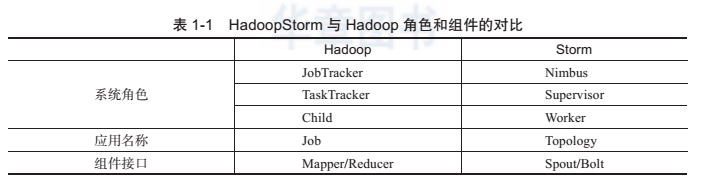
Storm与Hadoop的角色和组件比较

Storm集群和Hadoop集群表面上看很类似。但是Hadoop上运行的是MapReduce作业，而在Storm上运行的是拓扑Topology，这两者之间是非常不同的。一个关键的区别是：一个MapReduce作业最终会结束，而一个Topology拓扑会永远运行（除非手动杀掉）。下表列出了Hadoop与Storm的不同之处。



如果只用一个短语来描述Storm，可能会是这样：分布式实时计算系统。按照Storm作者的说法，Storm对于实时计算的意义类似于Hadoop对于批处理的意义。众所周知，根据Google MapReduce来实现的Hadoop提供了Map和Reduce原语，使批处理程序变得非常简单和优美。那么Storm则是在批处理之前，及时处理了数据。

Storm与其他大数据解决方案的不同之处在处理方式上。Hadoop在本质上是一个批处理系统。数据被引入HDFS并分发到各个节点进行处理。当处理完成时，结果数据返回到HDFS供始发者使用。Storm支持创建拓扑结构来转换没有终点的数据流。不同于Hadoop作业，这些转换从不停止，它们会持续处理到达的数据。

Hadoop专注于批处理。这种模型对许多情形（如为网页建立索引）已经足够，但还存在其他一些使用模型，它们需要来自高度动态来源的实时信息。为了解决该问题，就得借助Twitter推出的Storm。Storm不处理静态数据，但它处理预计会连续的流数据。考虑到Twitter用户每天生成1.4亿条推文，很容易看到此技术的巨大用途。

Storm不只是一个传统的大数据分析系统：它是复杂事件处理（CEP）系统的一个示例。CEP系统通常分为计算和面向检测两类，其中每个系统都可通过用户定义的算法在Storm中实现。例如，CEP可用于识别事件洪流中有意义的事件，然后实时处理这些事件。

Storm作者Nathan Marz提供了在Twitter中使用Storm的大量示例。一个最有趣的示例是生成趋势信息。Twitter从海量的推文中提取所浮现的趋势，并在本地和国家级别维护这些趋势信息。这意味着当一个案例开始浮现时，Twitter的趋势主题算法就会实时识别该主题。这种实时算法是使用Storm实现的基于Twitter数据的一种连续分析。

