ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

IMPROVING ETL/SQL EXECUTION THROUGH ANOMALY DETECTION

Prof. (Dr.) R. Kamatchi*[1] and Kunal Suri[2]
*[1]Prof. (Dr.) R. Kamatchi Amity University, Mumbai, India
[2]K.J. Somaiya Institute of Management Studies and Research , VidyaVihar, Mumbai-77, India

ABSTRACT

Data Warehouse is a business analyst's dream – a platform where data from multiple sources are collected. Different types of analysis are performed on the data obtained from the data warehouse. The process of loading data into a data warehouse is known as ETL (Extract, Transform, Load). It is a complex process that comprises of executing thousands of SQL queries. These queries lead to the creation of ETL execution trace. The potential of these concepts haven't been used to their full potential. Anomaly is one of the features of these data sets that haven't been completely utilized. The study and identification of Anomaly in the execution of SQL queries and ETL execution trace can help us increase the efficiency of the ETL processes by removing them. To accomplish this task, we tackle this problem in two stages:- In the first stage, we use Anomaly detection techniques on a rich collection of production queries. In the second stage, we apply the Anomaly detection technique on the execution logs. By following this process, we greatly reduce the domain of our detailed analysis. We also identify the clusters that have a genuine concern from the clusters that were created by a huge data store.

Categories and Subject Descriptors
Data Warehouse and repository, Data Mining, Clustering, Query formulation

General Terms
Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Theory

Keywords- Data Warehousing, Anomaly Detection, Clustering, Log Files, Query tuning, Tuning.

1. INTRODUCTION

RDBMS is a vital part of any OLTP/OLAP software. We would be focusing on the OLAP software due to the richness of transaction data and the complexity of the query processing involved. Even a small sized data warehouse contains the data up to 5 Terabytes. There are several OLTP providers such as Oracle, Microsoft, and Teradata. However, data warehouse is a niche market that is not penetrated by many companies. Some of the big players in this field are SAP, Oracle and Microsoft. Today, data warehouse has evolved from a traditional business intelligence platform to encompass operational data stores, analytics and performance management. The data warehouse is moving slowly towards a near-real-time analytics of data. In addition to these, there are several emerging trends as well such as columnar storage and in-memory OLTP software. Still, the demand for optimization techniques and performance remain the primary focus area of research in the companies creating data warehousing. Another factor to consider while measuring the performance of OLAP software is the workload. Most of the data warehouses support mixed workload. Both of these factors can be studied using Anomaly detection.[1][2]

Background study:

One of the main factors that influence the performance of the loading and retrieval of data is the database engine. For example the hash based data distribution results in "shared nothing" architecture. This architecture is very popular and used by companies like Teradata. This architecture focuses on rewriting of queries, materialized views and various kinds of partitioning and indexing strategies. Another popular architecture is the columnar architecture. In this architecture, the data base uses the high compression rate. This rate can be leveraged because the domain of the values is similar. The underlying hardware and software form the second part. The database designers who decide the partitioning and indexing strategies form the third part that influences the performance of loading and retrieval of data. There is also a fourth part which decides the query performance – adherence of the users (basic and advanced) to the best practices of the industry. [3][4]

ISSN: 2394-7659
IMPACT FACTOR- 2.789

International Journal of Engineering Researches and Management Studies

Here, we focus on the third and fourth factor in query performance management. To monitor the performance of the queries, we generally use the "white box" approach. In this approach, we use a set of processes such as checklists, tools and reviews. However, they are manual and hence error prone. Also, they tend to be very basic and hence are easily missed out.

In this paper we propose another method. All the execution related attributes of the SQL queries are stored in the system tables of the databases. In our approach, we view the queries as groups. We try to find out patterns, behavior of the queries and observe that which queries are outlying from the crowd. We call this approach as the "Black-box approach".

Once the data is loaded into the OLTP software, it needs a way by which it can be sent to the data warehouse. This is achieved through ETL (Extract, Transform, Load). ETL is one of the most important phases in the data warehousing. Companies spend billions of dollars every year to procure the best ETL tools. This is because about 70% of the effort and time is spent in extraction, cleaning, conforming, transforming and loading the data. The implementation of the ETL can be either hand coded or tool based. Some big companies involved in creating tools for ETL are Informatica, SQL Server Integration Services (SSIS), Pentaho, etc. This paper the focuses majorly on the tool based ETLs.

All of these tools generate a huge amount of execution traces (logs) by a well-defined system. Depending on the granularity of the tracing level, the size of the logs generated will vary. However, irrespective of the size of the logs, they can be used to generate a lot of insights about various dimensions of the data. The analysis of this data leads to an efficient information pipeline which is of huge importance to any organization. Despite the apparent importance of the data, there has been very less research on the methods of applying data mining methods on this data.[5][6]

The logs generated are generally quite large and represents two kinds of problems – 1. Deciding the most critical features of the ETL that should be focused upon. 2. After identification of the features, how do we extract their information from the ETL log in an automated manner? For the first problem, we surveyed a large number of ETL developers. For the second problem, we used a simple text parsing tool based on Python. We would also like to add that all the analysis of the done belongs to a particular ETL tool. However, since the features mentioned in this paper are quite generic, there should be no problem in extending this method to other tools as well. [7][8]

The organization of the paper is as follows:- Section 2 focuses on the prior work that has been done in this area. Section 3 introduces the basic concepts of Anomaly detection. Section 4 contains the summary of our work done in ETL/SQL execution and ETL execution traces. Section 5 provides a brief overview on the query optimizations and section 6 details on the experimental setup and results. Section 7 discusses the challenges, future course of action and conclusion.[9][10]

2. RELATED WORK

This paper contains work related to three domains – 1. Query Optimization techniques, 2. Representing the ETL Log file 3. Selection of the appropriate Anomaly techniques. There has been extensive research in the fields of high dimensional data using Principal Component Analysis (PCA), using the sensor data for online detection. In addition to these applications, Anomaly detection has also been used in the fields like fraud detection, network intrusion detection and medical health.

Query optimization is field that dates back to 1970s. Query optimization is a very extensively researched field where the current state-of –the-art is "Query Hint Framework" or optimization of the XML Queries.[11][12]

On the other hand, there has been no significant research on ETL log representation. There have been very less research on meta-model based formalism for the ETL processes and the taxonomy of the spectrum of the ETL activities. ETL has also been extended using UML. However, none of these resources touches upon the primary characteristics of an ETL Log.

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

The results of the executed queries are not analyzed in terms of various cost components, even though the query optimization takes place during and pre execution. We use various Anomaly detection techniques to calculate the cost of the query execution. We also use the analogy with the process mining to mine the logs for information. Even though the major focus of the process mining is to build the process model from the trace by finding the temporal and casual relationships between the tasks. Data mining techniques have also been applied on the logs of the intrusion detection and network anomaly detection. Even though there are applications of Anomaly detection on process logs, the focus is primarily on the structural pattern, the relationship among the activities and then finding the Anomaly. In this paper, we do not compare Anomaly detection algorithms. On the hand, we use them to analyze the ETL logs.

3. FOUNDATION

ANOMALY

The statistical intuition says that the normal data objects follow a "generating mechanism", such as some given statistical process. An abnormal object deviates from this generating mechanism. A few examples of these detections are fraud detection, medicine, public health, sports statistics, detecting measurement errors.

The data is usually multivariate. There is usually more than one generating mechanism/statistical process underlying the data. The anomalies represent a different class of objects, so there may be a large class of similar objects that are the anomalies. Consequently, a lot of models and approaches have evolved in the past years to exceed these assumptions and it is not easy to keep track with this evolution.[13][14]

It is often noticed that many clustering algorithms account for noise objects. So, we look for anomalies by applying one of the algorithms and retrieving the noise set. There are several classification approaches to this problem – Global versus local detection, labeling versus scoring detection and modelling properties. In addition to these approaches, we have proximity based approaches. Sample approaches are the most appropriate method. Some of the examples are Distance-based approaches and density-based approaches. Another approach is the Angle-based approach whose rational is to examine the spectrum of pairwise angles between a given point and all other points. The points having a spectrum featuring high fluctuation are anomalies. In the figure given below, N1 and N2 are normal regions where as O1 is an Anomaly.

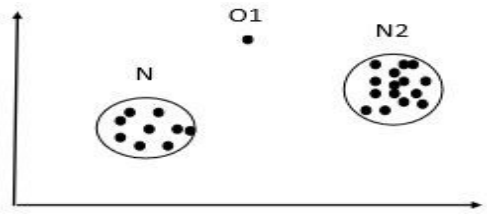


Fig. 1 Outlier in 2D Space

The Anomaly can be classified into two categories:- One where a single instance in an Anomaly and second where a sequence of observations is an Anomaly. We can use the spatial and temporal attributes to gain an additional context for the individual Anomaly. The input data can have any data type - binary, nominal, ordinal discrete and continuous. The output of the Anomaly detection task can be either a level (Anomaly or normal) or a score. The score is a preferred method because it gives an idea of the outlay. [15][16]

Most of the popular ways to detect an Anomaly are classification based or Nearest Neighbor based or clustering and statistics based. Nearest neighbor can be either distance based or density based. The statistics based techniques will

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

be either parametric or non-parametric. These methods are applied in several fields like fraud detection, intrusion detection, medical data, sports, novel topic detection and so on.

ETL TASKS

ETL is a process in data warehousing responsible for extracting data out of the source systems and putting it into a data warehouse. ETL includes the following tasks:-

1. Extracting the data from the source systems, data from different sources is converted into one consolidated data warehouse format which is ready for transformation processing.
2. Transforming the data may involve the following tasks :-
   - **a.** Applying business rules (so-called derivations e.g. calculating new measures and dimensions)
   - **b.** Cleaning (e.g mapping NULL to 0 or "Male" to "M" and "Female" to "F" and so on)
   - **c.** Filtering (e.g selecting only certain columns to load)
   - **d.** Splitting a column into multiple columns and vice versa.
   - **e.** Joining together data from multiple sources (eg. Lookup, merge)
   - **f.** Transposing rows and columns
   - **g.** Applying any kind of simple or complex data validation (eg. If the first 3 columns in a row are empty then reject the row from processing)
3. Loading the data into a data warehouse or data repository of other reporting applications

ETL Logs

ETL execution logs contain the information about the flow of data from the source system to the data warehouse and vice versa. This information has several components such as the source systems, mapping for the source systems to fields in the data warehouse, staging areas between the cubes and the source data, transformations, look-ups and joiner transformations, run-times and so on. All of these factors are important if we have to know the bottleneck in the execution of an ETL. However, we cannot expect to include all of these factors in the data mining process. So we try to include the ones that are favored by the community of ETL developers.[18..20]

4. FEATURE SELECTION

4.1 ETL OPTIMIZATION

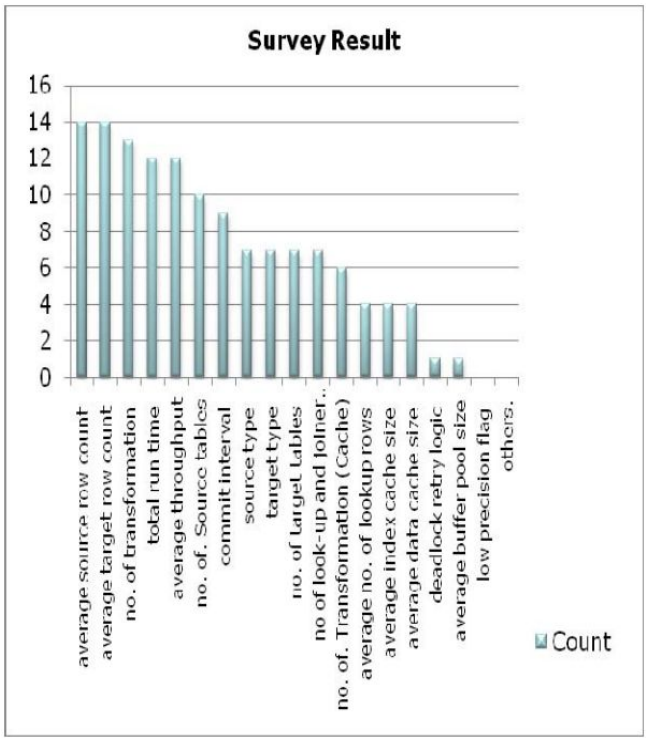
An ETL execution includes many features such as source systems, staging areas, transformations, look-ups, joins, target systems and the type of networks over which the data is transferred. In order to perform a feature selection, we conducted a survey among the ETL developers and asked them about the features that they think should be included in the data mining process.

We created an online intranet form which contained all the features of an ETL process and asked the 20 ETL developers to respond to the survey. Based on their responses, we shortlisted the following factors:- number of source rows, number of target rows, transformations, total run time, average throughput and number of source tables (as seen below).

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies



Most of the fields in the above are derived fields. So, we break them down into primary fields. Now, we have the following atomic features:-

1. Number of source rows, 2. Number of target rows, 3. Total runtime, 4. Number of Transformations

The above fields are also common in other ETL tools because they are very generic head.

4.2 SQL OPTIMIZATION

The performance of any query depends on four factors- 1. Database engine 2. The hardware and software being employed 3.database designers 4. Practices followed by basic and advanced users.

The architecture of a database engine and the hardware/software being used play a pivotal role but they still aren't a critical factor in the query performance. This is because the database can always be enhanced by adding more memory and processors. However, the database design and the query formulation by the user play a bigger role in the query performance. This is because they are dependent on various standards that are always evolving and manual in nature. This means that they are prone to errors. [21][22]

The last two factors are the ones with greatest potential for improvement. We try to identify the queries with resource heavy behavior by using detection techniques.

5. EXPERIMENTS AND RESULTS

5.1 ETL OPTIMIZATION

We selected more than 500 logs from a server. However, we observe that if we include logs from different servers which do not have similar hardware/software configuration, we run into a risk of identifying several legit records as anomalies. This happens because the logs generated by different servers have different mechanisms of generation.

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

We create a simple VBA script to compare the strings in the server logs and extract the four chosen parameters. This script creates a CSV file which contains the value of the four parameters for every log file contained in a folder (as can be seen below).

	A	B	C	D	E
1	Name	Source Row	Target Row	Time	No .of Transformation
2	X1.log	159833	159833	168	2
3	X2.log	4719112	4719112	9278	2
4	X3.log	16178	16178	1423	2
5	X4.log	20715	20715	338	2
6	X5.log	494	494	82	2
7	X6.log	160	160	35	2

From the above table, it is evident that the number of transformation that are not changed by a big margin. So, even though we have included it initially, we do not use it in our clustering algorithm. The code snippet for the number of rows can be seen as:-

```
If InStr(strContent, strPat2) > 0 And flagsrc = 0 Then
  strRow = Mid(strContent, InStr(1, strContent, "[") + 1,
InStr(1, strContent, "]") - InStr(1, strContent, "[") - 1)
  strWrite = strWrite + "," + strRow
  'tsw.WriteLine (strWrite)
  flagsrc = 1
End If
```

It can be clearly seen that our code is hard coded with the underlying pattern of the log. We did this because we have selected the log from a single server and hence do not have the risk for heterogeneous logs.

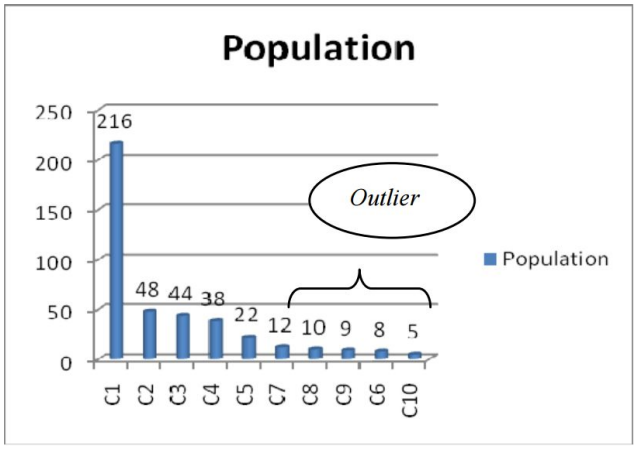
AS far as records with corrupt values are concerned, there were several records which had no source records or no target records or no time entries. We removed these entries to have a consistent data set. Once we cleaned up our data, we were left with 350+ logs which is a good number of logs to run our clustering algorithm. We used the scalable Expectation Maximization algorithm with default values of the parameters. The underlying assumption of this algorithm is that every cluster follows a normal distribution.

The clusters are distributed as shown below:-

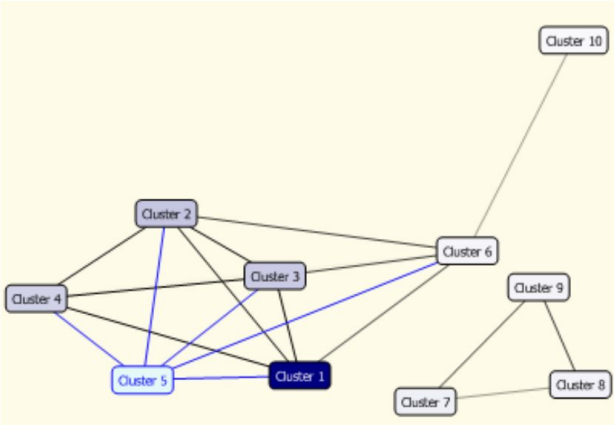
ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies



The clusters having the least population are the anomalies and hence are marked. The cluster viewer give us the following visualization of these clusters:-



The similarity between different clusters can be visualized by the weight of the lines connecting them.

In this experiment, we identify a cluster as an “” if it has less than five percent records of the overall population.

5.2: ANALYSIS OF THE RESULTS

We see that there are 5 members in the 10th cluster. Out of these 5 members, 3 members face serious issues with query where as the other 2 are not affected much. When inspected in detail, it was found that they had issues because of faulty method of writing queries. 6th cluster has 8 members out of which only 4 had major issues and the other 4 had issues with connection types used for data load. This was because a generic relational loader which was used instead of a customized external loader. In the above figure, clusters 7, 8, 9 have population of less than 20 members. This makes them anomalies as well. However, they become anomalies because they had a large number of source and target rows. It is also evident that these clusters are similar because the weight of the lines joining them is also similar.

From the above observations we can see that this whole method helped us in bringing down the number of logs under investigation from more that 500 to 44. Out of the 5 identified anomalies, 2 clusters are outliers because of the

ISSN: 2394-7659
IMPACT FACTOR- 2.789

International Journal of Engineering Researches and Management Studies

connection between source system and target system. The rest 3 clusters are anomalies due to large number of source and target rows.

From the above analysis, we see that even though the clusters 7,8,9 are anomalies, they do not pose any risk in query execution. It is because the number of rows is larger than any other source/target table. This is a normal thing to happen and cannot be treated as query performance degradation. So, we need to implement an automated system that gives us the ability to monitor these queries at regular intervals. The main highlight of this experiment was the reduction of the clusters to be studied from 500 to 44.

5.2 QUERY OPTIMIZATION

To study the query optimization, we use a collection of more than 25000 database queries in a production environment.

The features for clustering have been selected based on the query cost model. These features are: - SQLtext of the query, CPU Cost, IO Cost, Memory need for temporary tables that store the intermediate results and the number of records affected. These factors are some of the major components in our query cost model. All the major databases contain system tables that include the execution statistics of the queries. This ensures that retrieving the statistical information from these tables is possible across all major vendors.[25][27][28]

To perform the data mining, we use an ensemble of several unsupervised learning methods. These methods are used because the data being used is unlabeled. The success of this ensemble is determined by the diversity of the detectors. We use the intersection of false positive and false negative when the cost of false positive is too high. On the other hand, we use the union of false positive and false negative if the cost of false negative is very high.

5.2.1 DISTANCE BASED APPROACH

To start with this approach, we normalize all the data points. This is done because all the five attributes being used have different weightage and hence they have different importance as well. Median and inter-quartile range are used for normalization because of the higher breakdown point than the mean and the standard deviation. The algorithm to achieve to perform the normalization is:-

1. The median and quartiles for each of the features is calculated.
2. The feature value for all the data points are normalized.
3. Euclidean distance is calculated for each of these points.
4. All the distances are sorted in the descending order.
5. The top r% of these values are selected.

Mahalanobis distance can be used to improve the factor independence among different attributes. It is very difficult to calculate the precise value of r. The summary of the distribution can be found below:-

Distance Band	No. of Observation
1-50	26141
51-100	265
101-150	53
151-200	33
201-250	8
250+	8
Grand Total	26508

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

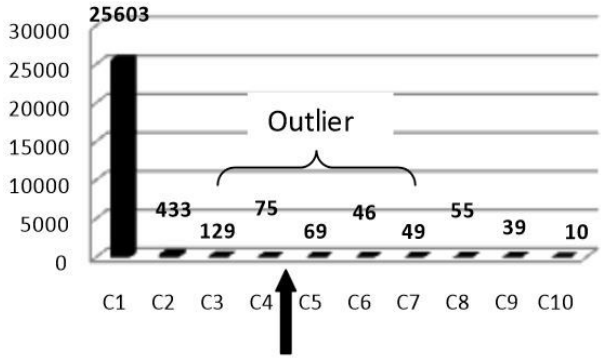
International Journal of Engineering Researches and Management Studies

Based on the data in the above table, we can draw two conclusions – 1. Large number of records during execution can also make a query an anomaly. 2. If a query is in the neighborhood of the origin, we cannot say that query conforms to the best practices. However, we can safely ignore these queries because their effect is not very high.

Several times it happens that a data point becomes an anomaly when another point is removed from the data set. This phenomenon is called masking because the second data point was masking the first data point. On the other hand, sometimes it also happens that a data point becomes an anomaly only when another data point is present. This phenomenon is called swamping. In the next approach, we will remove the data point with larger anomaly so that the masked data points can come forward.

5.2.2 CLUSTERING BASED APPROACH

The main idea behind performing clustering is that it requires unlabeled data. So, when it creates clusters of data, we can find the anomaly by looking at the number of members in a cluster. The cluster containing minimal numbers of elements are taken as anomalies. To perform the clustering, we have several algorithms such as K-means, Nearest Neighbor, and PAM and so on. In our approach we use the K-means algorithm where K=10. The following figure shows the bar chart with the anomalies:-



The arrow serves as a requirement driven slider.

5.2.3 AVERAGE DISTANCE BASED APPROACH

The next approach we take is the calculation of distance between all the data points. After calculating the distance between all the points, we calculate the average of the all the distances. The data point lying farther than the average distance is declared as the anomaly. The algorithm to cluster data points based on this approach is:-

1. Normalize the values of all the features.
2. Distance between all the points are calculated.
3. The average distance with all the neighbors is calculated.
4. Sort the distances in ascending order and pick the top n% values.

The distribution for every distance range can be seen below:-

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

Average Distance Band	No. of Observation
0-50	26309
51-100	144
101-150	40
151-200	9
200+	6
0-50	26309
Grand Total	26508

5.2.4 DENSITY BASED APPROACH

Many times it is possible that a cluster of densely packed data points are located away from the origin. In that case, the distance based approach would declare those data points as anomaly. This approach tries to solve this problem by looking for regions with dense or sparse population of the data points. Anomaly is declared for regions with sparse density. The identification of the dense/sparse region is done by using Local Outlier Factor (LOF). This method is optimized from identifying global and local anomaly by identifying conditions where clusters of different density exist. To execute this algorithm, we take K = 10.

Step 1:- A distance D is found for every data point such that there are at least 10 neighbors within that distance.

Step 2:- Find all the neighbors for each data point such that their distance from the point is less than D. Cardinality of such a set can be more than 10 as well.

Step 3:- After this, Local Reachability Density for each data point is calculated. This is the inverse of average reachability distance with its neighbors.

Step 4:- The ratio between the neighbors' reachability index with its own is calculated as the LOF.

The results are analysed after breaking this in multiple tables. The value of K was 10. At this point in time, a query ID 26508 was not detected as an anomaly because there are many data points in its neighborhood which have similar local reachability density. Hence, we can see that K plays a huge role in this algorithm.

6. CONCLUSIONS

A very different approach has been taken in this paper for query and ETL optimization. Standard techniques are applied while executing them both, the real cost lies in their design and the practices followed by the user. We propose an ensemble of data mining algorithms to identify the SQL queries and ETL process which have the highest performance load.

As far as the analysis of the ETL logs are concerned, the primary challenge was creating an algorithm which can be applied to any ETL tool. The log file considered for our study is a text file. However, in real life most of the log files are created in XML format which conform to a particular schema specifications. This specification can help us in avoiding the hard coding of data in the function. This would also enable us to run these jobs periodically and hence monitor the ETL execution at regular intervals.

REFERENCES

[1] Kimbal, Ralph and Caserata, Joe,"The Datawarehouses ETL Tool Kit." 1. s.l. : Wiley. p. 528. 978-0764567575.

[2] J. Hodge (vicky@cs.york.ac.uk) and Jim Austin , "A Survey of Outlier Detection Methodologies" Victoria Artificial Intelligence Review, 2004.

ISSN: 2394-7659
IMPACT FACTOR- 2.789

IJERMS

International Journal of Engineering Researches and Management Studies

[3] *Vassiliadis, Panos, Simitsis, Alkis and Skiadopoulos, Spiros. s.l. "Conceptual modeling for ETL processes." ACM, 2002. DOLAP '02 Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP. 1-58113-590-4.*

[4] *ARINDAM BANERJEE, VIPIN KUMAR ,"Outlier Detection: A Survey." ACM Computing Surveys.*

[5] *Vassiliadis, Panos, Simitsis, Alkis and Baikousi, Eftychia," A taxonomy of ETL activities" Hong Kong, China : ACM, 2009. DOLAP '09 Proceeding of the ACM twelfth international workshop on Data warehousing and OLAP. 978-1-60558-801-8.*

[6] *Trujillo, Juan and Luj an-Mora, Sergio. "A UML Based Approach for Modeling ETL Processes in Data Warehouses."s.l. : Springer, 2003.*

[7] *Yu, Barnett, V. and Lewis, T.: 1994, "Outliers in Statistical Data." John Wiley & Sons.,3 edition..*

[8] *Anh Duong Hoang Thi and Binh Thanh Nguyen, "A Semantic Approach towards CWM-based ETL Processes." Graz, Austria : s.n.,2008. Proceedings of I-SEMANTICS '08.*

[9] *W.M.P. van der Aalst and A.J.M.M, " Process Mining: A Research Agenda", Computers in Industry archive. Volume 53 , Issue 3 (April 2004)*

[10] *Wenke Lee and Salvatore J. Stolfo ,"Learning Patterns from Unix Process Execution Traces for Intrusion Detection" AAAI Workshop on AI Approaches to Fraud Detection, 1997*

[11] *Hawkins D,"Identification of Outliers", Chapman and Hall, 1980*

[12] *Han, Kamber, "Data Mining Concepts & Techniques"*

[13] *Mehmed Kantardzic, "Data Mining—Concepts, Models, Methods, and Algorithms "*

[14] *Margaret.H. Dunham, S.Sridhar, "Data Mining Introductory & Advanced topics "*

[15] *Moh'd Belal Al- Zoubi," An Effective Clustering-Based Approach for Outlier Detection", European Journal of Scientific Research, ISSN 1450-216X Vol.28 No.2 (2009), pp.310-316*

[16] *Kenji Yamanishi et al, "Dynamic Syslog Mining for Network Failure Monitoring", KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*

[17] *Lucantonio Ghionna et. al ,"Outlier detection techniques for process mining applications," ISMIS'08: Proceedings of the 17th international conference on Foundations of intelligent systems"*

[18] *Donald Feinber g, M ark A. Bey er , Gartner RAS Core Research Note G00209623, 28 January 2011*

[19] *V Chando la, A B anerjee, B Kuman, Anomaly Detection: A Survey, ACM Comp uting Survey 2009*

[20] *P. Filzmoser, R. M aronna, and M. Werner, "Outlier identification in high dimensions", Computational Statistics and Data Analy sis,Volume 52, Issue 3, 1 January 2008, Pages 1694-1711*

[21] *S. Subramanian et. al, "Online outlier detection in sensor data using non-parametric models", VLDB '06 Proceedings of the 32nd international conference on Very large data bases*

[22] *Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler, "A Comprehensive Survey of Data M ining-based Fraud Detection Research", arXiv:1009.6119v1*

ISSN: 2394-7659
IMPACT FACTOR- 2.789

International Journal of Engineering Researches and Management Studies

[23] Wenke Lee and Salvatore J. Stolfo ,"Learning Patterns from Unix Process Execution Traces for Intrusion Detection" AAAI Workshop on AI Approaches to Fraud Detection, 1997

[24] Surajit Chaudhuri , "An overview of query optimization in relational systems" , PODS '98 Proceedings of the seventeenthACMSIGACT-SIGM OD-SIGART symposium on Principles of database system

[25] Matthias Jarke and Jurgen Koch, "Query Optimization in Database Systems" , ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 16 Issue 2, June 1984

[26] Nicolas Bruno, Surajit Chaudhuri and Ravi Ramamurthy, "Power Hints for Query Optimization", Data Engineering, 2009. ICDE '09. IEEE 25th International Conference

[27] Hoang Vu Nguyen, Hock Hee Ang and Vivekanand Gopalkrishnan, "M ining Outliers with Ensemble of Heterogeneous Detectors on Random Subspaces", Database Systems for Advanced Applications, 2010.

[28] Markus M . Breunig et. al , "LOF: identifying density -based local outliers", SIGM OD '00 Proceedings of the 2000 ACM SIGM OD international conference on Management of data

[29] Rui Xu;Wunsch, D., II, "Survey of clustering algorithms", IEEE Transactions on Neural Networks

[30] Ujjwal Das Gupta, Vinay Menon, Uday Babbar., "Detecting the number of clusters during Expectation- M aximization clustering using Information Criterion", 2010 Second International Conference on Machine Learning and Computing

[31] Irad Ben-Gal, "Outlier Detection", Data Mining and Knowledge Discovery Handbook, 2010

[32] Lucantonio Ghionna et. al ,"Outlier detection techniques for process mining applications," ISM IS'08: Proceedings of the 17th international conference on Foundations of intelligent systems"

[33] Kimbal, Ralph and Caserata, Joe,"The Datawarehouses ETL Tool Kit." 1. s.l. : Wiley. p. 528. 978-0764567575.

[34] Yuqing Wu, Jignesh M . Patel and H. V. Jagadish, "Structural Join Order Selection for XM L Query Optimization", 19th International Conference on Data Engineering (ICDE'03)