Apache 旗下与大数据相关的开源项目列表 – UNIX2GO

unix2go.com/big-data-projects-under-apache-asf/

Apache 软件基金会（ASF）旗下与大数据相关的开源项目列表，这是一个非常庞大的生态系统，要列出“所有”项目几乎不可能，因为项目的活跃度、相关性以及“大数据”的定义本身就有弹性。不过，我可以整理一个尽可能详尽的列表，包含许多在业界被广泛使用或具有重要影响力的大数据相关 Apache 项目，并进行功能和场景的对比。

更全面的 Apache 大数据相关项目对比

项目 (Project)	功能特点 (Core Function/Features)	使用场景 (Use Cases)
Airflow	工作流编排、调度和监控平台。	定义、调度和监控复杂的数据管道（ETL、ELT）、机器学习流程等。
Arrow	跨语言、跨平台的列式内存数据格式规范和库。	加速不同大数据系统（如 Spark, Pandas, Drill）之间的数据交换，减少序列化/反序列化开销。
Atlas	Hadoop 生态系统的数据治理和元数据管理框架。	数据发现、血缘追踪、元数据集中管理、数据分类、安全策略实施。
Avro	基于 Schema 的数据序列化系统。	RPC 调用和持久化存储中的数据序列化，尤其在 Hadoop 生态内广泛使用。
BookKeeper	可扩展、容错、低延迟的持久化日志存储服务。	为分布式系统（如 Pulsar）提供可靠的 Write-Ahead Logging (WAL) 或消息存储。
Cassandra	分布式、高可用、高可扩展的 NoSQL 宽列数据库。	需要高写入吞吐量、多数据中心部署、无单点故障的场景，如物联网数据、用户活动日志。
Doris (Incubating)	MPP（大规模并行处理）分析型数据库。	实时数据仓库、交互式商业智能(BI)报表、Ad-hoc 即席查询。
Drill	Schema-free（模式自由）的分布式 SQL 查询引擎。	直接查询多种异构数据源（NoSQL、文件系统、云存储），无需预先定义 Schema 或进行 ETL。
Druid	面向列存储、支持实时摄入的高性能分布式数据存储（OLAP 引擎）。	对大规模（特别是时间序列）事件数据进行快速聚合和探索性分析，常用于实时监控仪表盘。
Flink	分布式流处理引擎，以流为核心，但也支持批处理，提供精确一次处理语义。	对实时性、状态管理和准确性要求高的流处理场景，如实时监控、欺诈检测、复杂事件处理。
Flume	分布式、可靠、高可用的海量日志数据收集、聚合和传输服务。	从各种来源（如 Web 服务器）收集日志数据，并将其传输到中央存储（如 HDFS, Kafka）。
Gobblin (Incubating)	通用数据摄取框架。	从各种来源（数据库、Kafka、文件系统等）统一摄取数据到目标系统（如 HDFS, Kafka）。

Hadoop	分布式文件系统(HDFS)、资源管理(YARN)和计算框架(MapReduce)的基础平台。	大规模数据集的存储和批处理，是大数据生态系统的基石。
HBase	构建于 HDFS 之上的分布式、可伸缩、面向列的 NoSQL 数据库。	需要对海量数据进行实时随机读/写访问的场景，如用户画像、实时推荐、消息存储。
Hive	基于 Hadoop 的数据仓库工具，提供 SQL 查询接口(HiveQL)来分析存储在 HDFS 上的数据。	大规模数据的批处理 ETL、数据汇总、报表分析，适合对延迟不敏感的分析任务。
Hudi (Hoodie)	数据湖平台，在 HDFS/云存储上提供记录级别的插入、更新、删除和增量处理能力。	构建事务性数据湖，管理缓慢变化维度（SCD），进行增量数据处理和流式摄取。
Iceberg	用于大型分析数据集的开放表格式（Table Format）。	在数据湖（如 S3, HDFS）上管理大型表，提供 ACID 事务、模式演进、时间旅行等功能，优化查询性能。
Impala	高性能、低延迟的 MPP（大规模并行处理）SQL 查询引擎，主要查询 Hadoop 中的数据。	对存储在 HDFS、HBase、Kudu 中的数据进行快速的交互式 SQL 分析和 BI 报表。
Kafka	分布式、高吞吐量的事件流平台（消息队列/发布订阅系统）。	实时数据管道、日志收集、消息队列解耦、流处理引擎的数据源/汇、事件溯源。
Knox	Hadoop 集群安全访问的应用程序网关。	为 Hadoop 服务（REST/HTTP API）提供统一的访问入口、认证、授权和代理。
Kudu	分布式列式存储引擎，针对快速分析和实时数据进行了优化。	需要快速扫描分析与低延迟随机读写并存的场景，如实时报表、时序数据存储、需要更新的数据分析。
Kylin	基于 Hadoop/Spark 的分布式 OLAP（在线分析处理）引擎。	通过预计算（Cube）技术，为超大规模数据集提供亚秒级的 SQL 查询响应能力，支持 BI 报表。
Lucene	高性能、功能齐全的文本搜索引擎库（被 Solr, Elasticsearch 使用）。	底层搜索引擎库，为应用程序提供索引和搜索功能。

NiFi	可视化、易于使用的数据流处理和自动化系统。	在不同系统间自动化数据流转、转换、路由和处理，构建实时 ETL 和数据集成管道。
Oozie	Hadoop 作业（MapReduce, Pig, Hive, Spark 等）的工作流调度器。	在 Hadoop 集群中编排和调度一系列相互依赖的作业。 (Airflow 更现代化)
ORC	(Optimized Row Columnar) 高效的列式存储文件格式。	优化 Hadoop 生态中的数据存储和查询性能，特别适合 Hive。
Parquet	高效的列式存储文件格式，支持复杂的嵌套数据结构。	优化 Hadoop 生态及云存储上的分析查询性能，被 Spark 等引擎广泛支持。
Phoenix	HBase 的 SQL 接口层。	通过标准的 SQL 和 JDBC 接口访问 HBase 数据，支持低延迟的 OLTP/操作型查询。
Pig	基于 Hadoop MapReduce 的高级数据流语言和执行框架 (Pig Latin)。	使用脚本语言简化 Hadoop 上的复杂数据转换和 ETL 任务。（近年来 Spark/Flink 更常用）
Pinot	实时分布式 OLAP 数据存储。	面向用户的实时分析、仪表盘等场景，提供对海量数据的超低延迟查询。
Pulsar	分布式、云原生的发布订阅消息系统和流平台，支持多租户和跨地域复制。	企业级消息队列、流数据处理、微服务间通信，提供统一的消息传递和流处理模型。
Ranger	提供 Hadoop 平台及其组件集中安全策略管理和监控的框架。	对 HDFS, Hive, HBase 等组件进行统一的认证、授权管理和审计。
RocketMQ	分布式消息队列和流处理平台，强调金融级的可靠性和顺序消息。	订单处理、交易系统、日志流处理等需要高可靠、事务消息或严格顺序消息的场景。
Samza	分布式流处理框架（通常与 Kafka 配合使用）。	构建实时数据处理应用，支持大规模状态管理。 (与 Flink, Spark Streaming 竞争)
Solr	基于 Lucene 的企业级开源搜索平台。	全文搜索、分面搜索、实时索引、数据库集成搜索、地理空间搜索等。
Spark	快速、通用的大数据处理引擎，支持批处理、交互式查询、流处理、机器学习和图计算。	多功能大数据处理平台，覆盖 ETL、交互式分析(Spark SQL)、实时流处理(Structured Streaming)、机器学习(MLlib)、图计算(GraphX)。

Sqoop	在 Hadoop 与关系型数据库（RDBMS）之间进行批量数据传输的工具。	从 RDBMS 导入数据到 HDFS/Hive/HBase，或从 Hadoop 导出数据到 RDBMS。
Storm	分布式实时计算系统（流处理框架）。	实时分析、在线机器学习、持续计算，对低延迟要求高的流式任务。（近年来 Flink 更流行）
Tez	基于 YARN 的应用框架，优化 Hadoop 数据处理流程（如 Hive, Pig）。	提升 Hive 等依赖 MapReduce 的任务的执行效率，通过 DAG（有向无环图）优化执行计划。
ZooKeeper	为分布式系统提供高可用的集中协调服务。	分布式锁、配置管理、命名服务、领导者选举、集群成员管理（被 HBase, Kafka 等广泛依赖）。

重要提示:

- **项目状态:** 这个列表包含活跃项目和一些仍在孵化 (Incubating) 或虽然不如以前流行但仍有使用的项目。Apache 项目的状态会随时间变化。
- **并非全部:** Apache 旗下还有更多与数据处理、中间件、服务器等相关的项目，这里主要聚焦于与“大数据”处理强相关的项目。可能仍有遗漏。
- **生态整合:** 这些项目通常不是孤立使用的，而是相互集成构成复杂的大数据解决方案。
- **技术选型:** 选择哪个项目取决于具体的业务需求、数据特性、团队技能、性能要求、运维成本等多种因素。