重磅开源KSQL：用于Apache Kafka的流数据SQL引擎

原创 2017-08-29 薛命灯 大数据杂谈

编辑｜薛命灯
策划｜Natalie

> Apache Kafka 是数据流水线架构中涉及数据处理的关键组件。 8 月 28 日，Kafka Summit 大会在旧金山召开，同一天，Confluent 宣布 Kafka 在变得无处不在的道路上一记新的里程碑：KSQL。

Kafka 的作者 Neha Narkhede 在 Confluent 上发表了一篇博文，介绍了 Kafka 新引入的 KSQL 引擎——一个基于流的 SQL。推出 KSQL 是为了降低流式处理的门槛，为处理 Kafka 数据提供简单而完整的可交互式 SQL 接口。KSQL 目前可以支持多种流式操作，包括聚合（aggregate）、连接（join）、时间窗口（window）、会话（session），等等。

与传统 SQL 的主要区别

KSQL 与关系型数据库中的 SQL 还是有很大不同的。传统的 SQL 都是即时的一次性操作，不管是查询还是更新都是在当前的数据集上进行。而 KSQL 则不同，KSQL 的查询和更新是持续进

行的，而且数据集可以源源不断地增加。KSQL 所做的其实是转换操作，也就是流式处理。

KSQL 的适用场景

1. 实时监控

一方面，可以通过 KSQL 自定义业务层面的度量指标，这些指标可以实时获得。底层的度量指标无法告诉我们应用程序的实际行为，所以基于应用程序生成的原始事件来自定义度量指标可以更好地了解应用程序的运行状况。另一方面，可以通过 KSQL 为应用程序定义某种标准，用于检查应用程序在生产环境中的行为是否达到预期。

2. 安全检测

KSQL 把事件流转换成包含数值的时间序列数据，然后通过可视化工具把这些数据展示在 UI 上，这样就可以检测到很多威胁安全的行为，比如欺诈、入侵，等等。KSQL 为此提供了一种实时、简单而完备的方案。

3. 在线数据集成

大部分的数据处理都会经历 ETL（Extract—Transform—Load）这样的过程，而这样的系统通常都是通过定时的批次作业来完成数据处理的，但批次作业所带来的延时在很多时候是无法被接受的。而通过使用 KSQL 和 Kafka 连接器，可以将批次数据集成转变成在线数据集成。比如，通过流与表的连接，可以用存储在数据表里的元数据来填充事件流里的数据，或者在将数据传输到其他系统之前过滤掉数据里的敏感信息。

4. 应用开发

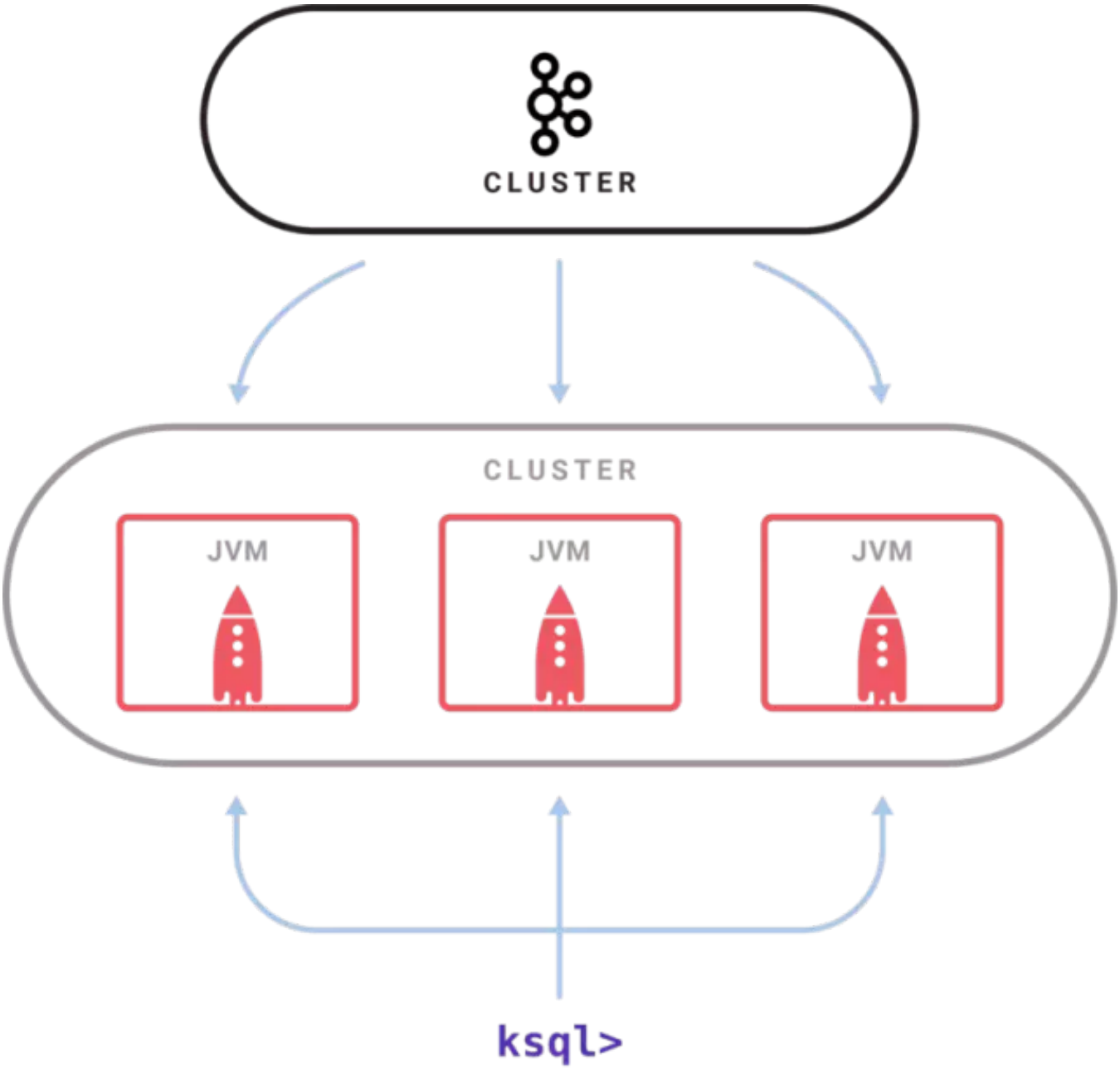
对于复杂的应用来说，使用 Kafka 的原生 Streams API 或许会更合适。不过，对于简单的应用来说，或者对于不喜欢 Java 编程的人来说，KSQL 会是更好的选择。

KSQL 的核心抽象

KSQL 是基于 Kafka 的 Streams API 进行构建的，所以它的两个核心概念是流（Stream）和表（Table）。流是没有边界的结构化数据，数据可以被源源不断地添加到流当中，但流中已有的数据是不会发生变化的，即不会被修改也不会被删除。表即是流的视图，或者说它代表了可变

数据的集合。它与传统的数据库表类似，只不过具备了一些流式语义，比如时间窗口，而且表中的数据是可变的。KSQL 将流和表集成在一起，允许将代表当前状态的表与代表当前发生事件的流连接在一起。

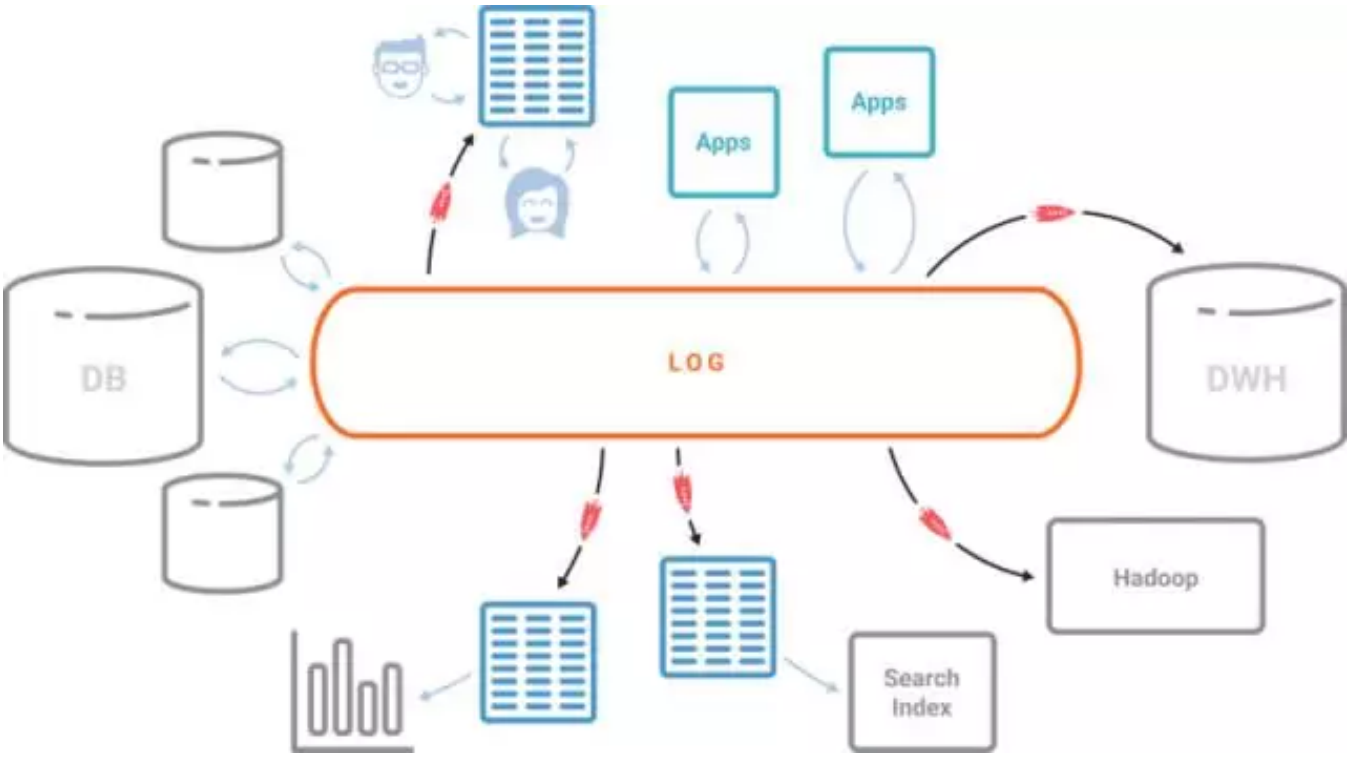
KSQL 架构



KSQL 是一个独立运行的服务器，多个 KSQL 服务器可以组成集群，可以动态地添加服务器实例。集群具有容错机制，如果一个服务器失效，其他服务器就会接管它的工作。KSQL 命令行客户端通过 REST API 向集群发起查询操作，可以查看流和表的信息、查询数据以及查看查询状态。因为是基于 Streams API 构建的，所以 KSQL 也沿袭了 Streams API 的弹性、状态管理和容错能力，同时也具备了仅一次（exactly once）语义。KSQL 服务器内嵌了这些特性，并增加

了一个分布式 SQL 引擎、用于提升查询性能的自动字节码生成机制，以及用于执行查询和管理的 REST API。

Kafka+KSQL 要颠覆传统数据库



传统关系型数据库以表为核心，日志只不过是实现手段。而在以事件为中心的世界里，情况却恰好相反。日志成为了核心，而表几乎是以日志为基础，新的事件不断被添加到日志里，表的状态也因此发生变化。将 Kafka 作为中心日志，配置 KSQL 这个引擎，我们就可以创建出我们想要的物化视图，而且视图也会持续不断地得到更新。

KSQL 的未来

KSQL 目前还处于开发者预览阶段，作者还在收集社区的反馈。未来计划增加更多的特性，包括支持更丰富的 SQL 语法，让 KSQL 成为生产就绪的系统。

这里有 KSQL 的快速入门指南和一个演示程序。可以在 Slack 的 #KSQL 频道上向作者提供反馈信息，或者如果发现 Bug，可以在 GitHub 上提出来。

参考文章：

https://www.confluent.io/blog/ksql-open-source-streaming-sql-for-apache-kafka/

快速入门指南：

https://github.com/confluentinc/ksql/tree/0.1.x/docs/quickstart#quick-start

演示程序：

https://github.com/confluentinc/ksql/blob/0.1.x/docs/demo.md#demo

GitHub 地址：

https://github.com/confluentinc/ksql/issues

今日荐文

点击下方图片即可阅读

崛起的 GPU 数据库大揭秘：多数据流实时分析，如何做到快如闪电？

CNUTCon 全球运维技术大会将于 9 月 10-11 日在上海举行，大会主题是“智能时代的新运维”，并特设“大数据运维”专场，邀请了来自腾讯、苏宁等公司大咖分享他们在最新运维技术实践过程中遇到的坑与经验，更有 Google、Uber、eBay、BAT 等一线技术大牛现场为你解