5-4-cn

1 本节我们讨论流数据处理模型
2 数据处理系统提供大数据计算处理能力和应用开发平台。
从计算架构的角度，将数据处理系统分为算法层、计算模型层、计算平台和引擎层等。
与大数据相关的计算算法包括机器学习算法和数据挖掘算法。
计算模型是指不同类型的大数据在不同场景下的处理方式，
包括批处理、流计算、结构化数据的大规模并发处理(MPP)模型、内存计算模型和数据流图模型。
就计算平台和引擎而言，通常具有代表性的大数据处理平台有 Hadoop、Spark、storm、Pregel 等，
本节我们以 Storm 平台为例讨论下流数据处理模型

3 流计算是一种处理实时动态数据的计算模型。传统企业数据库存放的是历史数据也即静态数据，
即数据在进行计算处理前必须全部进入数据库，技术人员可以对数据库进行查询、更新等操作，
并利用数据挖掘和 OLAP 等分析工具从静态数据中找到有价值的信息支持企业决策分析。
在互联网应用中（用户网页点击追踪、在线实时推荐系统等）、智能交通系统、无线传感器网络监控等领域，其数据产生方式与数据特征有以下特点和计算要求
数据不再是分批次到达而是动态连续不断到达
计算分析要求实时性、快速响应、低延迟
数据量大，但不看重数据的存储，而强调数据的即时处理分析
看重数据整体的计算分析结果，而不关注个体数据
数据元素到达的顺序和时序无法预测或控制，计算程序要能够做出应对
MapReduce 针对已经进入数据库的静态数据进行离线批量计算出来，其计算结果也存入静态数据库；而流计算则是针对动态连续性数据流进行实时分析计算，获得计算结果后，数据要么导入静态数据库，要么丢弃，即一次性使用。
要支持这种数据流的计算模式，流计算框架一般包括三个步骤，数据实时采集，数据实时计算数据实时查询服务

4 分布式系统中常用有向非循环图（DAG, Directed Acyclic Graph）来表征计算流程或计算模型。
如下图就表示了分布式系统中的链式任务组合，图中的不同颜色节点表示不同阶段的计算任务（或计算对象），
而单向箭头则表示了计算步骤的顺序和前后依赖关系。

5 Storm 是一种 Native Stream Processing System，即对流数据的处理是基于每条数据进行，其并行计算是基于由 Spout（数据源）和 Bolt（处理单元）组成的有向拓扑图 Topology 来实现。
Topology：定义了并行计算的逻辑模型（或者称抽象模型），也即从功能和架构的角度设计了计算的步骤和流程。

6 Storm 的计算体系也采用了主从（Master/Slave）架构，主要有两类节点：主节点 Master 和工作节点 Slave
主节点上运行一个 Nimbus 守护进程，类似于 Hadoop 的 JobTracker，负责集群的任务分发和故障监测。Nimbus 通过一组 Zookeeper 管理众多工作节点；每个工作节点运行一个 Supervisor 守护进程，监听本地节点状态，根据 Nimbus 的指令在必要时启动和关闭本节点的工作进程。

7 流计算有两种方式，以 Storm 为代表的原生流处理系统和以 Spark 为代表的微批处理系统，
微批处理是为了对数据采取行动(处理)而以小组(“批次”)收集数据的实践。与原生流处理系统相比，微批处理不是真正的实时处理，而是接近实时处理，因为它可以等到 Micro-batch processing 组成后再进行处理，

因此可以节省计算成本。
与此相比，传统的“批处理”通常意味着对一大组数据采取行动。Micro-batch processing 处理是传统批处理的变体，因为数据处理发生得更频繁，所以处理的新数据组更小。
在 Micro-batch processing 处理和传统批处理中，在任何处理发生之前，数据都是根据预先确定的阈值或频率收集的。

8 流计算应用
在实际应用中，有时将批处理和流计算相结合，完成如图所示的历史记录分析和实时分析。
Kafka 代理从数据生产者获取数据，并将其交付到流计算 Storm 集群和批计算 Hadoop 集群中，Storm 的处理输出可存储在 NoSQL Cassandra DB 中，批处理的输出可以存储在 Hbase 中。
在对数据进行处理后，可以根据批处理和流处理的结果进行虚拟化、决策、预测、OLAP 和推荐等数据分析。

9 本节我们简单学习了流处理技术，并以 Storm 为例。今天内容就到这里，谢谢大家