RDBMS is the heart for both OLTP and OLAP types of applications. For both types of applications thousands of queries expressed in terms of SQL are executed on daily basis. All the commercial DBMS engines capture various attributes in system tables about these executed queries. These queries need to conform to best practices and need to be tuned to ensure optimal performance. While we use checklists, often tools to enforce the same, a black
Outlier Detection Techniques for SQL and ETL Tuning

box technique on the queries for profiling, outlier detection is not employed for a summary level understanding. This is the motivation of the paper, as this not only points out to inefficiencies built in the system, but also has the potential to point evolving best practices and inappropriate usage. Certainly this can reduce latency in information flow and optimal utilization of hardware and software capacity. In this paper we start with formulating the problem. We explore four outlier detection techniques. We apply these techniques over rich corpora of production queries and analyze the results. We also explore benefit of an ensemble approach. We conclude with future courses of action. The same philosophy we have used for optimization of extraction, transform, load (ETL) jobs in one of our previous work. We give a brief introduction of the same in section four.

Reference

- Donald Feinberg, Mark A. Beyer, Gartner RAS Core Research Note G00209623, 28 January 2011
- J. Hodge (vicky@cs.york.ac.uk)* and Jim Austin, “A Survey of Outlier Detection Methodologies” Victoria Artificial Intelligence Review, 2004
- V Chandola, A Banerjee, B Kuman, Anomaly Detection: A Survey, ACM Computing Survey 2009
- P. Filzmoser, R. Maronna, and M. Werner, “Outlier identification in high dimensions”, Computational Statistics and Data Analysis, Volume 52, Issue 3, 1 January 2008, Pages 1694-1711
- S. Subramaniam 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
- Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler, “A Comprehensive Survey of Data Mining-based Fraud Detection Research”, arXiv:1009.6119v1
- 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
- Surajit Chaudhuri, “An overview of query optimization in relational systems”, PODS ’98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database system
- Matthias Jarke and Jurgen Koch, “Query Optimization in Database Systems”, ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 16 Issue 2, June 1984
- Nicolas Bruno, Surajit Chaudhuri and Ravi Ramamurthy, “Power Hints for Query Optimization”, Data Engineering, 2009. ICDE ’09. IEEE 25th International Conference
- Yuqing Wu, Jignesh M. Patel and H. V. Jagadish, “Structural Join Order Selection for XML Query Optimization”, 19th International Conference on Data Engineering (ICDE’03)
- Samiran Ghosh, Saptarsi Goswami, Amlan Chakrabarti, “Outlier detection from ETL Execution trace”, 2011 International Conference on Network and Computer Science (ICNCS 2011)
- Barnett, V. and Lewis, T.: 1994, “Outliers in Statistical Data.” John Wiley and Sons.,3 edition.
- Hawkins D, "Identification of Outliers", Chapman and Hall, 1980
- Kimbal, Ralph and Caserata, Joe,"The Datawarehouses ETL Tool Kit." 1. s.l. : Wiley. p. 528. 978-0764567575.
- Lucantonio Ghionna et. al ,"Outlier detection techniques for process mining applications,"
ISMIS’08: Proceedings of the 17th international conference on Foundations of intelligent systems
   - Irad Ben-Gal, “Outlier Detection”, Data Mining and Knowledge Discovery Handbook, 2010
   - Ujjwal Das Gupta, Vinay Menon, Uday Babbar., “Detecting the number of clusters during Expectation-Maximization clustering using Information Criterion”, 2010 Second International Conference on Machine Learning and Computing
   - Rui Xu; Wunsch, D., II, “Survey of clustering algorithms”, IEEE Transactions on Neural Networks
   - Markus M. Breunig et. al , “LOF: identifying density-based local outliers”, SIGMOD ’00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data
   - Hoang Vu Nguyen, Hock Hee Ang and Vivekanand Gopalkrishnan, “Mining Outliers with Ensemble of Heterogeneous Detectors on Random Subspaces”, Database Systems for Advanced Applications, 2010.

Index Terms
   Computer Science
   Database Systems

Key words
   Outlier Detection
   ETL Tuning
   Query Tuning
