白皮书

制造数据分析指南

为什么需要制造数据平台

作者:凯睿德制造首席执行官
Francisco Almada Lobo

Critical
manufacturing
an ASM PT company

目录

简述

数据就是新的石油

Clive Humby，英国数学家、乐购会员卡架构师，被广泛认为于2006年首次提出了这个说法："数据就是新的石油。极具价值的数据，如果未经处理，就无法被有效利用。石油只有在转换为天然气、塑料、化学品等形式后，才能成为一种有价值的材料，产生商业利润；同样，我们必须对数据进行解构、分析，才能使其产生价值。" [1]

毋庸置疑，我们正在见证一个技术革命的进程。我们可以称之为工业4.0、物联网（IoT）、大数据或人工智能（AI），推动工业4.0的解决方案正在不断迭代出新。

无论我们从何种角度审视这一快速的技术演进，都可以发现一种不变的核心要素：数据。

然而，不同于以往以制造业为中心的工业革命，在这场以数据转换为中心的基础性技术变革中，制造业的步伐明显滞后。相比其它行业，对于这些技术价值的认识和应用，制造业人士一直持保守态度，并迟迟不肯付诸于行动。现在，他们已经开始追赶，但速度仍然很慢。

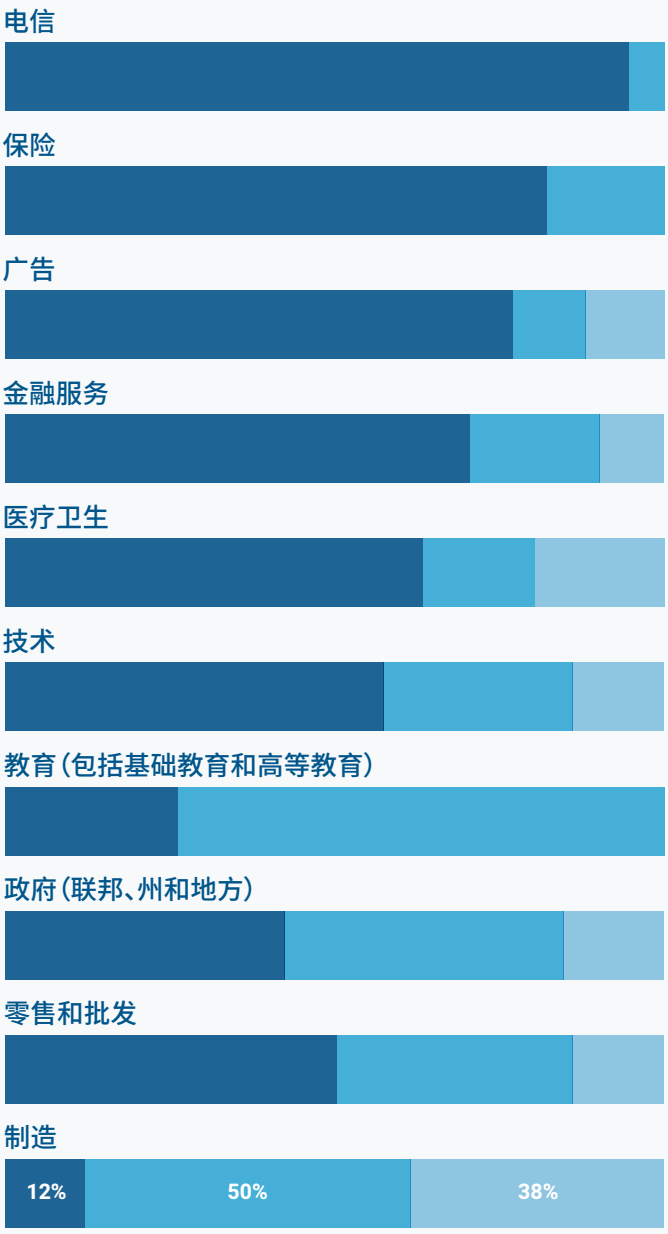
举例而言，1999年，Kevin Ashton在宝洁公司的一次演讲中，首次提出了"物联网"的概念；然而，20多年后，制造业仍在学习物联网的含义，制定应用物联网的相关战略。

本白皮书旨在逐步解析"数据平台"这一全新的软件系统，如何作为一种解决方案帮助制造商获取并利用数据，释放巨大的生产潜力。

1. https://medium.com/project-2030/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294

不同行业中的大数据使用情况

- 是的，我们正在使用大数据
- 我们以后可能会使用大数据
- 不，我们不打算使用大数据



第1章

数据革命因何而起?

三大基础技术

1. **电子产品的微型化和可获取性**
   1965年,戈登·摩尔提出摩尔定律:单位空间内可容纳的晶体管数量每隔两年便会增加一倍。虽然定律中的周期逐渐缩短,可容纳的晶体管数量现在已趋近饱和,但在过去几十年中,这一假设均得到了验证。

2. **硬件和软件实现重大突破**
   数据处理和存储性能以及经济性显著提升,如云技术。虽然貌似新技术,但CompuServe早在1983年就提供了少量存储文件的磁盘空间。

3. **人工智能兴起,特别是机器学习(ML)**
   这并不是新方法,深度学习早在1943年便问世,当时沃尔特·皮茨(Walter Pitts)和沃伦·麦卡洛克(Warren McCulloch)就建立了一个基于大脑神经网络的计算机模型。

不过,值得一提的是,大数据和分析领域中使用的工具,在很大程度上是受到需要处理社交网络产生的大量数据的需求的驱动。事实上,早期的数据技术演进正是起源于社交媒体。当时,随着非结构化数据阵列的日益增多,现有的数据存储和分析功能无法处理或捕获大量数据。技术处理和实时分析的需求推动了从批处理到流式处理的转变,此后,大数据技术便进入了全新的发展阶段。

物联网数据平台的诞生

单一解决方案能否解决所有问题?

数据生成的速度越来越快,数据量也越来越大,这就催生了新的需求和挑战;另一方面,新一代科学家们希望利用这些数据,使用诸如图形可视化、人工智能等先进分析解决方案。在这些因素的共同驱动下,物联网数据平台产生了。

在很多数据处理方式上,各类数据平台都大同小异,但人们应当根据不同领域的具体制造要求进行重新思考。虽然可以将这些平台与已有的跟踪和控制软件解决方案同时部署,但这种方式下,平台产生的整体价值仅是冰山一角。反之,如果将两者相融合,充分利用彼此的能力,那么这种模式产生的价值将是巨大的。

在下一章中,我们将探讨更广泛的数据分析主题,如数据平台存在的原因,以及如何在制造业中利用数据平台来发挥现有先进解决方案的优势,如现代制造执行系统(MES)。

第2章

暗数据

任何技术革命或趋势在造福人类的同时，也会不可避免地产生副作用。过去几年中，市场上涌现了大量的数据存储和处理工具、平台和解决方案，帮助公司提取信息进行分析、预测或提供有价值的建议。

但问题是，大多数公司（制造商也不例外）产生的数据是无效的。在生产中，机器、材料、运输系统、仓库和员工会以日志或数据库形式记录大量数据，但大部分可能永远都不会被用到或看到！

这种数据没有直接价值，即使在重组后，也不能转化为有用信息，因此，我们称之为暗数据[2]。一般而言，大多数公司都有海量暗数据：65%的暗数据隐藏于机器、网络和员工方面[3]。

美国卡耐基梅隆大学亨氏信息系统和公共政策学院教授Rahul Telang表示，暗数据约占公司总存储数据的90%[4]（不是生成的数据，而是真正存储的数据）。

如此体量的数据不可能被视而不见。虽然硬件和存储系统的价格持续走低，但保存这些数据的成本依然非常高昂。为什么呢?因为在大多数情况下，数据不是简单地以原始格式存储，而是需要经过精心编排和组织。所以，为什么公司要保留这些数据呢?

虽然暗数据可能永远都用不到或产生价值，但出于以下两个原因，公司一般会保留它们：

1. **降低风险和责任**
   主要出于合规目的，诉讼需要也是因素之一。特别是有严格监管要求的行业（如食品饮料、医疗保健或半导体），合规性问题可能会影响业务的正常开展。将纸质信息数字化，即使会产生大量暗数据，也有助于优化工作效率。

2. **备不时之需**
   虽然数据无法立即产生价值，但未来可能会提供有价值的信息。事实上，许多制造商逐渐意识到，生产过程中产生的数据不仅在现阶段有助于制定决策，而且未来还能用于训练机器学习算法，通过生成的预测模型预测行为和性能，进而增强竞争优势。

保存暗数据时，无需对数据执行任何转换、编排或整理操作。它以“保留模式”存储，供将来使用。暗数据对存储性能或分析没有明确要求，只要存储可靠并且在需要时可以调用，就可以按较为简易、经济的格式保存。

因此，能以原始格式存储事件的解决方案很受欢迎。这种数据存储方式不仅可靠，而且成本较低（如有需要，可以永久存储），便于用户随时提取数据。我们稍后会看到，Kafka就是这样的一种解决方案。

理解暗数据时，重要的一点是，必须意识到它并不“暗”。暗数据生成分析建议后，就不再“暗”了，可以发挥出自身价值。多数情况下，制造商只是没有意识到暗数据的存在。因此，我们应当让暗数据重回视野，让大家认识到暗数据蕴藏的价值。此后，才有必要配置多分析暗数据的平台。

2. https://www.gartner.com/en/information-technology/glossary/dark-data
3. https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/datumize-dark-data-in-manufacturing-and-logistics-solution-brief.pdf
4. https://www.forbes.com/sites/tomtaulli/2019/10/27/what-you-need-to-know-about-dark-data/

第3章

边缘解决方案

对于数据平台来说，在数据到达平台之前，挑战便已到来。在纯粹的集中式解决方案中，数据必须从设备发送到服务器上，然后处理后的输出再被传送回设备中。

边缘解决方案能在数据生成位置附近完成部分处理工作，从而创建出“混合”解决方案。

边缘计算非常重要，原因有很多：

1. 设备和中央处理器之间的网络可能会产生高延迟，并产生不良后果，特别是在需要快速响应或决策的情况下。

2. 如果数据生成的速率很高，在数据生成位置附近处理数据，可以降低在网络和处理层中获取和分析大量数据的成本。

3. 在本地保留一定级别的数据有助于规避合规和安全漏洞，特别是在数据敏感的情况下。

4. 调试也更方便，甚至还能减轻设备故障引发的后果。

5. 在许多物联网部署中，网络连接并不可靠。边缘解决方案可以在恢复连接之前暂时保存数据，有效解决网络中断的问题。

就制造业而言，特别是在离散领域，虽然有将简单的物联网设备添加到传统设备上，以收集一些基本信息的案例，但大部分数据集都是通过本地现有的过程控制解决方案发送到数据平台，例如可编程逻辑控制器（PLC）。

因此，边缘解决方案必须能与控制解决方案的所有接口进行通信。虽然一般来说，开放平台通信（OPC）[5]是工业自动化领域安全可靠数据交换的互操作性标准，但它并非适用于所有应用。部分制造领域特有的接口数量非常多，例如半导体行业的SECS/GEM，或用于电子组装的IPC CFX。有鉴于此，我们又添加了其他物联网特定的协议，如MQTT或BLE，以及其他较旧且功能较差的接口，如TCP/IP、数据库或基于文件的接口。边缘解决方案能处理各种复杂接口，并将这些细节与更高的处理层隔离开来。

在制造业中，软件也会控制制造过程，而不是简单地收集数据进行分析，这是制造业与其他行业的另一个明显不同点。虽然有多种控制回路，但更多基于软件的控制回路会与在更高级别运行的其他软件解决方案（如MES）进行交互。这种现象在部分行业中被称为软件联锁，需要具备一定级别的处理和工作流功能，同时也必须在边缘运行。

最后，考虑到解决方案需要与其他系统交互，工厂中使用的边缘解决方案数量也在不断增加，用户必须要能对此类解决方案进行集中管理，包括监测、调试、模拟和自动部署。

鉴于上述原因，现有软件应用程序（有时在业界称为“设备集成”应用程序）将继续使用，但必须加以改进，以满足生成分析建议的特定数据驱动需求。

5. https://opcfoundation.org/about/what-is-opc/

第4章

流分析的发展

不久前，生产系统处理的大多是事务性工作。MRP或MES系统根据存储在关系数据库中的数据执行事务处理。

随着数据库发展日趋完善，许多公司意识到数据分析有助于提高工厂绩效，数据的时效性越强，对他们的决策就越有价值。不过，这些分析系统也会增加作为业务主干的事务系统数据库的负载。[6]

由于市场对数据分析的需求逐步增长，数据库负载不断增加，工业企业纷纷开始改变这种模式，并于2000年初[7]创建了数据仓库。用户可利用ETL工具（提取、转换和加载）从事务数据库中批量加载数据。除提取外，目标是对不同系统的数据进行标准化处理，使其转变为仓库内的单一表示和单一结构。

在加载数据之前，系统会先过滤和清理数据，生成存储在在线分析处理（OLAP）技术中的“信息”。经过预计算和聚合处理后的数据能加快分析过程。OLAP数据库被划分成一个或多个多维数据集（如时间和日期），以便于数据分析。

随着数据源的规模和种类持续增长，整个流程会变得更加复杂。由于业务数据更加规范，因此数据仓库实现项目的周期会变得非常长，而且成本也将更加高昂。

同时，纳入非结构化数据的需求也会增加。照片或多媒体文件、数据地图、测试结果均没有预定义的数据模型，传统数据库系统很难处理此类数据。

此外，在企业对数据仓库的依赖性逐渐提高后，它们面临的挑战也越来越大。由于基础性技术是在单机中扩展，随着业务规模的扩大，单机变得越来越大，成本也变得越来越高，减少数据存储量的需求迫在眉睫。

6. https://imply.io/post/how-we-got-to-streaming-analytics
7. https://www.dataversity.net/brief-history-data-warehouse/

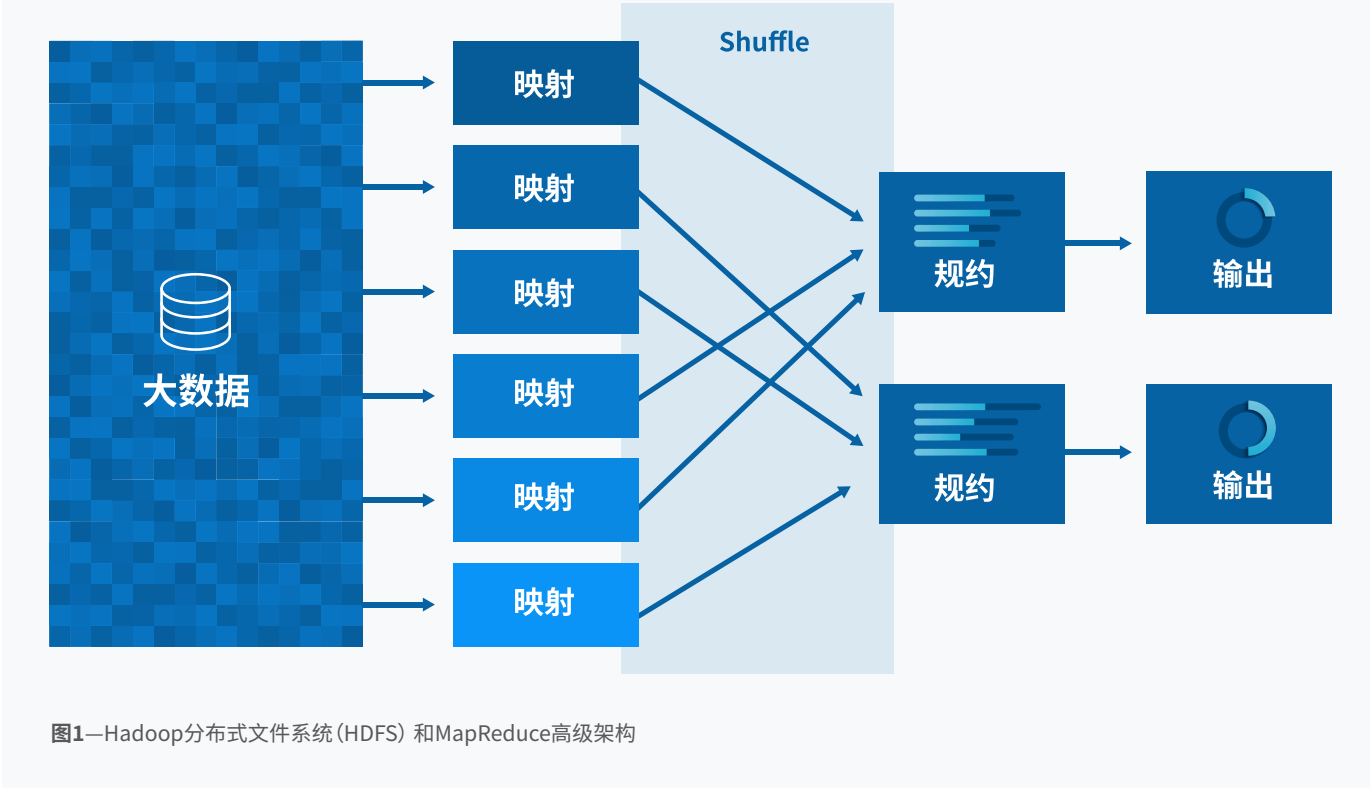


图1—Hadoop分布式文件系统（HDFS）和MapReduce高级架构

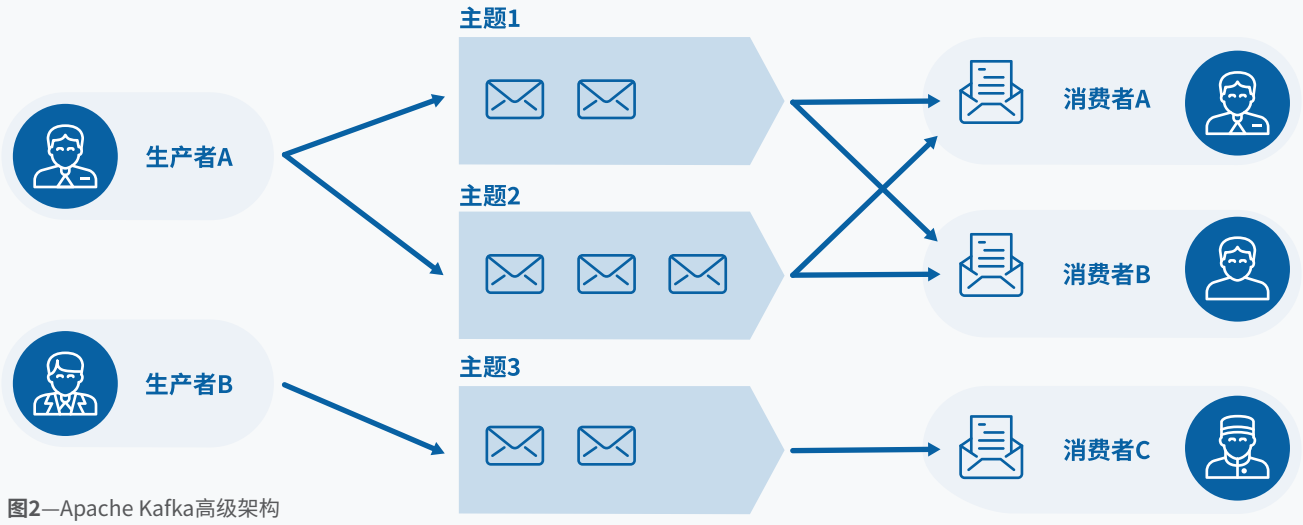


图2—Apache Kafka高级架构

在寻找更为经济的数据存储方式的过程中，很多公司发现了一个名为Hadoop的开源框架。借助Hadoop分布式文件系统（HDFS）和名为MapReduce[8]的批数据处理框架（见上一页的图1），该开源框架可利用商品硬件存储大量数据，从而大幅降低成本。

Hadoop为“数据湖”架构奠定了基础。在这个新范例中，存储和处理之间有明确的界限。存储是通过基于分布式文件系统的集中存储位置来实现的，它包含结构化、非结构化和半结构化数据——数据湖。

目前有若干种用于优化处理各类数据的处理引擎，这些引擎与数据湖相连，能检索查询所需数据，并计算这些查询的答案。

Hadoop打造出了一个高度可扩展的存储平台。它能以经济的方式将数据长期存储在数百个低成本服务器上，这些服务器并行运行的开销很低，几乎可以线性扩展。

数据湖架构能解决许多问题，并已用于大多数私有云中，但它们并不能解决所有问题。尽管它们成本低、可扩展，但它们无法为查询快速编写和读取数据。

8. https://research.google/pubs/pub51/
9. https://blog.linkedin.com/2011/01/11/open-source-linkedin-kafka
10. https://www.confluent.io/blog/okay-store-data-apache-kafka/

Apache Kafka的贡献

Apache Kafka项目始于LinkedIn，后来成为了一个开源项目[9]。创建的解决方案容错率高，并且能以固定方式长期（如有需要，可以无限期）存储数据。与数据湖不同，该方案的吞吐量更高，能以极低延迟（毫秒范围内）每秒处理数千条消息。

Kafka在一个或多个服务器上作为集群运行，该服务器可以跨多个数据中心工作，中心可将记录流存储在名为“主题”的类别中。添加额外节点能进一步提升可扩展性。Kafka还利用了复制和分区等功能。

Kafka用生产者应用程序编程接口（API）替换了消息代理，应用程序可利用API将记录流发布到一个或多个主题；也可以利用consumer API订阅一个或多个主题并处理生产者创建的记录流（参见图2）。但与消息代理不同，Kafka的持久性更佳。[10]

数据河

数 据管理中有不少与“水”有关的概念。除了数据湖，还有数据河。这是为什么呢?不同位置的数据源像溪流一样汇总形成主数据河。当一家公司准备利用分析建议来改进任何流程时，就会建起一道水坝。

第5章

处理

数据平台的处理层非常神奇。数据处理主要有两个层次：流处理和批处理。

流处理能实时处理数据，从开始接收数据后便能快速检测情况，并作出相应反应。它还能在数据生成后立即将数据输入分析工具，以接收即时结果。统计过程控制（SPC）控制图便是一例，它需要在控制或规范设定点出现异常后立即作出反应。

批处理能处理在给定时间段内存储的数据块，例如，处理在生产班次内执行的所有MES物联网事务，以更新数据仓库指标。这种处理特别适用于不需要实时指标但数据量大、处理时间长、计算量繁重的任务。

在统计处理或数据分析解决方案中，为提取和分析数据需要做部分转换处理，这些都是简单的数据准备步骤。

但其他转换工作的复杂度较高，操作难度较大。机器学习便是一例。机器学习需要大量的历史信息（训练集），这些信息均通过批处理收集（详见下文）。所有分析结束后，生成的结果就是算法，将算法应用于传入的数据后，便能使用流处理。

顶级数据科学家都能够充分利用数据平台。他们使用库处理并行数据（Apache Spark[11]是一款广泛应用的解决方案），并运行用R、Python或Scala等编程语言编写的代码。Spark能构建出可靠性高、内存计算快的流处理平台。

总而言之，像Spark这样的解决方案（见图3），可以在Kafka提供的高效持久层和分发层上提升容错性和数据处理效率。开源外部库越来越多，不少新库都新增了SQL查询（Spark SQL）、流处理、机器学习（MLib）和图分析（GraphX）等数据分析选项。

除了这些标准的数据分析任务之外，数据科学家还控制着数据扩充的过程，相关内容将在下一章中介绍。数据扩充可以决定对来自特定数据源的、对某项分析有用的数据的丰富程度。它们能使特定数据成为标准元素，之后，数据会通过数据接收器传输到关系数据库中，由商业智能分析人员进行分析。

11. https://spark.apache.org/

图3——Apache Spark生态系统

第6章

数据扩充

数据扩充是一种特殊的处理方式。数据扩充用其他数据扩充现有数据或信息，以创建新信息或洞察力，或使其更加完整。[12]

这一过程广泛应用于客户关系管理（CRM）[13]等领域，在这些领域中，现有记录中会有其他内外部数据并入，使潜在/现有客户信息更加详实。数据扩充也用于商业智能应用程序中，它能为数据添加相关维度，让数据更具价值，这里我们将其称为情境化。

在制造领域，数据扩充还属于新型概念，而且与传统方式有明显区别：使用外键创建关系数据库模型，将分散信息（例如，过程执行数据与设备数据或测试操作结果）关联起来。速度是最大的困难：这种模式无力持续增加新信息源，也跟不上这些信息源生成处理和存储数据的节奏。

正如我们在“第2章——暗数据”中所说，企业会出于安全原因存储未使用的数据，可能当前不会用到，但未来可能会用到。虽然数据扩充必须按照业务需求开展，但数据扩充的要求可能一时还无法掌握，或者会随着时间推移发生巨大变化。我们不仅要处理更多数据，还要处理非结构化的数据类型，如长文本或多媒体内容，这些数据类型很难以表格的形式（包括行和列）与关系数据库相匹配。

物联网和数据扩充

在制造业和物联网领域，这个话题的相关度更高。物联网系统以非常原始和非文本化的形式生成数据，系统越来越简易，成本越来越低。物联网系统需要在边缘，或稍后在数据河某处添加更多情境信息，使其真正具有利用价值。

在寻找最适合数据扩充的阶段时，目标之间常常会互相冲突。一方面，数据扩充应尽量在后期完成，因为用户可能并不知道自己想从这些数据中得到哪些洞察力。在业务目标不明确的情况下，盲目扩充数据只会浪费时间和资源。

另一方面，为了确保结果有效，扩充应在边缘或摄取过程中及时完成。这是因为如果在云中执行下一个处理和/或储存阶段，扩充过程的成本可能会大幅上升。靠近边缘进行扩充能减少源数据流的复杂性，或其“数字排放”，并将其简化为其他应用程序可以理解和利用的形式。

12. https://www.redpointglobal.com/blog/what-is-data-enrichment/
13. https://www.vainu.com/blog/data-enrichment/

因此，我们要采用多层次的方法来扩充数据

第一级是需要立即分析的数据。这些工作可以尽早完成。与业务目标相对应的其他高复杂度数据可以留待以后处理。

物联网数据的一级扩充与缺乏主数据有关。在车间系统中，主数据通常存储在MES系统中。在某些情况下，当主数据十分复杂时，专为管理主数据而创建的专用解决方案也并不少见。因此，MES和嵌入式主数据堪称制造企业的支柱。利用主数据扩充物联网的现象越来越普遍，这是正确利用分析解决方案，得出有用洞察力的一项基本要素。

然而，如果仅扩充主数据，那么数据扩充的更多优势将无法得到充分利用。

对于**第二级**扩充，在过程控制日趋完善后，仅凭情境化原始数据点所获得的优势很快就会减少。关联和发现不同来源的复杂因果分析能挖掘出更多价值。MES可在这方面发挥重要的作用；除情景化主数据外，MES还是信息基础设施的支柱，是主数据和运行时数据的锚，能连接车间中发生的物理和业务流程。

第二级扩充还有一项特点，即它可以在稍后时间应用于先前收集的数据点中。用户可以借助类似于Kafka的解决方案来“回放”和扩充原始数据，以便利用这些扩充后的数据点发布新的主题。这些新创建的数据流随后可用于下游分析解决方案。而这正好弥补了用于数据管理（数据湖、数据河）的水流模拟的不足之处。

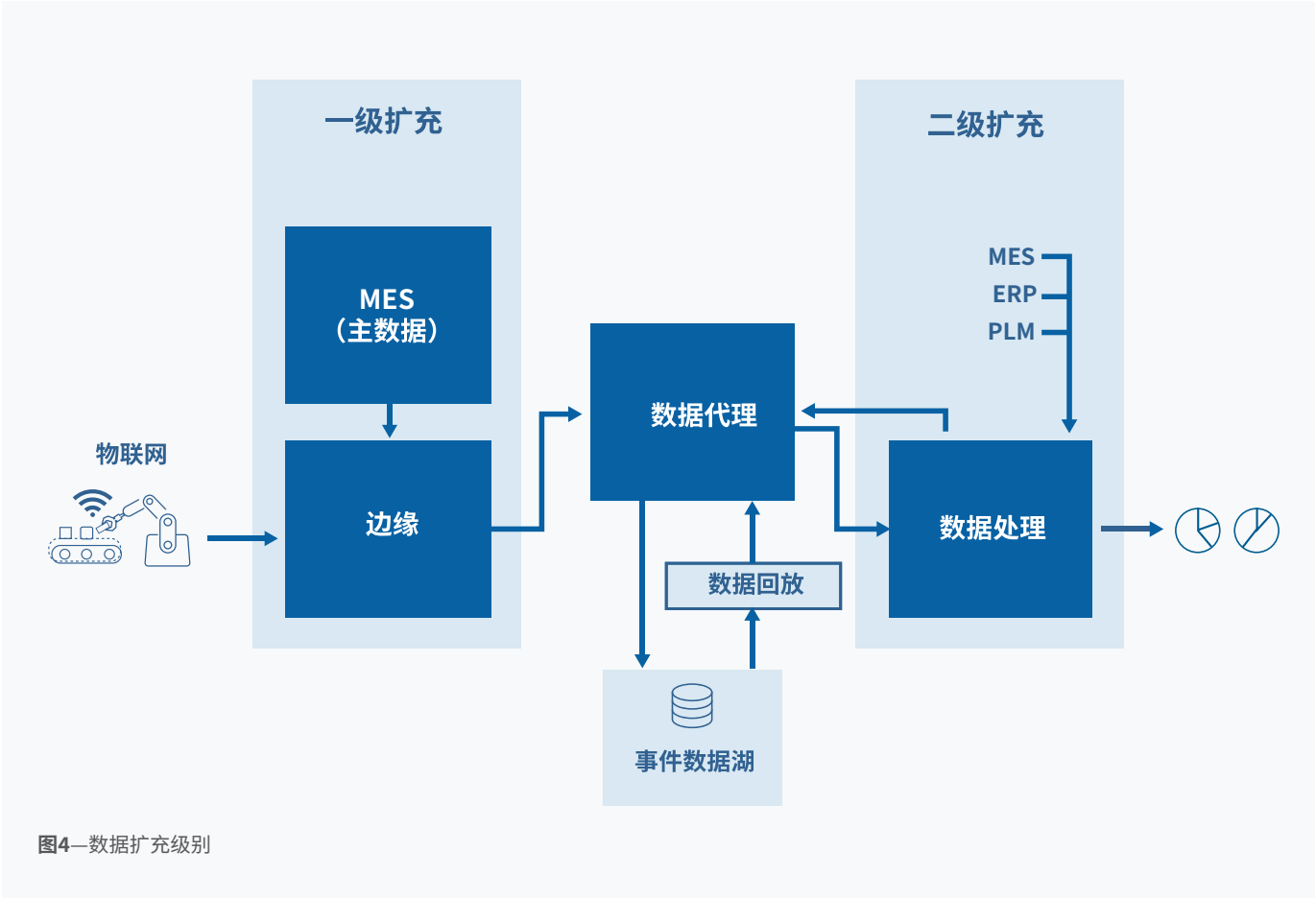


图4—数据扩充级别

第7章

数据分析

数据科学和分析趋势是当下的热门话题，虽然机器学习和人工智能炙手可热，但许多数据分析任务仍是由Microsoft Excel和SQL[14]完成。这至少能说明一件事：人们对这些应用程序太过熟悉，养成了默认使用的习惯；新技术还不成熟，无法解决数据科学家的需求。

此外，另一个挑战是数据分析涵盖大量活动，它们的复杂程度各不相同，而且通常由组织内的不同角色执行。常见分类如下：

- 商业智能（BI），它能解决企业在经营中遇到的常见问题。在制造业中，常见问题涉及生产量、产出量、质量、周期时间、库存等。解决这些问题的数据常以图表、仪表板或报告的形式直观呈现。

- 高级分析会用到更复杂的统计技术。在这里，机器学习发挥的作用将会越来越大。训练神经网络需要用到大量数据集，以便根据当前数据生成预测。在现有资产基础上，制造商的目标将是基于过去的数据（包括预测模型和近期的规范模型）开发未来绩效和行为模型。预测性维护是高级分析中的一个重点领域，它的投资回报周期更短，适用性更广，但与此同时，也有不少类似模型正在开发，以期全面提升工厂性能和效率。

由于商业智能和高级分析的需求、受众和所用工具都明显不同，因此关键是要将这两项功能区分开。部分数据分析员（特别是BI分析员）接受过基于表格格式思考和分析数据的培训。他们可能不熟悉流的概念，他们当前的工具可能也无法分析连续的数据流。为了解决这种问题，他们抓取快照（他们希望它是表格格式），然后运行SQL语句，应用统计分析，创建图表和报告。

对于他们来说，理想的数据存储库是关系数据库。缺点是，他们访问的数据必须经过某种程度的转换，对此他们没有控制权，他们只能对这个数据集执行分析。

不同类型的分析

在分析情境中，需要按复杂程度和优势将分析行为分成多种类型。分析可分为四类[15]，从最简单的开始依次为：描述性、诊断性、预测性，以及最复杂的规范性。

描述性分析

描述性分析是应用最广泛的基本分析类型。公司报告是常见的描述性分析案例，它能简单地介绍一个组织的运营、销售、财务、客户和利益相关者的历史信息。

对于工厂而言，物联网事件数据可视化的常规方法是利用历史数据记录器输入数据，并通过折线图实现数据可视化，观察折线图即可感知一个或多个变量随时间变化的情况。

用户可以利用描述性分析查看这些变量是否超过了阈值，并了解是否有趋势发生。警报也属于描述性分析，一旦变量超过特定阈值，就会触发报警，提示已出现必须解决的特定问题或异常情况。

描述性分析是最常用的分析类型，它与某些行业长期使用的历史数据记录器等应用程序没有太大区别。它显示过去发生的事情，所以通常它能提供“发生了什么?”这个问题的答案。如果您想在经营中保持前瞻性，那么通过此类分析获得的见解可能比较局限。

14. https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html
15. https://medium.com/@Magora_ltd/4-types-of-data-analytics-to-boost-your-business-b3e1d4d41903

诊断性分析

诊断性分析是高级分析的一种形式，它能检查数据或内容，并使用统计模型、数据发现和数据挖掘回答“为什么会发生这种情况?”的问题。它通常用于分析预测水平。

借助假设和变量之间的关系，诊断性分析可以洞察降级/故障或其他类型事件的情况，便于用户识别引发特定问题的根本原因。不过，诊断性分析也可能会非常复杂且耗时，这具体取决于变量的数量及它们之间的关系。统计知识和领域知识都很重要。

诊断性分析比描述性分析更具相关性，因为了解过根本原因后，用户未来可以采取措施规避这些问题。但更重要的是，诊断性分析为预测性分析奠定了基础，因为前者能减少因果分析中无意识的偏见和对相关性的误解，这对于缩小备选方案范围和迈入后续阶段至关重要。

预测性分析

在预测性分析中，人工智能会处理数据。它能回答“会发生什么?”的问题。经过前面的描述性和诊断性分析，我们掌握了变量行为及其关系的信息，纵观所有步骤，最有价值的是预测阶段。

预测性分析是数据科学家的主要领域，需要开展许多活动才能有效利用。首先，我们要确定预测性分析可以解决什么问题，以及哪些变量会影响结果。 然后，我们要确定将用于训练数据集的数据源。驱动因素（自变量）和结果（因变量）能否生成足够的数据，是该阶段面临的主要问题之一。

然后，还有确定所用分析类型或算法的关键阶段。在机器学习的情境中，可能会用到回归模型、决策树、神经网络、朴素贝叶斯算法。这里没有适用于所有情况的通用型模型。选择前，用户应全面了解不同模式在不同背景下的适用性和有效性。

最后，我们需要验证结果，验证它们是否有意义，是否可以大规模应用；也就是说，所得到的算法是否可以用于预测因驱动因素变化而导致的依变量的行为。需要注意的是，预测应以概率或估计来表述，而非准确结果。

规范性分析

规范性层面建立在上述所有层面的基础上，负责回答“应该怎么做?”的问题。如果预测模型的示例结果表明机器很可能发生故障，或剩余使用寿命较短，则规范性分析应推断出需采取哪些措施来纠正这种情况。

规范性分析从预测中得到结果，并结合规则和基于约束的优化，得出最终决策。决策与准确性和风险相关，结果可以自动化并立即应用，也可以仅提供建议，由用户做决定。

如今，达到规范性分析水平的公司仍在少数。据Gartner集团统计，2019年，只有3%的受访企业使用了规范性分析，而大量使用预测分析工具的企业则达到了30%。

这些问题大多与战术选择有关，特别是在短时间内需要做出大量决策时。如果问题更棘手，那么操作会变得更复杂，也可能更容易出错。目前，这一领域仍在高速发展，规范性分析的应用范围也在不断扩展。

第8章

分析用例--预测性维护

虽然在许多制造区域[16]中都可以依靠预测模型应用机器学习，但是在使用采用MES系统的物联网数据平台时，预测性维护一直是最受关注的话题。

预测性维护之所以受追捧，是因为它的投资回报周期更短。如果您能预测机器何时会出现故障，何时需要进行纠正性维护，您便可以在机器停止工作前实行预防性措施，避免产生负面影响。您可以避开不必要的预防性维护，现阶段的预防性维护通常取决于机器正常运行时间、生产的工件数量、机器运行的小时数，甚至特定的日历日。不必要的维护会产生高昂的成本，并导致真正需要维护的机器缺乏维护资源。

然而，预防性维护模型应用起来并不轻松。首先，我们需要可靠的数据，然后要正确建模，最后还要评估所做预测的质量。

但是，预测性分析最根本的问题可能是获取准确的历史数据，以便捕捉相关信息，确定是什么事件导致了故障。在制造设备故障领域创建强大人工智能模型也需要用到数据，但与社交网络等平台上的海量数据不同，这类数据很少。用户必须有一组深度足够的故障相关传感器的历史数据，而且故障要能够与收集的参数变化或特定事件相关联。否则，在最好的情况下，它仍然只是一个有根据的猜测。

设备故障的好消息和坏消息

随着制造商不断改良运营方式，灾难性的设备故障已经十分罕见。计划外机器停机通常会造成严重后果，此类事件不仅会影响相关机器运行，而且会导致整条生产线瘫痪，引发生产停工、工期重排等连锁反应。因此，制造商通常会采取预防性行动，有时比推荐周期更加频繁，以杜绝此类事件再次发生。

如果我们从多个传感器收集到了足够的数据，那么，设计问题框架就显得极为重要。在这方面，数据科学家需要和掌握设备操作相关知识的关键内部员工携手合作。这些学科专家了解机器的物理特性、操作条件，通常了解故障原因，具备相关专业知识（大多数故障本质上是机械故障）。

另一方面，从物联网传感器收集的数据必须具备丰富的情境信息，以便验证可能对故障产生较大影响的变量，这一点也很重要。因此，机器给定部分传感器的相关数据（如温度、压力、速度、声音、电流或振动）得到了完善，例如对原始数据点进行附加标记，包括设备、配方/程序、附加参数读数、当前生产的产品，距上次保养过去的时间等。

这些情境数据经过完善后，我们便能用其描述故障类型（例如，急性故障或降级），然后选择建模策略，例如用于预测剩余使用寿命的回归模型，或用于计算随时间变化的失效概率的生存模型。

16. https://www.ibmbigdatahub.com/blog/how-improve-manufacturing-roi-prescriptive-analytics

结合用于维护管理的预测性和规范性模型

为创建和测试各类预测模型，数据科学家可以使用物联网数据平台的批处理和机器学习库来创建训练集。入选模型经过应用和细化后，结果算法就可以永久使用。使用物联网数据平台的流处理特性收集每个新数据点时，该算法会调用API（例如，维护管理应用程序的API）来更新信息，以供维护技术人员使用。

维护管理通常是现代MES的一个模块，它可以使用这些数据向技术人员提供更多信息和建议，因此，该模块可以预测高故障率设备的维护活动（减少意外停机），并推迟低故障率设备的维护活动（避免不必要的维护，节省成本）。

维护也是一个可以使用规范性模型的领域。确定基于给定机械变量的机器故障预测结果后，便能确定操作内容：例如，更换给定零件。

展望未来，充分发挥技术人员的经验知识、充分利用强化学习（机器学习分支）也有助于持续提高模型的准确性。技术人员可以确认标记的维护执行需求，给算法“奖励”，然后算法便能相应地自我调整。缺少历史数据可能会产生较差或不完整的初始结果，而这种方法则是解决此类问题的一种有效方法。

第9章

If This, Then That

(若这样，就那样)

“If This, Then That”是一种鲜明的数据处理方式，起源于一个名为IFTTT[17]的家庭自动化相关商业解决方案。相反，IFTTT名称出自编程条件语句“if this, then that（若这样，就那样）”。

前提是，用户在不知道编码的情况下，可以创建简易脚本（配方）来自动化任务，其中一个设备或服务中的某类事件会自动触发另一个设备或服务中的某个操作。IFTTT具有互操作性，可与其他解决方案和供应商兼容。

将同样的IFTTT概念应用于流式制造数据有显著的优势。在许多情况下，离散事件或变量达到一定阈值时，会触发特定行动。

由于所有制造数据源的数据平台都是相同的，因此在处理来自机器MES/ERP事务的数据或事件时，都可以以通用方式使用其功能。IFTTT也不例外。

当涉及到事件时，我们会对特定的机器事件采取行动（例如机器停机、机器启动、特定类型的故障、可用的报告）。对于MES，我们会在批次暂停、放行或报废时采取行动；对于ERP，我们会对计划的新生产订单，或为特定产品更改的新BOM结构采取行动。

监控变量的变化趋势，并在其达到某个阈值时采取行动，此外，还可能发生其他事件，例如产量损失或循环时间超过限值，或者给定参数超过或低于限值（例如，当传感器的读数超过安全限值）。

在操作层面，难度自警报/通知到触发更复杂的操作（如工作流）依次递增。最常见的响应是发送电子邮件、文本消息或其他类型的警报，目的是通过调用API与现有系统和功能交互来扩展潜在选项的数量（例如，暂停或停止机器运行）。

这种IFTTT系统的基本特征是，它与操作执行系统分离（因此仅适用于动作可以异步触发的情况），而且易于配置，因此可以由制造工艺人员来操作，就像IFTTT在家庭自动化中的工作方式一样。

第10章

服务，还是输出

数据处理的最后一个块是服务，它包含处理完成后的输出或进一步的操作。

最简单的服务是通过数据接收器获得结果。数据接收器可用于接收其他设备的传入事件。通过多种类型的接收器，用户能够以文件、SQL表格形式，或使用OData[18]等多种方式处理输出数据（OData是一种标准方法，可确保与处理结果显示系统的互操作性。）。OData意为开放数据协议（opendataprotocol，ODP），是定义简易构建和使用API的标准。例如，使用OData可将数据导入Microsoft Power BI中。

其他输出包括在数据处理时触发活动或执行功能。这些活动包括可视化、调用外部系统的API、触发警报、运行机器学习算法等。

同样需要注意的是，当贵公司已经投资于商业智能或分析解决方案（例如Power BI、Tableau或SAS）时，这种服务和输出选项（尤其是使用SQL数据接收器的选项）支持无中断实施。所有已完成的工作都可以逐步使用数据平台，而不是直接访问源，这样有助于提升系统的可扩展性，并将平台转换为所有分析功能的单一信息源。

17. https://www.techrepublic.com/article/ifttt-the-smart-persons-guide/

18. https://www.odata.org/

在集成平台中，此类输出完成了可能生成数据的软件应用程序的闭环。例如，生成的设备集成数据点可能会因为某个值超出了可接受的范围而触发API，使设备停止运行；或者MES可能会根据解决方案中完成的事务处理而显示一个KPI的图表。

第11章

MES和数据平台

根据前几章中对数据平台的介绍，这些解决方案都非常有效。

很多人以为它们是解决所有问题的灵丹妙药。甚至还有人认为数据平台可以取代MES。然而事实并非如此，因为这两个系统的性质、运作方式、应对数据的颗粒度和实现的目的都有很大差异。不过，虽然两者不尽相同，但结合两种解决方案可以产生非常显著的协同效益。

MES是一种解决方案，旨在映射与制程，物理或业务相关项目，以确保流程可见、受控、符合法规要求。它是一个事务系统，可在步骤发生变化之前对其进行验证，然后注册和存储事务信息，以便后续跟踪。在数据收集方面，MES记录的数据粒度更高（例如，控制图和合规性/可追溯性）。由于MES的事务性属性，在商业智能方面，它只能用于查找记录和分析历史信息（在数据仓库中对聚合数据进行报告或交互）。

而数据平台则是一种设计用于接收、存储和处理不同来源的大型和高频数据集的解决方案，便于用户分析实时数据点和/或大量存储的单个数据点。记录的数据主要用于实时或后期转换/计算和分析。就BI而言，数据平台通常支持许多流处理或批处理的分析选项，包括数学、统计和人工智能功能（如机器学习）。

不过，如果能结合这两种解决方案，就会产生更多优势。首先，MES也可以作为数据发生器——所有MES交易事务都可以作为事件被发送到数据平台。此外，如数据扩充一章所述，MES是一种非常特殊的数据源，它能为其他非情境化数据点提供情境化信息，这些数据点来自设备传感器等其他数据生成设备。然而，MES不仅仅是一个情境化层，它更是一种强大的公共或规范数据模型（CDM）。

在数字时代大获成功的公司往往都改进了他们的数据集成方式，而不是局限于收集和挖掘数据。顾名思义，CDM能够更加顺畅地集成独立的系统，这不仅能优化流程，也能简化数据分析和数据挖掘工作。它还能使不同的应用程序（基于Spark或其他框架构建）理解模型并执行操作或分析。大公司需要从零开始创建自己的CDM，这是一项高难度的长期工作，过程中会经历数个试错周期，并且需要将数据从其他系统转换到CDM中。不过，一旦MES创建完毕，其模型的一个子集即可用作数据平台的规范模型，但会随时间进一步完善。这样一来，各类分析，或者任何在数据平台上工作的应用程序都能有一个公共的数据字典，以便理解查询系统所需的架构。

最后，MES是制造工厂中不同人员使用的操作工具，而数据平台则是自动运行的解决方案，而非设置或配置时运行的方案。因此，数据平台的所有输出都可以在MES层面实现可视化；数据 平台的结果可在MES中实现可视化；或者由MES触发新的数据平台分析。这样就完全闭合了MES和数据平台之间的环路，使组合解决方案真正达到协同的效果。

第12章

融合贯通

在前面的章节中，我们已经带您了解了数据平台所涉及的制造业通用概念。尤其要指出的是，重用现有组件（如边缘解决方案）和添加其他制造软件（如MES）的上下文信息，方能发挥数据管理的真正价值。相比单独应用这些技术，制造商整合使用多种解决方案将能获得更多效益。

以下是上述主要内容的概括。现代数据平台架构是围绕流的概念构建的，由一系列执行不同功能的组件组成。各个实施方案可能略有不同，但通常包括以下内容（参见图5）：

1. **边缘处理**——首先从传感器或其他来源捕获数据。第一级处理接近需要提高响应时间、节省带宽和提高系统安全性的位置。

2. **数据摄取**——最明显的问题之一（在物联网出现之前，这个问题在系统架构中的优先级并不高）就是数据摄取的问题，因为数据源的数量和收集数据点的频率呈急剧上升态势。数据摄取计划通过提供一个中心来解决这个问题，使不同的边缘系统能够连接到这个中心，确保事件接收和数据验证安全高效

3. **数据代理**——确保数据分发到订阅了某些数据的所有进程。数据代理必须性能出众，以便实时分析和响应，同时也要具备容错功能。这些解决方案可以处理流处理和批处理需求，并根据请求从任何一点开始“回放”数据。应用最广的数据代理解决方案是开源项目Kafka。

4. **事件持久化**——对于处理大量可能无法立即使用的数据的数据平台来说，以低成本、容错的方式对原始数据进行事件持久化十分重要。Kafka类的技术用于数据代理时，可保证原始数据长期存储，打造出事件数据的数据湖（持久层）。

5. **数据处理**——从数据代理处收集必要的数据，涉及操作或汇总流数据或批处理数据，包括转换和数据汇总。Apache Spark是数据处理时可用的一款解决方案，它也是一个开源项目，可提供高度可靠、快速的“内存”计算和流处理平台。Spark类的解决方案可以在Kafka提供的高效持久层和分发层上提升容错性和数据处理效率，增加各种分析、图形和机器学习功能，使分析堆栈更加完善。

6. **数据扩充**——数据扩充是指在到达平台的原始事件中添加信息的过程，让足够的情境信息或相关数据点关联起来，生成更好的分析建议。

7. **行动/输出**——最后一个块（也称为服务）包括触发活动或数据处理的执行功能。这些活动包括可视化、存储、数据接收器、调用外部系统的API、触发警报、运行机器学习算法等。

显然，除了本白皮书中提到的开源元素外，微软、AWS、谷歌、IBM等公司也为物联网平台提供了多种商业化的解决方案。各公司使用的技术各不相同；有些是专有的，有些是开源的，适用于堆栈的所有领域；可选定价方案有多种，具体取决于连接设备数量、消息数量、分析数据数量等方面。

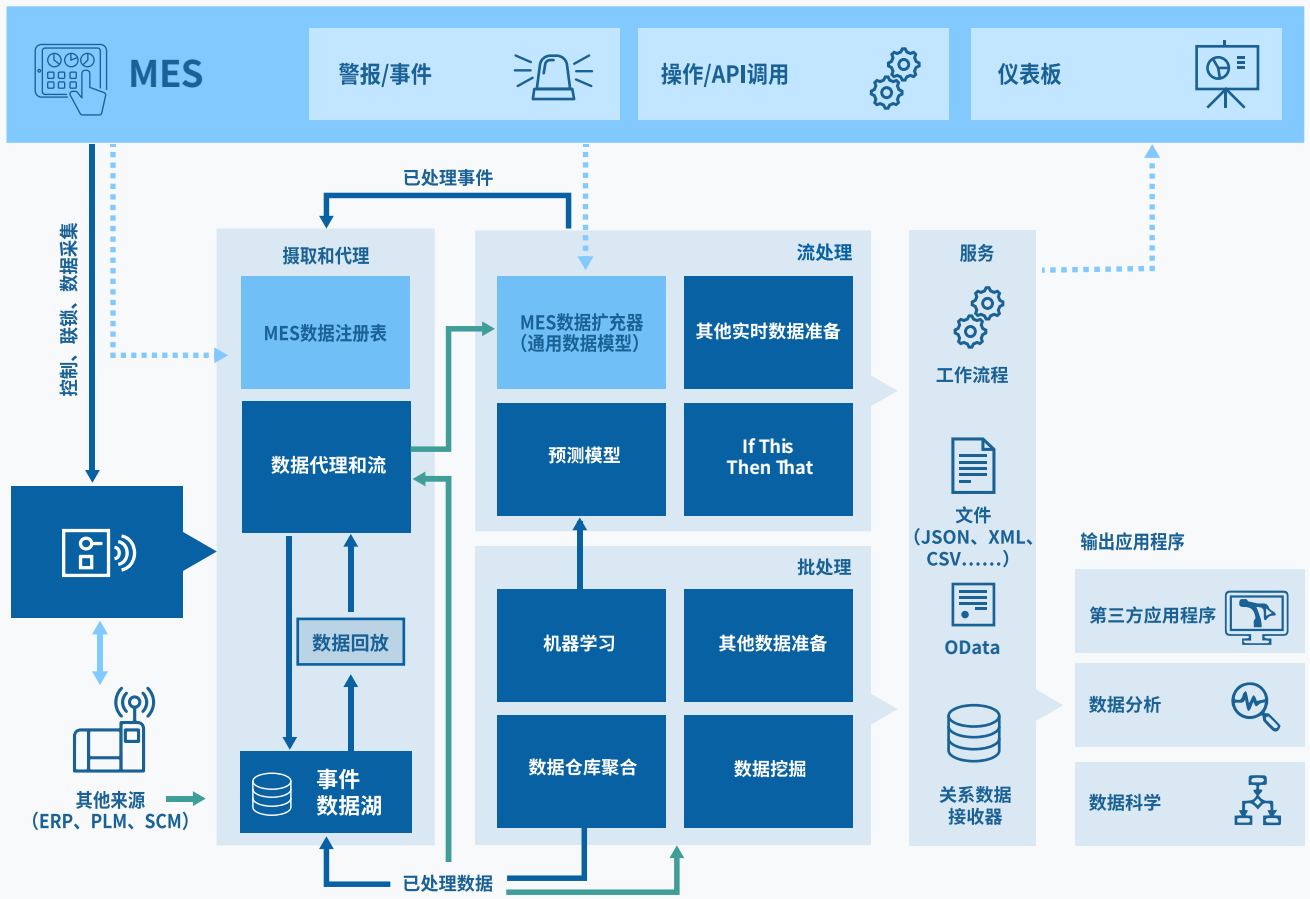


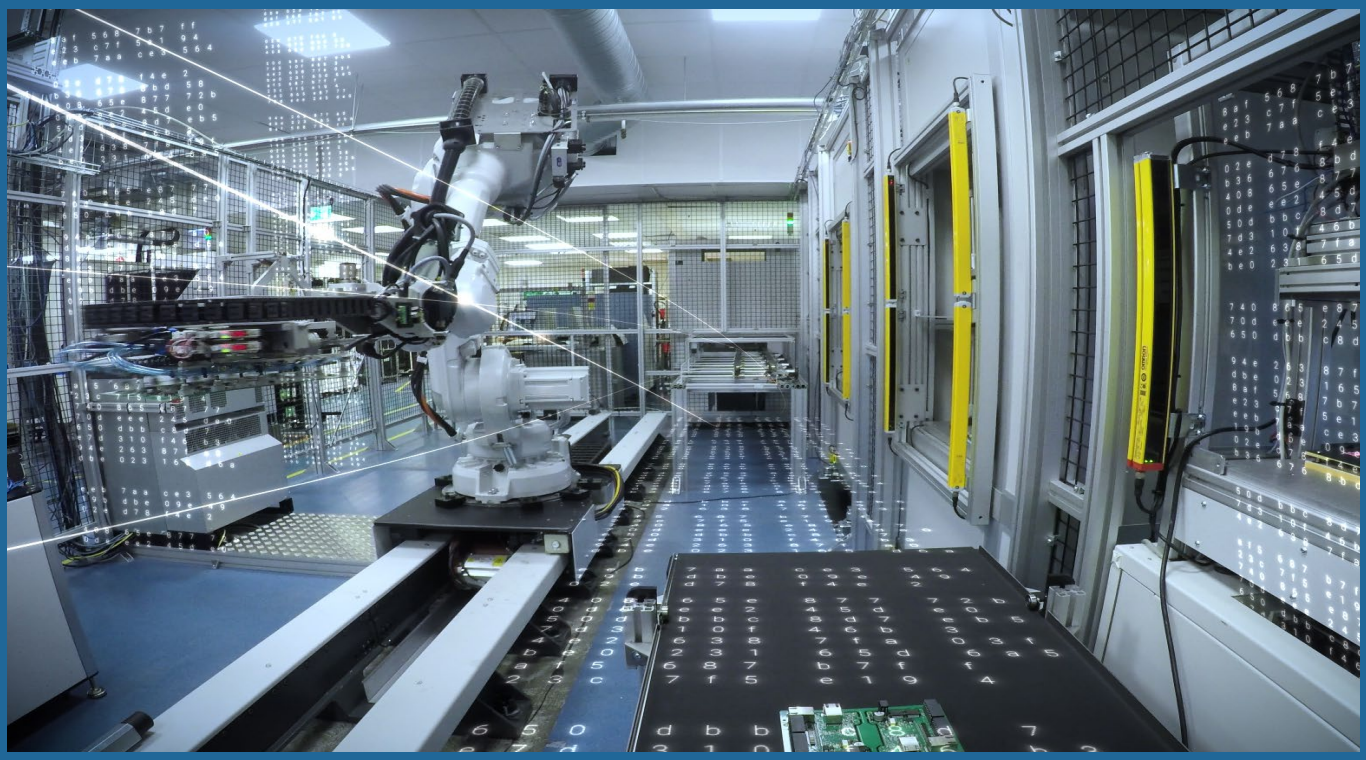
图5—支持MES的物联网数据平台

要点回顾

了解数据能为企业做什么，应该采用什么类型的技术来访问和存储，应该收集哪些数据，以及可以部署哪些策略来使用这些数据十分重要，正确操作不仅可以提升企业运营效率，而且还能更有效、迅速地应对市场压力。

数据平台如果与MES协同工作，势必将发挥出更大的价值。此类平台可以为大量用例创造价值，而这些用例也将在制造企业的投资下不断发展壮大。

凯睿德制造致力于持续深耕MES领域，打破车间解决方案之间的壁垒，通过融合智能、运营和自动化技术，为工业4.0赋予更多商业价值。



Francisco Almada Lobo拥有波尔图大学的工商管理硕士和电子工程学位。他的职业生涯始于CIM研发机构，1997年，他入职了Siemens Semiconductor。在西门子、英飞凌和奇梦达工作期间，他在多个制造领域积累了丰富的经验，并在2004年领导了一个运行中大体量MES系统的迁移工作。

2010年，他正式出任凯睿德制造首席执行官，成功引领公司发展为MES领域的龙头企业。他是福布斯技术委员会成员、200M Fund Investment Committee成员、SEMI智能制造技术委员会欧洲分会执行委员会成员。

凯睿德制造可提供灵活度高、易于配置的现代制造执行系统（MES）。凯睿德制造MES能帮助制造商满足严苛的产品可追溯性和合规性要求；以固有的闭环质量降低风险；与企业系统和工厂自动化无缝集成，并为全球生产运营提供深度智能、可视性佳的解决方案。

因此，我们的客户已经为工业4.0做好了准备。即便市场需求、机遇或客户要求发生变化，他们也可以随时随地轻松调整自己的业务，从而更有效地开展竞争并获得收益。

如需了解更多信息，敬请访问：www.criticalmanufacturing.com/cn

凯睿德制造软件（苏州）有限公司
电话: 400 666 3830
地址：江苏省苏州市苏州工业园区世纪金融大厦411室
Email: lisayang@criticalmanufacturing.com

关注凯睿德官方微信公众号
获取更多MES资讯