An Optimized Model for Deploying Data Warehouse in Cloud Environment

Shilpa Bohra[1], Pravin S. Metkewar[2]

[1] MBA-IT, Student, SICSR, Affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India

[2] Associate Professor, SICSR, Affiliated to Symbiosis International University (SIU), Pune, Maharashtra, India

---***---

Abstract - *Cloud computing is defining a new trend for IT by providing a set of services that appear to have infinite capacity, at par with deployment and high availability at feasible cost. The business world is moving towards the cloud for many enterprise applications. Cloud computing is in great limelight in the entire media and blogosphere. In this paper, we will focus on how the current cloud infrastructures support data warehouse elasticity. We will speculate important parameters which needs to be taken into consideration before deploying data warehouse model in to the cloud environment. We will also design a model for road mapping successful deployment of data warehouse in cloud.*

***Key Words*: Cloud Computing, Deployment, Elasticity, Amazon EC2, Data Warehouse elasticity, Amazon Redshift.**

1.INTRODUCTION

The business world, now is slanting towards the cloud for some undertaking applications. Most concerned point for them is "the place to store essential information". For these organizations, security and different concerns might keep them from receiving cloud foundation for information warehousing. Is this sensible? In this day and age, information and examination are vital to business. All extensive undertakings have assembled Data Warehouse for reporting and examination purposes utilizing the information from an assortment of sources, including their own particular transaction processing frameworks and different databases. Be that as it may, building and running an information stockroom—a focal vault of data originating from one or more information sources—has dependably been entangled and costly. Most information warehousing frameworks are mind boggling to set up, cost a huge number of dollars in forthright programming and equipment costs, and can take months in arranging, acquirement, execution, and sending forms. After you have made the underlying speculations and set your Data warehouse up, you need to contract a group of database overseers to keep your questions running quick and ensure against information misfortune. Certainly, there are signs that those worries are dying down. In an examination report illustrating its IT forecasts for 2012 and past, industry investigator firm Gartner expressed that, "At year-end 2016, more than 50 percent of Worldwide 1000 organizations will have put away client delicate information in people in general cloud." It's convoluted however the single reason organizations might need to utilize the cloud is the exceptionally same reason they won't utilize it. A study by the Aberdeen bunch finds that associations are progressively utilizing cloud-based investigation to pick up favorable circumstances, for example, four times quicker business knowledge organization times and have half more clients effectively drew in with examination. There are a number of factors relating to the making, developing and maintaining of a data warehouse system. Like, setting up a data warehouse can take a long time. Over-provisioning (over-estimating the needs of the system to meet a certain service level at peak workloads) can lead to high costs. Organizations may lack the expertise needed to set up and maintain a data warehouse. System failure and crashes, downtime, uptime or system overload can have numerous consequences for an organization. However, there is a potential solution for these issues: Elastic cloud computing & Amazon Redshift.

1.1 Background Information

As organizations move toward Cloud processing, one vital variable for achievement is embracing multi-occupant Software-Defined Networking (SDN) arrangements in data centers. Hyper-V Network Virtualization (HNV) is a key empowering agent for a multi-occupant SDN arrangement. It is additionally crucial for executing a half breed cloud environment where inhabitants can bring their own IPs, as

well as their whole system topology. This is conceivable on the grounds that the virtualized systems are dreamy from the basic system fabric. System virtualization all in all and HNV specifically, are generally new ideas. Not at all like server virtualization which is generally develop and a broadly comprehended innovation, system virtualization still does not have a more extensive commonality. The test for data warehouse modelers in the time of Cloud computing is to make and convey models that influence these realigned mechanical administrations. Whether those data warehouse models are for little or medium size organizations, they should be executed for information and a more extensive perception framework. That foundation must address the vital business operational parts of investigation and reporting. In this new data warehouse model, the advantages to shopper incorporate versatility as the organizations can rake in on the other advantage of cloud reusable parts. In spite of all the great things about Cloud computing innovation's numerous administrations, there likewise exist the unavoidable inconveniences, workload administration is not adaptable as assets are shared and oversaw. There is additionally the vulnerability of high volume of information development to and from the cloud in non-secure situations. These may abuse consistence necessities. There are many difficulties when deploying data warehouses into the cloud

- Importing the data required for the data warehouse into the cloud for warehousing can be a test, because when using the cloud, a client is subject on the web and the foundation of the cloud supplier. This might be both an execution and cost issue.

- Getting a lot of data from cloud storage to virtual hubs given by the cloud supplier for processing can be an execution issue.

- Getting the data warehouse to execute as desired can be a challenge because cloud suppliers tend to offer low-end hubs (e.g. virtual machines) for calculations, while local data warehousing frameworks have a tendency to be very much provisioned as far as CPU, memory and disk bandwidth is concerned.

- Applications running in the cloud experience WAN idleness.

- Loss of control can prompt issues including security and trust.

- There is a requirement for comprehensive cloud-based framework organization and data lifecycle administration.

When we send our organization's data stockpiling into the cloud, there are a couple of parameters to be considered while selecting a supplier:

A. Uptime

B. Security

C. Administrative Consistence

D. Extra Administration

E. Straightforwardness

F. Speed

G. Interoperability with business insight apparatuses

1.2 Related work

Three models of Cloud computing are accessible today. There is Software as a Service (SaaS), Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). SaaS model is unified with which most people are well known regardless of the fact that they don't comprehend its hidden innovations. Google's Gmail for instance, is a standout amongst the most broadly known and ordinarily utilized SaaS stages. SaaS Big Data Applications (BDAs) exist at the most elevated amount of the cloud stack. Customers are prepared to utilize them out-of-the-crate. There is no perplexing costly framework to set up or programming to introduce and oversee. In this same model, Salesforce has extended its offerings through a progression of acquisitions including Radian6 and Buddy Media. Salesforce now offers cloud based social, information and examination applications. More current participants like AppDynamics, BloomReach, Content Analytics, New Relic and Rocket Fuel all arrangement with expansive amounts of cloud-based information. Both AppDynamics and New Relic take information from cloud based applications and give bits of knowledge to enhance their execution. BloomReach and Content Analytics utilize Big Data to enhance hunt discoverability down e-trade locales. Rocket Fuel then again, utilizes Big Data to enhance the ads it renders to potential merchants. IaaS is a model through which the administration supplier economically profits the important equipment assets to run a client's applications as system, process and capacity. At the least level, IaaS makes it simple to store information in the cloud. Here we discover administrations as the Amazon Elastic Compute Cloud (EC2), the Simple Storage Service (S3) and the Google Cloud Storage. Maybe more than some other organization, Amazon spearheaded the general population cloud space with its Amazon Web Services (AWS) advertising. Different suppliers as AT&T, IBM, Microsoft and Rackspace have additionally kept on growing their cloud framework offerings.

At last, PaaS is best portrayed as an improvement situation facilitated on third-party infrastructures. They encourage quick application configuration, testing and sending. PaaS environments are regularly utilized as application sandboxes. Utilizing those stages, designers are allowed to make and in some sense, ad lib in a domain where the expense of expending assets is incredibly diminished. Google App Engine, Google Compute Engine, VMware's SpringSource and Amazon's AWS are regular case of well-known PaaS offerings. Subsequent to turning out late in this journey, Microsoft has kept on growing its Azure cloud advertising. The organization's Azure HDInsight is an Apache Hadoop offering in the cloud which empowers Big Data clients to turn Hadoop bunches here and there on interest. All the more as of late, Google presented Google Cloud Dataflow which it's situating as a successor to MapReduce. Dissimilar to MapReduce, which takes a batch-based approach to deal with preparing information, Cloud Dataflow can deal with both batch and streaming data. An elastic data warehousing framework in the cloud would consequently increment or decline the quantity of hubs utilized, permitting one to spare cash. The potential for elasticity is essential to us, because it gives a justifiable reason (budgetary points of interest) for associations to consider data warehousing in the cloud. In the following parts we will address important issues in deploying data warehousing in order to analyze the potential for data warehousing in the cloud. We will additionally address practical prerequisites (functional requirements) that data warehousing systems will largely have to agree to with a specific end goal to make data warehousing frameworks more cloud-friendly. This paper is intended to answer the accompanying principal questions:

A. Is it achievable to deploy elastic data warehouses on to the cloud?

B. What might an elastic data warehouse on the cloud resemble?

Amazon Redshift has modified how enterprises look at data warehousing by drastically reducing the investment and effort linked with deploying data warehouse systems without degrading the features and performance. Amazon Redshift is a dynamic, self-managed, petabyte-scale data warehousing solution that makes it easy and cost-effective to analyze bulk of data using existing business intelligence (BI) tools like Tableau, Micro strategy etc. As mentioned in [13], With Amazon Redshift, you can get the performance of columnar data warehousing engines that perform massively parallel processing (MPP) at a tenth of the cost. You can start small for $0.25 per hour with no commitments and scale to petabytes for $1,000 per terabyte per year. A Massively Parallel Processing Architectures (MPP) enables you to use all of the resources present in the cluster for processing data, thereby drastically improving performance of petabyte scale data warehouses. MPP data warehouses allow you increase performance and efficiency by simply adding more nodes to the cluster. Amazon Redshift, Druid, Vertica, GreenPlum, and Teradata Aster are some of the data warehouses built on an MPP architecture. Open source frameworks such as Hadoop and Spark also support MPP.

2. LITERATURE REVIEW

All through this area we will look at existing endeavors to consolidate data warehousing and cloud computing including other applicable endeavors, permitting us to see regardless of whether positive results have as of now been accomplished inside this field, or what the vision of others with respect to information warehousing or database frameworks when all is said in done moving towards the cloud is. Most work has been done in the field of moving OLTP databases towards the cloud. To the extent we know there are no critical endeavors of moving expansive scale information warehousing frameworks into the cloud. D.J. Adabi explores the limitations and opportunities of moving data management into the cloud in [2]. Conclusions that come forward from this paper are that current database systems are not yet particularly suited for moving into the cloud. However, the author argues that decision support systems (which includes data warehousing systems) are the most likely database systems to take advantage of the cloud. Reasons to support this claim are:

A. A shared nothing architecture works well for analytical data managements as well as for the cloud

B. ACID (atomicity, consistency, isolation, durability) properties are important for relational databases and while these properties are hard to guarantee in the cloud, they are typically not needed for data warehouses

C. Sensitive data in terms of security can often be left out of the analysis, making security less of an issue.

In [4], M. Brantner et al. attempted to build a database system on top of Amazon's S3, while focusing on OLTP. Though the paper does not focus on data warehousing, it is interesting and relevant to see whether a simple database system can be built on S3. M. Brantner et al. do not achieve the ACID properties, but these are typically not important for data warehousing [4]. Some promising results are presented. The paper is limited in the sense that it does not explore the possibilities of Amazon's EC2 service in detail. In [5], D. Lomet et al. proposed an architecture for making transactional databases (OLTP) more suited for deployment in the cloud. It is proposed that the database system be split up into 2 types of components: transaction components (TC's) and data components (DC's), supposedly making the database more suited for operating inside the cloud due to for example flexibility. The data components do not know anything about the transactions. The architecture, called ElasTraS (Elastic Transactional relational database), is meant to be deployable on the cloud while being fault tolerant and self-managing. The architecture consists out of the following components:

A. Distributed Fault-tolerant Storage (DFS) for the distributed storing of data.

B. Owning Transaction Managers (OTM's) that can exclusively 'own' multiple partitions in the DFS.

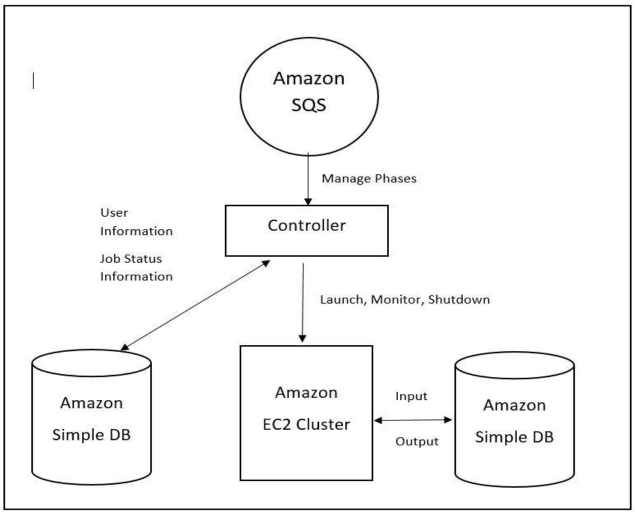
C. TM Master responsible for assigning partitions to OTM's as well as monitoring OTM's including load balancing, elastic scaling and recovering from failures.

D. Metadata Manager (MM) maintaining the mapping of partitions to their owners.

While in ci A. Aboulnaga et al. describe some of the challenges involved in deploying databases onto the cloud. Challenges empathized are placement of VM's across physical machines, the partitioning of resources across VM's and dealing with dynamic workloads. These challenges, if not solved, could impact customer willingness to move their data warehousing system towards the cloud. A solution to the partitioning of CPU capacity is provided. A way to provide adaptive workload execution in the cloud (elasticity) is provided in [7]. More about workload management can be found in [8], where it is argued that query interactions should be taken into account when managing workloads. I/O performance for database systems is analyzed in [9]. In [10], machine learning techniques are proposed to predict query execution times, which can be important regarding workload management in for example distributed and parallel database systems. A technique using VM's to separate database instances is introduced in [11], which is relevant for example for multi-tenant hosting in the cloud (more customers operating on a single physical machine). Distributing databases can be important while using multiple nodes in the cloud in order to run a data warehousing system.

3. PROPOSED ARCHITECTURE/MODEL

Figure 1. Amazon EC2- Virtual Server Hosting



In the above architecture we have used an Amazon SQS (Simple Queue Service) for quicker, solid, versatile and completely oversaw message lining administration. SQS makes it basic and savvy to decouple the parts of a cloud application. SQS can be utilized to transmit any volume of information, at any level of throughput, without the loss of messages or 24x7 accessibility of different administrations.

Then, the messages are being exchanged to the controller which helps in overseeing distinctive stages. The segment Amazon Simple DB being utilized as a part of the proposed design is an extremely open NoSQL data store that offloads the work of database association. It oversees client data and occupation status data. The reason for utilizing Amazon Simple DB is a direct result of its excessive elements like

A. Central with simple accessibility

B. Zero organization

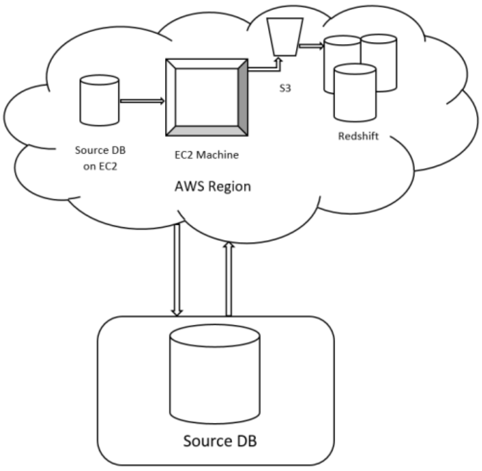
C. Secure

D. Reliable

E. Cost productive

Amazon EC2 cluster is a new instance of high performance computing applications. It interacts with Amazon simple DB to produce desired outputs.

Figure 2. Deploying Data warehouse using Amazon Redshift



As a columnar MPP innovation, Amazon Redshift offers key advantages for performant, savvy data warehousing including proficient pressure, lessened I/O, and lower stockpiling prerequisites. It depends on ANSI SQL, so you can run existing queries with practically zero change. Subsequently, it has turned into a well-known decision for enterprise data warehouses and data marts today. In this segment, we jump further into Amazon Redshift and examine more about its abilities. Amazon Redshift conveys quick inquiry and I/O execution for all intents and purposes any information size by utilizing columnar capacity, and by parallelizing and appropriating inquiries over numerous hubs. It mechanizes the majority of the basic managerial assignments connected with provisioning, arranging, checking, moving down, and securing a data warehouse, making it simple and cheap to oversee. Utilizing this computerization, you can construct petabyte-scale data warehouse in minutes rather than the weeks or months taken by conventional on-premises usage.

A. Speed and Uptime

Amazon Redshift utilizes columnar capacity, data compression, and zone maps to diminish the measure of I/O expected to perform queries. Interleaved sorting empowers quick execution without the overhead of keeping up records or projections. Amazon Redshift utilizes a MPP design to exploit every single accessible asset by parallelizing and appropriating SQL operations. The hidden equipment is intended for high performance data processing, utilizing local attached storage to amplify throughput between the CPUs and drives, and a 10 GigE network system to boost throughput between hubs. Execution can be tuned taking into account your data warehousing needs: AWS offers Dense Compute (DC) with solid-state drives furthermore Dense Storage (DS) choices. Continuous deployment of programming updates conveys progressing execution enhancements with no user intervention.

B. Security

To give information security, you can run Amazon Redshift inside a virtual private cloud in view of the Amazon Virtual Private Cloud (Amazon VPC) administration. You can utilize the software-defined networking model of the VPC to define firewall rules that restrict traffic based on the rules you configure Amazon Redshift bolsters SSL-empowered associations between your customer application and your Amazon Redshift information stockroom group, which empowers information to be scrambled in travel. The Amazon Redshift figure hubs store your information, yet the information can be gotten to just from the cluster's leader node. This disengagement gives another layer of security. Amazon Redshift incorporates with AWS CloudTrail to empower you to review all Amazon Redshift API calls. To keep your information, secure very still, Amazon Redshift scrambles every piece utilizing equipment quickened AES-256 encryption as every square is composed to plate. This encryption happens at a low level in the I/O subsystem; the I/O subsystem scrambles everything kept in touch with disk, including intermediate query results. The blocks are backed up as is, which means that backups are also encrypted.

C. Consistency, Durability and Availability

To provide the best possible data durability and availability, Amazon Redshift automatically detects and replaces any failed node in your data warehouse cluster. It makes your replacement node available immediately and loads your most frequently accessed data first so that you can resume querying your data as quickly as possible. Because Amazon Redshift mirrors your data across your cluster, it uses the data from another node to rebuild the failed node. The cluster is in read-only mode until a replacement node is provisioned and added to the cluster, which typically takes only a few minutes. Amazon Redshift clusters reside within one Availability Zone. [14] However, if you want to a Multi-AZ setup for Amazon Redshift, you can create a mirror and then self-manage replication and failover. With just a few clicks in the Amazon Redshift Management Console, you can set up a robust disaster recovery (DR) environment with Amazon Redshift. You can keep copies of your backups in multiple AWS Regions. In case of a service interruption in one AWS Region, you can restore your cluster from the backup in a different AWS Region. You can gain read/write access to your cluster within a few minutes of initiating the restore operation.

D. Scalability and Elasticity

With a few clicks in the console or an API call, you can easily change the number and type of nodes in your data warehouse as your performance or capacity needs change. [16] Amazon Redshift enables you to start with as little as a single 160 GB node and scale up all the way to a petabyte or more of compressed user data using many nodes. For more information, see About Clusters and Nodes in the Amazon Redshift Cluster Management Guide [15].

While resizing, Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel. During this process, you pay only for the active Amazon Redshift cluster. You can continue running queries against your old cluster while the new one is being provisioned. After your data has been copied to your new cluster, Amazon Redshift automatically redirects queries to your new cluster and removes the old cluster. You can use Amazon Redshift API actions to programmatically launch clusters, scale clusters, create backups, restore backups, and more. Using this approach, you can integrate these API actions into your existing automation stack or build custom automation that suits your needs.

E. Cost Model

Amazon Redshift requires no long-term commitments or upfront costs. This pricing approach frees you from the capital expense and complexity of planning and purchasing data warehouse capacity ahead of your needs. Charges are based on the size and number of nodes in your cluster. There is no additional charge for backup storage up to 100 percent of your provisioned storage. For example, if you have an active cluster with two XL nodes for a total of 4 TB of storage, AWS provides up to 4 TB of backup storage on Amazon S3 at no additional charge. Backup storage beyond the provisioned storage size, and backups stored after your cluster is terminated, are billed at standard Amazon S3 rates. There is no data transfer charge for communication between Amazon S3 and Amazon Redshift. For more information, see Amazon Redshift Pricing.

F. Interfaces

Amazon Redshift has custom Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers that you can download from the Connect Client tab of the console, which means you can use a wide range of familiar SQL clients. You can also use standard PostgreSQL JDBC and ODBC drivers. For more information about Amazon Redshift drivers, see Amazon Redshift and PostgreSQL in the Amazon Redshift Database Developer Guide. [17] You can also find numerous examples of validated integrations with many popular BI and ETL vendors. [18] In these integrations, loads and unloads execute in parallel on each compute node to maximize the rate at which you can ingest or export data to and from multiple resources, including Amazon S3, Amazon EMR, and Amazon Dynamo DB. You can easily load streaming data into Amazon Redshift using Amazon Kinesis Firehose, enabling near real-time analytics with existing BI tools and dashboards. You can locate metrics for compute utilization, memory utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster by using the console or Amazon Cloud Watch API operations. We are seeing a strategic shift in data warehousing as enterprises migrate their analytics databases and solutions from on-premises solutions to the cloud to take advantage of the cloud's simplicity, performance, and cost-effectiveness. This paper offers a comprehensive account of the current state of data warehousing on AWS. AWS provides a broad set of services and a strong partner ecosystem that enable you easily build and run enterprise data warehousing in the cloud. The result is a highly performant, cost-effective analytics architecture that is able to scale with your business on the AWS global infrastructure.

CONCLUSION

Throughout this paper we have discussed that data warehousing systems in the cloud have great potential, due to potential for elasticity, scalability, deployment time, reliability and reduced costs (due to e.g. elasticity). This paper presents a model, Amazon Elastic Compute Cloud (Amazon EC2), which takes out our need to put resources into equipment in advance, so that we can create and send applications with high speed. We can utilize Amazon EC2 to dispatch the same number of or as couple of virtual servers as we need, arrange security and organizing, and oversee capacity. Amazon EC2 facilitates to scale up or down to handle changes in necessities or spikes in notoriety, decreasing your need to forecast traffic. Also we saw that the Amazon Redshift administration is an undeniable poke at Oracle. However, it's changing the game, as far as what's conceivable around the development of data warehouses in the cloud. Redshift can give quick query execution by utilizing columnar capacity methodologies and innovation, a lot of which is taken from enterprise database innovation.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to teachers, friends and member of organization for their kind cooperation and encouragement.

REFERENCES

[1] http://www.Amazon web services

[2] S. Das, S. Agarwal, D. Agrawal, A.E. Abbadi: ElasTraS: An Elastic, Scalable, and Self Managing Transactional Database for the Cloud. In USENIX HotCloud, 2009

[3] D. Abadi: Data Management in the Cloud: Limitations and Opportunities. In Data Engineering, 2009

[4] M. Brantner, D. Florescu, D. Graf, D. Kossmann, T. Kraska: Building a Database on S3. In SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data.

[5] D. Lomet, A. Fekete, G. Weikum, M. Zwilling: Unbundling Transaction Services in the Cloud. In CIDR Perspectives, 2009

[6] A. Aboulnaga, K. Salem, A.A. Soror, U.F. Minhas, P. Kokosielis, S. Kamath: Deploying Database Appliances in the Cloud. IEEE Data Eng. Bull., 2009

[7] N.W. Paton, M.A.T. de Aragão, K. Lee, A.A.A. Fernandes, R. Sakellariou: Optimizing Utility in Cloud Computing through Autonomic Workload Execution. IEEE Data Eng. Bull, 2009

[8] M. Ahmad, A. Aboulnaga, S. Babu: Query Interactions in Database Workloads. In DBTest '09 Proceedings of the Second International Workshop on Testing Database Systems, 2009

[9] W.W. Hsu, A.J. Smith, H.C. Young: I/O Reference Behavior of Production Database Workloads and the TPC Benchmarks - An Analysis at the Logical Level. In ACM Trans. Database Syst., 2001.

[10] A. Ganapathi, H.Kuno, U.Dayal, J.L. Wiener, A. Fox, M. Jordan, D. Patterson: Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning. In Proceedings of the 2009 IEEE International Conference on Data Engineering, 2009.

[11] A.A. Soror, U.F. Minhas, A. Aboulnaga, K. Salem, P. Kokosielis, S. Kamath: Automatic Virtual Machine Configuration for Database Workloads. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008.

[12] Cloud Standards Customer Council. Practical Guide to Cloud Computing Version 2.0, April, 2014.

[13] http://aws.amazon.com/redshift/pricing/

[14] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html

[15] http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes

[16] http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html

[17] http://docs.aws.amazon.com/redshift/latest/dg/c_redshift-and-postgres-sql.html

[18] http://aws.amazon.com/redshift/partners/

BIOGRAPHIES

Shilpa Bohra, Pursuing MBA-IT in Data warehousing and Business Intelligence from Symbiosis Institute Computer science and Research