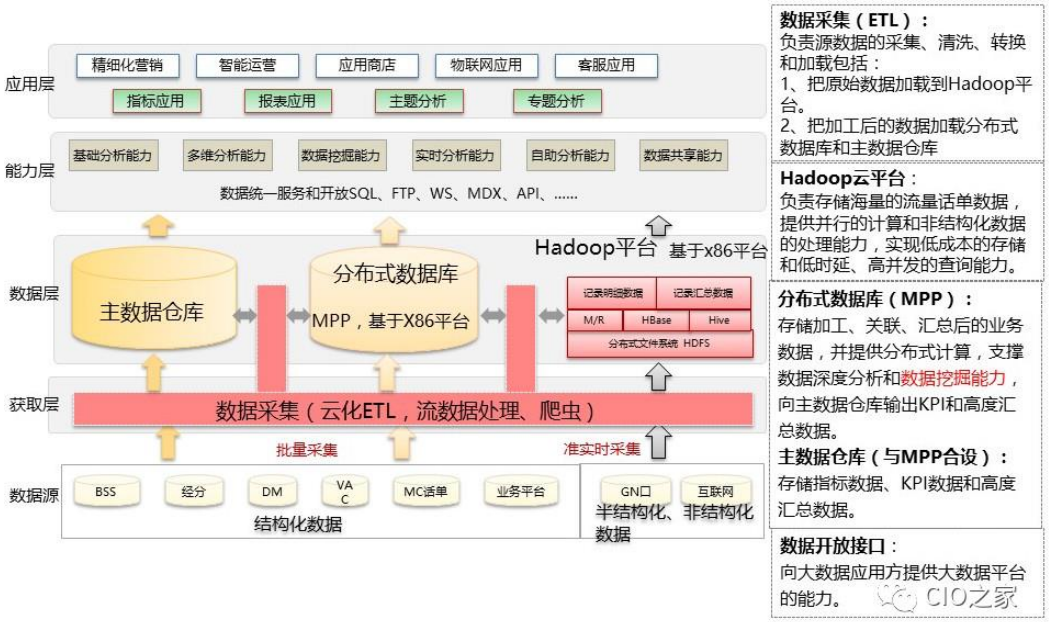
# 如何构建一个企业的大数据分析平台

（来源：CIO 之家公众号，2020-01-09）

面对海量的各种来源的数据，如何对这些零散的数据进行有效的分析，得到有价值的信息一直是大数据领域研究的热点问题。

大数据分析处理平台就是整合当前主流的各种具有不同侧重点的大数据处理分析框架和工具，实现对数据的挖掘和分析，一个大数据分析平台涉及到的组件众多，如何将其有机地结合起来，完成海量数据的挖掘是一项复杂的工作。在搭建大数据分析平台之前，要先明确业务需求场景以及用户的需求，通过大数据分析平台，想要得到哪些有价值的信息，需要接入的数据有哪些，明确基于场景业务需求的大数据平台要具备的基本的功能，来决定平台搭建过程中使用的大数据处理工具和框架。

**大数据平台目标架构**



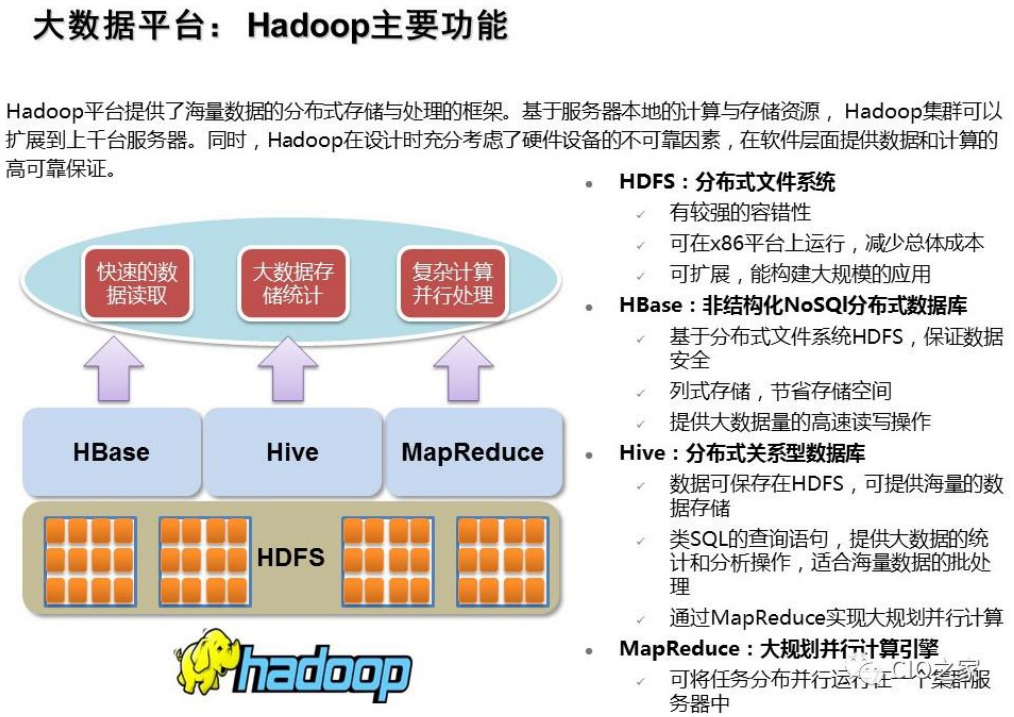

（1）操作系统的选择操作系统一般使用开源版的 RedHat、Centos

或者 Debian 作为底层的构建平台，要根据大数据平台所要搭建的数据分析工具可以支持的系统，正确的选择操作系统的版本。

（2）搭建 Hadoop 集群 Hadoop 作为一个开发和运行处理大规模数据的软件平台，实现了在大量的廉价计算机组成的集群中对海量数据进行分布式计算。Hadoop 框架中最核心的设计是 HDFS 和 MapReduce，HDFS 是一个高度容错性的系统，适合部署在廉价的机器上，能够提供高吞吐量的数据访问，适用于那些有着超大数据集的应用程序；MapReduce 是一套可以从海量的数据中提取数据最后返回结果集的编程模型。在生产实践应用中，Hadoop 非常适合应用于大数据存储和大数据的分析应用，适合服务于几千台到几万台大的服务器的集群运行，支持 PB 级别的存储容量。Hadoop 家族还包含各种开源组件，比如 Yarn，Zookeeper，Hbase，Hive，Sqoop，Impala，Spark 等。使用开源组件的优势显而易见，活跃的社区会不断的迭代更新组件版本，使用的人也会很多，遇到问题会比较容易解决，同时代码开源，高水平的数据开发工程师可结合自身项目的需求对代码进行修改，以更好的为项目提供服务。

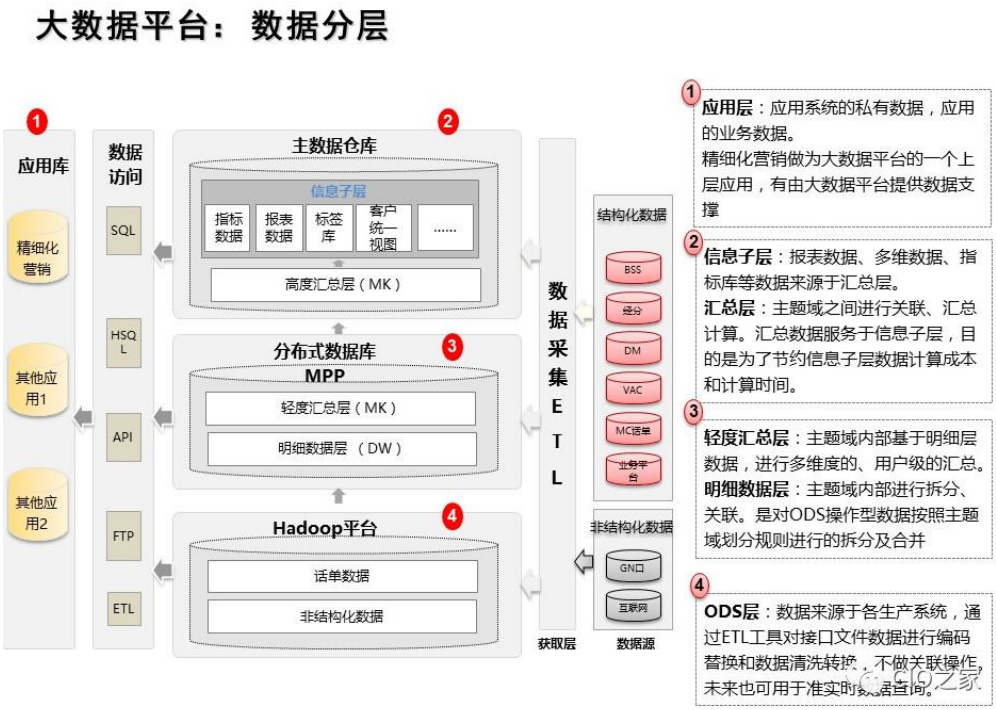
（3）选择数据接入和预处理工具面对各种来源的数据，数据接入就是将这些零散的数据整合在一起，综合起来进行分析。数据接入主要包括文件日志的接入、数据库日志的接入、关系型数据库的接入和应用程序等的接入，数据接入常用的工具有 Flume，Logstash，NDC（网易数据运河系统），sqoop 等。对于实时性要求比较高的业务场景，比如对存在于社交网站、新闻等的数据信息流需要进行快速的处理反馈，那么数据的接入可以使用开源的 Strom，Spark streaming 等。当需要使用上游模块的数据进行计算、统计和分析的时候，就需

要用到分布式的消息系统，比如基于发布/订阅的消息系统 kafka。还可以使用分布式应用程序协调服务 Zookeeper 来提供数据同步服务，更好的保证数据的可靠和一致性。数据预处理是在海量的数据中提取出可用特征，建立宽表，创建数据仓库，会使用到 HiveSQL，SparkSQL 和 Impala 等工具。随着业务量的增多，需要进行训练和清洗的数据也会变得越来越复杂，可以使用 azkaban 或者 oozie 作为工作流调度引擎，用来解决有多个 hadoop 或者 spark 等计算任务之间的依赖关系问题。




（4）数据存储除了 Hadoop 中已广泛应用于数据存储的 HDFS，常用的还有分布式、面向列的开源数据库 Hbase，HBase 是一种 key/value 系统，部署在 HDFS 上，与 Hadoop 一样，HBase 的目标主要是依赖横向扩展，通过不断的增加廉价的商用服务器，增加计算和存储能力。同时 hadoop 的资源管理器 Yarn，可以为上层应用提供统一的资源管理和调度，为集群在利用率、资源统一等方面带来巨大的

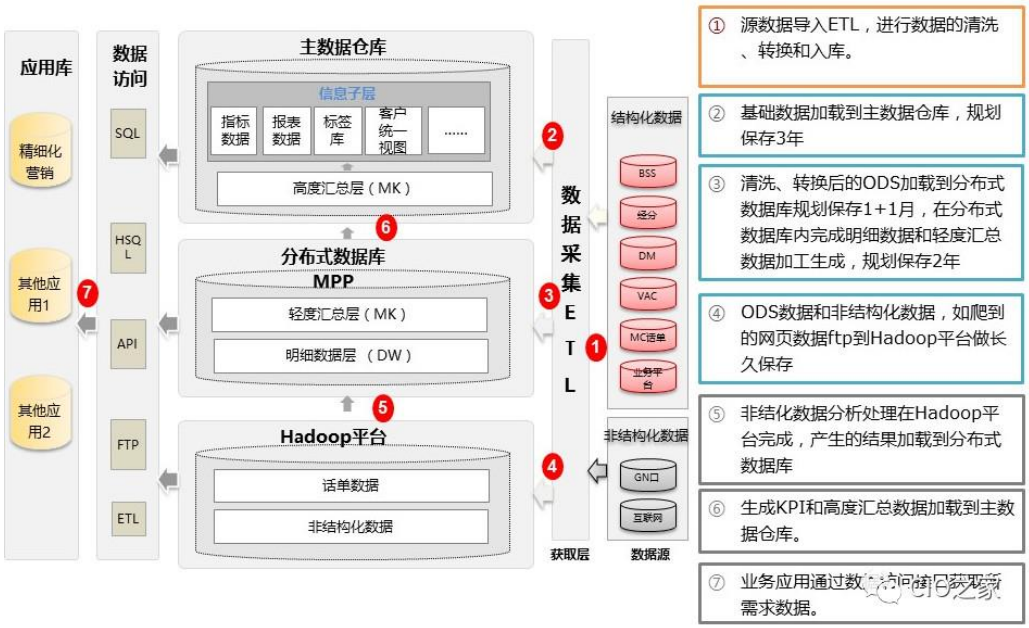
好处。Kudu 是一个围绕 Hadoop 生态圈建立的存储引擎，Kudu 拥有和 Hadoop 生态圈共同的设计理念，可以运行在普通的服务器上，作为一个开源的存储引擎，可以同时提供低延迟的随机读写和高效的数据分析能力。Redis 是一种速度非常快的非关系型数据库，可以将存储在内存中的键值对数据持久化到硬盘中，可以存储键与 5 种不同类型的值之间的映射。



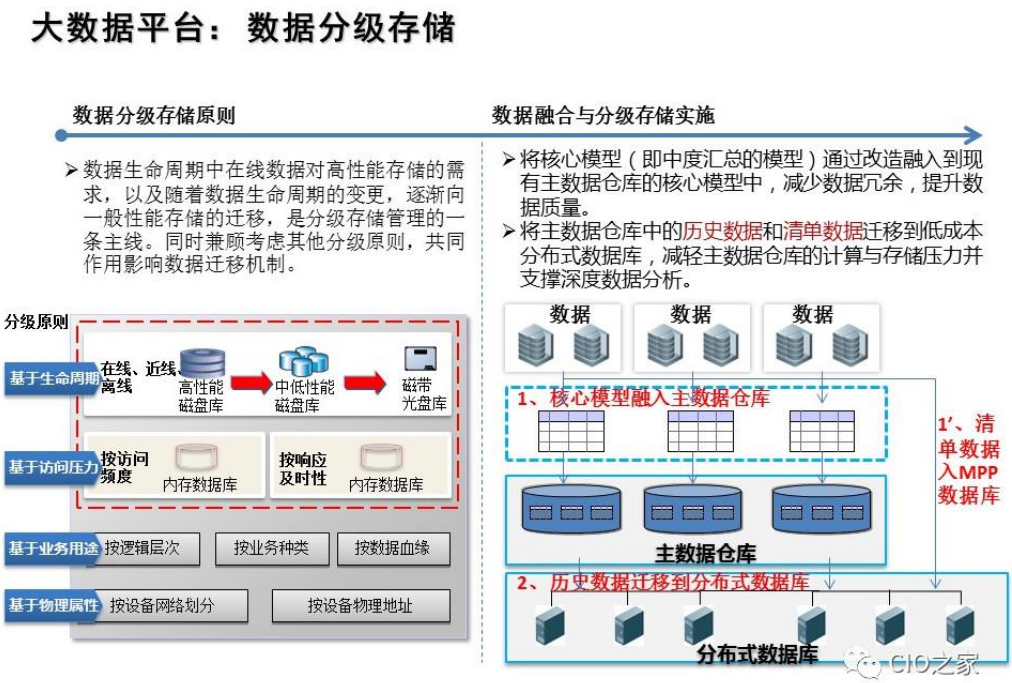

（5）选择数据挖掘工具 Hive 可以将结构化的数据映射为一张数据库表，并提供 HQL 的查询功能，它是建立在 Hadoop 之上的数据仓库基础架构，是为了减少 MapReduce 编写工作的批处理系统，它的出现可以让那些精通 SQL 技能、但是不熟悉 MapReduce、编程能力较弱和不擅长 Java 的用户能够在 HDFS 大规模数据集上很好的利用 SQL 语言查询、汇总、分析数据。Impala 是对 Hive 的一个补充，可以实现高效的 SQL 查询，但是 Impala 将整个查询过程分成了一个执行计划树，而不是一连串的 MapReduce 任务，相比 Hive 有更好的并发性和

避免了不必要的中间 sort 和 shuffle。Spark 可以将 Job 中间输出结果保存在内存中，不需要读取 HDFS，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。Solr 是一个运行在 Servlet 容器的独立的企业级搜索应用的全文搜索服务器,用户可以通过 http 请求,向搜索引擎服务器提交一定格式的 XML,生成索引，或者通过 HTTP GET 操作提出查找请求，并得到 XML 格式的返回结果。还可以对数据进行建模分析，会用到机器学习相关的知识，常用的机器学习算法，比如贝叶斯、逻辑回归、决策树、神经网络、协同过滤等。

**大数据平台：数据处理流程**




**大数据处理的需求和特点**




（6）数据的可视化以及输出 API 对于处理得到的数据可以对接主流的 BI 系统，比如国外的 Tableau、Qlikview、PowrerBI 等，国内的 SmallBI 和新兴的网易有数（可免费试用）等，将结果进行可视化，用于决策分析；或者回流到线上，支持线上业务的发展。成熟的搭建一套大数据分析平台不是一件简单的事情，本身就是一项复杂的工作，在这过程中需要考虑的因素有很多，比如：稳定性，可以通过多台机器做数据和程序运行的备份，但服务器的质量和预算成本相应的会限制平台的稳定性；可扩展性：大数据平台部署在多台机器上，如何在其基础上扩充新的机器是实际应用中经常会遇到的问题；安全性：保障数据安全是大数据平台不可忽视的问题，在海量数据的处理过程中，如何防止数据的丢失和泄漏一直是大数据安全领域的研究热点。

原文链接：https://mp.weixin.qq.com/s/lXcdhvfIR4aePOfU_FNUmw，转载请注明。