The Definitive Guide to the Data Lake

1. **引言**

如果说围绕数据湖的炒作是轻描淡写的，也许这是数据行业对流行语需求的一个固有结果：一个术语在没有明确的定义和可重复的商业价值之前就流行起来，这种情况并不少见。当一些如“大数据”，“数据池”，甚至“数据仓库”等概念初次在工业界出现的时候，我们已经看到这种现象很多次了。这个术语既接受了专家们对数据沼泽喋喋不休的审查，同时也经受了对数据湖将会对企业数据架构产生深刻影响的潜力。随着数据湖开始离开炒作周期，并且面临着务实的IT和业务利益相关者的压力，对于数据湖清晰的定义，用例和最佳实践的需求还在持续增长。

本文旨在通过结合数据和信息管理的基本原理和现有实现的经验，阐明数据湖的概念，以说明当前的数据体系结构将如何转变为现代数据体系结构。现代的数据架构包括Hadoop以及它的生态系统，集成了数据仓库，面向发现的环境， 高度专业化的分析或操作数据技术。因此，数据湖已经称为企业数据管理转型的隐喻，并将根据既定的原则、驱动因素和最佳实践发展数据湖的定义，这些原则、驱动因素和最佳实践将很快作为事后的领悟并应用在公司中。

1. **数据湖的定义和视角**

随着数据湖概念成为核心数据架构的一部分，问题是数据湖是一种架构策略还是一种架构目标，答案是两者都是。通过将数据湖定义为各种数据工作负载（内部和部分）的企业数据的集中存储库，我们解决了最终状态体系结构，并为数据体系结构相关决策在实现数据湖临界质量的过程中提供了指导。许多公司智慧地将数据湖的概念应用到拉尔夫·沃尔多·爱默生的名言“生命是一个旅程，而不是一个目的地。”并且认识到数据湖是一个独特旅程的终极目的地，它根据采用公司的驱动因素和优先级，遵循自己的节奏和方向。基于我们正在公司中看到的架构应用模式，我们可以把数据湖的应用之旅分为四个商业应用过程。第1阶段包括“踢轮胎”大数据试点项目，这些项目侧重于特定的业务成果，作为使用和管理Apache Hadoop环境的介绍。第二阶段是一个更为“反动”的方法，因为公司开始集中精力利用Hadoop的优点来解决现有架构低效性，并隔离出清晰，快速和可衡量的商业价值的机会。其中的一种模式是利用Hadoop的可扩展性，低开销的持久层或者它进行大数据处理和分析的能力。一个例子，最近接受了将历史数据仓库数据重新定位到数据湖中，作为一个在线存储扩展，具有较低的服务级别协议，以减少数据仓库的总体规模和管理。其他部署可以提供潜在的长期价值，例如将操作源数据采集的数据仓库暂存区重新定位到数据湖中，以便数据仓库集成可以将获取的数据作为长期原始数据（以及数据未使用的附加数据）源化和持久化仓库）以实现数据科学家和分析开发所需的可承受性和数据处理能力。

在第三阶段，组织从一个被动的方法转化为一个主动的方法。这一阶段包括进一步整合大数据和分析项目数据的计划，并利用Hadoop可负担的可扩展性，来使得大量的社交网络平台，物联网数据，在一个单一的，集中的仓库中的非结构化数据，它关注并鼓励重用但是又尽量避免数据重复。最后，第四阶段是连续优化阶段。不可避免地，当数据湖成为IT战略和计划的核心组成部分的时候，将会产生一个爆发点。操作应用程序和分析应用程序之间的间隔被重新思考为单个的企业平台，他利用了每个架构组件的优势。考虑到第四阶段是Hadoop实现企业数据体系结构策略的基础组件，并通过持久性和数据引擎层支持更多的操作、分析和大数据工作负载。

最终，数据湖的目标是构建一个高效的企业数据架构，带有一个单一的企业数据仓库，可以满足不同的企业应用程序负载，通过数据的可重用性和一致性来提高效率。数据湖将通过整合在成熟阶段之后继续增长（以避免在第1阶段中出现多个Hadoop集群筒仓），并努力寻求关键的质量和动量。然而，任何公司加快采用数据湖的速度都会受到数据监管，安全，角色和访问权限等相关的建设速度的限制。

拥有一个高效的现代数据架构所带来的益处的同时，必然会伴随着产生相应的风险。因此，如果数据湖注定要成为公司核心数据架构的一部分，那么它必须满足这些数据管理的每一个严格要求。数据湖战略应该将数据监管扩展到包括大数据，数据发现，数据科学用例和角色。推动业务用例的发展，Hadoop作为发现平台需要访问新的和现有的数据源。然而，通过不受限制地访问所有数据来发现数据对于数据治理来说是很难接受的。因此，可信和可靠的安全和监视方法需要特权、角色和数据组之间的关系，这些数据组遵循数据所有者和责任的规则。

而且，企业需要认识到，他们自己基于原则的数据治理和安全方法将需要更多的工作来维护，直到广泛的企业采用开始建立最佳实践和成功的案例研究。这也是数据湖的领导者将为数据湖的跟随者创造最佳的实践，随着技术不断快速发展以满足企业需求。

1. **明晰数据湖组织**

Hadoop的灵活性和可扩展性使得数据湖的概念续力到下一代，可以被企业使用的中心化的数据仓库。作为在数据架构中的持久层，所有形式的数据都可以通过Hadoop数据引擎的公共访问层进行维护、编目、探索和利用，而Hadoop的数据引擎又是由Hadoop的交互式和多代理的另一个资源协商者（YARN）所驱动的。防止数据湖变成数据沼泽的方法是数据监管（带有元数据和数据世系支持的数据所有权策略），平衡数据发现和操作使用的是数据管理。 组织数据和加强安全性被认为是敏捷性、自由性和数据发现的对立面。然而，在风险、回报和优化之间取得平衡是每个公司的个人决定。进一步加强数据管理的核心原则， 例如最小化数据重复和实现数据重用。数据湖必须包含多租户和总体资源管理，这些管理可以按业务优先级进行逻辑处理，包括数据分类、各种数据应用程序类型和其他特殊注意事项。随着时间的推移，包含操作系统工作负载的科学工作负载，从而开始将数据移出数据池的过程最小化，而不是将工作转移到数据池中的数据。 企业面临的一个挑战将是确保数据湖保持单一，外围Hadoop集群和数据存储不会损害完整的多租户和健壮的单一的数据湖环境的全面资源管理的好处。

1. **按数据分类组织**

首先，数据湖中的所有数据将分为三种基本分类之一：原始、派生或聚合。 原始的数据可以被用于初始的工程，然而，不能总是预期来自同一原始数据的持久化历史的未来需求。通常地，获取原始数据会通过进一步的分类、派生或增强进行扩展，从而有利于后续活动，如简化数据集成或增强发现或数据科学。出于这个原因，一定要尽可能多地收集历史记录，以获取详细的原始数据。（例如，将操作系统数据提取到Hadoop中，可以使得数据仓库和数据市场获益成为一个中转地），可以作为一个回退数据集如果数据集成代码被发现有缺陷或者更改的话。第二种分类是派生数据，涉及基于数据集成、清理、计算或业务规则的任何数据，这些数据可以使企业中的其他人受益于可重用性。最后，聚合的数据表示结果集，可消费信息和从消费者视角的上下文—商业用户、消费者和应用程序。

1. **按用途组织**

用途的主要区别是在数据湖中的发现环境和生产环境。在基本的数据分类之后，下一个子组织成为用于发现和评估的原始数据的新数据源。这些新的子分类可以用于个人或者合作团队的分析，这是探索新数据源的地方，然后，如果认为有价值，就考虑纳入整个数据湖。其他的操作和分析系统的用例已经被很好地理解了。（例如，派生或集成的数据仓库的历史数据可以被保存为优化的模式用于数据访问。）最终地，数据监管决定可以将哪些数据带入数据湖，以及可以组合哪些数据进行发现和分析。此外，现代数据体系结构将减少复制和移动操作数据的需要应用中采用YARN作为数据操作系统和单一数据域来满足数据湖中的操作和分析应用程序工作负载。YARN处理引擎(如Storm和Spark)的出现和成熟支持了这种范式转换的采用。清晰的，有逻辑的数据组织将有利于Hadoop工程师在物理分区、压缩或者压缩数据文件和文件类型以符合使用模式方面的技术决策。

1. **影响数据湖组织的特殊考虑**

可以影响数据湖的特殊考虑包括邻近性，遵从性和安全性。邻近性的概念是在网络间最小化数据移动，如果主要的数据摄入是基于云的公开数据，一些企业可能会考虑将数据湖逻辑上分为一个云公开的部分和一个企业内部的部分。同样关键的是企业的合规性需求。这可能需要维护在室内的，在数据湖中一部分的，带有安全性需求的私有数据进行强制地隔离。对数据安全性的进一步要求可能不仅包括访问和监视，还包括数据加密和持久化的数据与最终用户之间的传输。

无论是在云中还是在本地，发现环境都可以用于评估新的或外部数据，这些数据稍后将被治理并合并到数据湖中。在一些情况下，公司发现按需的Hadoop环境是一种更容易的方法，可以在将新数据合并到更大的数据湖之前对其进行研究。Hadoop作为数据采集门槛低的优势是采用数据湖的前驱。这样做的代价是，这些随需应变的环境不能从数据湖中提供所有数据所带来的完整数据多产性中获益。

1. **数据湖整合数据发现，数据科学和企业BI**

数据湖的一个好处是，上面列出的三种数据分类能够基于集中的数据存储库为多种数据应用程序类型提供服务。除了Hadoop满足大数据应用程序的需求之外，数据湖满足了企业数据架构支持企业BI需求（无论是通过使用依赖于Hadoop的内部或外部数据仓库的持久暂存。）数据发现和数据科学，统计功能和机器学习算法的原始数据（以及大数据应用和非结构化分析的明显便利）。企业特别感兴趣的是企业BI和数据发现之间的关系。经过验证的BI方法基于性能管理的指标和业务主题领域的上下文，以实现业务预定义的一致性。因为这些定义被转化为数据模型和模式，然后就可以将操作数据转换为这些指标。过去的十年我们看到，敏捷BI方法作为对业务的响应，并不总是能够完全预定义或理解操作源数据的质量。敏捷BI方法提供了一个应对业务波动和敏捷性的更好的方法。使用数据湖可以使数据仓库暂存区更加详细地持久化，操作系统的原始事务历史，可以提取并转换为独立企业数据仓库和数据集市的详细业务事务、度量和指标，新项目还可以从中获得历史的原始数据。此外，除了大数据主题领域之外，数据湖还可以作为保存历史上很少访问的数据仓库中数据的地方。企业的BI方法，在BI需求分析阶段对发现、分析和可行性进行快速数据探索时，可以从数据湖获得更多好处。最后，对于原始的、派生的和聚合的数据属于大数据领域的用例，数据湖也很容易为那些生产环境提供便利。

对于今天在竞争性的商业环境中快节奏的变化，公司正在认识到发现的关键作用，发现他们不知道的问题的答案与预先定义的度量标准、业务规则和企业中当前存在的定义同等（甚至更）重要。这促使分析师直接处理数据，以发现对业务有意义的定义。数据湖中健壮且高效的发现环境有助于发现新的业务定义、见解和分析，这些可以通过治理作为企业或部门的视角加以制度化。

1. **数据湖成功的关键因素**

数据湖随着业务的增长而产生复合价值，并成为整个企业数据战略的一部分。然而，当前的数据管理思想要求对数据的价值合理化，而不是量化数据的潜在价值。换句话说，一种更长期的数据存储方法决定了一种新的思维方式，即通过数据湖实现临界质量。以数据和分析为中心的公司对获取和分析尽可能多的数据的痴迷所驱使。

1. **重新考虑长期数据**

几个原则来获取数据湖的临界质量可以被应用到每一个数据工程。首先，每一个数据工程都应该考虑和认识到其数据在未来的应用和发现中具有可重用性，同时理解接下来的数据需求可能是未知的。这是平衡数据治理的需要，一直决定数据湖保留后，当前强化的业务驱动框架流程已经消耗了原始数据。第二，考虑到与所有数据相关的风险管理以及数据湖对数据吸收的较低门槛，企业中的信息政策通过考虑获取新数据，从而定义和分配数据所有者，以便以后数据管理员和专家可以随时获得更多的历史记录。这些正在进行中的当局项目和代表团现在应该扩大到可以从历史数据中获益的项目。第三，作为中心化策略的一部分，通过增加数据使用的多样性及其数据湖的概念，将避免在整个数据中存储多种重复数据和移动数据的成本高昂的固有能力，数据治理框架本身应该是企业的。Hadoop 2.0 YARN的到来，从根本上改变了Hadoop从一个独占的、面向批处理的作业系统，变成能够管理访问HDFS中数据的许多数据处理引擎的总体资源（支持批处理、交互式和流式应用程序）。 虽然集群管理非常需要健壮的资源管理，但YARN的意义在于它可以在单一环境中简化当前和未来的数据应用引擎。虽然集群管理非常需要健壮的资源管理，但YARN的意义在于它可以在单一环境中简化当前和未来的数据应用引擎。对数据湖的持续投资促进了计算和内存资源的成功管理所有用户和应用程序。数据引擎和单个数据持久性的好处使数据操作系统的概念得以实现，从而减少了整个企业中的数据重复和移动。

1. **首先建立数据治理**

数据治理的需求一直在通过定义和分配数据所有者、数据管理员和专家来加强与企业中所有数据和信息策略相关的业务驱动因素和风险管理框架，这些所有权和代理权现在应该扩展到数据湖。

虽然数据湖由于数据使用的多样性增加以及其存放多种数据的内在能力而变得复杂，但数据治理框架本身应该从根本上保持不变。因此，对于数据湖的成功增长产生贡献的第二大因素是在数据湖策略开始时建立数据治理。正如在早期的四阶段成熟度模型中所概述的，公司通常从一个试验项目开始，然后在向前移动到主动活动之前的案例。如果在数据湖策略开始时没有建立数据治理，在必须对数据治理进行翻新（如果有可能的话）时，可能会出现严重的并发症。最坏的情况是公司必须创建第二个数据湖，以解决数据治理和安全性问题。并在数据被复制到整个HDFS（文件删除比较困难）之前，销毁在接收时丢失数据安全性的原始尝试。

数据治理为每个人将如何使用数据湖中的数据建立了一个共同的理解，并减轻了对数据湖势头有害的疑虑和猜测。已建立的数据治理计划将需要扩展当前围绕数据所有者、管理者、角色和委派权限的定义，以支持大数据、数据发现和数据科学，所有这些都包含在一个数据湖中。分析模型和见解的分布也应在数据湖治理中给予特别考虑。最后，数据治理将固有地定义其他关键技术需求，如访问控制、移动性和安全性。

1. **提前解决安全需求**

以数据为中心的安全方法提供了一个从创建到消费的广泛的视角来考虑数据。数据安全的基石将首先是身份验证，然后是授权。身份验证必须首先通过使用强密码、双通道身份验证或安全令牌来验证用户是否是他们声称的用户。一旦验证，一个安全的、集中的数据访问凭证存储库就可以供管理应用程序调用。因此，以数据为中心的安全方法的关键因素是了解每个数据元素级别的数据，以定义将被带入数据湖的内容的访问权限。更进一步，确定数据可能需要最初加密的点，以及加密程度将决定安全技术如何协同工作。在安全性方面，发生在访问数据的不同应用层之间。生产数据湖必须非常密切地关注物理入口点的管理，例如命令行和用户对集群的网络访问。

安全方法中同样重要的是了解数据使用模式，以识别如何正确地保护和加密数据湖中的数据。