# Google BigQuery 搭配 Tableau：最佳做法

# 简介

Tableau 与 Google BigQuery 珠联璧合，让用户可以通过易用的可视化界面分析海量数据，快速获得问题解答。搭配使用这两款工具，您可以收获以下益处：

- 让普通用户也能够运用 Google BigQuery 的强大功能，快速进行交互式分析。
- 使用可视化分析工具在数秒之间就能分析完数十亿行数据，不需要编写任何代码，也完全不用操心服务器端管理工作。
- 只需数分钟时间即可制作出令人惊叹的仪表板，这些仪表板连接到您的 Google BigQuery 数据，让您的组织随时掌握最新情况。
- 使用 Tableau Server 和 Tableau Online 在网络上分享报告和真知灼见，让人人都可通过任何设备访问。
- 结合 Google BigQuery 的云端敏捷性与 Tableau 快如闪电的速度，更快看到项目的价值。

优化这两种技术的组合将大大提升性能、缩短设计周期，有助于用户和组织取得更大的成功。在本白皮书中，我们将讨论数据建模和查询构建方面的优化技巧，以便最大限度提升可视化的响应迅捷性。我们还将讨论在结合使用 Tableau 和 BigQuery 时，可通过哪些技巧来最大限度提高成本效益。

# 作者

Pierce Young，Tableau 产品经理

Vaidy Krishnan，Tableau 高级产品经理

Riley Maris，Tableau 高级产品营销专家

Babu Prasad Elumala，Google 解决方案工程师

Seth Hollyman，Google 技术计划经理

Tino Tereshko，Google 企业解决方案工程师

Mike Graboski，Google 解决方案工程师

# 目录

# 技术概述

## Google BigQuery

BigQuery 使用纯 SQL 就可以在数秒内处理完数千万亿字节 (PB) 的数据，不需要用户进行微调或掌握专门的技能组合。BigQuery 依托 Dremel（Google 用于分析大量数据集的革命性技术），以每十亿字节 (GB) 几角钱的费用，提供大型企业之前需要支付数百万美元才能获得的性能水平。

BigQuery 是一个数据仓库，最适合用来针对大量结构化和半结构化数据集运行 SQL 查询。例如，适合采用它的用例和数据集有：

- 临时分析
- 网络日志
- 计算机/服务器日志
- 物联网数据集
- 电子商务客户行为
- 移动应用数据
- 零售分析
- 游戏遥测
- Google Analytics Premium 数据
- 所有采用传统 RDBMS 时需要数分钟（乃至数小时）才能完成一次批量查询的数据集

BigQuery 完全不需要运营和维护，并与 Google Cloud Platform 集成在一起。与其他云端分析解决方案不同的是，BigQuery 不需要提前配置服务器群集。BigQuery 会在运行时确定处理群集的规模并配置它们。

随着数据量的增加，BigQuery 将自动增加处理能力，但处理每十亿字节 (GB) 数据的价格不变。

### 旧版 SQL 与标准 SQL

Google BigQuery 升级了其 API，现在不仅可使用 BigQuery SQL（现称"旧版 SQL"），还可使用标准 SQL；Tableau 则相应地升级了 Google BigQuery 连接器，以支持这种向标准 SQL 的转变。标准 SQL 让 BigQuery 用户受益良多，包括使用详细级别表达式、加速元数据验证以及为连接服务选择收费项目。本指南假设读者使用标准 SQL。

如需详细了解如何从旧版 SQL 迁移到标准 SQL，请参阅有关从旧版 SQL 进行迁移的在线帮助指南。

## Tableau

Tableau 帮助人们查看并理解数据。我们的现代分析平台基于斯坦福大学开发的技术，让普通用户也能充分发挥数据的功用。这样，广大用户都可以与自己的数据互动、提出疑问、解决难题、分享洞见、创造具有变革意义的价值。用户不管能否自如地使用传统商业智能 (BI) 工具，都可以快速学会使用 Tableau 的直观、拖放式用户界面，来制作和探索内容丰富的交互式可视化以及功能强大的仪表板。

最近，我们进一步扩展了 Tableau 平台的功能，现在它包含直观、简捷、智能的数据准备功能，以及使用自然语言查询已发布数据源的功能。

## Tableau 本身提供的优化功能

**数据源连接器** – Tableau 本身就有一个经过优化的连接器可连接到 Google BigQuery，该连接器既支持实时数据连接，也支持内存中数据提取。通过 Tableau 的数据混合功能，用户可将来自 BigQuery 的数据与来自其他 67 个受支持数据源的数据融合起来。对于使用 Tableau Server 或 Tableau Online 发布到云端的可视化，可以与 Google BigQuery 保持直接连接。

**并行查询** – Tableau 可以充分利用 Google BigQuery 的功能以及其他数据源来同时执行多个查询，所支持的并发查询总数最多可达 16 个。如果查询结果尚未缓存，系统会将多批独立且经过消重的查询合为一组，并发送给 BigQuery。由于 BigQuery 采用横向扩展架构，用户应该会体验到并行查询带来的大幅性能提升。

**查询融合** – 在可行的情况下，Tableau 会将工作薄和仪表板中的多个查询融合在一起，从而减少发送到 BigQuery 的查询数量。首先，Tableau 会找出类似的查询，排除那些返回的列存在差异的查询。然后，它会合并只有聚合级别或用户计算存在差异的查询。

**外部查询缓存** – 如果基础数据源在您上次运行同一查询后未发生过更改，Tableau 将自动从之前保存的查询缓存中读取数据，这样数据几乎可以瞬时加载完毕。

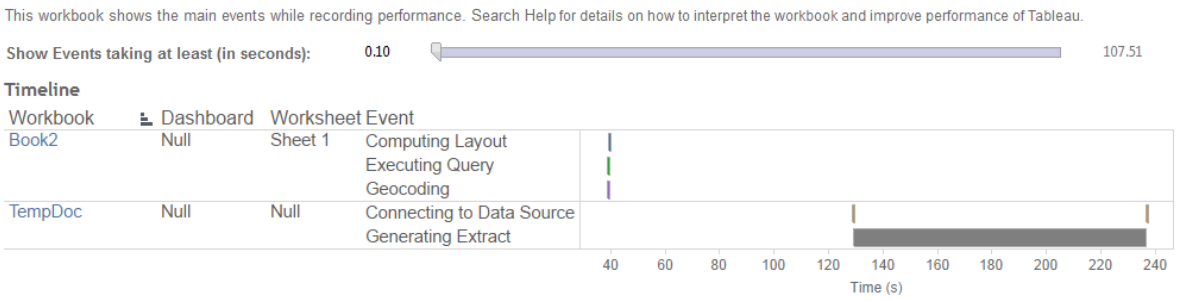
**Tableau Desktop 中的按需连接** – 当您打开已发布的工作簿时，Tableau Desktop 仅连接到显示当前工作表的数据所需的数据源。也就是说，您可以更快地看到数据。

# 优化性能的最佳做法：Tableau

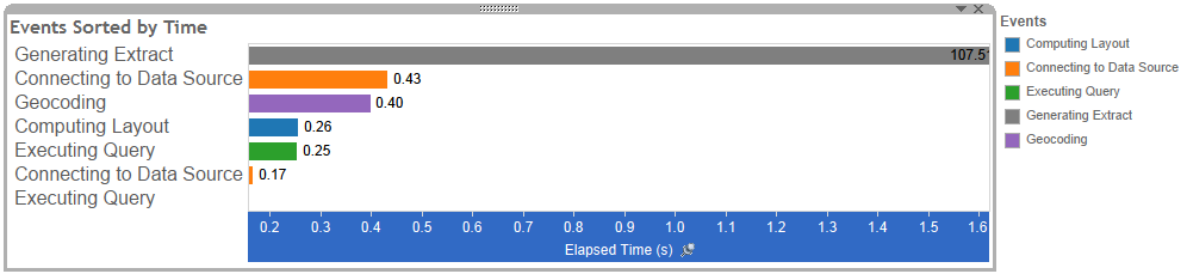
在介绍其他工具和自定义设置前，我们首先建议您在可行的情况下持续更新自己的 Tableau 部署。这样，您就可以充分利用我们在产品最新版本中引入的性能改进。

## 性能记录器

性能记录器是一款强大的内置工具，让您可以找出运行缓慢的查询并优化工作簿，以最大限度地提升性能。其方法是跟踪具体工作簿执行一次查询和计算布局所用的时间。将鼠标悬停在下方某个绿条之上，用户便可查看正在针对 BigQuery 生成的查询。发现运行缓慢的查询后，您通常可以通过重新审视数据模型来解决性能问题。




在“时间范围”视图中，可以通过“工作簿”、“仪表板”和“工作表”列明确事件的上下文。




如果您想加快工作簿的运行速度，可以根据哪些事件的持续时间较长来判断应先从何处着眼。

如需详细了解如何创建或解读性能记录，请访问以下帮助文章：

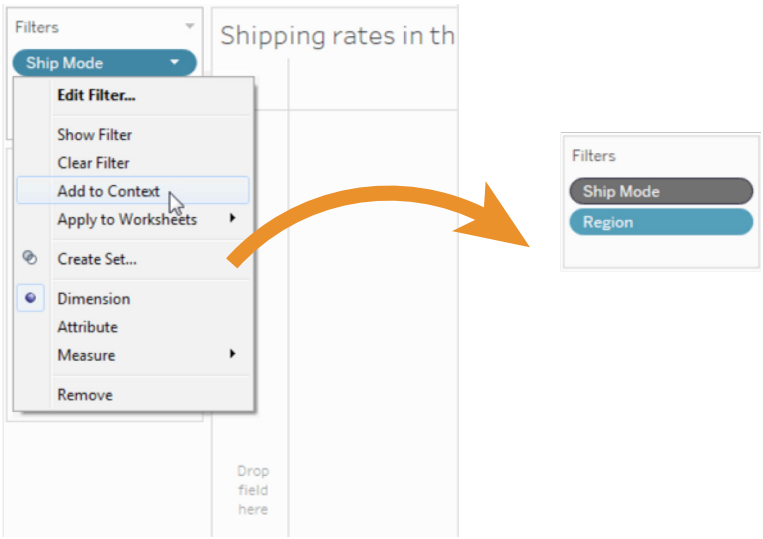
Tableau Desktop 上的性能记录器（创建）
Tableau Server 上的性能记录器（解读）

## 上下文筛选器

如果您要将筛选器应用于大型数据源，则可以通过设置上下文筛选器提升性能。可以先用一个上下文筛选器来筛选数据源，这样只需对筛选出的记录再应用其他筛选器。采用这种顺序可避免将每个筛选器都应用于数据源中的每条记录。

如果要设置的筛选器可大大降低数据集规模，并打算将这些筛选器用于众多数据视图，则应将这些筛选器设为上下文筛选器。

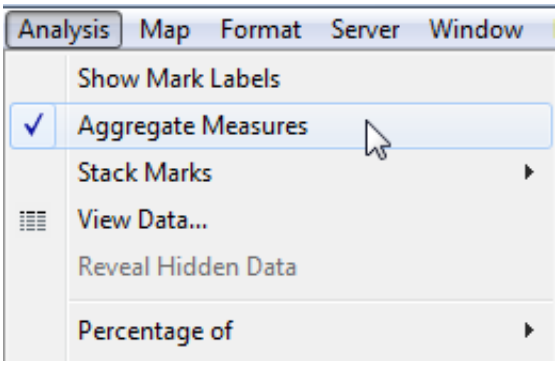
有关详细信息，请参阅我们在线帮助指南中的“使用上下文筛选器提高视图性能”一文。




您可以设置一个或多个上下文筛选器来提升性能。

## 聚合度量

如果您创建的视图运行缓慢，请确保您使用的是聚合度量，而不是离散度量。视图运行缓慢通常意味着，您正在尝试同时查看众多行数据。您可以通过聚合数据来减少行数。



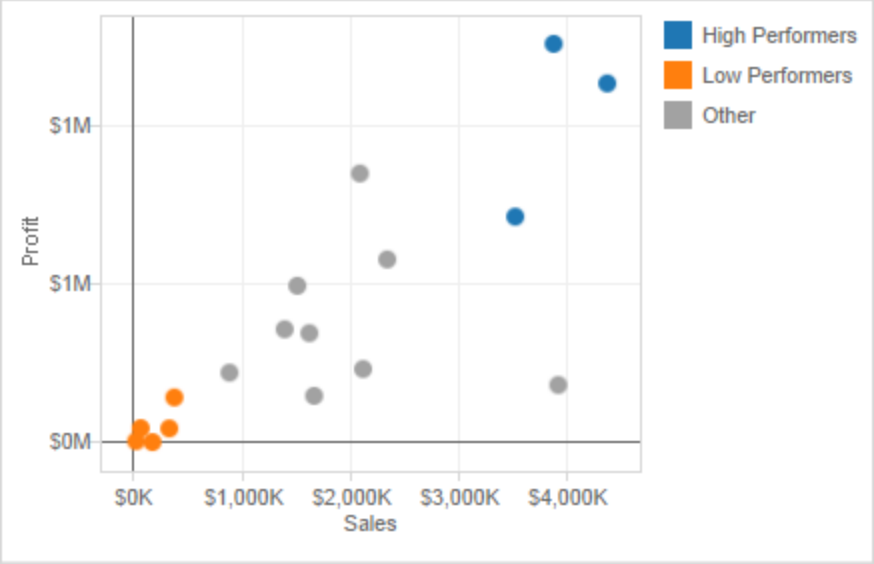

在“分析”菜单中可以查看度量是否为聚合度量。您还可以为度量设置默认聚合。

有关详细信息，请参阅我们在线帮助指南中介绍数据聚合的文章。

## 集与组

如果您想要对某个维度进行筛选，以便根据度量值的范围移除一些成员，则应该创建一个集，而不是使用定量筛选器。例如，可创建一个集，仅返回某维度中的前 50 项（而不是该维度中所有项）。

当您基于所选的内容创建组时，请确保仅包含感兴趣的列。集中每多出一列，性能便会降低一分。




在 Tableau 中创建组时，可以选择将所有剩余的成员都归入“其他”组。

有关详细信息，请参阅我们在线帮助指南中的“创建集”一文以及“创建组”一文。

## 先添加筛选器

如果您要处理的是大型数据源，并且停用了自动更新，那么在向视图中添加筛选器后创建的查询可能会相当缓慢。不应先生成视图再指定筛选器，而应该先指定筛选器再将字段拖到视图中。这样，您运行更新或启用自动更新后，系统将会先进行筛选。

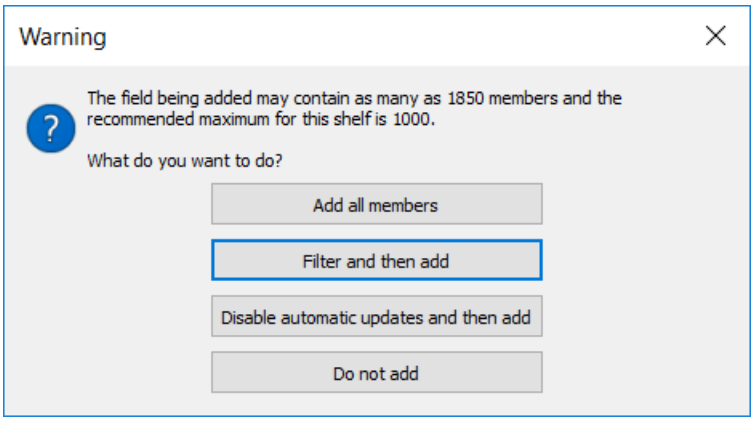
## 停用自动更新

当您将字段放入功能区后，Tableau 会通过自动查询数据源来生成视图。如果您创建的数据视图包含大量数据，则查询可能会非常耗时并且大大降低系统性能。在这种情况下，您可以指示 Tableau 在您生成视图的过程中停用查询。之后，当您准备好查看结果时，可以再重新启用查询。

有关详细信息，请参阅我们在线帮助指南中介绍自动更新和性能的文章。

## 留意警告

如果您试图将某个大型维度（拥有众多成员）放入任一功能区，Tableau 会显示性能警告对话框。该对话框将提供四个选项，如下图所示。如果您选择"添加所有成员"，则性能可能会大幅下降。




将大型维度放入功能区时，如果可能会严重影响性能，Tableau 会发出警告。

### 优化并行查询

您可以使用自定义属性来提高从 BigQuery 返回到 Tableau Online 和 Tableau Server 的大型结果集的性能；在 Tableau Desktop 上，则可以通过配置并行查询做到这一点。可以将这些自定义属性纳入到您的已发布工作簿或数据源中，只要您在将工作簿或数据源发布到 Tableau Online 或 Tableau Server 前先指定这些属性即可。

有关详细信息，请参阅我们在线帮助指南中"Google BigQuery"一文的"使用自定义属性来提高查询性能"部分。

## 优化成本和性能的最佳做法：Google BigQuery

为提升查询性能和降低成本，通常最好避免使用数据位于 Google Cloud Storage 等外部数据源中的联合表。在这种情况下，如果要对数据集执行迭代查询，应使用查询 API 实现 BigQuery 中的数据（独立于 Tableau），以便通过 Tableau 对数据集进行高性能查询。

### 非标准化与预联接

BigQuery 支持超大型联接，而且联接性能极佳。不过，BigQuery 是列式数据存储，针对非标准化数据集使用时可以发挥出最优性能。

云技术所带来的一大好处就是可以将存储资源与计算资源分离，让用户可以对这两种资源分别进行扩展和付费。由于 BigQuery 存储价格便宜、扩展性强，通过非标准化和预联接将数据集转换为同构表通常可节省资金。实质上，这意味着您可以少用一些计算资源，而多用一些存储资源（后者更有利于提升性能，也更经济实惠）。由于 BigQuery 属于列式存储，可以有效压缩数据，因此减少计算资源而增加存储资源是个不错的选择，而且成本也可能更低。

BigQuery 是卓越的 ETL 工具，可让您快速且高效地执行大规模变换和传递。实现 128 MB 以上的数据集时，请务必启用"允许大型结果"。

如需详细了解如何准备待加载的数据和如何使用 BigQuery 的 SQL 方言来查询数据，请参阅以下文章：

加载非标准化的嵌套数据和重复数据
写入较大的查询结果

## 按日期划分表

将一个表分割成多个较小的分区也就是我们所谓的"划分"，这样做有助于简化数据管理工作、提高查询性能和降低成本。而且，BigQuery 也支持对已分区的表进行聚类分析；如果数据已经按日期或时间戳列进行了分区，或者您针对查询中的特定列创建了筛选器或聚合，这种支持就会派上用场。

有些数据天生就适合按日期进行分区：例如日志数据，或者记录中包含单调递增时间戳的任何数据。在这种情况下，不妨按日期划分 BigQuery 表，并在表名称中加入日期。若要利用这一点，需在 Tableau 中使用自定义 SQL。

有关详细信息，请参阅在线帮助指南中的"连接到自定义 SQL 查询"一文。

例如，可以为表指定下面这样的名称：mytable_20170501、mytable_20170502，依此类推。

之后，如果要运行按日期筛选的查询，可以使用 BigQuery 通配符表的函数：

```
SELECT
 name

FROM
 `myProject.myDataSet.mytable_*'

WHERE
 age >= 35
```

上例将自动包括带有 mytable_ 前缀的所有表。

若要使用通配符，必须按照以下模式命名表：[任意前缀]YYYYMMDD。

一些其他的数据库系统依靠划分来提升性能。实际上，按日期划分表的做法对 BigQuery 的性能影响可以忽略不计，这样做主要是为了降低成本。由于您处理的数据减少了，因此为每次查询支付的费用便会降低。

请注意，如果决定按分钟级别划分表，可能因分片过多而直接影响性能。必须多加注意，以免同时产生过多分片。"每天"以下的划分精度通常都是可接受的。

如需详细了解划分，请参阅以下文章：

分区表简介
聚簇表简介
使用通配符表查询多个表

### 运行多个类似查询时指定目标表

虽然您运行多个相同的查询时，查询缓存非常有用，但是当您运行略微不同的类似查询（例如，两次运行的查询相比，只是 WHERE 子句中的值有所变化）时，查询缓存作用不大。在这种情况下，可以对源表运行一次查询，将您打算重复查询的记录写入到一个新的目标表。然后，针对您创建的新目标表运行查询即可。

**例如，假设您打算运行三个查询，它们分别包含不同的 WHERE 子句：**

```
WHERE col1 = "a"
WHERE col1 = "b"
WHERE col1 = "c"
```

**这种情况下可以针对源表运行一次查询，将输出记录写入到目标表中：**

```
SELECT col1
FROM source
WHERE col1 = "a" OR col1 = "b" OR col1 = "c"
```

通过使用"OR"运算符连接各个 WHERE 子句，我们可以获取到所有相关记录。我们的新目标表可能远远小于原始的源表。由于 BigQuery 根据查询中处理的数据量收费，与直接针对源表运行查询相比，针对新的目标表运行后续查询可节省资金。将来必须注意要清理这些表，以防其存储费用增加。

## 使用 Tableau 将 Google BigQuery ML 模型结果可视化

借助 BigQuery ML，用户可以利用嵌入的机器学习技术，根据 BigQuery 中存储的数据来训练模型。不过，就像处理任何其他数据时一样，直接查询数据库并非总是探索模型输出的理想方法。

而利用 Tableau 的 BigQuery 连接器，您就可以轻松操纵预测模型的结果，同时这种操纵方式还有助于直观地理解数据。此外，Tableau 还让用户可以轻松与他人共享自己的模型及其结果，使其他人也能受益于自己的所有工作成果。

如需查看示例，请阅读 Using BigQuery ML with Tableau to predict housing prices（结合使用 BigQuery ML 与 Tableau 来预测房价）。

虽然在这项练习中调用 Google BigQuery ML 中的机器学习算法必须要用到自定义 SQL，但在其他情况下使用自定义 SQL 来代替 Tableau 本身的连接时，需要考虑性能方面的因素。应尽可能使用 Tableau 本身的数据源连接，以便获得最优性能。

如需详细了解如何使用标准 SQL 查询来创建和执行机器学习模型，请参阅 Google Big Query ML 在线帮助指南。

# 案例研究：zulily 结合使用 Tableau 和 Google BigQuery 实现自助式分析的要诀

zulily 是一家发展飞快的电子商务公司，它使用 Google BigQuery 作为企业数据仓库，再搭配 Tableau，打造了一个可实现数据访问和可视化分析的大数据平台。通过集成 BigQuery 与 Tableau，其分析团队在日常工作中就能快速获得、提取并使用数据来构建报告和模型，而无需依靠 IT 人员。此外，业务部门的用户也可以实时访问用于快速制定决策的关键数据，无需分析师参与即可获得基本的分析洞见。

## 下面列举了 zulily 采取的几项最佳做法：

**通过在 Google Compute Engine 上使用 Tableau Server 来降低延迟** - 传统的模式是每个区域分别作为独立的 VPC（虚拟专用云），而实际上有更好的做法：直接利用 Google 的专用主干网，无需接入互联网，也不需要进行额外的设置。这种做法还可以让您以合适的规模进行部署，避免过度配置。

**使用联合来源并将 BigQuery 作为 Tableau 的查询对象** - 对于 Google Cloud 内的数据，您应充分利用 BigQuery 查询外部数据源的功能，并将 BigQuery 作为您的数据湖。在某些情况下，您可以减少需要通过网络传送到 Tableau 进行分析的数据量。

**通过与 BigQuery 的实时连接来处理大型数据集** - 充分利用 BigQuery 处理大型数据集的功能，只将结果通过网络传送出去。除非有明确的理由要提取数据，否则不妨将与 BigQuery 建立的默认 Tableau 连接设为"实时"连接。

如需了解更多信息以及所有 10 大要诀，请参阅以下由两部分组成的系列博文：

**第一部分：** Zulily 为何使用 Tableau 和 Google BigQuery 创建自助式营销分析平台

**第二部分：** zulily's top 10 tips for self-service analytics with Google BigQuery and Tableau（zulily 结合使用 Google BigQuery 和 Tableau 实现自助式分析的 10 大要诀）

# 结语

通过运用最佳做法，业务部门用户、数据分析师等都能够最大限度提高针对 Google BigQuery 生成的 Tableau 可视化的性能和响应迅捷性。结合使用这两种技术时，用户可以真正做到瞬间可视化数十亿行数据，实现随想随答。

## 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

Tableau 免费试用

Tableau Online 帮助指南：Google BigQuery

Tableau Server and Google Cloud Platform: rapid-fire business intelligence in the cloud（Tableau Server 与 Google Cloud Platform 珠联璧合：在云端实现快速便捷式商业智能）

Tableau 和 Google 的解决方案

Tableau 和大数据：概览

为何要在云端实现业务分析？

设计高效工作簿