

27 | 大数据从哪里来？

2018-12-29 李智慧

从0开始学大数据

[进入课程 >](#)



讲述：李智慧

时长 11:45 大小 10.77M



大数据就是存储、计算、应用大数据的技术，如果没有数据，所谓大数据就是无源之水、无本之木，所有技术和应用也都无从谈起。可以说，数据在大数据的整个生态体系里面拥有核心的、最无可代替的地位。很多从事机器学习和人工智能的高校学者选择加入互联网企业，并不是贪图企业给的高薪，而是因为只有互联网企业才有他们做研究需要用到的大量数据。

技术是通用的，算法是公开的，只有数据需要自己去采集。因此数据采集是大数据平台的核心功能之一，也是大数据的来源。数据可能来自企业内部，也可能是来自企业外部，**大数据平台的数据来源主要有数据库、日志、前端程序埋点、爬虫系统。**

从数据库导入


在大数据技术风靡之前，关系数据库（RDMS）是数据分析与处理的主要工具，我们已经在关系数据库上积累了大量处理数据的技巧、知识与经验。所以当大数据技术出现的时候，人们自然而然就会思考，能不能将关系数据库数据处理的技巧和方法转移到大数据技术上，于是 Hive、Spark SQL、Impala 这样的大数据 SQL 产品就出现了。

虽然 Hive 这样的大数据产品可以提供和关系数据库一样的 SQL 操作，但是互联网应用产生的数据却还是只能记录在类似 MySQL 这样的关系数据库上。这是因为互联网应用需要实时响应用户操作，基本上都是在毫秒级完成用户的数据读写操作，通过前面的学习我们知道，大数据不是为这种毫秒级的访问设计的。

所以要用大数据对关系数据库上的数据进行分析处理，必须要将数据从关系数据库导入到大数据平台上。上一期我提到了，目前比较常用的数据库导入工具有 Sqoop 和 Canal。

Sqoop 是一个数据库批量导入导出工具，可以将关系数据库的数据批量导入到 Hadoop，也可以将 Hadoop 的数据导出到关系数据库。

Sqoop 数据导入命令示例如下。

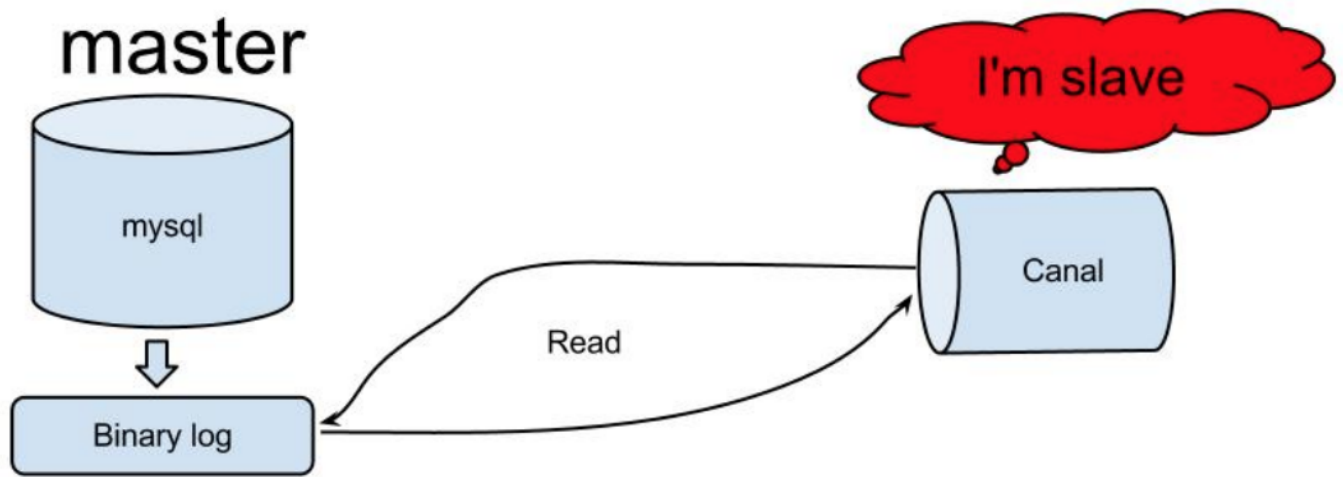
 复制代码

```
1 $ sqoop import --connect jdbc:mysql://localhost/db --username foo --password --table TE'
```

你需要指定数据库 URL、用户名、密码、表名，就可以将数据表的数据导入到 Hadoop。

Sqoop 适合关系数据库数据的批量导入，如果想实时导入关系数据库的数据，可以选择 Canal。

Canal 是阿里巴巴开源的一个 MySQL binlog 获取工具，binlog 是 MySQL 的事务日志，可用于 MySQL 数据库主从复制，Canal 将自己伪装成 MySQL 从库，从 MySQL 获取 binlog。



而我们只要开发一个 Canal 客户端程序就可以解析出来 MySQL 的写操作数据，将这些数据交给大数据流计算处理引擎，就可以实现对 MySQL 数据的实时处理了。

从日志文件导入

日志也是大数据处理与分析的重要数据来源之一，应用程序日志一方面记录了系统运行期的各种程序执行状况，一方面也记录了用户的业务处理轨迹。依据这些日志数据，可以分析程序执行状况，比如应用程序抛出的异常；也可以统计关键业务指标，比如每天的 PV、UV、浏览数 Top N 的商品等。

Flume 是大数据日志收集常用的工具。Flume 最早由 Cloudera 开发，后来捐赠给 Apache 基金会作为开源项目运营。Flume 架构如下。

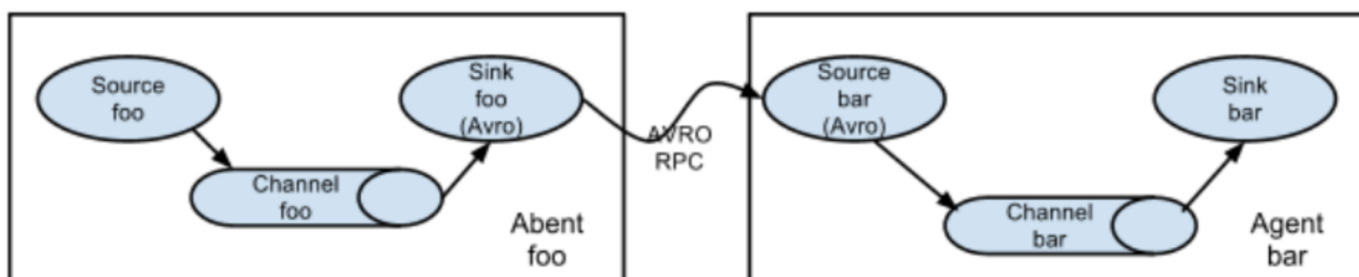


从图上看，Flume 收集日志的核心组件是 Flume Agent，负责将日志从数据源收集起来并保存到大数据存储设备。

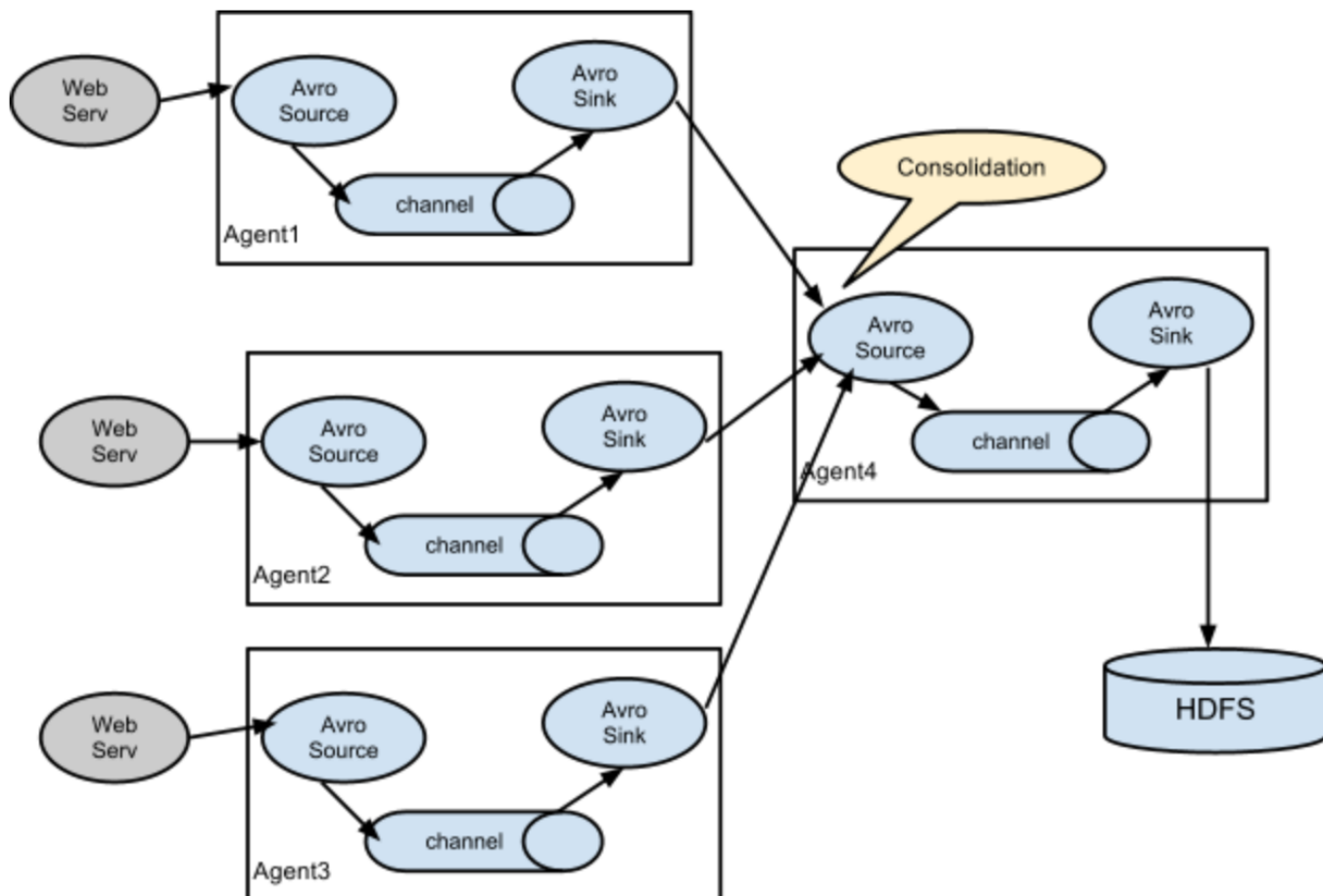
Agent Source 负责收集日志数据，支持从 Kafka、本地日志文件、Socket 通信端口、Unix 标准输出、Thrift 等各种数据源获取日志数据。

Source 收集到数据后，将数据封装成 event 事件，发送给 Channel。Channel 是一个队列，有内存、磁盘、数据库等几种实现方式，主要用来对 event 事件消息排队，然后发送给 Sink。

Sink 收到数据后，将数据输出保存到大数据存储设备，比如 HDFS、HBase 等。Sink 的输出可以作为 Source 的输入，这样 Agent 就可以级联起来，依据具体需求，组成各种处理结构，比如下图的结构。



这是一个日志顺序处理的多级 Agent 结构，也可以将多个 Agent 输出汇聚到一个 Agent，还可以将一个 Agent 输出路由分发到多个 Agent，根据实际需求灵活组合。



前端埋点采集

前端埋点数据采集也是互联网应用大数据的重要来源之一，用户的某些前端行为并不会产生后端请求，比如用户在一个页面的停留时间、用户拖动页面的速度、用户选中一个复选框然后又取消了。这些信息对于大数据处理，对于分析用户行为，进行智能推荐都很有价值。但是这些数据必须通过前端埋点获得，所谓前端埋点，就是应用前端为了进行数据统计和分析而采集数据。

事实上，互联网应用的数据基本都是由用户通过前端操作产生的，有些互联网公司会将前端埋点数据当作最主要的大数据来源，用户所有前端行为，都会埋点采集，再辅助结合其他的数据源，构建自己的大数据仓库，进而进行数据分析和挖掘。

对于一个互联网应用，当我们提到前端的时候，可能指的是一个 App 程序，比如一个 iOS 应用或者 Android 应用，安装在用户的手机或者 pad 上；也可能指的是一个 PC Web 前端，使用 PC 浏览器打开；也可能指一个 H5 前端，由移动设备浏览器打开；还可能指的是一个微信小程序，在微信内打开。这些不同的前端使用不同的开发语言开发，运行在不同的设备上，每一类前端都需要解决自己的埋点问题。

埋点的方式主要有手工埋点和自动化埋点。

手工埋点就是前端开发者手动编程将需要采集的前端数据发送到后端的数据采集系统。通常公司会开发一些前端数据上报的 SDK，前端工程师在需要埋点的地方，调用 SDK，按照接口规范传入相关参数，比如 ID、名称、页面、控件等通用参数，还有业务逻辑数据等，SDK 将这些数据通过 HTTP 的方式发送到后端服务器。

自动化埋点则是通过一个前端程序 SDK，自动收集全部用户操作事件，然后全量上传到后端服务器。自动化埋点有时候也被称作无埋点，意思是无需埋点，实际上是全埋点，即全部用户操作都埋点采集。自动化埋点的好处是开发工作量小，数据规范统一。缺点是采集的数据量大，很多数据采集来也不知道有什么用，白白浪费了计算资源，特别是对于流量敏感的移动端用户而言，因为自动化埋点采集上传花费了大量的流量，可能因此成为卸载应用的理由，这样就得不偿失了。在实践中，有时候只是针对部分用户做自动埋点，抽样一部分数据做统计分析。

介于手工埋点和自动化埋点之间的，还有一种方案是可视化埋点。通过可视化的方式配置哪些前端操作需要埋点，根据配置采集数据。可视化埋点实际上是可以人工干预的自动化埋点。

就我所见，在很多公司前端埋点都是一笔糊涂账。很多公司对于数据的需求没有整体规划和统一管理，数据分析师、商业智能 BI 工程师、产品经理、运营人员、技术人员都会在数据采集这里插一脚，却没有专门的数据产品经理来统一负责数据采集的规划和需求工作。很多需要的数据没有采集，更多没用的数据却被源源不断地被采集存储起来。

不同于业务需求，功能和价值大多数时候都是实实在在的。数据埋点需求的价值很多时候不能直观看到，所以在开发排期上往往被当作低优先级的需求。而很多埋点也确实最后没起到任何作用，加剧了大家这种印象。老板觉得数据重要，却又看不到足够的回报，也渐渐心灰意冷。

所以专业的事情需要专业对待，从安排专业的人专门负责开始。

爬虫系统

通过网络爬虫获取外部数据也是公司大数据的重要来源之一。有些数据分析需要行业数据支撑，有些管理和决策需要竞争对手的数据做对比，这些数据都可以通过爬虫获取。

对于百度这样的公开搜索引擎，如果遇到网页声明是禁止爬虫爬取的，通常就会放弃。但是对于企业大数据平台的爬虫，常常被禁止爬取的数据才是真正需要的数据，比如竞争对手的数据。被禁止爬取的应用通常也会采用一些反爬虫技术，比如检查请求的 HTTP 头信息是不是爬虫，以及对参数进行加密等。遇到这种情况，需要多花一点技术手段才能爬到想要的

小结

各种形式的数据从各种数据源导入到大数据平台，进行数据处理计算后，又将数据导出到数据库，完成数据的价值实现。输入的数据格式繁杂、数据量大、冗余信息多，而输出的数据则结构性更好，用更少的数据包含了更多的信息，这在热力学上，被称作熵减。

熵是表征系统无序状态的一个物理学参量，系统越无序、越混乱，熵越大。我们这个宇宙的熵一刻不停地在增加，当宇宙的熵达到最大值的时候，就是宇宙寂灭之时。虽然宇宙的熵在不停增加，但是在局部，或者某些部分、某些子系统的熵可以减少。

比如地球，似乎反而变得更加有序，熵正在减少，主要原因在于这些熵在减少的系统在吸收外部能量，地球在吸收太阳的能量，实现自己熵的减少。大数据平台想要实现数据的熵的减少，也必须要吸收外部的能量，这个能量来自于工程师和分析师的算法和计算程序。

如果算法和程序设计不合理，那么熵可能就不会下降多少，甚至可能增加。所以大数据技术人员在审视自己工作的时候，可以从熵的视角看看，是不是输出了更有价值、更结构化的数据，是不是用更少量的数据包含了更多的信息。

人作为一个系统，从青壮到垂老，熵也在不停增加。要想减缓熵增的速度，必须从外部吸收能量。物质上，合理饮食，锻炼身体；精神上，不断学习，参与有价值的工作。那些热爱生活、好好学习、积极工作的人是不是看起来更年轻，而整日浑浑噩噩的人则老的更快。

思考题

前面提到，爬虫在采集数据的时候，可能会遇到对方的反爬虫策略，反爬虫策略有哪些？如何应对这些反爬虫策略？

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。

 极客时间

从 0 开始学大数据

智能时代你的大数据第一课

李智慧
同程艺龙交通首席架构师
前 Intel 大数据架构师

新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 26 | 互联网产品 + 大数据产品 = 大数据平台

下一篇 所有的不确定都是机会——智慧写给你的新年寄语

精选留言 (10)

写留言



REAL_MADIR...

2018-12-31

8

利用熵增熵减原理来过好这一生

展开



纯洁的憎恶

2018-12-30

4

Sqoop适合离线批量导入关系数据库的数据，Canle适合实时导入关系数据库的数据。

flume是比较常用的大数据日志收集工具。

前端埋点采集。很多前端操作不会引发后端响应，但对于分析用户行为十分重要。...

展开



😊

2018-12-29

3

反爬虫技术：检查头浏览器信息；检查refer是否正常的流程链上的URL；对IP 或者 imei mac进行实时计算请求量高的；避免csrf攻击的办法也可以在这里调用接口检查ID；针对通过无界面浏览器的爬取行为要进行行为分析 比如简单的操作步骤间隔时间等

应对策略：对于疯狂的爬虫封禁。想对付的竞争对手进行真假数据混合。消磨对手排查...

展开



Creso

2019-02-20

1

- 1.请求头
- 2.ip地址
- 3.验证码
- 4.js加密
- 5.必须登录...

展开



杰之7

2019-01-02

👍 1

通过这一节的学习，用煽减来看待大数据平台。

整个过程通过初始的数据获取，包括从数据库导入数据，有Sqoop,cancel的方式，日志系统导入数据，有Flume将数据库导入到HDFS中，SDK从前端埋点获取数据，及爬虫系统获取数据。...

展开 ▾



balabala

2019-05-01

👍

关于数据从哪里来这个问题，在当前有这么多自动化数据导入、数据处理手段的前提下，数据获取、整理、清洗仍然存在很多不可避免的dirty work，怎么样看待和处理遇到的这种dirty work？

展开 ▾



小老鼠

2019-01-22

👍

大数据获取不断地写磁盘会不会影响系统的性能？

展开 ▾



John

2019-01-17

👍

請問老師 MySQL的binlog用Canel 那麼另一個特別流行的postgresql該用什麼工具呢 謝謝

作者回复: Sqoop是SQL操作，所以是通用的。

◀ ▶



hunterlodg...

2019-01-07

👍

“数据埋点需求的价值很多时候不能直观看到，所以在开发排期上往往被当作低优先级的需求。而很多埋点也确实最后没起到任何作用，加剧了大家这种印象。老板觉得数据重要，却又看不到足够的回报，也渐渐心灰意冷。”

大实话，我们今年的一个大项目也做了很多埋点，目的也是便于分析项目的上线效果，然而采集的大量数据并没有有效利用起来

展开 ∨



萧杰

2018-12-29



反爬虫如果企业单单是在http请求头上监听，可以用scrapy框架有支持很多类库，模拟真实用户浏览器渲染请求，现在我也发现电商网站在开始使用请求参数加密的方式，而作为一个爬虫者，从技术手段的角度怎么应对，请老师答疑解惑。