Tableau 和大数据：概述

目录

如今的大数据趋势 **3**

数据的演变和分析需求 3

大数据：机遇与风险并存 4

Tableau 处理大数据的方式 **5**

大数据，大格局 5

数据访问和连接 5

大规模地与所有数据快速交互 6

Tableau 和大数据分析生态系统 **7**

云基础结构 8

引入和准备 8

存储和处理 9

查询加速 10

数据目录 10

大数据分析架构 **10**

主要云提供商示例 11

Tableau 客户示例 12

通用模式 13

关于 Tableau 和其他资源 **14**

如今的大数据趋势

数据的演变和分析需求

如今，数据无处不在 – 访问和分析数据的需求也是如此。"大数据"作为一个流行词可能已经被大家广泛接受和使用，而大数据的"三个特点"（数量大、种类多和速度快）比以往任何时候都更适用于大数据分析用例。虽然这种看法有一定的主观性，但行业所讨论的这些或其他特点（如可变性、有效性、准确性等）提醒我们，如今大数据仍然仅仅是数据，但它会逐渐变得更加复杂，组织必须通过创新才能有效地收集、整理、理解并利用它。

数字化转型正在各个行业和各种规模的组织中发生，各种各样的"事物"创造出多种格式和来源的大量数据。组织需要收集、处理和分析比以往任何时候都更加多样化的数据。从无架构 JSON 到其他数据库（如关系和 NoSQL）中的嵌套类型，再到非平面数据（如 Avro、Parquet 和 XML 等），数据格式正在成倍增加。要利用这些数据，连接器显得至关重要。

组织通常拥有以下几种数据的组合：

- **结构化数据** – 带有针对特定问题的预计算聚合，可能作为内存中计算的数据提取，并聚合起来用于分析。这通常是组织拥有的最有条理且易于访问的数据。
- **半结构化数据**（或对象存储）– 可能位于关系数据库、数据仓库或数据集市中。通常，这些都是定期刷新的业务概念，专门用于实体分析（问题已知，答案未知）。例如，交易次数、商机或销售人员在出现商机时采取的操作。
- **非结构化原始数据** – 位于数据湖或云端存储中。这包括由社交网络源、IoT 设备等来源创建的流数据。数据科学家可能会挖掘和转换此类数据，但它的全部潜力仍然未知。

尽管尚未找到某些数据最具价值的用例，但所有这些数据都能满足知识工作者更广泛的需求，因为他们需要访问和分析这些数据来制定决策。用于数据分析和可视化的应用程序正逐渐向数据本身靠拢。这意味着向云端的大规模转移中，分析可以与强大的存储和数据处理服务同时进行，从而实现更高的灵活性和更大的规模。无论组织是拥有广泛的、基于云的大数据实践，还是目前只对数据进行了较少的分析，都可以向业务部门和 IT 部门人员赋予必要的能力来可视化模式并分析其中的见解，进而让整个组织获得显著的效益。

尽管现代分析为所有技能水平的更多业务用户提供了更广泛的功能，但要想让所有这些数据成为整个组织的有用资源，我们仍然面临着许多复杂的挑战。业务需要随着数据本身的变化而变化，因此需要一个具有敏捷性和适应性的大数据策略和架构。对于组织而言，与其构建专注于数据连接的单一功能平台，不如拓宽大数据商机的范围，并考虑其不断变化的分析用例。否则，它们可能会错失更大的机会。

大数据：机遇与风险并存

如今，数据资产正日益成为利润丰厚企业与苦苦挣扎公司之间的重要分水岭。但是，数据规模巨大、增长迅猛、极为多样，这一切远远超出了关系数据库管理系统的处理能力，处理费用也极为昂贵。除了通过预计算和共享计算节省硬件成本，客户还想要最大程度地减少数据移动。如果基础结构能够使客户以最灵活的方式移动数据，将有助于弥补非结构化原始数据与直接可供用户分析的数据之间的差距。

组织还面临连接和性能问题。即使具有实时连接或内存中分析选项，但在生成数据提取或与其他数据混合时，巨大的数据湖也会带来沉重的负担。现代自助式的分析方法在很多方面有望实现敏捷性，但是在这些数据集上进行大量联接会使系统无力应对。

IT 部门和业务部门必须合作，采取自下而上的方法，由领域专家创建元数据、业务规则和报告模式。必须持续迭代和改进这些进程，以满足业务不断变化的需求；在当今的数据化转型时代，业务不会停滞不前，因此大数据分析框架也不应止步不前。

Tableau 处理大数据的方式

大数据，大格局

在 Tableau，我们所做的一切都是为了践行我们的使命：帮助人们查看并理解数据。Tableau 是专为数字化经济打造的现代分析平台，因为我们从根本上相信，数据必然会实现大众化。了解数据的人员应能提出与数据相关的问题，这意味着所有技能水平的知识工作者都应该能够访问、分析任何位置的数据，并发现其中的见解。

就在许多客户着手各种各样的大数据技术时，我们已调整工程投资、生态系统中的合作关系以及整体愿景，使其与数据格局的变化保持一致。Tableau 在大数据领域有着丰富而超前的投资历史。这些投资包括与 Hadoop 和 NoSQL 平台的数据连接，以及大规模本地数据仓库和云端数据仓库。

> 起初我们只是为了满足一种范围非常有限的业务用例，结果一发而不可收，它快速传播开来。所有人都在谈论大数据分析，但 Tableau 做出了行动，将它化繁为简。
>
> - ASHISH BRAGANZA，联想全球商业智能总监

了解联想如何在 28 个国家/地区中将报告效率提升 95%

数据访问和连接

为实现任何规模或格式的数据分析，我们支持广泛的数据访问，无论数据位于何处。Tableau 目前支持超过 75 种本机数据连接器，通过我们的可扩充性选项，还可以支持数不胜数的其他数据连接器。随着新数据源不断出现并为我们的用户提供价值，我们会继续将供应商的连接器与 Tableau 进行集成和认证，将它们整合到我们的产品中，以减少访问数据过程中出现的摩擦。我们相信，无论是 Web 流量、数据库中的记录还是日志文件，始终有许多数据源是用户希望使用的。

- **基于 SQL 的连接** - Tableau 使用 SQL 与 Hadoop、NoSQL 数据库和 Spark 进行对接。 Tableau 生成的 SQL 根据 ANSI SQL-92 标准进行标准化。使用 SQL 能够实现强大的功能，这是因为它极其精简（只需一个表达式）、开源且依照标准构建，不依赖任何库，而且功能丰富、表达力强。例如，使用 SQL 可以表达联接操作、函数、条件、汇总、分组和嵌套操作。

- **NoSQL 接口** – 顾名思义，NoSQL("不仅仅是 SQL")数据库除了包含关系格式数据之外，还可以具有以非关系方式建模的数据，支持包括列、文档、键值和图形在内的其他存储类型。这也意味着它们支持类 SQL 接口。

- **ODBC** – Tableau 的驱动程序采用开放数据库连接 (ODBC) 编程标准，并作为 SQL 与这些大数据平台提供的类 SQL 数据接口之间的转换层。使用 ODBC，您可以访问支持 SQL 标准并实现 ODBC API 的任何数据源。对于 Hadoop，这包括 Hive 查询语言 (HiveQL)、Impala SQL、BigSQL 和 Spark SQL 之类的接口。为了尽可能达到最佳性能，我们以自定义方式调优 Tableau 生成的 SQL，同时将聚合、筛选和其他 SQL 操作下推到大数据平台。

- **Web 数据连接器** – 通过 Tableau Web 数据连接器 SDK，可以连接到现有连接器以外的数据。自助式分析用户可通过 HTTP（包括内部 Web 服务、JSON 数据和 REST API）连接到几乎任何可供访问的数据，从而提高针对外部数据的大数据分析能力。

大规模地与所有数据快速交互

我们希望用户能够大规模访问自身的所有数据、与其他数据集成并快速找到见解。为了帮助实现大数据的自助式可视化分析，Tableau 已经对多种开创性技术进行了投资。

- **Hyper 数据引擎** – Hyper 是我们的高性能内存中数据引擎技术，可帮助客户更快速地分析大型或复杂数据集。借助专有的动态代码生成技术和先进的并行技术，Hyper 可更好地利用现代硬件，与之前的 Tableau 数据引擎相比，能够实现高达 3 倍的数据提取创建速度和 5 倍的查询速度。Hyper 还可以扩大和加快较慢的数据源，方法是创建数据提取，并将其调入内存中。

- **混合数据架构** – Tableau 可实时连接数据源或将数据（或子集）调入内存中。您可以在这两种模式之间来回切换以满足您的需求。我们的混合式数据访问方法为用户提供了较高的灵活性，并且有助于优化查询性能。

- **VizQL™** – Tableau 的核心是一种专有技术，在这种技术的推动下，交互式数据可视化成为理解数据的过程中不可缺少的一环。传统的分析工具迫使您分析行列式数据，选择要显示的数据子集，将这些数据组织成表，然后才能根据此表创建图表。而 VizQL 跳过了这些繁琐步骤，直接为您的数据创建可视化表现形式，在您进行分析时提供可视化反馈。VizQL 可以让您无限制地探索数据以查找其最佳表现形式，并且您可以无限制地"撤消"操作，确保不会误入歧途。在此可视化分析周期中，用户在操作过程中学习、按需要添加更多数据，并且最终获得更深入的见解。与通过代码生成仪表板相比，这不仅仅是一项更加丰富的体验，其方式也更容易被所有技能水平的人员所接受。

> 使用 Tableau，您可以真正与数据集进行实时交互，并且能够分析数据，然后在数分钟内按照您所需的方式呈现它。
>
> - JAMIE FAN，GRAB 的产品分析主管

了解 Grab 如何分析数百万行数据来改善客户体验

Tableau 和大数据分析生态系统

Tableau 这样的现代分析平台可能是通过发现见解来解锁大数据潜能的关键所在，但它仍然只是完整的大数据平台架构的重要组成部分之一。组建完整的大数据分析管道可能本身就是一项挑战。好消息是，在开始之前您无需构建整个生态系统，也无需集成整个策略中的每一个组成部分。

Tableau 非常适合大数据模型，因为我们优先考虑灵活性，即跨平台移动数据、按需调整基础结构、利用新数据类型以及启用新用户和用例的能力。我们相信部署大数据分析解决方案不应干涉您的基础结构或策略，而是应该帮助您利用已做出的投资，包括对大数据生态系统内的合作伙伴技术的投资。

云基础结构

越来越多的组织将业务流程和基础结构迁移到云端。由于基于云的基础结构和数据服务已移除本地 Hadoop 数据湖面临的一些主要障碍，因此可以比以往更轻松地实施和管理基于云的大数据分析解决方案。

> **Hadoop 以其低成本、横向扩展存储（Hadoop 分布式文件系统 - HDFS）、基于特定用途的处理引擎（首先是 MapReduce，然后是 Hive、Impala 和 Spark）以及共享数据目录 (Hive Metastore) 的强大组合，为现代数据湖奠定了基础。**
>
> **如今，曾经的共置存储和计算服务可按需单独在云端中扩展。资源还可以更轻松地进行纵向和横向扩展，并且按需定价。总体而言，云端提供更高的效率、更好的管理和服务协调能力。**
>
> 阅读 AtScale 产品副总裁 Josh Klahr 的这篇优秀文章，了解更多信息。

Tableau 提供与组织使用的基于云的技术的关键集成，包括 Amazon Web Services、Google Cloud Platform 和 Microsoft Azure。

引入和准备

在现代"引入和加载"的设计模式中，任何规模或形态的原始数据通常都归于数据湖：一种存储库，能够以原有格式（结构化、半结构化和非结构化）存储大量数据。数据湖通过更快、更灵活的来引入和存储数据，满足现代大数据分析的要求，可供任何人以各种方式快速分析原始数据。

社交网络、智能电表、家用自动化设备、视频游戏和 IoT 传感器等不同位置的联网设备和应用不断生成流数据。通常，这种数据通过半结构化数据管道进行收集。虽然可对流数据应用实时分析和预测算法，但通常使用 lambda 架构将流数据按原始格式路由并保存在数据湖（如 Hadoop）中，以便用于分析。Lambda 架构是一种数据处理架构，旨在利用批处理和流处理方法处理大量数据。这种设计克服了在延迟、吞吐量和容错方面的挑战。现在有多种可用于流式处理数据的工具，包括 Amazon Kinesis、Storm、Flume、Kafka 和 Informatica Vibe Data Stream。

数据湖通过 API 或类 SQL 语言提供经过优化的处理机制，可通过"读取模式"功能转换原始数据。数据进入数据湖后，需要将其引入并做好准备，供分析使用。Tableau 拥有 Informatica、Alteryx、Trifacta 和 Datameer 等合作伙伴，可帮助完成此过程并与顺畅地 Tableau 协同处理。您也可以使用 Tableau Prep 进行自助式数据准备。

存储和处理

Hadoop 具有出色的恢复能力，成本低，提供横向扩展数据存储、并行处理和群集工作负载管理功能，已用于数据湖。尽管 Hadoop 常用作大数据平台，但它并不是数据库。Hadoop 是一个开源软件框架，用于在商用硬件的群集上存储数据和运行应用程序。它能大量存储任意类型的数据，具备执行大型处理能力，能够处理非常多的并行任务或作业。

在现代分析架构中，Hadoop 提供低成本存储和数据存档，可将陈旧的历史数据从数据仓库移入在线冷存储。它还可用于 IoT、数据科学和其他非结构化的分析用例。Tableau 提供与所有主要 Hadoop 分布的直接连接（通过 Impala 与 Cloudera 连接、通过 Hive 与 Hortonworks 连接，以及通过 Apache Drill 与 MapR 连接）。

数据和数据仓库始终在现代分析架构中占有一席之地，它们将继续发挥重要作用，在整个企业中提供受管控且维度一致的准确数据，以实现自助式报告。即使是采用其他技术（如 Hadoop 和数据湖）的公司，往往也会保留关系数据库作为混合数据源的一部分。Snowflake 是使用本机 Tableau 连接器、基于 SQL 的云端原生企业数据仓库的示例。

Amazon Web Services Simple Storage Service (S3) 和 NoSQL 数据库等对象存储具有灵活的架构，也可用作数据湖。Tableau 支持 Amazon Athena 数据服务连接到 Amazon S3，并提供可直接连接到 NoSQL 数据库的多种工具。常用于 Tableau 的 NoSQL 数据库的示例包括但不限于 MongoDB、Datastax 和 MarkLogic。

数据科学和工程平台 Databricks 基于 Spark（一种用于批量数据处理和交互式横向扩展数据处理的热门引擎）提供数据处理。通过本机连接器连接到 Spark，您可以在 Tableau 中可视化来自 Databricks 的复杂机器学习模型的结果。

查询加速

虽然您可以对大数据进行机器学习和情绪分析，但人们的第一个问题常常是：交互式 SQL 有多快？说到底，SQL 是服务于业务用户的管道，这些业务用户希望使用大数据获得更快捷、可重复性更高的 KPI 仪表板，实施探索性分析。

这种对速度的需求促进了对更快速数据库的采用，这些数据库利用内存驻留技术和大规模并行处理 (MPP) 技术，如 Exasol 和 MemSQL、基于 Hadoop 的存储（如 Kudu），以及通过预处理实现更快速查询的技术（如 Vertica）。使用 SQL-on-Hadoop 引擎（如 Apache Impala、Hive LLAP、Presto、Phoenix 和 Drill）以及 OLAP-on-Hadoop 技术（如 AtScale、Jethro Data 和 Kyvos Insights）时，这些查询加速器进一步模糊了传统数据仓库和大数据领域的界限。

数据目录

企业数据目录主要充当数据源和常用数据定义的业务术语表，使用户能够更轻松地从受管控和批准的数据源中找到用于制定决策的正确数据。通过扫描引入的数据源，数据目录将由来自表、视图和存储过程中的元数据填充。数据整理工作甚至可以包括知识库信息和 Web 链接，以帮助用户理解数据上下文，并实现更多智能分类和自动数据发现。

数据目录存在于可视化分析解决方案中，也可用作与 Tableau 无缝集成的独立服务。我们的一些数据目录合作伙伴包括 Informatica、Alation、Unifi、Collibra 和 Waterline。

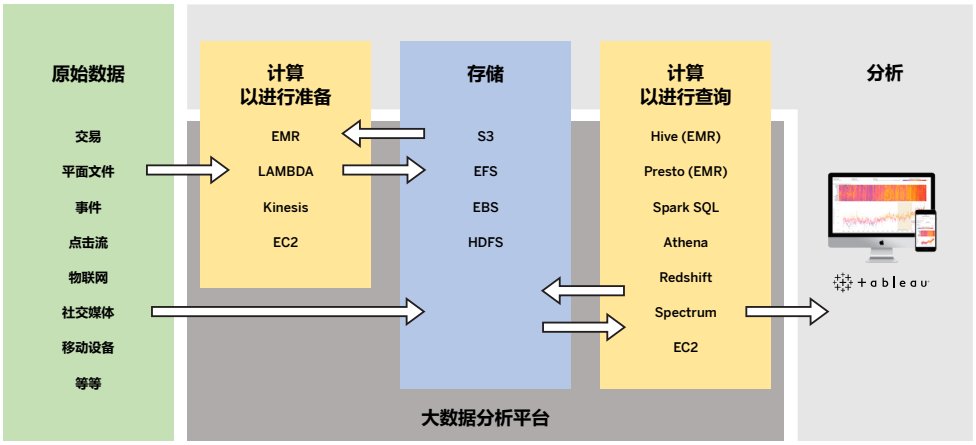
大数据分析架构

请务必记住，谈到成功的大数据架构时，没有一种方法是"万能"的。我们的客户针对自身的大数据分析需求定制了独特的解决方案，采用不同的平台和工具来构建数据管道。尽管如此，我们要进一步了解促进这些大数据分析平台取得成功的架构的通用组件。

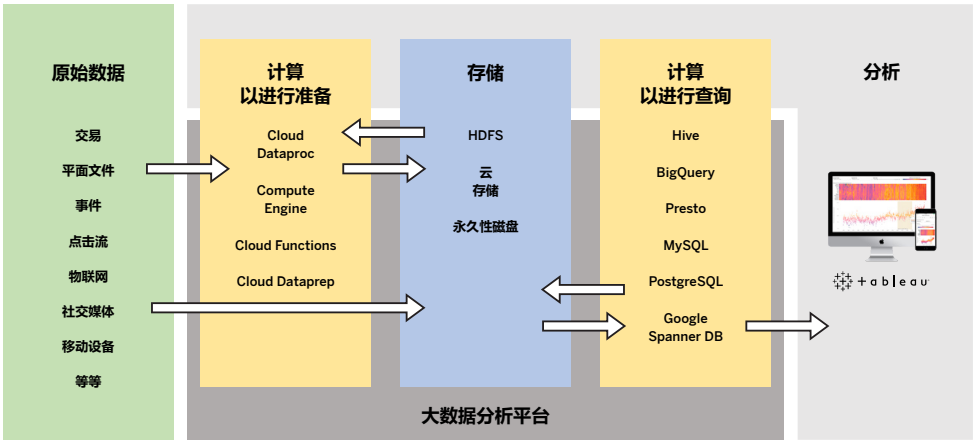
免责声明：请注意，以下示例是 *Tableau* 的解释，并非由它们所代表的云提供商或客户设计。我们在相应位置附上了指向原始说明的链接（如有）。这些流程图是经过简化的概览，旨在突出显示不同流程的关键元素中的相似之处。这些图表可能无法反映完整的大数据分析平台的每个部分，并且可能仅代表某些用例。另请注意，"计算以进行准备"最类似于"进程/目录"，而"计算以进行查询"最类似于"分析/模型"。

主要云提供商示例

Amazon Web Services



Google Cloud Platform



Microsoft Azure

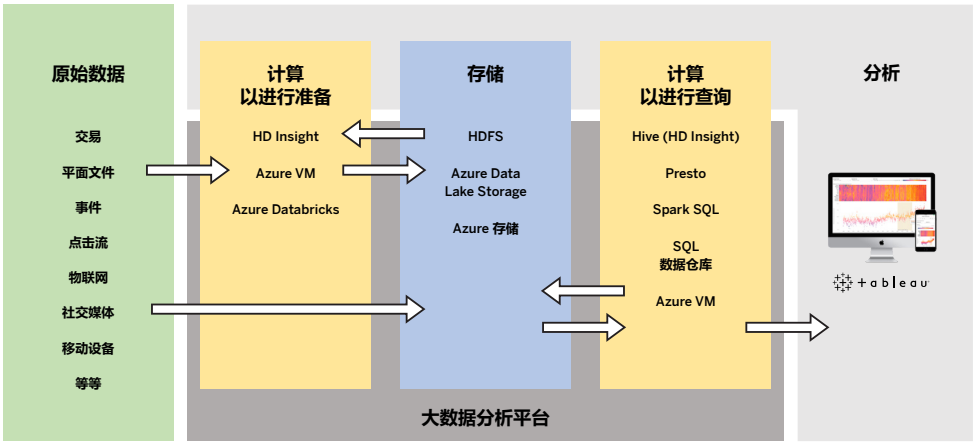
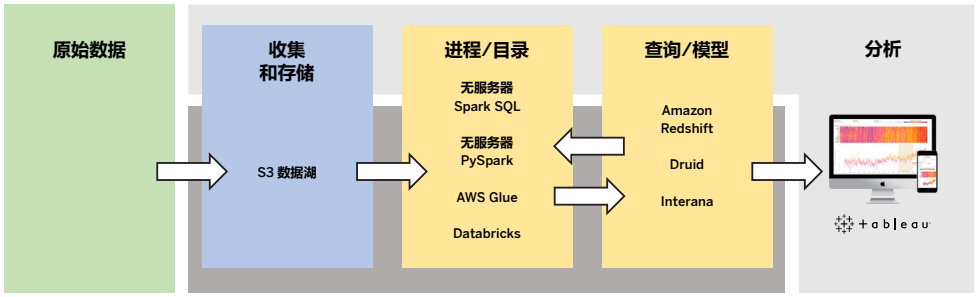
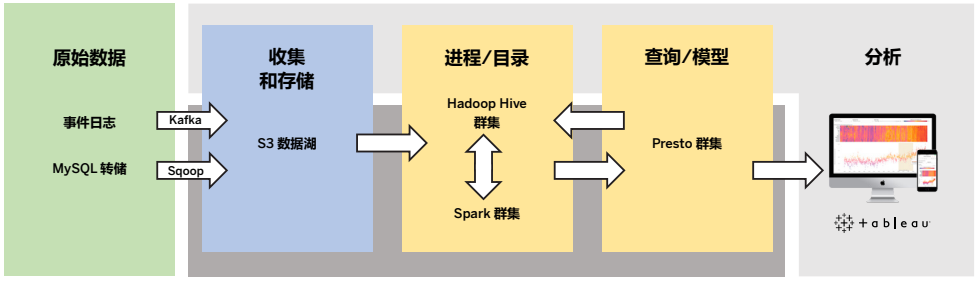


Tableau 客户示例

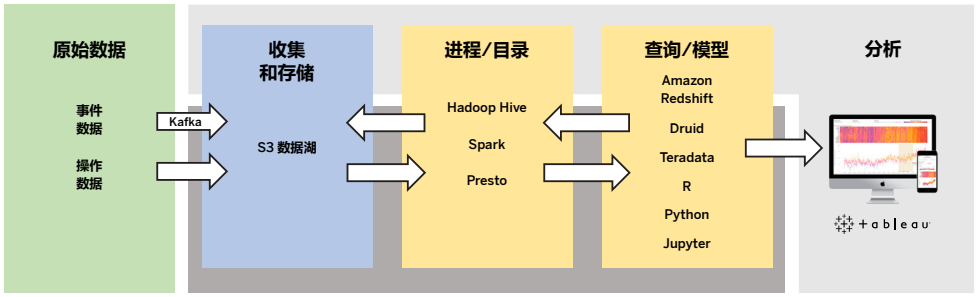
Edmunds – 了解更多信息



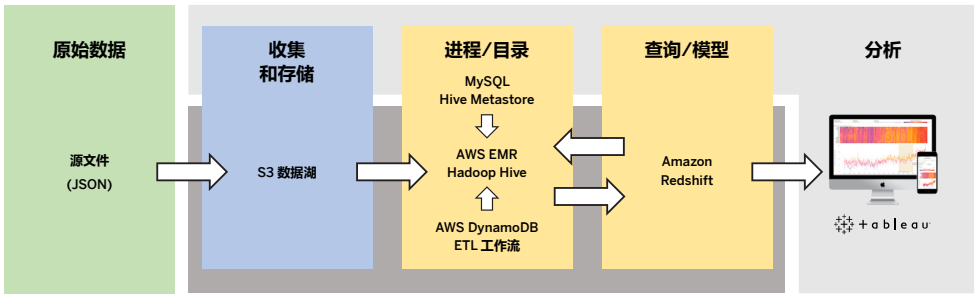
Airbnb – 了解更多信息



Netflix – 了解更多信息



Expedia – 了解更多信息



通用模式

虽然不存在两个相同的企业架构，但留意相似的模式以及它们的共同之处有助于为您自己的大数据分析平台制定策略。以下是我们在成功的大数据分析架构中始终能观察到的内容：

- **一个存储层** – 许多人将其称作数据湖。您的数据策略可能需要多个存储环境，但应包含结构化、半结构化和非结构化数据。
- **服务器和无服务器计算引擎** – 一些计算引擎用于数据准备和分析，另外一些用于查询。无服务器计算的动态性质可提供更高的灵活性和弹性，因为无需预先分配资源。
- **对数量、速度和多样化的支持** – 这不仅适用于数据，还适用于日益复杂、不断增多的用例，其中一些用例尚有待发现。
- **适用于作业的正确工具** – 请务必采用您的架构组件来构建独特的数据策略，但在业务需求变化时保持敏捷性也尤为关键。
- **企业级治理和安全性** – 虽然我们尚未深入了解这些领域的详细信息，但安全性和治理是确保可扩展性和正确使用数据的基础。
- **成本意识** – 在考虑大数据架构的必要功能和灵活性的同时，也要考虑成本。云端能够为发展提供较高的弹性，但是您需要考虑数据存储和处理、并行性、延迟、分析用例等费用问题。

大数据格局持续变化，但始终有一个主题贯穿所有挑战：企业需要能够使用通用的现代分析平台来访问任意位置、任意规模的数据。数据驱动型决策通过合适的平台、流程和程序来为人员提供支持，注定会是一笔宝贵的资产。

关于 Tableau

Tableau 是一个完整易用的可视化商业智能平台，可直接用于企业，通过大规模快速自助式分析帮助人们查看并理解数据。无论是在本地还是在云端，在 Windows 还是 Linux 上，Tableau 都能够充分利用您现有的技术投资，随着您数据环境的变化和增长来进行扩展。让您最为宝贵的两项资产充分发挥价值：数据物尽其用，员工人尽其才。

其他资源

现代分析平台的构成要素

IT 助力的 Tableau 企业分析

适用于企业的 Tableau：IT 概述

Tableau 免费试用

tableau