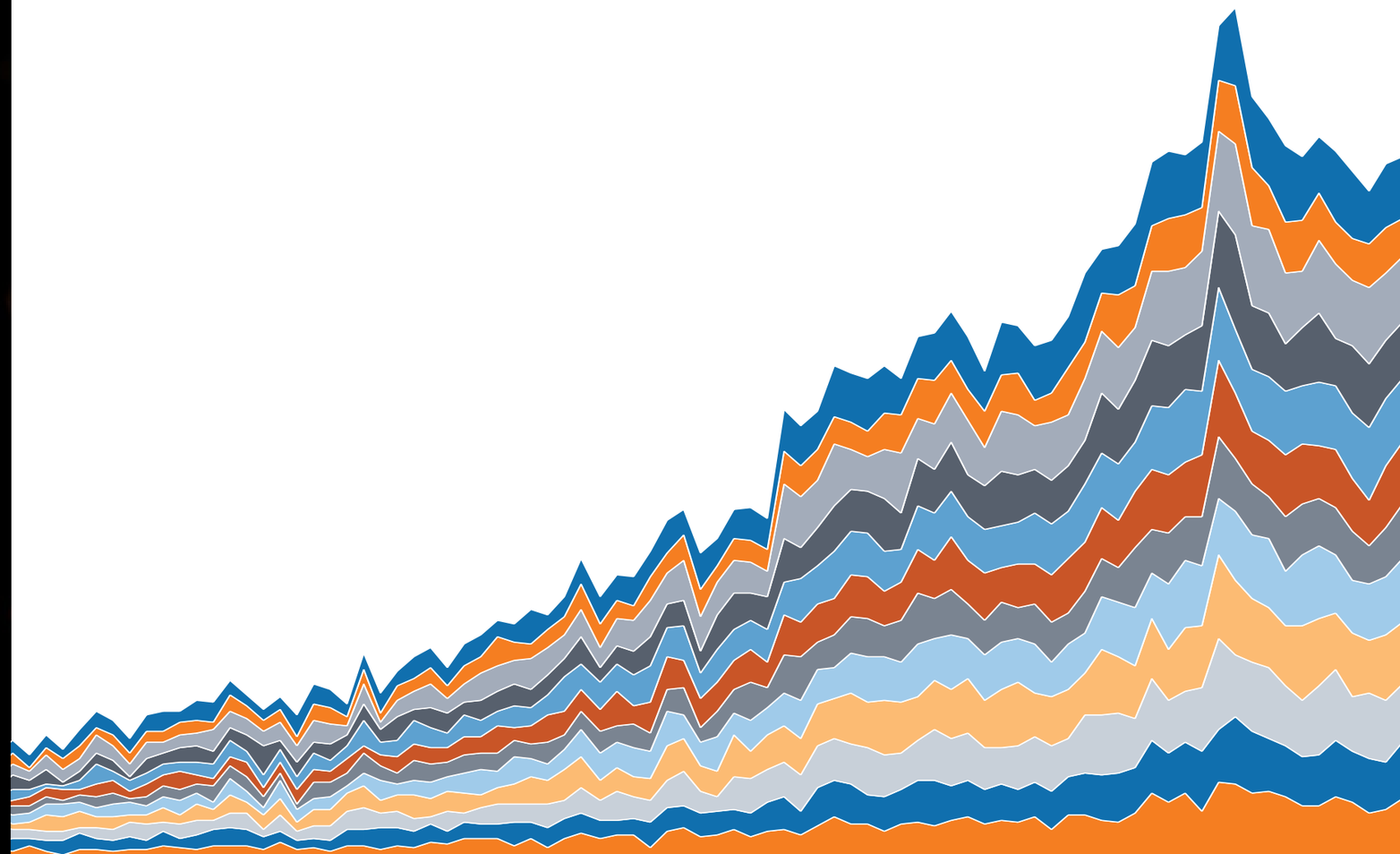
大数据

十大趋势

每年，Tableau 员工都会围绕行业动态开展讨论。相关讨论帮助我们整理出下一年中最显著的一系列大数据趋势。我们的预测如下：

大数据 10 大趋势

更多的组织将会存储和处理各种形式和规模的数据，并从这些数据中提取价值。支持大量结构化和非结构化数据的系统将继续获得更多关注。符合市场需求的平台一方面要能够帮助数据看护者管控和保护大数据，另一方面要能够为最终用户提供分析这些大数据的能力。这些系统将日益成熟，在企业 IT 系统内部按照标准良好地运行。了解我们针对明年做出的大数据发展预测。



1 大数据速度提高、门槛降低：选项增多，加快 Hadoop 速度

没错，您可以在 Hadoop 上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 Hadoop 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。

这种对于速度的需求促使更多的人采用 Exasol 和 MemSQL 等速度更快的数据库、Kudu 等基于 Hadoop 的存储以及可以提高查询速度的技术。使用 SQL-on-Hadoop 引擎（Apache Impala、Hive LLAP、Presto、Phoenix 和 Drill）与 OLAP-on-Hadoop 技术（AtScale、Jethro Data 和 Kyvos Insights）时，这些查询加速器正在让传统数据仓库和大数据世界更加难以区分。

2

大数据不再仅仅是 Hadoop：为特定用途构建的 Hadoop 工具变得过时

过去几年中，我们看到几种技术在大数据浪潮中水涨船高，这些技术可以满足在 Hadoop 上进行分析的需求。但是具有复杂和异源性环境的企业不再愿意采用仅适用于一种数据源 (Hadoop) 的孤立商业智能访问点。他们问题的答案埋藏在多种多样的数据源中，涵盖记录系统和云数据仓库，Hadoop 和非 Hadoop 数据源的结构化和非结构化数据。（另外，连关系数据库也越来越多地支持大数据。举例来说，SQL Server 最近添加了 JSON 支持。）

客户将要求获得对所有数据适用的分析功能。不区分数据和数据源的平台将会大行其道，而仅仅为 Hadoop 构建且无法跨用例部署的平台将受到冷落。Platfora 的退出预示了这一趋势。

3

为了获取价值，组织从起步阶段就开始利用数据湖

数据湖就像人造水库。您首先建造大坝将末端堵住（构建群集），然后让它充满水（数据）。一旦水库建成，您就开始将里面的水（数据）用于各种用途，例如发电、饮用和娱乐（预测性分析、ML、网络安全等）。

直到现在，将水引入湖中本身也是一个目的。这种情况将得到改变，因为 Hadoop 的业务理由将受到削弱。组织将会要求以可重复和敏捷的方式对湖进行使用，以便更快地获得答案。在对人员、数据和基础架构进行投资之前，他们将会仔细考虑业务效果。这会促使业务和 IT 部门建立更牢固的合作关系。作为大数据资产的利用工具，自助式平台将会在更大程度上得到认可。

4

日趋成熟的基础架构拒绝一成不变的框架

Hadoop 不再仅仅是数据科学用例的批处理平台。它已经成为临时分析的多功能引擎。甚至被用于关于日常工作量的运营报告 — 过去由数据仓库处理的那些报告。

组织将通过追求用例特定的基础架构设计来对满足这些混合需求。在敲定数据战略之前，它们会研究大量因素，包括用户角色、问题、量、访问频率、数据速度以及聚合水平。具备现代参考价值的基础架构将是需求驱动型架构。它们将合并最佳的自助式数据准备工具、Hadoop Core 和最终用户分析平台，并且合并方式可以根据此类需求的发展重新进行配置。这些基础架构的灵活性最终将驱动技术选择。

5

推动大数据投资的因素是多样化，而不是数量或速度

Gartner 用三个 V 来定义大数据：数量 (Volume) 大、速度 (Velocity) 快、种类 (Variety) 多。 虽然三个 V 都在增长，种类 (Variety) 却正在成为最重要的大数据投资驱动因素，New Vantage Partners 的一项近期调查印证了这一趋势。各公司将尝试集成更多数据源并专注于大数据的“长尾巴”，这种趋势也会持续加强。从无架构 JSON 到其他数据库（关系和 NoSQL）中的嵌套类型，再到非平面数据（Avro、Parquet 和 XML），数据格式正在增加，连接器正在变得更加关键。公司将继续根据分析平台能否实现与这些不同数据源的实时直接连接来对其进行评估。

6

Spark 和机器学习为大数据增辉

Apache Spark 曾经是 Hadoop 生态系统的组成部分，现在正在成为企业的首选大数据平台。在一项针对架构师、IT 经理、商业智能分析师的调查中，近 70% 的应答者更愿意选择 Spark，而不是当前市场份额最高的 MapReduce，后者侧重于进行批处理，对交互式应用程序或实时流处理帮助不大。

这些“大数据的大计算”功能提升了以计算密集型机器学习、人工智能和图算法为特点的平台。尤其值得一提的是，Microsoft Azure ML 因为易于入门和与现有 Microsoft 平台集成而大获成功。向大众开放 ML 的结果是出现更多的模型和应用程序，生成千万亿字节的数据。在机器学习和系统智能化的过程中，各方都将关注自

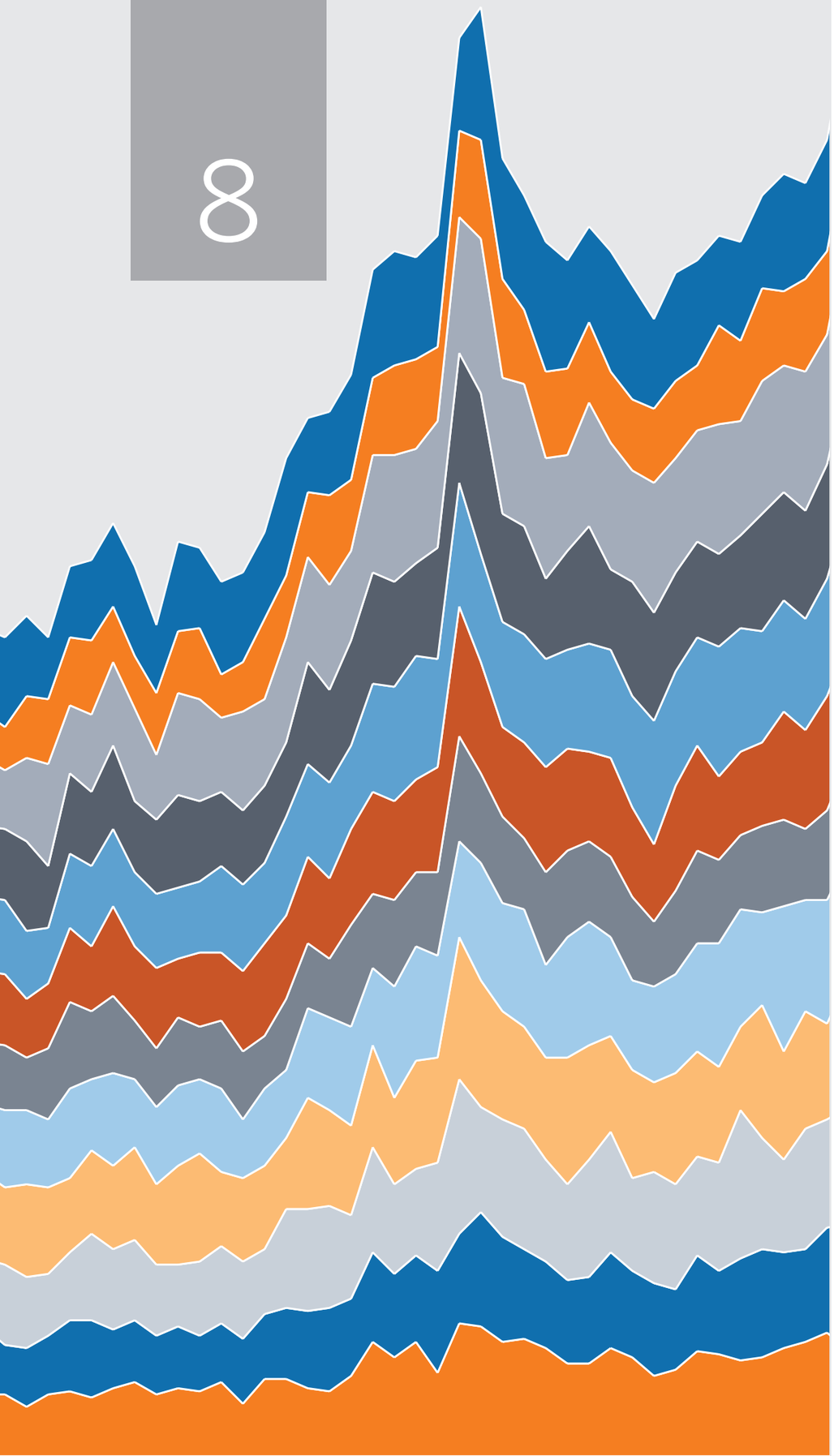
7

助式软件提供商如何让这些数据能够为最终用户所用。

物联网、云和大数据的汇合为自助式分析创造新的机会

一切事物都将有一个传感器，用于将信息发回处理中心。物联网正在生成大量结构化和非结构化数据，并且此类数据正在被越来越多地部署到云服务中。这些数据常常来源不同，并且分散在多个关系和非关系系统中，例如 Hadoop 群集和 NoSQL 数据库。虽然存储和托管服务领域的创新加快了获取流程，但对数据本身的访问和理解仍然是一个重大的“最后一英里”问题。因此，人们越来越需要可以无缝连接和合并多种云端托管数据源的分析工具。此类工具让企业能够对任何地点、任何类型的数据进行探索和可视化，从而帮助它们发现物联网投资中隐藏的机会。

8



自助式数据准备成为主流，最终用户开始影响大数据

这个时代面临的最大问题之一是如何让业务用户可以访问 Hadoop 数据。自助式分析平台的兴起使我们的境遇大为改观。但业务用户还想进一步降低分析所用数据的准备时间和复杂性。在需要处理多种多样的数据类型和格式时，这一点尤其重要。

敏捷的自助式数据准备工具不但可以在源位置准备 Hadoop 数据，还能够以快照方式提供数据，实现更快捷、更轻松的探索。我们已经在这个领域看到大量创新，这些创新来自于重点关注最终用户大数据准备的公司，例如 Alteryx、Trifacta 和 Paxata。这些工具可以为较晚采用 Hadoop 的用户和“落后者”降低门槛，并将持续受到欢迎。

9

大数据发展壮大：Hadoop 增强企业标准

我们看到，Hadoop 日趋成为企业 IT 格局中的核心环节。人们将围绕企业系统增加安全性和管控组件方面的投资。Apache Sentry 提供了一个系统，该系统可对存储在 Hadoop 群集上的数据和元数据实施细化、基于角色的授权。Apache Atlas 是数据管控计划的一部分，它让组织可以在整个数据生态系统中应用一致的数据分类方法。Apache Ranger 为 Hadoop 提供集中式安全管理。

客户开始希望从自己的企业级 RDBMS 平台获得这些类型的功能。这些功能正在移向新兴大数据技术的前沿，从而为其在企业中的采

10

用消除了又一个障碍。

元数据目录的兴起可以帮助人们找到值得分析的大数据

长期以来，公司因为无法处理过多的数据而将它们丢弃。

借助 Hadoop，它们可以处理大量数据。然而，数据通常并没有以易于查找的方式组织。

元数据目录可以帮助用户发现和理解值得使用自助式工具进行分析的相关数据。这种尚未满足的客户需求正在被 Alation 和 Waterline 之类的公司填补。这些公司使用机器学习来实现 Hadoop 数据的自动查找。它们使用标记来对文件进行目录分类，发现数据资产之间的关系，甚至通过可搜索 UI 提供查询建议。这可以帮助数据使用者和数据维护者减少验证、查找和准确查询数据所需的时间。来年，自助式探索的认知度和需求将得到提高。它将作为自助式分析的自然扩展实现增长。

关于 Tableau

将数据可视化集成到您的零售程序和流程中要比您想象的简单。

无论数据量有多大或是来源于多少系统，Tableau 都能帮助人们查看并理解数据。通过从 PC 到 iPad 的无缝体验快速连接、混合、可视化并分享数据仪表板。不需要编程技能，就能创建和发布带有自动数据更新功能的仪表板，并分享给同事、合作伙伴或客户。现在就开始免费试用。

TABLEAU.COM/ZH-CN/TRIAL