DataStream API编程

崔星灿 · York University / Apache Flink Committer

Aache Flink 在线教程 — 2019年03月26日

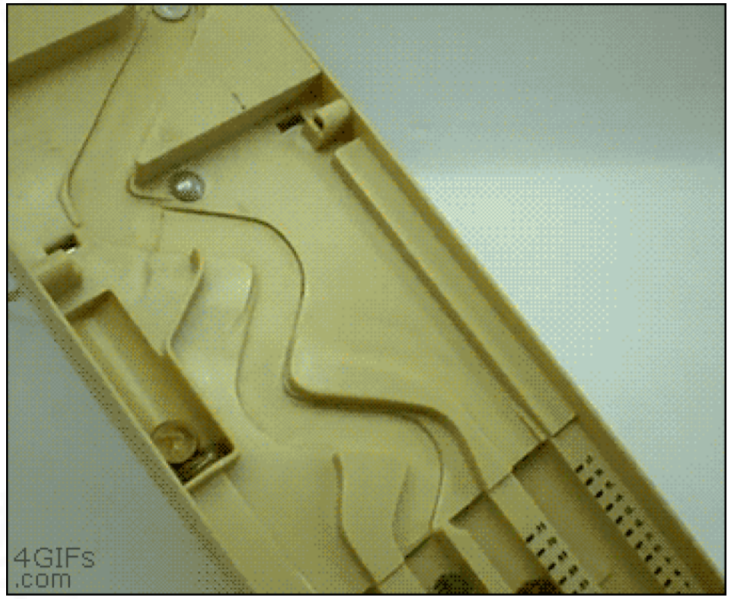
CONTENT

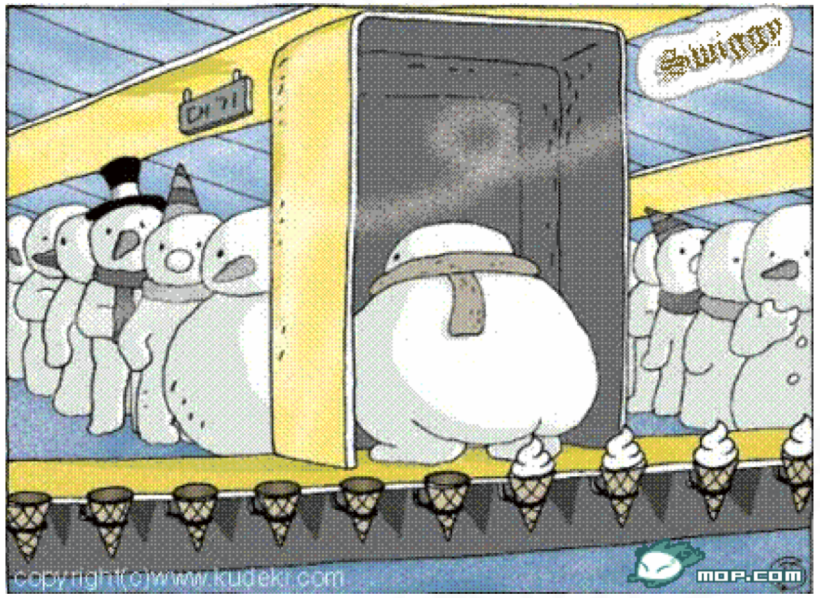
目录 >>

01

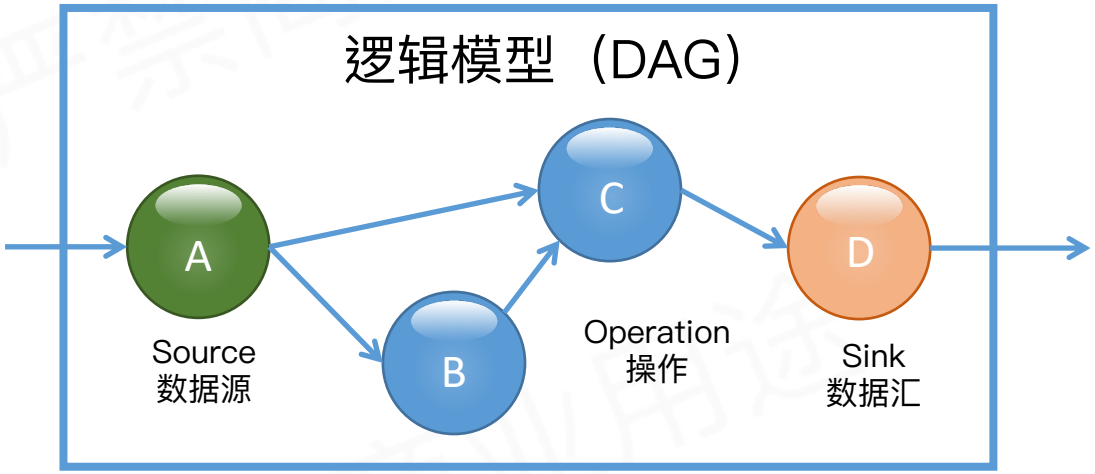
流式应用开发基础

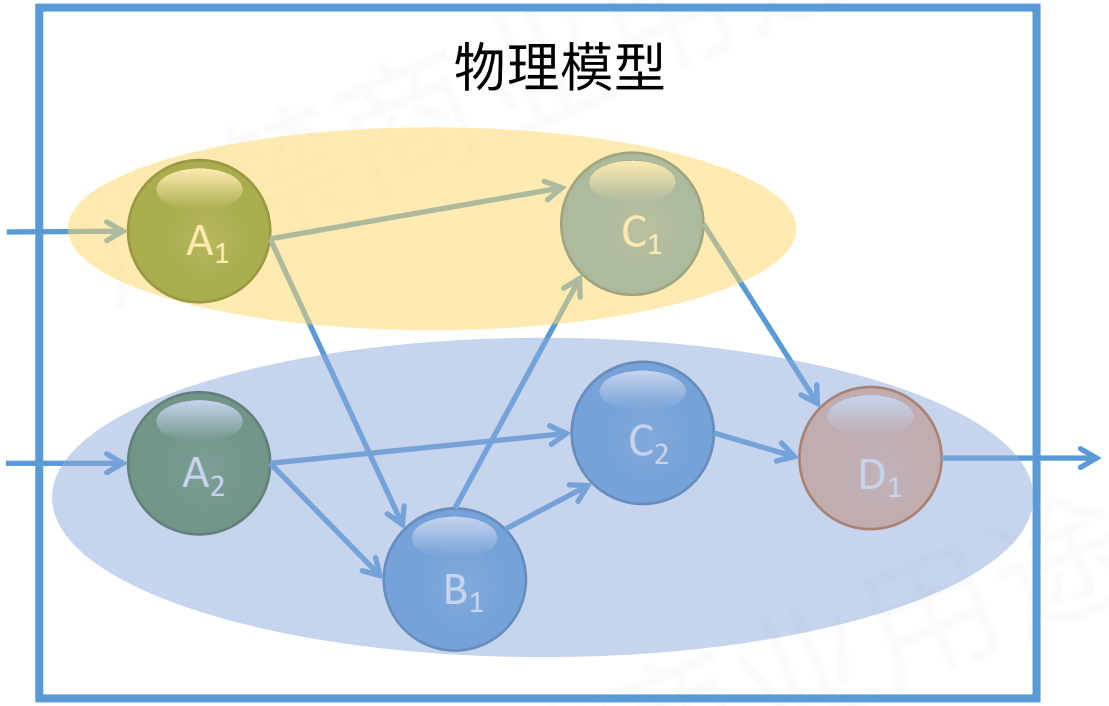
什么是流处理





分布式流处理的基本模型





流处理API的衍变

```
TopologyBuilder builder = new TopologyBuilder();

builder.setSpout("spout", new RandomSentenceSpout(), 5);
builder.setBolt("split", new SplitSentence(), 8).shuffleGrouping("spout");
builder.setBolt("count", new WordCount(), 12).fieldsGrouping("split", new Fields("word"));
```

Apache Storm

“面向操作”，低层

```
StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();

DataStream<String> text = env.readTextFile ("input");
DataStream<Tuple2<String, Integer>> counts = text.flatMap(new Tokenizer()).keyBy(0).sum(1);
counts.writeAsText("output");
```

Apache Flink

“面向数据”，高层

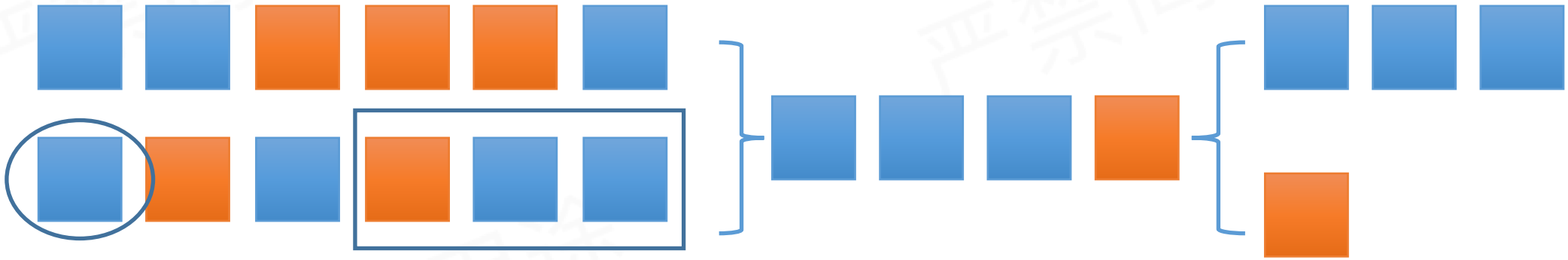
02

Flink DataStream API概览

Flink DataStream程序结构

```
//1、设置运行环境
StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
//2、配置数据源读取数据
DataStream<String> text = env.readTextFile ("input");
//3、进行一系列转换
DataStream<Tuple2<String, Integer>> counts = text.flatMap(new Tokenizer()).keyBy(0).sum(1);
//4、配置数据汇写出数据
counts.writeAsText("output");
//5、提交执行
env.execute("Streaming WordCount");
```

操作概览



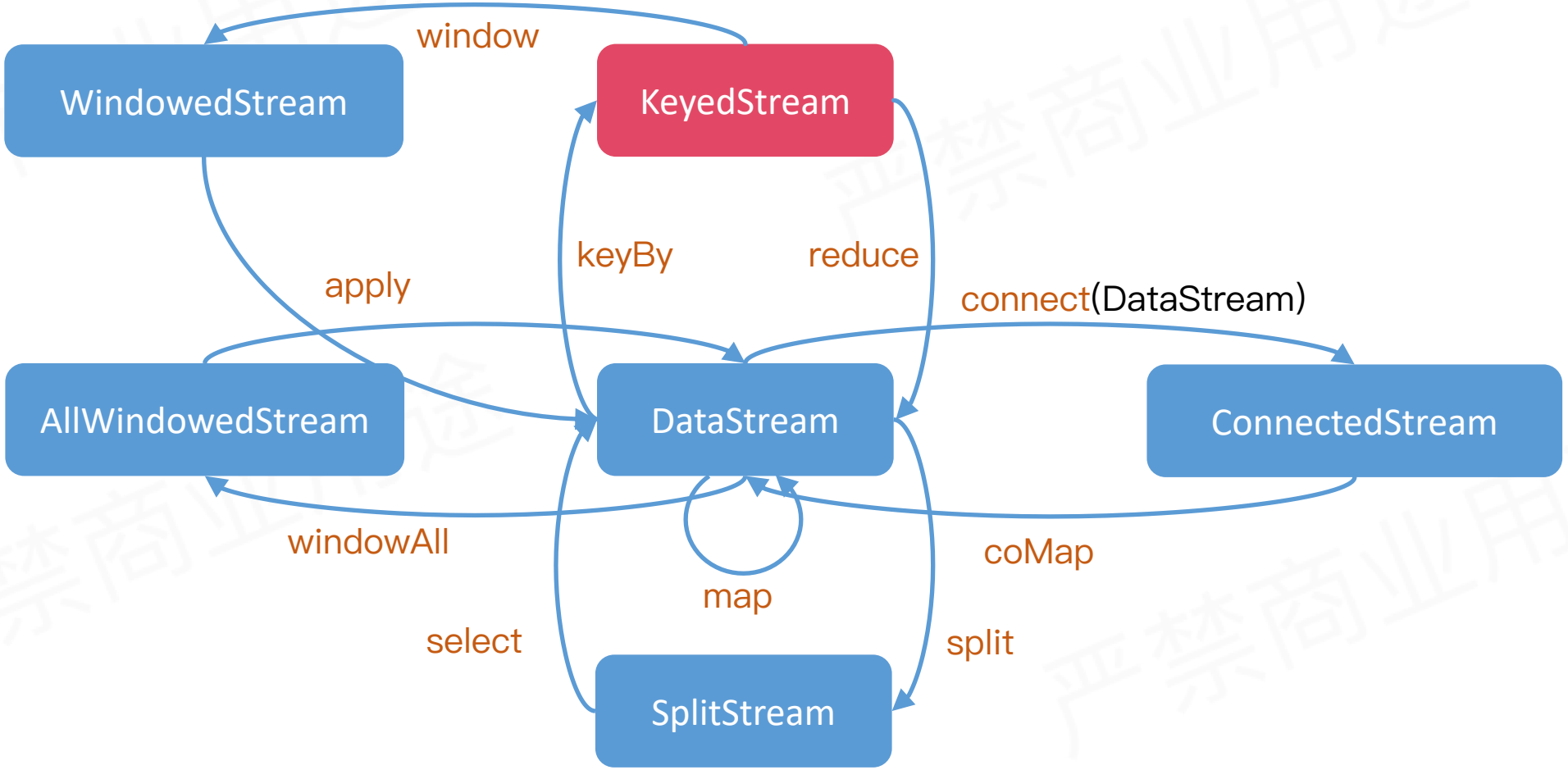
1、基于单条记录
filter、map

2、基于窗口
window

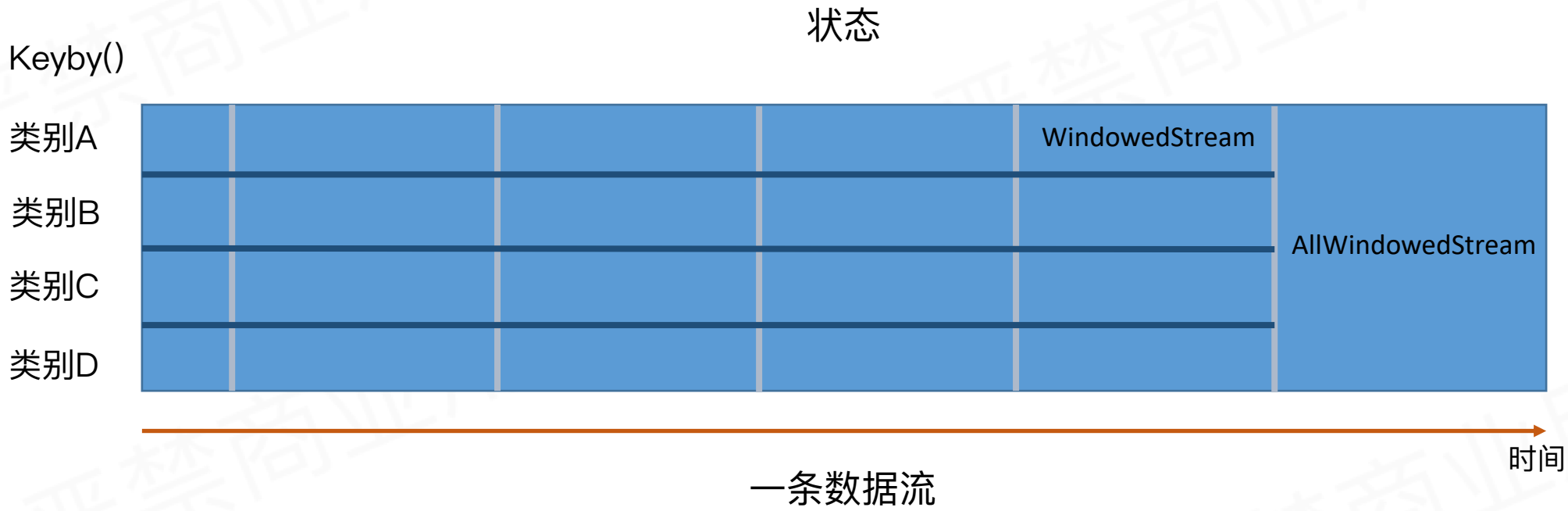
3、合并多条流
union、join、connect

4、拆分单条流
split

DataStream 基本转换



理解KeyedStream

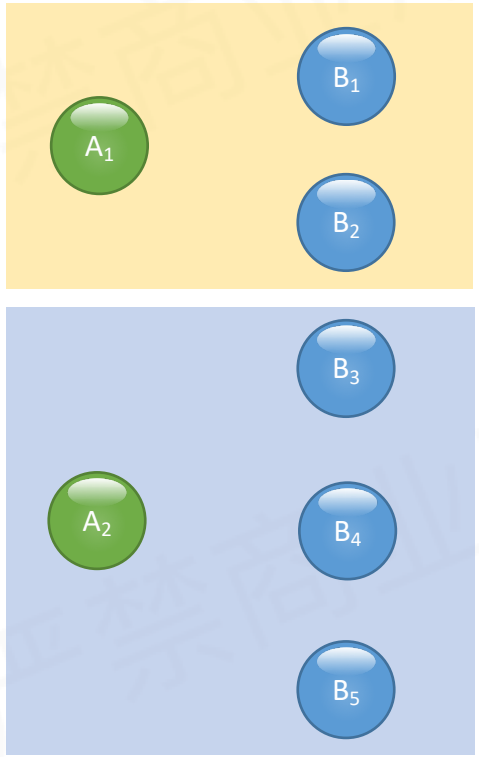


key数 >> 并发度

03

其他问题

物理分组



类型	描述
dataStream.global();	全部发往第1个task
dataStream.broadcast();	广播
dataStream.forward();	上下游并发度一样时一对一发送
dataStream.shuffle();	随机均匀分配
dataStream.rebalance();	Round-Robin（轮流分配）
dataStream.recale();	Local Round-Robin（本地轮流分配）
dataStream.partitionCustom()	自定义单播

类型系统

```
DataStream<String> text = env.readTextFile ("input");
DataStream<Tuple2<String, Integer>> counts = text.flatMap(new Tokenizer()).keyBy(0).sum(1);
```

TypeInformation

类型	说明
基本类型	Java的基本类型（包装类）以及void、String、Date、BigDecimal、BigInteger
复合类型	Tuple和Scala case class（不支持null）、Row、POJO
辅助、集合类型	Option、Either、List、Map等
上述类型的数组	
其他类型	自定义TypeInformation或Kryo处理，不推荐使用

04

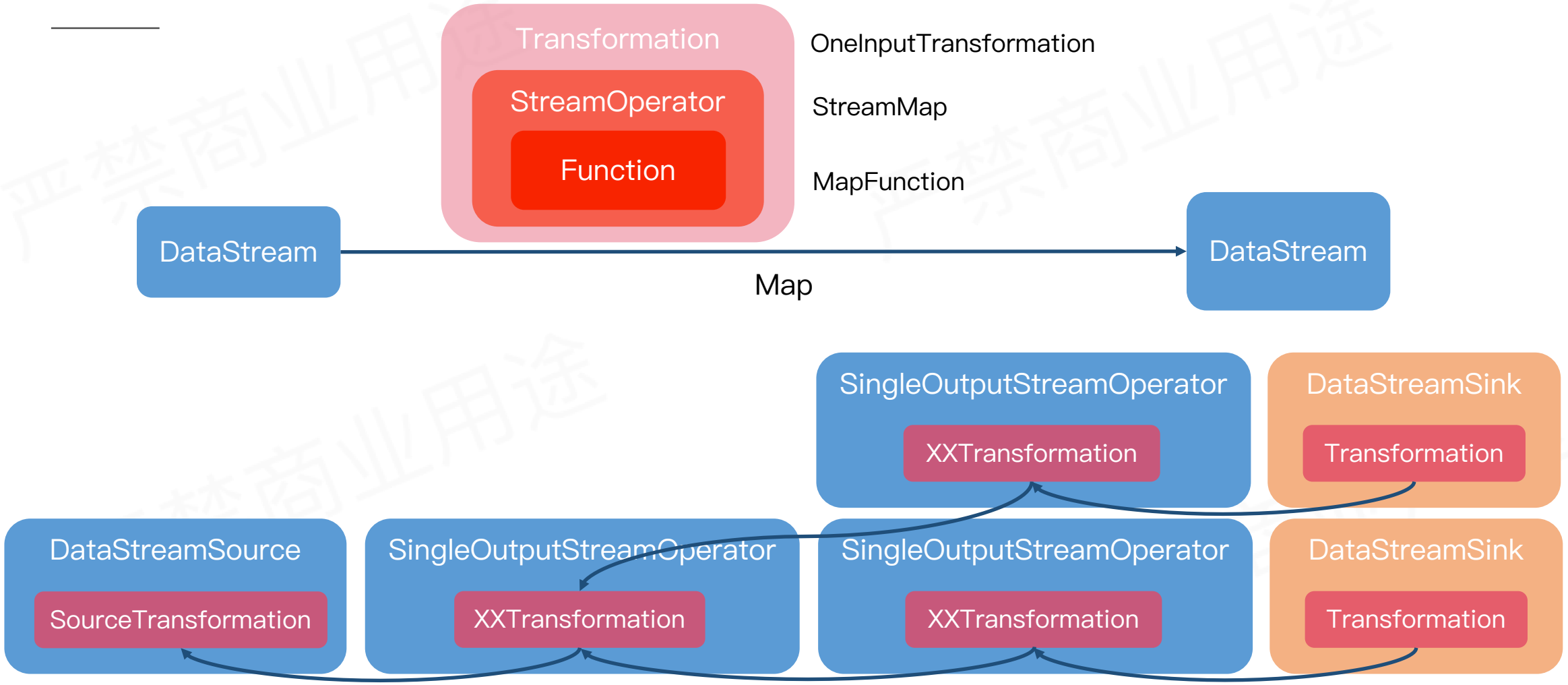
源码简析

示例：实时统计成交额

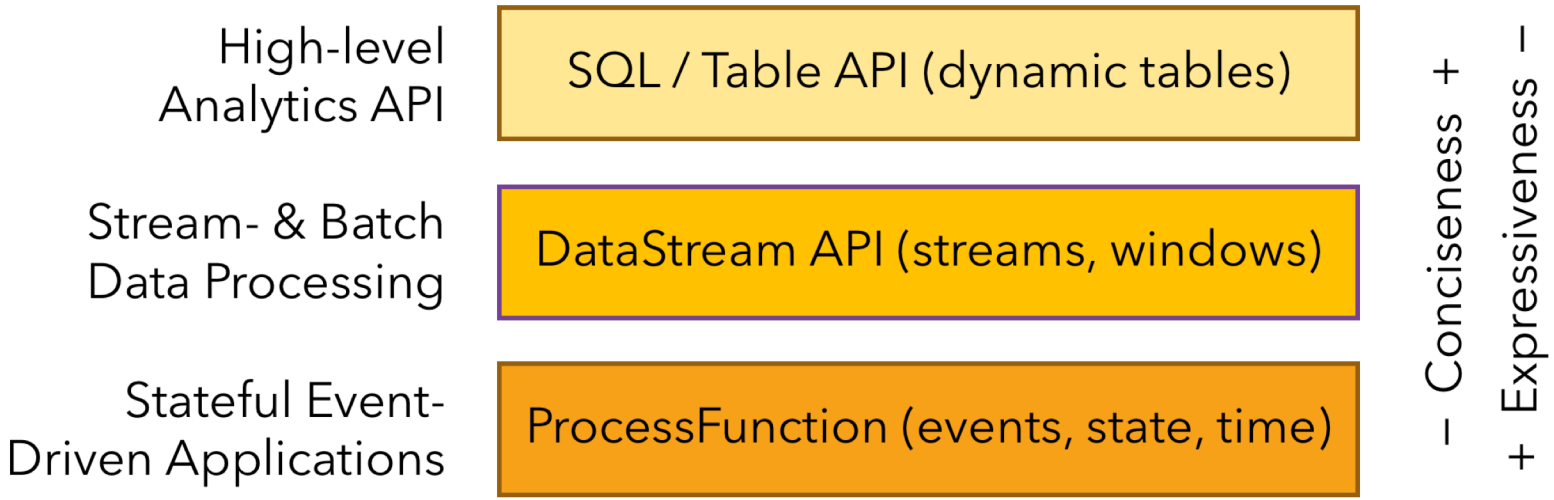
数据源：Tuple2<String, Integer> 商品类别 -> 成交额

任务：1、实时统计每个类别的成交额； 2、实时统计全部类别的成交额

API 原理



总结



THANKS

Flink China社区大群

扫一扫群二维码，立刻加入该群。