Apache Griffin：分布式系统的数据质量解决方案

Apache Griffin 是开源的大数据数据质量解决方案，支持批处理和流模式，其是基于 Apache Hadoop 和 Apache Spark 构建，由 eBay 开发，并于 2016年12月07日进入 Apache 孵化。Griffin 提供了一个可以处理不同的任务，如定义数据质量模型，执行数据质量测量，自动化数据分析和验证，以及跨多个数据系统的统一数据质量可视化的全面的框架，旨在解决大数据应用中数据质量领域的挑战。

项目产生背景

在eBay，当人们在处理大数据(Hadoop或者其它streaming系统)的时候，数据质量的检测是一个挑战。不同的团队开发了他们自己的工具在其专业领域检测和分析数据质量问题。于是我们希望能建立一个普遍适用的平台，提供共享基础设施和通用的特性来解决常见的数据质量问题，以此得到可信度高的数据。

目前来说，当数据量达到一定程度并且有跨越多个平台时（streaming数据和batch数据），数据数量验证将是十分费时费力的。拿eBay的实时个性化平台举个例子，每天我们都要处理大约600M的数据，而在如此复杂的环境和庞大的规模中，数据质量问题则成了一个很大的挑战。

在eBay的数据处理中，发现存在着如下问题：

- 当数据从不同的数据源流向不同的应用系统的时候，缺少端到端的统一视图来追踪数据沿袭(Data Lineage)。这也就导致了在识别和解决数据质量问题上要花费许多不必要的时间。
- 缺少一个实时的数据质量检测系统。我们需要这样一个系统：数据资产(Data Asset)注册，数据质量模型定义，数据质量结果可视化、可监控，当检测到问题时，可以及时发出警报。
- 缺乏一个共享平台和API服务，让每个项目组无需维护自己的软硬件环境就能解决常见的数据质量问题。

为了解决以上种种问题，Griffin 平台从而诞生了。其主要包含以下特性：

- 精确度检测：验证结果集数据是否与源数据是一致的
- 数据剖析：利用数据集的一致性、独特性和逻辑性，来进行统计分析和数值评估。
- 异常监测：利用预先设定的算法，检测出不符合预期的数据
- 可视化监测：利用控制面板来展现数据质量的状态。
- 实时性：可以实时进行数据质量检测，能够及时发现问题。
- 可扩展性：可以用于多个数据系统。
- 可伸缩性：工作在大数据量的环境中，目前运行的数据量约1.2PB(eBay环境)。
- 自助服务：Griffin提供了一个简洁易用的用户界面，可以管理数据资产和数据质量规则；同时用户可以通过控制面板查看数据质量结果和自定义显示内容。

Apache Griffin 架构

Apache Griffin 主要包含三层：数据收集处理层（Data Collection&Processing Layer）、后端服务层（Backend Service Layer）和用户界面（User Interface）。

数据收集处理层

在这一层，最关键的是模型引擎(Model Engine)，Griffin是模型驱动的解决方案。基于目标数据集（targetdata-set）或者源数据集（作为高真的基准数据源 –“golden reference data”），用户可以选择不同的数据质量维度来执行目标数据质量验证。我们有内置的程序库来支持以下检测方式：我们支持两种类型的数据源，batch数据和streaming数据。对于batch数据，我们可以通过数据连接器从Hadoop平台收集数据。对于streaming数据，我们可以连接到诸如Kafka之类的消息系统来做近似实时数据分析。在拿到数据之后，模型引擎将在我们的spark集群中计算数据质量。

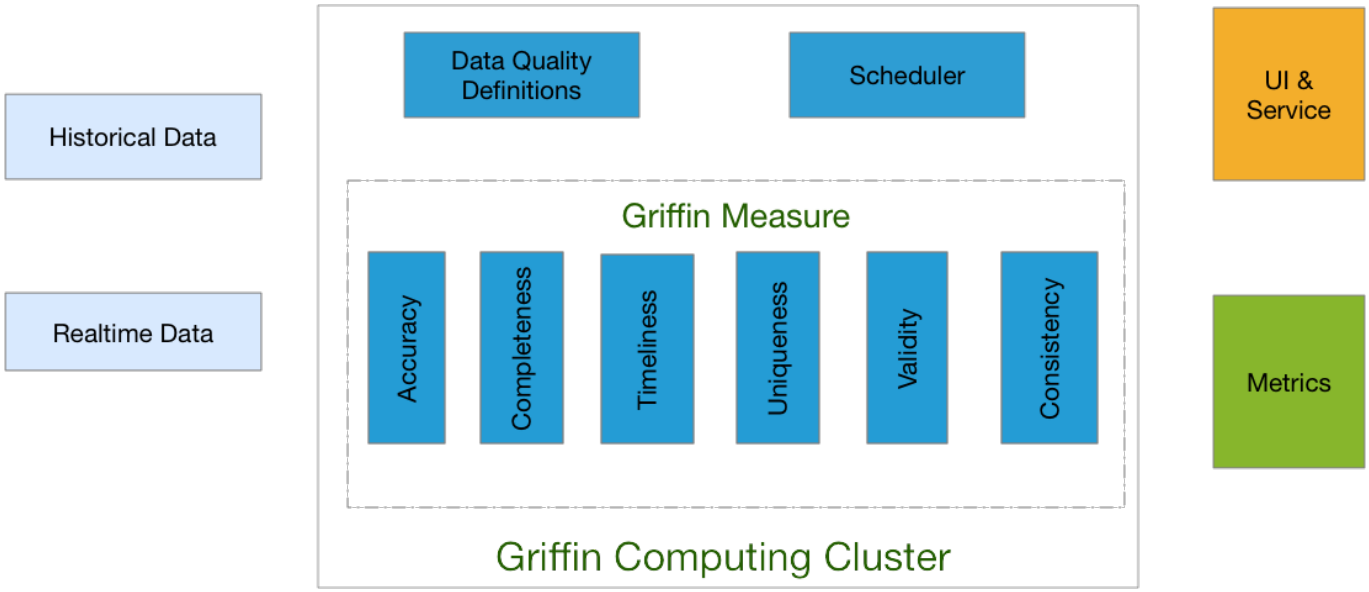
后端服务层

在这一层上，有三个关键组件：

- 核心服务：用来管理元数据，例如模型定义、订阅管理和用户定制等等。
- 作业调度：根据模型的定义创建并调度作业，然后触发模型引擎的运行并取得度量值结果，然后存储度量值，并在检测到数据质量问题时发送电子邮件通知。
- REST服务：我们提供了内置的REST服务来实现Griffin的各项功能，例如注册数据资产，创建数据质量模型，度量发布，度量检索，添加订阅等等。因此，开发人员可以基于这些web服务开发自己的用户界面。

用户界面

Griffin有一个内置的可视化工具，它是基于AngularJS和eCharts开发的web前端应用，可以很好地展现数据质量结果。
架构如下：



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

2018年12月12日，Apache Griffin 成功晋升成 TLP 的，参见这里。Apache Griffin 已经在 eBay, Expedia, 华为, 京东, 美团, PayPal, 平安银行, PPDAI, 唯品会, VMWare 等公司使用。