

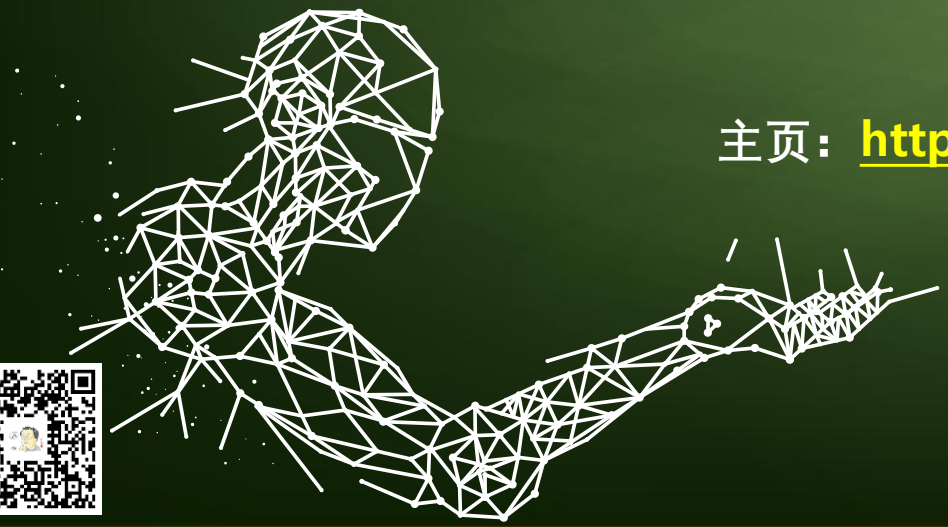
08

8.1 Structured Streaming 概述

授课教师：[李阳源]

ywlee1212@gmail.com

主页：https://sites.google.com/view/dalguri



SL-02:提纲

8.1 概述
8.2 编写Structured Streaming程序的基本步骤
8.3 输入源
8.4 输出操作
8.5 容错处理
8.6 迟到数据处理
8.7 查询的管理和监控

8.1 概述

8.1.1 基本概念

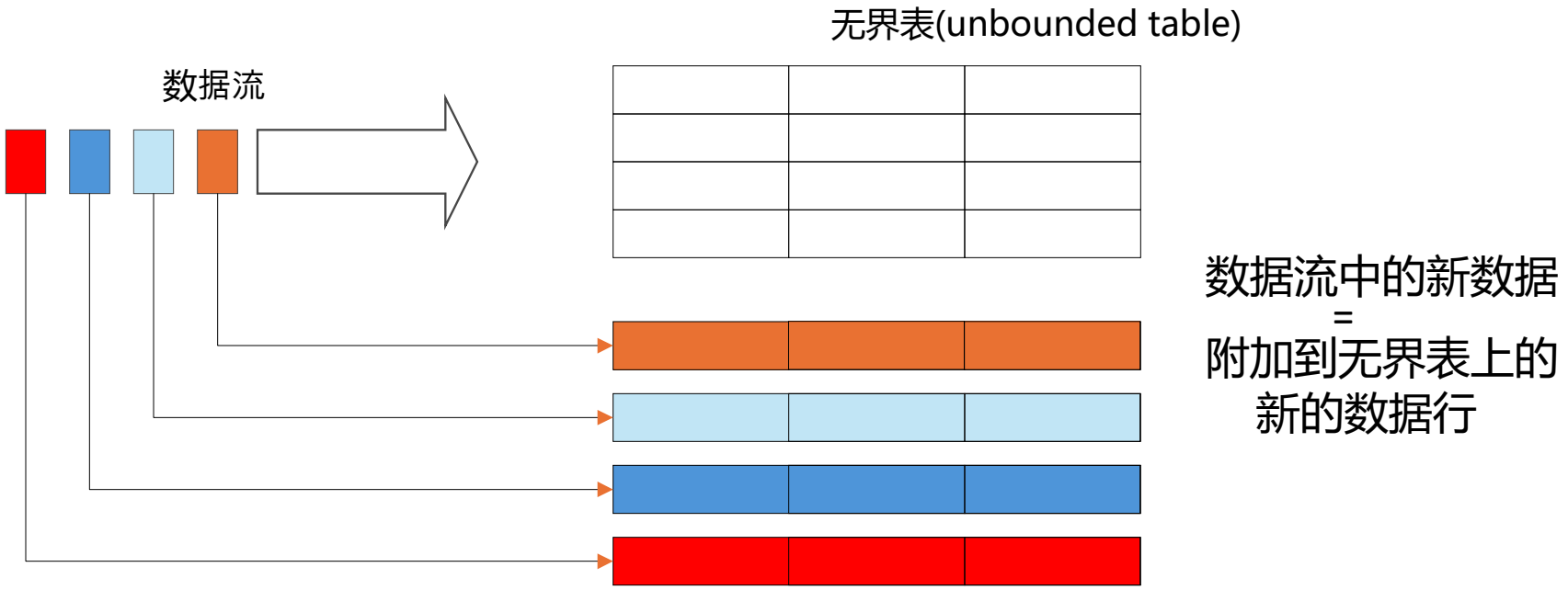
8.1.2 两种处理模型

8.1.3 Structured Streaming和Spark SQL、Spark Streaming关系

SL-04 8.1.1 基本概念



- Structured Streaming的关键思想是将实时数据流视为一张正在不断添加数据的表
- 可以把流计算等同于在一个静态表上的批处理查询，Spark会在不断添加数据的无界输入表上运行计算，并进行增量查询



SL-05 8.1.1 基本概念



- 在无界表上对输入的查询将生成结果表，
  系统每隔一定的周期会触发对无界表的计算并更新结果表

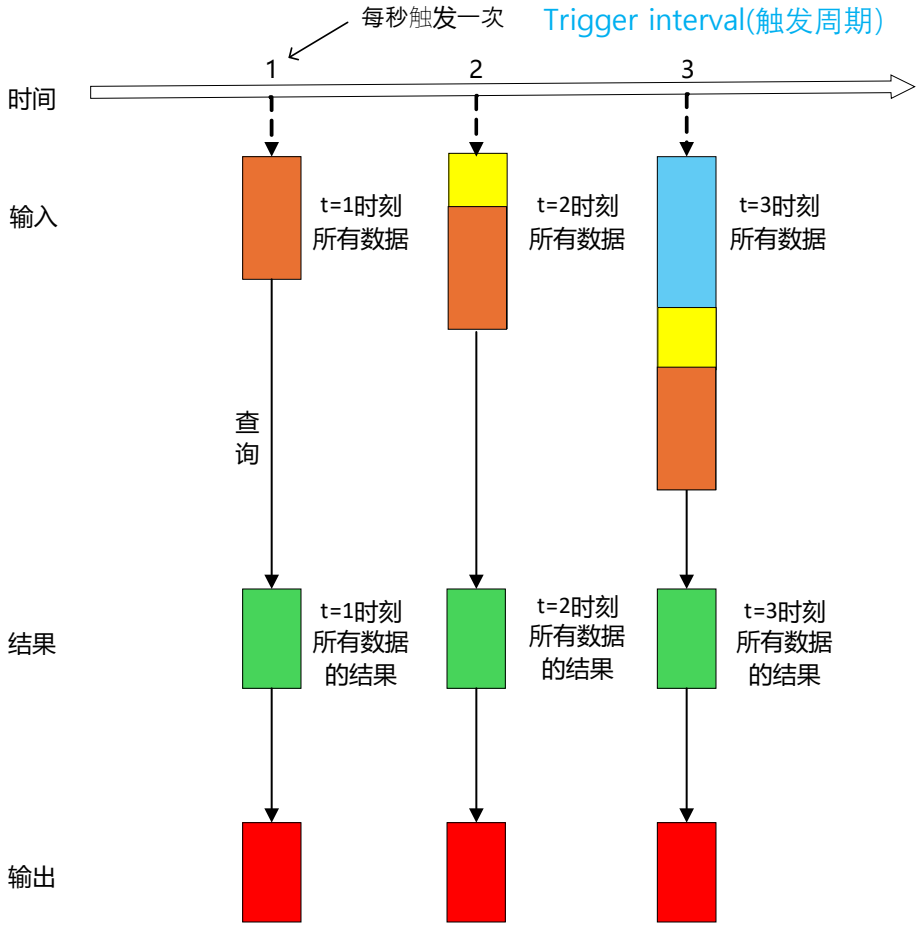


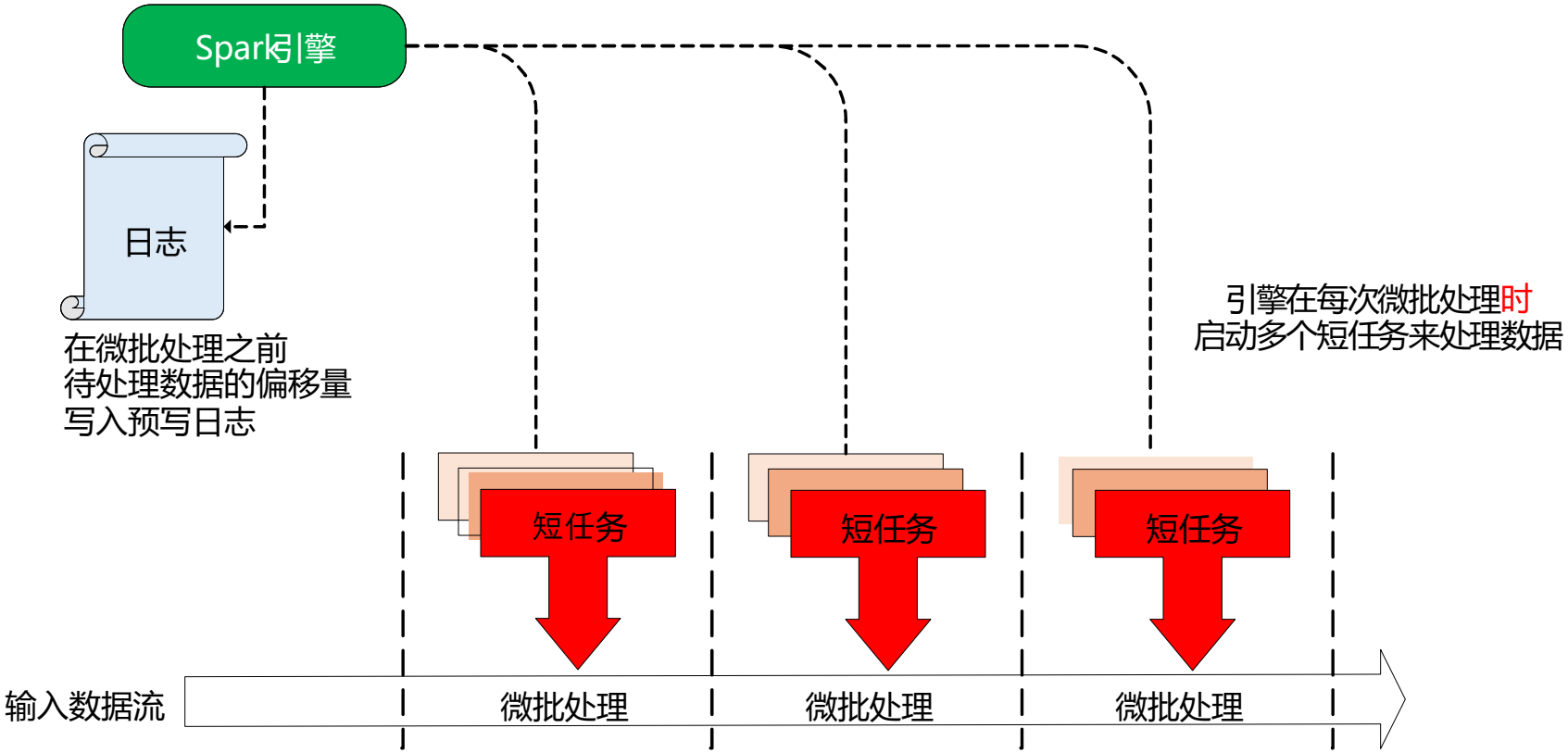
图 Structured Streaming编程模型

SL-06 8.1.2 两种处理模型



（1）微批处理(Micro-batch Processing)

- Structured Streaming默认使用微批处理执行模型，这意味着Spark流计算引擎会定期检查流数据源，并对自上一批次结束后到达的新数据执行批量查询
- 数据到达和得到处理并输出结果之间的延时超过100毫秒

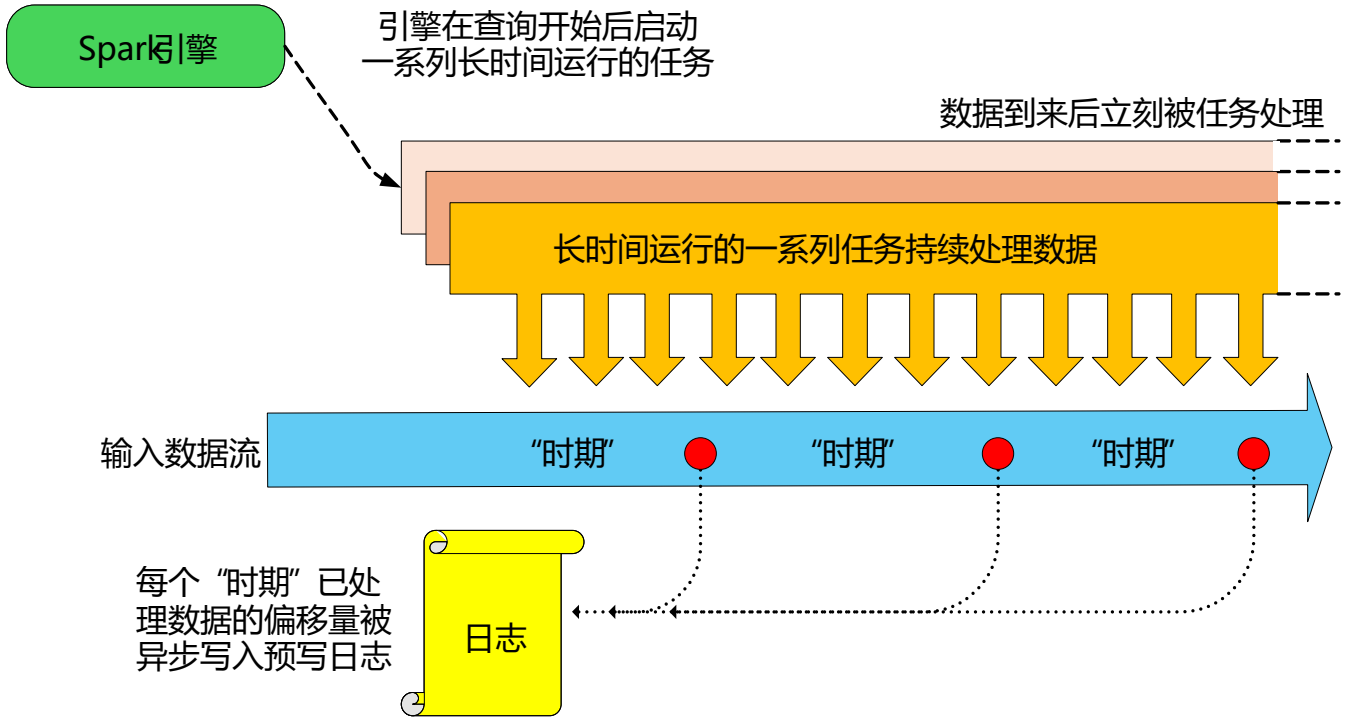


SL-07 8.1.2 两种处理模型



（2）持续处理

- Spark从2.3.0版本开始引入了持续处理的试验性功能，可以实现流计算的毫秒级延迟
- 在持续处理模式下，Spark不再根据触发器来周期性启动任务，而是启动一系列的连续读取、处理和写入结果的长时间运行的任务



SL-08 8.1.3 Structured Streaming和Spark SQL、Spark Streaming关系



Structured Streaming

Structured Streaming处理的数据跟Spark Streaming一样，也是源源不断的数据流，区别在于，Spark Streaming采用的数据抽象是DStream（本质上就是一系列RDD），而Structured Streaming采用的数据抽象是DataFrame。

Structured Streaming

Structured Streaming可以使用Spark SQL的DataFrame/Dataset来处理数据流。虽然Spark SQL也是采用DataFrame作为数据抽象，但是，Spark SQL只能处理静态的数据，而Structured Streaming可以处理结构化的数据流。这样，Structured Streaming就将Spark SQL和Spark Streaming二者的特性结合了起来。

SL-08 8.1.3 Structured Streaming和Spark SQL、Spark Streaming关系

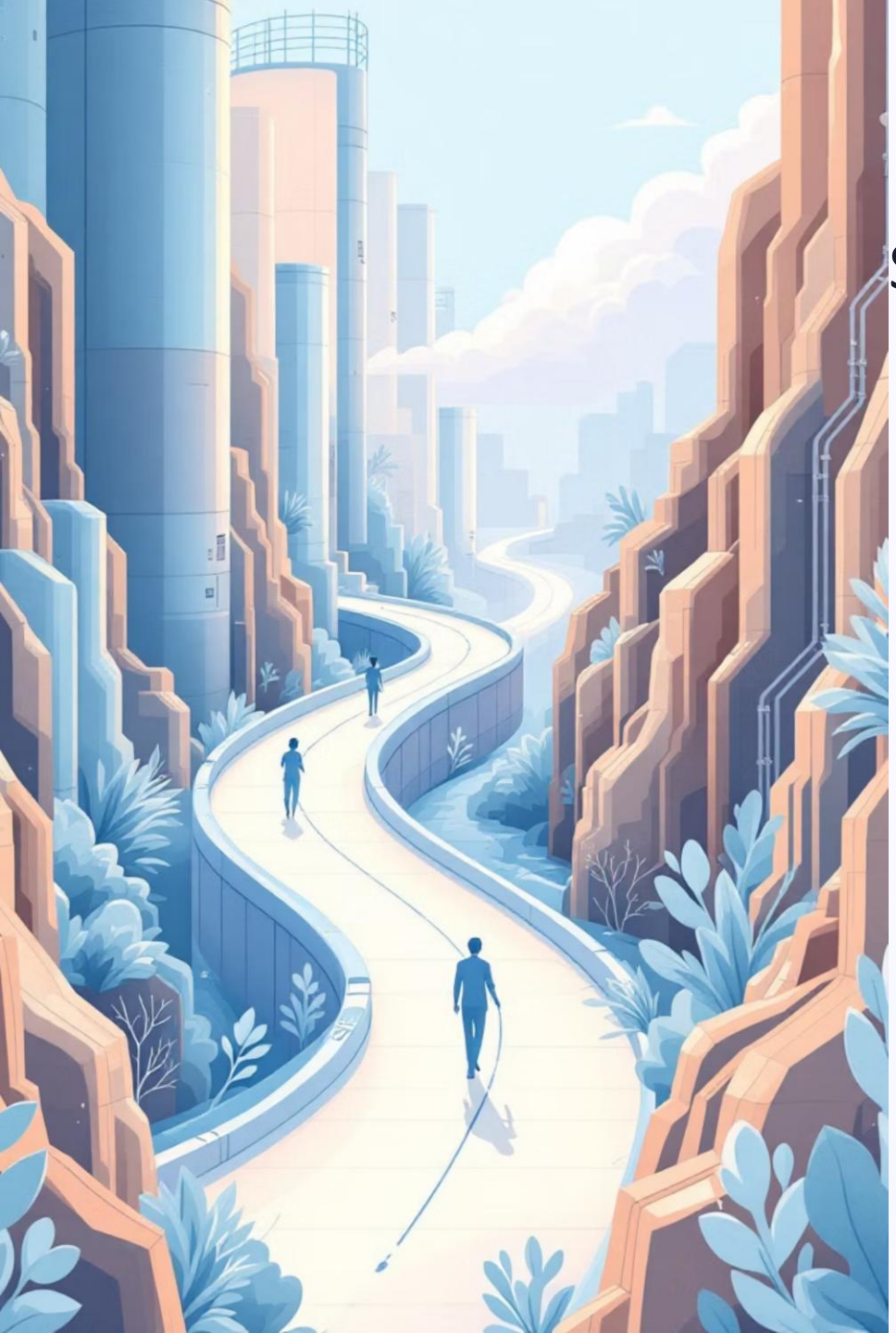


Structured Streaming

Structured Streaming可以对DataFrame/Dataset应用前面章节提到的各种操作，包括select、where、groupBy、map、filter、flatMap等。

Spark Streaming

Spark Streaming只能实现秒级的实时响应，而Structured Streaming由于采用了全新的设计方式，采用微批处理模型时可以实现100毫秒级别的实时响应，采用持续处理模型时可以支持毫秒级的实时响应。



Spark Streaming与Structured Streaming简介

Spark Streaming

基于DStream抽象，采用微批处理架构，实现秒级实时响应能力

Structured Streaming

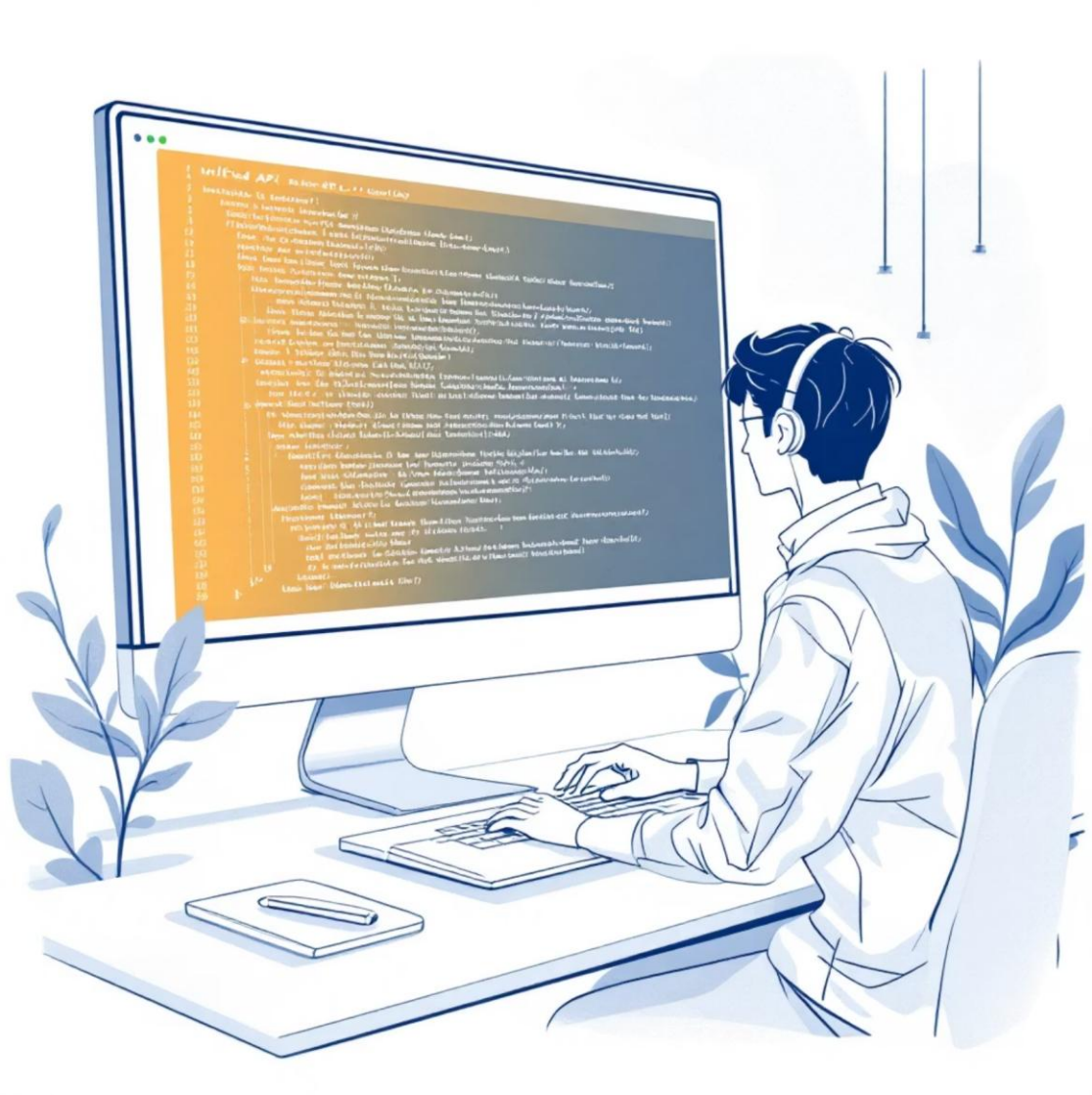
构建在Spark SQL引擎之上，支持持DataFrame/Dataset操作，提供供更简洁的声明式API

丰富的操作支持

支持select、where、groupBy、map、filter、flatMap等各种数据转换操作操作

Made with GAMMA

Structured Streaming的编程模型优势



01

统一API设计

批处理与流处理使用相同的API，大幅简化开发复杂度和学习曲线

02

自动增量计算

引擎自动管理状态，持续更新计算结果，开发者无需关注状态管理细节

03

高级功能支持

原生支持事件时间窗口、流与静态数据的join、水印机制等企业级功能

Made with GAMMA



响应速度的飞跃

1秒

Spark Streaming

秒级延迟

100毫秒

Structured Streaming微批处理

亚秒级延迟

1毫秒

Structured Streaming持续处理

毫秒级延迟

从秒级到毫秒级的性能提升，为实时业务决策提供更快的响应能力

Made with GAMMA

实时响应能力深度对比

1

Spark Streaming

采用微批处理架构，端到端延迟通常在秒级范围，适合对实时性要求不高的场景

2

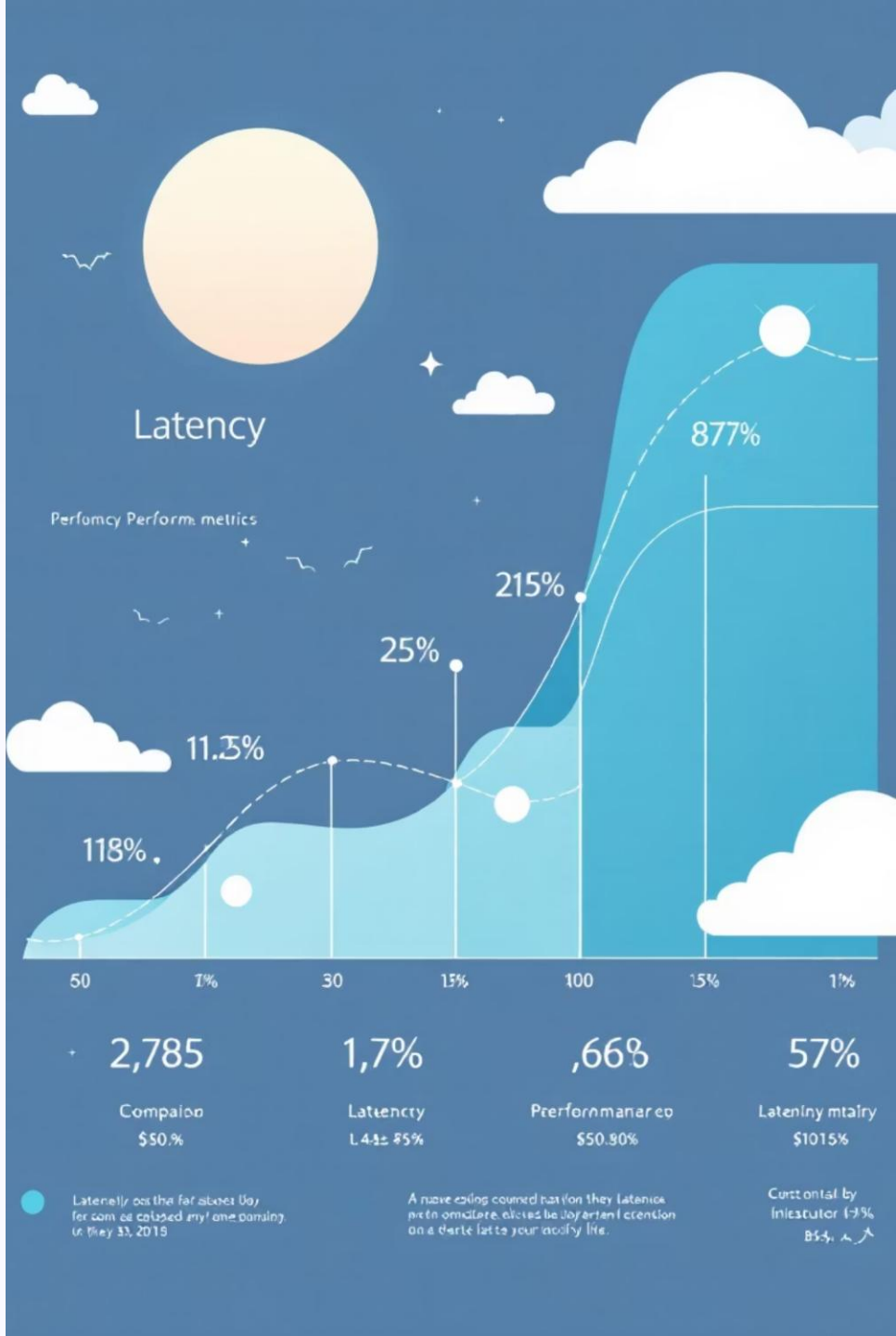
Structured Streaming微批处理模式

采用全新设计方式，端到端延迟可降低至100毫秒级别，性能提升显著

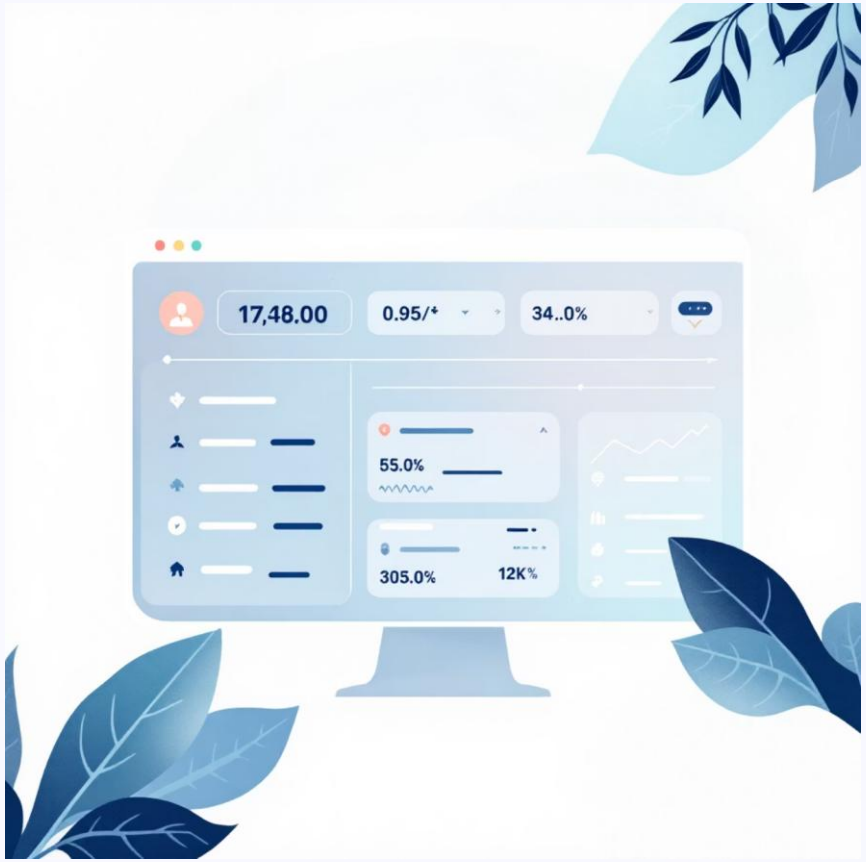
3

Structured Streaming持续处理模式

支持毫秒级实时响应，端到端延迟可低至1毫秒，满足超低延迟需求



Structured Streaming实时模式详解



超低延迟性能

Databricks Runtime 16.4及以上版上版本支持，端到端延迟可低至5毫秒

操作型工作负载

专为实时业务决策设计，如欺诈检测、个性化促销推荐等关键应用场景

并行调度优化

查询所有阶段并行调度执行，数据实时shuffle传递，最小化等待时间

Made with GAMMA



Spark Streaming与Structured Streaming对比总结

Spark Streaming

数据抽象

基于DStream与RDD，编程模型较底层

开发复杂度

功能成熟但API复杂，学习曲线陡峭

维护状态

处于维护模式，新特性开发较少

Structured Streaming

数据抽象

基于DataFrame，统一批流处理API

易用性

高级API设计，开发效率显著提升

未来趋势

Spark流处理主流方向，推荐新项目采用

Made with GAMMA



结语：Structured Streaming引领流处理新时代

简化开发范式

结合Spark SQL与流处理优势，统一批流处理编程模型，降低技术门槛

企业级特性

支持高吞吐量、低延迟、端到端容错的实时数据应用，满足生产环境需求

数据驱动未来

助力企业实现数据驱动的实时智能决策，构建现代化数据基础设施

> "从批处理到流处理，从DStream到DataFrame，Structured Streaming代表着Apache Spark流处理技术的未来方向"

Made with GAMMA