**大数据架构职业能力图谱**

* **计算机技术** 02
  + 操作系统
  + 数据库
  + 云计算
  + 分布式存储
  + 分布式计算
* **编程工具与技术** 03
  + Python编程语言
  + Excel
  + JavaScript和HTML
  + SQL
* **网络管理** 06
  + 网络设计
  + 云计算开发与运维
  + 高性能计算与可扩容性
  + 咨询技能
* **数据治理（Data Governance）** 09
  + 数据库架构
  + 数据摄入（Data ingestion）解决方案
  + 数据安全
  + 信息安全
* **数据分析** 10
* 数据项目质量控制
* 数据分析编程
* 数据分析项目设计与执行
* 数据分析类型
* 行业经验
* **机器学习** 08
  + 聚类分析（Clustering）
  + 判别归类分析
  + 决策树
  + 机器学习
* **数据沟通与可视化** 11
  + 数据准备
  + 可视化工具设计

计算机技术

新的想法诞生新的技术，让人应接不暇。这些新的技术，这些新兴应用和对应的IT发展趋势，使得数据分析师必须了解甚至掌握最新的IT技能。数据库开发和管理在大数据时代显得尤为重要，相关的数据库管理、运维和开发技术，将成为广大BI、大型企业和咨询分析机构特别看重的技能体现。代表着更多类型（尤其是非结构化类型）的海量数据的涌现，要求我们实时采集、分析、传输这些数据集，在对基础设施提出严峻挑战的同时，也特别强调了数据库开发和管理人员的挑战。比如分布式的、面向海量数据管理的数据库系统之一NoSQL，就是面向大数据领域的非关系型数据库的流行平台，高可用、大吞吐、低延迟、数据安全性高等应用特点成为了很多企业的看重的特点。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理  工程师 | | | 大数据  工程师 | | | 大数据架构  工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
|   **操作系统** |  |  |  |  |  |  |  |  |  |
|  Linux |  |  |  |  |  |  |  |  |  |
|  Windows Server |  |  |  |  |  |  |  |  |  |
|  UNIX |  |  |  |  |  |  |  |  |  |
|   **数据库** |  |  |  |  |  |  |  |  |  |
|  数据的种类和特点 |  |  |  |  |  |  |  |  |  |
|  传统的关系性数据库 |  |  |  |  |  |  |  |  |  |
|  数据库管理 |  |  |  |  |  |  |  |  |  |
|   **云计算** |  |  |  |  |  |  |  |  |  |
|  弹性计算 |  |  |  |  |  |  |  |  |  |
|  云存储：RDS、OTS、ADS |  |  |  |  |  |  |  |  |  |
|  资源管理 |  |  |  |  |  |  |  |  |  |
|   **分布式存储** |  |  |  |  |  |  |  |  |  |
|  Hadoop |  |  |  |  |  |  |  |  |  |
|  HDFS OR Maxcompute |  |  |  |  |  |  |  |  |  |
|  大数据整合 |  |  |  |  |  |  |  |  |  |
|  阿里云Data IDE |  |  |  |  |  |  |  |  |  |
|   **分布式计算** |  |  |  |  |  |  |  |  |  |
|  MapReduce |  |  |  |  |  |  |  |  |  |
|  Spark |  |  |  |  |  |  |  |  |  |
|  Spark Streaming |  |  |  |  |  |  |  |  |  |
|  Pig/Hive/HBase |  |  |  |  |  |  |  |  |  |

编程工具与技术

编程将是数据分析师日常工作的一个组成部分。这一关键技能将这个职业从一个传统的业务分析师或统计员区分出来。在日常工作中，或提取数据分析师需要编写程序从数据库中查询数据。根据职位的要求，有时也需要编写机器学习算法程序。数据编程语言有各自不同的特点，因此，数据分析师通常需要了解并掌握一个或多个编程语言，并对一些用于科学计算的模块有深刻的理解。

***注：大数据助理工程师必须具备SQL，Excel技能；大数据 工程师和大数据架构工程师必须具备Python技能，同时可选其他一种计算机语言（如JavaScript和HTML）***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理工程师 | | | 大数据  工程师 | | | 大数据架构工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
| £  **Python编程语言** |  |  |  |  |  |  |  |  |  |
| £ 数据操作 |  |  |  |  |  |  |  |  |  |
| £ 从外部文件导入数据 | ê |  |  |  | ê |  |  | ê |  |
| £ 连接数据库 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据类型 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据对象 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据类型转换 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据预处理 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据抽样 | ê |  |  |  | ê |  |  | ê |  |
| £ 数学函数、统计函数 | ê |  |  |  | ê |  |  | ê |  |
| £ 条件、循环、控制执行 | ê |  |  |  | ê |  |  | ê |  |
| £ 自编函数 | ê |  |  |  | ê |  |  | ê |  |
| £ 科学计算 |  |  |  |  |  |  |  |  |  |
| £ 统计模型 |  |  |  | ê |  |  | ê |  |  |
| £ 机器学习 |  |  |  | ê |  |  | ê |  |  |
| £ 高性能计算 |  |  |  |  |  |  |  |  |  |
| £ Python + Hadoop编写分布式程序 |  |  |  | ê |  |  | ê |  |  |
| £ Scala |  |  |  | ê |  |  | ê |  |  |
| £  **EXCEL** |  |  |  |  |  |  |  |  |  |
| £ 数据操作 |  |  |  |  |  |  |  |  |  |
| £ 从外部文件导入数据 |  | ê |  |  | ê |  |  | ê |  |
| £ 连接数据库 |  | ê |  |  | ê |  |  | ê |  |
| £ Pivotal Table |  | ê |  |  | ê |  |  | ê |  |
| £ 数学函数、统计函数 |  | ê |  |  | ê |  |  | ê |  |
| £ 数据可视化 |  | ê |  |  | ê |  |  | ê |  |
| £ 统计模型 |  |  |  |  |  |  |  |  |  |
| £  **JavaScript和HTML** |  |  |  |  |  |  |  |  |  |
| £ 数据操作 |  |  |  |  |  |  |  |  |  |
| £ 从外部文件导入数据 | ê |  |  |  | ê |  |  |  | ê |
| £ 连接数据库 | ê |  |  |  | ê |  |  |  | ê |
| £ 数据类型 | ê |  |  |  | ê |  |  |  | ê |
| £ 数据类型转换 | ê |  |  |  | ê |  |  |  | ê |
| £ 可视化编程 |  |  |  |  |  |  |  |  |  |
| £ 接CSS3可视化技术 | ê |  |  |  | ê |  |  |  | ê |
| £ CSS3转换和动画 | ê |  |  |  | ê |  |  |  | ê |
| £ SVG | ê |  |  |  | ê |  |  |  | ê |
| £ HTML5画板 | ê |  |  |  | ê |  |  |  | ê |
| £ JavaScript and jQuery编程 | ê |  |  |  | ê |  |  |  | ê |
| £ 条件、循环、控制执行 | ê |  |  |  | ê |  |  |  | ê |
| £ 自编函数 | ê |  |  |  | ê |  |  |  | ê |
| £ 可视化模块（D3.js，Node.js） | ê |  |  |  | ê |  |  |  | ê |
| £  **SQL** |  |  |  |  |  |  |  |  |  |
| £ 数据操作 | ê |  |  |  | ê |  |  | ê |  |
| £ 连接数据库 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据操作语言 (DML) | ê |  |  |  | ê |  |  | ê |  |
| £ 数据定义语言 (DDL) | ê |  |  |  | ê |  |  | ê |  |
| £ 数据类型转换 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据预处理 | ê |  |  |  | ê |  |  | ê |  |
| £ 数据抽样 | ê |  |  | ê |  |  | ê |  |  |
| £ 数学函数、统计函数 |  |  |  | ê |  |  | ê |  |  |
| £ 条件、循环、控制执行 |  |  |  | ê |  |  |  |  | ê |
| £ 宏 |  |  |  | ê |  |  |  |  | ê |

网络管理

作为数据分析师，对基本的代数知识应该有所了解，最起码不反感这一抽象的知识。在日常工作中，具体地说，数据分析师需要将业务问题转化为数学表达式，运算数学表达式或求解方程式。有时还需要利用一些高级的数学知识和抽象思维将不同类型的函数的功能区分开来，了解不同函数之间的关系，以及函数内部方程式的关系。

***注：该类目下《网络设计》知识点将在培训课程中提及，但不作为单独培训课程提供。***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理  工程师 | | | 大数据  工程师 | | | 大数据架构  工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
|   **网络设计** |  |  |  |  |  |  |  |  |  |
|  网络布线施工规范 |  |  |  |  |  |  |  |  |  |
|  网络硬件设备 |  |  |  |  |  |  |  |  |  |
|  应用系统的设计开发 |  |  |  |  |  |  |  |  |  |
|  数据通信 |  |  |  |  |  |  |  |  |  |
|  网络体系结构 |  |  |  |  |  |  |  |  |  |
|  局域网组网技术 |  |  |  |  |  |  |  |  |  |
|  TCP/IP网络协议 |  |  |  |  |  |  |  |  |  |
|  接入网与接入技术 |  |  |  |  |  |  |  |  |  |
|  网络安全技术 |  |  |  |  |  |  |  |  |  |
|  性能测试和优化技术 |  |  |  |  |  |  |  |  |  |
|  **云计算开发与运维** |  |  |  |  |  |  |  |  |  |
|  IAAS/PAAS/SAAS |  |  |  |  |  |  |  |  |  |
|  云计算网络搭建 |  |  |  |  |  |  |  |  |  |
|  云服务器初始配置 |  |  |  |  |  |  |  |  |  |
|  云存储设备初始配置 |  |  |  |  |  |  |  |  |  |
|  容量规划与实施 |  |  |  |  |  |  |  |  |  |
|  集群部署 |  |  |  |  |  |  |  |  |  |
|  数据中心容错 |  |  |  |  |  |  |  |  |  |
|  负载均衡和监控 |  |  |  |  |  |  |  |  |  |
|  虚拟技术 |  |  |  |  |  |  |  |  |  |
|  故障排除 |  |  |  |  |  |  |  |  |  |
|   **高性能计算与可扩容性** |  |  |  |  |  |  |  |  |  |
|  HPC并行、网格计算 |  |  |  |  |  |  |  |  |  |
|  伸缩性方案 |  |  |  |  |  |  |  |  |  |
|  评价指标 |  |  |  |  |  |  |  |  |  |
|   **咨询技能** |  |  |  |  |  |  |  |  |  |
|  需求分析 |  |  |  |  |  |  |  |  |  |
|  DevOps |  |  |  |  |  |  |  |  |  |
|  文档写作 |  |  |  |  |  |  |  |  |  |

数据治理（Data Governance）

对基础统计知识的理解对一个数据分析师是至关重要的。例如，业务经理可能会要求数据分析师运行一个A / B测试，了解统计知识将有助于解释经过处理后的信息和数据。有一些基本的统计概念是必须熟悉的，如统计测试，统计分布，最大似然估计等等。对统计知识的掌握，则会帮助模型分析师判断在一个特定的情况下，那一种技术才是最恰当的。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理  工程师 | | | 大数据  工程师 | | | 大数据架构  工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
| £ **数据库架构** |  |  |  |  |  |  |  |  |  |
| £ 数据库架构设计 |  |  |  |  | ê |  |  |  | ê |
| £ 数据质量控制框架 | ê |  |  |  | ê |  |  |  | ê |
| £ 元数据管理 | ê |  |  |  | ê |  |  |  | ê |
| £ 主数据管理（MDM） | ê |  |  |  | ê |  |  |  | ê |
| £ 数据库管理（DBA） | ê |  |  |  | ê |  |  |  | ê |
|  **数据摄入解决方案** |  |  |  |  |  |  |  |  |  |
|  ETL |  |  |  |  | ê |  |  |  | ê |
|  数据来源与获取 | ê |  |  |  | ê |  |  |  | ê |
|  数据提取 | ê |  |  |  | ê |  |  |  | ê |
|  数据融合 |  |  |  |  | ê |  |  |  | ê |
|  数据转换，浓缩 |  |  |  |  | ê |  |  |  | ê |
|  数据需求调查和发现 |  |  |  |  | ê |  |  |  | ê |
| £ **数据安全** |  |  |  |  |  |  |  |  |  |
| £ 身份认真和用户权限组 | ê |  |  |  | ê |  |  |  | ê |
| £ 保密性、可用性、完整性策略 | ê |  |  |  | ê |  |  |  | ê |
| £ 容灾备份与防攻击 | ê |  |  |  | ê |  |  |  | ê |
| £ 数据库加密 | ê |  |  |  | ê |  |  |  | ê |
| £ **信息安全** |  |  |  |  |  |  |  |  |  |
| £ 敏感数据管理策略 | ê |  |  |  | ê |  |  |  | ê |
| £ 数据传输加密 | ê |  |  |  | ê |  |  |  | ê |

数据分析

分类分析在大数据领域应用非常广泛。在营销、金融、电信、零售、健康等行业都有应用案例。数据分析师经常会被要求处理一些顾客行为分类的工作。根据业务的要求，对数据分析师掌握分类算法的程度也要求不同。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理工程师 | | | 大数据  工程师 | | | 大数据架构工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
| £  **数据项目质量控制** |  |  |  |  |  |  |  |  |  |
| £ 根据项目找到相关的数据源 | ê |  |  | ê |  |  |  | ê |  |
| £ 了解数据在数据库里所表达的意思 | ê |  |  | ê |  |  | ê |  |  |
| £ 评估数据的质量及其对项目的影响 | ê |  |  | ê |  |  |  | ê |  |
| £ 数据转换及清洗 | ê |  |  | ê |  |  | ê |  |  |
| £ 数据脱敏 | ê |  |  | ê |  |  |  | ê |  |
| £ 数据分析项目的质量控制 | ê |  |  | ê |  |  | ê |  |  |
| £  **数据分析编程** |  |  |  |  |  |  |  |  |  |
| £ 数据分析编程的流程（倒推法） | ê |  |  | ê |  |  | ê |  |  |
| £ 编程效率和程序运行效率 | ê |  |  | ê |  |  |  | ê |  |
| £ 基于阿里云技术的编程基础 | ê |  |  | ê |  |  | ê |  |  |
| £ 质量控制流程 | ê |  |  | ê |  |  |  | ê |  |
| £  **数据分析项目设计与执行** |  |  |  |  |  |  |  |  |  |
| £ 数据分析编程的流程（倒推法） | ê |  |  | ê |  |  |  | ê |  |
| £ 数据分析项目计划管理流程 | ê |  |  | ê |  |  | ê |  |  |
| £ 商业目标、背景调查、分析范围、分析方案 | ê |  |  | ê |  |  |  | ê |  |
| £ 项目前分析和项目绩效考评 | ê |  |  | ê |  |  | ê |  |  |
| £  **数据分析类型** |  |  |  |  |  |  |  |  |  |
| £ 报表 | ê |  |  | ê |  |  | ê |  |  |
| £ 顾客列表 | ê |  |  | ê |  |  | ê |  |  |
| £ 洞察 |  |  |  |  |  |  |  |  |  |
| £ 预测 |  |  |  |  |  |  |  |  |  |
| £ 优化 |  |  |  |  |  |  |  |  |  |
| £ 战略分析 |  |  |  |  |  |  |  |  |  |
| £  **行业经验（可选）** |  |  |  |  |  |  |  |  |  |
| £ 零售业 |  |  |  |  |  |  |  |  |  |
| £ 金融 |  |  |  |  |  |  |  |  |  |
| £ 电信 |  |  |  |  |  |  |  |  |  |
| £ 营销 |  |  |  | ê |  |  |  | ê |  |
| £ 广告 |  |  |  |  |  |  |  |  |  |
| £ 健康 |  |  |  |  |  |  |  |  |  |
| £ 文化传媒 |  |  |  |  |  |  |  |  |  |

分类分析（Classification）

分类分析在大数据领域应用非常广泛。在营销、金融、电信、零售、健康等行业都有应用案例。数据分析师经常会被要求处理一些顾客行为分类的工作。根据业务的要求，对数据分析师掌握分类算法的程度也要求不同。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理  工程师 | | | 大数据  工程师 | | | 大数据架构  工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
| £  **聚类分析（Clustering）** |  |  |  |  |  |  |  |  |  |
| £ 层次聚类（Hierarchical Clustering） | ê |  |  |  | ê |  |  |  | ê |
| £ K-Mean | ê |  |  |  | ê |  |  |  | ê |
| £ 高效聚类分析 |  |  |  |  |  |  |  | ê |  |
| £  **判别归类分析** |  |  |  |  |  |  |  |  |  |
| £ 贝叶斯分类器 |  |  |  |  |  |  |  |  | ê |
| £ K-Nearest Neighbors（KNN） |  |  |  |  |  |  | ê |  |  |
| £ 关联规则分析 | ê |  |  |  | ê |  |  | ê |  |
| £ 基于SQL分类算法 |  |  |  | ê |  |  | ê |  |  |
| £  **决策树** |  |  |  |  |  |  |  |  |  |
| £ 决策树 | ê |  |  |  | ê |  |  | ê |  |
| £ 随机森林 |  |  |  |  |  |  |  |  | ê |
| £  **机器学习** |  |  |  |  |  |  |  |  |  |
| £ 神经网络 |  |  |  |  |  |  |  | ê |  |
| £ 支持向量机 |  |  |  |  |  |  | ê |  |  |

数据沟通与可视化

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息。为了有效地传达思想概念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相当稀疏而又复杂的数据集的深入洞察。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 大数据助理  工程师 | | | 大数据  工程师 | | | 大数据架构  工程师 | | |
| 操作 | 应用 | 方案 | 操作 | 应用 | 方案 | 操作 | 应用 | 方案 |
| £  **数据准备** |  |  |  |  |  |  |  |  |  |
| £  KPI与其代表的意义 | ê |  |  | ê |  |  | ê |  |  |
| £  故事主线 | ê |  |  | ê |  |  | ê |  |  |
| £  DataV | ê |  |  | ê |  |  | ê |  |  |
| £  Excel | ê |  |  | ê |  |  | ê |  |  |
| £  BI报表 | ê |  |  | ê |  |  | ê |  |  |