Automation in cloud environment using cloud services and python script

MSc Research Project
MSc Cloud Computing

Mukul Anaspure
x21118141

School of Computing
National College of Ireland

Supervisor: Shivani Jaswal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mukul Anaspure		
Student ID:	x21118141		
Programme:	MSc Cloud Computing	**Year:**	Jan 2022
Module:	MSc Research Project		
Supervisor: **Submission Due Date:**	15th December 2022		
Project Title:	Automation in cloud environment using cloud services and python script		
Word Count: **6285**	**Page Count** **18**		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mukul Anaspure
Date:	15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Automation in cloud environment using cloud services and python script

Mukul Anaspure
x21118141

Abstract

Data backup and business continuity issues are crucial in networks as the value and importance of digital data are constantly increasing. It is vital to frequently backup the data but also at the same time, it is also necessary to keep the storage costs in mind. In the world of cloud computing, the data backups are handled with the help of policies that are provided by the cloud service providers like Azure, AWS. But if we consider unplanned backup requests or adhoc backup requests there is a probability of missing these requests and causing a disaster to business in the near future. In addition, whenever the data is backed up there are chances of similar data being uploaded to storage accounts causing risk of increased redundancy, high bandwidth and high storage costs and usage. This report mainly highlights the strategy to deal with such backup requests where the entire process of a user triggering the backup from cloud portal is automated with the help of automation tools like selenium web driver, scripting language and checking the duplicate multimedia files mainly audio files on a server. This research report will thus allow organizations to adopt such automotive strategy, saving organizations and users valuable time and money.

1 Introduction

Business continuity is crucial since a service disruption can have a negative impact on a company's objectives and cause significant losses in revenue, brand reputation, and market share. As a result, it is becoming increasingly important to create disaster recovery plans and backup data (Suguna and Suhasini, 2015). Regular data backups enable clients to recover their data in the event that a server crash or other incident occurs. According to recent statistics, the actual average cost of a data breach occurrence is $4.24 million. (Zhang *et al.,* 2014) In today’s world, many businesses use the cloud as their backup solution because cloud service providers store their client’s data on off-site servers in a secured environment. The backups can be scheduled with the help of backup policies, which may differ from server’s environment to environment. Users do not need to be concerned about their regular backups because, once the policy is defined for a specific server, the backup will be initiated at the designated time, regardless of any changes to the server's data. Here, there are two possibilities: Firstly, what if there are any unscheduled backup requests or unplanned backup requests. Secondly, there are chances of duplicate or similar data uploaded to cloud storage which leads to increasing the storage redundancy and storage costs.

Before continuing further, unscheduled, or unplanned backup requests involve the capturing of data at a specific time. For instance, whenever a client needs to perform an activity on a certain server from their end, they can ask the hosting team to initiate the server's backup from the cloud interface. The client's activity can be delayed if the team doesn't receive the mail, which might interfere with other scheduled operations. To reduce the storage redundancy and save storage costs data deduplication is a technique used to lower storage needs by removing the redundant data. Multimedia files, such images, audio, and videos, are well-known for being substantially larger in size and requiring more storage than standard files. Therefore, it is not feasible to have duplicate copies of these multimedia files. Multimedia files can benefit from the deduplication approach, which increases storage effectiveness. This report concentrates on the effectiveness of audio file storage in the cloud among multimedia files.

The standard procedure used by businesses to handle these ad hoc or unscheduled backup requests is entirely manual. It demands work from people. An engineer must trace such client emails before going to the cloud site and beginning the backup. Before responding to the client's email, it must be observed till it is finished after being triggered. If the backup fails, he must troubleshoot, retrigger the backup, watch it until it is finished, and then send an email of confirmation to the customer. This requires a lot of human effort, which wastes important corporate time. Also, this procedure can be automated which can help the business and employees save their valuable time. In addition, whenever the data is being stored on the cloud, there is a need to check the uniqueness of data because duplicate data are a major worry if they are present in a large volume of data in any type of storage, due to storage issues. A survey found that the amount of data in the cloud in 2021 will be between 40 and 50 trillion gigabytes, which presents several big data difficulties such as high volume, high velocity, and variety of data in the form of text, PDFs, photos, and videos. Numerous clients of various organizations export a significant amount of data that is identical. Duplicate data reduce the storage's effectiveness. The elimination of redundant data is crucial for cloud storage optimization. The only way to deal with this massive data increase is through the detection of duplicate data and its localization. Many third-party tools are used in many existing ways for this, but because the scalability of data is growing every day, existing tools are inadequate for offering the best answer.(Higazy et al., 2013)

The significance of the research is to keep the data on cloud storage in unique state and reduce the redundancy of storage and save the storage costs. The technique used for keeping the data unique is called data deduplication. This technique is used to increase the cloud storage efficiency. The previous studies were done on the multimedia files like video, text, image so the following research is done on audio file. Whenever an audio file will be uploaded to cloud storage then its uniqueness will be checked first and then only it will be uploaded to the cloud storage. By using this approach, the data will remain unique and there will be more space available on disk to store new data. As we know, data is increasing day by day hence this technique will be useful for the businesses to save their expenses on storage units. Also one more significance of the research is to save the valuable time of an employee and business by automating the unscheduled backup requests. As mentioned in above section, the entire manual process of handling such backup requests can be automated which will allow the business to invest their time in other valuable activities.

1.1 Research Question

To what extent automation can make enhancement in manual handling of unplanned backup requests and keeping the storage units efficient by eliminating duplicate copies of data.

1.2 Limitations

The main limitation of the current system is wasting time and money on storage units. The task of handling unscheduled backup requests is a repetitive task and can be automated which will save valuable time of the organization. Also, when the data is backed up it is important to check the uniqueness of data. If the duplicate copies are stacking up in the storage units, then it will definitely incur additional costs to the organizations. The current system (Leekha and Shaikh, 2021)focuses on the files like text files, videos or images but there is no such system for audio files where we can the uniqueness in audio files. The proposed system will help to overcome this limitation.

The report is divided into 6 sections. Section 1 has introduction to the research topic. Section 2 is all about related work about the topic. Section 3 and 4 highlights about the methodology and design specifications used in this research. Also, justification of used services is discussed in this section. Moving forward to section 5 which is about Implementation. How the research idea was actually implemented is discussed here. Section 6 is about the evaluations and experiments which are performed. Finally section 7 discusses about the conclusion and future work.

2 Related Work

The following section focuses on the previous work done on the different types of backup methods and data deduplication methods. In addition, the limitation of the related work is also mentioned in the section.

When it comes to automatic data backup and restoration, the field of cloud computing adds more value to its significance. In order for a business to operate efficiently, it must ensure that its data is regularly backed up and that it is kept in an unique state in order to save costs and the stress on storage units. The intervals vary depending on how crucial the business is. Data backups might be done every day, every week, or every month. This will benefit the company in situations where such sensitive data may be highly risky.(Menard, Gatlin and Warkentin, 2014) With the aid of restore points, the cloud service providers can assist in returning the business to normal while preventing further disruption to the customers.

Organizations need to pay attention to their routine data backup and disaster recovery strategy because disasters are occurring more frequently and with greater intensity. While no organization can completely prevent disasters, prudent planning can always help to lessen their impact. Cloud backup and disaster recovery aid in the continuance of the business in the event of a disaster in the world of IT, where the entire organization is dependent on the data. (Hamadah, 2019)

Although data backup is regarded as an essential component of all IT firms worldwide, many of them neglect to keep frequent backups. As a result of failing or nonexistent backups, many firms suffer data loss. Please take note that the reasons backups fail are covered in the next sections. Apricorn, a hardware company, conducted a study in October 2021 that found that approximately 60% of IT organizations did not intend to back up their company files and that more than 50% of IT organizations had experienced data loss. On the basis of a social media platform poll, the study was carried out (IT Pro,2021)Backing up is not a person's obligation, according to Jon Fielding, the managing director of Apricorn's EMEA region. Regardless of rank, it should be taken care of by every employee working for a specific company. Every person is accountable for their data backup, which minimizes downtime, protects finances and reputations, and enables a far faster response to a crisis situation leading to full restoration and recovery. (IT Pro,2021)

2.1 Methods of creating backup in cloud

An innovative notion for file backup is the HS-DRT, which combines a fast encryption technology with a successful ultra-widely dispersed data transmission technique. There are three components to it: The first three critical functions are the Data Centre, Supervisory Server, and components are involved nodes that admin has selected. The client nodes are made up of computers, mobile devices, Network Attached Storage, and storage services. They are connected to the Data Centre and a supervisory server through a secure network.(Ueno et al., 2010) The basic operation of the suggested network system is split into two sequences, the first of which is backup and the second of which is recovery. When the Data Centre receives the data to be backed up, it scrambles, fragments, and duplicates the data to varying degrees in order to achieve the required recovery rate in accordance with the established service level. The backup sequence is the name of this procedure. The second stage involves the Data Centre re-encrypting the fragmentations before distributing them to the client nodes in a random order. Additionally, the Data Centre provides the metadata required to successively decode the pieces.

When a service failure is detected, the Cold Backup Service Replacement Strategy (CBSRS) recovery process is started; it is not started when the service is available. A transcendental recovery technique is used in the Hot Backup Service Replacement Strategy (HBSRS) for service composition in dynamic networks. (Sun et al., 2011) It restores the service composition dynamically based on availability and the current state of the service composition prior to the services interruption. The backup services always remain in the activated states during service implementation, and the initial returned results of services are then adopted to guarantee the effective execution of service composition. The HBSRS lowered service recovery time when compared to the CBSRS.
However, when original services and backup services are run concurrently, the recovery cost rises correspondingly.

Technique mainly focuses on the router failure situation and significant cost savings (SBBR). The most crucial element it provides the network management system via multi-layer signalling is IP logical connectivity, which endures even after a router failure. However, there are certain discrepancies between logical and physical setups that could cause some performance issues when it comes to the cost reduction notion. Additionally, it demonstrates in(Palkopoulou, Schupke and Bauschert, 2011) how service-imposed maximum outage requirements directly influence the SBRR architecture's setup but it cannot combine the optimization principle with cost cutting.

Since cloud computing is still in its infancy, there is a lack of standardization and a desire to precisely define its essential components. The term "cloud computing" refers to a model for providing ubiquitous, practical, on-demand network access to a pool of configurable computing resources (such as networks, servers, storage, applications, and services) that can be quickly provisioned and released with little management work or service provider interaction. This definition was recently provided by NIST in a special publication 800-145 (Mell and Grance,). A cloud service should have the ability for on-demand self-service, which enables a user to unilaterally provision computing resources like server time and network storage as needed automatically without requiring human interaction with each service provider. This is one of the key characteristics outlined by NIST.

2.2 Data Deduplication techniques

The problem of big data, involving variety, velocity, veracity, and volume, is brought on by the collection of data from many devices. Deduplication, or equivalent data detection and removal, is a resource-appropriate and cost-effective approach.

The method proposed by (Zhang *et al.*, 2014)which includes detection of reuse of text at semantic level for continuous word embedding.
(Higazy *et al.*, 2013)has discussed various deduplication techniques for Arabic languages in which quality and complexity measures were highlighted. (Elmagarmid, Ipeirotis and Verykios, 2007)explored ways to increase the effectiveness and scalability of these algorithms as well as deduplication approaches and similarity measures used to identify similar terms. The author (JACK E) suggests a novel strategy called data profiling to improve the accuracy in huge datasets. (Cohen and Richman, 2002)combined web databases and discovered similarities between the linguistic names given to items across several datasets.
Nearly every location where data is stored or delivered in cloud storage can use data deduplication (Halevi *et al.*, 2011)Numerous cloud service providers offer disaster recovery (Javaraiah, 2011), and deduplication can be utilized to speed up replication time and reduce bandwidth costs by replicating data after deduplication. Data deduplication can also be used for backup and archive storage in the cloud to lessen physical capacity and network traffic (YOU, POLLACK AND LONG, 2005)(Tan *et al.*, 2010)
Deduplication can be used to store active data, like virtual machine pictures, in less space. How to strike a balance between performance impact and storage space savings are important considerations when employing deduplication in main storage (Wang *et al.*, 2012) Deduplication techniques, according to Mandagere, et al. represent the efficiency of deduplicated storage in terms of fold factor, reconstruction bandwidth, metadata overhead, and resource consumption. Data in cloud storage is dynamic by nature (Wang *et al.*, 2012) (Yang and Jia, 2013)For instance, the way that data is used in the cloud varies over time; some data chunks might be commonly read at one point in time but not at another. While some datasets may need a high amount of redundancy due to dependability requirements, others may be routinely accessed or modified by several users concurrently. Support for this dynamic feature in cloud storage is therefore essential. However, the majority of current techniques are static in nature, which restricts their full applicability to the dynamic nature of data in cloud storage.
Data deduplication technology, sometimes referred to as clever compression, single-instance storage, data reduction, and capacity optimized storage, is used to lessen the storage strain. (Geer, 2008)This deduplication technique compares and identifies data pieces using hashing. Comparing just a portion of the dataset is quicker and less effective than hashing, also known as fingerprinting. It employs the SHA-1 algorithm to build the hash value, which produces a 20-byte hash value to compare to the data already there. Logical pointers are made of the duplicate data if the comparison is false; otherwise, it assumes the data is unique and stores it by adding the hash information to the system.
The principal type of data deduplication is the topic of this research. (Ahmed El-Shim,2012), SHA-1 and the Rabin fingerprint are used to accomplish this. The file is divided into pieces of varying sizes using the Rabin fingerprint algorithm, and the chunked files are hashed using SHA-1. This system is solely designed to work with Windows operating systems. In this research (Ahmed El-Shim,)a private data deduplication protocol-based deduplication

technique for private data storage is developed and defined. In the simulation-based approach, the proposed private data deduplication protocol is demonstrably secure under the assumption that the underlying hash function is collision-resistant.

3 Research Methodology

The methodology for the research is described in this section. The research question mainly focuses on automation in cloud infrastructure. Scripting is a best possible option to automate the repetitive tasks in cloud environment. This research is regarding automating the backup requests in Azure cloud and finding the duplicate files on a sever. Python scripting is used in this research along with Selenium. When it comes to testing your scripts then Selenium is best possible option to test out the cases.

In addition, when there is a need to perform automation in cloud environment especially Azure cloud then Azure Logic Apps is the best option. It allows the users to design their workflows according to their needs. Moreover, Azure Logic Apps has an advantage of easy integration with other Azure services like Azure Functions. Azure Functions is an important component in this research. Basically, azure function is a trigger-based service. It can run a script based on variety of events without worrying or managing the infrastructure. But considering our research, Selenium is an important component and default azure functions do not support the selenium libraries. Therefore, a docker image has been created which consists of all the required selenium libraries and pushed to Azure Container Registry. And finally, an Azure function is created with the help of this docker image. Now the azure function can be used for Selenium python-based scripts. Following figure shows the general idea of this.

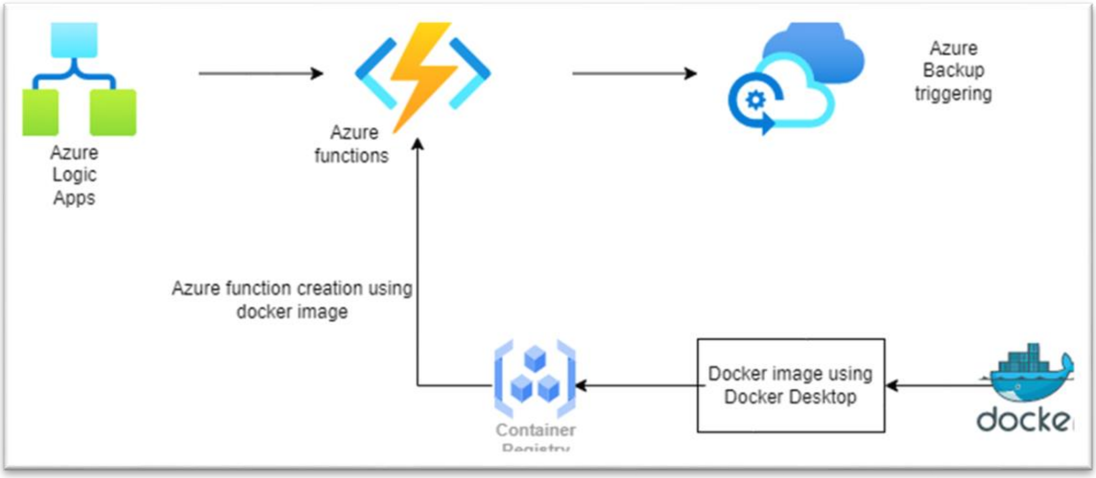


Figure: Architecture diagram for Backup Automation

For finding duplicate files from server python script is used along with hashing concept. Its very important to know unique hash values of each files and then compare it with each other. This will find the duplicate files on the server depending on the content of file. The script is scheduled on a server with the help of Windows Task Scheduler as this service is a free service for all windows users and best available option when there is a need to perform the automation on a server level.

3.1 Why Azure?

The cloud platform which is selected in this research is Microsoft Azure. The straightforward answer to this is Azure is more popular than other public cloud providers. It has many services and advantages that encourages the organizations to opt for Azure cloud.
In addition, more individuals may relate to cloud services/technologies through Azure, which offers a wider market reach. Azure is the undisputed winner in terms of working standards, principles, and even best-practices. When discussing Azure technologies in comparison to those of other cloud providers, a larger audience can be reached.

4 Design Specification

The design specifications of this research is discussed in the following section. The process workflow and the overview of tools and technologies used is also discussed in the below sections.

This study contributes to the automation of handling unscheduled backups and finding the duplicate records, files to save valuable time and storage cost of organizations The manual handling of backup requests is a repetitive task that consumes human effort. This task could be difficult to carry out every day if we perform it manually. Therefore, there are several techniques to search for automation. For automation process Azure logic app, Azure Functions are used to create a process flow. Python scripting language is used to automate the backup process and detect the duplicate records, files. Figure 1 shows a general picture of components in the Azure Logic App.

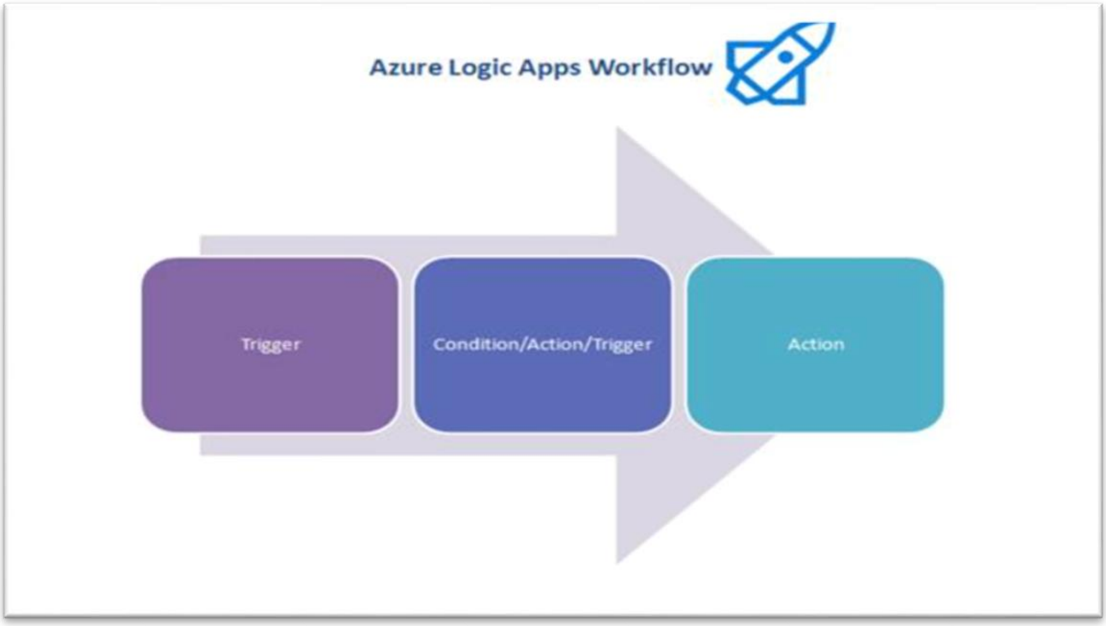


Figure: Overview of Azure Logic Apps

In this research, the trigger has been generated when an email comes in outlook application. In the next stage, the condition is mentioned of having an attachment in the mail body. If the condition is true, then the python script of triggering the backup is triggered.

4.1 Approach:

The approach followed in the research which directly affects the question is the use of ‘Azure Logic Apps’, which is feasible to the research description and the use case for the research. This methodology enables the user to automate your workflows with little to no code. With the help of this methodology, the tasks which occur on a repetitive basis in an organization can be automated by monitoring and integrating other services. With the help of a visual designer provided by the Azure Logic Apps service, the user can create their logical workflow according to their use cases. The workflows can vary from each other use cases. Considering the popularity of cloud computing services and their increasing demands in the IT sector, public cloud providers and third-party service providers have their alternatives to this service.

- Azure Logic Apps - On the cloud platform Azure Logic Apps, automated workflows may be built and executed with little to no coding. You can easily create a workflow that controls and combines your apps, data, services, and systems by utilizing the visual designer and prebuilt operations. (Logic App Service – IPaaS | Microsoft Azure,)
- Amazon Step Functions - In order to build distributed applications, automate procedures, orchestrate microservices, and build data and machine learning (ML) pipelines, developers can use AWS Step Functions, a visual workflow service.(Serverless Workflow Orchestration)
- Google Cloud Composer – The user can author, plan, and monitor pipelines that span hybrid and multi-cloud environments with the aid of Cloud Composer, a fully managed workflow orchestration solution based on Apache Airflow. The most recent version of Cloud Composer features autoscaling, which offers workflows with bursty execution patterns cost effectiveness and increased stability. (Cloud Composer | Google Cloud,)
- Microsoft Power Automate – This tool is also used for creating automated workflows between various apps and services. The main weakness of power automate is security reasons. The malicious actors can create malicious workflows which are customized with the help of Power Automate Capabilities. (Power Automate: Vectra AI,)

It is visually clear that all the services mentioned above can get the research job done. In the below section, it has been discussed that which service is best to get the job done and why it has been selected in this research methodology.

4.2 Why Azure Logic Apps:

The selection of best possible service by comparing among all the available services in market is very important.

Regarding the selection of service, the Azure Logic Apps is the best possible option for this research. Not only because Microsoft Azure is leading cloud service providers but also it allows integration of various services with each other. The major advantage of using Azure Logic App is because it can easily integrate with Azure Functions. The user can simply call their Azure Functions in their automated workflows and continue their task.

In this research report, the major advantage of using Azure Logic App is enormous and best possible to any other alternative. The basic and first requirement of Azure Logic App is to have a trigger point in the workflow and then the mentioned actions will be performed. The research initial phase is to monitor the incoming mails which is a trigger point for which Azure Logic Apps is the best option. Also, as the workflow moves forward the Azure Functions can be easily integrated in the workflow. This will help the script to trigger and perform the required task. In addition, Azure Log Apps can have multiple triggers as well.

The method selected for the research is best possible option to get the output but there are some limitations of Azure Logic Apps which have been experienced by some users. (Top Azure Logic Apps Likes & Dislikes 2022)According to some users they faced an issue of automation delay and some irrelevant output is displayed which will make the job fail misguiding the process designer. Also, some users complained about the inconsistency in code execution. The same code sometimes take 10-12 seconds for execution and sometimes it will take only milliseconds to execute.

4.3 Advantages:

The organizations can get benefit of this research methodology in reducing the manual efforts which are required to handle the unscheduled backup requests in Azure. Also, this methodology can be used by the users to know about the duplicate data present in the server. The users will get to know how many duplicate files are present in the server. This will allow the organization to save their storage costs and can use the storage units to store only unique values.

5 Implementation

The implementation part is not only restricted to running a python script or PowerShell script, but it is also important to setup an environment considering the limitations of Microsoft Azure. The configuration module which is associated with this report gives a clear idea regarding the implementation part. The main motivation behind this research to automate the process of handling unscheduled backup requests and finding the duplicate files which are present on the server. For handling unscheduled backup requests following are the core steps:

1. Creating an Azure Logic App
2. Preparing a script using selenium
3. Creating an Azure Function that contains the prepared in the above step.
4. Calling the Azure function in Azure Logic App.
5. Sending an automated mail after backup is triggered.

Considering the above steps, it seems to be very easy and straightforward. But there are some limitations of Azure Functions. The default Azure functions do not have the required dependencies that Selenium requires. And Selenium is important part of this research. So, to overcome this a Docker image has been created and deployed as Azure Function with Selenium libraries. In brief, A docker image is created using Docker desktop, which will contain all the libraries required for Selenium. The Docker image is then pushed to Azure Container Registry. And finally, an Azure Function is created and the docker image is deployed from Azure Container Registry.

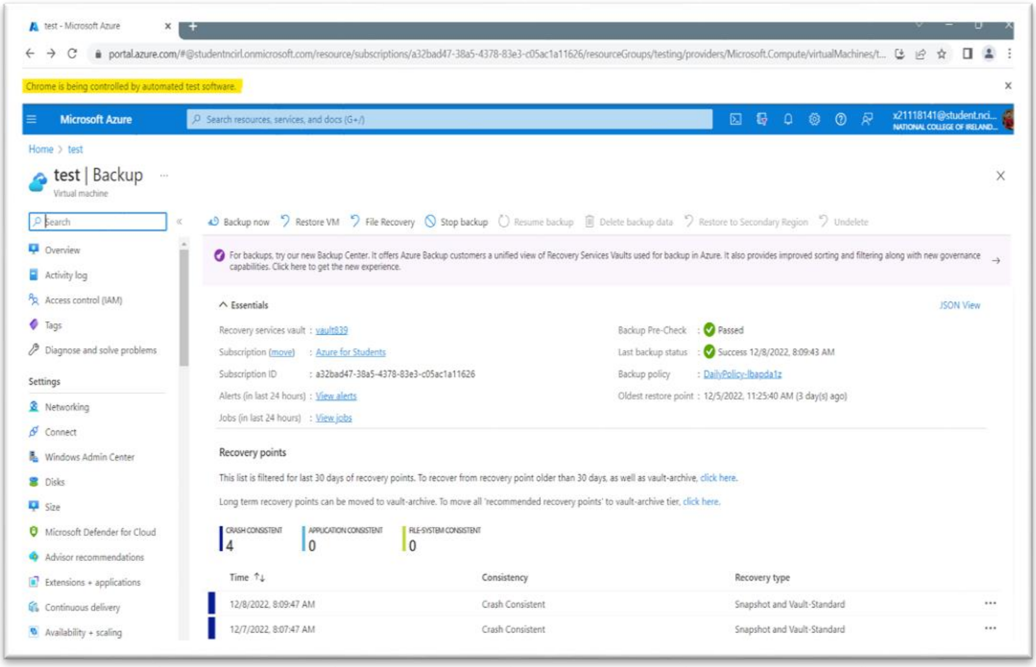
For finding duplicate files in a server, there are many software's in the market which will give the required results. But many of the third-party tools are either commercial or they are not suitable for automatic scenarios. The files should not be compared just based on their names because they could have different names yet the same content. It is preferable to obtain the hashes of every file and look for duplicates among them. In this research, the script is created which will look for duplicate audio files in a specified path. A package called “Tkinter” is used. This package is used for creating Graphical User interfaces. The script is scheduled at a specified time with the help of Task Scheduler. At that specified time, the script will run automatically, and it will pop up and window asking for the folder in which the duplicate values need to be checked. Once the folder is selected then the script will calculate the hash values of each files, files which are there in sub-folders and suggest which files have multiple copies. Once the duplicate files are known, an automated mail will be sent to the user suggesting deleting the required files.

6 Evaluation

In any cloud environment if the work is related to writing scripts and performing automations of daily repetitive tasks then it is always reliable to test the scripts and check if there are any obstacles in user experience. In the below section three experiments which are performed are discussed briefly.

6.1 Experiment 1

The first experiment is to use the script for automatically triggering the server (Virtual Machine) backup operation from the Azure portal. This trigger would happen when a mail is received in the Microsoft Outlook application for a server backup request. The script is written in Python with the help of Selenium Web drivers because it is the best available option for testing your scripts. Eventually, when the mail is received in Outlook Application, then it would start the workflow which we have mentioned in Azure Logic Apps. So, receiving mail in the outlook application is working as a trigger for Azure logic Apps. Figures 2 and 3 below highlights the successful working of the script. Also, it can be seen in the top left side where it says, “Chrome is being controlled by automated software” which means the script is working and there is no intervention or any manual effort. Finally, once the backup is triggered then an automatic mail is triggered to the user saying backup is successfully triggered. So eventually the entire process which was previously done with entire manual efforts has been automated including sending a mail confirmation to the user.



portal.azure.com/#@studentncirl.onmicrosoft.com/resource/subscriptions/a32bad47-38a5-4378-83e3-c05ac1a11626/resourceGroups/testing/providers/Microsoft.Compute/virtualMachines/t...

Chrome is being controlled by automated test software.

test | Backup

Virtual machine

For backups, try our new Backup Center. It offers Azure Backup customers a unified view of Recovery Services Vaults used for backup in Azure. It also provides improved sorting and filtering along with new governance capabilities. Click here to get the new experience.

Essentials

Recovery services vault : vault839
Subscription (move) : Azure for Students
Subscription ID : a32bad47-38a5-4378-83e3-c05ac1a11626
Alerts (in last 24 hours) : View alerts
Jobs (in last 24 hours) : View jobs

Backup Pre-Check : Passed
Last backup status : Success 12/8/2022, 8:09:43 AM
Backup policy : DailyPolicy-lbapda1z
Oldest restore point : 12/5/2022, 11:25:40 AM (3 day(s) ago)

Recovery points

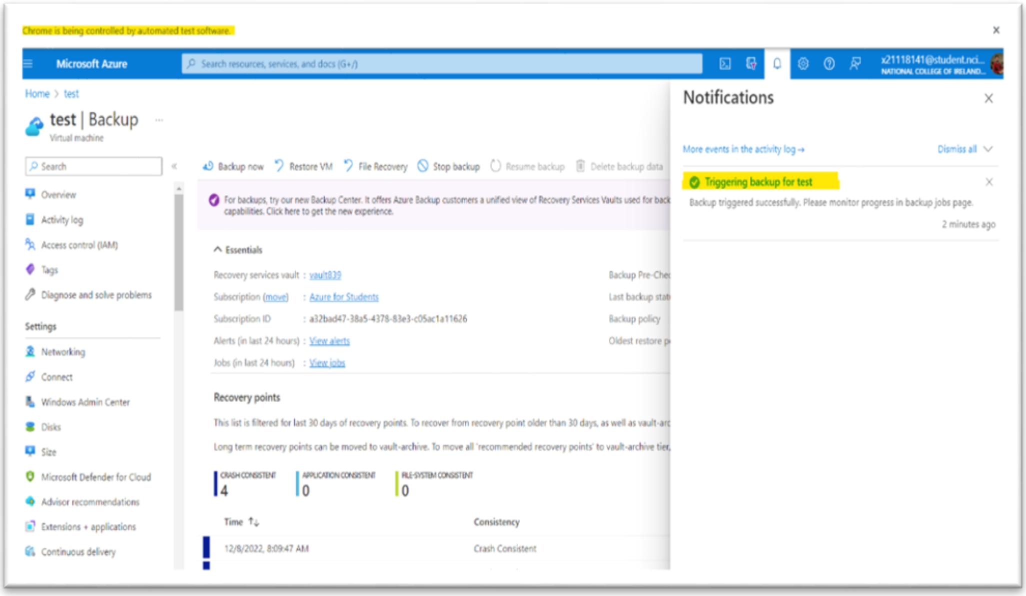
This list is filtered for last 30 days of recovery points. To recover from recovery point older than 30 days, as well as vault-archive, click here.

Long term recovery points can be moved to vault-archive. To move all 'recommended recovery points' to vault-archive tier, click here.

CRASH CONSISTENT	APPLICATION CONSISTENT	FILE-SYSTEM CONSISTENT
4	0	0

Time ↑↓	Consistency	Recovery type
12/8/2022, 8:09:47 AM	Crash Consistent	Snapshot and Vault-Standard
12/7/2022, 8:07:47 AM	Crash Consistent	Snapshot and Vault-Standard

Figure 2:Working version of script



Chrome is being controlled by automated test software.

test | Backup

Virtual machine

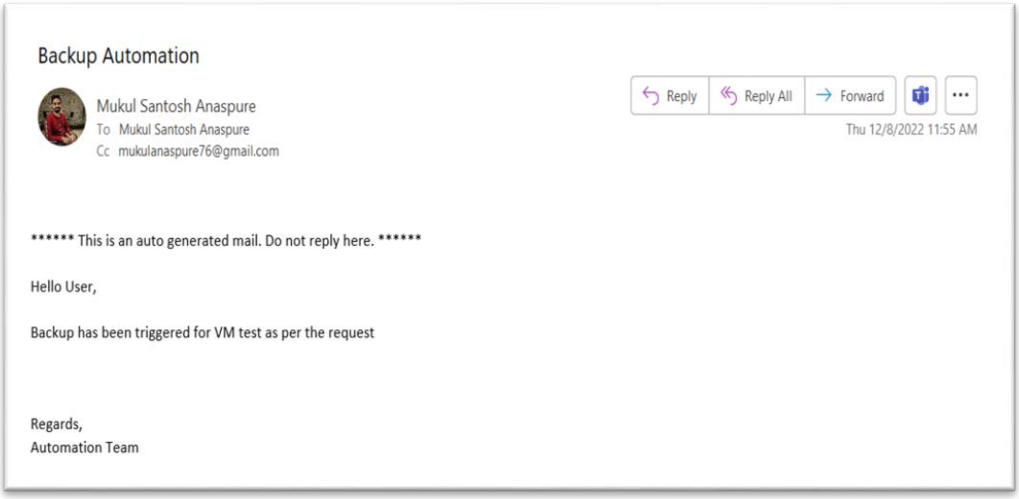
Notifications

Triggering backup for test

Backup triggered successfully. Please monitor progress in backup jobs page.

2 minutes ago

Figure 3: Automated backup triggering



Backup Automation

Mukul Santosh Anaspure
To Mukul Santosh Anaspure
Cc mukulanaspure76@gmail.com

Thu 12/8/2022 11:55 AM

****** This is an auto generated mail. Do not reply here. ******

Hello User,

Backup has been triggered for VM test as per the request

Regards,
Automation Team

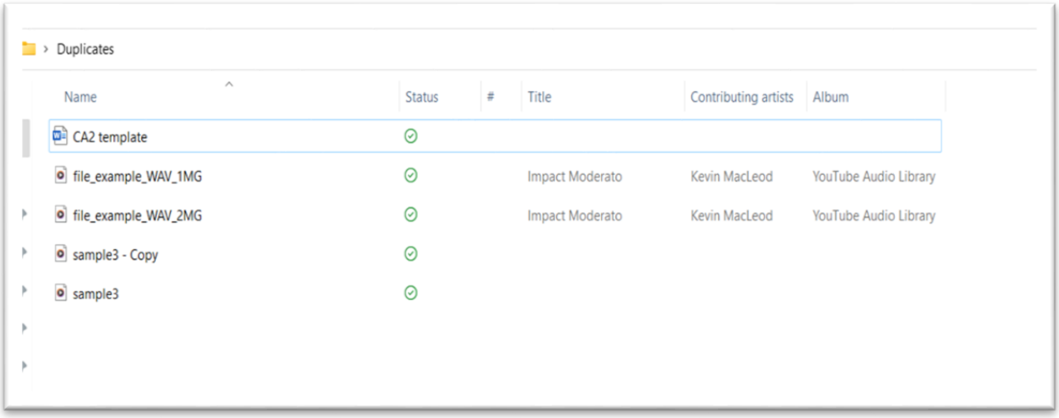
Figure 4: Sending an automated confirmation mail

After the backup is triggered automatically, then an automated mail is sent to the user as shown in above figure.

6.2 Experiment 2

The second experiment was to find duplicate audio files on the server. It is necessary to back up the data and not miss any backup requests but at the same time is also necessary to the keep the data copies unique. In cloud environments and organizations that use cloud platforms works in a team that sits worldwide. It's very obvious that multiple users are logged in to a server to carry out their daily tasks. There is a possibility of users storing same copy of files increasing burden on storage units and eventually rising the organization's budget in storage units. Also, there are many third-party tools available in market but scripting is the best possible option where we can automate this tasks frequently. In this experiment, the script is prepared in python and tkinter package is used. This package is used to create graphical user interfaces in python. The servers drive can have complex folder-files combinations like a main folder can have files and subfolders and the subfolders can have multiple files or folders. With the help of script, it is easy to calculate hash values of each files within fraction of time. So even if same copy of file is present in a main folder and subfolder, it can be found out and notify to the user with an email. For this experiment, .wav audio files are used as they are uncompressed audio files. In addition, we have created a simple folder-file structure where there is one main folder containing multiple audio files and different types of files. When the script is executed, it calculates the hash value of each file and then returns the name of duplicate files. In this experiment we have main folder called 'Duplicates' and in this folder, we have multiple files.When the script is executed, it will calculate the hash values of each file and return the name of files which are duplicate.

In figure 5, we can see the folder structure as mentioned in above paragraph.

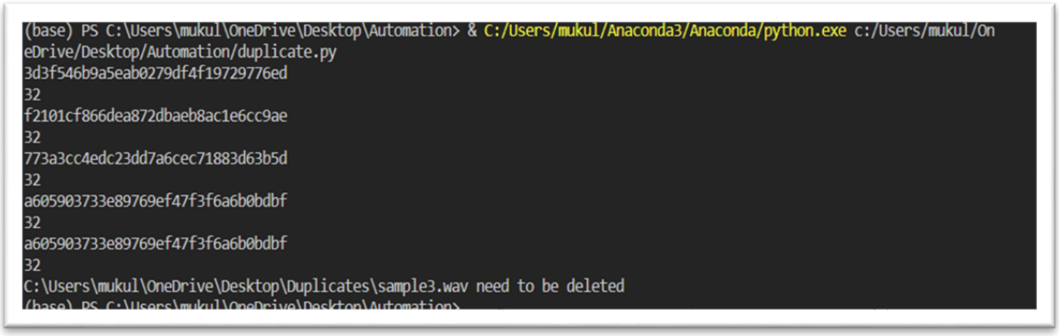


Duplicates

Name	Status	#	Title	Contributing artists	Album
CA2 template	✓				
file_example_WAV_1MG	✓		Impact Moderato	Kevin MacLeod	YouTube Audio Library
file_example_WAV_2MG	✓		Impact Moderato	Kevin MacLeod	YouTube Audio Library
sample3 - Copy	✓				
sample3	✓				

Figure 5 : Folder-file structure

In figure 6, we can see the hash values calculated for each file in the folder ‘Duplicate’. There are 5 files: 1 word file and 4 audio files in which sample3.wav is a duplicate one. We can see the script has detected the duplicate file and suggested to delete it.

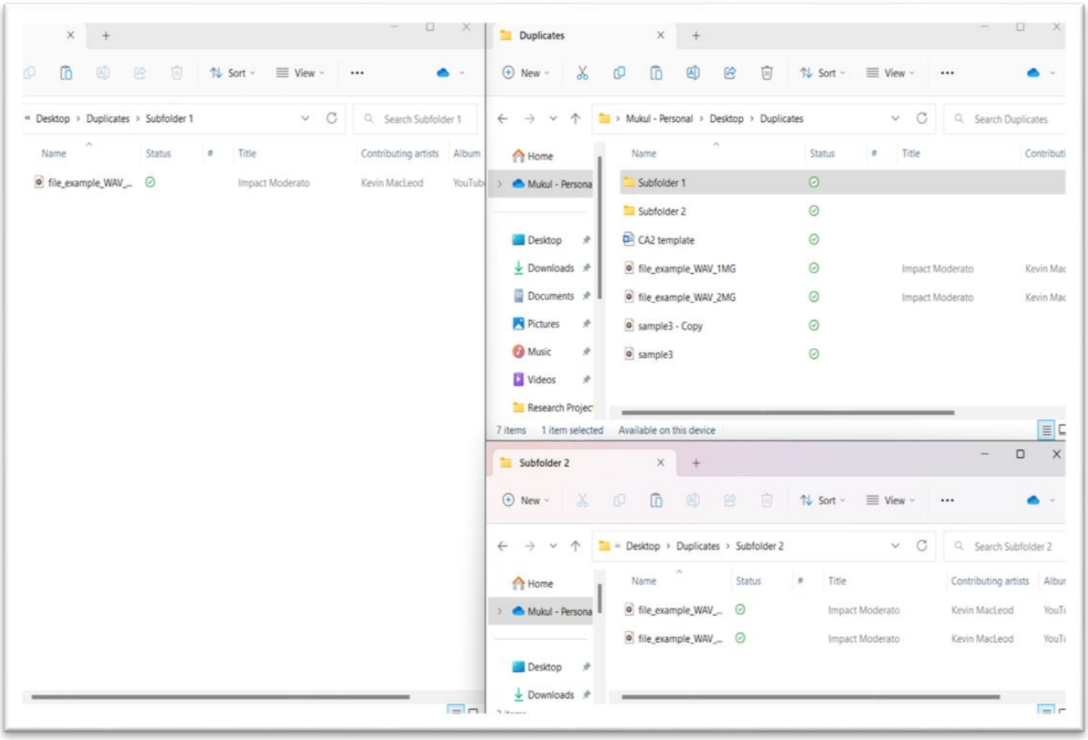


```
(base) PS C:\Users\mukul\OneDrive\Desktop\Automation> & C:/Users/mukul/Anaconda3/Anaconda/python.exe c:/Users/mukul/On
eDrive/Desktop/Automation/duplicate.py
3d3f546b9a5eab0279df4f19729776ed
32
f2101cf866dea872dbaeb8ac1e6cc9ae
32
773a3cc4edc23dd7a6cec71883d63b5d
32
a605903733e89769ef47f3f6a6b0bdbf
32
a605903733e89769ef47f3f6a6b0bdbf
32
C:\Users\mukul\OneDrive\Desktop\Duplicates\sample3.wav need to be deleted
(base) PS C:\Users\mukul\OneDrive\Desktop\Automation>
```

Figure 6: Script calculating unique hash values

6.3 Experiment 3

In this experiment, we have changed the folder-file structure and made it bit complex because in many cloud environment scenarios servers are being used by many users and every user can have different folder-file structure. In below experiment, we have a main folder ‘Duplicates’ and in this we have 2 subfolders (Subfolders 1 and Subfolders 2) and files. The subfolders are also having files in it. Please refer to figure 7. When the script is executed, we can see in below figure 8 that script has calculated 5 hash values. They are compared with all the hash values of files present in main folder ‘Duplicates’ and suggested to delete the duplicate ones. The script also suggests the destination folder in which the duplicate file is present making it easier to delete or track the file.



Desktop › Duplicates › Subfolder 1

Name	Status	#	Title	Contributing artists	Album
file_example_WAV_...	✓		Impact Moderato	Kevin MacLeod	YouTub

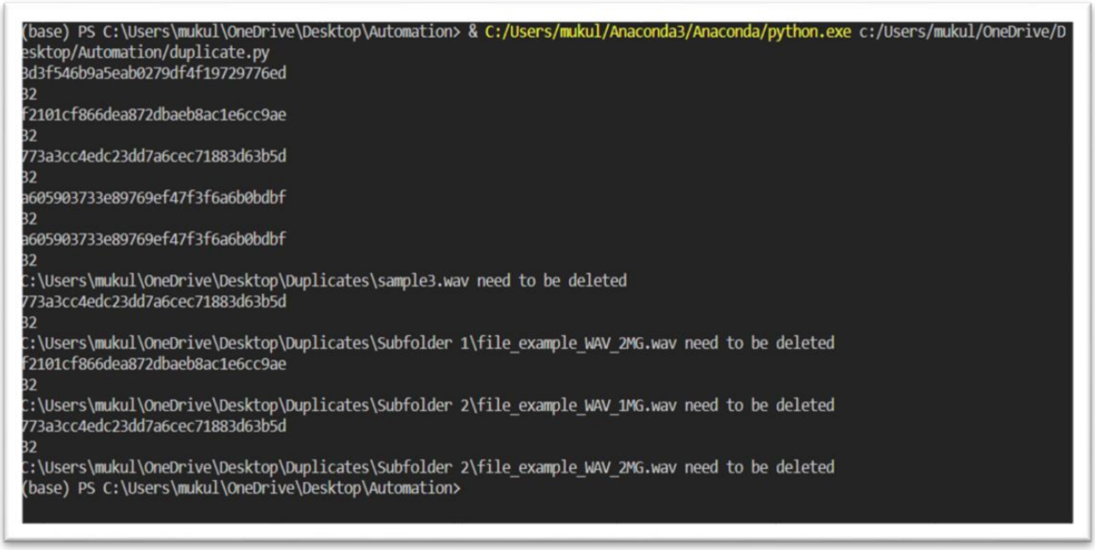
Mukul - Personal › Desktop › Duplicates

Name	Status	#	Title	Contributi
Subfolder 1	✓			
Subfolder 2	✓			
CA2 template	✓			
file_example_WAV_1MG	✓		Impact Moderato	Kevin Mac
file_example_WAV_2MG	✓		Impact Moderato	Kevin Mac
sample3 - Copy	✓			
sample3	✓			

Desktop › Duplicates › Subfolder 2

Name	Status	#	Title	Contributing artists	Albur
file_example_WAV_...	✓		Impact Moderato	Kevin MacLeod	YouTi
file_example_WAV_...	✓		Impact Moderato	Kevin MacLeod	YouTi

Figure 7: Complex level folder-file structure



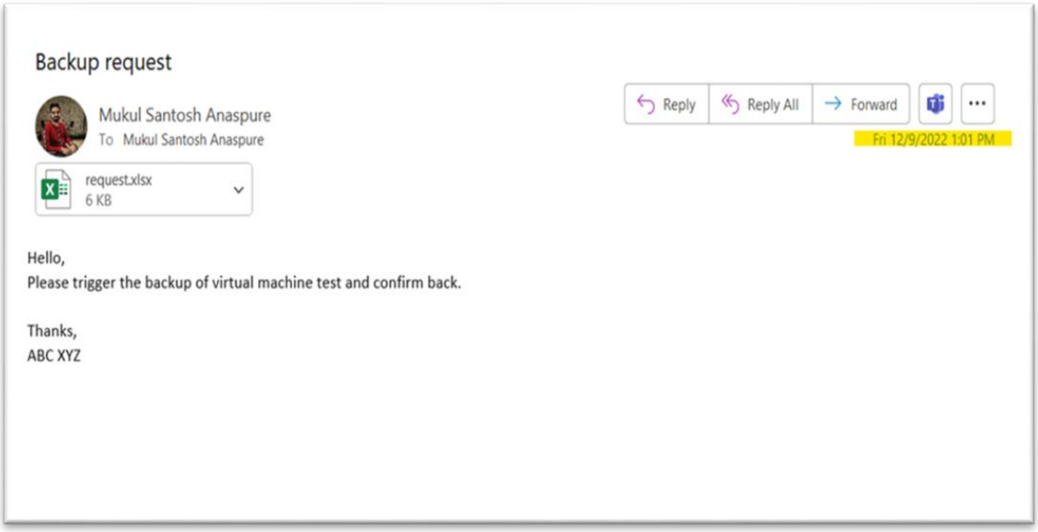
```
(base) PS C:\Users\mukul\OneDrive\Desktop\Automation> & C:/Users/mukul/Anaconda3/Anaconda/python.exe c:/Users/mukul/OneDrive/D
esktop/Automation/duplicate.py
3d3f546b9a5eab0279df4f19729776ed
32
f2101cf866dea872dbaeb8ac1e6cc9ae
32
773a3cc4edc23dd7a6cec71883d63b5d
32
a605903733e89769ef47f3f6a6b0bdbf
32
a605903733e89769ef47f3f6a6b0bdbf
32
C:\Users\mukul\OneDrive\Desktop\Duplicates\sample3.wav need to be deleted
773a3cc4edc23dd7a6cec71883d63b5d
32
C:\Users\mukul\OneDrive\Desktop\Duplicates\Subfolder 1\file_example_WAV_2MG.wav need to be deleted
f2101cf866dea872dbaeb8ac1e6cc9ae
32
C:\Users\mukul\OneDrive\Desktop\Duplicates\Subfolder 2\file_example_WAV_1MG.wav need to be deleted
773a3cc4edc23dd7a6cec71883d63b5d
32
C:\Users\mukul\OneDrive\Desktop\Duplicates\Subfolder 2\file_example_WAV_2MG.wav need to be deleted
(base) PS C:\Users\mukul\OneDrive\Desktop\Automation>
```

Figure 8: Script calculating unique hash values

6.4 Discussion

1. Experiment 1 made it very clear that VM backup can be triggered automatically from the portal with the help of python selenium script. Also, the automated mail is sent once the backup is successfully triggered. In addition, the condition which we have mentioned in Azure Logic Apps is working fine. The condition is whenever mail of

backup requests is received in outlook application then the designed workflow would be performed. In below figure 9,10 we can see that whenever a mail is received in outlook then the Azure Logic App is automatically triggered which makes the script to run.



Backup request

Mukul Santosh Anaspure
To Mukul Santosh Anaspure

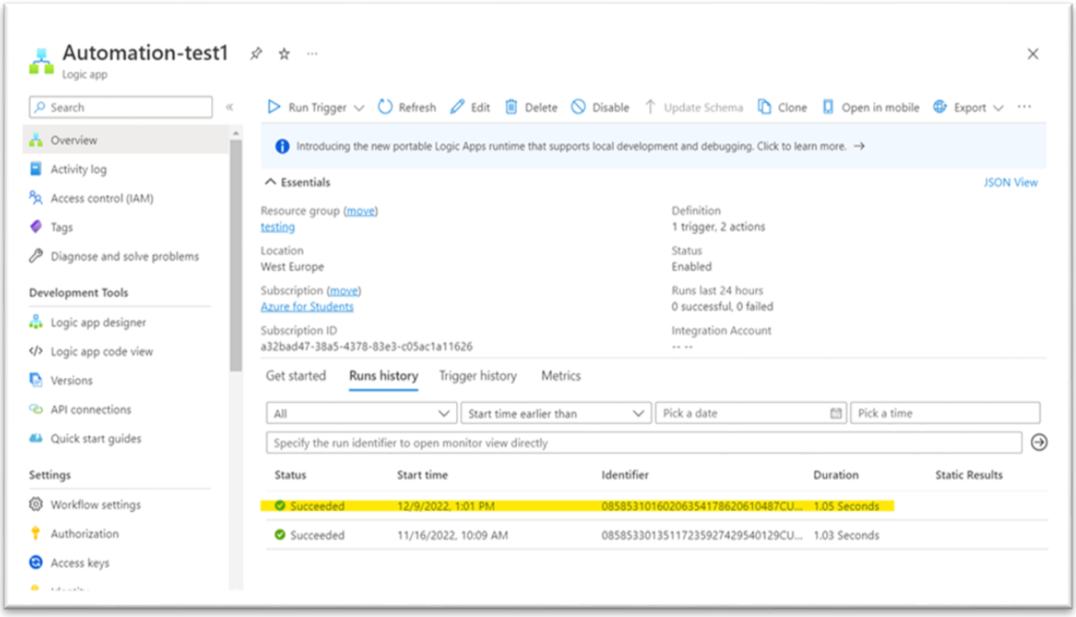
Fri 12/9/2022 1:01 PM

request.xlsx
6 KB

Hello,
Please trigger the backup of virtual machine test and confirm back.

Thanks,
ABC XYZ

Figure 9: Initial mail from user acting as a trigger



Automation-test1
Logic app

Introducing the new portable Logic Apps runtime that supports local development and debugging. Click to learn more. →

Essentials

Resource group (move)
testing

Location
West Europe

Subscription (move)
Azure for Students

Subscription ID
a32bad47-38a5-4378-83e3-c05ac1a11626

Definition
1 trigger, 2 actions

Status
Enabled

Runs last 24 hours
0 successful, 0 failed

Integration Account
-- --

Status	Start time	Identifier	Duration	Static Results
Succeeded	12/9/2022, 1:01 PM	08585310160206354178620610487CU...	1.05 Seconds	
Succeeded	11/16/2022, 10:09 AM	08585330135117235927429540129CU...	1.03 Seconds	

Figure 10 – Logic Apps workflow

2. In experiment 2 and 3 we have performed data deduplication on audio files with the help of python script. We have scheduled that script with the help of Windows Task Scheduler at a particular time. In the first trial we have kept the folder-file structure in a very simple manner where there is only one mail folder which consists of multiple audio files, word files etc. So, when the script is run it calculates the hash values for all the files present in folder. But the scenario is no longer same when it comes to servers as multiple users are using it from different locations across the globe. The folder-file structure can be very complex sometimes and becomes difficult to track the duplicate

files and delete them. Scripting is a best possible option to automate this task. We have used hashing concept in this approach. We have taken content of each file and pass that content through a hash function as shown in figure 11. This hash function will generate a unique string of hash values. Each file will have a unique string generated by the hash function. The length of unique string is always a fixed length that is 32 characters because we have used MD5 hash function. The length of string will be always 32 characters long irrespective of the file size. The unique hash values are compared with each other and if there is any duplicate value then its corresponding file is duplicate file and it can be deleted.

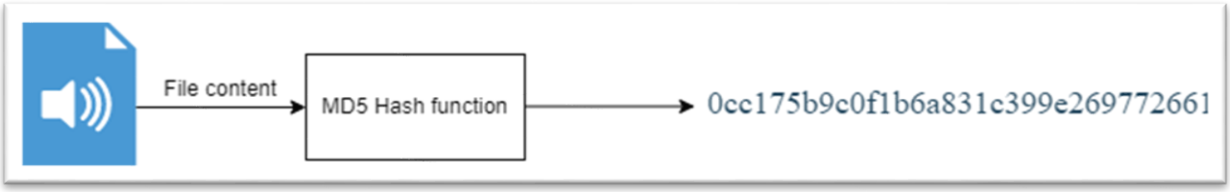


Figure 11 – Hash function

6.4.1 Improvements:

There is always space for improvement. Average practices are fast losing relevance in today's competitive environment. In this research, improvement is a need in case of multiple server's backup request. How can the script be modified if there is a backup request for multiple servers. Also, if there are any user suggested inputs like retention period. Finally trying to execute the technique one after another which means the backup will be automatically triggered for requested servers and when backup is completed then the deduplication script can fetch the duplicate files from the same servers and then send an email confirming about the backup trigger and number of duplicate files present. Finally trying to execute the technique one after another which means the backup will be automatically triggered for requested servers and when backup is completed then the deduplication script can fetch the duplicate files from the same servers and then send an email confirming about the backup trigger and number of duplicate files present.

Finally trying to execute the technique one after another which means the backup will be automatically triggered for requested servers and when backup is completed, then the deduplication script can fetch the duplicate files from the same servers and then send an email confirming about the backup trigger and number of duplicate files present.

7 Conclusion and Future Work

This research has proposed how to automatically handle the backup triggers in a cloud environment and find the duplicate audio files on a server using scripting and hashing function. In addition, organizations can save valuable time and invest time in some productive tasks. The deduplication technique would always be efficient for finding the duplicate audio files on a regular basis. In a cloud environment, most of the cost is incurred in storage units hence it is always vital to keep the data in a unique state. The proposed method can help organizations to save their annual budgets from a storage cost perspective.

The future work will be to reduce the possibility of backup failure in the cloud. For example, the backup is often failed in the cloud and most of the time, the reason behind the failure is an unstable state of VSS writers. The VSS services should always be in a stable state when the backup is being triggered. A future script in which it will first restart the services on the server level and then trigger the backup from thee portal. Restarting the service does not affect the server performance. Automating this entire process would save precious time for engineers and the cost of the organization and help automate similar tasks in a cloud environment.

References

(Azza Higazy, Tarek El Tobely, Ahmed H. Yousef and Amany Sarhan, "Web-based Arabic/English duplicate record detection with nested blocking technique", Proceedings - 2013 8th International Conference on Computer Engineering and Systems ICCES 2013, pp. 313-318, 2013.)
A.El-Shimi, R. Kalach, A. Kumar, A. Oltean, J. Li and S. Sengupta, "Primary data deduplication - large scale study and system design", ACM Conference on Annual Tech USENIX, pp. 1-12, 2012.

Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis and Vassilios S. Verykios, "Duplicate record detection: A survey", IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1-16, jan 2007.

D. Geer, "Reducing the storage burden via data deduplication", IEEE Computer Journal, vol. 41, pp. 15-17, Dec. 2008.

E. Palkopoulou, D. A. Schupke and T. Bauschert, "Recovery Time Analysis for the Shared Backup Router Resources (SBRR) Architecture," 2011 IEEE International Conference on Communications (ICC), 2011, pp. 1-6, doi: 10.1109/icc.2011.5963411.

Hamadah, Siham. (2019). ICIC Express Letters ICIC International©2019 ISSN. ICIC Express Letters. 13. 593-599. 10.24507/icicel.13.07.593.

https://aws.amazon.com/step-functions/

https://azure.microsoft.com/en-us/products/logic-apps/

https://cloud.google.com/composer

https://www.gartner.com/reviews/market/enterprise-integration-platform-as-a-service/vendor/microsoft/product/azure-logic-apps/likes-dislikes

https://www.vectra.ai/learning/power-automate

Jack E. Olson, Data Quality: The Accuracy Dimension, Elsevier Inc., 2003.
K. Yang and X. Jia, An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing, Parallel and Distributed Systems, IEEE Transactions on, vol. PP, pp. 1-1, 2012.

L. L. You, K. T. Pollack, and D. D. E. Long, "Deep Store: An Archival Storage System Architecture, " presented at the Proceedings of the 21st International Conference on Data Engineering, 2005.

L. Sun, J. An, Y. Yang and M. Zeng, "Recovery strategies for service composition in dynamic network," 2011 International Conference on Cloud and Service Computing, 2011, pp. 60-64, doi: 10.1109/CSC.2011.6138553.

Mell, P. and Grance, T., 2011. The NIST definition of cloud computing.

S. Suguna and A. Suhasini, "Overview of data backup and disaster recovery in cloud," International Conference on Information Communication and Embedded Systems (ICICES2014), 2014, pp. 1-7, doi: 10.1109/ICICES.2014.7033804.

Philip Menard, Robert Gatlin & Merrill Warkentin (2014) Threat Protection and Convenience: Antecedents of Cloud-Based Data Backup, Journal of Computer Information Systems, 55:1, 83-91, DOI: 10.1080/08874417.2014.11645743

Qi Zhang, Jihua Kang, Jin Qian and Xuanjing Huang, Continuous Word Embeddings for Detecting Local Text Reuses at the Semantic Level, pp. 797-806, 2014.)
SNIA, Advanced Deduplication Concepts, 2011.

T. Yujuan, J. Hong, F. Dan, T. Lei, Y. Zhichao, and Z. Guohui, "SAM: A Semantic-Aware Multi-tiered Source De-duplication Framework for Cloud Backup, " in Parallel Processing (ICPP), 2010 39th International Conference on, 2010, pp. 614-623.

V. Javaraiah, "Backup for cloud and disaster recovery for consumers and SMBs, " in Advanced Networks and Telecommunication Systems (ANTS), 2011 IEEE 5th International Conference on, 2011, pp. 1-3.

W. Cong, W. Qian, R. Kui, C. Ning, and L. Wenjing, "Toward Secure and Dependable Storage Services in Cloud Computing, " Services Computing, IEEE Transactions on, vol. 5, pp. 220-232, 2012.

W. Ng, Y. Wen and H. Zhu, "Private data deduplication protocols in cloud storage", ACM Symposium on Applied Computing, pp. 441-446, Mar. 2012.

William W. Cohen and Jacob Richman, "Learning to match and cluster large high-dimensional data sets for data integration", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 475-480, 2002.

Y. Ueno, N. Miyaho, S. Suzuki and K. Ichihara, "Performance Evaluation of a Disaster Recovery System and Practical Network System Applications," 2010 Fifth International Conference on Systems and Networks Communications, 2010, pp. 195-200, doi: 10.1109/ICSNC.2010.37.