RESEARCH ARTICLE OPEN ACCESS

Data Ware House System in Cloud Environment

Mr. Krishna Prasad Bajgai[1], Mr. Amit Kumar Asthana(HOD)[2]
[1]Student (M.Tech-CSE), Subharti Institute of Technology & Engineering
[2]Swami Vivekanand Subharti University, Meerut

ABSTRACT
To reduce Cost of data ware house deployment , virtualization is very Important. virtualization can reduce Cost and as well as tremendous Pressure of managing devices, Storages Servers, application models & main Power. In current time, data were house is more effective and important Concepts that can make much impact in decision support system in Organization. Data ware house system takes large amount of time, cost and efforts then data base system to Deploy and develop in house system for an Organization . Due to this reason that, people now think about cloud computing as a solution of the problem instead of implementing their own data were house system . In this paper, how cloud environment can be established as an alternative of data ware house system. It will given the some knowledge about better environment choice for the organizational need. Organizational Data were house and EC2 (elastic cloud computing) are discussed with different parameter like ROI, Security, scalability, robustness of data, maintained of system etc.
Keyboard – EC2, Iaas, Saas, & Paas cloud coumputing ,Dw.

I. INTRODUCTION

Database Technology is well developed and accepted by every small to large scale organization in all over the world . Database technology has collected Vast Amount of data in the organizational storage . it will be a great serve to the society, if this data can be managed well for strong, then the organizational decision support system . A Concept to manage large Data Storage of Data based System is call the Data were house System .[1]

- Data were house system concept is also implemented in many organization across the world . It is very easy to deploy data were house in the Organization compare to earlier days.

The maintaining very large amount of data cross –platform data transformation & loading, integration and data retrieval are main stages of data were house system. The concept of DW, in earlier time, people find difficulties in working with large data and cross platform data integration .[2]

Dw System Concept is still in growing stage but its components are well defined to serve as good decision support system .

Integration of various system can be done by defining well-structured meta data and to achieve faster retrieval rate can be done by various types of data mining algorithm in a data were house System .

Another emerging technology that sought more in ICT is cloud Computing . Cloud Computing is a duster of scalable and vertical resources like computers, storages, System S/W etc. The users are required to have the internet enable devices to access services of service provider for the implementation of cloud computing. The services providers Provided all remaining requirements.

They are required to maintain various computers , servers, database S/W, System S/W and networking systems . There are mainly three types of services according to there Standard.

− (IaaS) Intertexture o a sources ,
− Platform as a services (Paas)
− Software as a services (SaaS).,

All these three types of service are again divided in to Public and private services.

DW System meanly services to top management people due to its decision making ability . Organization are running to world the data were house system by having their own system or adopt service of cloud computing that is EC2.[2]

This parts of paper describe about importance of both system using relevant scenario. It also explain . brief architecture of data were house system & cloud computing System . By using different example and case studies. It is also discuses by functionality & comparison of in-house data were house system & EC2 System. In conclusion of this paper have all pros & cons about data ware house and EC2. Conclusion will a way for developed & user of ICT to choose the best for their applications.

II. FUNCTION OF THE SYSTEM :-

Data from different department of organization are collected & stored together in one place. The large amount of data which come from

heterogeneous sources can not able to manage by conventional data base system .[4]
OLTP (Online transaction processing)
These data base system are made to perform small transactions for OLTP, while. DW System are used for complex analysis that some times have more then two or three dimensions. These system are known as OLAP (Online Analytical Processing).

There are three types of computational resources in terms of s/w or computational power which made available thronged a computer network (e.g. the internet).[5] EC2 refer to computational resources are available as per requirement on rent . s/w can be provided as a services to customer in the cloud is SaaS (s/w as a services) as per example :-

Data base tools, ERP tools, Utility tools, etc . infrastructure can be provided as a services to customer in the cloud is called IaaS.
(Infrastructure as a services) as per example cluster of processing unites file server , Data bas server etc.

Computational recourse can be periods to a customer to use the cloud as a platform like operating System , system s/w etc.

This is known as Paas (Platform as a services) [5] we are interested in running a DW system by using the cloud as a plate form we are mostly interested in PaaS service through this paper.

(A) Data Ware House

DW system is also organization . It Collected data from organization for at least five year or more. Data collected by DW System were generated on various OS and Data base system.
There are different internal & external format. Same time meta data about data is quite different when it reach to centralize storage system . for Exp:- Data is generated in hexadecimal format which in centralize repository it has character data type . Conversation of data from one system to another is done precisely for data for maintain consistently & reliability.

Conversation and Integration are main two tasks for loading data in to data ware house system.[5]
Further it also needed for generate meta data of stored data to make proper utilization of data ware house repository. Only meta data is not enough for faster and efficient usage but retrieval algorithms also have very important role in the system.

Data ware house are the enterprises most valuable assets in what concerns critical business information, making them an appending target for malicious inside and out side attackers.

DW are mainly database storing consolidated historical and current business data for decision support system. [6]

In figure no. 1 DW system having three layers. Electronic data extraction layer is first , transformation according to data warehouse systems requirements is second layer and loading cleaned data in to DW storage is third layer of DW. Transformation layer also do cleansing data of various formats as per systems required format.
In figure no. 2 shows the architectural diagram of DW system. In figure no. 2 meta data is one of the key elements for retrieval of data. Meta data repository is used by retrieval algorithms to dig up, dig down and dig across from internal storage.[7]
These algorithms work on the structure of the objects and get the internal format of storage of data. Meta data also contains information about the indexes, clusters and other objects created on a basic storage objects.

These objects are used in retrieval algorithm to calculate best suitable execution path for a given query. Primary constraints and integrity constraint of a basic storage objects in meta data helps retrieval algorithms to find data by using relationship defined with in an object or between two different objects.

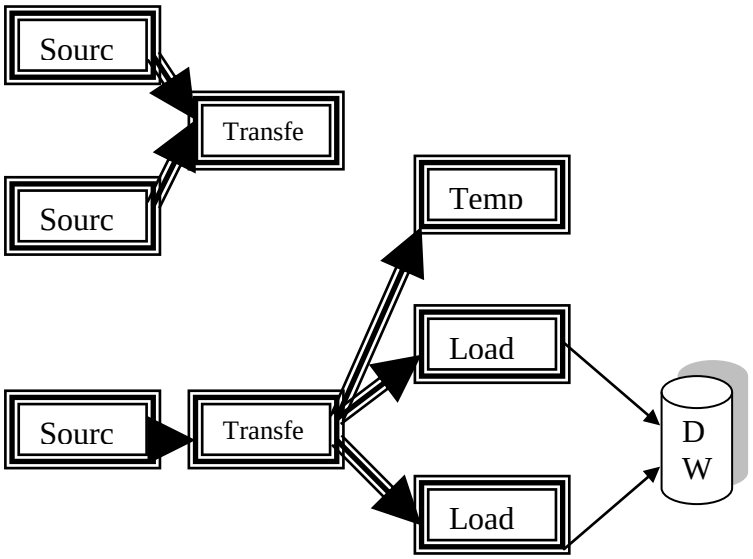
Retrieval algorithms or data mining algorithms are another important element of data ware house system [7].

Using meta data , these algorithm select suitable path for large data retrieval. Large amount of data needs to have very well maintained meta data because most of the summary field in a DSS report are pre-calculated in the system. For summary generation, this steps is also very necessary. In the internal storage, if summary fields are not pre-calculated and preserved then all retrieval process of data will do summarization process. This is redundant process and one can save big amount of time by having all necessary summaries are pre-calculated with basic storage unit. Thus meta data should be structured such that it is easy to access the summary the fields and whenever there is any change in a dependent data then meta data automatically refreshes all those summary fields affected by that[1].

DW's data many times come from the departmental level where it is called a data mart. All data marts work together and build centralized data marts work together and build centralized data work house system.

In DW system various resources such as servers, storage devices , different kind of platform , system software and network connectivity is managed in house by organization's people. Now system software means database system ,

optimizer, retrieval algorithms and meta data repository needs to be taken care for authorities.



Fg. 1 Layer of data ware housing system for collection of data.

Rules for meta data management and best selection path for optimizer in retrieval algorithm are written as per system connectivity and network band width in the system. It also calculates decay for traversing data from one storage unit to another storage unit and sum the amount for data retrieval cost. Such calculation helps in the retrieval methods and boosts the per formation of the process. This system is tightly bound with the resource and it's utilization. [7]

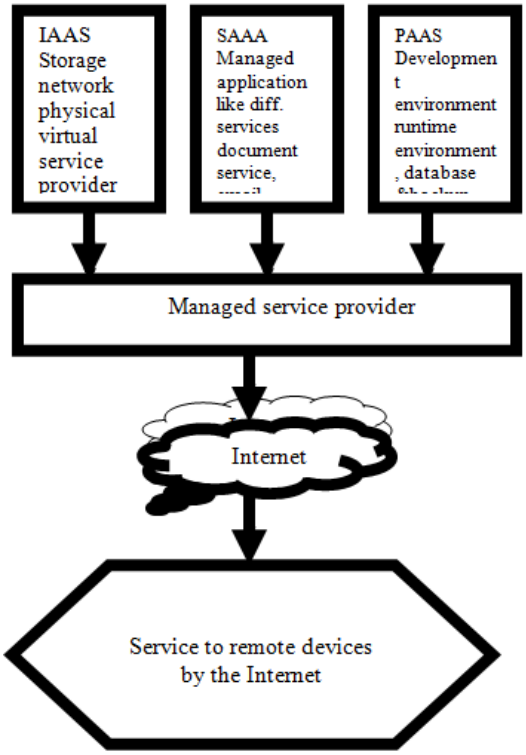
(B) **Elastic Cloud Computing:-**

In EC2 (Elastic cloud computing) service providers provides different kind of services on rent. It includes Iaas, Paas and Saas as stated above. Users of the EC2 does not require to establish computing environment in their house. Implementation of the S/W, deployment of system and maintenance of data warehouse work are performed by the service providers. Hardware & Software maintenance are also dedicated from their responsibility, thus it saves not only cost of the system but complexity of the tedious task including managing man power.[8] Fig.-3 shows the different kind of services provided by cloud computing.

Computing resources like server , storage and network are provided as a service through the internet. These resources are available on different OS or database platform or any middle ware services. It also provides special S/W services like document management , meta data management , data retrieval S/W etc.[7]

It is mandatory for services provider to provide uninterrupted services with high scalability , robustness and security. A consumer need to purchase computing power and other services prescribed in the list as per their requirement without bothering about computing resources and manpower investment. There are some reason that makes the EC2 (Elastic cloud computing) to consider as an alternative of in house data warehouse concept and technology. Some of among them are given below.

- Scalability :- With clouding computing there is an illusion of infinite computing resources. If a customer want more resources , he/she can rent these resources and more capabilities will become available to the customer almost instantly .
- Speed of deployment :- Offering of full-fledged services by cloud providers can reduce deployment time compared to in house deployment.
- Reliability :- A cloud providers can achieve high reliability , Theoretically . This can be achieved not only by making backups , but also by having more resources ; for example multiple data centers.
- Elasticity :- In cloud computing a pay –per –use payment model is generally applied , meaning that you only pay for the resources you actually use. This model ensures that deployment costs and costs due to over – provisioning are avoided.
- Reduced costs :- Costs can be reduced because users can hire services as per requirement. The cost of effort is also reduced because one can instantly acquires services as per demand.



Given are major benefit of EC2 that attract people to have services on rent rather by having in house system for Organization. Some time Theoretical & Practical ration may not match but overall services are efficient per quality.

III. EC2 AND DATA WAREHOUSES:-

As describe above to develop in house data warehouse system in the Organization required long span of time and sizeable Capital investment in hardware & software. Lost of technical reserves and technical main power required to build on environment .[2]

Hugest data management become a key issue for developer and administrator for such system . system must work with good retrieval rate and data should be summarize with more then two or three dimensions for effective DSS Reports . EC2 should provide these services on rent . Here, neither technical resources and required to establish in environment nor very high skilled people are needed to employ for management. EC2 is more Convenient for Small and medium scale Organization . Unlike EC2, Organization data were house is generated for their in-house data only though it very complex process but it provide more consistent, robust and scalable approach to the user . in house data were house storage is granted for betterment of Organizational DSS that enforces organization's standards and reduces redundancy. Consistent standard , structures and secured transactions are important in in-house data warehouse system. Later, if organization wants to switch over from one platform to another platform they can do the migration and there is no redundancy in data format and data values. [2]

In ECC, user may have to wait for availability of services and some time for the compatibility of H/W and S/W. Migration of data from one system to another is not easy in EC2 so for whole data is under control of services provider. Here, one can have access to generate data; This data can be send and retrieve by user but how data has been stored in a system or what types of internal structure is assigned to the data is unknown to service taking organization. Small and medium scale organization can find these services very suitable to them. EC2 can reduce the cost of capital in investment as well as maintenance of devices and other resources. While in-house data warehouse demands very high capital investment and other running cost. According to the Darrel M. West , minimum 40% of cost reduction is estimated in EC2 for different cases compare to in-house data warehouse system[5]. This is very beneficial for organizations that want to cut off there capital cost.

Data ware house has in-house storage allocation so optimizer tool can give extra value to device a cost of query execution. Data administrator can alter the command of query according to storage location in the network , band with of establish network and distance between two or more storage units. User of data warehouse can

have control on retrieval process. If centralize data warehouse system is not working then data marks of department is able to provide service on department data.

Table -1 Approximate cost of data warehouse system in organization.

SNo	Description	Approximate Estimated Cost	Approximate of given set of Values	Cumulative cost for 5 years.
1	Capital cost of H/W	$20000 to 40000	$30000	$30000
2	Operational cost of H/W per Annum	$1000 to $2000	$1500	$7500
3	Capital cost of DBMS	$10000 to $26000	$18000	$18000
4	Operational Cost of DBMS per Annum	$400 to $1200	$800	$4000
5	Capital cost of ETL tools	$4000 to $8000	$6000	$6000
6	Operational cost of ETL tools per Annum	$400 to $600	$500	$25000
7	Capital cost of Retrieval and data mining tools	$4000 to $ 16000	$10000	$10000
8	Operation cost of retrieval and data mining	$1000 to $5000	$3000	$15000

Total Approximate amount spent for organizational data warehouse system: $69800

It is not necessary requirement for centralized data ware house system should be up for the transaction and / or DSS reports. Later these transaction can be merged the centralized data ware house system as and when transformation batch get started.[3]

In EC2, our own optimizer routine can not work with service providers utilities. Services provider can be here there internal optimizer to optimize data belong to it's storages for different organizations. If certain service are down in cloud computing then it effort the inter system. Some time user are unable to access there local data too. In EC2, user can not compute retrieval time and cost of data processing.

Security is always and issue when we are working with the internet. In EC2 , VPN is a good solution for security of our transactions on the internet. Where as data warehouse is private with in within an organization , it is secured in organization environment.[7]

Implementation of organization (In house) data warehouse and EC2 are different in there own way. Complete implementation of organizational data ware house can be done with in 1-2 years or in same case it may takes more the two year of periods.

Implementation and deployment is complex process compare to other task of the system. Unlike organization data warehouse system , EC2 can start working within 2-3 months. In very few cases it may takes up to 6 month of period of complete management of necessary tools to starts up producing data for data ware system. [5]

A study of both data warehouse system for cost and performance point of view has been done for more then 50 industry people. Table-1, show the capital cost and operational cost of organizational data ware house system. In-house data warehouse system naturally having large amount for deployment of new H/W and S/W. this cost is one time cost and consider as a capital cost of data ware house system while to maintain all H/W and S/W need additional cost that is operational cost of data warehouse .[3] Table – 1 shows capital cost and operational cost for consecutive five year as per current rate.

The cost of data ware house system for EC2 is shown in table -2 as per current market rate service provider are changes different rate for there different services like file server usage, CPU usage, RAM usage , Band with usages etc.

Table-2 Approximation cost of EC2 data warehouse system .

Description	Approximate estimated cost
Cost of CPU hour usage Cost o RAM hours usage Cost of H/W based networking usages. Cost of out going band with Cost of cloud file storage usage. Cost of operating system platform usage.	All these utilities are provided to user by $150 to $600 per uer per month rate.
Cost of DBMS and other related product usage.	In SAAS rate of utilities start with $200-$500 per user per month rate.

Given Table-2 is prepared by considering range of rate available in the market for users. Later on mean of the range will calculate effective cost for EC2 data ware house system.

Table-1 shows set of values according to current market. Capital cost is spent for once at the time of deployment and operational cost is for every year. Last column shown the cumulative cost of consecutive five years operational cost and capital cost. The approximate total cost is $115500 for organization data warehouse system.

In EC2 , as discussed earlier it is service based architecture, charges are given as per services. Table-2 shows cost of service charges. There are many kinds of package available by service provider. These package start with $150 and go up to $600. The average of the cost is $375. Given charges are per user per service for infrastructure and services used by user of provider. Similarly charges for data warehouse products like ETL tools, meta data management tools, and data mining tools , also include in the cost. Generally this charges starts with $200 and go up to $500. The average of cost is $350 per user per month.

IV. CONCLUSION

It can be said that there is no denying about bugs in the equipment's and technology which makes our system fail. However to know current trends of technology is beneficial. Similarly data ware house and EC2 are in a growing stage of ICT. One should understand the requirement and budget of their application.

EC2 is work very better in small scale and medium scale organization but some facts about its working condition are also useful to know for users. Unlike EC2 , in-house data warehouse requires large time span to deploy system successfully. There were many unsuccessful development of data warehouse in earlier days. It also demand large amount of manpower and infrastructure to work behind it. Manpower with the skill of current technology and trends is always a hazards for any organization. [2]

Currently , in the market there are not lack of hardware and software. Various types of tools are available for an application . Prices are getting down so storage media ,network band width , processing speed can be available at affordable rate. Connection to world wide web is easy even at remote place. Now people are more concern for high reliability and security for their data. If services are provided with less hazard and high security then they will definitely choose services at provided rate. [8]

The best way to choose an approximate solution for organization is to define requirements. Fix the time span and budget allocated for the system . if requirement need to be solved within short period of time and budget low then cloud computing is very good . security and scalability of data is vital then in-house data warehouse is more preferable.

REFERENCE

[1]. Buddhadav B, Shah Neepa(2009). Efficient data access method in a hierarchical way for data warehouse . The National Journal of Computer Science and Technology , vol. I, Issue I, Jan-Jan, SV-ACRID,pp. 26-30

[2]. S. Chaudhary, U. Dayal; An overview of data warehousing and OLAP Technology. In ACM Sigmod record, 1997

[3]. Kimball Ralph , Inmon W.H. (1996) , The Data Warehouse toolkit practice technique for building dimension data warehouses, John Wiley & Sons, Inc.

[4]. N.W. Patan , M.A.T. de Aragao , K. Lee, A.A.A. Fernades, R sakellariou: Optimizing utility in cloud computing through Autonomic workload execution . IEEE Data Eng. Bull ,2009

[5]. West Darrell M. "Saving money through cloud computing ", Government studies at Brookings, aprial ,07,2010

[6]. Wyld, David , "Moving to the cloud : An Introduction to cloud computing in Government ", IBM center for the Business of Government E-Government series, 2009

[7]. Helfer , Markus (2001), Managing and Measuring Data Quality in Data Warehousing , Word Multi conference on systematic , cybernetic and informatic, 22-jul-01-25-jul-01, Orlando, Florida, USA

[8]. William Meknight (2000) , The CRM-Ready Data Warehouse , DM Review enterprise Column.

[9]. Alford, Ted and Gwen Morton , " The Economic of Cloud Computing: addressing the Benefits of Infrastructure in the cloud," Booz , Allen , and Hamilton, 2009

[10]. Optimus information http://www.optimusinfo.com/blog/2011/09/24 data warehousing in the cloud.html.

[11]. Amazon web services http://aws.amazon.com/

[12]. IBM(General US web site) http://www.ibm.com/us/en/

[13]. Oracle (General US web site) http://www.oracle.com/us.index.html

[14]. Teradata (General website) http://www.teradata.com

[15]. MySql, http://www.mysql.it/
[16]. Tomcat, http://www.tomcat.apache.org/
[17]. The global world net association , http://www.globalworldnet.org