数据安全

第4章 数据采集与传输安全

计算机与大数据学院 刘延华

福州大学
FUZHOU UNIVERSITY

数据安全

自动驾驶数据安全白皮书

（2020）

与传统数据类似，自动驾驶数据安全的特性也主要表现为机密性、完整性、可用性。自动驾驶数据的机密性是指用户隐私数据、测试场景数据、人机交互数据等不泄露给未授权的个人、实体、进程，并保证其不会被利用的特性。自动驾驶数据的完整性是指自动驾驶决策与控制数据、动态交通环境数据等没有遭受以未授权方式所作的更改或破坏，保证自动驾驶车辆信息数据的正确生成、存储和传输的特性。自动驾驶数据的可用性是指已授权的个人、实体一旦需要就可以访问和使用自动驾驶数据和资源的特性。

第4章 数据采集与传输安全

数据采集安全的概述

数据采集的完整性：

在客户端采集数据，为了保证尽量不影响用户体验，所以在采集数据时，一般不会同步发送数据，而是在本地先做缓存，然后再整体压缩、打包并在网络好时一起通过公网进行传输。

- 如果客户端一直网络不好，传输失败时，则会累计在本地，而本地缓存会有限额；在缓存数据全部发送前，客户端若被卸载，则都会丢掉部分数据。
- 在 Web 端传输数据时，虽然是同步发送，不过由于公网传输的网络问题，一般也会有 3% 到 7% 的数据丢失，并且基本难以避免。

数据采集安全的概述

数据采集的隐私性：

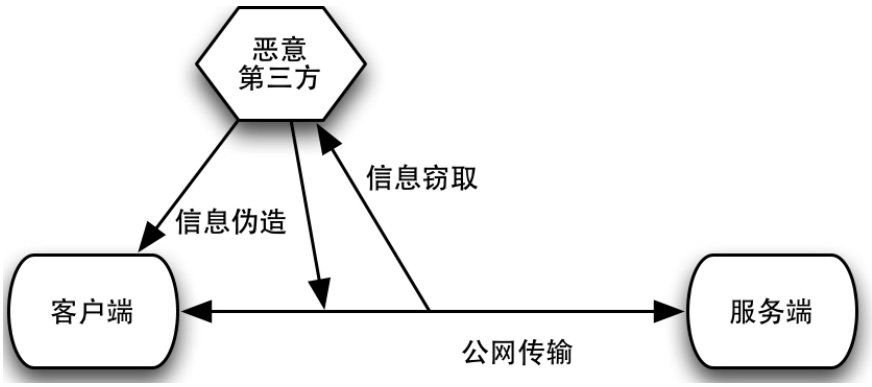
第三方可能会在传输过程中截获传输的数据，从而拿到传输的这些用户行为数据。这些用户数据都是体现用户在客户端的一些具体的用户行为，蕴含着用户的隐私。

数据采集安全的概述

数据采集的准确性：

第三方可能会在传输过程中伪造数据，从而让后台的分析结果不准确。这种伪造可能是直接调用传输的 API，可能是在多个模拟器上运行 App，甚至可能是直接人工作在真实设备上操作 App，都会导致传输到服务端的数据不准确。

数据采集安全的概述



数据采集与传输的安全

数据采集安全的概述

常用的技术解决方案：

1、使用 HTTPS 作为传输协议

HTTPS 是一种网络安全传输协议，采用HTTP、SSL/TLS来对数据包进行加密与传输。

- 能提供对服务器的身份认证，保护交换数据的隐私与完整性。
- 由于数据采集在客户端，HTTPS不能解决数据完整性的问题；
- 同样，HTTPS 不能阻止数据的伪造。

数据采集安全的概述

常用的技术解决方案：

2、传输内容加密

- 不仅通过传输协议加密，对于传输的内容也进行加密。
- 可以阻止恶意第三方拿到传输协议，从而没有办法通过直接调用 API 的方式来进行数据伪造；
- 但对于客户端的数据伪造，依然是无能为力；
- 仅对传输内容进行加密，不能改变是在客户端采集数据，以及通过公网传输数据的本质，所以并不能解决数据完整性的问题。

数据采集安全的概述

人工操作刷数据很难识别

数据采集安全的概述

常用的技术解决方案：

3、采集后再进行数据清洗

- 基于统计信息的数据清洗：例如，把那些流量明显大于平均值的设备或者 IP 的用户行为过滤掉，把那些行为频率明显超过正常人限度的用户行为过滤掉等；
- 基于用户行为特征的数据清洗：采用一些机器学习的手段，通过对整体的用户行为进行训练，找到异常用户和用户的异常数据；
- 基于设备真实性的数据清洗：资产通常身份哈希码，通过识别一个设备是一个真实的设备，还是一个模拟器，来解决数据造假问题。 *刷数据*

数据采集安全的概述

数据采集安全是数据安全生命周期的第一个过程，是对数据来源安全的管理，是数据安全能力成熟度模型(Data Security Capability Maturity Mode,简称DSMM)的基础阶段，是后续工作的基础。

该过程包含四个过程域，分别为：数据分类分级、数据采集安全管理、数据源鉴别及记录、数据质量管理。

数据采集安全

数据采集过程
+
标准规范
（视角）

- 数据分类分级
- 数据采集安全管理
- 数据源鉴别及记录
- 数据质量管理

数据采集要遵循最小够用和可用不可见原则，前者要求在数据采集使用方面要求确保数据专事专用、最小够用，杜绝过度采集、误用、滥用数据，切实保障数据主体的数据所有权和使用权。
--2020年央行发布《金融业数据能力建设指引》

数据分类分级

- 基于法律法规，以及业务需求确定组织机构内部的数据分类分级方法，对生成或收集的数据进行分类分级标识。
- 数据分类分级是数据采集阶段的基础工作，也是整个数据安全生命周期中最基础的工作，它是数据安全防护和管理中各种策略制定、制度落实的依据和附着点。

数据分类分级

具体要求：

- 组织建设：

组织机构设立负责数据分类分级工作的管理岗位和人员，主要负责定义组织机构整体的数据资产分类分级的安全原则以及相关能力提供。

数据分类分级

- 制度流程：

① 建立数据资产分类分级原则、方法和操作指南。

② 对组织机构的数据资产进行分类分级标识和管理。

③ 对不同类别和级别的数据建立相应的访问控制、数据加解密、数据脱敏等安全管理和控制措施。

④ 建立数据分类分级变更审批流程和机制，通过该流程保证对数据分类分级的变更操作及其结果符合组织机构的策略要求。

数据分类分级

- **技术工具：**

建立数据分类分级打标或数据资产管理工具，实现对数据资产的分类分级自动标识、标识结果发布、审核等功能。

在技术层面需要建立数据管理平台，按照数据分类分级原则和制度要求对数据打标签，进行数据分类和分级区分，并依据此设置访问控制策略和加解密策略，还要能够对新增数据根据要求进行自动打标签处理。

数据分类分级

- **人员能力：**

负责该项工作的人员应了解数据分类分级的合规要求、能够识别哪些数据属于敏感数据。

数据分类分级

数据分类					数据分级
序号	业务类别	数据一类	数据二类	重要性分类	
1	在线业务	用户信息	用户昵称、在线状态	一般	公开
2			用户姓名、手机号、地址	重要	秘密
3		订单信息	订单号、订单价格、订单地址	重要	秘密
4			订单进账、内部统计分析	重要	机密
5	人事业务	xxxx	xxxx	一般	公开
6		xxxx	xxxx	重要	秘密

数据分类分级

数据分类分级的常用技术：

1、正则表达式匹配：正则表达式匹配的是敏感信息字段，针对某些定义字段的匹配；

2、机器学习：采用机器学习算法，对样本进行学习，并用测试集进行验证，注意学习样本和测试样本不能混淆，机器学习对于数据分类分级的方法更为通用。

数据采集安全管理

- 在采集外部客户、合作伙伴等相关方的数据的过程中，需明确采集数据的目的和用途，确保数据源的真实性、有效性和最少够用等原则要求，并规范数据采集的渠道、数据的格式以及相关的流程和方式，从而保证数据采集的合规性、正当性和执行上的一致性，符合相关法律法规要求。

数据采集安全管理

- 数据采集过程中涉及包含个人信息及商业数据在内的海量数据，现今社会对于个人信息和商业秘密的保护提出了很高的要求，需要防止个人信息和商业数据滥用，采集过程需要信息主体授权，并应当依照法律、行政法规的规定和与用户的约定，处理相关数据；另外还应在满足相关法定的规则的前提下，在数据应用和数据安全保护之间寻找适度的平衡。

数据采集安全管理

- 法律要求

采集的数据及采集过程严格按照《网络安全法》、《个人信息安全规范》等相关国家法律法规和行业规范执行。

数据采集安全管理

- 基本要求

a) 采集的数据信息，包括但不限于数据、文本、文件、图片、音频和视频等；

b) 采集数据的的传输方式，包括但不限于有线通讯传输、无线通讯传输和数字通讯传输等方式；

c) 数据采集者（信息系统服务方）应设置专人负责信息生产或提供者的数据审核和采集工作；

d) 数据采集者（信息系统服务方）应明确数据来源、采集方式、采集范围等内容，并记录存档；

数据采集安全管理

- 基本要求

e) 数据采集者（信息系统服务方）应制定标准的采集模板、数据采集方法、策略和规范，采集策略参数配置应包括采集周期、有效性、检测时间、入口地址和采集深度等；

f) 对于初次采集的数据，应采用人工与技术相结合的方式根据其来源、类型或重要程度进行分类；

g) 最小化采集数据，仅需要完成必须工作即可；

h) 对采集的数据进行合理化存储，依据数据的使用状态进行及时销毁处理。

数据采集安全管理

- 采集方式

数据采集包括实时监测收集（系统运行数据、威胁数据等）和系统生产基础数据（人员信息、财务账单、采购供应商等）。

可包括手工录入填报、权限获取、传感器收集、格式化的数据导入及数据ETL等。

ETL，一种数据仓库技术，用来实现将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。**ETL**一词较常用在数据仓库，但其对象并不限于数据仓库。

数据采集安全管理

- 采集周期

数据采集周期分为两种：

1) 对于实时监测数据，采集周期应按照实际工作条件下，系统连续进行10次采集，10次采集时间的平均值作为系统的数据采集周期；

2) 对于系统生产基础数据采用固定期限加动态调整。变化不大的数据信息采集周期为6个月，涉及数据信息变动的调整的可根据需要动态调整。

数据采集安全管理

- 技术工具

1) 加密：在数据采集前端和采集传输路径安全方面，至少对秘密级以上数据采用加密措施，包括但不限于采集程序本身的加密（如DES、3DES）、传输过程加密（SSL）、网络层加密（VPN）、链路加密（专线）等方式；

2) 完整性：在数据采集前后采取校验码等技术对数据完整性进行校验，包括但不限于：数字签名、Hash算法校验、文件大小比对、人工复验等方式；

数据采集安全管理

- 技术工具

3) 匿名：对采集数据在采集和传输过程及存储过程中涉及展示的情景下，对数据进行脱敏和匿名模糊，包括但不限于数据信息替换、数据内容截取、模糊处理等方式；

4) 审计日志：数据从采集开始的整个过程，提供所有采集操作的日志记录，日志记录内容包括但不限于日期、时间、操作类型（动作）、主体（操作者）、客体（被操作对象）、状态等；

5) 断网自动保护：在进行采集的过程中，如遇网络中断，需将已采集的数据缓存在采集前端设备，保证15天内继续对数据进行采集且系统不丢失数据，待网络恢复后自动续传采集的数据。

数据采集安全管理

- 风险评估

在对数据进行采集的过程中，应组织风险评估小组，对采集过程进行风险评估，评估内容包括但不限于：

a) 采集过程是否合规：是否有采集负责人进行审核等相关采集操作、采集的数据是否最小化、采集等；

b) 采集过程过程安全要求：是否采用了加密、完整性校验、匿名、日志和断网保护等措施；

数据源鉴别及记录

- 对产生的数据源进行身份鉴别和记录，防止数据仿冒和伪造。

数据源鉴别是指对收集或产生数据的来源进行身份识别的一种安全机制，防止采集到其它不被认可的或非法数据源（如机器人信息注册等）产生的数据，避免采集到错误的或失真的数据；

数据源记录是指对采集的数据需要进行数据来源的的标识，以便在必要时对数据源进行追踪和溯源。

数据源鉴别及记录

- 技术工具：

采取技术手段对外部收集的数据和数据源进行识别和记录，即通过数据溯源的机制，保证数据管理人员能够追踪与其加工和计算数据相关的数据源。

对关键溯源数据进行备份，并采取技术手段对溯源数据进行安全保护。

数据源鉴别及记录

- 人员能力：

负责该项工作的人员应理解数据源鉴别鉴别标准和组织机构内部数据采集的业务场景，能够结合实际情况执行。

1) 在进行数据采集时，需要专人或团队对数据源进行鉴别和溯源管理，提供数据源管理策略和方案。

2) 在进行数据采集时，需要对数据采集源进行识别和标识。可采取数据标签的形式，确保数据唯一性。

3) 对数据采集源进行身份鉴别，防止数据源假冒和伪造。包括但不限于使用用户名/口令认证、指纹识别、人脸识别、动态口令卡、短信（语音）验证码、USB-Key等鉴别方式。

数据源鉴别及记录

- 人员能力：

4) 在数据生命周期整个过程中，需要对采集的数据进行溯源管理，将数据每次操作前后的情况和状态进行日志记录和保存，以便对数据进行溯源。可采用源数据管理系统Apache Atlas、数据血缘管理工具Cloudera Navigator Data Management等。

5) 对溯源数据进行传输和存储时，需要采取加密和完整性校验技术保证数据安全。包括但不限于SSL、VPN、MD5、RSA、RC4等。

6) 在溯源数据过程中，需要对关键溯源数据进行备份，并采取加密和完整性校验技术进行安全保护。

数据质量管理

- 建立组织机构的数据质量管理体系，保证对数据采集过程中收集/产生的数据的准确性、一致性和完整性。

数据安全保护的对象是有价值的数据，而有价值的前提是数据质量要有保证，所以必须要有数据质量相关的管理体系。

目的是保证对数据采集过程中收集和产生的数据的准确性、一致性和完整性。

数据质量管理

- **技术工具：**

利用技术工具实现对关键数据进行数据质量管理和监控，实现异常数据及时告警或更正。

1、对数据资产进行分类和等级划分；

2、对在线数据的质量监控，比如针对业务数据库实时产生的数据，这就要求需要对业务数据进行定义并对流程进行改造实现实时监控；

3、对离线数据质量监控，比如针对数据仓库或数据开发平台的离线数据；

4、提供数据质量事件的处理流程，一旦发现数据质量异常及时进行告警和上报，积极采取纠正措施。

数据质量管理

- 人员能力：

负责该项工作的人员对数据质量管理规范有一致性理解，能够基于组织机构的实际数据质量管理需求开展相关工作。

1) 数据质量进行管理要贯穿数据全生命生命周期。

2) 需要设置专门的岗位和人员，负责制定数据质量管理规范及对数据质量进行管理和监控。

3) 需要对数据完整性进行定义和监控。如人员信息要完整覆盖姓名、性别、年龄等，保证没有遗漏。

4) 需要对数据规范性进行定义和监控。如日期信息都以yyyy-mm-dd格式存储，保证数据规范统一。

5) 需要对数据一致性进行管理和监控。如同一个人的性别信息在从不同的数据库表中取过来应该是一致的。

数据质量管理

- 人员能力：

6) 需要对数据准确性进行定义和监控。如人员信息的年龄应该在0-120，超出此范围即为不合理不准确。

7) 需要对数据唯一性进行管理和监控。如同一个ID应该没有重复记录，确保数据唯一不重复。

8) 需要对数据关联性进行管理和监控。如两张数据库表建立的关联关系存在，不丢失数据。

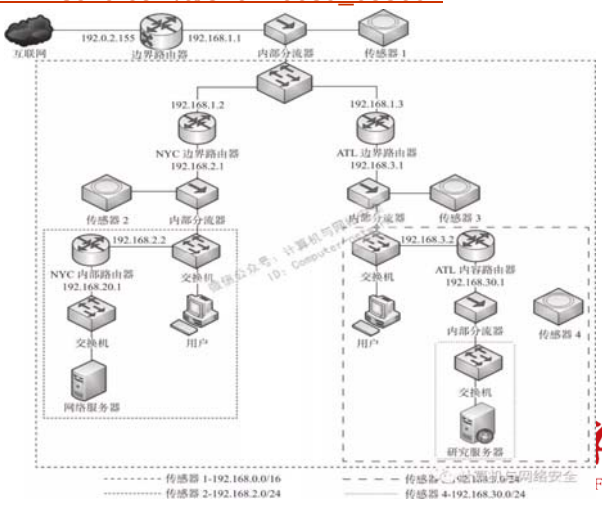
9) 应尽量避免用户自己输入，尽量提供选择，设定字典表。如人员性别设置男、女选择菜单等。

10) 需要设置数据质量校验和监控方法。如人工比对、程序比对、统计分析等。

11) 需要设置数据质量异常上报流程。如监控发现-上报-评估-更正-监控。

数据采集案例

- 网络安全态势感知之数据采集
  https://www.sohu.com/a/329776639_653604



谢谢大家，一起交流学习！

QQ: 10068 0 2383