

基于大数据的智能工厂数据平台架构设计与研究

常镜洳

(大连东软信息学院软件工程系, 辽宁 大连 116023)

摘 要: 在分析智能工厂国内外研究现状基础上, 对基于大数据的智能工厂数据平台架构技术展开研究, 为智能生产的运行分析、预测、决策调控以及数字孪生信息物理融合提供技术参考。探讨了智能工厂定义与内涵, 以及智能工厂大数据来源和特征, 采用Hadoop、Spark、Storm热门开源大数据计算引擎, 提出了数据来源层、数据传输层、数据存储层、资源管理层、处理分析层以及业务应用层构成的智能工厂大数据平台技术架构, 有效解决智能工厂大数据多源复杂性和实时性的要求和难点。所提数据平台技术架构将对智能制造和智能工厂的实现具有重要的借鉴价值。

关键词: 大数据; 智能工厂; 数字孪生

中图分类号: TP311.5 **文献标识码:** A

Design and Research of Data Platform Architecture for Intelligent Factories Based on Big Data

CHANG Jingru

(Department of Software Engineering, Dalian Neusoft University of Information, Dalian 116023, China)

Abstract: Based on the analysis of the existing research status of intelligent factories at home and abroad, this paper studies the data platform architecture technology for intelligent factories based on big data, providing technical reference for the operation analysis, prediction, decision control and the fusion of information and physics of digital twins of intelligent production. The paper not only discusses the definition and connotation of intelligent factory, intelligent factory big data source and characteristics, but also puts forward by using Hadoop+Spark+Storm the big data platform technology architecture of intelligent factory which includes the data source layer, the data transfer layer, the data storage layer, the resource management layer, the processing and analysis layer, and the business application layer. The technology architecture of the data platform will have important reference value for the realization of intelligent manufacture and intelligent factories.

Keywords: big data; intelligent factory; Digital Twin

1 引言(Introduction)

随着物联网、云计算、大数据等新兴技术快速发展, 制造业掀起了信息化、自动化、智能化为发展方向的新一轮革命浪潮, 近年来引起了世界各国的极大关注。

以美国和德国为首的欧美发达国家在新形势下制定了一系列制造业规划。美国自2009年陆续推出一系列工业战略规划和工业互联网概念, 以通用(GE)公司为代表, 并于2012年发布了《工业互联网: 打破智慧与机器》产业政策报告; 2013年汉诺威工业博览会上德国政府提出为了“工业4.0”战略落地, 西门子公司已将工业4.0概念引入其生产控制系统和工业软件开发中。

针对以上挑战, 中国政府于2015年提出“中国制造2025规划”, 促进以云计算、大数据等新一代信息化技术与现代制造业的深度融合与创新。学术研究方面, 吉旭等提出高分子材料行业云计算与大数据云制造架构和关键技术^[1]; 张洁等提出大数据驱动的“关联+预测+调控”的智能车间分析、决策新模式^[2]; 王建民分析工业大数据来源并归纳大数据管理分

析关键技术^[3]; 陶飞等提出基于大数据传输与连接的数字孪生五维模型及其10大应用领域^[4]。

基于以上背景, 本文针对智能工厂数据平台问题展开研究, 深入探讨基于大数据的智能工厂数据平台技术框架体系, 以求从海量数据中挖掘有价值的信息来指导智能工厂的运行和优化。

2 智能工厂(Intelligent factory)

2.1 智能工厂的定义

相对于传统制造业而言, 智能制造是其转型和升级, 即在制造过程中进行分析、推理、判断和决策等智能化活动, 以人为核心地位的同时实现人机一体化; 实现智能制造的关键是智能工厂的建立和实施。智能工厂的定义最早在德国“工业4.0”中提出, 即利用物联网、云计算、大数据等新兴技术, 通过传感器、网络等将机器、设备、人员和软件程序连接起来并逐渐融合, 以高效监控、采集、处理和分析数据, 实现加工过程、信息管理以及服务的智能化, 构建高效、节能、绿色的人性化工厂。

2.2 智能工厂的实施

实现智能工厂的第一要点是使用物联网将“黑暗数据”即未使用的数据关联起来,并将其转化为有用信息系统,从而快速响应消费者需求变化和市场的突变,实现敏捷式需求驱动型制造模式。

智能工厂实施的第一步是将车间的系统连接起来;由于在物联网技术出现之前,企业的多数现有设备已部署完毕,且无内置连接功能,目前缺乏通用的通信标准支持传统设备与物联网进行互操作,但是在物联网开放平台上将传统设备与中间件连接的生态系统方案已成功上市,且OPAF、IIC等组织和英特尔等IT巨头厂商正在加紧制定开放的物联网标准。

为了应用高级分析发掘数据的巨大价值,智能工厂实施的第二步为采集、存储、预处理、分析数据;多数现有设备可生成海量数据,但由于数量过多,这些数据无法发回数据中心进行快速分析;为了提取最重要的信息,在终端分析或雾计算策略中,在流数据采集点、网关、云端或它们之间的任何位置部署算法来实施高性能计算,从而评估需要的数据和不需要的数据。

3 大数据(Big data)

3.1 大数据概念

在舍恩伯格编写的《大数据时代》中,大数据指不用随机分析法(抽样调查)这样捷径,而采用所有数据进行分析处理;具有大量化、高速化、多样化、价值化等多重属性。

大数据在不同层面需要不同技术;数据采集层需要利用ETL工具将分布、异构数据源中数据抽取到临时中间层,完成清洗、转换等数据预处理工作后加载入数据仓库或集市;数据存储层需利用分布式文件系统、数据仓库、关系/非关系数据库等,实现各种数据的存储、管理;数据分析和处理层需结合机器学习和挖掘算法,利用计算框架引擎实现对大规模数据的分析和处理,并可视化呈现分析结果。其两大核心关键技术包括分布式存储和分布式处理,从而容忍时间内有效地处理大规模数据。

3.2 智能工厂的大数据来源

智能工厂的大数据来源主要包括:企业信息化数据、工业物联网数据、“跨界”数据^[5],具体如图1所示。

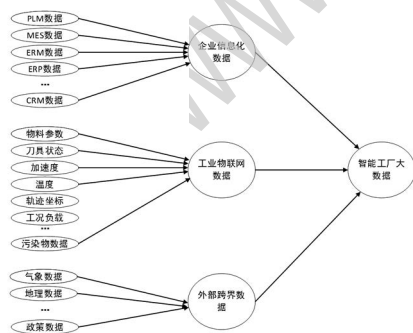


图1 智能工厂大数据源

Fig.1 Intelligent factory big data source

作为传统的工业数据资产库,企业信息化数据主要来自PLM(产品全生命周期管理系统)、MES(制造执行系统)、ERP(企业资源管理系统)、PQM(产品质量管理系统)、CRM(客户关系管理系统)等,这部分数据从原材料入库、加工、检验到出厂流通,贯穿产品全生命周期和整个价值链,往往具有高价值密度。

随着数控机床、工业机器人、传感器、RFID等智能设

备和感知网络的广泛使用,工业物联网数据不仅包括物料参数、刀具状态、工况负载等生产实时数据,还包括污染物、有害气体等作业环境实时数据,实现所有生产数据和过程数据的统一管理。

随着互联网与工业制造的不断融合,除了来自制造业的内部大数据之外,还包括影响企业生产和运营的外部“跨界”数据,比如气象数据、地理信息数据、政策数据、经济数据、法律法规数据等。

3.3 智能工厂的大数据特征

智能工厂数据多样化、多元化、异构化,且处理场景复杂多变,因此其数据呈现典型的大数据3V特征:大量化、多样性、高速性;具体体现如下:(1)大量化,以乳制品质检为例,依据质检信息管理系统中1400种质检方法的电子记录和计算,一包牛奶从原料入库到消费者手中,经过35个工序和105个检测环节,每天有超过1GB的数据如潮水般涌入质检系统中流转;(2)多样性,以乳制品生产为例,既存在灌装机、杀菌机等设备的运行时间、转速等运行参数和乳品液体流量、压力、温度等结构化数据,也包括像乳制品特殊材料结构表、可编程逻辑控制器(PLC)控制程序等半结构化数据,还有像码垛机器人的三维立体模型、乳品质检图表、物流运输小车监控视频等非结构化数据;(3)高速性,智能工厂中PLC、传感器等设备在极短时间窗口内对生产过程进行不间断采样,产生的数据流按时间序列如潮水般涌入数据库中,以乳制品灌装生产线为例,PLC按照1S的采样间隔不断产生灌装盒定位、液体流量和罐装容量等监控数据。

4 智能工厂数据平台技术架构(Technology architecture of intelligent factory data platform)

大数据处理包括复杂批数据处理、基于历史数据的交互查询和挖掘、实时数据流处理、图结构数据处理,智能工厂实际应用同时存在以上几种场景,因此采用当前热门Hadoop、Spark、Storm开源计算引擎,也是Apache软件基金会最重要三大分布式框架。

基于以上内容,形成如图2所示(Ambari和Zookeeper管理涵盖的范围)的智能工厂数据平台架构。自底向上分为数据来源层、数据传输层、数据存储层、资源管理层、处理分析层和业务应用层,数据从下到上流动;下面就每层分别展开详细介绍。

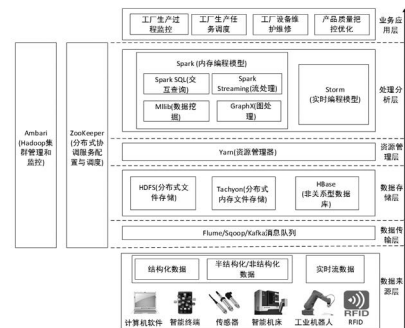


图2 智能工厂数据平台技术架构

Fig.2 Technology architecture of the intelligent factory data platform

4.1 数据来源层

数据来源层即数据提供者,主要负责实时可靠的产生、采集、获取智能工厂范围内的多源异构数据;由于工业生产处于高速运转,工业设备产生数据类型多是非结构化数

据,且对数据的实时性要求也更高。该层面面向工厂物理制造资源,主要包括智能机床、工业机器人、计算机软件、智能终端等,通过在制造资源上安装和配置工业传感器、RFID标签、二维码、条形码采集数据,并通过有线互联网、无线互联网、GSM网、红外、蓝牙等基础网络设施连接这些生产制造资源,按照物联网协议进行数据传输和交换。

4.2 数据传输层

数据传输层位于数据来源层与数据存储层之间,是智能工厂各种数据源的数据进入大数据系统的第一步,负责将外部数据源的数据导入像HDFS、HBase的持久层中;常用的技术包括:Flume、Sqoop、Kafka等。

Flume和Sqoop主要负责各种静态数据的采集与传输,允许用户将数据从关系型数据库抽取到Hadoop中,一旦生成最终的分析结果,Sqoop便可以将这些结果导回数据存储层。例如数控机床加工车间管理系统负责生产制造单元一个阶段的动作规划并分配资源,通常是一至几个小时,从而处理批量生产;其信息数据包括制造单元的位置、其间路由、批量制造所需的工具、材料清单、零件库存量以及单元操作状态等。

由于制造实时数据流传输需要很强的可用性的输入管道,优先使用Kafka负责实时流数据的传输和导入。例如智能机床内置的传感器传感监测刀具的运动速度、加速度、轨迹坐标、温度等实时数据的传输;智能车间的网络监控实时数据,快速持续到达,必须采用实时采集和计算,且响应时间为秒级甚至毫秒级。

4.3 数据存储层

数据存储层负责存储海量的非结构化和半结构化的松散数据,对大型数据实现随机和实时的读写访问。该层包括Hadoop的核心组件分布式文件存储系统HDFS,以及Spark生态系统的分布式内存文件存储Tachyon、Hadoop的实时查询框架HBase。

HDFS作为Hadoop生态圈的基础,适合运行于廉价计算集群上,以一次写入、多次读取的流数据形式存储超大文件。

Tachyon是一个高性能、高容错、基于内存的开源分布式存储系统,能够为集群框架Spark提供内存级速度的跨集群文件共享服务;被部署在计算平台Spark之下和文件存储系统HDFS之上。

HBase是一个高可靠、高性能、面向列的NoSQL分布式数据库;利用HDFS作为其文件存储系统,利用Zookeeper作为协同服务。由于智能工厂的数据复杂性、来源的多样性,在数据库表的设计上,不同的数据源产生的数据,存放在不同的数据表中;同时为了提高数据访问性能并实现毫秒级响应,根据业务需求和存储要求设计每张表行键。

4.4 资源管理层

为了实现计算框架统一部署和运行,它们都部署在Hadoop 2.0的资源管理框架YARN上;YARN作为一种通用资源管理系统,为数据处理层提供资源的统一管理和调度;从而实现计算资源的按需弹性伸缩(支持一万个计算节点和二十万个内核集群)、不用负载应用混搭、通过共享底层存储而避免数据跨集群迁移等。

4.5 处理分析层

处理分析层负责大数据的处理工作,利用各种计算框架编写代码模型实现智能工厂大数据预处理、数据关联分析、数据挖掘等,从而揭示和实现制造数据时序演化规律、制造

设备性能预测、任务决策调控等;分布式并行计算框架包括内存计算框架Spark、流计算框架Storm。

Spark负责智能工厂中企业信息化数据离线批量处理工作,任务时间跨度一般为月/天/时级;Spark SQ负责智能工厂中企业信息化数据和互联网“跨界”数据中基于历史数据的交互式查询,组件GraphX负责图结构数据的处理工作;MLib负责数据挖掘且提供各种算法模型,例如分类中逻辑回归、贝叶斯等算法,聚类中K-Means、模糊K均值等算法,机器学习中神经网络、最小二乘法、深度学习等算法,任务时间跨度一般为分钟/秒级;Spark Streaming库负责秒级数据流的处理与分析。

Storm是一个免费、开源的分布式实时计算系统,可以简单、高效、可靠地处理流数据;由于其毫秒级实时响应速度,Storm主要负责智能工厂中工业物联网流数据的并行计算工作,例如机床刀具运行状态的实时分析、车间管理系统的实时报警等任务。

4.6 业务应用层

大数据已经应用于人类社会的各行各业,拥有大数据不是目的,应用高级分析发掘数据的巨大价值才是关键。由企业信息化数据、工业物联网数据和跨界数据汇聚成的工业大数据蕴含着许多工业生产规律,犹如一块“金矿”蕴藏着无限财富价值;工业大数据在工业生产过程监控、生产任务调度、工业生产设备故障诊断及预测、网络协同制造等方面具有广泛应用。

例如作为“工业之母”的高端数控机床,通过传感器感知的温度、运动坐标、速度、加速度、声发射信号等实时数据可动态控制和优化下一步切/削/铣进刀参数、预测刀具破损和磨损状态、分析能耗情况等。

5 结论(Conclusion)

本文在分析智能工厂大数据来源、特征,以及各种复杂处理场景基础上,基于目前流行的Hadoop、Spark、Storm三大分布式开源计算框架设计和研究了智能工厂大数据平台技术架构,并提出了数据来源层、数据传输层、数据存储层、资源管理层、处理分析层以及业务应用层。后续将立足于提高企业生产率、提高产品质量和设备可靠性,以智能化程度较高高端数控机床加工工厂为背景,逐步搭建、实现以及优化以上大数据技术平台,从而高效采集、处理和分析数控机床的主轴震动、温度、噪声、转速、进给量、信号量^[5]等,以诊断故障发生部位、判断故障性质和类型、分析故障原因,并预测故障风险概率和不良产品概率,为数控机床的可靠性和设备维修提供重要借鉴价值。

参考文献(References)

- [1] 吉旭,钟淦基,于洋,等.高分子材料行业云制造的关键技术及应用[J].计算机集成制造系统,2015,21(11):3072-3078.
- [2] 张洁,高亮,秦威,等.大数据驱动的智能车间运行分析与决策方法体系[J].计算机集成制造系统,2016,22(5):1220-1228.
- [3] 王建民.工业大数据技术[J].电信网技术,2016,8(8):1-5.
- [4] 陶飞,刘蔚燃,刘检华,等.数字孪生及其应用探索[J].计算机集成制造系统,2018,24(1):1-17.
- [5] 郭安,于东,胡毅.信息物理融合技术在机床故障诊断系统的应用研究[J].小型微型计算机系统,2017,4(4):896-900.

作者简介:

常镜如(1983-),女,博士,讲师.研究领域:智能工厂生产调度,大数据。