Day 9
实时流计算-快速上手

实时大数据：流数据普遍，但没有充分产生价值

物联网/边缘/传感器

车联网

StreamingML

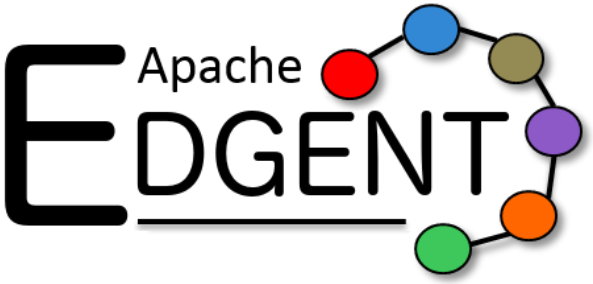
没有好用的实时流计算平台？

- 实时故障检测
- 危险驾驶行为分析
- 电子围栏
- 偏航检测
- 超速检测
- ...

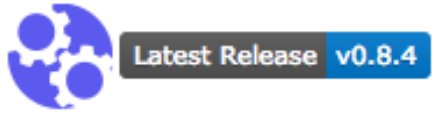
- 实时异常检测
- 实时分类
- 实时反欺诈
- 实时推荐

HUAWEI

开源流计算框架



稳了！！



Apache Gearpump

什么是实时流计算

实时：实时处理，计算框架按事件逐条实时处理

流：one by one的数据流

计算：数学运算、数据分析、算法模型执行等

实时流计算：实时处理当下正在发生的流数据，逐条大数据分析或运行机器学习算法

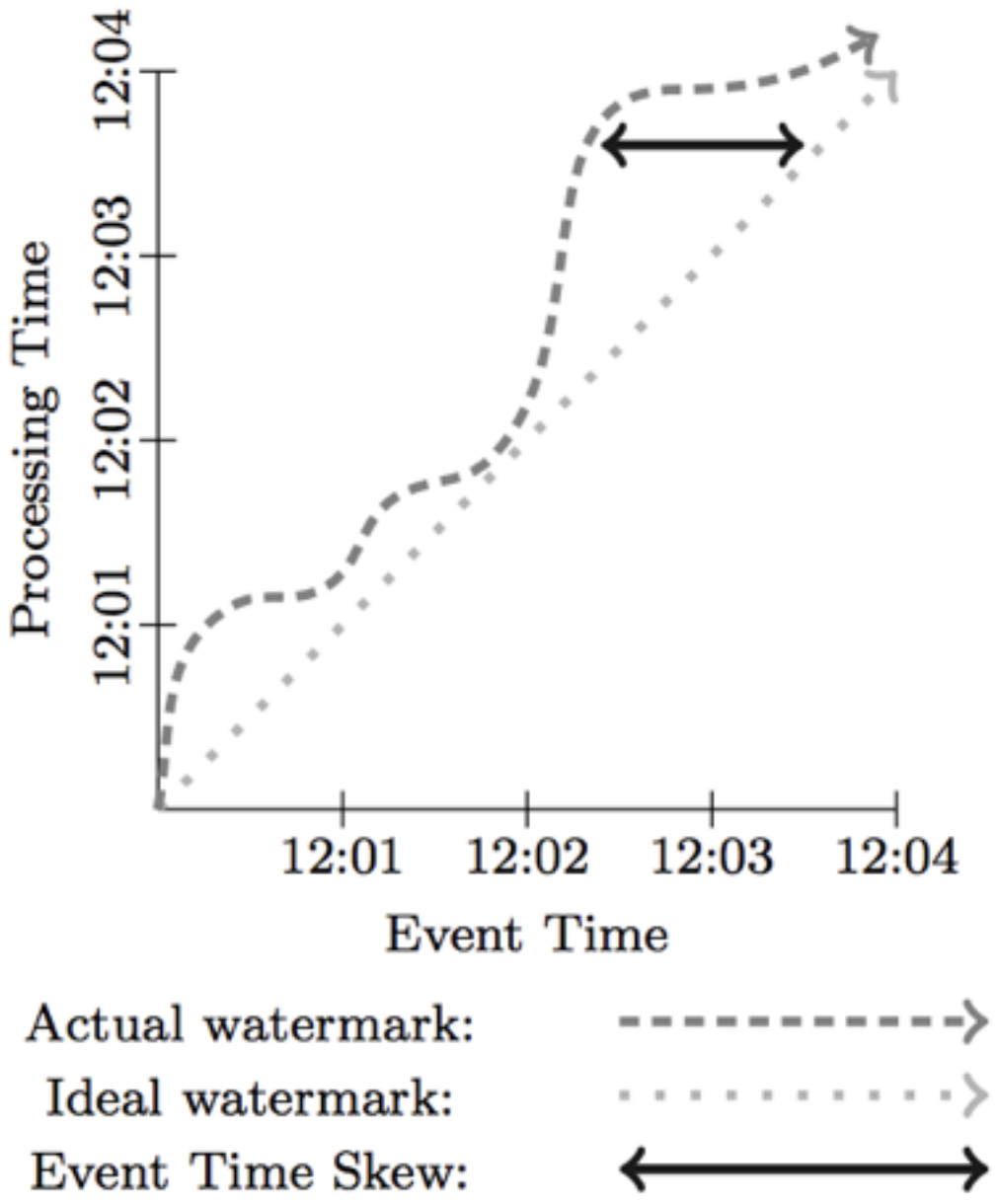


Figure 2: Time Domain Skew

大数据AI：越实时越有价值

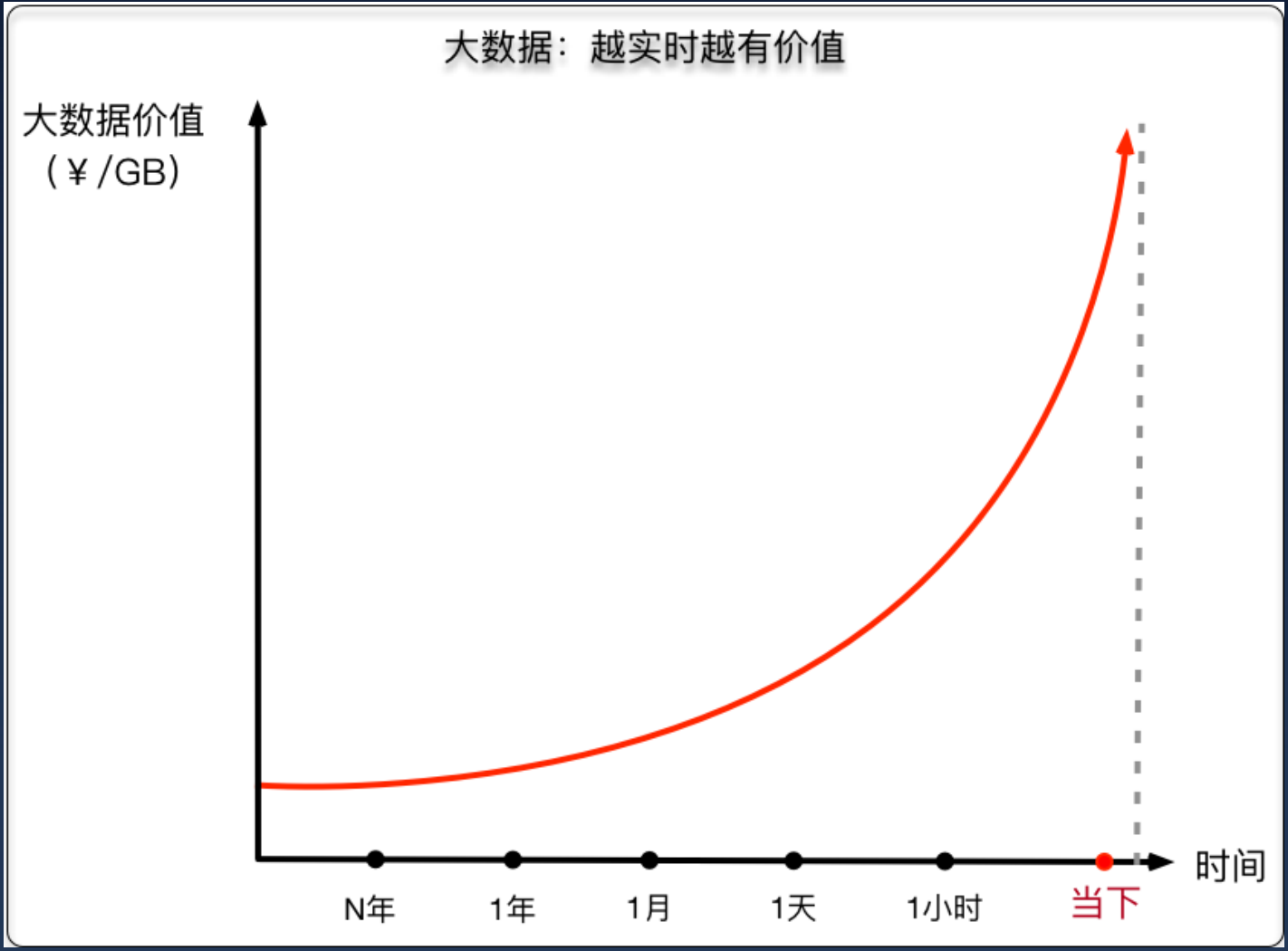
- 事件发生：one by one
- 事件主体：人、机器
- 技术发展：加速度-颠覆
- 人的耐心程度：加速度-降低
- 大数据的增速：加速度-增长

所以

- 实时流计算快速驱动业务
- 实时流计算最大限度挖掘数据价值

适用场景：

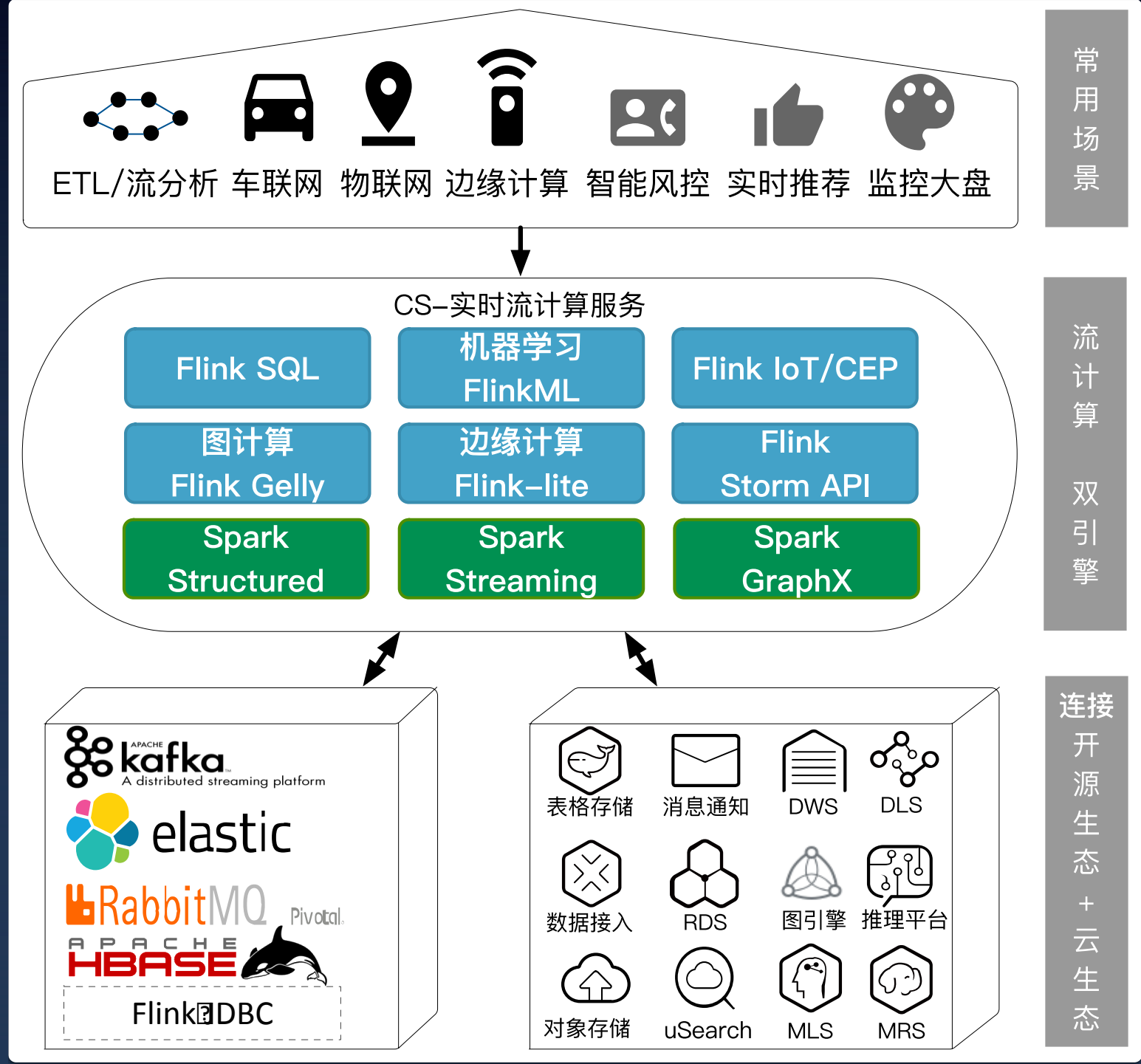
实时推荐(商品/广告)、实时监控大盘、打车、金融风控、异常检测、交通、物流、外卖...



概览

实时流计算服务（Cloud Stream Service, 简称CS）
提供实时处理流式大数据的全栈能力, 简单易用, 即时执行Stream SQL或自定义作业。无需关心计算集群, 无需学习编程技能。完全兼容Apache Flink和Spark API

https://www.huaweicloud.com/product/cs.html

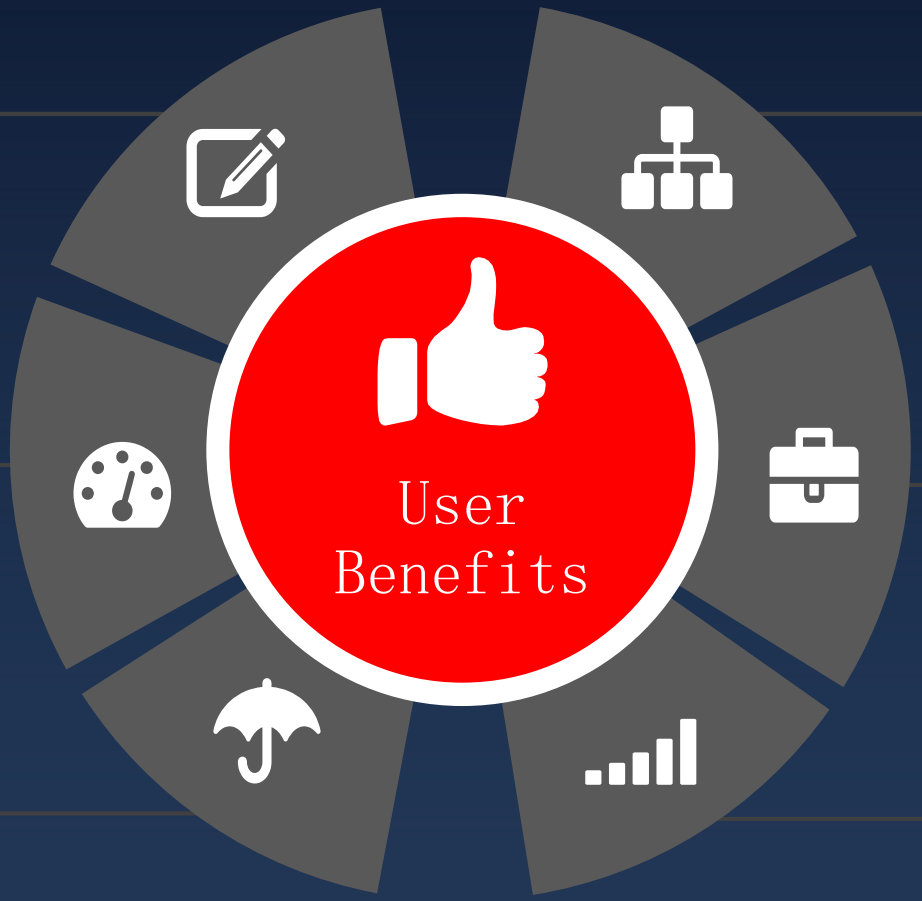


特点

易用
- StreamSQL编辑器
- StreamSQL可视化拖拽
- 在线调试
- 作业可视化监控

按需计费 & 包年包月
- 按实际使用量计费
- 用多少付费多少
- 1 SPU/小时 = 0.5元
- 包年包月更优惠

作业即服务Job as a Service
- 作业输出流可视化
- 作业输出流可订阅
- 作业输出流提供Restful API



低延时 高吞吐
- 毫秒级时延
- 每秒处理百万消息

完整生态 开箱即用
- 开源生态
- 连通云存储和AI服务
- 无需关心基础设施
- 即时执行业务作业

安全可靠
首创完全托管的独享集群

- 物理隔离和安全策略
- 华为软件安全加固

SQL编辑器

特点：

1. 所见即所得
2. SQL满足80%业务
3. 强大的SQL特性
4. SQL连接一切

保存 | 另存为 | 语义校验 | 调试 | 提交 | 设为模板

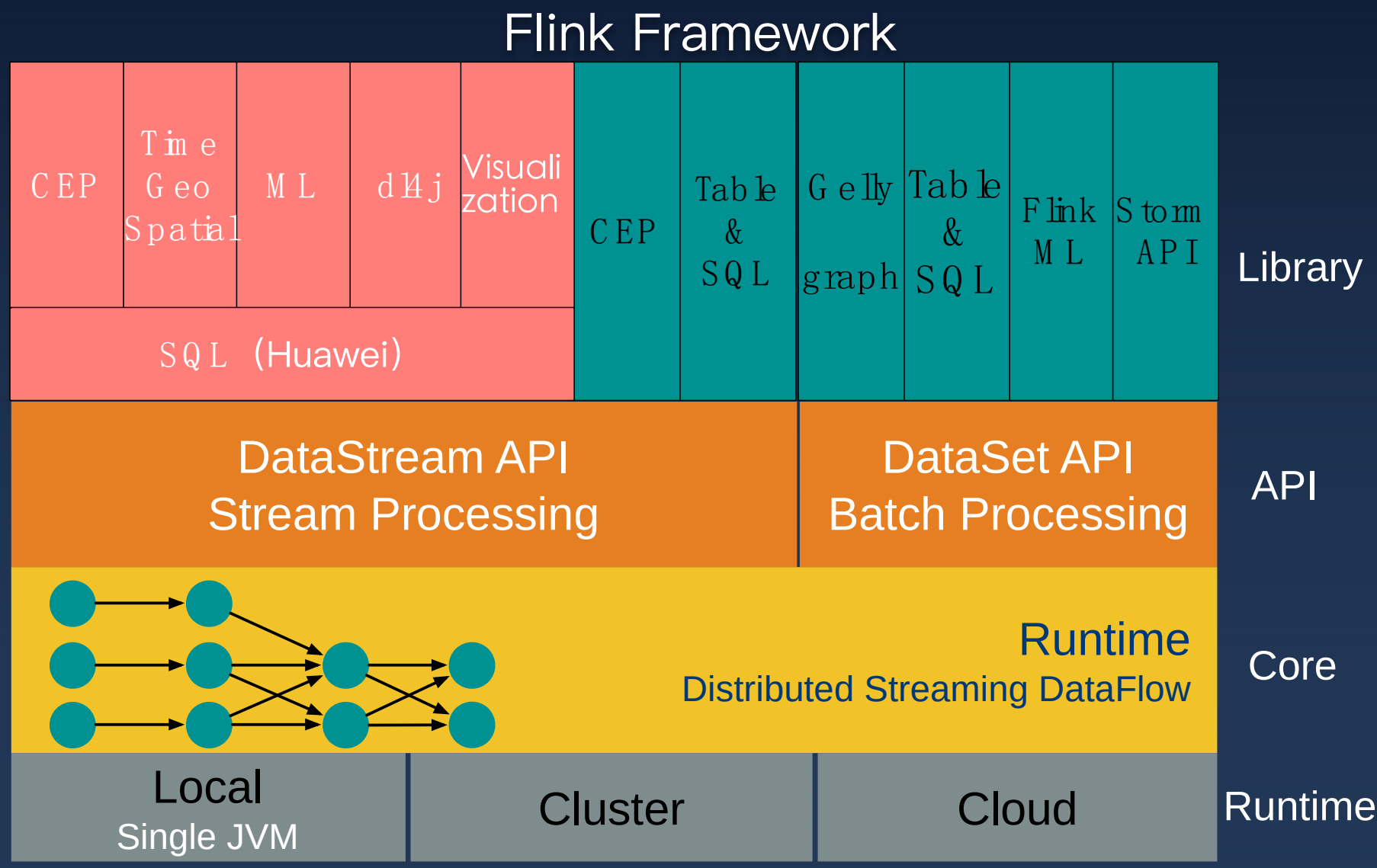
```sql
  * encode:  结果编码方式，可以为csv或者json
  * field_delimiter: 当编码格式为csv时，属性之间的分隔符
  **/
CREATE SINK STREAM fake_licensed_car (
  car_license_number STRING,
  first_zone String,
  second_zone String
)
WITH (
  type = "dis",
  region = "cn-north-1",
  channel = "csoutput",
  partition_key = "car_license_number",
  encode = "csv",
  field_delimiter = ","
);

/** 输出套牌车信息 **/
INSERT INTO fake_licensed_car
SELECT * FROM camera_license_data MATCH_RECOGNIZE
(
  PARTITION BY car_license_number
  ORDER BY proctime
  MEASURES A.car_license_number as car_license_number, A.camera_zone_number as first_zone,
  ONE ROW PER MATCH
  AFTER MATCH SKIP TO LAST C
  PATTERN (A B+ C+)
  WITHIN interval '5' minute
  DEFINE
    B AS B.camera_zone_number <> A.camera_zone_number,
    C AS C.camera_zone_number = A.camera_zone_number
) MR;
```

错误: 0　　行 78, 列 6

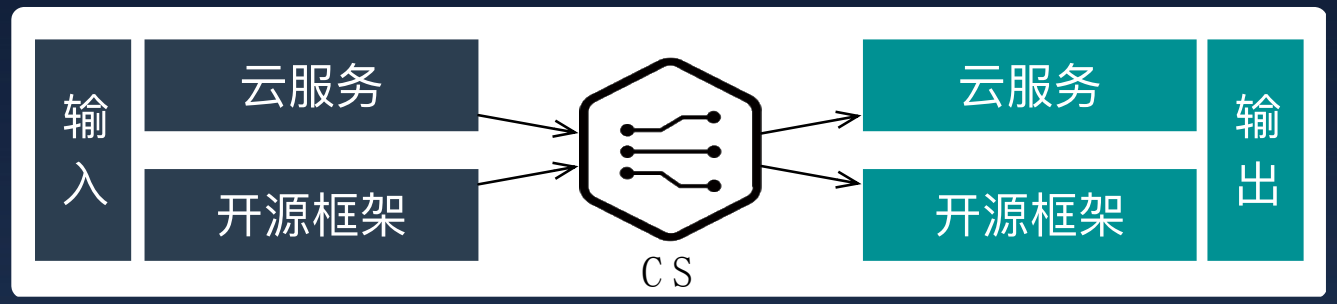
流计算双引擎：Flink + Spark

Spark ecosystem:

1. Structured Streaming
2. Spark Streaming
3. GraphX
4. Spark ML



全生态支持：开源生态+华为云生态



示例：

```
CREATE SINK STREAM yaw_warning (
  MessageContent STRING /* 偏航消息内容 */
)
WITH (
  type = "smn",
  region = "cn-north-1",
  topic_urn = "urn:smn:cn-north-1:a77d6595e37d443fab32d1db9739ed23:Yaw_alarm",
  message_subject = "Yaw_alarm",
  message_column = "MessageContent"
);
```

https://support.huaweicloud.com/usermanual-cs/cs_01_0002.html

车联网-Time GeoSpatial实践地理分析

DDL for Time Geospatial – 基本元素

1. ST_Point(latitude, longitude) 纬度和经度构成点
2. ST_Line(array[point1...pointN]) 多点构成线
3. ST_POLYGON(array[point1...point1]) 多点构成多边形
4. ST_CIRCLE(point, radius) 点和半径构成圆

SQL Geospatial Scalar Functions - 基本操作

1. ST_DISTANCE 计算两点间距离
示例： select *ST_DISTANCE(ST_POINT(x1, y1), ST_POINT(x2, y2))* FROM input
2. ST_PERIMETER 计算多边形周长
示例： Select *ST_PERIMETER(ST_POLYGON(ARRAY[ST_POINT(x11, y11), ST_POINT(x12, y12), ST_POINT(x11, y11)])* FROM input
3. ST_AREA (polygon) 计算多边形面积
示例： Select ST_AREA(ST_POLYGON(ARRAY[ST_POINT(x11, y11), ST_POINT(x12, y12), ST_POINT(x11, y11)]) FROM input
4. ST_OVERLAPS (polygon1, polygon2) 多边形是否相交
5. ST_INTERSECTS 检查两条线是否相交
6. ST_WITHIN 检查一个点是否被包含在一个几何形状中
7. ST_CONTAINS 检查一个多边形是否包含另一多边形
8. ST_COVERS 检查一个多边形是否被另一多边形覆盖
9. ST_DISJOINT 检查两个多边形是否不相交
10. ST_BUFFER(geometry, distance) 在给定距离的参考多边形周围创建一个多边形
11. ST_INTERSECTION (geometry, geometry) 创建一个多边形用于限定两个输入多边形的交集区域
12. ST_ENVELOPE(geometry) 创建一个包含输入的多边形的最小矩形

SQL Time Geospatial – 高级操作，在窗口中的GEO函数

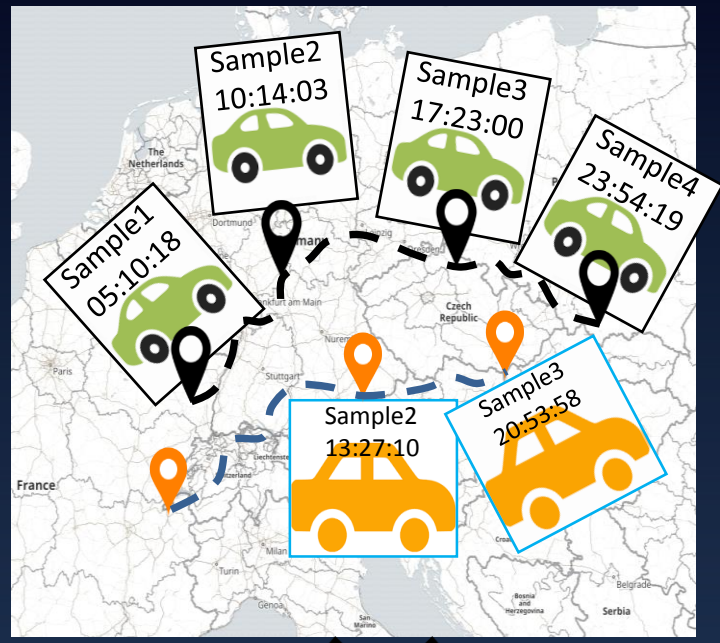
1. AGG_DISTANCE(point) 计算窗口时间内覆盖的距离
示例： SELECT *AGG_DISTANCE*(ST_POINT(x,y)) OVER (ORDER BY proctime RANGE BETWEEN INTERVAL '10' MINUTE PRECEDING AND CURRENT ROW) FROM input
2. AVG_SPEED 计算窗口时间内的速度
示例： SELECT *AVG_SPEED*(ST_POINT(x,y)) OVER (PARTITION BY user ORDER BY rowtime ROWS BETWEEN 10 PRECEDING AND CURRENT ROW) FROM input

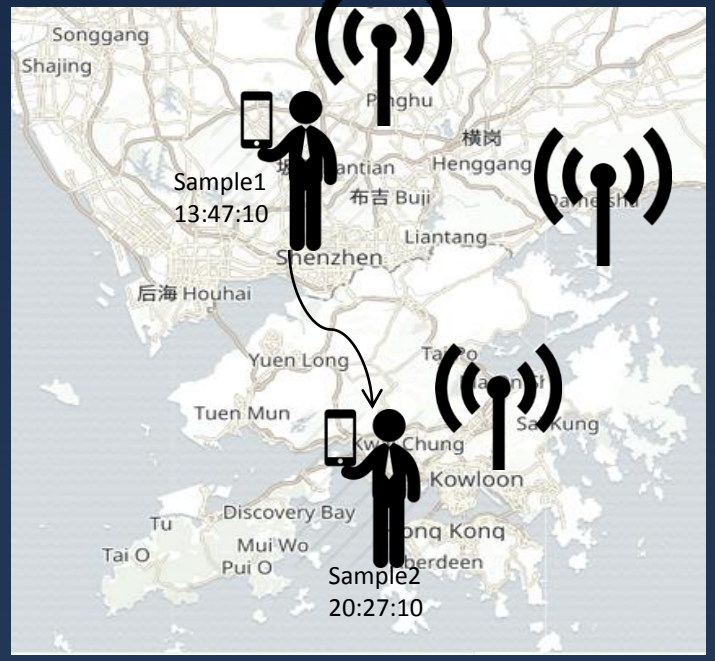
上述函数支持一下窗口：

1. on HOP/TUMBLE/OVER/SESSION windows
2. on count/time windows
3. on rowtime/proctime windows

应用：

1. 偏航告警
2. 区域检测
3. 距离/相交/包含关系
4. 多种窗口的平均速度和距离



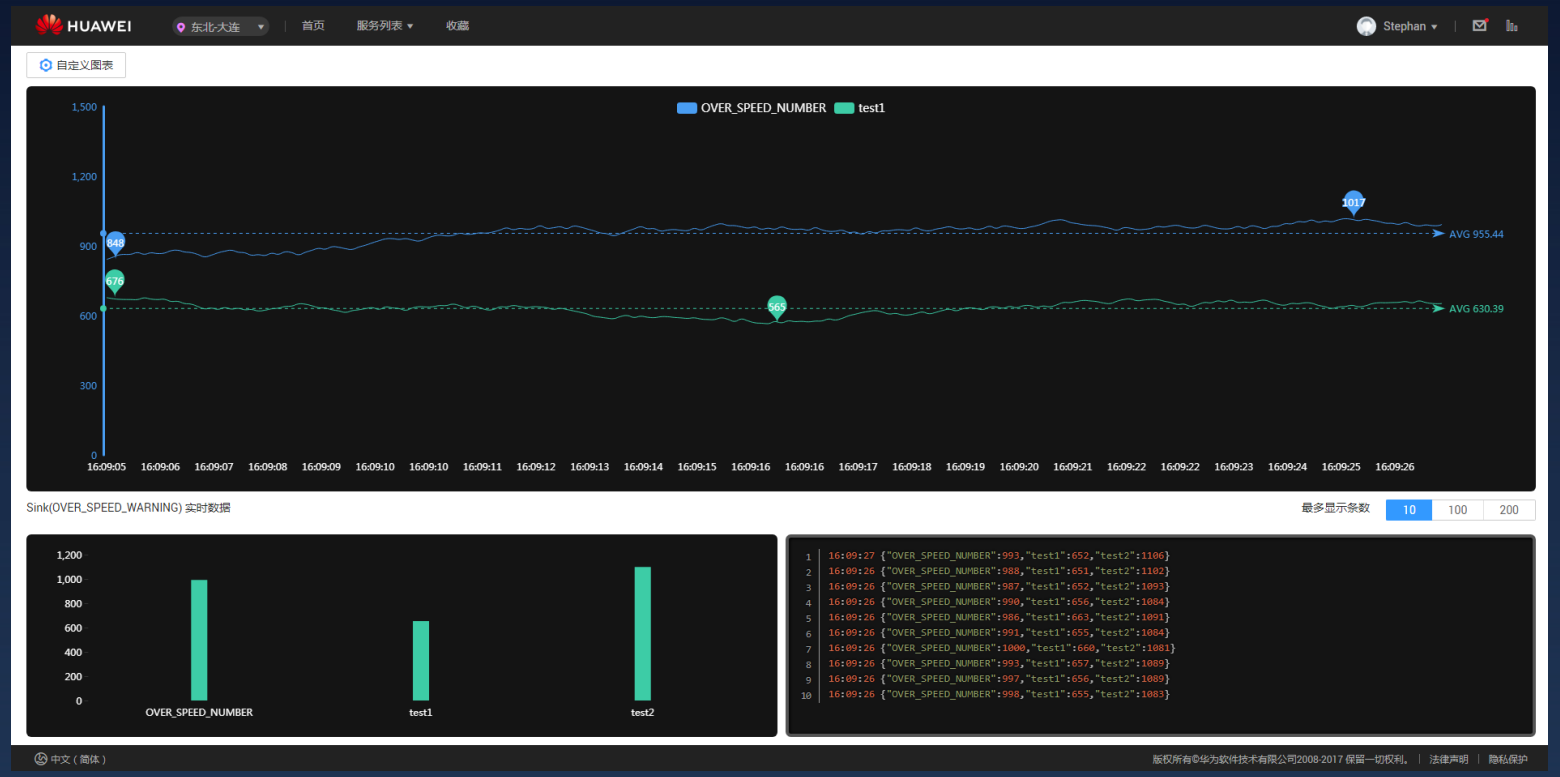


更多资料见 这里

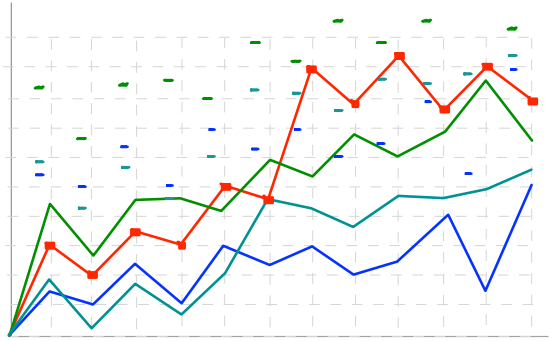
实时监控大盘

输入流数据经Flink计算后，输出流同时支持：

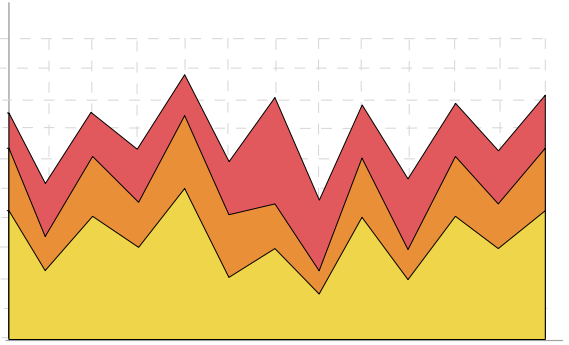
- OpenTSDB持久化：写入CloudTable服务
- **输出流可视化**
  1. 作业提供Restful API，用户订阅输出流
  2. 实时监控大盘



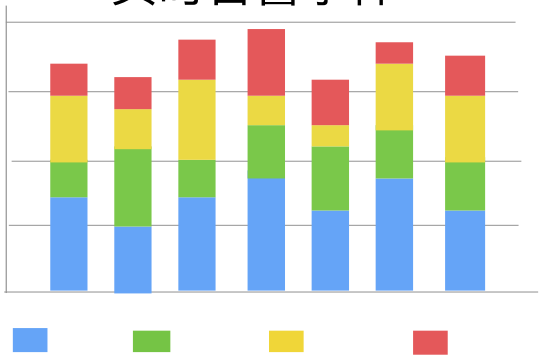
实时–异常检测



实时统计



实时告警事件



实时流计算服务Console页面

实时流计算服务
v1.10.5

总览
作业管理
作业模板
集群管理

功能菜单

作业管理 › test-flinkSql-1030 › 编辑

执行作业

保存 | 另存为 | 语义校验 | 调试 | 提交 | 设为模板

StreamSQL参数设置

```
1  /**
2    * 该示例为流分析场景通用模板。数据的输入源和输出通道均由DIS服务提供，需先开通DIS服务并创
3    * >>>>>>>>>请务必确保您的账户下已在数据接入服务（DIS）里创建了您配置的通道<<<<<<<<<<
4    *
5    * >>>>>样例输入<<<<<
6    *  流名：car_infos(car_id,car_owner,car_brand,car_price):
7    *  1,lilei,bmw320i,28
8    *  2,hanmeimei,audia4,27
9    * >>>>>样例输出<<<<<
10   *  流名：audi_cheaper_than_30w(car_id,car_owner,car_brand,car_price):
11   *  2,hanmeimei,audia4,27
12   **/
13
14 /** 创建输入流，从DIS的csinput通道获取数据。
15   *
16   * 根据实际情况修改以下选项：
17   * channel：数据所在通道名
18   * partition_count：该通道分区数
19   * encode：数据编码方式，可以是csv或json
20   * field_delimiter：当编码格式为csv时，属性之间的分隔符
21   **/
22 CREATE SOURCE STREAM car_infos (
23   car_id STRING,
24   car_owner STRING,
25   car_brand STRING,
26   car_price INT
27 )
28 WITH (
29   type = "dis",
30   region = "cn-north-1",
31   channel = "csinput",
32   partition_count = "1",
```

StreamSQL编辑器

错误: 0　　行 86, 列

运行参数设置　调试参数设置

* SPUs	2
管理单元SPU数: 1，计算单元SPU数: 1	
* 并行数	1
开启Checkpoint	☑
Checkpoint间隔(s)	10
Checkpoint模式	AtLeastOnce
保存作业日志	☑
* OBS桶	changetest-obs-nodele...
	OBS已授权
作业异常告警	☑
* 主题名称	test
异常自动重启	☑
空闲状态保留时长	1 小时

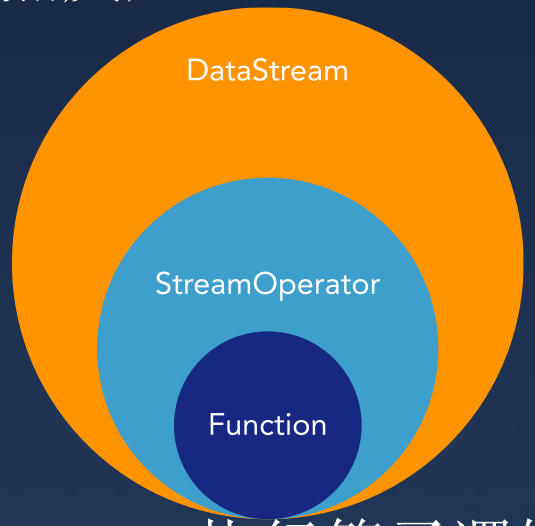
以下内容为延伸阅读

Flink SQL 原理及使用入门
https://mp.weixin.qq.com/s/o_E4KVMAvkt41lRdeUWrEw
（包含**Get Started**、**架构原理**、**语法等**）

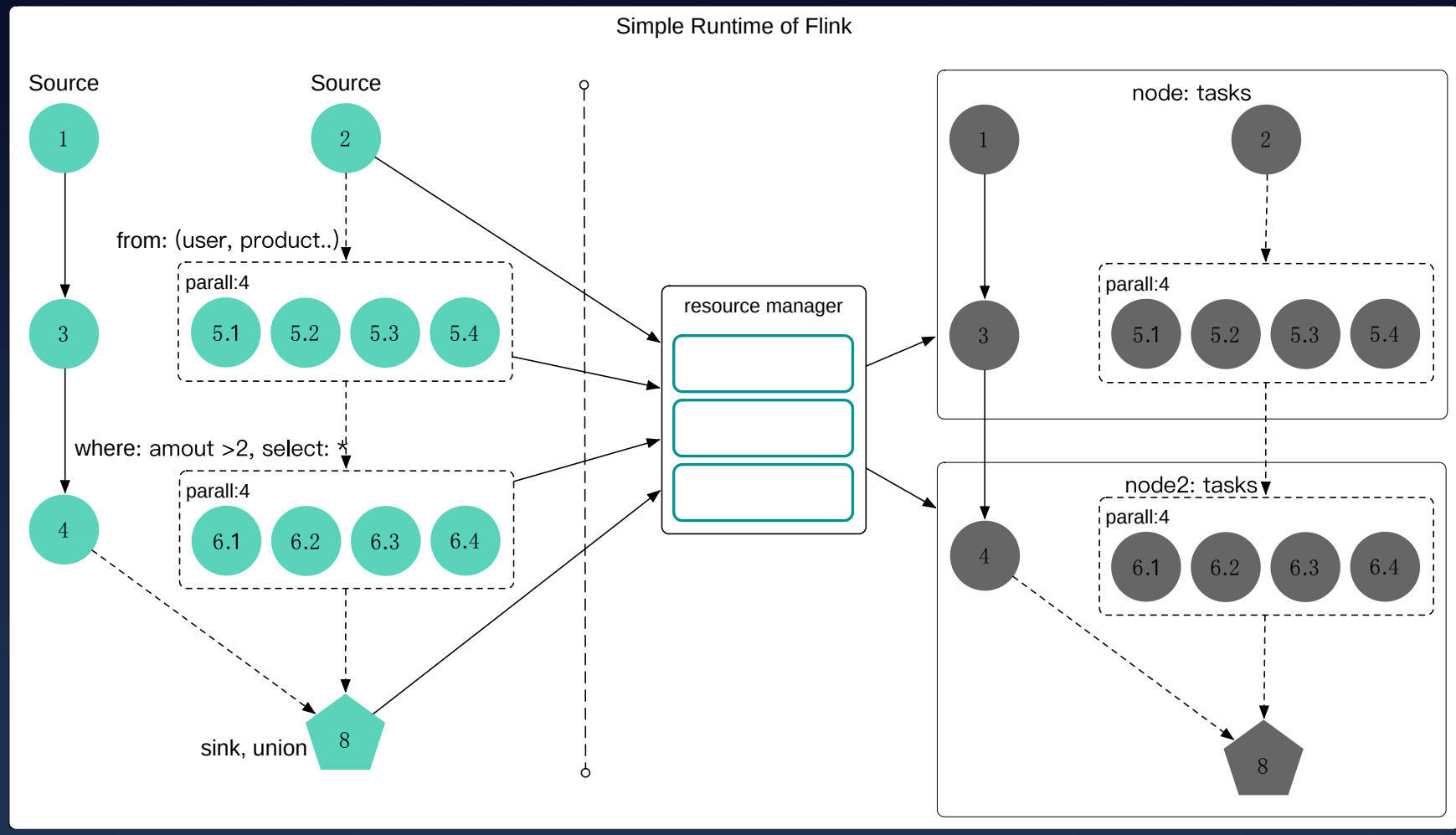
Flink运行时示意图

API或函数的调用链，由逻辑上的算子表示，经过逻辑优化，调度到物理节点上执行。

算子启动顺序为：dataflow中最后一个节点先起，依次向上启动；算子之间通过netty通信，由网络buffer实现自然反压。

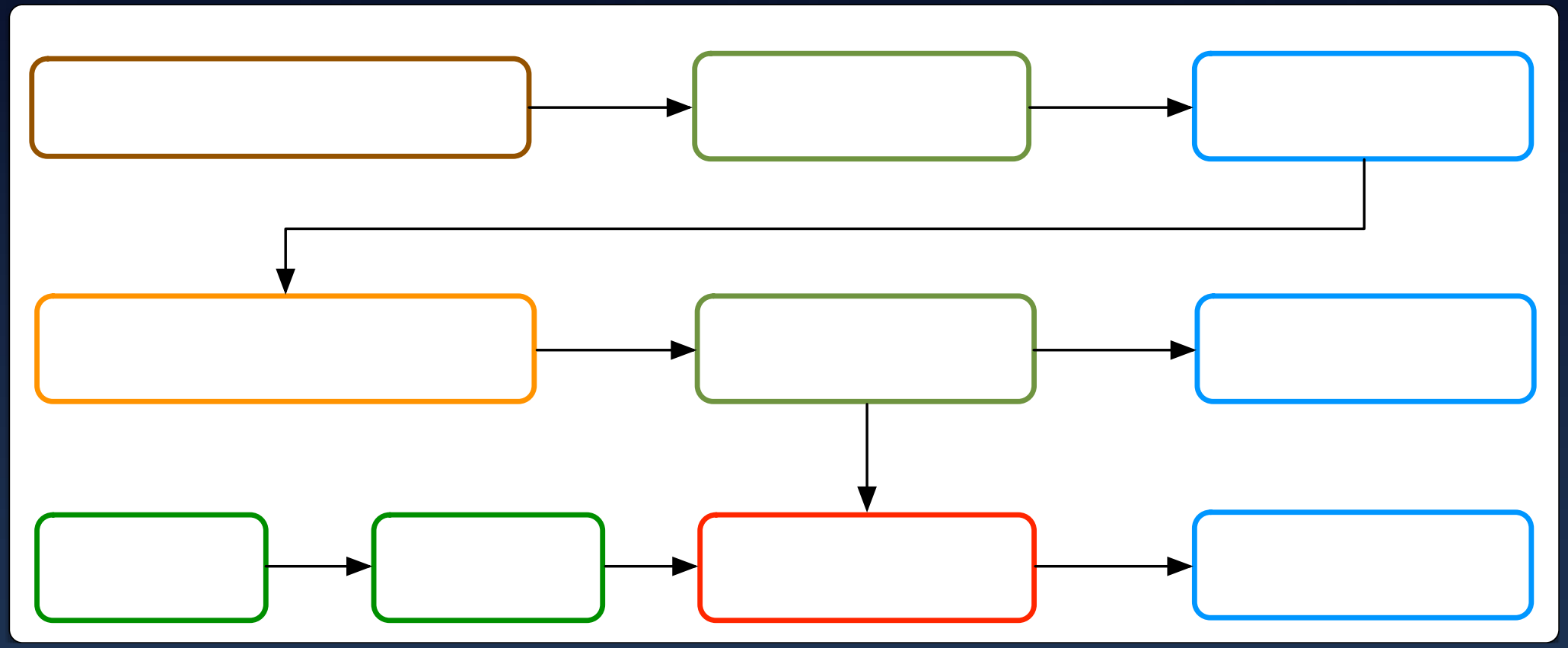


apply() 执行算子逻辑



HUAWEI

Flink SQL运行时的过程



Flink SQL parse, optimize, codegen, pre-compile => DataStream

HUAWEI

Flink vs. Spark

Flink：
- dataflow模型
- 丰富易用的Stream API
- 功能完善：SQL、Table、CEP、ML、Graph

Spark：
- Structured Streaming时延缩短
- 社区活跃
- 生态完善
- Spark StreamSQL

各有所长，距离快速缩小

Thank You.

华为实时流计算服务CS