Hadoop Ecosystem

数 据 嗨 客

核心组件

Hadoop HDFS | 一个可复制的，可扩展的Hadoop底层分布式文件系统，同时提供对应用程序数据的高吞吐量访问和提高MapReduce的数据输入性能.

MapReduce | 用于在集群上处理并发，把工作分解成多个任务并同时处理大量数据的分布式编程模型和软件框架.

Hadoop Common | 框架的基础与核心,提供底层操作系统及其文件系统的抽象基本服务和基本过程支持，其他Hadoop模块的常用工具.

Hadoop YARN | 一个用于并行处理大型数据集的基于YARN的系统.

系统部署

Apache Ambari | 用于创建、管理和监控Hadoop集群的工具，可以很方便的安装、调试Hadoop集群.

Cloudera Manager | 基于浏览器的Hadoop管理器，减轻处理和监控大型Hadoop集群的负担，帮助安装和配置Hadoop软件.

Apache ZooKeeper | 维护配置信息，命名，提供分布式同步和提供组件服务，解决分布式应用中经常遇到的一些数据管理问题，包括集群管理、统一命名和配置同步等.

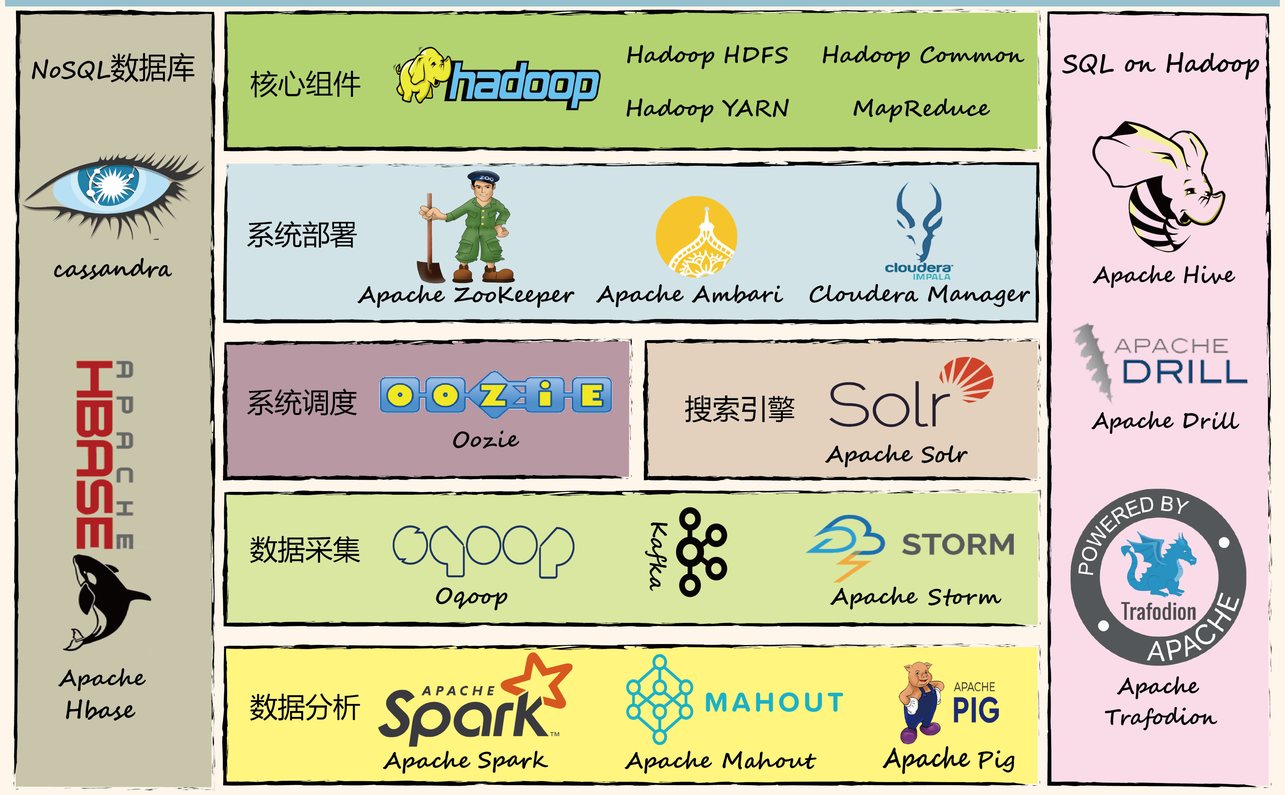
系统调度

Apache Oozie | 在Hadoop生态系统中，Oozie可以把多个MapReduce作业组合到一个逻辑工作单元中，从而完成更大型的任务；Oozie是一种Java Web应用程序，它运行在Java Servlet容器(Tomcat)中，并使用数据库来存储以下内容：

✓ 工作流定义.

✓ 当前运行的工作流实例，包括实例的状态和变量.

Hadoop生态系统



NoSQL 数据库

Apache HBase | 非关系性分布式数据库，且允许使用HDFS进行随机读取和写入.

Cassandra |混合型的非关系的数据库，类似于Google的BigTable，同时具有以下几个特点：

✓ 容错性：数据自动复制到多个节点实现容错.

✓ 高性能：在测试和实际应用方面优于流行的NoSQL数据库.

✓ 分散化：没有单节点失败，没有网络瓶颈，集群中的每个节点都是相同的.

✓ 可扩展：允许以线性方式来高度扩展的巨大NoSQL数据库.

✓ 耐用性：即使整个数据中心出现故障，也不会丢失数据.

SQL On Hadoop

Apache Hive | 定义了一种类似SQL的查询语言HQL，它能够将SQL转化为MapReduce任务在Hadoop上执行。Hive数据仓库软件有助于使用SQL读取，写入和管理驻留在分布式存储中的大型数据集，可以将结构投影到已存储的数据上，提供了一个命令行工具和JDBC驱动程序来将用户连接到Hive.

Apache Drill | Drill是用于大数据探索的SQL查询引擎。在大数据应用中，它能去兼容，并且高性能地去分析结构化数据和变化迅速的数据，同时，还提供业界都熟悉的标准的查询语言(即：ANSI SQL).

Apache Trafodion | 构建在Hadoop/HBase基础之上的关系型数据库，能够完整地支持ANSISQL，并且提供ACID事务保证.

数据采集

Sqoop | 批量数据传输工具，可以将关系数据库的数据转储放置在Hadoop中，也能将MapReduce工作输出的数据移回至关系数据库中.

Kafka | 分布式发布/订阅工具，将系统分离允许多个订阅者发布数据。以容错方式存储记录,在发生记录时处理记录数据流.

Storm | 分布式计算框架，Storm可以轻松地处理无限数据流，实时处理Hadoop为批处理所做的事情.

数据分析

Apache Spark | 不仅有快速的执行能力，丰富的编程API接口，还能把工作分解成多个任务并同时处理，比MapReduce有更多的内置功能(比如SQL)的通用处理框架.

Apache Pig | 用脚本语言来分析、处理大型数据集的平台，通常配合Hadoop使用，同时具有以下优点：

✓ 易于编程.

✓ 可扩展.

✓ 优化强.

Mahout | 使用预先编写的库在MapReduce上运行机器学习算法，可以不用重写机器学习算法就能使用MapReduce的机器学习库.

搜索引擎

Apache Solr | 支持许多世界上互联网站点的搜索和导航功能，具有高可靠性，可扩展性和容错性，可提供分布式索引，复制和负载平衡查询，自动故障转移和恢复，集中式配置等功能.