提取并加载数据湖

Last modified: October 30, 2019 • Reading Time: 7 minutes

要将数据放入数据湖（Data Lake）中，您首先需要通过 SQL 或某些 API 从源中**提取（Extract）**数据，然后将其**加载（Load）**到数据湖中。此过程称为提取和加载-简称为“ EL”。

有许多出色的现代 EL 供应商（有时称为 ELT 供应商），例如 Fivetran, Stitch Data 或 Blendo。

这些 EL 提供程序为最流行的 API 构建了详细的 Extract 脚本，并提供了将数据提取并将其加载到数据湖的简单体验。该过程通常涉及建立管道的过程，在该过程中既为目标数据又为数据源提供了凭据，并提供一些轻度转换（例如，选择要同步的表和字段，出于隐私原因而隐藏一些值等）的配置。

在大多数情况下，只需最少的工程量即可完成设置。

提取（Extract）选项

提取是通过 API 或 SQL 从数据源提取数据的阶段。我们可以对所有可用数据进行完整提取，或者每次运行同步时都可以进行增量提取。完整提取将从数据源中提取所有数据。增量提取将仅从数据源中提取更新的记录。

完全提取

完全提取是最简单的方法，因为不需要进行配置，但是有两个很大的缺点。

1. 您最终会在数据湖中获得大量重复数据
2. 您增加了分析堆栈中一些步骤的复杂性

您将必须弄清楚数据湖中实际需要哪些数据，因此它将需要更复杂的逻辑和更多的处理。

增量提取

首选的替代方法是进行增量提取。这更具挑战性，因为您需要检查新增或修改的行，并为更改的数据模式负责。但是这种方式通常是首选的，因为在数据湖中需要处理和更新的数据少得多。所有云 ELT 供应商都支持从您的资源中进行增量提取。

增量提取的主要缺点是数据源中的删除。通常情况下，检测和实施删除并不容易。在大多数情况下，ELT 提供程序不能保证删除时的一致性，在某些情况下可以做到这一点，也可以由源实现，例如，数据从不删除，而是标记为 is_deleted 。

完整的转储将确保您始终具有源状态的精确副本。请记住，在一般情况下，在分析中这并不重要，但是保留删除的记录也可能是必需的。

加载（Load）选项

无论您如何从数据源中提取数据，您都需要决定如何将这些更改反映在目标数据上。您可以将更改推送到数据湖中的现有数据，也可以将此新数据与现有数据分开存储。

推送变更

如果将数据库系统用作数据湖，则可以使用推送的更改来更新数据。最终将获得从源数据到数据湖的紧密数据副本，并优化存储。

分开存放

另一种方法是保存更改而不更新记录。如果您使用文件系统并且不想在数据湖上增加很多复杂性，那么这几乎是唯一的方法。这样做的好处是您拥有数据上发生的所有更改的历史记录。

多模式

如果您使用的是文件系统，大多数 EL 供应商会将每个源作为新的模式或文件夹插入到 Lake 中。这是理想的选择，因为您的数据仍将按源进行组织，并且通用命名表不会相互覆盖。

这意味着在查询这些模式时，除表名外，还需要记住指定模式名。

```
SELECT * FROM "salesforce"."_user" AS "SFuser" JOIN "zendesk"."user" AS "ZDuser" ON "SFuser"."email" == "ZDuser"."email"
```

其他提取和加载路线

传统 ETL

由于我们在此处概述的原因，我们建议使用 ELT 而不是 ETL。但是，如果您仍然想以传统方式做事，可以在将数据载入数据湖之前进行转换，您可以使用 Xplenty 或 Amazon Glue。

如果执行此操作，则实际上是一次完成数据湖和数据仓库阶段，而跳过了推荐堆栈中的数据湖部分。这听起来像是一件了不起的事情，但需要注意的是，它不会为您节省任何金钱或时间-实际上，这可能会花费更多。

自己动手做

如果您珍惜时间，金钱，理智和数据完整性，请不要自己动手编写 EL 脚本。如果您 DIY，您会将宝贵的工程资源投入于云解决方案可以以很少的成本和时间完成的工作。您的数据工程师本可以做与您的整体数据基础架构和产品有关的更重要的数据项目。

有时，您可能需要为不受广泛支持的源创建自定义代码。如果确实需要，请至少使用 Apache AirFlow 之类的框架。您想要的最后一件事是随意地部署一堆脚本和 cron 任务。

Written by: Kostas Pardalis
Reviewed by: Dave Fowler , Matt David