数据合规人要懂的100个产品/技术常识

一、基础技术名词

1. 什么是数据发现?

一旦数据收集完成，下一步就是数据发现。数据发现是识别可用于数据分析和/或数据整合的数据集的过程。这一阶段对于评估数据质量至关重要，因为数据发现工具可以浏览数据或应用高级分析来监测模式和异常值。

帮助商务用户进行日常操作和业务决策这一目的，可以通过可视化分析来实现。数据准备这一关键阶段往往被忽视，然而在正确处理数据之前了解数据意味着数据在共享使用时会更加可靠。数据发现过程使用工具及专业知识，在数据专家的协助下、从收集到的数据中观测到模式或趋势。

数据发现可以分为两大类：

- 手工数据发现是传统的商业智能过程，由数据专家凭借他们渊博的应用案例知识、理解力和丰富经验，手工绘制数据图。这种方法完全依赖个人理解力，由数据专家构思并勾勒出数据图表，用以关联和理解数据。
- 智能数据发现是一种更现代的商业智能形式，使用机器学习的自动化过程来揭示数据价值、并提供高质量的商务见解。使用人工智能的优点是，耗时少，可以准备、构思、整合和共享相关联的数据。也可以编制数据可视化，呈现隐藏的模式和有价值的见解。

2. 什么是分级分类?

分类：更多是从业务角度出发，在企业理清数据家底后，明确知道哪些数据（其实应该是元数据，更贴切一些应该是字段）属于哪个业务范畴，也就是类别。这个业务范畴囊括的范围可大可小，完全依托于企业前期基于业务的梳理结果。举个例子：身份证号这一类数据，既可以属于个人信息范畴，也可以属于个人基本信息范畴，前者的范围明显大于后者。也许有朋友会发出疑问，给业务划分类别当然是越细越好。这就是笔者要在此处强调的，做数据分类，并不是业务越细分越好，因为很有可能细分业务之后，最终却发现无数据可进行归类，这是典型分类失败的体现。当然反过来也成立，分类少了，数据归不进去，也是分类失败的体现。

分级：不同于数据分类，对于大多数企业来说，更多是从满足监管要求的角度出发。数据分级属于数据安全领域，或许称呼它为敏感等级更为贴切。企业中的数据有的密级程度高、有的低、有的可公开、有的不可公开，敏感等级不同的数据对内使用时受到的保护策略不同，对外共享开放的程度也不同。如果企业对自己内部的数据没有一个明确地认识，先不说是否可以满足监管要求，对于自身的运营来说都是严重的隐患，因为很可能一不小心就将内部的敏感信息泄露了出去。

3. 什么是元数据?

元数据（Metadata）中的『元（Meta）』可以理解为事物或对象，『数据（data）』当然就是指该对象的相关数据。你可能接触过照片的元数据，其中包括图像尺寸、拍摄时间或者是光圈和快门信息、GPS数据，对于视频文件也一样，比如画面的尺寸、视频和音频的编码、时长等等。

实际上你可以理解成，关于该文件或对象的一切信息都是元数据，无论是技术相关的信息还是内容信息的一切。

技术型元数据

技术型元数据通常涵盖了从相机或摄像机获得的信息范围，这很自然，因为这些数据主要就是由其拍摄和生成的。其中除了包括前面提到的图像大小、帧速率、编码以外，还可能（取决于相机和来源）包括镜头型号、焦距、白平衡、相机硬件序号、镜头硬件序号等等。取决于摄像机型号的不同，一些基本数据会跟随数字媒体文件的生成被嵌入到媒体文件内部，而另一些可能会被单独存储在一个称为『Sidecar』的文件中，这通常是一个XML文件，而且带有硬件生产商的特定属性。

内容型元数据

这通常是更有用的元数据，因为它包括的范围更广。内容型元数据除了能用来描述媒体或片段的内容以外，还可以被用来对素材进行管理、分类、纳入上下游操作流程，甚至也能提供一些技术型元数据的功能。

基本上，你可以把内容型元数据直接理解成关键字，但它有除了可以是字符型内容以外，还可以是评分、勾选框等类型。

内定型元数据可以非常简洁，也可以非常冗长，但相同点在于，它们目前来说都是由工作人员在制作流程中手动添加和修改的——也许是摄像师在拍摄中添加的场号、镜号，也许是DIT人员添加的卷号、样片号、机位编号、景别附注、外观附注，也许是剪辑助理或VFX艺术家添加的各种注释等等。

4. 什么是数据映射（Data Mapping）？

给定两个数据模型，在模型之间建立起数据元素的对应关系，将这一过程称为数据映射。 数据映射是很多数据集成任务的第一步，例如：数据迁移（data migration）、数据清洗（data cleaning）、数据集成、语义网构造、p2p信息系统。

5. 什么是数据迁移？（data migration）

数据迁移是指将数据从一个位置转移到另一个位置，从一种格式转换为另一种格式，或从一个应用程序移动到另一个应用程序的过程。数据迁移通常是为数据引入新系统或位置的结果。业务驱动因素通常是应用程序迁移或整合，在这种迁移或整合中，原有系统会被共享同一数据集的新应用程序所取代或增强。如今，随着企业从内部基础架构和应用程序迁移到基于云的存储和应用程序以优化或转变公司，数据迁移即开始。

6. 什么是数据清洗？（data cleaning）

对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。数据清洗从名字上也看的出就是把"脏"的"洗掉"，指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。

7. 什么是数据治理？（Data Governance）

数据治理是指为确保数据安全、私有、准确、可用和易用所执行的所有操作。它包括人们必须采取的行动、必须遵循的流程以及在整个数据生命周期中为其提供支持的技术。数据治理意味着设置适用于收集、存储、处理和处置数据的内部标准，即数据策略。它规定了谁可以访问哪些数据以及哪些数据应受治理。数据治理还涉及遵循行业协会、政府机构和其他利益相关者设定的外部标准。

数据治理包含以下几方面内容

1、确保有效助力业务的决策机制和方向；

2、确保绩效和合规进行监督；

3、确保信息利益相关者的需要评估，以达成一致的企业目标，这些企业目标需要通过对信息资源的获取和管理实现。

8. 什么是机器学习?

机器学习 (ML) 是人工智能 (AI) 的一个分支，旨在构建能够根据所使用的数据进行学习或改进性能的系统。人工智能是一个宽泛的术语，指的是模仿人类智能的系统或机器。机器学习和人工智能这两个术语经常被相提并论，有时甚至互换使用，但它们的含义并不相同。其中一个重大区别是，所有的机器学习都是 AI，但不是所有的 AI 都是机器学习。

如今，机器学习无处不在。当我们与银行交互、在线购物或使用社交媒体时，机器学习算法会发挥作用，让我们获得高效、顺畅和安全的体验。目前，机器学习及其相关技术正迅速发展，对于它的强大功能，我们只是略知一二而已。

9. 什么是人工智能?

人工智能的范围可以说很大、很泛，从表面上可以理解为机器的智能化，让机器像人一样能解决思考解决问题。其实人工智能核心技术包括很多的方面：推理、知识、规划、学习、交流、感知、移动和操作物体的能力等。可以说机器学习和深度学习都是人工智能这个大主题下的一部分吧，深度学习又可以归为机器学习的一部分。简而言之，机器学习和深度学习是人工智能的两个关键的技能，看人工智能的发展历史，人工智能三大研究内容：计算机模仿人类的思考，对环境的感知和动作的实现是人工智能的三大研究内容。

即:人工智能>机器学习>深度学习

10. 什么是接口（API）？

我们去餐厅看着菜单点菜，点好菜后，服务员会根据你的菜单，给你上菜。其中点菜就是餐厅提供的一种服务，这个服务的输入是菜单名，输出就是做好的菜。

小结：所以说API就是给客户提供服务的一种方式，它还需要入参和出参。

再举几个我们工作中的常见例子：

- 例子1：微信开放平台给其他开发者提供了微信扫码登录的API，开发者只要调用这个API就可以实现扫码登录。这个API的入参是登记在微信开放平台的一个appid和密钥，出参则是用户的openid等信息。
- 例子2：腾讯云给其他开发者提供了发短信的API，开发者只要调用这个API就可以发短信。这个API的入参是用户的手机号码和短信内容，出参则是发短信。
- 例子3：这是最常见的例子。我们的后台会暴露很多API给到前端调用，也就是HTTP接口。比如说一个查询商品的接口，入参是商品名称，出参是商品详情。

11. 什么是SDK?

SDK全称是软件开发包，常见的比如百度地图SDK、微信支付SDK等。SDK是软件开发商封装自己的一些基础服务后，对外提供的一种软件开发工具包。目的在于省去第三方应用开发者的开发成本，使用现成的软件能力来服务于自己的产品。

- 例如：百度地图SDK，提供了完整的地图展示、导航、定位等功能。作为第三方开发者，只需要调用SDK里的接口来使用这些服务即可，不需要自己从头开始来开发这些功能，极大的降低了开发成本，而对于SDK厂商来说，扩展了自己的生态圈，也丰富了用户群。