数据治理总体解决方案

1大数据治理体系与数据治理体系的联系与区别

大数据治理是一项系统工程，大到大数据技术平台的搭建、组织的变革、政策的制定、流程的重组，小到元数据的管理、主数据的整合、各种类型大数据的个性化治理和大数据的行业应用。

组织必须治理全部大数据，将大数据治理定义如下：

大数据治理是广义数据治理计划的一部分，即制定与大数据有关的数据优化、隐私保护与数据变现的政策。 将上述大数据治理的定义分解为以下部分：

大数据是广义数据治理计划的一部分

数据治理机构必须采取以下措施，以将大数据整合到既有的数据治理框架中：

扩展数据治理宪章的外延，将大数据治理纳入其中；
拓宽数据治理委员会成员的范围，将数据科学家等大数据的超级用户吸纳进来；
任命处理社交媒体等特定大数据的主管；
将大数据与元数据、隐私、数据质量和主数据等数据治理准则结合。

大数据治理关乎**政策制定**

政策包括人们在特定情形下如何作为的成文和非成文的宣告。譬如，大数据治理政策可能申明，未经顾客知情并同意，组织不得将顾客的 Facebook 资料整合到其主数据记录中。

大数据必须**优化**

考虑一下组织是如何将现实世界的准则应用到大数据治理中的。公司设计了精致的企业资产管理计划，对机器、飞机、交通工具和其他资产进行妥善管理。与对实物资产进行登记类似，组织必须对大数据进行如下优化：

元数据——建立大数据类别信息；
数据质量管理——像公司对实物资产进行定期检修一样，定期净化大数据；
信息生命周期管理——对大数据进行存档，并在没必要继续保存某些数据时，将其删除。

大数据隐私至关重要

组织同样必须建立旨在防止大数据误用的适当政策。组织在处理社交媒体、地理定位、生物计量学和其他形式的个人可识别信息 (PII) 时，必须考虑涉及的声誉、规制和法律风险。

大数据必须**变现**

所谓变现，就是将数据等资产转化为现金的过程，变现的方式可以是将数据卖给第三方，也可以是利用数据开发新的服务。

在当下，公司意识到，必须将大数据视为具有财务价值的企业资产。例如，运营部门可以通过传感器数据，根据定期检修计划，提高设备正常运行时间。呼叫中心可以分析客户代表的记录，通过了解顾客呼叫的原因，降低呼叫量。此外，零售商可以使用主数据激活 Facebook 的应用程序，提升顾客忠诚度。

2.1大数据治理框架

大数据治理框架由三大部分组成：大数据类型、大数据治理领域、行业与功能

传统的数据治理领域同样适用于大数据

产业和功能场景

	情绪分析	患者监测	索赔分析	基因测试	电子病历
医疗保健业					
公共事业		智能仪表			
零售业	Facebook忠诚度计划	RFI日志		人脸识别	
电信业	客户流失分析	位置服务	客户流失分析		
保险业	索赔调查	车载通信技术	索赔欺诈分析	承保	
客户服务					通话质量保证
IT		IT日志分析			

业务流程整合
数据质量
隐私
元数据
组织

web和社交媒体数据 | 机器对机器的数据 | 大体量交易数据 | 生物计量学数据 | 人工生成的数据

大数据类型

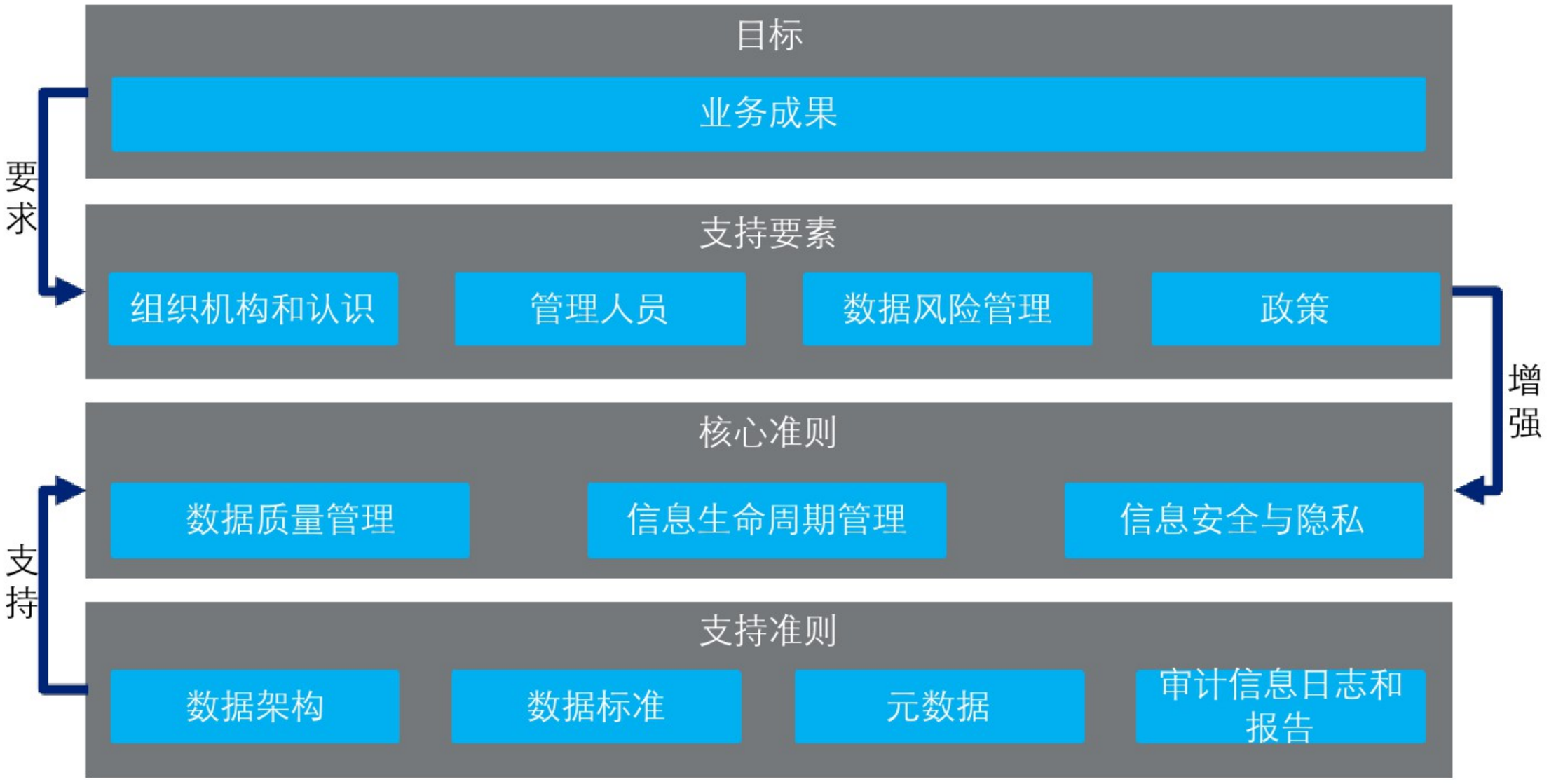
大数据治理需要高度聚焦于数据本身。

信息治理

大数据分析受用例驱动，用例的具体情况因产业和功能而异。

3.1大数据治理成熟度模型

实施大数据治理的第一步，是评估大数据治理成熟度的当前状态和期望的未来状态。现将某信息治理委员会的成熟度模型用于成熟度评估。该模型设立了4个领域的11个大数据治理成熟度指标。

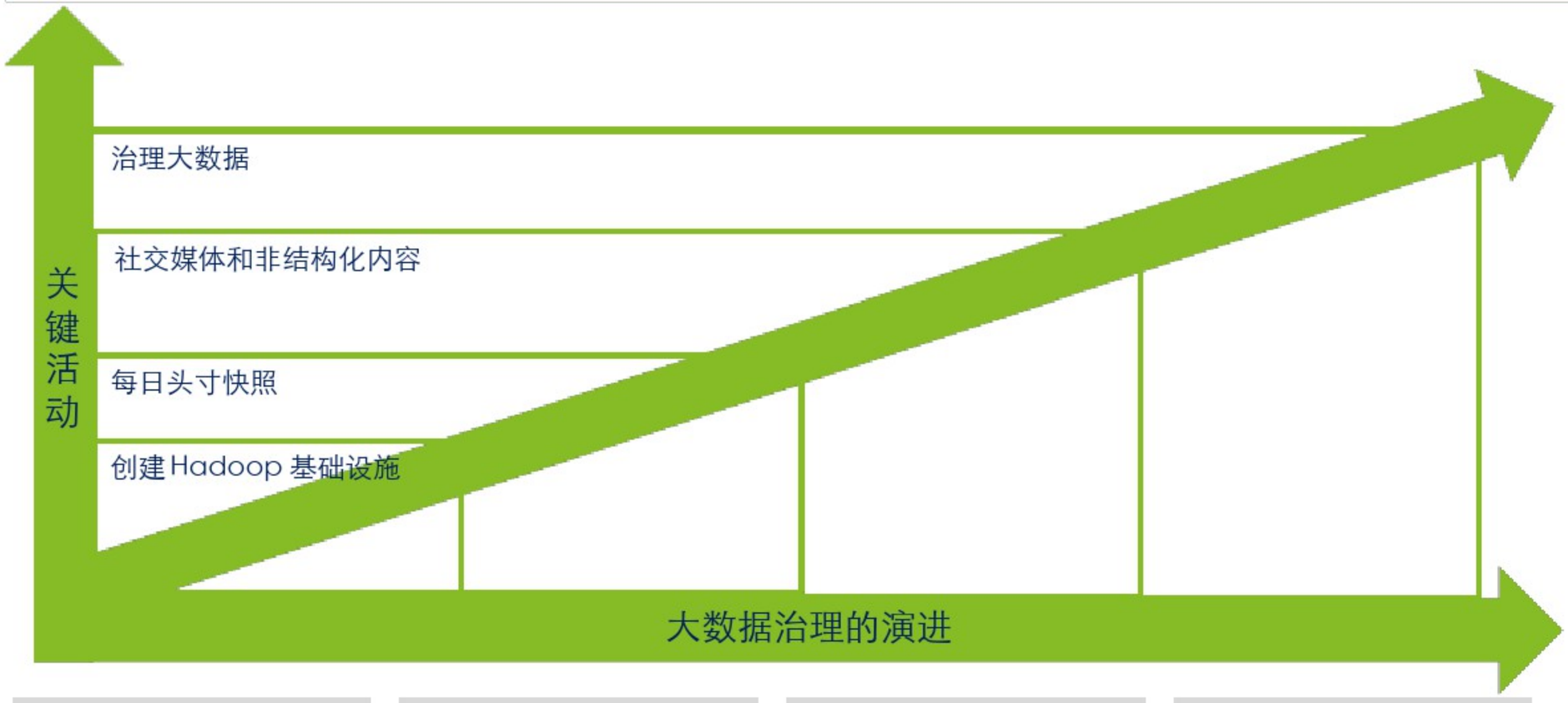


3.1.1大数据治理成熟度模型介绍及问题示例

	模型介绍	问题示例
目标	✍ 目标指信息治理计划的预期结果。目标倾向于关注降低风险与提升价值，这反过来又受降低成本和提高收入的驱动。 ✍ 业务成果：代表信息治理计划的目标和目的。	业务成果： A是否已经确定了大数据治理计划的关键业务关联方? B是否对大数据治理可带来的财务收益进行了量化?
支持要素	✍ 组织结构和认识：指业务部门和 IT部门间的相互责任，以及对治理不同管理层次中数据的信托责任的认识。 ✍ 管理人员：旨在保证数据监护，实现资产增值、风险消解和组织控制的质量控制准则。 ✍ 数据风险管理：据以识别、保留、量化、规避、接受、消解和转嫁风险的方法论。 ✍ 政策：期望得到落实的组织行为的书面表达。	✍ 数据结构和认识：如关键角色的职位说明中，是否包含大数据治理，如配备首席数据官和信息治理官? ✍ 管理人员：是否已经建立了责任分配（ RACI）矩阵，以定义针对大数据关键属性的角色和责任? ✍ 数据风险管理：是否在大数据治理与风险治理之间建立了联系? ✍ 政策：是否已经归档了一组大数据治理政策?
核心准则	✍ 数据质量管理：指测量、提高和保证产品数据、测试数据和归档数据的质量和集成性的方法。 ✍ 信息生命周期管理：有关信息采集、使用、保留和删除的系统化的、基于策略的方法。 ✍ 信息安全与隐私：组织用于消解风险和保护数据资产的策略、实践和控制手段。	✍ 数据质量管理：对于与大数据相关的质量问题（数据价值不高或不显著），是否达成了一致意见? ✍ 信息生命周期管理：是否制定了流程，根据法律和业务要求合法处理不再需要的大数据? ✍ 信息安全和隐私：首席信息安全官是否是大数据治理计划的关键支持者?
支持准则	✍ 数据架构：结构化和非结构化数据系统及应用的架构式设计，用于实现数据的可用性，并将数据分配给合适的用户。 ✍ 元数据：指用于创建常见的语义定义、 IT术语、数据模型和数据库的方法和工具。 ✍ 审计信息日志和报告：指监测和测量数据价值、风险和信息治理有效性的组织流程。	✍ 数据架构： Hadoop 、NoSQL 以及与当前架构相关的其他新兴大数据技术的共存战略是怎样的? ✍ 分类和元数据：业务词库是否包含与大数据相关**的关键业务术语（如针对点击流数据的“独立访客”）**? ✍ 审计信息日志和报告：企业如何检测特权用户对医保索赔和通话详单等敏感大数据的访问?

案例5.2 某大型金融机构资金管理部的大数据治理路线图

某大型金融机构的资金管理部，为大中型企业提供现金管理和流动性管理的综合服务。该部门处于部署大数据计划的早期阶段，其最初的大数据治理路线图如右图所示：



第1-6个月
构建技术基础设施，获得Linux服务器和Apache Hadoop发行版。
由于大数据是一个新事物，在切入业务前，必须设计一个可行的用例，并进行财务可行性论证。

第6-12个月
引入详细的交易记录，以分析每日头寸快照。
受传统基础设施成本高昂的影响，以往的金融机构从未进行这样细致入微的分析。

第12-24个月
将社交媒体数据和其他非结构化内容引入Hadoop环境。
由于金融机构的大多数客户是大企业，对交易对手的10-K和10-Q归档等非结构化内容，进行探索性分析。

第24-36个月
资金管理部已经有了现成的聚焦于大企业客户的主数据的信息治理计划。

此外，组织要认真审视数据管理的传统方面：
怎样将数据导入并导出Hadoop?
Hadoop中的数据质量如何?
大数据的元数据是怎样的?
如何将大数据整合到未来12个月将要部署的主数据管理数据库中?

大数据已经成为主流媒体的热门词汇，高管层至少很有可能同意支持一个大数据试点项目。因此，数据治理团队需要及时更新路线图，将与大数据有关的人员、流程和技术计划纳入其中。

大数据处理框架的组成

大数据类型

大数据治理需要高度聚焦于数据本身。我们将大数据分为五种：web和社交媒体数据、机器对机器的数据、大体量交易数据、生物计量学数据和人工生成的数据。

信息治理准则

传统的信息治理准则，同样适用于大数据，相关准则包括组织、元数据、隐私、数据质量、业务流程整合、主数据整合和信息生命周期管理。

产业与功能

大数据分析是受例驱动的，用例的具体情况因产能和功能而异。限于篇幅，我们只列出了部分的产业和功能。

大数据的类型

大数据大体可分为五种类型

6.1职责分配（RACI）所代表的内涵

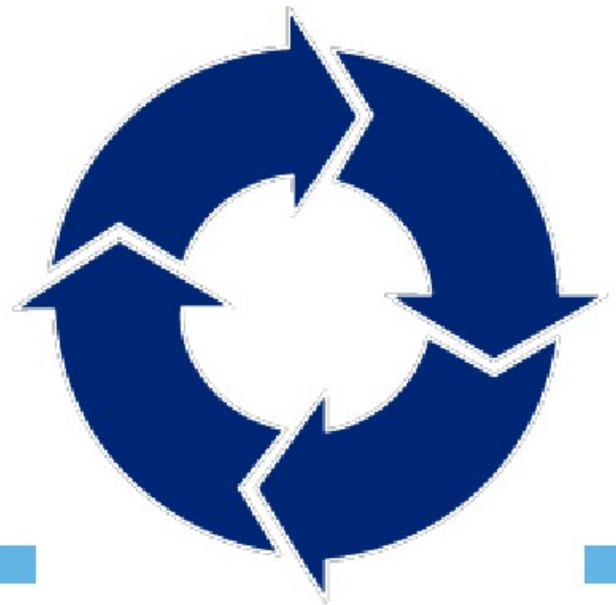
应负责方（Responsible）

指授权管理某属性的人。

（一种属性可有多个负责人）

最终负责方（Accountable）

指数据属性承担最终责任的人。



咨询方（Consulted）

指通过双向沟通接受咨询的某人或某些人。

被告知方（Informed）

指通过单向沟通被告知的某人或某些人。

7大数据治理计划需要实施的最佳实践

元数据是描述数据产品特征的任何信息，如名字、位置、可感知的、重要性、质量、对企业的价值，以及与企业认为值得管理的其他数据产品的关系等。元数据决定信息架构的如何满足业务需求，因此元数据是信息治理计划的关键。

1

创建一个体现关键大数据术语的业务定义的词库。

2

理解对Apache Hadoop中元数据的持续支持。

3

对业务词库中的敏感大数据进行标记。

4

从相关的大数据存储中输入技术元数据。

5

将相关的数据元与业务词库中的术语进行链接。

6

使用运营元数据监测大数据的流动。

7

保留技术元数据，以支持数据血统和影响分析。

8

从非结构化文件中采集元数据，支持企业搜索。

9

扩展既有的元数据角色，将大数据纳入其中。

7.1业务词库

业务词库

业务词库是企业用于传达其对信息的认识的语言。创建并维护该层业务元数据，对表达要求的含义和描述IT系统可用的信息至关重要。

业务词库保证了信息开发的准确性和速度。

术语代表着企业和业务层面对信息的理解，所以许多组织倾向于自下而上创建数据词典，对已有的信息进行归类。

在处理大数据时，业务驱动的数据定义和数据目录之间的区别尤为重要。

7.3对业务词库中的敏感数据

进行分类大数据治理计划需要对社会保险号码等敏感数据进行分类。分类应来自业务词库模型并被传承到不同数据库中数据的所有物理实例中。

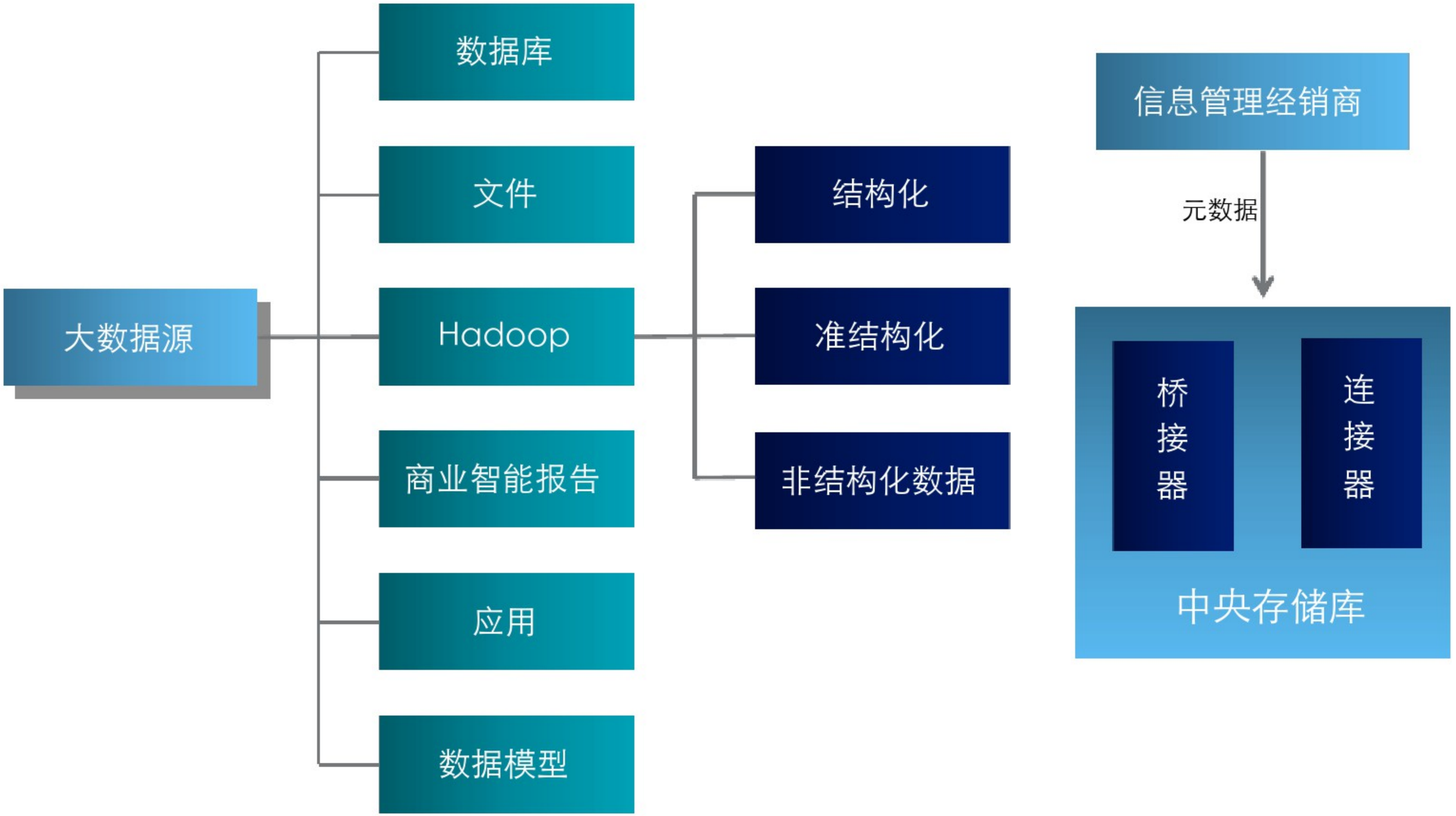
敏感的大数据可能隐藏在非结构化文本中。大数据治理计划应考虑数据分析工具的利用，以便自动发现非结构化字段的敏感数据。

首席信息安全官制定有关敏感数据的政策。只有在识别到敏感数据的位置时，组织才能执行政策，因此，在业务词库中标记敏感数据就非常关键。

大数据治理团队可以通过使用数据分析工具发现敏感的大数据，以监督对政策的遵从度。

从相关的大数据存储中输入技术元数据

在创建业务词库后大数据治理团队需要从 大数据源 中采集合用的、相关的元数据。



7 元数据


元数据
创建一个体现关键大数据术语的业务定义的词库
理解对Apache Hadoop中元数据的持续支持
对业务词库中的敏感大数据进行标记
从相关的大数据存储中输入技术元数据
将相关的数据元与业务词库的术语进行链接
使用运营元数据监测大数据的流动
保留技术元数据，以支持数据血统和影响分析

从非结构化文件中采集元数据，支持企业搜索

创建非结构化数据的索引，也是元数据的一种形式，许多企业的搜索供应商已开发相应工具。

保险业	通过向呼叫人员提供客服关怀、告警、保单和客户信息文件等多个文件库的可搜索访问，可将平均处理时间减少三秒，年节约数百万美元。
制药业	通过提供对EMC Documentum、文件系统、微软Share-Point、内网和外部数据库中客户、患者和研究数据的快速访问，加快科研进程。
医疗保险业	让临床医生可访问来自医学刊物和其他文件库的最新研究成果。

7.9 拓展既有的元数据角色，将大数据纳入其中

信息治理团队可能安排许多与原数据相关的角色。组织需考虑这些角色进行拓展，以将大数据治理纳入进来。

业务词库管理者	本角色负责保管应将大数据术语包含在内的业务词库。
元数据管理者	本角色负责在相关数据源识别和输入技术元数据。
数据血统管理者	数据血统管理者与数据管理者配合，确保数据血统分析中数据源之间的数据流可得到准确地反映。
数据主管	本角色参与大数据特别是关键业务术语定义的管理。
数据架构师	本角色监督元数据模型的创建及其与企业数据模型的连接。
数据科学家	本角色缩短了大数据原始卷和使其有用的业务洞察间的距离，其通过创造力和想象力创建原型，以揭开大数据中的秘密。

9 大数据质量

数据质量管理是测度、提高、验证质量以及整合组织数据的方法等一套行为准则。体量极大、速度极快和多样的特点，决定了大数据质量所需的处理有别于传统信息治理计划的质量管理。

维度	传统数据的质量	大数据的质量
处理频率	处理是面向批量的	处理是实时的或面向批量的
数据多样性	数据格式大部分是结构化的	数据格式可能是结构化的、准结构化的或非结构化的
置信度	数据需处于原始阶段，以方便数据仓库的分析	糟糕的数据质量可能会阻碍分析工具获得业务洞察
数据进化的时间选择	在下载到数据仓库前数据需要进化	数据的体量和速度可能要求采取流式的、内存中的分析来进化数据、从而降低存储要求
关键数据元素	评估客户地址等关键数据元素的数据质量	数据可能被模糊定义或错误定义，关键数据元素可能会反复变化
分析位置	数据迁移到数据质量和分析引擎	数据质量和分析引擎可进入数据中，以保证可接受的处理速度
管理工作	数据主管可管理大部分数据	由于体量大和速度快，数据主管只能管理相对更小的数据

大数据治理计划必须采取的实践

9.1 与商业上的利益攸关者协作，建立并测度大数据质量的置信区间

9.2 利用准结构化和非结构化数据，提高人口稀疏的结构化数据的质量

9.3 使用流数据分析技术解决内存中的数据质量问题，无需将中间结果输入硬盘

9.4 任命对信息治理委员会负责的主管，由其负责提高

10 业务流程整合

10.1

? 识别将会受到大数据治理影响的关键流程

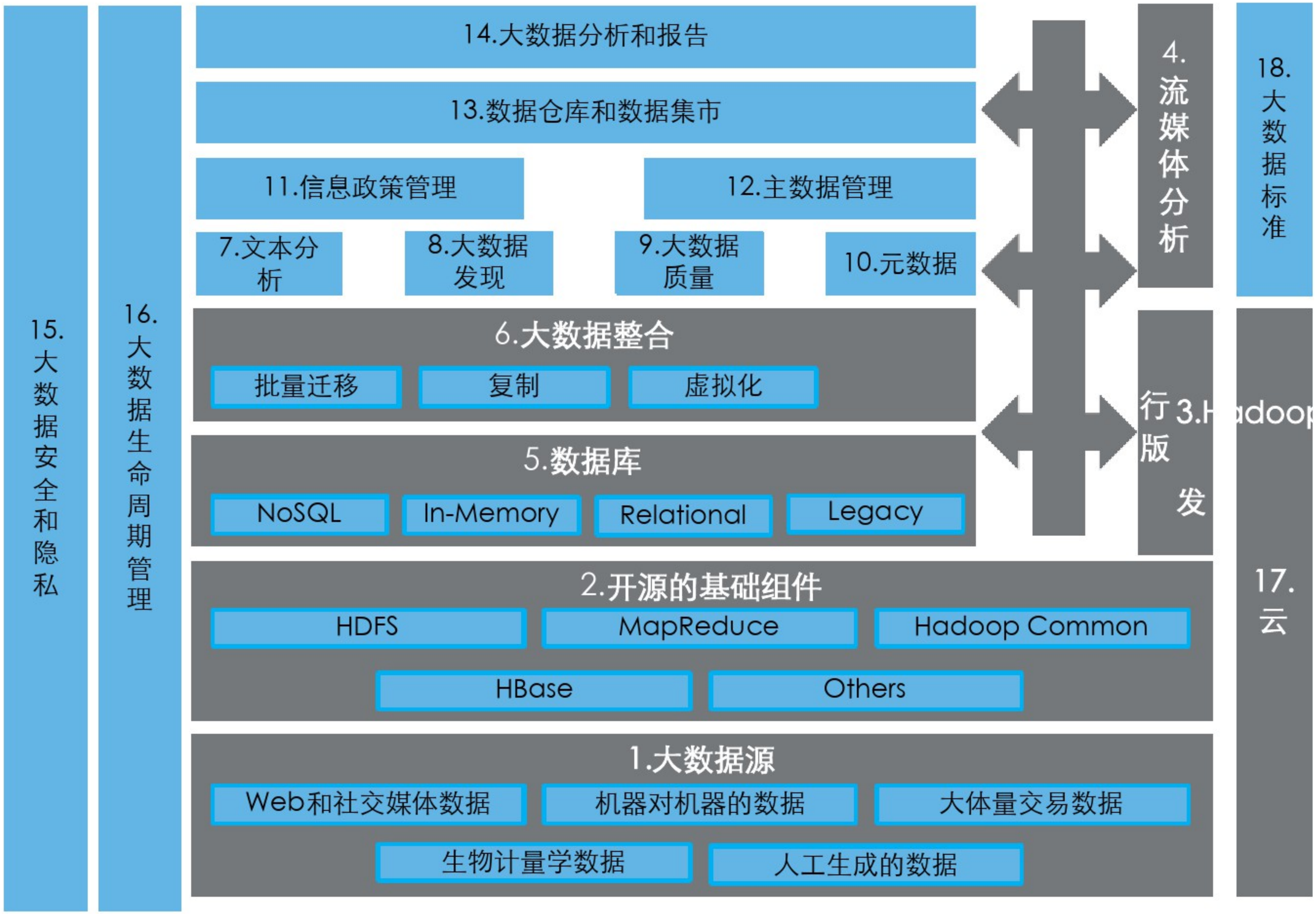
10.2

? 建立关键合同的流程图

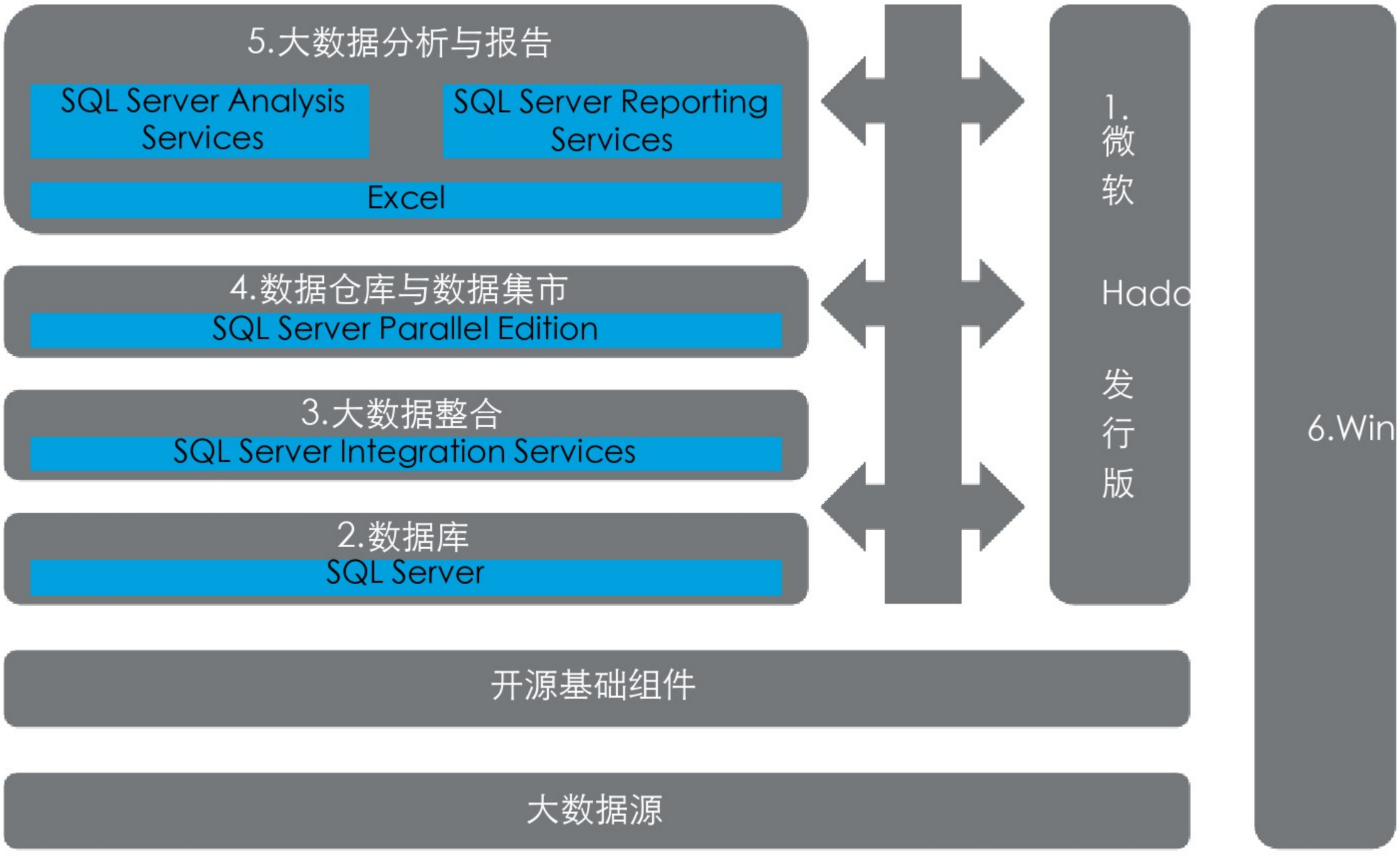
10.3

? 针对业务流程中的关键步骤，制定大数据治理政策

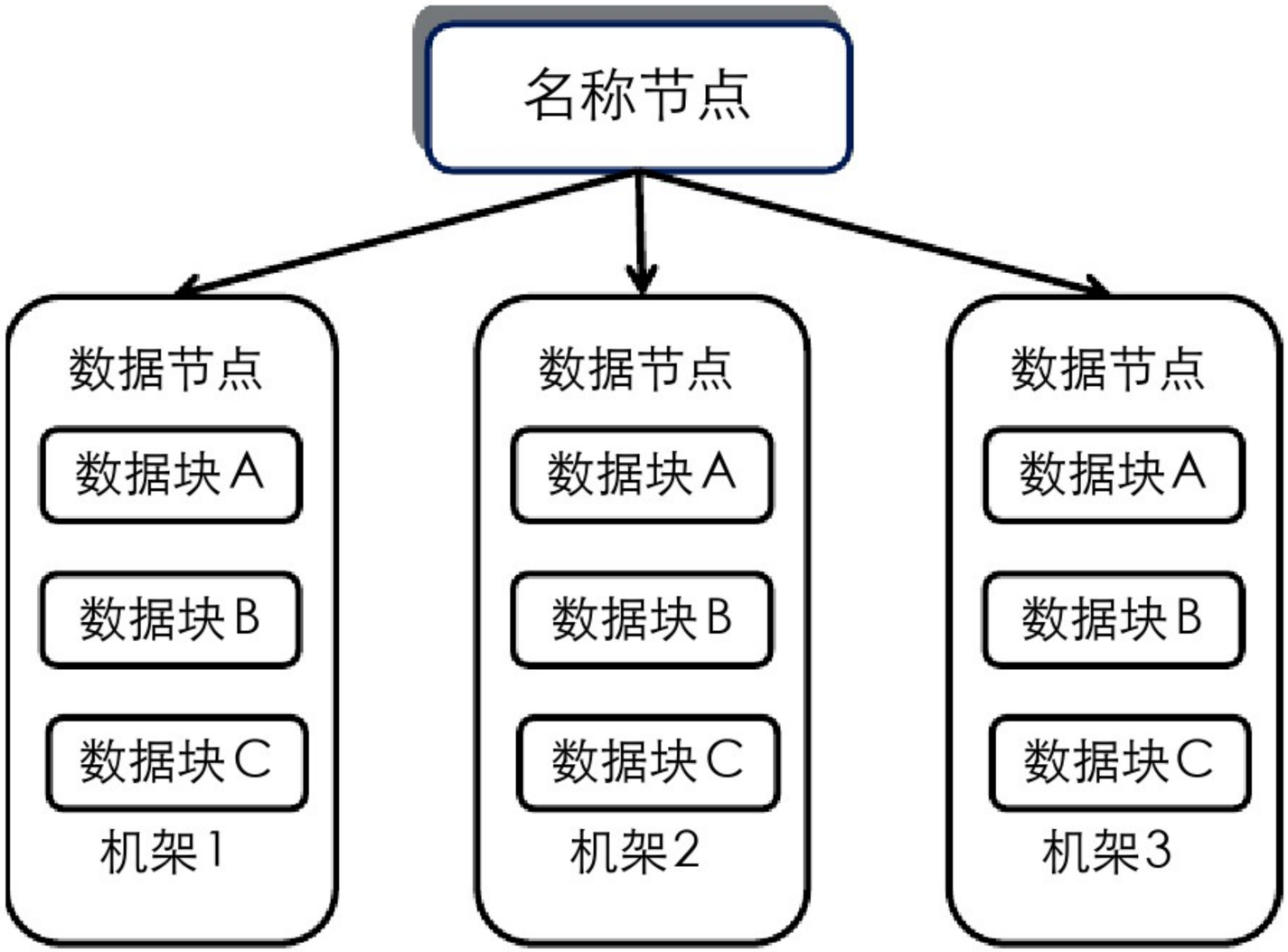
图21.1 大数据技术参考架构



微软的大数据平台


5.大数据分析与报告
SQL Server Analysis Services
SQL Server Reporting Services
Excel
4.数据仓库与数据集市
SQL Server Parallel Edition
3.大数据整合
SQL Server Integration Services
2.数据库
SQL Server
1.微软发行版
Hado
6.Win
开源基础组件
大数据源

理解对Apache Hadoop中元数据的持续支持



图：Hadoop分布式文件系统（HDFS）的技术构架

作为Hadoop关键支持要素的元数据

如图Hadoop分布式文件系统（HDFS）是一个带单个名称节点和多个数据结点的主/从架构。

单点故障

因为HDFS很容易受到名称节点故障的损害，所以Hadoop经销商建议管理者存储一些不同本地硬盘的备份

可拓展性

随着数据存储动能的扩大，主服务器名称节点可能出现可拓展性的问题，主服务器名称节点必须将所有元数据保存在内存中。

HCatalog

Hcatalog项目是Apache孵化器的 一部分，旨在解决Hadoop中缺乏元数据支持的问题。

大数据安全与隐私

部分漏洞

- HDFS没有授权系统，注册用户可以在群中读写任何数据
- Hadoop注册用户通过“whoami”命令访问，这是不安全的
- Hbase没有访问控制，Hadoop群中任何工作运行均可以访问群中任何数据
- ……

变通方案

- 不要在Hadoop中存储任何敏感数据
- 对敏感数据进行加密，包括隐藏文本和非结构欧化领域的内容
- 将每个数据置于自己的群中，以便用户仅可以访问被授权的数据
- ……

Hadoop是一项新技术，我们预计随着大公司和供应链的介入，上述问题将被得到解决。