Hybrid Computation

Leveraging Cloud Computational Resources with On-Premesis Data

Tanvi Arora

Mahesh Kuklani

Examiner: Dr. Sohail Rafiqi

Southern Methodist University

Data Science

Cloud Computing

November 28th, 2018

Writing period

09. 22. 2018 – 11. 28. 2018

Project Proposal

The world we live in today is one that is increasingly data-centric. Along with our ability to generate more data must also come an increase in our capacity to make sense of that data. As organizations attempt to scale their environments, bottlenecks often arise within the underlying relational database. Some organizations use this as a reason to migrate to a cloud or hybrid environment, while others, whether from necessity or preference, remain in an on-premises environment. In this paper, we have sought to explore the transition to cloud. The design underpinning this analysis was constructed to allow examination of the feasibility and performance of utilizing cloud computational resources to augment the throughput of local relational database systems while avoiding the need for additional hardware and minimizing disruption of the existing code base.

While startups and personal endeavors are typically small and agile, it is the larger enterprises that struggle against inertia and must come to grips with the long tailed transitions that would come along with cloud adoption. Through this project our team will explore reasons to migrate to cloud while also try not to get caught up in the hype. While everyone sees the cents per service unit , are there actual savings in the long run? With this research we want to explore why's and ifs to be considered before moving conventional systems to cloud.

Why this Project?

The potential benefits from that would come along with the ability to scale a local database in a way that is flexible and ultra-low cost are obvious: low barrier to entry for small organizations, mitigation of security concerns around cloud storage.

Solution Specifics

Databases provide a prime use case for on-premises private and hybrid cloud models. For our research we will use on-prem relational database and explore hybrid models that can either re-use any of the on-prem resources for storage and compute with the flexibility to be scalable to cloud on need basis or look for a potential model that may be cost effective and provide cloud advantages while moving storage and compute entirely on cloud. To test this team will explore use of APIs and serverless compute services offered by cloud providers.

Contents

List of Figures

List of Tables

Introduction

In past few years there has been a huge chatter about cloud computing. Every Organization in this era is at least reviewing or looking at resources to see if moving to cloud would same them time and efforts. Cloud as we are aware makes provisioning of new resources quickly so an organization can concentrate their efforts on tasks that create more value for them rather than concentrating their resources on procuring hardware or provisioning servers. At the same time companies do not want to invest in hardware that is hardly utilized for about 3-4 hours in a day thus making a classic case for moving to cloud. In cloud you utilize resources for the time you need and terminate the instance when your work is complete and paying only for the time a resource is up and utilized.

Businesses are often confused by the thought of moving to cloud. While they may have valid concerns, cloud supporters have advantages to show , but ultimately it is the business who has to decide if the advantages outweigh their concerns. [1]

1 Concerns

Concerns and what do cloud supporters have to say about them :

1.1 Security

Cloud environments experience – at a high level – the same threats as traditional data center environments, the threat picture is the same. Both run softwares, softwares have vulnerabilities, and there is someone out there waiting to exploit these vulnerabilities. However security on cloud is a shared responsibility model of security. While cloud provider takes care of the security of the cloud , some aspects of security remain sole responsibility of the consumer. Effective cloud security depends on knowing and meeting these consumer responsibilities.

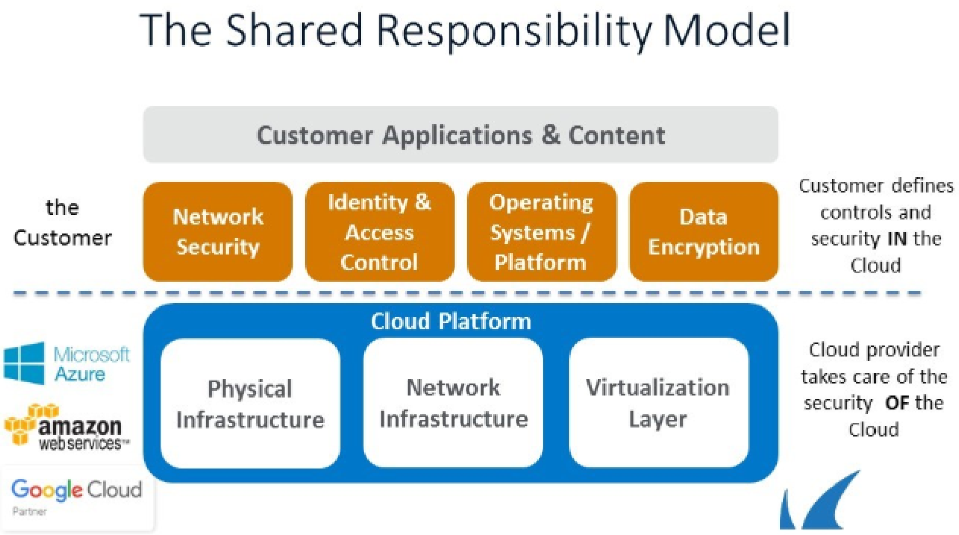

The Shared Responsibility Model
Customer Applications & Content
the Customer
Network Security
Identity & Access Control
Operating Systems / Platform
Data Encryption
Customer defines controls and security IN the Cloud
Cloud Platform
Physical Infrastructure
Network Infrastructure
Virtualization Layer
Cloud provider takes care of the security OF the Cloud
Microsoft Azure
amazon webservices
Google Cloud Partner

Figure 1: Shared responsibility model
[2]

1.2 Data Privacy

Cloud computing involves the dispersal of data across servers located anywhere in the world. Like globalization of networks. By crossing borders, involves considering countries with restrictive privacy and protection laws . This is somewhat covered as part of customer responsibility towards cloud security. For this corporations need to understand what kind of data will they load into cloud and who will have access to this data. [3]

Other aspect of data privacy is handling sensitive data. Yes one of the problems is that this technology is light years ahead of the law and there are questions that need to be answered. Who owns the data, consumer or the hosting cloud provider? Can a cloud deny the consumer access to their own data or can it share this data with marketing firms Obviously , the safest approach is data privacy is more a consumer responsibility. Keep data under proper control and apply data encryption methods. Regarding the question about laws, each company wants to protect their reputation. To get more clients and maintain them, cloud providers would uphold your data privacy. There are getting more managed and include all the necessary provisions that one should take while setting up their data on cloud. [4]

1.3 ROI

ROI or Return of Investment is widely a measure of financial success and can be a measured in a variety of ways. If you move to public cloud, you generally decrease invenstment but increase cost. With private cloud, it is vice-versa. But what matters is value to business, customer value, seller value, market brand value, corporate value. In case of cloud services, these relate to productivity, speed, size and quality. [5]

1.4 Implementation Cost

1.5 Integration Issues

Most enterprises would apply an incremental model of implementation. It is less risky than big-bang. This requires integration of services.The risk of not being able to integrate is critical. If you cannot build a system, you cannot use it. This also adds to the cost of including glue-softwares to connect various interfaces .It could involve rewrite of code or existing process models. Not to forget significant skills are required to assemble and customize multiple cloud services , requiring the applications to be loosely coupled, programmed to perform in an integration layer instead of underlying infrastructure.

1.6 System Quality

- Performance
- Functionality
- Manageability
- User satisfaction

1.7 Resource Comparisions

Storage

Table 1: Storage characteristics

Storage Features	S3	GCloud
Durability	99.999999999%	99.999999999%

Continued on Next Page...

Table 1 – Continued

Storage Features	**S3**	**GCloud**
Availability	data is automatically distributed across minimum of 3 physical Azs(57 Azs across 19 geographic regions)	99.9% SLA for regional and 99.95% SLA for multi-regional
Scalability	Maximum size limit of 5TB. Largest object uploaded in single PUT is 5GB. Objects > 100MB, should be uploaded via multipart upload capability	Maximum size limit of 5TB. Objects > 5MB should be uploaded via multipart or resumable uploading
Costs	Pricing based on data storage levels, based on frequency of access, size of data. Pricing model varies as per size tier. PUT, COPY , etc. operations are priced for frequent and infrequent access per 1000 transactions as well as for retrieval from archive.	Pricing model is different for different regions, frequency of access and size of data. PUT, COPY , GET operations are priced per 10000 transactions. It does have minimum days of storage and charges penalties for early deletion for cold storage
Security	Supports 3 different forms of encryption. Supports security standards	Encryption is automatic and no customer action is required. More than 1 encryption mechanisms used.

Continued on Next Page…

Table 1 – Continued

Storage Features	**S3**	**GCloud**
Compliance	Consumer retains complete control and ownership over the region in which their data is physically located, making it easy to meet regional compliance requirements. Compliance certifications, including PCI-DSS, HIPAA/HITECH, FedRamp, EU Data Protection Directive, FISMA, etc.	Compliance certifications include PCI-DSS, HIPAA/HITECH, SOC* , FedRamp, GDPR, etc.
Query in place	Allows to run sophisticated Big Data analytics on your data w/o moving the data into a separate analytics system	Allows to run Big Data analytics on data (Big Query)
Flexible Management	Storage administrators can classify, report and visualize data usage trends to reduce costs and improve service levels. Objects can be tagged based on their features to control storage consumption, cost and security	Pricing modes based on different regions . Does not offer volume discounts

Table 2: Server Based Compute

Server Based Computing	**AWS EC2**	**Google Compute Engine**
Elastic web-scale computing	enables to increase/decrease capacity within minutes with the additional feature of auto scaling	managed instances can be set up for auto-scaling
Completely controlled	consumer has complete control over instances including root access	consumer has complete control of systems and unlimited flexibility
Flexible cloud hosting services	variety of instance shapes available	variety of instance shapes available
Integrated	integrated with multiple Google services to provide end to end cloud solution	integrated with multiple Google services to provide end to end cloud solution
Reliable	The Amazon EC2 Service Level Agreement commitment is 99.99% availability for each Amazon EC2 Region.	Google cmpute engine guarantees 99.99% monthly uptime percentage to customers
Secure	Amazon EC2 works in conjunction with Amazon VPC to provide security and robust networking functionality for your compute resources	Google Compute Engine has ISO 27001, SOC 1, SOC 2, SOC 3, SSAE-16 certifications, exhibiting commitment to information security

Continued on Next Page…

Table 2 – Continued

Server Based Computing	**AWS EC2**	**Google Compute Engine**
inexpensive	pay for what you use. Various options available to reduce cost for low risk or less available operations requirements(ex- spot instances)	after a 10 min minimum charge, pay for what you use
Easy to start	Free to start	Free to start

Table 3: Serverless Compute

Serverless Compute	AWS Lambda	Google Cloud Function
No Server Management	Automatically runs your code w/o requiring to provision or manage servers.	code executes in fully managed environment, no need to provision any infrastructure or worry about managing any servers
Continuous Scaling	auto-scaling enabled	auto-scaling enabled
Pricing	charged for every 100 ms of code execution	charged for the time code runs
Portable	supports Node.js, python , C# and Go	can be written in Node.js or python making them portable

Background

To achieve this we are planning to use the below Amazon resources like S3, DynamoDB, EC2 instances and Lambda function.

1 S3

S3 - stands for Simple Storage Service. This service could be utilized to collect, store and analyze huge amounts of data. Data stored in S3 could be retrieved from anywhere. It provides comprehensive security and compliance capabilities. S3 is designed to deliver 99.999999999% durability.

1.1 Advantages:

[6]

1. Unmatched Durability, Availability, & Scability
2. Most comprehensive security & compliance capabilities
3. Query in place
4. Flexible management
5. Most supported by partners, vendors, & AWS serices
6. Easy, Flexible data transfer

2 Dynamo DB

[7] Amazon DynamoDB is a nonrelational database that delivers reliable performance at any scale.

2.1 Advantages:

1. Performance at scale
2. Fully managed
3. Enterprise-ready

3 Elastic Compute Cloud (EC2)

[8] Amazon Elastic Comput Cloud is a web service that provides secure, resizable compute capacity in the cloud. EC2 has changed the economics of computing by allowing companies to pay only for the capacity that is being utilized.

3.1 Advantages:

1. Elastic web-scale computing
2. Completely controlled
3. Flexible cloud hosting services
4. Integrated
5. Reliable
6. Secure
7. Inexpensive
8. Easy to start

4 Elastic Compute Cloud (EC2)

[9] AWS Lambda can run code without provisioning or managing servers. We have to pay only for the compute time consumed.

4.1 Advantages:

1. No servers to manage
2. Continuous scaling
3. Subsecond metering

5 Azure SQL Database

This is a relational database-as-a-service (DBaaS) with the latest version of Microsoft SQL Server Database Engine. It is high performance, reliable and secure database on which data-driven applications and websites can be built in the programming language of choice without needing to manage infrastructure. [10]

5.1 Advantages:

1. Fully managed
2. Advanced security
3. Built-in intelligence

Approach

1 Using AzureDataFactory and Azure SQL Database

Use cloud resources for compute with database on-premises.

1. Create Azure SQL DB on Azure cloud.
2. Create SQL Database and SSIS DB database
3. Create Azure Data Factory for serverless compute on Azure cloud.

2 Using S3 –> Lambda –> DynamoDB

1. Create S3 bucket
2. Create Lambda Function
3. Create a table in Dynamo DB

3 Using AWS/Google –> AzureSQLDB

1. Create EC2/Google compute instance
2. Install unixODBC
3. Install Microsoft ODBC version 16 or 17
4. Connect to SQL Database on Azure

Experiments

Cloud computing is a change and if used in the right way, can be a great accelerator of collaborated architecture. Cloud is an Internet phenomenon and trying to use cloud the traditional enterprise way will not achieve real returns from the cloud. In short Enterprises need to embrace new architecture.

One approach is using an incremental approach. This is less risky than big-bang and also gives transition time to employees. This approach uses a hybrid cloud where existing on-prem resources are connected to new resources on cloud. New external cloud services can be incorporated in the in-house solutions, leading way to gradually get rid of on-prem resources by their end of life.

Consider a basic on-prem architecture.

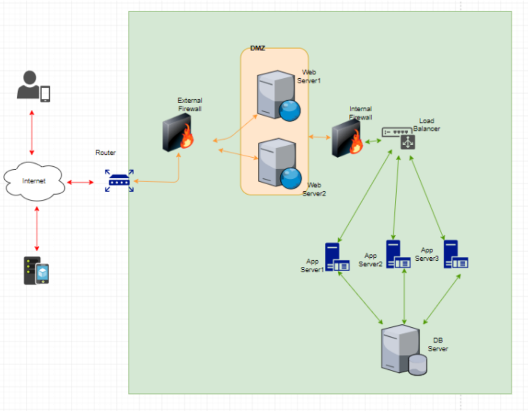


Figure 2: General Cloud Architecture

1 Azure Data Factory and SQL Database

As part of our experiment we want to move compute to Azure Cloud by using Azure Data Factory. Azure Data Factory is a fully managed service provided by Microsoft that composes data storage, processing and movement services into streamlined, scalable, and reliable data production pipelines. To utilize compute in Azure Cloud we have created table sales_records in Azure SQL Database Cloud and an SSIS service in AzureDataFactory. Once SSIS service is deployed to Cloud it is scheduled to be run daily which computes sales data for every quarter and sales for every product based on region. We are utilizing database in cloud due to the limitations of creating VPN or direct-connect as these seem to be costly options for an individual, but any corporation or industry will have VPN or direct-connect set to reduce latency. Corporation can use VPN or direct-connect to connect to on-premises database.

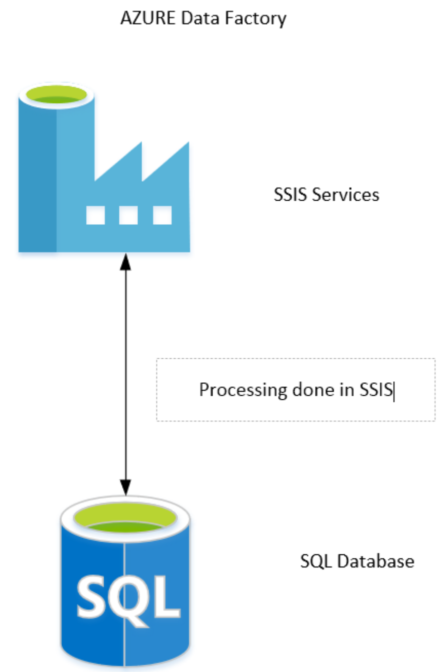


Figure 3: SSIS Azure Data Factory

1. Create SQL Database.
   a) Login to Azure portal and click create SQL Database
   b) Create Cloud Computing database with a blank database
   c) Once DB is created, get the connection string for this database.
2. Create Azure Data Factory.
   a) Login to Azure portal and click Data + Analytics and click Data Factory
   b) Create unique name for SSIS Data Factory
   c) Click Author and Monitor.

d) Click Configure SSIS Integration Runtime tile.

e) On the SQL Settings page enter the configuration of the above SQL Database.

f) Select Catalog Database Server Endpoint to host SSISDB.

2 S3 –> Lambda –> DynamoDB

As part of this experiment we upload a csv file to S3. AWS Lambda has a trigger whenever a new item is added to S3. Lambda function kicks in, it does its processing and uploads the data in csv file in S3 to DynamoDB. DynamoDB can process this data for end user or any analytic requirement. AWS Lambda is a serverless architecture which utilizes resources required for processing the file in S3 to DynamoDB. Once the processing is complete the client is not charged unless the Lambda function is triggered again. To achieve this we followed the following steps: [11]

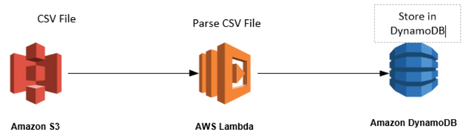


Figure 4: AWS Lambda S3 to DynamoDB

1. Create Policy in IAM.
   a) Select S3, then select all actions and all resources
   b) Add additional permissions for DynamoDB, select all actions and all resources
   c) Add additional permissions for Cloudwatch, select all actions and all resources.
2. Create Role
   a) Attach the above policy to this role
3. Create Lambda Function
   a) Create function, author from scratch
   b) Give function name

c) Select runtime as Python 3.0

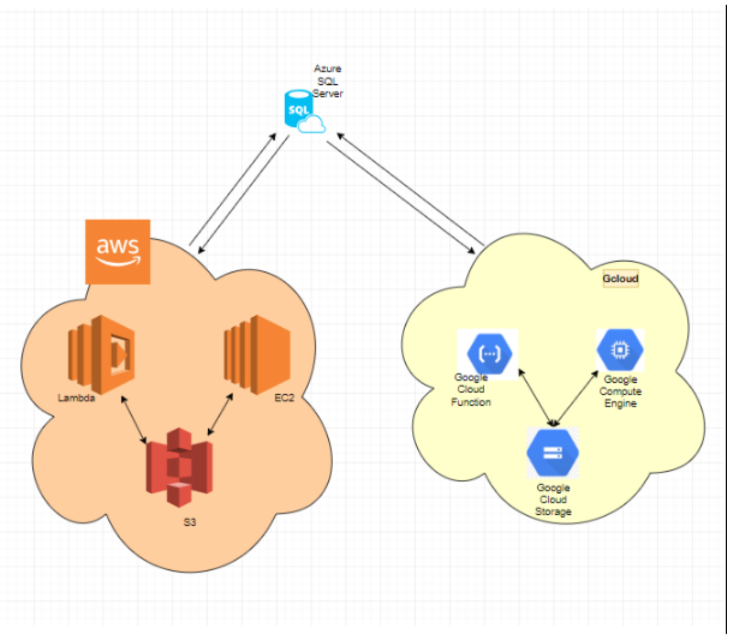
d) Choose existing role

e) Select role created above

f) Add trigger for S3, select bucket, event type (object created), prefix and filter if any.

3 AWS/Google Cloud –> Azure SQL Database

As part of this experiment we connected AWS EC2/Google Compute instance to Azure SQL Database.



3.1 Operations

1. Connect to Azure SQL Server

2. Load a file

3. Read a table from Azure SQL server

Conclusion

1. Cloud computing allows to choose multiple cloud services from multiple clouds to be integrated for your use-case

2. Use multi-cloud structure for backup and recovery.

Bibliography

[1] [Online]. Available: https://www.marutitech.com/5-reasons-why-cloud-can-transform-your-business/

[2] [Online]. Available: https://smartermsp.com/wp-content/uploads/2017/05/Shared-Security-Model.jpg

[3] [Online]. Available: https://legal.thomsonreuters.com/en/insights/articles/understanding-data-privacy-and-cloud-computing

[4] [Online]. Available: https://www.privacyrights.org/blog/privacy-implications-cloud-computing

[5] [Online]. Available: http://www.opengroup.org/cloud/cloud_for_business/p6.htm

[6] A. S3, "Amazon s3." [Online]. Available: https://aws.amazon.com/s3/?nc2=h_m1

[7] A. DynamoDB, "Amazon dynamodb." [Online]. Available: https://aws.amazon.com/dynamodb/?nc2=h_m1

[8] A. EC2Doc, "Amazon ec2." [Online]. Available: https://aws.amazon.com/ec2/?nc2=h_m1

[9] A. A. Lambda, "Amazon aws lambda." [Online]. Available: https://aws.amazon.com/lambda/?nc2=h_m1

[10] [Online]. Available: https://docs.microsoft.com/en-us/azure/sql-database/

[11] Try2Catch, "How to read csv file and load to dynamodb using lambda function." [Online]. Available: https://www.youtube.com/watch?v=5nuUmlezrHs