# Data Acquisition on a Virtual Machine: Three Scenarios

Saritha Narahari
California State University, Sacramento
1550 IronPoint Road apt711
Folsom, ca-95630

## ABSTRACT
Virtual machine computing is becoming more and more prevalent. Companies are providing virtual desktops for their employees and using virtual machines to run server software. Additionally, the use of Infrastructure as a Service has placed virtual machines within the reach of even more people. Virtual machines can pose a challenge because of their transient nature. Also, the nature of how virtual machines store data could prove problematic.

## Keywords
Data Acquisition, Virtual Machine

## 1. INTRODUCTION
The Virtual Machine is the concept where a software application is behaved as a host machine. Virtual machine application runs on the host (Actual) machine Operating System as the Guest which runs inside the Physical computer. A Hypervisor is used to create the virtual machine and manage by allocating the machine requirements on a host. There are two types of Hypervisors. In Type-1 Hypervisor virtualization has direct access to the resources. In this type the performance is like the actual performance of the system. Whereas in the Type-02 Hypervisor, the Virtualization incurs overhead as the host's Operating System takes the requests from VM and allocates the resources.

"Performing Forensic investigation on the Virtual Machine is the challenge because the nature of the systems that's is virtualized and isolated from the host must be analyzed prior before performing investigation."

Performing Forensic Investigation on a host machine involutes four steps to recognize, acquire and analyze a virtual machine

- Forensic Image Creation
- Sensitive Information Identification and recovery
- Virtual machine analysis
- Documentation

Forensic Image Creation phase involves creation of the image which has the record of all the activities performed inside the virtual machine." While creating the image we need to ensure that the data is complete, and the data is not modified. After acquiring the Image investigation is done on it as investigation on the original disk data is not recommended which may result in data modifications, therefore it is important to create the disk image of the original data."

"Sensitive information Identification and recovery phase, Operating Systems create the keep logs of the activities I.e., debugging, management, record purposes".The Investigator should understand and be aware of the host operating system before performing analysis which helps in identifying the sensitive information and recover the traces of the VM and the illegal activities. The File associations in the registry reveal the information about the applications installed and used in the host machine. Even if the hypervisor is uninstalled, the .vbox, .vmdk, .vmx etc files in the host machine confirm the usage of virtual machine.

Investigators face some challenges in the recovery of the data deleted and files corrupted. Sometime deleted files can be recovered from the temporary locations." Deleted snapshots, VM configuration files etc can be recovered by using some application such as UNDELETE, Handy Recovery etc., which can be analyzed once they are recovered investigation can be continued". They are some limitations in the file recovery like file encryption, Physical Destruction, Degaussing, Gutmann Method which cause file corruption.

Virtual Machine Analysis Phase, Virtual Machine analysis consume more time compared to the normal machine analysis. In order to analyze the VM, we need to get access to it." The virtual machine is analyzed by mounting it as the hard drive in another machine or by using it with a hypervisor to get access into the virtual environment". Once the disk image is extracted from the original disk, it is analyzed with the tools when the VM access is granted. Most of the physical machine forensic tools support the virtual machine with the Virtual machine operating system compatible software.

"In Documentation phase, every record of the investigation is documented, and all the activities related to analysis, evidence transfer, validation, storage must be documented so that everything will be available for the further investigation. It is important to have report forms in each phase for the documentation."

## 2. PROJECT GOAL
The goal of the project was to analyze the use of a traditional forensics data acquisition to acquire data in a virtual machine environment. Here the main sought is to compare the integrity of data gathered to see if it would reflect the true actions of a scenario enacted in each environment. The choosen three data acquisition scenarios to enact are

1. Perform data acquisition from within a local Virtual Machine environment. This scenario is meant to simulate an instance where the company has complete control of the virtual environment and wished to acquire data from the virtual machine given to the suspect.
2. Perform data acquisition from a disk image of a local virtual machine. This scenario is meant to simulate an instance where the company controls the storage device upon which the virtual image disk is stored but may not have complete access to get inside of the VM while it is running.
3. Perform data acquisition from within a cloud Virtual Machine. This scenario is meant to simulate an instance where the company or individual is using

Infrastructure as Service from a cloud provider. Access to the actual disk image or storage device is not possible, however access to the virtual environment itself is.

## 3. IMPLEMENTATION

To illustrate the differences between the results the three different scenarios we performed same set of activities in the cloud and virtual environments and Autopsy is the forensic tool used for investigating the differences.

Here these environments are seeded with data that could later attempt to acquire. 1. Did a web search on Spiderman and downloaded few similar images and deleted one image. 2. created a document with the text written 'Trump Trump Trump' so that a keyword search can be performed using the keyword 'Trump'. Using Autopsy, the aim is to find out the traces of these files and other interested data that is present.

In the First Scenario, the Oracle VM Virtual Box, which is freely available on internet is selected and installed for investigation. We also download a Windows virtual machine from Microsoft Developer to act as the OS within our Virtual box environment. Within the Virtual box virtual environment, Virtual disk image must be seeded with data so that the record of activities is stored. In this environment, the activities are listed and created the text document as the first step. In the second step, Autopsy forensic tool is installed, and investigation is performed on the Virtual machine.

For the Second Scenario, we extracted disk image of the virtual machine which has the record of the activities performed in the first scenario's first step. The disk image was extracted immediately after performing the "seeding" steps, before Autopsy was installed and run on the virtual machine. The extracted disk image is in the .vmdk file format. Autopsy doesn't recognize .vmdk files so a file conversion was done. For the file conversion qemu-img software is used, through which .vmdk disk image is converted to .raw file. Thereby Autopsy can perform analysis on the .raw file.

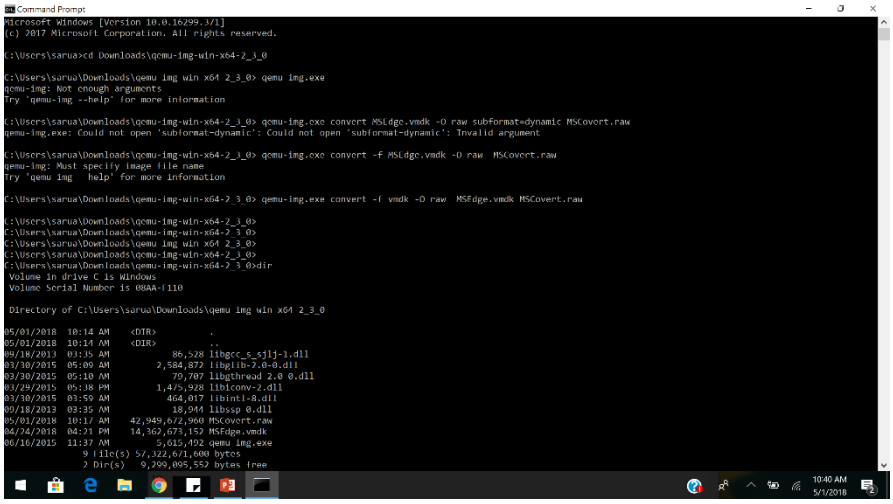


**Fig 1: vmdk file converted to Raw file using qemu-img software**

Raw file is added as the data source selected as disk image in autopsy.

In the Third Scenario, using the Microsoft Azure Cloud Services, a premium account is created and MS Windows Virtual Machine is setup using their standard process. The standard MS Azure virtual machine for a premium comes with 127GiB storage on SSD. The disk is not encrypted in this type of account. Additionally, MS Azure does not guarantee to persist local SSD data. If persistent data is desired, one must sign up for a different type of MS Azure File Storage account in addition to the virtual machine account. The environment was "seeded" following the steps outlined previously. Autopsy was then installed on the virtual machine. It took two days between seeding the data and running Autopsy to perform the acquisition.

## 4. RESULT

During the data acquisition, we examined each environment for the following data:

1. Keyword search for "Trump".
2. Searches for Spiderman related data/images and traces of our Spiderman browsing.
3. Records of email address.
4. Records of bookmarks
5. Record of the deleted files, including the Spiderman image that was deleted.

Comparison between the scenario 1 and scenario 2 results:

1. The number of Deleted files in both scenarios' is divergent. As analysis is performed on the same the virtual disk image there is a time lapse between the data acquisition in virtual environment.
2. There are some files with the modified date and time listed as 0000-00-00 00:00, as we used a freely available virtual box and the disk image from online there are the chances we need to expect from other usages
3. The Recent Accessed files are similar in the both cases.
4. There are some email addresses listed which are cached and the Bing.url as saved bookmark in the both scenarios.
5. While browsing for spider man images, there are few .html files stored in the cache memory, some of them with metadata and some with unknown metadata. With the spider man keyword search we were able to retrieve the same set of files in both cases.
6. With Trump keyword search we retrieved files which common in the both scenarios.

## 4.1 Scenario 1 Results

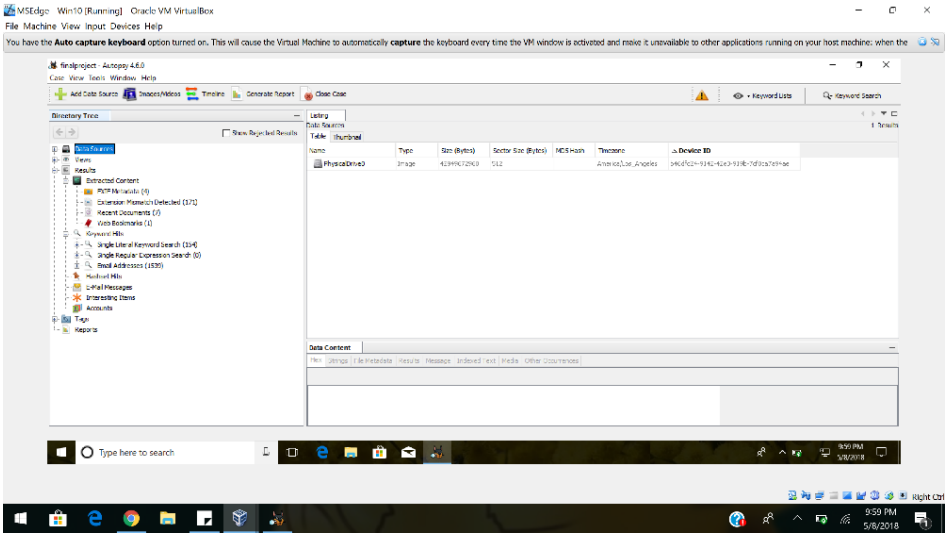

**Fig 2: Disk image added as the data source in the virtual environment:**

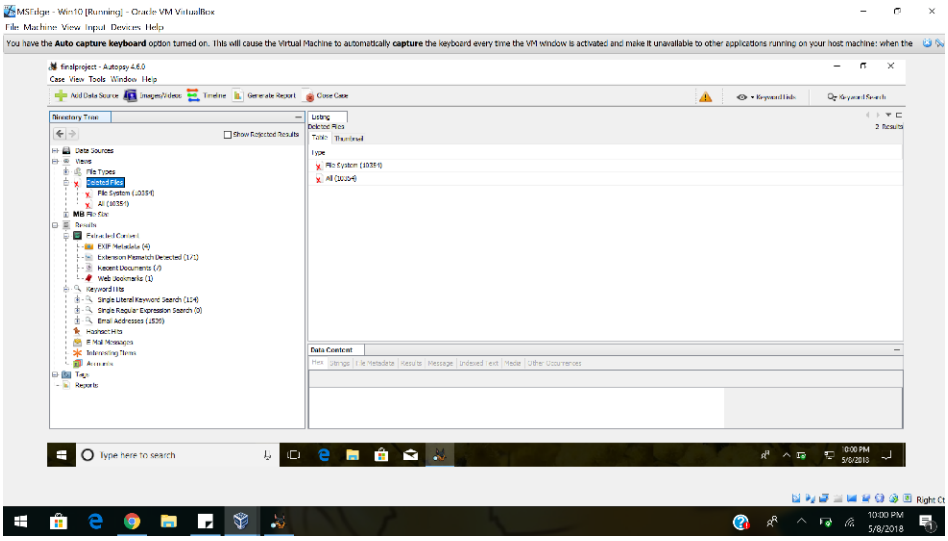

**Fig 3: Deleted Files retrieved from virtual machine analysis**

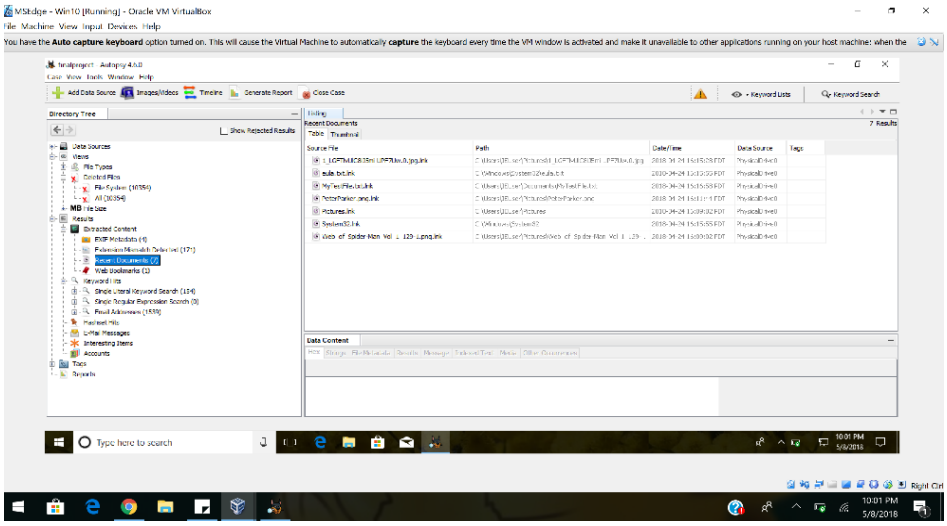

**Fig 4: Recent Documents accessed in the virtual environment**

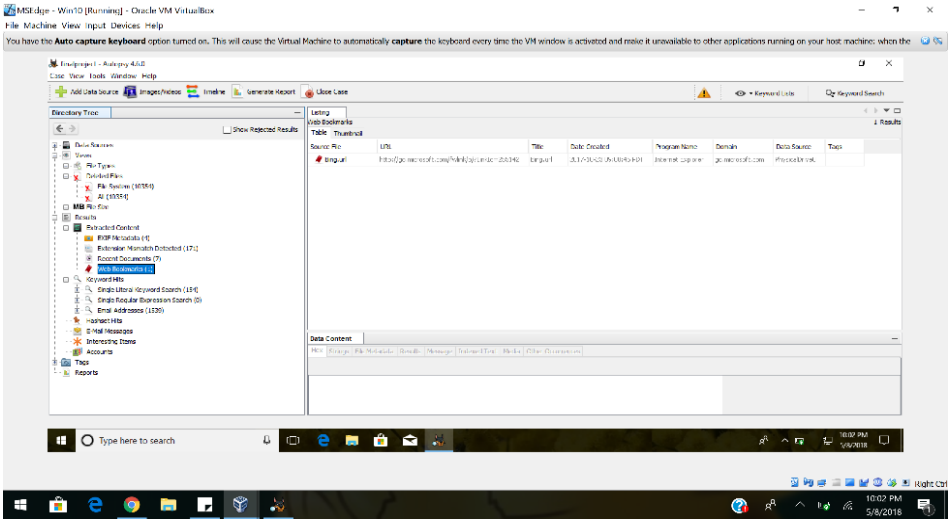

**Fig 5: Web Bookmarks saved in the virtual machine's browser**

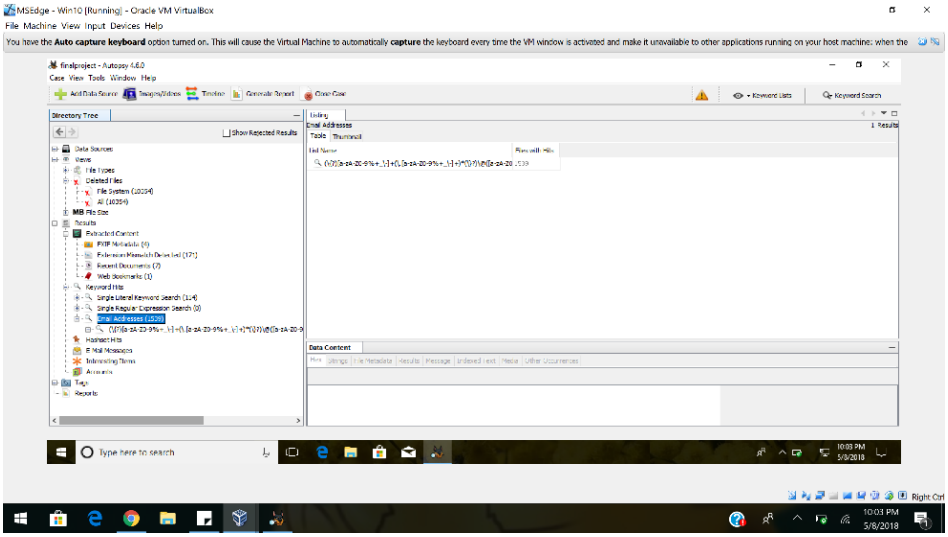

**Fig 6: Email addresses stored**

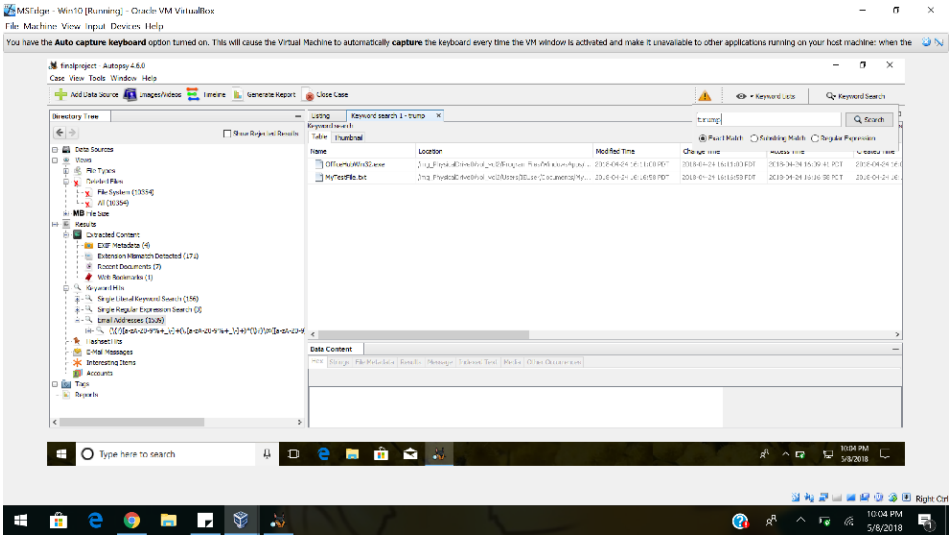

**Fig 7: Trump keyword search returning the files with trump in it**

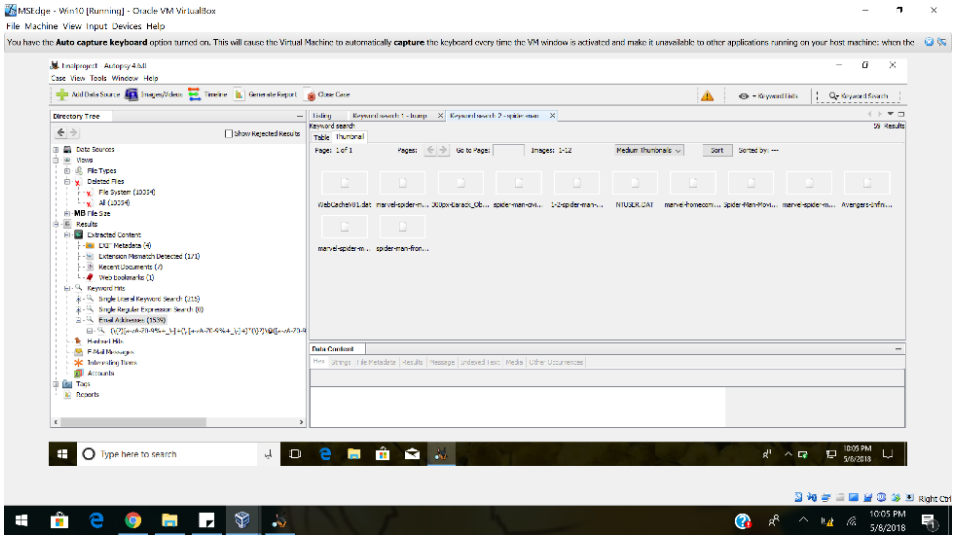

**Fig 8: Spiderman keyword search retries the files stored in the cache and the virtual storage**

## 4.2 Scenario 2 Results

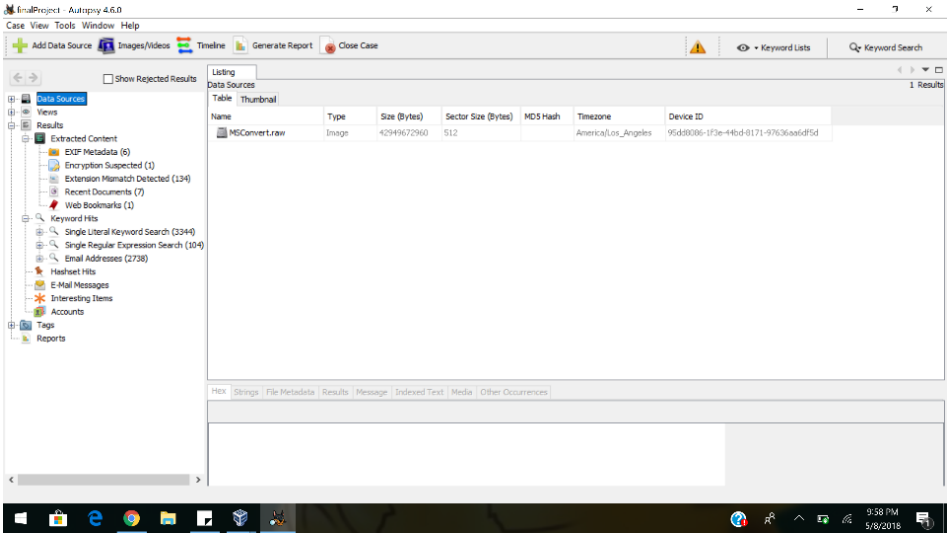

**Fig 9: MSConvert.raw (the converted disk image) as the data source**

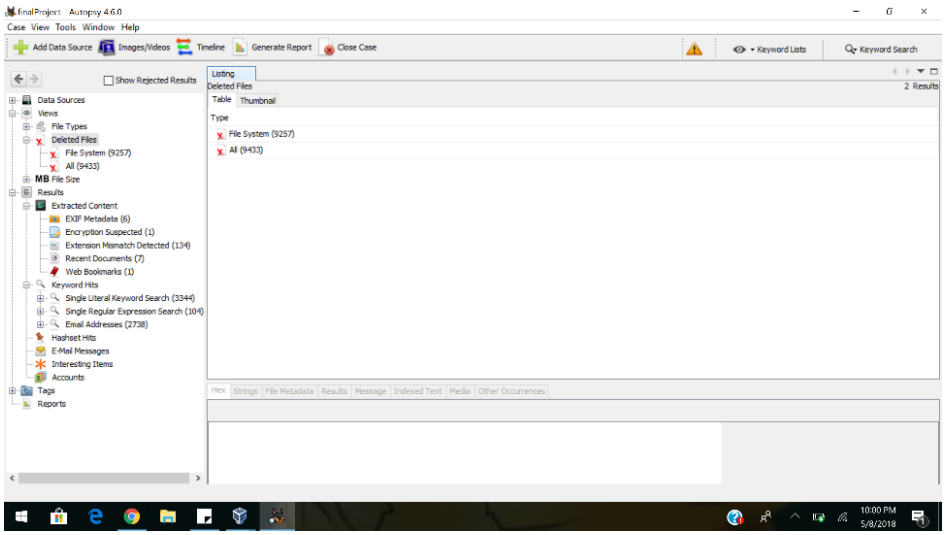

**Fig 10: Deleted number files analyzed from the disk image**

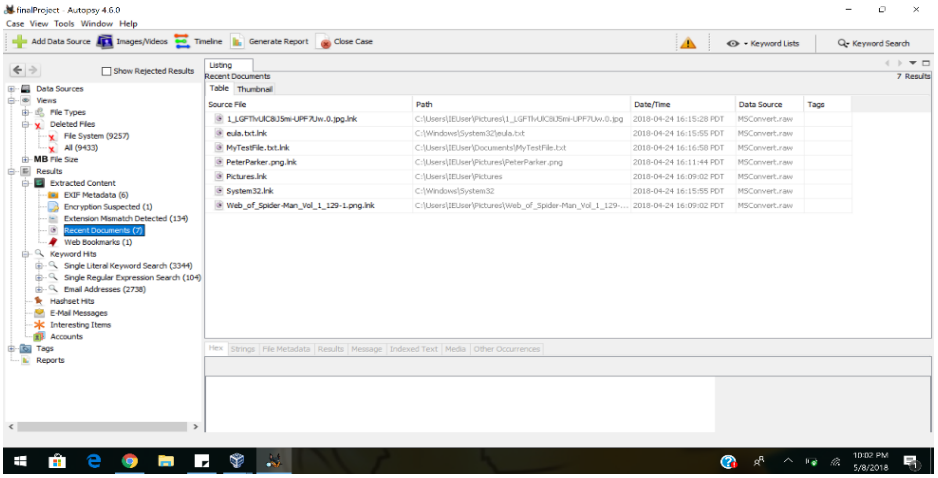


**Fig 11: Recent Documents illustrated from the disk image**

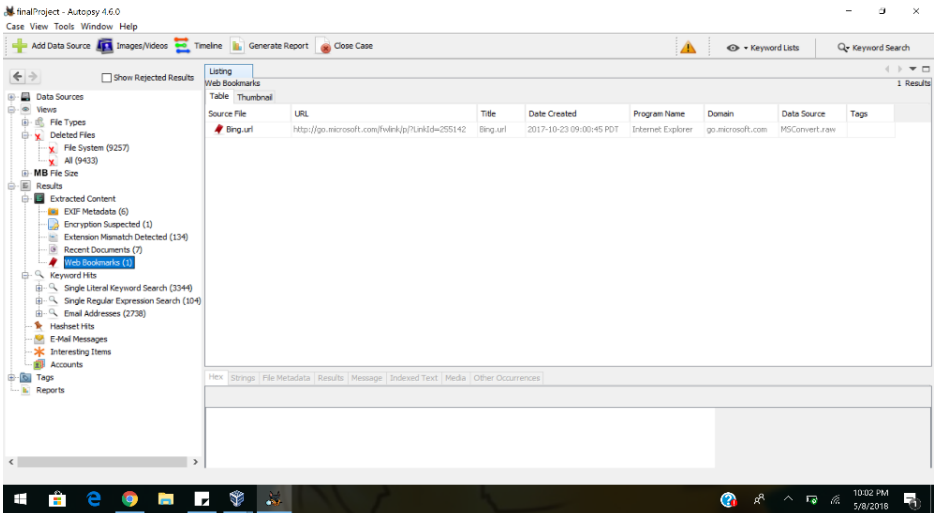


**Fig 12: Web Bookmarks saved on the virtual disk image**

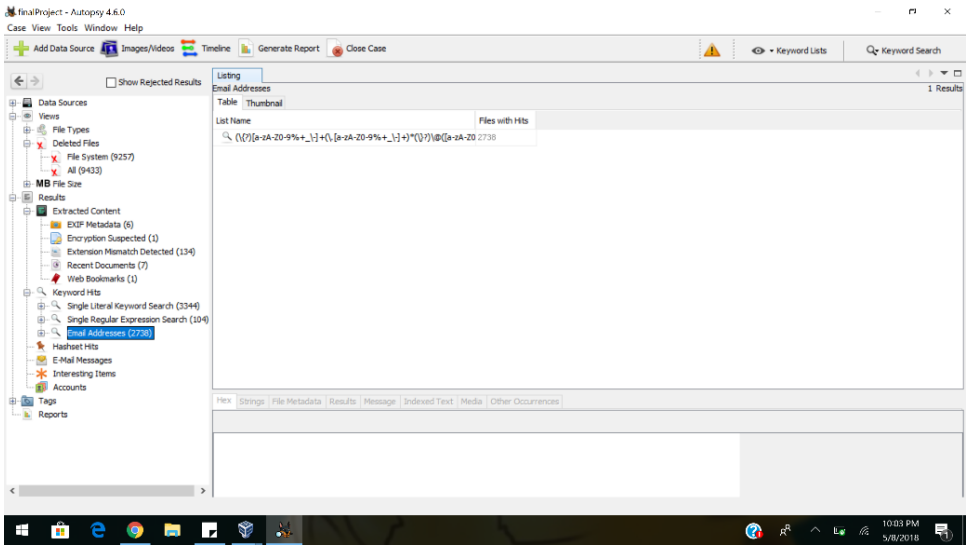


**Fig 13: Email Addresses accessed over virtual disk image**

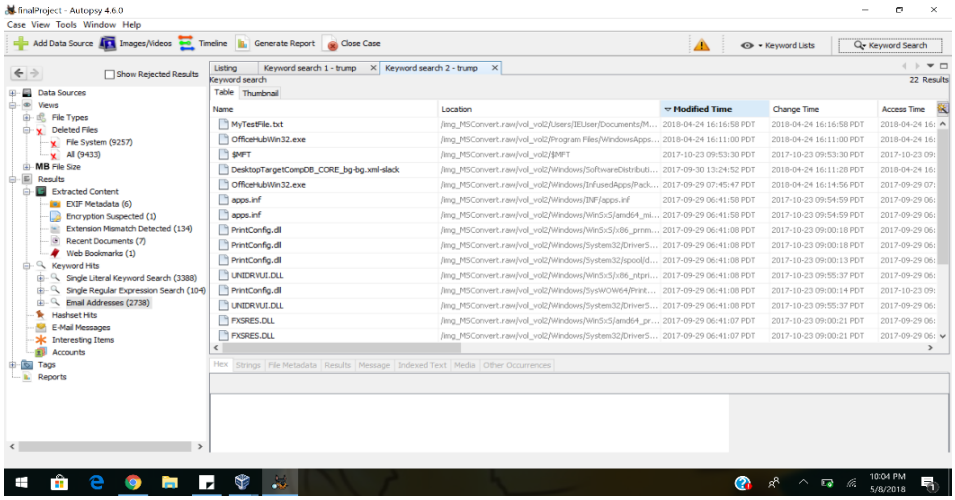

**Fig 14: Keyword search Trump resulting all the files with key word**

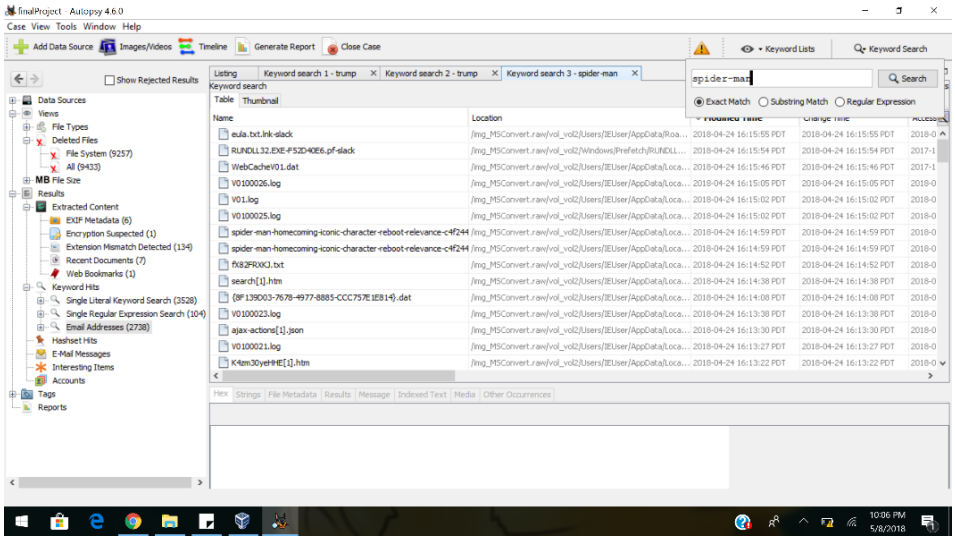

**Fig 15: Spiderman Keyword search which retries all the files with the Spiderman along with images**

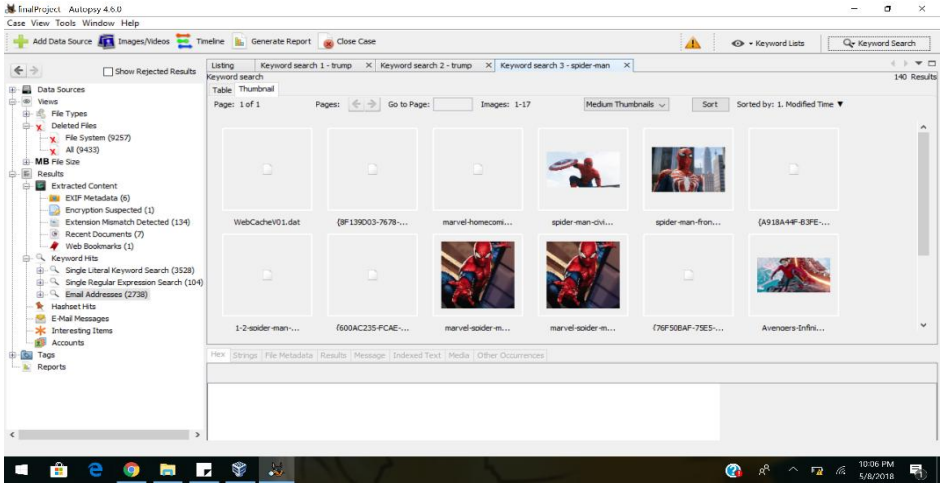

**Fig 16: Spiderman Images**

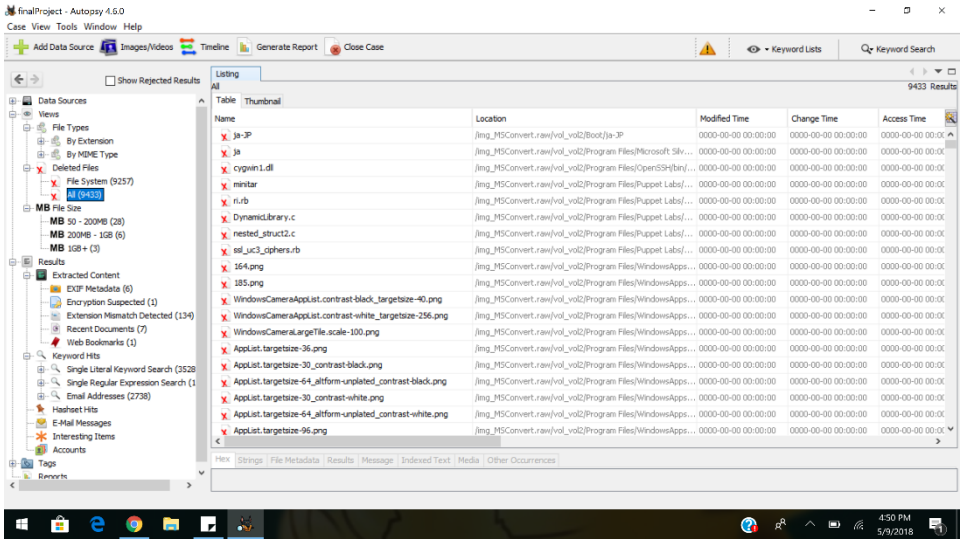

**Fig 17: List of the Deleted files: (with 0000-00-00 00:00 modified time)**

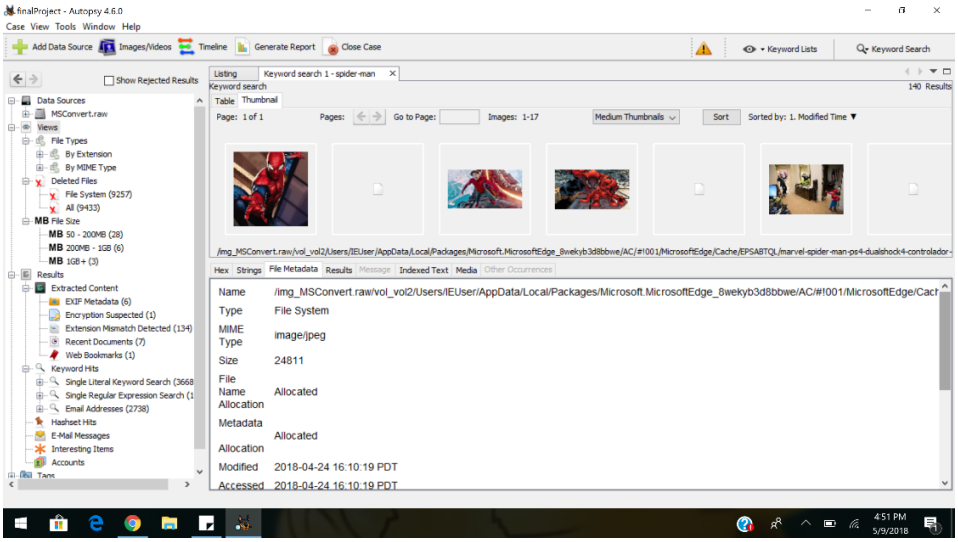

**Fig 18: Images with Metadata**

## 4.3 Scenario 3 Results

1. It is easy to find the file with the "Trump" keyword, using Autopsy's keyword search. The metadata did not include date/time information for the file.

2. There were several references found referring to .html files that were browsed during our Spiderman web browsing. These files were in the cache for the Microsoft web browser. We were able to locate the Spiderman images that we downloaded off the internet. These images did not have any date metadata associated with them. For dates/time metadata the value was 00-00-0000 00:00:00.

3. The emails acquired by Autopsy from this VM are interesting. There were 160+ email addresses listed. Since we did not conduct any email activity or store any email addresses during our "seeding" we were surprised to see so many emails. We had a few theories regarding how these emails came to reside in the VM, such that Autopsy would acquire them. 1. The email addresses were gathered from various files that were browsed during our web searching. 2. The email addresses reside(d) in files that came with the Virtual Machine image that we installed from MS Azure.

4. There were no bookmarks acquired other than the standard Bing bookmark included with MS browser software.

5. There were 0 deleted files recovered by Autopsy. In our "seeding" process we deleted one file. This file was not show, nor any other files as with the local VM scenarios. There are few theories about why there were zero deleted files and gathered various theories from our CSC 253 classmates during our presentation. 1. Because the file storage system is SSD, the gap between seeding the data and acquiring the data was too long (two days). A know fact regarding SSD acquisitions is that the data should be gathered as soon as possible because SSD wear leveling will not necessarily preserve deleted data. Additionally, because the SSD drive is a shared drive amongst several Azure VMs, wear leveling might be even more prevalent and allow for an even shorter period that data could be recovered. 2. Cloud services use multiple distributed hard disk drives. Given MS Azures own caveat about the lack of persistence for

standard VM data, there is no reason to assume that deleted files would be stored, or that were, in fact, accessing the same physical disk that the file had been deleted from.

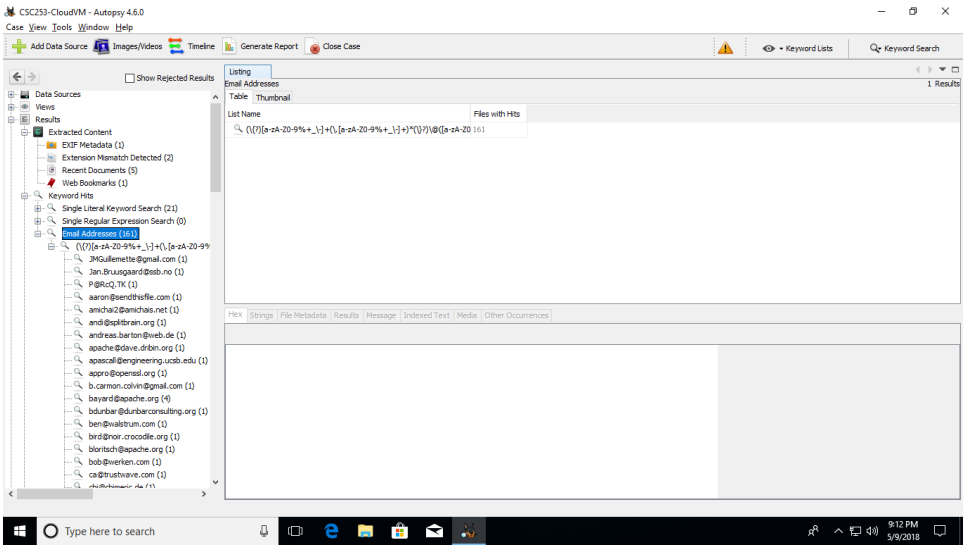


**Fig19: Email addresses - MS Azure**

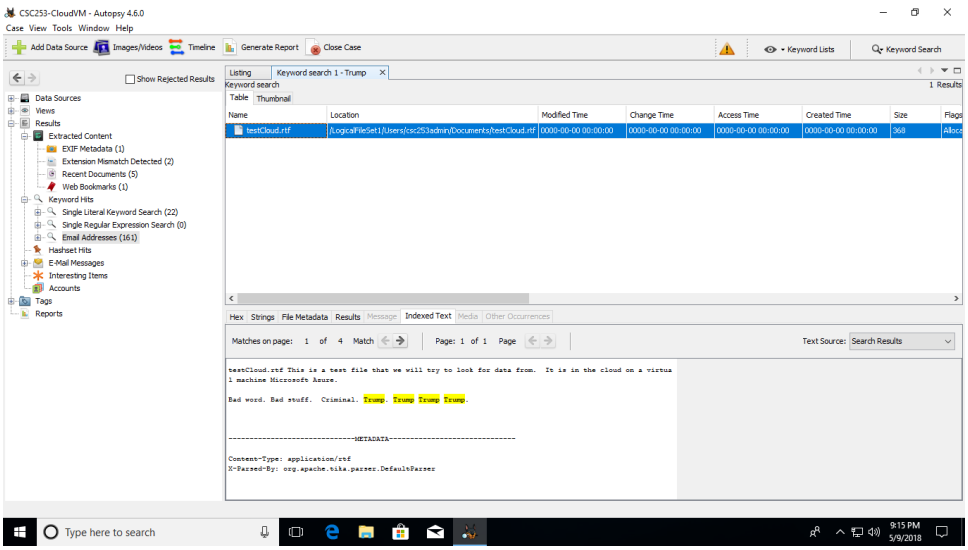


**Fig 20: Trump Keyword - MS Azure**

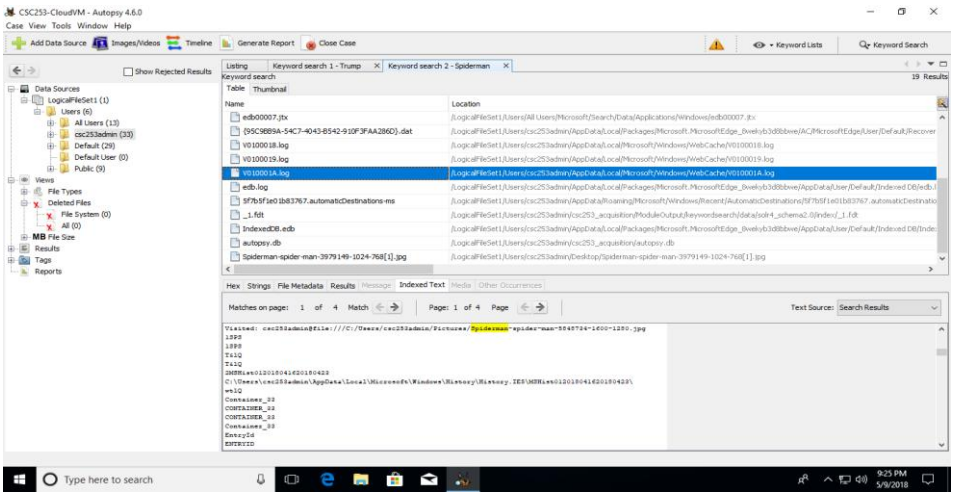


**Fig 21: Cache of Spiderman activity - MS Azure**

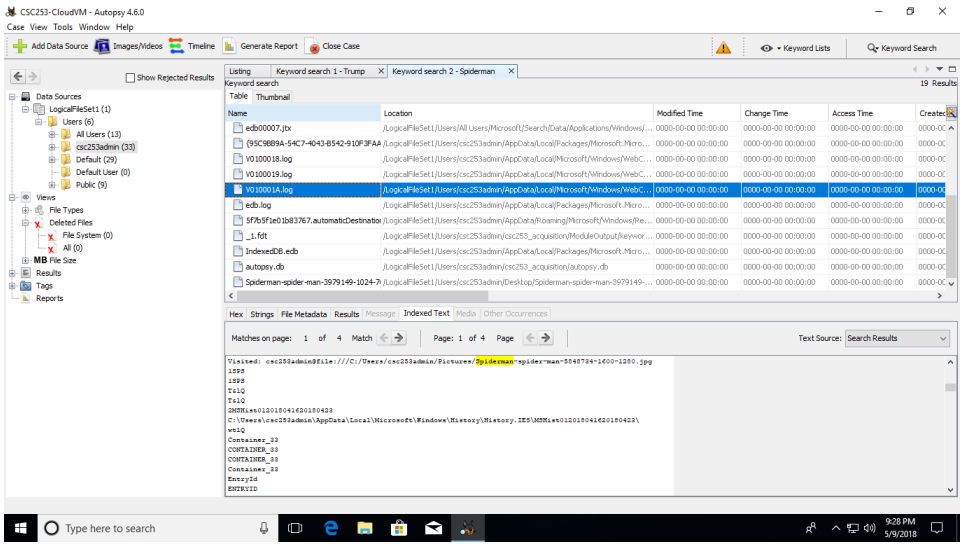

**Fig 22: Dates Metadata Missing from cache information - MS Azure**

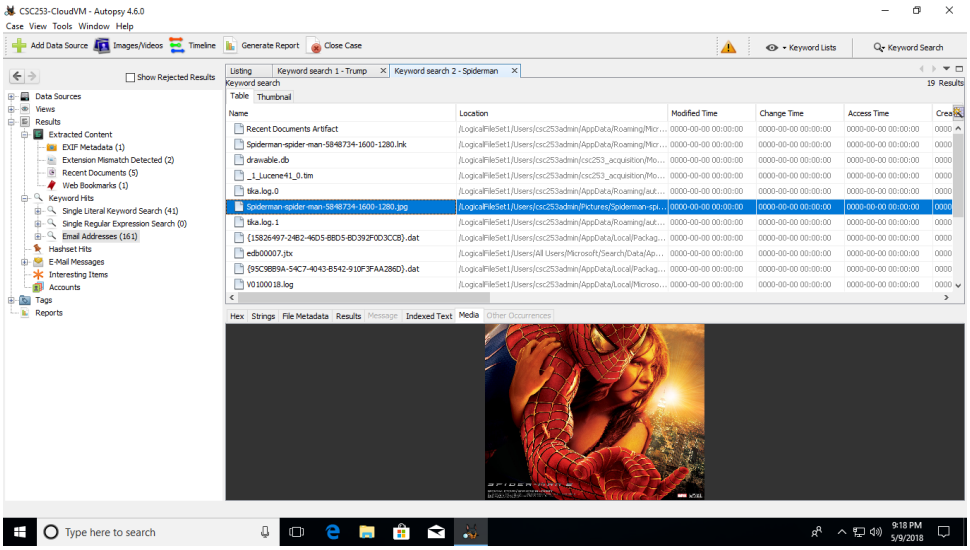

**Fig 23: Spiderman image file - MS Azure**

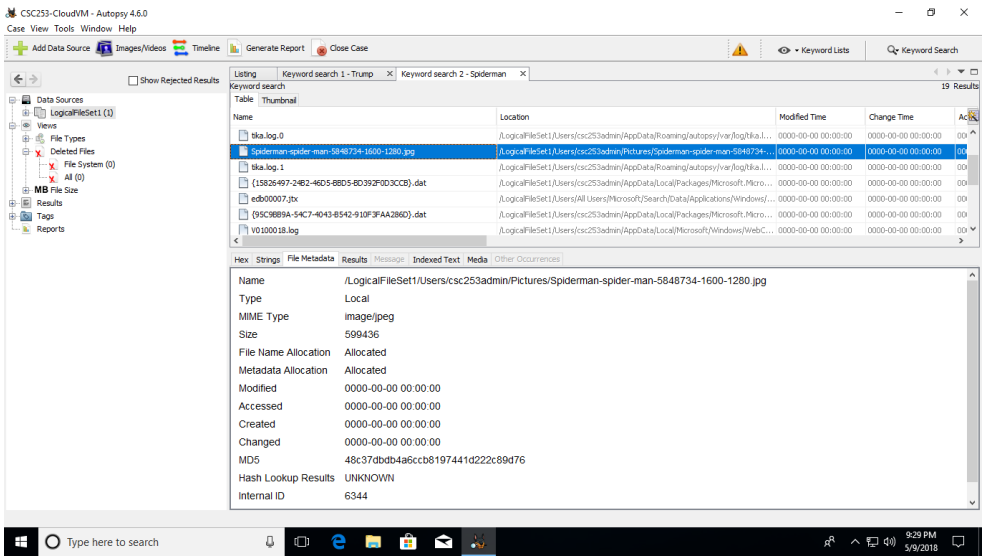


**Fig 24: Date metadata missing from Spiderman image file - MS Azure**

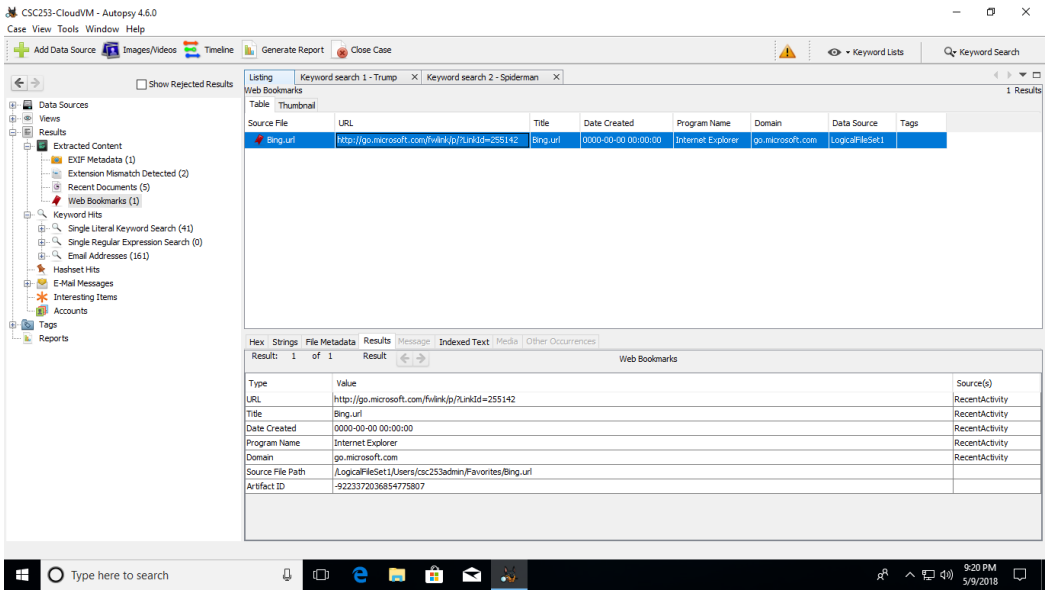


**Fig 25: Bookmarks - MS Azure**

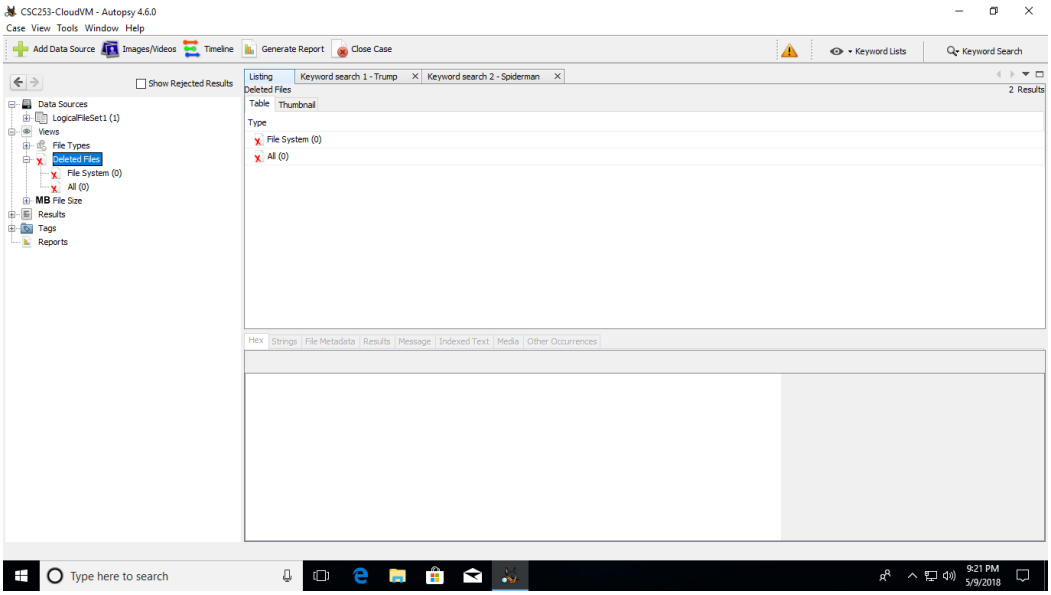


**Fig 26: Deleted Files - MS Azure**

# 5. CONCLUSION

The goal of the project was to analyze the use of a traditional forensics data acquisition to acquire data in a virtual machine environment. The sought to compare the integrity of data gathered to see if it would reflect the true actions of a scenario enacted in each environment. In each environment, the results found both expected results and discrepancies between the expectations to find and acquired. These discrepancies lead us to believe that establishing the integrity of data acquired from a Virtual Machine using traditional forensic software was not possible for our project.

Expected results.

1. In all scenarios we found the file containing the word "Trump" that was seeded in the environment.
2. In all scenarios we were able to find trails of our web surfing activity located in the browser cache. These trails included date metadata.
3. In all scenarios we were able to locate the files that we downloaded, which were not deleted. However, only the local VM(Scenarios 1 and 2) retained data metadata for the file. For MS Azure, non-standard settings must be applied to the configuration to ensure that date metadata is stored for files. If one wanted to ensure that metadata for files would be stored in the cloud VM, that configuration setting would need to be made upon setting up the VM, otherwise such data would not be accessible upon later investigation.
4. In the case of the local VM(Scenarios 1 and 2), were able to acquire the deleted Spiderman image.

The discrepancies fell into two broad categories.

1. **The presence of unexpected data**

In all three scenarios Autopsy software found emails that could not have been placed there by the user (us). There were 1,500+ emails acquired in Scenario 1, 2,700+ emails in Scenario 2 and XXX emails in Scenario 3. Despite the web browsing that we performed in order to

seed these environments, it is unlikely that we came across that many email addresses during our process. These emails may have already been present on the Virtual Image and were inherited by us as the user. Therefore, in a real investigative case, it could be difficult to establish that the suspect was, in fact, in contact with any of the listed emails.

Additionally, one would expect the number of emails from Scenarios 1 and 2 to be the same since, essentially, it was the same environment. The excess number of email in Scenario 2 might be attributable to extra data located inside the .vmdk image that is necessary to operate it. Further investigation is needed to determine the exact reason for the discrepancy.

In Scenarios 1 and 2, there were also deleted images present in the Autopsy results. These are not images that we deleted as the user and we have no prior knowledge of these images. Some of the deleted images contained metadata describing their usage date and some did not. In the case of a true forensic investigation, deleted criminal images would not necessarily be attributable to the suspect. How can we prove that deleted images without accurate metadata belong to suspect activity? On that same note, if some of the data acquired from the forensic investigation cannot be relied upon, how can any of it be? Even those images that contain date metadata could be brought under suspicion in a court of law. Further investigation is needed to determine if/how can truly distinguish between the user's deleted files and deleted files that came as part of the virtual image.

As with the emails, there is also a discrepancy between the number of deleted files reported by Scenario 1 (9000+) and Scenario 2 (10000+). The excess number of email in Scenario 2 might be attributable to extra data located inside the .vmdk image that is necessary to operate it. Further investigation is needed to determine the exact reason for the discrepancy.

**2. The absence of expected data.**

We had expected a record of the deleted file from the cloud VM in Scenario 3. However, there was no record of any deleted files in this environment. The standard configuration of VM storage for MS Azure does not guarantee persistence of data. While there are options for persisting VM data/files via MS Azure we cannot be sure that deleted files would be preserved. Further investigation is called for to understand if/how deleted VM data could be made to persist for a distinct period.

It was expected that date metadata would exist for the Spiderman images that was downloaded from the web. This metadata was not present. Again, the standard configuration of MS Azure VM local storage does not guarantee file metadata preservation. If this type of record is needed, it needs to be configured during VM storage setup.

# 6. SUMMARY

It is clear from our project that more work needs to be done to solidify how forensic data acquisition from virtual machines can be made trust-worthy. Any company or service that utilizes virtual machines as a regular part of their business will need to decide if there is any anticipated or possible need for future forensic data acquisition. If future forensic investigation is to be accommodated for, a few conclusions can be made as a result of our project.

The local virtual machine, which is completely under the control of the owning organization, seems to be more reliable than the cloud VM machine. To compensate for the discrepancies discovered during our investigation, creators of local virtual images should ensure that the image is clean and free of any historical data. Theoretically, a clean virtual image could be trusted to contain only data placed there by the current user, and not data present from the image itself. This hypothesis warrants further investigation.

For MS Azure cloud virtual machines, standard configurations do not facilitate accurate data acquisition. Further investigation needs to be made into how cloud VMs can be configured in such a way to preserve data should a forensic investigation be necessary at a future data.

# 7. REFERENCES

[1] Digital Forensics on a Virtual Machine; University of Alabama website, undated; Juan Carlos Flores Cruz, Travis Atkinson http://atkison.cs.ua.edu/papers/ACMSE11_JF.pdf

[2] Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques; Digital Investigation Volume 9, Supplement, August 2012, Pages S90-S98; Josiah Dykstra, Alan T.Sherman https://www.sciencedirect.com/science/article/pii/S1742287612000266

[3] Evolution of Traditional Digital Forensic in Virtualization; ACMSE '13 Proceedings of the 51st ACM Southeast Conference Article No. 30; Savannah, Georgia — April 04 - 06, 2013; Juan Carlos Flores Cruz, Travis Atkinson https://dl.acm.org/citation.cfm?id=2500078.