大数据开发之Flink

技术背景：如今的大数据技术应用场景，对实时性的要求已经越来越高。作为新一代大数据流处理框架，由于非常好的实时性，Flink 独树一帜，在近些年引起了业内极大的兴趣和关注。spark本身是基于批处理模式的原因，并不能完美高效处理原生的数据流，因此对流式处理计算支持较弱。Flink能够提供毫秒级别的延迟，同时保证了数据处理的低延迟、高吞吐和结果的正确性，还提供了丰富的时间类型和窗口计算、Exactly-once 语义支持，另外还可以进行状态管理。Flink 在实时分析领域的优势，使得越来越多的公司开始将实时项目向 Flink 迁移，其社区也在快速发展壮大。目前，Flink 已经成为各大公司实时领域的发力重点，特别是国内以阿里为代表的一众大厂，都在全力投入，不少公司为 Flink 社区贡献了大量源码。如今 Flink 已被很多人认为是大数据实时处理的方向和未来，很多公司也都在招聘和储备了解掌握 Flink 的人才。

流式计算

1、数据的时效性

工作中，我们一般会先把数据存储在表，然后对表的数据进行加工、分析。既然先存储在表中，那就会涉及到时效性概念。

如果我们处理以年，月为单位级别的数据处理，进行统计分析，个性化推荐，那么数据的的最新日期离当前有几个甚至上月都没有问题。但是如果我们处理的是以天为级别，或者一小时甚至更小粒度的数据处理，那么就要求数据的时效性更高了。比如：

对网站的实时监控

对异常日志的监控

这些场景需要工作人员立即响应，这样的场景下，传统的统一收集数据，再存到数据库中，再取出来进行分析就无法满足高时效性的需求了。

2、流式计算和批量计算

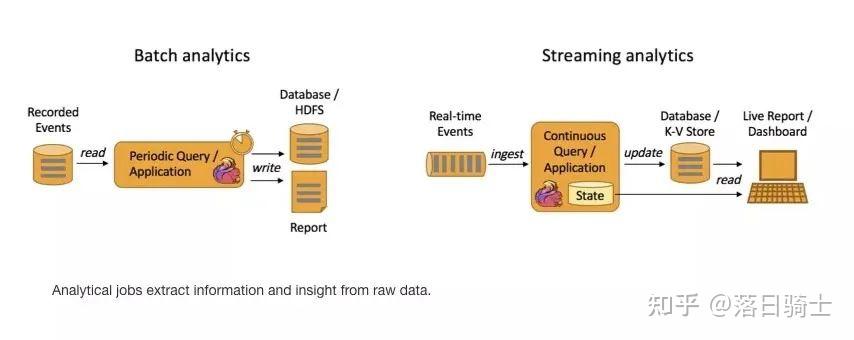
上面说到的：统一收集数据->存储到DB->对数据进行批量处理，就是我们说到的批量计算。而流式计算，顾名思义，就是对数据流进行处理，是实时计算。

主要原理：

（1）与批量计算那样慢慢积累数据不同，流式计算立刻计算，数据持续流动，计算完之后就丢弃。

（2）批量计算是维护一张表，对表进行实施各种计算逻辑。流式计算相反，是必须先定义好计算逻辑，提交到流式计算系统，这个计算作业逻辑在整个运行期间是不可更改的。

（3）计算结果上，批量计算对全部数据进行计算后传输结果，流式计算是每次小批量计算后，结果可以立刻实时化展现。



说明：（1）左边是Batch Analytics，右边是 Streaming Analytics。Batch Analysis 就是传统意义上使用类似于 Map Reduce、Hive、Spark Batch 等，对作业进行分析、处理、生成离线报表。

（2）Streaming Analytics 使用流式分析引擎如 Storm，Flink 实时处理分析数据，应用较多的场景如实时大屏、实时报表。

3、流式计算流程和特性

（1）流程：

提交流计算作业

等待流式数据触发流计算作业

计算结果持续不断对外写出

（2）特性：

实时，低延迟

无界，数据是不断输出无终止的

连续，计算连续进行，计算之后数据就会被丢弃