

开篇词 | 攻克实时流计算难点，掌握大数据未来！

你好，我是周爽，相熟的人都叫我“爽哥”。

我曾任职于华为 2012 实验室高斯部门，负责实时分析型内存数据库 RTANA、华为公有云 RDS 服务的研发工作。目前，我专注于移动反欺诈解决方案的研发。针对公司业务需求，我开发了一个实时流计算系统，并在此基础上完成了风控系统的研发。最终，这个系统被一个独角兽收购。

最近这两年，越来越多的业务和数据分析对实时性提出更高的要求，与之对应解决实时计算问题的流计算框架，也开始流行起来。

因为工作原因，常有人问我有关实时流计算系统的问题。整体观察下来我发现：很多时候，他们并非不知道这些框架，也并非不熟悉这些框架的 API 和工作原理，而是不清楚如何将框架，运用到实时业务中去，也就不能很好地解决落地问题。

业务功能要求实时，我该怎么落地？

在此之前，我想先说一点：如果你的业务相对简单，通过查数据库的方式，就能够做到毫秒级返回，那也没有必要去研究更复杂的技术。正所谓，“如无必要，勿增实体”，保持一切简单就好。

但是当请求非常多、数据量非常大，并且对请求时延要求非常严格时，比如，必须在毫秒甚至微秒级返回，那么问题就变得复杂了。比如这些场景：

实时检测异常的反欺诈或风控系统；
实时展示业务报表的大屏系统；
实时计算用户兴趣偏好的推荐系统；
实时统计过车流量的智能交通系统。



@拉勾教育

面对以上业务场景，如果按照传统数据库增删查改的方法，需要将数据全部记录在数据库中，然后在查询时，再即时遍历和计算。很明显，这种方案不管是存储空间，还是计算时间，成本都非常高，已经不能有效地进行实时计算了。

因此，原本习惯了做增删查改业务逻辑开发的人员，在初次接触实时流计算业务场景时，不可避免地会遇到种种难题，比如以下几点。

1. 需要统计的时间窗口很长，数据量也很大。比如“相同设备在 3 个月内注册事件的次数”，此时，如果你想实时计算获得结果，就不能够通过遍历数据库的方式来实现了。
2. 需要统计的变量，其值域非常大。比如“同一用户在 6 个月内使用不同 IP 个数”，如果是数亿用户和数亿 IP，你还能够用集合来记录这些不同的值吗？更何况，还需要在指定的时间范围内进行计算。
3. 一次完整的业务，可能需要计算数十个甚至数百个特征。比如，实时风控系统中，风控模型的输入便是如此。为了保证用户体验，风控系统必须在数秒甚至数百毫秒内返回。
4. 有些问题的算法，天然就很复杂，数据量又很大，如何做到实时计算呢？比如，社交网络的二度关联分析，还有许多复杂的统计学习和机器学习模型。
5. 甚至有些时候，产品和开发人员都不清楚，是否需要或者能够，使用实时流计算技术。或许难以置信，但这样的公司和开发人员，真的不在少数。

如果想切实解决这些难题，就需要透过现象看本质。我认为，之所以会出现上面的种种难题，主要是因为以下**五种原因**：

- 一是，缺乏对实时流计算技术以及它的适用场景的整体认识；
- 二是，不知道如何用“流”来实现各种业务逻辑的异步和高并发计算；
- 三是，不知道如何针对“流”这种独特的数据模式，设计实时算法；
- 四是，对各种流计算框架的认识只停留在 API 调用层面，而没有理解其背后的设计原理，也就是“流”这种计算模式的，核心概念和关键技术点；
- 五是，缺少对一些已有案例的借鉴和思考。

如何解决实时流计算问题？

既然明确了问题，接下来我们应该怎样克服呢？我认为可以从**系统架构**和**实时算法**两个方面来突破。

系统架构

从架构师的角度看，要为产品设计一个好的实现方案，既要有足够的技术储备，也要充分理解具体的业务问题。通过分析各类实时业务场景，我们可以发现，大多数方案都是基于“流计算”技术的。

“流计算”本质上是一种“异步”编程方法。业务数据像“流水”一样，通过“管道”，也就是“队列”，持续不断地流到各个环节的 subsystem 中，然后由各个环节的 subsystem 独立处理。所以，为了更快地处理“流”，可以通过**增加管道的数量，来提高流计算系统的并行处理能力**。

目前，开源的流计算框架虽然有许多（比如 Storm、Spark Streaming、Samza 和 Flink），但其实这些主流框架背后，都有着一套类似的设计思路和架构模式。它们都涉及流数据状态、流信息状态、反向压力、消息可靠性等概念。**先行理解这套设计思路和架构模式，可以帮助你快速掌握，所有主流流计算框架的工作原理。**

实时算法

系统架构提供了整体的计算框架，但要实现具体的业务功能，还需要针对“流数据”设计合适的算法。毕竟，与传统“块数据”相比，“流数据”需要连续不断并且实时地进行处理。

对于实时流计算中的算法，最最核心的问题，在于解决“大数据量”和“实时计算”之间的矛盾。数据量一大，几乎所有事情都会变得复杂和缓慢。“大数据量”的问题，集中在四个方面：时间窗口很长、业务请求量很大、内存受限、数据跨网络访问。

为了实现“实时计算”的效果，需要你针对算法做非常精心的设计。所幸的是，**这些算法的设计和实现也是有规律可循的**。你只需要掌握几种特定类型的算法，比如计数、求和、均值、方差、直方图、分位数、HyperLogLog 等。而对于更加复杂的算法，如果不能直接进行实时计算，那我们可以通过 Lambda 架构来解决！

课程设计思路

本课程就是从“**系统架构**”和“**实时算法**”这两个方面，来带你理解实时流计算系统。为此，我为你设计了以下学习路径。（注意，模块三为“**实时算法**”部分，其余模块为“**系统架构**”相关。）

模块一，实时流计算入门。我将介绍流计算系统的整体架构和使用场景，以及入门流计算前，需掌握的编程基础，比如 NIO 和异步编程，以及异步系统中的 OOM 和反向压力问题。

借此，你会对实时流计算系统有个整体的认识，并对“流”的本质有个初步理解。

模块二，自己动手做一个流计算框架。我将介绍如何从 JDK 里最基础的工具类，一步步开发出一个分布式流计算框架。

通过这种自己动手的方式，希望帮助你理解流计算系统的核心概念及实现原理。

模块三，核心技术篇。我将详细讲解流计算能够解决哪些类型的问题，包括流数据操作、时间维度聚合计算、关联图谱分析、事件序列分析、模型学习和预测等。此外，还将讨论流计算过程中非常重要的状态管理问题，并带你思考如何最终将前面的流计算框架扩展为分布式系统。

借此，你会掌握实时流计算中涉及的各种算法，这些算法会有助于你解决各种实时业务场景中的问题。

模块四，开源流计算框架原理解析及实战。我将详细对比和分析，各种开源流计算框架的具体实现，来巩固你对流计算核心概念和技术的理解，并带你正确理解这些框架的 API 设计，以便你在各种业务场景下，能够灵活地使用它们，最终实现各种复杂的业务逻辑。

此外，我还会通过两个案例，也就是实时风控和实时数据同步，来带你理解如何将开源流计算框架，运用到具体的业务场景中。

《21 讲吃透实时流计算》大纲

开篇词 | 攻克实时流计算难点，掌握大数据未来！

模块一 实时流计算入门

1 实时流计算的通用架构

2 异步和高并发：为什么 NIO 是异步和高并发编程的基础？

3 反向压力：如何避免异步系统中的 OOM 问题？

4 流与异步：为什么说掌握流计算先要理解异步编程？

模块二 自己动手做一个流计算框架

5 有向无环图（DAG）：如何描述、分解流计算过程？

6 CompletableFuture: 如何理解 Java 8 新引入的异步编程类?

7 死锁: 为什么流计算应用突然卡住, 不处理数据了?

8 性能调优: 如何优化流计算应用?

模块三 流计算到底在计算什么

9 流数据操作: 最基本的流计算功能

10 时间维度聚合计算: 如何在长时间窗口上实时计算聚合值?

11 关联图谱分析: 如何用 Lambda 架构实现实时的社交网络分析?

12 事件序列分析: 大家都在说的 CEP 是怎么回事?

13 模型学习和预测: 如何检查流数据异常?

14 状态管理: 为什么说流计算是有“状态”的?

15 扩展为集群: 如何实现分布式状态存储?

模块四 开源流计算框架原理解析及实战

16 Apache Storm: 最早的开源流计算框架

17 Apache Spark: 批流处理亦在这处理

17 Spark Streaming：从批处理走向流处理

18 Apache Samza：最简洁的开源流计算框架

19 Apache Flink：最惊艳的开源流计算框架

20 场景案例：如何用 Flink 实现实时风控引擎？

21 场景案例：如何用 Flink SQL CDC 实现实时数据同步？

彩蛋

彩蛋 1 | 竟然还有分布式的 JVM？

彩蛋 2 | 穷途末路的选择：Lambda 架构

结束语 | Java 程序员的成长之路和从业方向

@拉勾教育

讲师寄语

本课程对实时流计算技术的关键点，做了提纲挈领的分析和讲解，期望你能够从点到面而知全局，迅速领悟大多数流计算框架的本质，在方案选型和软件开发时，做到胸有成竹。

在流计算技术尚未在国内兴起之前，我就根据公司业务需要，从头开始设计并实现了自己的流计算框架。这是我的实战经验总结，它经得起事实验证。

未来，实时流计算技术必然会成为大数据的主流模式，数据不仅以“流”的方式被处理，还以“流”的方式被存储。希望这个课，给你切实的帮助。

拉勾教育 · 互联网人实战大学

大数据高薪训练营

PB 级企业大数据项目实战 + 拉勾硬核内推

5 个月全面掌握大数据核心技能

[> 点击图片，立即查看 <](#)

@拉勾教育

PB 级企业大数据项目实战 + 拉勾硬核内推，5 个月全面掌握大数据核心技能。点击链接，全面赋能！