29 | 盘点可供中小企业参考的商业大数据平台

2019-01-03 李智慧

从0开始学大数据 进入课程 >

讲述：李智慧
时长 10:42 大小 9.82M

专栏前面我讲了，稍具规模的互联网企业都会搭建自己的大数据平台。但是有同学会问，对于更多的中小企业和初创公司而言，自己搭建大数据平台的成本是不是有点高。确实，拿一个开源的软件搭建自己的大数据平台，对于中小企业来说，无论是人才储备还是服务器成本，似乎都有点难以承受。所幸，还有商业大数据平台可供选择。

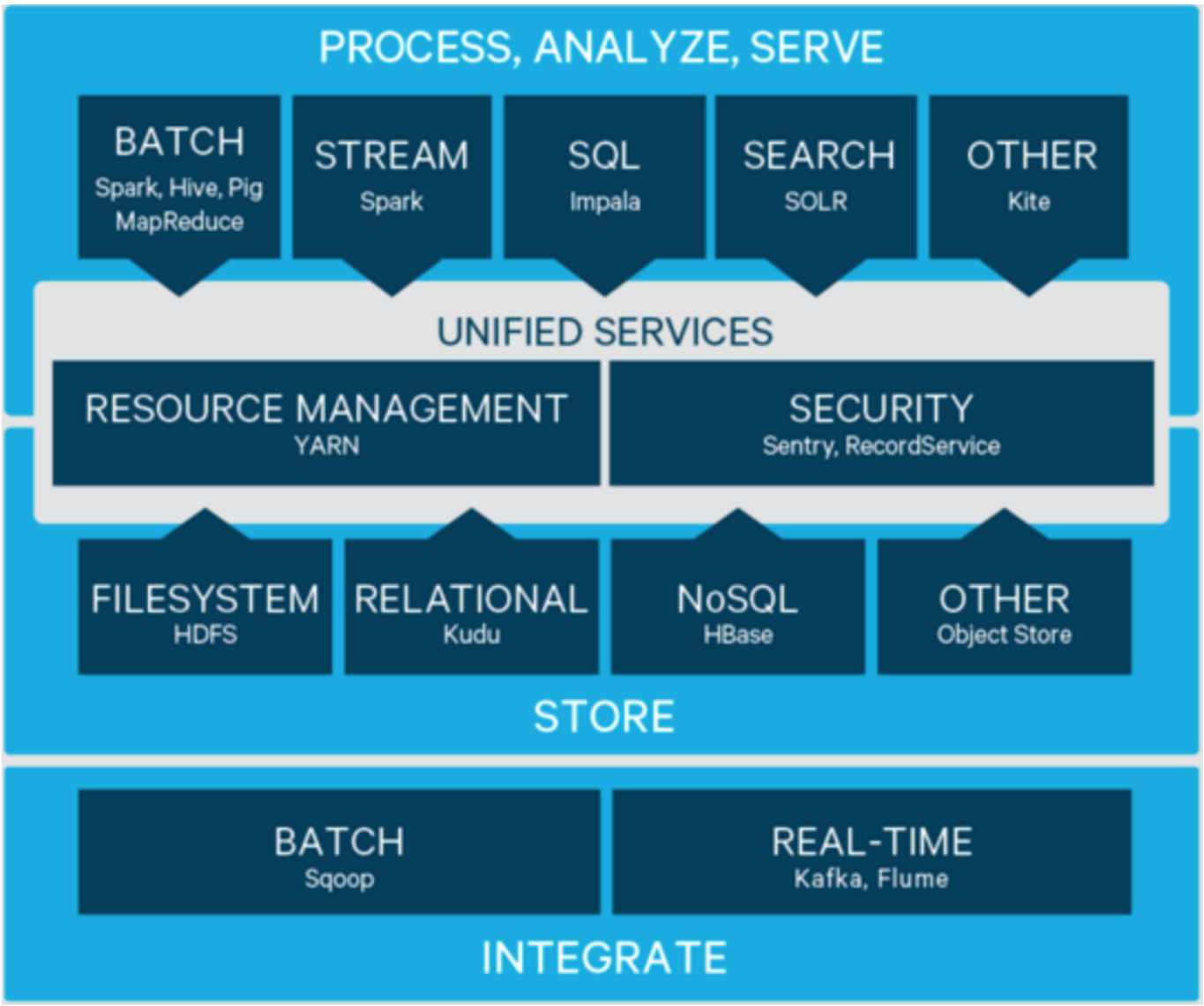
今天我就来和你盘点一下可供中小企业参考的商业大数据平台。

大数据解决方案提供商

Hadoop 作为一个开源产品，关注的是大数据技术实现和产品功能。但是要把 Hadoop 这样的技术产品在企业真正应用起来，还有很多事情要做：企业目前的技术体系如何与

Hadoop 集成起来，具体的解决方案如何实现？如何去做 Hadoop 的部署、优化、维护，遇到技术问题该怎么办？企业需要的功能 Hadoop 不支持怎么办？

Cloudera 是最早开展商业大数据服务的公司，面向企业提供商业解决方案，也就是支持企业解决我上面所说的问题。Cloudera 提供技术咨询服务，为企业向大数据转型提供技术支持。同时 Cloudera 也开发了自己的商业产品，最主要的就是 CDH。



CDH 是一个大数据集成平台，将主流大数据产品都集成到这个平台中，企业可以使用 CDH 一站式部署整个大数据技术栈。从架构分层角度，CDH 可以分为 4 层：系统集成，大数据存储，统一服务，过程、分析与计算。

1. 系统集成：数据库导入导出用 Sqoop，日志导入导出用 Flume，其他实时数据导入导出用 Kafka。

2. 大数据存储：文件系统用 HDFS，结构化数据用 Kudu，NoSQL 存储用 HBase，其他还有对象存储。

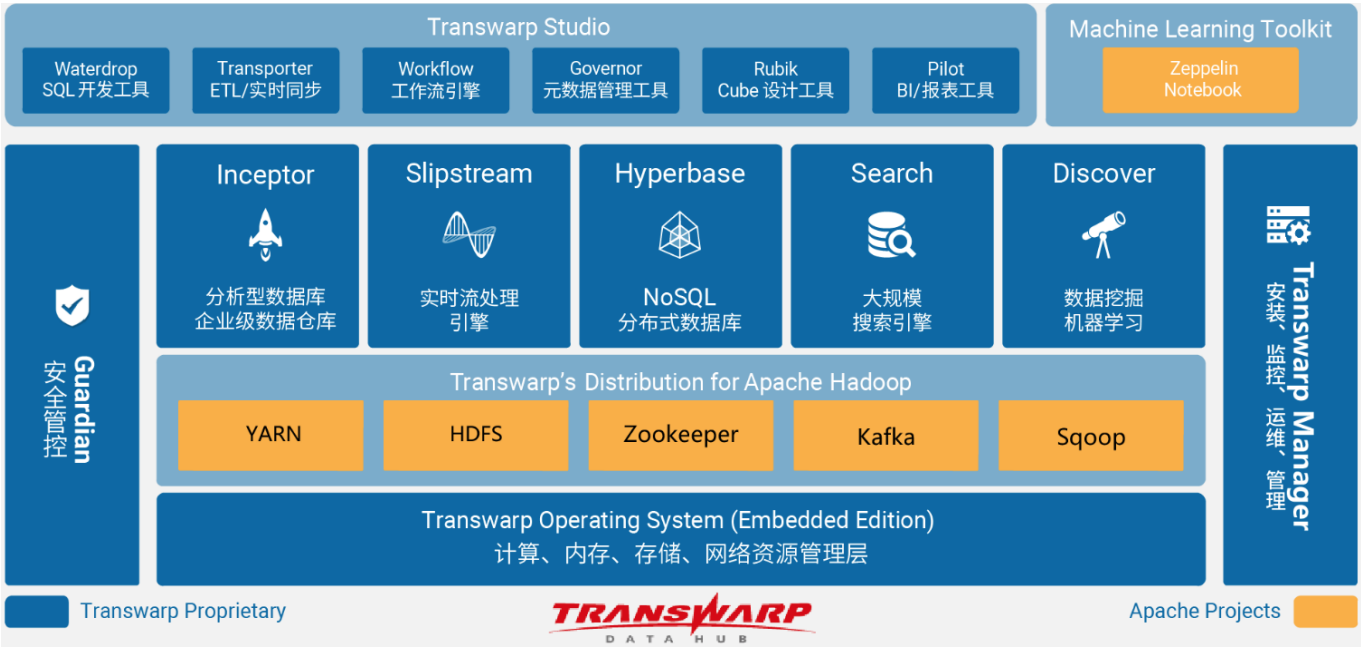
3. 统一服务：资源管理用 Yarn，安全管理用 Sentry 和 RecordService 细粒度地管理不同用户数据的访问权限。

4. 过程、分析与计算：批处理计算用 MapReduce、Spark、Hive、Pig，流计算用 Spark Streaming，快速 SQL 分析用 Impala，搜索服务用 Solr。

值得一提的是，Cloudera 也是 Apache Hadoop 的主要代码贡献者。而开源产品也需要大的商业开发者的支持，如果仅仅就是零零散散的个人开发者，这样的开源产品的发展将很快失控。而商业公司也需要参与开源产品的开发，保证开源产品的发展路径和自己的商业目标保持一致。

除了 Cloudera，还有一家比较大的大数据商业服务公司叫 HortonWorks。近期（2018 年 10 月），Cloudera 和 HortonWorks 宣布合并，这样全球范围内大数据商业服务的格局基本已定。这或许意味着大数据技术领域的创新将进入微创新阶段。

国内本土和 Cloudera 对标的公司是星环科技，商业模式和 Cloudera 一样，主要是为政府和传统企业向大数据转型过程中提供技术支持服务。核心产品是类似 CDH 的 TDH，如下图所示。



面向企业提供解决方案是早期 IT 服务厂商的主要商业模式，通过产品、服务、技术支持等方式向企业收费。IBM、微软、Oracle 都是基于这样的商业模式赚得盆满钵满。早期的 Cloudera 也是基于这样的商业模式，并很快崛起。但是技术时代的变革来的实在是太快了，幸福的日子很快就过去了。

大数据云计算服务商

Oracle、微软这样的传统 IT 企业主要服务对象是企业和政府，营收和利润自然也主要来自企业和政府。所以当互联网开始崛起的时候，虽然以 Google 为代表的互联网公司很快就在技术领域取代了微软们的领先地位，但是大家的商业模式不同，井水不犯河水，倒也相安无事。

后来，Google、亚马逊这样的互联网公司发展出云计算这样的商业模式，企业无需购买、部署自己的服务器，只需要按需购买云服务，就可以使用各种各样的计算资源，比如虚拟主机、缓存、数据库等。相比以往自建数据中心，企业可以以更低的成本、更简单的方式、更灵活的手段使用云计算。随着云计算的快速发展，阿里巴巴等互联网企业也快速跟进，侵蚀以往 IT 巨头的企业领域市场，让 Oracle 这样的 IT 大厂感受到前所未有的压力。

现在所有应用程序都部署在云上，数据也产生在云端，这样自然而然的，大数据也在云上处理处理即可，主流的云计算厂商都提供了大数据云计算服务。

云计算厂商将大数据平台的各项基本功能以云计算服务的方式向用户提供，例如数据导入导出、数据存储与计算、数据流计算、数据展示等，都有相应的云计算服务。我以阿里云为例，一起来看看云计算厂商的主要大数据服务。

1. 数据集成：提供大数据同步服务，通过提供 reader 和 writer 插件，可以将不同数据源（文本、数据库、网络端口）的数据导入、导出。

2.E-MapReduce：集成了 Hadoop、Spark、Hive 等主要大数据产品，用户可以直接将自己的 MapReduce、Spark 程序或者 Hive QL 提交到 E-MapReduce 上执行。

3. 分析性数据库 AnalyticDB：提供快速低延迟的数据分析服务，类似 Cloudera 的 Impala。

4. 实时计算：基于 Flink 构建的流计算系统。

我们看阿里云提供的这些服务，从技术栈角度看，几乎和 Cloudera 的 CDH 一样，这是因为人们的需求就是这样，只是提供的方式不同。Cloudera 通过 CDH 和相关的技术支持，支持企业部署自己的大数据集群和系统。而阿里云则将这些大数据产品都部署好了，使用者只要调用相关 API 就可以使用这些大数据服务。

阿里云将这些大数据基础服务和其他大数据应用服务整合起来，构成一个大数据产品家族，这就是阿里云的数加。数加功能体系如下。

大数据 SaaS 服务商

大数据存储和计算固然有难度和挑战，也因此有了不少解决方案提供商。但是大数据的采集、分析、展现也有一定的门槛和难度，能不能帮企业把这一部分也实现了呢？这样企业无需关注任何技术细节，甚至不需要做任何技术开发，就可以拥有大数据采集、处理、分析、展示一套完整的大数据平台。

如果说云计算厂商把大数据服务当作基础设施（基础设施即服务，IaaS）和平台（平台即服务，PaaS）提供给企业使用，那么还有一些企业，直接把大数据服务当作软件提供给企业（软件即服务，SaaS）。

对于像友盟、神策、百度统计这样的大数据 SaaS 服务商来说，你只需要在系统中调用它提供的数据采集 SDK，甚至不需要调用，只要将它提供的 SDK 打包到自己的程序包中，就可以自动采集各种数据，传输到他们的大数据平台。

然后你登录到他们的大数据平台上，各种数据统计分析报告已经自动生成，甚至和行业同类产品的对比数据也已经生成。此时你只需要查看、分析这些数据就可以了，几乎不需要做任何开发。

当然这类大数据 SaaS 厂商提供的服务比较简单，如果需要精细化、定制化地进一步采集数据、分析数据，还是需要自己调用接口进行开发。

但是，即使是不做进一步的开发，对于很多初创互联网产品而言，百度统计这类大数据服务提供的数据分析也是极有价值的。

大数据开放平台

除了上面提到的这几类商业大数据平台，还有一类大数据商业服务，就是大数据开放平台。

这类平台并不为用户提供典型的数据处理服务，它自身就有大量的数据。比如各类政府和公共事业机构、各类金融和商业机构，它们自己存储着大量的公共数据，比如中国气象局有海量的历史天气数据、中国人民银行有大量的客户征信数据、阿里巴巴有海量的电子商务数据。

如果这些数据是公共所有的，那么使用者就可以直接提交计算请求到这些大数据开放平台上进行计算。如果这些数据涉及保密和隐私，那么如果在不涉及用户隐私的情况下，也可以计算出有意义的结果，比如使用阿里巴巴的数据可以统计出区域经济繁荣指标和排名。

还有一种风控大数据开放平台，结合用户数据和自身数据进行大数据计算。金融借贷机构将借款人信息输入风控大数据平台，大数据平台根据自己的风控模型和历史数据进行风险分析，给出风险指数。金融借贷机构根据这个风险指数决定用户贷款额度和利率等，而风控大数据平台又多获得了一个用户数据，可以进一步完善风控模型和数据库。

小结

大数据已经进入成熟期，大数据技术和应用的各种垂直领域也被逐渐细分，并有越来越多的商业公司进入，继大数据技术生态之后，大数据商业生态也逐渐成型。

对于企业而言，大数据只是实现自己商业目标的工具，如果能借助商业大数据平台，更快实现自己的商业价值，事实上是更划算的事。作为技术人员，能利用自己的大数据知识，做好

商业大数据方案的选型，将商业解决方案更好地应用到自己所在的企业，对自己和公司都是非常有价值的。

思考题

你了解的商业大数据平台和解决方案还有哪些？这些平台和方案的技术特点和商业价值是什么？

欢迎你点击“请朋友读”，把今天的文章分享给好友。也欢迎你写下自己的思考或疑问，与我和其他同学一起讨论。

上一篇　28 | 知名大厂如何搭建大数据平台?

下一篇　30 | 当大数据遇上物联网

精选留言 (14)

写留言

My dream

2019-01-04

我想搭建基于sql查询的自己的大数据平台，要些什么条件啊，我不想用什么阿里腾讯的，他们都收费，而且好贵

观弈道人

2019-01-03

4

行文流畅。现在想转型做大数据技术，似乎没啥意义，了解点普及型大数据知识吧。

Zach_

2019-01-05

2

1. 我们公司就是属于借贷公司，看到后台有神策埋点数据的采取，也知道有神策数据这个公司，但是不确定这个数据是不是上送到神策的大数据平台去了，可以和我们总监确认；

2.看了今天的专栏，有提到风控模型，还需要确认一件事就是：我们的风控模型也是在大数据平台吗？

展开

作者回复: 风控一般有规则引擎和机器学习模型两种，后者通常用大数据平台训练。

修行者

2019-01-03

2

我知道的是华为的 FusionInsight 的大数据解决方案：
前期主要针对的是电信行业，电信运营商提供的大数据解决方案，
现在逐步扩展到平安城市，与政府（公安）合作提供智慧城市相关服务；
也有金融方面的应用。

sunlight00...

2019-01-03

2

如果是企业内网使用大数据的话，还是需要自建的，感觉没有成熟的方案呢

展开

Sam.张朝

2019-01-04

1

一篇文章的内容感觉有点少，特别是最近这几篇，偏重于行业境况。

展开

纯洁的憎恶 1

2019-01-03

很实用，有利于定位和挑选合作伙伴。

展开

作者回复: 👍

smalldemon 1

2019-01-03

公司用的ambari

展开

朝晖 1

2019-01-03

Talend这样通过各种组件图形化操作完成数据的清洗、计算
偏向于有开发经验的人使用
老师 对这样的软件怎么看呢

展开

张国宇

2019-04-03

老师您好，请问大数据开放平台中的风控应用场景中，金融机构客户会上传客户信息到平台，平台侧如何保证数据安全隔离，并且让客户相信这一点。另外平台侧保留金融机构的客户数据是否需要征得金融机构的同意？

展开

小老鼠

2019-01-22

医院有许多传统的手写病案，有没有什么工具可把这些传统数据传成电子数据。

木白
2019-01-09

最近面了一个公司，说是要基于Hadoop做类似阿里云数加平台这样能够开箱即用的大数据产品，主要是面向to B的客户。如您所说，在目前大数据基础设施和开源工具都比较成熟的情况下，您觉得目前做这个东西的前景怎么样？如果只是在既有工具的上层进行一些封装，提供一些工具的话，对技术的提升大吗？

展开

作者回复: 要看公司产品定位和技术深度，总得方向不错的。

Zach_
2019-01-05

嗯，那我就去研究一下我们用的规则是啥。

展开

阿拉丁
2019-01-03

用过wcms，其中保存视频/图片等非结构化数据，大数据平台能支持吗？

展开