法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象学院
  - 新浪微博：小象AI学院

大纲

- SparkStreaming概述
- SparkStreaming工作原理
- SparkStreaming程序设计

大纲

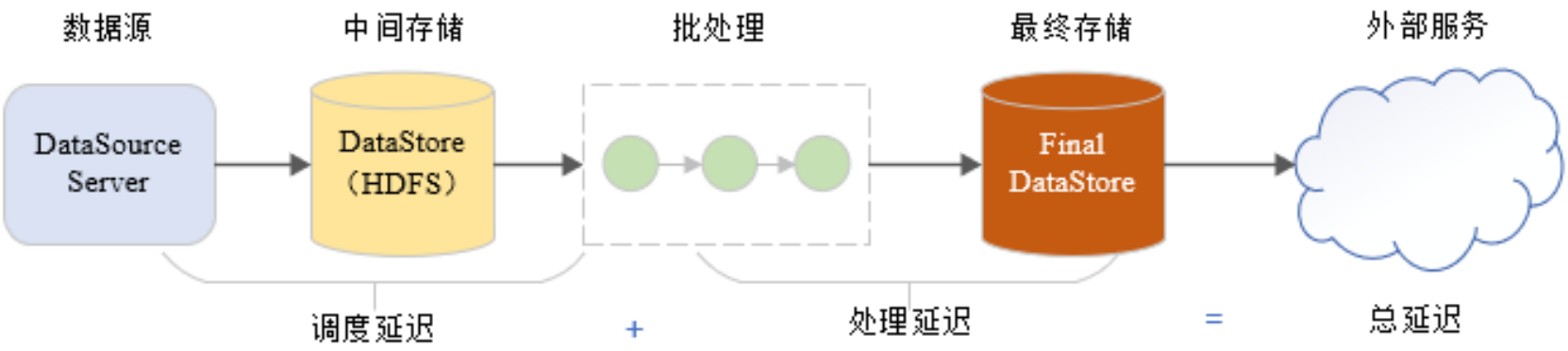
SparkStreaming概述

SparkStreaming设计动机

- 很多重要的应用要处理大量在线流式数据，并返回近实时的结果
  - 社交网络趋势跟踪
  - 电商网站指标统计
  - 广告系统
- 具备分布式流式处理框架的基本特征
  - 良好的扩展性
  - 低延迟（秒级别）

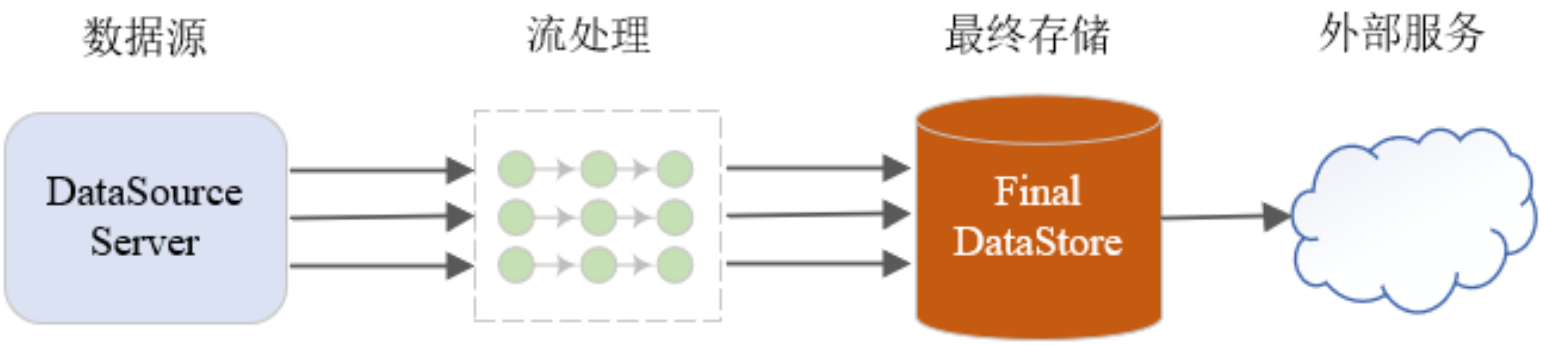
批处理

- 调度延迟
- 处理延迟

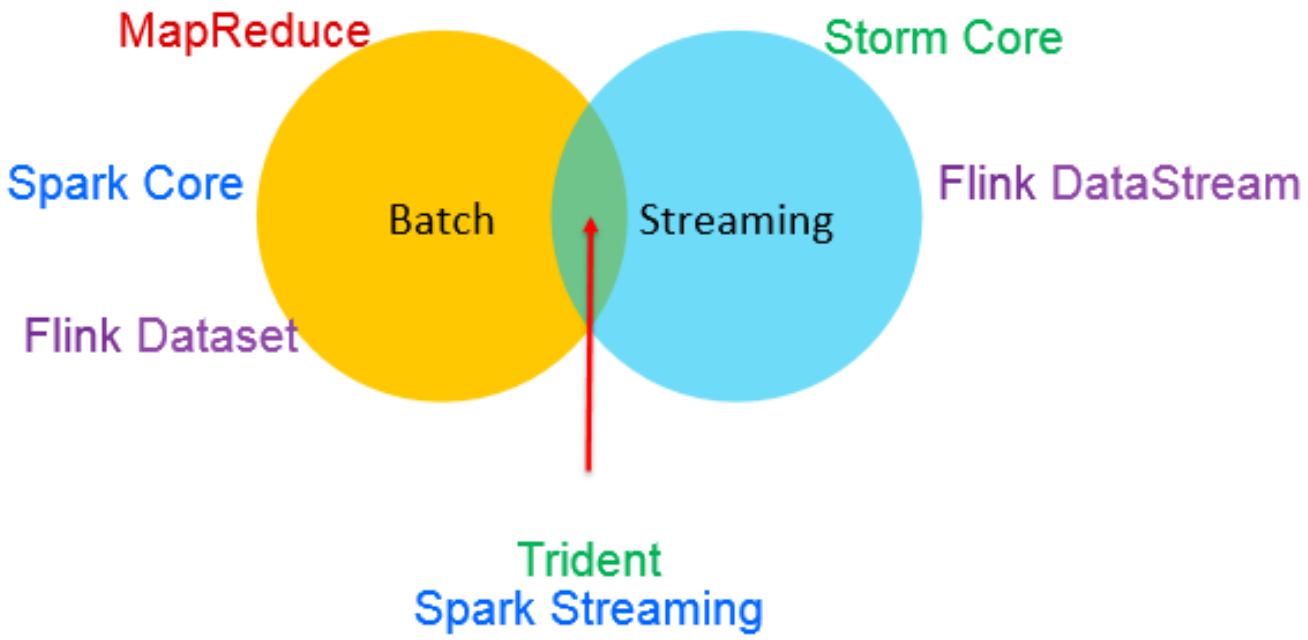


流式处理

- 流式处理
- 低延迟



流式计算框架


MapReduce
Storm Core
Spark Core
Batch
Streaming
Flink DataStream
Flink Dataset
Trident
Spark Streaming

什么是SparkStreaming

- 将Spark扩展为大规模流处理系统
- 可以扩展到100节点规模，达到秒级延迟
- 高效且具有良好的容错性
- 提供了类似批处理的API，很容易实现复杂算法

SparkStreaming特点

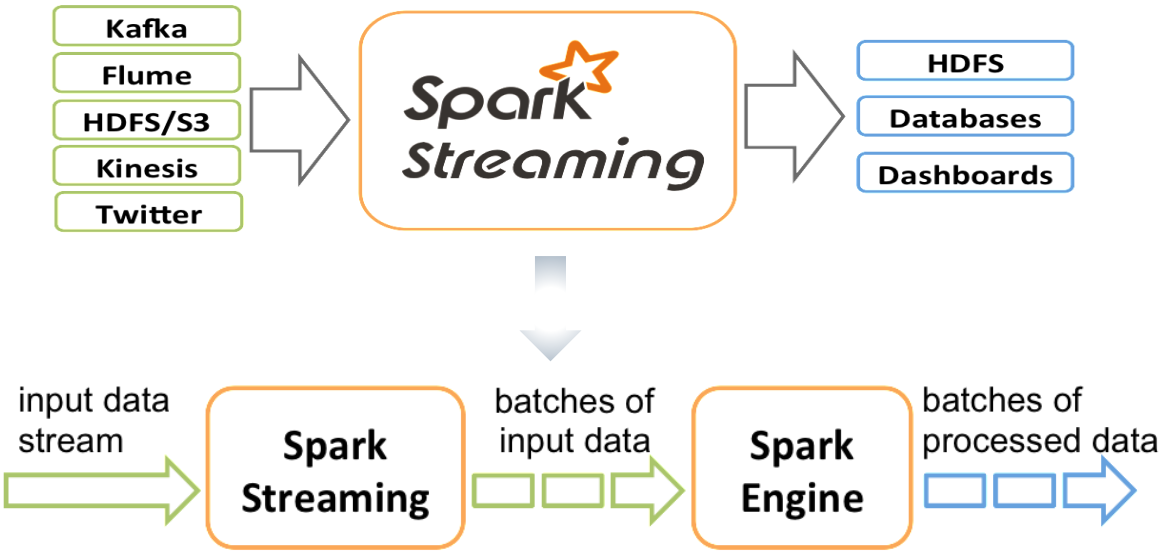
- 易用性好
  - 提供很多高级算子，实现复杂运算非常简单
  - 流式API和批处理API很类似，学习成本低
- 平台统一
  - 不需要维护两套系统分别用于批处理和流式处理
  - 可以自由调用Spark的组件，如SparkSQL、Mllib
- 生态丰富
  - 支持各种数据源和数据格式
  - 社区活跃，发展迅猛

大纲

SparkStreaming工作原理

SparkStreaming原理

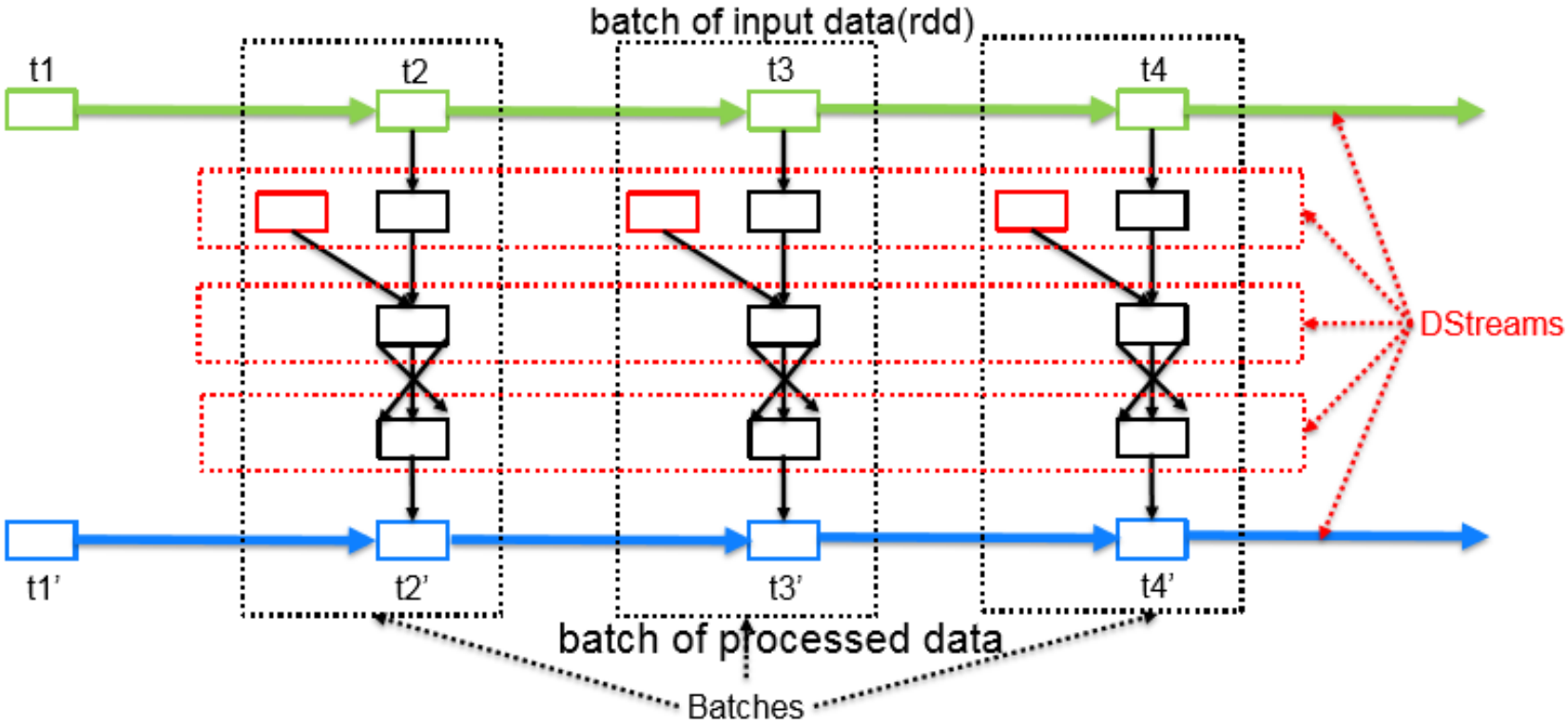
- 将流式计算转化为一批很小的、确定的批处理作业（micro-batch）
  - 以X秒为单位将数据流切分成离散的作业
  - 将每批数据看成RDD，使用RDD操作符处理
  - 最终结果以RDD为单位返回（写入HDFS或者其他系统）



Spark组件之间数据集类比

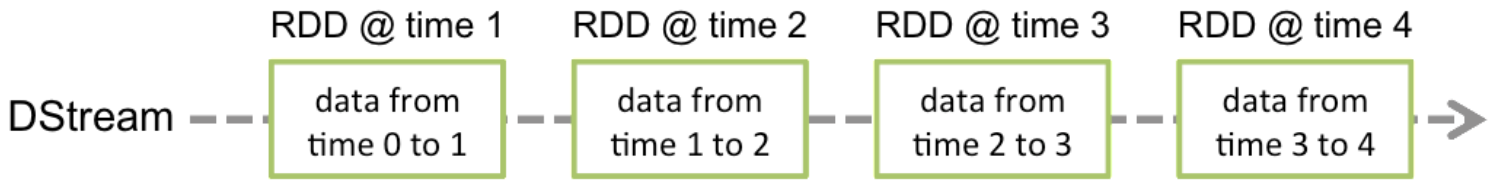
	Data Set	transformation
Spark Core	RDD	RDD -> RDD
Spark SQL	DataFrame/DataSet	DataFrame/DataSet ->DataFrame/DataSet
SparkStreaming	DStream	Dstream -> DStream

核心概念-Dstream & Batch

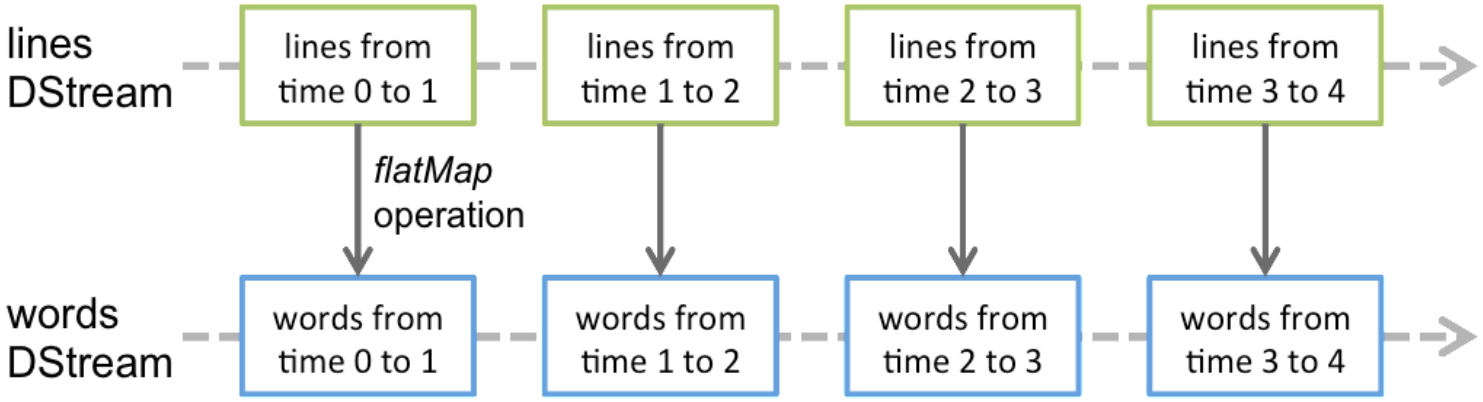


核心概念-DStream

- 将连续的数据进行离散表示
- DStream中每一个离散的片段都是一个RDD



- DStream可以转换成另外一个DStream



Stream Data Source

- 内置数据源
  - socketTextStream
  - textFileStream
  - 其他
- 外部数据源
  - Kafka
  - Flume
  - ZeroMQ
  - 其他

Stream Transformation

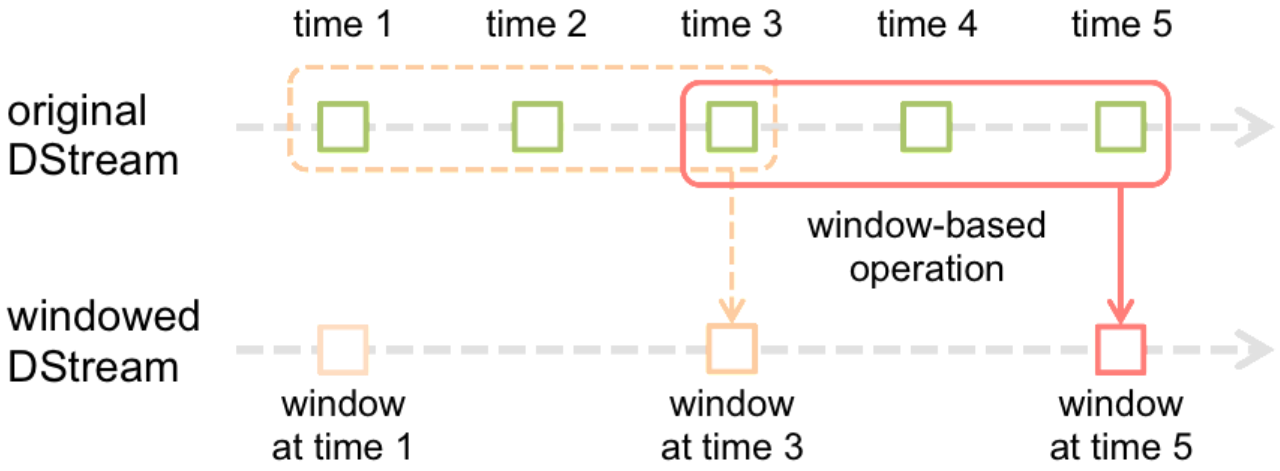
- 类RDD转换
  - map、flatMap、filter、reduce
  - groupBykey、reudceByKey、join
- Streaming独有转换
  - window
  - mapWithState

Stream Output

- 将处理过的数据输出到外部系统
- 内置输出
  - print
  - saveAsTextFiles
- 自定义输出
  - foreachRDD

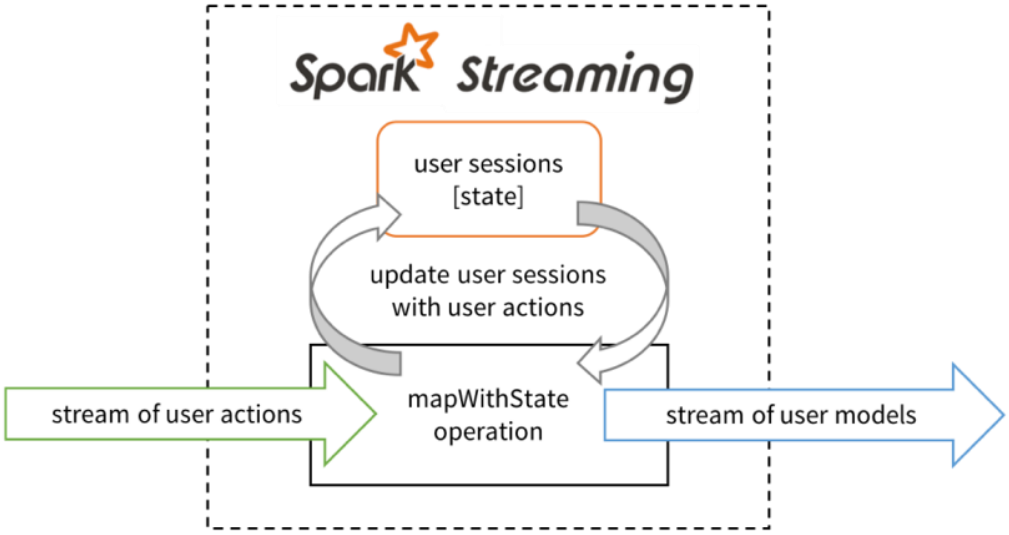
基于窗口的transformation函数

- window、countByWindow、reduceByWindow等
- window length：窗口长度
- sliding interval：滑动窗口时间间隔
- 示例：pairs.reduceByKeyAndWindow((a:Int,b:Int) => (a + b), Seconds(3), Seconds(2))



mapWithState

- 由Spark Streaming自己维护状态信息，不需要借助外部的存储系统
- 相对updateStateByKey性能提升10倍左右
- 相对updateStateByKey维护的key状态多10倍
- 接收参数为StateSpec对象，返回一个新的DStream



Stream Checkpoint

- 可以checkpoint的两种类型数据
  - Metadata checkpointing，针对Driver中的元数据设置检查点，包括配置信息、DStream一系列操作、提交了job但未完成的batch等
  - Data checkpointing，保存stateful带状态操作的数据
- Checkpoint局限性
  - Application 重新编译后，从checkpoint中恢复会失败，需要清空checkpoint

大纲

SparkStreaming程序设计

Spark Streaming程序设计

```
val conf = new SparkConf().setMaster("local[2]")

val ssc = new StreamingContext(conf,Seconds(5))

val ds = ssc.socketTextStream("192.168.183.100",8888)

val rs = ds.flatMap(_.split(" ")).map((_,1))

            .reduceByKey(_ + _)

rs.print()

ssc.start()

ssc.awaitTermination()
```

流式上下文

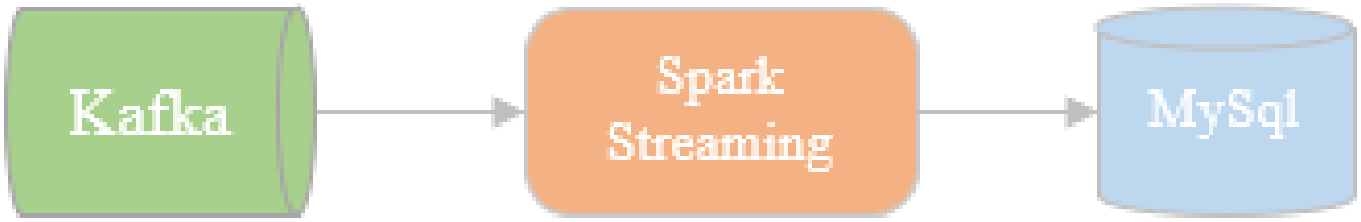
流式数据输入

流式转换

流式数据输出

启动流式处理

实时流处理系统设计与实现



消息队列　　流式实时计算　　结果保存

用户行为分析系统需求

- 用户行为分析系统处理流程
  - 用户使用的客户端会收集用户行为事件（以点击事件为例），将数据发送到Kafka
  - 后端基于SaprkStreaming的实时分析系统从kafka中消费数据，进行实时分析
  - 实时系统分析完成的数据写入到外部存储MySql，可以实时获取用户的行为数据，并可以导出进行离线统计分析

用户行为分析系统数据源

- 数据源
  - Kafka订制主题
  - 一个事件包含4个字段：
    - ✓ deviceId：软件设备版本号
    - ✓ deviceType：软件设备类型
    - ✓ time：事件发生的时间戳
    - ✓ click：点击次数
  - 数据格式：deviceId|deviceType|time|click

疑问

- 小象问答官网
  - http://wenda.chinahadoop.cn

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象学院
- 新浪微博：小象AI学院