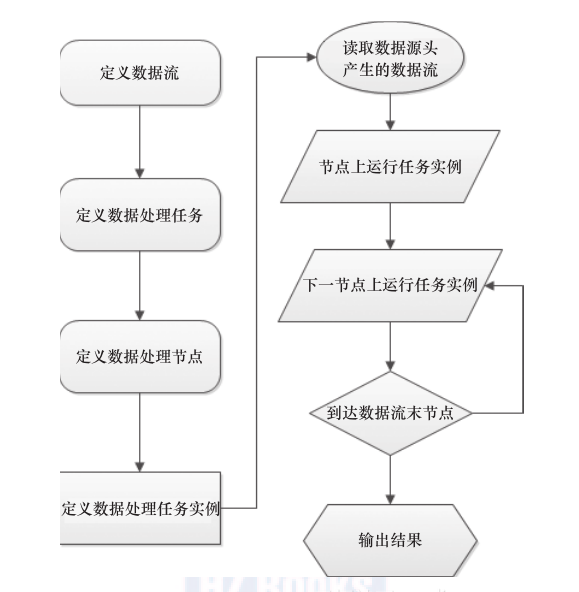
Storm数据流模型

数据流模型是由数据流、数据处理任务、数据节点、数据处理任务实例等构成的一种数据模型。如下图所示。



分布式流处理系统由多个数据处理节点（node）组成，每个数据处理节点上运行有多个数据任务实例，每个数据任务实例属于一个数据任务定义。任务实例是在任务定义的基础上，添加了输入流过滤条件和强制输出周期属性后，可实际推送到数据处理节点上运行的逻辑实体；数据任务定义包含输入数据流、数据处理逻辑和输出数据流属性。

数据流模型简介

首先介绍数据流模型中的一些重要概念。

1. 数据流

数据流是时间分布和数量上无限的一系列数据记录的集合体。数据记录是数据流的最小组成单元，每条数据记录包括三类数据：数据流名称（stream name）、标识数据（key）和具体数据处理逻辑所需的数据（value）。

1. 定义数据处理任务

定义数据处理任务只是定义一个数据处理任务的基本属性，任务还无法直接执行，必须将其实现为具体的任务实例。数据处理任务的基本属性包括输入流、输出流和数据处理逻辑。

1. 输入流（可选）

输入流描述该任务依赖哪些数据流作为输入，是一个数据流名称列表；数据流产生源不会依赖其他数据流，可忽略该配置。

1. 输出流（可选）

输出流描述该任务产生哪个数据流，是一个数据流名称；数据流处理链末级任务不会产生新的数据流，可忽略该配置。

1. 数据处理逻辑

数据处理逻辑描述该任务具体的处理逻辑，如由独立进程执行的外部处理逻辑。

1. 数据处理节点

数据处理节点是可容纳多个数据处理任务实例运行的实体机器，每个数据处理节点的IPv4地址必须保证唯一。

1. 数据处理任务实例

对一个数据处理任务定义进行具体约束后，可将其推送到某个处理节点上运行具体的逻辑实体。数据处理任务基本属性包括数据处理任务定义、输入流过滤条件、强制输出周期。下面进行具体介绍。

1. 数据处理任务定义

数据处理任务定义指向该任务实例对应的数据处理任务定义实体。

1. 输入流过滤条件

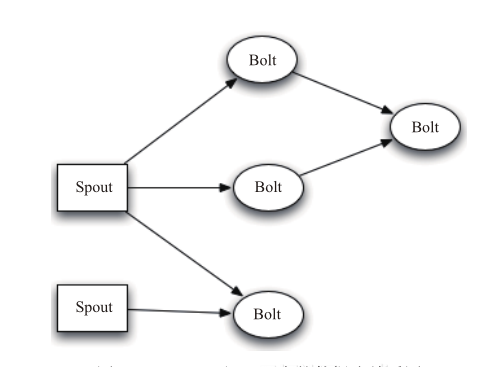
输入流过滤条件是一个布尔类型表达式列表，描述每个输入流中符合什么条件的数据记录可以作为有效数据交给处理逻辑。若某个输入流中所有数据记录都是有效数据，则可直接用true表示。

1. 强制输出周期（可选）

强制输出周期描述以什么频率强制该任务实例产生输出流记录，可以用输入流记录或间隔时间作为周期。如果忽略该配置，则输出流记录产生周期完全由处理逻辑自身决定，不受框架约束。

Storm数据流模型

数据流（Stream）是Storm中对数据进行的抽象，它是时间上无界的Tuple元组序列。在Topology中，Spout是Stream的源头，负责为Topology从特定数据源发射Stream（Spout并不需要接收流，只会发射流）；Bolt可以接收任意多个流作为输入，然后进行数据的加工处理过程，如果需要，Bolt还可以发射出新的流给下级Bolt处理。Topology内部Spout和Bolt之间的数据流关系图如图所示。



Topology中每一个计算组件（Spout和Bolt）都有一个并行执行度（Task），在创建Topology时可以指定，Storm会在集群内分配对应并行度个数的线程来同时执行这一组件。Storm提供了若干种数据流分发（stream grouping）策略来解决在两个组件（Spout和Bolt）之间发送Tuple。在定义Topology时，需要为每个Bolt指定接收什么样的流作为其输入。