European University Cyprus

School of Sciences | Department of Computer Science and Engineering

Cloud Storage Data Connectors

[enter any applicable subtitle here]

Bachelor in Computer Science

Theodoros Meras

Acknowledgements

Write some text for someone that you would like to say thank you for the completion of this work

Abstract

Write a page of text explaining the project, what you wanted to do, what was done, how it was done and what has been achieved.

Contents

Chapter 1

Introduction

1.1 Introduction

Introduce the project briefly.

1.2 Aims and Objectives

Mention what you aim to do for project. For every aim you have you must also mention what are the objectives that you must achieve in order to fulfil the aim.

1.3 Structure of the thesis

Explain to the reader what is included in the remainder of the thesis.

1.4 Summary

Summarise what was said in this chapter and link with the next chapter.

Chapter 2

Background

2.1 Cloud computing overview and definition

Cloud computing is commonly seen as a revolution in the world of distributed computing, offering easy online access to storage, applications, and other utilities, making it very attractive to individuals and enterprises alike. This concept of accessing computing resources without a local IT infrastructure took the world by storm around the early 2000s, when "the cloud" started becoming a well-known term. The use of cloud services has been rapidly increasing ever since, with the worldwide cloud computing market size expected to be around $500 billion in 2022 [4] and most IT decision-makers claiming that, in a few years, 95% of all workloads will be carried out in the cloud [3].

So, what exactly is cloud computing? While no common conclusion has been reached among experts in the field [29], the National Institute of Standards and Technology (NIST) [24] at the U.S. Department of Commerce has proposed a comprehensive formal definition that seems to be in line with many people's perceptions regarding Cloud Computing.

NIST states that:

> *Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

Five elements that characterize cloud computing are also mentioned [24, 21, 29]:

I *On-demand self-service:* A customer can gain access to resources like storage and applications automatically when needed, without interacting with the Cloud Service Provider (CSP)

II *Broad network access:* Computer resources can be conveniently provisioned online via mechanisms that allow the use of (heterogeneous) client devices (e.g., laptops, smartphones etc.)

III *Resource Pooling:* CSP resources are gathered in a common pool using virtualization, allowing multiple customers to share a single resource (multi-tenancy) and dynamically assigning physical and virtual resources to match demand. Using such a pool-based model has many benefits like higher economies of scale, speed and availability but users do not control or are aware of where the provisioned resources are located. In some cases, the location at a high abstraction level (i.e., country, state etc.) can be specified.

IV *Rapid Elasticity:* Access to cloud capabilities, which seem limitless to the consumer, does not require any up-front commitment as they can be elastically availed and released exactly when scaling up or down is deemed necessary according to demand.

V *Measured Service:* Despite the fact that customers share a common pool of resources, there are mechanisms in place to measure individual usage for each customer, hence providing transparency to both CSPs and consumers and allowing for resource control and optimization. Accordingly, cloud customers are commonly charged with a pay-per-use pricing model.

2.2 Service Models

Cloud providers mainly offer three service models to consumers: [24, 21, 22]:

Software as a Service (SaaS) SaaS allows the customer to use an application provided and hosted online by a CSP. The application can be accessed using client devices via either a program interface or a thin interface like a web-browser. SaaS consumers do not have any control over the underlying cloud infrastructure, with the exception of some possible configuration changes that can be made to the application. Some characteristic examples of SaaS include SalesForce and Zoom. [10, 16]

Platform as a Service (PaaS) A customer using PaaS can access a development environment to build and deploy applications using services, tools, programming languages and more provided by the CSP (or by another source). Once again, the customer does not control the cloud infrastructure but can control any applications they deployed and can possibly make configuration changes to the hosting environment. An example is Google's App Engine [7].

Infrastructure as a Service (IaaS) With IaaS, the customer can directly control IT resources such as networks, servers and storage. The customer has control over storage, OSs and deployed apps and possibly restricted control over host firewalls and other networking elements. Amazon's Elastic Compute Cloud (EC2) is an IaaS example. [1]

2.3 Deployment Models

There are three common types of cloud deployment models: [24, 21, 12]

Public cloud This type of cloud, which is the most popular choice, is openly available for any individual to access. The infrastructure is owned and managed exclusively by the cloud provider, and is located on their premises. Public clouds often involve lower costs but access to computing resources is shared with other users (tenants). An example of public cloud is Microsoft Azure.

Private cloud The resources offered by private clouds are exclusively dedicated to and used by a single organization using a private network and the infrastructure can be hosted on-site or on a third-party's premises. This model allows more flexibility when customizing the environment and is more secure, as there is no multi-tenancy.

Hybrid cloud Hybrid clouds involve a combination of different environments. The most common example is a combination of a public cloud and a private (on premises infrastructure) cloud. With hybrid clouds, companies can use on-premises infrastructure for workloads where privacy and low latency is essential, while also scaling up as needed by availing extra resources from the public cloud.

2.4 Challenges of Cloud Computing

While the cloud can offer organizations a great many benefits such as elasticity, cost-efficiency, and convenience for accessing computing resources, adopting the cloud also involves certain challenges that prospective customers need to be aware of: [21, 17]

Security One of the biggest deterring factors when it comes to cloud computing is the issue of security and confidentiality. After all, using the cloud typically involves sharing data to an unknown third party and becoming exposed to both external and internal threats. External dangers are no different than those that any large data center faces but, in the cloud, security responsibilities are distributed among different parties (users responsible with application security, while CSPs responsible with physical security etc.) In addition, cloud consumers face internal threats, namely the other users. Any error or security hole during the virtualization process might allow one user to access sensitive data of another and reputation fate sharing is also a possibility if resources are shared with someone with a criminal mind. Finally, users often rely on contracts to protect themselves against provider malfeasance but accidents can still occur (e.g., data permissions bug rendering sensitive information visible)

Availability and Performance Nowadays, CSPs promise extremely high availability for their services as part of the Service Level Agreement (SLA) (Azure SLAs claim up to 99.99% availability, or around 1 hour downtime per year [20]). However, with the complexity of cloud computing systems, even the biggest CSPs cannot guarantee that unforeseen outages will not occur, which could devastatingly affect an organization if it occurs during a period of peak traffic. Furthermore, performance unpredictability is also a big challenge due to the heterogeneous nature of cloud environments [28], where the machines on which customers host their applications, data etc. can be significantly different from each other (e.g. in terms of I/O performance).

Vendor Lock-In Another challenge that cloud consumers often face is vendor lock-in, which occurs when they become dependent on a single provider's infrastructure and cannot move to another without significant cost, compatibility issues and legal considerations. The factors that contribute to vendor lock-in include proprietary Application Programming Interface (API)s and lack of open standards for VN and data interchange formats and service deployment interfaces. Consequently, the customers become vulnerable to any changes the providers make regarding their services (e.g. payment model). [26]

Migration Planning Migrating to the cloud can greatly benefit an organization by relieving them of the cost and time necessary for setting up and maintaining infrastructure and by offering great flexibility, scalability and convenience for provisioning computing resources. It must be mentioned, though, that before migrating to the cloud, extensive planning must be carried out to ensure that it is a worthwhile and cost-efficient endeavor, otherwise failure is highly likely [18]. Specifically, [23] states that the company must first analyze their IT infrastructure and determine its complexity, size and quality, so that a decision can be made on what to migrate and in what order. Additionally, meetings with stakeholders and employees are required to determine the operational requirements like how long the downtime during the migration can be and if the firm's network can handle the process. Assets will need to be categorized based on their criticality to decide on which will remain in-house for better security. It is also essential to carefully decide which of the 6 migration strategies to use for each application (rehosting, replatforming, repurchasing, refactoring, retiring or retaining) [27]. Finally, the management costs before and after the migration will need to be calculated and compared, and a document detailing potential obstacles should be prepared.

2.5 Summary

Summarise what was said in this chapter and link with the next chapter.

Chapter 3

Cloud Storage Services Comparison

3.1 Introduction

Cloud storage solutions are rapidly rising in popularity among individuals and enterprises, as the benefits of using such services in day-to-day life and business become more and more evident. Cloud storage applications like Google Drive are used by almost 40% of households in certain areas while, by 2025, 100 *Zettabytes* of data (50% of all data) will be stored in the cloud [25].

Customers nowadays have dozens of options when it comes to storage providers, some of the most popular ones being established providers like Dropbox [5] and tech giants such as Amazon and Microsoft, which offer their Simple Storage Service (S3) and Azure Blob services [2, 9], respectively. All storage services, however, differ in certain characteristics like pricing, capacity and performance. It becomes important, then, to investigate these differences in order to make an educated decision on which service to adopt.

The first distinction that needs to be made is between cloud *object* storage and cloud *file* storage [14, 15]:

Object storage This technology allows data to be stored and managed in an unstructured format called *objects*. Objects include the data that make up a file, user-created metadata and a unique identifier. They are stored in a flat data environment which enables fast scaling and application can easily retrieve objects of any data type, using the metadata and the identifier.

As businesses expand, they are tasked with handling larger and larger volumes of data from a variety of sources that is used by both applications and users. This data is often unstructured and in many different formats and storage media, making it hard to store it in a central repository. Therefore, cloud object storage can solve this issue as it offers vast scalability and cost-efficient storage tiers for storing all kinds of data natively in a single virtual repository, accessible from anywhere. An example of cloud object storage is Amazon S3 [2]

Some common use cases for object storage are:

- ***Data lakes***, which are centralized repositories for storing data of any scale, structured or unstructured, rely on object storage to operate.[**?**] They offer encryption, access control and great scalability, allowing for easy and dynamic storage expansion of up to petabytes of content with a pay-as-you-go charging model.
- The virtually limitless data that is stored in cloud object storage is often used by businesses to perform ***big data analytics***, in an effort to better understand their customers, operations and market.
- Cloud object storage provide flexible data storage for ***cloud-native applications***, accessible via an API, thus aiding developers and allowing for easier and faster deployment.
- An excellent use of object storage is ***data archiving***, replacing on-site archive infrastructure and providing enhanced data security, durability, accessibility and near-instant retrieval times.
- Object storage systems can used to replicate data across many systems, regions and data centers, ensuring that ***backups*** are ready to be used for recovery from hardware failures.
- Replication on a global scale can be achieved with cloud object storage, vastly reducing the storage costs and increasing availability for ***rich media*** like music, videos and digital images.
- Since ***Machine Learning (ML)*** models generate inferences after being trained over billions of data items, object storage becomes a necessity to handle such scale in a cost-efficient manner.

File Storage In contrast to object storage, file storage utilizes a hierarchical structure for storing data, where files are stored inside of folders which are grouped into directories and so on, and strict protocols like Network File System (NFS) are used. Consequently, it is more difficult to locate a specific piece of data among billions, compared to object storage, where each object is uniquely identified. In addition, the inherent hierarchy and pathing limits the potential scalability of cloud file storage, whereas object storage provides near-limitless scaling.

File synchronization is one of the major features that have contributed to the popularity of these systems. File storage allows users to automatically synchronize their files across a variety of devices, making it very convenient and simple to continue their work as they left it. Another popular use case of file storage collaborative work, as the most up-to-date version of a file can be easily shared among a group of users.[19] The user interface and experience of cloud file storage services are also significantly more user-friendly, making them suitable for casual and experienced users alike. An example of cloud file storage is Dropbox. [5]

3.2 Object Storage Services

Two of the most popular object storage services are Amazon S3 and Microsoft Azure Blob Storage. [2, 9]

In ***Amazon S3***, objects are stored in buckets, and prefixes (shared names) are used to organize the latter. **S3 object tags**, which are a set of up to 10 key-value pairs, can also be appended to objects to make storage management easier, as these tags can be added and edited at any point in the object's lifetime for stronger access control and other use cases. In addition, **S3 inventory** reports can be generated manually or automatically to provide information on the objects in a bucket, to make working with lots of data easier and to allowing gather information on encryption, replication and more. Amazon provides an online Management Console, a Representational State Transfer (REST) API, a Command-Line Interface (CLI) and a Software Development Kit (SDK) in various languages for accessing S3 features. [6]

Microsoft Azure *Blob Storage* uses three types of related resources. The first are storage accounts, which provide a unique namespace for the user's data and is used in the address of each object. There are different kinds of storage accounts, each tailored for different uses and with different redundancy options: [11]

1. *General-purpose v2* is the standard account for storing blobs (data). It is recommended for most use cases.

2. *Premium block blobs*, for block blobs and append blobs. Makes use of Solid-state drive (SSD) technology, making it ideal for workloads that require low latency or that involve many transactions.

3. Premium page blob, which is similar to a premium block bob account, but it's only used for page blobs.

Containers are the second resource, which organize a set of blobs. An unlimited number of containers can be included in a storage account, and an unlimited number fo blobs can be stored in a container. The third and final resource kind are blobs,of which there are three types:

1. *Block blobs* are made up of blocks of data, with each block being uniquely identified. Block blobs are ideal for large amounts of text or binary data, as up to 50.000 variable size blocks can be included in a single block blob.

2. *Append blocks* are also made up of blocks of data, but they are optimized for append operations, making them ideal for use cases like Virtual Machine (VM) data logging.

3. *Page blobs* are composed of 512-byte pages and are suitable for random write and read operations. They support Azure virtual machines by acting as disks

As with S3, to help the management of large pools of data, Blob storage provides the ability to use **blob index tags**, key-value pairs that can be used to categorize blobs, find specific blobs across an entire account, set permissions and more. With **Azure Cognitive Search**, blob content are imported as search documents that are indexed, allowing for the information stored in the blob data themselves to be searched. Also, with the help of Artificial Intelligence (AI) enrichment, text from images can be extracted. Furthermore, **Blob inventory** automatically generates report with user-defined rules to provide an overview on containers, snapshots, blobs, blob versions and their properties. Blob storage object are accessible via REST API, Azure PowerShell (a set of cmdlets fo Azure resources)[13], CLI or a client library in one of many languages. [8]

3.2.1 Storage Classes

3.3 Conclusions

Chapter 4

Analysis and Design

4.1 Introduction

Introduce what this chapter is going to present.

4.2 Next section title

4.3 Next section title

4.4 Summary

Summarise what was said in this chapter and link with the next chapter.

Chapter 5

Implementation and Testing

5.1 Introduction

Introduce what this chapter is going to present.

5.2 Implementation

5.2.1 Next subsection

5.2.2 Next subsection

5.3 Testing

5.3.1 Next subsection

5.3.2 Next subsection

5.4 Summary

Summarise what was said in this chapter and link with the next chapter.

Chapter 6

Evaluation

6.1 Introduction

Introduce what this chapter is going to present.

6.2 Discussion

6.3 Difficulties phased

6.4 Knowledge acquired

6.5 Future work

6.6 Conclusions

Bibliography

[1] Amazon ec2. https://aws.amazon.com/ec2. Accessed 17-October-2022.

[2] Amazon s3. https://aws.amazon.com/s3. Accessed 20-October-2022.

[3] Cloud 2025: The future of workloads in a cloud-first, post-covid-19 world. Accessed 23-October-2022.

[4] Cloud computing market size, share and trends analysis report by service (iaas, paas, saas)... https://www.grandviewresearch.com/industry-analysis/cloud-computing-industry. Accessed 23-October-2022.

[5] Dropbox. https://www.dropbox.com. Accessed 20-October-2022.

[6] General s3 faqs. https://aws.amazon.com/s3/faqs/. Accessed 30-October-2022.

[7] Google app engine. https://cloud.google.com/appengine. Accessed 17-October-2022.

[8] Introduction to azure blob storage. https://learn.microsoft.com/en-us/azure/storage/blobs/storage-blobs-introduction. Accessed 30-October-2022.

[9] Microsft azure blob. https://azure.microsoft.com/en-us/products/storage/blobs/#overview. Accessed 20-October-2022.

[10] Salesforce. https://www.salesforce.com/eu/. Accessed 17-October-2022.

[11] Storage account overview. https://learn.microsoft.com/en-us/azure/storage/common/storage-account-overview?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json. Accessed 30-October-2022.

[12] What are public, private, and hybrid clouds? https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-private-public-hybrid-clouds. Accessed 23-October-2022.

[13] What is azure powershell? https://learn.microsoft.com/en-us/powershell/azure/what-is-azure-powershell?view=azps-9.0.1. Accessed 30-October-2022.

[14] What is object storage? https://aws.amazon.com/what-is/object-storage/. Accessed 26-October-2022.

[15] What is object storage? https://cloud.google.com/learn/what-is-object-storage. Accessed 26-October-2022.

[16] Zoom. https://zoom.us. Accessed 17-October-2022.

[17] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, 2010.

[18] D. Aspinall. Cloud migration risks: 5 reasons why cloud migrations fail. https://www.jeffersonfrank.com/insights/top-reasons-cloud-migrations-fail. Accessed 19-October-2022.

[19] E. Bocchi, I. Drago, and M. Mellia. Personal cloud storage benchmarks and comparison. *IEEE Transactions on Cloud Computing*, 5(4):751–764, 2015.

[20] M. Brandwin. How azure slas can help you to achieve optimal uptime and availability. https://www.viacode.com/azure-slas-for-uptime-and-availability. Accessed 23-October-2022.

[21] T. Dillon, C. Wu, and E. Chang. Cloud computing: issues and challenges. In *2010 24th IEEE international conference on advanced information networking and applications*, pages 27–33. Ieee, 2010.

[22] I. Foster, Y. Zhao, I. Raicu, and S. Lu. Cloud computing and grid computing 360-degree compared. In *2008 grid computing environments workshop*, pages 1–10. IEEE, 2008.

[23] H. Malouche, Y. B. Halima, and H. B. Ghezala. Enterprise information system migration to the cloud: Assessment phase. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE, 2016.

[24] P. Mell, T. Grance, et al. The nist definition of cloud computing. 2011.

[25] S. Morgan. The world will store 200 zettabytes of data by 2025. `https://cybersecurityventures.com/the-world-will-store-200-zettabytes-of-data-by-2025/`. Accessed 20-October-2022.

[26] J. Opara-Martins, R. Sahandi, and F. Tian. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. *Journal of Cloud Computing*, 5(1):1–18, 2016.

[27] S. Orban. 6 strategies for migrating applications to the cloud. `https://aws.amazon.com/blogs/enterprise-strategy/6-strategies-for-migrating-applications-to-the-cloud/`. Accessed 19-October-2022.

[28] B. Teabe, P. L. Wapet, A. Tchana, and D. Hagimont. Dealing with performance unpredictability in an asymmetric multicore processor cloud. In *European Conference on Parallel Processing*, pages 332–344. Springer, 2017.

[29] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner. A break in the clouds: towards a cloud definition, 2008.

Appendices

Appendix A

Write a few words about what is included in this appendix

Appendix B

Write a few words about the contents of the appendix

Manuals

Installation Manual

User Manual