INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
www.ijrcar.com
Vol.8 Issue 6, Pg.: 1-9
June 2020

BIG DATA STREAMS ANALYSIS TO IMPROVE DECISION-MAKING PROCESS IN EGYPTIAN INSURANCE COMPANIES

Ehab K. Elhadad[1], Mohamed M. El Hadi[2]

[1] *Department of Computer Sciences and Information Systems Sadat Academy for Management Sciences, ehabelhdad@yahoo.com*

2 Department of Computer Sciences and Information Systems Sadat Academy for Management Sciences, mohamed.m.elhadi@gmail.com

Author Correspondence: Sadat Academy for Management Sciences, Cairo, Egypt, ehabelhdad@yahoo.com

Abstract: - One of the significant facts in insurance industry is the explosive growth of insurance data .these data are increasing rapidly. Data mining is a technique to extract hidden useful information from large database. Massive data indicate too many problems cannot help insurance underwriters optimize their decision making processes.

Frequent pattern stream mining associated with accessing large amounts of data of various types and from a variety of new sources in order to discover useful patterns and trends, beside meeting the need for timely insights, mining streams requires fast, real-time processing to get results in short response times.

The aim idea is that mining more important patterns more efficiently over data streams by using weighted maximal frequent pattern based on sliding window for improving the insurance underwriting Decision-making. This paper will build a proposed Model of weighted maximal frequent pattern mining over data streams based on sliding window (WMFP-SW) for improving insurance decision making.

Keywords: Big data, Frequent pattern mining, Decision-making**.**

1- INTRODUCTION

In complex companies decisions are made in continual basis, such decisions may be more or less circuital have long or short term effects, involve people and roles, the ability of making decision is one of primary factors that influence the performance and competitive strength of companies.

Persistent overcapacity in the insurance industry has reached the point where more competition on price would be suicidal for carries.

The insurance-underwriting process is central to the acceptance of any risk – and the defining and documenting of the terms of that acceptance. Thus, an underwriter seeks to achieve predictability and certainty of results across the organization by enhancing ancillary services such as risks control and claim handling, customizing products for individual industries and individuals clients, Fine-tuning of existing products or developing new ones to address emerging and nontraditional risks. Therefore underwriters must have access to all of this data and be able to explore, analyze, and respond to the patterns, trends, and relationships that they discover in it. (Kholghi, M., 2011)

Fast and continuous development of advanced database systems, data collection technologies, and the World Wide Web, makes data grow rapidly in various and complex forms. Therefore, mining of such complex data becomes an important task in data mining realm.
The remainder of this paper contains five sections: the literature review where the definition of big data, big data processes and big data stream mining techniques in section 2. The Main characteristics of big data stream in Egyptian insurance companies in section 3. Section 4 presents deep discussion Proposal Model of Frequent Pattern Mining over Big Data Streams for improving insurance decision-making .Finally, to summarize, section 5 provides a concluding disction.

2- LITERATURE REVIEW

This section briefly reviews big data phenomena

2.1 - Big data

2.1.1. Big Data Concepts

In Gartner's IT Glossary Big Data is defined as "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". [Gartner, 2011] Similarly, TechAmerica Foundation defines big data as follows: "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information".

2.1.2. The basic characterization of big data

The three basic features of big data: Volume, Variety, and Velocity. Other organizations and big data practitioners (e.g., researchers, engineers, and so on), has extended this 3V model to a 4V model by including a new "V": Value.

• **Volume**: refers to large amounts of any kind of data from any different sources, including mobile digital data creation devices and digital devices..

• **Velocity**: refers to the speed of data transfers.

• **Variety:** refers to different types of data collected via sensors, smart phones or social networks, such as videos, images, text, audio, data logs, and so on.

• **Value**: refers to the process of extracting valuable information from large sets of social data and it is usually referred to as big data analytics. (Abawajy, J., 2015)

• **Veracity**: refers to the correctness and accuracy of information..

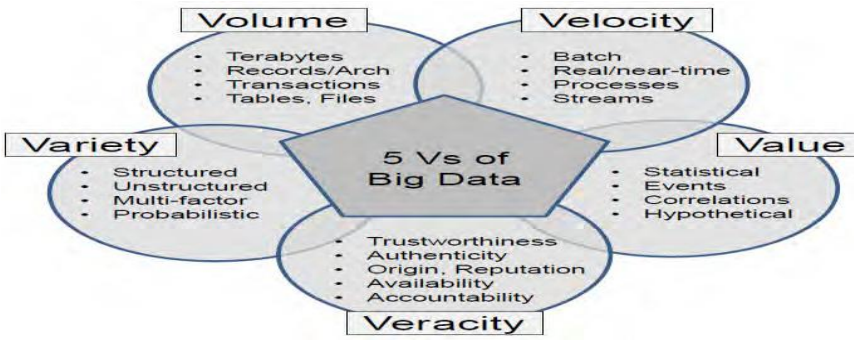

Figure 1: The basic features of big data "5V".

Table (1): The difference between Traditional data and Big data

	Traditional data	Big data
Volume	GB	Constantly updated (TB or PB currently)
Generated Rate	Per hour, day,	More rapid
Structure	Structured	Semi- Structured or un- Structured
Data Source	centralized	Fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS , NoSQL
Access	interactive	Batch or near real-time

2.2 Big data processes

The overall process of extracting insights from big data can be broken down into five stages .These five stages form the two main sub-processes: data management and analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis.

Analytics, on the other hand, refers to techniques used to analyze and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process in the overall process of 'insight extraction' from big data. (Chardonnens, T., 2013).

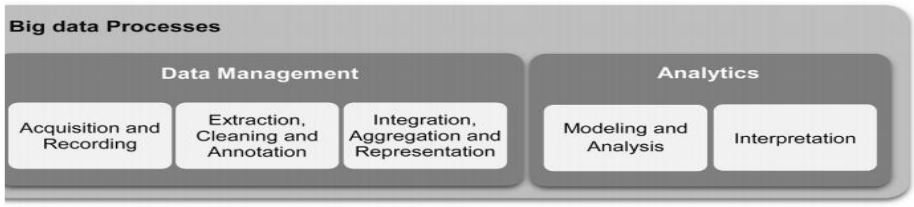

Figure2: Big data processes.

2.2.1 Big Data Management

The data life cycle consists of the following stages: collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery.

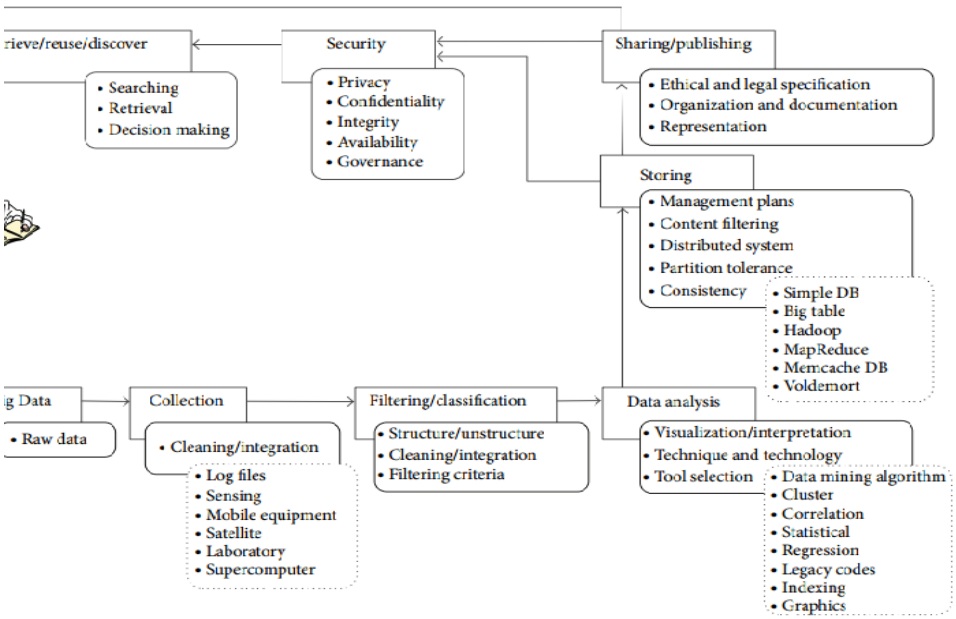

Figure 3: Big data life cycle management.

2.3 Big Data Mining Techniques

Some methods are proposed to solve data stream mining challenges and problems.

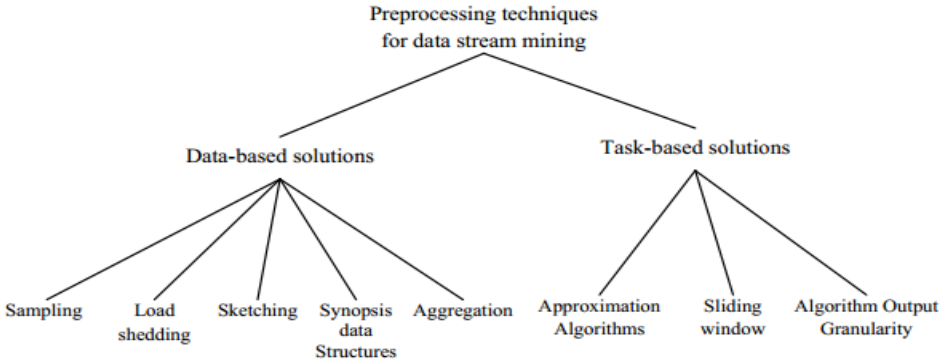

Figure 4: Big data mining techniques.

These Techniques are:

2.3.1. Data based Techniques:

These techniques are used to outline the whole dataset or choosing a subset of the incoming stream to be analyzed. Sampling, load and sketching techniques are used to summarize the data set. While synopsis of data structures and aggregation are used to select subset of data set.

2.3.2. Task based Techniques:

These methods change existing techniques or invent new methods to solve the computational challenges o f data stream processing. Approximation algorithms, sliding window and algorithm output granularity represent this category.

2.3.3. Data Stream Processing Models

There are three types of data stream processing models namely, Landmark Model, Damped Model, Sliding Window Model.

Landmark Model: processes the entire history of stream data over the some specific point in the past and in the present. In this model, summary data is to be maintained in the data structure.

Sliding Window Model: maintains and processes the part of the stream data in the current window. The result from sliding window model reflects the recent frequent itemsets. The old transactions are deleted when the new transactions arrived into the current window for processing due to unbounded, high speed characteristic of data in nature. The size of the window depends on the application and the system resources.(Ghatage, R.A., 2014)

Damped Window Model: processes the stream data based on the weight assigned to each transaction. Here, the older transactions are assigned by less weight towards the itemset frequencies and higher weight for recent data.

2.3.4. Big Data Stream Mining Algorithms

Provides an overview to the different mining algorithms we will discuss the key stream mining problems and will discuss the challenges associated with each problem. (Zahir Irani, 2017)

Data Stream Clustering.

Clustering is a widely studied problem in the data mining literature. However, it is more difficult to adapt arbitrary clustering algorithms to data streams because of one-pass constraints on the data set.

Data Stream Classification.

The problem of classification is perhaps one of the most widely studied in the context of data stream mining. The problem of classification is made more difficult by the evolution of the underlying data stream. Therefore, effective algorithms need to be designed in order to take temporal locality into account.

Frequent Pattern Mining.

The problem of frequent pattern mining was first introduced in , and was extensively analyzed for the conventional case of disk resident data sets. In the case of data streams, one may wish to find the frequent itemsets either over a sliding window or the entire data stream.

Change Detection in Data Streams.

The patterns in a data stream may evolve over time. In many cases, it is desirable to track and analyze the nature of these changes over time.

In addition, data stream evolution can also affect the behavior of the underlying data mining algorithms since the results can become stale over time. There are different methods for change detection data streams. The effect of evolution on data stream mining algorithms.

3- The Main characteristics of big data stream in Egyptian insurance companies

Some of the key analytics and data related challenges faced by Egyptian Insurance companies are:

Lack of quality data increases risks: A huge amount of data, if not maintained properly, poses challenges. Insurers also need to be regularly audited to adhere to data standards established. There is also a lack of consistency. poor quality data adds to the problem.

Lack of a single, enterprise-wide view: For a business to make the right decision and achieve optimum performance, information has to flow across functional boundaries such as sales, marketing, claims, underwriting, and operations. Information also needs to be consolidated from both structured and unstructured sources within and outside of the organization.

Inability to derive intelligent information from available data:

The massive volume of data spread across the enterprise needs to be turned into intelligent and actionable information so insurers can effectively use it to analyze key business parameters, and adjust the strategic roadmap to successfully meet business objectives.

Inability to forecast key business metrics and detect fraudulent claims:

Many insurers still lack the ability to study the past trends and forecast the future. Insurers also lack the ability to detect fraudulent patterns and identify potential fraudulent claims in the life-cycle, which results in increased claims leaks.

Because of Big data stream requirements, still there is need of general purpose model with smaller space complexity, smaller time complexity and high performance in nature for improving the insurance Decisions-Making.

4- Proposal Model of Frequent Pattern Mining over Big Data Streams for improving insurance decision-making

The Big data streams model assumes that input data are not available for random access from disk or memory, The overall process of data stream mining System has to perform in single pass , selection of some data from stream, pre-process the data, learn the model inclemently and finally represent the knowledge in form of patterns or model.

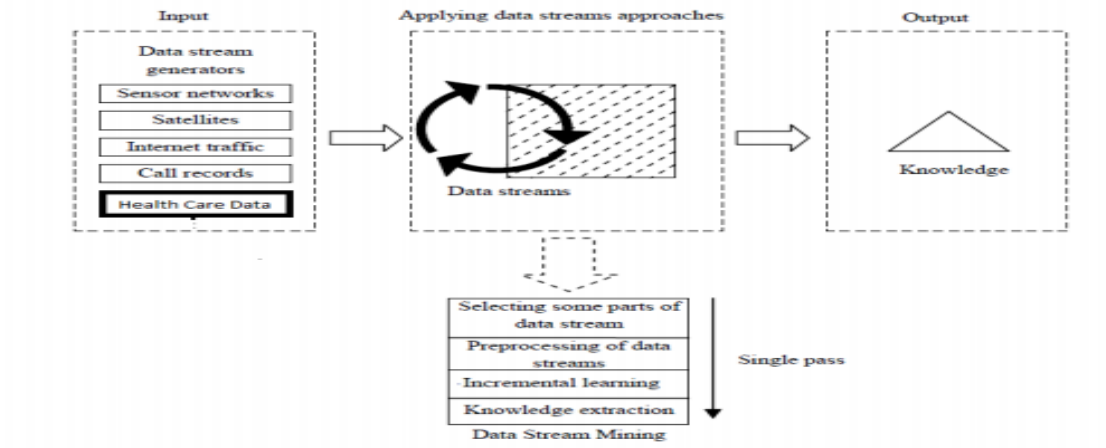

Figure 5: Big Data Mining (WMFP-SW) Model

Big Data streams mining methods are particularly effective in situations where deep and predictive insights need to be uncovered from data sets that are large, diverse and fast changing—Big Data. Across these types of data, Big Data mining easily outperforms traditional methods on accuracy, scale, and speed. (Kholghi, M., 2011).

4.1 (WMFP-SW) Approach

Weighted maximal frequent pattern mining over data streams based on sliding window model (WMFP-SW) is the first approach for mining weighted maximal frequent patterns (WMFPs) over sliding window model-based data streams.

4.1.1 Sliding window-based frequent pattern mining over data streams

In Big data streams, although a certain item is currently infrequent, it can become frequent one according to addition of new transaction data. However, those two scan-based methods must read databases from the first again since they already eliminated infrequent items in the previous step.

To solve this, mining methods suitable for Big data streams have been proposed, and they can perform mining tasks with only one database scan, thereby responding to changes of data streams immediately.

After that, sliding window-based frequent pattern mining approaches have been proposed, which can mine frequent patterns considering the latest transaction data of large data streams.

The method divides data streams into windows composed of a set of constant-sized transactions and finds frequent patterns from recently generated windows, where the size of windows and the number of them can be assigned as various values by users.

Through the sliding window-based approach, we can always obtain frequent patterns reflecting recent information. (Lee,G., 2014).

4.1.2 Maximal frequent pattern mining over data streams

In sliding window-based data stream mining, since the remaining parts except for the latest windows are not considered, the overheads can be reduced, but we cannot still avoid causing them if the size of windows or the number of them becomes large.

For this reason, the MFP notation, which can compress generated frequent patterns into a small number of compressed forms, can be utilized in the mining process, and a variety of MFP mining methods have been proposed .

Consequently, this technique not only can reduce tree traversal operations effectively but also can enhance pruning efficiency by preventing generation of needless conditional trees.

4.1.3 Applying weight conditions into frequent pattern mining over data streams

Weights of items in data streams are used in the mining process after they are converted into normalized values within a certain range.

The reason is that if a weight of any item is too large, it is hard to denote its weighted support as a finite number of digits.

The main challenge of applying weights is to maintain the anti-monotone property.(Yun,U., Lee, G., Ryu, K.H.(2014)

The framework of the algorithm,(WMFP-SW) is based on the state-of-the-art MFP mining algorithm, FPmax and the outstanding tree restructuring technique, BSM.

4.2-The opportunities of Big data streams Mining in Egyptian insurance companies

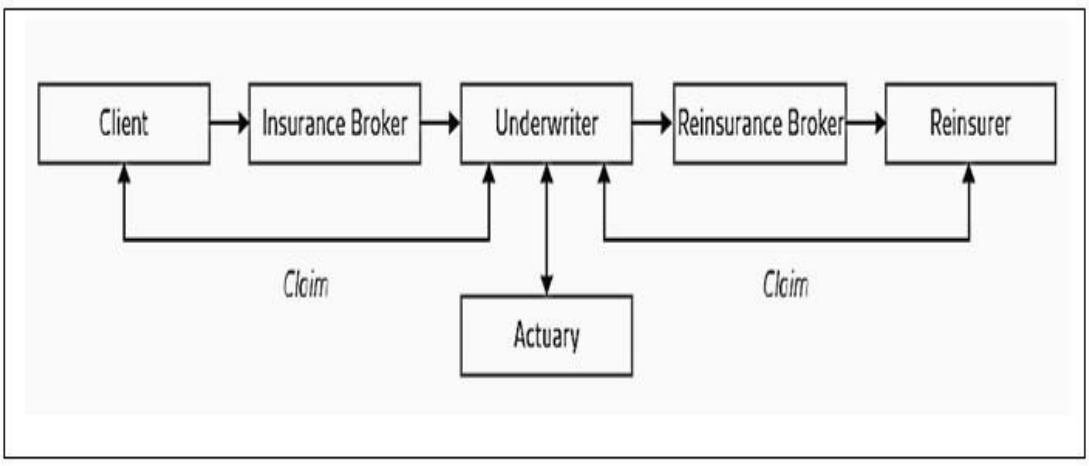

Figure 6: The opportunities of Data stream mining in insurance

Advanced analytical methods provide insurers opportunities to obtain client-specific insights, enabling them to create innovative, value-added services for their customers. Advanced analytics also enables them to speed up processes and to streamline business practices in areas such as portfolio management,

underwriting, claims management and fraud detection. It enabled them to optimize their own insurance portfolio mix, reduce risks and minimize claims. (Piotr Stolarski , 2015)

Leveraging information derived from big data supports insurers in creating models that enable better risk assessments and pricing of products. Another important area in the detection of fraud which is often part of the claims management process. Big Data streams mining and predictive modeling is the way forward for insurers for improving pricing, Segmentation and increasing profitability.

5- Conclusion

Frequent pattern mining over data streams is currently one of the most interesting fields in data mining. Weighted maximal frequent pattern mining over data streams based on sliding window model (WMFP-SW) is the first approach for mining weighted maximal frequent patterns (WMFPs) over sliding window model-based data streams.

These big data streams mining applications in insurance industry result in improving the effectiveness and efficiency of the insurance processes . hence, big data mining is considered as the most suitable technology in giving additional insight into insurance entities as well as big data mining acts as an active automated assistant in helping many insurance personal to make better decisions on their insurance activities.

References

[1] Abawajy, J. (2015). **Comprehensive analysis of big data variety landscape**. International Journal of Parallel, Emergent and Distributed Systems,

Vol. 30(1),pp. 5–14,.www.tandfonline.com/doi/abs/10.1080/17445760.2014.925548

[2] Chardonnens, T.(2013):**Big data analytics on high velocity streams.** Software Engineering Group, Department of Informatics, University of Fribourg (Switzerland), Retrieved from [hsttp//:www. Exascale.info/students projects/Chardonnens.pdf].

[3] Gartner. (2012), "IT Glossary – Big Data", Retrieved from [http//: www.gartner.com/it-glossary/big-data/] .

[4] Ghatage, R.A., Rokade, A.D., & Chavan, R.V.(2014).**A survey on sliding window based weighted maximal frequent mining over data streams**.IJARCET.vol.3, no. 10.

[5] kholghi, M., & keyvanpour, M.(2011).**An analytical framework for data stream mining techniques based on challenges and requirements**.(IJEST).vol.3, no.3 .

[6] Lee,G., Yun, U., & Ryu, K.H .(2014).**Sliding window based weighted maximal frequent pattern mining over data streams**. Expert systems with applications, vol. 41, pp. 694-708.

[7] Piotr Stolarski ,(2015**): Method of insurance premium models extraction from web sources**, Poznań University of Economics Faculty of Informatics and Electronic Economy. https://ue.poznan.pl/data/upload/articles/20151023/.../piotr-stolarski-en.pdf
Viewed on 20/4/2015

[8] Yun,U., Lee, G., Ryu, K.H.(2014).**Mining maximal frequent patterns by considering weight conditions over data streams**. Knowledge-based systems, vol. 55, pp.49-65.

[9] 9-Zahir Irani,(2017) : **Critical analysis of Big Data challenges and analytical methods**, Journal of Business Research 70 , pp. 263–286.
www.sciencedirect.com/science/article/pii/S014829631630488X Viewed on 20/4/2017.

A Brief Author Biography

Ehab K. Elhadad – PhD. Candidate, Department of Computer Sciences and Information Systems, Sadat Academy for Management Sciences.

Mohamed M. El Hadi – Professor of Computers and Information Systems, Sadat Academy for Management Sciences. President of Egyptian Society for Information Systems and Computer Technology.