

工业大数据技术综述

王建民

清华大学软件学院, 北京 100084

摘要

工业大数据是工业数据的总称, 包括信息化数据、物联网数据以及跨界数据, 是工业互联网的核心要素。分析了工业大数据的发展背景和主体来源, 剖析了工业大数据、企业信息化和工业互联网之间的相互关系, 阐述了工业大数据“多模态、高通量、强关联”的数据特点以及“跨尺度、产业链、跨界”多源数据融合的应用特点, 探讨了工业大数据软件系统架构, 给出了制造环节的大数据应用实例。

关键词

工业大数据 ; 工业互联网 ; 大数据软件 ; 系统架构

中图分类号 : TP399

文献标识码 : A

doi: 10.11959/j.issn.2096-0271.2017057

Survey on industrial big data

WANG Jianmin

School of Software, Tsinghua University, Beijing 100084, China

Abstract

Industrial big data is a collection of data in the industry application, which consist of business data in the information system, machine data in the industrial IoT system, and related data in the related websites. Industrial big data is a very important component of industrial internet. The main resources of the industrial big data were analyzed with introduction of the industrial internet. Then the relationships among “industrial big data, enterprise digitalization and industrial internet” were discussed. The “multi-typed, high-throughput, and strong-correlated” characteristics of industrial big data were presented. A typical industrial big data application scenario with “cross-time-scale, cross-supply-chain, and cross-organization” was shown. Furthermore, Industrial big data software architecture was discussed and a real-life application case was shown.

Key words

industrial big data, industrial internet, big data software, system architecture

①

[https://www.
predix.io/](https://www.predix.io/)

②

[https://www.
siemens.com/
global/en/
home/products/
software/
mindsphere.html](https://www.siemens.com/global/en/home/products/software/mindsphere.html)

③

[http://new.abb.
com/abb-ability](http://new.abb.com/abb-ability)

④

[https://www.
schneider-
electric.com/](https://www.schneider-electric.com/)

⑤

[http://www.
rootcloud.com/](http://www.rootcloud.com/)

⑥

[http://www.
hannovermesse.
de/product/
cosmoplatform/
2399901/R567621](http://www.hannovermesse.de/product/cosmoplatform/2399901/R567621)

⑦

[http://www.
casicloud.com/](http://www.casicloud.com/)

1 工业大数据的发展背景

当前,以大数据、云计算、移动物联网等为代表的新一轮科技革命席卷全球,正在构筑信息互通、资源共享、能力协同、开放合作的制造业新体系,极大扩展了制造业创新与发展空间。新一代信息通信技术的发展驱动制造业迈向转型升级的新阶段——工业大数据驱动的新阶段,这是在新技术条件下制造业生产全流程、全产业链、产品全生命周期数据可获取、可分析、可执行的必然结果。

工业互联网正处在高速发展时期,国际工业和软件龙头企业正在迅速布局工业互联网技术与产品。2015年8月,美国通用电气公司率先宣布推出首个面向工业数据和分析的云平台Predix^①,如图1所示。2016年2月,Predix正式对开发者开放;在2016年4月的汉诺威工业博览会上,西门子公司宣布将推出云平台MindSphere^②,到了2017年4月的汉诺威工业博览会,MindSphere成为西门子公司展区最核心的展出内容,如图2所示。

在2017年汉诺威工业博览会上,总部

位于瑞士的ABB集团宣布推出产品ABB Ability^③,它集成了ABB集团从设备到边缘计算再到云服务的跨行业、一体化的数字化能力。总部位于法国的施耐德电气有限公司也推出了EcoStruxure平台^④,它综合了施耐德电气有限公司的互联互通产品、边缘控制以及应用、分析和服务。

国内三一重工股份有限公司、海尔集团、中国航天科工集团公司等企业也开始打造自己的平台,分别推出了三一树根物联平台^⑤、COSMO平台^⑥、航天云网平台^⑦等工业互联网平台产品,如图3所示。

工业大数据是工业互联网的核心要素。《中国制造2025》规划中明确指出,工业大数据是我国制造业转型升级的重要战略资源,需要针对我国工业自己的特点有效利用工业大数据推动工业升级。一方面,我国是世界工厂,实体制造比重大,但技术含量低、劳动密集、高资源消耗制造的比重也大,实体工厂和实体制造升级迫在眉睫;另一方面,我国互联网产业发展具有领先优势,过去十多年消费互联网的高速发展使互联网技术得到长足发展,互联网思维深入人心,需要充分发挥这一优势,并将其与制造业紧密结合,促进制造业升级和生产性服务业的发展。

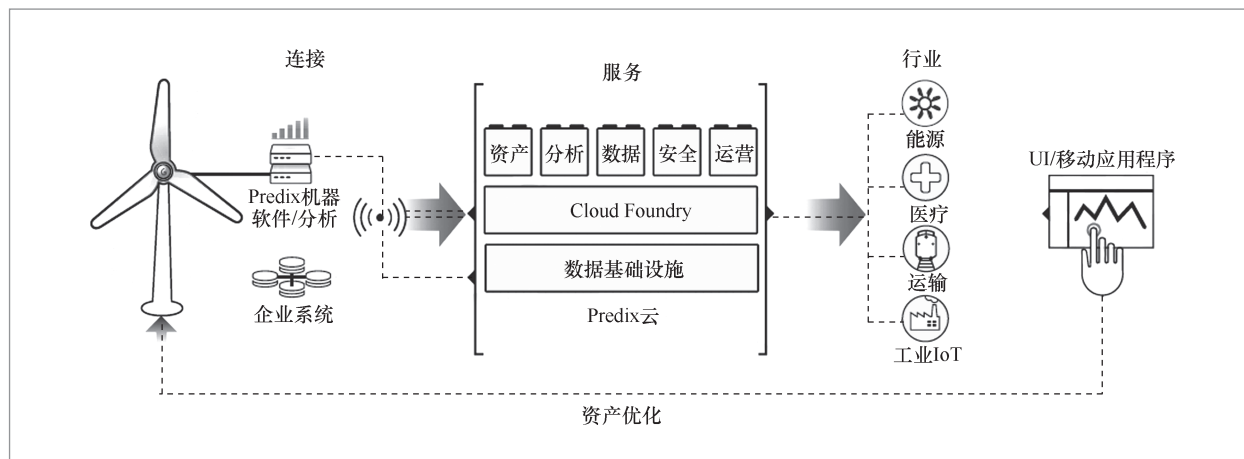


图1 通用电气 Predix 平台 MRO 5.0

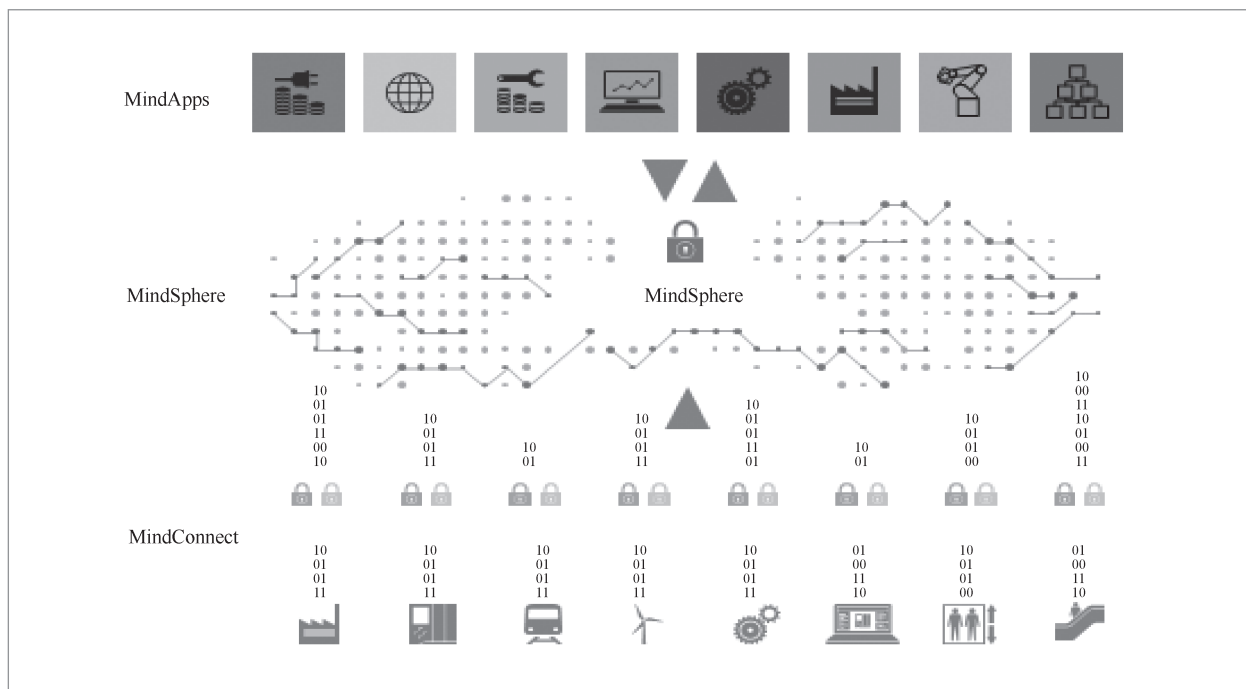


图2 西门子 MindSphere 平台

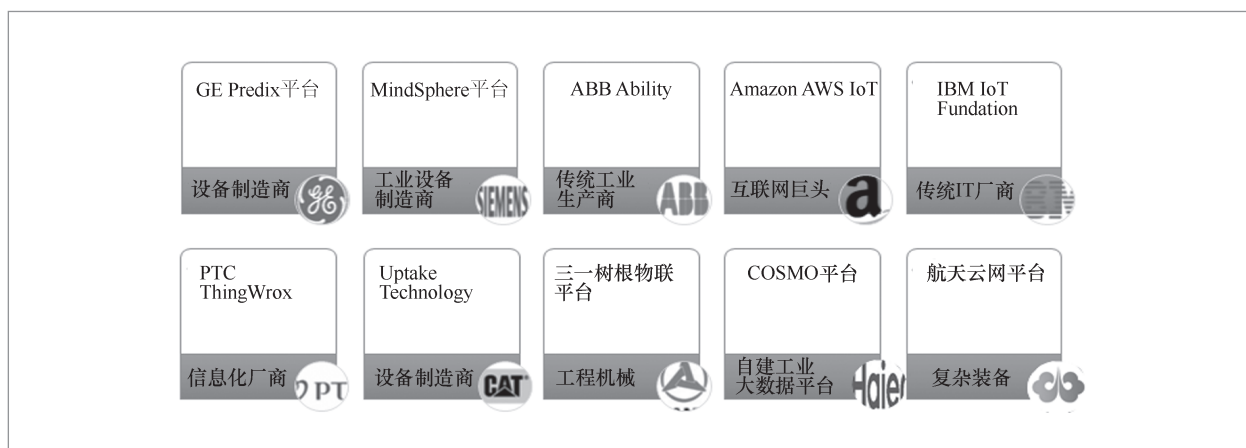


图3 工业互联网技术与产品举例

因此,我国在推进工业大数据的应用过程中,要突出智能互联产品创新^[1],同时兼顾智能制造和制造服务,用数据驱动制造全生命周期从设计、制造到交付、服务、回收各个环节的智能化升级,推动制造全产业链智能协同,优化生产要素配置和资源利用,消除低效中间环节,整体提升中国制造业发展水平和世界竞争力。

2 工业大数据的内涵

如图4所示,工业大数据即工业数据的总和,其来源主要包括企业信息化数据、工业物联网数据、“跨界”数据^{⑧[12,31]}。

企业信息系统存储了高价值密度的核

⑧

<http://www.aai-alliance.org/>

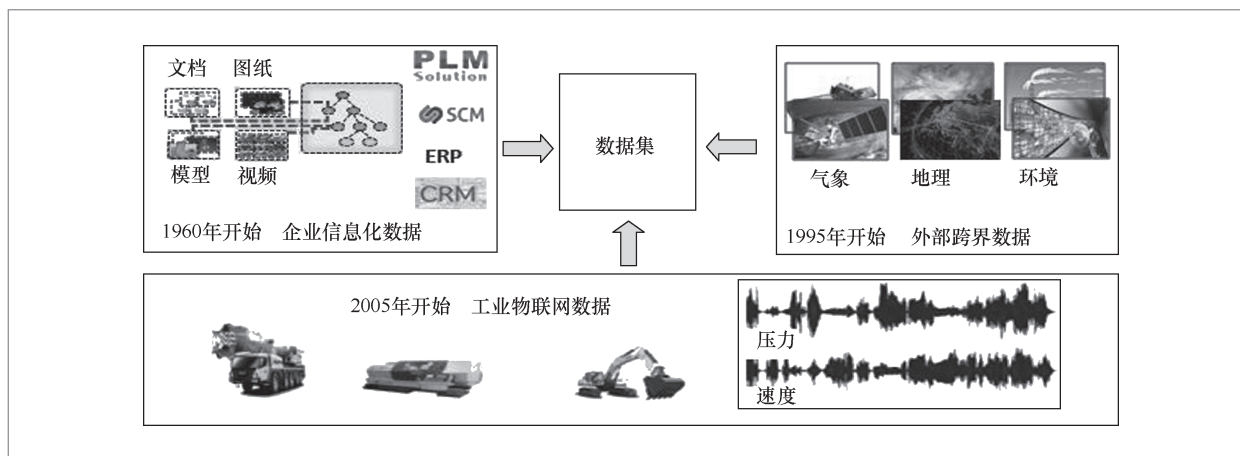


图4 工业大数据的来源

心业务数据,积累的产品研发数据、生产制造数据、供应链数据以及客户服务数据存在于企业或产业链内部,是工业领域传统数据资产。近年来,物联网技术快速发展,工业物联网成为工业大数据新的、增长最快的来源之一,它能实时自动采集设备和装备运行状态数据,并对它们实施远程实时监控。互联网也促进了工业与经济社会各个领域的深度融合,人们开始关注气候变化、生态约束、政治事件、自然灾害、市场变化等因素对企业经营产生的影响,因此外部跨界数据已成为工业大数据不可忽视的来源。

人和机器是产生工业大数据的主体。人产生的数据是指由人输入计算机中的数据,例如设计数据、业务数据等;机器数据是指由传感器、仪器仪表和智能终端等采集的数据。近年来,由人产生的数据规模的比重正逐步降低,企业信息化和工业物联网中机器产生的海量时序数据是工业数据规模变大的主要来源,机器数据所占的比重将越来越大。

3 工业大数据的特征

工业大数据作为对工业相关要素的数

字化描述和在赛博空间的映像,首先符合大数据的4V特征^[4],即大规模(volume)、速度快(velocity)、类型杂(variety)、低质量(veracity)。相对于其他类型大数据,工业大数据还具有反映工业逻辑的多模态、强关联、高通量等新特征。

多模态是指工业大数据必须反映工业系统的系统化特征及其各方面要素,包括工业领域中“光、机、电、液、气”等多学科、多专业信息化软件产生的不同种类的非结构化数据。比如三维产品模型文件不仅包含几何造型信息,还包含尺寸、公差、定位、物性等其他信息;同时,飞机、风机、机车等复杂产品的数据又涉及机械、电磁、流体、声学、热学等多学科、多专业。

强关联反映的是工业的系统性及其复杂动态关系,不是数据字段的关联,本质是指物理对象之间和过程的语义关联。包括产品部件之间的关联关系,生产过程的数据关联,产品生命周期设计、制造、服务等不同环节数据之间的关联以及在产品生命周期的统一阶段涉及的不同学科不同专业的数据关联。

高通量即工业传感器要求瞬时写入超大规模数据。嵌入了传感器的智能互联产品已成为工业互联网时代的重要标志,用

机器产生的数据代替人产生的数据,实现实时的感知。从工业大数据的组成体量上来看,物联网数据已成为工业大数据的主体。以风机装备为例,根据IEC61400-25标准,持续运转风机的故障状态,其数据采集频率为50 Hz,单台风机每秒产生225 KB传感器数据,按2万台风机计算,如果全量采集,则写入速率为4.5 GB/s。总体而言,机器设备产生的时序数据的特点包括海量的设备与测点、数据采集频度高(产生速度快)、数据总吞吐量大、7×24 h持续不断,呈现出“高通量”的特征。

4 工业大数据与企业信息化

企业信息化数据是工业领域传统数据资产,也是工业大数据的第一个来源。在传统制造业企业信息化系统中,广义产品生命周期管理(product lifecycle

management, PLM)系统主要支持产品开发,企业资源计划(enterprise resource planning, ERP)系统负责“人财物、产供销”,供应链管理(supply chain management, SCM)系统协调供应链,客户关系管理(customer relationship management, CRM)系统服务于企业客户和用户,制造企业生产过程执行系统(manufacturing execution system, MES)负责制造执行,维护、维修和运营(maintenance, repair and operations, MRO)系统支持售后服务^⑨等。这些企业信息化系统积累的产品研发数据、生产制造数据、物流供应数据以及客户服务数据多以结构化数据的形态存储在企业数据库系统中,计算机辅助设计(computer aided design, CAD)、仿真等研发软件产生的非结构化工程数据则通过文件方式管理。

在企业信息化领域,信息系统是数据和流程的综合体,二者联系紧密。图5是波

^⑨ <http://mro.thss.tsinghua.edu.cn/>

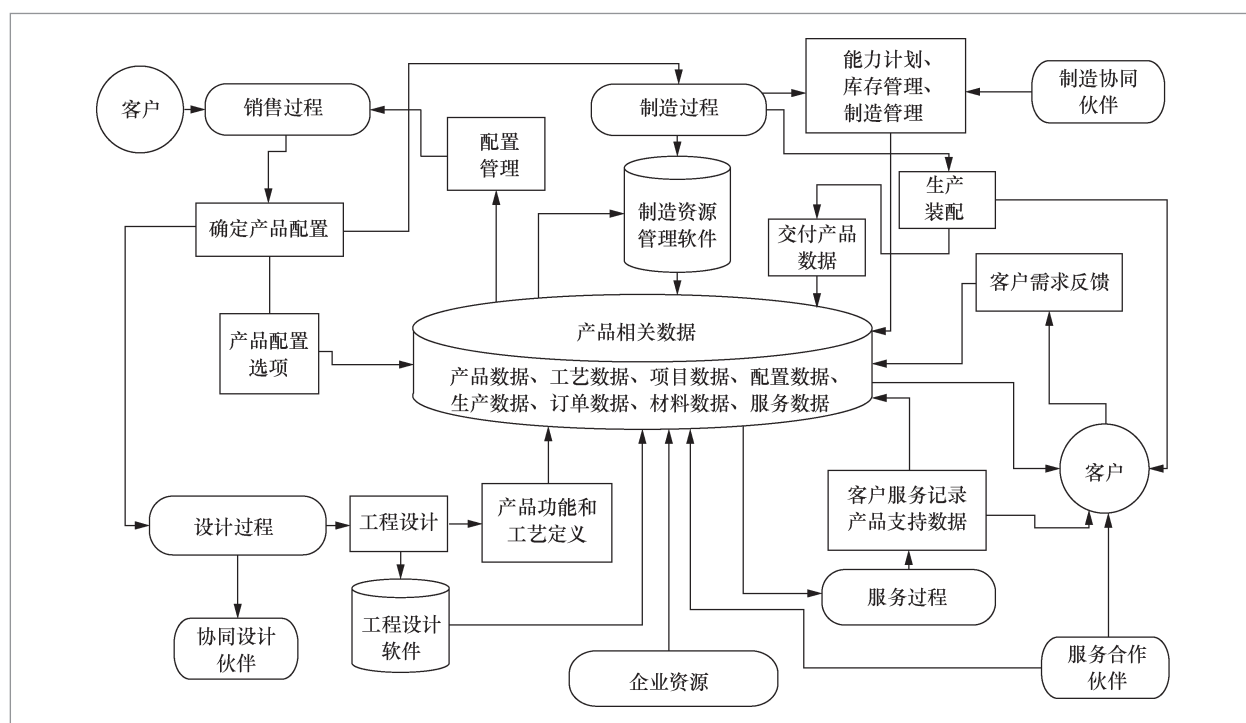


图5 波音公司“以过程为中心”与“以数据为中心”信息化对比分析

音公司20多年前“以数据为中心”的信息化架构。20世纪90年代,随着空中客车公司的崛起,波音公司危机四伏,订单下降,美誉度也随之下降。危难之际,波音公司当时新上任的董事长武达德特别注重波音公司信息系统建设,成功解决了“信息孤岛”问题,极大地改善了波音公司业绩。具体方法就是以数据为中心,把产品数据、工艺数据、项目数据等有机融合在一起,构成逻辑上的单一产品数据源,然后在数据集上重构业务流程。从客服到客户,提供端到端的服务模式,让业务流程围着数据转。由此可见,流程是产生数据的来源,也是消费数据的归宿,数据是支持流程无缝、高效运行的原材料,二者不可分离。

从数据流动的视角来看,企业信息化解决了工业领域“有数据”的问题,网络化解决了“能流动”的问题,工业大数据要解决数据“智能化”的问题。“信息化”能够把正确的数据在正确的时间以正确的方式传递给正确的人和机器,“智能化”则把海量的工业数据转化为信息,信息转化为知识,知识转化为科学决策,以应对和解决制造过程的复杂性和不确定性等问题,其目标在于不断提高制造资源的配置效率。

企业信息化主要解决的是数据单元传递问题,工业大数据则主要是基于数据集分析问题。如图6所示,这是一个制造业复杂装备阶段活动示意^[5],业务活动沿实线部分从上游往下游传递,它主要反映了订单、票据等数据是否正确,这是信息化过程中需要解决的核心问题。虚线主要是反馈部分,通过分析数据集发现业务规律和决策准则,然后反馈给前面的各个环节使用,从而形成数据全生命周期的闭环,这就是信息化和大数据智能化的区别,然而两者又是不可分割的。

5 工业大数据与工业互联网

工业互联网可以从网络、数据和安全3个方面理解。其中,网络是基础,即通过工业全系统的互联互通,促进工业数据的无缝集成;数据是核心,即通过工业数据全周期的应用,实现机器弹性生产、运营管理优化、生产协同组织与商业模式创新,推动工业智能化发展;安全是保障,即通过构建涵盖工业全系统的安全防护体系,保障工业智能化的实现。工业互联网的发展体现了多个产业生态系统的融合,是构建工业生态系统、实现工业智能化发展的必由之路。

工业大数据是智能制造与工业互联网的核心,其本质是通过促进数据的自动流动解决控制和业务问题,减少决策过程带来的不确定性,并尽量克服人工决策的缺点。随着互联网与工业的深度融合,机器数据的传输方式由局域网络走向广域网络,从管理企业内部的机器拓展到管理企业外部的机器,支撑人类和机器边界的重构、企业和社会边界的重构,释放工业互联网的价值。

6 工业大数据典型应用场景

工业大数据的应用主要是实现制造业企业生命周期的智能化水平提升,以智能化生产为核心,涵盖了从设计研发、生产制造、经营管理到售后服务的整个流程,实现提质增效。随着工业互联网的发展,企业的数据从内部数据实现了跨界,应用也随之拓展到“互联网+用户”“互联网+产业链”“互联网+服务”等场景。工业互联网产业联盟发布的《工业互联网体系架

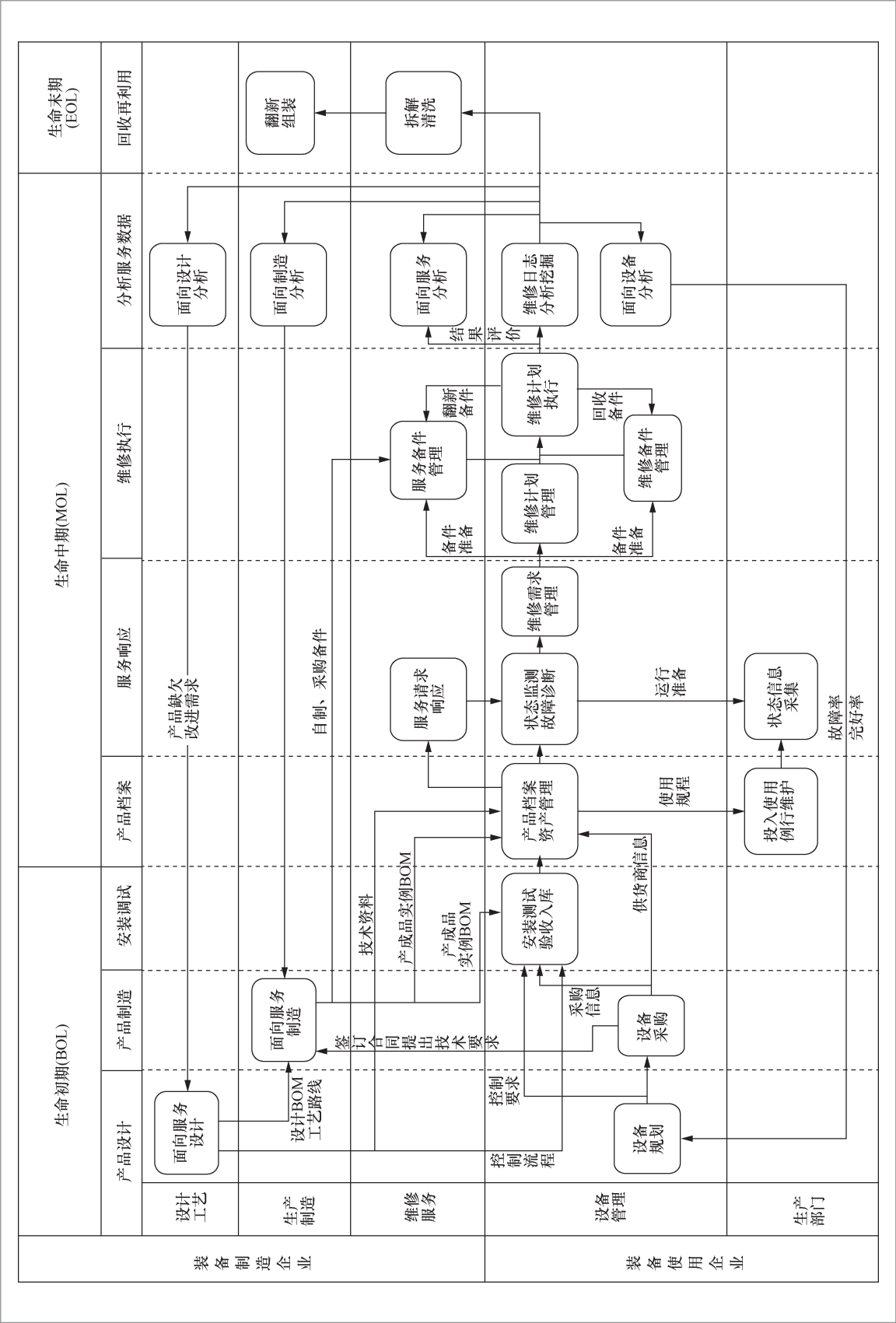


图 6 信息化与大数据：数据单元传递与数据集合分析

构》对此进行了总结,将其分成了4类典型场景:智能化生产、个性化定制、网络化协同、服务化延伸。

以采矿为例,图7是未来智慧矿山的场景。传统矿山经营过程中有很多装备,一般是按照数月前拟定的计划生产,不能很好地满足市场的实际需求,经常会存在过度生产或者生产不足的问题^[6]。未来可以通过工业大数据直接感知市场需求,通过市场分析可以知道哪一种铁矿石配比在当前市场上适销,据此确定各种铁矿石的生产需求,并制定生产计划,然后实时将操作命令下达到相应的智能化工程装备,指挥这些工程装备协同工作,这就是跨尺度的信息集成和优化。也就是说,把当天的市场需求通过大数据挖掘出来,直接传达到设备上,变成设备的行动和操作。还可以通过设备网络化,实时获取装备工况数据。当工程装备出现问题或异常时,及时地发现问题,找到问题的原因。另外,还可

以通过大数据构建基于规则或案例的故障预测系统,对装备状态进行预测,更好地保障设备健康。

由此可见,工业大数据是提升工业生产效率,降低能耗,转变高耗能、低效率、劳动密集的粗放型生产面貌的必要手段。结合数控机床、工业机器人等自动生产设备的使用,并建立从经营到生产系统贯通融合的数据流,做到数据全打通和数据流通不落地,可以提升企业整体生产效率,降低劳动力投入,有效管理并优化各种资源的流转与消耗。大数据也是实现工业企业从制造向服务转型的关键支撑技术。工业领域智能服务的本质就是智能产品加上感知控制能力和大数据分析,通过对产品使用过程中的自身工作状况、周边环境、用户操作行为等数据进行采集和分析,可以提供在线健康检测、故障诊断预警等服务以及支持在线租用、按使用付费等新的服务模型。



图7 未来智慧矿山的场景

7 工业大数据软件架构

图8是基于数据库视角的大数据系统软件技术架构,涉及物联网、数据库、数据处理、流处理、数据分析、批处理和机器学习等技术。关系数据库存放的是结构化的关系数据,是企业信息化支撑技术,数据库是先有模式后有数据,即先定义数据库表,然后才能插入数据;与数据库不同,互联网、物联网等产生的大多是非结构化数据,这些数据往往是设备产生的,先有数据,后有模式。

在工业的ERP、PLM、SCM、CRM等系统中,20%的“SQL小数据”具有80%的价值密度,而物联网或互联网中,80%的“非结构化大数据”密度只有20%的价值密度,甚至更低。因此,工业数据中20%

的工业“小数据”需要“引爆”80%的工业“大数据”的价值。工业大数据应用本质上就是要综合利用结构化数据和非结构化数据,从中提炼出隐含的模式,形成知识,从而使工业制造和服务更加智能化。

工业大数据更关注数据源的“完整性”,而不仅仅是数据的规模,因此对数据管理有较高要求。工业大数据应用需要实现数据在物理信息、产业链、跨界3个层次的融合。由于“信息孤岛”的存在,这些数据源通常是离散和非同步的,这与其他领域大数据集成具有明显差异,因此不仅需要从数据模型,更需要从制造过程、层次化物料表(bill of material, BOM)结构、运行环境等多类型工业语义层面对工业大数据进行一体化整合管理,其中BOM是产品全生命周期数据集成的关键手段。

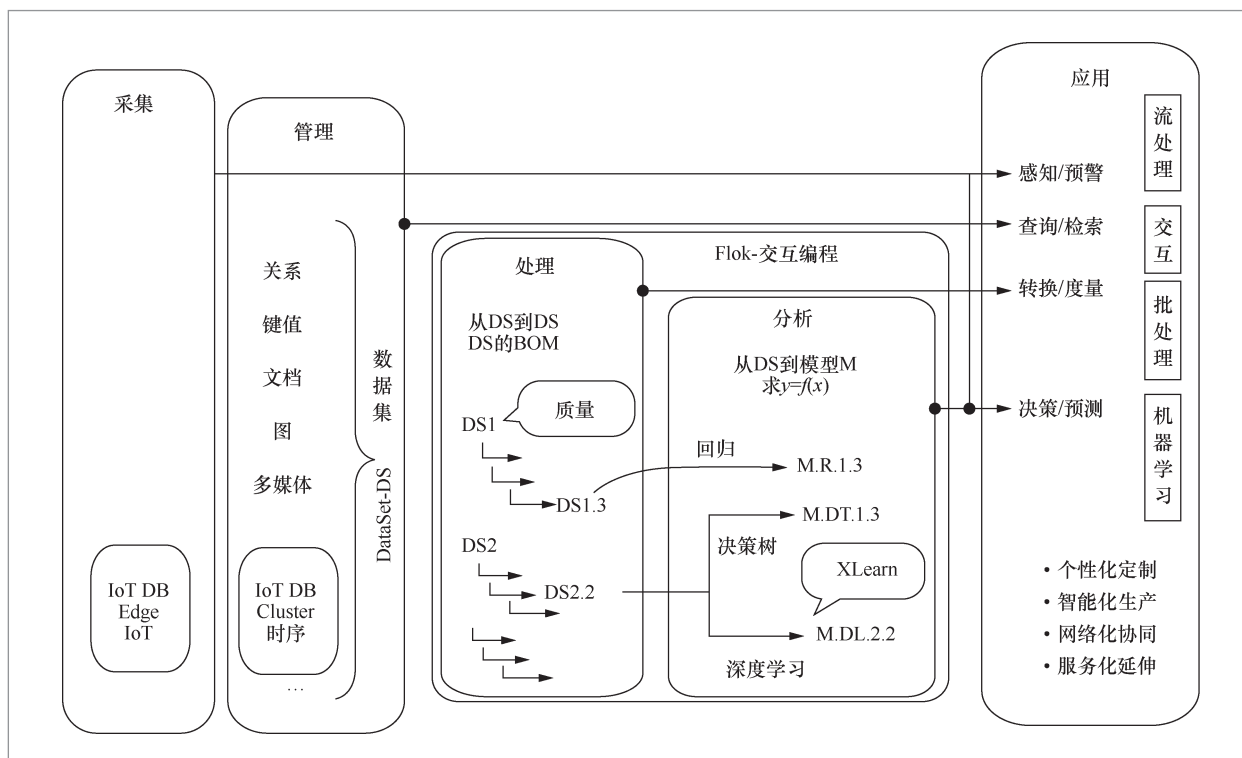


图8 基于数据库视角的大数据系统软件技术架构

工业大数据生命周期包括采集、管理、处理、分析和应用5个环节。工业大数据处理是一个数据集到另一个数据集的“物理”过程。工业大数据分析就是要利用机器学习、深度网络等人工智能算法,从训练数据集里学习未知的模型,这个模型就是一个基于数据集 (X,Y) ,求 $Y=F(X)$ 计算式中 F 函数的过程。简而言之,工业大数据分析是从训练数据集到模型的“化学”过程。

工业大数据应用在工业互联网中的应用可分为4个层次:监视、控制、优化、自主。监视就是要能远程实时监测装备的运行状态;控制就是要实现操作者对机器的远程遥控,让机器能够执行操作者远程下达的操作指令;优化就是要基于海量工业大数据发现知识,提供在线运行调度、健康检测、故障诊断预警等装备在线运维服务;自主就是要实现装备的自主决策和装备集群的自主协同,通过“机器换人”实现生产运维的少人化和无人化。

8 工业大数据典型应用举例

图9是一个服务器主板生产流程。上料之后,第一个业务环节是锡膏印刷,然后是贴装电子元件,之后是回流焊,看起来是一个比较简短的流程,但生产过程中会出现器件偏移、立碑、覆盖件漏印、爬锡短路等故障。电子器件因此将会出现质量问题,会给生产厂商造成严重的经济和声誉损失。

为了解决这个问题,在生产线流程中增加了两个自动化质量检测环节,如图10所示。

相比于图9,图10增加了锡膏印刷检测(solder paste inspection, SPI)和自动光学检测(automatic optic inspection, AOI)环节。增加这两个检测环节后,需增加人工复判环节,而由于人工复判存在不稳定的问题,质量检测成功率并未达到预期要求,企业为此付出了较大的人力成本,甚至经济损失。

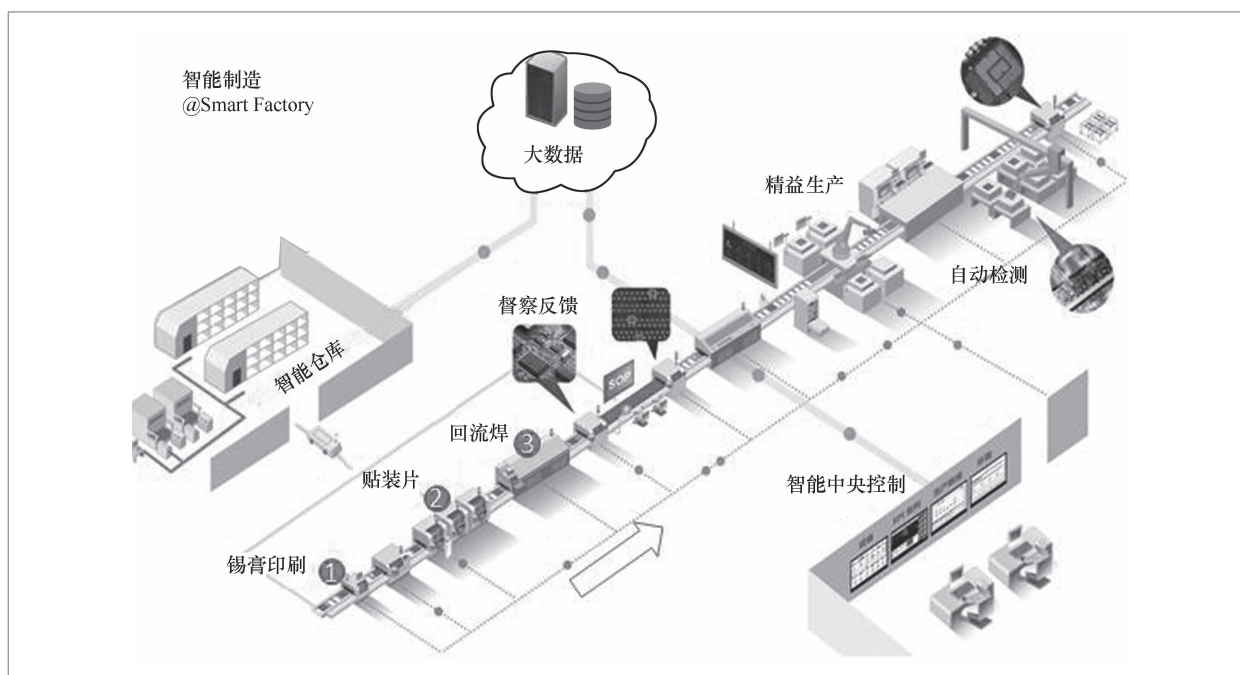


图9 电子元器件表面贴装生产流程

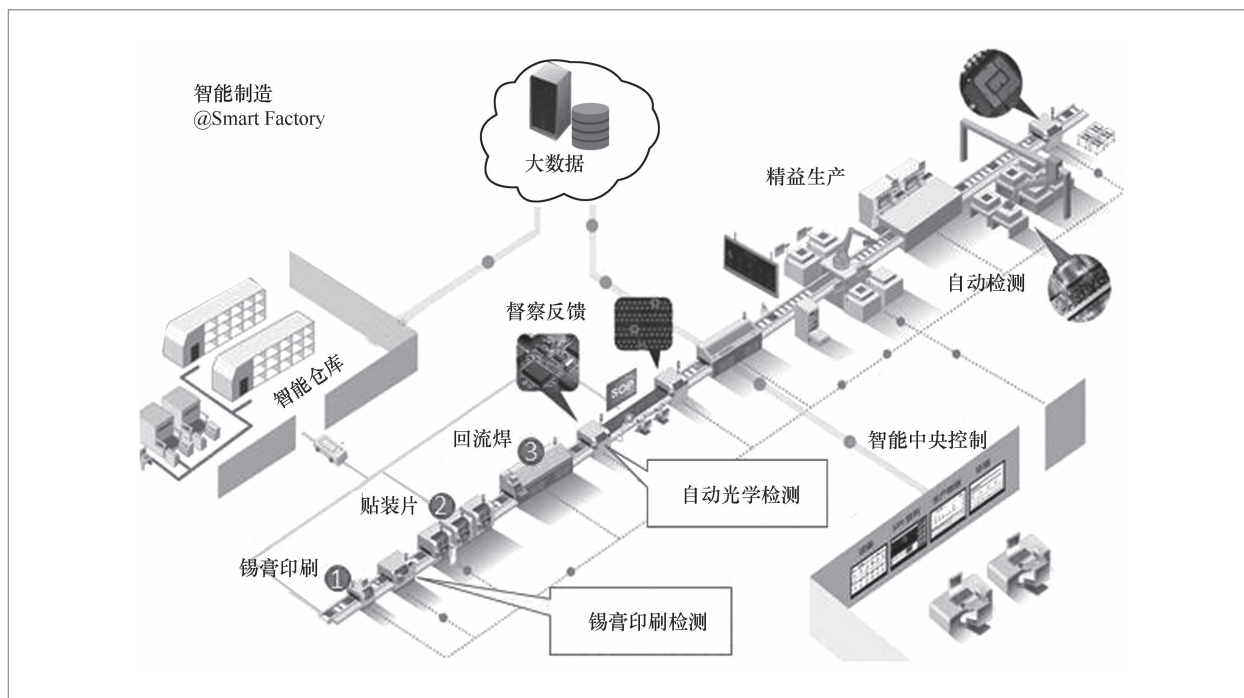


图 10 加入检测环节后的电子器件表面贴装生产流程

基于积累的主板焊点SPI质量检测数据和主板AOI质量数据，采用基于机器学习的智能复判算法，判定焊点质量合格率。由于这些不同类型故障的数据是相互交织在一起的，需要用机器学习的办法进行复杂的数据集分类，然后采用随机森林的办法形成了一个基于人工智能的电子器件表面贴装“故障复判机器人”。该算法只需要把5个参数输入，就可以判断服务器主板每个焊点是否合格，极大提升了服务器主板制造的品质稳定性与合格率^⑩。

9 结束语

工业大数据创造价值的过程才刚刚开始，工业大数据的价值不仅在于对现有业务的优化，更在于支撑企业、行业乃至全社会的创新、转型和发展。伴随着工业互联网的不断普及，工业大数据应用价

值将逐步释放，智能制造也将得到长足发展。当前，工业大数据仍处在高速发展的历史阶段，其概念内涵、技术方法、价值创造模式还在不断创新演化之中，需要努力和大胆地创新实践。应该针对智能制造和工业互联网的创新需求，把工业大数据的应用与工业自身提高质量、降低成本、提升管理水平的需求结合起来，特别要与中国工业的发展阶段和结构特点结合起来，走出有中国特色的工业大数据技术与产业创新路线，助力中国工业弯道取直与加速腾飞。

致谢

感谢清华大学软件学院、大数据系统软件国家工程实验室、北京工业大数据创新中心和工业大数据系统与应用北京市重点实验室的同事们。

^⑩ <http://video.e-works.net.cn/Video/VideoDetail.aspx?vid=3277&sid=268>

参考文献:

- [1] PORTER M E, HEPPELMANN J E. How smart connected products are transforming competition[J]. Harvard Business Review, 2014(11): 65-88.
- [2] 工业互联网产业联盟工业大数据特设组. 工业大数据技术与应用实践(2017)[M]. 北京: 电子工业出版社, 2017.
- [3] 王建民. 工业大数据技术[J]. 电信网技术, 2016(8): 1-5.
- [4] WANG J M. Technology of industrial big data[J]. Telecommunications Network Technology, 2016(8): 1-5.
- [5] MANYIKA J, CHUI M, BROWN B, et al. Big data: the next frontier for innovation, competition, and productivity[M]. [S.l.:s.n.], 2011.
- [6] 莫欣农. 跨生命周期数据管理平台助力中国制造服务转型[J]. 科技纵横, 2015(8): 58-60.
- [7] MO X N. Cross lifecycle data management platform promotes transformation of manufacturing services in China[J]. IEEE Spectrum, 2015(8): 58-60.
- [8] DAYAL U, GUPTA C, VENNELAKANTI R, et al. An approach to benchmarking industrial big data applications[C]// WBDB 2014, August 5-6, 2014, Potsdam, Germany. Berlin: Springer International Publishing, 2014: 45-60.

作者简介



王建民 (1968-), 男, 博士, 清华大学软件学院教授、博士生导师、院长, 信息学院副院长, 数据科学研究院副院长、管理委员会副主任, 大数据系统软件国家工程实验室执行主任, 工业大数据系统与应用北京市重点实验室主任。主要研究方向为大数据系统软件、工业大数据、产品全生命周期管理、业务过程管理等。

收稿日期: 2017-10-18

基金项目: “十三五”国家重点研发计划基金资助项目 (No.2016YFB1000700)

Foundation Item: The National Key Research and Development Program of China (No.2016YFB1000700)