International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

Intelligent Agent Based Predict System for Enterprise Service Platform

Narasimman S, Jayavarman V, Parandhaman P, Vasanth V, Umavathi. V
Bachelor Of Technology In Computer Science And Business Systems Er. Perumal Manimekalai College Of Engineering Anna University Chennai 600 025

Abstract- Rising storage and computational capacities have led to the accumulation of voluminous datasets. These datasets contain insights that describe natural phenomena, usage patterns, trends, and other aspects of complex, real-world systems. We propose greedy K-NN (K-Nearest Neighbor) data allocation strategies (across the agents) that improve the probability of identifying data leakages. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject "realistic but fake" data records to further improve our chances of detecting leakage and identifying the guilty party. Mining large data requires intensive computing resources and data mining expertise, which might be inaccessible to most of the users. With the regularly obtainable cloud computing resources, data mining tasks cannot be stimulated to the cloud or outsourced to the third party to save cost. In this new pattern, data and model confidentiality becomes the major unease to the data owner. Data owners have to understand the possible trade-offs among client-side costs, model quality, and confidentiality to justify outsourcing solutions. In this paper, we propose the RASP Boost framework to address these problems in confidential cloud-based learning. The RASP-Boost approach works with our previous developed Random Space Data Perturbation (RASP) method to protect data confidentiality and uses the boosting framework to conquer the complexity of learning high-class classifiers as of RASP disconcerted data. So, we have to build upsome cloud-client combined boosting algorithms. These algorithms need low client-side calculation and communication expenses. The client does not call for to stay online in the progression of learning models. So, we have methodically studied the confidentiality of data, model, and learning process under a realistic security model.

Keywords- Cloud Computing, Data Confidentiality, Data Leakage Detection, K-Nearest Neighbor (K-NN), Data Allocation Strategies, Privacy-Preserving Data Mining, Random Space Perturbation (RASP), RASP-Boost, Secure Machine Learning, Outsourced Data Mining, Data Security, Boosting Algorithms, Cloud-Based Learning, Privacy Protection, Confidential Computing.

I. INTRODUCTION

OLAP

A winning company nowadays has several decisions to formulate. The enhanced those decisions are entire, the further unbeaten, and gainful, the company is. To many leading decisionmakers, the ability to analyse faster and better than the competition resources better decisions, developed profitability, and more success. Optimization of the relational database (RDB) has enabled many companies to ably accumulate the data about dealings, giving decision maker extra knowledge to utilize. However, there is a higher boundary to the quantity of data that one can have in an RDB and at a standstill achieve a resourceful study on. The On-Line Analytical Processing (OLAP) allow user to carry out rapid and effectual study on great amounts of data. The data are stored in a multi-dimensional method that more openly models genuine business data. And OLAP too allows users to contact summary data earlier and easier. They can then drill down into the summary records to get more detailed data, if need be. For a more detailed explanation of the OLAP rules, refer to Appendix A.

In this paper we describe the present state-run of the art of OLAP technology. A number of the issues discussed are data storage, OLAP architectures, and what kind of business construction will promote the most from a particular OLAP characteristics. Influential in the OLAP field are analysed, and a significant outline of the product and/or product line is

specified. Furthermore, issues connected to data mining and data-warehouses are addressed.

OLAP System Workings

An OLAP structure is integrated of numerous works. A top-level sight of the system includes a data source, an OLAP server, and a client. The data source is actually the source of data going to be analysed. Data from the basis are transferred or mocked keen on the OLAP server, anywhere it is prearranged and prepared to offer short query times. The client is the user interface to the OLAP server. In this segment, the role of each part and the implication in the complete system is described.

The source in an OLAP system is the server that stores the data to be analysed. Depending on the use of the OLAP product, the source could be a data warehouse, a legacy database housing company data, a collection of spread sheets that holds business data, or a grouping of any of the above. The capability of an OLAP product to work with data from a number of sources is very important. Requiring that all source data is stored in a particular format or in a confident database is problematic for database administrators. Ittoo reduces the authority and suppleness of the OLAP product. Administrators and users find that OLAP products that allow data mining from not only a wide selection of sources, but multiple sources, are more flexible and useful than those that have additional needs.

It is the back end of an OLAP system of the OLAP server. This is what does all of the work(depending on the model of the system), and where data that is dynamically accessed is stored. Dissimilar philosophies preside over the architecture of the server. In particular, a major feature of an OLAP product is whether the server uses a multi-dimensional database(MDDB) to stock up the data, or a relational database (RDB). This section labels pros and cons to all approach.

MOLAP

The MOLAP stand meant for Multidimensional On-Line Analytical Processing. This way that the server uses an MDDB to stock up data. Because most OLAP products are based on an MDDB, the term OLAP regularly refers to MOLAP as well .The purpose for using an MDDB is evenly open. It can ably store data that are by temperament multidimensional, provided so as to a resource of fast querying of the database. Data are transmitted from a data source (as described above) into the multidimensional database, and then the database is aggregated. This recalculation is actually what allows the OLAP queries to be quicker, since the computation of summary data is formerly done. The query time converts a function solely of the time required to access one piece of data, as opposed to the time to access many pieces of data and performing the computation. This approach also chains the philosophy of undertaking the work once, and by means of the results completed Multidimensional databases are a quite new technology. All The uses of MDDBs bear the alike drawbacks that mainly new technologies do. Namely, they are not as robust as RDBs and are not as enhanced to the same extent. An added disadvantage is that most multidimensional databases are incapable to be used while aggregating data, so it frequently takes point in time for new information to adapt on hand for study.

ROLAP

The ROLAP stands meant for Relational On-Line Analytical Processing. The era ROLAP specifies that the OLAP server is based on a relational database. The source data are entered into a relational database, usually in a star or snowflake schema, which services in fast retrieval times. The server provides a multidimensional model of the data, via enhance d SQL queries. There are a lot of reasons to get a relational database for storage as dissimilar to a multidimensional database. RDBs are a well-conventional technology that has taken sufficiently of opportunities for optimization. Factual world use has directed to a more robust product. Moreover, RDBs bear huge amounts of data than MDDBs do. They are intended for huge amounts of data. A main argument next to RDBs is that querying a huge database with SQL to obtain summary data typically resulted in multifaceted queries. An unskilled SQL computer operator could easily connection up valued system resources attempting to implement a query that is very simple in a MDDB.

Application OLAP

This is by future the largest area and is generally what is thought of or referred to by the term OLAP. Application OLAP frequently consists of a multidimensional database that is to be accessed by an exacting application, or maybe multiple applications. There is Vendors in this area which is mainly offer clients for the database. The client can just be a viewer, or it can be a healthy application that provides the user a lot.

ICEBERG QUERY

Data mining is the development of evaluating data from different perspective and shortening it keen on useful information - the information that can be used to boost income, cuts costs, or together. Data mining software is one of a number of analytical tools used for analyzing data. Data mining software allows the users to analyze data from many diverse magnitudes or approaches, group it, and summarize the relationships acknowledged. Precisely, data mining is the process of ruling relationships or else patterns between dozens of fields in huge relational databases. Data mining software analyzes associations and patterns in stored transaction data based on open-over user queries. More than a few types of analytical software are accessible statistical, machine learning, and neural networks.

Data mining brings allotment of reimbursement to businesses, society, governments, sales, marketing, insurance, health care, transportation and medicine and so on.

Market Segmentation: Recognize the general features of clients who buy the same products from your company.

Fraud Detection: Discover which transactions are most likely to be falsified.

Direct Marketing: Make out which scenario must be integrated in a transmitting listing to acquire the uppermost response frequency.

Banking/Finance: Used to identify client consistency by analyzing the data of customer purchasing activities. Interactive marketing compute what each individual accessing a Web site is most likely interested in seeing.

Data Mining Relationship

Data mining covers of some of four types of relationships are sought Classes Stored data is used to find data in prearranged groups. For example, a restaurant sequence could source client purchase data to define when clients visit and what they typically order. This information might be used to enlarge traffic by having daily specials. Clusters Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to spot market segments or consumer affinities.

Associations: The Data can be mined to classify associations. The beer-diaper example is an example of associative mining Sequential patterns Data is mined to anticipate behavior patterns and leanings. For example, an outside tools seller might guess the probability of a backpack being purchased based on consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of special levels of analysis that exist such as artificial neural networks, genetic algorithms, decision trees, nearest neighbor method, rule induction, data visualization. The cube is used to signify data along with some measures of interest. Although called a "cube", it can be 2-dimensional, 3-dimensional, or higher-dimensional. Each dimension represents a few features in the database and the cells in the data cube signify the calculate of interest. For example, they could hold a tally for the number of times that attribute grouping occurs in the database, or the smallest amount, highest, sum or average value of some attribute. Queries are performed on the cube to recover decision hold up information. Lately, introduced the CUBE operator for appropriately supporting multiple aggregates in OLAP database. CUBE operator is the n-dimensional generation of group-by operators. It computes group-by consistent to all possible combinations of a list of features.

SEQUENCE DATABASES

This provides a multiplicity of ways to query the data and bioinformatics analysis tools to help facilitate genetic study. The basic association of these databases has bent the way computer-based molecular biology research is conducted. This chapter will put up an understanding of series databases by reviewing data storage space, ordinary tools and online resources linking to these resources. If the laboratory is the groundwork of untried biology, nucleotide sequence databases are the foundation of genomic bioinformatics. These databases provide raw genetic data, nucleotide sequence, and a variety of resources to extract information from it. Simple questions relating to subjects such as presence or absence of homologous sequences, amount of genetic data available for an organism, and literature related to genes can be answered through the nucleotide sequence databases. First and foremost, these databases provide a complete resource for publicly available nucleotide sequence data.

International Nucleotide Sequence Database Consortium (INSDC), the united name of the trio, jointly obtains, processes, and publishes data for public use. Labs all-inclusive generate sequence data submitted to the INSDC as genome projects or as a precondition for publication. The association has set up a mechanism to split data professionally so that each database is

constantly underneath the same data set. The collaboration among the three allows each to develop their own interfaces and data curation strategies to represent their research initiatives while maintaining the original data reliability. The INSDC has also utilized their skill to form data standards by increasing their services further than nucleotides to include gene expression data Collaboration is a colossal task requiring enough forethought to allow for the exponential enlargement of sequence data while incessantly updating (sequences are shared each day) three diverse data centers. A great deal of the success the INSDC has skilled can be accredited to its capability to decide upon and stick to a common data arrangement. Knowledge of the terminology and layout of the sequence files is useful as it provides an understanding of the structure of genetic data encountered in many online databases and bioinformatics research applications.

Features The most informative area of the GBFF is the Features Table. As described in the introductory section, the features in the table are used by all of members of the INSDC. One or more qualifiers may accompany each of the feature elements, which allows for a further description of it. The feature is aligned to the left side of the document with the corresponding sequence located directly across; the qualifiers are listed directly below separated by a forward slash To assist with annotations, data contributors are asked to provide as much of the feature information as possible before submitting the entry keen on the database. Web Feat at EBI and the Sequin Help documentation at Gen Bank, both listed in Table 1, assist with the annotation process by outlining the features and qualifiers needed for a successful database entry. Three important Features of the Features Table are Source, Gene, and CDS (Coding Sequence). Source is the only required feature of the table and one of the few features with mandatory qualifier. Sources depict the natural source of the series start with the necessary life form qualifier. Some of the optional qualifiers include cell line, plasmid, gender, strain, tissue type etc. Gene is described as a range in the nucleotide sequence that.

has been identified as a gene, for which there is a corresponding name. The gene qualifiers include allele, map, product and partial. CDS describes an area of a sequence that has been translated from a nucleotide chain to a sequence of amino acids.

Probabilistic Algorithm

The applicability of these probabilistic methods to strong control is currently restricted by the information that the sample generation is possible only in very special cases which include systems exaggerated by real parametric indecision bounded in rectangles or spheres. Sampling in more general indecision sets is normally performed during over bounding, at the expense of an exponential rejection rate. In this paper, randomized algorithms for solidity and performance of linear time invariant uncertain systems described by a general - pattern are considered. In particular, efficient polynomial-time algorithms for indecision structures consisting of a random number of full multifaceted blocks and tentative parameters are developed. The most frequently used configuration for robustness analysis and design of complex uncertain feedback systems consists of a given plant, which possibly includes weighting functions and a controller, and an unsure block diagonal matrix. With this feedback configuration, different design methodologies can be implemented, and several performance objectives can be analyzed.

In addition, since the structure of is very general, various sources of uncertainty, such as parametric, nonparametric, structured and unstructured, can be easily in use into account. For these reasons, the –configuration is a valuable tool for both practitioners and theoreticians, so that powerful and useful results have been developed in the last few years. However, classical nastiest case robust control has also shown some limits when the control system is exaggerated by common ambiguity structures. To examine these limitations, many papers focused on difficulty issues of feedback system. The most often used configuration for toughness analysis and design of complex unsure feedback system.

This model consists of a given plant, which maybe includes weighting functions and a controller, and an unsure chunk diagonal matrix. With this feedback configuration, different design methodologies can be implemented, and several performance objectives can be analyzed. In addition, because the arrangement of is very general, various sources of insecurity, such as parametric, nonparametric, structured and unstructured, can be simply in use keen on account. For these reasons, the –configuration is a valuable tool for both practitioners and theoreticians, so that powerful and useful results have been developed in the last few years. However, classical nastiest case vigorous control has also shown some limitations while the control system is exaggerated by general indecision structures.

The authors are with the Department Automatic e Informatics, The contribution of these papers is to demonstrate that several problems in linear robust control are NP-hard, which in twist implies that they are not practically biddable, if not the number of qualms toward the inside into the feedback system is incredibly limited. To avoid this drawback, many other contributions attacked the same problem following a parallel line of research, with the goal of computing upper and lower bounds (instead of the "true" value) of the toughness edge for very general feedback configurations. In extra terms, the local point of these papers is to develop either necessary or sufficient conditions for robust stability and performance. The nice feature of these bounds is that their evaluation generally requires the solution of convex programs which can be easily performed, in this setting, is indeed bounded in a given set, but it is also a random matrix with a given probability distribution. In this way, both probabilistic and deterministic information is captured. In this paper, the class of so-called radically symmetric distributions in excess of the insecurity set is studied.

Organization Profile

Newmark Technology is a multi-faced organization formed with a vision to become a leader in various domains such as Software Training, Medical Transcription, Software Consulting and Project Development and Guidance. Having established in 2007, it is backed by a strong team to keep pace with the fast growing and latest technologies, Newmark Technology is committed to provide the best service. The company has diverse experience in giving on campus Training, Seminars and guiding students with their projects.

A major step-up in the growth of Newmark Technology was the launch of a Research Project training division for Research fellows on IT and Electronics Engineering. With an aim to employ qualified and adequately skilled personnel, the company has forayed into providing Project Training and Development to Engineering graduates across all disciplines. The quality of development and training provided at Newmak Technology enables the candidates successfully making a career in responsible industry standards. NewMak Technology Manpower Recruitment Industry provides End-to-End Recruitment Solutions for all varieties of Industries. NewMak Technology Manpower is backed up by a team of vibrant youngsters full of zest and driven by the urge to succeed and be the best.

The Success story of NewMak Technology Manpower relies on its qualitative approach and industries best practices. Our dedication and enthusiasm that has helped us achieve so much within a short span of time and have a strong clientele to our credit. The company has its head office at Coimbatore and its development centre at Erode, Trichy and Chennai. NewMak Technology is both technically and managerially very strong.

Objectives

In this work author presented a tactic to efficiently answer joint queries on both structured and text types of data. The records which are in data warehouses are usually extracted from other database systems and consequently contain only what is known as structured data. A huge quantity of text document is insufficient for processing ingeniously joint queries over structured and text data., a proposal for providing quick approximate answers to the iceberg query is devised with the purpose of helping the user refine the threshold before issuing the "final" iceberg query with the suitable threshold. That is, it tries to eradicate the need of a domain expert or histogram statistics to make a decision whether the query will actually return the preferred "tip" of the iceberg. This tactic for upcoming with the exact threshold is complementary to the efficient processing of iceberg queries.

This paper gives passing introduction about data mining used data mining and its events. This paper presents a complete survey on the existing most important information about the evaluation of iceberg queries, the need for iceberg queries and algorithm employed for evaluation of iceberg queries. The objectives of iceberg queries are calculated in this paper. This gives us the future direction to work on efficient evaluation of iceberg queries. The main objective of Iceberg queries is to retrieve data quickly. Query optimization is the cleansing process in database administration, and it helps carry down speed of carrying out. Data mining techniques are often measured by their speed. The reason behind this is that the faster the tool can run and the larger the data set can be applied. The Iceberg queries are usually very elite to calculate as they require more than a few scans of relationships.

Importance Of Data Analytics

In recent years, the rapid growth of digital technologies has resulted in the generation of massive amounts of data across various domains such as business, healthcare, finance, and social media. Organizations are increasingly relying on data-driven decision-making processes to improve efficiency and

gain a competitive advantage. However, traditional data processing systems are not capable of handling such large-scale and complex data efficiently.

To address these challenges, advanced data analysis techniques such as data mining and machine learning have been introduced. These techniques help in extracting useful patterns, trends, and insights from large datasets. In addition, cloud computing provides scalable storage and processing capabilities, enabling organizations to manage and analyse data more effectively.

The integration of intelligent agents, cloud computing, and machine learning algorithms has opened new possibilities for building efficient prediction systems. These systems not only improve accuracy but also reduce processing time and computational complexity. Furthermore, data security has become a major concern, especially when sensitive information is stored and processed in cloud environments.

In this project, an intelligent agent-based prediction system is proposed to enhance enterprise service platforms. The system utilizes advanced algorithms and secure data processing techniques to ensure efficient and reliable performance. The main objective is to provide accurate predictions while maintaining data confidentiality and system scalability.

II. LITERATURE SURVEY

"SEQUENCE QUERY PROCESSING"

In this work P. Seshadri, M. Livny, and R. Ramakrishnanet.al[1] has proposed Many applications require the capability to manipulate sequences of data, we motivate the importance of sequence query processing, and existing a framework intended for the optimization of series queries based on top of several different techniques. These include query transformations, optimizations that use meta–data and collecting of intermediate results. Many physical life applications use data that is inherently sequential. Such information is intelligently seen and questioned as far as a grouping deliberation and is often physically put away as a succession. Databases ought to enable arrangements to be questioned in a definitive method, using the arranged semantics of the information, and exploit the open doors accessible for query optimization. Social databases are lacking in this see information accumulations are dealt with as sets, not sequences. Subsequently, communicating arrangement questions is tedious, and assessing them is wasteful.

"The Design And Implementation Of A Sequence Database System"

In this work P. Seshadri, M. Livny, and R. Ramakrishna et.al [2] has proposed discussing the design and implementation of SEQ, a databases system it supports our sequence data. SEQm models a sequence as an ordered collection of records and supports declaratives sequence query language based on an algebra of query operators, thereby permitting algebraic query optimization and valuation. SEQ has been worked as a segment of the PREDATOR database framework that gives support for social and different sorts of complex information as well as a segment of the PREDATOR multi-threaded, client-server database framework which underpins arrangements, as well as relations and different sorts of complex information.

The framework utilizes the SHORE stockpiling executive library for low-level database usefulness like cushion administration, concurrency control and recuperation. An alternate plan worldview provides query preparing support for different information composers, including both successions and relations Much genuine data contains intelligent requesting connections between information things. Sequence data alludes to information that is requested because of such a relationship. Customary relational databases provide no abstraction of ordering in the data model, and do not support queries based on the logical sequentially in the data. Financial management products provide special purpose systems for analyzing stock market data. Current general-purposed databases systems provide limited support for sequence data.

"Models And Issues In Data Stream Systems"

In this work B. Babcock, et al[3] has proposed In this overview paper we motivate the need for and research issues arising from a new model of data processing In this model, information does not appear as determined relations, yet rather touches base in different, nonstop, quick, time shifting information streams. Notwithstanding auditing past work pertinent to information stream frameworks and current undertakings in the zone, the paper finds themes in stream question dialects, new necessities and difficulties in inquiry preparing, and algorithmic issues. Recently another class of information escalated applications has turned out to be broadly perceived: applications in which the information is displayed best not as constant relations but rather very as transient information streams. Examples of such

applications include financial applications, network monitoring, security, telecommunications data management, web applications, manufacturing, sensor systems, and others. In the information stream demonstrate, singular information things might be social tuples, e.g., arrange estimations, call records, site page visits, sensor readings, et cetera. In any case, their nonstop entry in different, fast, time-shifting, conceivably capricious and unbounded streams seem to restore some on very basic level new research issues.

"**Niagaracq: A Scalable Continuous Query System For Internet Databases"**

In this work J. Chen, et al [4] has proposed Our gathering strategy is recognized from past gathering improvement approaches in the accompanying ways. To start with, we utilize an incremental gathering advancement methodology with dynamic re-gathering. New inquiries are added to existing inquiry gatherings, without regrouping as of now introduced questions. Second, we utilize an inquiry split plan that requires negligible changes to a universally useful question motor.

Third, Niagara bunches both change-based and clock-based inquiries consistently. To guarantee that NiagaraCQ is versatile, we have likewise utilized different methods including incremental assessment of constant inquiries, utilization of both force and push models for recognizing heterogeneous information source changes, and memory collecting. This paper displays the plan of NiagaraCQ framework and gives some exploratory outcomes on the system's introduction and adaptability.

"TEMPORAL MANAGEMENT OF RFID DATA"

In this work F. Wang, et.al [5] has proposed RFID innovation can be utilized to fundamentally expand the proficiency of business forms by giving the ability of programmed recognizable proof and information catch. This innovation presents numerous new difficulties with current information administration frameworks. RFID information is time-subordinate, powerfully changing, in expansive volumes, and pass on certain semantics.

RFID information administration frameworks need to adequately bolster such huge scale transient information made by RFID applications. These frameworks need a clear transient information display for RFID information to help with following and checking queries. Through the programmed information accumulation, RFID innovation can succeed in more noteworthy deceivability and item speed transversely source chains, more effective record administration, less demanding item following and observing, lessened item forging and burglary, and much diminished work cost.

Then again, there is an abyss between the physical world and the translated world through sensor perceptions. These perceptions should be naturally translated and semantically changed into business rationale information, before they can be coordinated into business applications, such as ERP and WMS.

"Warehousing And Analyzing Massive Rfid Data Sets"

In this work H. Gonzalez, J. Han, X. Li, and D.Klabja et.al[6]has proposed Radio Frequency Identification (RFID) applications are set to play an essential role in object tracking and supply chain management systems. In the near future, it is expected that every major retailer will use RFID systems to track the movement of products from suppliers to warehouses, store backrooms and eventually to points of sale. The volume of information generated by such systems can be enormous as each individual item (a pallet, a case, or an SKU) will leave a trail of data as it moves through different locations. Radio Frequency Identification (RFID) is a technology that allows a sensor (RFID reader) to read, from a distance and without line of sight, a unique identifier that is provided(via a radio signal) by an "inexpensive" tag attached to an item. RFID offers a possible alternative to barcode identification systems and it facilitates applications like item tracking and inventory management in the supply chain.

"**Flowcube: Constructing Rfid Flowcubes For Multi-Dimensional Analysis Of Commodity Flows"**

In this work H. Gonzalez, J. Han, and X. Li et.al [7] has proposed The volume of data generated by a typical RFID application will be enormous as each item will generate a complete history of all the individual locations that it unavailable at every point in time, possibly from a septic production mark at a given factory, passing through multiple warehouses, and all the way to a particular checkout countering a store. The development ways of such RFID information frame enormous product chart speaking to the areas and spans of the way organizes navigated by both things. This item stream contains rich multi-dimensional data on the attributes, trends, changes and exceptions of product developments. In this paper, we propose a strategy to develop a warehouse of ware streams, called flow cube. As in standard OLAP, the model will be made

out of cuboids that total thing streams at a given deliberation level.

"Olap On Sequence Data"

In this work E. Lo, Baio, W.- S. Ho, C.- K. Chui, and D. Cheung,et.al[8] has proposed The main distinction of S-OLAP from customary OLAP is that an arrangement can be described not just by the attributes estimations of its establishing items, but likewise by the subsequence/substring designs it has. This paper thinks about numerous angles identified with Sequence OLAP. The ideas of arrangement cuboids and succession information 3D shape are presented. model S-OLAP framework is worked so as to approve the proposed concepts.
The model can bolster 'pattern-based' grouping and total, which is presently not upheld by any OLAP framework. The execution points of interest of the model systems well as exploratory outcomes are presented. Since applications frequently include both social information and succession information, the Devise system was proposed to demonstrate groupings arranged relations. By putting away succession information utilizing ordinary relations, it is substantially less demanding to inquiry a mix of social tables and information arrangements. This approach enables more integrated optimization and evaluation. SRQL is an extension of SQL. It is used in the Devise system for supporting queries on mixtures of sequences and relations. However, Devise did not address the issues of warehousing and efficient analysis of sequence data. Moreover, SQLite is not expressive enough to express queries with complicated patterns such as recurring patterns.

"Fast Evaluation Of Iceberg Pattern-Based Aggregate Queries"

In this work Z. He, P. Wong, B. Kao, E. Lo, and R. Chengeta[9] has proposed A Sequence OLAP (S-OLAP) system provides a platform on which pattern-based aggregate (PBA) queries on a sequence database are evaluated. In its simplest form, a PBA query consists of a pattern template T and an aggregate function F.A pattern templates a sequence of variables both is defined over a domain. For example, the template T = (X,Y ,Y ,X) consists of two variables Xand Y . Every factor is instantiated with every single conceivable incentive in its comparing area to determine totally conceivable examples of the template. Sequences are assembled in light of the examples they possess.

The reply to a PBA inquiry is an arrangement cuboid (s-cuboids), which is a multidimensional cluster of cells. Every cell is related with an example instantiated from the query is design format. The esteem of each s-cuboids cell is acquired by applying the total capacity Foto the arrangement of information groupings that have a place with that cell.

Since an example layout can involve numerous factors and can be subjectively long, the actuated s-cuboids for a PBA question can be uveitis shown that while CB is suitable for computing a full s-cuboids from score, the II method is more efficient if the system has already materialized and cached a significant number of inverted lists, or if only a portion of the s-cuboids cells have to be computed. As we have mentioned, our approach to answering an IPBA query is to classify iceberg cells and compute only them, the II method is more suitable. In this sector we describe the inverted table structure and explain how to compute s-cuboids using inverted indices.

"On Synopses For Distinct-Value Estimation In Multiset Operations "

In this work K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla rt.al [10] has proposed the undertaking of approximating the quantity of particular qualities (DVs) in a huge dataset emerges in a wide assortment of settings in PC science and somewhere else. We give DV gauge strategies that are planned or use inside an adaptable and versatile "synopsis warehouse "architecture. In this setting, inward data is divided into partitions, and a synopsis is created for each partition each synopsis can then be used to quickly estimate the number of DVs in its comparing parcel. By consolidating and broadening various outcomes in the writing, we get both proper summaries and novel DV estimators to use related to these abstracts.

Our synopses can be created in corresponding and can then be easily combined to harvest synopses and DV estimates for arbitrary unions, intersections or differences of partition. Our synopses can also grip rem ovals of individual partition character. We employ the premise of command statistics to demonstrate to facilitate our DV estimators are impartial, and to set awake instant formula and pointed mistake limits. Based on a different boundary proposal, we can exploit results due to Cohen in order to select synopsis sizes when initially designing the warehouse. Fundamental issue of the above piece vector information structures, when used as summaries in our stockroom setting, is that association is the only bolstered standard activity. One must, fall back on the incorporation/prohibition equation to deal with set crossing points.

"Srql: Sorted Relational Query Language"
In this work R. Ramakrishnan, D. Donjerkovic, A. Ranganathan, K. S. Beyer, and M. Krishnaprasad et.al[11] has proposed Systematic data, or sequences, can be found in a wide rangeof commercial, statistical, and scientific applications. Theseapplications need DBMS hold up to store up, control, and query sequence powerfully, and such hold up is absent in RDBMSs as the relational model provides put of tuples as its just data structure. The most generally utilized inquiry dialect for social frameworks is unable of answering some regular inquiries presented by business and scientific applications, for example, moving aggregates. Sequence information is experienced in a wide assortment of scientific and business applications, e.g., trial follows, process advancement, satellite perceptions after some time, standard market costs, and compensation narratives. There is also great interesting maintaining a history of a user's queries, or a log of the changes made to a database, and analysing such trace data to identify interesting patterns of usage. Given these trends, the ability to analyse large sequences is becoming gradually important and DBMS vendors are beginning toad such capabilities. The work detailed here presents an appealing distinctive to the two primary existing approaches, which depend on ADTs and EADTs.

"Optimization Of Sequence Queries In Database Systems"
In this work R. Sadri, C. Zaniolo, A. Zarkesh, and J. Adibi et.al[12] has proposed the need to scan for mind boggling and repeating designs in database arrangements is shared by numerous applications. In this paper, we talk about how to express and bolster productively successive example questions in databases.

Thus, we first present SQL-TS, an expansion of SQL, to express these examples, and after that we ponder how to upgrade search queries for this dialect. We take the ideal content search algorithm of Knuth, Morris and Pratt, and sum it up to handle complex inquiries on sequences. In this paper, we see arranged relations as successions as in SRQL yet propose another and all the more ground-breaking SQL-like dialect for design looking, and propelled methods for optimizing questions in such a dialect.

"Data Cube: A Relational Aggregation Operator Generalizinggroup-By, Cross-Tab, And Sub-Totals"
In this work J.In this work J. Gray, et al., et.al[13] has proposed Data analysis applications typically aggregate data across many dimensions looking for anomalies or unusual patterns. The SQL total capacities and the GROUP BY administrator deliver zero-dimensional or one dimensional aggregate. Applications require the N-dimensional generalization of these administrators. This paper defines that administrator, called the information 3D square or simply cube. The block administrator sums up the histogram, cross-classification, move up, penetrate down, and sub-add up to builds found in most report authors.

The curiosity is that cubes are relations. Thus, the 3D square administrator can be imbedded in more unpredictable non-procedural data analysis programs. The solid shape administrator treats every one of the total traits as a measurement of N-space We address every one of these two issues in this segment. The answer to the principal question is that social frameworks show N-dimensional information as a connection with N-attribute domains. For instance, 4-dimensional (4D) earth temperature information is commonly spoken to by a Weather table(Table 1).

The initial four segments speak to the four dimensions: latitude, longitude, height, and time. Additional columns speak to estimations at the 4D points such as temperature, weight, moistness, and wind speed.

"Computing Iceberg Queries Efficiently"
In this work M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. D.Ullman, et.al [14] has proposed We propose efficient algorithms to evaluate iceberg queries using very little memory and significantly fewer passes over data, when compared to current technique employ categorization or hashing. We present a novel case learn by in excess of three Gigabytes of Web data to show the investments obtain by our algorithms. A lot of data mining queries are basically iceberg queries. For example, market place analyst carry out market hamper queries on big data warehouses that a mass client sales dealings.

The simplest way to answer an iceberg query is to maintain an array of counters in main memory, one counter for each unique target fixed, so we can answer the query in a solitary go within excess of the data. However, as we have already indicated, answering the query in a single pass is not possible in our

applications, since relation is usually several times larger than the obtainable reminiscence.

"**Bottom-Up Computation Of Sparse And Iceberg Cubes ”**
In this work K. S. Beyer and R. Ramakrishnan et.al [15] has planned We presented a fresh algorithm (BUC) for Iceberg-CUBE computation. BUC builds the CUBE bottom-up; i.e., it builds the CUBE by starting from a group-by on a single attribute, then a group-by on a pair of attributes, then a group-by on three attributes, and so on. This is the conflicting of all techniques projected previous for computing the CUBE and has a significant sensible benefit: BUC avoids computing the well-built group-byes that don't convene smallest amount support. The pruning in BUC is like the pruning in the Apriority calculation for affiliation rules, aside from that BUC exchanges some pruning for area of reference and reduced memory prerequisites. BUC utilizes a similar pruning strategy when figuring inadequate, finish Cubes. When the result of the cardinalities for a gathering by is large in respect to the quantity of tuples that really show up in the outcome, we say the gathering by is scanty. At the point when the quantity of meagre gathering byes is expansive with respect to the number of aggregate numbers of gathering byes, we say the CUBE is inadequate.

“Fast Computation Of Sparse Datacubes,”
In this work K. A. Ross and D. Srivastava, et.al [16] has proposed We propose a novel algorithm for the fast computation of data cubes over sparse relations, and demonstrate the efficiency of our algorithm using synthetic, standard and real-world data sets. When the relation fits in memory, our technique performs multiple in-memory sorts and does not incur any I/O beyond the input of the relation and the output of the data cube itself. When the relation does not fit in memory, a dividend-conquer strategy divides the problem of computing the data cube into several simpler computations of sub-data cubes. Frequently, all except one of the sub-data cubes can be computed in memory and our in-memory key applies. In that case, the total I/O overhead is linear in the number of CUBE BY attributes.

“Olap Over Uncertain And Imprecise Data”
In this work D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, et.al[17] has proposed We identify three regular query properties and use them to shed light on alternative query semantics. While there is much work on representing and querying indefinite date, to our knowledge this is the first paper to handle both imprecision and uncertainty in an OLAP setting. Trust that a key commitment of this paper is our methodology identifying natural criteria, for example, consistency, authenticity, and relationship safeguarding and utilizing them to think about elective inquiry semantics is an approach that can be connected outside the OLAP setting (and indeed, faithfulness and connection conservation). In this paper, we expand the multidimensional OLAP data model to stand for data vagueness, especially vagueness and in safety, and learn likely semantics for aggregation queries in excess of such data. While there is much work on representing and querying ambiguous data, and even some work in the context of OLAP, to our knowledge this is the first paper to identify criteria that just be satisfied by any approach to handling data ambiguity in an OLAP setting, and to use these criteria in a principled method to arrive at appropriate semantics for queries.

“**Olap On Search Logs: An Infrastructure Supporting Data-Driven Applications In Search Engines”**
In this work B. Zhou, D. Jiang, J. Pei, and H. Li et.al[18] has proposed in this paper, by observing that many data-driven applications in search engines highly rely on online mining of search logs, we develop an OLAP system on seek logs which fills in as a foundation supporting different information driven applications. An experimental examination utilizing genuine information of more than two billion query sessions exhibits the handiness and attainability of our design. Many past investigations were directed on breaking down search logs for different application situations.

A few examinations mod-peered toward the navigate data in seek logs as certain pertinence criticism, which can be utilized to enhance positioning calculations. For example, for instance Search logs, which record users' search behaviour, contain rich and up-to-date information about users' needs and preferences. While look for engines get backing command from the Web, users implicitly vote for or against the retrieved information as well as the services using their clicks.

“**E-Cube: Multi-Dimensional Event Sequence Processing Using Concept And Pattern Hierarchies,”**
In this work M. Liu , et al[19] has proposed Many modern applications including tag based quantity transit systems, RFID-based supply chain management systems and online financial feeds require special purpose event stream processing technology to analyse vast amounts of sequential Ulti-

dimensional data available in real-time data feeds. Traditional online analytical processing (OLAP) systems are not designed for real-time pattern-based operations, while Complex Event Processing (CEP) systems are designed for series discovery and don't hold up OLAP operations. We will demonstrate novel E-Cube model that combines CEP and OLAP techniques for multi-dimensional event pattern analysis at different concept levels. A London transportation situation will be known to demonstrate the usefulness and appearance of this future knowledge.

"Multi-Dimensional Regression Analysis Of Time-Seriesata Streams"

In this work C. Yixin, et.al [20] has proposed In this paper, we examine strategies for on-line, multi-dimensional relapse investigation of time-arrangement stream information, with the accompanying commitments our examination demonstrates that just a small number of packed relapse measures rather than the total stream of information should be enrolled for multi-dimensional straight relapse examination, to encourage on-line stream information examination, a mostly materialized data 3D shape display, with relapse as measure, and a tilt time span as its time dimension, is proposed to minimize the amount of data to be retained in memory or stored on disks, and an exception-guided drilling approach is developed for on line, multi-dimensional special case based relapse examination. In light of this outline, calculations are proposed for old analysis of time-arrangement information streams. The information distribution centre and OLAP innovation depends on the joining and union of information in multi-dimensional space to encourage ground-breaking and quick on-line information examination.

"Partitioning Algorithms For The Computation Of Average Iceberg Queries"

In this work J. Bae and S. Lee et.al[21] has proposed Iceberg queries are to compute aggregate functions over an attribute (or set of attributes) to and aggregate values above some identified threshold. It's difficult to execute these queries because the number of unique data is greater than the number of counter buckets in memory. However, previous research has the limitation that average functions were out of consideration among aggregate functions. So, in order to compute average iceberg queries anciently we introduce the theorem to select candidates by means of dividing and propose POP algorithm based on it. Iceberg CUBE problem introduced in is to compute only those group-by dividers with an aggregate value above some minimum support threshold. The basic CUBE problem is to compute all of the aggregates as efficiently as possible. However, both the pruning technique proposed in Apriority and coarse counting principle can't be applied to average functions.

"**Efficient Iceberg Query Evaluation Using Compressed Bitmap Index"**

In this work B. He, H.-I. Hsiao, Z. Liu, Y. Huang, and Y. Chen et.al [22] has proposed Decision support and knowledge discovery systems often compute aggregate values of motivating attributes by processing huge amount of data in very large databases and/or warehouses. In particular, iceberg query is a special type of aggregation query that computes aggregate values above a user-provided threshold. Usually, only a small number of results will please the threshold constraint. Yet, the results often carry very important and valuable business insights. On account of the little outcome set, icy mass queries offer numerous open doors for significant question improvement. In any case, most existing icy mass question handling calculations don't take advantage of the little outcome set property and depend extraordinarily on the tuple-check based methodology. This causes escalated circle gets to and computation, bringing about long preparing time particularly when information estimate is vast.

Bitmap record, which assembles one bitmap vector foreach trait esteem, is picking up prominence in both segments arranged and push situated databases in late years. Answering ice sheet inquiries and processing icy mass cube have distinctive streamlining territories. The focal point of answering iceberg questions is to accelerate the handling time of single icy mass inquiry. The focal point of registering icy mass solid shapes, such that of is to expand the common calculation to shorten the block age time.

On The Computation Of Multidimensional Aggregates

In this work S. Agarwal, et al[23] has proposed at the core of all OLAP or multidimensional data examination applications is the capacity to at the same time total through numerous sets of dimensions. Registering multidimensional totals is an execution bottleneck for these applications. This paper exhibits quick calculations for registering a gathering of gathering byes, Methods of processing single gathering byes have been very much concentrated however little work has been done on improving a collection of related aggregates. gives some rules of scan to be used in a client implementation of the cube operator. These incorporate the littlest parent optimization and dividing of information by characteristic qualities, which we

adopt in our calculations. In any case, the essential application in is on precluding the semantics from securing the solid shape administrator. There are reports of on-going re-look identified with the information shape in bearings reciprocal to our own: presents calculations for choosing what assemble byes to pre-register and record.

"Hashed Samples: Selectivity Estimators For Set Similarity Selection Queries"

In this work M. Hadjieleftheriou, X. Yu, N. Koudas, and D. Srivastava, et.al[24] has proposed We study selectivity assessment techniques for . A wide variety of similarity measures for sets have been proposed in the past. In this work we focus on the class of abstract similitude and outline selectivity estimators in view of from the earlier built examples. To begin with, we examine the entanglements related with direct applications of arbitrary inspecting and contend that consideration needs to be taken in how the examples are developed; uniform random sampling yields low precision, while inquiry delicate genuine time sampling is more costly than correct arrangements (both in CPU and I/O cost). We demonstrate to assemble hearty samples priori, in view of existing summaries for particular value estimation. We demonstrate the precision of our system hypothetically and confirm its execution tentatively.

The literature survey plays an important role in understanding the existing research and technologies related to the proposed system. It provides a clear idea about the current methods, their advantages, and limitations, which helps in developing an improved and efficient solution. In this project, various research works related to data mining, cloud computing, secure data processing, and machine learning algorithms have been studied in detail to design an intelligent agent-based prediction system.

Several researchers have focused on data mining techniques to extract meaningful information from large datasets. Traditional data mining methods provide useful insights but often face challenges in handling high-dimensional data and ensuring data security. With the rapid growth of cloud computing, many systems have started outsourcing data storage and processing to cloud environments. While cloud platforms offer scalability and flexibility, they also introduce security risks such as data leakage and unauthorized access.

To overcome these issues, encryption-based approaches have been proposed in recent studies. Techniques such as Order Preserving Encryption (OPE) allow operations to be performed on encrypted data without decrypting it. However, OPE alone is not sufficient to provide strong security against various attacks. Therefore, advanced methods like Random Space Perturbation (RASP) have been introduced, which combine multiple techniques such as random projection, noise injection, and encryption to enhance data confidentiality. These approaches ensure that sensitive data remains protected while still allowing efficient query processing.

In addition to security techniques, machine learning algorithms have been widely used for prediction and classification tasks. Among these, the K-Nearest Neighbour (KNN) algorithm is one of the most commonly used methods due to its simplicity and effectiveness. It works by identifying the nearest data points based on similarity measures and using them to predict the output. However, traditional KNN algorithms may face performance issues when dealing with large datasets. To address this, optimized versions such as Greedy KNN have been developed, which improve efficiency and reduce computational cost.

Many existing systems attempt to combine data security and machine learning techniques, but they often face trade-offs between performance, accuracy, and confidentiality. Some systems provide high security but suffer from low efficiency, while others offer better performance but compromise data privacy. Therefore, there is a need for a balanced approach that ensures both security and efficiency.

The proposed system is designed by considering the limitations of existing methods. It integrates secure data processing techniques with efficient machine learning algorithms to provide accurate predictions while maintaining data confidentiality. By analysing various research works, it is evident that combining cloud computing, encryption techniques, and intelligent algorithms can significantly improve system performance and reliability.

Overall, the literature survey highlights the importance of secure and efficient data processing in modern applications. It provides a strong foundation for the proposed system and justifies the need for developing an advanced prediction system that addresses the challenges faced by existing approaches. This study also helps in identifying the best techniques and

methodologies that can be used to achieve the desired objectives of the project.

III. SYSTEM ANALYSIS

Existing System

Sequence Databases .Before PREDATOR, traditional database systems prepare not formally support sequence data. PREDATOR stores sequence data based on the objective relational model. The DEV is system supports sequence data processing using the relational model. To inquiry arrangement information, build up an expansion to SQL, called SQL-TS, to express numerous sorts of example based questions. These frameworks don't straightforwardly bolster OLAP tasks, or the preparing of example based total inquiries. OLAP.The idea of information 3D shape to encourage OLAP activities. Their work has been followed up in many subsequent papers. A few examples include iceberg cube, bottom-up cube computation, and top-down cube computation. These OLAP studies, however, only focus on relational data. In recent years, the OLAP technology has been extended to unconventional data. TRIDENT is designed to assimilate data incrementally as it arrives, allowing both streaming and in-place datasets to be managed. The system employs a network design based on distributed hash tables (DHTs) to ensure scalability as new nodes are added to its resource pool and uses a gossip protocol to keep nodes informed of the collective system state.

When it is efficient to do so, we also provide built-in preprocessing and modelling facilities. Specific contributions of TRIDENT include:

A fast, query-driven approach to feature space exploration and model construction, with a rich set of queries that support retrieval of training data.

Modelling guidance provided by automated dimensionality reduction, bias-variance decomposition, and ad hoc creation of pilot models.

Training data management for fast, targeted retrievals that can be exported to formats of different datasets.

However, these studies do not address pattern-based analysis on sequence data. It extends the work in with a focus on search-log specific OLAP operations. There are also studies on OLAP engines for stream data, in which the focus is on continuously updating aggregates based on a small time window in stream processing.

Iceberg query and iceberg cube: The computational issue addressed in their work is the difficulty of housing a large multidimensional array in memory for effective computation of cuboids (and thus iceberg cells). Two techniques, namely, sampling and coarse counting are devised to identify candidate iceberg cells Full sweeps of the circle occupant information are expected to wipe out false positives and false negatives in the appropriate response. Interestingly, our methodology is to figure the icy mass cells of a s-cuboid through measurable tests. As we have shown in the experiments, disk accesses are mostly avoided, resulting in very fast processing. Since, numerous works has been done on iceberg query and its variant. For example, studied iceberg query using average as the measure. The majority of these systems are tuple-examine based, which require something like one sweep of the connection, until the point that an ongoing work in, which use the bitmap lists for question optimization.

In the context of data warehousing, focused on computing iceberg cube, which computes and materializes cells of a data cube whose measures satisfy a specified threshold. These works focus on selecting a smart order of computing aggregation over all combination of aggregate attributes, in order to maximize sharing of the computation. Answering iceberg queries and computing iceberg cube have different optimization goals. The focus of the former is to shorten the query latency where the focus of the latter is to shorten the cube computation time.

CEP: Complex Event Processing (CEP) systems demonstrate sophisticated pattern-matching capabilities in processing stream data. Yet, the state-of-the-art CEP systems do not support higher-level OLAP operations, such as pattern-based aggregation yet.

Sampling: It also adopts the estimator in to estimate the selectivity for set similarity selection queries (e.g., given a document d, estimate the number of documents that are similar Tod). Different from, which only uses for count estimation, we use it to derive statistical tests in our context.

Problem Statement

In modern enterprise environments, large volumes of data are generated continuously from various sources such as transactions, user interactions, and operational processes. Managing and analysing this data efficiently is a major challenge. Traditional systems are not designed to handle high-

dimensional and large-scale datasets, resulting in slow processing and reduced performance.

Another significant issue is data security. When data is outsourced to cloud environments, there is a risk of unauthorized access and data breaches. Existing systems do not provide sufficient mechanisms to ensure data confidentiality during storage and processing.

Moreover, many existing prediction systems lack accuracy and scalability. They fail to deliver reliable results when dealing with complex and dynamic datasets. High computational cost and memory usage further limit their practical applications.

Therefore, there is a need for a robust and efficient system that can handle large datasets, provide accurate predictions, ensure data security, and operate efficiently in cloud environments.

Drawbacks Of Existing System

- Low aggregated classification result.
- Poor in computation time.
- High error rate with misclassified value may appear.
- Does not support multiple datasets with separate cloud agents.
- It cannot bind with multiple data aggregation with same time.

Proposed System Greedy K-Nn Algorithms

In the Greedy K-NN, the key is the algorithms that can learn base learners from the perturbed data. We categorize the algorithms into two categories: the pool based and the seed based. For each category, we will investigate two types of base classifiers: random decision stumps and random linear classifiers. The new Greedy K-nn algorithms that can improve the predictive accuracy of recommendations. However, the quality of recommendations can be evaluated along a number of dimensions of datasets.

It also relying on the accuracy of recommendations alone may not be enough to find the most relevant items for each dataset. In particular, the importance of diverse recommendations has been previously emphasized in several studies.

These studies argue that one of the goals of recommender systems is to provide a user with highly idiosyncratic or personalized items, and more diverse recommendations result in more opportunities for users to get recommended such items.

Cost Analysis

The cost in the whole learning procedure consists of three parts: the cost of cloud-side processing, the amount of data transferred to the cloud, and the price of client-side dispensation. Excluding the initial cost of preparing and uploading the perturbed data, we are more interested in the client-side costs of preparing the base classifiers and transferring them to the cloud.

Confidentiality Analysis

Confidentiality guarantee consists of several parts: the confidentiality of perturbed data in the cloud, the confidentiality of queries and the learning process, and the confidentiality of generated models.

Data Confidentiality: Data confidentiality has been discussed in our paper on the RASP approach for outsourced databases. We include the key points here to make the paper self-contained. According to the threat model, the attacker may know only the perturbed data, i.e., the first level of prior knowledge, or the distribution of each dimension, i.e., the second level, which correspond to the brute-force attack, and the ICA attack, respectively.

Brute-Force Attack: This attack will examine each possible original matrix X according to the known Y .We show that this process is computationally intractable. The goal is to show the number of the valid X dataset in terms of a known perturbed dataset Y . Beneath we talk about a streamlined adaptation that contains no OPE segment - the OPE rendition has in any event the same level of security. ICA Attack. With the known distributional information, the aggressor can accomplish more on evaluating the first information than simple.

The known most important strategy is called Independent Component Analysis (ICA) . For a multiplicative perturbation = AX, the fundamental method is to find an optimal projection, wY , where w is a d+2 dimension row vector, to result in a row vector with its value distribution close to that of one original attribute. This goal is approximately achieved by examining the non-gaussianity2 characteristics of the original distribution - finding the projections by maximizing the non-gauss inanity of the result wY . The non-gauss inanity of the original attributions is crucial because any projection of a multidimensional normal distribution is still a normal distribution, which leaves no clue for recovery.

Advantages Of Proposed System:

- Data confidentiality is provided by the RASP method and its combination. It is mostly used to protect the multidimensional array of queries in secure method and with efficient query processing.
- The range query is used in database for retrieving the stored data's It recovers the record from the database where go assigns some esteem between upper and bring down limit.
- The KNN inquiry signifies K-Nearest Neighbour question for total. K means positive whole number and this question is utilized to discover the k closest neighbour esteems. Better classification result.
- Less in memory usage with better computation time.

Feasibility Study

Preliminary investigation examines project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running systems. All systems are feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

- Economic Feasibility
- Operation Feasibility
- Technical Feasibility

Economic Feasibility

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In economic feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs.

The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, there is nominal expenditure and economic feasibility for certain.

Operational Feasibility

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project including the following:

- Is there sufficient support for the management from the users?
- Will the system be used and worked properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So, there is no question of resistance from the users that can undermine the possible application benefits.

The well-planned design would ensure the optimal utilization of the computer resources and would help with the improvement of performance status.

Technical Feasibility

The technical issues usually raised during the feasibility stage of the investigation include the following:

- Does the necessary technology exist to do what is suggested?
- Do the proposed equipment's have the technical capacity to hold the data required to use the new system?
- Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
- Can the system be upgraded if developed?
- Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to cater to the needs of 'Secure Infrastructure Implementation System'. The current system developed is technically feasible. It is a web-based user interface for audit workflow at DB2 Database. Thus, it provides easy access to the users. The database's purpose is to create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission for the users would be granted based on the rules specified.

Therefore, it provides a technical guarantee of accuracy, reliability and security. The software and hard requirements for the development of this project are not many and are already available in-house at NIC or are available as free as open source. The work for the project is done with the current

equipment and existing software technology. Necessary bandwidth exists for providing fast feedback to the users irrespective of the number of users using the system.

To understand the basic working mechanism of the developed algorithms, we start with the basic boosting framework and map it to the setting of cloud-client collaborative learning. Algorithm 1 shows the boosting procedure in our approach. We map each step to the cloud or the client as shown in the comment.

$I(Pi == ti)$ is the indicator function, whichreturns 1 if the condition $Pi == ti$ is true, otherwise returns0. This framework enables two key features: (1) the disturbed data allows model evaluation and comparisons to be completed independently by the cloud; (2) the testing procedure is also independently done by the cloud. These features maximize the use of cloud and eliminate the cloud-client interactions in the iterations.

Algorithm 1 Cloud-Client Boosting on Greedy K-NN Perturbed Data

1: N: the number of records,$\omega_{1,k}$: the weight for the record i in k-th iteration, n: the number
of iterations, $\mathcal{Y}_1$: the perturbed records in the cloud.
2:$\omega_{1,0} \leftarrow \frac{1}{N}, i = 1 \ldots N$; //by cloud
3: prepare and send a set of base classifiers $\mathcal{H}_0$ encoded and protected with perturbation
parameters; //by client
4: extend $\mathcal{H}_0$ to a large set $\mathcal{H}_1$ with some algorithms(optional)// by cloud
5: **for** k from 1 to n **do**
6: search $\mathcal{H}_1$ to find a base classifier $h_{k(y)}$that minimizes the weighted error with
weights $\{\omega_{1,k}, i = 1 \ldots . N\}$; //by cloud
7: apply $p_i = h_{k(y_i)}$to each record y_i and generate the prediction $\{p_i, i = 1 \ldots N\}$; // by
cloud
8: compute the weighted error rate $\mathcal{E}_k = \sum_{i=1}^{N} \omega_{1,k-1} I(p_i == t_i)$; //by cloud
9: $\propto_k \leftarrow \frac{1}{2ln\frac{1-\epsilon_k}{\epsilon_k}}$; //by cloud
10: $\omega_{1,k} \leftarrow \omega_{1,k-1} exp^{-t_1 a_k h_{k(x_1)}}$, and $Z = \sum_{i=1}^{N} \omega_{1,k}$; //by cloud
11: normalize $\omega_{1,k} \leftarrow \frac{\omega_{1,k}}{Z}$; //by cloud
12: **end for**
13: repeat the above procedure for different parameters settings, such as n;// by cloud
14: download $\{\propto_1, h_1(\ ), i = 1 \ldots n\}$; // by client

The key steps include (1) Step 3: The client prepares and sends a customary of encoded base classifiers, (2) the optional Step 4: the cloud extends the set with an algorithm, and (3) Step 6: the cloud works with the pool of encoded base classifiers to find the base classifier hk that work seasonably well on the weighted examples.

The original AdaBoost algorithm in each iteration will search for one classifier hk that minimizes the weighted error rate ε for N examples in a family of weak classifiers H.Specifically, it is defined as

$$h_k = arg\max_{h_j \in \mathcal{H}} \epsilon_j, \text{where } \epsilon_j, = \frac{1}{N}\sum_{1=1}^{N} \omega_{1,k} I(h_j(y_i)! = t_i).$$

The search space H is often limited, for example, the entire set of decision stumps for all dimensions.

However, in the RASP-Boosting framework the client needs to encode and transfer the set H to the cloud, which is prohibitively expensive for a large set such as the entire set of decision stumps. Instead, we let the client prepare a malleolus base classifier, and the cloud tries to find an acceptable one from the pool in each iteration. Theoretically, this method still works because the furthering framework requires only weak base classifiers. The major problem is how the candidates should be selected and how large the pool should be to avoid the situation that all candidates in the pool give $\approx$ 0.5 weightederror rate (for two-class problem) in certain iteration, whichwill significantly reduce the model quality.

In the following, we will present the key idea of encoding the base classifiers for RASP-perturbed data. Then, we improve several algorithms for the client to generate the pool of base classifiers and for the cloud to (optionally extend and) search the pool.

Pool-Based Algorithms

In this set of algorithms, the client generates a pool of randomly selected linear classifiers, based on only the dimensional distribution of the training data. The cloud will select one from the pool. We will discuss two methods for the client to generate the pool, and the method for the cloud to utilize the pool.

Random Decision Stump Pool (DS Pool).In this method, the client randomly selects a set of decision stumps to encode and transfer. The key problem is to select effective decision send to the cloud. Again, we will investigate how the pool size affects the learning result.

Cloud-side Processing. The cloud side will search the pool to find the best one that gives the lowest weighted error rate.

Depending on the data distribution and the randomly generated candidates, there is a small probability that all the base learners in the pool give weighted error rates ~ 50%. This probabilitywill exponentially decrease with the increasing pool size. Wewill study the lower bounds of pool size for different datasetsin experiments.

Seed-Based Algorithms

In the pool based algorithms, the client needs to generate a set of randomly selected base classifiers, encode them, and send to the cloud. It might be costly, e.g., with hundreds of encoded classifiers. In the following, we consider reducing the client's work further by using the seed-based algorithms. The client will send a few randomly selected "seed classifiers", and the cloud will generate a pool based on these seeds. The following algorithms depend on the linearity property of the query matrix Q. with two randomly picked seed decision stumps on the same dimension, we can derive all decision stumps on the same dimension. However, not all of these decision stumps are effective for our use. Again, we hope the result will shatter around the canter of the population.

Derived Random Linear Classifier (Derived). The linearity property of query matrices can be extended to general linear classifiers on the OPE space.

Cloud-side Processing. The cloud side will randomly generate batch derived decision stumps or linear classifiers to find the best one. Similar to the pool based algorithms, the problem is the appropriate number of random trials. We will investigate this problem in experiments.

IV. SYSTEM SPECIFICATION

System specification defines the technical requirements and configuration needed for the proper functioning of the proposed system. It includes both hardware and software requirements that are essential for implementing and running the application efficiently. A well-defined system specification ensures smooth development, deployment, and maintenance of the system.

The hardware requirements for this project are minimal, making the system cost-effective and easy to deploy. A computer system with a minimum of Intel Pentium processor, 2 GB RAM, and sufficient storage space is required to run the application. Higher configurations such as improved processors and additional memory can enhance system performance, especially when handling large datasets. The system should also support basic input and output devices such as keyboard, mouse, and display monitor for user interaction.

The software requirements include the operating system, programming language, and development tools used in the project. The system is developed using Java programming language, which provides platform independence and flexibility. Java Virtual Machine (JVM) enables the application to run on different operating systems without modification. The application can be executed on Windows or Linux-based systems.

For development and execution, tools such as Eclipse or NetBeans Integrated Development Environment (IDE) can be used. These tools provide a user-friendly interface for coding, debugging, and testing the application. The system also uses standard Java libraries and APIs for implementing networking, input/output operations, and graphical user interfaces.

The system follows a client-server architecture, where communication between client and server is established using TCP/IP protocol. Socket programming is used to enable data transmission over the network. This ensures reliable and efficient communication between different components of the system.

In addition, the system requires a dataset for processing and analysis. The dataset is stored and managed efficiently to support quick retrieval and processing. Proper memory management techniques are used to handle large volumes of data without affecting system performance.

Hardware Configuration:

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 1 GB DDR2 RAM.

Software Specification:

- Operating system : Windows 7.
- Coding Language : Java
- IDE : NetBeans 8.1

Overall, the system specification provides a clear understanding of the resources and environment required to implement the proposed system. It ensures that the system operates efficiently, securely, and reliably under different conditions. This specification also helps in future upgrades and scalability of the system.

In addition to the basic requirements, the system is designed to be scalable and adaptable to different environments. This means that the application can be executed on systems with higher configurations to achieve better performance and faster processing speed. As the size of the dataset increases, the system can efficiently handle the load by utilizing improved hardware resources such as higher RAM capacity and multi-core processors.

The network requirements also play an important role in the functioning of the system, especially in a client-server architecture. A stable network connection is required to ensure smooth communication between the client and server. This helps in reducing data transmission delays and improves overall system responsiveness. In cloud-based environments, high-speed internet connectivity further enhances system performance.

Security requirements are also considered as part of system specification. Since the system deals with sensitive data, it is important to ensure that proper security mechanisms are in place. Techniques such as data encryption, secure key management, and user authentication are used to protect data from unauthorized access. These security features ensure data confidentiality and integrity throughout the system.

The system is also designed with flexibility in mind, allowing easy modification and future enhancements. New features and modules can be added without affecting the existing system functionality. This makes the system maintainable and adaptable to changing requirements.

Error handling and system reliability are also important aspects of system specification. The system includes mechanisms to handle runtime errors, invalid inputs, and system failures. Proper validation checks are implemented to ensure that only valid data is processed. This improves system stability and reduces the chances of unexpected failures.

Furthermore, the system supports efficient memory and resource management. It ensures optimal utilization of available resources, which helps in maintaining performance even when processing large datasets. The system is also tested under different conditions to verify its reliability and robustness.

Overall, the extended system specification highlights the importance of performance, scalability, security, and reliability in the proposed system. It ensures that the system is capable of handling real-world applications efficiently while maintaining high standards of quality and performance.

V. SOFTWARE DESCRIPTION

front end

The software requirement specification is created at the end of the analysis task. The capacity and execution allotted to programming as a feature of framework building are created by propelling a total data report as practical portrayal, a portrayal of framework conduct, a sign of execution prerequisites and plan imperatives, fitting approval criteria. Highlights OF Java stage has two parts: The Java Virtual Machine (Java VM) The Java Application Programming Interface (Java API) The Java API is a vast accumulation of instant programming contraptions that give numerous valuable abilities, for example, graphical UI (GUI) gadgets.

The Java API is assembled into libraries (bundles) of related segments. The accompanying figure delineates a Java program, for example, an application or applet, that is running on the Java stage. As the figure appears, the Java API and Virtual Machine protect the Java program from equipment conditions.

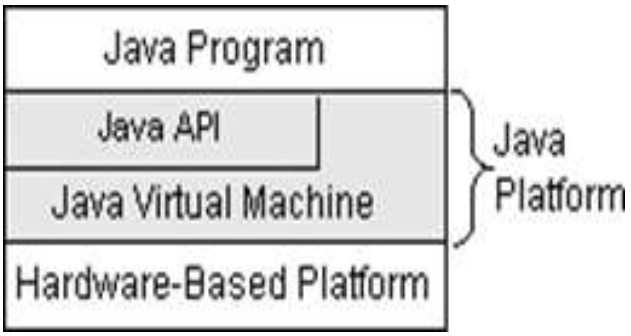

Figure 5.1 -. Java Platform Architecture

As a stage free condition, Java can be a bit slower than common code.

However, keen compilers, all around tuned mediators, and in the nick of time byte code compilers can convey Java's execution near that of local code without undermining movability. Attachment OVERVIEW: A system attachment is a great deal like an electrical attachment. Different fittings around the system have a standard method for conveying their payload. Anything that comprehends the standard convention can 'plug in' to the attachment and impart. Web convention (IP) is a low-level steering convention that breaks information into little bundles and sends them to a location through a system, which does not ensure to disperse said parcels to the goal. Transmission Control Protocol (TCP) is a more elevated amount convention that figures out how to dependably transmit information.

A third convention, User Datagram Protocol (UDP), sits beside TCP and can be utilized specifically to help quick, connectionless, variable transport of parcels.

Client/Server:
A server is anything that has some resource that can be shared. There are figure servers, which give processing power; print servers, which deal with an accumulation of printers; circle servers, which give arranged plate space; and web servers, which store site pages. A client is basically whatever other element that requirements to access a specific server.

A server process is said to "listen" to a port until a client connects to it. A server is allowed to accept multiple clients connected to the same port number, though each session is unique. To deal with numerous customer associations, a server procedure must be multithreaded or have some different methods for multiplexing the synchronous I/O. Held
SOCKETS: Once associated, a more elevated amount convention follows, which is subject to which port client are utilizing. TCP/IP holds the lower, 1,024 ports for particular conventions. Port number 21 is for FTP, 23 is for Telnet, 25 is for email, 79 is for finger, 80 is for HTTP, 119 is for Netnews- and the rundown goes on. It is up to each protocol to determine how a client should interact with the port.

JAVA AND THE NET: Java underpins TCP/IP both by covering the effectively settled stream I/O interface. Java underpins both the TCP and UDP convention families.

TCP is utilized for solid stream based I/O over the system. UDP underpins a less complex, subsequently quicker, point-to-point datagram-situated model.

Inetaddress:
The Inet Address class is utilized to epitomize both the numerical IP address and the area name for that location. Client associate with this class by utilizing the name of an IP have, which is more helpful and justifiable than its IP address. The InetAddress class shrouds the number inside. As of Java 2, variant 1.4, InetAddress can deal with both IPv4 and IPv6 addresses.

Industrial Facility Methods: The InetAddressclass has no obvious constructors. To make an InetAddressobject, client utilize one of the accessible manufacturing plant techniques. Factory methods are simply a tradition whereby static strategies in a class restore an occurrence of that class. This is done in lieu of over-burdening a constructor with different parameter records while having one of a kind strategy name makes the outcomes much clearer.

Three commonly used InetAddressfactory methods are:
1. Static InetAddressgetLocalHost ( ) throws UnknownHostException
2. Static Inet AddressgetByName (String hostName)throws Unknowns Hos tException
3. Static Inet Address [ ] getAll By Name (String host Name) throws Unknown Host Exception

The get Local Host ( ) method simply returns the Inet Address object that represents the local host. The get By Name ( ) method returns anInet Address for a host name passed to it. If these methods are unable to resolve the host name, they throw an Unknown Host Exception.

On the internet, it is common for a single name to be used to represent several machines. In the world of web servers, this is one way to provide some degree of scaling. The getAllByName ( ) factory method returns an array of InetAddresses that represent all of the add rises that a particular name resolves to. It will also throw an UnknownHostException if it can't resolve the name to at least one address. Java 2, version 1.4 also includes the factory method getByAddress ( ), which takes an IP address and returns anInetAddress object. Either an IPv4 or an IPv6 address can be used.

Instance Methods:
The InetAddress class likewise has a few different strategies, which can be utilized on the articles returned by the techniques just talked about. Here are probably the most regularly utilized. Boolean equivalents (Object other)-Returns genuine if this protest has a similar Internet address as other.

1. byte [ ] get Address ( )- Returns a byte array that represents the object's Internet address in network byte order.
2. String getHostAddress ( ) - Returns a string that speaks to the host address related with the InetAddress protest.
3. String get Hostname ( ) - Returns a string that speaks to the host's name related with the InetAddress question.
4. Boolean isMulticastAddress ( )- Returns genuine if this Internet address is a multicast address. Else, it returns false.
5. String toString ( ) - Returns a string that rundowns the host name and the IP address for accommodation.

Tcp/Ip Client Sockets: TCP/IP attachments are utilized to actualize dependable, bidirectional, steady, point-to-point and stream-construct associations between has with respect to the Internet. An attachment can be utilized to associate Java's I/O framework to different projects that may live either on the neighbourhood machine or on some other machine on the Internet. There are two sorts of TCP attachments in Java. One is for servers, and the other is for customers. The Server Socket class is intended to be a listener, which sits tight for customers to interface before doing anything. The Socket class is designed to connect to server sockets and initiate protocol exchanges.

The creation of a Socket object implicitly establishes a connection between the client and server. There are no methods or constructors that explicitly expose the details of establishing that connection. Here are two constructors used to make customer attachments: Socket (String hostname, I net port) - Creates an attachment interfacing the nearby host to the named host and port; can toss an Unknown Host Exception or an IO Exception.

Attachment (I net Address Ip Address, I net port) - Creates an attachment utilizing a pre-existing Inet Address protest and a port; can toss an IO Exception. An attachment can be analysed whenever for the location and port data related with it, by utilization of the accompanying strategies: Inet Address get InetAddress ( ) - Returns the InetAddress related with the Socket protest. I net get Port ( ) - Returns the remote port to which this Socket protest is associated. Integer Local Port ( ) - Returns the nearby port to which this Socket question is associated.

VI. PROJECT DESCRIPTION

Problem Definition
In this paper we calculated the problem of answering iceberg pattern-based aggregate (IPBA) queries. We put forward a synopsis-based solution, which samples and stores a small synopsis of a sequence database in main memory. We devised three statistical tests that process the synopsis to confidently classify a cell as iceberg or non-iceberg, and to confidently compute aggregate estimates of the iceberg cells. We proved two theoretical results related to the sample size properties of our synopsis-based approach. Our theorems state that the size of the synopsis needed to maintain a certain level of estimation accuracy and pruning effectiveness is independent of the database size.

These findings allow us to limit the synopsis size (and hence the memory requirement of the system) even when the sequence database grows indefinitely. The theorems also allow us to estimate a synopsis size with which a given accuracy requirement is guaranteed to be satisfied. From this observation, we devised two algorithms, namely, SBA ands+. While both are orders of level more efficient than previous approaches, SBA+ has the advantage of reducing I/O and CPU costs in cases where the accuracy test cannot be pleased by some iceberg cells.

Overview Of The Project
Multidimensional scaling (MDS) is an exploratory method used to pick out unrecognized dimensions affecting behaviour. Using MDS reduces huge amounts of statistics to notably simple, easy-to-visualize structures that reveal essential relationships in an economical manner and provides well-known answers to many problems in perception, emotion, and cognition, in which the stimuli are too complex to be quantified by way of other means.

To provide a basic advent to MDS for the no mathematician, in particular, the spatial distance model, which maps items as points in a multidimensional space.
Individual and aggregate analyses and individual differences scaling results are presented.

Module Description

- **Cloud Server Mapper Module:**

In this module the RASP denotes Random Space Perturbation. The RASP information irritation method is mix of OPE, arbitrary clamour expansion, and irregular projection, to give solid flexibility to assault on the bothered information and additionally questions. It likewise rations multidimensional exhibits, which permit introduced ordering strategies to be connected to accelerate extend inquiry handling.

OPE signify Order Preserving Encryption is a technique for encoding information so it's conceivable to make productive disparity correlations on the scrambled things without decoding them. Irregular projections are a great path for dimensionality decrease. Irregular projection is a procedure of anticipating novel high-dimensional information onto a lower dimensional information portrayal. Random noise injection is mostly used to adding noise to the input to obtain proper output when we compare it to the estimated power.

Task Management Module:

In this module the RASP strategy and its mix give secrecy of information, and this methodology is ordinarily used to ensure the multidimensional kind of questions in secure mode and furthermore with ordering and effective inquiry preparing will be finished. Grate has some imperative highlights. In RASP the utilization of lattice duplication does not safeguard the dimensional qualities so no compelling reason to experience from the dispersion based assault. Scratch does not protect the separations between procedures, so it keeps the information that are bothered from remove based assaults.

And furthermore, it won't ensure more troublesome structures it might be a framework and different parts. The range questions can be sent to the RASP annoyed information, and this range inquiry assign open limits in the multidimensional space. In Random space bother, the annoyance is utilized to do crumbling this procedure will occur as indicated by the key esteem that is determined by the proprietor. In this module the data owner has to register like owner and have to provide owner name as well as key value. And then the users have register and obtain the key value and data owner name from the owner to do access in the cloud. In this user can submit their query as variety query or KNN query and get their reply. We look at and illustrate the effect with encrypted and also in decrypted arrangement of the data for the query construct by the user.

Knn Query Management Module:

In this module KNN query denotes K-Nearest Neighbour query. This query is usually used to retrieve the nearest neighbour values of k. Here k is used to indicate optimistic integer value. KNN algorithm is mostly used for classification and regression. The use of KNN algorithm is to process the range query to KNN query. This algorithm consists of two methods. That is used to make interaction between the client and the server.

The client will send the query to the server with initial upper bound and lower bound. This upper bound range has to be more than the k points and the lower bound range have to be less than the k points.

Process And Result Estimation Module:

In this module the cloud service provider stores the data in the cloud database by changing the data to the troubled form to retrieve aggregate data result. It is a supplier of web facilitating administrations. The customer creates the underlying reach and sends its protected annoyed frame to the server.

1. The server chips away at the protected range questions and finds the inside range covering at any rate k focuses.
2. The customer interprets the safe bothered inside range from the server and extends it to the external range, which is sent back to the server.
3. The server restores the focuses in the external range.
4. The customer unscrambles the focuses and passages the k closest focuses. Abusing User Data Like programmer's cloud specialist organizations need to change access to their clients' information also, essentially as a result of accomplishing benefit. At that point they charge these organizations a little expense for demonstrating the client a notice on the web interface. Be that as it may, they are not just filtering for basically watchwords, likewise more modern techniques are utilized to acquire a wide range of measurements from a client's messages, records and so forth.

This is portrayed under the term information mining, which implies extraction, examination and utilization of information in a way that it was not initially put away for. The suppliers can interface a few various types of client information together to get exceptionally exact client profiles which would then be able to be utilized to do conduct expectation.

System Architecture

The system architecture represents the overall structure and working flow of the proposed intelligent agent-based prediction system. It illustrates how different components interact with each other to process user requests, ensure data security, and generate accurate prediction results. The architecture is designed in a modular and scalable manner to support efficient data processing and secure communication.

The system mainly consists of four major components: User Interface, Intelligent Agent Modules, Cloud Server, and Database. These components work together to perform various operations such as data input, processing, storage, and result generation.

The user interface acts as the entry point of the system where users can perform operations such as registration, login, data upload, and query execution. It provides a simple and user-friendly environment for interacting with the system. Users submit their queries and input data through this interface, which is then forwarded to the processing modules.

The intelligent agent plays a key role in managing and processing user requests. It consists of multiple sub-modules such as task management, KNN query processing, result processing, and security management. The task management module controls the workflow and ensures that each operation is executed in the correct sequence. The KNN query processing module applies the K-Nearest Neighbour algorithm to analyze the data and generate predictions based on similarity measures. The result processing module prepares the final output and ensures that it is presented in a meaningful format. The security management module handles encryption and ensures that sensitive data is protected during transmission and storage.

To ensure data confidentiality, the system uses an encryption module based on Random Space Perturbation (RASP). Before sending the data to the cloud server, it is converted into an encrypted format. This prevents unauthorized access and protects the data from potential attacks. The encrypted data is then transmitted securely to the cloud server for processing.

The cloud server acts as the central processing unit of the system. It is responsible for storing encrypted data, executing queries, and performing computations. The server includes components such as encrypted data storage, data processing engine, and database management system. It processes the incoming queries without exposing the original data, ensuring both security and efficiency.

The database stores all necessary information such as user details, datasets, encrypted data, and generated results. It ensures efficient data retrieval and management. The database is designed to handle large volumes of data and supports quick access to stored information.

After processing the query, the results are sent back to the user through the decryption module. This module converts the encrypted results into readable format. The final output is then displayed to the user through the user interface.

The architecture also ensures smooth data flow between components using secure communication protocols. It supports scalability, allowing the system to handle increasing data and user requests. The modular design makes it easy to update or modify individual components without affecting the entire system.

Overall, the system architecture provides a clear representation of how the proposed system operates. It ensures efficient data processing, strong security, and reliable performance. This architecture plays a vital role in achieving the objectives of the project and enables the system to be used effectively in real-world applications.

Figure 6.1 System Architecture

SYSTEM FLOW DIAGRAM

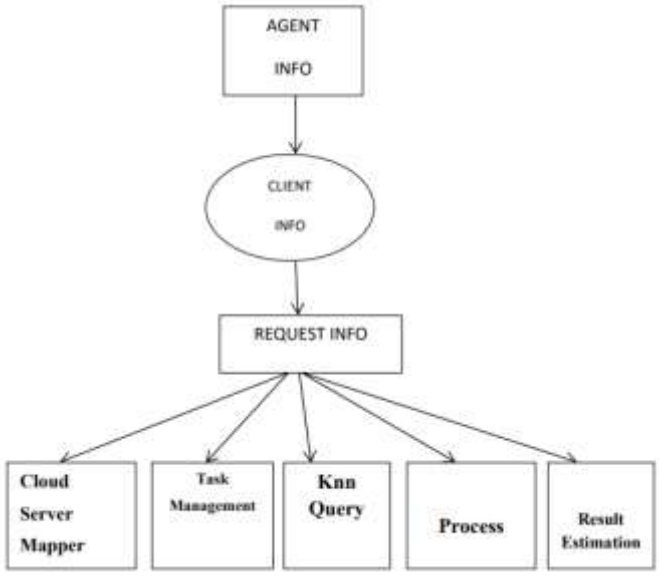

Figure 6.2 System Flow Diagram

DATA FLOW DIAGRAM

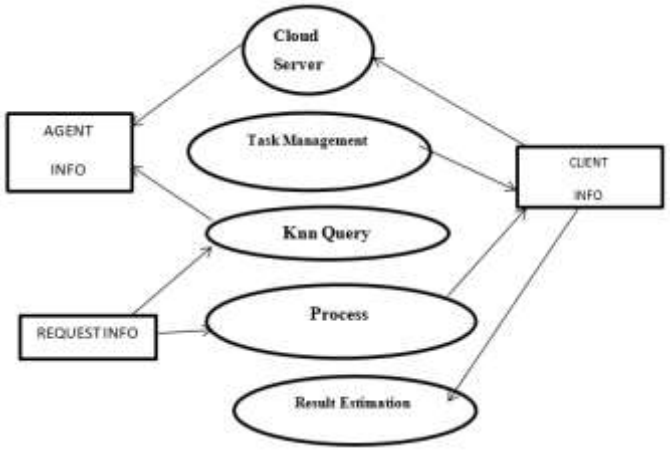

Figure 6.3 Data Flow Diagram

INPUT DESIGN

Input design is a part of overall system design, which requires careful attention. Input of data as designed as user-friendly and easier. Input design is a process of converting the user- oriented description of the input to the computer based information system into programmer- oriented specification. The objective of the input design is to create an input layout that is easy to follow and prevent operator errors.

Input Data Set: Collecting dataset from UCI Machine learning repository. the real-life data set, named Wisconsin Breast Cancer is used. The data set is publicly available on UCI machine learning repository and consists of 699 instances with nine continuous attributes. by removing some malignant instances to form a very unbalanced distribution has been employed. The resultant data set had 483 instances (39 (8 percent) malignant and 444 (92 percent) benign instances). The nine continuous attributes are not transformed into categorical attributes.

OUTPUT DESIGN

The output design refers to the results and information that are generated by the system for many end users. Efficient and intelligent output design improves the system relationships with the user and help in decision making. The output of the system is in the form of report. outliers are present among Wisconsin cancer samples, the distribution of gene expression values in cancer samples will have three sets. The Upper set corresponds to activated attributes results while the Lower indicates inactivated attributes result. the Kernel set named Kernel Set, that is a subset of the original data set, which is able to describe the original data set both in terms of data structure and of obtained results Consequently, this outlier issue can be addressed through the idea of detecting a “change point” or “break point” in the ordered gene expression values of the cancer group. A model related to fitting least squares should be effective for this goal remarkable note should be made for the definition of a new set, called kernel set, that has been demonstrated to be able to generate the “same” output results in terms of rough outlier set with time computational benefits.

VII. SYSTEM TESTING

System testing is one of the most important phases in the software development life cycle, as it ensures that the entire system works correctly and efficiently before it is deployed for real-time use. The main purpose of system testing is to verify that all the modules in the system are properly integrated and function together as expected. It also helps in identifying errors, bugs, and performance issues that may affect the system’s overall functionality.

In this project, system testing is carried out in a structured manner to ensure that the intelligent agent-based prediction system performs accurately and reliably. The testing process begins with verifying individual components of the system and gradually moves towards testing the complete system. Each

module, such as data input, data encryption, query processing, and result generation, is tested to ensure that it produces the correct output.

The testing process includes several levels such as unit testing, integration testing, validation testing, output testing, and user acceptance testing. In unit testing, individual modules are tested independently to ensure that each part of the system works as intended. This helps in detecting errors at an early stage and simplifies debugging. Integration testing is performed after unit testing, where multiple modules are combined and tested together to verify proper interaction and data flow between them. This ensures that there are no communication errors between modules.

System testing is then carried out on the complete system to evaluate its overall performance and functionality. It ensures that the system meets all the specified requirements and performs efficiently under different conditions. The system is tested using various input data to check its accuracy, reliability, and stability. Validation testing is also conducted to ensure that the system satisfies user requirements and produces correct results in real-world scenarios. This testing helps in confirming that the developed system meets the intended objectives.

Output testing plays a crucial role in verifying that the system generates results in the correct format and with high accuracy. The outputs are carefully analysed to ensure that they are meaningful, consistent, and easy to understand. Both on-screen outputs and stored results are tested to maintain quality and correctness.

User acceptance testing (UAT) is the final stage of testing, where the system is evaluated by the end users. In this stage, users interact with the system and perform various operations such as uploading data, executing queries, and viewing results. Feedback is collected from users to identify any issues or improvements needed. This ensures that the system is user-friendly and meets the expectations of real-world users.

In addition to functional testing, performance testing is also carried out to evaluate the system's efficiency in handling large datasets. The system is tested with different data sizes to measure execution time and response time. This helps in ensuring that the system performs well even under heavy workloads. Security testing is also performed to verify that the data is protected from unauthorized access. Encryption techniques used in the system are tested to ensure data confidentiality and integrity.

Overall, the system testing process confirms that the proposed system is stable, secure, and efficient. All identified errors are corrected, and the system is optimized for better performance. The successful completion of system testing ensures that the system is ready for deployment and can be effectively used in real-time applications.

A progression of tests is performed for the proposed framework before the framework is prepared for client acknowledgment testing. The testing steps are:

1. Unit testing
2. Integration testing
3. Validation testing
4. Output testing
5. User acceptance testing.

UNIT TESTING

Unit testing centres check endeavours around the littlest unit of programming plan, the module. This is otherwise called 'module testing' .The modules are tried independently.
This testing is done amid programming stage itself. In this testing step, every module is observed to work attractively as respect to the normal yield from the module.

INTEGRATION TESTING

Information can be lost over an interface; one module can adversely affect others; sub-capacities when consolidated may not deliver the coveted significant capacities; mix testing is an orderly testing for building the program structure.

While in the meantime directing to reveal mistakes related inside the interface? The goal is to take unit tried modules and to join them and test it all in all. Here redress is troublesome in light of the fact that the huge costs of the whole program confound the disconnection of causes. This is the joining testing step; every one of the mistakes experienced are rectified for the following testing step.

SYSTEM TESTING

System testing is the stage of implementation, which is aimed at ensuring that the system works accurately and proficiently before live activity begins. Testing is crucial to the achievement of the framework.

Framework testing makes a sensible supposition that if every one of the parts of the framework are right, the objective will be effectively accomplished. The hopeful framework is liable to an assortment of tests.

VALIDATION TESTING

Check testing runs the framework in a mimicked situation utilizing recreated information. This mimicked test is at times called alpha testing.

This reproduced test is fundamentally searching for mistakes and monitions in regard to end client and choices plan details cap were determined in the before stages yet not satisfied amid development. Approval alludes to the way toward utilizing programming in a live situation with a specific end goal to discover mistakes. The input from the approval stage for the most part delivers changes in the product to manage mistakes and disappointments that are revealed. Then a set of user sites is selected that puts the system in to use on a live basis. They are called beta tests.

The beta test suits use the system in day to day activities. They process live transactions and produce normal system output. The system is live in every sense of the word; except that the users are aware they are using a system that can fail. But the transactions that are entered and persons using the system are real. Validation may continue for several months. During the course of validating the system, failure may occur and the software will be changed. Continued use may produce additional failures and need for still more changes.

OUTPUT TESTING

After performing the validation, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the output generated or displayed by the system under consideration. Hence the output format is considered in two ways-one is on screen and another in printed format.

USER ACCEPTANCE TESTING

User acceptance of a system is the key factor for the success of any system. The system under consideration is tested for the user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes whenever required. This is done in regard to the following point:

An acceptance test has the objective of selling the user on the validity and reliability of the system .it verifies that the system's procedures operate to system specifications and that the integrity of important data is maintained. Performance of an acceptance test is actually the user's show. User motivation is very important for the successful performance of the system. After that a comprehensive test report is prepared. This report shows the system's tolerance, Performance range, error rate and accuracy.

VIII. SYSTEM IMPLEMENTATION

System implementation is a crucial phase in the software development process where the designed system is converted into a working application. It involves translating system design into actual code, integrating different modules, and deploying the system in a suitable environment. The main objective of system implementation is to ensure that the developed system operates smoothly, efficiently, and meets all the functional and non-functional requirements.

In this project, the implementation of the intelligent agent-based prediction system is carried out using a structured and modular approach. The system is developed using Java programming language, which provides platform independence and robust performance. The implementation follows a client-server architecture, where the client interacts with the system through a user interface and the server handles data processing, storage, and computation. Communication between the client and server is established using networking concepts such as TCP/IP and socket programming, ensuring reliable data transmission.

The implementation process begins with setting up the development environment and configuring the required tools and software. The dataset used for the project is collected, pre-processed, and prepared for analysis. Data preprocessing involves cleaning the dataset, removing inconsistencies, and converting it into a suitable format for processing. After preprocessing, the data is secured using encryption techniques such as Random Space Perturbation (RASP), which ensures that sensitive information is protected before being stored in the cloud.

The system is divided into multiple modules, each responsible for a specific functionality. The cloud server module manages data storage and handles

encrypted data securely. The task management module coordinates the execution of different operations and ensures smooth workflow within the system. The KNN query processing module is responsible for analysing the data and identifying the nearest neighbour values based on user queries. This module plays a key role in prediction and classification tasks. The result processing module generates the final output by decrypting the processed data and presenting it in a user-friendly format.

During implementation, each module is developed and tested independently before being integrated into the complete system. This modular approach helps in identifying and fixing errors at an early stage. Once all modules are integrated, the system is tested to ensure proper interaction between components and smooth data flow.

The system also includes a user interface that allows users to perform various operations such as registration, login, data upload, and query execution. The interface is designed to be simple and user-friendly, enabling easy interaction with the system. Users can input their queries, and the system processes these queries using the implemented algorithms to generate accurate results.

The implementation phase also focuses on performance optimization and system efficiency. The algorithms are designed to handle large datasets with minimal computational cost. The use of cloud computing provides scalability, allowing the system to manage increasing data volumes and user requests effectively.

Maintenance is another important aspect of system implementation. The system is designed in such a way that modifications and updates can be easily performed without affecting overall performance. This ensures that the system can adapt to future requirements and technological advancements.

Overall, the system implementation successfully transforms the proposed design into a fully functional application. The system demonstrates efficient data processing, secure data handling, and accurate prediction capabilities. The use of advanced technologies and structured implementation approach ensures that the system is reliable, scalable, and suitable for real-world enterprise applications. This phase plays a vital role in bringing the project from concept to reality and ensures that the system meets all intended objectives.

IMPLEMENTATION PROCEDURE:

It is the very basic function which describes effectively the very basic questions of how, where and when the objectives can be realized or it serves as a guiding framework. Planning equally involves a careful assessment of the available resources and the challenges which the team might have to encounter while reaching their business objectives/goals.

The objectives of this maintenance work are to make sure that the system gets into work all time without any bug. Provision must be for environmental changes which may affect the computer or software system. This is called the maintenance of the system. Nowadays there is the rapid change in the software world. Due to this rapid change, the system should be capable of adapting these changes. In this project the process can be added without affecting other parts of the system.

Maintenance plays a vital role. The system is liable to accept any modification after its implementation. This system has been designed to Favor all new changes. Doing this will not affect the system's performance or its accuracy.
All experiments were conducted on an Intel 2.8 GHz Pentium Core2Duo system with a 512 KB cache, an 800 MHz EPCI bus, and 2 GB DDR2 of RAM.

The system runs Windows 8.We present four arrangements of trial results to research the accompanying inquiries, correspondingly. (1) How costly is the RASP irritation? (2) How versatile the OPE improved RASP is to the ICA-based assault? (3) How effective is the two-arrange go question processing. Then, the arranged unique qualities are relatively apportioned by the objective pail dissemination to make the basins for the first conveyance. With the adjusted unique and target cans, a unique esteem can be mapped to the objective can and suitably scaled. Therefore, the data mining cost mainly comes from the bucket search procedure (proportional to log D, where D is the number of buckets). It shows the cost distributions for 20K records at different number of dimensions. The dimensionality has slight effects on the cost of RASP perturbation. Overall, the cost of processing 20K records is only around 0.1 second.

EXPERIMENTAL ANALYSIS

The previous sections have addressed several major aspects: the cloud-client algorithms, the client-side cost analysis, and the confidentiality analysis. The experiments will study how the cloud-client algorithms perform in terms of different

settings that may also involve the trade-off between model quality and client-side costs. Specifically, we will show the scalability of our approach with client-side computation and communication costs on real datasets. We will conduct a set of experiments to understand which of the four private learning methods is the best in terms of costs and model quality. We will evaluate model confidentiality with the proposed method.

Table 8.1 - Datasets for experiments

Dataset	Records	Dimensions	Link
German Credit	1000	20	https://goo.gl/lVy34O
Ozone Days	2536	73	https://goo.gl/Si6aDh
Spam base	4601	57	https://goo.gl/WPyXTi
Bank Marketing	45211	17	https://goo.gl/vvgj3M
Twitter Buzz	140000	77	https://goo.gl/Yfy80u

EXPERIMENT SETUP

Datasets: For easier validation and reproducibility of our results, we use a set of public datasets from UCI machine learning repository for evaluation, each of which has only two classes. These datasets have been widely applied in various classification modelling and evaluation. These datasets have been widely applied in classification modelling and evaluation. Table 1listsstatistics of the datasets. They cover different scales and dimensions to make the results more representative. In pre-processing, each dimension of the datasets is normalized with the transformation$(v - \mu j)/\sigma j$,where$\mu j$is the mean and$\sigma 2$is the variance of the dimension j.

Implementation: We implement the perturbation methods based on the algorithms in the corresponding papers. The Greedy K-NN is implemented based on the AdaBoost algorithm. The four learning algorithms are implemented as plugins to the framework. All these implementations use C++ and are thoroughly tested on an Ubuntu Linux server.

IX. CONCLUSION

This paper presents the Greedy K-NN that aims to provide practical confidential classifier learning with the cloud or a third-party mining service provider. Confidential cloud mining should address four aspects: data confidentiality, model confidentiality, model quality, and low client side costs. We use the RASP perturbation to guarantee the data confidentiality. However, it is difficult to learn a high quality classifier from the RASP perturbed data, as it only allows linear queries, which can be translated to non-optimal linear classifiers. We develop the boosting based Greedy K-NN to obtain high-quality classifiers with these non-optimal linear classifiers. The intuition is that boosting requires only weak base classifiers that are slightly better than random guesses.

Four algorithms are developed with the same working pattern: the client provides a set of encoded base classifiers, and the cloud computes a boosting model from the set. This pattern does not require the client to stay online during boosting iterations, which is convenient for the client. We have developed four such algorithms: DS Pool, LCPool, DerivedDS, andDerivedLC as the candidates. The confidentiality of data, query, learning process and models is formally analysed, and we show the confidentiality of data and query is satisfactorily guaranteed.

We have conducted an extensive evaluation of the proposed algorithms and studied the effect of the major factors in the Greedy K-NN. The result shows that DS Pool andDerivedDS can generate high-quality models with accuracy very close to the optimal boosting models. We also evaluate the concept of model confidentiality for real data and models and show that the model confidentiality is well preserved under the security assumption.

Future Enhancement

The future work on random space data aggregation will be focus offers numerous advantages over it have become possible to outsource large databases to database service providers and let the providers maintain the range-query service.

In any case, a few information may be touchy that the information proprietor does not have any desire to move to the cloud except if the information secrecy and inquiry security are ensured.

We need to centre the Random Space Encryption (RASP) approach that permits effective range seek with more grounded assault versatility than existing proficiency cantered methodologies. The arbitrary space bother (RASP) information irritation technique to give secure and productive range inquiry and KNN question administrations for ensured information in the cloud.

Future enhancement plays an important role in improving the capabilities and performance of the proposed system. Although the current system provides an efficient and secure prediction mechanism using intelligent agents and cloud-based processing, there are several areas where further improvements can be made to enhance its functionality, scalability, and real-world applicability.

One of the major enhancements that can be incorporated is the integration of advanced artificial intelligence and deep learning techniques. While the current system uses algorithms like K-Nearest Neighbour (KNN) for prediction, future versions can include more sophisticated models such as neural networks, decision trees, or ensemble methods to improve prediction accuracy and handle more complex datasets. These models can learn patterns more effectively and provide better decision-making support.

Another important enhancement is real-time data processing. Currently, the system processes data in a structured manner, but future improvements can enable real-time analytics where data is processed instantly as it is generated. This is especially useful in enterprise environments where quick decision-making is critical. Real-time processing can be achieved by integrating streaming technologies and high-performance computing techniques.

Scalability is another key area for improvement. As data volume continues to grow, the system can be extended to support big data technologies such as distributed computing frameworks. This will allow the system to process large-scale datasets efficiently without performance degradation. Cloud platforms can be further optimized to dynamically allocate resources based on workload, ensuring better performance and cost efficiency.

Security can also be enhanced by implementing more advanced encryption techniques and multi-layer security mechanisms. While the current system uses methods like Random Space Perturbation (RASP), future systems can incorporate hybrid encryption models, blockchain-based security, and stronger authentication protocols to provide higher levels of data protection. This is particularly important when dealing with sensitive enterprise data.

The user interface of the system can be improved to provide a more interactive and user-friendly experience. Future enhancements may include graphical dashboards, data visualization tools, and customizable reports that help users better understand the analysis results. This will make the system more accessible to non-technical users and improve overall usability.

Another possible enhancement is the integration of Internet of Things (IoT) devices for real-time data collection. This will allow the system to gather data from various sources such as sensors and smart devices, making the prediction system more dynamic and applicable to modern applications like smart cities, healthcare monitoring, and industrial automation.

In addition, the system can be extended to support multiple domains and applications. Currently, the system is designed for enterprise service platforms, but it can be adapted for use in healthcare, finance, education, and other sectors. This flexibility will increase the scope and usefulness of the system. Performance optimization techniques can also be applied in future versions to reduce processing time and improve efficiency. Techniques such as parallel processing, optimized algorithms, and caching mechanisms can be used to enhance system performance.

Overall, the future enhancements aim to make the system more intelligent, secure, scalable, and user-friendly. By incorporating advanced technologies and improving existing features, the system can be transformed into a more powerful and versatile solution capable of handling complex real-world challenges. These improvements will ensure that the system remains relevant and effective in the rapidly evolving technological landscape.

SOURCE CODE

The project is implemented using Java Swing and Socket Programming. The system contains different modules such as User Module, Agent Module, and Cloud Module.

USER MODULE

The User Module is responsible for interacting with the enterprise client. The user enters a unique user ID through the graphical interface. After receiving the user ID, the system creates a user frame for communication with the cloud server. The module also initializes the UserReceiver thread, which continuously listens for responses from the cloud and intelligent agents. This module acts as the front-end

communication layer between the enterprise user and the distributed predictive system

Sample Code: package user import javax.swing.JOptionPane; import javax.swing.JFrame; public class Main {
public static void main(String[] args) {
// Getting User ID from user int id = Integer.parseInt(JOptionPane.showInputDialog(
new JFrame(), "Enter User Id"));
// Opening User Frame
UserFrame uf = new UserFrame(id); uf.setTitle("User - " + id); uf.setResizable(false); uf.setVisible(true);
// Starting User Receiver Thread UserReceiver ur =
new UserReceiver(uf,id); ur.start();

AGENT MODULE

The Agent Module allows agents to select tasks and communicate with the cloud server.
Sample Code:

String aid=jTextField1.getText().trim();
String task=jComboBox1.getSelectedItem().toString(); DatagramSocket ds=new DatagramSocket();
String ms="AgentInfo#"+aid+"#"+task;

byte dd[]=ms.getBytes(); DatagramPacket dpt= new DatagramPacket(dd,
0,
dd.length, InetAddress.getByName("127.0.0.1"), 9000);
ds.send(dpt);

Explanation:

- Agent ID and task are collected from the interface
- UDP socket communication is established
- Task information is sent to the cloud server
- The cloud manages intelligent task allocation

Cloud Module

The Cloud Module acts as the central management system of the proposed enterprise predictive platform. It monitors all connected agents, client requests, and task allocations. The module contains multiple tables to display agent information, client details, and request processing status.

The cloud server dynamically manages communication between users and intelligent agents using distributed architecture. This module improves scalability, centralized monitoring, and intelligent task coordination within the enterprise environment.

Sample Code:

package cloud;
public class CloudFrame extends javax.swing.JFrame { public CloudFrame() {
initComponents();
}
@SuppressWarnings("unchecked") private void initComponents() {
jTabbedPane1 =

new javax.swing.JTabbedPane();
// Agent Information Table
jTable1 = new javax.swing.JTable(); jTable1.setModel(
new javax.swing.table.DefaultTableModel(new Object [][] {},
new String [] { "Agent ID", "Task"
}));

// Client Information Table
jTable2 = new javax.swing.JTable(); jTable2.setModel(
new javax.swing.table.DefaultTableModel(new Object [][] {},
new String [] { "Client ID"
}));
// Request Information Table jTable3 = new javax.swing.JTable(); jTable3.setModel(
new javax.swing.table.DefaultTableModel(new Object [][] {},
new String [] { "Client ID", "Task",
"Data", "Agent ID"

}));
// Adding Tabs
jTabbedPane1.addTab("Agent Info", jTable1);
jTabbedPane1.addTab("Client Info", jTable2);
jTabbedPane1.addTab("Request Info", jTable3);
}
public static void main(String args[]) {
java.awt.EventQueue.invokeLater(new Runnable() {
public void run() { new CloudFrame()
.setVisible(true);
}
});
}
public javax.swing.JTable jTable1; public javax.swing.JTable jTable2; public javax.swing.JTable jTable3;
private javax.swing.JTabbedPane jTabbedPane1;

}

The above program demonstrates Runtime Application Self Protection in the enterprise predictive system.

Working Process:

- User request is received during runtime
- The system continuously monitors input data
- Suspicious keywords are identified
- Malicious requests are blocked immediately
- Safe requests are allowed for processing

Security Features:

- Runtime threat detection
- Input validation
- Attack prevention
- Secure enterprise communication
- Intelligent runtime monitoring

RASP with KNN Integration

In the proposed system:

- RASP protects the application during execution
- KNN performs intelligent classification and prediction
- Intelligent agents process enterprise requests securely
- Cloud modules manage secure communication

Sample Screenshots

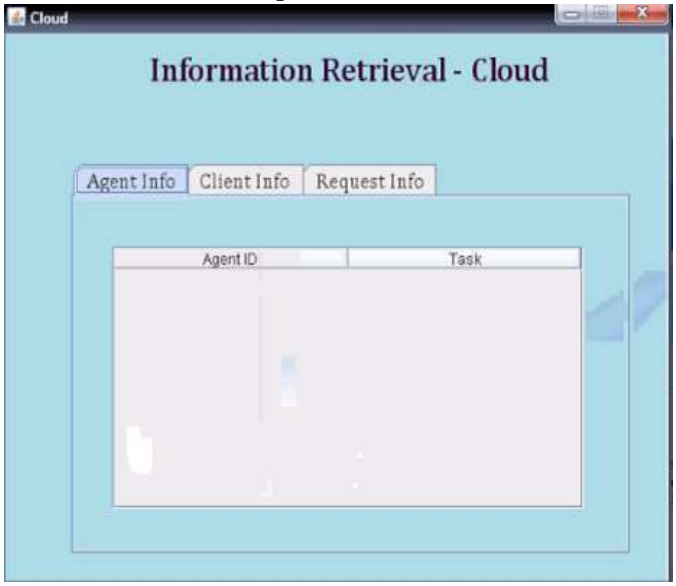

Figure – 10.1 Agent Info Panel Screenshot

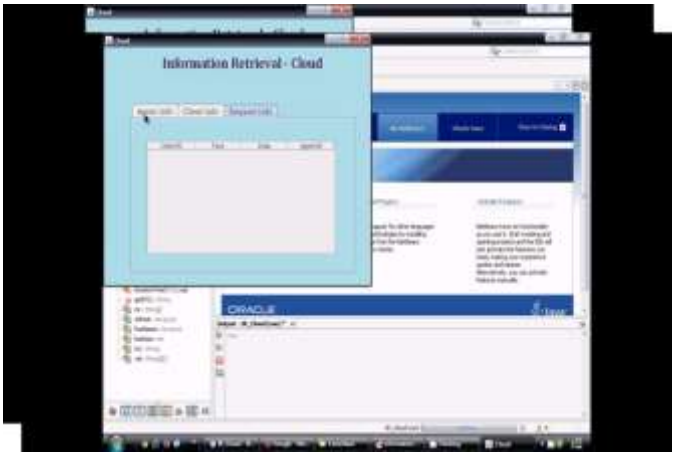

Figure - 10.2 System Home Page

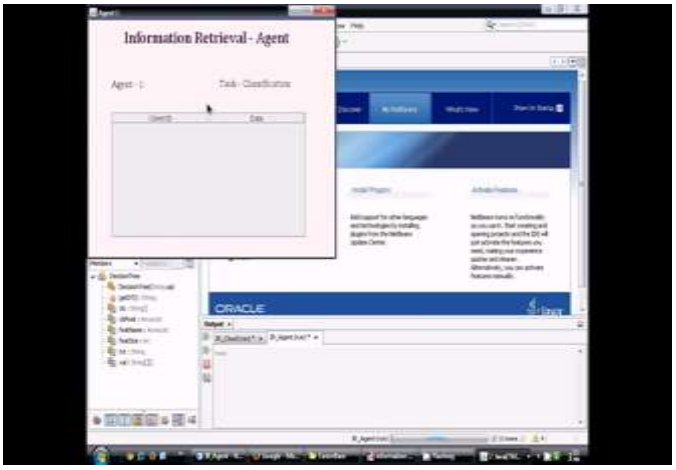

Figure - 10.3 Data Owner Registration page

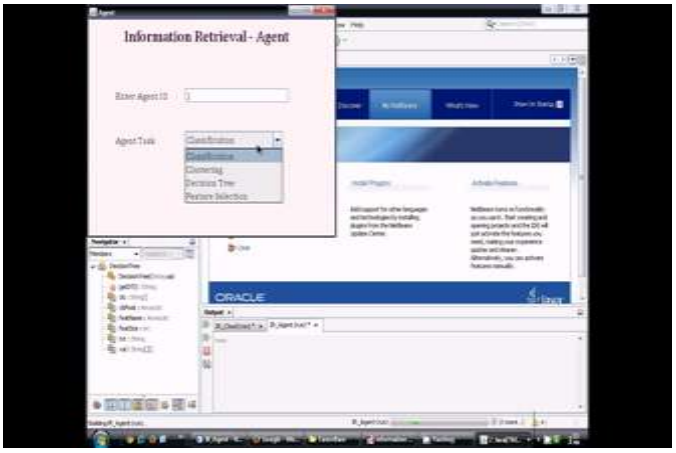

Figure - 10.4 Data Owner Login Page

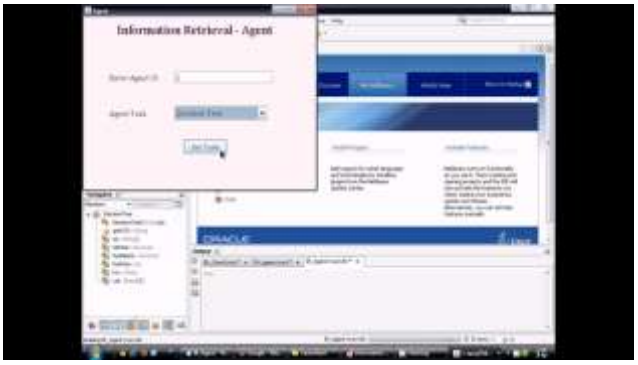

Figure - 10.5 User Registration Page

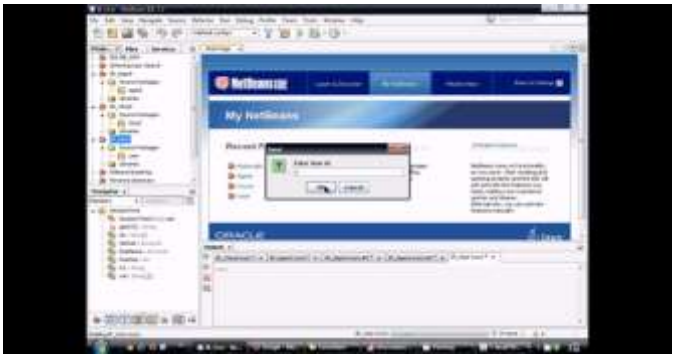

Figure - 10.9 Encrypted Data Storage in Cloud

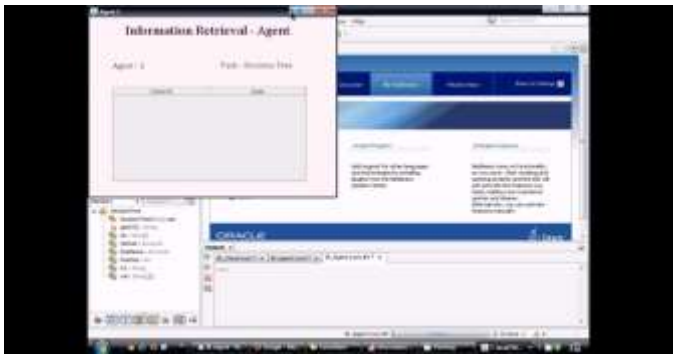

Figure - 10.6 User login Page

Figure - 10.10 Query Input Interface

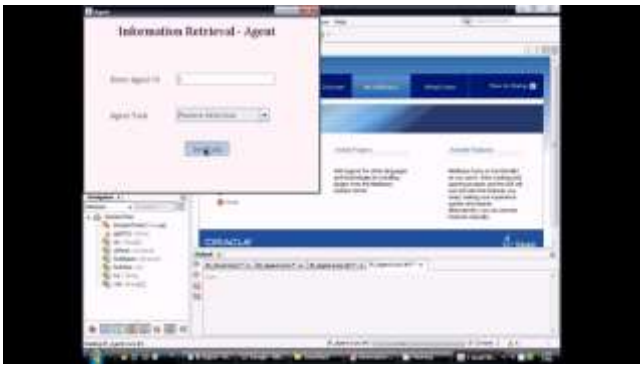

Figure - 10.7 Dataset Upload Interface

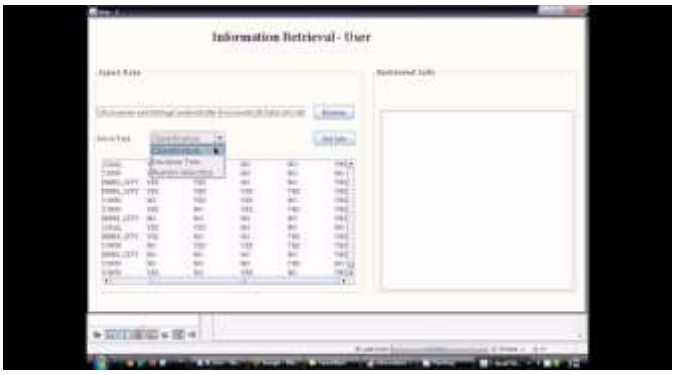

Figure - 10.10 Query Input Interface

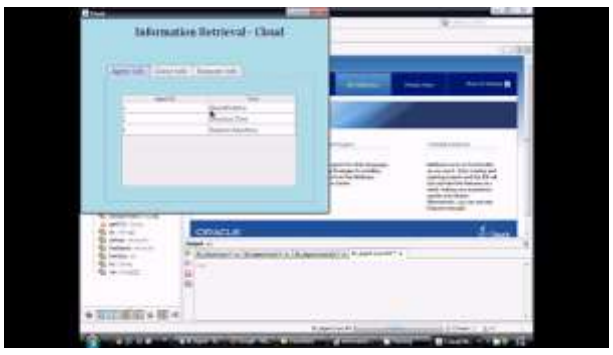

Figure - 10.8 Data Encryption Process Using RASP

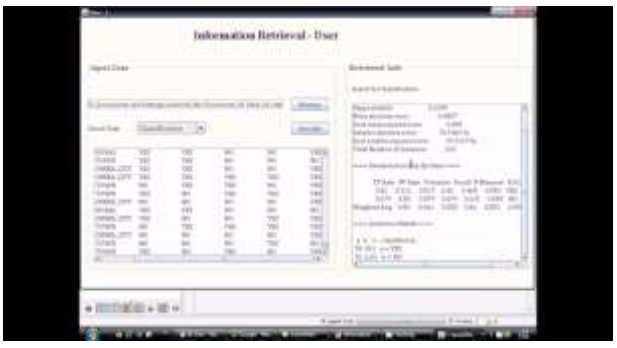

Figure - 10.11 Range Query Processing

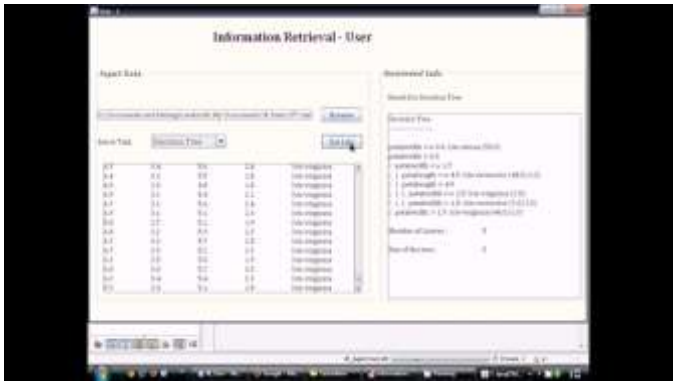

Figure - 10.12 KNN Query Processing

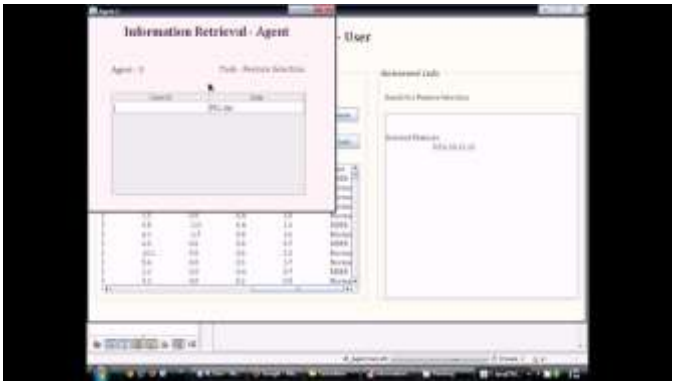

Figure - 10.16 Output Display Interface

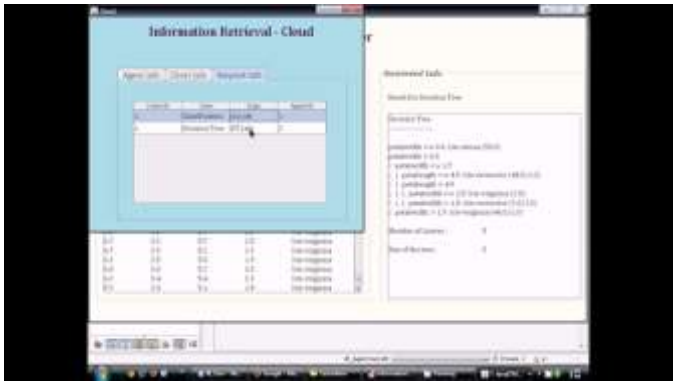

Figure - 10.13 Intermediate Result Processing

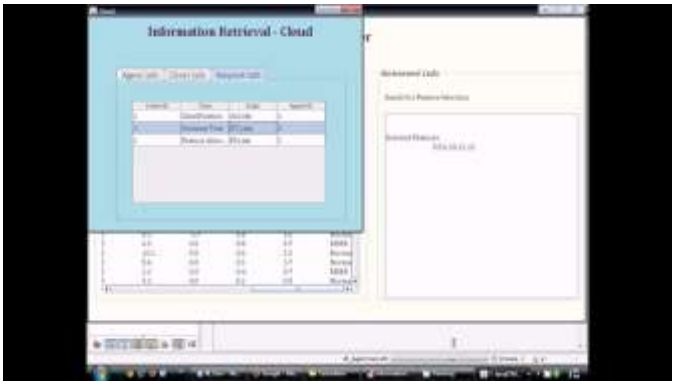

Figure - 10.17 System Workflow Completion

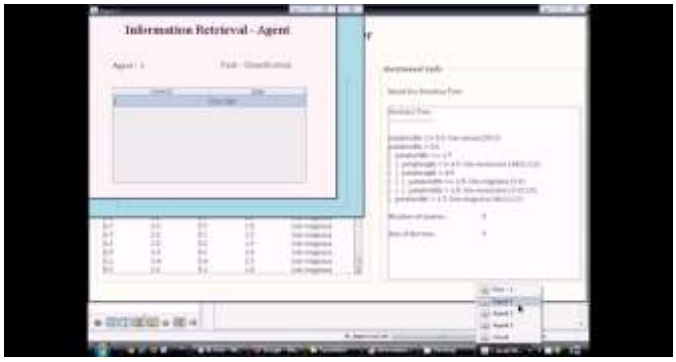

Figure - 10.14 Data Decryption Process

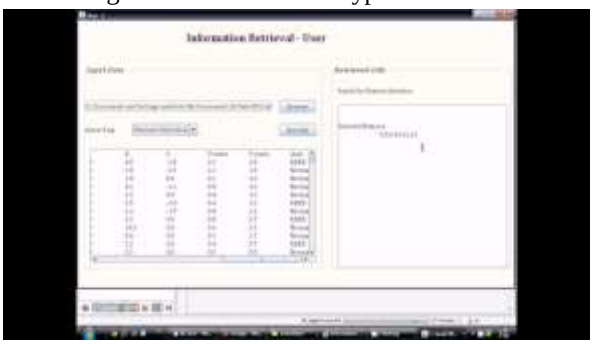

Figure - 10.15 Final Result Generation

REFERENCES

1. P. Seshadri, M. Livny, and R. Ramakrishnan, “Sequence query processing, “in SIGMOD, 1994.
2. P. Seshadri, M. Livny, and R. Ramakrishnan, “The design and implementation of a sequence database system,” in VLDB, 1996.
3. B. Babcock, et al., “Models and issues in data stream systems,” in PODS, 2002.
4. J. Chen, et al., “Niagara: a scalable continuous query system for internet databases,” SIGMOD Rec., 2000.
5. F. Wang, et al., “Temporal management of RFID data,” in VLDB, 2005.
6. H. Gonzalez, J. Han, X. Li, and D. Klabjan, “Warehousing and Analyzing Massive RFID Data Sets,” in ICDE, 2006.
7. H. Gonzalez, J. Han, and X. Li, “FlowCube: Constructing RFIDFlowCubes for Multi-Dimensional Analysis of Commodity Flows,” inVLDB, 2006.
8. E. Lo, B. Kao, W.-S. Ho, C.-K. Chui, and D. Cheung, “OLAP on sequence data,” in SIGMOD, 2008, pp. 649–660.

9. Z. He, P. Wong, B. Kao, E. Lo, and R. Cheng, "Fast evaluation of iceberg pattern-based aggregate queries," in CIKM, 2013, pp. 2219–2224.
10. K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla, "On synopses for distinct-value estimation under multiset operations," in SIGMOD, 2007, pp. 199–210.
11. R. Ramakrishnan, D. Donjerkovic, A. Ranganathan, K. S. Beyer, and M.
12. Krishnaprasad, "SRQL: Sorted Relational Query Language," in SSDBM, 1998.
13. R. Sadri, C. Zaniolo, A. Zarkesh, and J. Adibi, "Optimization of sequence queries in database systems," in PODS, 2001.
14. J. Gray, et al., "Data cube: A relational aggregation operator generalizinggroup-by, crosstab, and sub-totals," Data Mining and Knowledge Discovery., vol. 1, no. 1, pp. 29–53, 1997.
15. M. Fang, N. Shivakumar, H. Garcia-Molina, R. Motwani, and J. D. Ullman, "Computing iceberg queries efficiently," in VLDB, 1998, pp. 299–310.
16. K. S. Beyer and R. Ramakrishnan, "Bottom-up computation of sparse and iceberg cubes," in SIGMOD, 1999, pp. 359–370.
17. K. A. Ross and D. Srivastava, "Fast computation of sparse datacubes, "in VLDB, 1997, pp. 116–125.
18. D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "Olap over uncertain and imprecise data," in VLDB, 2005, pp. 970–981.
19. B. Zhou, D. Jiang, J. Pei, and H. Li, "OLAP on search logs: an infrastructure supporting data-driven applications in search engines," in SIGKDD, 2009, pp. 1395–1404.
20. M. Liu , et al., "E-cube: Multi-dimensional event sequence processing using concept and pattern hierarchies," in ICDE, 2010, pp. 1097–1100.
21. C. Yixin, et al., "Multi-dimensional regression analysis of time-series data streams," in VLDB, 2002.
22. J. Bae and S. Lee, "Partitioning Algorithms for the Computation of Average Iceberg Queries," in DaWaK, 2000.
23. B. He, H.-I. Hsiao, Z. Liu, Y. Huang, and Y. Chen, "Efficient iceberg query evaluation using compressed bitmap index," TKDE, vol. 24, no. 9, pp. 1570–1583, 2012.
24. S. Agarwal, et al., "On the computation of multidimensional aggregates," in VLDB, 1996, pp. 506–521.
25. M. Hadjieleftheriou, X. Yu, N. Koudas, and D. Srivastava, "Hashed samples: selectivity estimators for set similarity selection queries," PVLDB, vol. 1, no. 1, pp. 201–212, 2008.
26. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules,"
27. Proceedings of the 20th International Conference on Very Large Data Bases, 1994.
28. J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.
29. C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages and Programming, 2006.
30. K. Chen and L. Liu, "Privacy Preserving Data Classification with Rotation Perturbation," IEEE Transactions on Knowledge and Data Engineering, 2005.
31. B. Pinkas, "Cryptographic Techniques for Privacy-Preserving Data Mining," ACM SIGKDD Explorations, 2002.
32. I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.
33. T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, 1967.
34. L. Breiman, "Random Forests," Machine Learning Journal, 2001.
35. Y. Lindell and B. Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," Journal of Cryptology, 2009.
36. M. Armbrust et al., "A View of Cloud Computing," Communications of the ACM, 2010.
37. Amazon Web Services, "Overview of Cloud Computing," 2019.
38. Oracle, "Java Platform Documentation," 2020.
39. Sun Microsystems, "Java Networking and Socket Programming Guide," 2018.
40. UCI Machine Learning Repository, "Dataset Collection for Machine Learning," University of California, Irvine.
41. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 2010.
42. X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), 2006.