Storm和Spark Streaming框架对比

Storm和Spark Streaming两个都是分布式流处理的开源框架。但是这两者之间的区别还是很大的，正如你将要在下文看到的。

处理模型以及延迟

虽然两框架都提供了可扩展性(scalability)和可容错性(fault tolerance)，但是它们的处理模型从根本上说是不一样的。Storm可以实现亚秒级时延的处理，而每次只处理一条event，而Spark Streaming可以在一个短暂的时间窗口里面处理多条(batches)Event。所以说Storm可以实现亚秒级时延的处理，而Spark Streaming则有一定的时延。

容错和数据保证

然而两者的代价都是容错时候的数据保证，Spark Streaming的容错为有状态的计算提供了更好的支持。在Storm中，每条记录在系统的移动过程中都需要被标记跟踪，所以Storm只能保证每条记录最少被处理一次，但是允许从错误状态恢复时被处理多次。这就意味着可变更的状态可能被更新两次从而导致结果不正确。

任一方面，Spark Streaming仅仅需要在批处理级别对记录进行追踪，所以他能保证每个批处理记录仅仅被处理一次，即使是node节点挂掉。虽然说Storm的Trident library可以保证一条记录被处理一次，但是它依赖于事务更新状态，而这个过程是很慢的，并且需要由用户去实现。

实现和编程API

Storm主要是由Clojure语言实现，Spark Streaming是由Scala实现。如果你想看看这两个框架是如何实现的或者你想自定义一些东西你就得记住这一点。Storm是由BackType和Twitter开发，而Spark Streaming是在UC Berkeley开发的。

Storm提供了Java API，同时也支持其他语言的API。 Spark Streaming支持Scala和Java语言(q其实也支持Python)。

批处理框架集成

Spark Streaming的一个很棒的特性就是它是在Spark框架上运行的。这样你就可以想使用其他批处理代码一样来写Spark Streaming程序，或者是在Spark中交互查询。这就减少了单独编写流批量处理程序和历史数据处理程序。

生产支持

Storm已经出现好多年了，而且自从2011年开始就在Twitter内部生产环境中使用，还有其他一些公司。而Spark
Streaming是一个新的项目，并且在2013年仅仅被Sharethrough使用(据作者了解)。

Storm是 Hortonworks Hadoop数据平台中流处理的解决方案，而Spark Streaming出现在 MapR的分布式平台和Cloudera的企业数据平台中。除此之外，Databricks是为Spark提供技术支持的公司，包括了Spark Streaming。

虽然说两者都可以在各自的集群框架中运行，但是Storm可以在Mesos上运行, 而Spark Streaming可以在YARN和Mesos上运行。

本文翻译自：http://xinhstechblog.blogspot.com/2014/06/storm-vs-spark-streaming-side-by-side.html