Peer Group Analysis in Identity and Access Management to Identify Anomalies

Mayuri Dhamdhere, Shridevi Karande, Madhura Phatak
Department of Computer Engineering
Maharashtra Institute of technology, Pune, Maharashtra, India
mayu18dhamdhere@gmail.com, shridevi.karande@mitpune.edu.in, madhura.phatak@mitpunr.edu.in

Abstract: Nowadays companies need to manage user accesses across multiple channels such as mobile, social and cloud. Business is always about delivering value to customers, and Identity and Access Management is very essential part of ensuring that employees are both empowered to deliver that value and prohibited from damaging the security, business reputation. Organization can use Identity and Access data to detect the attacks within the organization. Data Analytics techniques can be used to identify the anomalies in the identity and Access Data, out of that Peer group Analysis is one technique. Peer Group Analysis focuses on the local pattern rather than the global models; anomaly may not be present when compared with whole data but there may be some outliers present when compared with its peer groups. Context anomaly is the data points that considered abnormal when compared with its peer group. In this research we are doing peer group analysis on Identity and Access Data using the various outlier detection algorithms. Various context anomaly detection algorithms are used to identify the abnormal user behaviour which gives the risk insights the organization. We are proposing peer group analysis which can able to locate entitlements that can be exclusive than other users of same group, which gives the risk of the user access.

Keywords: Anomaly detection, Context Anomaly, Identity and Access Management (IAM), Peer Group Analysis.

I. INTRODUCTION

A. Identity and Access Management Concept

In today's world organizations need to implement efficient and flexible business processes and information. Such processes require more reliable and accurate identity and access management solutions. Protected user's accesses play a very important role in the exchange of data and information. In addition, electronic data is becoming ever more valuable for most companies. Access protection must therefore meet increasingly strict requirements an issue that is often solved by introducing strong authentication. Modern IAM techniques allow users administration and management of their access rights more effectively. Cloud is used in IAM, which handles data and accesses. Prerequisite for that is all information has to be clearly defined and monitored. Data Analytics plays an important role to make the intelligent access management. Machine learning approach is used to find out the patterns, outliers in the IAM data which can effectively gives the list of risky users.

Top 5 Identity and Access Management Tools [18] are,

1. SailPoint
2. RSA
3. IBM
4. Oracle,
5. Courion

B. Peer Group Analysis-

Peer group analysis is the analysis technique which compares the target object with the other object which is identified as similar to the target object in some sense (peer group) [7]. The advance of PGA is different. User's profile is created based on the behaviour of some similar users where current outlier detection techniques over time include profiling for single user. PGA considers local patterns in the data rather than global patterns; a data sequence may not give results unusually when it is compared with the complete data set of sequences but it shows unusual properties when compared with its peer group. That is, it may begin to deviate in behaviour from objects to which it has previously been similar. PGA is a technique which has been developed to describe the analysis of the time development of a given data (the target) relative to other objects data that have been identified as similar initially. The PGA approach is different. The profile of user is formed based on the behaviour of several similar users. Current outlier detection techniques over time include profiling for single user.

As companies rely more heavily on computerized systems to run their big business, Such companies are facing increased complexity in professionally running user identities at each of these three stages of the identify lifecycle. This computerized system with less intelligence brings the risk of giving access more than need and job title. There need an intelligence which we can bring through various machine learning approaches. Peer group analysis is one of the techniques which identify the anomalies present in the Data which helps the organization to take right decision. Also there are very less research on the IAM and peer group analysis is very well technique to identify the fraud.

II. LITERATURE SURVEY

Current IAM framework is given by Ertem Osmanoglu in his Identity and Access management business performance through connected intelligence book [9] (55-80). Detailed IAM framework is described with the current state and the various challenges is given in this book. In future state roadmap they have given the intelligent IAM with risk based approach which can be achieved through four methods such as Peer group analysis, Role Analysis and Resource allocation and analysis. This book presents an action plan to help organizations to develop an efficient IAM strategy. Using this approaches organization can manage risk associated with the identity and access and reduce important business progression. They also pressure out the significance of identity and access management (IAM) when dealing with main business processes. Companies are able to detect extraordinary access and outliers forms are completely essential for a manager to be able to address many of the challenges in each of the IAMs major areas. They draw on the fact that identity management implies data analysis, reporting and ongoing monitoring, and modeling and efficient decision making processes in order to emphasize the importance of IAM and the necessity for specialized IT solutions. There are some bad provisioning practices which cause managers to have a limited understanding of the required entitlement level for their employee and over-assignation of entitlements occur. Therefore, the ability of an IAM team to provide a manager a risk-based report view showing the access requested is critical for the stakeholder of an application/service/department to understand the risks associated with the requested access.

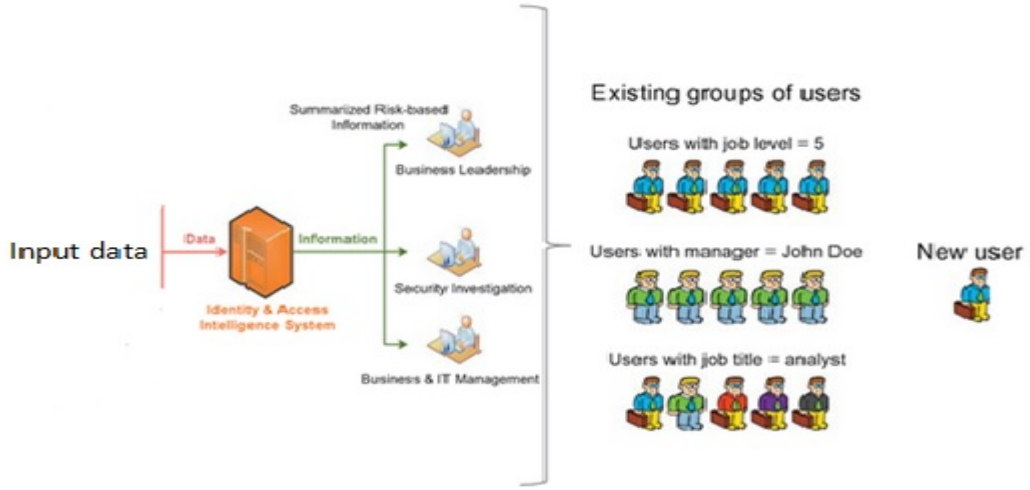

Fig 1: Peer Group Analysis in Identity and Access Management (Scenario) [9] (181)

Figure 1 describes the scenario of peer group analysis, Groups of users are created based on some attributes of users such as profile type and job title and then new user is assigned to the appropriate group. So that new user can have appropriate entitlement to do his/her job. But if new user is assigned to the one group and he had given the entitlement consistent with other group then that new user should identified as outlier through analysis.

To Identity the Outlier in the data various outlier algorithms studied by the chandola in his Anomaly Detection: A Survey [4]. In this survey basic anomaly types are given such as point anomaly, context anomaly, and collective anomaly. Various Techniques are given in this paper with its advantages and disadvantages. Application of these anomaly detection is well explained in this paper which includes the fraud detection, Intrusion detection, Activity monitoring, Network performance etc. Anomaly Detection for Discrete Sequences: A Survey is another paper which gives the techniques to detect anomaly in the discrete sequences [5].

David J. Weston , David J. Hand ,Niall M. Adamsin has given the Idea of Plastic card fraud detection using peer group analysis [6]. They demonstrated the existing plastic card transaction accounts that evolve sufficiently closely to enable fraudulent behaviour to be detected. Using real world data consisting of high transaction volume accounts, they showed three months of transaction history was adequate to produce peer groups that could usefully track a target for at least one further month. They have also shown screen accounts to determine which are more likely to be open to peer group analysis. Fraud detection methodologies have their own characteristic weaknesses and strengths. The particular strength of anomaly detection approaches such as peer

group analysis is adaptability to new types of fraud. Peer group analysis extends anomaly detection by borrowing strength from the population of accounts. Supervised methods have the ability to detect known patterns of fraud more reliably than anomaly detection methods. A fraud detection system, therefore, is unlikely to rely solely on one method. Rather than comparing relative performance of different methods we should be looking at ways to combine those methods [12]. In this regard, peer group analysis looks promising as a component in a fraud detection system since it monitors the data for anomalies in a completely different way to typical anomaly detection techniques. Finally the method by which they constructed synchronous time series delayed the fraud detection until the end of the day. This is an unrealistic approach to fraud detection since it is quite clearly something a fraudster could exploit. The time series should be constructed such that the peer group comparisons are performed immediately at the time of the actual transaction.

III. SYSTEM ANALYSIS

A. System Architecture

System Architecture of Peer Group Analysis describes how the modules and the component of the system are comprised and how they work together to build the overall system. Aim of the whole architecture is to identify the outlier in the identity and access data of the company. Outlier Analysis is used to find user entitlements that are different than other users with a similar profile.

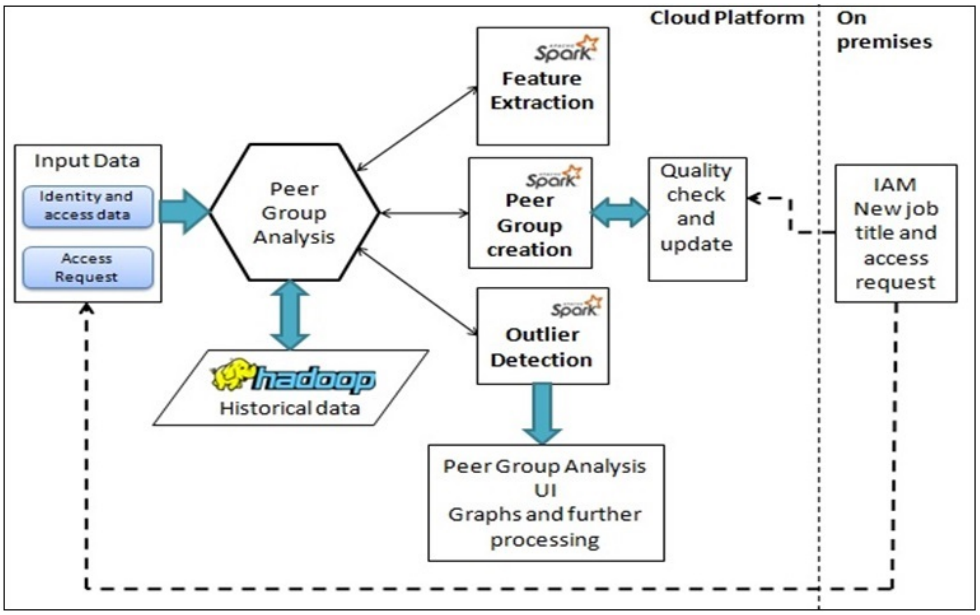

Fig 2: System Architecture

Figure 2 shows that the system is divided into the tree modules named feature extraction, peer Group creation and outlier detection. Identity and Access data set is given as input to the system and the outlier in the data is output of the system. Hadoop is used to store the historical data. All the modules are implemented on the Hadoop framework. Input data is taken from the on premises and which is given to the cloud for processing.

B. Input Data

Input data is the identity and access management data, which includes the files such as User, Entitlements, Account, Application, Organisation units. All the data is stored in the csv file and the metadata data of these files stored separately to validate data and dynamically add more columns to csv file. Input data also contains the access requests that are coming for outlier checking. All these input data is getting from the various tools which governs the Identity of the data in premises. Then this data is given to our system as input which is on cloud platform. This input data is stored on Hadoop for further processing. Figure 3 below shows the data model of the database. Data models define how the logical structure of a database is modelled. Data Models are fundamental entities to introduce abstraction in a DBMS.

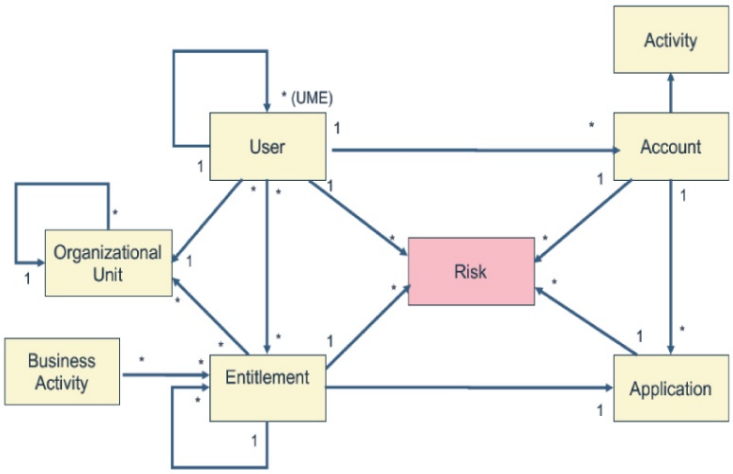

Fig 3: Data Model

IV. ANOMALY DETECTION

Anomaly detection can be done based on the type of anomaly. There are three types of anomaly, point anomaly, contextual anomaly and collective anomaly. For our system evaluation is based on the contextual anomaly because the outlier is present in the data specific to the particular context not specific to the global data.

A. Contextual Anomalies

Contextual anomalies are considered in applications where the dataset has a combination of contextual attributes and behavior attributes. These terms are sometimes referred to as environmental and indicator attributes [19]. There are generally four common ways in which we define contextual attributes, these are defined in Table 1. In this paper we are considering the contextual attribute of type profile, where we are creating the clusters of the data and then we are checking the anomaly specific to the cluster.

Table 1: Contextual Attribute types Definition

Term	Definition
Spatial	The records in the dataset include features which identify location information for the record. For example, a sensor reading may have spatial attributes for the city, province, and country the sensor is located in; it could also include finer-grained information about the sensors location within a building, such as floor, room, and building number [19].
Graphs	The records are related to other records as per some graph structure. The graph structure then defines a spatial neighbourhood whereby these relationships can be considered as contextual indicators [19].
Sequential	The records can be considered as a sequence within one another. That is, there is meaning in defining a set of records that are positioned one after another. For example, this is extremely prevalent in time-series data whereby the records are time stamped and can thus be positioned relative to each other based on time readings [19].
Profile	The records can be clustered within profiles that may not have explicit temporal or spatial contextualities. This is common in anomaly detection systems where a company defines profiles for their users; should a new record violate the existing user profile, that record is declared anomalous [19].

B. Contextual Anomaly Detection Algorithms

Contextual Anomaly is implemented by various algorithms which contain some sequential approach or machine learning approach. We are evaluating our system by implementing two algorithms which are K means clustering algorithm and Attribute value frequency algorithm

1. Hierarchical K-Means Clustering Algorithm-

In Hierarchical K-Means Clustering algorithm both content and contextual anomaly are detected. First the content anomaly is checked against whole data and then the data is divided onto the groups and the contextual anomaly is detected.

K-means clustering partitions the dataset into a set of clusters which minimizes the sum of squares distance between each data point. This can be done by iterating through the following steps [19]:

1. Randomly initiate K random clusters.
2. Partition the dataset into the K random clusters using below equation, replacing items into each cluster based on the smallest distance to the cluster.

$$min \sum_{i=1}^{k} \sum_{Xj} ||Xj - \mu i||$$

3. Re-calculate the centroids for each cluster.
4. Repeat Step 2 until Step 3 does not modify cluster centroids.

Advantages:

1. Clustering-based technique for the contextual anomaly detector has a relatively short testing time, and a longer training time.
2. For the contextual anomaly detector this is acceptable because training will happen online and testing will happen relatively infrequently.
3. This algorithm also exposes the parallelization.

Disadvantages

1. K-means clustering algorithm is computationally expensive and selection of k random cluster is itself is one disadvantage.
2. The anomaly detection assessments use the idea of data decrease by only evaluating the computationally more expensive contextual anomaly detector on a very small subset of the data. This is based on the assumption that an anomaly is a rare occurrence.

2. Attribute Value Frequency-

Attribute Value Frequency is the scalable and fast anomaly detection algorithm for categorical data. AVF algorithm scales linearly with the number of data points and attributes. AVF is relying on a single data scan. AVF is compared with a various representative outlier detection strategies that have not been considered against each other. Algorithmic steps for AVF are following:

1. Calculate the frequency of every possible value of each attribute
2. Average frequencies for each individual data point
3. Data points with lower scores are mostly considered as to be outliers because they will have infrequent values on average.

Advantages

1. Attribute Value Frequency Runs significantly faster than Greedy AVF needs one pass over data instead of *k.*

Disadvantages

1. Attribute Value Frequency algorithm is slower for larger datasets.
2. Higher dimensionalities result in increasingly larger item sets.

V. EXPERIMENTAL SETUP

All the Algorithms are implemented on Apache Spark 2.1.0 framework by using Scala 2.10. All the experiments are performed on the 3.2 GHz Intel Core Processor machine with 6 GB RAM.

Apache Zeppelin 0.7.0 is used to show interactive dashboard. Microsoft Redhat Linux 7.3 is used as an operating system.

Metrics of evaluation of the algorithms is based on the accuracy and the execution time. Resource utilization is increased due to the use of parallel processing and spark.

Dataset after Preprocessing is given the figure 4 which has the 7000 records and the 10 attributes.
Contextual attributes are User, Entitlements and Profile type.
The approximate results are

1. K- means clustering algorithm-
   After running the K means clustering algorithm on the above data set the partial result we had get are given in below table. The results are showed based on the Accuracy and Execution time we get. Algorithm we had used is given in the spark machine learning library.

Table 2: Result of k-means clustering algorithm

Accuracy	85.8%
Execution Time	2min 5sec

2. Attribute Value Frequency-
   This algorithm is implemented with the help of the spark machine learning library. Frequency, mean and standard deviation is calculated based on the spark library. The partial results that we had got by this algorithm are given below in the table number 3, which is in terms of the accuracy and the execution time.

Table 3: Result of Attribute Value Frequency algorithm

Accuracy	87.1%
Execution Time	1min 34sec

VI. CONCLUSION AND FUTURE WORK

Peer group analysis technique is used to analyze the data based on the data groups which are created based on some common attributes. In this research we are using the peer group analysis to identify the outlier present in the Identity and Access Data. Out of point, contextual and collective outlier; for our dataset we are considering the contextual anomaly. Identity and Access Data has the attribute of type profile, which can be used to create peer groups. User having the entitlements different from its peers is considered as outlier. Traditional way of doing peer group analysis is statistical way such as sorting method and regression method. In this research we have identified some data analysis algorithm such as k means clustering algorithm and attribute value frequency algorithm. Each algorithm has its own advantages and disadvantages which has listed in the paper. These two algorithms are implemented on the spark framework and the partial results are showed in the form of accuracy and the execution time. This analysis helps organization to manage the accesses inside the organization and improves the decision making process of the organization.

Future work of this research is to consider the contextual anomaly in the form of the graph attribute and then compared the results. Based on the outlier the risk score can be calculated which can be used to detect the risk insights the organization. Risk advisor can be implemented by using the machine learning classification algorithms.

VII. REFERENCES

[1] Ludwig Fuchs, Gnther Pernul,"Supporting Compliant and Secure User Handling - A Structured Approach for In-House Identity Management", 2007 IEEE Second International Conference on Availability, Reliability and Security, Germany.
[2] Sanizah ahmad, Habshah Midi ,"Outlier Detection in logistic regression and its application in medical data", 2012 IEEE Colloquium on Humanities, Science and Engineering Research (CHUSER 2012), December 3-4, 2012, Kota Kinabalu, Sabah, Malaysia
[3] M. I. Petrovskiy, "Outlier Detection Algorithms in Data Mining Systems", Programming and Computer Software, Vol. 29, No. 4, 2003, pp. 228237.
[4] Varun Chandola, Arindam Banerjee, "Anomaly Detection for Discrete Sequences: A Survey", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, May 2012
[5] Chen-Chia Chuang, Shun-Feng Su, "Robust Support Vector Regression Networks for Function Approximation with Outliers", IEEE Transactions on Neural Networks, Vol. 13, No. 6, November 2002
[6] David J. Weston,David J. Hand, Niall M. Adams, Christopher Whitrow, "Plastic card fraud detection using peer group analysis" ,Springer ADAC (2008), London 2:4562.
[7] Zakia Ferdousi and Akira Maeda, "Unsupervised Outlier Detection in Time Series Data" ,2006 IEEE Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), USA.
[8] Matthias Hummer, Michael Kunz, Michael Netter, Ludwig Fuchs and Gunther Pernul , "Advanced Identity and Access Policy Management using Contextual Data",2015 IEEE 10th International Conference on Availability, Reliability and Security, Germany.
[9] Ertem Osmanoglu, "Identity and Access management business performance through connected intelligence", Book, Elsevier 2013, Pages 177-190.
[10] Daisuke Mashima, Daisuke Mashima, "Using Identity Credential Usage Logs to Detect Anomalous Service Accesses", 5th ACM workshop on Digital identity management, Pages 73-80, Chicago, Illinois, USA, November 13 - 13, 2009.
[11] Robert Cowles, Robert Cowles, Robert Cowles, "Facilitating Scientific Collaborations by Delegating Identity Management", 2015 Workshop on Changing Landscapes in HPC Security, Portland, Oregon, USA June 16 - 16, 2015, Pages 15-19.
[12] Ion Petru, Catalin, Alexandru, Madalina Ecaterina, "Identity and Access Management – Risk Based Approach", 9th International Management Conference on Management and Innovation for Competitive Advantage, November 5th-6th, 2015, Bucharest, Romania.
[13] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, "Pregel: A System for Large-Scale Graph Processing", 2010 ACM SIGMOD International Conference on Management of data, Indianapolis, Indiana, USA June 06 - 10, 2010, Pages 135-146.
[14] Michael Kunz, Matthias Hummer, Ludwig Fuchs, Michael Netter and Gunther Pernul, "Analyzing Recent Trends in Enterprise Identity Management", 2014 IEEE 25th International Workshop on Database and Expert System Applications, Lisbon, portugal.
[15] Detlef Sturm, Detlef Sturm , "Permission Path Analysis Based on Access Intelligence", 18th ACM symposium on Access control models and technologies, Amsterdam, The Netherlands June 12 - 14, 2013, Pages 253-256.
[16] Yong Yu, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, Yuanshun Dai, and Geyong Min, "Identity-based Remote Data Integrity Checking with Perfect Data Privacy Preserving for Cloud Storage", 1556- 6013 ,2016 IEEE Transactions on Information Forensics and Security.
[17] William Aiken, "KaaSP: Keying as a Service Provider for Small and Medium Enterprises Using Untrusted Cloud Services", 9th International Conference on Ubiquitous Information Management and Communication, Article No. 20, Bali, Indonesia January 08 - 10, 2015.
[18] Top five Identity and Access Management tools, online available at-http://www.columninfosec.com/information-security/top-5-identity-access-management-tools.html
[19] M. A. Hayes, M. A. M. Capretz, "Contextual Anomaly Detection in Big Sensor Data", in Proc. of the 3rd Int. Congress on Big Data (IEEE BigData 2014), June 27-July 2, 2014, Anchorage, Alaska, USA.