storm的功能

Storm有许多应用领域：实时分析、在线机器学习、持续计算、分布式RPC（远程过程调用协议，一种通过网络从远程计算机程序上请求服务）、ETL（Extraction-Transformation-Loading的缩写，即数据抽取、转换和加载）等。

下面介绍Storm的三大主要应用

1. 流处理（stream processing）

Storm可用来实时处理新数据和更新数据库，兼具容错性和可扩展性，即Storm可以用来处理源源不断流进来的消息，处理之后将结果写入某个存储中。

1. 持续计算（continuous computation）

Storm可进行连续查询并把结果即时反馈给客户端，如把Twitter上的热门话题发送到浏览器中。

Storm能保证计算可以永久运行，直到用户结束计算进程为止。

1. 分布式远程程序调用（distributed RPC）

Storm可用来并行处理密集查询。Storm的拓扑结构是一个等待调用信息的分布函数，当它收到一条调用信息后，会对查询进行计算，并返回查询结果。例如，分布式RPC可以做并行搜索或者处理大集合的数据，通过配置DRPC服务器，将Storm的Topology发布为DRPC服务。客户端程序可以调用DRPC服务将数据发送到Storm集群中，并接收处理结果的反馈。这种方式需要DRPC服务器转发，其中DRPC服务器底层通过Thrift实现。适合的业务场景主要是实时计算，且扩展性良好，可以增加每个节点的Worker数量来动态扩展。

Storm绝对是一个相当“有内涵”的系统，能把那么复杂的事情抽象得很完美，能把数学用到极致，这是笔者使用Storm实现消息确保被执行一次之后最直接的感觉。例如互联网方向的实时推荐，海量的用户、数据，且需要实时的推荐算法计算，该如何实现？有了Storm，Spout从数据源取得数据，Bolt可以合并，也可以切分数据，最终完成整个业务逻辑处理并输出结果。分分合合，就像河水从曲曲折折的河道一直流向大海一样，所以也称这种数据处理方式为流式计算。