Flink中的时间

Flink凭借其高效的流处理能力如今备受推荐系统工程师的欢迎。不过在使用Flink的时候，很多同学都会被它的各种时间概念困扰，往往不是很清楚各种时间概念的具体区别，因此经常会无意识的造成很多潜在问题。
这节课中我们会和大家探讨一下Flink中不同的时间概念，希望能帮助大家更好的理解和使用Flink这个强大的工具。

2种时间概念

在Flink中首先最容易造成混淆的就是它的两种不同时间概念（在早期Flink版本种还会有3种）：**处理时间(Processing Time)** 和 **事件时间(Event Time)**。

- 处理时间：指的是在执行某个计算操作时的机器时间，通常为系统当前时间戳。
- 事件时间：指的是事件本身的触发时间，这个时间通常是由数据源采集上报给Flink的。

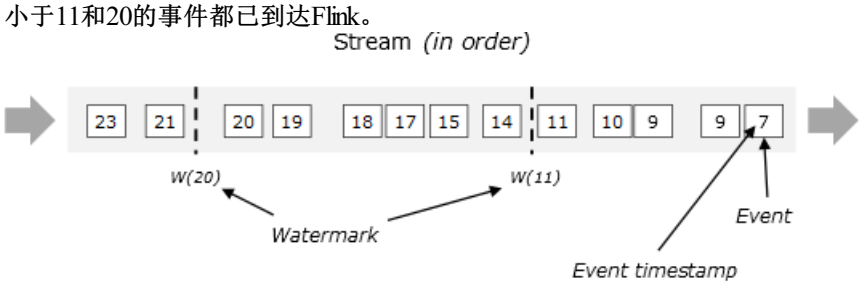
在实际工作中通常事件时间会比处理时间更常用，不过弄清楚这两种时间概念的区别仍然对于我们正确的使用FLink处理数据将有很大的帮助。

水位线

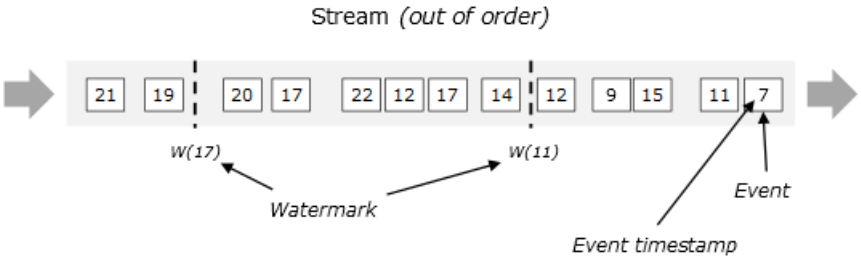
理想情况下，如果每个事件都是按照它的发生时间进入到Flink并且被我们处理的话，那么一切都没有什么问题，我们可以按照时间将它们切分成一个个的小的窗口然后分别处理。但是现实中往往会发生各种意外情况导致事件到达的顺序错乱，或者丢失个别事件。针对这种情况Flink提出了水位线的概念，帮助我们更好的处理乱序的事件。

我们可以简单把水位线理解为另一个时间戳：假设某个事件的水位线为twt_wtw，则Flink会认为当该事件到达的时候，所有事件时间t<=twt <= t_wt<=tw的事件都已到达。

下图是在不存在乱序的情况下事件时间和水位线的关系图。我们可以看到，分别在W(11)和W(20)这两个水位线下，所有时间小于11和20的事件都已到达Flink。



而在有乱序事件发生的时候，水位线就显得至关重要。通过下图我们可以看到，在水位线W(11)和W(17)处，Flink就认为所有时间小于11（或17）的事件都已到达，并会据此进行窗口的切分。



Flink中的时间和水位线问题虽然看起来并不复杂，但是还是需要在实践中不断思考和学习才能更好地掌握其精髓。