# 概述

参考：

<https://baijiahao.baidu.com/s?id=1684665476419300396&wfr=spider&for=pc>

## 大数据

大数据是指无法在一定时间范围内无法用常规软件工具进行捕捉、管理和处理的数据集合，需要使用新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。主要解决海量数据的存储和分析计算问题。

Hadoop一个用于分布式大数据处理的开源框架，由Apache基金会所开发的分布式系统基础框架，允许使用简单的编程模型在跨计算机集群的分布式环境中存储和处理大数据。同常是指一个更广泛的概念-Hadoop生态圈。

## 特点(5V)

Volume（大量）：数据存储量

Velocity（高速）：大数据区分于传统数据挖掘的最显著特征

Variety（多样）：数据分为结构化数据和非结构化数据

value（低价值密度）：低价值密度的高低与数据总量成反比

Veracity（真实性）

## 优点

高可靠性：Hadoop底层维护多个数据副本，即使某个存储出现故障，也不会导致数据的丢失；

高扩展性：在集群间分配任务数据，可方便的扩展数以千计的节点；

高效性：hadoop是并行工作的，以加快任务处理速度；

高容错性：能够自动将失败的任务从新分配。

# 原理

HDFS：分布式文件系统

MapReducer：分布式数据处理模型和执行环境（分布式计算）

YARN：job调度和资源管理框架

Common：支持其他模块的工具模块（辅助工具）

## HDFS

## MapReduce

## Yarn

### 原理

### Mesos

参考：

<https://www.jianshu.com/p/9c52edc12a9d>

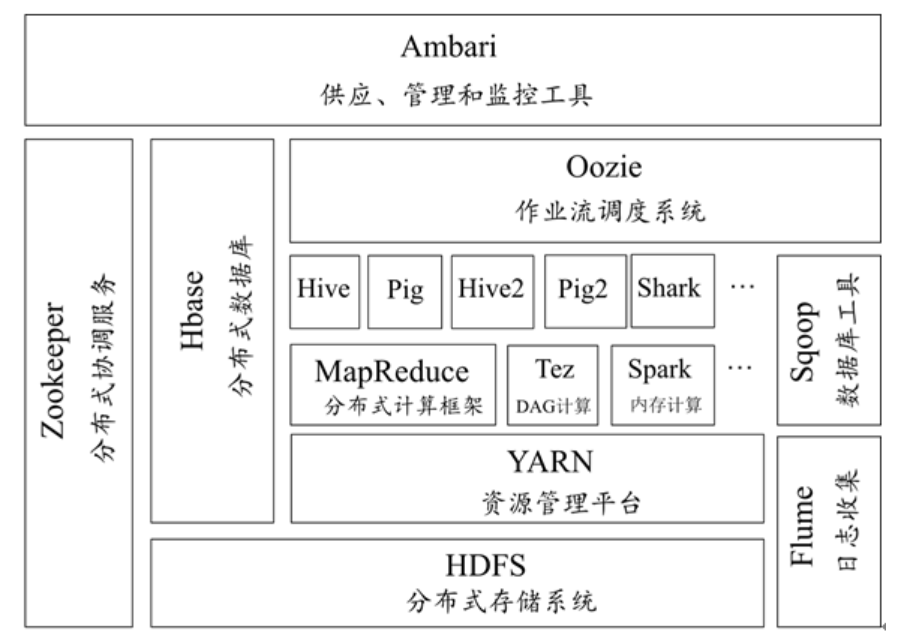
# 模式

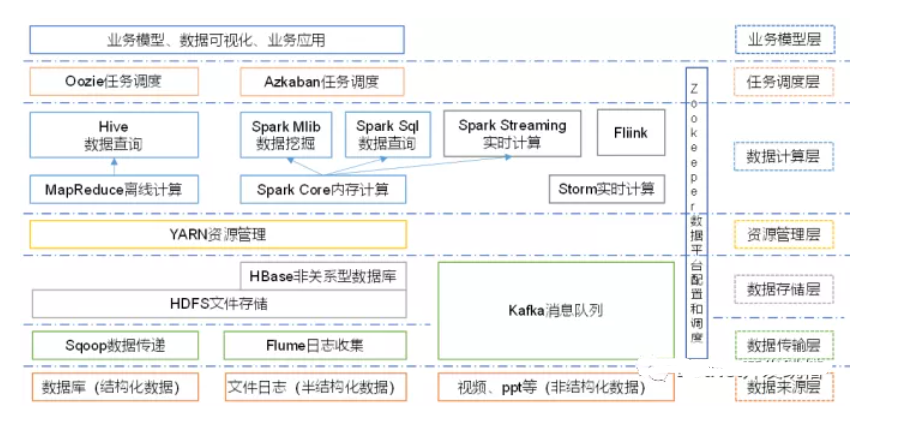
单机模式：不需要配置，Hadoop被认为单独的java进程，经常用来做调试；

伪分布式模式：可以看做只有一个节点的集群，在这个集群中，这个节点既是master，也是slave，既是namenode，也是datanode，既是jobtracker，也是tasktracker；

完全分布式模式：Hadoop的守护进程运行在由多台主机搭建的集群上，是真正的生产环境。

# 生态圈





## Sqoop

Sqoop数据迁移工具，是一款开源的数据导入导出工具，主要用于在Hadoop与传统的数据库间进行数据的转换，它可以将一个关系型数据库（例如MySQL、Oracle等）中的数据导入到Hadoop的HDFS中，也可以将HDFS的数据导出到关系型数据库中，使数据迁移变得非常方便。

## Flume

Flume日志收集工具，是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

说明：

比如在异构数据库的传输中，如果使用全量导出的方式，则Flume可以用来加载日志。

## Kafka

Kafka：一种高吞吐量的分布式发布订阅消息系统。

说明：

Kafka、Flume 是非常成熟的日志和事件流采集工具，为数据湖提供了实时接入的通道。

## Zookeeper

Zookeeper分布式协作服务Zookeeper是一个分布式的，开放源码的分布式应用程序协调服务，是Google的Chubby一个开源的实现，是Hadoop和HBase的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等用于构建分布式应用，减少分布式应用程序所承担的协调任务。

## HDFS

HDFS分布式文件系统HDFS是Hadoop分布式文件系统，它是Hadoop生态系统中的核心项目之一，是分布式计算中数据存储管理基础。HDFS具有高容错性的数据备份机制，它能检测和应对硬件故障，并在低成本的通用硬件上运行。另外，HDFS具备流式的数据访问特点，提供高吞吐量应用程序数据访问功能，适合带有大型数据集的应用程序。

## HBase

Hbase分布式存储系统，是Google Bigtable克隆版，建立在HDFS之上，它是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库。

和传统关系数据库不同，HBase采用了BigTable的数据模型：增强的稀疏排序映射表（Key/Value），其中，键由行关键字、列关键字和时间戳构成。

HBase提供了对大规模数据的随机、实时读写访问，同时，HBase中保存的数据可以使用MapReduce来处理，它将数据存储和并行计算完美地结合在一起。

## YARN

Yarn（Yet Another Resource Negotiator）资源管理框架，是Hadoop 2.0中的资源管理器，它可为上层应用提供统一的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处。

## MapReduce

MapReduce分布式计算框架，是一种计算模型，用于大规模数据集（大于1TB）的并行运算。

“Map”对数据集上的独立元素进行指定的操作，生成键值对形式中间结果；“Reduce”则对中间结果中相同“键”的所有“值”进行规约，以得到最终结果。 MapReduce这种“分而治之”的思想，极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。

## Hive

Hive基于Hadoop的数据仓库Hive是基于Hadoop的一个分布式数据仓库工具，可以将结构化的数据文件映射为一张数据库表，将SQL语句转换为MapReduce任务进行运行。

其优点是操作简单，降低学习成本，可以通过类SQL语句快速实现简单的MapReduce统计，不必开发专门的MapReduce应用，十分适合数据仓库的统计分析。

说明：

在数据湖中，存在一种使用场景，即关系型数据库将数据同步到Hive（经过MapReduce），然后HDFS存储后经过加工使用。

## Spark

一种基于内存的分布式计算框架，与Mapreducer不同的是Job中间输出结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的MapReduce的算法，内部提供了大量的库，如 Spark Sql、Spark Streaming等。

## Flink

一种基于内存的分布式计算框架，用于实时计算场景较多。

说明：

Spark 支持数据的批处理和 ETL 转换，广泛用于数据湖上的清洗、聚合和机器学习。

Flink 支持实时流处理，非常适合对实时数据管道的接入与分析，是现代数据湖中非常关键的组件。

## Storm

对数据流做连续查询，在计算时就将结果以流动形式输出给用户，用于“连续计算”。

## Tez

## Impala

用于处理存储在Hadoop集群中大量数据的MPP（大规模并行处理）SQL查询引擎，与Hive不同，不基于MapReducer算法。它实现了一个基于守护进程的分布式结构，负责在同一台机器上运行的查询执行所有方面，执行效率高于Hive。

### Mahout

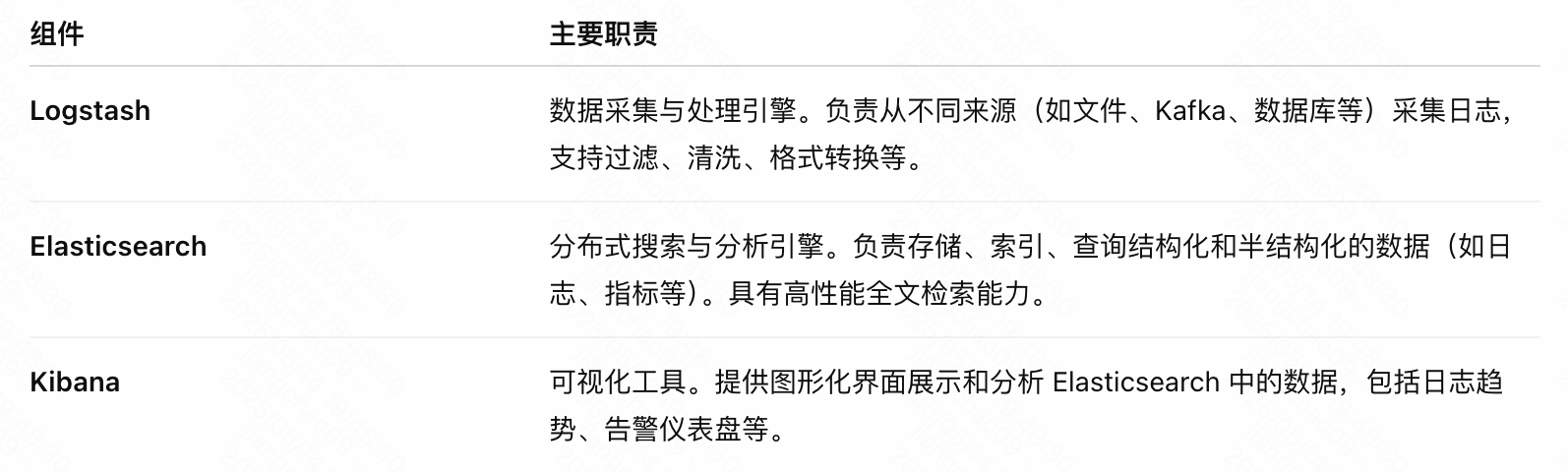
Mahout数据挖掘算法库Mahout是Apache旗下的一个开源项目，它提供了一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序。

Mahout包含许多实现，包括聚类、分类、推荐过滤、频繁子项挖掘。此外，通过使用 Apache Hadoop 库，Mahout 可以有效地扩展到云中。

### Spark Sql

### Spark Streaming

## ELK



### Filebeat

Filebeat是一个日志文件托运工具，在服务器上安装客户端后，Filebeat会监控日志目录或者指定的日志文件，追踪读取这些文件（追踪文件的变化，不停的读），并且转发这些信息到ElasticSearch或者Logstarsh中存放。  
 当你开启Filebeat程序的时候，它会启动一个或多个探测器（prospectors）去检测你指定的日志目录或文件，对于探测器找出的每一个日志文件，Filebeat启动收割进程（harvester），每一个收割进程读取一个日志文件的新内容，并发送这些新的日志数据到处理程序（spooler），处理程序会集合这些事件，最后filebeat会发送集合的数据到你指定的地点。

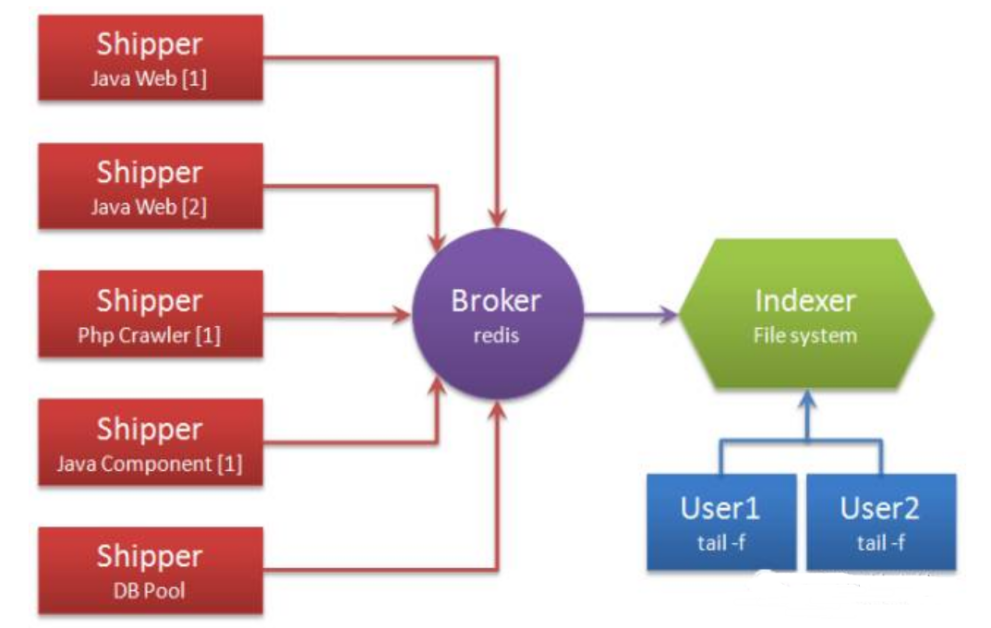
### Logstash

Logstash诞生于2009年8有2日，其作者是世界著名的虚拟主机托管商DreamHost的运维工程师乔丹西塞（Jordan Sissel）。

Logstash的设计非常规范，有三个组件，其分工如下：

1. Shipper负责日志收集。职责是监控本地日志文件的变化，并输出到Redis缓存起来；
2. Broker可以看作是日志集线器，可以连接多个Shipper和多个Indexer；
3. Indexer 负责日志存储。在这个架构中会从Redis接收日志，写入到本地文件。

因为架构比较灵活，如果不想用Logstash的存储，也可以对接到Elasticsearch，这也就是前面所说的ELK的套路了。



### Elasticsearch

### Kibana

## Oozie

一个管理hadoop job 的工作流程调动管理系统，用于协调多个MapReducer任务的执行。

## Atlas

# 数据湖

Hadoop生态圈与数据湖（Data Lake）之间有密切关系。可以说，Hadoop生态系统是最早实践数据湖理念的基础平台之一，很多数据湖的关键能力最初都是在 Hadoop 生态中发展起来的，比如分布式存储、批处理、元数据管理等。

## 定义与目标

数据湖是一种统一存储平台，可以以原始格式存储结构化、半结构化和非结构化数据，并支持各种计算引擎进行访问和分析。它的核心目标是：

高扩展性：支持海量数据（PB 级别）存储

多样性：支持各种数据格式（日志、文本、图片、JSON、CSV 等）

开放性：不依赖某一厂商，支持多种计算框架

低成本：基于廉价硬件和对象存储

分离存储与计算

## Hadoop



## Hadoop生态圈与数据湖的关系

1、HDFS 是数据湖的最早实现基础

数据湖需要一个可扩展、容错、低成本的存储层，HDFS完全符合这一点。

它为数据湖中的原始数据（如 JSON、日志、图片、音频等）提供了统一的存储空间。

2、Hive 为数据湖提供了元数据管理与 SQL 接入

Hive 建立在 HDFS 之上，提供了类 SQL 查询能力，使得存储在数据湖中的数据可以通过结构化查询语言（HiveQL）访问。

Hive 的Metastore是后续如 Presto、Spark、Trino、Iceberg 等数据湖工具的元数据来源。

3、Spark / Flink 提供数据湖的计算能力

Spark 支持数据的批处理和 ETL 转换，广泛用于数据湖上的清洗、聚合和机器学习。

Flink 支持实时流处理，非常适合对实时数据管道的接入与分析，是现代数据湖中非常关键的组件。

4、Kafka + Flume 等提供流式数据接入能力

数据湖需要能够从多源（数据库、应用、消息队列等）接入数据。

Kafka、Flume 是非常成熟的日志和事件流采集工具，为数据湖提供了实时接入的通道。

5、Hive / Ranger / Atlas 提供数据治理能力

数据湖通常用于企业级数据管理，必须支持数据权限、安全审计、数据血缘追踪等。

Hive + Ranger + Atlas 正好构成了数据湖中\*\*数据治理层\*\*的重要基础。

## Hadoop 生态工具在数据湖中的层次结构

数据接入层（Ingestion Layer）

└── Flume / Kafka / Sqoop

存储层（Storage Layer）

└── HDFS / HBase / S3（扩展到云）

元数据层（Metadata Layer）

└── Hive Metastore / Atlas

计算层（Compute Layer）

└── Spark / Flink / Tez / MapReduce

安全治理层（Governance & Security）

└── Ranger / Atlas

查询分析层（Query Layer）

└── Hive / Presto / Trino / Impala

## 对比现代数据湖架构的演化方向



## 总结

Hadoop 生态是数据湖的起点和基础，提供了从数据采集 → 存储 → 元数据管理 → 计算 → 查询 → 安全治理的完整闭环。

关系：

数据湖是一个概念框架，而 Hadoop 提供了其早期的技术实现。

数据湖中仍大量使用 Hadoop 工具（如 HDFS、Hive、Spark、Ranger 等）。

演进方向：

Hadoop 生态的工具正在向“云原生”、“实时化”、“轻量化”演进（如替代 HDFS 的 S3，对 Hive 的增强如 Iceberg）。

Hudi、Iceberg、Delta Lake 等是在 Hadoop 生态基础上发展起来的现代数据湖引擎。