empath™

Skills Intelligence Platform

TOMORROW'S WORKFORCE. TODAY.

Protecting Corporate Data Provided For Machine Learning Inferencing

Adam Blum

Co-Founder / CTO

2022

Empath, Inc.

Protecting Corporate Data Provided For Machine Learning Inferencing

Use of corporate data to build and train machine learning models that influence business decisions is exploding. Sometimes this data is provided to internal company data science teams. Other times the data is provided to external systems and software vendors.

Empath™ uses internal client data about employees and their activities to infer skills for companies. In the process, we have learned several best practices for performing this very sensitive activity. These practices aim to protect company data and IT security while still allowing for the building of accurate models. In this document we will describe these best practices. Many of these recommendations are not performed by most external vendors when handling sensitive customer data. This document will provide you with useful insights towards providing data to external vendors. We hope these practices become commonplace, as their use should help accelerate the trend of using external machine learning-oriented vendors for more critical business processes.

No Live Integration

Don't perform live real-time integration to retrieve corporate data from your operational systems for use in machine learning training. Machine learning models are not trained in real-time as new data arrives (though predictions and inferences based on trained models may well be close to real-time). Thus, live integration isn't necessitated by how the data is used. It is far better to deliver the data in batch form.

Insisting on live integration is an unnecessary risk for your corporate IT infrastructure. It will typically slow down getting data to the machine learning provider (the "data processor" in GDPR parlance). The ongoing cost and effort to keep the external vendor permissions current and monitored is simply not worth it if there are better alternatives.

Data Provision via External Data Sharing

Given the aforementioned, we recommend instead that you deliver data to the vendor via an external data sharing mechanism. The best default is via secure FTP (SFTP). You should use your public SSH key for access versus being granted a username and password. There are plenty of other secure alternatives including AWS S3 and Azure Data Share.

Vendor Security Policies

Once your data is being used externally, you need to ensure that it is being handled with care and respect by the vendor organization. That organization needs to have policies in place that carefully circumscribe who has access to the data and what they can do with it. Vendor processes need to include separation of responsibilities so that no single rogue employee can abuse access to the data without others becoming aware of it. The policies should ensure that the principle of least privilege is applied: employees should have the minimum rights necessary to perform their job roles. SOC 2 is a worthwhile policy framework to start with. You should examine the vendor's documented security policies and ensure that they meet the requirements laid out by SOC 2.

Dedicated Cloud Data Infrastructure

When the data is processed by the vendor it is typically loaded to some data infrastructure to preprocess it for machine learning training. That infrastructure, often called a "data lake", should ideally be dedicated to your company's data. This could be a hosted relational database, such as Postgres. It could also be a cloud data platform such as Databricks or Snowflake. The key practice recommended is to provide administrative rights to this infrastructure for a member of your IT staff.

Failing that, you will want to be able to perform an audit of the vendor's data lake and the progress of your data through it, including access rights. A certified SOC 2 audit ensuring the security policies discussed earlier is a good step towards this.

Encryption for Storage & Transmission of Training Data

You should ensure that the data is stored by the vendor in the data lake dedicated for your data in encrypted form (encryption at rest). In addition, data should always be encrypted in transit, for example encrypting postgres connections with SSL. Some key management services (examples include AWS Key Management Service and Azure Key Vault) should be used to manage the encryption keys themselves.

Anonymization

If your data contains personally identifiable information (PII) of employees, you should anonymize the data before supplying it to the vendor, wherever possible. Doing so will ease compliance with GDPR, German Works Council, and other data privacy regulations. PII includes names and email addresses, but other information such as phone numbers and employee IDs may need to be scrubbed as well. Your vendor may be able to supply you with scripts to execute the redact of such information. Remember that if you choose to use redaction methods you should always carefully examine any such scripts.

CONCLUSION

Following these guidelines when providing data to vendors who will perform machine learning on it should reduce data provision risk drastically. Such efforts will even increase data security versus the typical attempt to build internal machine pipelines. This is especially the case if such efforts happen to use cloud data and machine learning infrastructure.

Protecting Corporate Data Provided For Machine Learning Inferencing

empath™

Skills Intelligence Platform
TOMORROW'S WORKFORCE. TODAY.

info@empath.net
empath.net