CLOUDERA

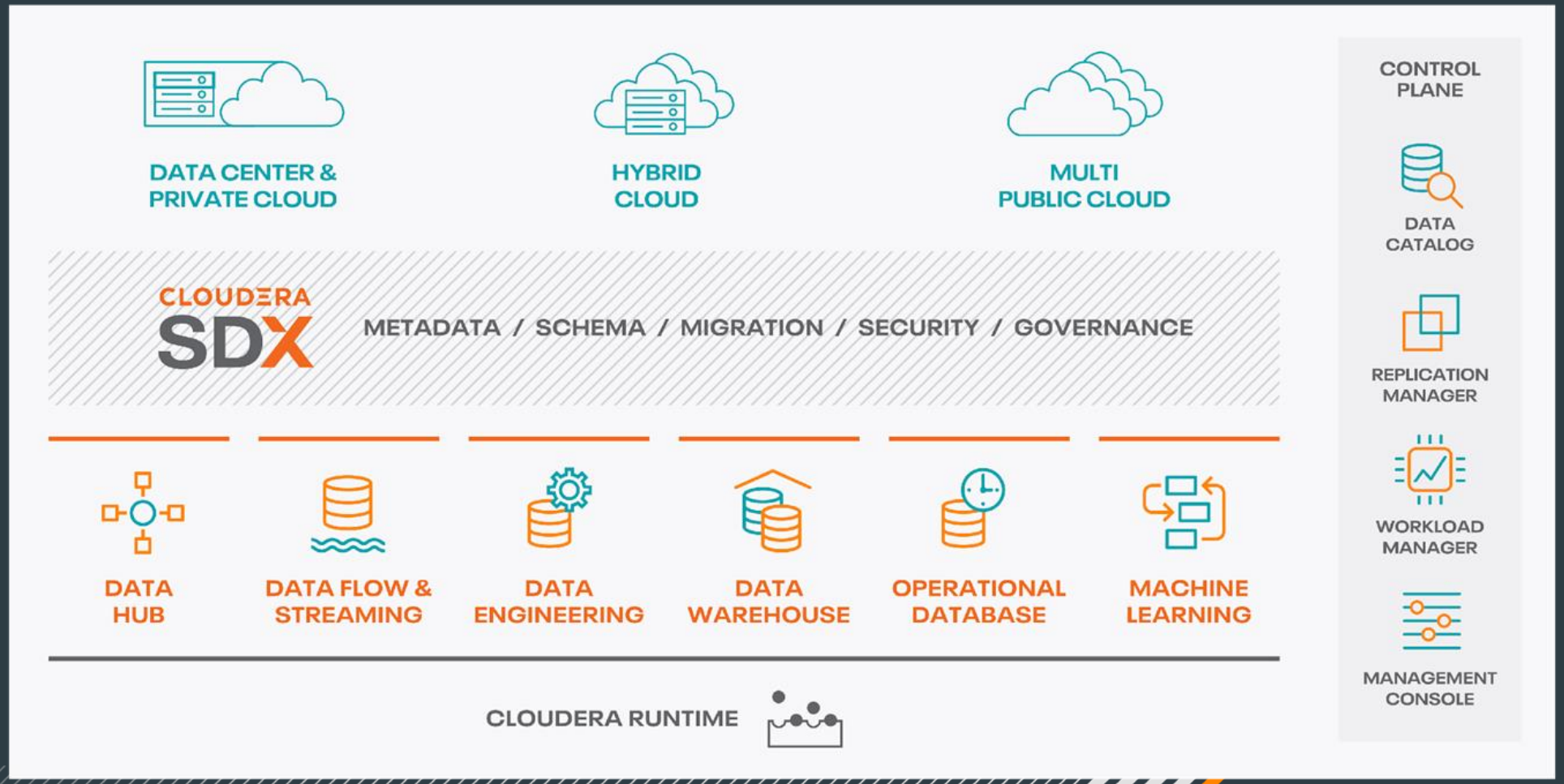
Cloudera Data Flow Workshop

15/09/2020



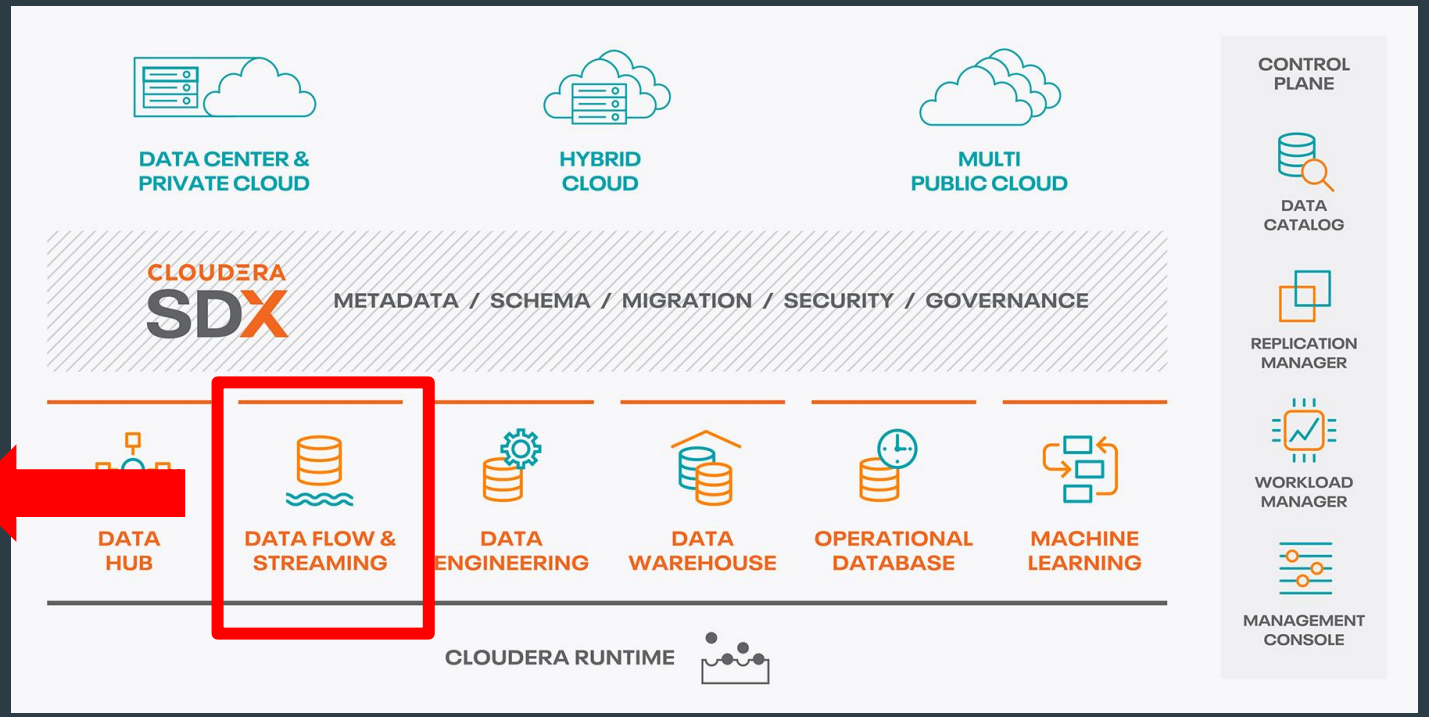

CLOUDERA DATA PLATFORM

全球第一个企业数据云



Cloudera DataFlow (CDF)

- Cloudera Flow



Cloudera DataFlow (CDF)

CDF IN Action 整体架构

MiNiFi

终端数据收集

US-West Fleet

C++ agent minifi

US-Central Fleet

C++ agent minifi

US-East Fleet

C++ agent minifi

Apache Kafka

数据中转平台

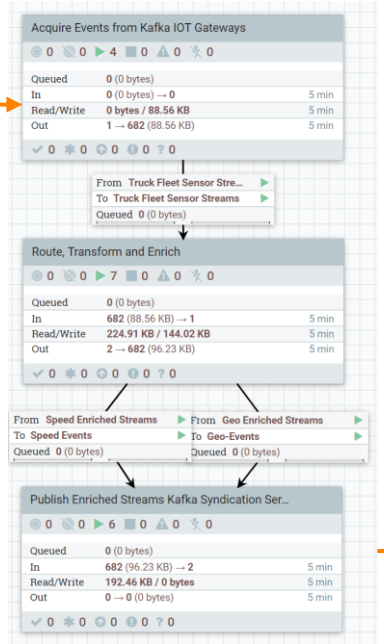
gateway-west-raw-sensors

gateway-central-raw-sensors

gateway-east-raw-sensors

Apache NiFi

数据链路管道



Apache Kafka

数据合并

Kafka Topic syndicate-transmission

Kafka Topic syndicate-speed

Kafka Topic syndicate-temp

Kafka Topic syndicate-geo

Kafka Topic syndicate-oil

Kafka Topic syndicate-breaks

Kafka Topic syndicate-battery

Kafka Topic syndicate-start/stop

Kafka Topic syndicate-acceleration

Kafka Topic syndicate-idle

Apache Flink

数据订阅和实时处理系统

PROCESSING APP 1

Apache Flink

PROCESSING APP 2

kafka

PROCESSING APP 3

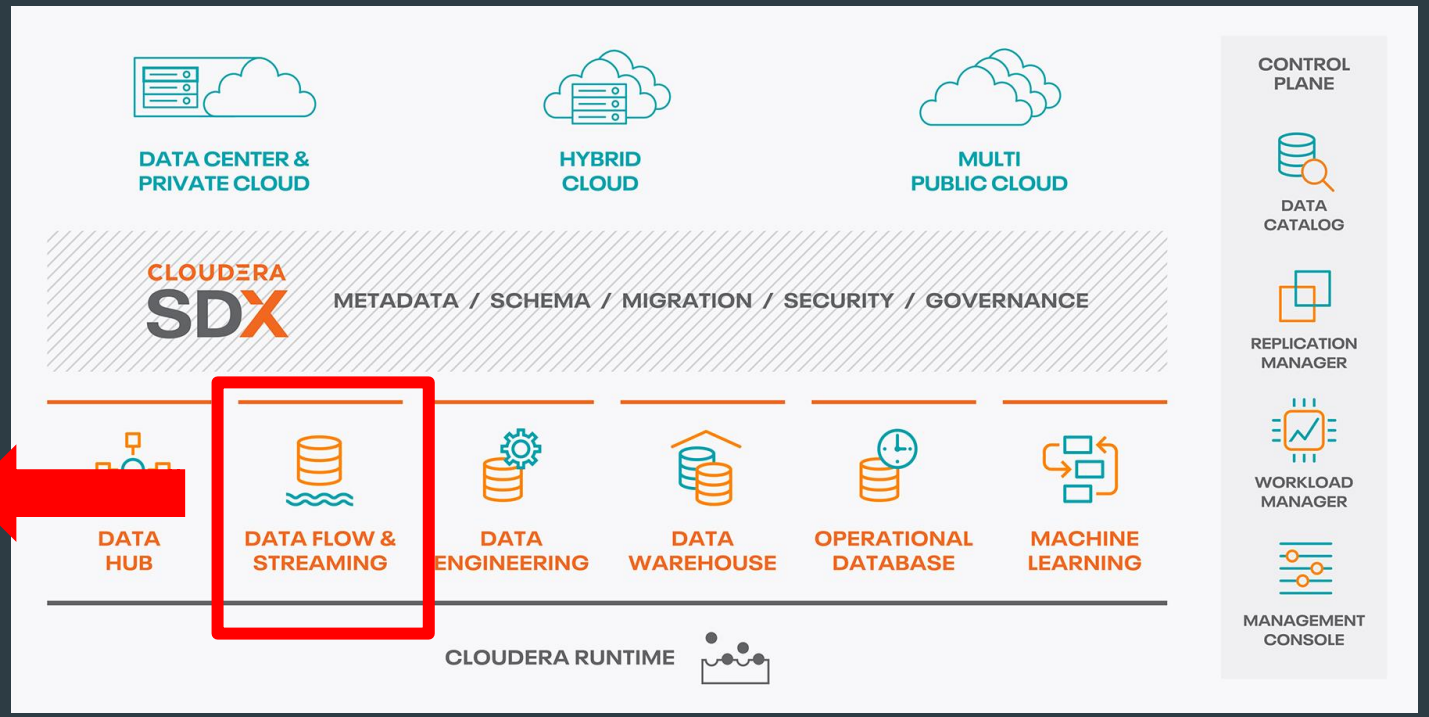
Apache Spark Structured Streaming

Cloudera Flow Management (CFM)

Apache NIFI 介绍

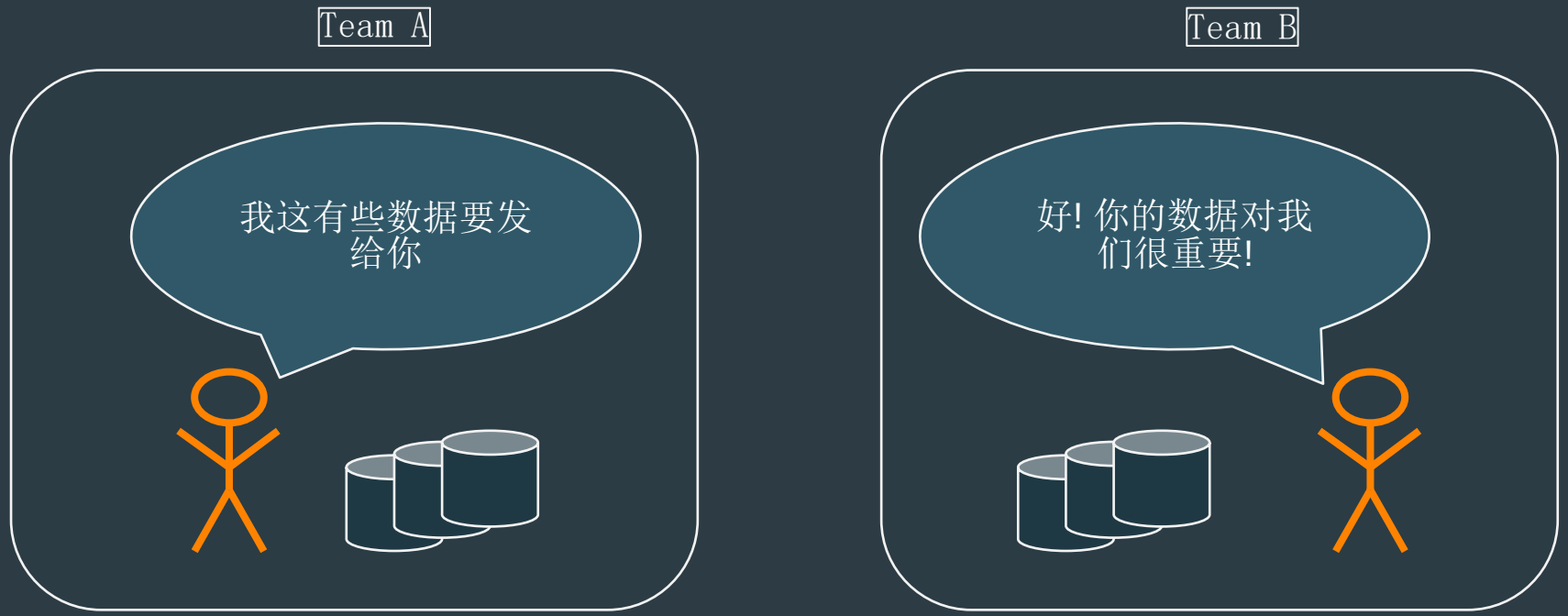
Cloudera DataFlow (CDF)

- Cloudera Flow



CFM

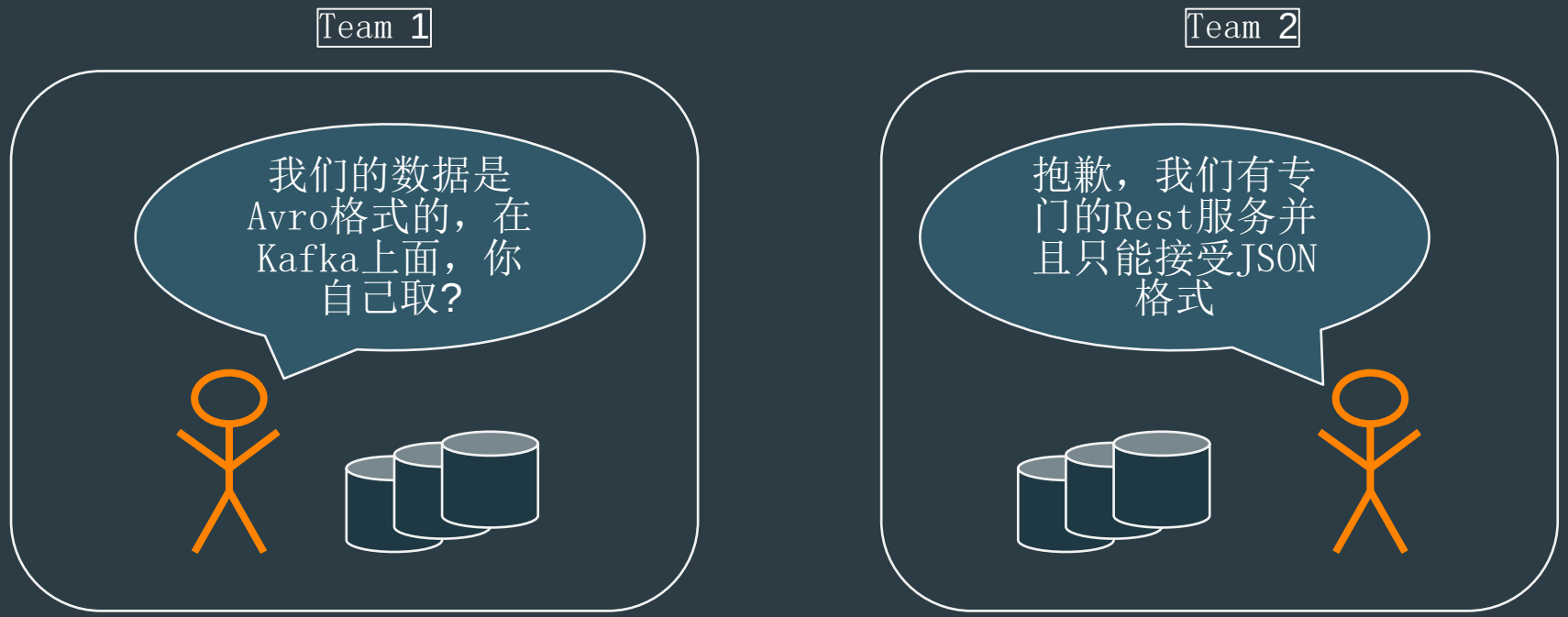
NIFI适用的场景



开发：很常见吧? 也应该很简单...

CFM

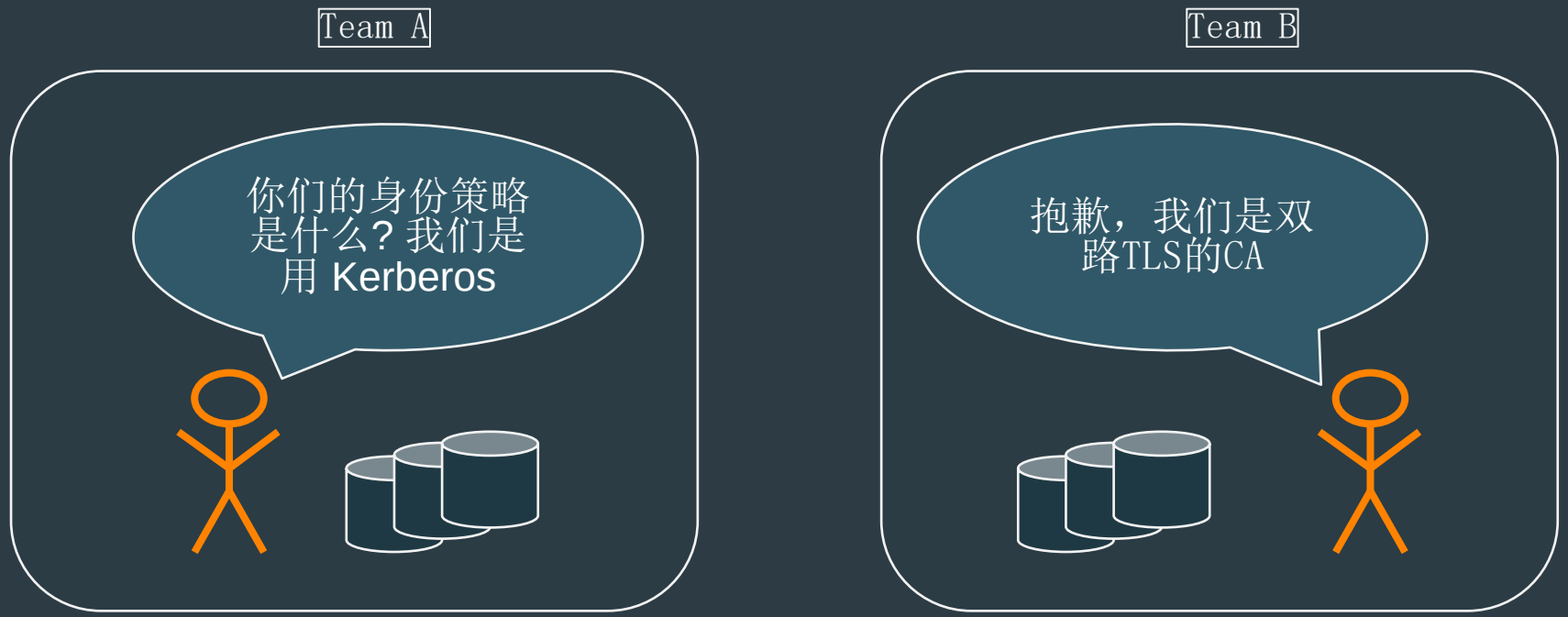
NIFI适用的场景



开发：写个程序，接Kafka，转JSON送给Rest

CFM

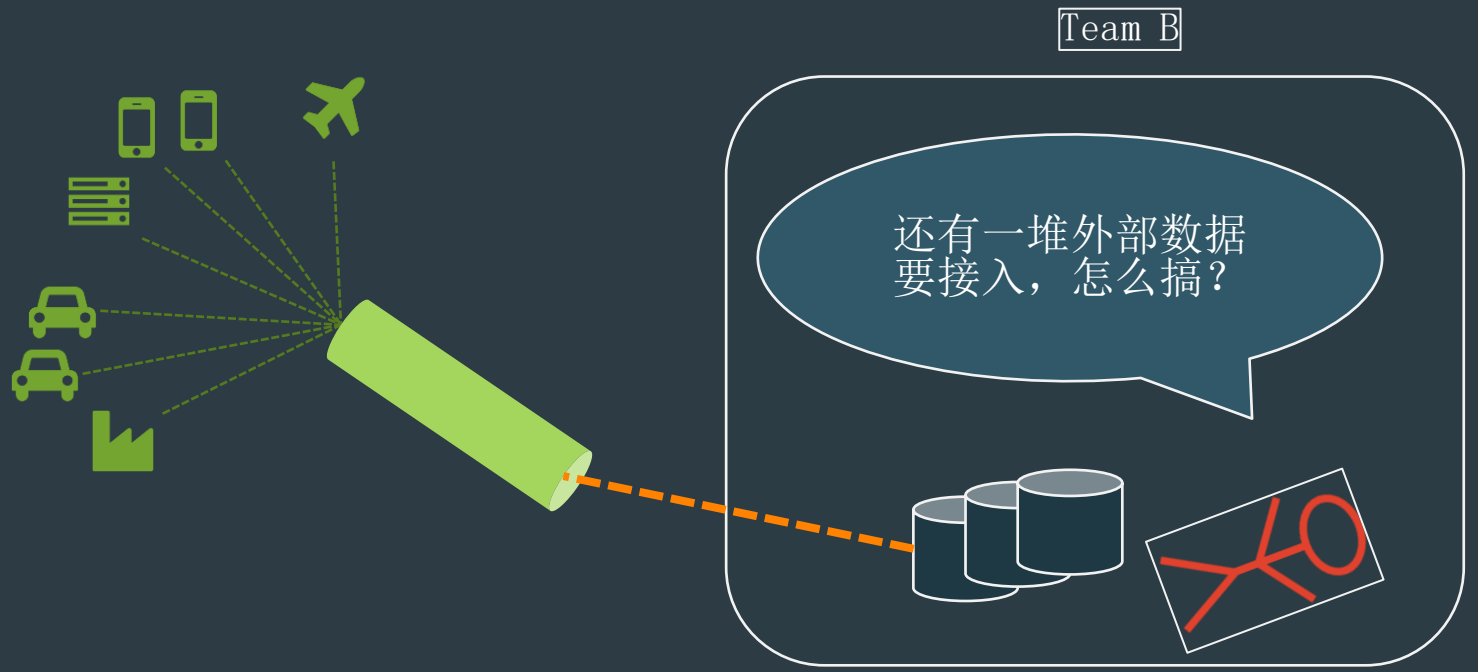
NIFI



开发：...加个功能，即支持Kerberos，还支持CA

CFM

NIFI



开发：每种写一套！

CFM

NIFI

运维部的同事?

可配置，可管理，可监控，自动告警，负载平衡的功能。

CIO?

云原生，基于人工智能的自动化运维

CFM

NIFI

产品经理?

需求变更

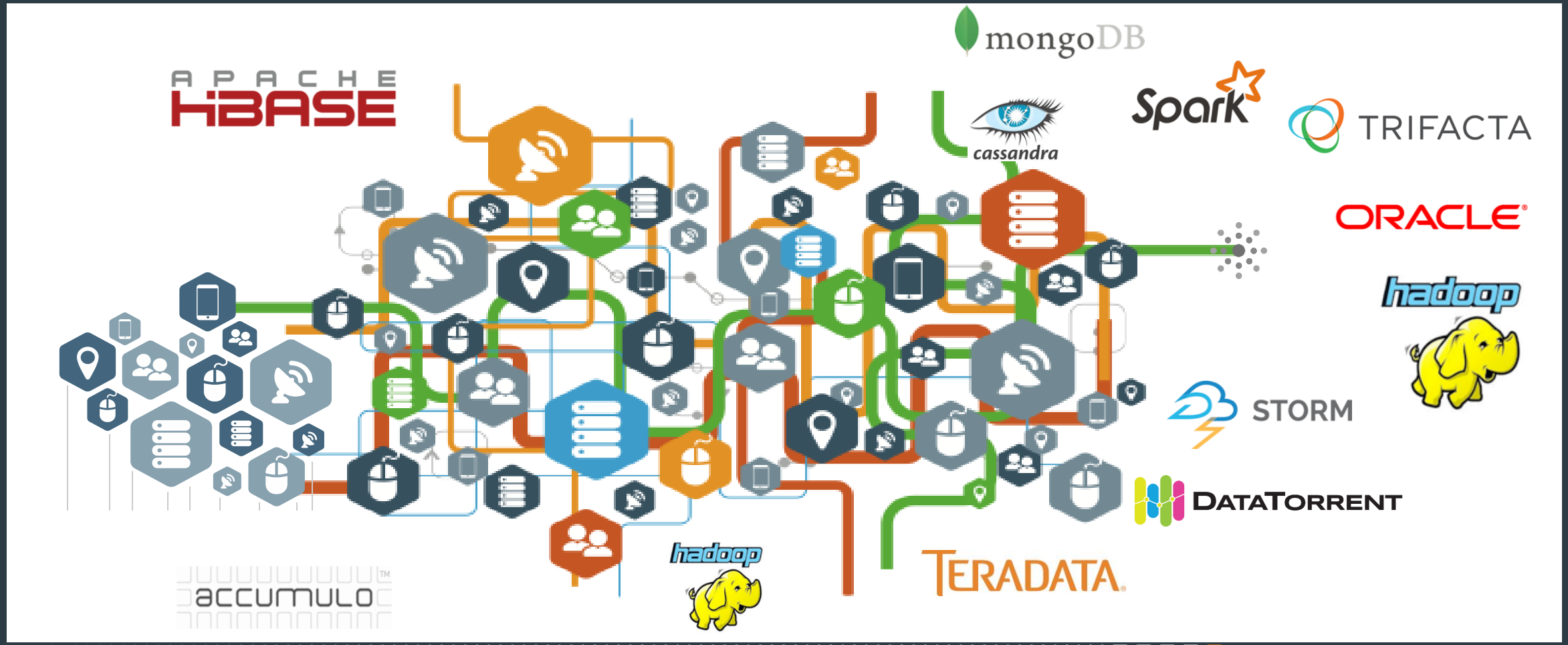
CFM

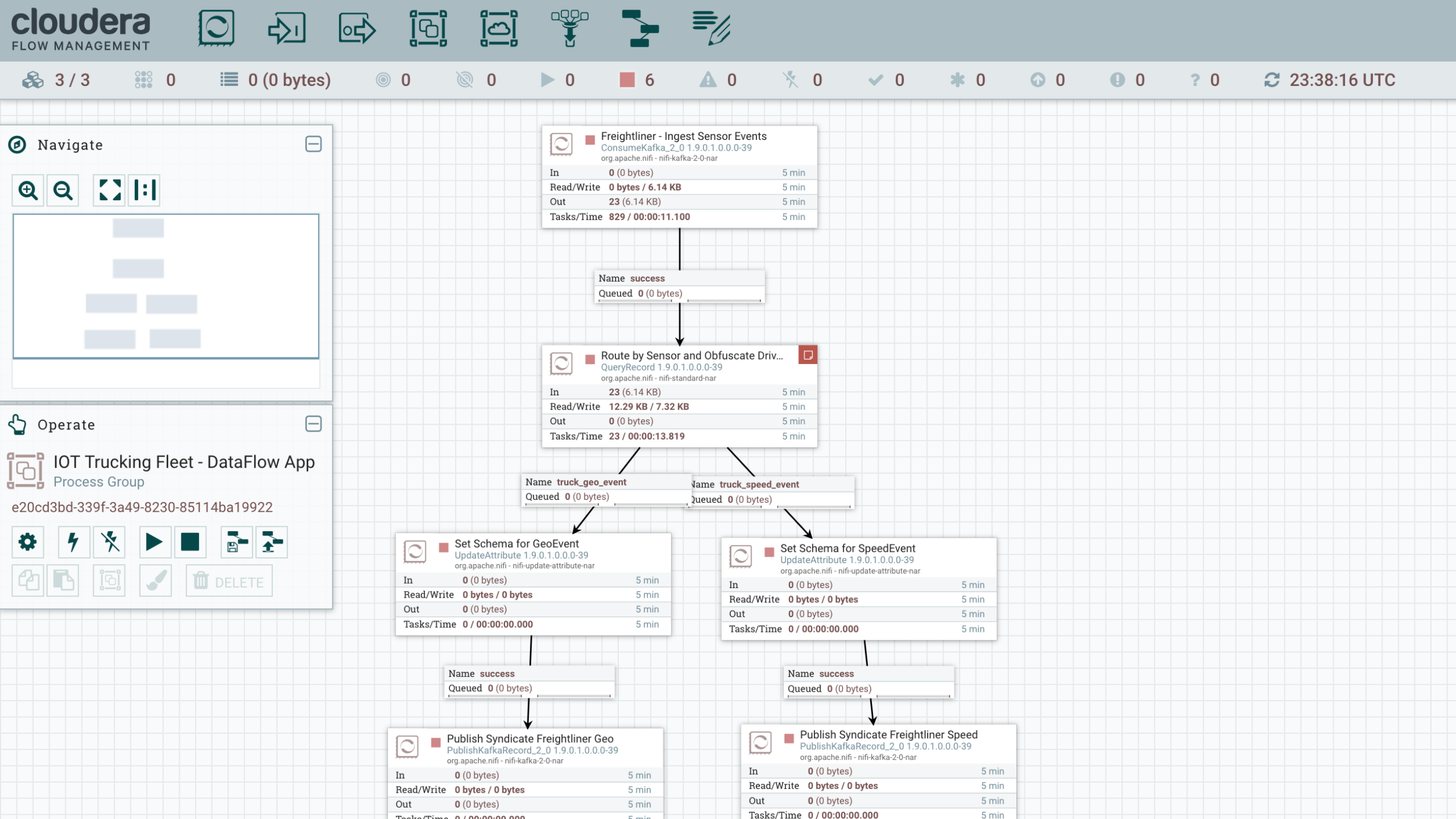
NIFI

自己做真的很简单么?

如此下去系统实际情况可能是这样的

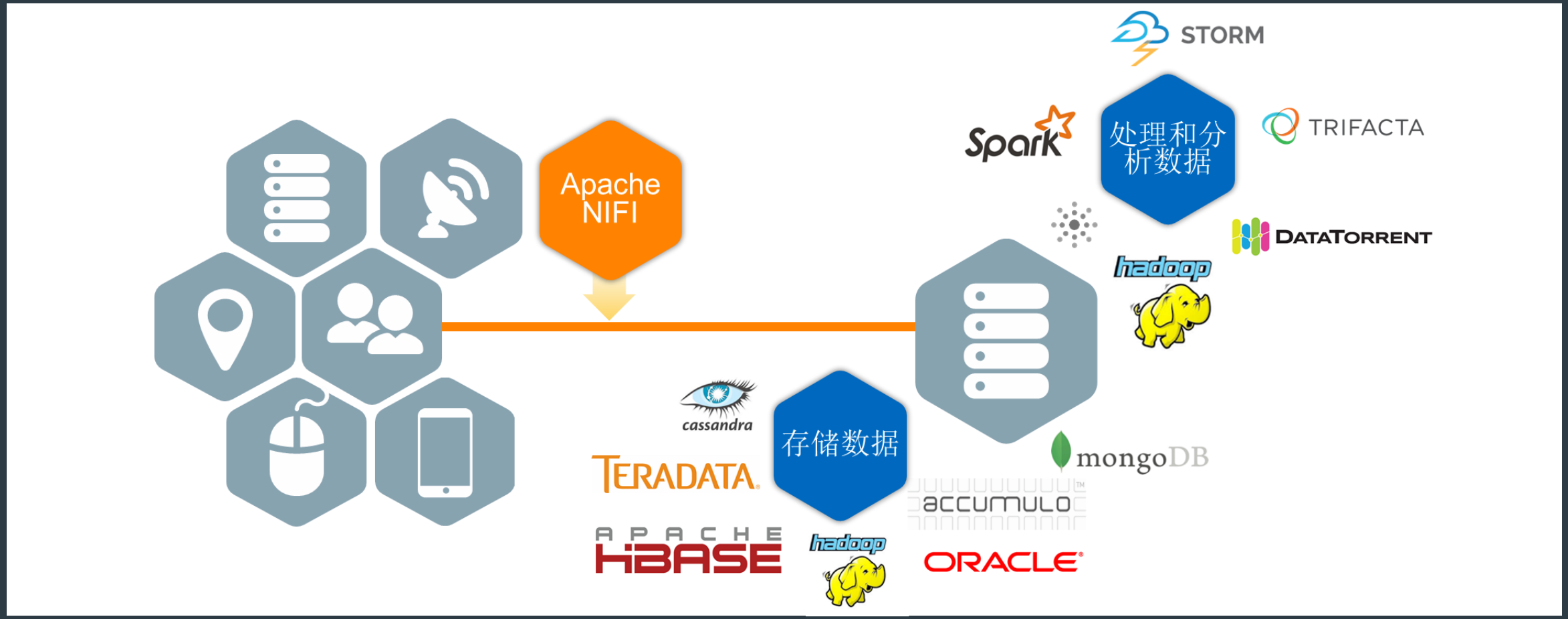
NIFI




cloudera
FLOW MANAGEMENT
3 / 3
0
0 (0 bytes)
0
0
0
6
0
0
0
0
0
0
23:38:16 UTC
Navigate
Operate
IOT Trucking Fleet - DataFlow App
Process Group
e20cd3bd-339f-3a49-8230-85114ba19922
DELETE
Freightliner - Ingest Sensor Events
ConsumeKafka_2_0 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-kafka-2-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 6.14 KB 5 min
Out 23 (6.14 KB) 5 min
Tasks/Time 829 / 00:00:11.100 5 min
Name success
Queued 0 (0 bytes)
Route by Sensor and Obfuscate Driv...
QueryRecord 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-standard-nar
In 23 (6.14 KB) 5 min
Read/Write 12.29 KB / 7.32 KB 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 23 / 00:00:13.819 5 min
Name truck_geo_event
Queued 0 (0 bytes)
Name truck_speed_event
Queued 0 (0 bytes)
Set Schema for GeoEvent
UpdateAttribute 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Set Schema for SpeedEvent
UpdateAttribute 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min
Name success
Queued 0 (0 bytes)
Name success
Queued 0 (0 bytes)
Publish Syndicate Freightliner Geo
PublishKafkaRecord_2_0 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-kafka-2-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Publish Syndicate Freightliner Speed
PublishKafkaRecord_2_0 1.9.0.1.0.0.0-39
org.apache.nifi - nifi-kafka-2-0-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

CEM - NIFI 要解决的问题

NIFI要解决的问题



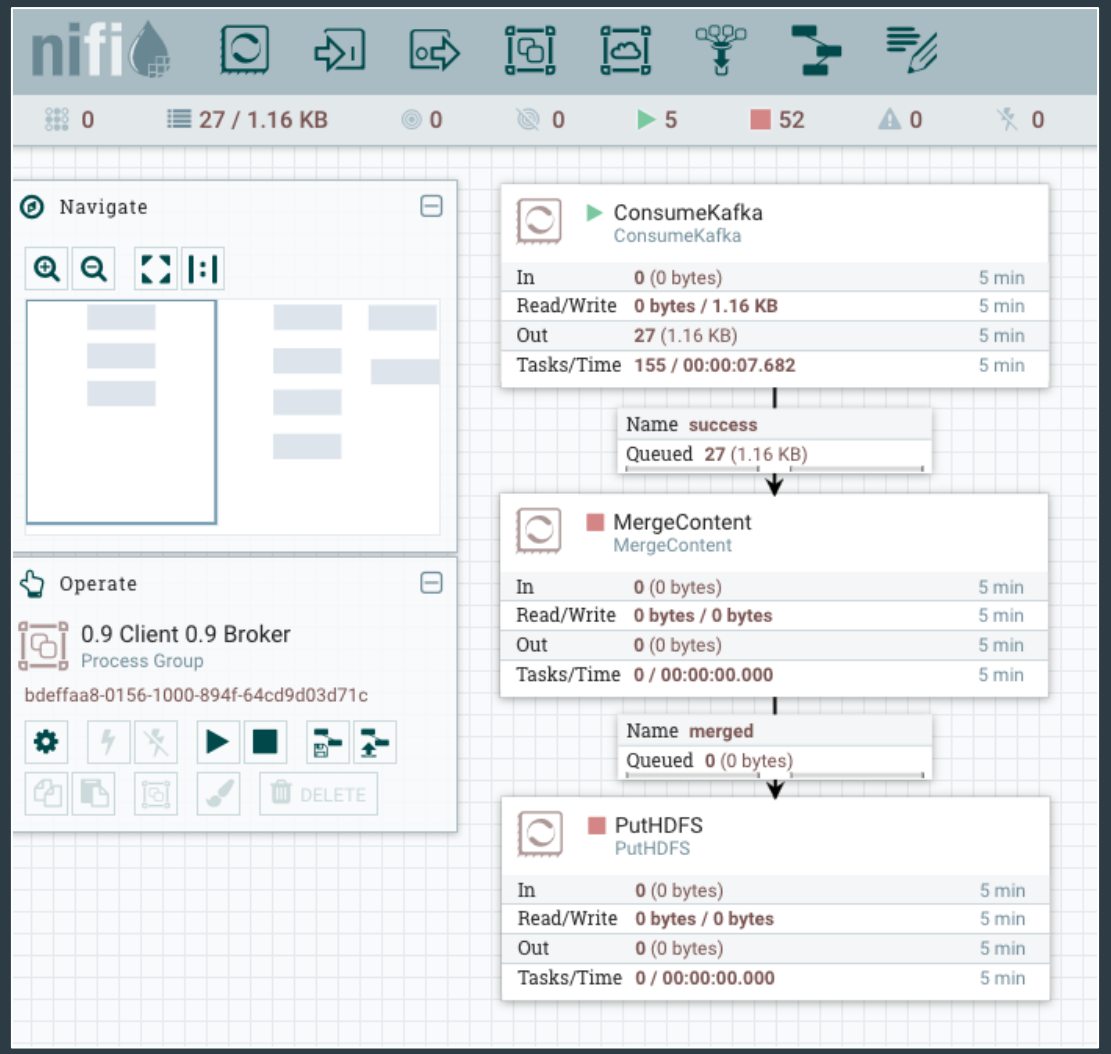
Cloudera Flow Management (CFM)

Apache NIFI 产品特点

CFM

NIFI-功能特点

- 1. 面向数据抓取,转换及推送。
- 2. 轻量化过滤分发。
- 3. 端到端的数据血缘。
- 4. 企业级数据流管理。



CFM

NIFI-功能特点

- 面向数据抓取，转换及推送。
  - 数据抓取

List	Get		Fetch	Listen	Execute	Query	Consume
AzureBlobStorage DatabaseTables File FTP GCSBucket HDFS S3 SFTP	AzureEventHub AzureQueueStorage CouchbaseKey DynamoDB File FTP HBase HDFS HDFSEvents HDFSFileInfo HDFSSequenceFile	HTMLElement HTTP IgniteCache JMSQueue JMSTopic Kafka Mongo RethinkDB SFTP SNMP Solr Splunk SQS TCP Twitter	AzureBlobStorage DistributedMapCache Elasticsearch Elasticsearch5 ElasticsearchHttp File FTP GCSObject HBaseRow HDFS Parquet S3Object SFTP	Beats GRPC HTTP Lumberjack RELP SMTP Syslog TCP TCPRecord UDP UDPRecord WebSocket	FlumeSource GroovyScript InfluxDBQuery Process Script SparkInteractive SQL SQLRecord StreamCommand	Cassandra DatabaseTable DatabaseTableRecord DNS ElasticsearchHttp Record Solr Whois	AMQP AzureEventHub EWS GCPubSub IMAP JMS Kafka Kafka_0_10 Kafka_0_11 Kafka_1_0 Kafka_2_0 KafkaRecord_0_10 KafkaRecord_0_11 KafkaRecord_1_0 KafkaRecord_2_0 MQTT POP3 WindowsEventLog

CFM

NIFI-功能特点

- 面向数据抓取，转换及推送。
  - 转换

Convert	Transform	Split	Parse	Others
ConvertAvroSchema ConvertAvroToJSON ConvertAvroToORC ConvertAvroToParquet ConvertCharacterSet ConvertCSVToAvro ConvertExcelToCSVProcessor ConvertJSONToAvro ConvertJSONToSQL ConvertRecord	JoltTransformJSON JoltTransformRecord TransformXml MergeContent MergeRecord ModifyBytes ModifyHTMLElement	SplitAvro SplitContent SplitJson SplitRecord SplitText SplitXml	ParseCEF ParseEvtx ParseNetflowv5 ParseSyslog ParseSyslog5424 PartitionRecord	CountText CreateHadoopSequenceFile CryptographicHashAttribute CryptographicHashContent HashAttribute HashContent IdentifyMimeType InferAvroSchema LogAttribute LogMessage LookupAttribute LookupRecord MonitorActivity MoveHDFS Notify

CFM

NIFI-功能特点

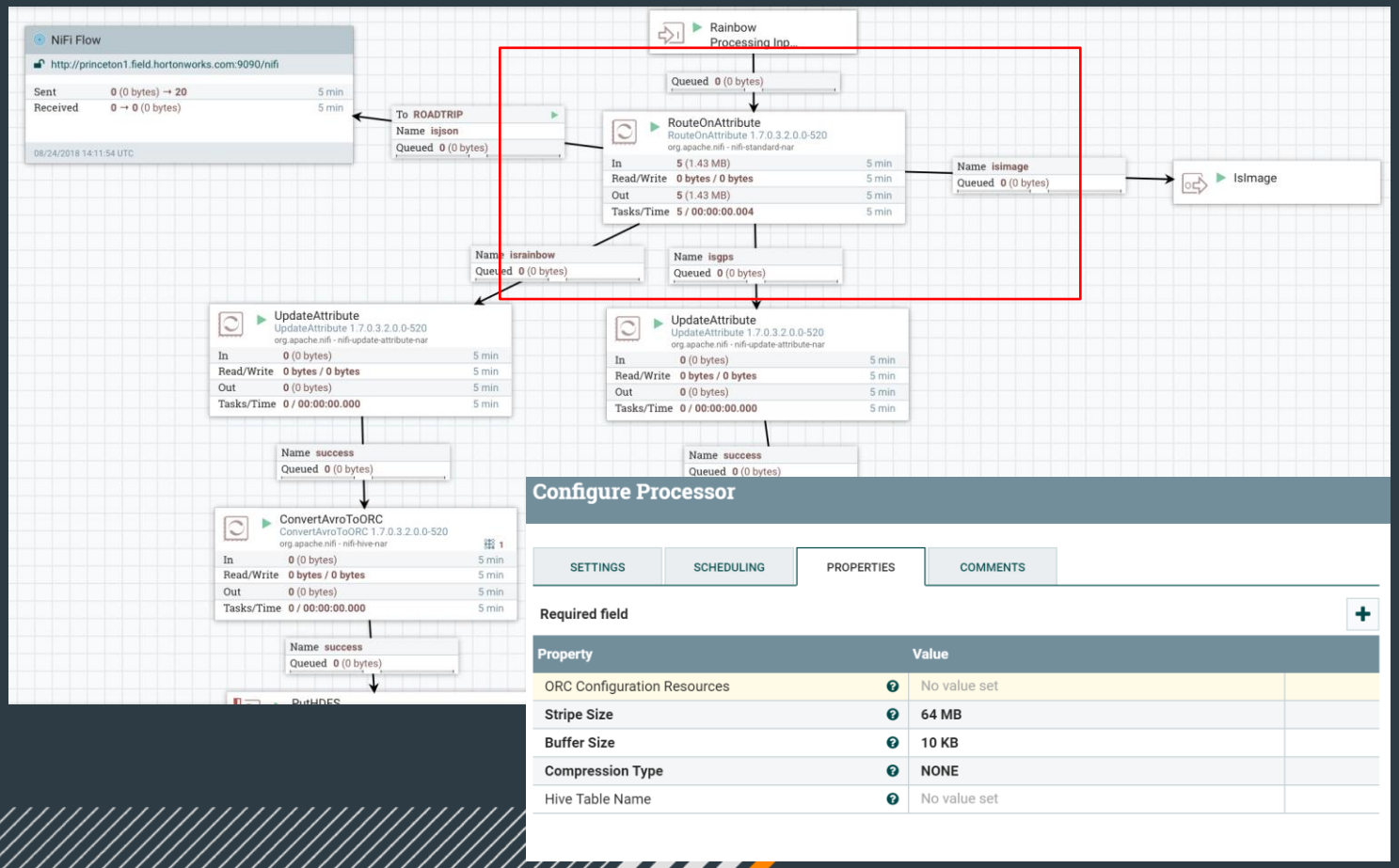
- 面向数据抓取，转换及推送。
  - 推送

Put			Delete	Execute	Publish/Post
SNS SolrContentStream SolrRecord Splunk SQL SQS Syslog TCP UDP WebSocket FTP GCSObject HBaseCell HBaseJSON HBaseRecord HDFS HiveQL	AzureBlobStorage AzureEventHub AzureQueueStorage CassandraQL CassandraRecord CloudWatchMetric CouchbaseKey DatabaseRecord DistributedMapCache DruidRecord DynamoDB Elasticsearch Elasticsearch5 ElasticsearchHttp ElasticsearchHttpRecord Email	HiveStreaming HTMLElement IgniteCache InfluxDB JMS Kafka KinesisFirehose KinesisStream Kudu Lambda Mongo MongoRecord Parquet RethinkDB Riemann S3Object SFTP Slack	AzureBlobStorage ByQueryElasticsearch DynamoDB Elasticsearch5 GCSObject HBaseCells HBaseRow HDFS Mongo RethinkDB S3Object SQS	FlumeSink	AMQP GCPubSub JMS Kafka Kafka_0_10 Kafka_0_11 Kafka_1_0 Kafka_2_0 KafkaRecord_0_10 KafkaRecord_0_11 KafkaRecord_1_0 KafkaRecord_2_0 MQTT HTTP

CFM

NIFI-功能特点

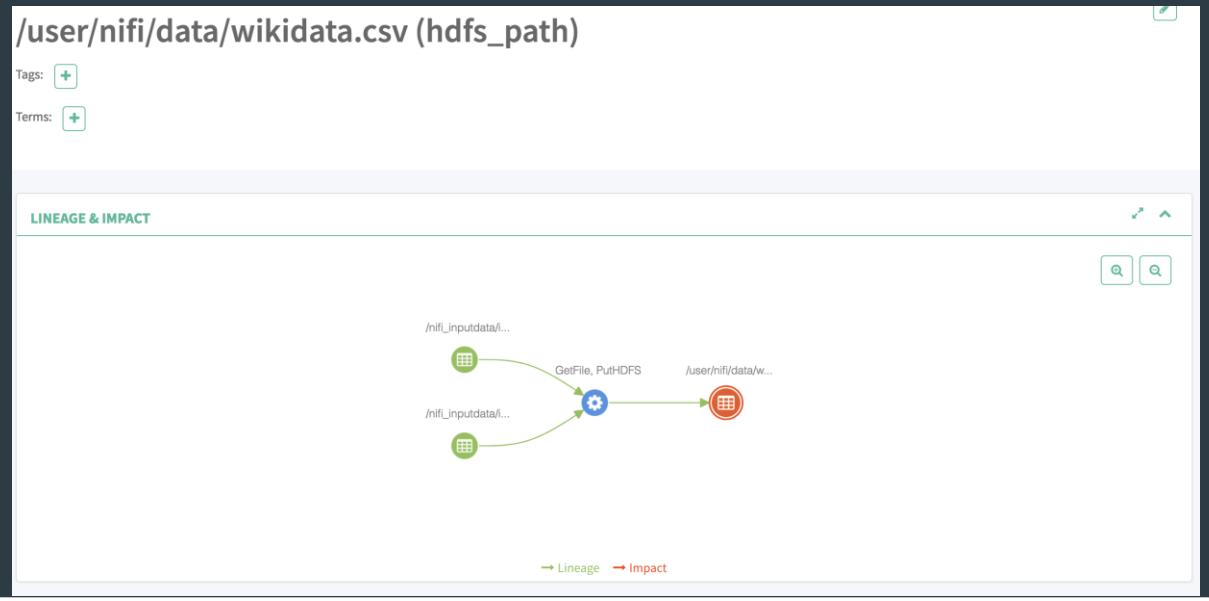
- 2.轻量化过滤分发。
  - 过滤及分发



CFM

NIFI-功能特点

- 3. 端到端的数据血缘。



Displaying 13 of 104

Oldest event available: 11/15/2016 13:34:50 EST

Showing the most recent events.

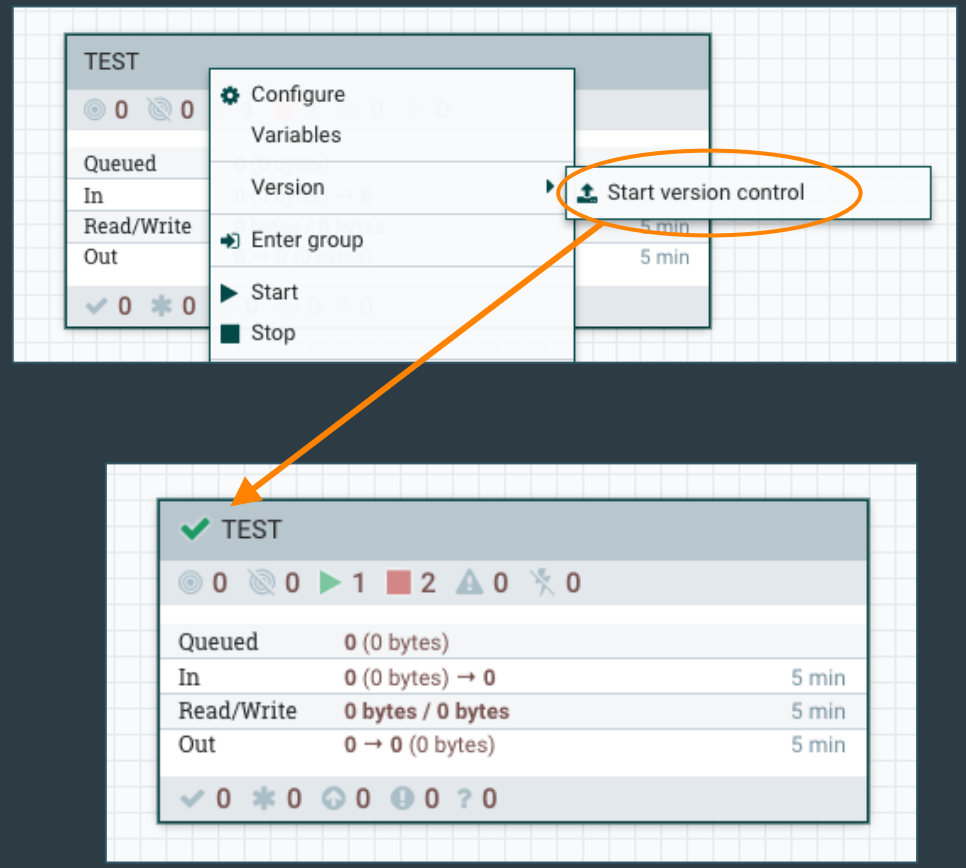
ConsumeKafka | by component name

Date/Time	Type	FlowFile Uuid	Size	Component Name	Component Type
11/15/2016 13:35:03.8...	RECEIVE	379fc4f6-60e0-4151-9743-28...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:02.7...	RECEIVE	78f8c38b-89fc-4d00-a8d8-51...	44 bytes	ConsumeKafka	ConsumeKafka
11/15/2016 13:35:01.6...	RECEIVE	2bcd5124-bb78-489f-ad8a-7...	44 bytes	ConsumeKafka	ConsumeKafka

CFM

NIFI-功能特点

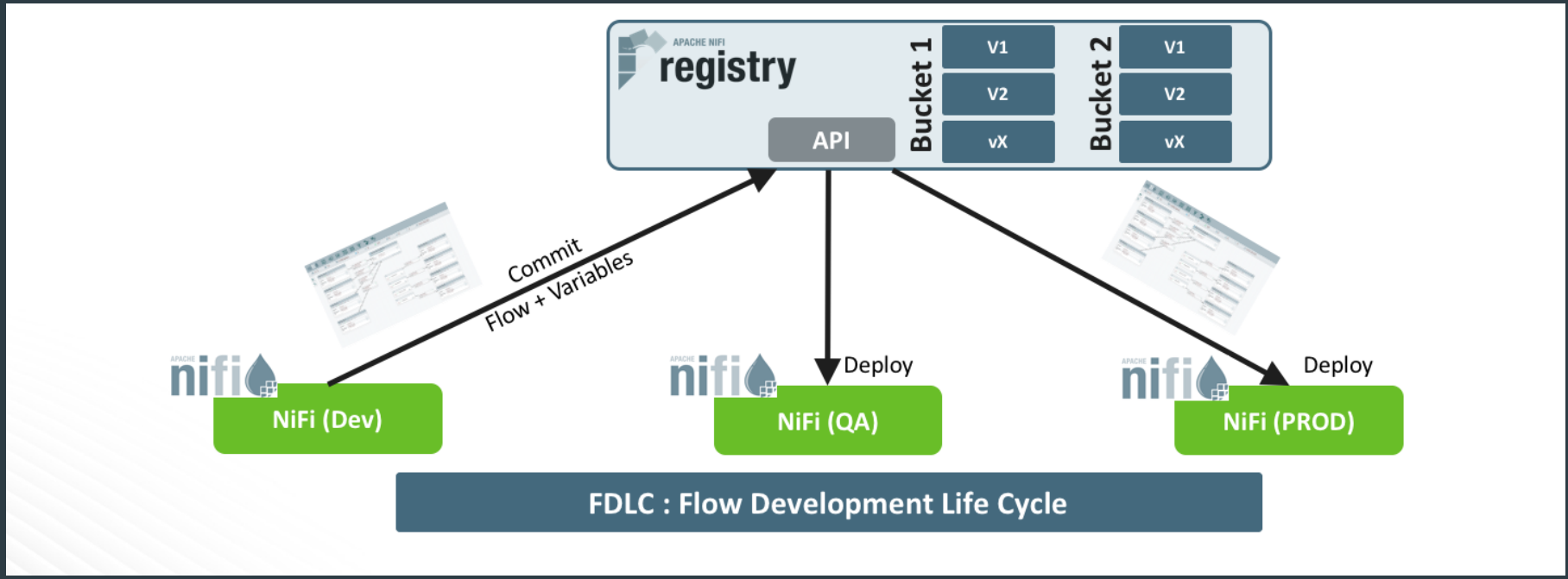
- 4. 企业级数据流管理。
  - 4.1 版本管理



CFM

NIFI-功能特点

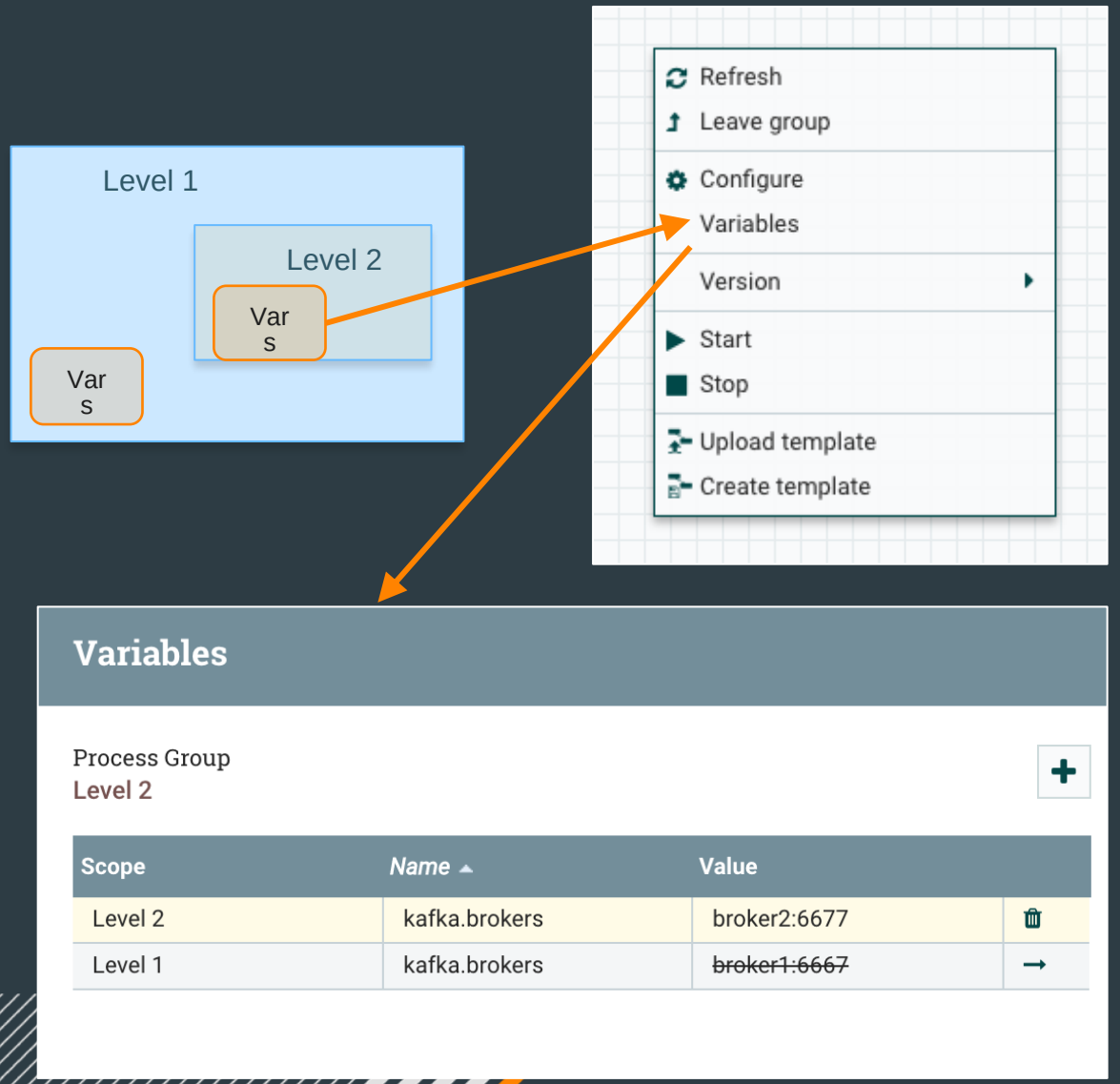
- 4. 企业级数据流管理。
  - 4.1 版本管理



CFM

NIFI-技术特点

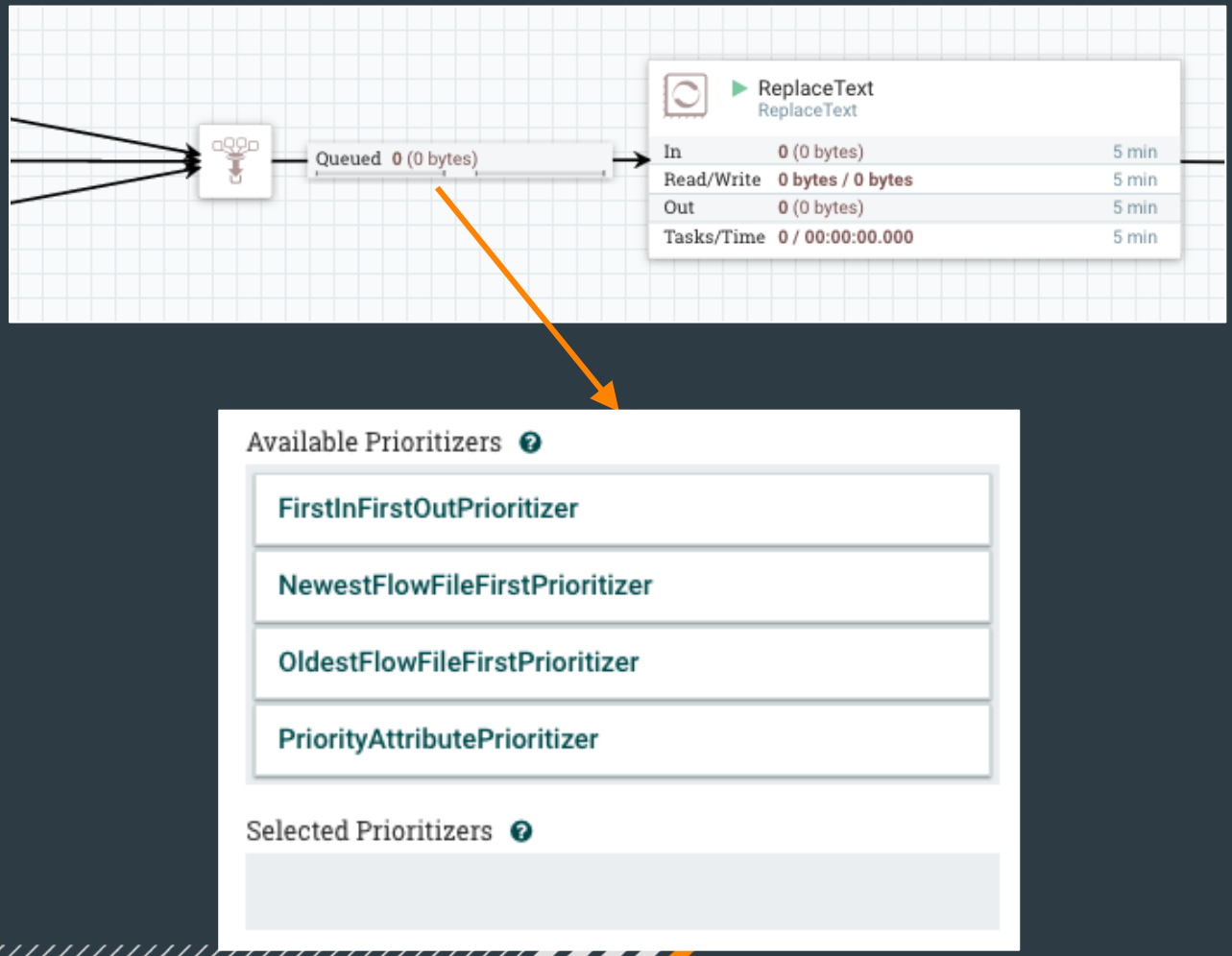
- 4. 企业级数据流管理。
  - 4.2 参数管理



CFM

NIFI-技术特点

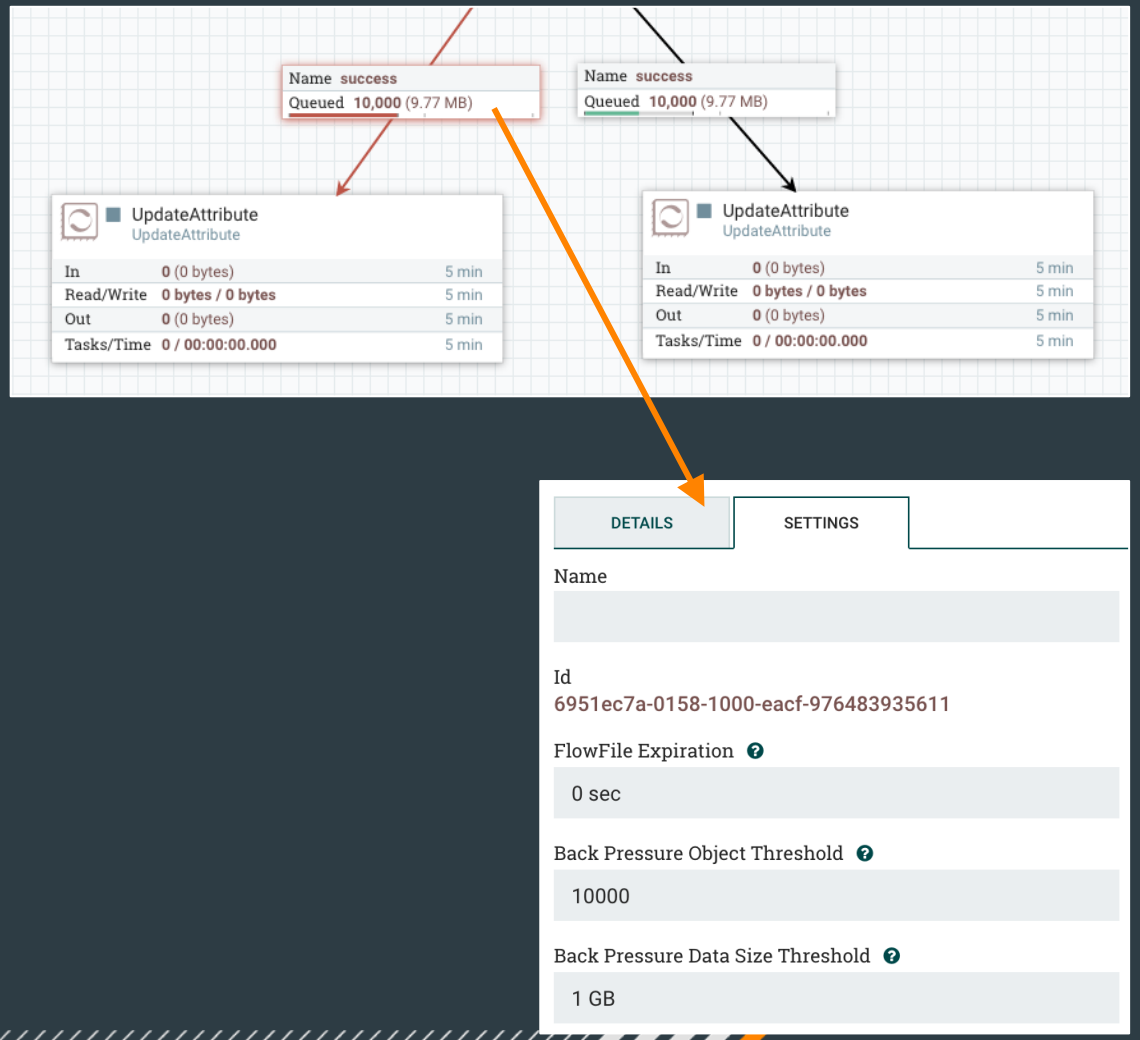
- 4. 企业级数据流管理。
  - 4.3 数据优先级管理



CFM

NIFI-技术特点

- 4. 企业级数据流管理。
  - 4.4 上下游吞吐管理

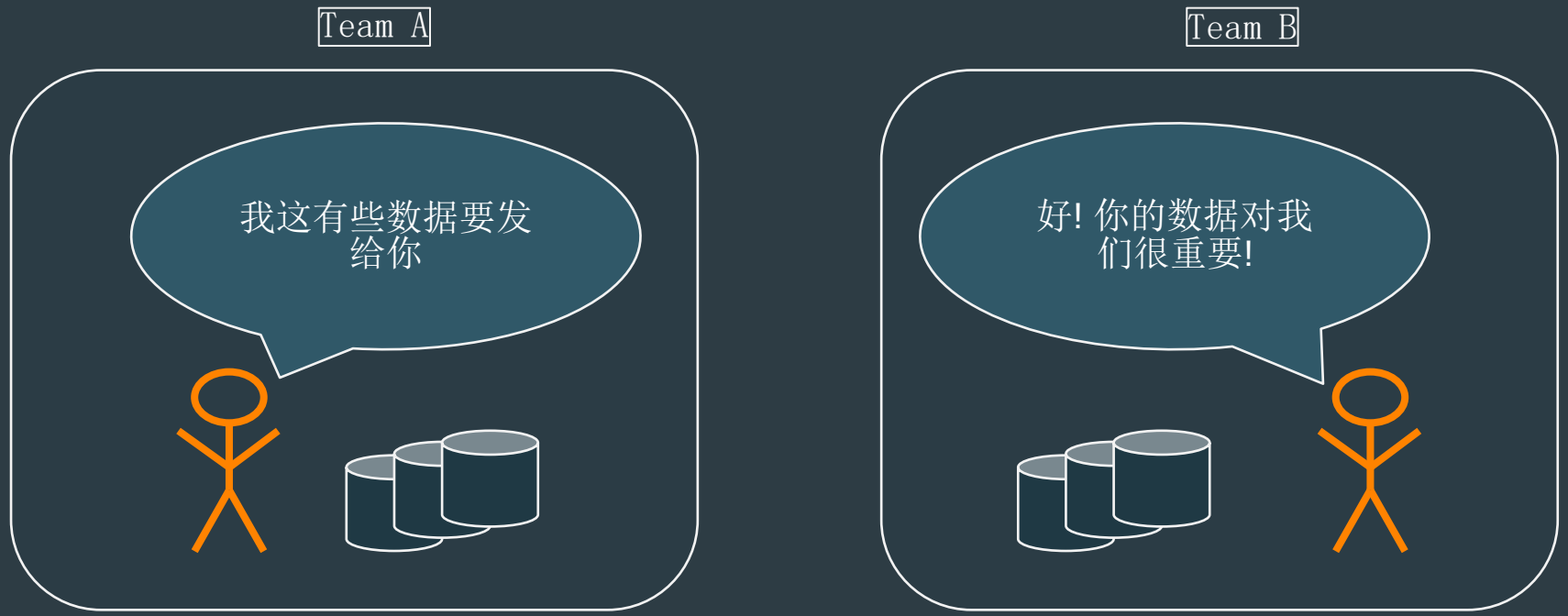


Cloudera Flow Management (CFM)

场景回顾

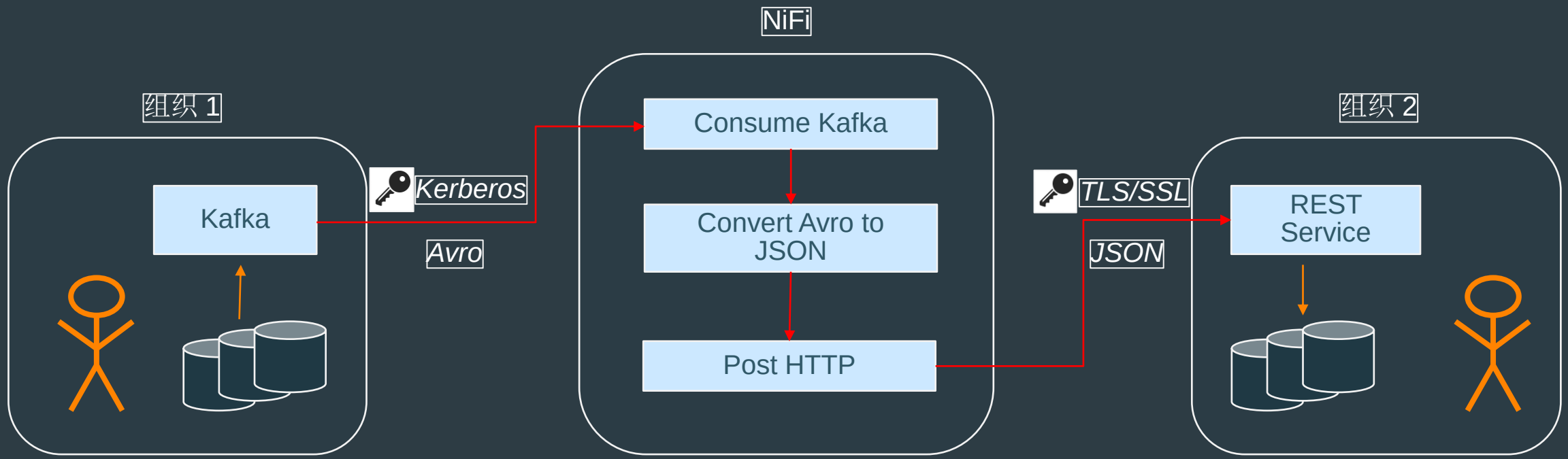
CFM

NIFI适用的场景



CFM

NIFI适用的场景

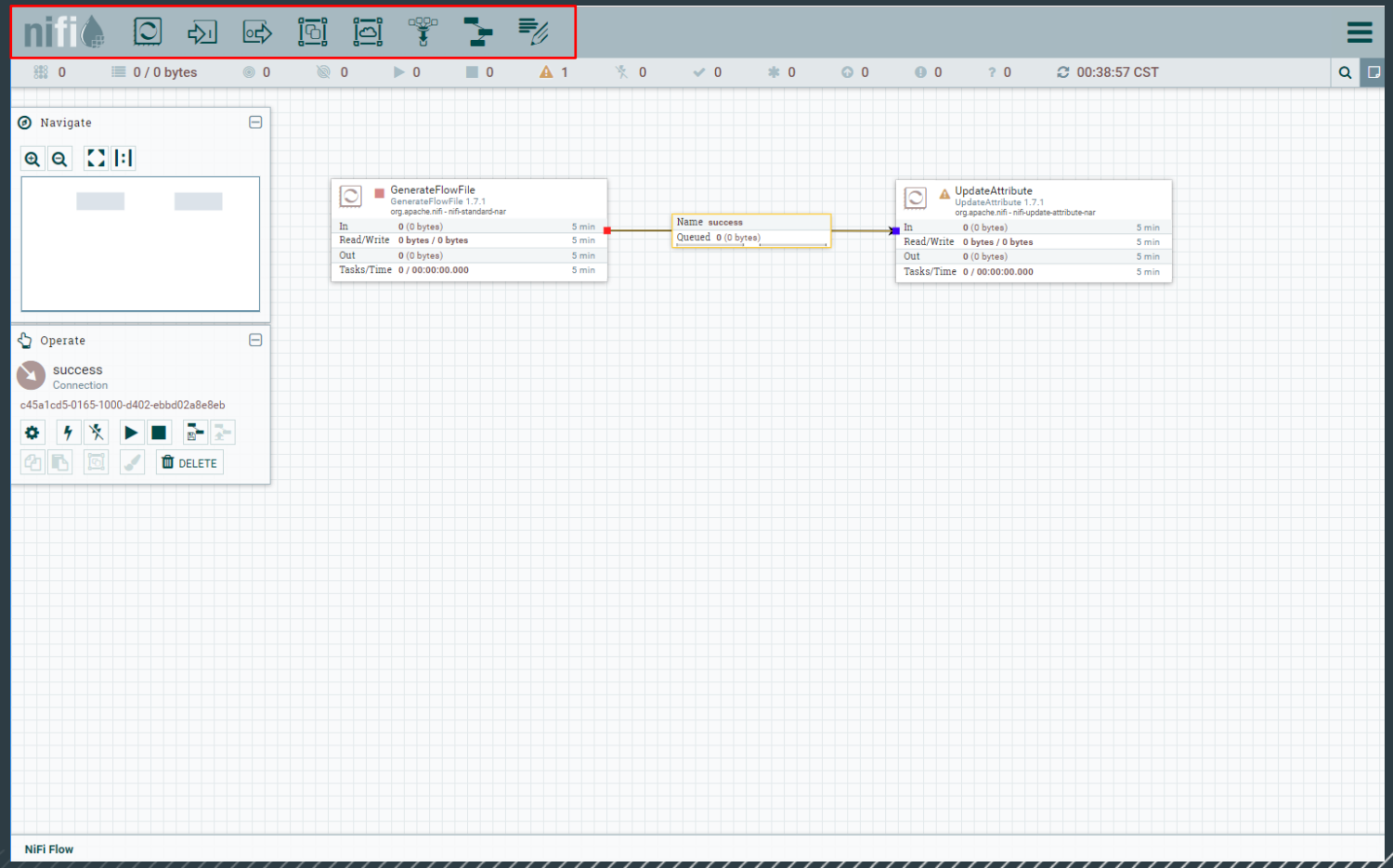


Cloudera Flow Management (CFM)

Apache NIFI UI 介绍

Apache NIFI UI 介绍

常用工具栏

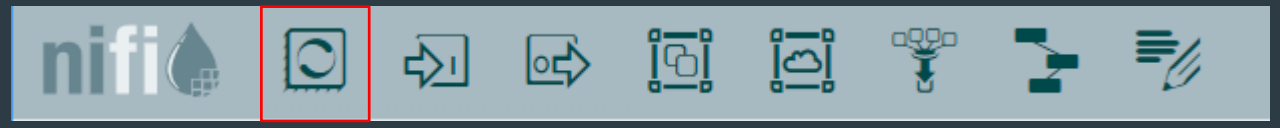


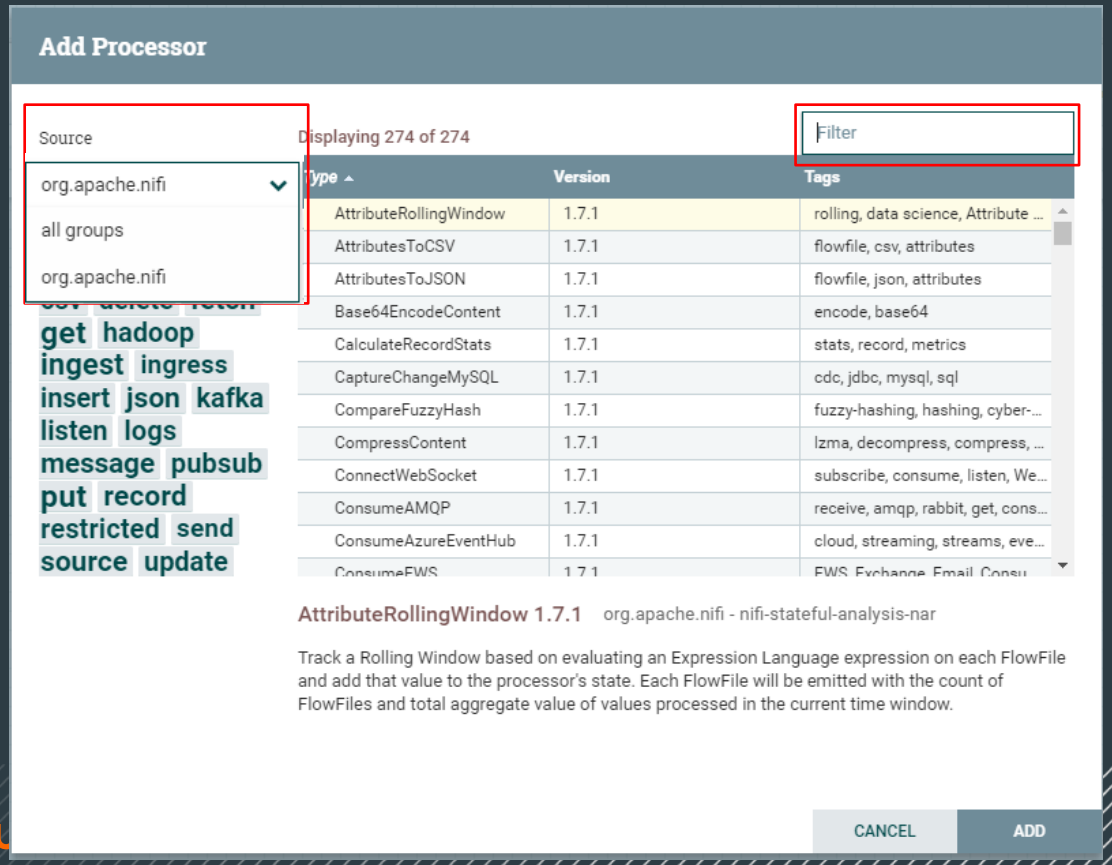
常用工具栏

包含常用的快捷方式

Apache NIFI UI 介绍

常用工具栏 – 添加处理器



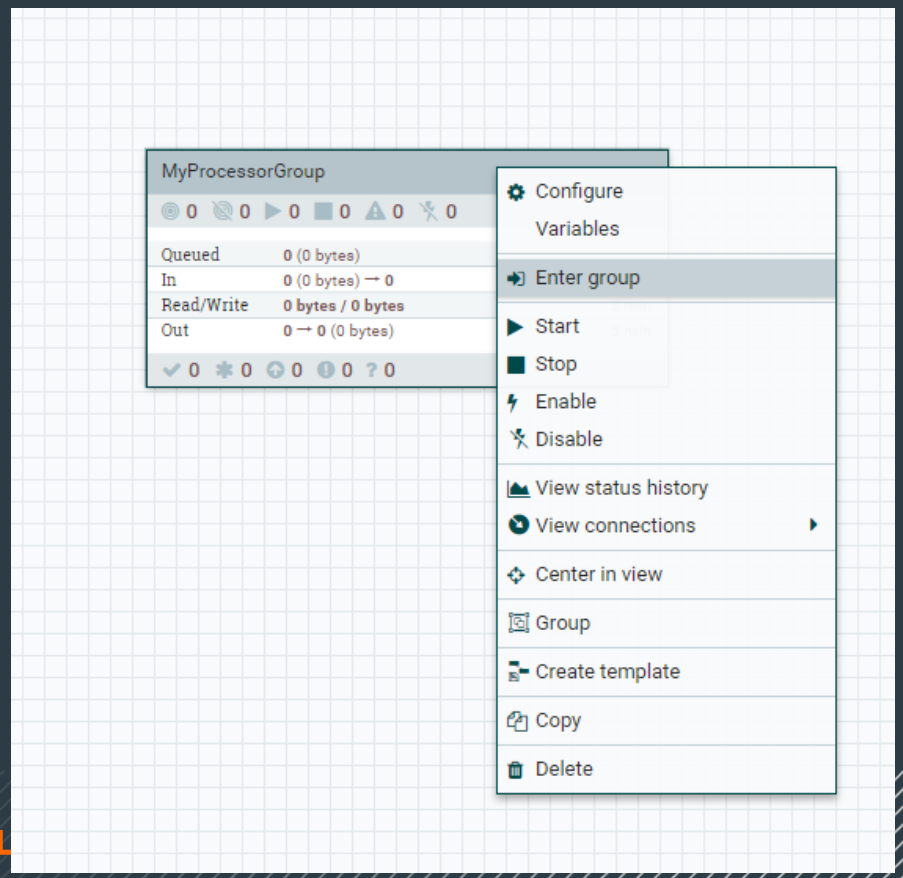


来源过滤-按处理器组属性过滤

名称过滤-按处理器名称过滤

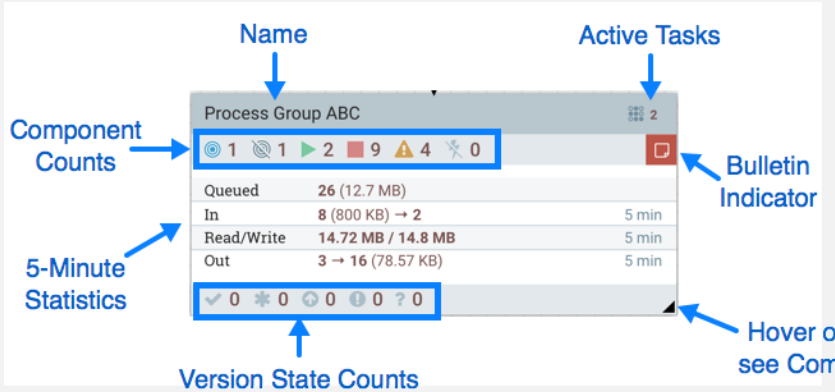
Apache NIFI UI 介绍

常用工具栏 -添加处理器组



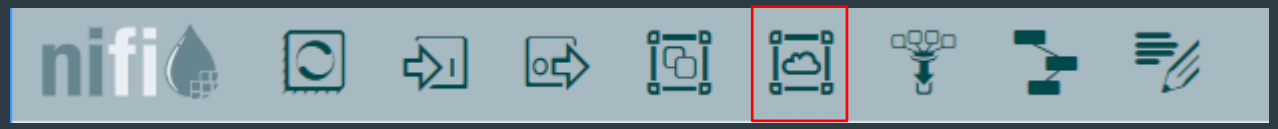
添加处理器组

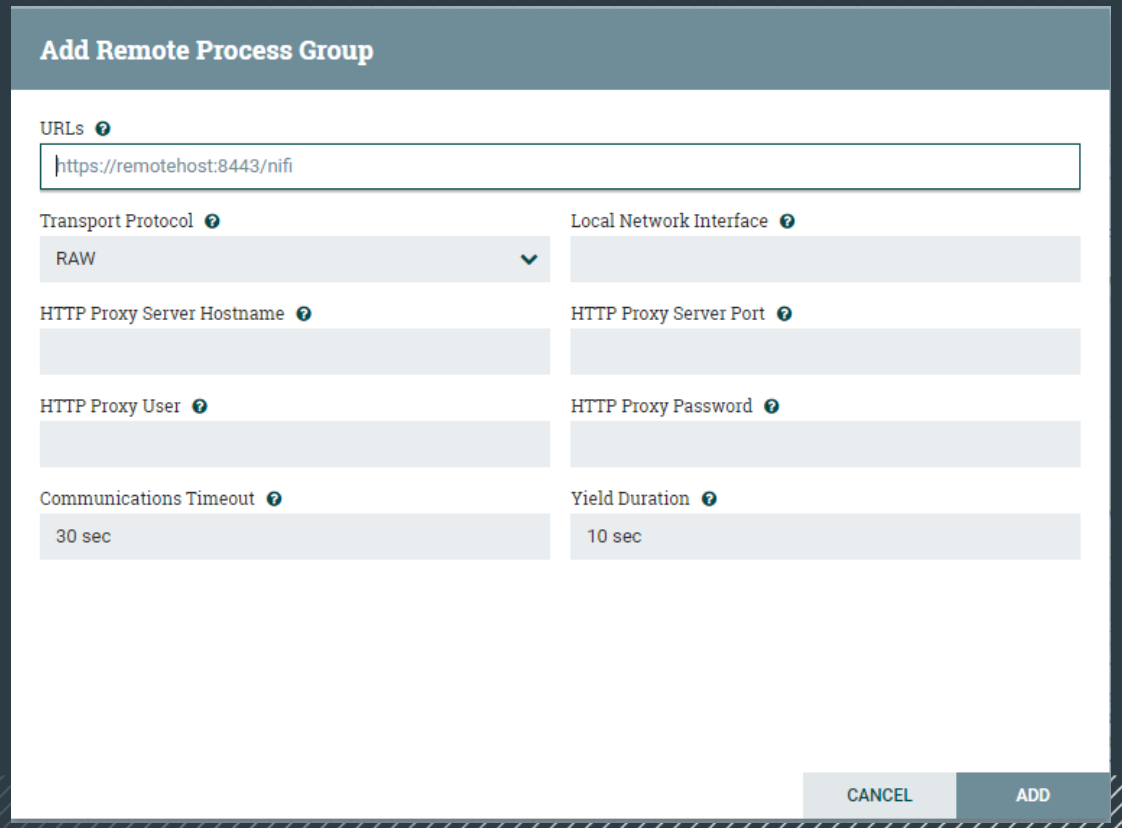
处理器组可以包含一批处理器来完成某一个任务，同时还可以批量对组进行一些控制



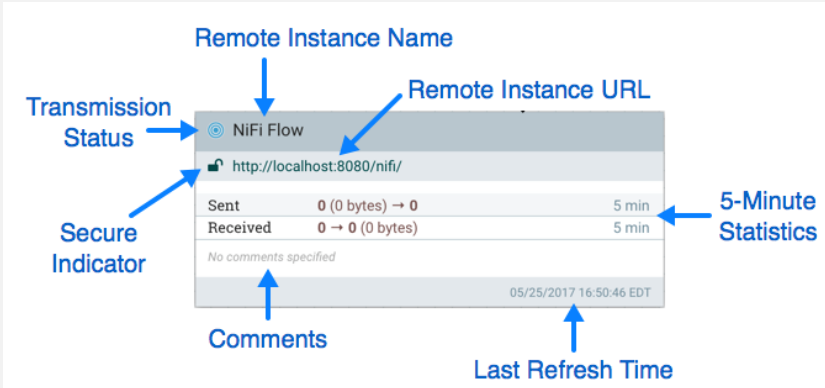
Apache NIFI UI 介绍

常用工具栏 – 添加远程处理器组



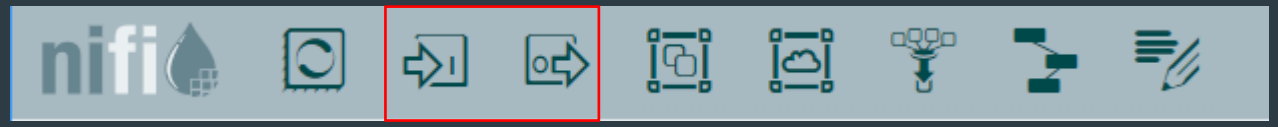


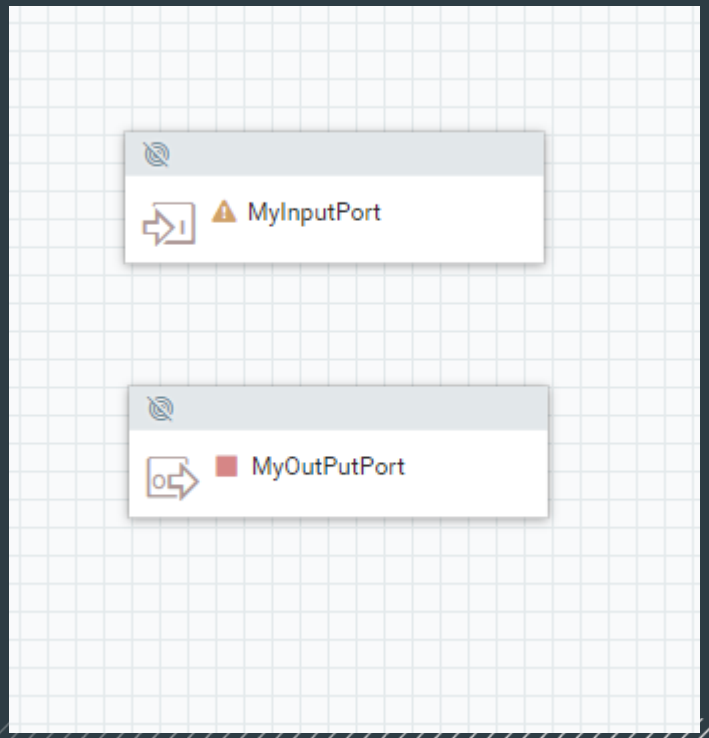
添加远程处理器组



Apache NIFI UI 介绍

常用工具栏 – 添加输入/输出端口





输入端口

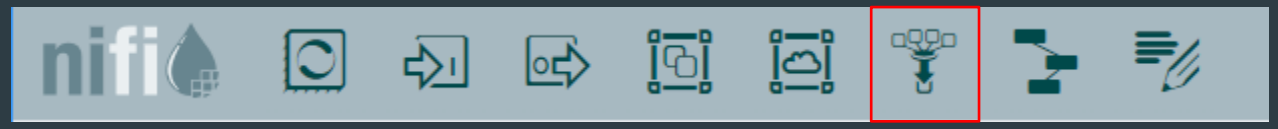
为处理器组提供了数据接入的接口

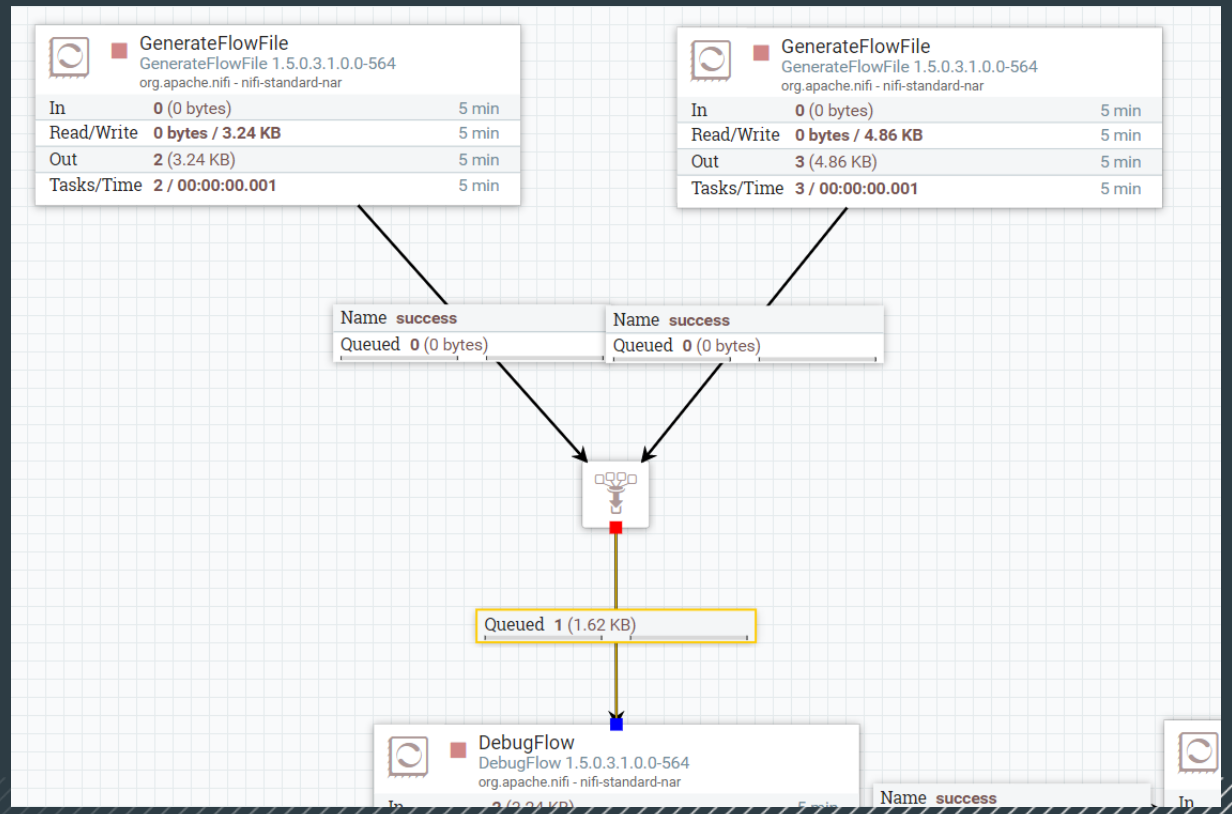
输出端口

为处理器组提供了数据外发的接口

Apache NIFI UI 介绍

常用工具栏 – 添加汇聚器



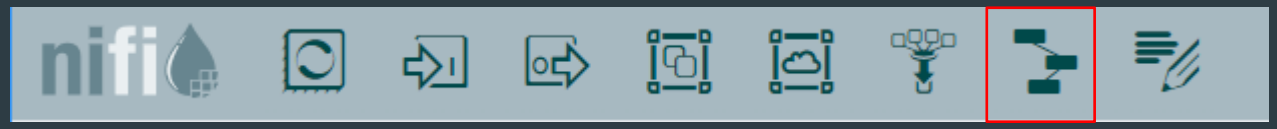


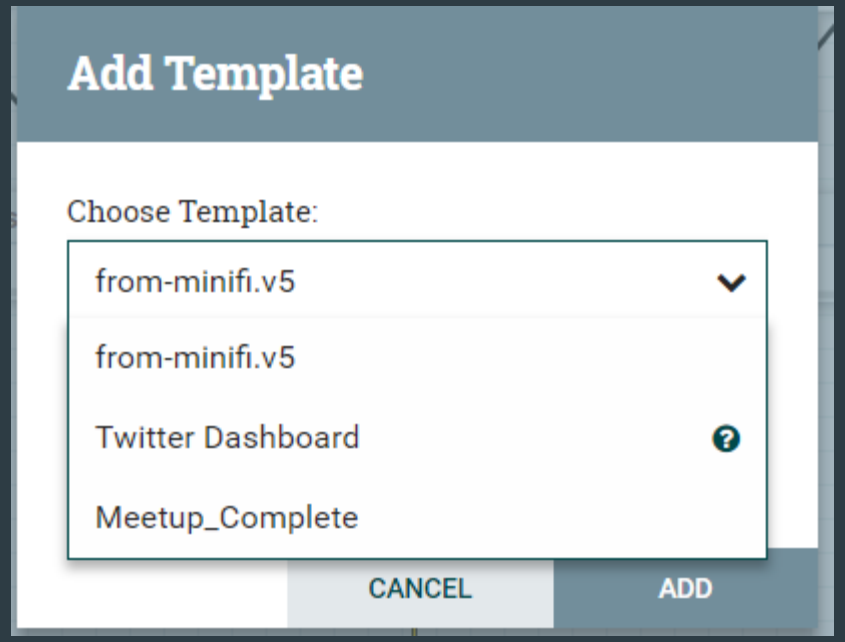
汇聚器

汇聚器可以使多个上游的输出统一进入下游队列，还可配置Flowfile出队优先级

Apache NIFI UI 介绍

常用工具栏 – 添加模板



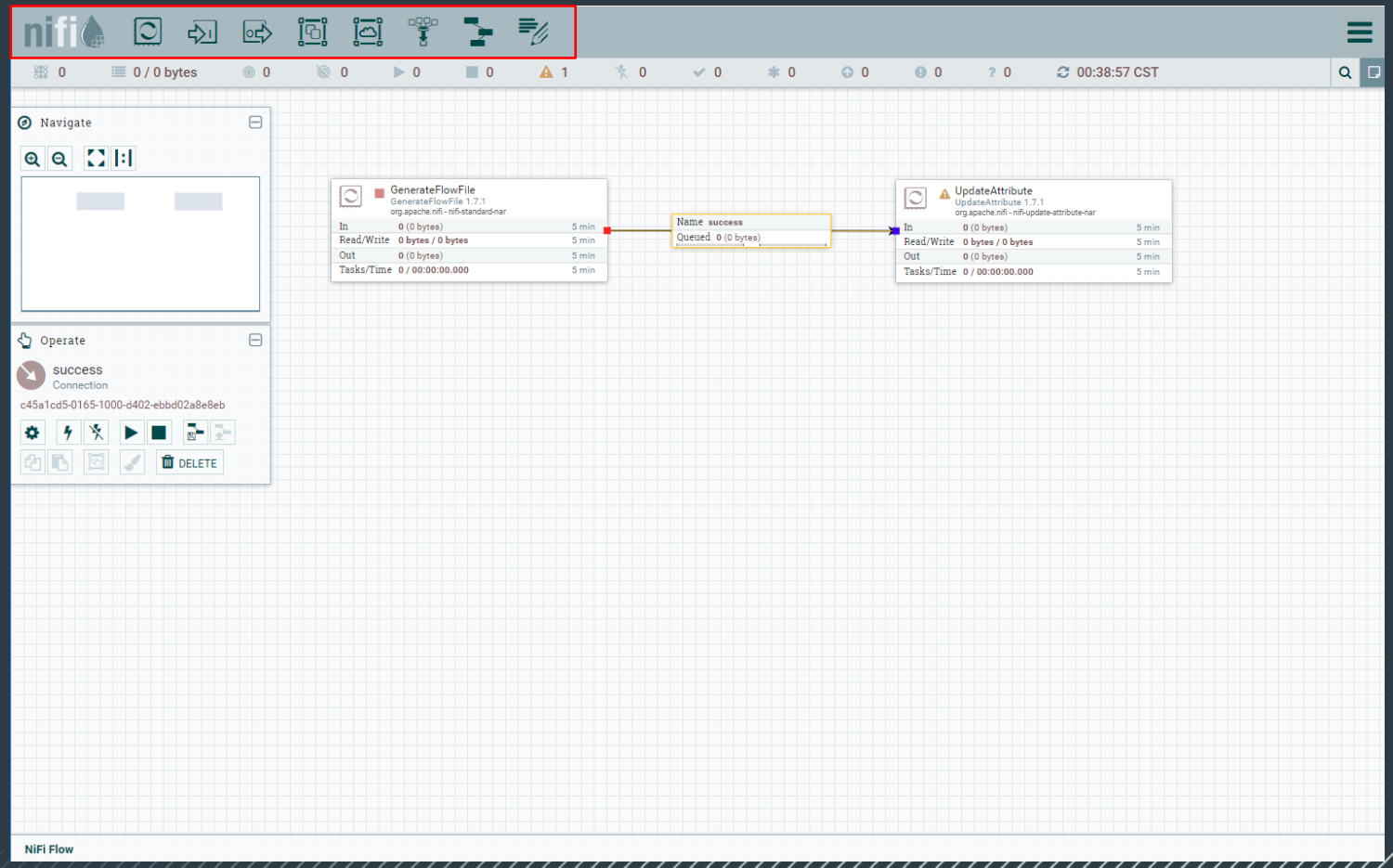


模板

可以将开发好的Dataflow以及相关的控制器进行模板化的保存.

Apache NIFI UI 介绍

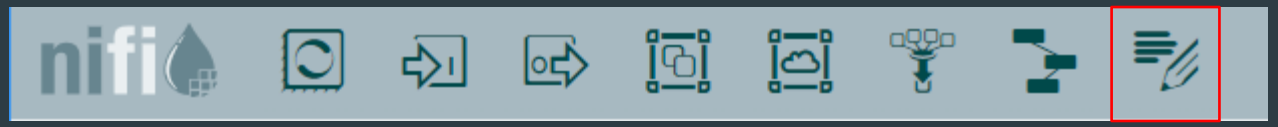
常用工具栏

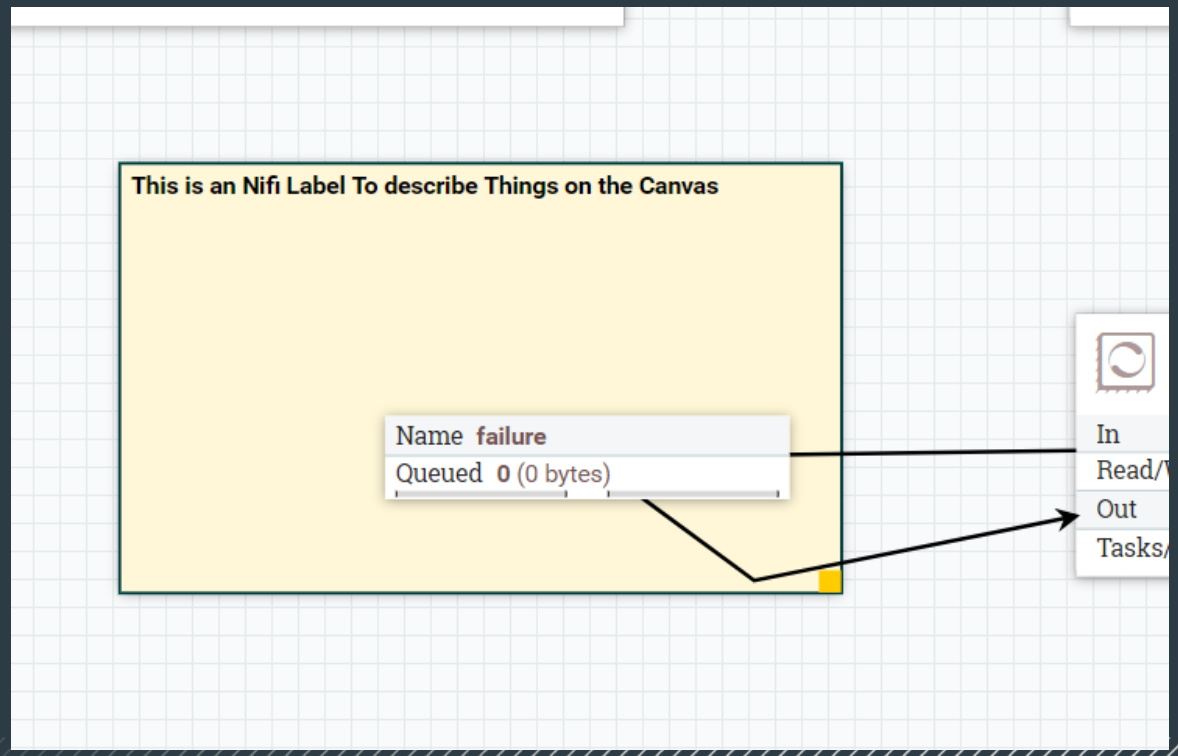


常用工具栏

Apache NIFI UI 介绍

常用工具栏 – 便利贴



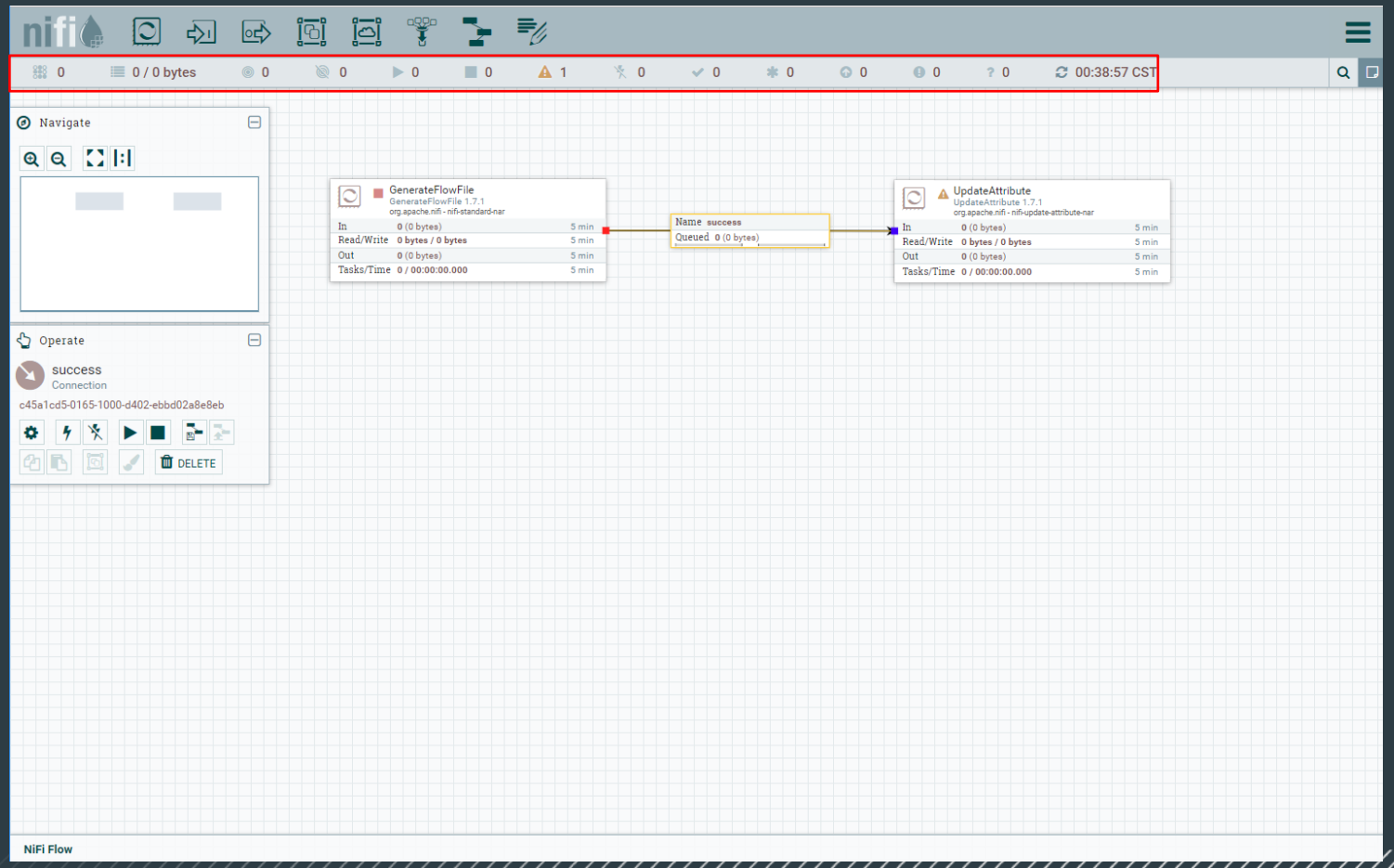


便利贴

在画布上对Dataflow做额外说明

Apache NIFI UI 介绍

统计信息及状态栏

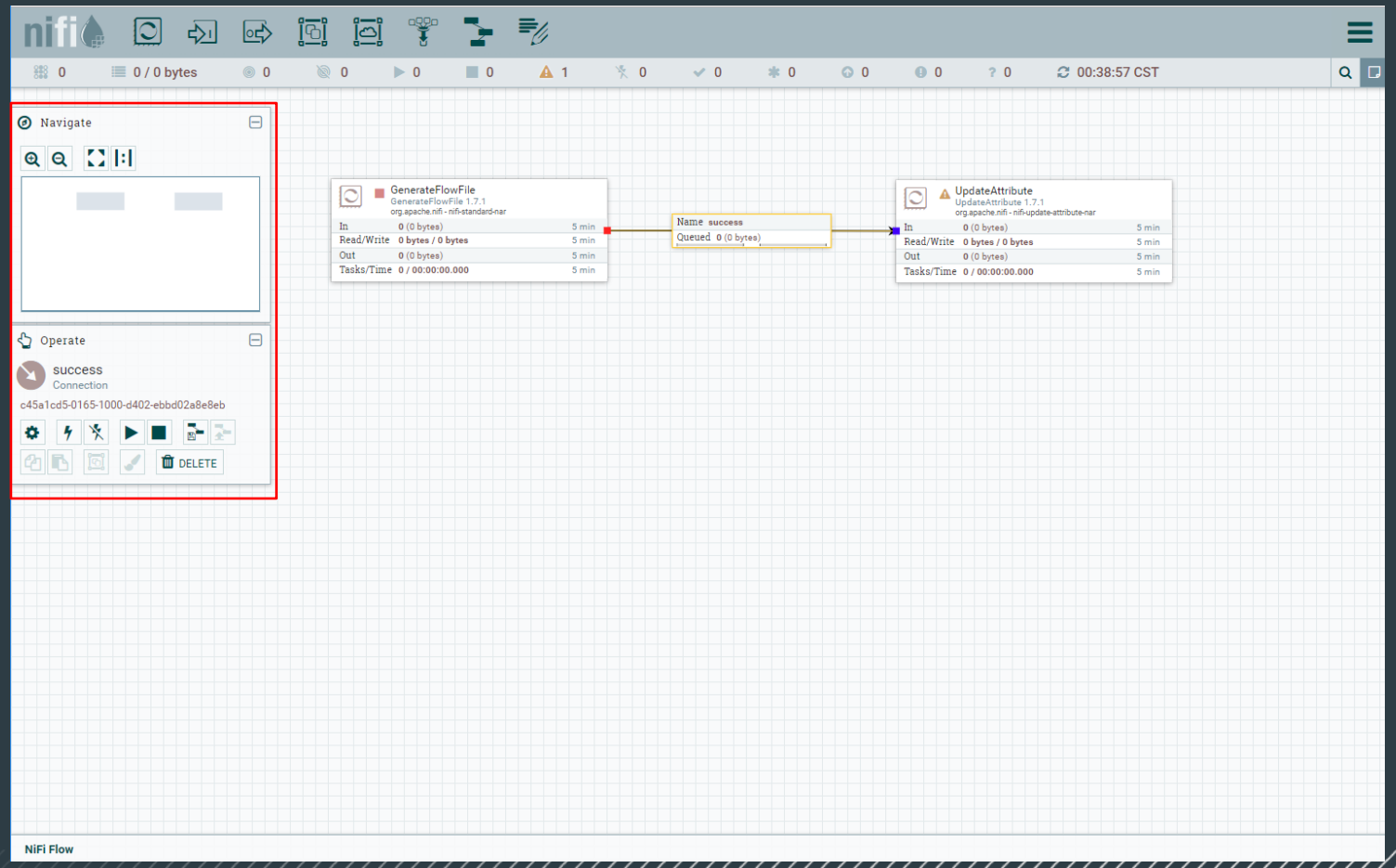


统计信息及状态栏

包含了一些常用的统计信息标识NIFI的当前状态

Apache NIFI UI 介绍

操作窗口



全局浏览窗

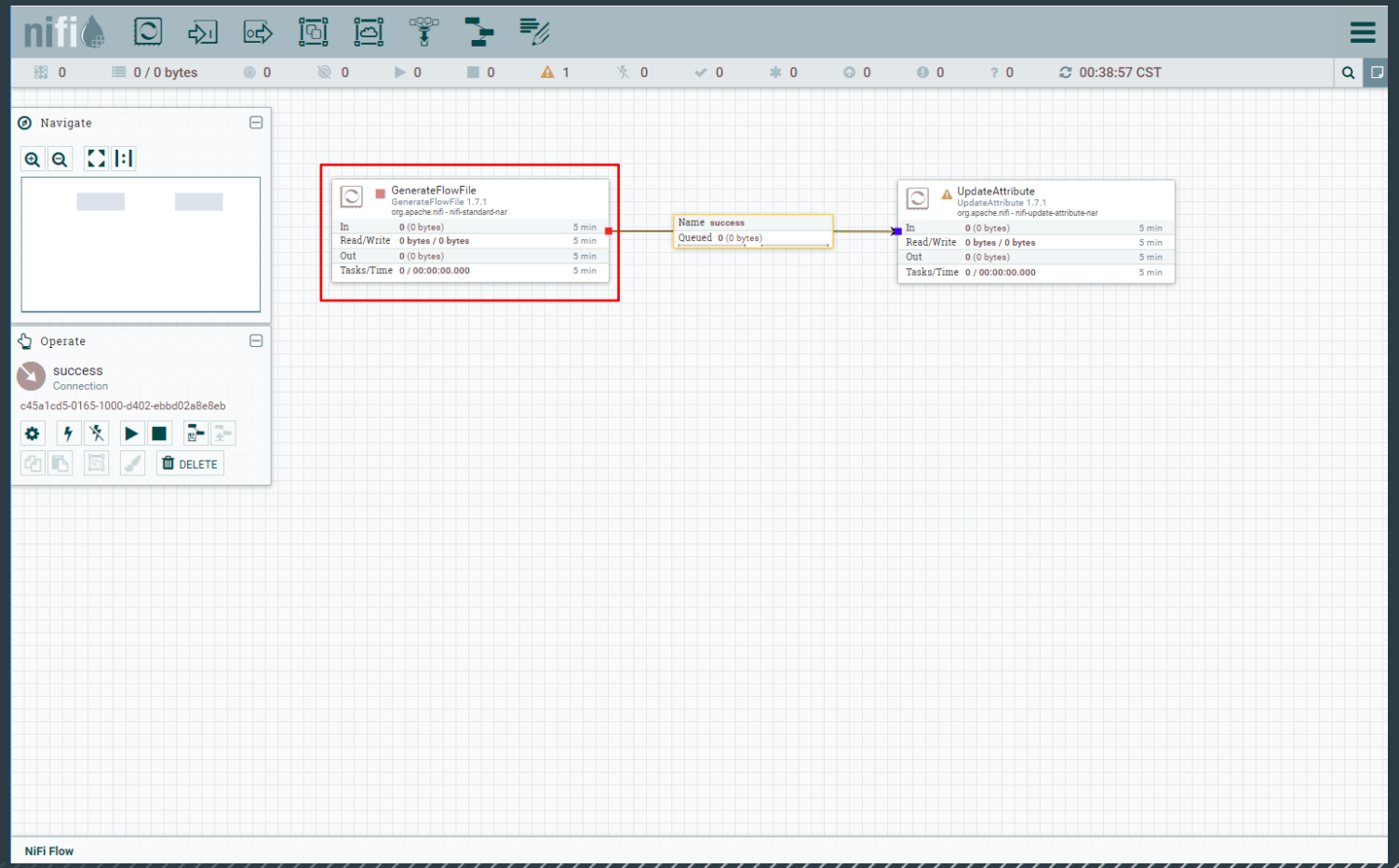
快速定位画布上有多少个处理器。

操作窗

对选中的NIFI元素的一些常见操作

Apache NIFI UI 介绍

处理器



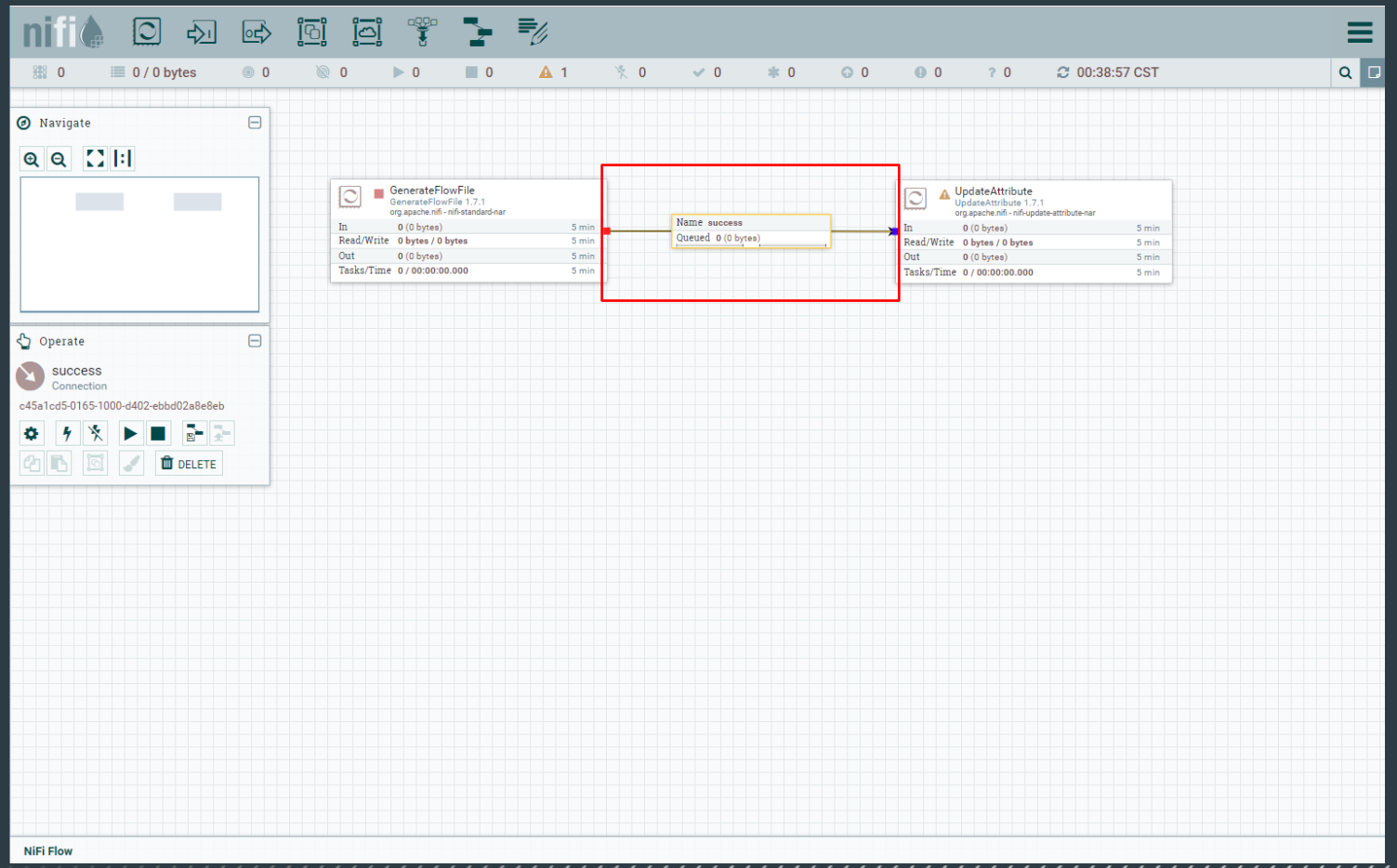
处理器

用于处理Dataflow的逻辑单元

- 监听数据源.
- 数据路由，转换.
- 推送数据至外部服务

Apache NIFI UI 介绍

处理器



处理器关系

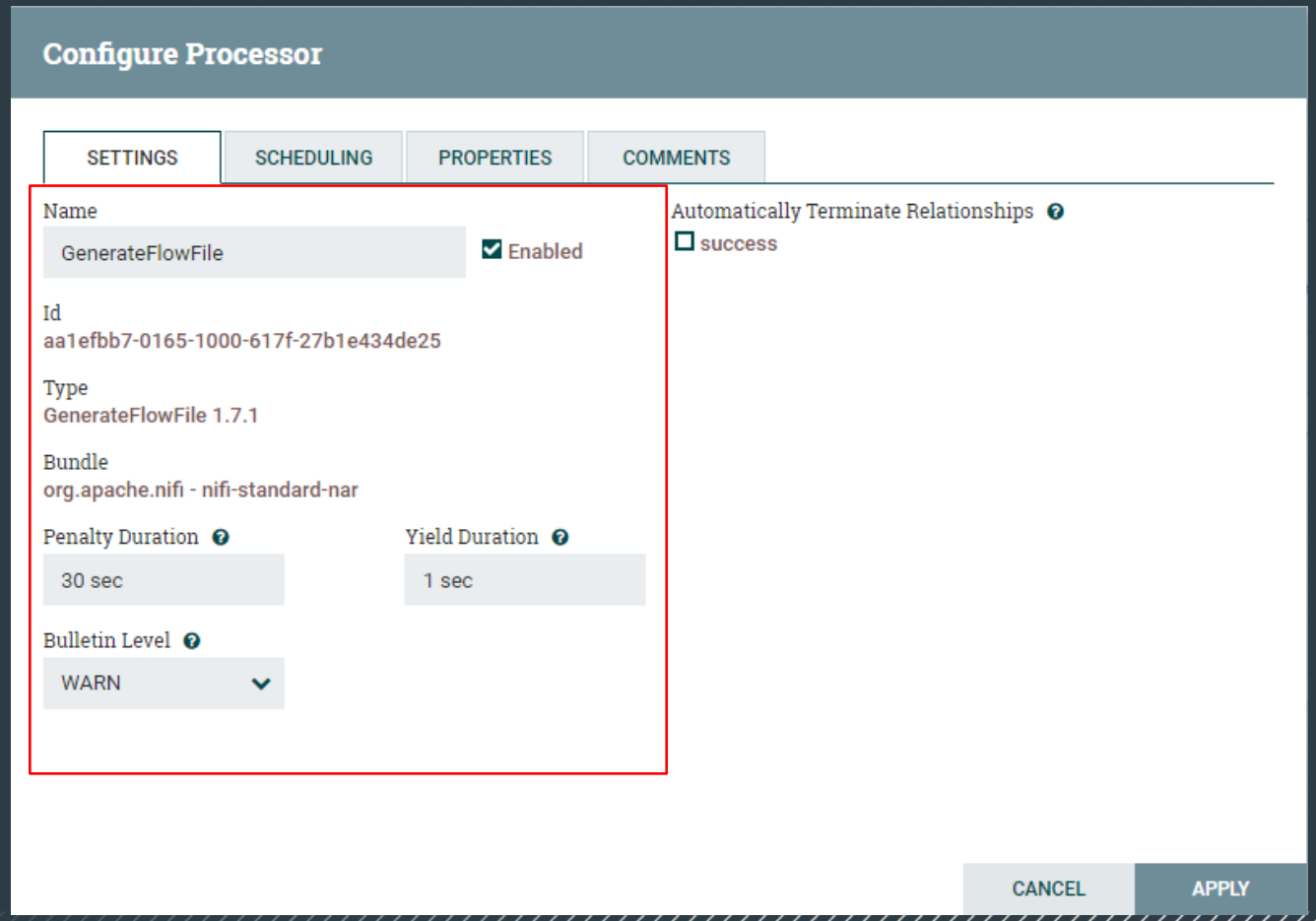
每一个处理器有0个或多个下游关系。例如成功流向，失败流向等，每个处理器不同

连接

每个连接可以包含一个或多个处理器关系

Apache NIFI UI 介绍

处理器配置



Name:处理器名称

ID:唯一的处理器ID

Type： 处理器类型

Bundle： Bundle包的名称

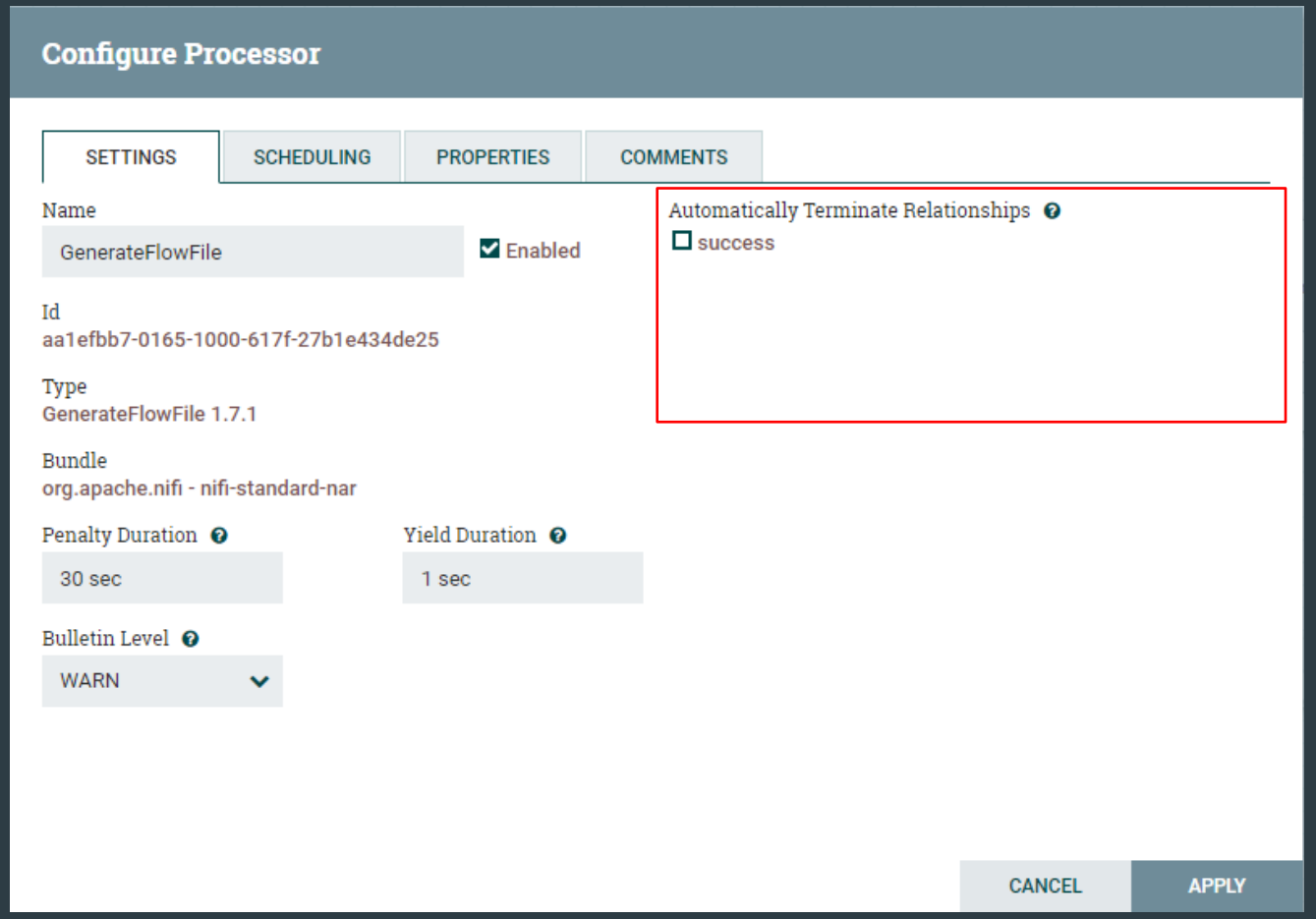
Penalty Duration： Flowfile故障延迟时间

Yield Duration： 下游故障停止时间

BulletinLevel:页面告警的日志级别

Apache NIFI UI 介绍

处理器

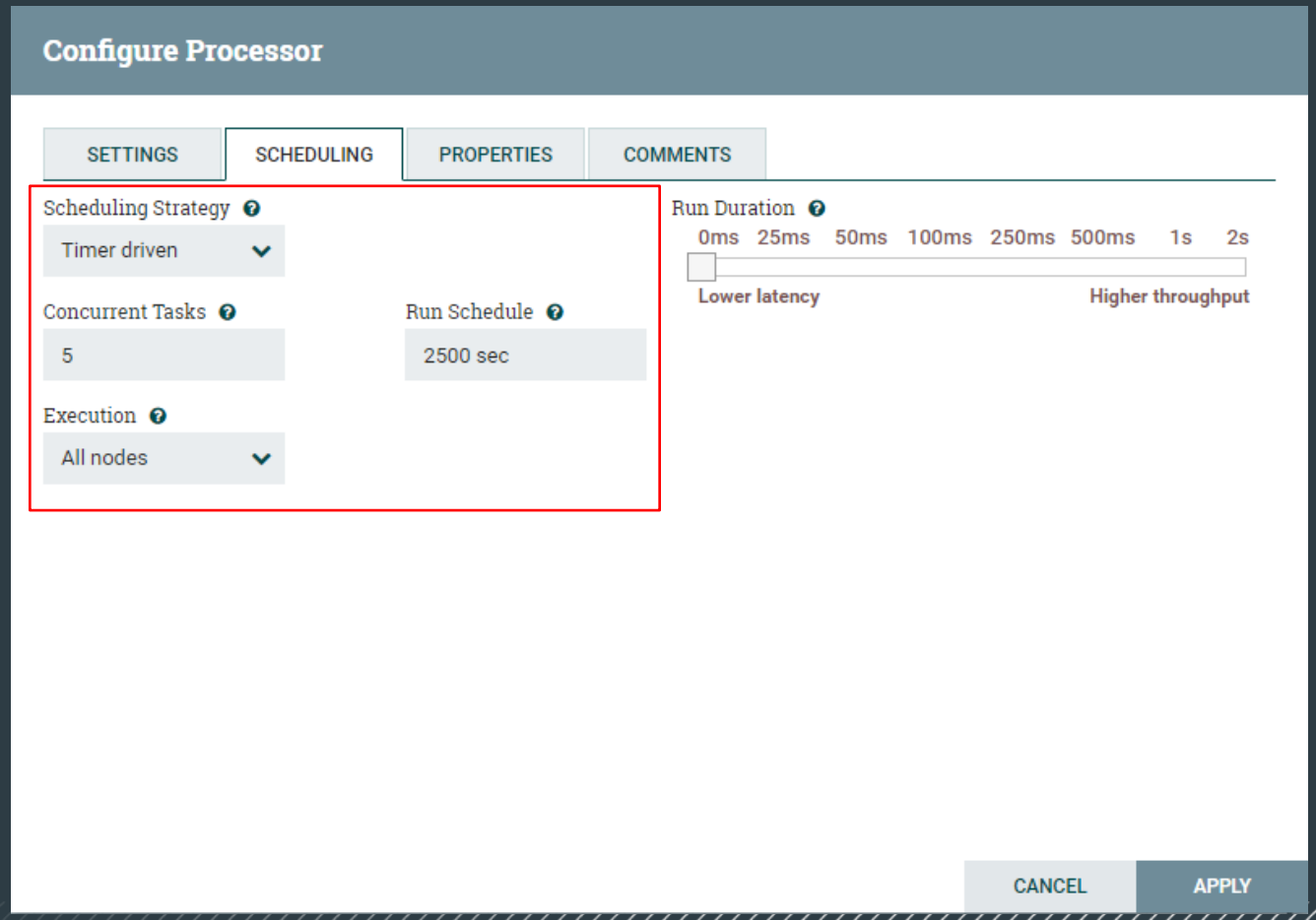


自动终止的关系

每一个处理器的下游流向需要完整配置，不需要的下游流向可以选择自动终止

Apache NIFI UI 介绍

处理器



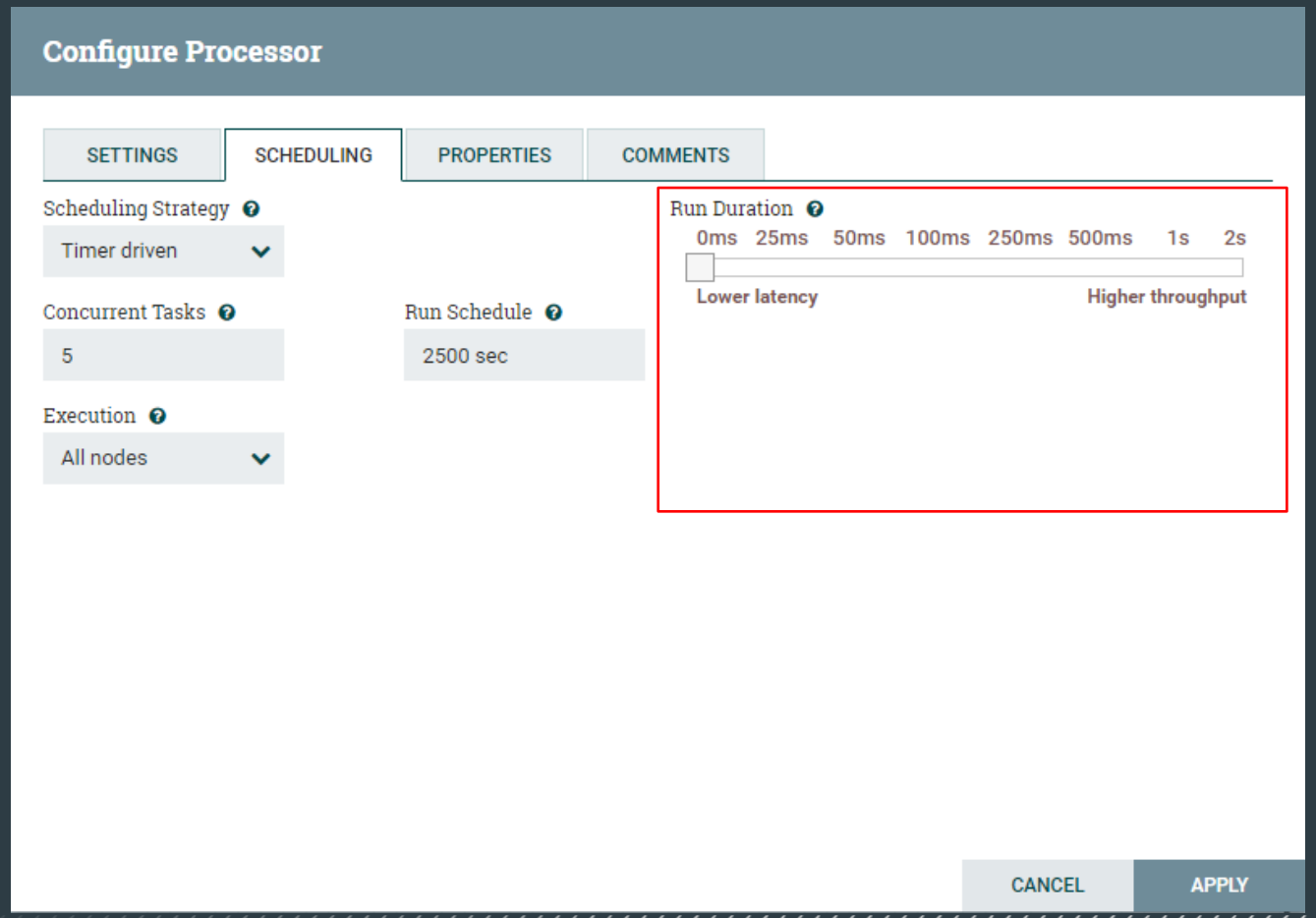
执行调度器计划

- Time Driven
- Event Driven
- CORN Driven
  - 0 0 13 * * ?
  - 0 20 14 ? * MON-FRI

ConcurrentTaks：并发线程
Run Schedule：执行间隔
Execution：执行节点

Apache NIFI UI 介绍

处理器

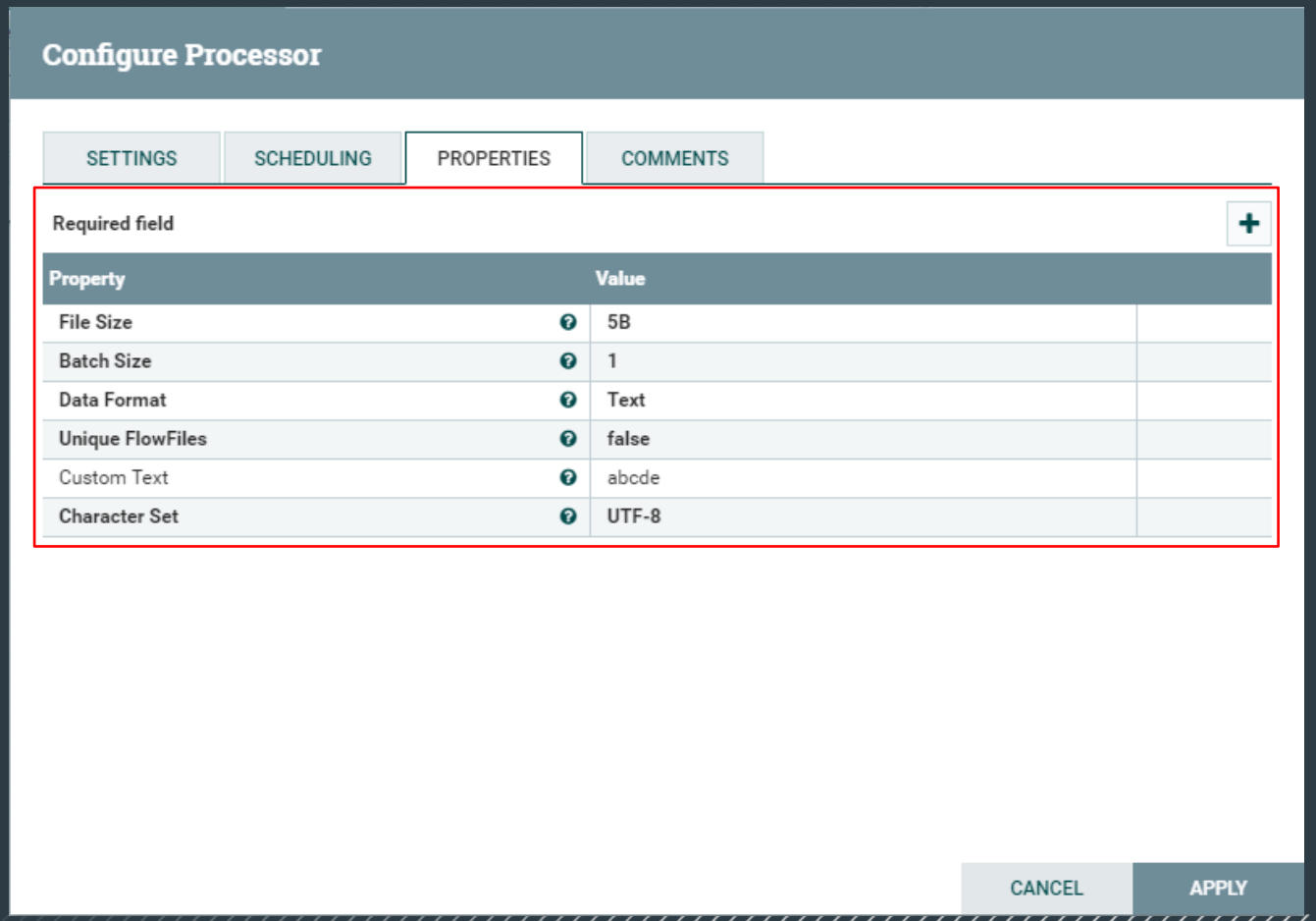


执行时间

- 执行时间短 ->
  低延时，低吞吐
- 执行时间长 ->
  高延时，高吞吐

Apache NIFI UI 介绍

处理器



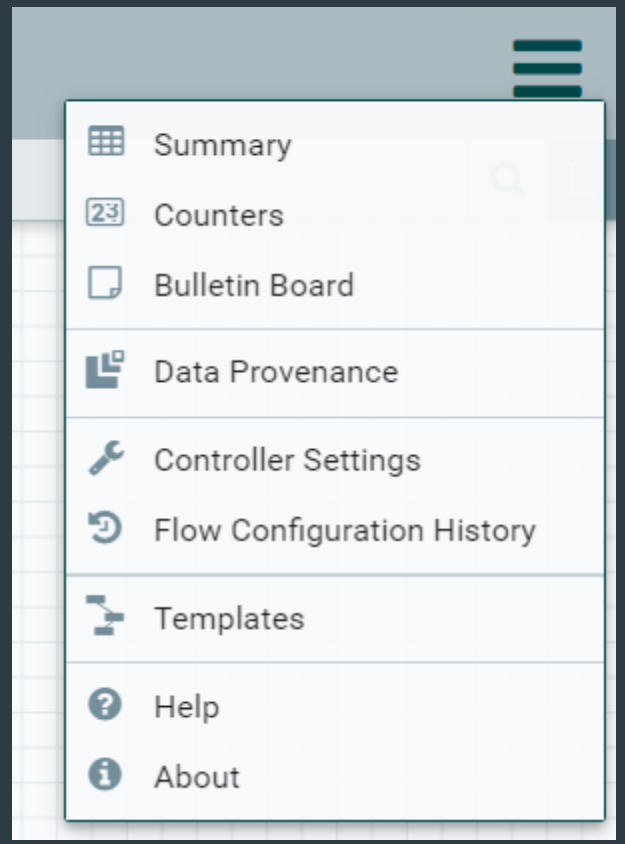
处理器相关配置

- 每个处理器不同
- 300 + 处理器约有共10，000个配置项。

Apache NIFI UI 介绍

服务管理及其他

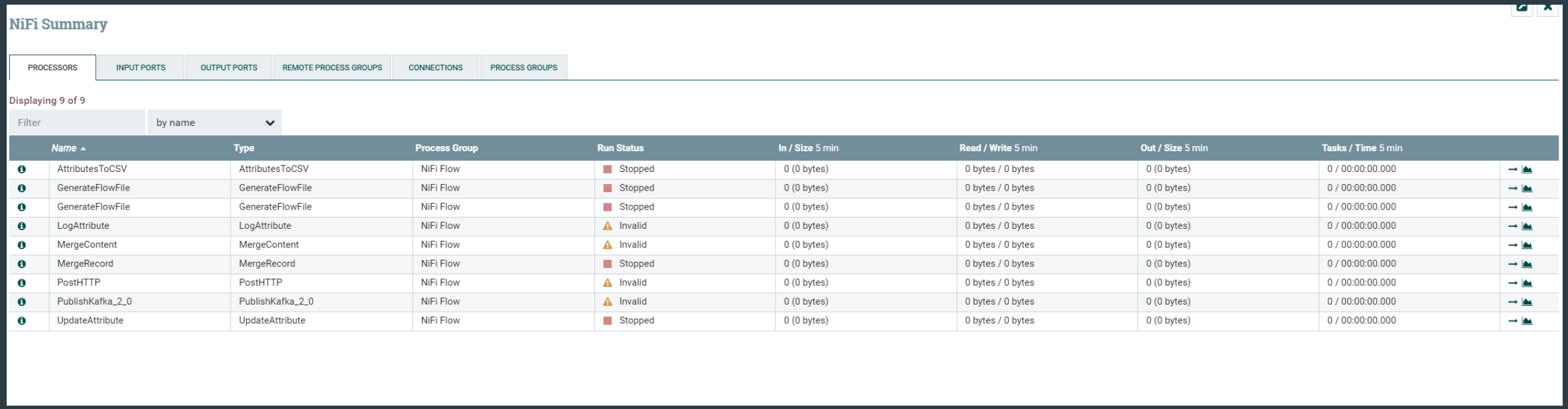
- 常用的管理性服务
  - Summary (处理器及关系汇总)
  - Counter
  - Bulletin Board (故障日志列表)
  - Controller Settings （公共服务配置）
  - …
  - Help （帮助文档）



Apache NIFI UI 介绍

服务管理及其他

- Summary(处理器及关系汇总)

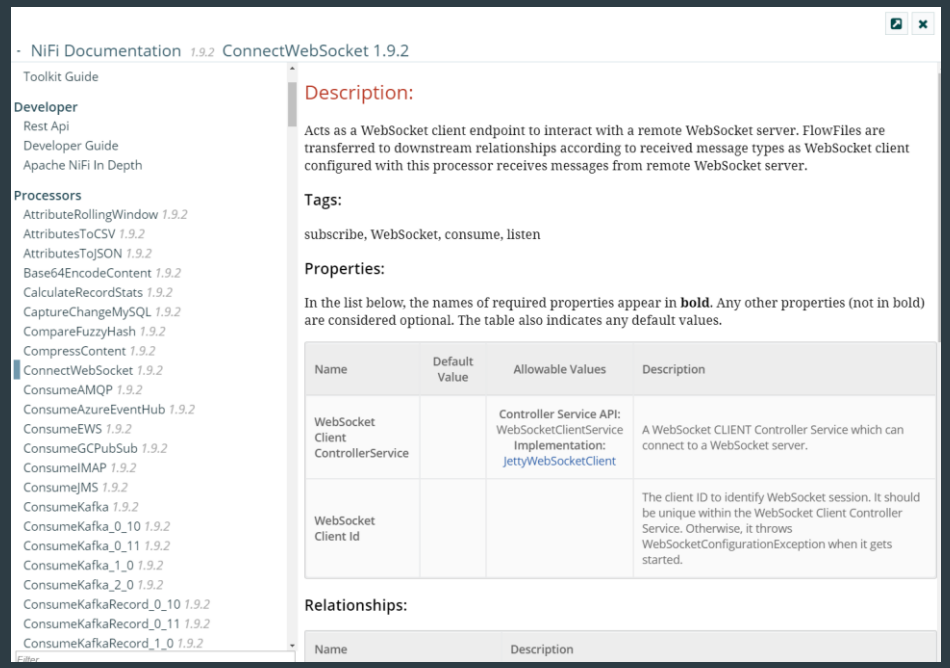


Name	Type	Process Group	Run Status	In / Size 5 min	Read / Write 5 min	Out / Size 5 min	Tasks / Time 5 min
AttributesToCSV	AttributesToCSV	NiFi Flow	Stopped	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
GenerateFlowFile	GenerateFlowFile	NiFi Flow	Stopped	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
GenerateFlowFile	GenerateFlowFile	NiFi Flow	Stopped	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
LogAttribute	LogAttribute	NiFi Flow	Invalid	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
MergeContent	MergeContent	NiFi Flow	Invalid	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
MergeRecord	MergeRecord	NiFi Flow	Stopped	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
PostHTTP	PostHTTP	NiFi Flow	Invalid	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
PublishKafka_2_0	PublishKafka_2_0	NiFi Flow	Invalid	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000
UpdateAttribute	UpdateAttribute	NiFi Flow	Stopped	0 (0 bytes)	0 bytes / 0 bytes	0 (0 bytes)	0 / 00:00:00.000

Apache NIFI UI 介绍

服务管理及其他

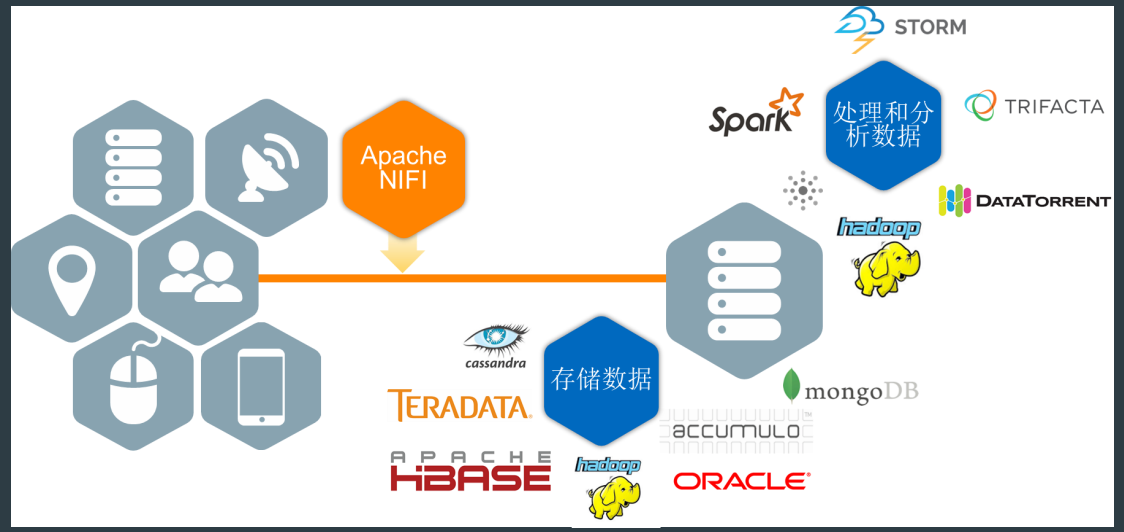
- Help(帮助文档)

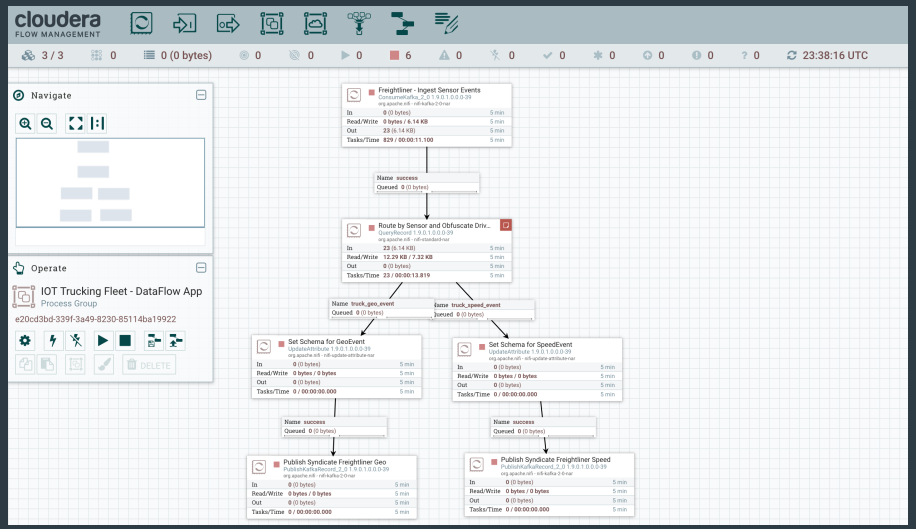


NIFI深度讲解

小结

- 1. CDF 产品介绍
- 2. NIFI 产品介绍
- 3. NIFI UI 介绍





动手实战

Lab 1 & Lab 2

https://github.com/wangyong23/CDF-Workshop-2020

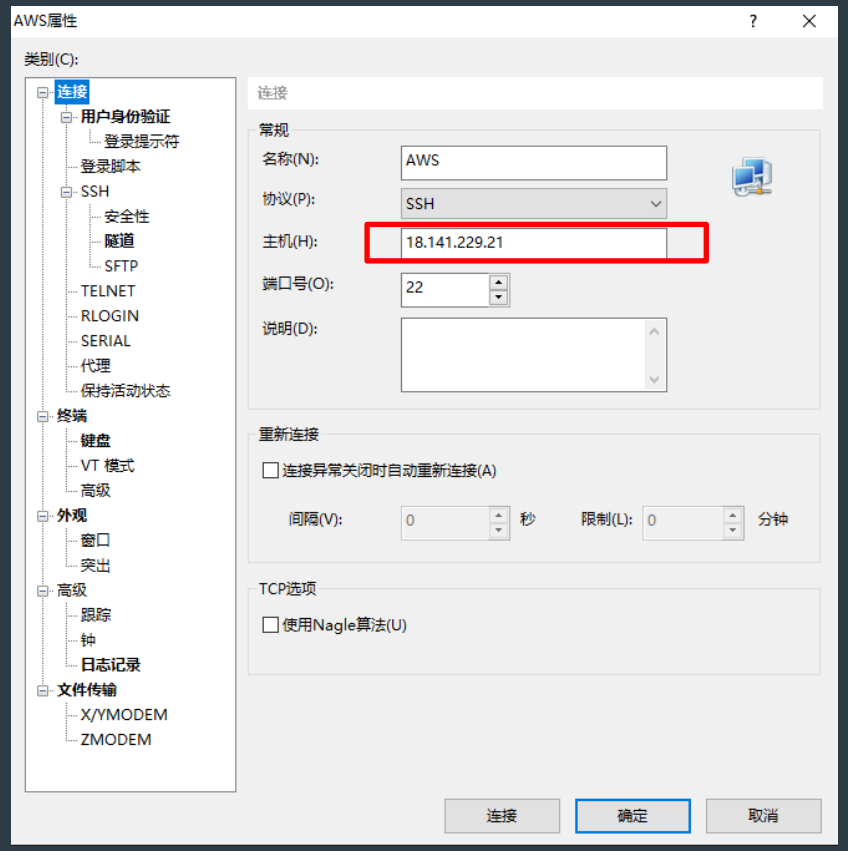
LAB 1

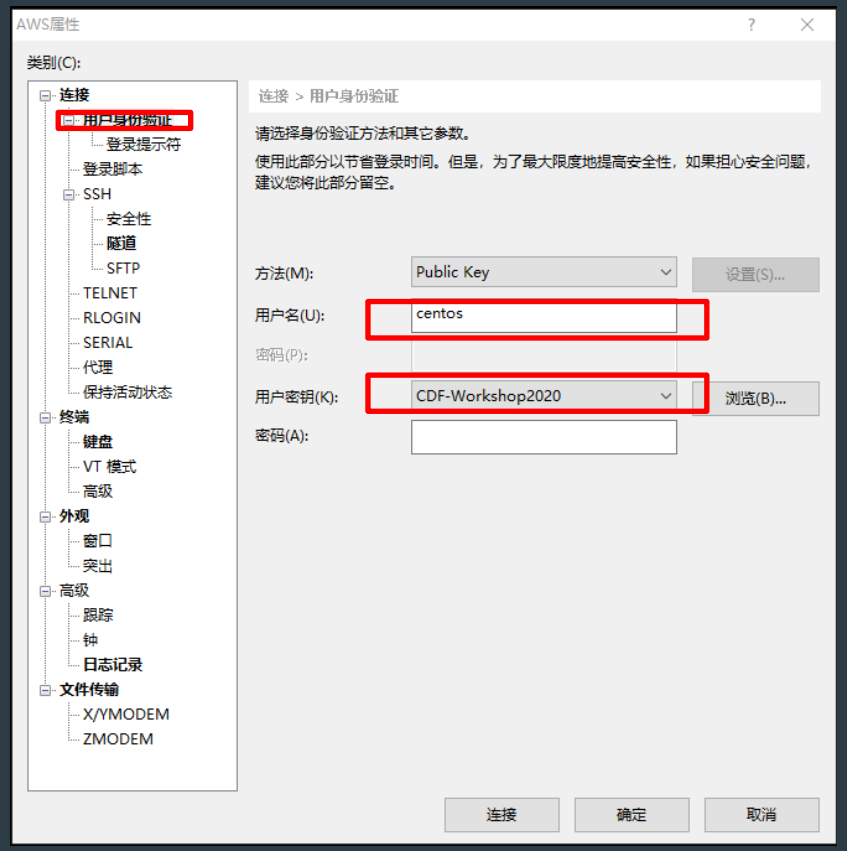
登录你的实验环境

- 配置SSH客户端
  - Mac，Linux用户-SSH
  - Windows用户-Putty 或其他SSH 客户端
- 下载Key
  - Max，Linux用户- 下载PEM KEY文件 - chmod 400
  - Windows用户- 下载PPK KEY文件

LAB 1

登录你的实验环境 - Xshell

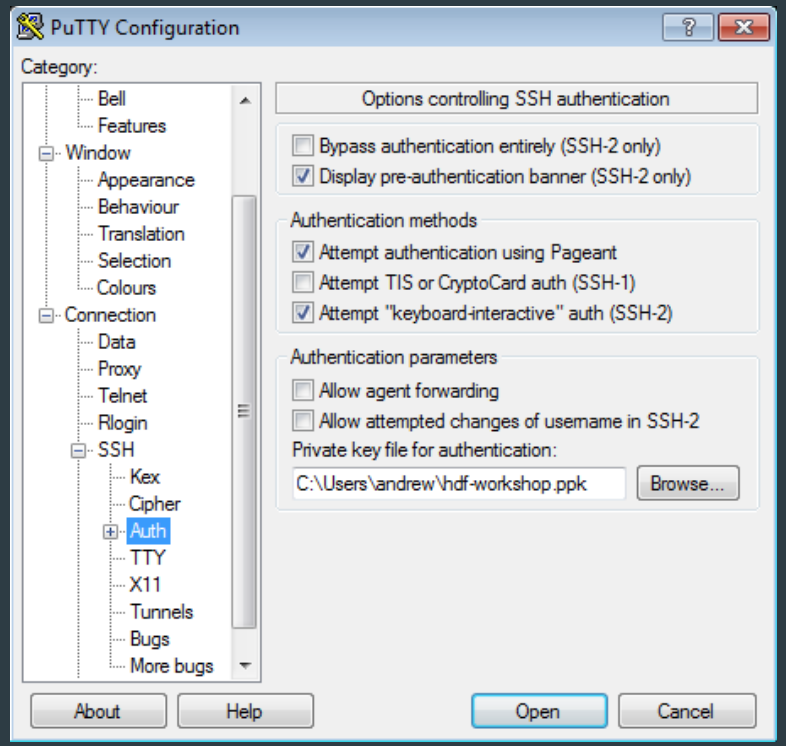


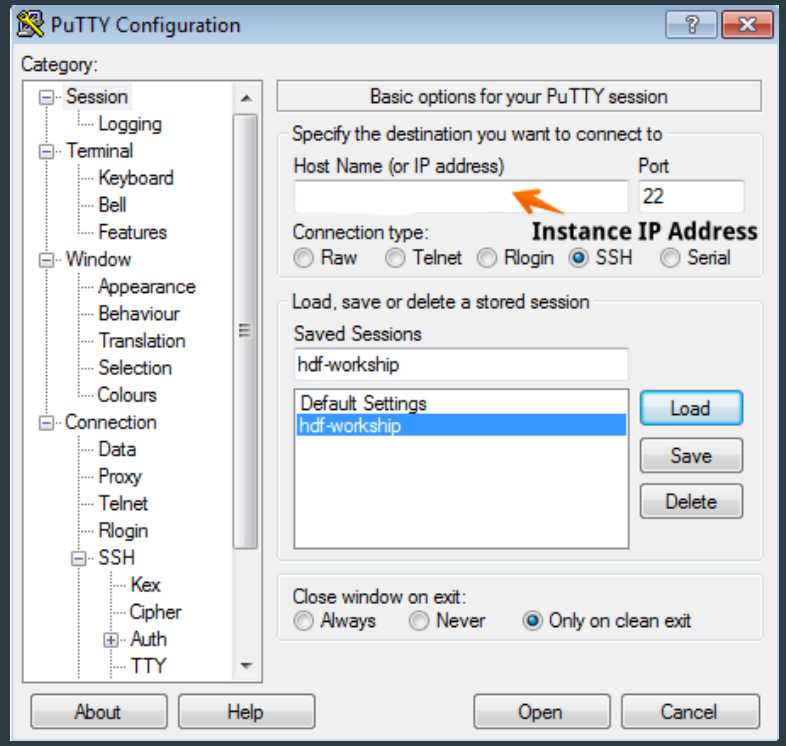


动手实战

Lab 1

- Windows - 进入OS环境





- Mac， Linux - 进入OS环境
- cp ~/Downloads/xxx.pem ~/.ssh/
- chmod 400 ~/.ssh/xxx.pem
- ssh -i ~/.ssh/xxx.pem centos@your.ec2.ip
- sudo su root

Lab 2

使用NIFI构建第一个数据流

meetup.com
Website
WebSocket API

NIFI

File


```
{
  "venue": {
    "venue_name":"Paris Baguette",
    "lon":-122.160645,
    "lat":37.447105,
    "venue_id":26054753
  },
  "visibility":"public",
  "response":"yes",
  "guests":0,
  "member": {
    "member_id":151326982,
    "photo":"https:\/\/secure.meetupstatic.com\/photos\/member\/4\/7\/f\/6\/thumb_230778422.jpeg",
    "member_name":"sylvie shand"
  },
  "rsvp_id":1807078527,
  "mtime":1568660772000,
  "event": {
    "event_name":"Practice speaking Spanish!!",
    "event_id":"hzslhryzmbzb",
    "time":1568944800000,
    "event_url":"https:\/\/www.meetup.com\/Palo-Alto-Practice-Spanish-Meetup\/events\/264447445\/"
  },
  "group": {
    "group_topics":[
      {
        "urlkey":"spanish",
        "topic_name":"Spanish Language"
      },
      {
        "urlkey":"spanish-culture",
        "topic_name":"Spanish Culture"
      },
      {
        "urlkey":"bilingual-spanish-english",
        "topic_name":"Bilingual Spanish\/English"
      },
      {
        "urlkey":"espanol",
        "topic_name":"Español"
      }
    ],
    "group_city":"Palo Alto",
    "group_country":"us",
    "group_id":18385793,
    "group_name":"Palo Alto Practice Spanish Meetup",
    "group_lon":-122.15,
    "group_urlname":"Palo-Alto-Practice-Spanish-Meetup",
    "group_state":"CA",
    "group_lat":37.44
  }
}
```

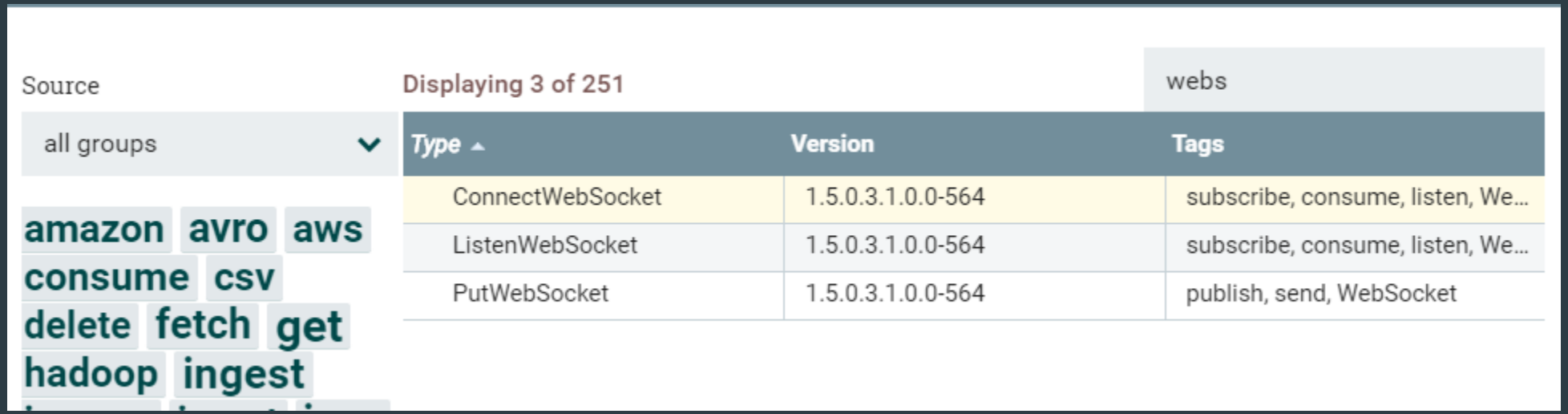
```
{
  "event_name":"Practice speaking Spanish!!",
  "event_url":"https://www.meetup.com/Palo-Alto-Practice-Spanish-Meetup/events/264447445/",
  "venue": {
    "lat":"37.447105",
    "lon":"-122.160645",
    "name":"Paris Baguette"
  },
  "group": {
    "group_city":"Palo Alto",
    "group_country":"us",
    "group_name":"Palo Alto Practice Spanish Meetup",
    "group_state":"CA",
    "urlkey":"spanish",
    "topic_name":"Spanish Language"
  }
}
```

```
{
  "event_name":"Practice speaking Spanish!!",
  "event_url":"https://www.meetup.com/Palo-Alto-Practice-Spanish-Meetup/events/264447445/",
  "venue": {
    "lat":"37.447105",
    "lon":"-122.160645",
    "name":"Paris Baguette"
  },
  "group": {
    "group_city":"Palo Alto",
    "group_country":"us",
    "group_name":"Palo Alto Practice Spanish Meetup",
    "group_state":"CA",
    "urlkey":"spanish-culture",
    "topic_name":"Spanish Culture"
  }
}
```

Lab 2

使用NIFI构建第一个数据流

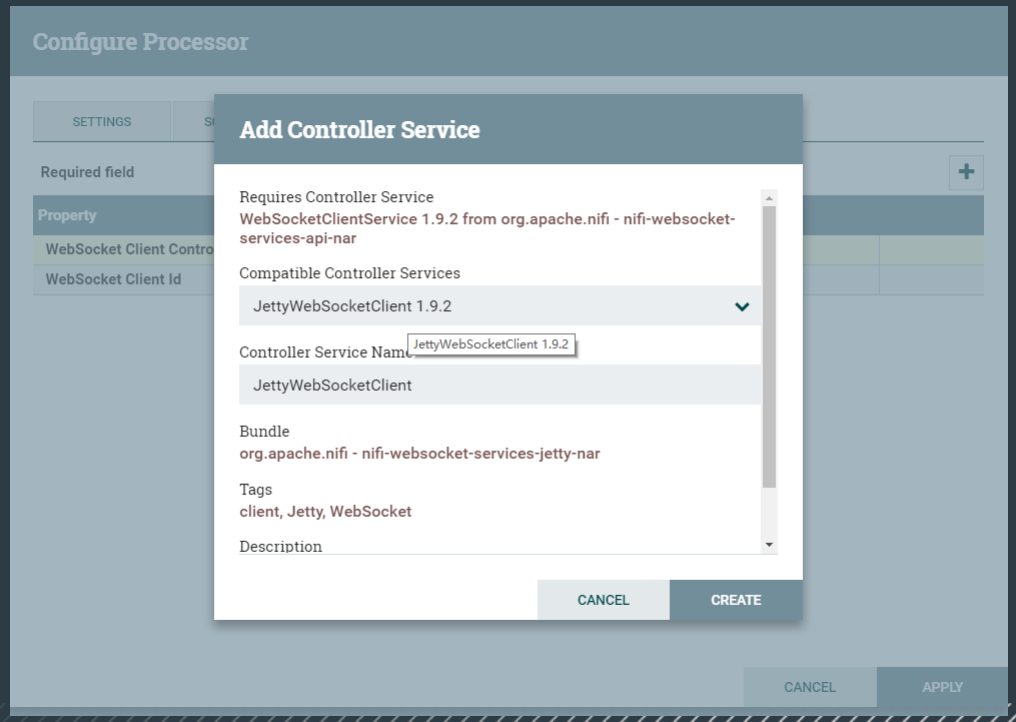
- 添加一个ConnectWebSocket 处理器

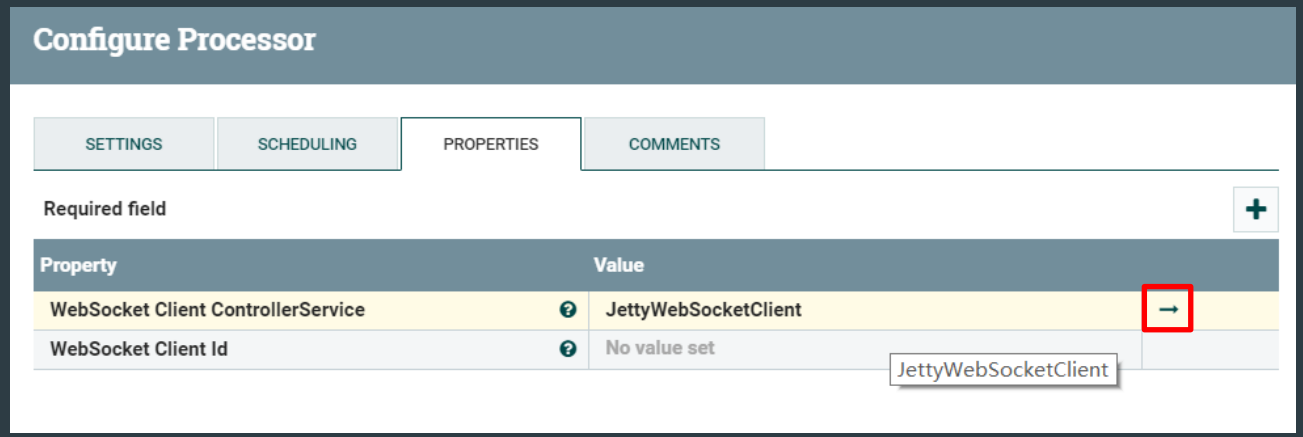


Lab 2

使用NIFI构建第一个数据流

- 选择Jetty客户端服务，并打开Jetty客户端服务配置

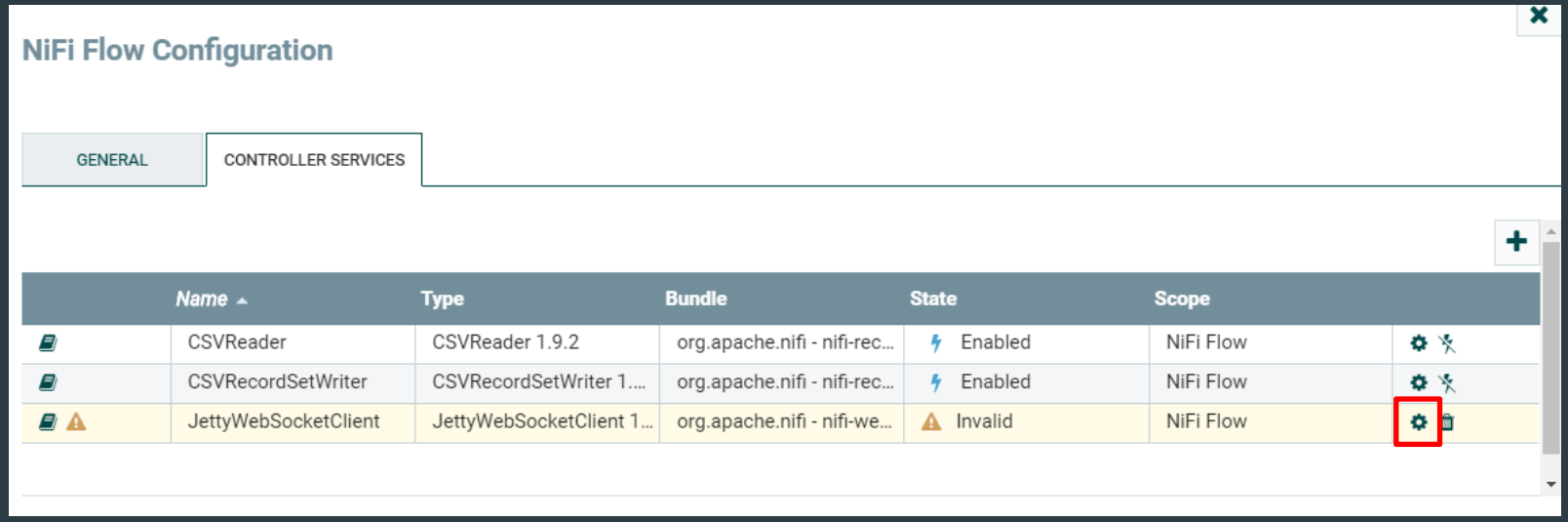




Lab 2

使用NIFI构建第一个数据流

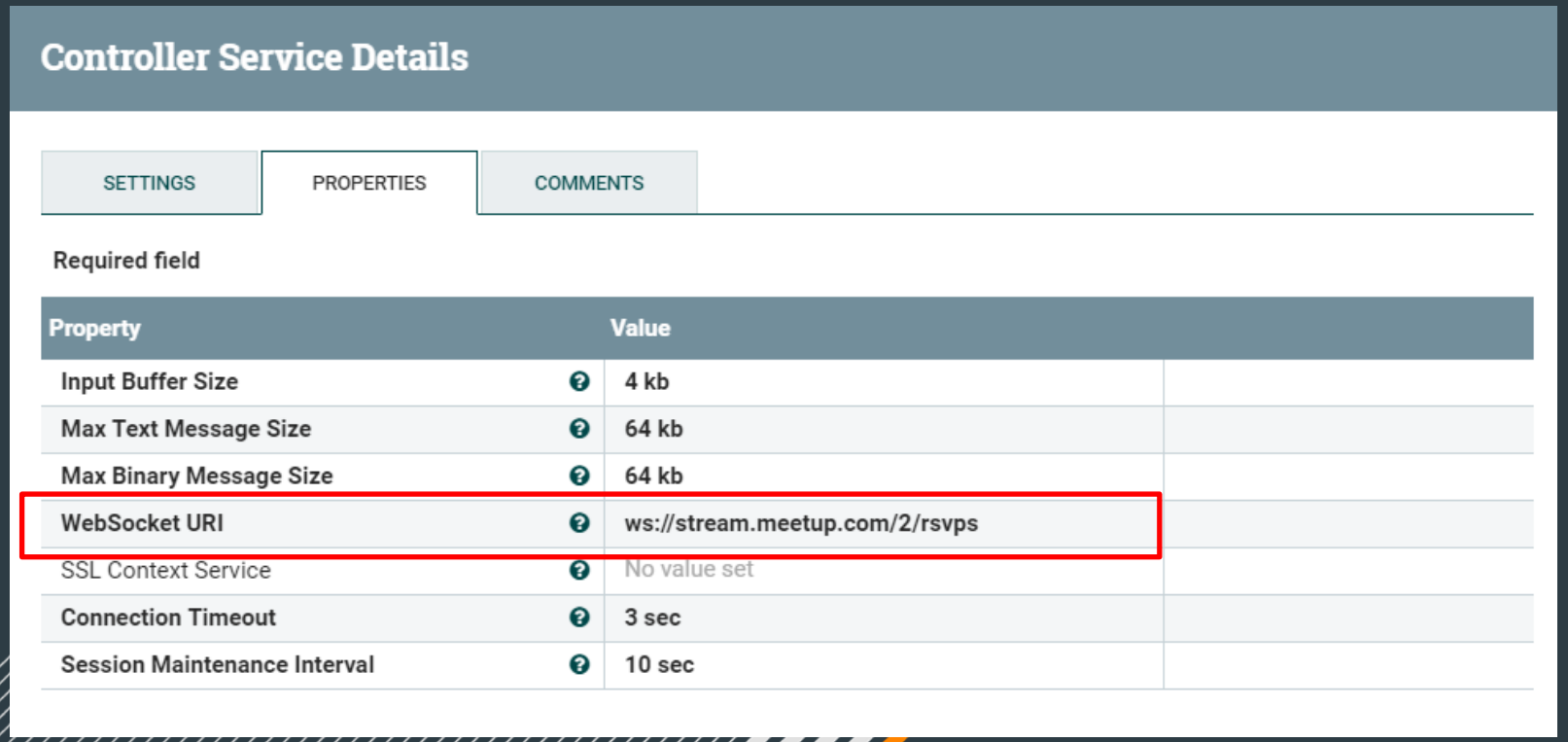
- 选择Jetty客户端服务，并打开Jetty客户端服务配置



Lab 2

使用NIFI构建第一个数据流

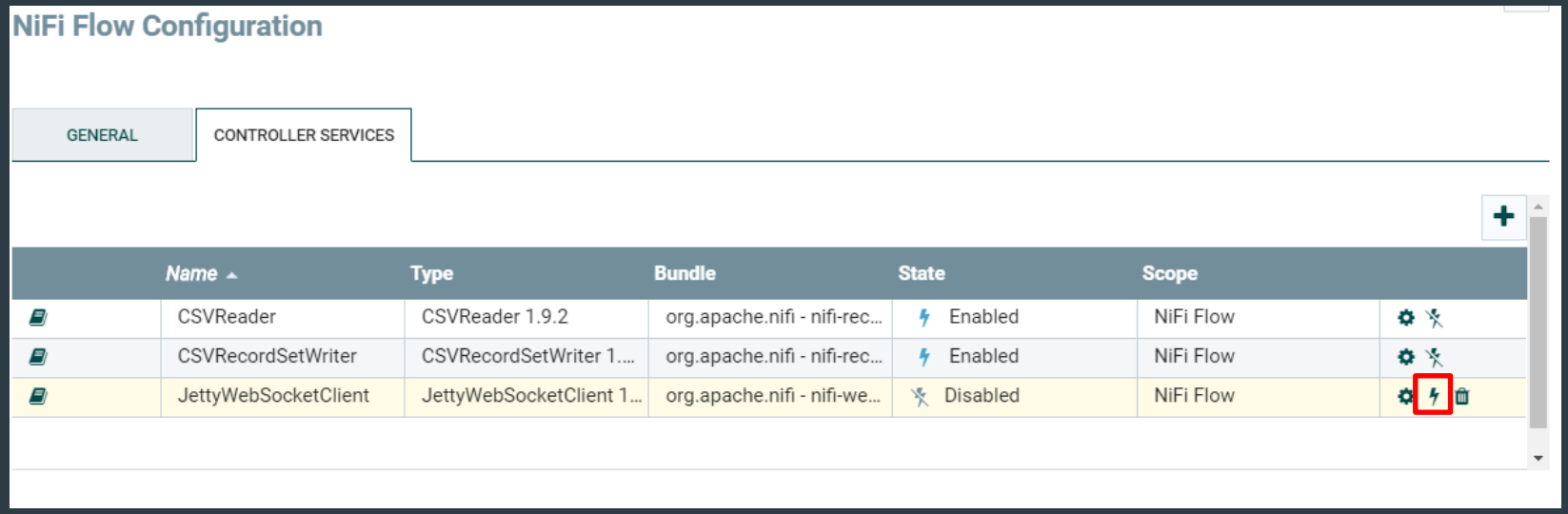
- 配置相应的Jetty客户端服务
  - WebSocketURI ws://stream.meetup.com/2/rsvps



Lab 2

使用NIFI构建第一个数据流

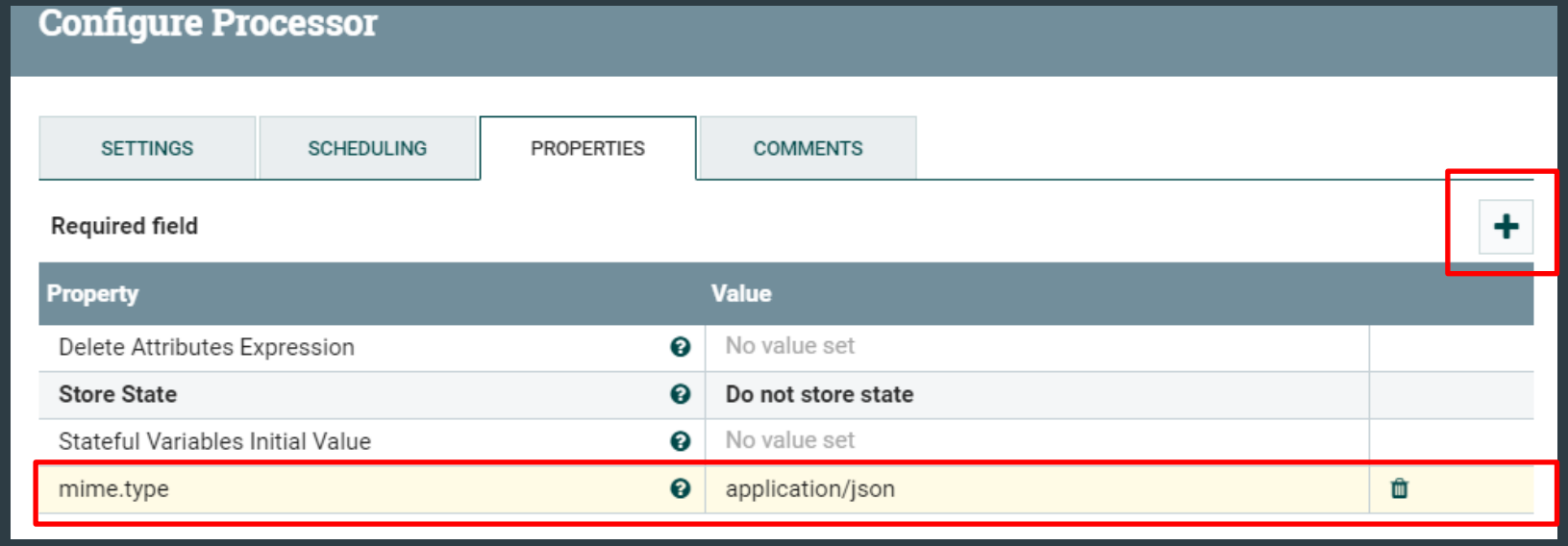
- 启用Jetty客户端服务



Lab 2

使用NIFI构建第一个数据流

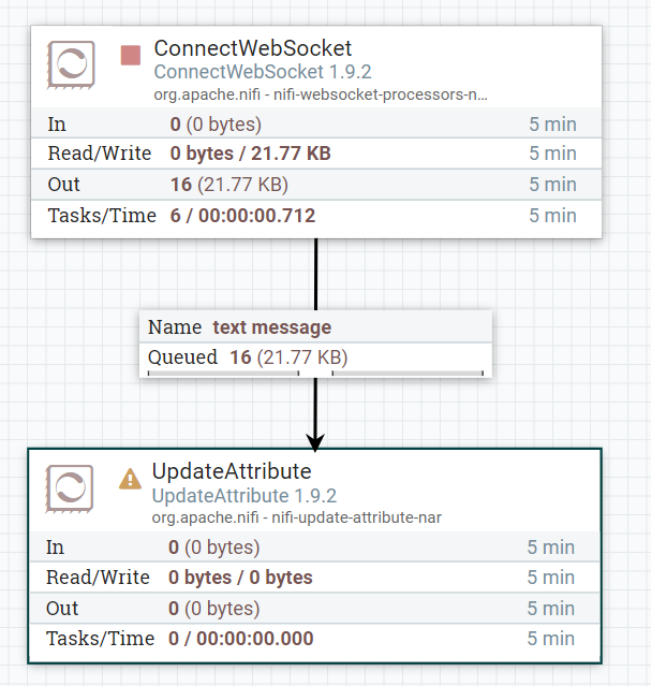
- 添加Update Attribute 处理器

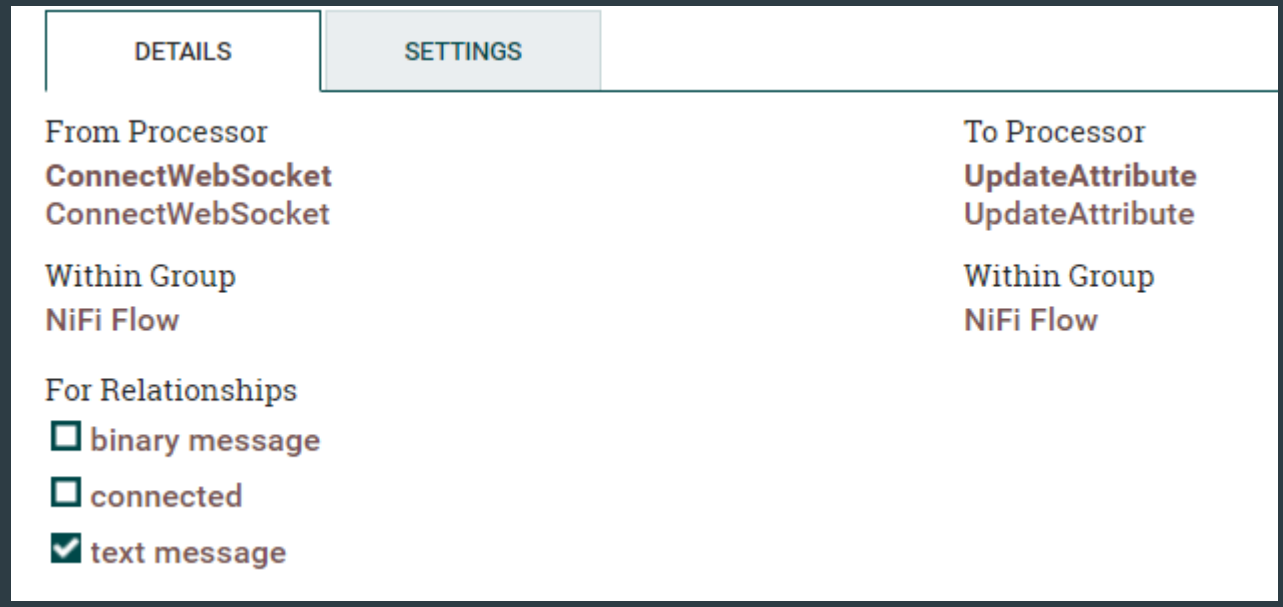


Lab 2

使用NIFI构建第一个数据流

- 配置两个处理器的关系

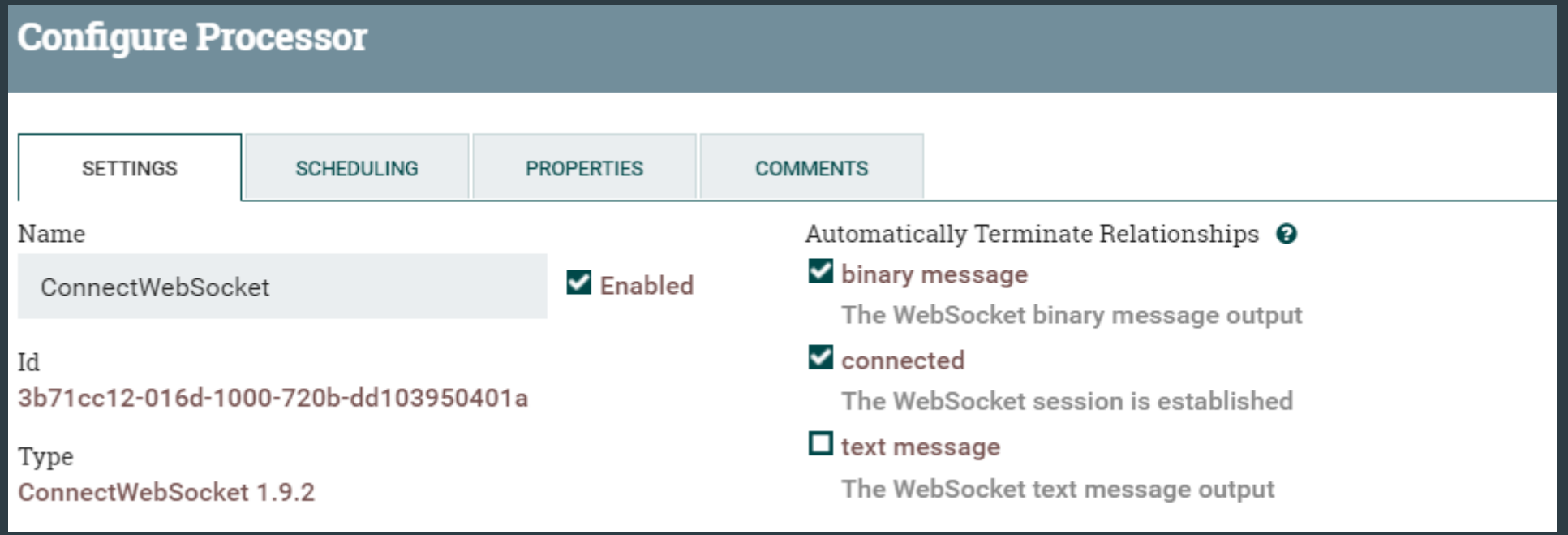




Lab 2

使用NIFI构建第一个数据流

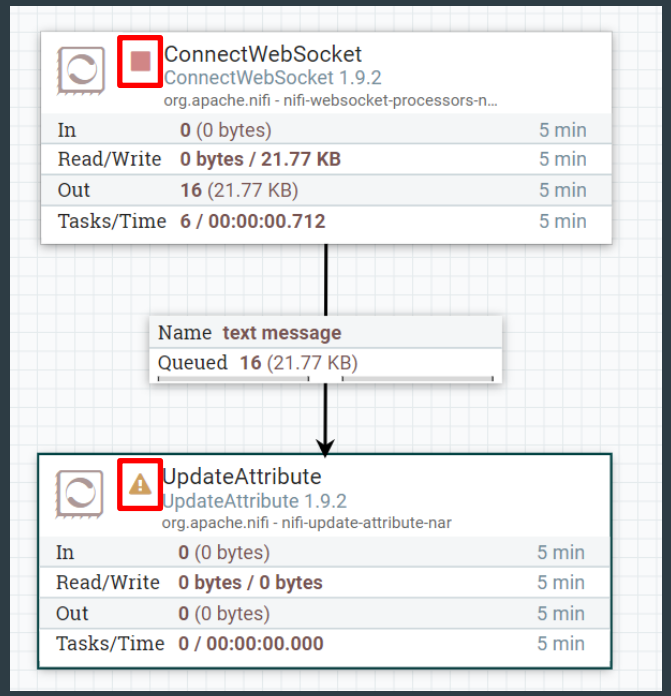
- 自动终止ConnectWebSocket处理器的其他输出关系



Lab 2

使用NIFI构建第一个数据流

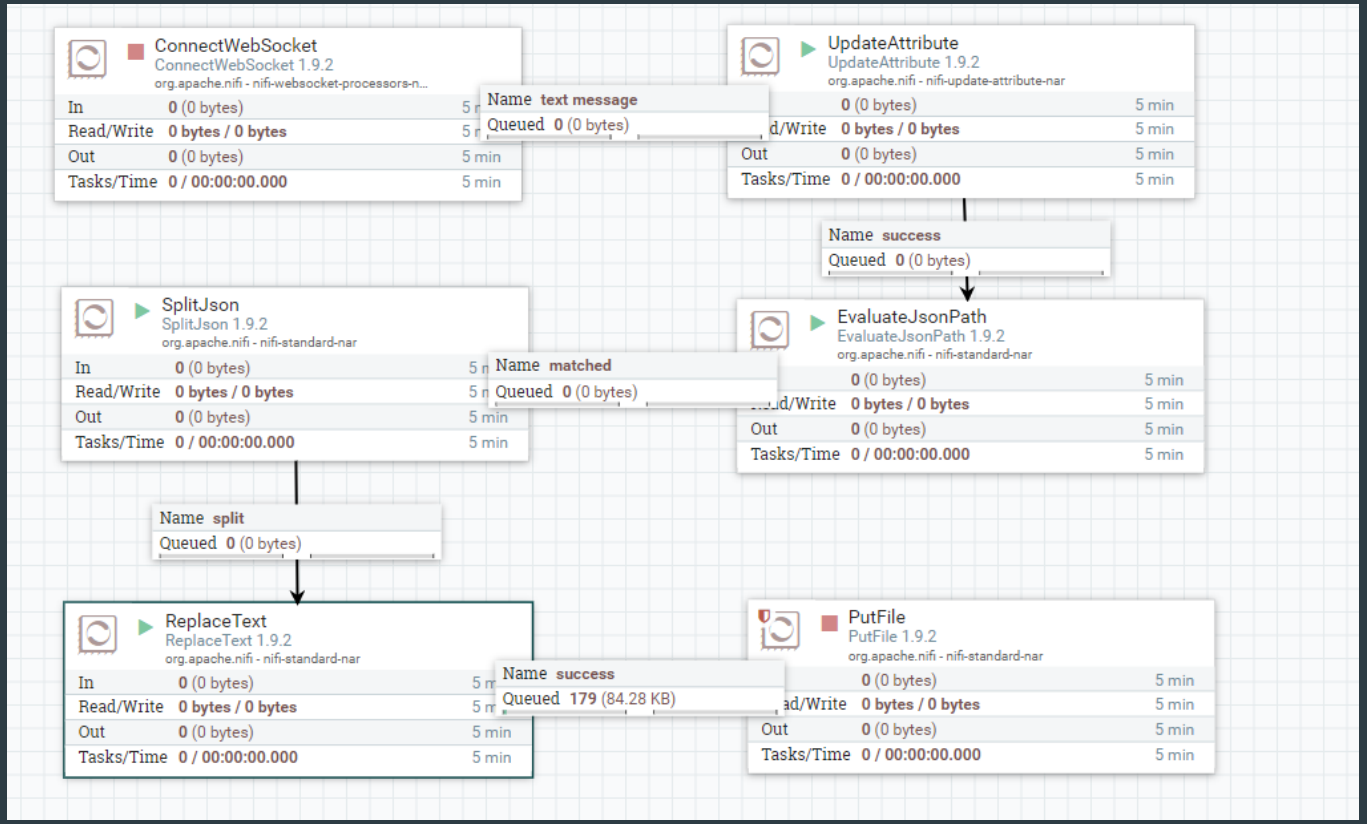
- 当ConnectWebSocket为红色方块时代表处理器基本配置已满足，可运行



Lab 2

使用NIFI构建第一个数据流

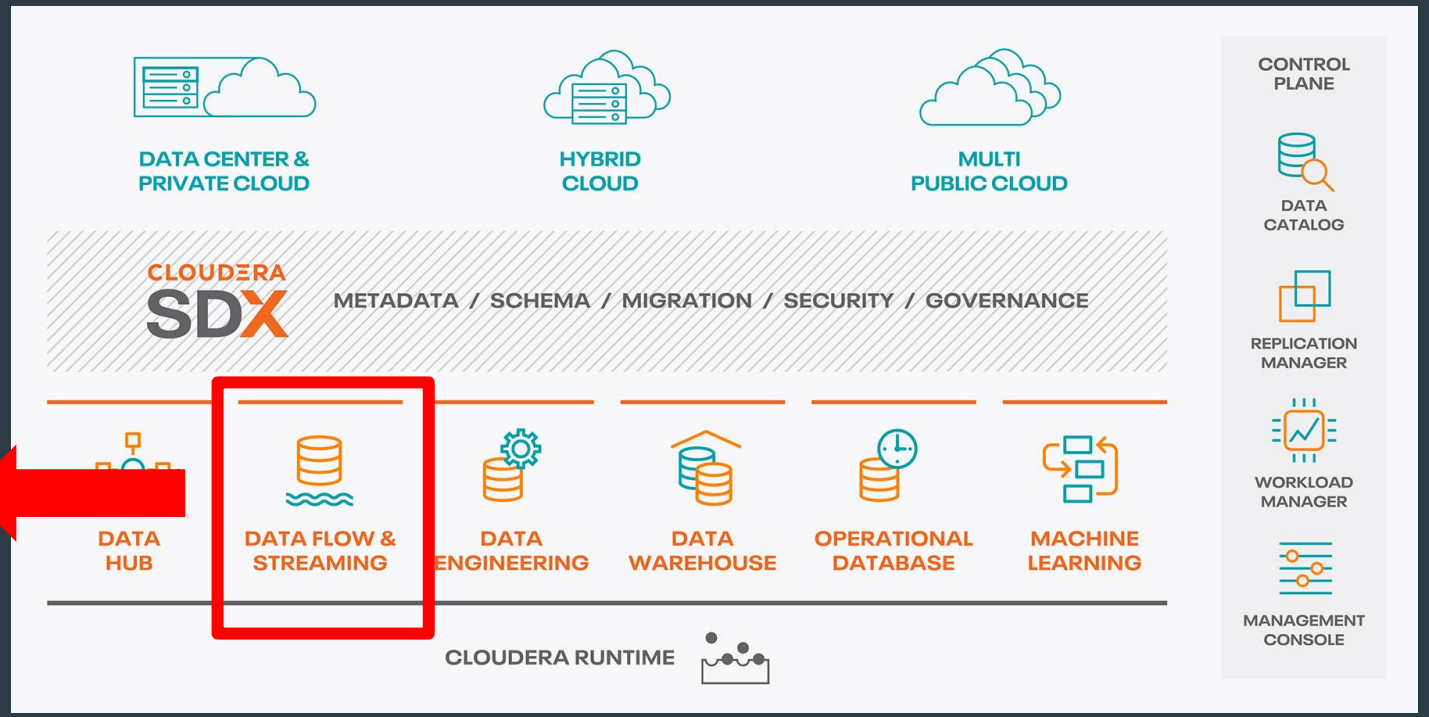
- 完成后续的数据流构建



Apache Kafka 介绍

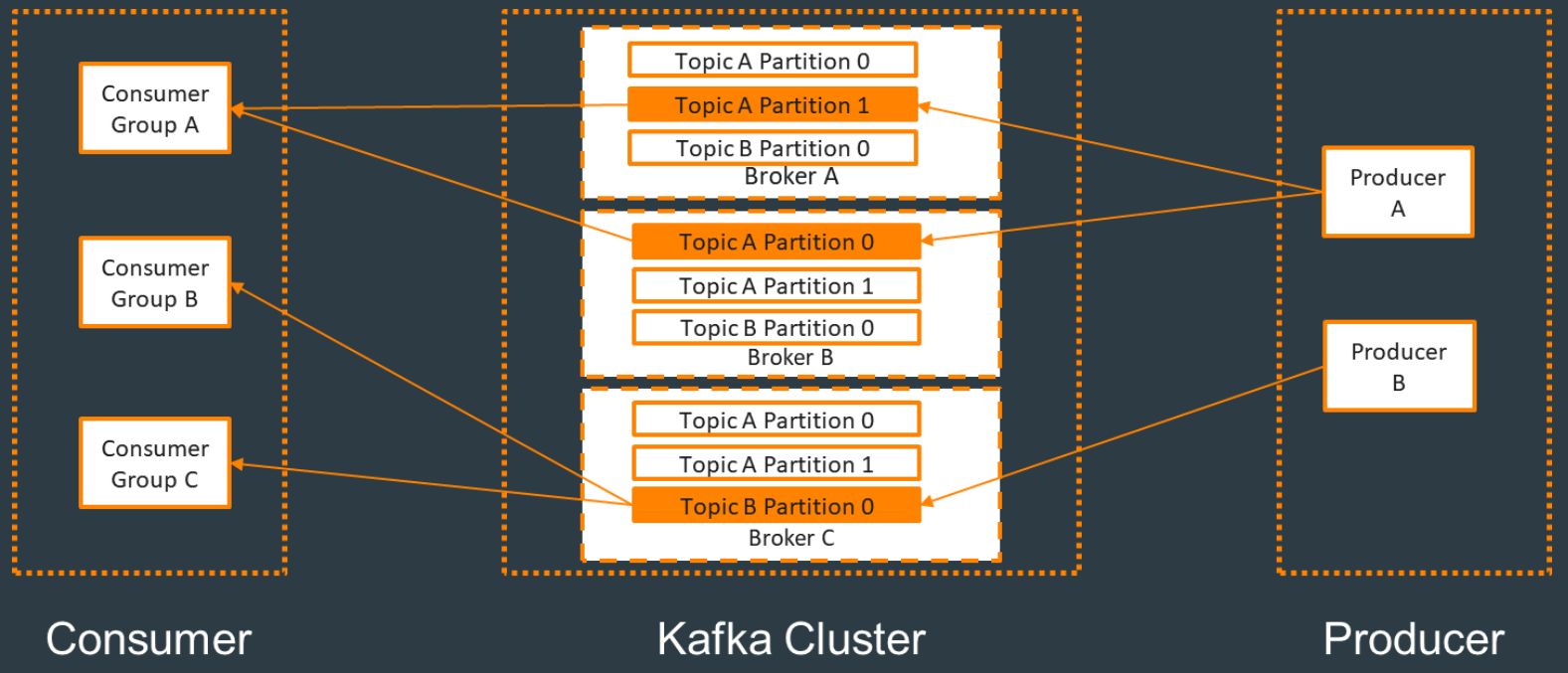
Cloudera DataFlow (CDF)

- Cloudera Flow



Apache Kafka介绍

Apache Kafka – 概览



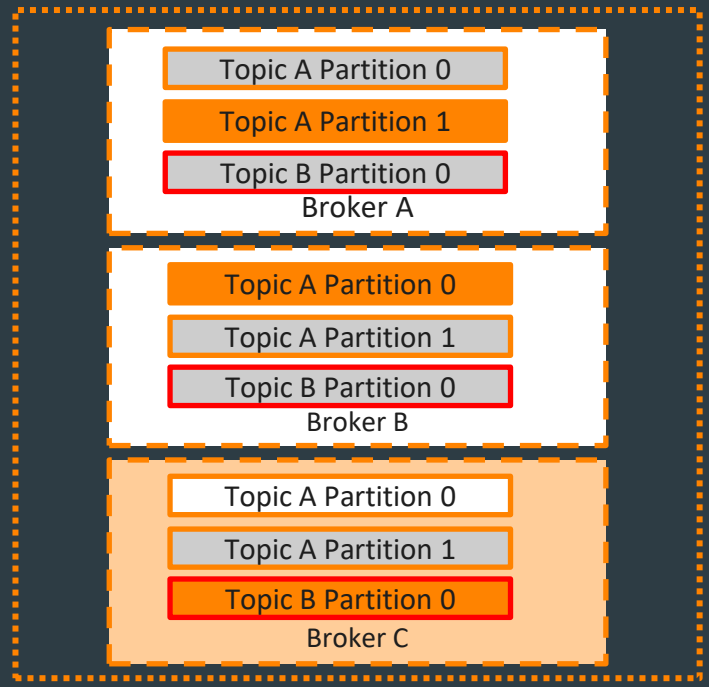
- 一个高吞吐，低延时，稳定可用的消息中间件及流数据实时处理系统

Apache Kafka介绍

Apache Kafka – 核心概念

- Broker
  - Topic 是发布/订阅的最小单位
  - Topic 可以被拆分成多个1 ~ N个Partition

Topic A 存在Partition 0， 1
Topic B 存在Partition 0

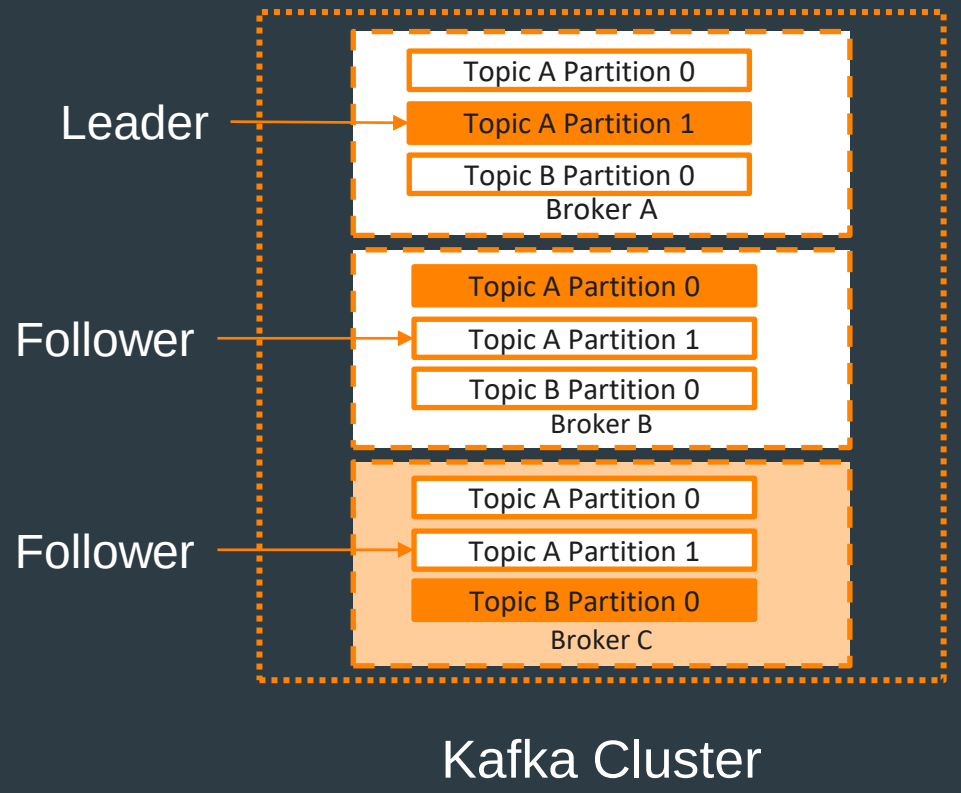


Kafka Cluster

Apache Kafka介绍

Apache Kafka – 核心概念

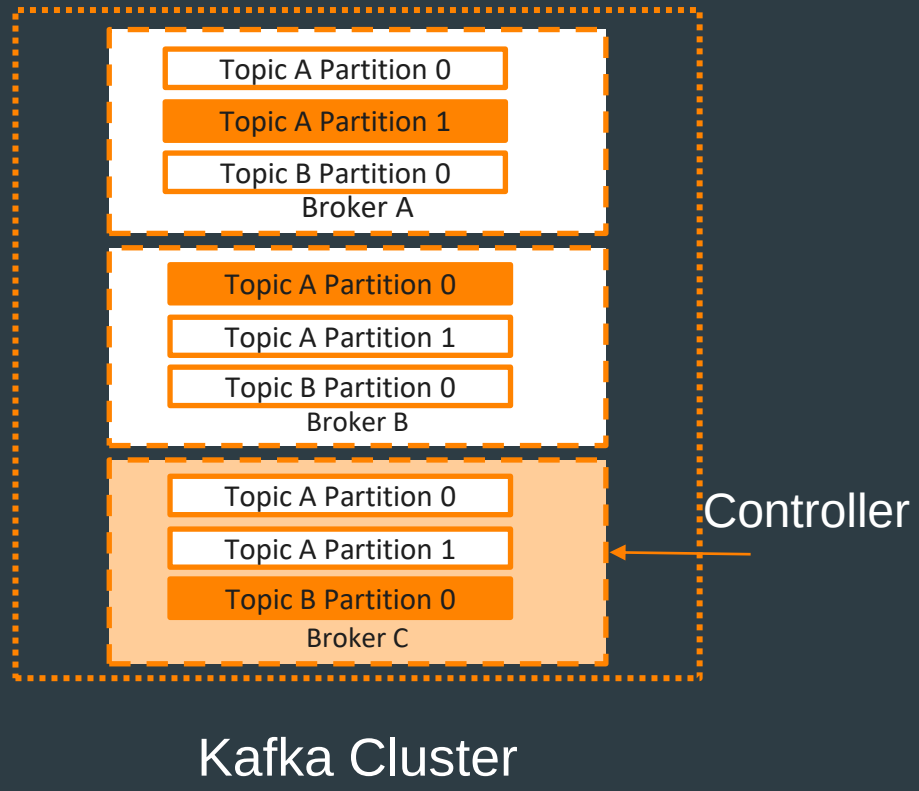
- Broker
  - Partition Replica
  - Partition Leader 和 Follower
  - ISR (In-Sync-Replica)



Apache Kafka介绍

Apache Kafka – 核心概念

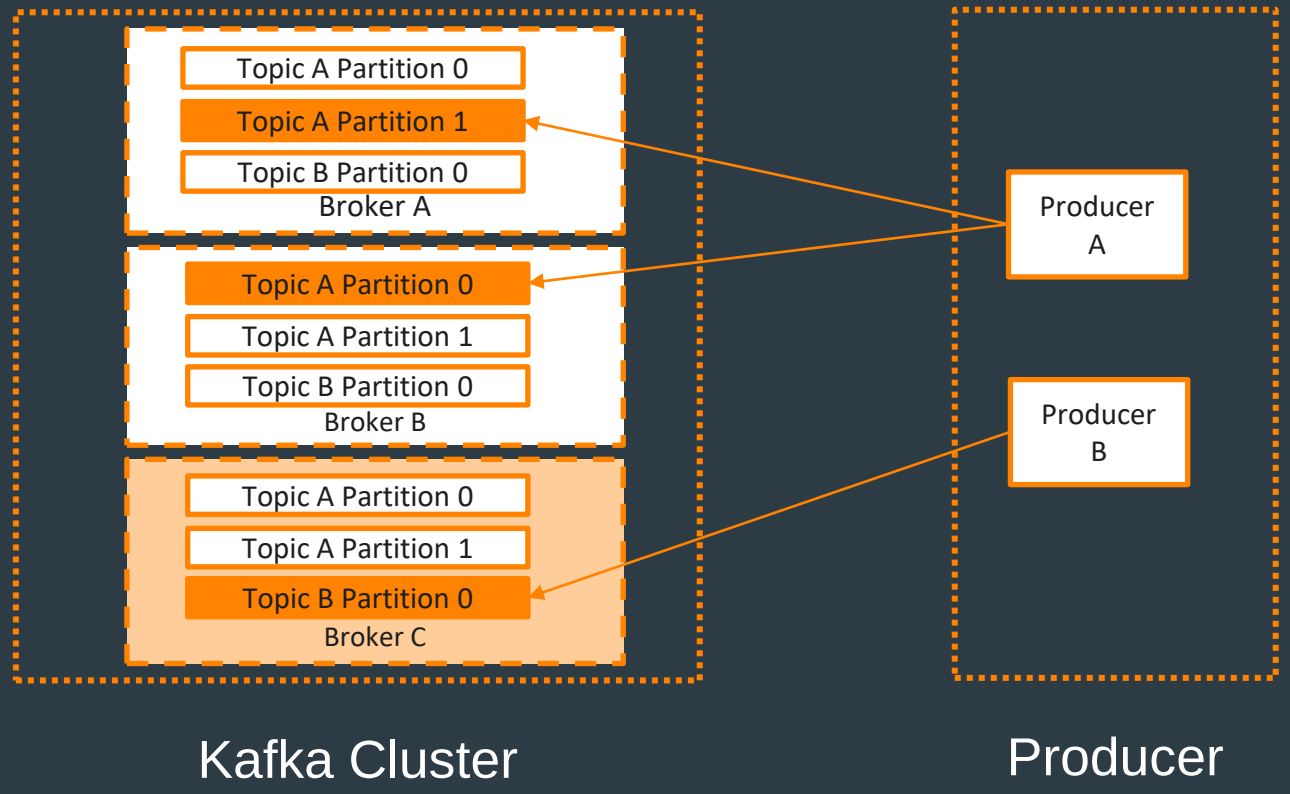
- Broker
  - Controller
  - 每一个Kafka 集群有且只有一个Controller
    - 负责监控其他Broker的活动
    - 维护Topic的CRUD
    - 选举新的Partition Leader
    - 同步新的Partition Leader至Broker



Apache Kafka介绍

Apache Kafka – 核心概念

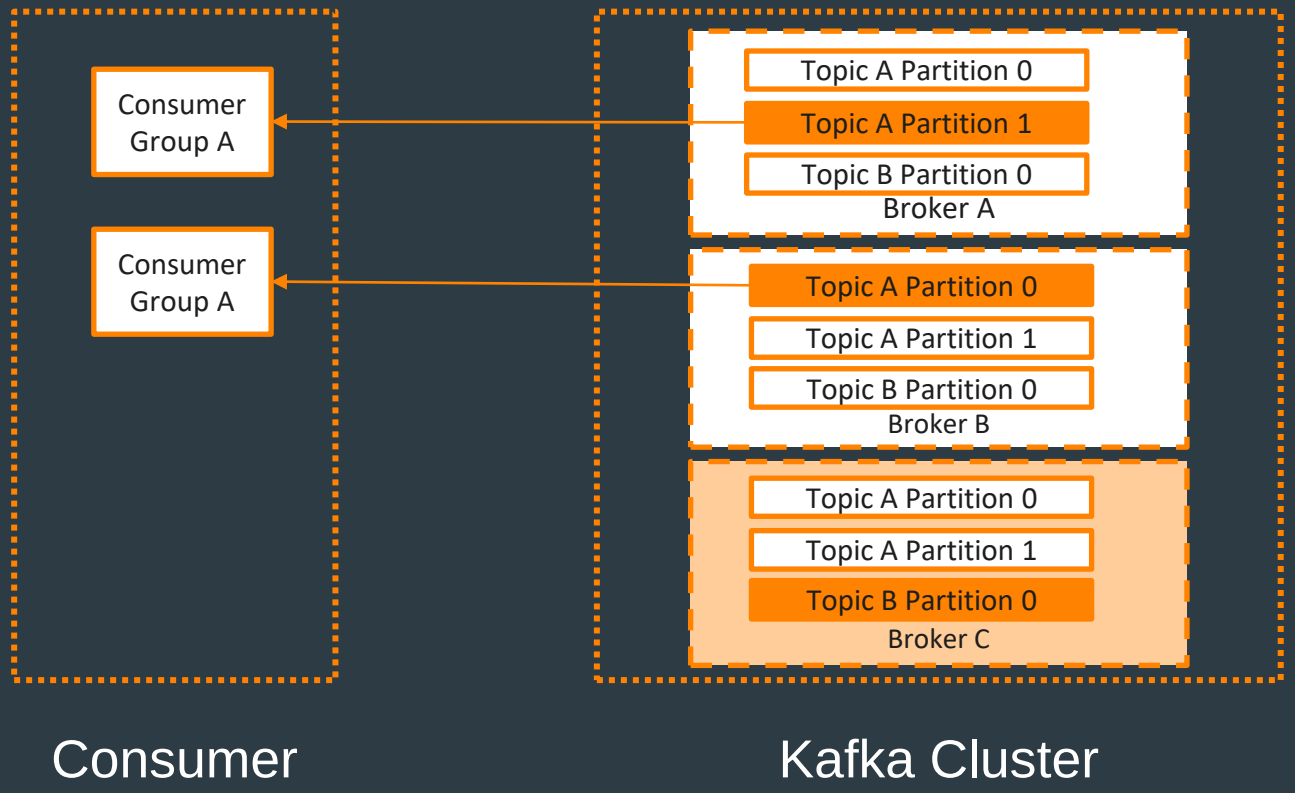
- Producer (客户端)
  - Partitioner
    - RoundRobin
    - Key
    - Custom
  - ACKs
  - Transaction



Apache Kafka介绍

Apache Kafka – 核心概念

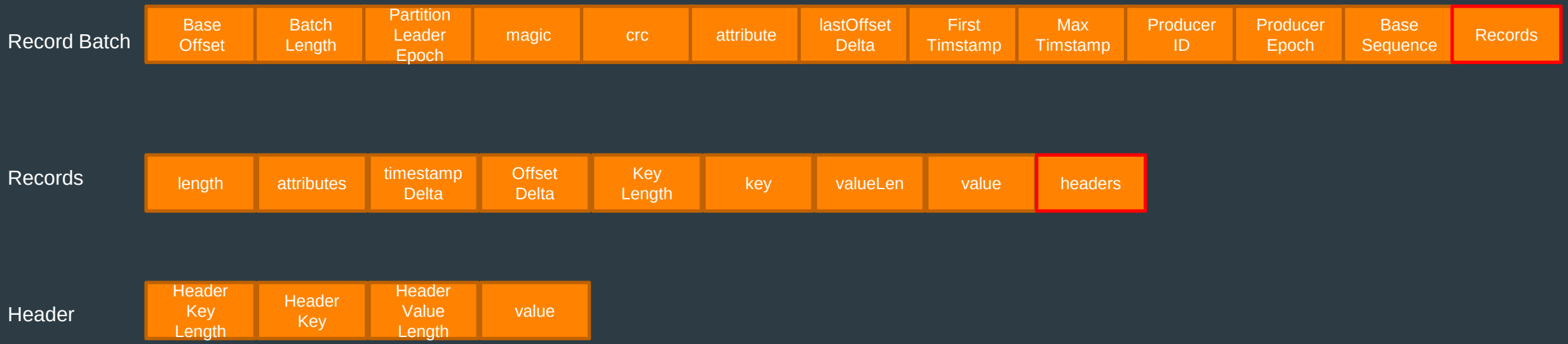
- Consumer (客户端)
  - Consumer Group
  - Consumer Rebalance
  - Consumer Lag
  - Read Committed



Apache Kafka介绍

Apache Kafka – 核心概念

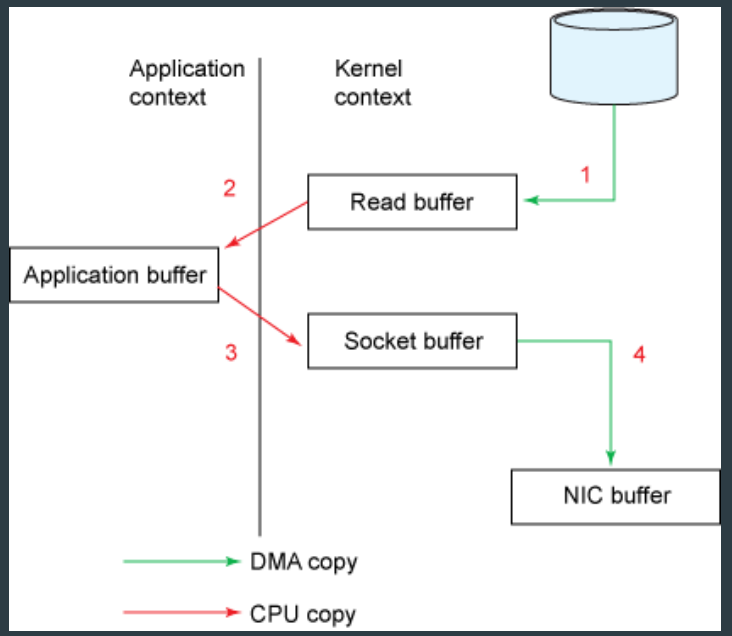
- Broker
  - Message

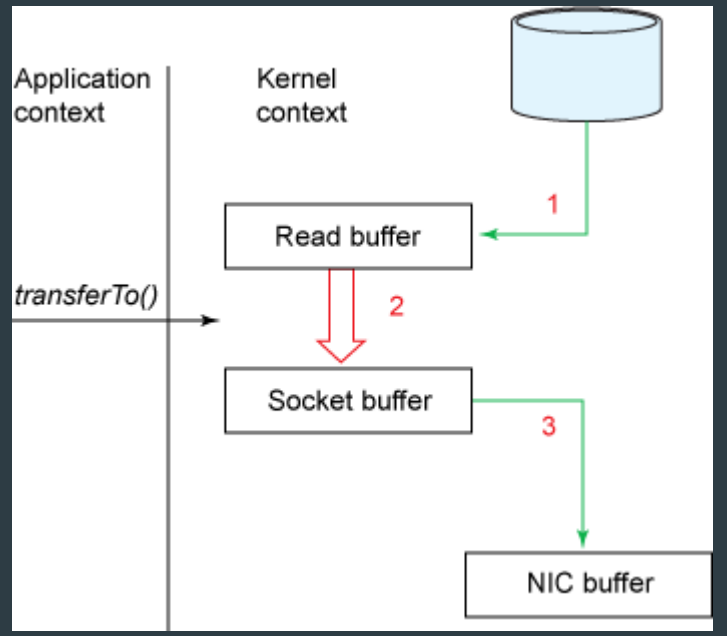


Apache Kafka介绍

Apache Kafka – 核心概念

- Kafka 为什么这么快
  - Zero Copy
  - Page Cache

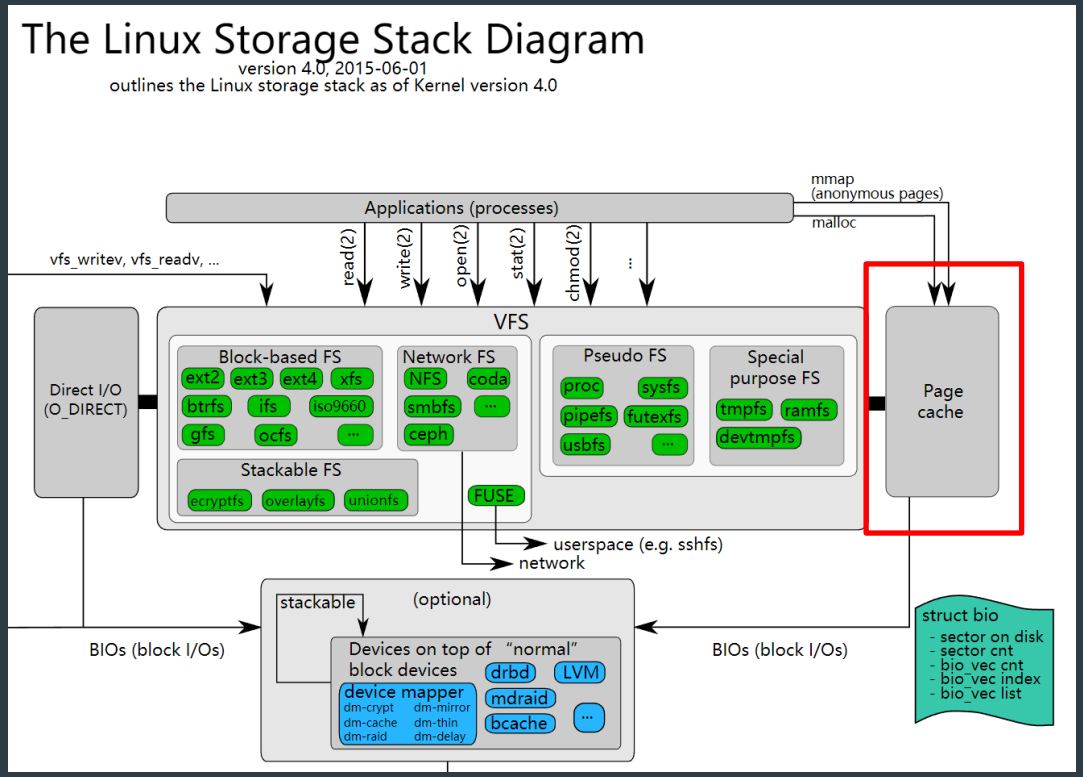




Apache Kafka介绍

Apache Kafka – 核心概念

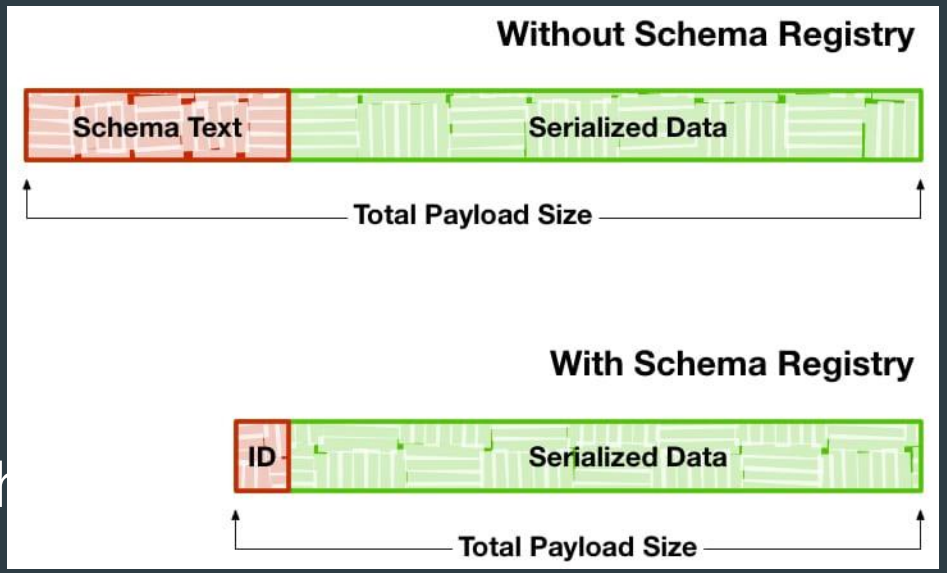
- Kafka 为什么这么快
  - Zero Copy
  - Page Cache



Schema Registry

Schema Registry概念

- 一个公共的Schema服务，用于保存或读取Schema
  - 1. 更好的元数据管理
    - 提供了可重用的Schema
    - 提供了Schema 版本管理
  - 2 运维效率
  - 避免了在消息本体中添加Schema
  - 消费者和生产者不需要完全统一固定Sch
  - 数据校验 Schema Validation



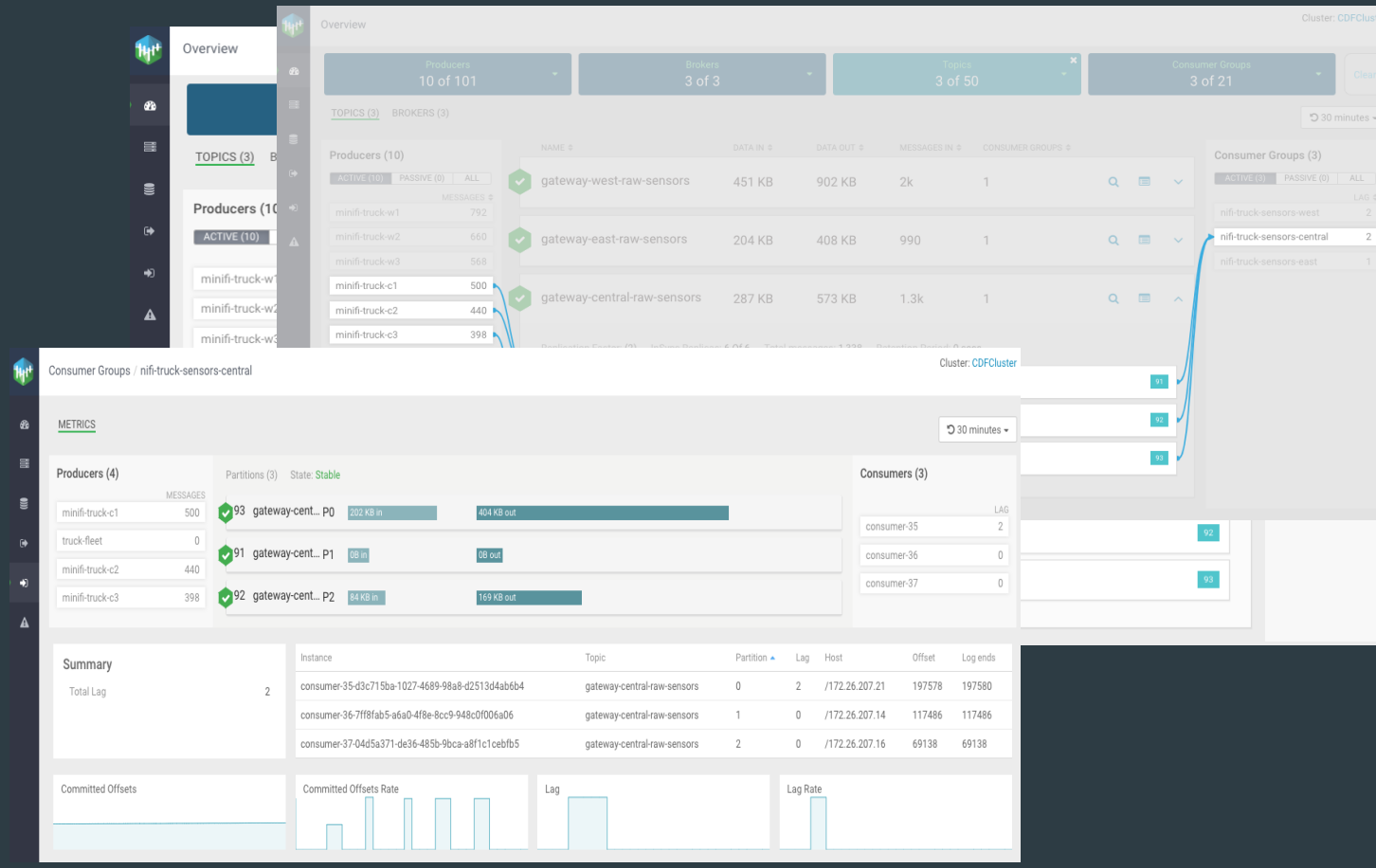
Stream Message Manager介绍

Stream Message Manager概念

- 交互式的可视化管理
- 监控所有的Kafka集群数据流向
- Producer/Consumer/Partition

的流向监控

- 自定义的过滤检索
- TOPIC CRUD
- Message 预览



动手实战

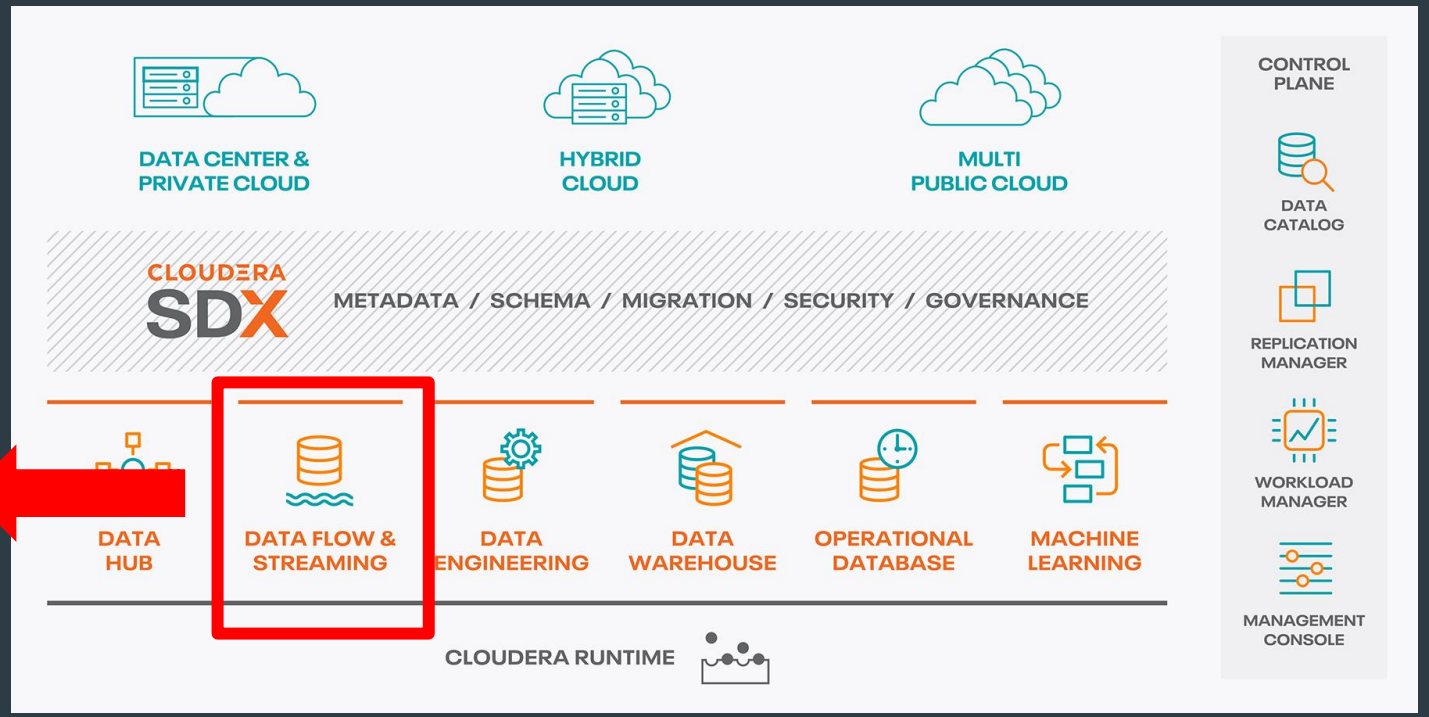
Lab 3 & 4

https://github.com/wangyong23/CDF-Workshop-2020

Apache Flink

Apache Flink 项目介绍

- Cloudera CSA



Apache Flink

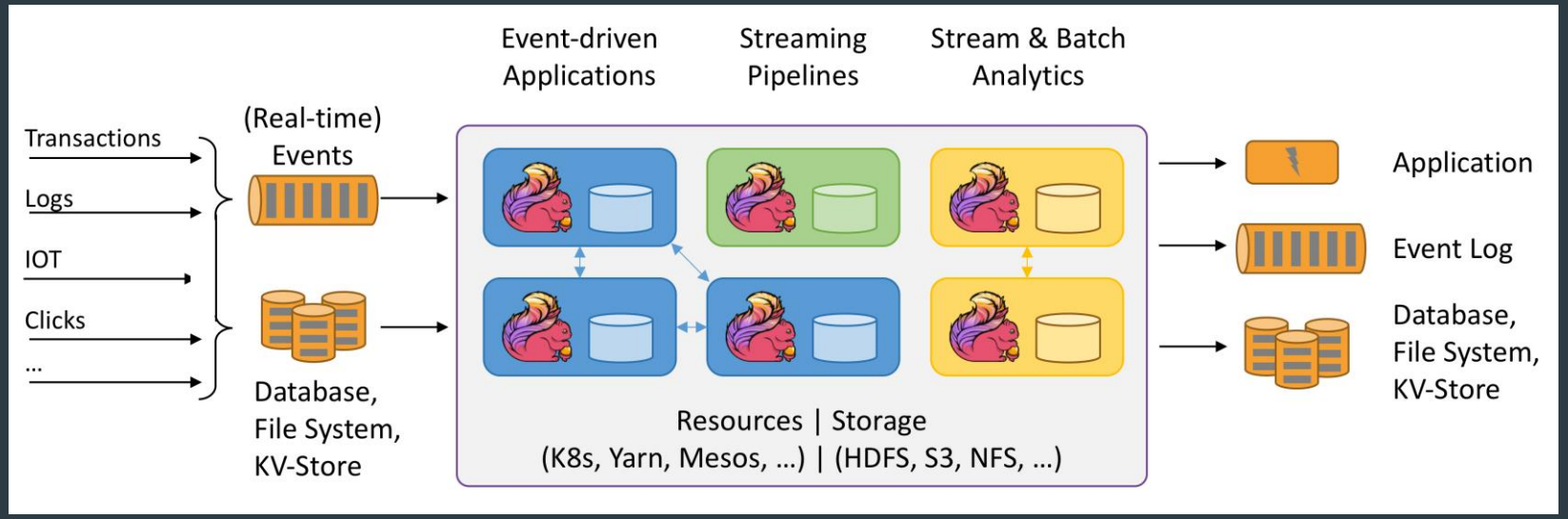
目录

- Apache Flink 项目介绍
- Apache Flink 基本能力介绍
- Apache Flink 适用场景介绍
- Cloudera CSA 模块介绍
- Apache Flink 未来展望

Apache Flink

Apache Flink 项目介绍

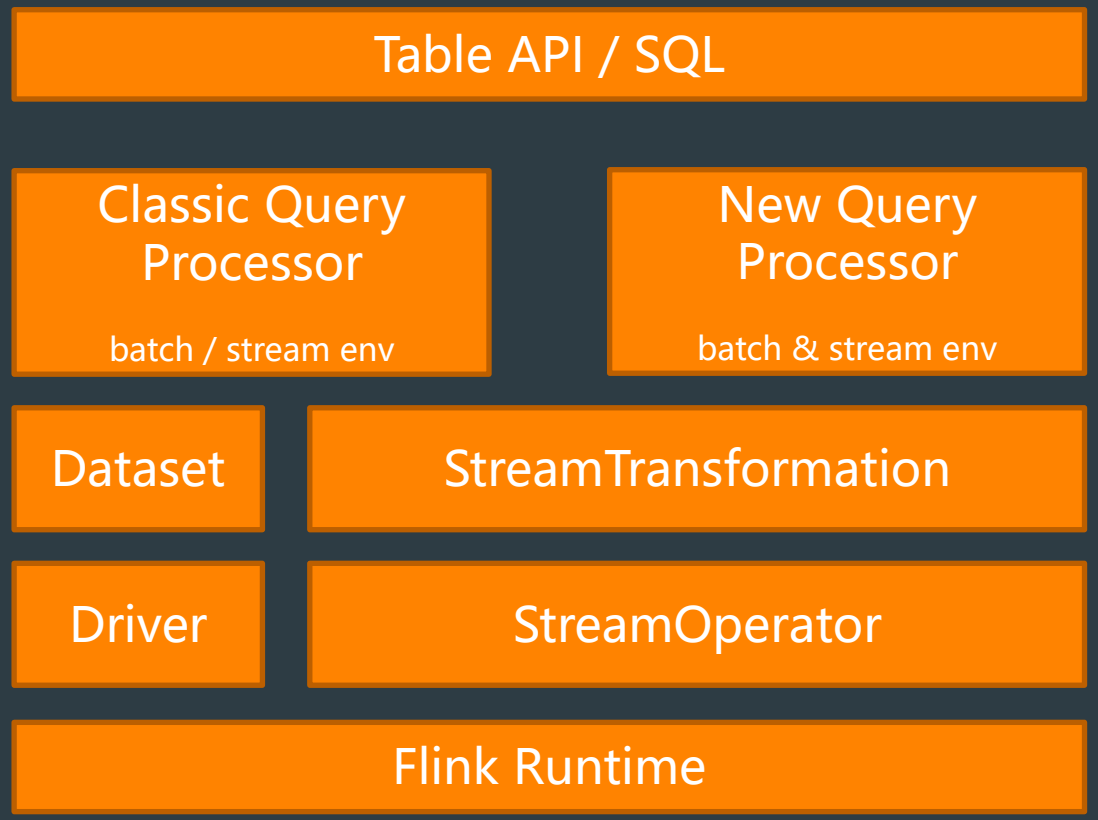
- Apache Flink 是一个用于解决实时数据处理的计算框架
  - 低延时
  - 高吞吐
  - 有状态
  - 分布式



Apache Flink

Apache Flink 项目介绍

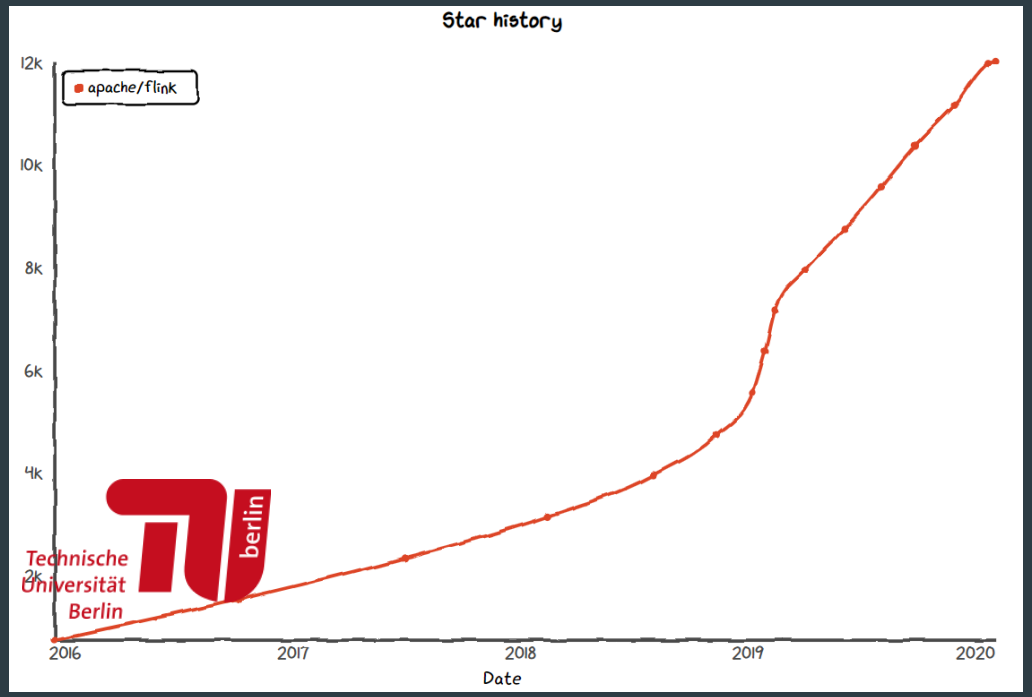
- Flink API
  - Stateful Stream Processing
  - DataStream/ DataSet API
  - Table API/ SQL



Apache Flink

Apache Flink 项目介绍

- 项目背景

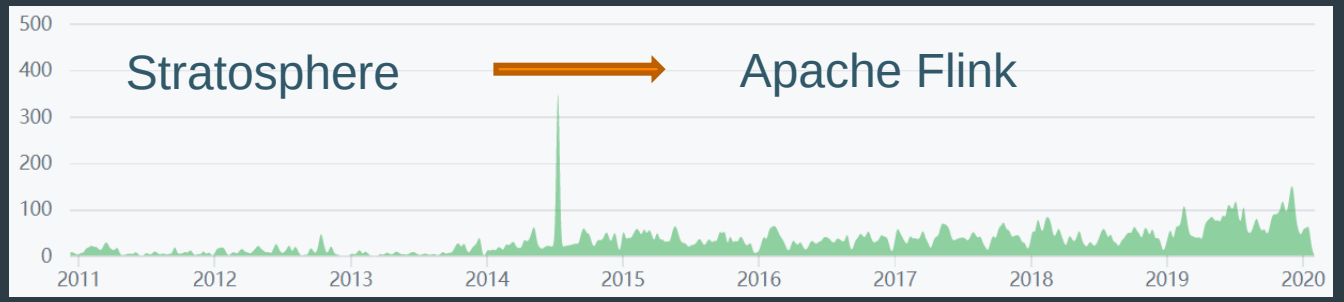


ververica

Alibaba.com

CLOUDERA

Technische Universität Berlin
berlin



Apache Flink

Apache Flink 项目介绍

Uber

实时数据处理，十亿消息处理和Stream SQL 应用

NETFLIX

3700+ 容器化的Flink实时处理服务平台。 1400+ 节点，2万+Core，数百个任务，3千亿事件每日，20TB状态数据

实时指标监控，实时ML, Stream SQL等
约1000+任务，10万+Core，10TB的状态

流式的反欺诈和实时分析平台
Cybersecurity

Apache Flink

Apache Flink 社区用户

https://flink.apache.org/poweredby.html

Apache Flink

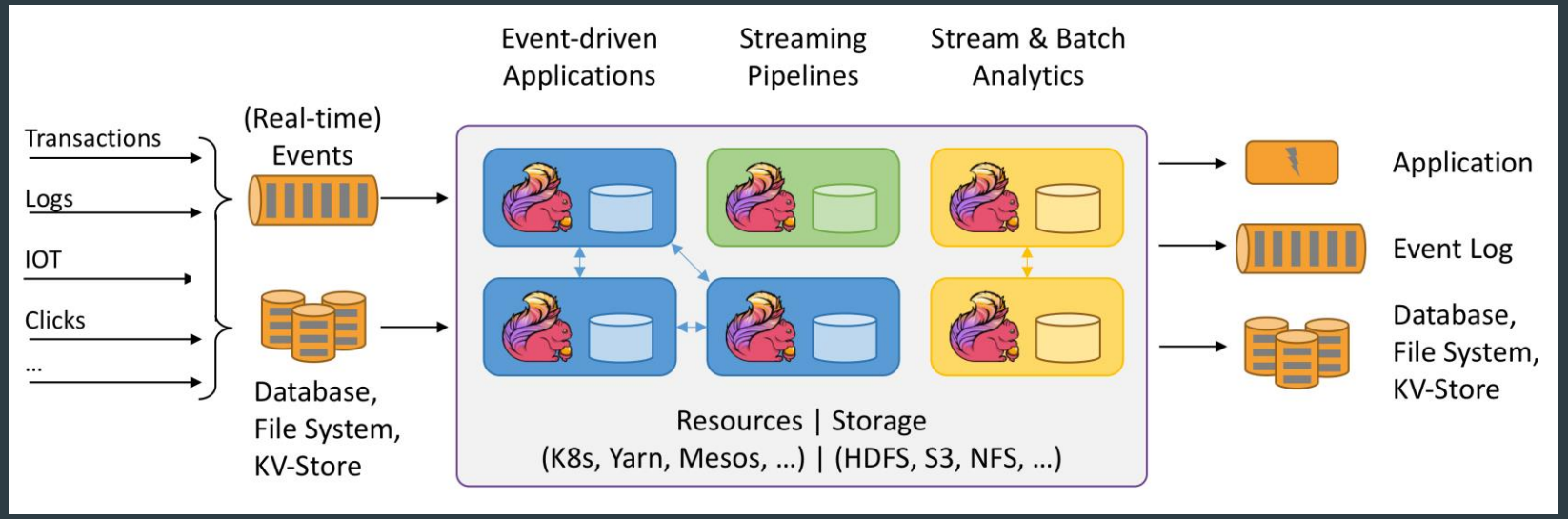
目录

- Apache Flink 项目介绍
- Apache Flink 基本能力介绍
- Apache Flink 适用场景介绍
- Cloudera CSA 模块介绍
- Apache Flink 未来展望

Apache Flink

Apache Flink 项目介绍

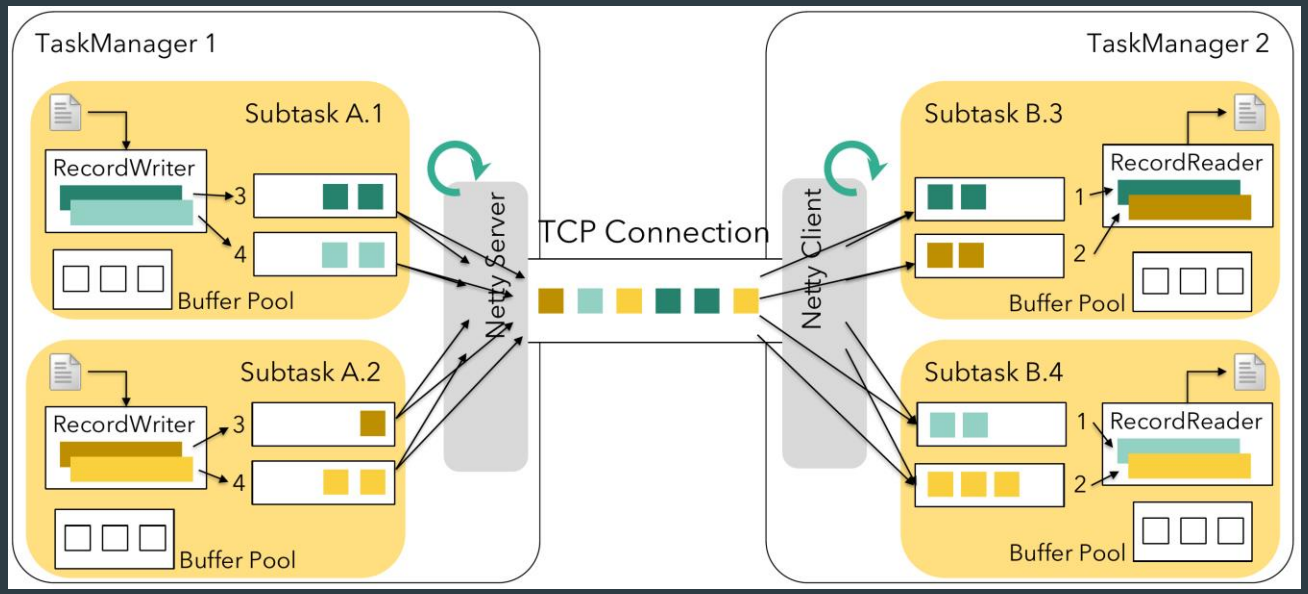
- Apache Flink 是一个用于解决实时数据处理的计算框架
  - 低延时
  - 高吞吐
  - 有状态
  - 分布式



Apache Flink

What is Apache Flink

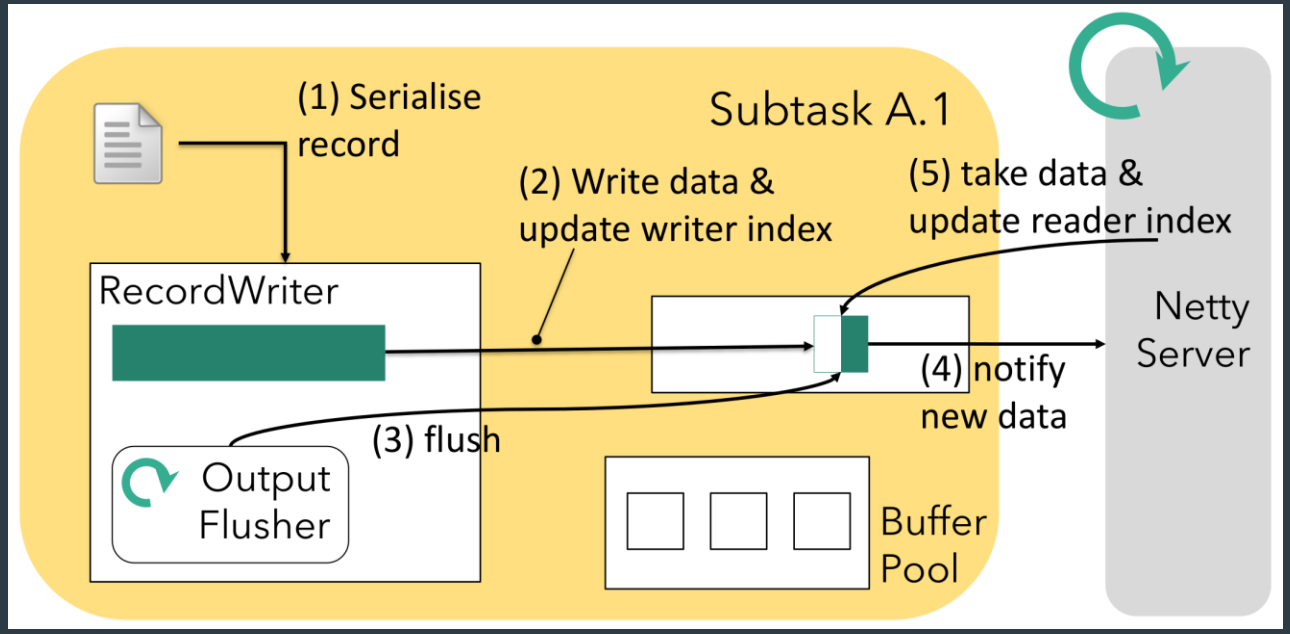
- 高吞吐
  - 每一个Task的subtask在TaskManager之间共享一个TCP链接



Apache Flink

What is Apache Flink

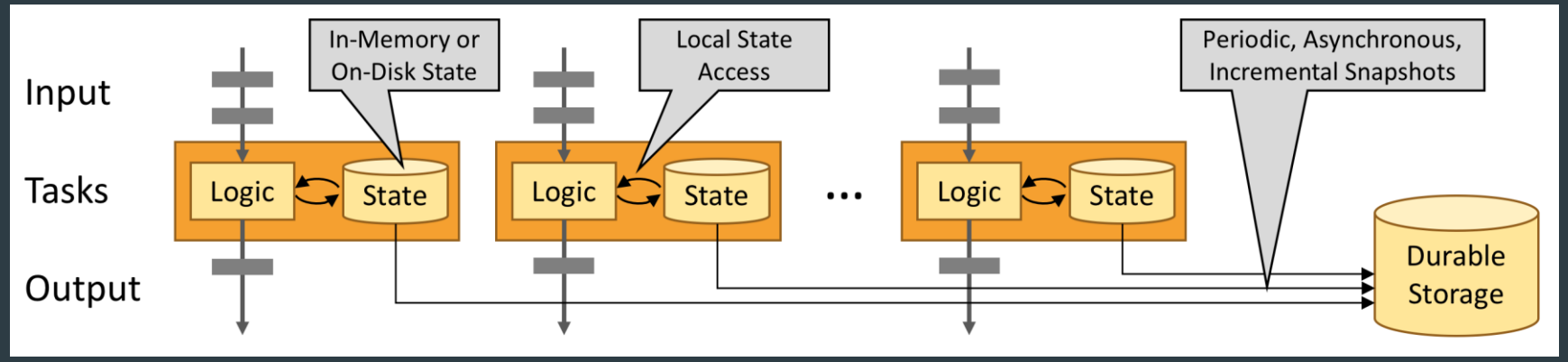
- 低延时
  - 可配置的缓冲区定时推送
  - setBufferTimeout = -1,0 ,x ms



Apache Flink

What is Apache Flink

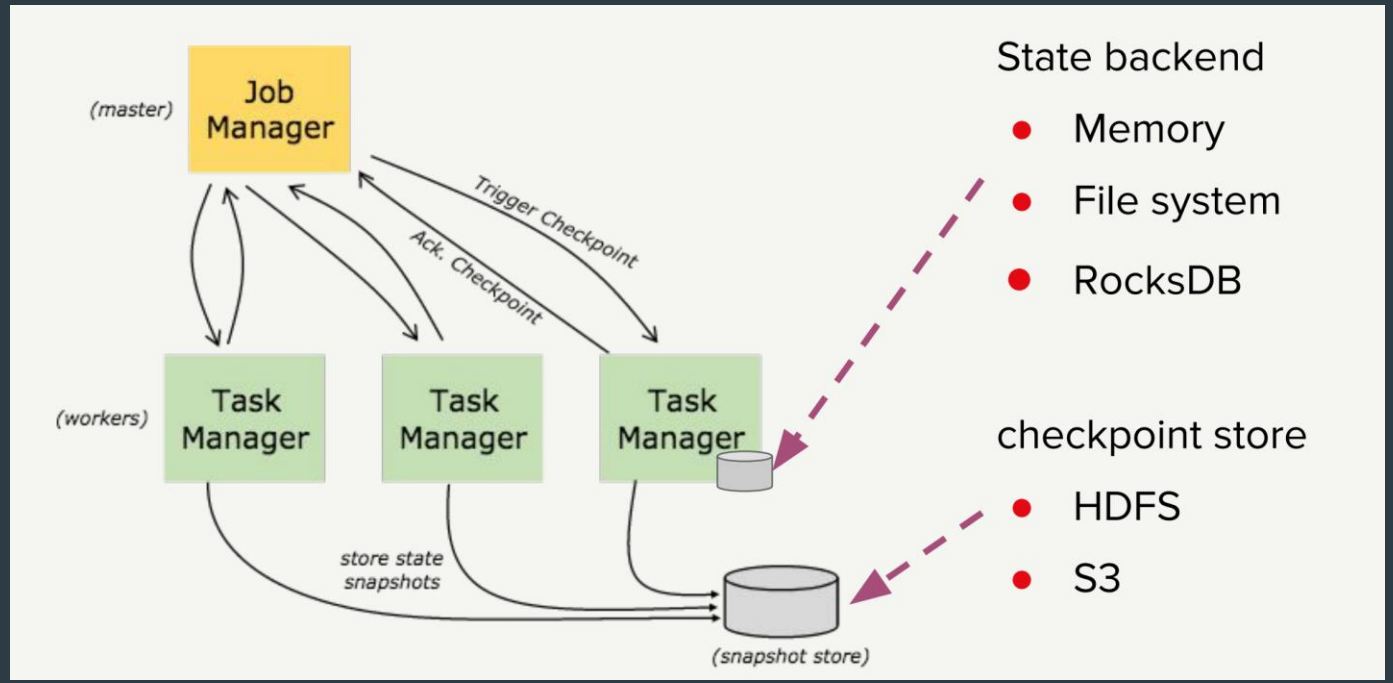
- 有状态



Apache Flink

What is Apache Flink

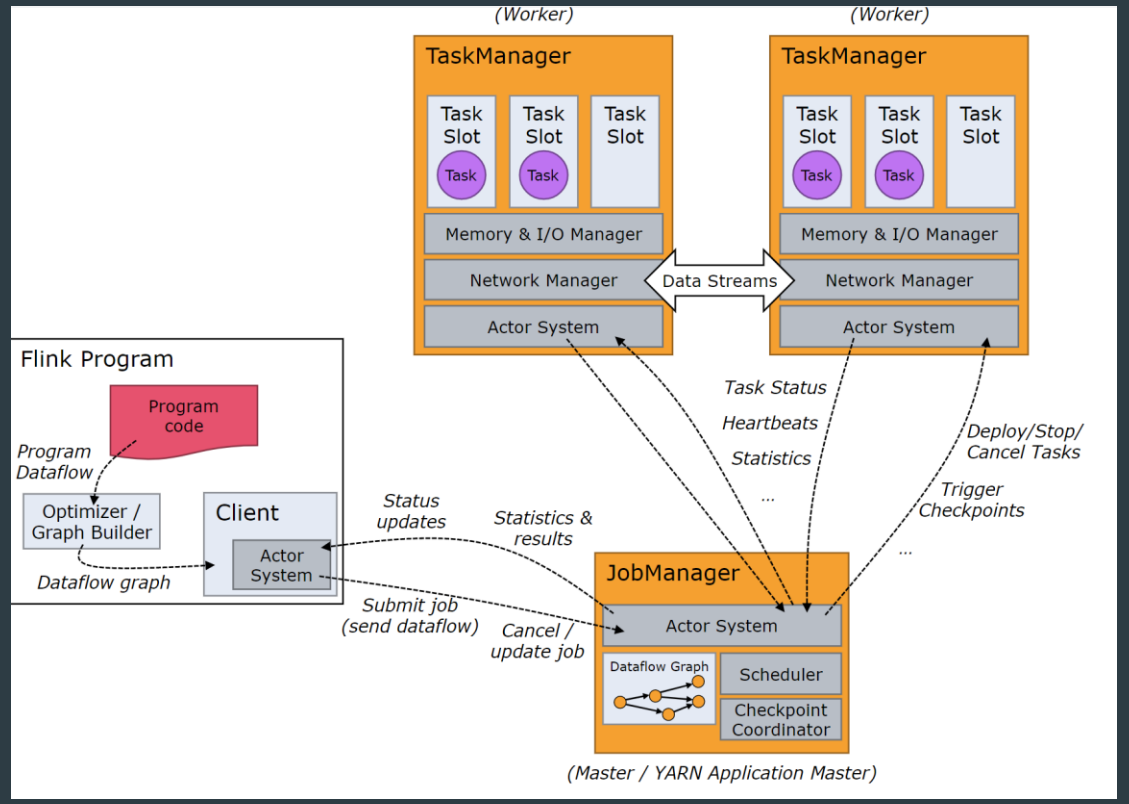
- 有状态
  - 内存
  - 文件
  - Rocks DB



Apache Flink

What is Apache Flink

- 分布式



Apache Flink

目录

- Apache Flink 项目介绍
- Apache Flink 基本能力介绍
- Apache Flink 适用场景介绍
- Cloudera CSA 模块介绍
- Apache Flink 未来展望

Apache Flink

What is Apache Flink

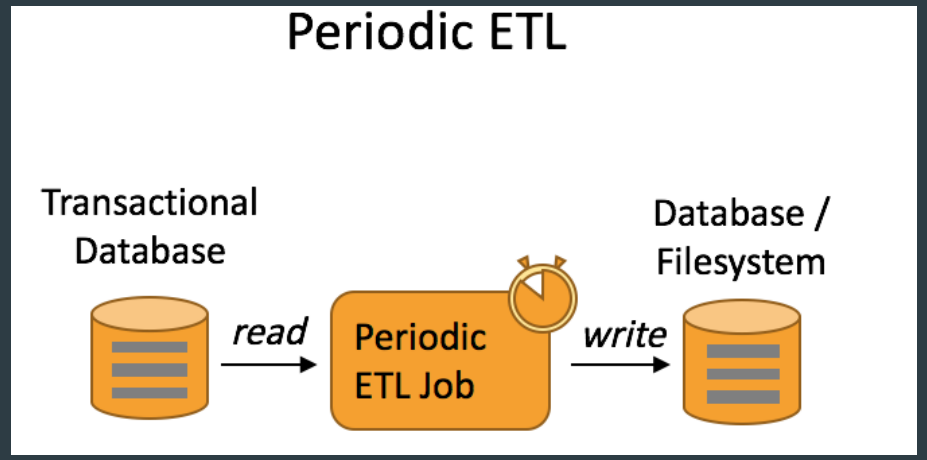
- 适用场景
  - 流式ETL
  - 实时数据分析
  - 事件驱动型的应用

有状态的实时计算框架

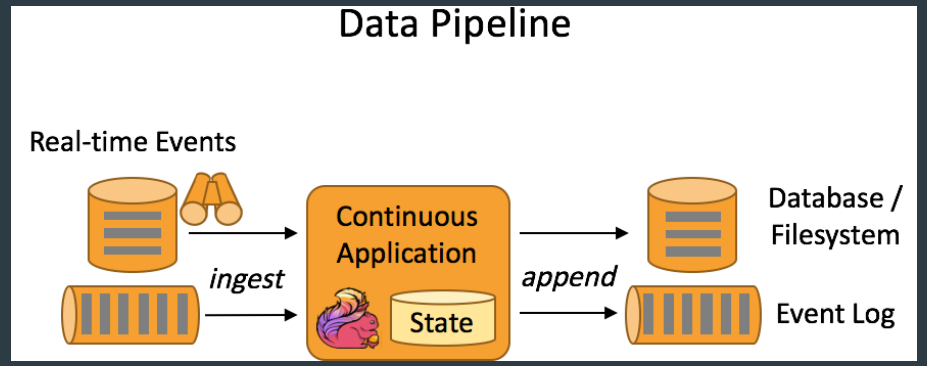
Apache Flink

适用场景 – 流式ETL

- 传统的ETL
  - 通过周期性的调用ETL脚本完成批处理的作业



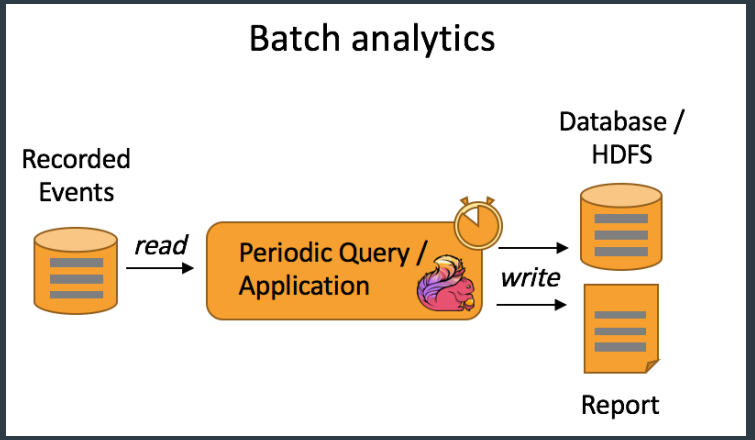
- 流式ETL
  - 低延时
  - 持续的数据转换



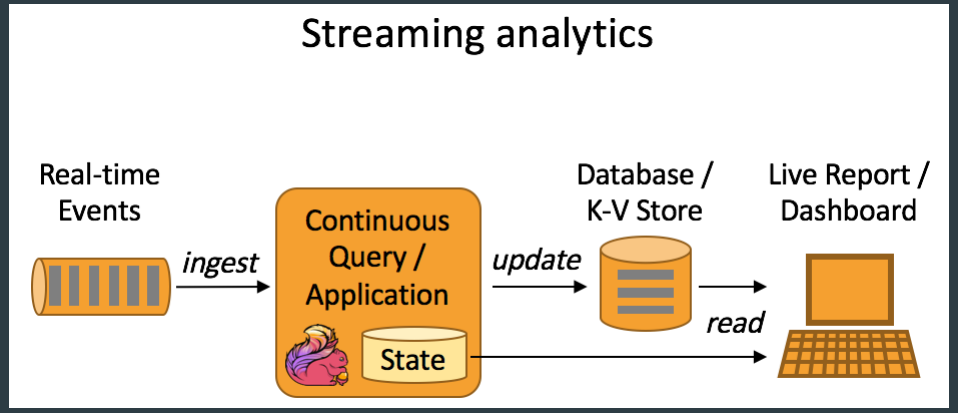
Apache Flink

适用场景 – 实时数据分析

- 批处理性的分析
  - ad-hoc 查询
  - 查询的变化远大于数据的变化
  - 交互性的分析和预分析



- 实时数据分析
  - 数据的变化远大于查询的变化
  - 实实时/低延迟的查询结果
  - 不需要Lambda架构的系统设计

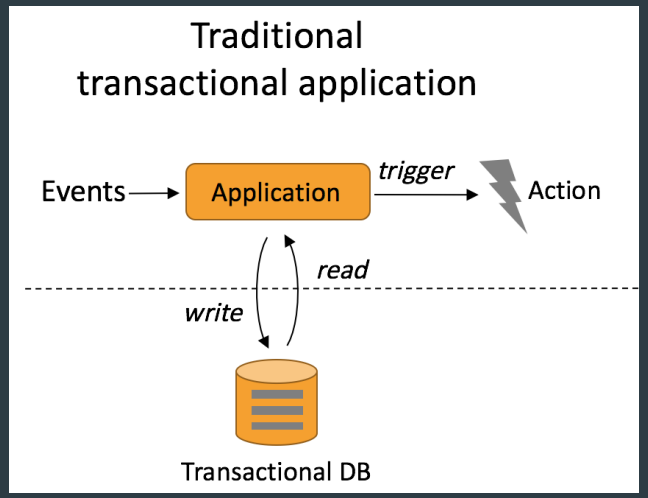


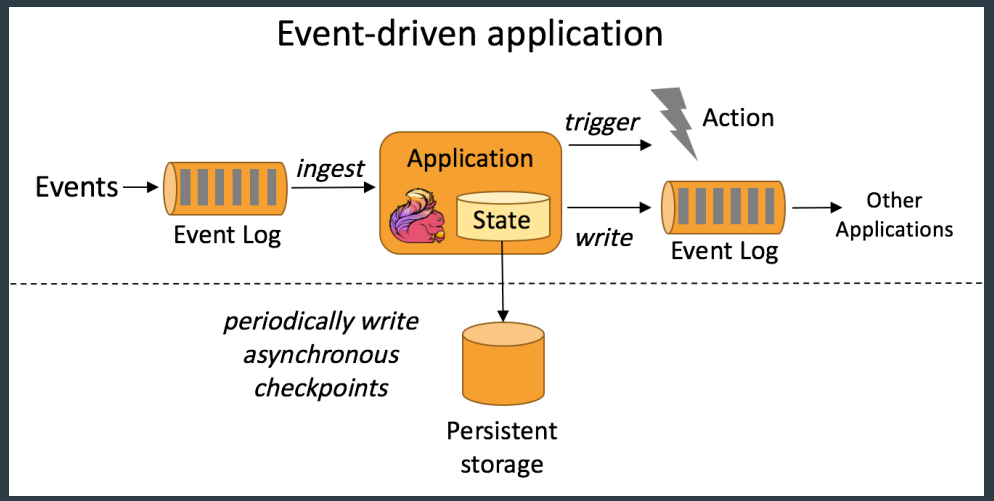
Apache Flink

适用场景 – 事件驱动型的应用

- 传统的事务处理应用
  - 计算/存储松散耦合
  - 对新到事件进行处理及响应
  - 应用状态保存在数据库中

- 事件驱动的处理应用
  - 状态保存在本地
  - 通过Checkpoint机制保证一致性
  - 计算/存储紧耦合（微服务架构）
  - 高可扩展性





Apache Flink

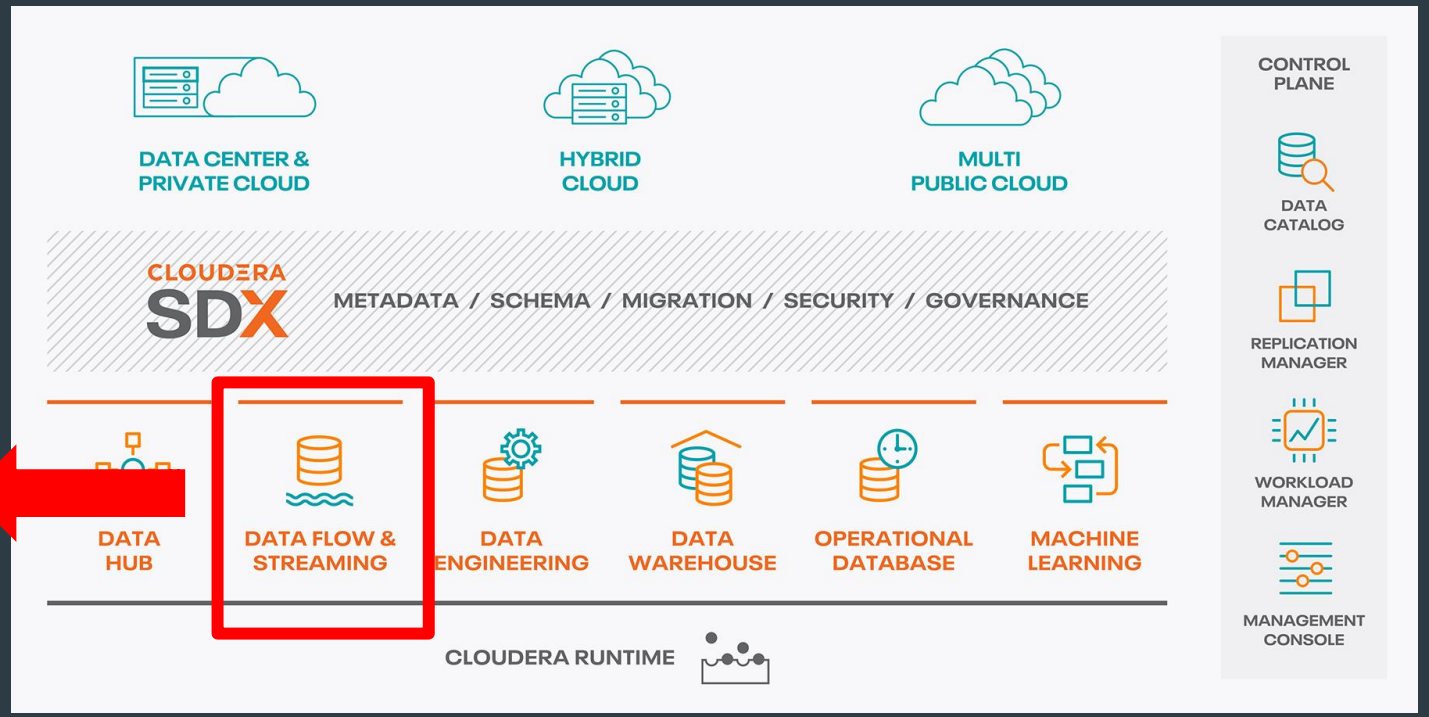
目录

- Cloudera CSA 模块介绍

Apache Flink

Cloudera CSA 模块介绍

- Cloudera CSA 模块介绍

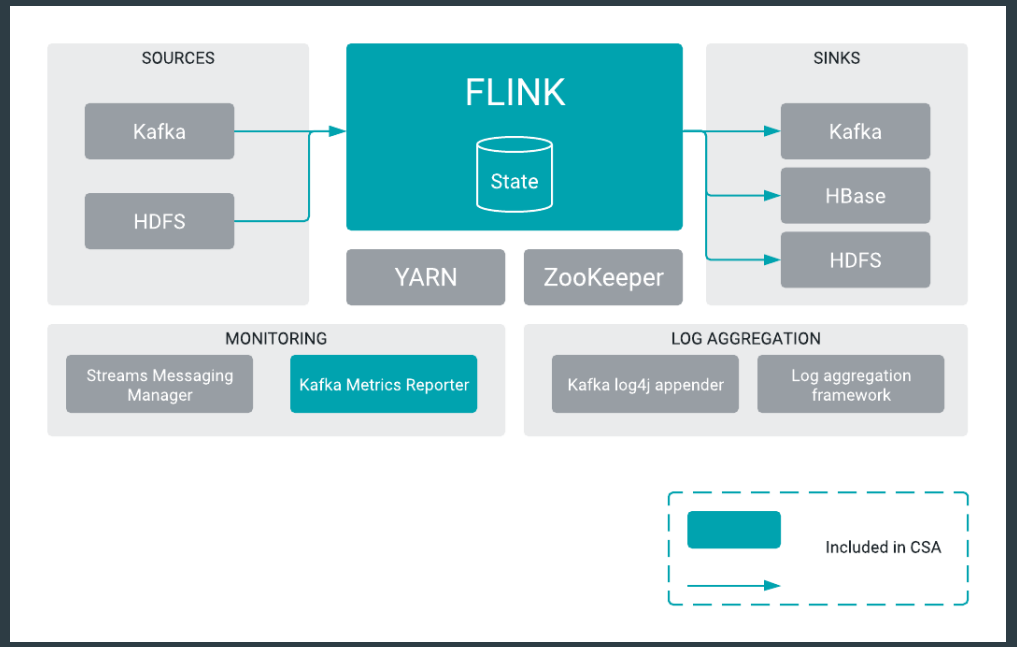


Apache Flink

Cloudera CSA 模块介绍

- Cloudera CSA 模块介绍
  - StableAPI：Flink Datastream API
  - Flink Kafka Connector
  - Flink HDFS Connector
  - Flink HBASE Connector

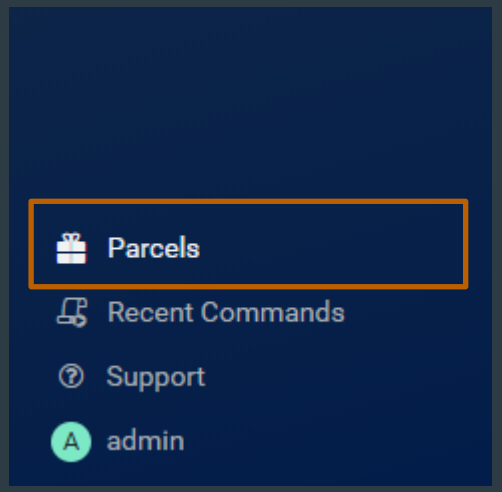
  - Flink Kafka Metrics Reports
  - Cloudera Schema Registry

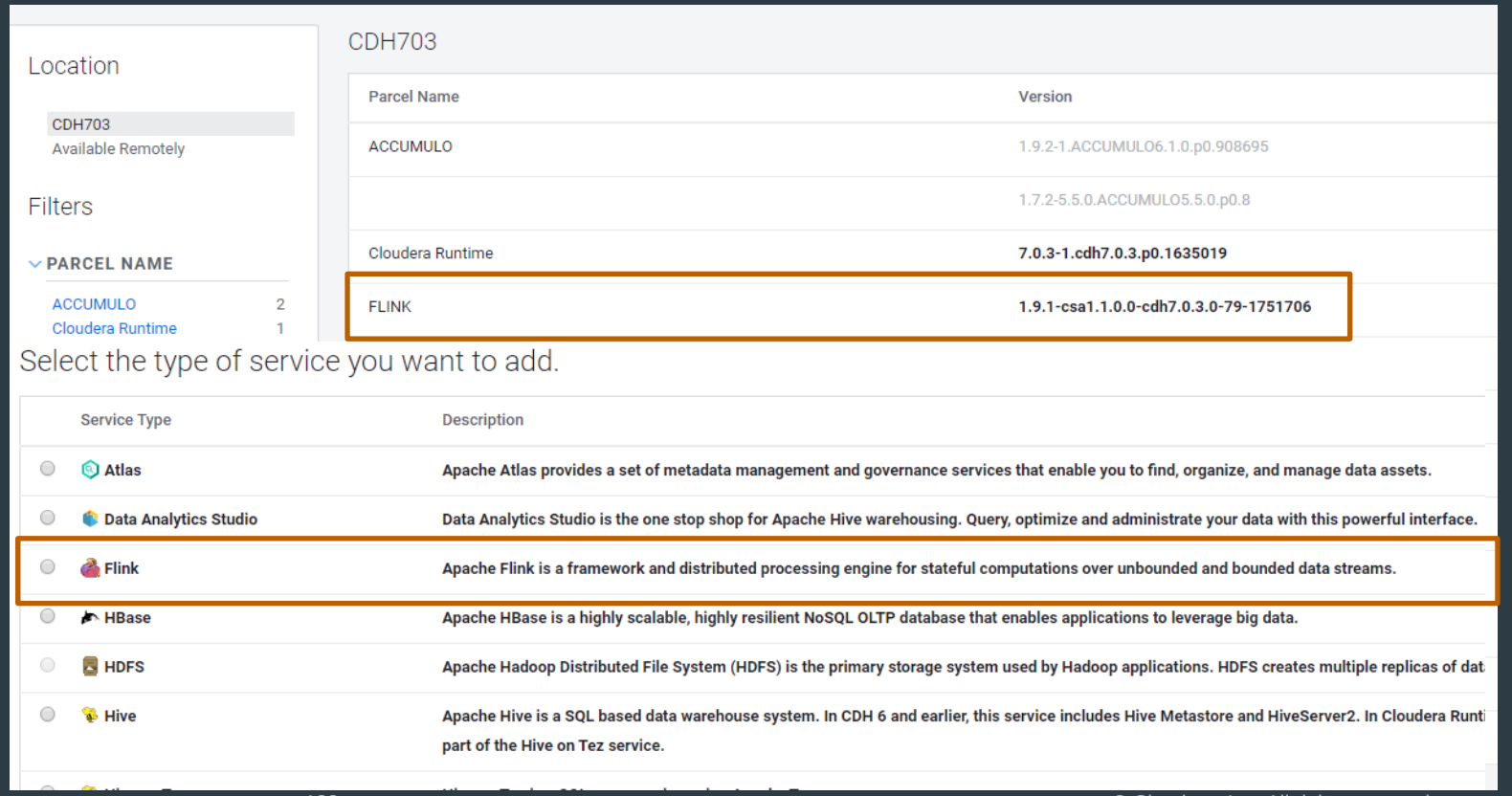


Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 模块下载



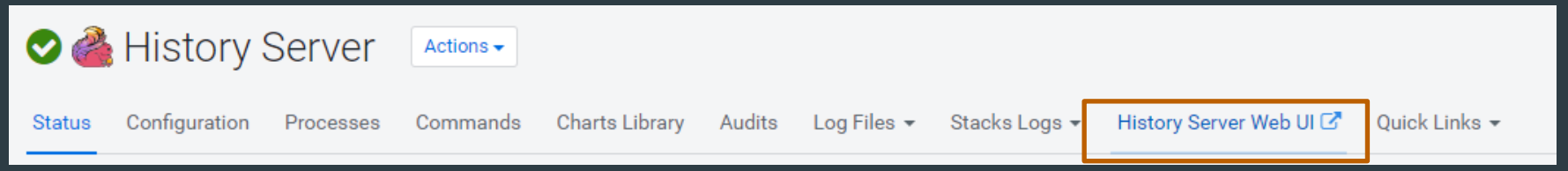


Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 模块安装
  - History Server
  - Gateway

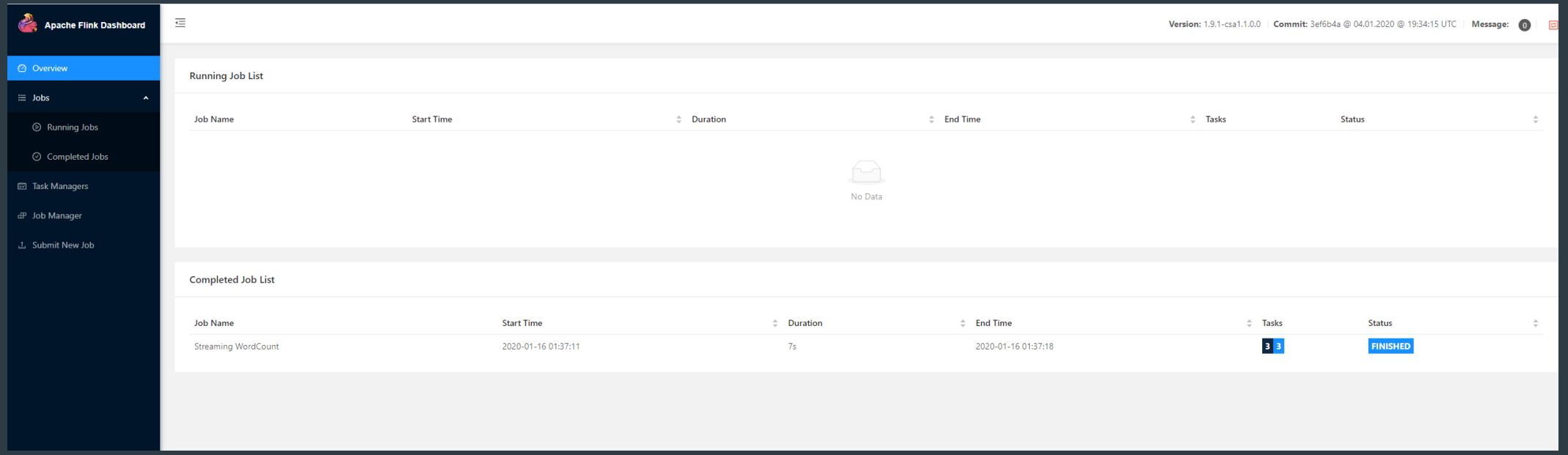
	Status	Role Type	State	Hostname
☐	✔	History Server	Started	cdpdc703.cloudera.com
☐	●	Gateway	N/A	cdpdc703.cloudera.com



Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 使用



Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 使用
  - 可选的使用ClouderaSchemaRegistry服务

```
<repository>
<id>snapshots</id>
<name>libs-release</name>
<url>https://repository.cloudera.com/cloudera/libs-release</url>
</repository>

<dependency>
    <groupId>org.apache.flink</groupId>
    <artifactId>flink-connector-kafka_2.11</artifactId>
    <version>${flink.version}</version>
</dependency>
<dependency>
    <groupId>org.apache.flink</groupId>
    <artifactId>flink-avro-cloudera-registry</artifactId>
    <version>1.9.1-csa1.1.0.0</version>
</dependency>
<dependency>
    <groupId>org.apache.avro</groupId>
    <artifactId>avro-compiler</artifactId>
    <version>${avro.version}</version>
</dependency>
<dependency>
    <groupId>org.apache.avro</groupId>
    <artifactId>avro-maven-plugin</artifactId>
    <version>${avro.version}</version>
</dependency>
<dependency>
    <groupId>org.apache.avro</groupId>
    <artifactId>avro</artifactId>
    <version>${avro.version}</version>
</dependency>
```

Apache Flink

下载/安装/使用 Flink

- Flink DataStream Program
  - Obtain an execution environment,
  - Load/create the initial data,
  - Specify transformations on this data,
  - Specify where to put the results of your computations,
  - Trigger the program execution

Apache Flink

下载/安装/使用 Flink

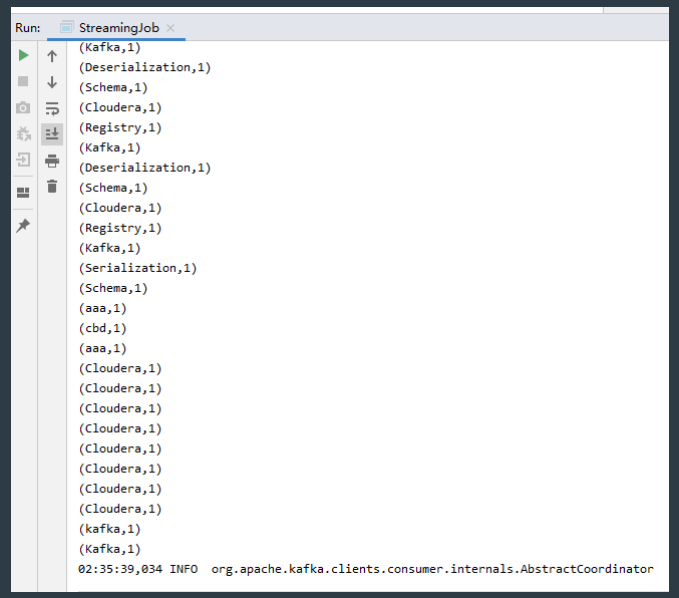
- Cloudera CSA 使用
  - 完成一个简单的Kafka Word Count

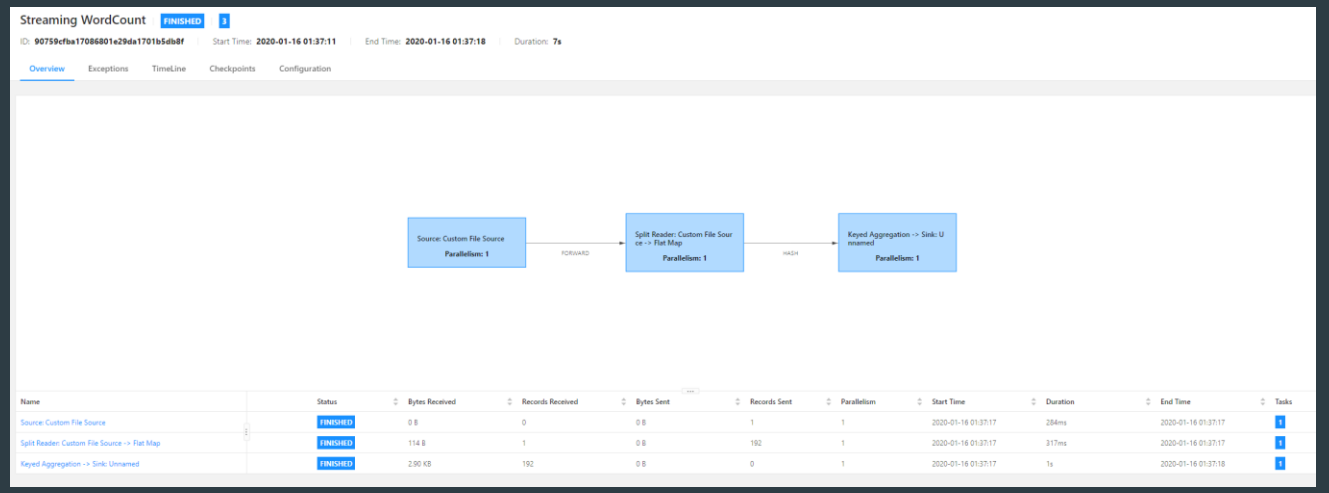
```
public static void main(String[] args) throws Exception {
    // set up the streaming execution environment
    final StreamExecutionEnvironment env = StreamExecutionEnvironment.getExecutionEnvironment();
    //env.setStreamTimeCharacteristic(TimeCharacteristic.ProcessingTime);
    //env.enableCheckpointing(5000);
    env.setParallelism(1);

    //ParameterTool params = ParameterTool.fromArgs(args);

    Properties properties = new Properties();
    properties.setProperty("bootstrap.servers", "192.168.1.64:9092");
    properties.setProperty("group.id", "testgroupA");

    FlinkKafkaConsumer<String> kafkaConsumer = new FlinkKafkaConsumer<>( topic: "test", new SimpleStringSchema(), properties);

    kafkaConsumer
            .setStartFromEarliest()
            .setCommitOffsetsOnCheckpoints(false);


    DataStream<Tuple2<String, Integer>> stream = env.addSource(kafkaConsumer) DataStreamSource<String>
            .flatMap(new Splitter()) SingleOutputStreamOperator<Tuple2<String, Integer>>
            .keyBy( ...fields: 0) KeyedStream<Tuple2<String, Integer>, Tuple>
            .sum( positionToSum: 1) SingleOutputStreamOperator<Tuple2<String, Integer>>
            ;

    stream.print();

    env.execute( jobName: "Kafka WordCount");
```

Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 使用
  - 本地测试或打包并提交Yarn集群运行





Apache Flink

Apache Flink 未来展望

- Apache Flink 未来展望

PAAS on CDP Public Cloud	Ranger集成	Atlas集成	Flink部署管理
• 快速构建Flink集群 • 提供可扩容和容灾的Flink应用管理 • 自动化的Kerberos/SSL/TLS配置 • 与Ranger和Atlas集成	• 在Ranger中提供Flink任务的授权和鉴权管理插件 例如，任务的提交权限，停止权限，和Kill权限等等	• 在Flink中提供元数据，血缘数据的捕捉的插件 • 例如，数据源，数据目标，业务元数据映射等等	• 任务以及任务Jar管理 • Checkpoint/Save point 管理 • A / B测试管理 • 状态演进管理（State Schema evolution）

Cloudera DataFlow (CDF)

CDF IN Action 整体架构

MiNiFi

终端数据收集

US-West Fleet

C++ agent

minifi

US-Central Fleet

C++ agent

minifi

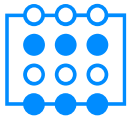
US-East Fleet

C++ agent

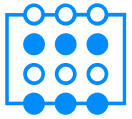
minifi

Apache Kafka

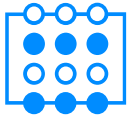
数据中转平台



gateway-west-raw-sensors



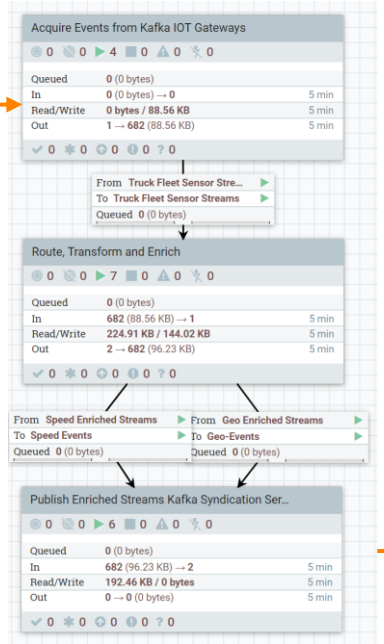
gateway-central-raw-sensors



gateway-east-raw-sensors

Apache NiFi

数据链路管道



Apache Kafka

数据合并

Kafka Topic syndicate-transmission

Kafka Topic syndicate-speed

Kafka Topic syndicate-temp

Kafka Topic syndicate-geo

Kafka Topic syndicate-oil

Kafka Topic syndicate-breaks

Kafka Topic syndicate-battery

Kafka Topic syndicate-start/stop

Kafka Topic syndicate-acceleration

Kafka Topic syndicate-idle

Apache Flink

数据订阅和实时处理系统

PROCESSING APP 1

Apache Flink

PROCESSING APP 2

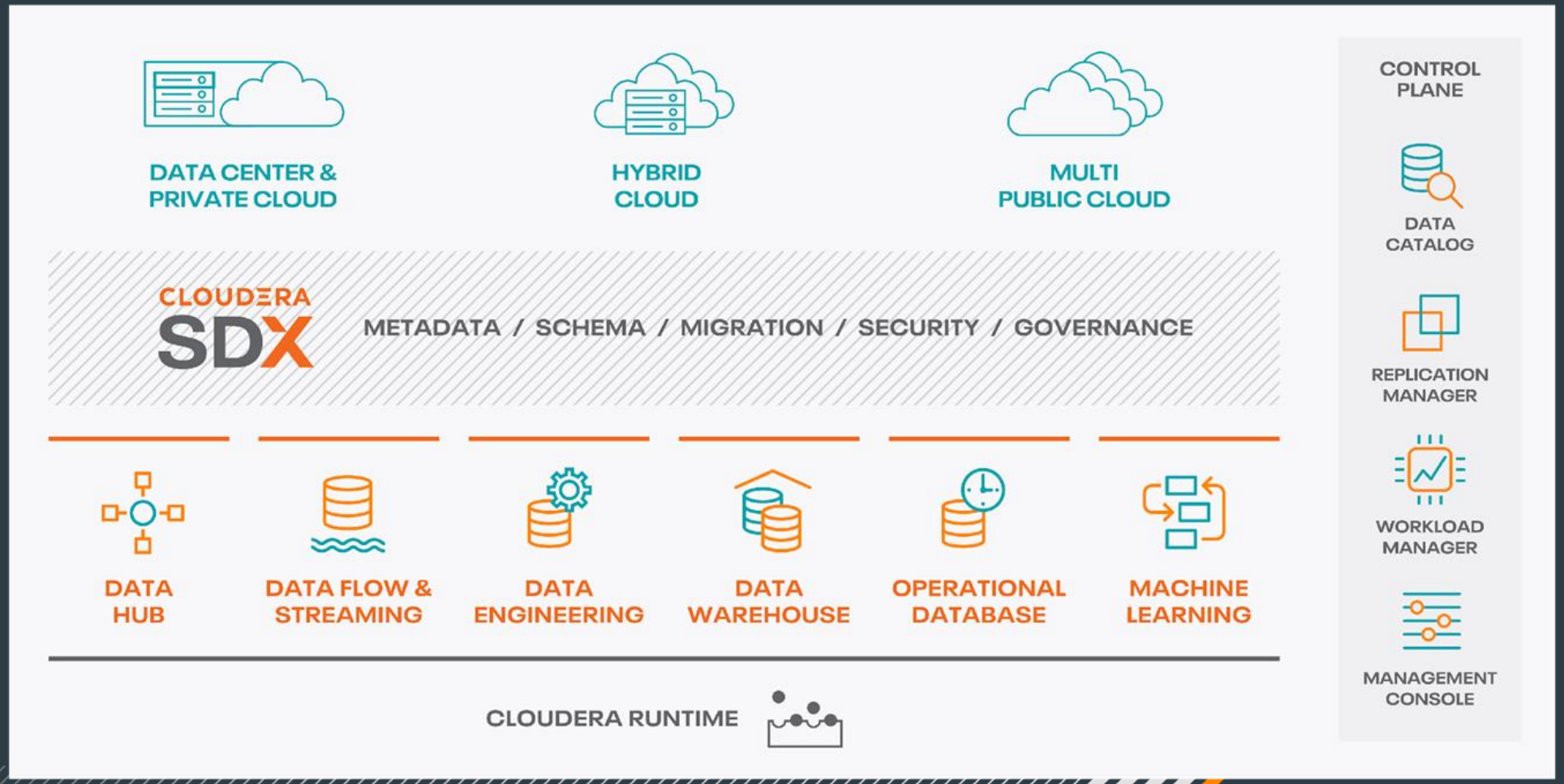
kafka

PROCESSING APP 3

Spark Structured Streaming

CLOUDERA DATA PLATFORM

全球第一个企业数据云



Apache Flink

Q & A