated items |
| Actual output | The data items are frequently generated |
| Remarks | pass |

Table 6.6 is used to check the association rule mining result by which is generated by using the transaction data sets.

**Table 6.7 to decrypt data in cloud**

|  |  |
| --- | --- |
| **Test case ID** | **Unit test case 6** |
| Test description | Decrypted data items |
| Input | Encrypted data |
| Expected output | The decrypted data for data owners. |
| Actual output | Decrypted data for clients |
| Remarks | pass |

Table 6.7 shows the decrypted data from the cloud and send the result back to owner.

**6.3 Declaration of excellence**

Auditing and exposure are used for management of declaration of quality. It is used to handle the required that is required for the quality of the project, so it will increase the close by and self-reliance of the product. Some quality declarations are:

* Analyzing, designing, coding and testing methods.
* In each software engineering the technical view.
* Testing methods for multiple areas.
* Able to organize the software and can modify it easily.
* Measure and coverage techniques.

**6.3.1 Excellence Factor**

The significant aim is to get high quality using the software quality and also it is improved by the changes that are done in the procedure and methodology. The factors can be divided into two forms:

* Directly calculated factors.
* Indirectly calculated factors.

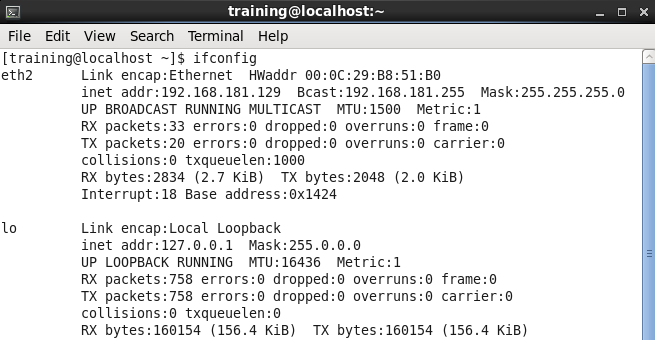
For every software products this forms have some important factors

* The features of operation.
* The capability of the changes it go through.
* Its ability to adjust to the new surroundings.
* The main aim is to get better performance efficiently.
* Time it is used to reach customer necessicities.

**CHAPTER 7**

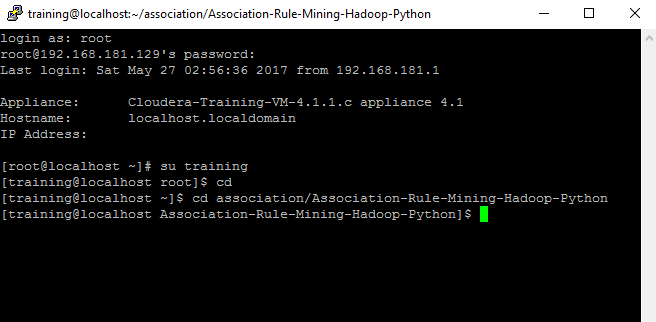
**EXPERIMENTAL ANALYSIS AND RESULTS**

By executing the step by step procedure for the project the following snapshots explains the output of the project and each and every screenshots explain about how the project is executing and the subsequent results.



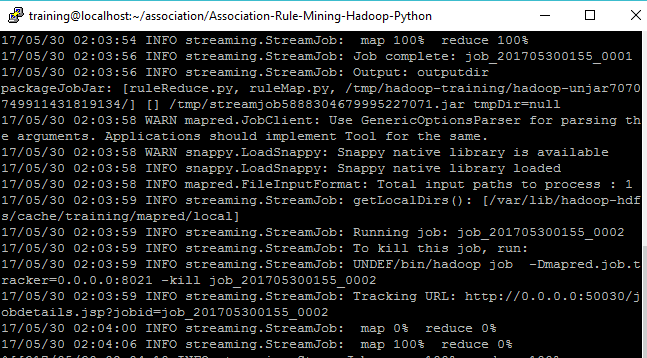
**Figure 7.1 hadoop tool configuration.**

Figure 7.1 shows the configuration of hadoop and the generation of ip address for the usage of putty and it will help to create session.

****

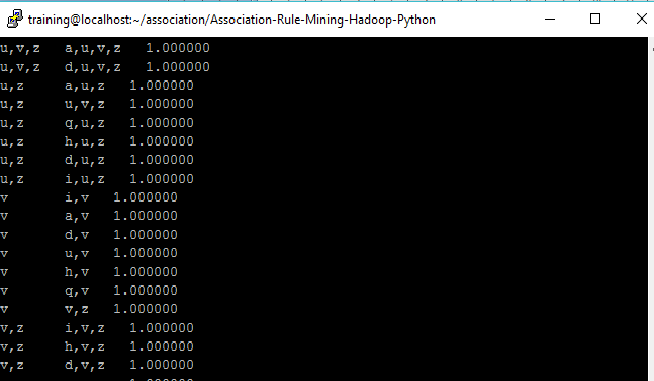
**Figure 7.2 putty session creation**

The putty session will be created and it will ask for user name and ask password and then the initialization of the hadoop with executing simple commands.

****

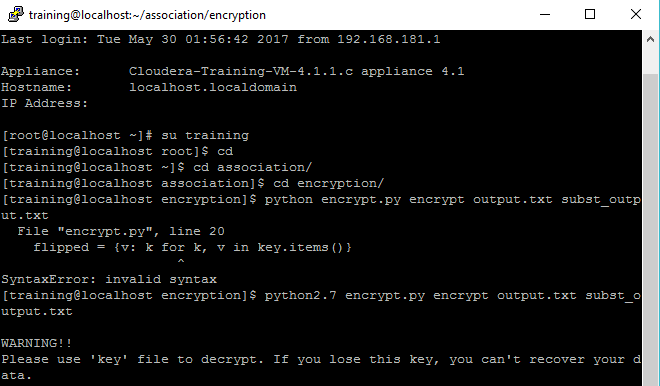
**Figure 7.3: Initialization of map reduce**

Figure 7.3 shows the execution of ruleGen.sh file which is used to delete the previously generated files and then generate the map reduce files which is of streaming information.



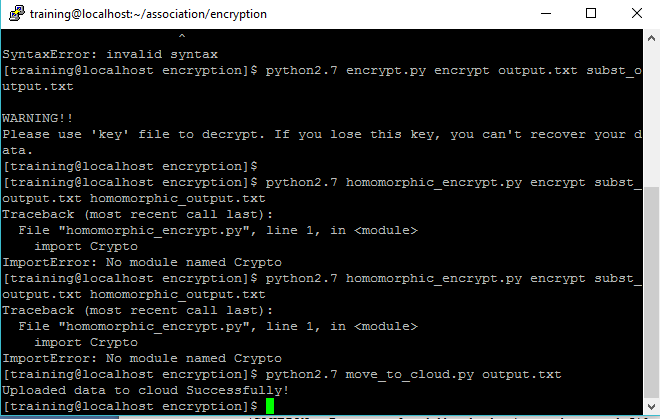
**Figure 7.4: frequent itemsets with association mining**

This figure 7.4 shows the itemsets in which are frequently generated using apriori algorithm and it will generate frequent items based on minimum support count.



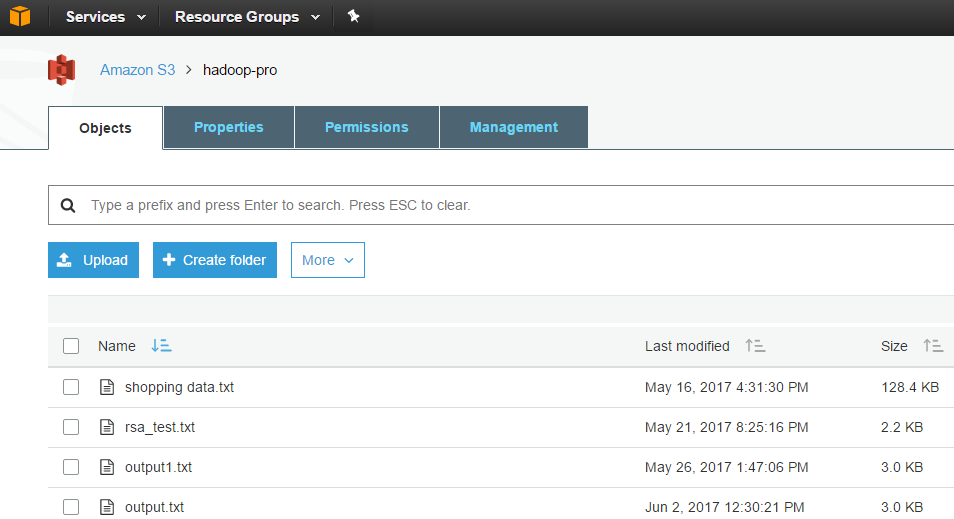
**Figure 7.5 : encrypt the file**

Figure 7.5 shows the substitution encryption for frequent itemsets and the key generation for the encrypted data. The key which is generated while encryption should be stored securely it will be used in decryption.



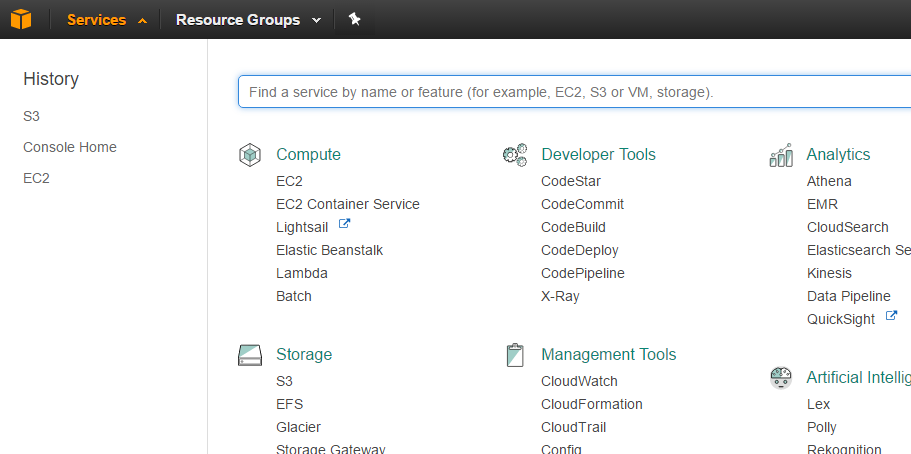
**Figure 7.6 upload the encrypted data**

Figure 7.6 shows how the encrypted data will be moved to cloud using move to cloud program.



**Figure 7.7: Cloud storage**

Figure 7.7 shows the encrypted data will be stored in AWS cloud . First AWS should create the bucket for the storage and then it will load the data to cloud.

****

**Figure 7.8 services provided by cloud**

Figure 7.8 shows the AWS service provided for the used to store, develop and to compute the data. It will provide S3 and EC2 services.

**CONCLUSION**

Mainly for security the privacy is provided for frequent itemsets which is of vertically partitioned. This lead to the data owners multiple data will be leaked this will lead to privacy issue. So to avoid that problem the privacy is given in cloud. By using cloud multiple data owner data is protected. These also provide mining result in cloud. When the raw data is compare with other data solution this solution provide less information and also this solution provide more efficiency so, this solution provide high privacy when the third party ready to access the raw data and it also avoid leakage. Mainly paper contain encryption technique which is used in cloud and then the cloud aided association comparison. These techniques have high potential for the applications.this type of solution is provided in this project.

**FUTURE ENHANCEMENT**

The Association rule mining is used to find vertically partitioned database using apriori, éclat methodology. By using this technology security is provided for multiple data owners and in future homomorphic encryption is used in cloud to provide high security and privacy. Homomorphic encryption is used to encrypt the data in plaintext and there will be symmetric and asymmetric encryption in this method.

**REFERENCES**

[1] Zhiqiang Yang, Rebecca N. Wright, “Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data*” IEEE Transactions on knowledge and data engineering,* vol 18, no. 9, September 2006.

[2] Jaideep Vaidya and Chris Clifton “Privacy preserving association rule mining in vertically partitioned data” *University west Lafayette,Indiana.*

[3] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale “Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases” *IEEE System journal, VOL. 7, NO. 3, September 2013.*

[4] Mahmoud Hussein, Ashraf El-Sisi and Nabil “Performance Tuning of Steganography Algorithm For Privacy Preserving Association Rule Mining in Heterogeneous Data Bas” ESRGroups France 2008.

[5] Madhuri N. .Kumbhar, Ms. Reena Kharat,” Privacy Preserving Mining of Association Rules on Horizontally and Vertically Partitioned Data: A Review Paper*” ieee* 2013.

[7] Mohammed J. Zaki “Scalable Algorithms for Association Mining” *IEEE Transactions on knowledge and data enginering, VOL. 12, NO. 3,2000*

[8] Duc H.Tran, Wee’Keong Ng,wei Zha “CRYPPAR: An Efﬁcient Framework for Privacy Preserving Association Rule Mining over Vertically Partitioned Data” *9th SIAM International Conference on Data Mining (SDM’09), 2009.*

[9] Yiqun HUANG, Zhengding LU, and Heping H “Privacy Preserving Association Rule Mining with Scalar Product” *HuazhongUniversity ofScience andTechnology Wuhan, Hubei 430074, P.R. China 2005.*

[10] Dragos¸ Trinc and Sanguthevar Rajasekaran “Towards a Collusion-Resistant Algebraic Multi-Part Privacy Preserving Association Rule Miningin Vertically Partitioned Data” *University of Connecticut Storrs, CT* 06269, 2000.

[11] Abdur Rahim Mohammad Forkan, Ibrahim Khalil “BDCaM: Big Data for Context-aware Monitoring - A Personalized Knowledge Discovery Framework for Assisted Healthcare” *IEEE Transaction on cloud computing, vol. x, no. x,* February 2015.

[12] M. J. Zaki, “Scalable algorithms for association mining,” *IEEE Trans. Knowl. Data Eng., vol. 1*2, no. 3, pp. 372–390, May/Jun. 2000.

[13] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in *Proc. ACM SIGMOD, pp*. 1–12, 2000.

[14] J.Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data,” in *Proc. SIGKDD, 2002.*

[15] M.Kantarcioglu and C. Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” *IEEE Trans.Knowl. Data Eng., vol.* 16, no. 9, pp. 1026–1037, Sep. 2004.

[16] Guoshuai Zhao, Xueming Qian “User-Service Rating Prediction by Exploring Social Users’ Rating Behaviors”*IEEE Transactions on multimedia, VOL. 18, NO. 3, march 2016.*

[17] Kevin Chiew “Analysis of Privacy-Preserving Mechanisms for Outsourcing Data Mining Tasks*” University* *of South Australia Adelaide, SA 5095, 2002.*

[18] Rachit V. Adhvaryu, Nikunj H. Domadiya “Privacy Preserving in Association Rule Mining On Horizontally Partitioned Database” *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014.*

[19] B. Raghuram, Jayadev Gyani “Privacy Preserving Associative Classification on Vertically Partitioned Databases”*IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) 2012.*

[20] Dennis E. H. Zhuang, Gary C. L. Li and Andrew K. C. Wong, Fellow “Discovery of Temporal Associations in Multivariate Time Series” *IEEE Transactions on Knowledge and Data Engineering 2005.*

[21] FIPS Publication 180-1: *Secure Hash Standard, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 1995.*

[22]FIPS Publication 180-2: Secure Hash Standard, Nat. Inst. Standards Technol.,  *Gaithersburg, MD, USA, 2002.*

[23]T. ElGamal, “A public key cryptosystem and a signature scheme based on discrete logarithms,*” IEEE Trans. Inf. Theory, vol. 31, no. 4, pp. 469–472, Jul. 1985.*

[24]N. Courtois, A. Klimov, J. Patarin, and A. Shamir, “Efﬁcient algorithms for solving overdeﬁned systems of multivariate polynomial equations*,” in Proc. EUROCRYPT, 2000.*

[25]P. Fournier-Viger. Real-life Datasets in SPMF Format, accessed on Apr. 6, 2*016.*

[26]P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. W. Wu, and V. S. Tseng, “SPMF: A Java open-source pattern mining library*,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 3389–3393, 2014*

[27]J.Vaidya and C. Clifton*, “Secure set intersection cardinality with application to association rule mining,” J. Comput. Secur., vol. 13, no. 4, pp. 593–622, 2005.*

[28]X. Ge, L. Yan, J. Zhu, and W. Shi, “Privacy-preserving distributed association rule mining based on the secret sharing technique*,” in Proc. SEDM, Jun. 2010.*

[29]R. Kharat, M. Kumbhar, and P. Bhamre, “Efﬁcient privacy preserving distributed association rule mining protocol based on random number,” *in Intelligent Computing, Networking, and Informatics. Raipur, Chhattisgarh, India: Springer, 2014.*

[30] C. Dong and L. Chen, “A fast secure dot product protocol with application to privacy preserving association rule mining,” *in Proc. 18th Paciﬁc-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD), Tainan, Taiwan, May 2014.*

[31] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, “Towards semantically secure outsourcing of association rule mining on categorical data,” *Inf. Sci., vol. 267, pp. 267–286, May 2014.*

[33] T. Fukasawa, J. Wang, T. Takata, and M. Miyazaki, “An effective distributed privacy-preserving data mining algorithm*,” in Proc. 5th Int. Conf. IDEAL, 2004.*