Cloud Computing with Data Warehousing

Vaibhav C.Gandhi, Jignesh A.Prajapati and Pinesh A.Darji

Parul Institute Of Engineering & Technology,
Gujarat Technological University,
P.O.Limda-391760,Vadodara, India

Abstract ***Cloud Computing is in developing state now days. It is very fast growing technology which is highly recommended by many organizations for several purposes. The cloud Infrastructure will be explained in this paper & most importantly the integration of data warehousing with cloud computing will be further explained with some reference applications like First, LogiXML, Lucidra ,Analytic SAAS and Some platforms like Google apps, IBM Blue Cloud, Amazon Elastic Cloud2***

Keywords: cloud computing, Elastic Computing, Warehousing

1. INTRODUCTION

Cloud computing is related to the grid, but different from it. The grid was supposed to be a distributed, parallel computing infrastructure along the model of the delivery of electricity – hence, the “grid” metaphor. Plug in and obtain CPU cycles on demand, exploiting the fact that most CPUs are busy between 2 and 10% of the time. Not that busy, so let’s find a way to capture the other 90 – 98% of the computing cycles. Hence, the grid. The grid envisioned combining heterogeneous operating systems, scheduling, authentication, storage and administration beneath a hardware/software abstraction layer that made service virtual. And here “service” means “web service.” Competing standards and lack of standards means the grid is a high bar to get over. Cloud computing faces many of the same challenges, but goes straight to the application level, letting the business demand for innovative computing services drive the infrastructure build out. Grid computing continues to be a work in progress in scientific and academic communities – such as NASA, Fermi Lab and related large governmental agencies – where levels of professionalism and authentication are high. Don’t laugh. The Internet was once a Defense Department research project. However, development latency is high and commercial business application results are years away. It is conceivable that the Amazon cloud (Elastic Compute Cloud [EC2]), Google cloud (App Engine), IBM cloud (Blue Cloud), as well as private enterprise corporate clouds built using 3Tera, Enomalism, or Kaavo design tools on web hosting networks such as Terremark, AT&T, OpSource or IBM’s many facilities in retail, finance, consumer goods, etc. could eventually be tied together to become the next high concept – the global grid, now re described as the commercial cloud.

2. ADVANTAGES

1. Given The abstraction of network, storage, database, security and computing infrastructure to the point of offering the image of an on-demand, virtual data center with all the flexibility implied in scalability and agility;
2. Choice of a retail-level interface suited to a business user or at least an interface suited to a high-level developer working with software components, not raw C code;
3. A pricing model that is retail in its conception – pennies per gigabyte, massive CPU cycles and bandwidth; and
4. A service level agreement (SLA) that a business person can understand and that accommodates data persistence, system reliability, redundancy, security and business continuity.

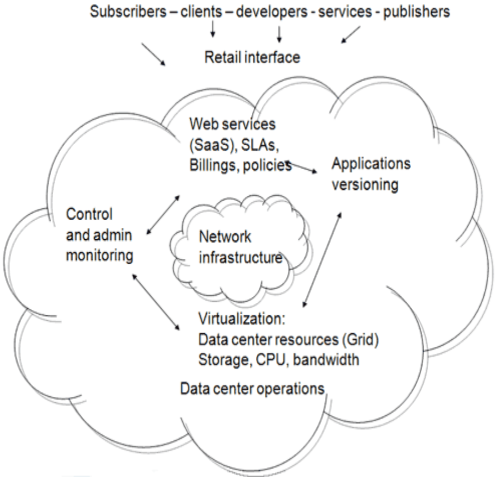
Subscribers – clients – developers - services - publishers
Retail interface
Web services (SaaS), SLAs, Billings, policies
Applications versioning
Control and admin monitoring
Network infrastructure
Virtualization: Data center resources (Grid) Storage, CPU, bandwidth
Data center operations

Figure 1: Different Partitioned Clusters

3. CAPABILITIES

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. Still, with cloud computing (as with web services), the service, not the database, is the primary data integration method.

Second, data warehousing in the cloud will push the pendulum back in the direction of data marts and analytic applications. Why? Because it is hard to image anyone moving an exiting 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 size. This will address the problem of data mart proliferation since it will make clear the cost and provide incentives for the business to throw it away when it is no longer needed.

Third, the inevitable hype around cloud computing will get a good dose of reality when it confronts the realities of data warehousing. Questions that a client surely needs to ask are: If I want to host the data myself, is there a tool to move it? Since this might be special project, how much does it cost? What are the constraints on tariffs (costs)? The phone company requires regulatory approval to raise your rates; but that is not the case with Amazon or Google or Layered Technology. Granted that strong incentives exist to exploit network effects (economies of scale and Moore’s Law like pricing). It is a familiar and proven revenue model to give away the razor and charge a little bit extra for the razor blade. Technology lock-in! It is an easy prediction to make that something like that will occur once the computing model has been demonstrated to be scalable, reliable and popular.

4. STRATEGIES

If you have been watching Apple, the silent move by Mr.Jobs to promote cloud and SaaS platform via how Apple devices will work is gaining success and has established a captive market. The day is not far off when SAP will run on a cloud and from an iPhone and iPad. Given that Exadata has emerged as a strong market and Fusion is slowly gaining momentum, Oracle’s decision to plunge into the Cloud market signifies many things

A counter move to Hadoop and Mapreduce – even Teradata and Microsoft are supporting these platforms

A move to ensure that SQL platforms continue to thrive

A move to thwart competition from Appliance vendors

A move to a greenfield opportunity To the CxO, this is very signifcant. An Oracle cloud will mean running Oracle software as SaaS, whereby both Capex and Opex can be managed. There will be no patches and upgrades to run, there will be minimal downtime. On the other hand, will you trust your business into Oracle’s hands? only time will tell

In the endgame, IBM, Teradata and Microsoft will need to establish their cloud presence or services or strategies, while strong challengers like AsterData are knocking at the customers doors. Even though consolidation is happening, the first mover advantage is going to be a little tough to overcome in the short order.

In the next set of announcements, we will hear Crowd sourcing, Social Media Support and much more from these providers, and this is a strong domain for Google and Yahoo. The way the industry is shaping up is very exciting and interesting.

5. AMAZON ELASTIC CLOUD

5.1 Services

ELASTIC – Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds or even thousands of server instances simultaneously. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

Completely Controlled – You have complete control of your instances. You have root access to each one, and you can interact with them as you would any machine. You can stop your instance while retaining the data on your boot partition and then subsequently restart the same instance using web service APIs. Instances can be rebooted remotely using web service APIs. You also have access to console output of your instances.

FLEXIBLE – You have the choice of multiple instance types, operating systems, and software packages. Amazon EC2 allows you to select a configuration of memory, CPU, instance storage, and the boot partition size that is optimal for your choice of operating system and application. For example, your choice of operating systems includes numerous Linux distributions, Microsoft Windows Server and OpenSolaris.

Designed for use with other Amazon Web Services – Amazon EC2 works in conjunction with Amazon Simple Storage Service (Amazon S3), Amazon SimpleDB and Amazon Simple Queue Service (Amazon SQS) to provide a complete solution for computing, query processing and storage across a wide range of applications.

Reliable – Amazon EC2 offers a highly reliable environment where replacement instances can be rapidly and predictably commissioned. The service runs within Amazon's proven network infrastructure and datacenters. The Amazon EC2 Service Level Agreement commitment is 99.95% availability for each Amazon EC2 Region.

Secure – Amazon EC2 provides numerous mechanisms for securing your compute resources.
Amazon EC2 includes web service interfaces to configure firewall settings that control network access to and between groups of instances.

Inexpensive – Amazon EC2 passes on to you the financial benefits of Amazon's scale. You pay a very low rate for the compute capacity you actually consume.

5.2 Features

Amazon EC2 provides a number of powerful features for building scalable, failure resilient, enterprise class applications, including:

Elastic Block Store – Amazon Elastic Block Store (EBS) offers persistent storage for Amazon EC2 instances. Amazon EBS volumes provide off-instance storage that persists independently from the life of an instance. Amazon EBS volumes are highly available, highly reliable volumes that can be leveraged as an Amazon EC2 instance's boot partition or attached to a running Amazon EC2 instance as a standard block device.

Multiple Locations – Amazon EC2 provides the ability to place instances in multiple locations. Amazon EC2 locations are composed of Regions and Availability Zones. Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and provide inexpensive, low latency network connectivity to other Availability Zones in the same Region.

Elastic IP Addresses – Elastic IP addresses are static IP addresses designed for dynamic cloud computing. An Elastic IP address is associated with your account not a particular instance, and you control that address until you choose to explicitly release it. Unlike traditional static IP addresses, however, Elastic IP addresses allow you to mask instance or Availability Zone failures by programmatically remapping your public IP addresses to any instance in your account.

REFERENCES

[1] The Enterprise Data Cloud White paper by Merv Adrian May 2009 http://www.itmarketstrategy.com
[2] Cloud Computing with data warehousing and Analysis Market June 2009
[3] www.Amazon.com/Elastic cloud