ICS 35.240.70
L67

中华人民共和国国家标准

GB/T XXXXX—XXXX

信息技术 大数据 技术参考模型

Information Technology — Big Data — Technical Reference Model

（征求意见稿）

（在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

目　次

前　言

本标准按照GB/T 1.1-2009给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由全国信息技术标准化技术委员会（SAC/TC28）提出并归口。

本标准起草单位：　。

本标准主要起草人：　。

信息技术 大数据 技术参考模型

1 范围

本标准规范了大数据的基础通用模型，包括大数据角色、活动和功能组件以及它们之间的关系。

本标准适用于理解大数据领域的复杂操作，是讨论需求、结构和操作的有效工具，并为大数据系列标准的制定提供了依据。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T AAAA 信息技术 大数据 术语

3 术语和定义

GB/T AAAA 规定的以及下列术语和定义适用于本文件。

3.1 大数据技术参考模型 Technical Reference Model

一种新的数据应用系统范式的抽象。

3.2 大数据参考架构 big data reference architecture

一种用作工具以便于对大数据内在的要求、设计结构和运行进行开放性探讨的高层概念模型。

注：比较普遍认同的大数据参考架构一般包含系统协调者、数据提供者、大数据应用提供者、大数据框架提供者和数据消费者等5个逻辑功能构件。

3.3 系统协调者 System Orchestrator

大数据参考架构中的一种逻辑功能构件，它定义所需的数据应用活动并将它们整合到可运行的垂直系统中。

注：系统协调者可以是人、软件或这二者。

3.4 数据提供者 Data Provider

大数据参考架构中的一种逻辑功能构件，它将新的数据或信息引入大数据系统。

3.5 大数据应用提供者 Big Data Application Provider

大数据参考架构中的一种逻辑功能构件，它执行数据生命周期操作，以满足系统协调者定义的需求以及安全和隐私保护需求。

3.6 大数据框架提供者 Big Data Framework Provider

大数据参考架构中的一种逻辑功能构件，它建立一种计算框架，在此框架中执行转换应用，同时保护数据完整性和隐私。

3.7 数据消费者 Data Consumer

大数据参考架构中的一种逻辑功能构件,它是使用大数据应用提供者提供的应用的末端用户或其他系统。

4 符号和缩略语

POSIX：可移植操作系统接口（Portable Operating System Interface）

5 大数据技术参考模型的目的和目标

本标准中的大数据技术参考模型提供了一个体系框架,用于有效描述大数据角色、活动和功能组件。

大数据技术参考模型目的包括：

——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；

——为解决相似的问题集提供一致的技术实现方法。

大数据技术参考模型的目的是为了方便对大数据复杂性操作的认识。它不代表一个特定的大数据系统的系统架构；相反，它是一种工具，使用通用的架构来描述、讨论和开发特定系统的架构。

大数据技术参考模型是一个通用的大数据系统概念模型，对于讨论大数据需求、结构合操作，它是一种有效的工具。该模型不依赖于任何特定的产品和服务供应商，也不定义规范的解决方案。

大数据技术参考模型支持以下标准化目标：

——在一个与供应商和技术无关的大数据高层概念模型语境下，增进对大数据构件、处理过程及系统的理解；

——为政府部门、相关机构和其他用户在理解、讨论、分类和比较大数据解决方案的过程中提供技术参考；

——促进对大数据互操作性、可移植性、可重用性和可扩展性的备选标准的分析。

6 大数据技术参考模型概述

本标准定义的技术参考模型为大数据标准化提供了基本参考点,为大数据系统的基本概念和原理提供了一个总体框架（如图2）。

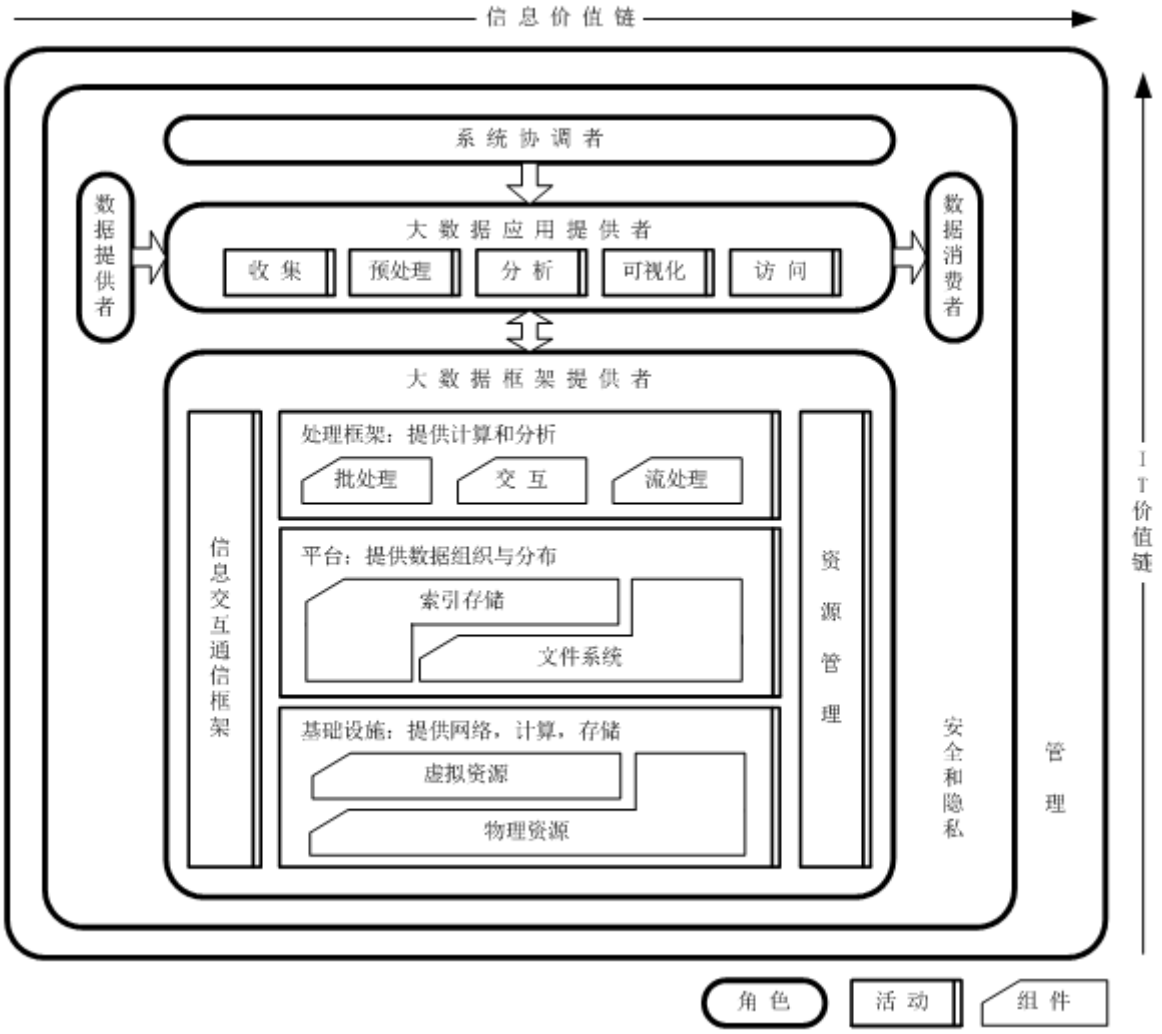

图1　大数据技术参考模型

大数据技术参考模型围绕代表大数据价值链的两个维度组织展开：信息价值链（水平轴）和IT价值链（垂直轴）。信息价值链表现大数据作为一种数据科学方法对从数据到知识的处理过程中所实现的信息流价值。信息价值链的核心价值通过数据收集、预处理、分析、可视化和访问等活动实现。IT价值链表现大数据作为一种新兴的数据应用范式对IT技术产生的新需求所带来的价值。IT价值链的核心价值通过为大数据应用提供存放和运行大数据的网络、基础设施、平台、应用工具以及其他IT服务实现。大数据应用提供者位于两个价值链的交叉点上，大数据分析及其实现为两个价值链上的大数据利益相关者提供特定价值。

大数据技术参考模型提供了一个构件层级分类体系，用于描述技术参考模型中的逻辑构件以及定义逻辑构件的分类。大数据技术参考模型中的逻辑构件被划分为三个层级，从高到低依次为角色、活动和组件。最顶层级的逻辑构件是代表大数据系统中存在的五个角色，包括系统协调者、数据提供者、大数据应用提供者、大数据框架提供者、数据消费者五个角色。另外两个非常重要的逻辑构件是安全和隐私以及管理，它们为大数据系统的五个角色提供服务和功能。第二层级的逻辑构件是每个角色执行的活动。第三层级的逻辑构件是执行每个活动需要的功能组件。

该模型可以用于表示由多个大数据系统组成的堆叠式或链式系统，其中一个系统的数据消费者可以作为后面一个系统的数据提供者。

该模型支持各种商业环境，包括紧密集成的企业系统和松散耦合的垂直行业，有助于理解大数据系统如何补充并有别于已有的分析、商业智能、数据库等传统的数据应用系统。

7 角色

7.1 系统协调者

系统协调者职责在于规范和集成各类所需的数据应用活动，以构建一个可运行的垂直系统。

系统协调者具体功能包括：配置和管理大数据技术参考模型中其它组件执行一个或多个工作负载，以确保各工作项能正常运行。负责为其它组件分配对应的物理或虚拟节点并对各组件的运行情况进行监控，并通过动态调配资源等方式来确保各组件的服务质量水平达到所需要求。

具体实现层面，系统协调者的功能可由管理员、软件或二者的组合以集中式或分布式的形式实现。

7.2 数据提供者

数据提供者的职责是将数据和信息引入到大数据系统中，供大数据系统发现、访问和转换。

其具体活动包括：

——收集、固化数据。

——创建描述数据源的元数据。

——发布信息的可用性和访问方法。

——确保数据传输质量。

数据提供者和应用提供者的接口涉及三个阶段：开始、数据传输和终止。

7.3 大数据应用提供者

大数据应用提供者的职责是通过在数据生命周期中执行的一组特定操作，来满足由系统协调者规定的要求，以及安全性、隐私性要求。

大数据应用提供者包括收集、预处理、分析、可视化和访问五个活动。

7.3.1 收集

负责处理与数据提供者的接口。

7.3.2 预处理

包括数据验证、清洗、移除异常值、标准化、格式化和封装。

7.3.3 分析

基于数据科学家的需求或垂直应用的需求，确定处理数据的算法来产生新的分析，解决技术目标，从而实现从数据中提取知识的技术。

7.3.4 可视化

提供给最终的数据消费者处理中的数据元素和呈现分析功能的输出。

7.3.5 访问

与可视化和分析功能交互，响应应用程序请求，通过使用处理和平台框架来检索数据，并响应数据消费者请求。

7.4 大数据框架提供者

大数据框架提供者的职责是为大数据应用提供者在创建具体应用时使用的资源和服务。

大数据框架提供者包括基础设施、平台、处理框架、信息交互/通信和资源管理五个活动。

7.4.1 基础设施

为大数据系统中的所有其他要素提供必要的资源，这些资源是由一些物理资源的组合构成，这些物理资源可以控制/支持相似的虚拟资源。这些资源分为下面几类：

——网络：从一个资源向另一个资源传输数据的资源。

——计算：用于执行和保持其他组件的软件的实际处理器和存储器。

——存储：大数据系统中保存数据的资源。

——环境：在建立大数据实例的时候必须考虑的物理厂房资源（电力、制冷等）。

7.4.2 平台

包含逻辑数据的组织和分布，支持文件系统方式存储和索引存储方法。

——文件系统：实施某种级别的 POSIX 标准以获取权限，进行相关的文件操作。

——索引存储：无需扫描整个数据集，便可以迅速定位数据的具体要素。

7.4.3 处理框架

提供必要的基础设施软件以支持实现应用程序能够满足数据数量、速度和多样性的处理。包括批处理、流处理，以及两者的数据交换与数据操作。

7.4.4 信息交互/通信

包含点对点传输和存储转发两种通信模型。在点对点传输模型中，发送者通过信道直接将所传输的信息发送给接收者；而在后者中，发送者会将信息先发送给中间实体，然后中间实体再逐跳转发给接收者。点对点传输模型还包括多播这种特殊的通信模式，在多播中，一个发送者可将信息发送给多个而不是一个接收者。

7.4.5 资源管理

计算、存储及实现两者互联互通的网络连接管理。主要目标是实现分布式的、弹性的资源调配，具体包括对存储资源的管理和对计算资源的管理。

7.5 数据消费者

通过调用大数据应用提供者提供的接口按需访问信息，誉其产生可视的，事后可查的交互。

7.6 安全和隐私

在安全和隐私管理模块，通过不同的技术手段和安全措施，构筑大数据平台全方位、立体的安全防护体系，实现覆盖硬件、软件和上层应用的安全保护，从网络安全、主机安全、应用安全、数据安全四个方面来保证大数据平台的安全性。

——网络安全：通过网络隔离，保证数据处理、存储安全和维护正常运行。

——主机安全：通过对集群内节点的操作系统安全加固等手段保证节点正常运行。

——应用安全：具有身份鉴别和认证、用户和权限管理、数据库加固、用户口令管理、审计控制等安全措施。

——数据安全：从集群容灾、备份、数据完整性、数据分角色存储、细粒度数据加密等方面保证用户数据的安全。

同时应提供一个合理的灾备框架，提升灾备恢复能力，实现数据的实时异地容灾功能，跨数据中心数据备份，克服单数据中心带来的数据风险。

7.7 管理

提供大规模集群的统一的运维管理系统，能够对包括数据中心、基础硬件、平台软件（存储、计算）和应用软件进行集中运维、统一管理，实现安装部署、参数配置、监控、告警、用户管理、权限管理、审计、服务管理、健康检查、问题定位、升级和补丁等功能。

具有自动化运维的能力，通过对多个数据中心的资源进行统一管理，合理的分配和调度业务所需要的资源，做到自动化按需分配。同时提供对多个数据中心的 IT 基础设施进行集中运维的能力，自动化监控数据中心内各种 IT 设备的事件、告警、性能，实现从业务纬度来进行运维的能力。

对主管理系统节点及所有业务组件中心管理节点实现高可靠性，双机机制，采用主备或负荷分担配置，有效避免了单点故障场景对系统可靠性的影响。

大数据时代的隐私性主要体现在不暴露用户敏感信息的前提下进行有效的数据挖掘，这有别于传统的信息安全领域更加关注文件的私密性等安全属性。根据需要保护的内容不同，隐私保护又可以细分为位置隐私保护、标识符匿名保护和连接关系匿名保护等。