Scaling Up for High Dimensional Data in Data Stores and Streams

G. V. Sam Kumar
Assistant Professor,
Dept. of Info. Technology,
Sathyabama University

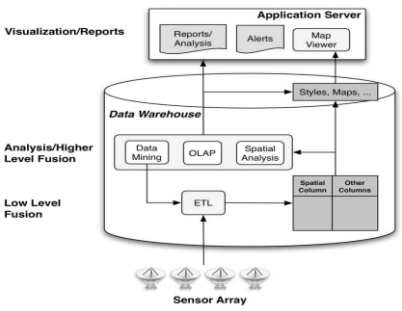
S. Ramakrishnan
B.Sc. M.C.A
Web Developer
Tamilnadu

ABSTRACT
The data in engineering and science has been on a massive scale and stored in gigantic storage devices. The data is moved in and out in the form of data streams. Data storage levels are reaching Yottabytes in terms of storage. Science and engineering transforms such data into rich and resourceful data. Intensive methods have been researched for high dimensionality. Science also uses high speed images and video data types in applications where data streams are reaching huge volumes with dynamic data distribution. Storage and computing such data is a challenging activity and especially in terms of system interactions and communications. Mining data streams is extracting knowledge in non stopping data streams. Research in this area has gained attraction due to the importance of its applications and the increasing potential of enhancement in streaming information. This paper discusses these challenges of data mining with a focus on issues like domain specific data integration, mining unstructured data, mining data streams.

General Terms
Data Streams, Stream Mining, Unstructured Data.

Keywords
Data Streams, Unbound Data, Data Structures.

1. INTRODUCTION
The field of computer and information technology has grown in handling massive high dimensional and high speed data, transforming data-poor resources to increasingly data-rich resources in the last decade. There is demand to research and analyze fast and large data streams in terms of capturing trends, patterns, and exceptions. The most essential task in analysis is to develop techniques and tools for mining this massive data. The data is the basic part of data mining to predict future decisions by mining hidden from data stores in databases and data warehouses or repositories [1]. Mining process becomes interesting when databases dynamically change their contents [2]. Data mining is defined as knowledge discovery in databases (KDD) [3]. Mining Data processes involves many disciplines and techniques like statistics, machine learning, Databases [4]. Data can have hidden relationships and mining is discovering meaningful patterns from this data [5]. Mining is also the analysis of data sets to find unsuspected relationships and summarize them [5]. The steps performed in data preparation before mining data involves data selection, cleaning, pre-processing, and data transformation [6]. There are several other tasks in Data Mining like Time series analysis, Association analysis, Classification, Regression, Cluster analysis and Summarization. The storage of data can be from various formats including independent sensors devices or connected applications. There are many available algorithms for data extraction like CARMA, ARMOR, PARTITION and quantum algorithm to name a few. Data mining is a dynamic research field which changes due to the advancement in computer hardware and software technologies, making organizations dependent on technologies. The current data flows could outpace the processing capabilities. For example Google index has above one trillion web pages. Web data processing has introduced developments like semantic web and the deep web. The semantic web data in standards are web extensions and the deep Web is that part of the web which stores HTML pages generated on demand from databases. The Analytic Data Warehouse (ADW) is a data warehouse with analytic capabilities. The scalability can be achieved through an engineered ADW-based architecture. ADW-based architecture typically includes major components like Sensor arrays, Data Extraction, transformation and loading of data, Centralized data warehousing, Analytic modules, Visualization and reports. Figure 1 depicts ADW-based information fusion architecture. Sensor arrays produce streams of data that need to be fused and analyzed. Sensor data is processed and loaded into a centralized data repository. The data which stored can be used for data mining, model generation and undergoes deployment factors across different database instances. The database infrastructure maintains the required analytical methods. The model monitors the incoming sensor data and alerts suspicious activity when detected. The benefits of using an integrated approach include improved security, speed of data management and access and ease of application development. The individual components in ADW relate to Data Fusion Process models. Sensor Arrays are the Sources of Information, ETL in Levels 0 and 1 are for processing data. Analytic Processing Modules exist at Levels 1, 2, and 4. Reports, Visualization, and Alerts are treated as Level 5 processing. The database-centric approach in an ADW facilitates monitoring and analysis of user-system interactions and can be leveraged effectively for supporting Level 5 activities. This study makes an attempt to improve data searches with optimized memory performances



. **Fig 1: ADW-based information fusion architecture**

2. ISSUES IN DIMENSIONAL HIGH SPEED DATA

The challenges in data mining are data integration of domain specific data, mining unstructured data or web data.

2.1 Data Integration of Domain Data

Data needs to be maintained for maintenance and retrieval of information and Data Preparation before mining is an often neglected but extremely important step in the data mining process. Each scientific discipline has its defined data sets with special mining requirements. An in-depth investigation of each problem domain and development of dedicated analysis tools are essential to the success of data mining in respective domains. When large data sets of data are collected via automatic methods, non-analysis of data to be integrated can produce highly misleading results.

- Solution: Reduction of data to an aggregate or normalization of information contained in large datasets into smaller sets should be the goal in mining data. The reduction methods can be simple tabulation, descriptive statistics or clustering. Important business information is stored in the form of text and unlike numeric data it is difficult to deal with. Text mining consists of the analysis of multiple documents, extracting key phrases or concepts and preparing the text processed for further analyses with numeric data mining techniques. Feature Selection on data is an important preliminary stage in the process of data mining and should be applied to text mining. Efficient mining can be achieved by including more variables for selecting documents in the initial and even actual model building phase. Drilling down the data with a few variables of interest like Gender, region, age etc. Statistical data can be computed for each group and the major group can also be classified in sub-groups based on the next level of variables for macro level results. The lowest level is the raw data based on which group summary or results can be mined. The summaries could be stored, the first time and every time when the data is mined the summaries can be referred first and returned if found, to minimize aggregations from the raw data which is time consuming

2.2 Unstructured Data

Any corporate information that is not in a database is called unstructured data and can be textual or non-textual. The information sets lack a pre-defined data model or not comprehensible in relational tables. Unstructured data is usually for humans. Textual unstructured data is a collection of emails, Documents, PowerPoint presentations and instant messages. Non-textual unstructured data is from media like images, audio/video files. Businesses need to rapidly analyze volumes of unstructured or structured data for acquiring knowledge from their own information management systems. The sheer volume of unstructured data within an enterprise can be costly in terms of storage and difficult to handle for traditional database. Where Surveys run the risk of incorrectly representing true customer service experience, unstructured data like Face book updates, Tweets, blogs, give consumers the opportunity to share their true feelings about services or products. The Problem with Unstructured Data is the feasibility in transforming unstructured data to structured data. Business interactions require heavy resources to sift through and extract the necessary elements. An amalgamation of ICT and mobile technologies with a decrease in the costs of data storage and transmission has resulted in an increased and immediate need for the unstructured data mining of virtual information.

- Solution: Unstructured data can be mined by establishing smaller representatives of the unstructured information. Since unstructured data is vast, sorting this data is next to impossible. Smaller miniatures of the vast data help achieve specific business decisions, while mining unstructured information. The analyses can be to a reasonable level of accuracy with a faster turnaround and lowered cost. The resulting decisions can be the base for deeper diving into the information for expected or favorable decisions. Monitoring the unstructured information to regularly review and purge data improves organization of data by disposal of irrelevant data. Unstructured data could also be categorized with abstracts and type of document (PDF, PPT, XML, etc.). The mining searches can qualify the necessary documents through customizable filters...

2.3 Mining Data Streams

Stream data refers to the data that flows into and out of the system like streams, where the data flow is fast, continuous and unbounded. Streams produce volumes of data that is multi-dimensional and contains both online and offline data. Technological advancements have enabled the capture of data in a wide range of fields where the measurements are generated continuously and in a very highly fluctuating data rates like sensor networks, network traffic or web logs. Querying and mining of such dynamic data is a computationally challenging task. Also the limited time scope to rescan or perform a multi-scan as in traditional data mining algorithms on the data. The available memory space is also not sufficient to store online processing stream data. Analysis of stream data can result in scientific to business decisions. Many Algorithms have been proposed to address challenges in streaming data like Google searches or Credit card transactions or real time data from sensors. Traditional data mining techniques have been found wanting in addressing the needs of data stream mining

- Solution: The pattern analysis in data streams approximating frequency counts. The frequency on historical data is calculated and the most frequent k items in the continuously arriving data are tracked. The corresponding cluster centers and their counts are updated after examination of memory constraints and time complexity. . Since the system can read a data stream once and in sequential order, dimensionality reduction is an imminent requirement. The data should be scanned once and memory used compactly. The creation of a macro cluster from the cluster centers which stores summarized statistics helps a historical analysis based on the number of clusters and time frame. Addition of User preferences in cluster centers can help handle high dimensional data and updating of mining results regularly need to be kept as an incremental process. The mining process can then adapt itself to available resources and still be accurate on mining results.

3. Analytics and Data Mining

Compact structures in data mining need to be the base to process queries like distinct elements in a (cardinality), most frequent elements, frequencies of the most frequent elements, elements belonging to a specified range and data set membership. A cardinality computation may require at least 4MB Table for 1000000 values. Frequency counting and range query processing may require table of 7MB Table for 10000000. A number of probabilistic data structures can be used in Data Mining. The applicability of a data structures is limited by the queries but structures populated by different data sets can be combined to process complex queries. Cardinality Estimation can be done with a simple technique called Linear Counting where a liner counter is set with a bit set and each element in the data set is mapped to a bit. If the ratio of a number of distinct items in the data set to mask m is less than 1, number of collisions will be low and weight of the mask will be a good estimation of the cardinality. If the ratio is very high (100), then all bits will be set to 1 and it will be impossible to obtain a reasonable estimation of the cardinality on the basis of the mask m. Frequency Estimation with Count-Min Sketch data structures estimate frequencies of elements, find top-K frequent elements, perform range queries Bloom Filter data structure is similar to Linear Counting and designed to maintain an identity of each item, not statistics. Bloom filter maintains a bit set where each value is mapped to a fixed number of bits with hash functions. Bloom filter is widely used as a preliminary probabilistic test that allows one to reduce a number of exact checks and can be applied to the cardinality estimation.

4. Conclusions and Summary

The existing stream data mining methods define the parameter before their execution but while running they do mention to user that how to adjust to these parameters. The idea in data stream is still in its infant stage and solution to problems will force users to analyze and predict from data streams. Data streams have a major role in the business world in the near future. When mining approaches adapt to available resources stream mining results will be more accurate. This paper has discussed the issues raised for high dimensional and speed data streams and suggested probable solutions for overcoming them in a naïve way

5. REFERENCES

[1] Gaber, M.M., Krishnaswamy, S. and Zaslavsky, A. (2004). Cost-Efficient Mining Techniques for Data Streams. In Proc. Australasian Workshop on Data Mining and Web Intelligence (DMWI2004), Dunedin, New Zealand. CRPIT, 32. Purvis, M., Ed. ACS.

[2] Papadimitriou, J.Sun, C.Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005,p697-708, Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. "Mining Data Streams: A Review", VIC3145, Australia, ACM SIGMOD Record Vol. 34, No. 2; June 2005

[3] Maria Halkidi, "Quality assessment and Uncertainty Handling in Data Mining Process" http://www.edbt2000.unikonstanz de/phd-workshop/papers/Halkidi.pdf

[4] Fayyad, U. M., G. P. Shapiro, P. Smyth. "From Data Mining to Knowledge Discovery in Databases", 0738-4602-1996, AI Magazine (Fall 1996): 37–53

[5] Jiawei Han, Micheline Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Champaign:CS497JH, fall 2001, www.cs.sfu.ca/~han/DM_Book.htm

[6] Claude Seidman. "Data Mining with Microsoft SQL Server 2000 Technical Reference", ISBN: 0-7356-1271-4,amazon. com/Mining-Microsoft-Server-Technical-Reference/dp/0735612714

[7] Gaber, M, M., Zaslavsky, A., and Krishnaswamy, S., A Cost-Efficient Model for Ubiquitous Data Stream Mining, Accepted for publication in the Tenth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004), Perugia Italy, July 4-9.

[8] Gaber, M, M., Zaslavsky, A., and Krishnaswamy, S., (2004), Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments, Accepted for publication in the Proceedings of Sixth International Conference on Data Warehousing and Knowledge Discovery - Industry Track (DaWak 2004), Zaragoza, Spain, 30 August - 3 September, Lecture Notes in Computer Science (LNCS), Springer Verlag.

[9] Garofalakis M., Gehrke J., Rastogi R.: Querying and mining data streams: you only get one look a tutorial. SIGMOD Conference 2002: 635. (2002).

[10] Ganti V., Gehrke J., Ramakrishnan R.: Mining Data Streams under Block Evolution. SIGKDD Explorations 3(2): (2002) 1-10.

[11] Jiawei Han, Micheline Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Champaign:CS497JH, fall 2001, www.cs.sfu.ca/~han/DM_Book.htm

[12] David Hand, Heikki Mannila, Padhraic Smyth. "Principles of Data Mining", ISBN: 026208290 MIT Press, Cambridge, MA, 2001.

[13] He, B., Patel, M., Zhang, Z., Chang, K.C.: Accessing the deep Web: A survey. Communications of the ACM, 50(2):94{101, 2007.. Sa□_s, F., Pernelle, N., Rousset, M.C.: Combining a logical and a numerical method for data reconciliation. J. Data Semantics, 12:66{94, 2009

[14] World Wide Web Consortium. W3C Semantic Web Activity, 1994. http://www.w3.org/2001/sw/

[15] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. "Mining Data Streams: A Review", VIC3145, Australia, ACM SIGMOD Record Vol. 34, No. 2; June 2005.

[16] Fernando Crespoa, Richard Weberb. "A methodology for dynamic data mining based on fuzzy clustering", Fuzzy Sets and Systems 150 (2005) 267–284.