ISSN : 2231-0711
K. Kala Bharathi et al | International Journal of Computer Science Engineering and Technology(IJCSET) | Feb 2015 | Vol 5, Issue 2,31-34

Recent Developments on Data Warehouse and Data Mining In Cloud Computing

K. Kala Bharathi* and K. Sandhya Sree

Department of Computer Science,
St. Pious X Degree & P. G. College for Women,
Nacharam, Hyderabad, India-500 076

Abstract:This paper expresses the use of data warehousing and data mining in cloud computing. Data warehouse is a centralized, persistent data store and data mining is a process of extracting potential information from raw data. Cloud computing is a new paradigm for hosting and delivering services over the internet and attracted business people, as it eliminates the requirement for users to plan ahead for provisioning and allows enterprises to start from the small and increase resources only when there is rise in service demand. Additionally, we compared how retrieval of data from data warehouse in the cloud environment reduces time, infrastructure and storage, over traditional method.

Keywords: Data Warehousing, Data Mining, Cloud Computing.

INTRODUCTION:

Cloud Computing:

Cloud computing is the next stage in the internet's evolution. Cloud computing is typically defined as a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. In cloud computing, the word cloud is used as a symbol for "the Internet," so the expression cloud computing means "a type of Internet-based computing," where different services such as servers, storage and applications are delivered to an organization's computers and devices through the Internet.
Why there is a need for cloud?
Means it is bigger, better, faster and cheaper.

- It is faster because it provides infrastructure on demand in terms of APIs.
- It is cheaper because reduced need for huge investment in purchasing hardware and software i.e. barrier to entry is much lower.
- It is better because no need to worry about infrastructure it is someone else's problem and we can focus on core business.

"Cloud computing is a model for on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction". Before cloud computing traditional business applications have always been very complicated and expensive. The amount and variety of hardware and software required to run them are scary. We need a whole team of experts to install, configure, test, run, secure, and update them. With cloud computing, we can eliminate those headaches because we are not managing hardware and software-that's the responsibility of an experienced vendor. Present is the age of information technology. The aspect of work and personal life are moving towards the concept of availability of everything online. Understanding this trend, the big and massive web based companies like Google, Amazon, and Salesforce.com came with a model named "Cloud Computing" the sharing of web infrastructure to deal with the internet data storage, scalability and computation. The shared infrastructure means it works like a utility. We only pay for what we need. Upgrades are automatic and scaling up or down is easy.

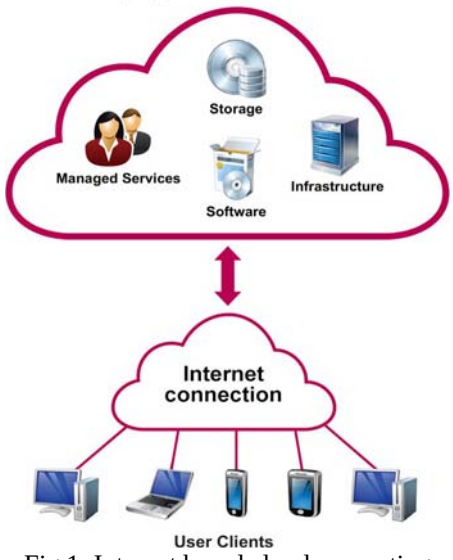


Fig 1: Internet based cloud computing

Cloud-based apps can be up and running in days or weeks, and they cost less. With a cloud app, we just open a browser, log in, customize the app, and start using it. Anything from basic word processing to collaboration to e-mail to multimedia processing can be accomplished more efficiently using cloud computing than using one's personal computer. Cloud is classified into four types Public, Private, Community and Hybrid. These four are also called as four-pillars of cloud.

Data warehouse:

A data warehouse is defined as a "subject-oriented, integrated, time variant, non-volatile collection of data that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. In data

warehouses historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be much larger than operational databases. Most queries on data warehouses are ad hoc and are complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Bill Innmon, the acknowledged "father" of the data warehouse, enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries like manufacturing for order shipment and customer support, retail for user profiling and inventory management, financial services for claims analysis, risk analysis, credit card analysis, and fraud detection, transportation for fleet (navy) management, telecommunications for call analysis and fraud detection, utilities for power usage analysis, and healthcare for outcomes analysis[1].

The data in a data warehouse have the following characteristics:

- Subject oriented: The data are logically organized around major subjects of the organization, e.g., around customers, sales, or items produced.
- Integrated: All of the data about the subject are combined and can be analyzed together.
- Time variant: Historical data are maintained in detail form.
- Non-volatile: The data are read only, not updated or changed by users.

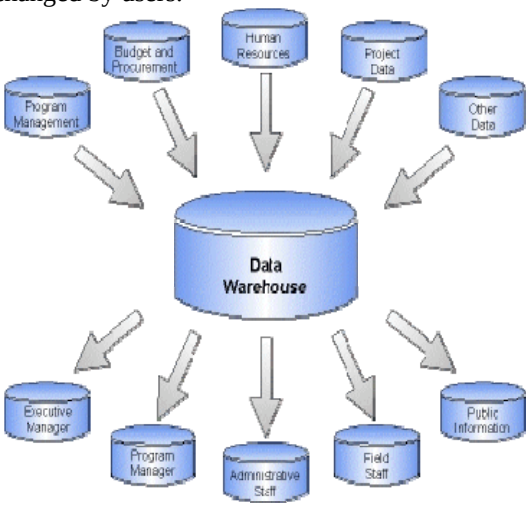


Fig 2: Collection of Data from various departments into warehouse

A data warehouse draws data from operational systems, but is physically separate and serves a different purpose. Operational systems have their own databases and are used for transaction processing. A data warehouse has its own database and is used to support decision making. Once the warehouse is created, users like analysts, managers access the data in the warehouse using tools that generate SQL (structured query language) queries or through applications such as a decision support system or an executive information system. "Data warehousing" is a broader term than "data warehouse" and is used to describe the creation, maintenance, use, and continuous refreshing of the data in the warehouse.

Advantages with Data Warehouse:

- One of the best advantages to using a data warehouse is that users will be able to access a large amount of information.
- One powerful feature of data warehouses is that data from different locations can be combined in one location.
- Another advantage of data warehouses is that they can create a structure which will allow changes within the stored data to be transferred back to operational systems.

Disadvantages with Data Warehouse:

- Before data can be stored within the warehouse, it must be cleaned, loaded, or extracted. This is a process that can take a long period of time.
- Another problem with the data warehouse is that it is difficult to maintain[2].

Data Mining:

Data Mining is the extraction or "Mining" of knowledge from a large amount of data or data warehouse. To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge from stored data. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use i.e. transforming data into information. Data mining is the process of applying intelligent methods to extract data patterns or model. This is done using the front-end tools. The traditional (reactive) Business intelligence tools to extract data are reactive but in contrast, data mining is proactive. Instead of having the end user define the problem, select the data, and select the tools to analyze the data, data mining tools automatically search the data for anomalies and possible relationships and there by identifying the problems. Data mining describes a new breed of specialized decision support tools that automate data analysis. Example banks and credit card companies use knowledge based analysis to detect fraud, thereby decreasing fake transactions.

Therefore, in the evolution from business data to business information, each new step has built upon the previous one.

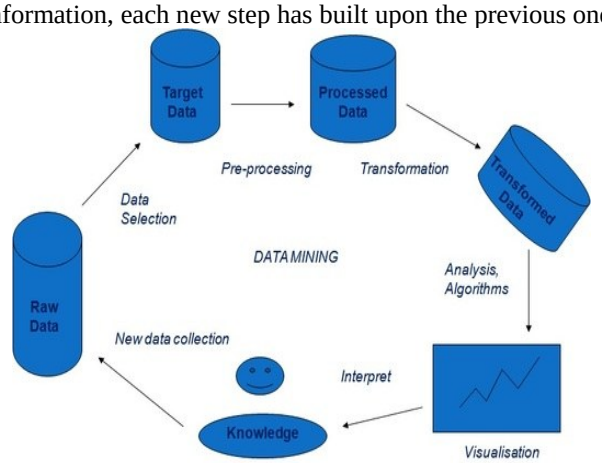


Fig 3: Knowledge base data extraction

Advantages / benefits with Data mining:

- Predict future trends, customer purchase habits
- Help with decision making
- Improve company revenue and lower costs
- Market basket analysis
- Fraud detection

Disadvantages/Barriers with Data mining:

- Amount of data is overwhelming
- Great cost at implementation stage
- Possible misuse of information
- Possible in accuracy of data

Data warehouse in Cloud Computing:

First, data warehousing raises the bar on cloud computing. Capabilities such as data aggregation roll up and related query intensive operations may usefully be exposed at the interface whether as Excel-like functions or actual API calls. Cloud computing is the opposite of traditional data warehousing. Cloud Computing wants data to be location independent, transparent and function shippable. Whereas, the data warehouse is a centralized, persistent data store. Run-time metadata will be needed so that data sources can be registered get on the wire and be accessible as a service. In the race between computing power and the explosion of data, large volumes of data continue to be stuffed behind I/O subsystems with limited bandwidth. Growing data volumes are winning. Second, data warehousing in the cloud will push the pendulum back in the direction of data marts and analytic applications. Because it is hard to image anyone moving an existing multi terabyte data warehouse to the cloud. Such databases will be exposed to intra-enterprise corporate clouds, so the database will need to be web service friendly. In any case, it is easy to imagine setting up a new ad hoc analytic app based on an existing infrastructure and a data pull of modest (unexceptional) size. This will address the problem of data mart rise since it will make clear the cost and provide incentives for the business to throw it away when it is no longer needed. Thus, with the rapid development of processing storage technologies and the success of the internet, cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications and services those can be rapidly provisioned and released with minimal management effort or service provider interaction. Cloud computing is basically for storing and accessing of applications from the computer (Remote). Whereas Data warehousing refers to the combination of many different databases across an entire enterprise used to store the information and generate the query regarding the required data. The sources will help to access the information, save downloads & update the information viz. suppose your one file has some data and you want to add same data or updating the data in a file, so these sources will help you. We want to see the data so we can access the data or generate a query for getting the information. One of the advantage is the data on cloud environment can be recycle, reuse, reduce & recover. Thus, responsible usages of the cloud take to provide a better environment for all[3, 4].

Data mining in cloud computing

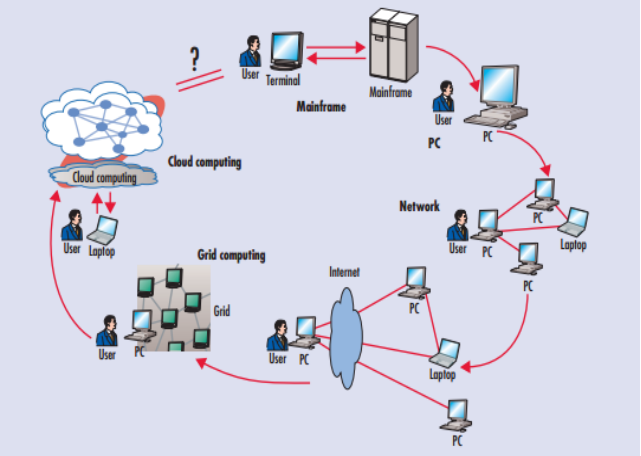
Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining. "Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way. The main effects of data mining tools being delivered by the Cloud are:

• The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive.

• The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.

"Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources. The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.



In above figure illustrated the computing paradigm shift on the last half century through distinct phases[5, 6].

CONCLUSION:

Data mining technologies provided through Cloud computing is an absolutely necessary characteristic for today's businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviors predicted. As the need for data mining tools is growing every day, the ability of integrating them in cloud computing becomes more and more tough.

Data warehousing over the cloud computing has potential for elasticity, scalability, deployment time, reliability and reduced costs. The capabilities of the Data warehousing over cloud computing is high, parallel and distributed. High Security issues will involved in the decision of moving a data from Data warehousing or data marts into the cloud. It involves easier controlling the environment Data warehousing over the cloud is largely hypothetical. Hence, with the increasing of data and growing of technology, in future warehouse maintenance and data extraction also become arduous and they may replaced by some new technologies.

ACKNOWLEDGMENT:

We thankful to The Principal, St. Pious X Degree & P.G. College for Women for providing literature facilities and also we thankful to our colleagues for encouraging us in this work.

REFERENCES:

1) Inmon, Bill and chuck Kelley. "The twelve rules of data warehouse for a c/s world," Data Management Review, 4(5), May 1994, pp.6-16.
2) http://www.exforsys.com/tutorials/data-warehousing/advantages-and-disadvantages-to-using-a-data-warehouse.html
3) Vaibhav C.Gandhi, Jignesh A.Prajapati and Pinesh A.Darji. Cloud Computing with data warehousing and Analysis Market June 2009, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 1, Issue 3, September – October 2012
4) William H. Inmon, Book, "Building the Data Warehouse". Page. 1-30
5) Bhagyashree Ambulkar and Vaishali Borkar, "Data Mining in Cloud Computing", MPGI National Multi Conference 2012 (MPGINMC-2012), 7-8 April 2012,
6) ORACLE, "Oracle Data Mining Mining Techniques and Algorithms", Link: http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/odm-techniques-algorithms-097163.html.