Storm中文简介

本节内容：

* Apache Storm是什么
* Apache Storm核心概念
* Storm原理架构
* Storm集群安装部署
* 启动storm ui、Nimbus和Supervisor

**一、Apache Storm是什么**

Apache Storm是自由开源的分布式实时计算系统，擅长处理海量数据，适用于数据实时处理而非批处理。

批处理使用的大多是鼎鼎大名的hadoop或者hive，作为一个批处理系统，hadoop以其吞吐量大、自动容错等优点，在海量数据处理上得到了广泛的使用。但是，hadoop不擅长实时计算，因为它天然就是为批处理而生的，这也是业界一致的共识。否则最近几年也不会有s4,storm,puma这些实时计算系统如雨后春笋般冒出来啦。

举个搜索场景中的例子，当一个卖家发布了一条宝贝信息时，他希望的当然是这个宝贝马上就可以被卖家搜索出来、点击、购买啦，相反，如果这个宝贝要等到第二天或者更久才可以被搜出来，估计就会有不少损失了。 再举一个推荐的例子，如果用户昨天在淘宝上买了一双袜子，今天想买一副泳镜去游泳，但是却发现系统在不遗余力地给他推荐袜子、鞋子，根本对他今天寻找泳镜的行为视而不见，这样商家的利益就有所损失。这是因为后台系统做的是每天一次的全量处理，而且大多是在夜深人静之时做的，那么客户今天白天做的事情要到明天才能反映出来。这也就是为什么需要实时处理的原因。

<http://storm.apache.org/>

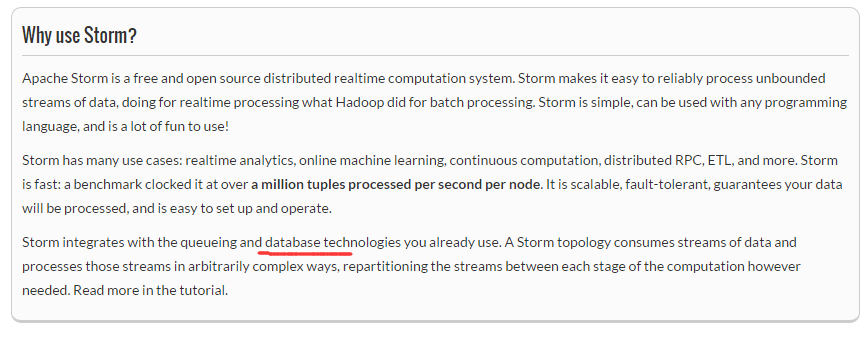


Storm是一个分布式计算系统，但是它跟mr不一样，Storm是实时的，Mr是离线批处理的。

离线mr主要是做数据挖掘、数据分析、数据统计和br分析。

Storm主要是在线的业务系统。数据像水一样，源源不断的来，然后在流动的过程中，就要把数据处理完。比如，将一些解析业务系统里采集的日志信息、报文解析成某一种格式，像解析过来的xml格式然后落到一个SQL或NoSQL数据库中。

当然，hadoop也可以做，但是它是离线批量的。



Storm是不作任何存储的，数据有地方来，结果有地方去。一般是结合消息队列或数据库来用的，消息队列是数据源，数据库是数据目的地。

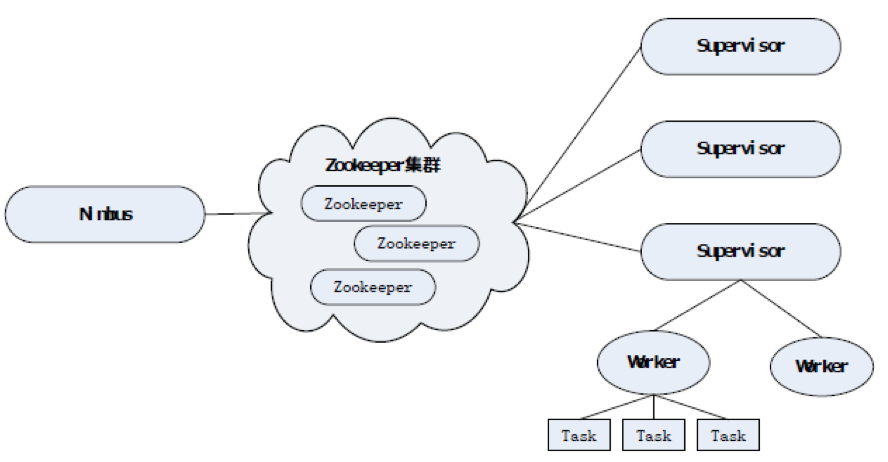


**二、Apache Storm核心概念**

* Nimbus：Storm集群主节点，负责资源分配和任务调度。我们提交任务和截止任务都是在Nimbus上操作的。一个Storm集群只有一个Nimbus节点。
* Supervisor：Storm集群工作节点，接受Nimbus分配任务，管理所有Worker。
* Worker：工作进程，每个工作进程中都有多个Task。
* Task：任务，每个Spout和Bolt都是一个任务，每个任务都是一个线程。
* Topology：计算拓扑，包含了应用程序的逻辑。
* Stream：消息流，关键抽象，是没有边界的Tuple序列。
* Spout：消息流的源头，Topology的消息生产者。
* Bolt：消息处理单元，可以过滤、聚合、查询数据库。
* Stream grouping：消息分发策略，一共6种，定义每个Bolt接受何种输入。
* Reliability：可靠性，Storm保证每个Tuple都会被处理。

**三、Storm原理架构**

**1. Storm集群架构图**



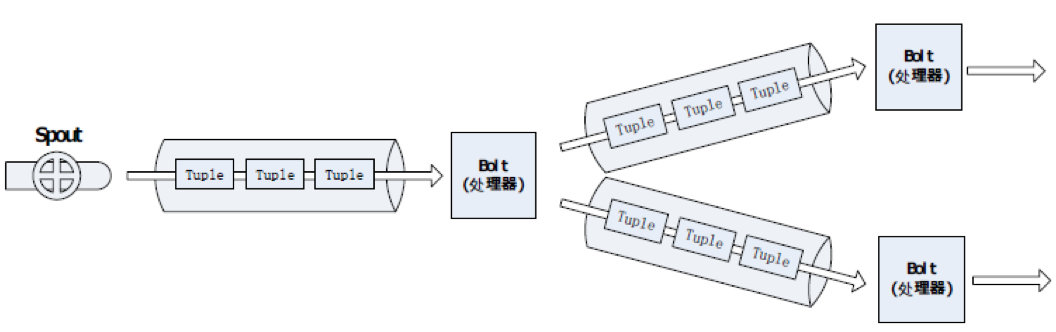
Zookeeper集群在Storm集群中的作用：

Zookeeper集群负责Nimbus节点和Supervior节点之间的通信，监控各个节点之间的状态。比如通常我们提交任务的时候是在Nimbus节点上执行的，Nimbus节点通过zk集群将任务分发下去，而Supervisor是真正执行任务的地方。Nimbus节点通过zk集群监控各个Supervisor节点的状态，当某个Supervisor节点出现故障的时候，Nimbus节点就会通过zk集群将那个Supervisor节点上的任务重新分发，在其他Supervisor节点上执行。这就意味着Storm集群也是高可用集群，如果Nimbus节点出现故障的时候，整个任务并不会停止，但是任务的管理会出现影响，通常这种情况下我们只需要将Nimbus节点恢复就可以了。Nimbus节点不支持高可用，这也是Storm目前面临的问题之一。不过一般情况下，Nimbus节点的压力不大，通常不会出现问题。

一般情况下，Zookeeper集群的压力并不大，一般只需要部署3台就够了。Zookeeper集群在Storm集群中逻辑上是独立的，但在实际部署的时候，一般会将zk节点部署在Nimbus节点或Supervisor节点上。

**2. 数据处理流程图**

storm处理数据的特点：数据源源不断，不断处理。



**3. 拓扑图分析**

storm中是没有数据存储结构的，我们需要自己设计数据落地接口，指明数据存储到哪一部分中。Storm本身是不存储数据的。

