Practical Privacy Preserving of data using Bigdata

Shubhangini L, Azhar Baig M.
Ballari Institute of Technology and Management, Ballari, Karnataka, India.

Abstract

Clustering techniques have been widely adopted in many real world data analysis applications, such as customer behavior analysis, medical data analysis, digital forensics, etc. This is because cloud computing offers not only reliable services with performance guarantees, but also savings on in-house IT infrastructures. However, as datasets used for clustering may contain sensitive information, e.g., patient health information, commercial data, and behavioral data, etc, directly outsourcing them to any distributed servers inevitably raise privacy concerns. In this paper, we propose a practical privacy-preserving K-means clustering scheme that can be efficiently outsourced to cloud server.

Keywords: Machine Learning, K-Means, Clustering Techniques, Map Reduce.

I. Introduction:

Clustering is one major task of exploratory data mining and statistical data analysis, which has been ubiquitously adopted in many domains, including healthcare, social network, image analysis, pattern recognition, etc. Meanwhile, the rapid growth of big data involved in today’s data mining and analysis also introduces challenges for clustering over them in terms of volume, variety, and velocity. To efficiently manage large-scale datasets and support clustering over them, public cloud infrastructure is acting the major role for both performance and economic consideration. Nevertheless, using public cloud services inevitably introduces privacy concerns. This is because not only many data involved in data mining applications are sensitive by nature, such as personal health information, localization data, financial data, etc, but also the public cloud is an open environment operated by external third-parties. For example, a promising trend for predicting an individual’s disease risk is clustering over existing patients health records, which contain sensitive patient information.

Therefore, appropriate privacy protection mechanisms must be placed when outsourcing sensitive datasets to the public cloud for clustering
II. Related works:

Cryptanalysis of a Homomorphic Encryption Scheme:

Homomorphic encryption allows making specific operations on private data which stays encrypted. While applications such as cloud computing require to have a practical solution, the encryption scheme must be secure. we are able to mount three attacks.

The first attack enables to recover a secret plaintext message broadcasted to multiple users. The second attack performs a chosen cipher text key recovery attack and it was implemented and verified. The last attack is a related chosen plaintext decryption attack. Privacy concern has become an important issue in data mining. A novel algorithm for privacy preserving in distributed environment using data clustering algorithm has been proposed. As demonstrated, the data is locally clustered and the encrypted aggregated information is transferred to the master site. This aggregated information consists of centroids of clusters along with their size.

Theoretical and experimental analysis illustrates that the proposed algorithm can effectively solve privacy preserving problem of clustering mining over distributed data and achieve the privacy-preserving aim.

Cloud computing is certainly revolutionary model which enables flexible, on-demand and low-cost usage of computing resources. Benefits of Cloud Computing are the sources of privacy and security problems, which emerge since the data owned by different consumers are kept in some cloud servers instead of under their own control. The privacy challenge is to design cloud services in such a manner as to decrease privacy risk and give compliance that is legal. For example, a business storing data about individual consumers using the cloud that is prominent company is sometimes high risk. Clients may not be in a position to sue enterprises if their privacy rights tend to be violated. This paper surveys privacy issues and security in cloud computing.

III. Data and methodology:

Existing System

Existing cloud computing system offers not only reliable services with performance guarantees, but also savings on in-house IT infrastructures. However, as datasets used for clustering may contain sensitive information, e.g., patient health information, commercial data, and behavioral data, etc, directly outsourcing them to public cloud servers inevitably raises privacy concerns.

Disadvantages of Existing System

- Less Efficiency
- Security is less
- Speed of Transmission is low
Proposed System

We are proposing a system K-means clustering over Large-scale Dataset using Map Reduce technique. First we are initializing trained data set for every different cluster which is related to medical Information. After, the clustering algorithm divide file into number of chunks and for every chunks hash code is generated for the security purpose. Before storing into Cloud System, classification algorithm classifies that file belong to which cluster category.

Advantages of the Proposed System

- More Efficiency
- Security is more because of Hash code generation
- Speed of Transmission is high because of deduplication concept is used while uploading file to the Cloud storage

IV. Implementation details

1 Logical Design

Design for WebApps encompasses technical and non-technical activities. The look and feel of content is developed as part of graphic design; the aesthetic layout of the user interface is created as part of interface design; and the technical structure of the WebApp is modeled as part of architectural and navigational design.

This argues that a Web engineer must design an interface so that it answers three primary questions for the end-user:

1. Where am I? – The interface should (1) provide an indication of the WebApp has been accessed and (2) inform the user of her location in the content.

2. What can I do now? – The interface should always help the user understand his current options- what functions are available, what links are live, what content is relevant.

3. Where have I been; where am I going? – The interface must facilitate navigation. Hence it must provide a “map” of where the user has been and what paths may be taken to move elsewhere in the WebApp.

Design goals – the following are the design goals that are applicable to virtually every WebApp regardless of application domain, size, or complexity.
1. Simplicity
2. Consistency
3. Identity
4. Visual appeal
5. Compatibility.

Design leads to a model that contains the appropriate mix of aesthetics, content, and technology. The mix will vary depending upon the nature of the WebApp, and as a consequence the design activities that are emphasized will also vary.

**The activities of the Design process:**

1. Interface design-describes the structure and organization of the user interface. Includes a representation of screen layout, a definition of the modes of interaction, and a description of navigation mechanisms. Interface Control mechanisms- to implement navigation options, the designer selects form one of a number of interaction mechanism;

   - **a.** Navigation menus
   - **b.** Graphic icons
   - **c.** Graphic images

   Interface Design work flow- the work flow begins with the identification of user, task, and environmental requirements. Once user tasks have been identified, user scenarios are created and analyzed to define a set of interface objects and actions.

2. Aesthetic design-also called graphic design, describes the “look and feel” of the WebApp. Includes color schemes, geometric layout. Text size, font and placement, the use of graphics, and related aesthetic decisions.

3. Content design-defines the layout, structure, and outline for all content that is presented as part of the WebApp. Establishes the relationships between content objects.

4. Navigation design-represents the navigational flow between contents objects and for all WebApp functions.
5. Architecture design identifies the overall hypermedia structure for the WebApp. Architecture design is tied to the goals establish for a WebApp, the content to be presented, the users who will visit, and the navigation philosophy that has been established.

a. Content architecture, focuses on the manner in which content objects and structured for presentation and navigation.

b. WebApp architecture, addresses the manner in which the application is structure to manage user interaction, handle internal processing tasks, effect navigation, and present content. WebApp architecture is defined within the context of the development environment in which the application is to be implemented.

J2EE uses MVC Architecture

Component design-develops the detailed processing logic required to implement functional components
Flow diagram:
V. Results and conclusion:
VI. Conclusion:

In this project, we proposed a privacy-preserving Map Reduce based K-means clustering scheme in cloud computing. Thanks to our light-weight encryption design based on the LWE hard problem, our scheme achieves clustering speed and accuracy that are comparable to the K-means clustering with privacy-protection. Considering the support of large-scale dataset, we securely integrated Map Reduce framework into our design, and make it extremely suitable for parallelized processing in cloud computing.
environment. Our prototype implementation over 5 million data objects demonstrates that our scheme is efficient, scalable, and accurate for K-means clustering over large-scale dataset.

References:


