Data Mining in Cloud Computing

Hamza Ahmed

Abstract: This discussion explains how through cloud computing, the process of data mining is facilitated. Through data mining, information that is potentially useful can be retrieved from raw data. People often face the need for targeted advertising, whereby data mining techniques give businesses greater efficiency, hence helping to lower costs. In the sector of cloud computing, data mining has become of great importance. The facilitation of data mining through cloud computing will enable users to obtain useful information via virtual warehouse of integrated data, helping to lower expenses of storage, technical staff, and purchase of infrastructure.

Key terms: Data mining, cloud computing

Introduction

The importance of the internet in our personal as well as our professional lives cannot be overstated as can be observed from the immense increase of its users. It therefore comes as no surprise that a lot of businesses are being carried out over the internet. Cloud computing may be one of the greatest advancements in information technology over the recent past. Cloud computing entails the use of hardware and software computer resources delivered over the internet as a service. IT drawings often depict the cloud as a cloud. Many organizations are opting to be hosted on large servers of third parties, whereby the organizations can access their software and information over the internet, rather than building IT infrastructures of their own to host their data and software. The low cost, great availability, and mobility of cloud computing has caused its use to gain a lot of popularity. Conversely, it brings along privacy threats to the data and information of acompany/organization (Naskar & Mishra, 2014). Data mining techniques have also advanced significantly in recent years and are being increasingly used in knowledge discovery in various databases, and are gaining a lot of popularity in fields which include: business, engineering, science, spatial data, and medicine (Naskar & Mishra, 2014). Users of the emerging trends in cloud computing can benefit from novel access to data of great value which can be transformed into valuable insight which can assist them to achieve their goals and objectives.

Aspects regarding cloud computing

Software and hardware delivery as a service over the internet is what constitutes cloud computing (Rountree & Castrillo, 2013). The concept of cloud computing represents computing as a utility and has recently attracted great attention. On the last half century, the paradigm shift of computing has undergone six unique phases (Stefania, 2014). In the first phase, connections to powerful mainframes were made through terminals.Many users shared these mainframes. In the second phase, stand-alone personal computers could satisfy the daily work of a user due their increased power (Stefania, 2014). In the third phase, computers could connect to one another, this being made possible by computer networks (Stefania, 2014). During the fourth phase, it became possible for different local networks to connect, making the network become more global (Stefania, 2014). In the fifth phase, the electronic grid made it possible to have computing power resources as well as computing storage resources that could be shared (Stefania, 2014). In the sixth phase, cloud computing has facilitated a simple and scalable way to exploit all the resources found on the internet (Stefania, 2014).

The National Institute of Standards and Technology defines cloud computing as: "A model for enabling ubiquitous, convenient, on-demand

network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"(Stefania, 2014). The institute further plains that the cloud model has 5 important characteristics, 4 deployment models, and 3 service models. In cloud computing, the essential characteristics are: resource pooling, measured service, self-service, rapid elasticity, and broad network access (Stefania, 2014). Cloud computing's service models are: Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS) (Communications.gov.au, 2014).

SaaS is a technology platform which facilitates a user to access applications through the internet as services which one hires as/when needed, rather than purchasing separate software which must initially be installed on the user's computer (Baun, 2011). As a variation of SaaS, PaaS as a service facilitates environment development (Rountree & Castrillo, 2013). PaaS enables the user to build applications of his/her own that run on the infrastructure of the service provider(Baun, 2011). The users receive applications via the servers' interface which can be accessed through the internet. IaaS facilitates the use of computer infrastructure (especially virtual platforms) (Salesforce.com, 2014). The users do not purchaseservers, software programs, network, or data storage equipment. When these resources are needed, payments are made for external services (Rountree & Castrillo, 2013).

As aforementioned, cloud computing has 4 deployment models which are: the hybrid, private, community, and public clouds. Regardless of the kind of model for service delivery (Whether SaaS, IaaS, or even Paas), implementation of cloud computing services is done using the four fundamental models. The public cloud platform is available to the public that is, to both individuals and organizations (Vrbić, 2012). The private cloud deployment model is a cloud computing infrastructure that is available to only one organization (Vrbić, 2012). The organization itself manages it or can have someone else do that for it (out-sourcing). The community cloud is a model that enables sharing of the same cloud computing structure by more than one organization (Vrbić, 2012).Special communities with shared interests and security requirements can be supported by this infrastructure. The hybrid cloud is a model that comprises a combination of two or more of the previously discussed models which remain distinct entities but possess a reciprocal link so as to facilitate data mobility between them (Vrbić, 2012).

Some of the best companies providing cloud computing services today include: IBM Dynamic Infrastructure, Google App Engine, AT & T synaptic hosting, Microsoft Azure, Salesforce, and Sun Microsystems Sun Cloud. Representation of virtually every resource on the internet is possible through cloud computing due to its infinite computing power (Vrbić, 2012). Additionally, the fact that the concept of cloud computing is supported by companies like Google, Cisco, Microsoft, and Oracle, which are the largest and most successful IT companies is a clear indicator of the direction the sector of IT is moving towards in the near future (Vrbić, 2012).

Data mining

Data mining involves identification of important trends or patterns through huge amounts of data (Han & Kamber, 2006).Data mining can be defined as a form of database analysis that aims to discover vital relationships as well as patterns in a given group of data ("Anderson," 2014). Advanced statistical techniques like cluster analysis, and in some instances, artificial intelligence and neuronal network techniques are used in the data analysis processes (Han &Kamber, 2006). Discovery of relationships among the data that were previously unknown especially when the data originates from different databases is a principal objective of data mining ("Anderson," 2014). There are several vital data mining techniques and a few will be described here. The first is classification which is often used to predict specific outcomes like response or no response; likely or not likely to buy; and high, medium or low value customer (Han & Kamber, 2006). The second one is association which is used to find rules affiliated with items that have co-occurrence (Stefania, 2014). It is used for root cause analysis, analysis of the market

basket, and root cause analysis (Stefania, 2014). It is also vital in defect analysis, product bundling as well as store placement. The third technique is clustering which is important for data exploration and identification of natural groupings (Stefania, 2014). Members of the same cluster are more similar with one another than they are to members of another cluster. Common applications include discovery in life sciences and identification of new customer categories. The fifth technique is anomaly detection which entails identification of unique cases on the basis of their deviation from the norm (Stefania, 2014). Examples include tax compliance and fraud in healthcare. The seventh technique is regression which is used in the prediction of continuous numerical outcomes like house value and a client lifetime value (Stefania, 2014).

With the availability of various data mining techniques combined with the necessity for identifying patterns in data that provide knowledge that is unobtainable through other means, it is no surprise that data mining finds great use in various activities (Berson, Smith, & Thearling, 2014). Businesses are capable of predicting how well a commodity may sell or create new campaigns for advertising on the basis of the new relationships provided by data mining algorithms (Stefania, 2014). Data mining helps in better analysis of geographical data and is also useful in the medical sector. Data mining techniques can also enable governments to identify illegal activities by associations, other governments, or by individuals (Stefania, 2014). The bottom line is that data mining has found great application in many fields of activity.

Data mining in cloud computing

Cloud computing refers to a new trend in delivery of services over the internet which relies on server clouds to fulfill tasks (Vrbić, 2012). Data mining in cloud computing refers to the process of obtaining structured data from internet data sources that are unstructured or partially-structured (Vrbić, 2012). There is great potential for useful data analysis and extraction facilitated in various sectors of human activities like medicine, marketing, finance, genetics, and banking (Griffith, 2014). By applying this technology it should be possible that by clicking the mouse a few times, a user can obtain the required useful data regarding customers, their interests, habits, frequency of buying certain commodities, their location, and other similar information (Spector, 2014). Services that were a reserve only for the wealthy companies is now available even to smaller ones, this being made possible by the cloud. By renting cloud services, smaller companies that may not be able to afford investing in expensive systems have the capability of analyzing large amounts of their internal data of even external data that may be of importance to them (Vrbić, 2012).

Through cloud technology large amounts of data, whose processing through standard technologies and methods cannot be achieved with efficiency, or relatively lower costs, can be easily processed meeting both affordability and efficiency (Stefania, 2014). Technically, data mining is a complex process and requires unique infrastructure founded on application of novel technologies for storage, processing, as well as handling (Vrbić, 2012). Currently,Big data or Hadoop is the hype in the sector of data processing (Vrbić, 2012). There are several solutions for huge data analysis and processing via the technologies and algorithms created by leading internet companies.

Big Data and NoSQL

The term Big Data is relatively new and refers large volumes of complex data sets whose processing and maintenance requirements cannot be achieved through traditional tools commonly used to manage databases (Kaur & Mann, 2014). NoSQL database is used for Big Data and can store data of great volume in distributed systems. Strict principles form the basis of relational databases which translated to the insurance of failure resistance, reliability, and stability. However, due to the need for a fast, extensible, and reliable

database, the problem can be addressed using the non-relational model (Vrbić, 2012). Major internet companies like Google and Amazon usually deal with vast amounts of data and have developed a technology for its processing and storage within the cloud so as to maintain database scalability and distributed systems. The databases they have developed are apparently non-relational. NoSQL is based on scalability, replication, and data partitioning (Vrbić, 2012). NoSQL databases have developed due to scalability, increased data production, storage and processing efficiency among others (Vrbić, 2012).

Apache Hadoop

This is a framework for creating applications that are scalable and distributed that function with large data volumes (in petabytes) (Vrbić, 2012). It is based on the Hadoop Distributed File System (HDFS) originating from Google's File System, and the Map Reduce algorithm (also from Google)(Vrbić, 2012). Hadoop is a cross-platform product created in Java. The Hadoop framework processes unstructured data that is "naked", and also facilitates the performance of a large number of calculations (Vrbić, 2012). As a segment of its Azure Cloud Platform, Hadoop uses Google, Yahoo, Microsoft, IBM, and Adobe among others (Vrbić, 2012).

Map Reduce

Map-reduce is a parallel programming modelintroduced by Google, which gives users a distributed file system. It can be described as a system that facilitates query executions in the background (Geng & Yang, 2014). In map reduce, two functions are involved in processing and are map and reduce. In the map stage, the distributed file system's "raw data" is read transparently and sorted, after which pairs of key are made (Geng & Yang, 2014). In the reduce stage, integration of the sorted pairs from the map stage occurs after which an output key is generated (value format) (Geng & Yang, 2014). Map reduce is widely used and it is a very efficient parallel programming model even with its simplicity (Geng& Yang, 2014).

Application of data mining in cloud computing

Due to the fact that data mining in cloud computing is still something new, the completed solutions available to users are still limited (Knobbe, 2006). New products will however be available in the near future and a greater number of data mining solutions exploiting cloud will be available (Vrbić, 2012). Examples of some solutions that are currently in existence include: Google Big Query, Amazon Elastic Map reduce, and the cloud's SQL Server Data Mining (Vrbić, 2012).

Conclusion

In the contemporary world we are living in, information is one of the most important and expensive resources. The vast amounts of data produced everyday often hide information that is potentially useful. A lot of the data does not only come from the information systems of organizations but also originates from "on-line" sources and may be useful to both companies and individuals. Examples of information from such data are the interests of customers and purchasing preferences. The great potential of cloud computing to store and process data, as well as data mining techniques which have shifted to the cloud establishes a great platform to analyze the large volume of data produced everyday which has useful information hidden in itself. This hidden information can be used to make better business decisions. The accessibility of the service from anywhere creates a decentralized system which is a great advantage. Lastly, data mining in the cloud also offers the advantage of smaller companies being able to benefit from this hidden potentially useful information whose access earlier was mainly a reserve for large companies that could afford to purchase the resources required to obtain it.

References

Anderson.ucla.edu. (2014). Data Mining: What is Data Mining?. Retrieved 18 November 2014, from

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

Baun, C. (2011). Cloud computing. New York: Springer.

Berson, A., Smith, S., & Thearling, K. (2014). An Overview of Data Mining Techniques.Thearling.com. Retrieved 18 November 2014, from http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Communications.gov.au,.(2014). Cloud Computing | Department of Communications. Retrieved 18 November 2014, from http://www.communications.gov.au/digital_economy/cloud_computing

Geng, X., & Yang, Z. (2014). Data Mining in Cloud Computing Xia Geng ,Zhi Yang.Webcache.googleusercontent.com. Retrieved 17 November 2014, from http://webcache.googleusercontent.com/search?q=cache:h6OaEiqpuxQJ:www.atlantis-press.com/php/download_paper.php%3Fid%3D9547+&cd=1&hl=en&ct=clnk

Griffith, E. (2014). What Is Cloud Computing?. PCMAG. Retrieved 18 November 2014, from http://www.pcmag.com/article2/0,2817,2372163,00.asp

Han, J., & Kamber, M. (2006). Data mining. Amsterdam: Elsevier.

Kaur, I., & Mann, D. (2014). Data Mining in Cloud Computing. Webcache.googleusercontent.com. Retrieved 17 November 2014, from http://webcache.googleusercontent.com/search?q=cache:-hY1zdCfe_UJ:www.ijarcsse.com/docs/papers/Volume_4/3_March2014/V4I3-0601.pdf+&cd=1&hl=en&ct=clnk

Knobbe, A. (2006). Multi-relational data mining. Amsterdam: Ios Press.

Naskar, A., & Mishra, M. (2014). Realisation of Resourceful Data Mining Services Using Cloud Computing. Webcache.googleusercontent.com. Retrieved 17 November 2014, from http://webcache.googleusercontent.com/search?q=cache:85ZdNFEWHKkJ:www.enggjournals.com/ijcse/doc/IJCSE13-05-07-028.pdf+&cd=1&hl=en&ct=clnk

Rountree, D., & Castrillo, I. (2013). The basics of cloud computing. Burlington: Elsevier Science.

Salesforce.com,. (2014). What is Cloud Computing Technology? - salesforce.com. Retrieved 18 November 2014, from http://www.salesforce.com/cloudcomputing/

Spector, B. (2014). Cloud computing and data consultancy | DataMine Lab. Dataminelab.com. Retrieved 18 November 2014, from http://dataminelab.com/services/

Stefania, R. (2014). Data mining in Cloud Computing. Webcache.googleusercontent.co

m. Retrieved 17 November 2014, from http://webcache.googleusercontent.com/search?q=cache:CP3zXb4RpTEJ:www.dbjournal.ro/archive/9/9_7.pdf+&cd=2&hl=en&ct=clnk

Vrbić, R. (2012). Data Mining and Cloud Computing. JITA - Journal Of Information Technology And Applications (Banja Luka) - APEIRON, 4(2). doi:10.7251/jit1202075v