Apache® Pinot™：开源分布式实时大数据分析基础设施

Apache Pinot 是一个分布式实时分布式 OLAP 数据存储，旨在以高吞吐量和低延迟提供可扩展的实时分析。该项目最初于 2013 年由 LinkedIn 创建，2015 年开源，于 2018 年 10 月进入 Apache 孵化器，2021年08月02日正式毕业成为 Apache 顶级项目。

Apache Pinot 可以直接从流数据源（例如 Apache Kafka 和 Amazon Kinesis）中提取，并使事件可用于即时查询。它还可以从批处理数据源（例如 Hadoop HDFS、Amazon S3、Azure ADLS 和 Google Cloud Storage）中提取。该系统的核心是列式存储，具有多种智能索引和预聚合技术以实现低延迟。这使得 Pinot 最适合面向用户的实时分析。同时，Pinot 也是其他分析用例的绝佳选择，例如内部仪表板、异常检测和临时数据探索。

Apache Pinot 的特点

Apache Pinot 主要有以下特点：

- 面向列的数据库：具有各种压缩方案，如 Run Length，Fixed Bit Length；
- 可插拔索引技术：支持排序索引（Sorted Index），位图索引（Bitmap Index），倒排索引（Inverted Index,），StarTree 索引，Bloom 过滤器，范围索引（Range Index），文本搜索索引（Lucence/FST），Json 索引，地理空间索引（Geospatial Index ）；
- 具有基于查询和 segment 元数据优化查询/执行计划的能力；
- 支持从 Kafka、Kinesis 等流系统近实时的摄取数据，也支持从 Hadoop、S3、Azure、GCS 等批处理系统摄取数据；
- 类似 sql 的查询语言，支持对数据进行选择、聚合、过滤、分组、排序和 distinct 查询；（过往记忆大数据备注：Apache Pinot 使用 Presto 实现了 ANSI SQL 查询语言，支持 JOIN 等操作。）
- 支持多值字段
- 支持水平扩展和容错

Apache Pinot 设计理念

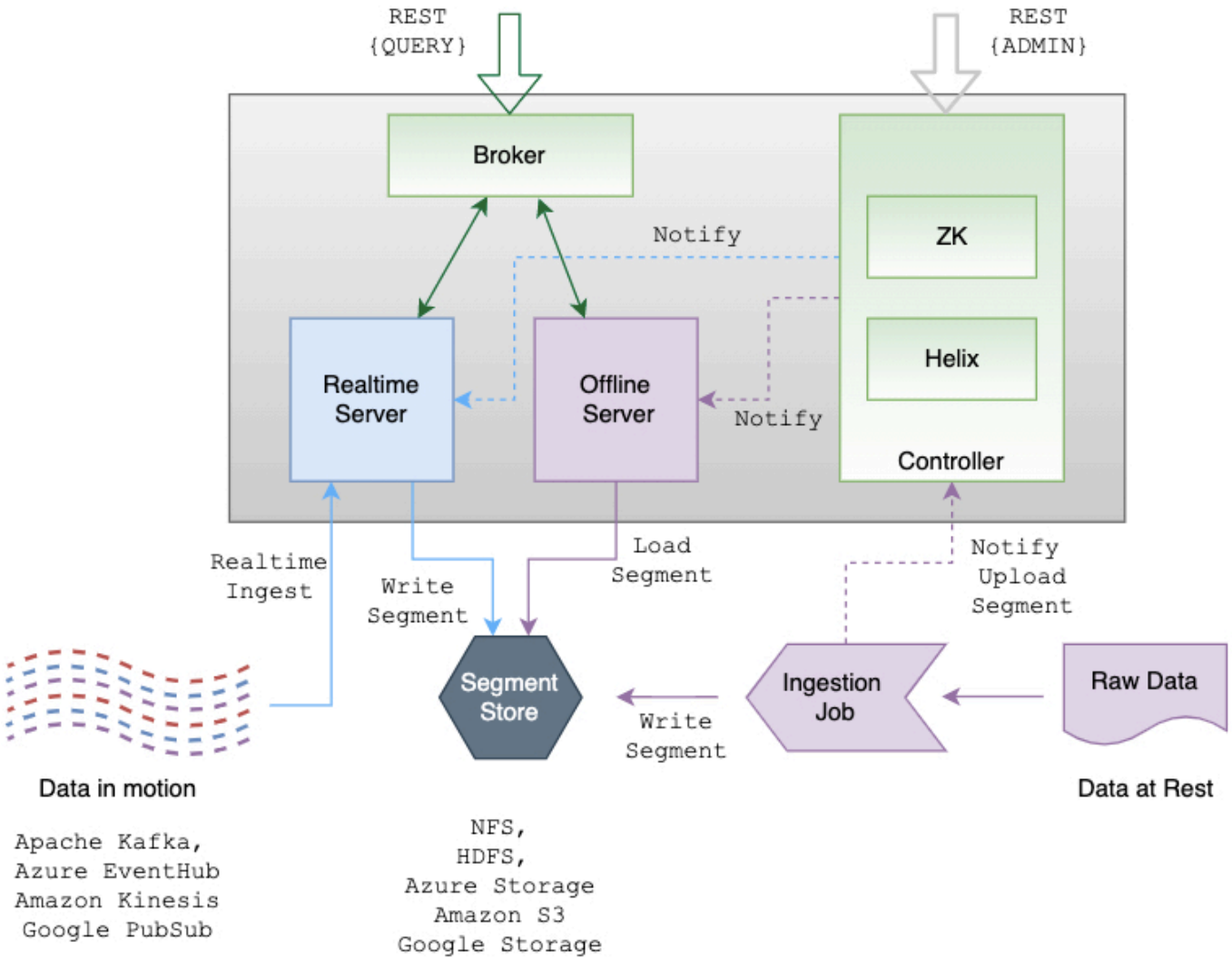
Pinot 是 LinkedIn 和 Uber 的工程师共同设计的，可以根据集群中的节点数量来扩展查询性能。随着添加更多节点，查询性能总是会根据期望的每秒查询量配额提高。为了在不降低性能的情况下实现无限数量节点和数据存储的水平可伸缩性，Pinot 遵守以下设计原则：

- 高可用性：构建 Pinot 是为了为客户应用程序提供低延迟的分析查询。根据设计，Pinot 没有单点故障。当节点故障时，系统继续提供查询服务。
- 水平可伸缩：在工作负载发生变化时通过添加新节点进行伸缩的能力；
- 延迟 vs 存储：构建 Pinot 是为了在高吞吐量的情况下提供低延迟。为此，开发了段分配策略（segment

assignment strategy）、路由策略、星树索引（star-tree indexing）等特性来实现这个功能。
- 不可变数据：Pinot 假设所有存储的数据都是不可变的。对于 GDPR 遵从性，我们提供了一个附加解决方案来清除数据，同时提供性能保证；
- 动态配置更改：必须在不影响查询可用性或性能的情况下执行添加新表、扩展集群、摄取数据、修改索引配置和重新平衡等操作。

Apache Pinot 的架构

下面是 Apache Pinot 的架构：



如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：过往记忆大数据

从上图可以看出主要有 Controller, Broker, Server, 以及 Minion 等组件。

Apache Helix && Apache Zookeeper

Pinot 使用 Apache Helix 进行集群管理。Helix 作为代理（agent）嵌入到不同的组件中，并使用 Apache Zookeeper 来协调和维护整个集群状态和运行状况。所有的 Pinot servers 和 brokers 都由 Helix 管理。Helix
是一个通用的集群管理框架，用于管理分布式系统中的分区和副本。可以将 Helix 看作是一个事件驱动的发现服务，它具有推和拉通知功能，可以将集群的状态驱动到理想的配置。

Controller

Pinot 的 Controller 充当集群整体状态和运行状况的驱动程序。由于它的角色是 Helix
的参与者（participant）和旁观者（spectator），它驱动其他组件的状态，所以它通常是在 Zookeeper 之后启动的第一个组件。启动 Controller
需要两个参数：Zookeeper地址和集群名称。如果集群还不存在，Controller 将自动通过 Helix 创建一个集群。

Broker

Broker 的职责是将给定的查询路由到适当的 server 实例。Broker 将收集并合并来自所有 server 的响应，并将其发送回请求客户机。broker 提供接收 SQL 查询并以 JSON 格式返回响应的 HTTP 端点。

Server

Server 管理 segments，并在查询处理期间完成大部分繁重的工作。Pinot 有两种 servers：实时 server 和离线 server，但 server 并不真正知道它将是实时 server 还是离线 server 。server 的职责取决于表分配策略（table assignment strategy）。

Minion

Minion 是一个可选的组件。Minion 用于从 Pinot 集群中清除数据（比如出于英国的 GDPR 遵从性等原因）。

关于 Apache® Pinot™ 的更多介绍可以到 https://pinot.apache.org/ 查看。