 DOI: https://doi.org/10.12349/iser.v6i1.4481

Research on High throughput Multi source Heterogeneous Network Security Data Collection Technology for Manufacturing Industry

Jianing Yang Jiawei Wang* Zhanhui Gang

Electronic Technology Information ResearchInstitute of MIIT, Beijing, 100040, China

Abstract

In the context of the swift advancement of information technology,data has emerged as a crucial strategic asset within modern society.In the Internet era,data access has become more diverse,including both traditional structured data,text information,video materials,video clips and other unstructured and semi-structured data.In the field of industrial manufacturing,collecting information from various network security devices is a complex and important task,and the complexity of data poses higher requirements for data collection technology.Based on a series of changes and challenges,this article proposes a high-throughput multi-source heterogeneous data collection technology for the manufacturing industry,which achieves high-throughput network security data collection for the manufacturing industry through ETL technology and data warehouse creation methods.

Keywords

Multi source; heterogeneous data; data collection; data processing

面向制造业的高通量多源异构网络安全数据采集技术研究

杨佳宁 王嘉伟* 刚占慧

国家工业信息安全发展研究中心，中国 · 北京 100040

摘 要

随着信息技术的迅速发展，数据已成为现代社会的重要战略资源。互联网时代的数据获取途径愈发多元，既有传统结构化数据，也有文字信息、影像资料以及视频片段等非结构化与半结构化数据。在工业制造范畴内，各种网络安全设备信息收集是一项较为复杂且重要的工作，数据的繁杂性给数据采集技术提出更高要求。基于一系列变动与挑战，本文提出了一种面向制造业的高通量多源异构数据采集技术，通过ETL技术和数据仓库创建等方法，实现对制造业的高通量网络安全数据采集。

关键词

制造业；多源异构；数据采集；数据处理

1 引言

多源异构数据采集技术是指采集来自不同领域或不同形式的数据，数据通常包括传感器数据、流量数据、设备日志等，具有多样的结构、格式和来源。随着人工智能、物联网（IoT）和大数据分析等技术的进步，数据采集与融合技术已成为现代信息技术的重要研究方向。

多源异构数据采集技术广泛应用于制造、钢铁等制造业领域，且随着企业数字化转型的发展，多源异构数据采集问题逐渐显现。国内许多企业和研究机构在多源异构数据采集和分析上投入大量研究。盛刚，李勇等人[1]通过对多种数据源进行分析和协议的抽象，实现了特钢企业 MES 系统架构下的多源异构数据的统一采集与效率提升。付楠，程光等人[2]设计了容器网络的跨层拓扑发现框架，使用 Barnacle 挖掘服务依赖关系可以获得更完备的网络流量和服务种类；訾立强，王庆伟[3]等人提出一种对工业控制信息设备数据采集技术；孙林[4]提出了一种针对能源互联网流量的数据采集技术；刘三平，龚伟[5]等人提出了一种针对制造行业现场数据采集技术。

目前，面向制造业的高通量多源异构数据采集技术面临两个主要问题：一是数据治理问题，由于数据源不同，难以保证质量和完整性；二是实时性问题，在大规模多源数据的实时采集与处理过程中，计算资源消耗大，如何平衡计算

【基金项目】国防基础科研计划（项目编号：JCKY2022608C001）。

【作者简介】杨佳宁（1990–），男，中国北京人，硕士，高级工程师，从事工控安全研究。

资源与实时性。基于以上挑战，本文提出了一种面向制造业的高通量多源异构数据采集技术。

随着信息技术在工业控制与办公网络中的深度融合，双网协同监测的重要性日益凸显。然而，在此过程中，生产网络监测数据向办公网络的传输面临着安全合规与高效性的双重挑战。为攻克这一难题，本文聚焦多源数据采集技术的探索与实践，旨在构建一个能够实时、精准采集工业控制网、内部办公网以及双网互联区各类海量、异构安全数据，并确保其安全、高效传输与汇聚的技术体系。

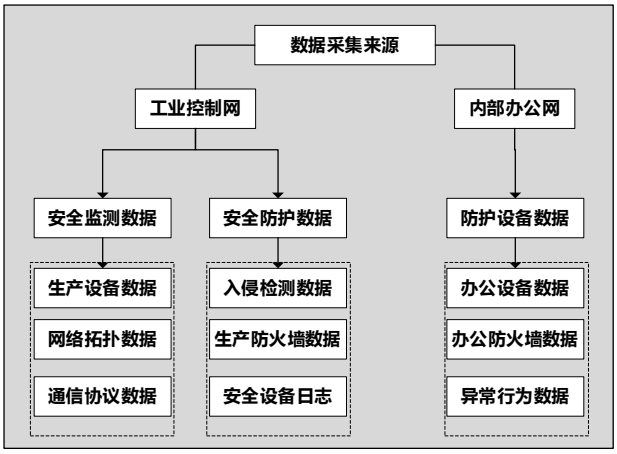

图 1 多源数据采集对象图

2 面向制造业的高通量多源异构数据采集技术研究

制造业高通量多源异构数据采集建设主要涉及业务、技术与管控三个层面。其一，采集工作需以需求为指引，按照业务数据需求明确采集范畴；其二，为防止重复采集情况出现，规划恰当的采集策略十分重要；其三，结合应用场景选取适宜的技术方法，在复杂场景中可综合运用多种技术手段。

2.1 数据采集

数据采集步骤通过数据感知技术进行优化，借助传感器、人工智能算法和大数据技术，实现对环境、物体和人体信息的感知与识别。通过硬件设备如摄像头、麦克风以及语音识别、图像识别等技术，将物理世界的信号转换为数字信息，并进一步提升至可认知的层次，如记忆、理解、规划和决策。

针对生产企业内部办公网络和工业控制网络中的各种类型、格式的数据，多源数据采集负责数据的采集、汇聚；数据的集成、预处理；数据的存储、计算等。

2.2 数据接口

2.2.1 HTTPS 接口

认证方式通过在 JSON 体中增加 token 和 timestamp（unix ms 级），其中 token=md5（from+timestamp+KEY），HTTPS 请求方式为 POST；推送频率：默认 15 分钟；接口描述：定期推送，传输内容为频率时间内产生的数据信息；传输形式：文件以二进制流形式读入，然后用 gzip 压缩，做 base64 操作后进行传输。

2.2.2 SFTP 接口

将待导入数据相关字段信息填写到 txt 文件中，数据中每行为一条数据，数据的所有属性以 JSON 格式进行存储。将填写数据信息的文件，采用 SFTP 方式上报，单次上传文件不超过 1024M，如超过则分多次上传。

2.3 ETL 技术

ETL（即 Extract-Transform-Load）是构建数据仓库的关键环节，其核心任务是从分散的异构数据源里提取数据，将之交付给后续转换流程予以处理，最终存入数据仓库，从而为联机分析处理与数据挖掘提供支撑。

2.3.1 数据抽取

在 ETL 的三个环节——数据抽取、转换和加载中，数据抽取直接从分散、异构的数据源获取数据，主要有全量抽取和增量抽取两种方式。全量抽取是将数据源中的所有数据提取并加载到数据仓库，类似于数据迁移或复制，原数据表不做修改，且转换为 ETL 可识别的格式，流程相对简单。

增量抽取在全量抽取之后开展，重点在于提取上次抽取之后新增或有修改的数据，其应用范围相较全量抽取更为宽泛。于 ETL 进程中，增量抽取的难点在于精准抓取变化数据，所采用的方法需要平衡数据准确性与系统性能，不仅要精准抽取变动数据，又不能给系统造成过重负担。当下，增量数据抽取主要有触发器、时间戳、全表对比以及日志表等途径。一般会在目标表设置三个触发器，分别对应数据的插入、修改与删除操作。并且，还需创建临时表，一旦目标表数据出现变化，触发器便会把变更信息写入临时表，待抽取结束后，对数据进行标记或清除处理。时间戳模式即在源表添加时间戳字段记录变更时间，根据比对情况确定抽取数据。部分数据库可自动更新时间戳字段，而不具备自动更新功能的数据库则需手动更新。全表删除插入法即每次先清空数据后进行数据抽取，一般用于数据量较小的情况。全表比对方式则是先创建临时表，记录主键和 MD5 的校验码，在数据抽取时，先对比两表的 MD5 校验码，若不同则进行更新，若无该主键值，则实施插入操作。日志表方法是在数据库构建业务日志表，当监测到业务数据有变动，业务系统程序就要对日志表进行更新与维护工作，增量抽取时，数据加载及其方式依据日志表信息而定，受业务系统生产数据库方式影响，且日志表的维护由业务系统代码承担。

2.3.2 数据转换

异构数据源之间常常存在不一致问题，数据转换过程旨在解决这个问题。ETL 根据不同的业务规则计算各企业的数据指标，并将其存储到数据仓库中，以便后续分析。

2.3.3 数据预处理

在数据转换流程的定义环节，用户有必要剖析各数据

源的结构特性以及可能出现的脏数据种类，依据剖析所得结果来规划数据转换规则，其中主要涵盖脏数据的检测及处理、数据粒度的转换以及商务计算规则等方面。脏数据的检测与处理规则相对繁杂，包含对不一致数据、不完整数据以及错误数据的处置。本文把脏数据的检测与处理从数据转换流程里独立出来，提前实施单独的处理操作，以此来精简转换流程。相较于数据清洗，数据预处理只是针对单一数据源里的脏数据进行处理，重点关注数据质量与格式问题，并不涉及重复数据的处理工作。

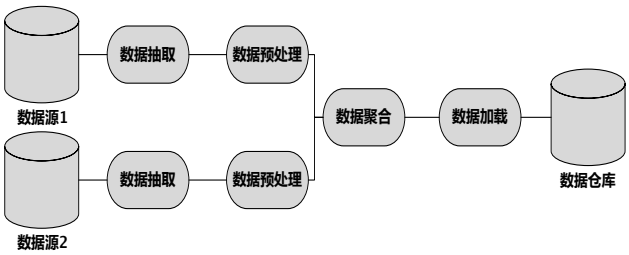

图 2 多源数据预处理流程图

数据预处理按照检测规则检查数据源中抽出的数据，检测出有问题的数据后，就按照对应的处理规则将其修正，然后将没有问题的数据交由后面的转换过程进行处理。对于遇到检测规则中没有定义的脏数据类型，记录到日志等待用户判定处理。

2.4 数据仓库构建

数据仓库作为数据的存储中枢由于其特性不可随意修改，基于制造业数据的特质，资源池构建了工控设备资产数据仓库与安全数据仓库。资产数据仓库重点存储企业工控设备资产相关信息，像工控设备的属性、地理位置以及功能作用等信息；安全数据仓库主要容纳工控安全相关的数据资料，诸如工控安全流量状况、安全日志、威胁情报、漏洞情形、病毒信息以及安全态势等内容。

2.4.1 数据预备域

来自应用系统的操作型数据由数据仓库接收并载入预备域，除了接收数据，预备域还需要进行检查、清洗和过滤工作。在数据仓库中数据处理数据首先进入的是预备域。

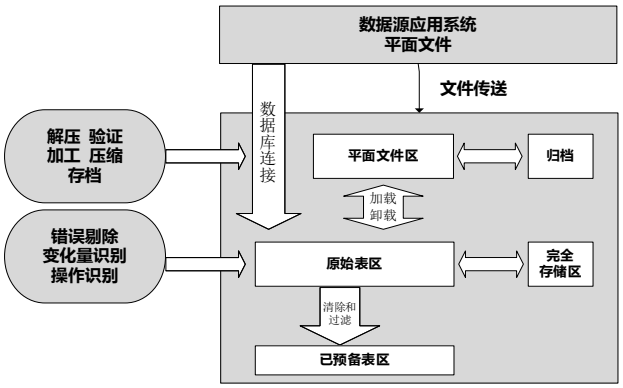

图 3 数据预备域结构图

数据从数据源中被获取，而后会根据不同的方式对获取的数据预处理，从而确保之后能够继续进行数据处理。数据预备域是数据载入数据仓库进行的第一次检验，作为首次进行处理数据的区域，该区域处理数据的质量将会影响数据仓库建设的效果。数据预备域由平面文件区、原始表区、已预备表区构成，平面文件区从数据源获取平面文件，如果出现平面文件不在操作平台上，平面文件区还会对这些文件进行解压、加工、压缩等工作。原始表区使用 DBMS 来管理数据预备域的原始表区而并非平面文件区的操作系统。为了确保后续处理顺利进行，原始表区必须完成大量工作，其中一部分工作非常繁琐，另一些则涉及处理海量数据。预备表区存储所有处理的最终结果，由 DBMS 进行管理。数据预备域是数据源加载到数据仓库的首次检验，该区域的数据质量直接影响数据仓库的建设效果。

2.4.2 数据存储域

为了便于数据分析器进行调查研究，将已被处理和已准备好的分析型数据存储在存储域。预备域的输入来源于各种源应用系统，因此预备域结构的功能以及设计的选择主要取决于源应用系统的特征。

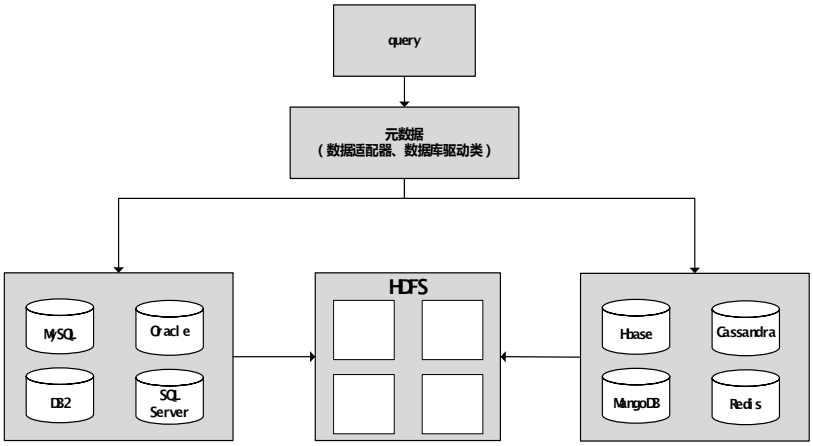

图 4 平台数据存储结构图

2.4.3 Hive 数据仓库

Hive 数据仓库包含数据层和逻辑层。数据层包含 Hadoop 平台及其组件，像 Sqoop、Hive、MapReduce 和 HDFS。原始数据经 ETL 导入后，以文件形式存储在 HDFS 上，为逻辑层提供数据支持。在逻辑层对数据仓库内原始数据进行分析之前，需开展前置处理工作。利用 Python 脚本获取远程映射数据，再依据业务需求展开查询与分析活动。

2.4.4 前置处理

数据源中不存放中文字段值，全部用 uuid 来代替，主要原因有两点，地理区域维度涵盖国家、省市县乃至乡镇村等多个层面，中文字符表述时长度较长，会占据大量存储空间，若采用 32 位 uuid，则仅需占用 4 个字节存储空间。第二点，不同根站点的子站点可能会有重名，为了解决此类二义性问题，使用 uuid 可以更方便地比对某几个站点是否属于同一站点。在预处理阶段，可以通过 Python 脚本，使用 webservice 的方式请求拉取这些键值对至本地目录，并缓存为文本文件，当正式分析过程中需要查找中文时，只要从缓存文件中读取即可。

2.4.5 查询分析处理

Hive 数据仓库查询分析过程涉及 HiveQL 输入、HiveQL 转换成 MapReduce 作业、Hive 查询结果反馈三个部分，HiveQL 输入在开发阶段有程序员编码实现，测试时可以手动输入，但该设计 HiveQL 不由用户直接手动输入，而是集成到自动化处理脚本中。同时，采用 SHELL 命令调用 HiveShell 的过程中，使用 Python 脚本进行数据预处理和根据业务需求进行数据查询分析处理，利用 Hive 的 transform 关键字，可以将 HiveQL 的初步查询结果作为 Python 脚本的输入，待 Python 脚本对这些数据逐行分析结束后，输出的结果集可以作为最终结果，也可以作为 Hive 查询分析语句的输入，此时 Hive 可以在 Python 输出结果集上调用聚合函数，常用的有 SUM、COUNT、AVG 等。HiveQL 转换成 MR 作业在 HiveShell 中输入 HiveQL 查询语句后，Hive 会自动将该查询语句转换成 MapReduce 作业交给 Hadoop 去处理，而对于复杂的数据查询分析处理，Hive 中没有现成的函数可供调用，此时需要用户自己实现 Hive 使用的 MapReduce 脚本，本项目中使用的是 Python 去实现。

Hive 查询结果反馈会将结果显示到控制台上。在实际处理过程中，这部分显示在控制台上的数据需要被自动化脚本追加到相应文件并最终导出至数据库，此功能仍然可以通过 SHELL 命令去完成。

2.4.6 后置处理

后置处理在数据查询分析处理结束后运行，同样，借助 Python 脚本处理程序执行日志。可以理解为一个脚本监控机制，对整个系统中脚本执行情况作出及时反馈，使管理员可以迅速发现存在的任何隐患。另外，当需要删除 Hive 分区数据时，也可以在后置处理中添加相应代码。

3 结论

基于制造业多源数据采集的难点，本文提出了一种面向制造业的高通量多源异构数据采集技术，该技术首先使用 ETL 技术对数据进行抽取、转换及预处理，之后通过数据仓库的构建来进一步对数据进行采集处理，将采集到的数据传入数据库并进行接入解析，为工业生产过程提供指导，使用该技术可以更直接反映制造业的高通量数据信息，通过对其进行处理得到更便于分析的结构化数据，有效高通量数据的杂、乱和脏等问题，为制造业企业的健康发展添砖加瓦。

参考文献

[1] 盛刚,李勇,张效华.某特钢企业MES系统多源异构数据采集架构设计与应用[J].仪器仪表用户, 2022, 29(12):16-19.

[2] 付楠,程光,滕跃,等.基于非侵入式数据采集的微服务依赖关系发现方法[J].网络空间安全科学学报,2023,1(02):112-121.

[3] 訾立强,王庆伟,王锐.工控系统信息设备数据采集建设[J].工业信息安全,2022,(02):51-59.

[4] 孙林.面向区域能源互联网的数据采集技术分析[J].集成电路应用,2024,41(04):160-161.

[5] 傅芳.网络视听节目监管系统中的数据采集关键技术[J].电视技术,2024,48(06):211-213.