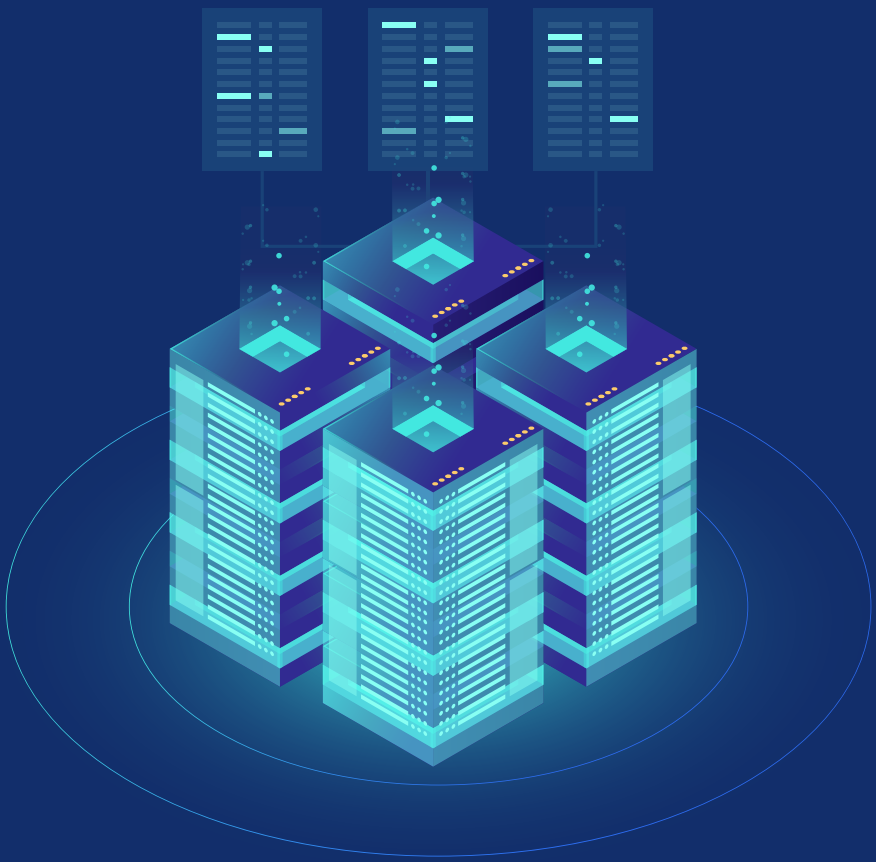
TalkingData

数据应用工程成熟度模型

Data Application Engineering Maturity Model



目录

01 引言

背景

在现代社会，随着企业的发展产生了大量的数据，生产部门有生产制造的数据记录，业务运营部门有营销数据，财务部门有经营数据，数据无处不在，数据又时时刻刻影响着企业运转中每个环节的决策。数据已经成为除了资金和人才以外企业新的资产价值增长点。

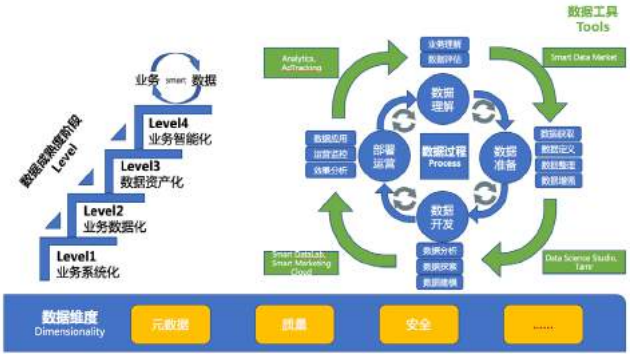
数据本身并不代表价值，数据仅仅是以一定格式对事实进行记录，是原始材料；只有结合环境和上下文的数据才有意义，这就是信息；伴随着信息的积累，我们从趋势和关系的挖掘中总结出了规律，这些规律就变成了知识；然后依据知识在企业经营中进行决策和行动，能进一步促进企业的良性循环。

数据产生价值的过程需要经历获取、存储、评估、整理、增强、分析、应用等多个环节，在小数据时代这些过程都相对简单和成熟。随着近些年数据收集方式的增多、传感设备数量的增加，计算能力的增强和存储方式的改进，导致了人们可感知的数据量急剧增多；按照摩尔定律，数据生成和存储的生长速度一直在呈现指数增长。大数据应运而生，带来了俗称的大数据 4V 特征：数量多（即数据集的规模）、多样性（即来自多种数据仓库、领域或类型的数据）、速度快（数据的流速）、多变性（在不同特征里的变化）。大数据的到来，使得在数据系统的演化进程中，人们对于高经济效益以及高效率的数据分析需求迫使现有技术不断变化。

伴随着大数据革命，必须考虑如下四个方面的相互作用：数据集的特征、对数据集的分析、数据处理系统的性能以及对经济效益的商业考虑。这些决定了数据应用的价值效果。通过不断的实践，我们总结出了当前大数据环境下的数据应用工程 - 成熟度（LPDT）模型。

模型概述

数据应用工程 - 成熟度（LPDT）模型（以下简称“成熟度模型”）主要针对大数据环境下的数据应用工程提供方法论依据。可以用来指导企业评估自身所处的数据应用成熟度阶段，也可以用来指导企业如何晋级到更高阶的成熟度阶段。成熟度模型分为成熟度阶段（Level）、过程（Process）、维度（Dimensionality）和工具（Tool）四个方面展开。



数据应用工程 成熟度模型（LPDT）

成熟度阶段（Level）分为 Level1 业务系统化、Level2 业务数据化、Level3 数据资产化、Level4 业务智能化共四个阶段；除此以外，还有一个更高阶的隐藏阶段为 Level5 企业智能化。成熟度阶段（Level）主要代表了企业在业务运转中应用数据能力的高低，可以通过过程（Process）、维度（Dimensionality）和工具（Tool）等多个维度去评估。本成熟度模型不设置详细的打分机制，只提供部分阶段的特征供企业自评估参考。

数据应用过程（Process）分为数据理解（Understand）、数据准备（Prepare）、数据开发（Develop）、部署运营（Operation）四个阶段，基本涵盖了所有数据应用过程，其中每个过程还会细分子过程、入输出及操作项，这些会在后续的章节详细阐述。数据应用过程可以理解为一个数据应用的最小迭代原型，也可以理解为一个大项目或企业的整体数据应用，其中过程与过程之间也可能发生小的迭代和回溯。使用时应在抽象理解的基础上与企业的实际情况映射。

数据维度（Dimensionality）是指贯穿于数据应用全过程的一些数据领域维度，是数据应用过程中必须考虑的方面，当前我们只考虑“元数据”、“质量”和“安全”三个维度展开，由于每个维度单独展开都是一个很大的话题，本文档中只结合数据应用过程有限地展开阐述。数据应用过程中还有很多其他维度本次暂不涉及，也欢迎各位使用者反馈。

数据工具（Tool）是指结合数据应用过程和数据维度各个环节会用到的工具，可能是开源的，可能是定向开发的，可能是 SaaS 的，可能是私有化部署，可能是免费的，可能是付费的。数据工具与过程和维度是相辅相成的关系，三者一起为成熟度阶段提供评估依据。

适用对象

数据应用工程 - 成熟度模型可适用于如下场景：

企业管理决策层（CEO、CIO 等）可以参考该模型评估企业数据应用的阶段，进行业务相关的数据战略决策，进一步规划数据在业务中的应用思路。

业务部门可以使用该模型优化业务流程，参考该模型系统进行数据应用，挖掘数据价值，提高效率，提升业务效果。

数据部门可以使用该模型更系统地建立数据管理和应用的流程机制，为更多的业务部门提供数据应用支撑能力。

02
成熟度模型

01 业务系统化
02 业务数据化
03 数据资产化
04 业务智能化
05 成熟度进阶

数据应用工程 - 成熟度模型（LPDT）主要从数据管理和应用的角度来衡量企业应用数据的能力，并将其分为以下多个成熟度阶段。针对不同的阶段，从企业管理、数据应用过程、数据维度、技术 / 工具等多个方面不同特征进行参照判定。

业务系统化

定义

业务系统化阶段是指：企业的业务流程清晰，且业务过程都已经通过 IT 系统实现，IT 系统的实现以业务为导向，可能有少量数据记录，但并没有以数据为导向积累数据。

特征

“业务系统化”阶段主要有以下特征：

企业管理：该阶段的企业战略以纯业务角度驱动；整个公司无数据意识，业务实施过程中无数据积累及数据优化业务的理念；企业的组织架构中无数据相关部门和职位的设置。

数据应用过程：该阶段的企业只是使用业务系统中必备的数字进行业务和财务的统计管理和分析。尚未开始理解业务链条背后各个环节的数据，也没有考虑使用技术工具进行数据积累。每次基于业务目标的数据统计都需要定制化开发处理。

数据维度：该阶段的元数据只涉及业务元数据，可能只在业务系统中使用，但并未统一所有的元数据术语，各业务线的业务单元分散管理。质量方面，可能会有一些测试和质检，但是并未从质量保证和质量控制角度设计质量管理指标和质量评价体系。数据安全层面只界定了财务数据，尚未对数据的分等定级和数据安全保密级别进行设计和划分。

技术 / 工具：开始使用平台 / 系统管理部分业务或整个业务线，但业务系统间并未打通和串联，各业务系统无数据沉淀，业务系统背后的数据未被收集或处于散乱无序的未管理状态。

业务数据化

定义

业务数据化是指企业在业务系统化的基础上开始建立数据理念，开始基于单业务各个环节进行数据的收集、管理、分析，并反馈优化该业务，数据体量相对单一，可能有业务的 BI 报表进行闭环的业务分析和迭代。该阶段是基于业务目标去收集数据和分析数据。

特征

“业务数据化”阶段主要有以下特征：

企业管理：该阶段的企业开始建立数据的理念，在业务过程中注重数据的积累，战略上开始考虑通过数据来分析和解决业务问题；组织架构中有数据相关的部门和数据分析师等相关的职位来支撑。

数据应用过程：该阶段开始考虑在业务系统中设置功能进行数据收集，有专门的团队对收集的数据进行管理和分析，挖掘数据对业务的优化；有数仓进行数据管理，有系统的 dashboard 等工具向决策层数字化反馈业务情况。各产品线和各个环节的数据孤立进行，管理和分析的数据主要是小体量的指标数据，很少涉及大量底层日志数据。

数据维度：该阶段有专门的系统平台功能进行元数据管理，统一了术语，该阶段分析处理的数据主要以结构化数据为主，可能有少量非结构化数据。质量方面开始设计质量监控指标，实施过程质量控制，建立对应的质量保证体系，并且实现平台化管理监控。数据安全维度对数据做了基本的分等定级，明确了涉密与非涉密的划分。

技术 / 工具：研发部门针对数据收集管理建立了专门的数据仓库或类似的技术架构进行业务线的数据沉淀，可以在系统上针对已存储的数据进行 ETL 处理和挖掘分析。

数据资产化

定义

数据资产化是指企业在业务数据化的基础上，考虑将数据作为资产去挖掘其价值。在该阶段会将所有的数据汇聚管理起来，实现不同的数据联通，跨界考虑数据价值的应用。且该阶段处理的数据维度更多源化，数据体量更大，不仅要处理内部数据，还需要考虑基于业务场景，如何与外部数据对接连通。该阶段是先收集数据，再从海量数据中去挖掘可能的价值。

特征

“数据资产化”阶段主要有以下特征：

企业管理：该阶段企业战略层面已经将数据作为企业的资产，将其与资金和人才一起同等考虑。设立了专门的数据部门来管理企业生态内外的所有数据汇集、管理、分享。组织架构方面，在管理层设置了数据管理委员会或者首席数据官（CDO）来负责决策层的数据管理，职位方面除了有数据分析师以外，还设置了数据科学家等角色基于数据、统计和行业知识的综合高阶角色。

数据应用过程：该阶段开始考虑将公司内部所有可数据化的环节数据化，将各个业务产品线的底层日志到业务指标多个维度数据收集、融合、打通，并统一管理，数据准备阶段还会考虑从外部获取合作或购买的数据进行数据增强 / 放大，

在数据开发过程中更多的使用算法技术进行数据价值挖掘和业务优化迭代。

数据维度：该阶段处理数据维度广，数据体量大，数据结构复杂。元数据管理要同时考虑结构化和非结构化数据。质量方面更多的面对非标准化数据、非结构化数据的质量问题，要考虑面对未知领域数据快速确定质量情况的能力。数据安全维度要更多的考虑同一类数据在不同场景过程中的安全保密级别，其中要充分考虑数据连通后的隐私被挖掘的可能性。鼓励数据不动，算法向数据靠拢，计算出合规结果后输出的模式。

技术 / 工具：有专门的技术团队基于 hadoop 等开源系统开发大数据收集、ETL 处理的工具，有数据目录等数据资产展现的工具，有沙箱环境等数据探索的平台和能力。可以处理大量非结构化数据；基于业务场景可实现实时处理或离线处理。

业务智能化

定义

业务智能化是在数据资产化的基础上，结合企业内外部的全域数据进行分析挖掘，使用 ML（机器学习）和 AI（人工智能）等技术自动化地处理数据，优化迭代业务。

特征

“业务智能化”阶段主要有以下特征：

企业管理：该阶段企业战略层面开始更多的关注 AI（人工智能）在企业业务中的应用，让更多数据自动化的优化业务。组织架构方面会建立专门的数据科学部门或研究院，考虑企业未来的业务，可能会有未来架构师等职位角色出现。

数据应用过程：该阶段数据理解、数据准备（含收集）、数据开发（含探索应用）、运营部署等各个环节都考虑如何将AI融入其中，提高每个环节的效率和效果。更自动智能的处理、分析和应用数据。

数据维度：该阶段元数据管理范围更大，需要考虑企业内包含业务、财务、人力等全域数据的元数据管理。数据质量在数据资产化的基础上，更多地考虑基于 AI 的数据模型的质量如何度量和管理。数据安全角度，更多地考虑区块链等先进技术在安全领域的应用。

技术 / 工具：在企业业务的大数据场景下，基于人工智能算法的数据探索平台，数据智能业务模型的挖掘成为主要的方向。更多地考虑在原有技术平台处理数据的环节中 AI 能力的集成和应用。

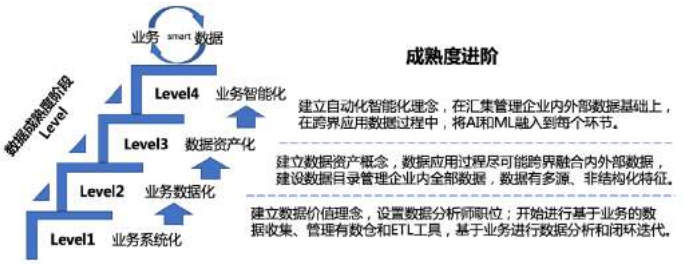
成熟度进阶

基于每个成熟度阶段，整理特征概要如下：

成熟度级别		1	2	3	4
成熟度阶段		业务系统化	业务数据化	数据资产化	业务智能化
战略组织和人员	战略 / 理念	企业无数据略，纯业务驱动	企业树立数据价值理念，开始注重数据积累和应用	企业开始将数据作为一种战略资产考虑，关注数据的连通性	企业有自动化智能化理念
	组织架构	企业无数据相关部门和职位设置	企业有数据部门或者职位中有数据分析师	公司决策层有 CDO 或数据管理委员会、职位有数据科学家	企业中有数据或 AI 研究院，有大量数据科学家，开始出现未来架构师

数据工程应用过程	数据理解	只考虑业务相关的数据指标	开始挖掘业务背后数据并评估收集可能性	开始评估所有业务环节和终端的数据采集与连接	开始评估业务、财务、人力等企业内全域数据的收集管理
	数据准备	业务未收集数据，无数据积累	有团队进行数据的收集和处理	数据准备可以通过外部数据源来增强数据能力	引入 AI（人工智能）和 ML（机器学习）在数据获取、整理和增强中的使用
	数据开发	基于业务的指标需要单独开发系统	有基于业务的报表系统或 dashboard 将分析结果展示	数据开发探索融合内外部数据基础上使用不同算法能力	引入 AI（人工智能）和 ML（机器学习）在数据探索分析中使用
	运营部署	只有单纯的业务指标分析	部署系统实现单业务线的数据闭环分析	数据应用更多考虑跨界数据的使用	引入 AI（人工智能）和 ML（机器学习）在数据应用和监控使用
数据应用基础能力	元数据管理	业务元数据分散且不统一	以结构化数据管理为主，有系统管理	数据管理中出现大量非结构化数据	考虑全域数据的元数据管理
	数据质量	有简单测试和质检，无质量体系	开始设计质量监控指标控制过程和结果质量	数据质量更多面对多维度、大体量、非结构化的质量度量方法	有 AI 和 ML 在质量管理中的应用
	数据安全	无数据分等定级和安全级别划分	对数据做了基本的分等定级区分涉密级别	需要考虑数据联通的安全隐私保护问题	更多考虑 AI 和区块链在数据安全领域的应用
数据应用技术工具	技术 / 系统 / 工具	有基于业务线的 IT 系统，但各条业务系统相对孤立	有专门的团队负责数仓和 ETL 并开发相关工具	开始构建数据资产目录管理企业所有数据；有沙箱环境	AI 模型优化原有系统各个技术环节，同时建立 AI 的模型库

基于每个成熟度阶段，企业要想进阶升级，可以从以下角度考虑。

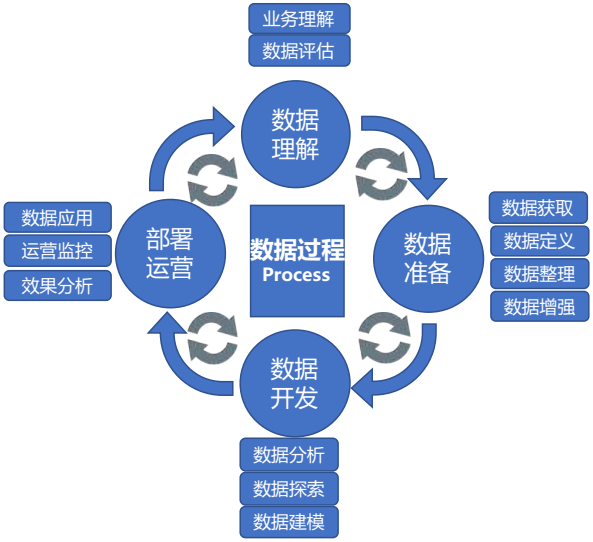


成熟度阶段 Level——进阶路线

03
数据应用过程

01 数据理解
02 数据准备
03 数据开发
04 部署运营

数据应用过程（Process）是数据应用工程 - 成熟度模型的核心组成部分，其阐述了一般数据应用经历的完整过程。数据应用过程从工程思维的角度将数据从收到用的过程进行梳理定义，包含数据理解（Understand）、数据准备（Prepare）、数据开发（Develop）、部署运营（operation）四个大的过程，每个过程还有细分子过程。数据应用过程可以理解为一个数据应用的最小迭代原型，也可以理解为一个大项目或企业的整体数据应用。



数据理解

概述

数据理解是指充分理解企业业务和数据，在此基础上定义数据要解决的业务问题，并评估其关联关系和可行性的过程。很多数据应用案例中曾出现无法正确辨别实际业务问题，而导致数据和业务之间的价值链断裂的情况。为了避免这类问题同时减少时间和资源的浪费，在工程开始之前就要清晰的识别出业务和数据之间的关联关系，明确范围、职责和最终目标。这就是数据理解环节要解决的问题。数据理解包含业务理解和数据评估两个子过程。

业务理解

业务理解是从商业角度全面理解客户想要达到的目标或者要解决的问题，划定业务目标和问题范围。

【任务项】：

了解业务：了解公司业务发展、企业或需求部门提出的数据应用场景、需求背景、要解决的问题和商业机会。

梳理需求：分析干系人的需要以明确需求，通常企业中不同部门、不同职级人员的需

求不同，应分别访谈并记录，以备后续步骤讨论使用。

明确目标：比较企业内部不同人员对数据应用结果的预期，数据分析师应协助企业需求方梳理预期达成的目标。如果有多个目标，可以将其分排优先级，规划在项目的各期逐步实现。

量化标准：与业务人员、项目干系人共同讨论本次数据应用工程结果的评判标准，在商业层面确定成功、失败的度量方式。如果项目周期较长，建议在项目执行中的重要环节设置验收标准。

【成果】：

业务调研报告：包括行业发展趋势、企业现状、业务现状 / 流程、业务问题等。

需求文档: 包含业务需求提出人、要解决明确的业务问题（需求目标）、预期结果是什么、计划在项目的什么阶段完成并交付、最后的评判标准。

数据评估

数据评估是从工程角度评估可行性，评估中需要包含数据的可获取性、技术的可行性、业务可行性（即是否能真正解决业务问题）、资源评估（人员和设备）、成本分析、风险分析等。

【任务项】：

数据可行性评估：梳理企业内部所有的数据源，数据内容（schema），数据质量情况，数据间的血缘联通关系，历史数据的使用情况，数据存储位置，数据是否都可以获取到，需要什么流程获取。企业外部可以获取什么数据资源，什么时候能获取到，详细内容和质量如何。基于所有可能获取的数据中哪些与目前项目有关联，数据是否需要再处理。主要从以上角度对企业内外部的数据情况及其可用性进行评估。

技术可行性评估：了解企业内部数据存储的技术环境，数据处理 / 分析 / 探索使用的平台系统及程序语言，基于不同环境下的数据转移或打通的技术可行性。

资源与风险评估：基于项目目标和数据现状评估整个项目的人力资源、存储 / 计算等软硬件资源需求。与需求部门和数据部门一起明确项目的约束、限制和风险，重点确认前期数据准备复杂度、中期数据分析颗粒度、后期模型 / 产品交付过程中各个环节风险点及备用方案。

工作项分解：在数据应用目标明确的前提下，结合各项评估结果，将项目分拆各个子目标，并制定对应的工作计划安排。

【成果】：

可行性分析报告：包含数据可行性评估结果、技术可行性评估结果、资源与风险评估结果、基于数据应用后业务问题的预期效果。

项目实施方案：包含但不限于以下内容，基于可行性分析结果设计的数据应用项目方案，各个环节的分拆实施方案，资源和计划安排，验收方案（包含效果和质量等）。

关键点 & 难点

“数据理解”过程的关键点主要集中在“明确目标”和“数据可行性评估”两个环节。

解决问题前先要明确问题是什么，很多数据应用最终效果不佳都在于初始的问题定义就不清晰，比如说“希望通过数据应用在营销环节提升效果”就是一个不明确的目标，需要明确到“将营销中的哪个指标提高到什么程度”，所以“明确目标”变的至关重要，我们要用 SMART 原则清晰地量化目标。

“数据可行性评估”十分重要，是因为只有了解了数据的现状，才能使项目方案在落地时顺利执行。如果只知道有数据，到了使用数据时才发现数据不可获取，或者数据的质量无法支撑项目使用，或者数据量匹配率极低，这样会导致项目执行中造成极大的风险，甚至无法完成预计目标。所以需要在数据理解阶段做好充分的“数据可行性评估”。

“数据理解”过程中还可能会出现如下一些难点也是要重点关注的：首先通常企业内部会认为此类项目是 IT 部门的职责，实际上 IT 部门只是技术、设备的保管和协助部门，数据应用工程应该由明确的数据治理或数据分析部门来管理，由多个业务和职能部门协助完成。其次，在考虑约束和风险的时候，除了时间和资源风险，还需要重点考虑数据安全以及相关法律法规。

数据准备

概述

数据准备是从各种数据源处获取原始数据，按照预期的业务需求定义数据应用的目标数据，将所有原始数据抽取、清洗、融合、转换、处理成为预期待分析挖掘的目标数据的过程。数据准备是所有数据应用工程都必须经历的过程，可以理解为是为数据分析或建模准备数据集的过程，我们将数据准备划分为数据获取、数据定义、数据整理、数据增强四个子过程，这四个子过程并不一定都是必须的，也并不要求有强顺序关系。

数据获取

数据获取是指用系统的方法，收集和测量各种来源的信息，以获得完整、准确的数据内容。获取的数据可以是结构化的，也可以是非结构化，可以是数字、文字、语音、图表等。在传统数据阶段，是需要什么数据才去设计收集的方法和指标；而大数据阶段正好相反，是先将所有的数据（比如日志数据）收集起来，再想办法从中抽取或挖掘想要的数据内容。

【任务项】：

数据源确认：一般获取数据的渠道有内部业务系统产生/收集，公开网络获取、外部购买、合作交换等。从不同数据源获取数据时的关注点不同，如果是内部业务系统，更多的是原始日志数据，可能很多是 rawdata，重点关注数据抽取和质量问题；如果是公开网络获取数据，一般都是整理好的数据或指标，需要重点关注其元数据和更新时间；如果是购买或交换的数据，最好是能让提供方给出数据说明或质量报告等。

数据内容分类标注：由于数据源不同，内容也差异很大，可能很多不同类型的内容是

混在一起的，需要将其按照主题域分类标注。从内容来看可以分为企业内部数据、宏观与行业数据、交互数据、检测数据和自然数据。从生成方式看可以分为人、机器和自然三个方面。从数据形式看可以分为符号、文字、数字、图像、语音等。

获取质量及稳定性确认：一般获取的数据在后期业务中是否能被应用起来，“质量及稳定性”至关重要，分析获取数据的质量可以采用先抽样评估，然后按照质量评价维度的各种指标的方式进行。数据获取频率要明确是一次性数据获取，还是周期性固定频率，还是以 API 接口的方式持续不定期的实时获取，不同获取频率需要关注的技术指标不同。

【成果】：

数据源目录: 记录了企业中可获取的所有数据源、负责人、获取的数据集内容（schema）、量级、格式、类型、频率、质量、历史使用记录等，可以通过元数据管理系统体现。

数据质量监控 / 评估系统：若是企业内部日志数据，需要有专门的监控系统来监控收集数据的各项质量指标；若是外部购买或合作交换的数据，需要对收到的数据进行质量评估后报告。

数据定义

数据定义是指在获取的数据源基础上，按照数据应用的业务目标定义目标数据集，并设计数据处理流程方案的过程。系统性来看，可以理解为定义数据处理流程、定义数据处理流程中每个环节的数据集，定义每个环节数据集的数据字段。该过程是数据准备过程的核心，相当于设计的过程，后续的工作都是在此基础上展开的。

【任务项】：

数据流程设计：是基于数据评估过程中的工作分解结果，确定数据应用的目标数据集，设计如何从不同的数据源中抽取需要的数据，进行必要的转换和清洗，生成目标数据集的处理路径，具体的处理方法。该工作相当于针对原始数据的重新组织，或者说制定数据处理的路线图。

数据集定义：定义明确的目标数据集和中间数据集，并清晰定义出每个数据集中的数据字段及其约束条件，数据的血缘关系图（即上游来源数据集和下游使用数据集）。关于目标数据集的定义，一般从“人”和“业务”两个维度来考虑，比如零售企业一般的业务从“商品”、“客群”、“交易”等三个大方面组织数据。

【成果】：

数据处理方案：一般包含了整体数据处理的流程，每个环节数据集的数据处理方法，数据监控的指标，数据集的质量评估方案，是后续数据整理和增强的指南。

数据资产目录：是一个元数据管理系统或者元数据管理表，其中包含了数据集定义任务项中的所有内容。

数据整理

数据整理是按照数据定义的处理方案，对数据源进行抽取、清洗、转换等加工处理，使之系统化、条理化，重新组织生成目标数据集的过程。数据整理是为了更方便进行后续的统计分析和数据挖掘，同时为数据消费者提供统一数据视图的集成方式，也可以理解为传统的 ETL 过程。

【任务项】：

数据抽取：将各渠道数据源获取的数据装载至适当的数据处理工具中，挑选适合本项目的数据。选择的标准包括与数据工程目标的相关性、质量和技术限制。数据抽取可以是选择特定的数据集、数据字段、特定属性范围内的数据记录（如一段时间、一个区域、一类商品）。可将抽取的数据装载到新的存储位置，构建新的数据集，为下一步做好准备。

数据清洗：即数据标准化的过程，抽样分析已有数据的质量，制定数据有效性策略和数据清洗策略，如针对特定字段格式的清洗、数据格式统一等。清洗后抽样验证数据质量是否达到既定标准、是否符合当地法规的数据安全标准。

【成果】：

数据集：清洗完成的标准化的新数据集

数据资产目录：对新生成的数据集的名称、schema、格式、血缘关系、质量情况等各种数据集信息。

数据操作文档：记录了数据处理的每一步操作策略，处理的代码，处理的效果及质量情况、处理过程中发生的问题记录等。这些信息可以在不同的系统或文档中体现。

数据增强

数据增强是将标准化后的多个单来源数据，以某一个字段进行数据联通，将联通的不同数据融合在一起生成新的字段更全的数据集，或者将通过联通后的数据交叉映射补充空缺字段内容。数据增强是多个数据集之间运算生成新的数据集的过程。

【任务项】：

数据联通匹配：是指针对不同的数据集，相同数据字段进行交集运算，确定不同数据集之间匹配率的过程，匹配率的高低决定了后续数据融合或者数据分析应用的最终效果。数据联通匹配也是很多企业与企业之间进行数据合作或进行数据业务开展的关键指标。比如：某电商想通过手机号匹配运营商数据确定其客户在不同区域的分布量及其通信消费情况，前提是该电商数据与运营商数据的联通匹配率高，其结果才能有价值。

数据融合：不同的数据集有相同的字段，以相同的字段将多个数据集从记录角度或从字段角度进行并集或交集运算的过程。在融合以前需要先将数据清洗标准化，并确定融合字段的共有性。

【成果】：

数据集：数据增强后生成的新数据集

数据资产目录：对新生成的数据集的名称、schema、格式、血缘关系、质量情况等各种数据集信息。

数据增强文档：记录了数据增强的策略，各数据源的匹配率，增强的效果，记录了数据增强的每一步操作，处理的代码，处理过程中发生的问题记录等。这些信息可以在不同的系统或文档中体现。

关键点与难点

数据准备是整个数据应用过程中最耗时的阶段，一般会消耗 80% 的时间或者资源，也是最苦最累最需要细心的过程。数据准备的每个环节都对最终的结果有很大的影响，其中有很多关键点需要重视：

数据源的稳定性和数据鲜度（时效性）

数据集定义中的处理策略和约束条件

数据处理过程中每个环节的质量监控

数据联通匹配率的高低（漏斗有效性）

数据准备过程中的难点主要集中在技术系统和数据孤岛两个问题上。数据整理和增强都会涉及到技术工具，海量数据对技术要求很高，各技术系统之间的适配和耦合能力至关重要，对该环节工作人员的技术要求也很高。在大公司数据可能被不同的生产线 / 部门拥有，要想使用数据，必须将数据汇聚或者开放数据访问的权限，统一数据过程就需要拆掉部门墙，这是很多生态型企业的难点，需要高层战略层面强力支持。

数据开发

概述

数据开发是以工程思维的角度将数据应用的关键实施过程进行演绎，是指在目标明确、数据集已整理完备的基础上进行分析挖掘探索数据应用模式的过程。数据开发是数据在业务领域最基本的模型探索，可以认为是小范围抽样实验的过程，并不包含在业务生产线上的工程化。数据开发包含数据分析、数据探索和数据建模三个子过程，三个子过程并不都是必须的，也并不具有强关联或顺序关系。

数据分析

数据分析是指在一定的商业场景或模型下，使用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息，形成结论，并对数据加以详细研究和概括总结的过程。数据分析更多的偏重于问题的答案是封闭式的一些业务场景，面对的主要是结构化数据。

【任务项】：

分析方法： 有两个维度的选择，

一个维度是很多业务领域有自己的场景和业务模型，比如：3A3R 模型、客群生命周期模型、流失预警模型、价值分级模型，只需要按照对应的模型指标，分析其数据分布、质量情况，以及与模型的符合度等；

第二个维度是数据统计方法的选择，包含描述性统计、回归分析、方差分析、假设检验、相关分析、聚类分析、因子分析、主成分分析等。这些都取决于分析的业务模型要解决的问题，随后的模型建立阶段也会包含其中部分内容，但更多是从算法角度深入切入，当前数据分析阶段主要使用的是描述统计、相关分析和方差分析等简单分析方法居多。

工具使用：一般情况需要依据数据量、维度，以及需要使用的分析方法和程序语言判断哪一款工具更适合分析场景的使用，同时也需要考虑程序复杂度与对使用者的知识和技能要求。

分析可视化：选定了分析方法和使用的工具基础上，调用准备好的数据执行分析的过程，将分析结果以图表等可视化形式展现。

效果验证：所有的分析结果都需要进行效果验证，一般可以与预期目标进行验证，可以使用历史数据进行验证，也可以使用真实结果进行验证，进一步评估分析的效果。

【成果】：

分析报告：需要包含选择的分析方法、分析工具、分析的过程、可视化分析成果、分析结论、效果验证方法及结果。

数据探索

数据探索：与数据应用相似，有两点不同，一是数据探索面对的以半结构化和非结构化数据居多；二是数据探索更多的是从业务场景中解决开放性的问题，可能探索没有明确的对错，可能探索没有明确定量定性的目标而只有一个方向或要解决的问题。

【任务项】同“数据分析”。但更多的强调业务逻辑上的创新。

【成果】同“数据分析”。

数据建模

数据建模：是指在准备好的数据集基础上，基于业务要解决的问题，设定假设，特征提取，使用算法构建模型，并迭代验证的过程。

此处的数据建模是狭义的建模，可以理解为我们经常说的数据科学。从经验来看一般的数据科学问题可以分为 5 大类，分别是：分类、异常检测、回归预测、聚类和强化学习。分析方法和工具也与这 5 类相对应，例如常见的预测类问题，当获得的数据可靠时，依靠数据做出决策会变得简单；其次面对一个全新领域，需要找出领域内实体间关系，则聚类分析和关联分析可以使数据之间的关系清晰化；偏差分析法可以为质量管理和异常检测提供分析的理论基础。实际操作时参与人员需要根据问题有针对性的选择。

【任务项】：

提出假设：包含业务假设和算法假设（选择）两个方面，业务假设是指基于业务要解

决的问题设置预期的假设目标，可以是枚举或是非，可以是趋势方向，可以是指标，可以是规律；算法假设是指解决该业务问题预期使用的算法或方法，可以是分类、异常检测、回归、聚类和强化学习等任何一种。

特征工程：所有的数据集都可以从描述性角度确认其特征，依据要解决的业务问题和选择的算法，从数据集中抽取需要的特征数据，并按照算法要求进行格式化处理的过程。

模型训练：将准备好的特征数据集带入算法，进行运算训练，并不断调整算法参数和特征数据进行迭代，直至训练达到预期效果，或者在现有算法和特征数据下的最优结果。若训练中一直未达到预期效果，可以调整算法或者特征数据。

模型评估：一般在进行模型训练时，会将特征数据分为两份或者多份，将其定义为训练集和验证集，在训练集上进行模型训练达到预期效果的情况下，使用验证集数据进行模型的评估，从召回率和正确率等多个角度综合评估。

【成果】：

模型设计方案：包含假设提出，数据集的选择，ground truth 的准备，特征工程的过程（包含特征描述、特征分析等），算法描述，训练结果，部分可能还涉及升维或降维等。

模型代码：可以存储在 git 或企业内部自有平台上。

模型评估报告：包含模型评估的方法及其效果。

关键点与难点

整个数据开发的关键点在于建模的过程，无论特征选择还是算法选择都不是一蹴而就的，都是一个不断试错和迭代的过程，这其中如果维度不够或者过多还会涉及到升维或者降维的过程。

难点在于假设的提出，要有充分的业务经验积累才能针对业务问题提出好的假设，所以开发的过程一定要有业务专家的参与才行。

部署运营

概述

部署运营是指将开发阶段模型成果在业务线上部署，在生产环境中例行化，并跟踪其运行效果的过程。一个算法模型在小量数据的实验阶段对系统和性能的压力都比较小，但是一旦将其部署在真正的业务线上，对存储和计算资源等算力的要求就会很大，对性能要求就很苛刻。运营部署过程可以分为数据应用、运营监控和效果分析三个子过程。

数据应用

数据应用是指将算法模型在生产环境下部署的过程。

【任务项】：

模型部署评估：主要是根据业务情况和模型要求，对正式业务环境下的数据存储、数据计算、性能指标及业务线影响进行综合评估。并制定正式部署方案和容灾备份方案。

模型部署操作：按照模型评估的结果进行部署操作。

【成果】：

部署评估报告

业务系统模型部署

运营监控

运营监控是指在业务系统部署模型后设置业务或质量监控指标，持续监控业务运营效果。

【任务项】：

指标设定：基于业务运营效果设置监控指标，可以是以往已有的监控指标，也可以是基于新的模式新增的指标，其中包含业务指标和性能指标。

指标监控：针对设置指标设置监控周期，在周期内进行对应指标的监控并记录。前期指标监控可以是人工参与，但当确定后，后期需要将其在监控系统自动化。

【成果】：

指标定义说明

指标监控结果：可以在某个监控系统上，也可以在例行报告或邮件中。

效果分析

效果分析是基于运营监控效果的数据，与数据理解阶段设置的数据目标和业务目标进行闭环分析。确认整个数据应用工程效果的效果。

【任务项】：

数据效果分析：基于“运营监控”中的数据指标进行模型应用前后的历史对比分析，同时与预期目标进行对标分析，在分析过程中系统性的评审“数据理解”、“数据准备”、“数据开发”和“部署运营”整个大闭环阶段所有数据处理环节的效果。

业务效果分析：是指将业务真实运行情况做预期目标对标分析，同时可以考虑与竞品的对标分析，对此次数据应用效果进行评价，同时基于现状提出将来的方向和目标。

【成果】：

数据分析报告

业务分析报告

关键点与难点

部署运营阶段的关键点在于要数据分析与业务分析紧密结合来说明数据应用的效果，且要

将整体“数据理解”、“数据准备”、“数据开发”和“部署运营”每个过程中的所有数据效果进行系统性的分析，而不只是最终效果的历史对标。

部署运营阶段的一个难点在于对于某些业务场景调整后的效果是受到多种因素的影响，而非仅仅是数据的影响，这种情况下应该尽量减少其他变量的调整，另外从变化因子的影响程度进行评估。另外一个难点在于真实效果数据的获取，这个要在业务理解阶段就需要考虑最后的评价方式。

04
数据维度

数据维度概述

数据维度（Dimensionality）是数据应用工程 - 成熟度模型的关键组成部分，是将数据应用工程中每个环节都会涉及到的同类型工作，从系统性的角度来整体考虑，类似的事情可能很多，本版本先从最重要的元数据、数据质量和数据安全三个维度来展开。

元数据管理

元数据概述

业界有专门针对元数据的全面介绍，本体系不做详细展开，仅将涉及数据应用工程相关的主要内容做必要阐述。

元数据相当于一个数据环境中的目录卡，在这个受控的数据环境中，元数据是描述数据的标签或数据的上下文背景。元数据为用户展示了在哪里可以找到什么类型的数据和信息，还提供了这些数据从哪里来，是如何处理的，相关数据转换规则和数据质量要求等详细信息，有助于理解数据的真实含义和对数据进行解释说明。可以理解为元数据就相当于数据的目录和字典。所以做好元数据管理在数据应用工程中至关重要，一般的元数据字典信息可以存储在专门的元数据，或者元数据文档，或者我们称之为“数据目录”的数据管理系统。并且一个公司最好有一个统一的元数据管理体系，这样能让公司所有员工在同一个语言频道上沟通和交流，业务人员和技术人员可以方便的理解数据。

元数据定义及分类

元数据（Meta Data）是指描述数据的数据，是关于一个企业所使用的物理数据、数据规则和约束、数据的物理与逻辑结构的信息。

元数据通常可以分为业务元数据、技术元数据和操作元数据。

业务元数据包括规则、定义、数据、术语表、运算法则和系统使用业务语言等，主要使用者为业务用户。

技术元数据主要用来定义数据应用过程中各个组成部分元数据结构，具体包括各个系统表和字段结构、属性、出处、依赖关系等，以及存储过程、函数、序列等各种对象。这其中描述的对象既包含结构化数据也包含非结构化数据。

操作元数据主要是指应用程序运行的信息，比如频率，记录数以及各个组件的分析和其他统计信息等。

如何管理元数据

1、明确元数据管理策略及架构

为了支撑数据工程，构建智慧的分析洞察，企业需要实现贯穿整个企业的元数据集成，建立完整且一致的元数据管理策略。这需要明确企业元数据管理的需求、目标、约束和详细策略，依据企业现状制定元数据管理的实施路线，确定元数据管理的安全策略、版本策略、

访问推送策略等等。

在策略确定后进行体系架构设计，体系架构从技术架构和数据架构两个角度考虑。技术架构方面，一般的元数据集成体系可以分为：点对点的元数据体系结构、中央辐射式元数据体系结构、分布式元数据集成体系结构和层次 / 星型元数据集成体系结构。数据架构方面，可以从以下几个角度展开：数据源角度、主题域的角度、实体的角度和业务角度等多方面。

2、实施元数据管理

创建业务术语词库。考虑到企业可以获取数据的容量和多样性，应该创建一个体现关键数据业务术语的业务定义词库（本体），该业务定义词库不仅仅包含结构化数据，还可以将半结构化和非结构化数据纳入其中。

创建技术和操作元数据库。基于数据应用过程的所有环节，参照数据架构和数据策略建立包含数据源、主题域、数据处理过程的元数据库。

建立长效支持机制。及时跟进和理解各种数据技术中的元数据，提供对其连续、及时地支持，比如 MPP 数据库、流计算引擎、Apache Hadoop/ 企业级 Hadoop、NoSQL 数据库以及各种数据治理工具如审计 / 安全工具、信息生命周期管理工具等。

打通元数据链路。将业务元数据和技术元数据进行链接，可以通过操作元数据监测数据的流动；可以通过数据血缘关系分析在整个信息供应链中实现数据的正向追溯或逆向追溯，了解数据都经历了哪些变化，查看字段在信息供应链各组件间转换是否正确等；可以通过影响分析了解具体某个字段的变更会对信息供应链中其他组件中的字段造成哪些影响等。

扩充元数据管理角色。扩展企业现有的元数据管理角色，比如可以扩充元数据管理者、数据主管、数据架构师以及数据科学家的职责，加入数据治理的相关内容。

数据质量

数据质量概述

质量是产品或工作的优劣程度，从字面意思分拆来看是指品质和数量，品质代表了可用性，数量代表了可测量性，所以质量管理需要更多的关注可用性和可度量性。

数据是 DIKW 模型（Data-Information-Knowledge-Wisdom）中的最基础层，只有数据被准确的保存记录才有后续有效信息的分析，才有知识规律的总结。

将“数据”和“质量”两个词组合在一起就可以看出数据质量的重要性。它是数据业务的基石。数据从收集、整理、分析到应用会受到多个环节的影响，所以要想使最后数据应用环节的数据质量效果好，必须保证前序各个环节的数据质量。所以数据质量不是单点的管理，是全方位的管理，是持续的管理。需要所有部门一起付出努力才能保证最后数据应用产品的质量。

数据质量维度

数据质量维度包括：

1. 准确性 (Accuracy)：数据准确性是指数据准确反映其所建模的“真实世界”实体的程度。通常，度量数据值与一个已确定的正确信息参照源的一致性可以度量准确性，如：将数据值与来自数据库或其他数据表的正确数据集比较，根据动态计算的数值进行检查，有时可能需要手工检查数值的准确性；

2. 完整性 (Completeness): 完整性的要求之一是一个数据集的特定属性都被赋予了数值。完整性的另一个要求，是一个数据集的全部行记录都存在。要对一个数据集的不同约束类型的属性应用完整性规则，如：必须有取值的必填属性，有条件可选值的数据元素，以及不适用的属性值。还可以认为完整性包括数据值的可用性和适当性；

3. 一致性 (Consistency): 一致性是指确保一个数据集的数值与另一个数据集的数值一致。一致性的概念相对宽泛，可以包括来自不同数据集的两个数值不能有冲突，或者在预定义的一系列约束条件内定义一致性。可以将更正式的一致性约束作为一系列定义一致性关系的规则，这些规则可以应用于属性值之间，记录或消息之间，或某一属性的全部数值之间。需要注意的是，不能将一致性与准确性或正确性相混淆。一致性可以定义在同一条记录中的一个属性值集合与另一个属性值集合之间 (记录级一致性)，或定义在不同记录中的一个属性值集合与另一个属性值集合之间 (跨记录一致性)，还可以定义在同一条记录中但在不同时间点的同一属性值集合之间 (时间一致性);

4. 时效性 (Currency): 数据时效性是指信息反映其所建模的当前真实世界的程度。数据时效性度量了数据的“新鲜程度”以及在时间变化中的正确程度。可以根据数据元素刷新的频率度量数据的时效性，从而验证数据是最新的。数据时效性规则定义了一个数据值在失效或需要更新之前已经历的“寿命”；

5. 精确度 (Precision): 精确度是指数据元素的详细程度。数值型数据可以有若干精确数位。例如，对数据取整或截断可能会产生精确度错误；

6. 有效性 (Validity): 有效性是指数据实例的存储、交换或展现的格式是否与数据值域一致，是否与其他相似的属性值一致。有效性确保了数据值遵从于数据元素的多个属性：数据类型、精度、格式、预定义枚举值、值域范围及存储格式等。为确定可能取值而进行有效性验证不等同于为确定准确取值而进行真实性验证。

如何进行数据质量管理

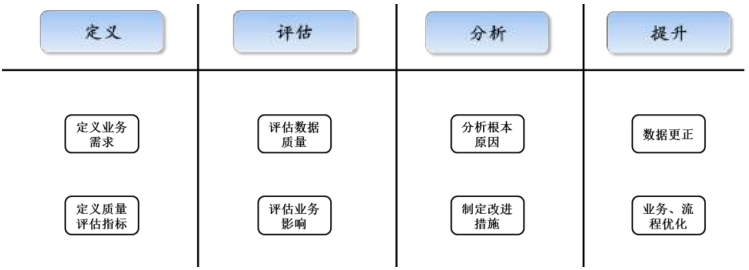
“如何进行数据质量管理”面对这个问题不同的企业和个人会给出多种不同的答案，有正向的质量控制方法，有逆向的质量保证方法，有丰田的 QCC，有问题管理导向的 5-WHY 和 8D。但综合来看，一般数据质量管理可以从质量理念，质量管理方法和质量工具三个角度入手。

1、理念——戴明环（PDCA）

在此仅介绍一个简单的理念

2、方法——数据质量管理提升方法

结合数据管理的生命周期定义以及戴明环（PDCA）理论，数据质量管理的生命周期可以分为四大阶段，八个工作步骤，具体定义如下：



数据质量管理提升方法论

1. 定义业务需求。定义和明确数据质量管理的目标和范围，以指导数据质量管理整个阶段的工作，数据的业务管理需求是数据质量规则的重要体现，在本阶段需要明确数据质量管理的目标以及业务需求，为后续的工作提供指导。

2. 定义质量评估指标。根据数据质量的管理目标以及业务规则，结合数据相关的信息技术环境分析，选取适合本部分数据的数据质量评价指标。

3. 评估数据质量。针对适用于本部分数据的数据质量评价指标，结合数据质量评价方法和数据质量评估工具,综合评估数据质量。评估结果为未来步骤提供基础。例如: 确定根本原因、需要的改进和数据更正等等。

4. 评估业务影响。使用各种方法、技术来评估劣质数据对业务、经济的影响。该步骤为建立改进业务案例，获取数据质量支持、确定适当的信息资源投资提供依据。

5. 分析根本原因。从业务、流程、信息系统等多方面来分析引起数据质量的真实原因。基于数据质量根本原因的分析可以帮助制定并执行数据质量的提升方案。

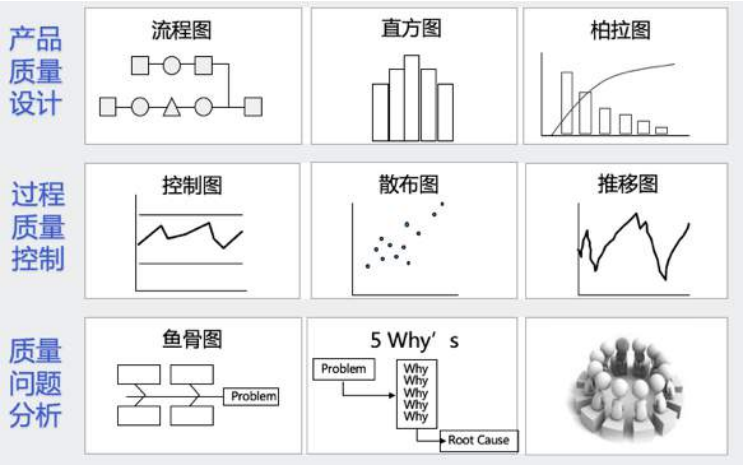
6. 制定改进措施。根据数据质量的原因分析，制定数据质量提升的行动计划和建议。基于这些计划和建议可以进行数据的更正。

7. 数据更正。对存在问题的数据进行更正或者提升，并且对数据更正的过程进行监控和确认，确保业务规则和目标得到满足。

8. 业务、流程优化。根据数据质量原因的分析、业务影响的分析、业务规则的分析等多方面的因素，对当前的业务、流程以及相关的信息环境进行优化，预防未来类似数据问题的出现。同时，对典型案例进行总结，形成数据质量管理知识库。

3、工具

传统的质量领域经常会提到质量七工具，他们是流程图、直方图、柏拉图、控制图、散布图、推移图、鱼骨图。从产品质量设计、过程质量控制和质量问题分析角度可以对应使用。



由于数据质量是存在于数据相关的所有环节，所以现在各类数据相关产品中都设置有质量模块 / 功能，比如在数据资产管理中的数据目录中就可以设置对每个数据集的数据质量指标进行监控。

综上，在数据质量这个大范畴中只是举例说明了质量理念、数据质量管理方法和数据质量工具。更多的数据质量管理还需要结合企业的发展阶段、数据场景和技术能力做出有效的选择。

数据安全

数据安全概述

数据安全是指通过建立和采用技术和管理的安全保护，保护数据不因偶然和恶意的原因遭到破坏、更改和泄露。

数据安全存在着多个层次，如：制度安全、技术安全、运算安全、存储安全、传输安全、产品和服务安全等。对于数据安全来说：制度安全治标，技术安全治本，其他安全也是必不可少的环节。数据安全不仅关系到个人隐私、企业商业隐私；而且数据安全技术直接影响国家安全。

如何做好数据安全管理

数据安全管理是计划、制定、执行相关安全策略和规程，确保数据和信息资产在使用过程中有恰当的认证、授权、访问和审计等措施。有效的数据安全策略和规程要确保合适的人以正确的方式使用和更新数据，并限制所有不适当的访问和更新数据。数据安全管理过程可以从要求、策略、实施、审计四个方便开展。

1、明确数据安全管理要求

数据安全管理目标是保证数据和数据主体信息保密性、数据和数据主体信息真实性和数据可用性。创建合适的数据安全策略，建立相应的机构组织结构和安全管理体系，包括系统和数据资产清单、组织和人员管理、符合业务流程的数据供应链管理体系、满足数据安全服务的元数据体系、各种安全合规性规范等数据服务基础安全能力的相关要求。

针对数据生命周期管理相关的数据活动，形成数据服务安全规范、控制措施、管理流程等数据活动安全能力，目的是降低各种数据活动的安全风险，保障数据安全。从规划、开发、部署到系统运维的生命周期各阶段对数据平台和应用采取必要的安全技术和管理安全措施，目的是建立安全的数据服务环境，降低运行安全风险。

2、定义数据安全策略

在实施数据安全管理前，首先要参考数据安全管理要求及相关依据建立数据安全策略，可以参考如下一些依据进行：

国家的法律法规：

以个人信息安全为例，欧盟颁布了《GDPR》，中国先后颁布了《网络安全法》、《个人信息和重要数据出境安全评估办法》

国家标准或行业监管要求：

以个人信息安全为例，有《信息安全技术个人信息安全规范》、《信息安全技术个人信息去标识化指南》、《数据出境安全评估指南》

数据的分类分级标准：

数据分类维度可以按照数据主体分类、主题分类、行业分类。分级可以将非涉密数据分为公开、敏感数据

大数据安全管理原则：

职责明确原则、意图合规原则、最小授权原则、数据保护原则、可审计原则等

业务规则要求及权限管理策略

3、实施数据安全管理

数据安全策略和要求的落实主要靠企业的安全制度流程及相关平台工具进行，在数据应用工程从数据源、网络传输、采集、存储、处理计算、应用等各个维度都需要进行安全策略的控制。如下图为简单列举。

4、审计数据安全

数据安全审计是一项控制活动，负责经常性分析、验证、讨论、建议数据安全管理相关的政策、标准和活动。数据安全审计的目标是为管理层和数据工程委员会提供客观中肯的评价、合理可行的建议。数据安全策略声明、标准文档、实施指南、变更请求、访问监控日志、报表输出，以及其他电子和书面记录等形成审计的基础。

数据安全审计包括：分析数据安全策略和标准；分析实施规程和实际做法，确保数据安全目标、策略、标准、指导方针和预期结果相一致；评估现有标准和规程是否适当，是否与业务要求和技术要求相一致；验证机构是否符合监管法规要求；检查安全审计数据的可靠性和准确性；评价违背数据安全行为的上报规程和通知机制；评审合同、数据共享协议、确保外包和外部供应商切实履行他们的数据安全义务，同时保证组织要履行自己应尽的义务；向高级管理人员、数据管理专员以及其他利益相关者报告数据安全状态，以及数据安全实践成熟度；推荐数据安全的设计、操作和合规改进工作。

对于有效的数据安全管理而言，没有什么可以替代数据安全审计工作。审计是一个支持性、可重复的过程，应当有规律的、高效的持续执行数据安全审计工作。

以上是依据数据安全管理的思路举例，数据安全多个层次的管理侧重点不同，企业要依据自己的业务特点确定数据安全的侧重点，但安全策略必须考虑法律和行业监管的要求。

05
数据工具与技术

综述

工欲善其事必先利其器，无论方法和理论多么的高深，要落地和实施必须依靠技术的力量，必须依靠工具和平台。随着移动互联网的发展及计算科学的不断演进，各种数据处理框架、大数据工具也都应运而生。盘点所有数据工具与技术发现以下特点：

1. 数据处理工具从收费走向开源。传统的数据存储和分析工具基本上以收费为主。比较典型的包括数据库类软件 Oracle、数据处理分析类工具 Excel、建模类工具 Matlab 以及 SPSS Modeler 等。但是在大数据场景下，最常用的工具往往是开源的工具。例如，目前广泛使用的 Hadoop、Spark、Hive 都是优秀的开源工具，Python、R 等编程语言也提供例如 Pandas、Scikit-Learn 等易用的开源库。同样，很多产品供应商提供的平台类产品都是基于这些开源工具构建的。因此，大数据处理场景下，常用的工具和平台都趋于开源化。

2. 大数据处理工具更为多样化。相比传统的数据软件，当前的大数据工具更具有多样性，在数据处理的各个阶段（数据采集、数据存储、数据探索、数据分析、数据建模、数据发布等）都浮现出了众多的工具，有的工具专注于做强某一个方面，有的贯穿了整个数据处理的流程，探索统一的大数据工具平台。

3. 大数据处理工具更为智能化。从功能上讲，当前大数据处理工具相比传统处理工具最令人兴奋的是智能化趋势，工具的智能化在贯穿数据处理的整个过程。例如，Tamr 主要用机器学习的方式来解决数据的关联、唯一实体的识别；DataRobot 等工具实现了建模过程、调参的智能化。这些都极大的简化了数据处理的过程。

大数据工具列表

每个公司都有自己的技术特点和工具体系，以下列出市面上常用的工具列表，供数据应用过程中参考使用。

常用主要开源工具

名称	说明
Apache Hadoop	开发可靠、可扩展、分布式计算的开源软件
Cloudera	Cloudera 的开源 Apache Hadoop 发行版，亦即（Cloudera Distribution including Apache Hadoop，CDH），面向 Hadoop 企业级部署。
Apache Spark	相对于 Hadoop 的 MapReduce 会在运行完工作后将中介数据存放到磁盘中，Spark 使用了内存内运算技术，能在数据尚未写入硬盘时即在内存内分析运算。Spark 在内存内运行程序的运算速度能做到比 Hadoop MapReduce 的运算速度快上 100 倍，即便是运行程序于硬盘时，Spark 也能快上 10 倍速度。
Apache Storm	Apache Storm 是一个免费的开源分布式实时计算系统。
Kafka	Kafka 是一种高吞吐量的分布式发布订阅消息系统，它可以处理消费者规模的网站中的所有动作流数据。
Hive	基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，可以将 SQL 语句转换为 MapReduce 任务进行运行。其优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析。
Elasticsearch	基于 Lucene 的搜索服务器。提供了一个分布式多用户能力的全文搜索引擎，基于 RESTful web 接口。用于云计算中，能够达到实时搜索、稳定、可靠、快速、安装使用方便。

MongoDB	MongoDB 是一个基于分布式文件存储的数据库。由 C++ 语言编写。旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。MongoDB 是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。
Hbase	开源的非关系型分布式数据库（NoSQL），它参考了谷歌的 BigTable 建模。Hadoop 项目的一部分，运行于 HDFS 文件系统之上，为 Hadoop 提供类似于 BigTable 规模的服务。

数据仓库与数据管理工具

名称	说明
Informatica PowerCenter	PowerCenter 构成了所有数据集成计划的基础，包括分析和数据仓库，应用程序迁移或整合与数据治理。
WhereHows	LinkedIn 开源了的一个元数据中心工具，已经在 LinkedIn 内部长期使用。方便内部员工发现公司内部的数据，跟踪数据集的移动和查看各种内部工具和服务的动向。
Atlas	一个可扩展和可扩展的核心基础治理服务集 - 使企业能够有效地和高效地满足 Hadoop 中的合规性要求，并允许与整个企业数据生态系统的集成。
Waterline Data	一款自动化的数据发现平台，可帮助数据架构师大规模自动清理 Hadoop 中的所有数据，并将数据安全地提供给业务用户，并使数据自动进行分析，无需手动探索每个文件。水线数据还有助于自动发现沿袭和业务元数据，并管理元数据。

数据清洗、集成和 ETL 工具

名称	说明
OpenRefine	有力的处理混乱数据的工具。可以从一种格式清洗、转化为另一种格式。Google 目前不在支持该工具的迭代。
Talend Open Studio	支持 ETL 过程的所有方面。直观的流程建模工具使企业利益相关者能够参与最初的 ETL 设计工作。提供了超过 800 个内置连接器。
Informatica PowerCenter Big Data Edition	使用可视化开发环境构建在 Hadoop 上本地运行的 ETL 数据流。数据流可以重复使用，并与其他开发人员和分析师通过集成开发环境（IDE）进行协作。允许访问包括 RDBMS，OLTP，OLAP，ERP，CRM 等。PowerCenter Big Data Edition 在 Hadoop 上提供了一个预建的转换功能库，包括数据类型转换和字符串操作，支持高性能缓存的查找，筛选器，连接器，分类器，路由器，聚合等等。
Jaspersoft ETL	Jaspersoft ETL 易于部署，用于从交易系统提取数据，以创建用于报告和分析的整合数据仓库或数据集市。
Apache Airflow	Apache Airflow 是一个用于编写，安排和监控工作流程的开源工具。它具有丰富的用户界面，可以方便地查看生产中运行的管道，监视进度，并在需要时排除故障。

BI 与可视化工具

名称	说明
Tableau	可以帮助用户快速的看到并理解数据。帮助任何人快速的分析、可视化并分享信息。
Power BI	微软为非技术人员提供的自服务式的 BI 解决方案。
Sisense	为复杂的数据简化了 BI 操作，包括大数据集以及分散的多个数据集。支持 R 语言。
Qlik	自助服务可视化使用数据可视化应用程序驱动洞察发现。组织中的每个人都可以轻松创建灵活的交互式可视化，并做出有意义的决策。使用简单的拖放界面来创建灵活的交互式数据可视化。利用智能可视化技术探索数据，自动适应您设置的参数 - 无需开发人员，数据科学家或设计人员。
ECharts	百度开源的可视化图表工具。遵循 BSD 开源协议，免费商用。满足各种可视化需求。
HighCharts	用纯 JavaScript 编写的一个图表库， 能够很简单便捷的在 web 网站或是 web 应用程序添加有交互性的图表，并且免费提供给个人学习、个人网站和非商业用途使用。HighCharts 支持的图表类型有曲线图、区域图、柱状图、饼状图、散状点图和综合图表。

数据建模与数据科学工具

名称	说明
IBM Data Science Experience（DSx）	一站式的数据科学工作空间，支持各种数据科学家需要的开源的工具集，支持 R, Python, Scala，并且与 Rstudio, Spark, Python 进行了集成。DSx 可以访问 Watson 数据平台提供的数据集，并且支持云端、私有部署以及桌面版本。
Azure Machine Learning	完全托管在云上的机器学习服务。可以轻松构建、部署和分享预测分析解决方案。
KNIME	功能完备并且富有弹性的针对数据科学家的平台。KNIME 可以通过 KNIME Bigdata Extension 去进行大数据的计算和机器学习。KNIME 分析平台提供了比较强大的数据准备能力，包括连接和混合数据、验证数据质量、值汇聚、平滑、数据集分区以及特征生成和选举等等。
SAS Visual Analytics Suite	SAS 公司一系列可视化分析的套件，包括 Visual Statistics, Visual Data Mining and Machine Learning。具备很好的数据访问能力以及数据准备能力，同时具备卓越的数据可视化和探索能力。支持不同的数据源和数据类型，并且具备非常好的社区的支持。
RapidMiner	平台包含大量的算法、灵活的模型能力、数据源集成能力以及数据准备的能力。平台易于使用，能够非常快速的进行模型的开发，并且具备良好的开源支持。
Dataiku	注重协作、对用户友好的数据科学平台。支持业务人员点击分析处理数据，支持数据分析师、数据科学家写脚本处理数据的能力，主要支持 Python\R\Spark\Hive 等脚本语言。同时，支持快速模型的能力。数据连接能力也很强悍。支持私有化部署。
Domino Data Lab	可以自己订制开发环境的平台，主要面向数据科学家，基本全部的操作都基于代码。只能在云端运行。
ANACONDA	一个由 Python 支持的开放式数据科学平台。Anaconda 的开源版本是 Python 和 R 的高性能版本，包括超过 100 种用于数据科学的最受欢迎的 Python，R 和 Scala 软件包。还可以访问超过 720 个软件包，可以很容易地安装 conda 包。
DataRobot	DataRobot 为所有技能水平的数据科学家提供了一个机器学习平台，可以在过去几分之一的时间内构建和部署准确的预测模型。该技术通过改变预测分析的速度和经济性来解决数据科学家的严重短缺问题。
H2O	H2O 是智能应用（深度学习、梯度提升、随机森林、广义线性建模（Logistic 回归、弹性网络）、K 均值等）的开源快速可伸缩机器学习 API。

06
附录

【附录 1】术语

元数据（MetaData）：是指描述数据的数据，是关于一个企业所使用的物理数据、数据规则和约束、数据的物理与逻辑结构的信息。

数据（Data）：可以是数字、文字、图像、符号等，它直接来自于事实，可以通过原始的观察或度量来获得。

信息（Information）：信息，指音讯、消息、通讯系统传输和处理的对象，泛指人类社会传播的一切内容。

数据集（DataSet）：是一种由数据组成的集合。

数仓：即数据仓库，英文名称为 Data Warehouse，可简写为 DW 或 DWH。数据仓库，是为企业所有级别的决策制定过程，提供所有类型数据支持的战略集合。它是单个数据存储，出于分析性报告和决策支持目的而创建。为需要业务智能的企业，提供指导业务流程改进、监视时间、成本、质量以及控制。

结构化数据：是由二维表结构来逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

非结构化数据：数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据。包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频 / 视频信息等等。

ETL（Extract-Transform-Load ）：用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。

数据理解：充分理解企业业务和数据，在此基础上定义数据要解决的业务问题，并评估其关联关系和可行性的过程。

数据准备：从各种数据源处获取原始数据，按照预期的业务需求定义数据应用的目标数据，将所有原始数据抽取、清洗、融合、转换、处理成为预期待分析挖掘的目标数据的过程。

数据开发：是以工程思维的角度将数据应用的关键实施过程进行演绎，是指在目标明确、数据集已整理完备的基础上进行分析挖掘探索数据应用模式的过程。

部署运营：将开发阶段模型成果在业务线上部署，在生产环境中例行化，并跟踪其运行效果的过程。

业务理解：是从商业角度全面理解客户想要达到的目标或者要解决的问题，划定业务目标和问题范围。

数据评估：是从工程角度评估可行性，评估中需要包含数据的可获取性、技术的可行性、业务可行性、资源评估、成本分析、风险分析等。

数据获取：是指用系统的方法，收集和测量各种来源的信息，以获得完整、准确的数据内容。

数据定义：是指在获取的数据源基础上，按照数据应用的业务目标定义目标数据集，并设

计数据处理流程方案的过程。

数据整理：是指对数据源进行抽取、清洗、转换等加工处理，生成目标数据集的过程。

数据增强：是将标准化后的多个单来源数据，进行数据联通，将联通的不同数据融合在一起生成新的字段更全的数据集，或者将通过联通后的数据交叉映射补充空缺字段内容。

BI（Business Intelligence）：商业智能，又称商业智慧或商务智能，指用现代数据仓库技术、线上分析处理技术、数据挖掘和数据展现技术进行数据分析以实现商业价值。

数据分析：是指在一定的商业场景或模型下，采用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

数据挖掘：一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

AI（Artificial Intelligence）：人工智能是研究使计算机来模拟人的某些思维过程和智能行为（如学习、推理、思考、规划等）的学科，主要包括计算机实现智能的原理、制造类似于人脑智能的计算机，使计算机能实现更高层次的应用。

ML（Machine Learning)：机器学习是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。是人工智能的核心。

CDO（Chief Data Officer)：是随着企业不断发展而诞生的一个新型的管理者。其主要是负责根据企业的业务需求、选择数据库以及数据抽取、转换和分析等工具，进行相关的数据挖掘、数据处理和分析，并且根据数据分析的结果战略性地对企业未来的业务发展和运营提供相应的建议和意见。

SaaS （Software-as-a-Service）（软件即服务）：一种通过 Internet 提供软件的模式，用户不用再购买软件，而改用向提供商租用基于 Web 的软件，来管理企业经营活动，且无需对软件进行维护，服务提供商会全权管理和维护软件。

数据源：数据的来源，是提供某种所需要数据的器件或原始媒体。一般获取数据的渠道有内部业务系统产生 / 收集，公开网络获取、外部购买、合作交换等。

数据质量管理：是指为了满足信息利用的需要，对信息系统的各个信息采集点进行规范，包括建立模式化的操作规程、原始信息的校验、错误信息的反馈、矫正等一系列的过程。

数据孤岛：数据孤岛分为物理性和逻辑性两种。物理性的数据孤岛指的是，数据在不同部门相互独立存储，独立维护，彼此间相互孤立，形成了物理上的孤岛。逻辑性的数据孤岛指的是，不同部门站在自己的角度对数据进行理解和定义，使得一些相同的数据被赋予了不同的含义。

分类: 根据预先确定的系统对数据进行分类，结果目录用于提供易于访问和检索的概念框架。

异常检测：异常检测是指利用定量的方式来描述可接受的行为特征，以区分和正常行为相违背的、非正常的行为特征来检测入侵。

回归预测：就是把预测的相关性原则作为基础，根据相关因素和预测目标建立模型，用模型预测目标。

聚类分析：是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，让在同一个子集中的成员对象都有相似的一些属性。

强化学习：又称再励学习、评价学习，是一种重要的机器学习方法，在智能控制机器人及分析预测等领域有许多应用。强化学习系统学习的目标是动态地调整参数，以达到强化信号最大。

特征工程：其本质是一项工程活动，目的是最大限度地从原始数据中提取特征以供算法和模型使用。

模型训练：将准备好的特征数据集带入算法，进行运算训练，并不断调整算法参数和特征数据进行迭代，直至训练达到预期效果，或者在现有算法和特征数据下的最优结果。

模型评估：一般在进行模型训练时，会将特征数据分为两份或者多份，将其定义为训练集和验证集，在训练集上进行模型训练达到预期效果的情况下，使用验证集数据进行模型的评估，从召回率和正确率等多个角度综合评估。

ground truth：指的是用于有监督训练的训练集分类的准确性。

数据探索：对数据进行初步研究，以便更好的理解它的特殊性质。

数据建模：是指在准备好的数据集基础上，基于业务要解决的问题，设定假设，特征提取，使用算法构建模型，并迭代验证的过程。

运营监控：是指在业务系统部署模型后设置业务或质量监控指标，持续监控业务运营效果。

数据目录：是指一种数据管理系统。是一种用户可以访问的记录数据库和应用程序元数据的目录。

主题域：是联系较为紧密的数据主题的集合。

DIKW 模型（Data-Information-Knowledge-Wisdom）：是一个解释数据（Data）、信息（Information）、知识（Knowledge）和智慧（Wisdom）之间的关系的模型，这个模型描述了数据一步步转化为信息、知识、乃至智慧的方式。

QCC（品管圈）：是由相同、相近或互补的工作场所的人们自动自发组成数人一圈的小圈团体（又称 QC 小组，一般 6 人左右），全体合作、集思广益，按照一定的活动程序来解决工作现场、管理、文化等方面所发生的问题及课题。

5-why 分析法：又称“5 问法”，对一个问题点连续以 5 个“为什么”来自问，以追究其根本原因。

8D：是解决问题的 8 条基本准则或称 8 个工作步骤。D0：征兆紧急反应措施。D1：小组成立。D2：问题说明。D3：实施并验证临时措施。D4：确定并验证根本原因。D5：选择和验证永久纠正措施。D6：实施永久纠正措施。D7：预防再发生。D8：小组祝贺

戴明环（PDCA）：PDCA 的含义是将质量管理分为四个阶段，即计划（Plan）、执行（Do）、

检查（Check）、调整（Action）。

GDPR：欧盟发布的《一般数据保护条例》（General Data Protection Regulation）。

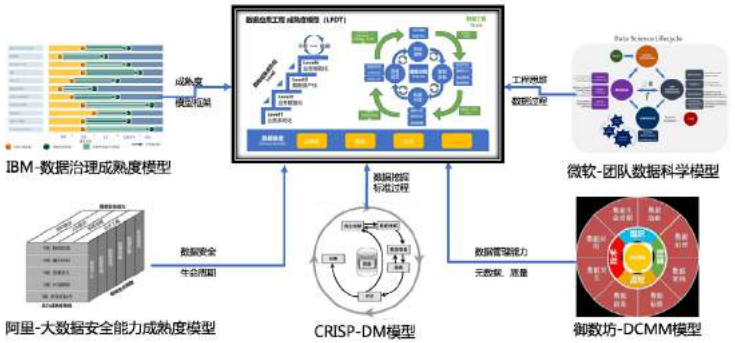
数据安全：通过技术及管理进行安全保护，保护数据不因偶然和恶意的原因遭到破坏、更改和泄露。

数据安全审计：是一项控制活动，负责经常性分析、验证、讨论、建议数据安全管理相关的政策、标准和活动。

数据集成：把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。

【附录 2】溯源与关系

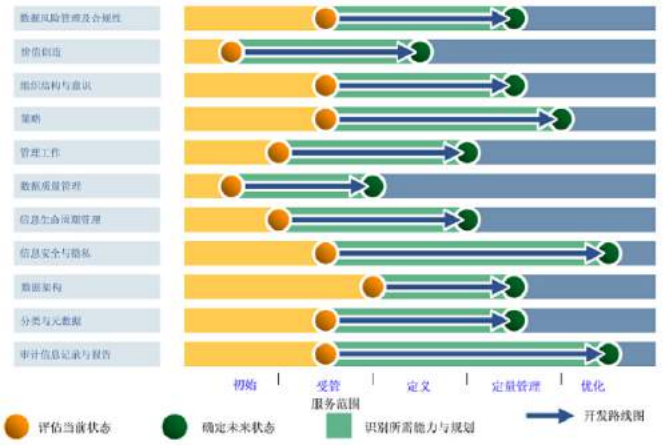
数据应用工程 - 成熟度模型（LPDT）借鉴了国内外数据管理 / 治理、数据挖掘、数据科学等众多模型和理论，结合 TalkingData 的数据实践经验，总结得出数据应用工程理论。其中借鉴了：IBM- 数据治理成熟度模型、微软 - 团队数据科学模型、阿里 - 大数据安全能力成熟度模型、CRISP-DM 模型、御数坊 -DCMM 模型、中国信息标准委员会各类标准材料、美国商务部的 NIST 框架等等素材。以下简单介绍几个主要模型的特点和内容。



一，IBM- 数据治理成熟度模型

2010 年 IBM 在《IBM 数据治理统一流程》中，结合 Software Engineering Institute (SEI) 在 1984 年开发的容量成熟度模型 (Capability Maturity Model, CMM)，提出了 IBM 数据治理成熟度评估模型，如图所示，主要分为 5 个等级，11 个功能模块。

1、11 个功能模块隶属于 4 个大组：



a. 成果：数据风险管理及合规性、价值创造

数据风险管理及合规性：确定数据治理与风险管理关联度，用来量化、跟踪、避免或转移风险等。

价值创造：确定数据资产是否帮助企业创造更大价值。

b. 支持条件：组织结构与意识、管理工作、策略

组织结构和意识：主要用来评估企业针对数据治理是否拥有合适的数据治理委员会、数据治理工作组和全职的数据治理人员，是否建立了数据治理章程以及高级主管对数据的重视程度等。

管理工作：是指质量控制规程，用来管理数据以实现资产增值和风险控制等。

策略：为企业如何管理数据在高级别指明方向。

c. 核心规程：数据质量管理、信息生命周期管理、信息安全与隐私

数据质量管理：主要指用来提高数据质量，保证数据准确性、一致性和完整性的各种方法。

信息生命周期管理：主要指对结构化、半结构化以及非结构信息化全生命周期管理相关的策略、流程和分类等。

信息安全与隐私：主要指保护数据资产、降低风险的各种策略、实践和控制方法。

d. 支持规程：数据架构、分类与元数据、审计信息记录与报告

数据架构：是指系统的体系结构设计，支持向适当用户提供和分配数据。

分类与元数据：是指用于业务元数据和技术元数据以及元模型、存储库创建通用语义定义的方法和工具。

审计信息记录与报告：是指与数据审计、内部控制、合规和监控超级用户等有关的管理流程。

2、五个等级：

初始级，临时的流程，整体环境不够稳定；

受管级，可重复流程，但可能无法针对组织中所有项目重复流程，存在基本的项目管理和流程规则，但仍有超出预期成本和时间的风险；

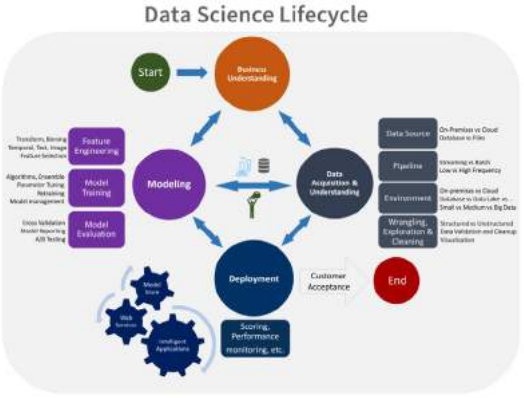
定义级，建立了标准流程集，通过组织的标准流程集定制标准、流程描述和项目过程，以适应特定项目或组织单位；

定量管理级，对流程进行定量度量和控制，所选的子流程大大提高了整体流程绩效；

优化级，在该级明确了组织的定量流程改进目标，并不断优化以适应变化的业务目标。

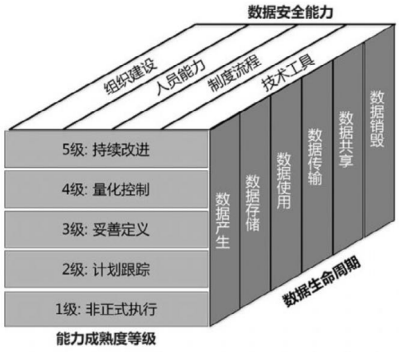
二、微软 - 团队数据科学模型

微软提供了一种有助于团队协作和学习的敏捷的迭代式数据科学方法 -Team Data Science Process (TDSP)。Team Data Science Process (TDSP) 提供了用于构建数据科学项目开发的生命周期。生命周期概述了执行项目时，其从开始到结束所遵循的步骤。该生命周期概述了项目通常执行项目时主要遵循的几个阶段：了解业务、数据采集和理解、建模、部署、客户验收。



三、阿里 - 大数据安全能力成熟度模型

阿里巴巴 - 大数据安全能力成熟度模型（Data Security Maturity Model, DSMM）：2016 年阿里巴巴基于组织数据的全生命周期（数据产生、数据存储、数据使用、数据传输、数据共享、数据销毁），从组织和人员、流程和操作以及技术和工具三个能力维度，针对组织的结构化数据的数据安全过程管理，提出规范性的大数据安全能力成熟度模型。



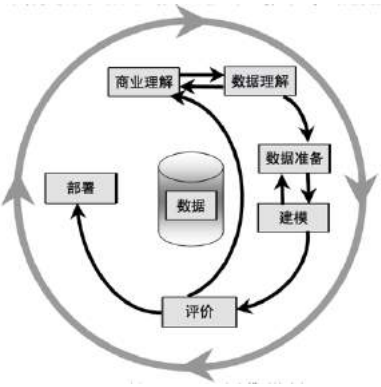
阿里 - 大数据安全能力成熟度模型

成熟度等级	评述	特征
等级 1: 非正式执行	仅依据特定业务需求来开展数据安全工作，未形成明确的工作内容定义	随机，被动的安全过程
等级 2: 计划跟踪	在已固化的数据安全工作内容的基础上，主要依赖人工执行相关工作，缺乏工具化和系统化的支撑业务的能力	主动，非正式的安全过程
等级 3: 妥善定义	经过对流程、人员的妥善管理，数据安全工作的开展已经能够保证对安全风险的规范化管理	正式，规范的安全过程
等级 4: 量化控制	数据安全工作的开展与业务发展的方向一致，并具备持续量化和跟踪的自动化能力	安全过程可控
等级 5: 持续改进	数据安全的整体管理已达到可持续感知，改进的状态，实现对业务发展的驱动	安全过程可调整

数据安全能力成熟度等级

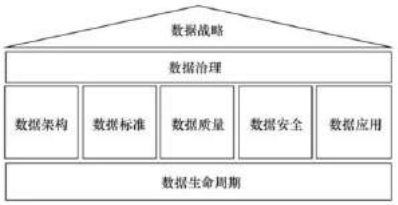
四、CRISP-DM 模型

CRISP-DM 将数据挖掘过程分为六个主要阶段。阶段的顺序并不严格，总是需要在不同阶段之间来回移动。流程图中的箭头表示阶段之间最重要且频繁的依赖关系。图中的外圈表示数据挖掘本身的循环性质。数据挖掘过程在解决方案部署后继续进行。在这个过程中吸取的经验教训可能会引发新的、往往更加重点突出的业务问题，随后的数据挖掘过程将受益于以前的经验。并描述了各阶段的一般任务（粗体）和其输出（斜体）。



五、御数坊 -DCMM 模型

2017 年御数坊借鉴国内外成熟度相关理论，结合数据生命周期管理各个阶段的特征，提出了数据管理能力成熟度模型（Data Management Capability Maturity Model, DCMM），包含了 8 个能力域，对应的 29 个能力项，并对其评价划分为 5 个等级。数据管理能力成熟度模型是通过一系列的方法、关键指标和问卷来评价某个对象的数据管理现状，从而帮助其查明问题、找到差距、指出方向，并且提供实施建议。



能力域	能力项
数据战略	数据战略规划
	数据职能框架
	数据战略实施
	数据任务效益评价
数据治理	数据治理组织
	数据制度建设
	数据治理沟通
数据架构	数据模型
	数据分布
	数据集成与共享
	元数据管理
数据应用	数据分析
	数据开放共享
	数据服务
数据安全	数据安全策略
	数据安全管理
	数据安全审计
数据质量	数据质量需求
	数据质量检查
	数据质量分析
	数据质量提升
数据标准	业务术语
	参考数据和主数据
	数据元
	指标数据
数据生存周期	数据需求
	数据设计和开发
	数据运维
	数据退役

六、NIST- 大数据架构

2015 年，NBD-PWG 参考架构小组开发了一个独立于供应商、技术方案和基础结构的大数据架构的概念模型。这个称作 NBDRA 的概念模型如图所示，它中立于供应商，展示了一个由相互关联的接口 / 服务相连接的功能组件组成的独立于技术和基础设施的大数据系统，为大数据标准化提供基本参考点，为大数据系统的基本概念和原理提供了一个总体框架，为各种利益相关者提供一种交流大数据技术的通用语言，鼓励大数据实践者遵守通用标准、规范和模式。

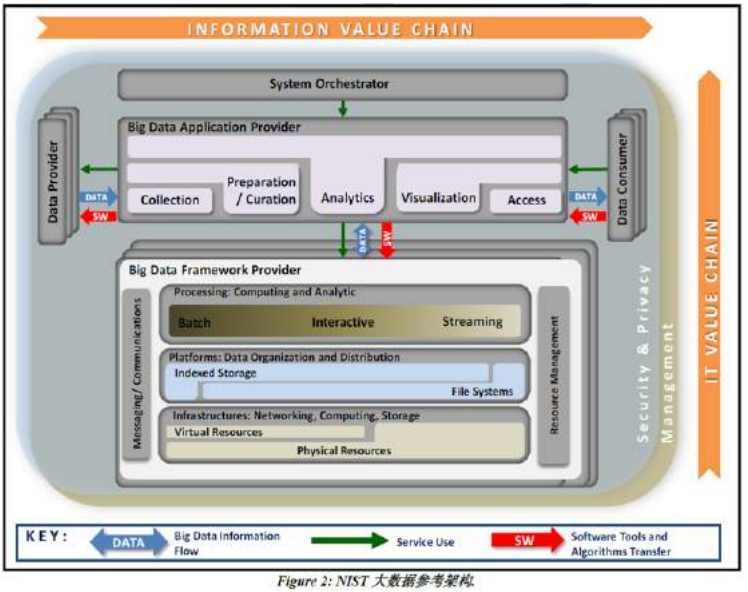


Figure 2: NIST 大数据参考架构

NBDRA 围绕两大价值链的两个轴线展开：信息价值链（水平轴）和 IT 价值链（垂直轴）。信息价值链的核心价值通过数据收集、数据准备 / 集成、数据分析、数据可视化及访问等应用产生。IT 价值链的核心价值由提供网络、基础设施、平台、应用工具和其他 IT 服务产生，这为大数据处理应用程序提供了托管和操作的支持。大数据应用程序提供商组件位于两个价值链的交叉点，这意味着数据分析及其实现在这两个价值链中都处于重要地位。

五个主要的架构模块代表在每个大数据系统中存在的不同技术角色：

系统协调者：定义和集成所需的数据应用活动到垂直操作系统中来

数据提供者：将数据和信息引入到大数据系统中

大数据应用提供者：执行一个生命周期，以满足安全性和隐私需求，也包括系统协调者定义的需求

大数据框架提供者：建立一个计算结构，在其中执行某些应用程序转换，同时保护隐私和数据的完整性

数据消费者：包括最终用户或其他系统利用大数据应用提供者的结果

安全和隐私问题会影响 NBDRA 中的所有组件。管理的作用是总体控制系统的执行，部署和维护。

【附录 3】参考文献

[1] DAMA InternationalD，AMA 数据管理知识体系指南，北京：清华大学出版社，2017

[2] Danette McGilveray，数据质量工程实践，北京：电子工业出版社，2010

[3] 车品觉，数据的本质，北京：北京联合出版公司，2017

[4] CRISP-DM 联盟（www.Crisp-DM.org），跨行业数据挖掘标准过程 -1.0 版本 (CRISP-DM 1.0)，2005

[5] 桑尼尔 索雷斯，大数据治理，北京：清华大学出版社，2015

[6] 阿里巴巴数据技术与产品部，大数据之路：阿里巴巴大数据实践，电子工业出版社，2017

[7] Óscar Marbán, Gonzalo Mariscal, Javier Segovia. A Data Mining & Knowledge Discovery Process Model. Universidad Europea de Madrid Spain, 2009

[8] Microsoft Azure, 什么是团队数据科学过程，

https://docs.microsoft.com/zh-cn/azure/machine-learning/team-data-science-process/overview

[9] NIST 大数据工作组（NBD-PWG），大数据定义，https://bigdatawg.nist.gov/，2017

[10] IBM, 大数据治理统一流程参考模型，

https://www.ibm.com/developerworks/cn/data/library/bd-1503bigdatagovernance4/index.html

[11] IBM, 大数据治理，https://www.ibm.com/developerworks/cn/bigdata/governance/index.html

[12] 全国信息安全标准化技术委员会，信息安全技术 个人信息安全规范（征求意见稿），

https://www.tc260.org.cn/，2017

[13] 全国信息安全标准化技术委员会，信息安全技术 个人信息去标识化指南（征求意见稿），

https://www.tc260.org.cn/，2017

[14] 全国信息安全标准化技术委员会，信息安全技术 数据出境安全评估指南指南（征求意见稿），https://www.tc260.org.cn/，2017

[15] 张群、吴东亚、赵菁华，大数据标准体系，中国电子技术标准化研究院

http://www.360doc.com/content/17/0913/13/39716884_686759660.shtml

[16] 崔晓波，https://mp.weixin.qq.com/s/OSCOfTF3LVK26bm7CYB51g，“TalkingData 服务过很多金融和互联网客户，金融企业利用大数据的话，一般要经过“四化”。

[17] 托雷，从数据来源、数据生态、数据技术、数加平台等方面，漫谈阿里大数据

http://bigdata.51cto.com/art/201608/516406.htm

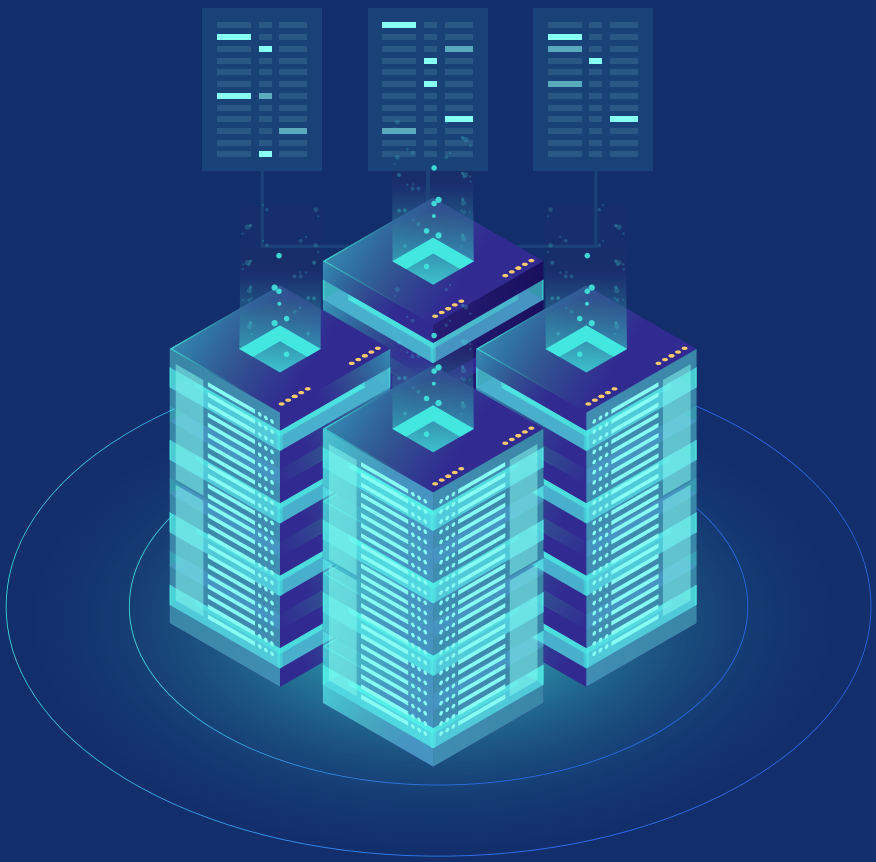
[18] TOP 50 bigdata platforms and bigdata analytics software

https://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/

TalkingData

数据应用工程成熟度模型

Data Application Engineering Maturity Model



T 400-870-1230
M support@tendcloud.com
关注TalkingData官方微信，获取最新资讯