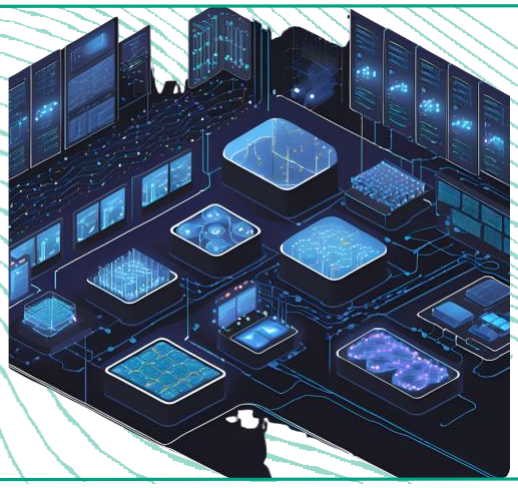
人工智能数据全链路安全



第一步：了解你的数据、你的人工智能以及你所面临的风险

▢库存您的人工智能全景：

- ▢ **识别所有AI举措**：全面盘点贵组织内的所有AI项目和应用。

  示例
  这包括内部开发的人工智能模型、第三方人工智能服务，甚至现有软件中由人工智能驱动的简单功能。

- ▢ **对人工智能应用进行分类**：根据其类型和目的对每项人工智能计划进行分类。

  示例
  它是像聊天机器人或内容生成器那样的生成式人工智能工具吗？还是用于风险评估的预测模型？了解每种人工智能应用的具体能力和局限性是评估其潜在风险的关键。

- ▢ **文档数据来源与流动**：对于每项人工智能（AI）计划，需明确其使用的数据来源、处理的数据类型（如客户数据、财务信息、专有代码）以及这些数据在不同系统间的流动方式。

  示例
  人工智能应用程序是否访问存储在云存储、本地数据库或外部应用程序编程接口（API）中的数据？使用图表可视化这些数据流有助于发现潜在的漏洞。

▢ 确定你的AI应用范围：

- ▢ **利用生成式人工智能安全范围矩阵**：该矩阵在参考资料中有所介绍，它提供了一个框架，根据贵组织对人工智能模型及相关数据的控制程度，将人工智能使用情况划分为五个范围。

示例
使用像ChatGPT这样的公共人工智能服务属于范围1，而从头开始构建和训练一个定制的人工智能模型则对应于范围5。每个范围都有不同的风险特征，了解你所处的范围对于制定合适的安全策略至关重要。

▢ 绘制数据流图：

- ▢ **创建可视化表示**：超越简单的文档记录，为每个AI应用开发可视化数据流图。
- ▢ **突出潜在漏洞**：精确识别敏感数据传输、存储或处理的区域。这些区域是实施安全控制的首选目标。

示例
如果您的数据流图显示人工智能聊天机器人会访问包含客户个人身份信息的数据库，那么您需要确保该数据库具备强大的访问控制和加密措施。

步骤2：逐步构建你的AI安全防线

▢ 身份与访问管理（IAM）：

- ▢ **最小特权原则**：这一原则至关重要。仅授予用户和系统执行其任务所需的最小访问权限。过度宽松的访问权限是灾难的根源。
- ▢ **强认证**：实施稳健的认证机制，包括多因素认证（MFA），以在授予用户访问人工智能系统和数据的权限之前验证其身份。
- ▢ **细粒度授权**：不要仅依赖基本身份验证。利用基于角色的访问控制（RBAC）为不同的用户角色定义细粒度权限。这确保用户只能执行与其工作职能相关的操作和访问相关数据。
- ▢ **控制模型推理端点的访问**：限制对允许与人工智能模型交互的应用程序编程接口（API）的直接访问。就像你不希望未经授权的用户直接连接到你的生产数据库一样，你也需要保护你的AI模型免受未经授权的推理请求的侵害。

示例
如果您使用的是Amazon Bedrock，请利用AWS身份与访问管理（IAM）来管理调用模型推理的权限。

- ▢ **安全API密钥**：如果您的AI应用程序依赖API密钥进行身份验证，请将这些密钥视为高度敏感信息。请安全地存储它们，定期更换它们，并实施机制以撤销已泄露的密钥。

▢ 数据保护，始终如一：

- ▢ **加密静态和传输中的数据**：加密是您最好的朋友。对存储在数据库、云存储和任何其他位置的敏感数据进行加密。同样重要的是对传输中的数据进行加密，尤其是在通过网络传输信息或与人工智能应用程序交互时。

示例
如果您使用Amazon S3来存储人工智能模型工件，请为您的S3存储桶启用加密功能。对于用于模型微调的敏感数据，考虑使用客户管理的AWS KMS密钥以增加额外的控制层。

- ▢ **数据最小化**：不要收集或存储非人工智能项目绝对需要的数据。你拥有的数据越少，你的攻击面就越小。

示例
如果您的AI模型仅需处理客户姓名和电子邮件地址，则无需存储如社会保险号或信用卡详细信息等敏感信息。

- ▢ **去标识化**：在可能的情况下，从数据集中移除或屏蔽个人身份信息（PII）。这对于用于训练人工智能模型的数据尤为重要，因为它降低了敏感信息泄露的风险。

▢ 人工智能时代的威胁建模：

- ▢ **了解人工智能特定威胁**：传统的威胁建模技术需要不断发展，以应对人工智能系统特有的脆弱性。

示例
提示注入、后门攻击和数据投毒只是针对人工智能应用的一些具体威胁示例。

- ▢ **应对提示注入**：提示注入是一种关键威胁，攻击者通过精心设计的恶意输入来操纵人工智能的输出。为降低此风险，可采取以下措施：
  - ▢ **用户输入的净化**：实施输入验证和净化技术，以过滤掉潜在的恶意字符或命令。
  - ▢ **限制模型对敏感信息的访问**：设计应用程序以防止人工智能模型访问或泄露不应访问的信息。资料中提供了一个示例，即在将敏感信息（如客户的欺诈评分）纳入提供给模型的上下文之前，对其进行编辑处理。
  - ▢ **使用亚马逊基石防护工具**：基石防护工具允许您定义策略和过滤器，以检测和阻止用户输入和人工智能输出中的不良内容。
- ▢ **保障供应链安全**：若您正在使用预训练的人工智能模型或第三方人工智能服务，请确保模型产物的完整性和真实性。

示例
验证模型的来源，使用校验和或数字签名来检测篡改，并扫描模型以查找已知漏洞。

第三步：赋能员工，促进协作，永不停歇地学习

▢ 让安全成为每个人的责任：

- ▢ **定期安全意识培训**：针对人工智能相关风险和最佳实践，开展持续的安全意识培训。向员工传授以下内容：
  - ▢ **数据处理政策**：明确传达关于敏感数据处理（包括人工智能训练所用数据和人工智能应用生成的数据）的政策。
  - ▢ **人工智能特定威胁**：阐述与人工智能相关的风险，如提示注入和数据投毒，并提供这些攻击如何发生的示例。
  - ▢ **报告程序**：建立清晰的渠道，用于报告可疑活动、潜在数据泄露或对人工智能安全的担忧。
- ▢ **培养安全文化**：鼓励员工树立安全至上的心态，让他们敢于识别和报告潜在问题，而不必担心遭到报复。

▢ 拥抱社区的力量：

- ▢ **行业合作**：参与行业论坛，出席会议，并与同行交流，分享有关新兴人工智能安全威胁、最佳实践以及经验教训的信息。
- ▢ **保持知情**：人工智能安全领域在不断发展。订阅相关的安全博客、新闻通讯和威胁情报源，以掌握最新动态。

▢ 持续监测与改进：

- ▢ **实施监控工具**：部署安全信息和事件管理（SIEM）系统、入侵检测与防御系统（IDPS）以及其他安全监控工具，以检测并响应人工智能基础设施中的可疑活动。
- ▢ **定期安全审计**：定期进行安全审计和渗透测试，以发现人工智能系统和数据管道中的漏洞。
- ▢ **适应与提升**：利用从监控、审计和事件响应中获得的洞察，持续改进您的人工智能安全态势。