什么是实时流计算

所谓实时流计算，就是近几年由于数据得到广泛应用之后，在数据持久性建模不满足现状的情况下，急需数据流的瞬时建模或者计算处理。在这种实时计算的应用实例有金融服务、网络监控、电信数据管理、Web应用、生产制造、传感检测，等等。在这种数据流模型中，单独的数据单元可能是相关的元组（Tuple），如网络测量、呼叫记录、网页访问等产生的数据。但是，这些数据以大量、快速、时变（可能是不可预知）的数据流持续到达，由此产生了一些基础性的新的研究问题 —— 实时计算。实时计算的一个重要方向就是实时流计算。

1. 实时流计算背景

数据的价值随着时间的流逝而降低，所以事件出现后必须尽快对它们进行处理，最好事件出现时便立刻对其进行处理，发生一个事件进行一次处理，而不是缓存起来成一批处理。例如商用搜索引擎，像Google、Bing和Yahoo!等，通常在用户查询响应中提供结构化的Web结果，同时也插入基于流量的点击付费模式的文本广告。为了在页面上的最佳位置展现最相关的广告，通过一些算法来动态估算给定上下文中一个广告被点击的可能性。上下文可能包括用户偏好、地理位置、历史查询、历史点击等信息。一个主搜索引擎可能每秒钟处理成千上万次查询，每个页面都可能会包含多个广告。为了几十处理用户反馈，需要一个低延迟、可扩展、高可靠的处理引擎。

对于这些实时性要求很高的应用，若把持续到达的数据简单地放到传统数据库管理系统DBMS中，并在其中进行操作，是不切实际的。传统的DBMS并不是为快速连续地存放单个的数据单元而设计的，而且也不支持“持续处理”，而“持续处理”是数据流应用的典型特征。另外，现在人们都认识到，“近似性”和“自适应性”是对数据流进行快速查询和其它处理（如数据分析和数据采集）的关键要素，而传统DBMS的主要目标恰恰与之相反：通过稳定的查询设计，得到精确的答案。

另外一些方案是采用MapReduce来处理实时数据流。但是，尽管MapReduce做了实时性改进，也很难稳定地满足应用需求。这是因为Hadoop MapReduce框架为批处理做了高度优化，典型的是通过调度批量任务来操作静态数据，任务不是常驻服务，数据也不是实时流入；而数据流计算的典型范式之一是不确定数据速率的事件流流入系统，系统处理能力必须与事件流量匹配。

1. 实时计算应用场景
2. 实时计算处理流程
3. 实时计算框架