DAMA-DMBOK

数据管理知识体系指南CDGA/CDGP认证

第14章 大数据和数据科学（完整课程视频请扫描二维码）

第14章 大数据和数据科学

01

引言

定义、业务驱动因素、目标和原则、基本概念

定义

定义：对多种不同类型的数据进行收集（大数据）和分析（数据科学、分析、可视化）以此来为在分析的初始阶段未知的问题找到答案。

业务驱动因素

期望抓住从多种流程生成的数据集中发现的商机，是提升一个组织大数据和数据科学能力的最大业务驱动力。

目标和原则

原则：
组织应仔细管理与大数据源相关的元数据，以便对数据文件及其来源和价值进行准确的清单管理。

目标：
发现数据和业务的联系
支持将数据源迭代集成到企业中
发现和分析可能影响到业务的因素
利用可视化技术，以恰当的、可靠的且合乎道德规范的方式来发布数据

1、数据科学

数据科学将数据挖掘、统计分析和机器学习与数据集成整合，结合数据建模能力，去构建预测模型、探索数据内容模式。

数据科学依赖于：

1）丰富的数据源：具有能够展示隐藏在组织或客户行为中不可见模式的潜力

2）信息组织和分析：用来领会数据内容，结合数据集针对有意义模式进行假设和测试的技术

3）信息交付

4）展示发现和数据洞察：分析和揭示结果，分享洞察观点

2、数据科学的过程

1）定义大数据战略和业务需求

每一步输出是下一步输入。可衡量的需求

2）选择数据源

3）采集和提取数据资料

4）设定数据假设和方法

5）集成和调整数据进行分析

模型的可行性部署取决于源数据的质量。

6）使用模型探索数据

对集成的数据应用统计分析和机器学习算法进行验证、训练，并随着时间的推移演化模型。

7）部署和监控

可以将产生有用信息的那些模型部署到生产环境中，以持续监控它们的价值和有效性。

3、大数据

1）数据量大（Volume）
2）数据更新快（Velocity）
3）数据类型多样/可变(Variety、Variability)
4）数据黏度大（Viscosity）
数据使用或集成的难度比较高
5）数据波动性大（Volatility）
数据更改的频率，以及由此导致的数据有效时间短
6）数据准确性低（Veracity）
数据的可靠程度不高

4、大数据架构组件

5、大数据来源

6、数据湖

1）数据科学家可以挖掘和分析数据的环境
2）原始数据的集中存储区域，只需很少量的转换
3）数据仓库明细历史数据的备用存储区域
4）信息记录的在线归档
5）可以通过自动化的模型识别提取流数据的环境

7、基于服务的架构

基于服务的体系架构（Services-Based Architecture，SBA）正在成为一种立即提供数据的方法，并使用相同的数据源来更新完整、准确的历史数据集。

1）批处理层

数据湖作为批处理层提供服务，包括近期的历史和数据

2）加速层

只包括实时数据

3）服务层

提供连接批处理和加速层数据的接口

8、机器学习

探索了学习算法的构建和研究，它可以被视为无监督学习和监督学习方法的结合。

无监督学习通常被称为数据挖掘，而监督学习是基于复杂的数学理论，特别是统计学、组合学和运筹学

机器学习三种类型

1）监督学习

基于通用规则（如将SPAM邮件与非SPAM邮件分开）

2）无监督学习

基于找到的那些隐藏的规律（数据挖掘）

3）强化学习

基于目标的实现（如在国际象棋中击败对手）

9、语义分析

使用自然语言NLP分析短语或句子、语义察觉情绪，并揭示情绪的变化，以预测可能的情景

10、数据和文本挖掘

数据挖掘是一种特殊的分析方法，它使用各种算法揭示数据中的规律。

1）剖析

剖析尝试描述个人、群体或人群的典型行为，用于建立异常检测应用程序的行为规范，如欺诈检测和计算机系统入侵监控。剖析结果事许多无监督学习组件的输入。

2）数据缩减

数据缩减是采用较小的数据集来替换大数据集，较小数据集中包含了较大数据集中的大部分重要信息。

3）关联

关联是一种无监督的学习过程，根据交易涉及的元素进行研究，找到它们之间的关联。

4）聚类

基于数据元素的共享特征，将它们聚合为不同的簇。

5）自组织映射

是聚类分析的神经网络方法，有时被称为Kohonen网络或拓扑有序网络，旨在减少评估空间中的维度，同时尽可能地保留距离和邻近关系，类似于多维度缩放

降维就像从等式中移除一个变量而不影响结果，使得这些问题变得更容易被解决，数据更容易被展示出来，

11、预测分析

预测分析是有监督学习的子领域，用户尝试对数据元素进行建模，并通过评估概率估算来预测未来结果。

最简单形式是预估。有许多基于回归分析做预估并从平滑算法中受益的技术，平滑数据的最简单方法是通过移动平均值，甚至是加权平均值。

12、规范分析

规范分析比预测分析更进一步，它对将会影响结果的动作进行定义，而不仅仅是根据已发生的动作预测结果。

13、非结构化数据分析

结合了文本挖掘、关联分析、聚类分析和其他无监督学习技术来处理大型数据集。监督学习技术也可用于在编程过程中提供方向、监督和指导，利用人为干预在必要时解决歧义问题。

14、运营分析

运营分析也称为运营BI或流式分析，其概念是从运营过程与实时分析的整合中产生的。包括用户细分、情绪分析、地理编码以及应用于数据集的其他技术，用于营销活动分析、销售突破、产品推广、资产优化和风险管理

15、数据可视化

可视化是通过使用图片或图形表示来解释概念、想法和事实的过程。数据可视化通过视觉概览来帮助理解基础数据。

16、数据混搭

Data Mashups 将数据和服务结合在一起，以可视化的方式展示见解或分析结果。许多虚拟化工具通过一些功能实现混搭，通过公共数据元素关联数据源，这些元素最初用于将名称或描述性文本关联到存储的代码。

02 活动

1、定义大数据战略和业务需求

大数据战略必须包括以下评估标准：

1）组织试图解决什么问题，需要分析什么
2）要使用或获取的数据源是什么
3）提供数据的及时性和范围
4）对其他数据结构的影响以及与其他数据结构的相关性
5）对现有建模数据的影响。包括扩展对客户、产品和营销方法的知识。

2、选择数据源

了解以下基本事实：

1）数据源头
2）数据格式
3）数据元素代表什么
4）如何连接其他数据
5）数据的更新频率

管理数据源：

1）基础数据
2）粒度
3）一致性
4）可靠性
5）检查/分析新数据源

3、获得和接收数据源

迭代地识别当前数据资产基础和这些数据源的差距，使用分析、可视化、挖掘或其他数据科学方法探索这些数据源，以定义模型算法输入或模型假设。

4、制定数据假设和方法

数据科学能够发现数据的意义和其中蕴含见解的答案集。制订数据科学方案需要构建统计模型，找出数据元素和数据集内部以及二者之间的相关性和趋势。

模型的效果取决于输入数据的质量和模型本身的及安全性。

5、集成和调整数据进行分析

准备用于分析的数据包括了解数据中的内容、查找各种来源的数据间的链接以及调整常用数据以供使用。

一种方法是使用共有键整合数据的通用模型；另一种方法是使用数据库引擎内的索引扫描和连接数据，以获得相似性和记录连接的算法和方法。

6、使用模型探索数据

1、填充预测

使用历史信息预先填充配置预测模型，这些信息涉及模型中的客户、市场、产品或模型触发因素之外的其他因素。

2、训练模型

需要通过数据模型进行训练。训练包括基于数据重复运行模型以验证假设，将导致模型更改。训练需要平衡，通过针对有限数据文件夹的训练避免过度拟合。

转换到生产之前，必须完成模型验证。通过训练和验证的模型偏移量来解决任何填充失衡或数据偏差问题。

3、评估模型

将数据放入平台并准备分析后，数据科学就开始。

需要用到数据科学实践中的一个道德组件

4、创建数据可视化

设定可视化的目的和参数：

时间点状态、趋势与异常、移动部分之间的关系、地理差异及其他

7、部署和监控

1、揭示洞察和发现

通过数据可视化来展示和发现和数据洞察是数据科学研究的最后一步，洞察应与行动项目相关联，这样组织才能从数据科学工作中受益。

2、使用附加数据源迭代

从特定的一组数据源中学习的过程，通常会导致需要不同的或额外的数据源，以支持得到的结论并向现有模型中添加洞察。

03 工具和方法

其他改变查看数据和信息方式的技术：

1）数据库内的高级分析
2）非结构化数据分析（Hadoop，MapReduce）
3）分析结果与操作系统的集成
4）跨多媒体和设备的数据可视化
5）链接结构化和非结构化信息的语义
6）使用物联网的新数据源
7）高级可视化能力
8）数据扩展能力
9）技术和工具集的协作

1、MPP无共享技术和架构

大规模并行处理（MPP）的无共享数据库技术，已成为面向数据科学的大数据集分析标准平台。

在MPP数据库中，数据在多个处理服务器之间进行分区，每个服务器都有自己的专用内存来处理本地数据。处理服务器之间的通信通常由管理节点控制，并通过网络互联进行。因为该架构没有磁盘共享，也不发生内存争用，因此称作“无共享”。

该技术还支持数据库内分析功能——在处理器级执行分析功能（如K-Means，回归分析的的能力）。

2、基于分布式文件的数据库

基于文件的解决方案中使用的模型称为MapReduce。该模型有三个主要步骤：

1）映射（Map）
识别和获取需要分析的数据

2）洗牌（Shuffle）
依据所需的分析模式组合数据

3）归并（Reduce）
删除重复或执行聚合，以便将结果数据集的大小减少到需要的规模。

3、数据库内算法

数据库内算法（In-database algorithm）使用类似MPP的原则，MPP无共享架构中的每个处理器可以独立运行查询，因此可在计算节点级别实现新形势的分析处理。

4、大数据云解决方案

5、统计计算和图形语言

R语言是用于统计计算和图形的开源脚本语言环境。它提供了各种各样的统计技术，如线性和非线性建模、经典统计检验、时间序列分析、分类和聚类。

6、数据可视化工具集

这些工具的优势：

1）复杂的分析和可视化类型
2）内置可视化最佳实践
3）交互性，实现视觉发现

1、解析建模

要通过其他应用程序共享和执行模型，需查找支持预测模型标记语言（PMML）的工具，这是一种基于XML的文件格式。

利用API接口直接进入存储层HDFS，可以提供各种数据访问技术，如SQL、内容流、机器学习和用于数据可视化的图形库，

解析模型与不同的分析深度相关联：

1）描述性建模以紧凑的方式汇总或表示数据结构。这种方法并不总能验证因果假设或预测结果，但确实能够使用算法定义或改善变量之间的关系，从而为这种分析提供输入。

2）解释性建模是数据统计模型的应用，主要是验证关于理论构造的因果假设。虽然它使用类似于数据挖掘和预测分析的技术，但其目的不同。它不能预测结果，只是将模型结果与现有数据相匹配。

预测分析的关键是通过训练模型来学习，学习方法的效果取决于它在测试集上的预测能力。

避免过度拟合——这种情况发生在用于训练模型的数据集不具有代表性，模型过于复杂，或者将少量噪声数据具有的特性当做大部分数据的共性时。

训练误差会随着模型复杂性的提高而持续降低，并且可以降至零。数据集随机分成三个部分：训练集、测试集和校验集。训练集用于拟合模型，测试集用于评估最终模型的泛化误差，校验集用于预测选择的误差。

2、大数据建模

大数据建模是一项技术挑战，对想要描述和管控数据的组织而言至关重要。

对数据仓库进行物理建模的主要驱动因素是为查询性能而启用数据填充。

04 实施指南

就绪评估/风险评估、组织和文化变革

管理大数据：确保数据源可靠、具有足够的元数据以支持数据使用、管理数据质量、确定如何整合来自不同源的数据，以及确保数据安全且受到保护。实施大数据环境的差异与一组未知问题有关：如何使用数据、哪些数据有价值、需要保留多长时间。

1、战略一致性

战略交付成果应考虑管理以下要素：

1）信息生命周期
2）元数据
3）数据质量
4）数据采集
5）数据访问和安全性
6）数据治理
7）数据隐私
8）学习和采用
9）运营

2、就绪评估/风险评估

1）业务相关性
2）业务准备情况
3）经济可行性
4）原型
5）可能最具挑战性和决策将围绕数据采购、平台开发和资源配置进行。
6）数字资料存储有许多来源，并非所有来源都需要内部拥有和运营。
7）市场上有多种工具和技术，满足一般需求僵尸一个挑战
8）及时保护具有专业技能的员工，并在实施过程中留住顶尖人才，可能需要考虑替代方案，包括专业服务、云采购或合作。
9）培养内部人才的时间可能会超过交付窗口的时间

3、组织与文化变迁

跨职能角色：

1）大数据平台架构师

硬件、操作系统、文件系统和服务

2）数据摄取架构师

数据分析、系统记录、数据建模和数据映射，提供或支持将源映射到Hadoop集群以进行查询和分析

3）元数据专家

元数据接口、元数据架构和内容

4）分析设计主管

最终用户分析设计、最佳实践依靠相关工具集指导实施，以及最终用户结果集简化

5）数据科学家

提供基于统计和可计算性的理论知识，交付适当的工具和技术，应用到功能需求的架构和模型设计咨询。

05

治理

过程控制、解决方案文档、标准和指南

大数据需要业务和技术控制，解决以下问题：

1）寻源：来源有哪些，什么时候接入源，什么是特定研究的最佳数据来源

2）共享：组织内部和外部要签订的数据共享协议和合同、条款和条件

3）元数据：数据在源端意味着什么，如何解释输出端的结果

4）丰富：是否丰富数据，如何丰富数据，以及丰富数据的好处

5）访问：发布什么，向谁发布，如何以及何时发布

1、可视化渠道管理

2、数据科学和可视化标准

标准包括：

1）分析范例、用户团体、主题域的工具标准

2）新数据的请求

3）数据集流程标准

4）采用中立的、专业的陈述过程，避免产生有偏见的结果，并确保所有要素都以公平一致的方式完成，包括：

①数据包含和排除

②模型中的假设

③结果统计有效性

④结果解释的有效性

⑤采用适当的方法

3、数据安全

4、元数据

元数据特征化数据的结构、内容和质量，包括数据的来源、数据的血缘沿袭、数据的定义、以及实体和数据元素的预期用途。技术元数据可以从大数据工具中获取，包括数据存储层、数据整合、MDM甚至源文件系统。考虑实时数据、静态数据和计算性数据元素，就要明确源端的数据沿袭关系。

5、数据质量

数据质量是与预期结果偏差的度量：差异越小，数据满足期望越好，质量就越高。

高级数据质量工具集的功能：

1）发现：信息驻留在数据集中的位置
2）分类：基于标准化模式存在哪些类型的信息
3）分析：如何填充和构建数据
4）映射：可以将哪些其他数据集与这些值匹配

度量指标

1、技术使用指标

使用技术分析手段查找数据热点（最常访问的数据），以便管理数据分发和保持性能。

2、加载和扫描指标

定义了提取率以及与用户社区的交互。

3、学习和故事场景

常用的测量方法包括：

1）已开发模型的数量和准确性

2）已识别的机会中实现的收入

3）避免已识别的威胁所降低的成本

本章完结 感谢观看

完整课程视频请扫描二维码咨询