December 2023

Virtual Machine Images Preconfigured with Security Scripts for Data Protection and Alerting

Brandon Maltzman

Scott Tyler Ellis

Hauke Vagts

Jim Miller

Devon Beanish

See next page for additional authors

Inventor(s)

Brandon Maltzman, Scott Tyler Ellis, Hauke Vagts, Jim Miller, Devon Beanish, and Assaf Namer

Virtual Machine Images Preconfigured with Security Scripts for Data Protection and Alerting

ABSTRACT

Developers use interactive development environments (IDEs) to create and share documents that contain live code, equations, visualizations, narrative text, etc. as part of the artificial intelligence/ machine learning (AI/ML) development process. Virtual machines that run IDEs may have access to private and/or sensitive data used during model training or use. For data security and compliance, it is necessary to highlight and track the VMs that have been in contact with sensitive information. This disclosure describes techniques to automatically identify and label the presence of sensitive data in virtual machines and disks as part of machine learning. Custom VM images are provided that include data scanning scripts that can identify the presence of sensitive data during or after usage, e.g., by a developer using an IDE. The scripts can automatically log the presence of data and generate alerts. Users of such virtual machines are provided additional controls to perform the training process in a secure and confidential manner in compliance with applicable data regulations.

KEYWORDS

- Data security
- Data governance
- Data loss prevention
- Data scanning script
- Sensitive data
- De-identification
- Personally identifiable information (PII)
- Protected health information (PHI)

BACKGROUND

Interactive development environments (IDEs) enable developers and data scientists to create and share documents that contain live code, equations, visualizations, and narrative text. Data scientists use such tools for collecting, formatting, and analyzing data as part of the artificial intelligence/ machine learning (AI/ML) development process.

To build and train AI models to be effective and accurate, access to unobfuscated, cleartext data is necessary. As part of the AI/ML development process, virtual machines that run IDEs may have access to private and/or sensitive data, such as personally identifiable information (PII) and/or protected health information (PHI). Virtual machines (VMs) that run IDEs may have ephemeral or permanent disks attached to which such data may be written. Such data may also be present in temporary checkpoint files that are not visible to the user. While the disks may have encryption at rest enabled, there is still a need, from a security operation standpoint, to highlight and track the VMs that have been in contact with sensitive information. While cloud service providers that provide VMs to run IDEs provide data deletion processes, in many cases, the providers retain data for a long period of time, e.g., 100+ days.

Data governance and observability of data in a cloud environment is critical to ensure compliance with regulations, protect customer privacy, and mitigate risks. While data that is used to train models can be obfuscated or de-identified to avoid exposure of sensitive or user confidential data, these techniques are not always employed, for example, due to specific business requirements or even misconfiguration. Multiple layers of defense and auditability where sensitive data is accessed and processed are needed to establish sufficient data security.

DESCRIPTION

This disclosure describes techniques to automatically detect and label the presence of sensitive data in virtual machines and disks as part of machine learning. Virtual machines that are used for training models are usually offered and controlled by cloud service providers (CSP). Custom VM images are provided that include data scanning scripts that can identify the presence of sensitive data during or after usage, e.g., by a developer using an IDE. The scripts can automatically log the presence of data and generate alerts. Users of such virtual machines are provided additional controls to perform the training process in a secure and confidential manner in compliance with applicable data regulations.

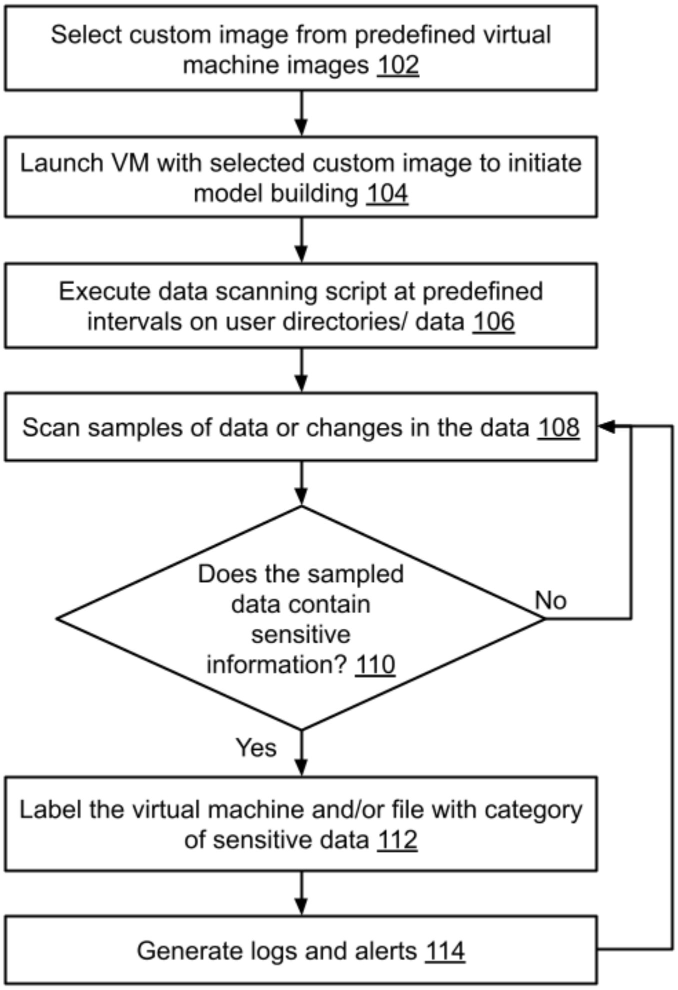
Select custom image from predefined virtual machine images 102
Launch VM with selected custom image to initiate model building 104
Execute data scanning script at predefined intervals on user directories/ data 106
Scan samples of data or changes in the data 108
Does the sampled data contain sensitive information? 110
No
Yes
Label the virtual machine and/or file with category of sensitive data 112
Generate logs and alerts 114

Fig. 1: Identification of sensitive data usage in machine learning model building

Fig. 1 illustrates an example process for identification of sensitive data usage during or post machine learning model development processes using a custom VM image. Pre-built custom virtual machine images are created and provided for users to select from (102) to utilize while building new machine learning models using IDEs. Users are required (by process, policy, or technical controls) to only use one of the custom images for such a task. Each image includes a script, e.g., as part of the image itself, as part of the image metadata, or by a scheduler job, for example with Serverless architecture, or other event driven platform, etc. The script is configured to scan user directories and components of the VM for sensitive information.

Upon user selection of a custom image, a VM is launched (104) based on the selected image to enable the user to initiate model building. The data-scanning script is executed (106) at predefined intervals on user directories/datasets. The script can be a start-up and/or shutdown script that executes when those events occur, a manual script that can be launched on demand, and/or a scheduled job that runs at pre-configured intervals.

The script can be provided by users and used in the training workflow. Since the VMs that execute the training process are managed by a CSP, and users have no control over or access to the VMs, the provisioning of custom images with scripts provides users the ability to add control to the training process (that is invisible to the users) and helps ensure compliance. The scripts can also include logic to send publish/subscribe or log messages if any sensitive or confidential information is identified during the model training/tuning process. This additional control can provide users valuable confirmation about the classification of data used by the CSP, giving customers another layer of data governance and observability.

Data scanning scripts can be initiated in multiple ways such as VM startup script, VM shutdown script, manually launched or automated script, etc. The script can leverage a data loss prevention (DLP) service and/or a sensitive data protection service. To reduce costs and improve efficiency, sample data from user directories and/ or other components of a virtual machine can be analyzed (108) as opposed to performing a complete data scan. The script can be set up to scan only the changes in data from a previous iteration of the scan.

If sensitive data is detected (110) in the sample data, a label can be applied (112) to the virtual machine. The label can indicate the type of sensitive data found. Labels can be assigned at a higher access control permission to prevent tampering. Logs can be generated (114) indicating type of information found, VM details, etc. An alert can also be triggered for a designated contact or system, e.g., via publish/subscribe or other techniques, to enable further action when sensitive data is found. The logs can be matched to a VM, including VM shutdown logs.

Scanning is continued even after a label has been assigned and logs are generated. For example, at the initial time the scan is performed, the assigned label (112) may indicate the category of data found (e.g., category 1) and logging and altering (114) is performed. When scanning is continued, it is possible that a subsequent sample may be identified as including category 2 data. At this iteration, the additional label (e.g., category 2) is added and further logging and alerting is performed. Note that the additional label may also be for a previously found category if a subsequent sample includes data for that category. In this manner, by continuing the iterative scanning process, additional instances of data samples may be identified and labeled, with additional logs/ alerts generated as applicable.

The use of VMs that are preconfigured with automated scripts can generate valuable information about the use of data during model training and provide an additional layer of data governance for both service providers and users. Data governance and observability of sensitive data in the cloud environment can ensure compliance with regulations, protect user privacy, and mitigate risk. By implementing the described techniques, organizations can gain further visibility into data use, identify use/storage of sensitive data, and monitor usage of the data for purposes such as model training. This information can then be used to enforce data protection, including actions such as creating an inventory of processed sensitive data, encrypting, masking, restricting access to certain data, monitoring and preventing unauthorized access or use of data, and more. Additionally, the generated logs can be used to track the type of data used in a model training pipeline to ensure proper use and to support auditing.

Organizations can, thus, develop an understanding of the usage, movement, and storage of data and can take action upon potential inappropriate use. This can help enable the use of appropriate data encryption or de-identification techniques by automatically surfacing insights that indicate where these techniques should be applied as well as equip the organization with the ability to take action on any detected misuse of data. For example, the techniques can automatically create an alert in a security information and event management (SIEM) tool when sensitive data is used in the training process.

As the use of AI/ML models is on the rise, the risk to organizations due to inappropriate data use is also increasing. The described techniques provide organizations mechanisms to address compliance requirements and protect their users. The techniques provide organizations a mechanism to automatically identify, alert, and take action on sensitive information being used in their environment and to implement safeguards to protect sensitive data.

Cloud service providers that offer the described techniques will enable their customers with enhanced visibility to make more informed decisions about the data they process and how it's used and protected. Implementation of the techniques by a CSP helps their customers with regulatory compliance, incident response, and data governance tasks.

CONCLUSION

This disclosure describes techniques to automatically identify and label the presence of sensitive data in virtual machines and disks as part of machine learning. Custom VM images are provided that include data scanning scripts that can identify the presence of sensitive data during or after usage, e.g., by a developer using an IDE. The scripts can automatically log the presence of data and generate alerts. Users of such virtual machines are provided additional controls to perform the training process in a secure and confidential manner in compliance with applicable data regulations.

REFERENCES

1. Zhang, Tianwei, Zecheng He, and Ruby B. Lee. "Privacy-preserving machine learning through data obfuscation." *arXiv preprint arXiv:1807.01860* (2018).
2. “Cloud Data Processing Addendum (Customers)” available online at https://cloud.google.com/terms/data-processing-addendum accessed Dec 8, 2023.
3. “Custom training overview | Vertex AI | Google Cloud” available online at https://cloud.google.com/vertex-ai/docs/training/overview accessed Dec 8, 2023.
4. “Using startup scripts on Linux VMs | Compute Engine Documentation | Google Cloud” available online at https://cloud.google.com/compute/docs/instances/startup-scripts/linux accessed Dec 8, 2023.

5. “Running shutdown scripts | Compute Engine Documentation | Google Cloud” available online at https://cloud.google.com/compute/docs/shutdownscript accessed Dec 8, 2023.

6. “Overview | Resource Manager Documentation | Google Cloud” available online at https://cloud.google.com/resource-manager/docs/labels-overview accessed Dec 8, 2023.