Data Analytics for Frequently used Data items

[1] Srinivash R C
[1] Student, Department of Information Technology,
Sathyabama University
Chennai, India

[2] Justin Samuel
[2] Professor, Department of Information Technology,
Sathyabama University
Chennai, India

Abstract*—Many Business worth billions and trillions of dollars have gone waste in the past due to lack of proper security techniques used for Analyzing frequently used data sets resulting into potential thread to the firms and customers privacy data. The organizations are in dire need of getting the frequently used data analytics so that they can get the insight of the data and better security plan and get maximum benefit from that. Organizations are providing enormous software solutions to solve real world problems. In order to keep the business alive and excellence in the upcoming era, a deep understanding of change in underlying technology trends (cloud, big data and etc.,), Analytics and change in business dynamics. Frequently used data is the key to the success of new and existing Business strategy. The ability to capturing the frequently used data quickly for system critical application using right data analytics and protecting them with proper security will save organizations fame, time and money. In this case study will discuss about DFDM. Data Analytics Frequently used Data Management (DFDM) is a solution that helps to capture the accurate frequently used data and protect the data for any applications.*

***Keywords*—** *Data Analytics Frequently used Data Management (DFDM);*

I. INTRODUCTION

At the beginning of a database life cycle the designer is obliged to make assumptions about the data quantity and the way the data will be used. He must also foresee how the data will evolve. The design of the database is based on such assumptions [1]. The same assumptions are carried over into a hand coded database schema in which the tables are defined and linked to one another by foreign keys and indexes. If the assumptions prove to be right [2], the database can evolve with no problems. If not, the database structure will soon be obsolete and the form will no longer fit to the content. It will also become increasingly difficult to adapt [6]. If the data volume grows much more than was expected (example: Big Data) and the frequently used data usage turns out to be quite different from what was originally foreseen, it becomes necessary to reengineer the existing static data analytics stuff over the database [3].

II. CHALLENGE AND SOLUTION

. As new applications came along, the structure was altered or extended to accommodate them. These structural changes were often made in an adhoc manner, so that the complexity of the structure grew and the quality sank, much like software systems according to the laws of software evolution. In addition to this structural deterioration, the contents of the database became corrupted by erroneous programs storing incorrect values and deleting essential records. With large databases it is difficult to recognize such quality erosion, but over time it spreads like a cancerous infection causing more and more system failures, that leads to the frequently used data analytics ineffective. Thus not only the data analytics suffers but also frequently used data quality suffers from erosion. If the data volume grows much more than was expected and the data usage turns out to be quite different from what was originally foreseen, it becomes necessary to reengineer the database and data analytics dynamically. The Data Analytics Frequently used Data Management (DFDM) framework can be well integrated with the Dynamic Data Analytics activities performed across comprehensive application life cycles, thus yielding optimization, efficiency, Quality and security of existing tools, resources and processes.

This paper also attempts to focus on:

1. How to map Skill set and cross utilizes Business Intelligence & DFDM groups and resources?

2. What are the benefits of starting DFDM activities at the early stages of testing life cycle validation?

3. How structured DFDM helps reduce security threat and costs?

4. DFDM Maturity and its impact on Business Dynamics

III. ANALYTICS

Analytics is the study of data via various quantitative methods such as statistics, simulation, optimization along with descriptive and predictive data mining to yield insights which are unlikely to be discovered using the usual methodologies of business intelligence (BI) (for example query and reporting).

Analytics can be broadly categorized into 4 types:

- Descriptive Analytics
- Diagnostic Analytics

- Predictive Analytics
- Prescriptive Analytics

A. *Descriptive Analytics*

Descriptive Analytics is used to answer the "What" part of the business, i.e. "What happened and what is happening?" This uses various data mining and data aggregation techniques to provide the insight about the past and the present business situation. Some of the common examples are the reports from which one can get the insights of the company's production, operation, inventory, etc.

B. *Diagnostic Analytics*

Diagnostics Analytics is used to answer the "Why" part of the business, i.e. "Why something happened in the business?" One cannot fix something is he does not know why that thing happened. So, one can use the diagnostic analytics to drill down to the lowest level to get the insight of why that thing happened. Various visualizations can be used to uncover the correlations and patterns and provide the reasons for why the revenues of the company are down or why the sales are high, etc. One can spot the essential factors which directly or indirectly affect the business.

C. *Predictive Analytics*

Predictive Analytics is used to predict the future, i.e. "What is expected to happen next?" This is studying the historical and real-time data, identifying the patterns and hence predict the future. One can use predictive analytics to identify the upcoming risks and opportunities and hence help grow the business.

D. *Prescriptive Analytics*

Prescriptive Analytics helps to advice the business users on the possible outcomes, i.e. help users "prescribe" various possible actions and further guide them towards a solution. On top of "what would happen" it also provides information on "why it would happen" and recommendations to what should be done next.

IV. Analytic Techniques

Various Data Mining techniques are used to perform analytics on the set of data and discover the hidden information. Data mining is the combination of statistical Analysis and Artificial Intelligence to discover the "hidden" information from the data. Once the hidden information is revealed important business decisions can be taken. Data mining can be broadly categorized into 3 categories:

- Classification/Clustering
- Association Rules
- Sequence Analysis.

A. *Clustering*

Clustering refers to combining or grouping similar objects together. The data set is analysed and grouped on the basis of certain rules to categorize upcoming information. The significant problem in grouping is to determine the rules which help partition the data into categories.

- Nonexclusive vs. Exclusive
- Extrinsic vs. Intrinsic
- Hierarchical vs. Partitioned

B. *Association Rules*

A set of various if-then statements used to discover the hidden relationship between unrelated datasets, relational databases, etc. are known as Association rules. These set up relationships between the objects which are commonly used together. For e.g. if a person buys a bread than he would also buy a butter, etc. The association rules follow basic confidence-support criteria, i.e. the association rules usually satisfy the user supplied minimum support and confidence at the same time.

C. *Apriori Algorithm*

Level wise search strategy is used in this algorithm, in which k item sets are used to explore k+1 item sets. This process continues as long as the datasets appear often in the database. The method used to implement Apriori algorithm is BFS (Breadth First Search) and the structure used is hash tree type.

D. *Fast Distributed Mining (FDM)*

It is the distributed data mining algorithm which follows the same principle as the Apriori algorithm. However, in this technique an association between the global data sets and the local data sets is determined which helps generate less number of candidate sets which further reduces the messages to be passed.

E. *Sequence Analysis*

The process of extracting various sequential patterns with the support of these patterns exceeding the minimal threshold support is known as Sequential pattern mining. It helps to extract those sequences which reflect the most frequent behavior of the sequence database. The data being considered for extracting the sequence is the one that appears in separate transactions.

F. *Apriori-all vs Apriori-some*

Apriori-all and apriori-some algorithm uses the concept of count-all/some algorithm in which either a person include maximal and non-maximal sequences or only the maximal sequence. Apriori-all is count-all algorithm in which both the maximal and non-maximal sequence are included, while apriori-some is a count-some algorithm in which only the maximal sequences are considered. However, one need to

consider that not too many longer sequences are considered which don't have minimal support.

G. *The k-means algorithm*

This is an iterative algorithm to partition the given data set into k (user specified) clusters. Initially the k points are chosen as the centroids from the input data set having n data points. Then the below two steps iterates till convergence:

- The entire data set is partitioned into n clusters by assigning each data point to its closest centroid. The closeness of the data point is determined by the Euclidean distance between the data point and the centroid.
- Each centroid is relocated to the center of all the data points of that cluster.

The algorithm converges once no relocation of centroid occurs.

H. *PageRank Algorithm*

It is a search ranking algorithm which utilizes the hyperlinks on web. This algorithm uses the link structure of the web to determine the overall page quality. The rank prestige is used to determine the pagerank algorithm:

- Depending upon the number of in links to a page the prestige is determined. The more the in links, more is the prestige value of that page p.
- The pages pointing to page p have their own prestige value also. So, a page with higher prestige vale pointing to page p is more important than the page with low prestige value pointing to page p.

I. *DBSCAN Algorithm*

DBSCAN is a density based clustering algorithm which help separate regions of low density from another regions of high density. Below is the algorithmic details for implementing DBSCAN:

- Mark the data points as core, border and noise points.
- Once the points are identified, eliminate the points which are noise points.
- Mark the edge between all the core points which are within the specified distance.
- Segregate all the groups of core points into separate clusters.
- Each border point is assigned to one of the clusters of the core points

V. Business Data Analysis

A simple Business data that is passed in business operation consists of several data Items.

1. Add Process Data Items One by One
   - Data type
   - Data Item name
   - Display layout
2. Select Data Reading Level in each Step
   - Editable
   - Display Only
   - Hidden

3) Deleting Process Data Items

- Deleting Data Items before the release of process Model
- Deleting Data Items after releasing the process Model

4) Understanding Distinctive Data Types

- Table type
- File type
- Discussion type
- User type
- Organization type
- Guide Panel type

For example, in the case of "Leave Application flow", it consists of Frequently used Data Items such as

- Absence Start date,
- Absence End date,
- Reason for Leave,
- Superior's Comment, etc.,

Not only common Data types such as [String type] and [Numeric type], but also [File type] and [Discussion type] are available to be defined.

Business Problem:

The loss of frequently used data quality

Failure to foresee data evolution

Lack of data independence

Typical rule violations

Content complexity

View complexity

Access complexity

Relational complexity

Structural complexity

Storage complexity

A dynamic task scheduler cum auditing tool to check such rules, volume and to report their violations dynamically will improve the dynamic selection of data analytics accordingly.

VI. Discussion and Conclusion

Data Analytics Frequently used Data Management (DFDM) framework can be well integrated with the Dynamic Data Analytics activities performed across comprehensive application life cycles, thus yielding optimization, efficiency, Quality and security of existing tools, resources and processes. Also it Leverage the investment and tools in Data mining for DFDM maturity and its impact on Business implementation. This approach has the potential of significantly saving the manual effort, data collection from multiple sources, synchronization, distribution, etc. can be a challenge. DFDM will provide a real-time tracking of the frequently used data request status and any changes made to it. It facilitates effective daily status reporting including generation of daily analytics selection and frequently used data reports for various stakeholders. Security and customization can be enabled using the application admin functionality, i.e. the type of fields, security access to different stakeholders, etc.

A. *Authors and Affiliations*

1) Srinivash R C, M.Tech student, at Department of Information technology, Sathyabama University, Chennai. He has 6 + Years of IT experience. His current research include Data warehouse, Data mining, Business Intelligence, Big Data and Cloud Computing.

References

[1] C. Dwork, "Differential privacy," in ICALP, 2006.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Base Syst, 2002.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in ICDE, 2006.

[4] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in VLDB, 1994.

[5] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in SIGMOD, 2000.

[6] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," in VLDB, 2012.

[7] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in KDD, 2002.

[8] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," TKDE, 2004.

[9] W.K.Wong,D.W.Cheung,E.Hung,B.Kao,and N.Mamoulis, "Security in outsourcing of association rule mining," in VLDB, 2007.

[10] W.K.Wong,D.W.Cheung,E.Hung,B.Kao,andN.Mamoulis, "An audit environment for outsourcing of frequent itemset mining," in VLDB, 2009.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in KDD, 2002.

[12] Maurizio Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi, "Anonymity preserving pattern discovery," VLDB Journal, 2008.

[13] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in KDD, 2010.

[14] N. Li, W. Qardaji, D. Su, and J. Cao, "Privbasis: frequent itemset mining with differential privacy," in VLDB, 2012.

[15] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in FOCS, 2007.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in TCC, 2006.

[17] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," in VLDB, 2011.

[18] X. Zhang, X. Meng, and R. Chen, "Differentially private setvalued data release against incremental updates," in DASFAA, 2013.

[19] L. Bonomi and L. Xiong, "A two-phase algorithm for mining sequential patterns with differential privacy," in CIKM, 2013.

[20] E. Shen and T. Yu, "Mining frequent graph patterns with differential privacy," in KDD, 2013.

[21] R. Chen, B. C. M. Fung, and B. C. Desai, "Differentially private transit data publication: A case study on the montreal transportation system," in KDD, 2012.

[22] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in CCS, 2012.

[23] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," SIAM Journal on Computing, 2012.

[24] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," SIGKDD Explorations, 2004.

[25] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Statistical Mechanics: Theory and Experiment, 2008.

[26] N. Guttmann-Beck and R. Hassin, "Approximation algorithms for minimum sum p-clustering," Discrete Applied Mathematics, 1998

[27] Parag Deoskar, Divakar Singh, Anju Singh " An Efficient Support Based on Ant Colony Optimization Technique for Lung Cancer Data " ,International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 2, Issue 9,September 2014.

[28] Vikram Garg ,Anju Singh , Divakar Singh "A Hybrid Algorithm for Association Rule Hiding using Representative Rule", International Journal of Computer Application , Vol .97, No. 9 , July 2014.

[29] T. Karthikeyan1 and N. Ravikumar, "A Survey on Association Rule Mining"International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014