HongKe 虹科

HK-Acceldata
集成指南

acceldata

1 HK-Acceldata Pulse集成指南

HK-Acceldata Pulse集成

与分布式数据系统集成以观察和改进数据处理

HK-Acceldata Pulse 允许数据团队观察和改进跨云、混合云和本地生态系统的数据处理。它为开源分布式文件系统、数据库、数据仓库、查询引擎和流平台提供连接器，如图 1 所示。

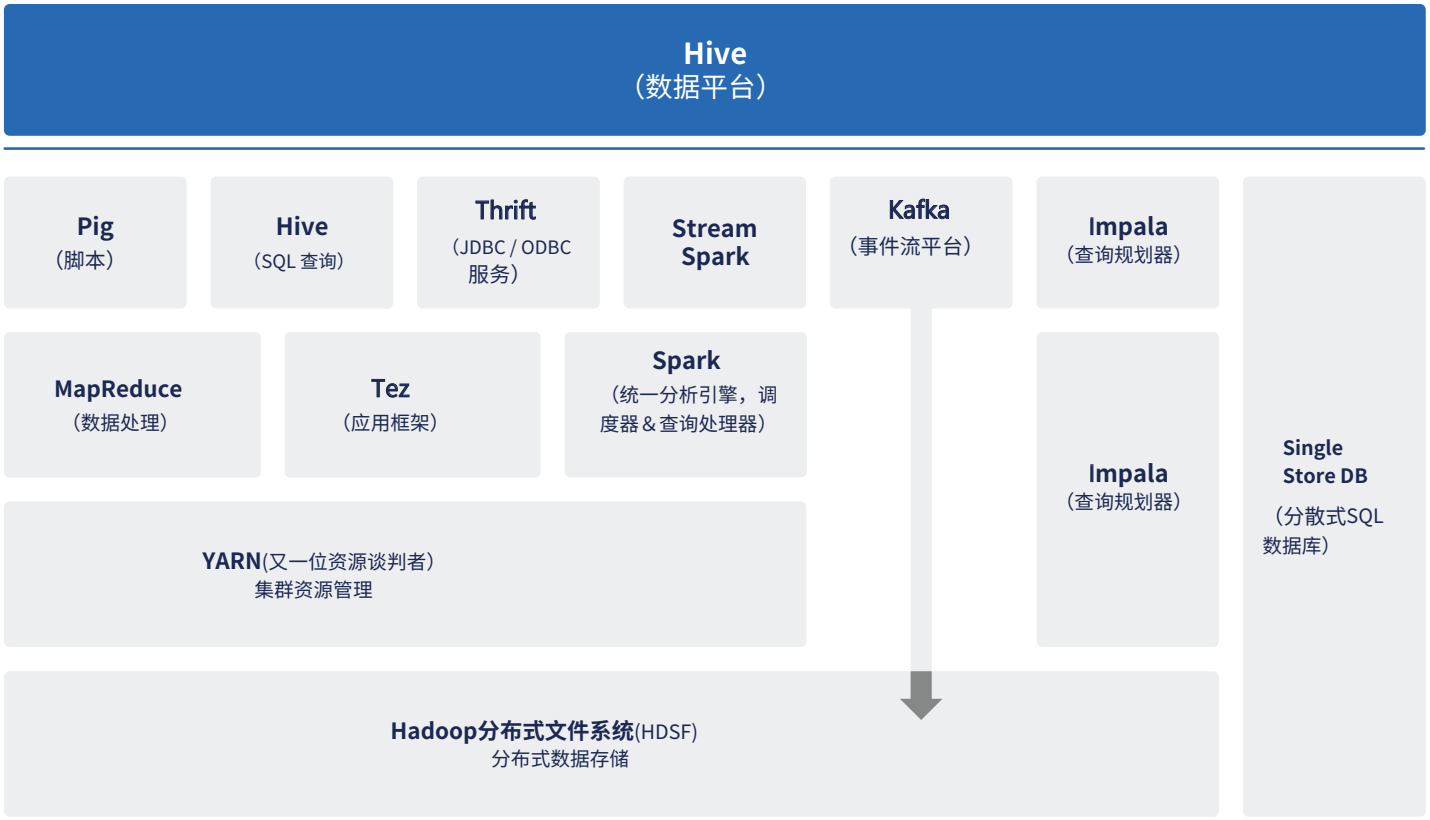

图1

Pulse 不仅监控单个组件；它将跨基础设施、数据和应用层的事件关联起来。因此，数据专家和数据工程师可以减少数据停机时间、加速数据消耗并优化容量规划以提高数据性能。以下两页中的表格提供了 Pulse 支持的数据源示例。定期添加新的连接器。

Pulse 支持的数据源

数据源	描述	使用Pulse监控
Apache Hadoop 分布式文件系统 (HDFS)	分布式文件系统，提供对商品硬件上数据的高吞吐量访问。	集群节点、存储容量、CPU、内存、客户端、文件、操作和进程的性能指标。按类型、大小、温度、上次修改、上次访问和用户跟踪文件
Apache Hadoop YARN	集群资源管理和作业调度优化的框架。负责管理和监控工作负载。	不同应用程序（YARN、MapReduce、SPARK 或 Tez）的性能
Apache Hadoop MapReduce	驻留在硬件节点集群中的大型数据集的并行数据处理	用户、查询、CPU 和内存的作业性能
查询引擎、数据库和数据仓库		
Apache Spark	用于批处理 (SQL)、流式传输、机器学习和图形数据的大规模数据处理的统一分析引擎、调度程序和查询优化器 在现代应用中	用户、查询、CPU 和内存的作业性能。跟踪用户数量、运行作业的队列、作业状态。跟踪工作趋势、配置和异常情况。
Apache Spark Thrift	使 JDBC 和 ODBC 客户端能够执行 Spark SQL 查询的应用程序服务	用户、查询、CPU 和内存的作业性能
Apache Hive on Spark	一个利用 Apache Spark 支持在 Apache Hadoop 中查询和分析数据的数据仓库。Apache Hive 提供对存储在 HDFS 或 HBase 中的文件的访问。	用户、查询、CPU 和内存的作业性能。
Apache HBase	可扩展的分布式 NoSQL 数据库，支持商品硬件集群中超大型表的结构化数据存储，支持对大数据的随机、实时读/写访问	HBase 集群、区域、表和读/写请求的容量和性能。发现集群和快照中的异常（HDFS 文件和 HBase 表在特定时间点的备份）
Apache Impala	Apache Hadoop 的查询计划器、协调器和引擎，可以对存储在 HDFS 或 HBase 中的数据进行 SQL 查询。	查询的性能、表中的错误以及守护程序识别的异常。
Apache Hive LLAP（Live Long and Process)	通过缓存频繁执行的查询或数据以加快执行速度来优化 Hive 查询性能的组件。	用户和查询的作业性能。跟踪 JVM 守护进程和进程。
Apache Tez	用于构建基于 YARN 的数据应用程序的应用程序框架，具有改进的 MapReduce 性能、优化的资源管理和动态流决策	用户、查询、CPU 和内存的作业性能。
SingleStore DB（以前的 MemSQL）	一个分布式的关系型 SQL 数据库	数据库、数据管道、查询、用户登录和审计的性能。

数据源	描述	使用Pulse监控
流媒体解决方案		
Apache Spark Streaming	支持实时数据流的可扩展、高吞吐量流处理的 Spark 扩展	用户、状态和正在运行作业的队列的作业性能。根据 CPU 和内存使用情况监控输入速率、流时间和输出。
Apache Kafka	用于高性能数据管道、流分析和关键任务应用程序的开源分布式事件流平台	跨生产者、主题和消费者流式传输的数据量
Apache NiFi	用于自动化和管理系统间数据流的数据流平台	NiFi Java 虚拟机指标，例如内存使用、线程和垃圾收集时间；NiFi 进程组指标，包括名称、活动线程、发送、排队、传输、接收的数据量；和 NiFi 流文件指标，包括文件数量、文件大小和处理持续时间。

2 HK-Acceldata Torch集成指南

HK-Acceldata Torch集成

连接数据源以观察和提高数据质量

HK-Acceldata Torch 是一种数据可观察性工具，可在数据处理和通过复杂数据管道时监控数据的质量。它包括一个软件开发工具包(SDK)，数据工程师可以使用它来了解复杂管道的完整性。Torch 的预构建连接器支持在云、混合云、流媒体和本地观察数据源的能力。数据工程师可以根据需要为数据源创建新的连接器。

数据源	描述
数据库	
微软 Azure SQL 数据库	用于事务和分析工作流的完全托管的 RDBMS 数据库服务。在 Azure 云中作为平台即服务 (Paas) 运行。
MySQL 数据库	用于事务、分析和机器学习工作负载的开源 RDBMS。在云端或本地运行。
Oracle数据库	用于事务、分析和混合工作负载的 RDBMS。在云端或本地运行。
PostgreSQL 数据库	支持和扩展 SQL 以支持事务、分析和机器学习工作负载的开源 RDBMS 系统。在云端或本地运行。
SingleStore DB（以前的 MemSQL）	用于事务、分析和机器学习工作负载的分布式 RDBMS。在云端、本地或容器中运行。

数据仓库/数据湖/查询服务	
Amazon Redshift	完全托管的数据仓库服务。它是一个 OLAP、面向列的数据库，使用大规模并行处理 (MPP) 快速处理大量结构化和非结构化数据，以支持实时分析、日志分析、BI 和机器学习工作负载。在 AWS 云中运行。
Apache HBase	开源 NoSQL 分布式存储，支持对大数据的随机、实时、读/写访问。在 Apache Hadoop 文件系统 (HDFS) 之上本地运行。
Apache Hive	A开源数据仓库，用于使用 SQL 分析驻留在分布式存储中的大型数据集。它提供对存储在 HDFS 或 HBase 中的文件的访问，并通过 Apache Spark、Tez 或 Map Reduce 提供查询执行。
Databricks	利用 Apache Spark 的可扩展数据仓库和数据湖平台。它支持 SQL 查询、数据流、图形处理和机器学习工作负载，并包括 ETL，与云和本地的许多数据源集成。在 Amazon AWS、Google Cloud 和 Microsoft Azure 上的云中运行。
Google BigQuery	用于 BI、大数据分析、地理空间分析和机器学习的全托管数据仓库。在 Google Cloud Platform 中作为平台即服务 (Paas) 运行。
MongoDB	一个 NoSQL 面向文档的数据库，用于构建高可用性和可扩展的互联网应用程序。在 AWS、Google Cloud Platform 或 Microsoft Azure 上的本地或云中运行。
Presto	开源分布式 SQL 查询引擎，用于在数据所在的数据源上运行交互式查询。Presto 查询可以组合来自多个来源的数据。它支持计算和存储分离，可以部署在云端或本地。
Snowflake	用于数据仓库、数据湖、数据共享、数据工程、数据科学和数据应用程序的云计算数据平台。Snowflake 架构的三层（存储、计算和云服务）作为软件即服务在 AWS、谷歌云平台或 Microsoft Azure 上的云中运行。
Teradata	用于开发大型数据仓库以支持分析工作负载的开源 RDBMS。在云端或本地运行。

文件系统/查询服务	
Apache Hadoop 分布式文件系统 (HDFS)	分布式文件系统，提供对商品硬件上数据的高吞吐量访问。旨在可靠地存储非常大的数据集，并将这些数据集以高带宽流式传输到应用程序。在本地运行。
Amazon S3（简单存储服务）	对象存储服务，用于为数据湖、网站、云应用程序、备份/档案、机器学习和分析存储任意数量的数据。在 AWS 云上运行
Amazon Athena	交互式查询服务，用于使用标准 SQL 查询和分析 Amazon S3 中的大数据。与 AWS Glue 数据目录的开箱即用集成。
Google Cloud Storage	对象存储服务，用于存储和访问 Google Cloud Platform 中的数据湖、网站、图像和流式视频数据、备份/存档。

HK-Acceldata Torch 的连接器

一体化	
Amazon Web Services (AWS) Glue	完全托管的数据集成服务，用于为应用程序、分析和机器学习发现、准备和组合数据。包括用户可以访问和搜索的目录。提供与 Amazon Redshift、Amazon RDS、Amazon DocumentDB、Amazon DynamoDB、Amazon Kinesis、Amazon S3、Apache Kafka 和 MongaDB 的连接。在 AWS 云上运行。

流式传输（动态或批处理数据）	
Apache Kafka	用于实时流数据管道、流分析和可适应动态数据流的应用程序的开源分布式事件流平台。它结合了流处理、消息传递和存储，可以分析历史和实时数据。在本地运行。

分析	
Tableau	数据可视化和分析平台，用于将多个数据源与基本 ETL 操作相结合，以支持探索性分析、仪表板和报告。在本地或云端运行。

通过 HK-Acceldata 软件开发人员套件实现的数据管道连接器

使用 SDK，开发人员可以将 Acceldata Torch 集成到 Apache Airflow 等开源工作流平台中。或者，开发人员可以利用 API 直接 与数据源和转换（ETL 或 ELT）工具集成，以构建自定义工作流。

工具	描述
Apache Airflow	开源工作流平台，使开发人员能够以编程方式创作、安排和监控工作流。Airflow 提供与 Google Cloud Platform、Amazon Web Services、Microsoft Azure、Apache Spark、Apache HDSF、Postgres、MySQL 和许多其他第三方服务的强大集成。
Apache Spark	用于大数据处理的开源统一分析引擎。
Amazon EMR	一个托管集群平台，可简化在 AWS 上运行大数据框架（例如 Apache Spark 和 Apache Hive）来处理和分析大量数据。EMR 还可以将数据转换和移入和移出其他 AWS 存储，例如 Amazon Simple Storage Service (S3) 和 Amazon DynamoDB。
Cloudera 数据平台	混合数据云，用于管理跨主要公共云和私有云的数据生命周期，同时无缝连接到本地环境。
Databricks	可扩展的数据湖库平台，用于结构化和非结构化数据，具有自动化和可靠的 ETL，并支持机器学习
Google Dataproc	一种完全托管且高度可扩展的服务，用于运行 Apache Spark、Presto 和 30 多个开源工具和框架。与 Google Cloud 完全集成，以支持数据湖现代化、ETL 和数据科学。
Presto	开源分布式 SQL 查询引擎，用于在数据所在的数据源上运行交互式查询。Presto 查询可以组合来自多个来源的数据。它支持计算和存储分离，可以部署在云端或本地。
Snowflake	用于数据仓库、数据湖、数据共享、数据工程、数据科学和数据应用程序的云计算数据平台
Tableau	用于发现、准备和组合应用程序、分析和机器学习的数据的可扩展和无服务器数据集成服务。包括用户可以访问和搜索的目录。

HongKe 虹科

虹科电子科技有限公司

www.hongcloudtech.com
hongcloudtech@hkaco.com

广州市黄埔区神舟路18号润慧科技园C栋6层

T (+86) 400-999-3848
M (+86) 155 2866 3362

各分部：广州 | 成都 | 上海 | 苏州 |
西安 | 北京 | 台湾 | 香港 | 美国硅谷

联系我们

行业交流群

获取更多资料

hongcloudtech.com