



Cloudera Data Flow Workshop

15/09/2020



CLOUDERA DATA PLATFORM

全球第一个企业数据云

HYBRID &
MULTI-CLOUD

SECURITY &
GOVERNANCE

ANALYTICS
EDGE TO AI

OPEN
DISTRIBUTION



DATA CENTER &
PRIVATE CLOUD



HYBRID
CLOUD



MULTI
PUBLIC CLOUD

CLOUDERA
SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE



DATA
HUB



DATA FLOW &
STREAMING



DATA
ENGINEERING



DATA
WAREHOUSE



OPERATIONAL
DATABASE



MACHINE
LEARNING

CLOUDERA RUNTIME



CONTROL
PLANE



DATA
CATALOG



REPLICATION
MANAGER



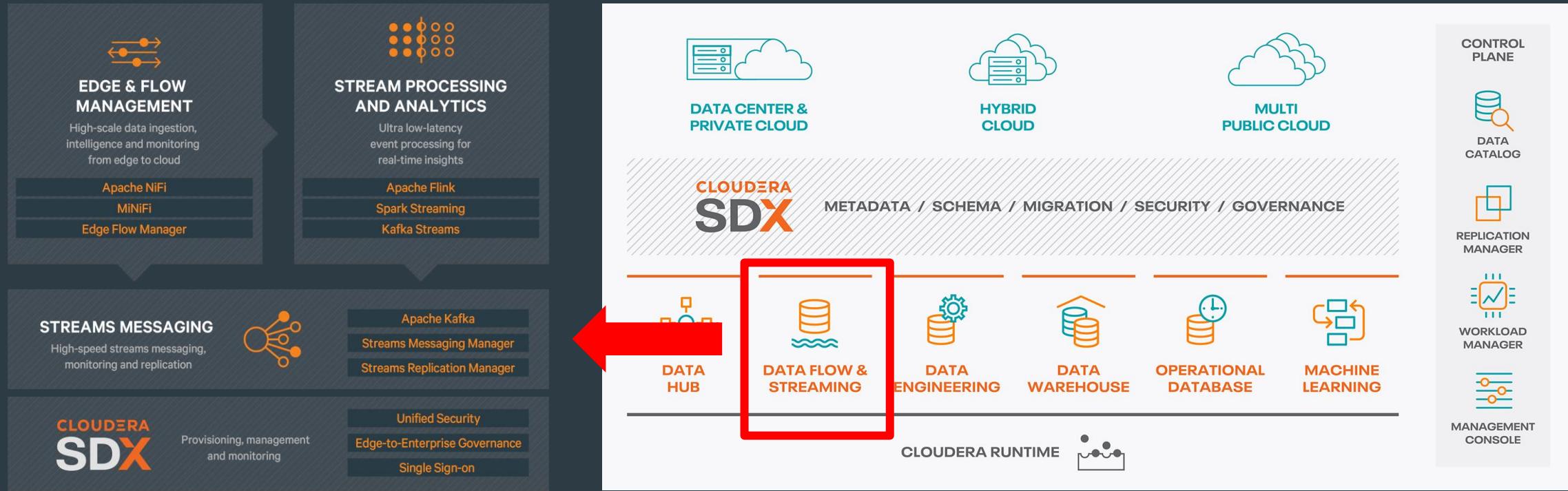
WORKLOAD
MANAGER



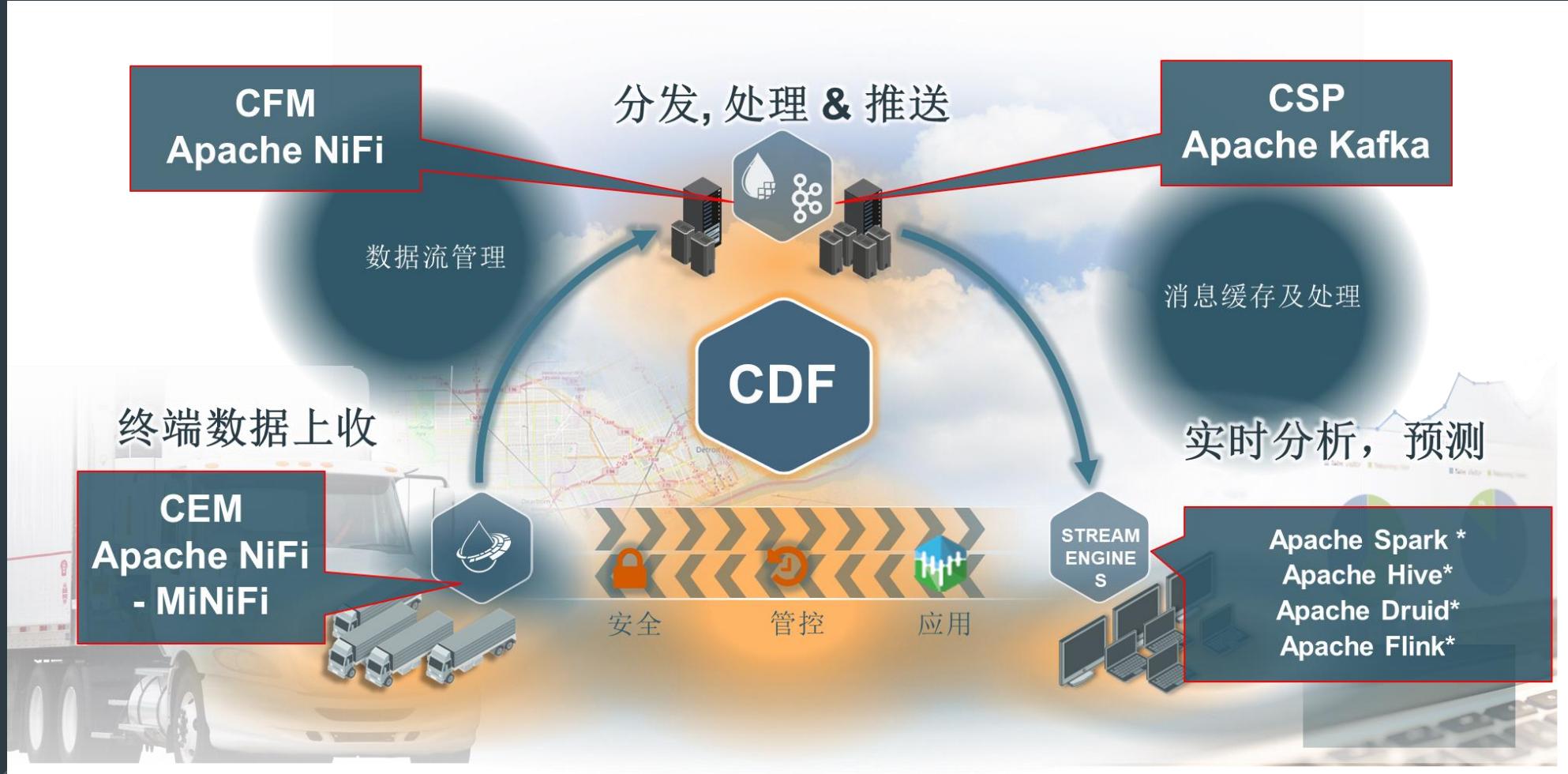
MANAGEMENT
CONSOLE

Cloudera DataFlow (CDF)

- Cloudera Flow



Cloudera DataFlow (CDF)

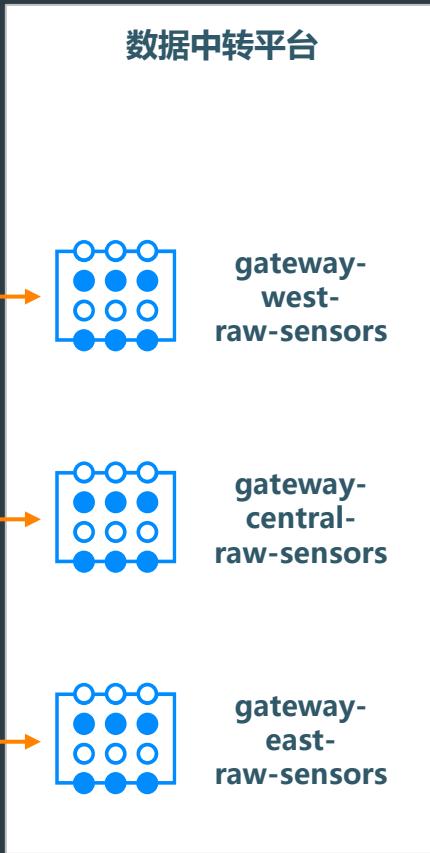


CDF IN Action 整体架构

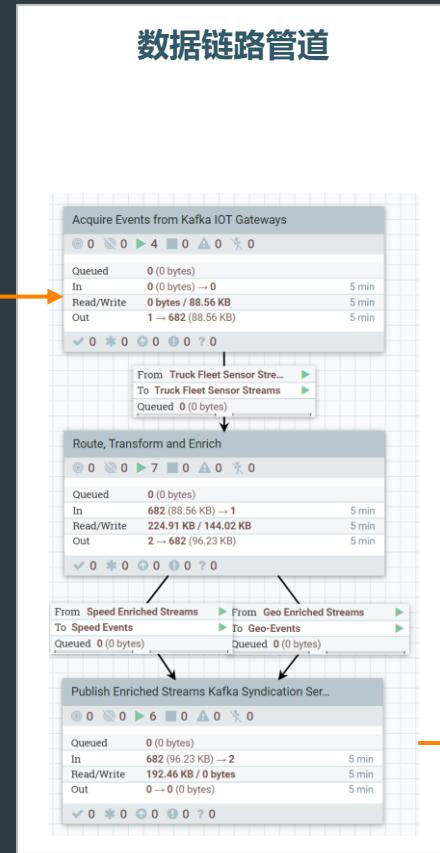
MiNiFi



Apache Kafka



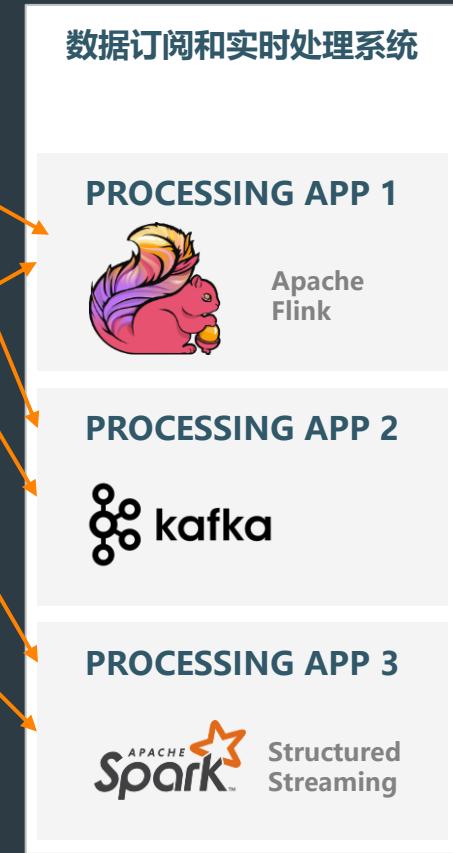
Apache NiFi



Apache Kafka



Apache Flink

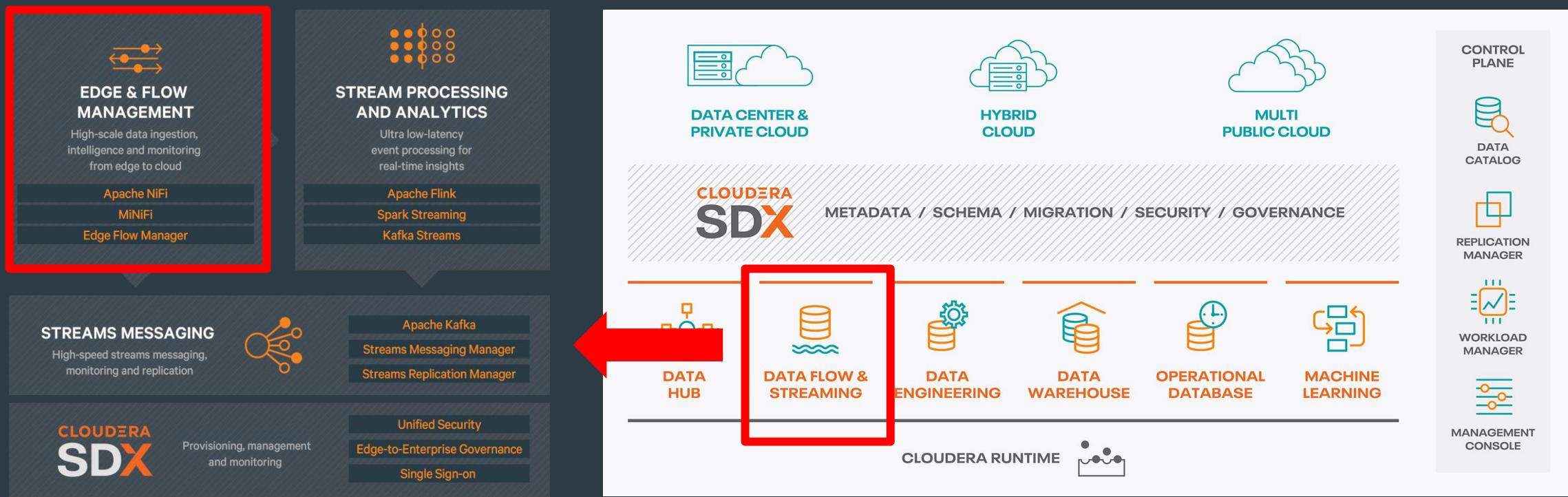


Cloudera Flow Management (CFM)

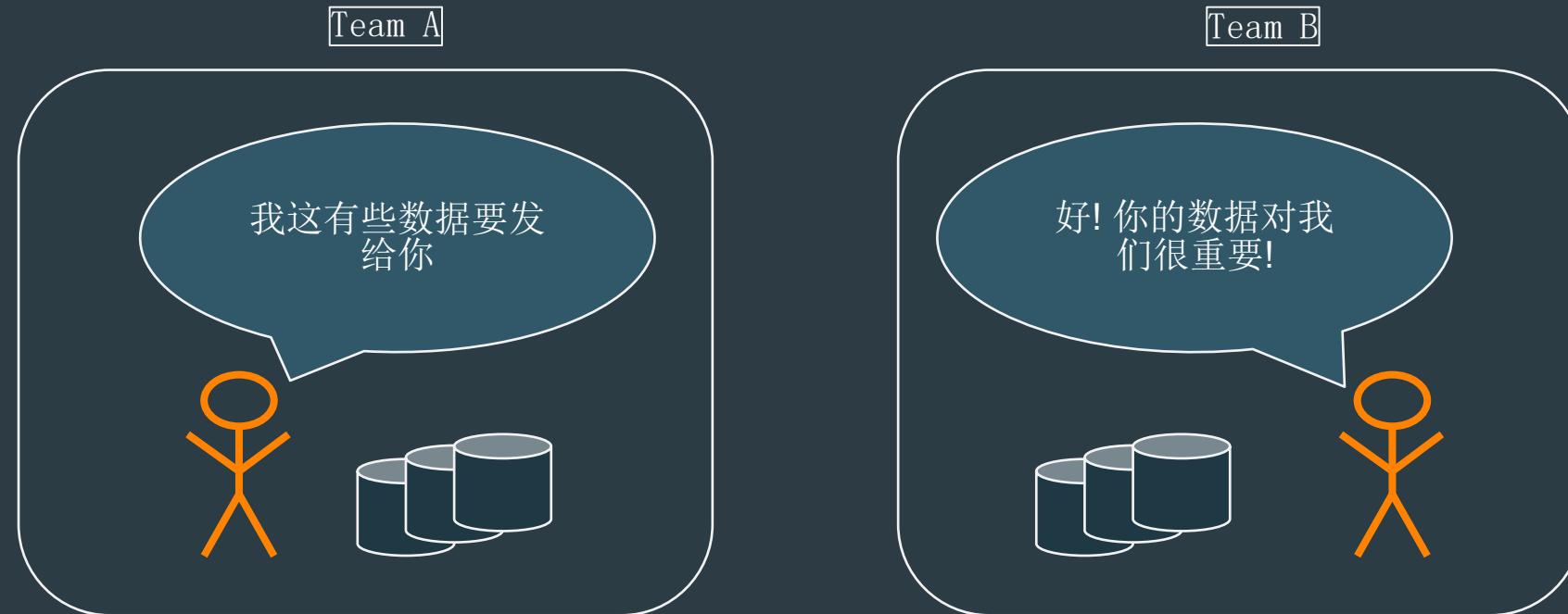
Apache NIFI 介绍

Cloudera DataFlow (CDF)

- Cloudera Flow



NIFI适用的场景



开发：很常见吧？也应该很简单...

NIFI适用的场景

Team 1

我们的数据是
Avro格式的，在
Kafka上面，你
自己取？



Team 2

抱歉，我们有专
门的Rest服务并
且只能接受JSON
格式



开发：写个程序，接Kafka，转JSON送给Rest

CFM

NIFI

Team A

Team B

你们的身份策略
是什么？我们是
用 Kerberos



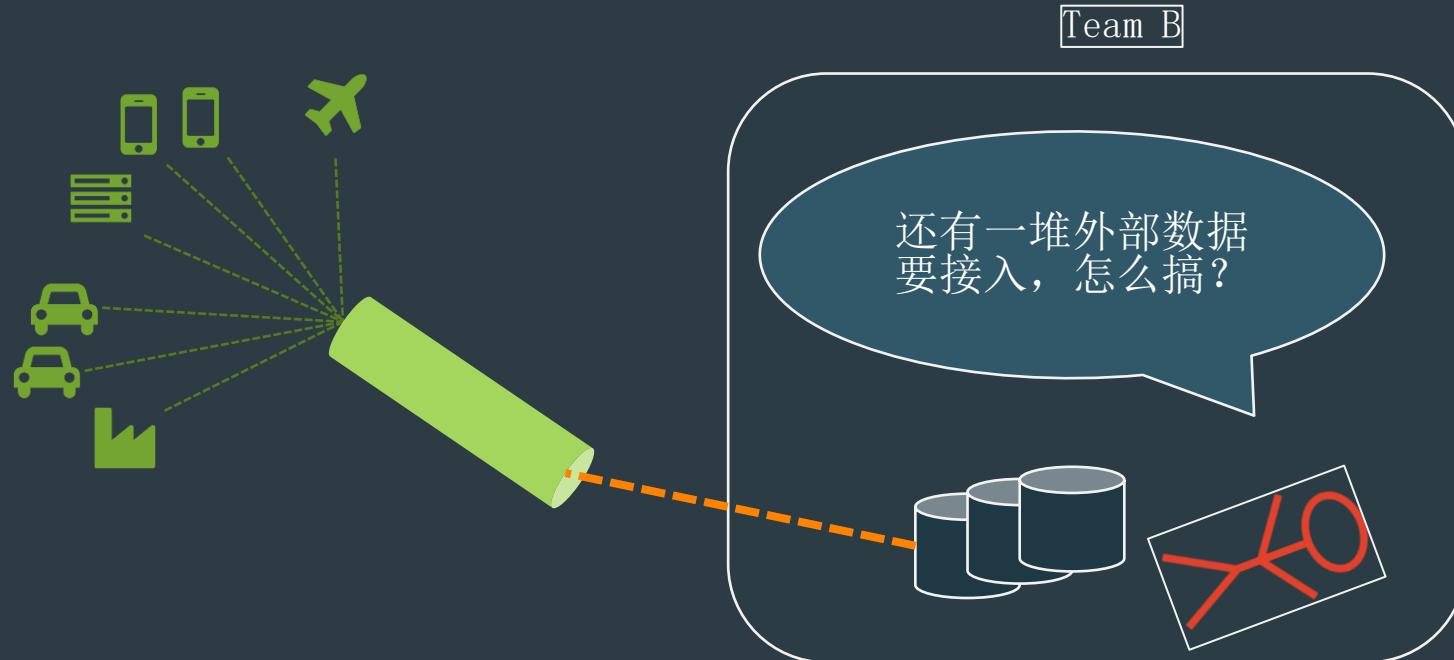
抱歉，我们是双
路TLS的CA



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。

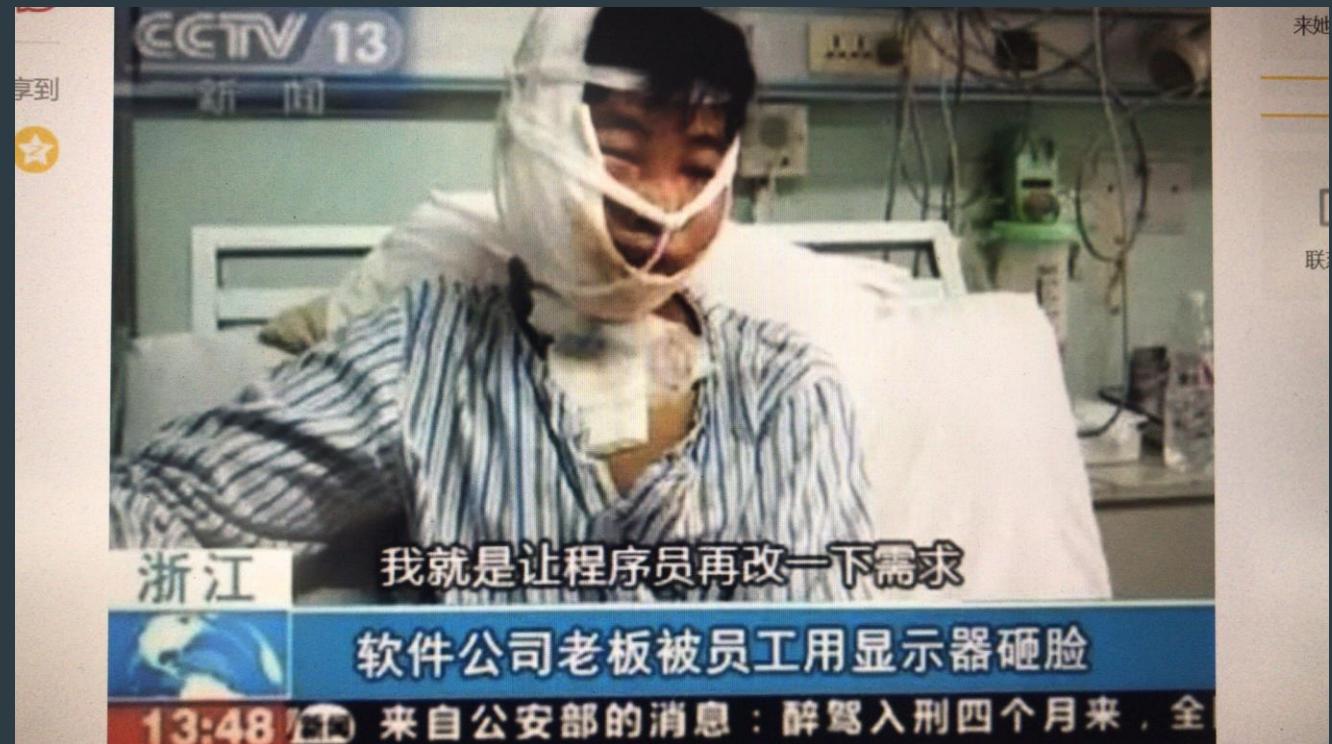
CIO?
云原生，基于人工智能的自动化运维

CFM

NIFI

产品经理？

需求变更



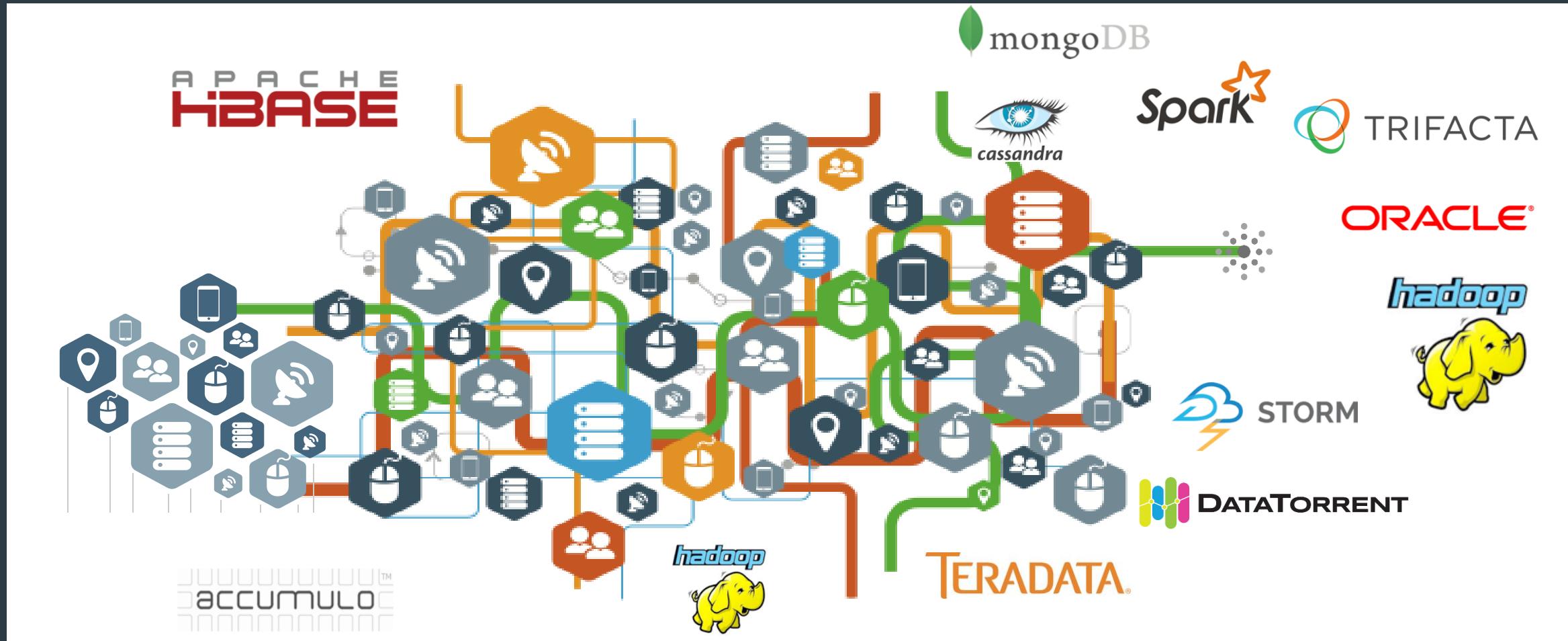
CFM

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

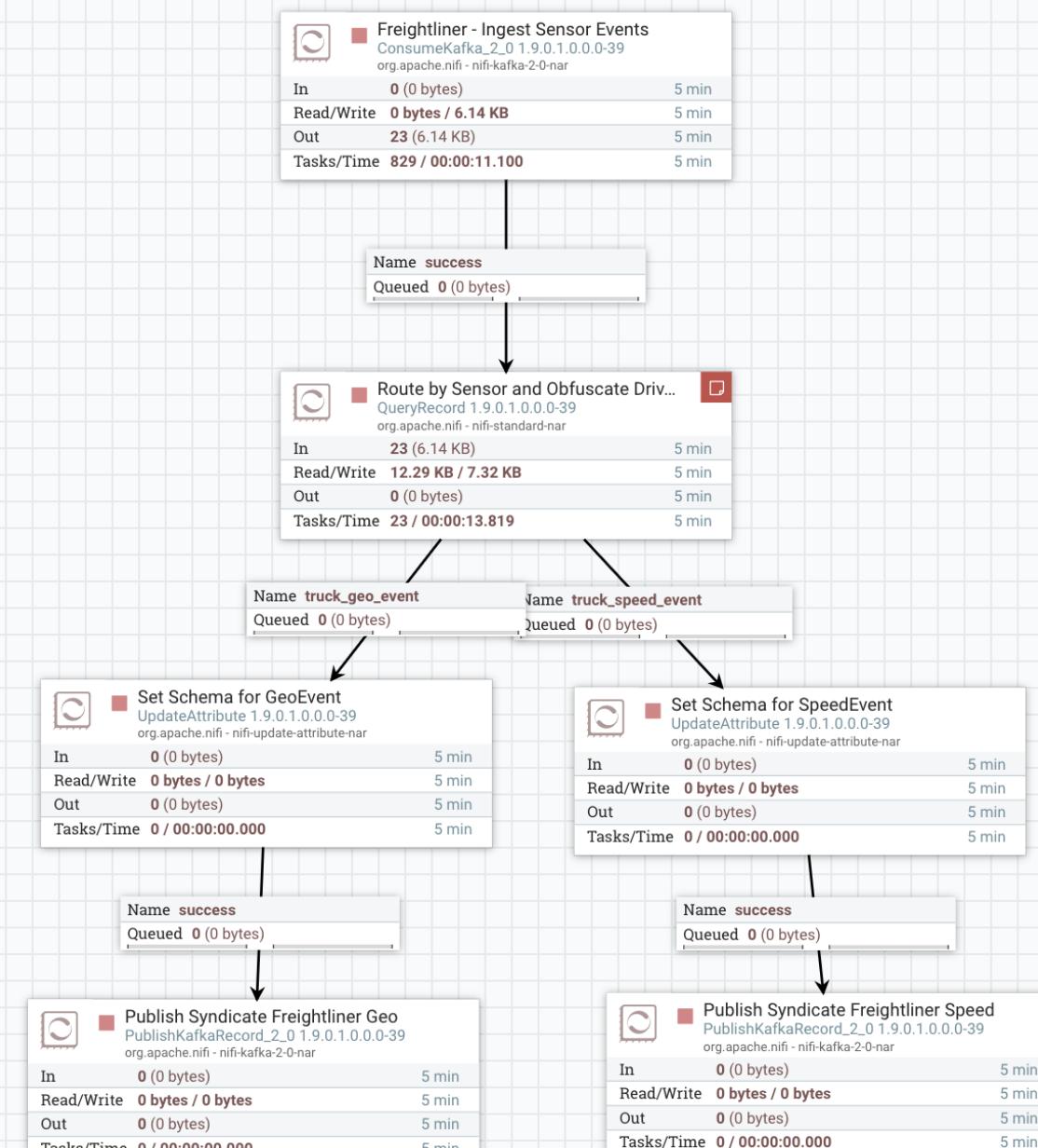
NIFI



IOT Trucking Fleet - DataFlow App
Process Group

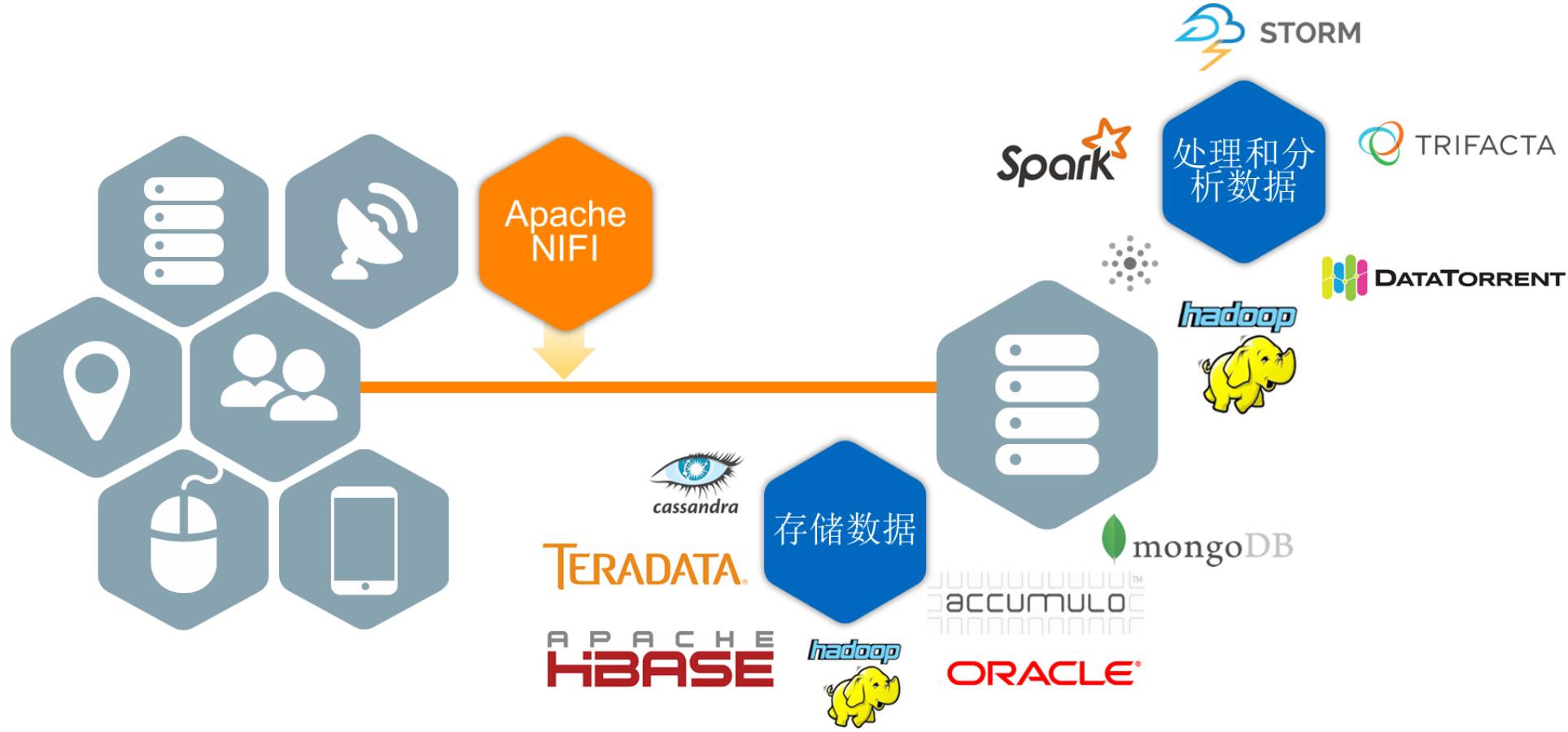
e20cd3bd-339f-3a49-8230-85114ba19922

DELETE



CEM - NIFI 要解决的问题

NIFI要解决的问题

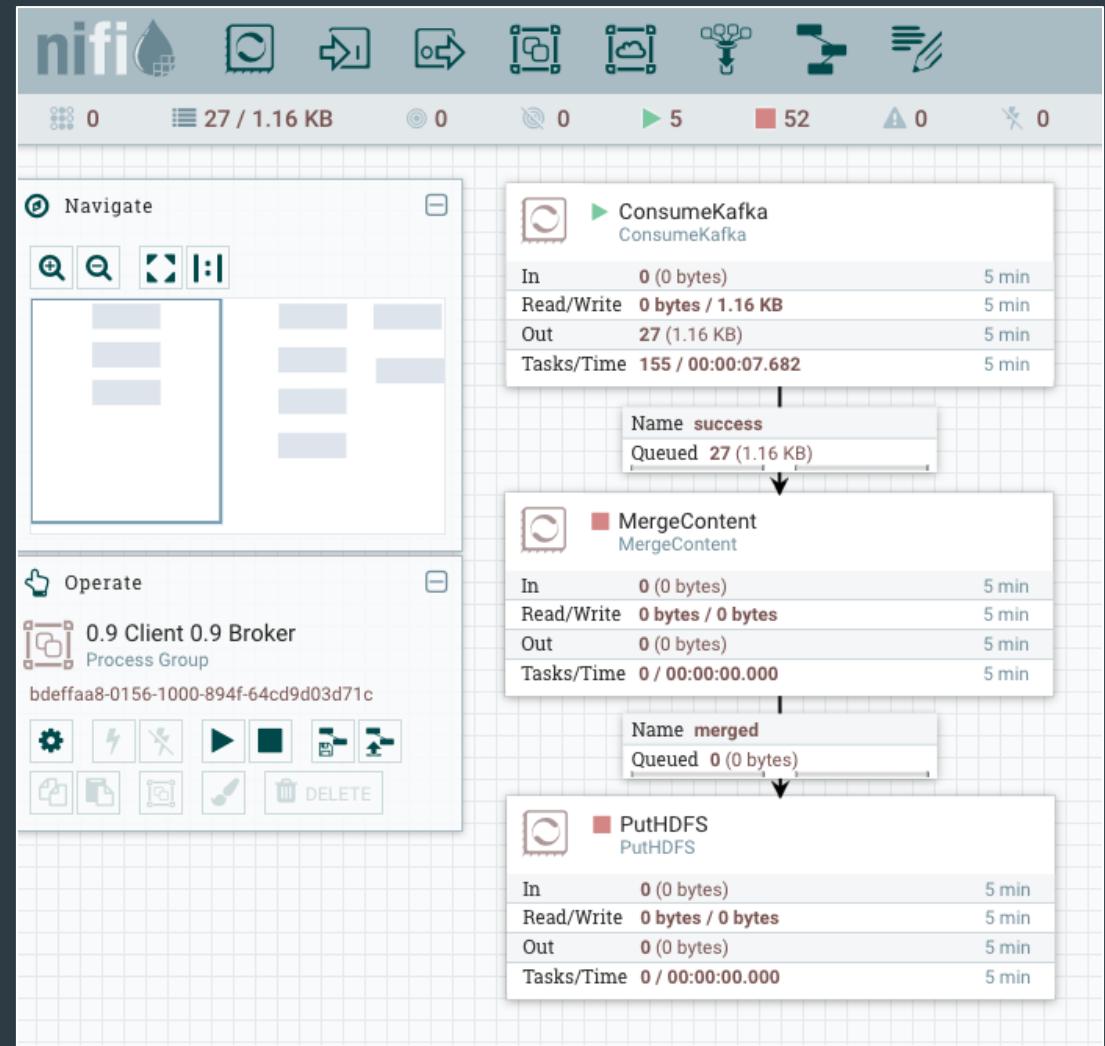


Cloudera Flow Management (CFM)

Apache NIFI 产品特点

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。



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	AzureEventHub	AzurElement	AzureBlobStorage	Beats	Cassandra	AMQP
DatabaseTables	AzureQueueStorage	HTTP	DistributedMapCache	GRPC	DatabaseTable	AzureEventHub
File	CouchbaseKey	IgniteCache	Elasticsearch	HTTP	DatabaseTableRecord	EWS
FTP	DynamoDB	JMSQueue	Elasticsearch5	Lumberjack	DNS	GCPubSub
GCSBucket	File	JMSTopic	ElasticsearchHttp	RELP	ElasticsearchHttp	IMAP
HDFS	FTP	Kafka	File	SMTP	Record	JMS
S3	HBase	Mongo	FTP	Syslog	Solr	Kafka
SFTP	HDFS	RethinkDB	GCSObject	TCP	Whois	Kafka_0_10
	HDFSEvents	SFTP	HBaseRow	TCPRecord		Kafka_0_11
	HDFSFileInfo	SNMP	HDFS	UDP		Kafka_1_0
	HDFSSequenceFile	Solr	Parquet	UDPRecord		Kafka_2_0
		Splunk	S3Object	WebSocket		KafkaRecord_0_10
		SQS	SFTP			KafkaRecord_0_11
		TCP				KafkaRecord_1_0
		Twitter				KafkaRecord_2_0
						MQTT
						POP3
						WindowsEventLog

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

NIFI-功能特点

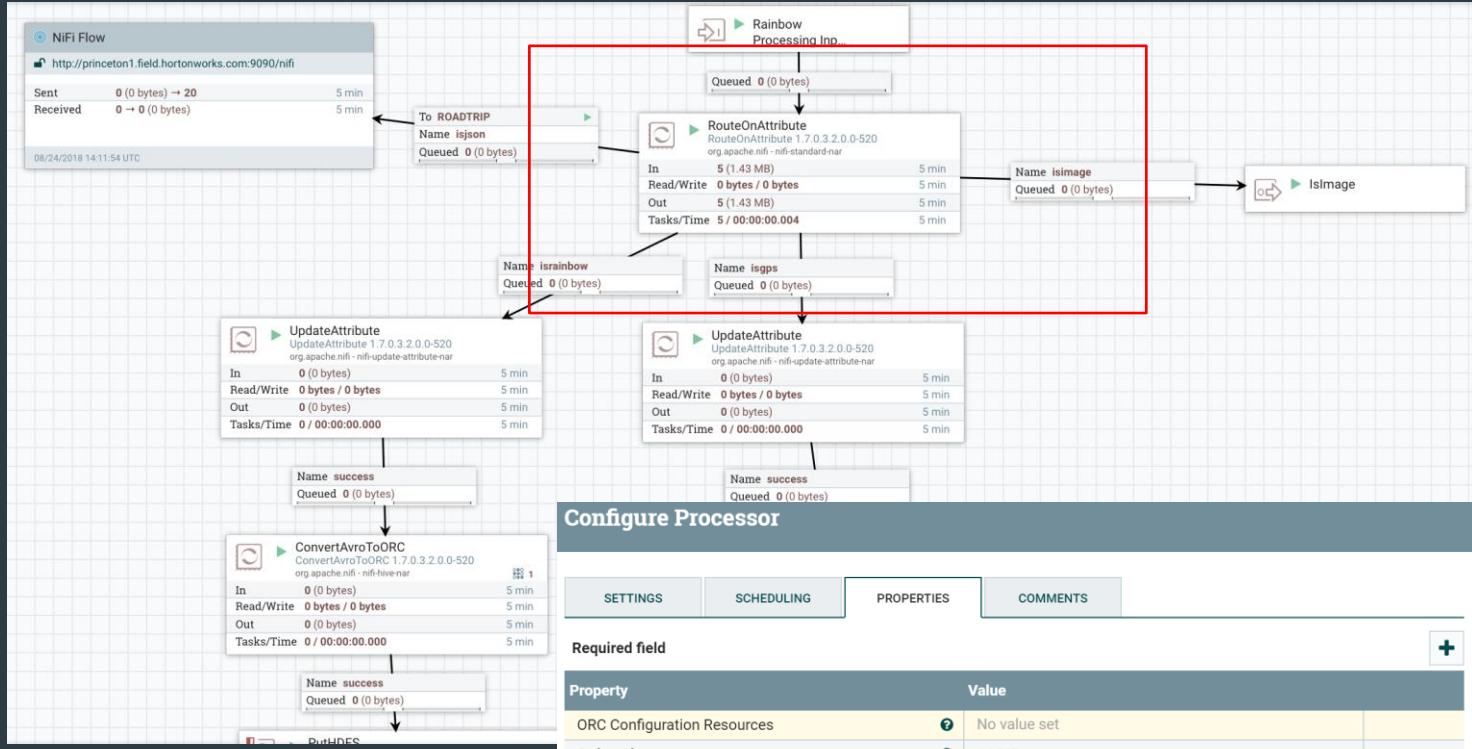
- 面向数据抓取，转换及推送。
 - 推送

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



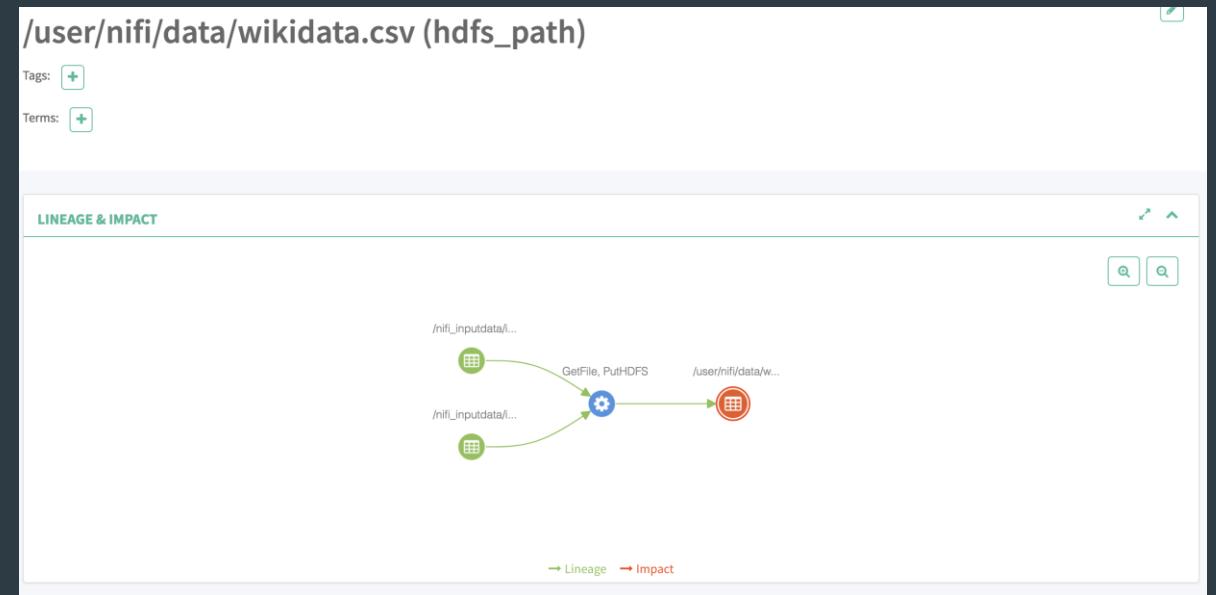
Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Required field			
Property			
ORC Configuration Resources	No value set		
Stripe Size	64 MB		
Buffer Size	10 KB		
Compression Type	NONE		
Hive Table Name	No value set		

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	FlowFileUuid	Size	Component Name	Component Type	Actions
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	

NIFI-功能特点

- 4. 企业级数据流管理。
 - 4.1 版本管理



The screenshot shows two panels of the Apache NIFI interface. The top panel is a context menu for a node named "TEST". The menu items include: Configure, Variables, Version (with a dropdown showing "5 min"), Enter group, Start, and Stop. The "Version" item is highlighted with a green checkmark icon. An orange arrow points from this menu to the bottom panel. The bottom panel displays the metrics for the "TEST" node. The metrics are:

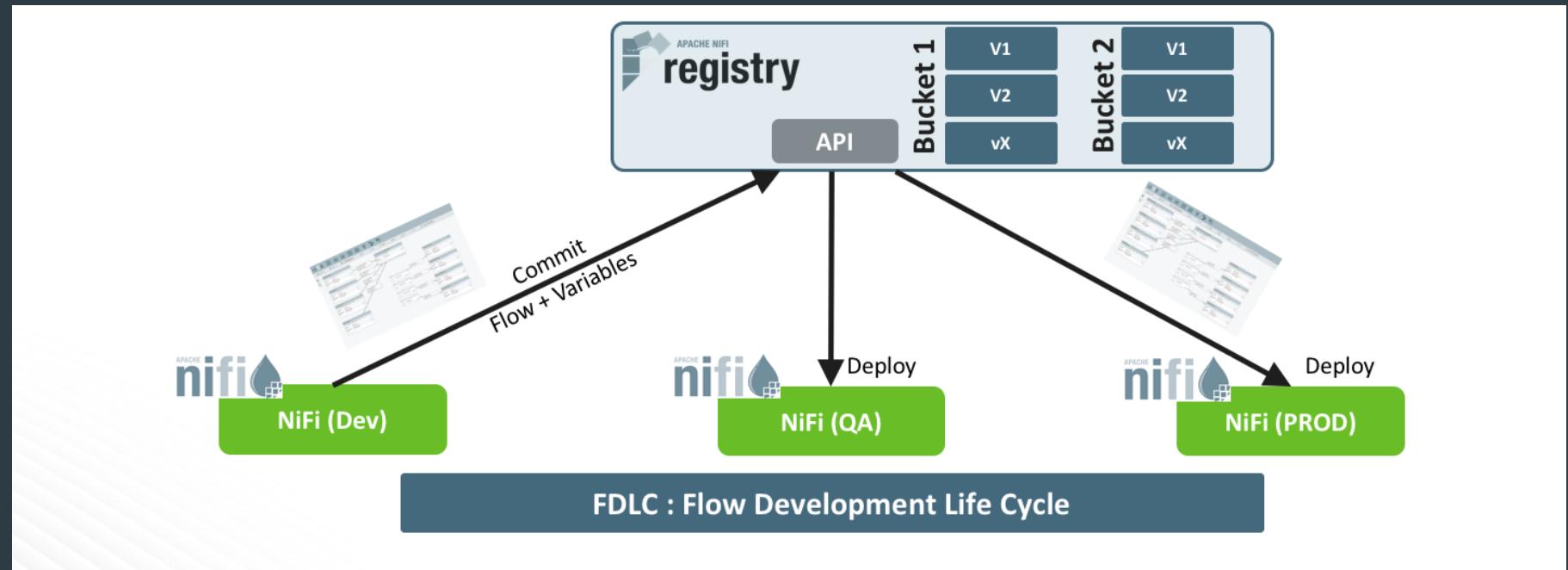
Category	Value	Time
Queued	0 (0 bytes)	5 min
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min

Below the metrics, there are status icons: a green checkmark for TEST, and counts for In (0), Out (0), Read/Write (0), and Stop (0).

CFM

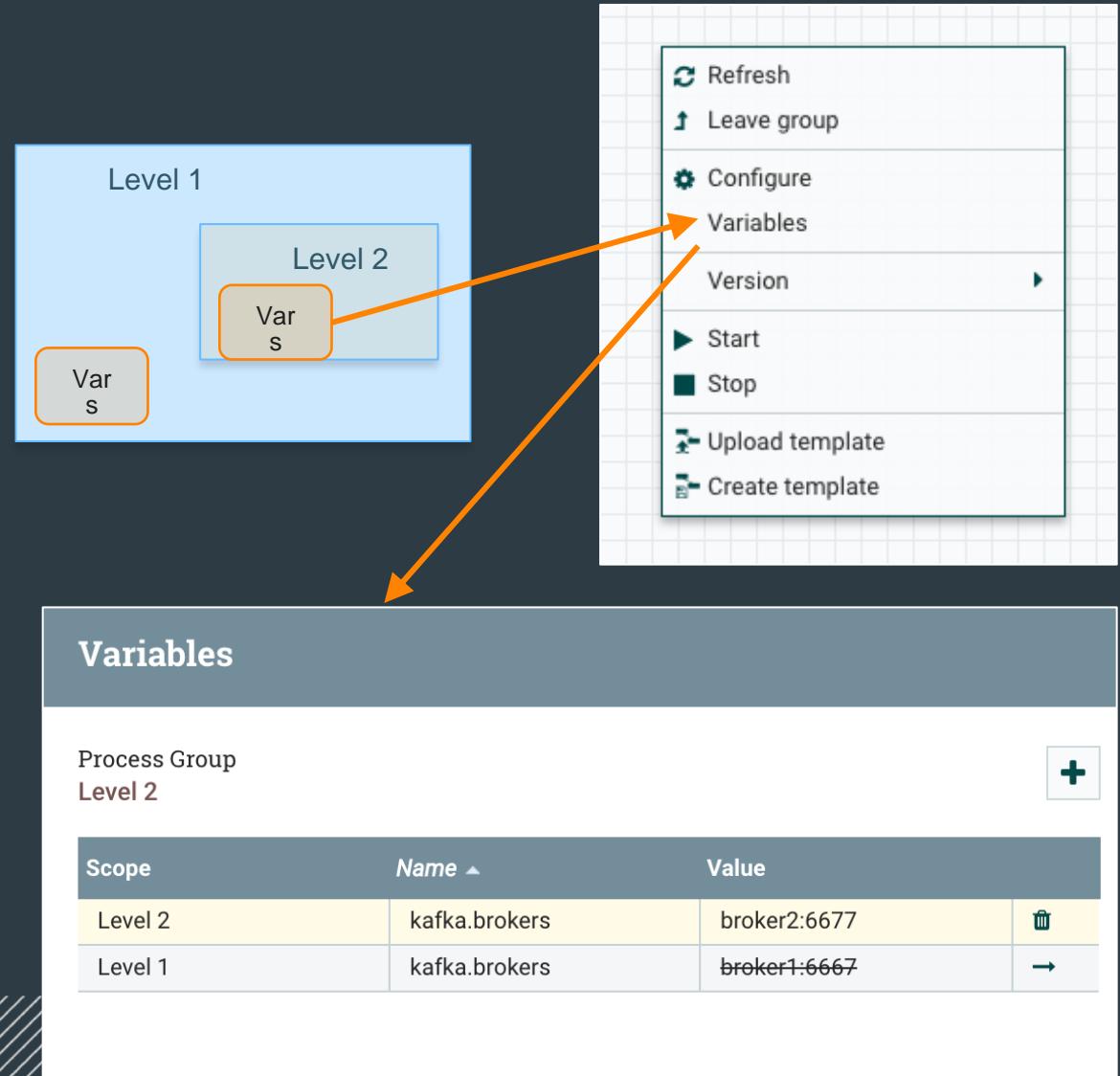
NIFI-功能特点

- 4. 企业级数据流管理。
 - 4.1 版本管理



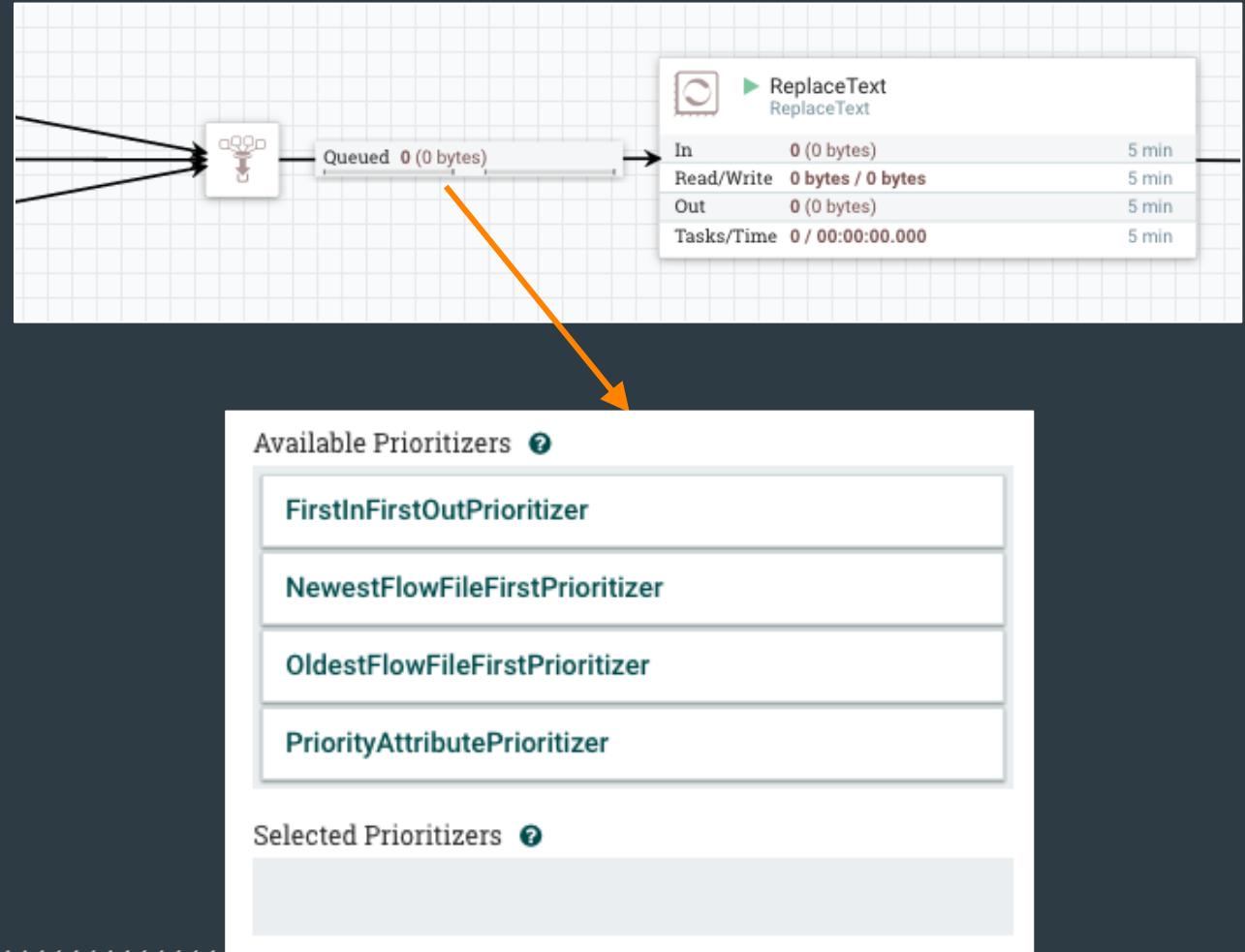
NIFI-技术特点

- 4. 企业级数据流管理。
 - 4.2 参数管理



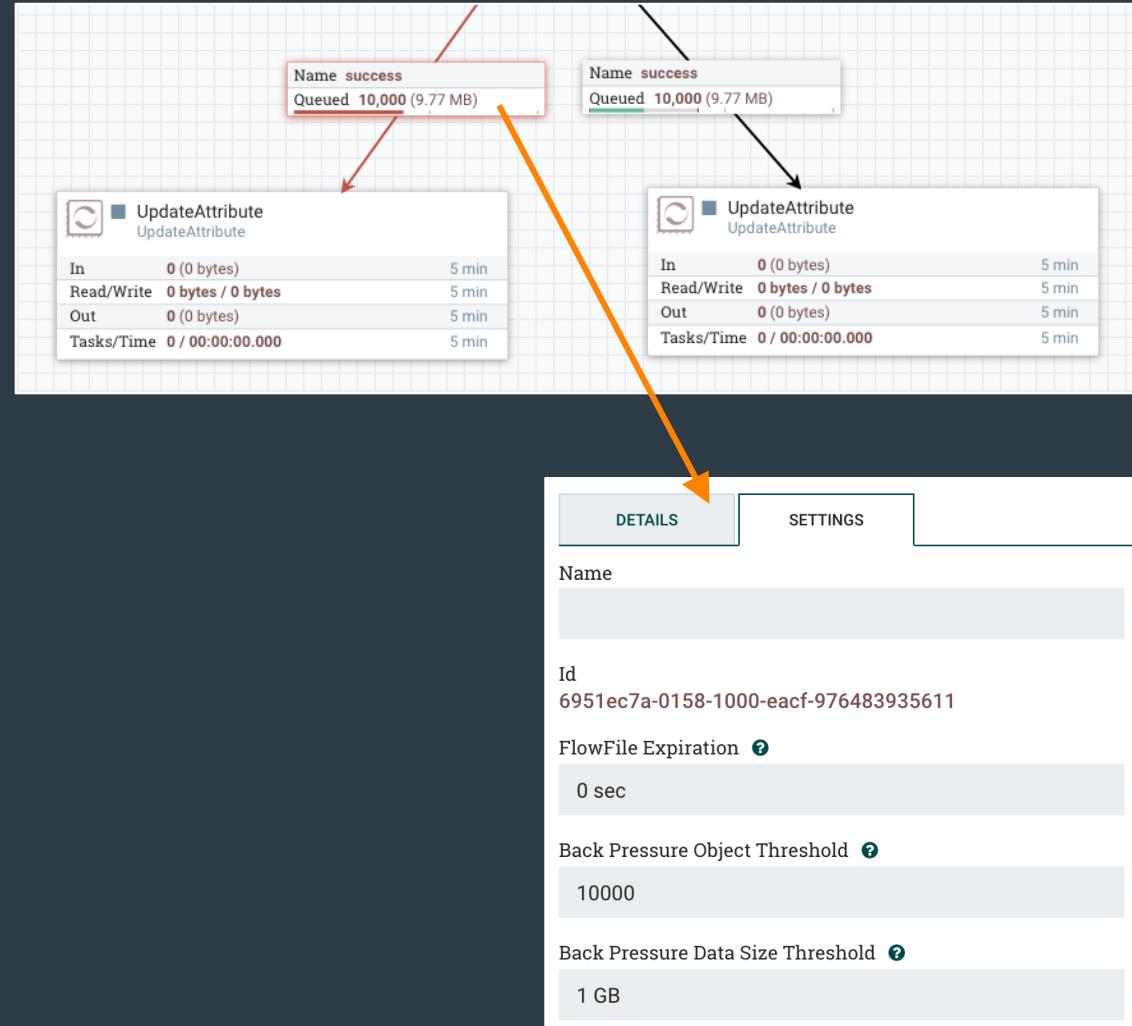
NIFI-技术特点

- 4. 企业级数据流管理。
 - 4.3 数据优先级管理



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

Team A

我这有些数据要发
给你



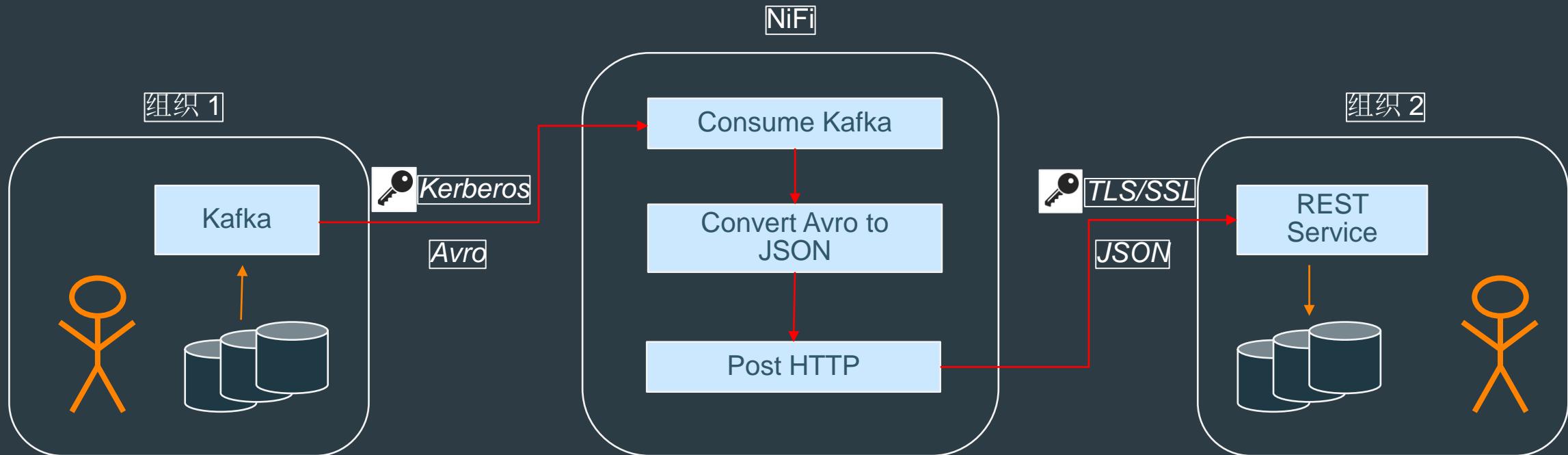
Team B

好! 你的数据对我
们很重要!



CFM

NIFI适用的场景

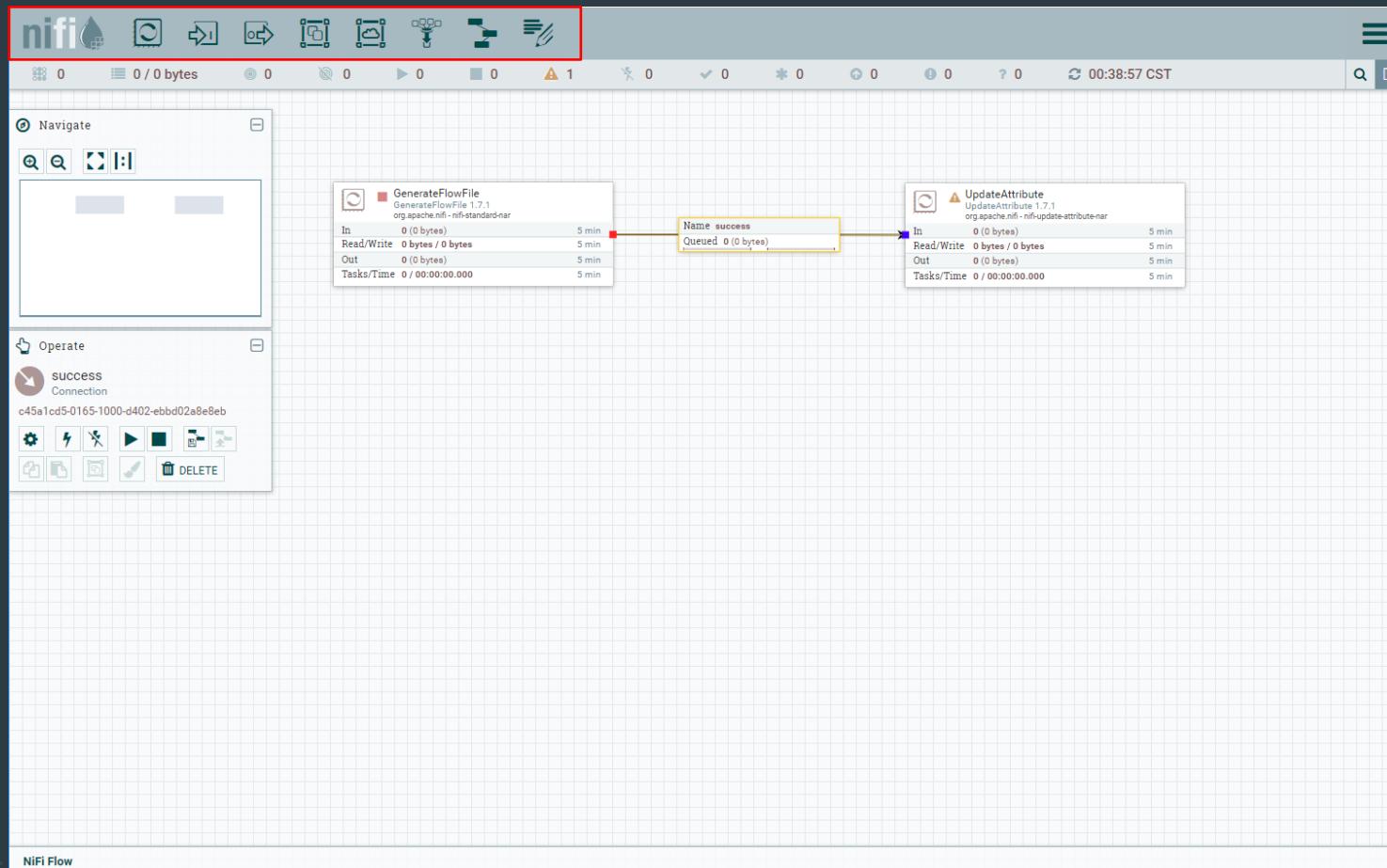


Cloudera Flow Management (CFM)

Apache NIFI UI 介绍

Apache NIFI UI 介绍

常用工具栏

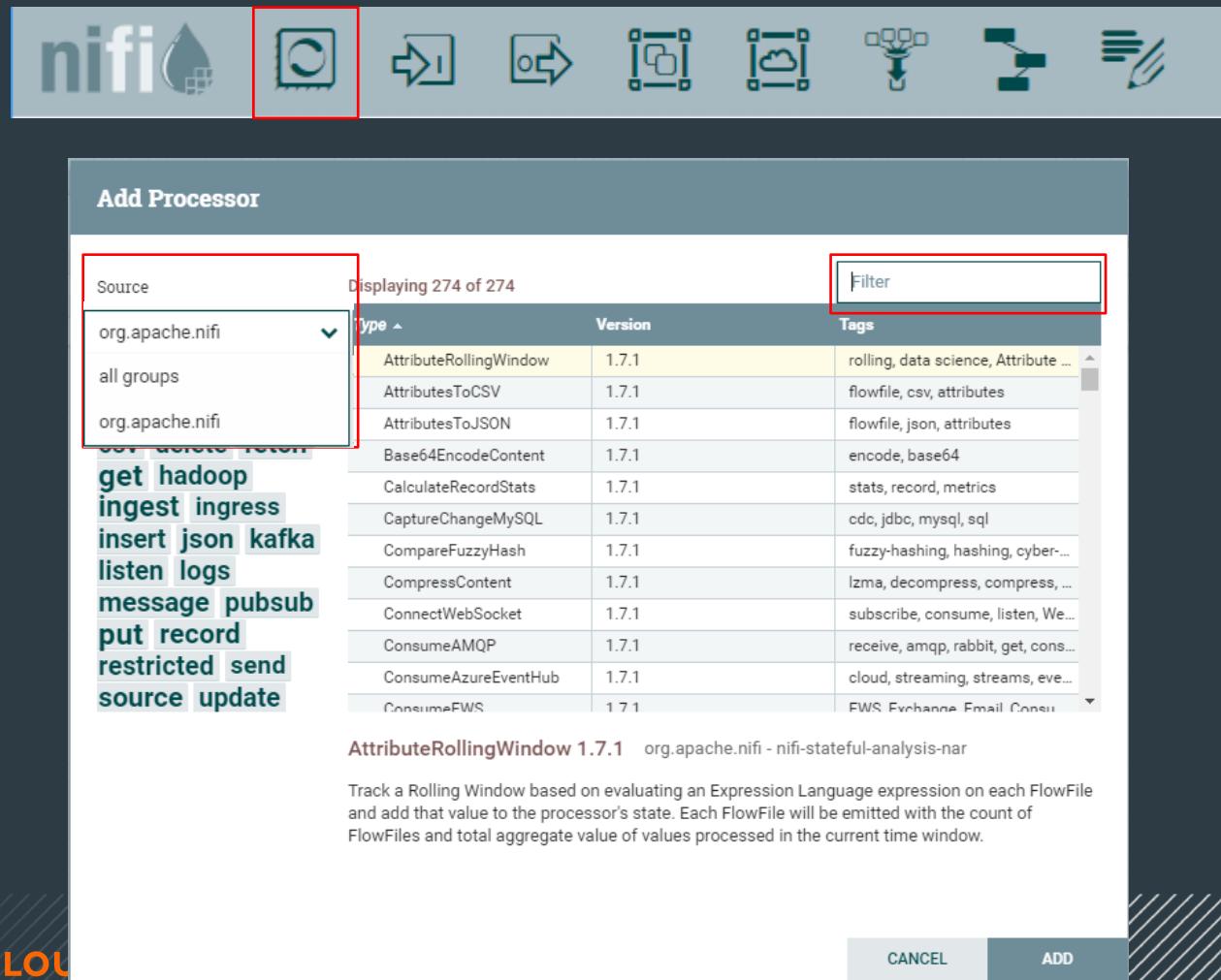


常用工具栏

包含常用的快捷方式

Apache NIFI UI 介绍

常用工具栏 – 添加处理器



The screenshot shows the 'Add Processor' dialog in the Apache Nifi UI. At the top, there is a toolbar with several icons: a 'nifi' logo, a 'Processor Group' icon (highlighted with a red box), a 'Get' icon, a 'Put' icon, a 'Listen' icon, a 'Consume' icon, a 'Merge' icon, and a 'Split' icon.

The main area is titled 'Add Processor'. It includes a 'Source' dropdown menu with options like 'org.apache.nifi' (highlighted with a red box) and 'all groups'. Below the dropdown is a list of processor names, each with its type, version, and tags. A 'Filter' input field is also present at the top of this list.

On the left side of the dialog, there is a sidebar with various processor categories listed as buttons:

- get
- hadoop
- ingest
- ingress
- insert
- json
- kafka
- listen
- logs
- message
- pubsub
- put
- record
- restricted
- send
- source
- update

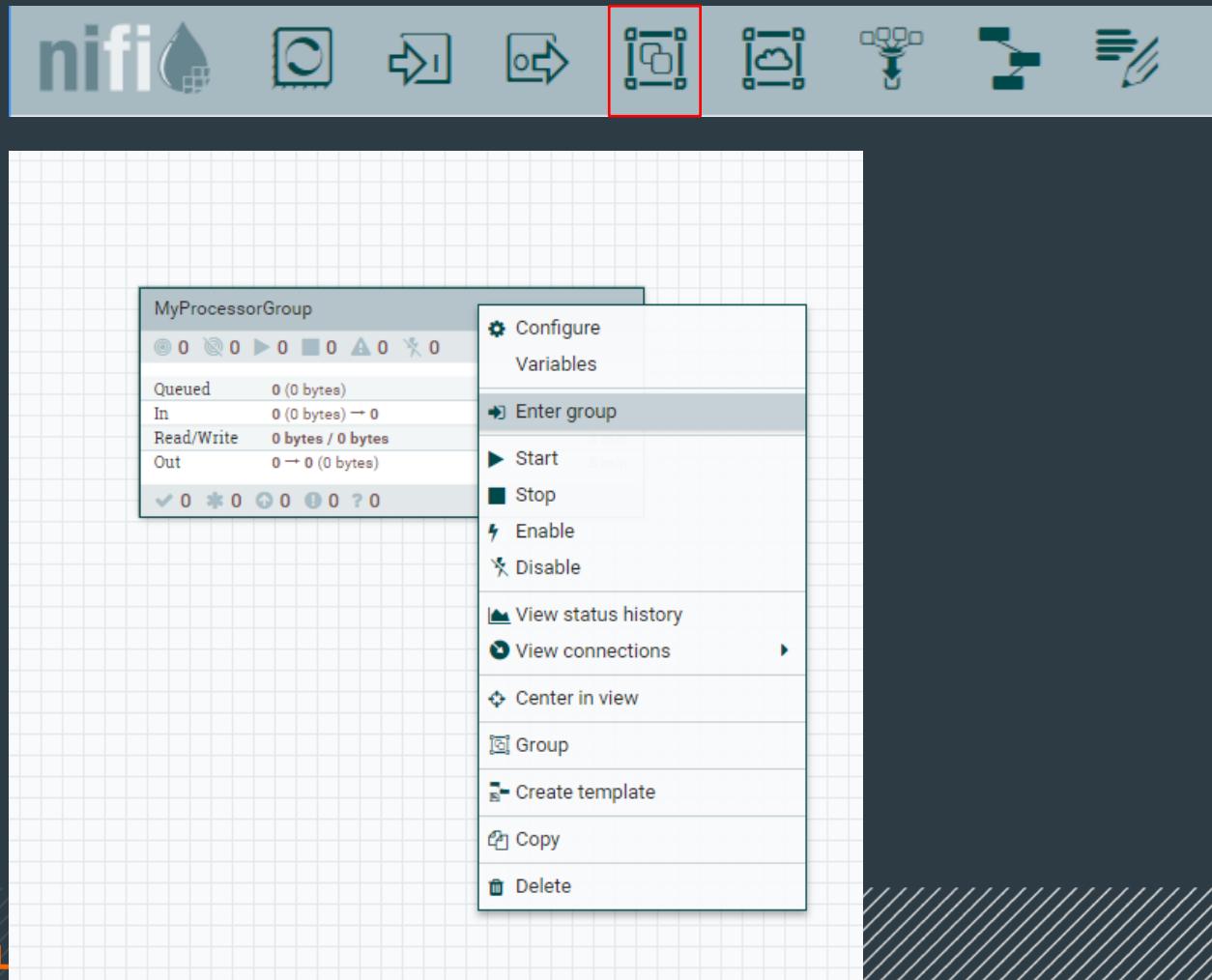
At the bottom of the dialog, there is a detailed description of the selected processor, 'AttributeRollingWindow 1.7.1 - org.apache.nifi - nifi-stateful-analysis-nar', and a note about its functionality. There are also 'CANCEL' and 'ADD' buttons at the bottom.

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

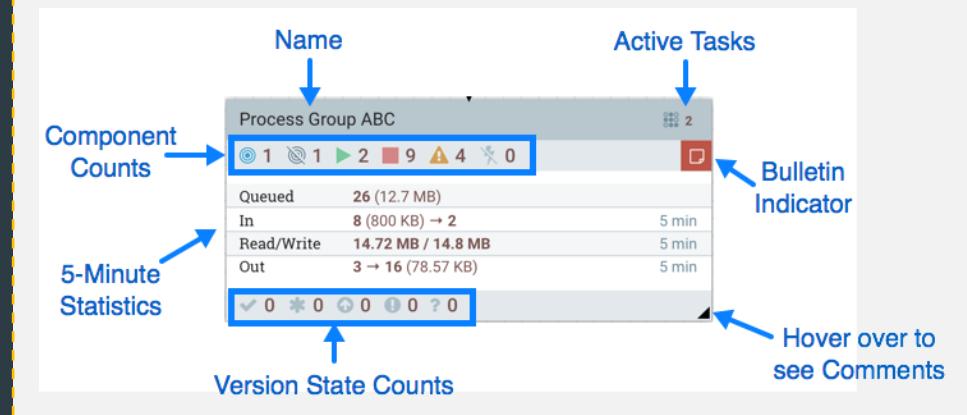
Apache NIFI UI 介绍

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

Add Remote Process Group

URLs <https://remotehost:8443/nifi>

Transport Protocol **RAW**

Local Network Interface

HTTP Proxy Server Hostname

HTTP Proxy Server Port

HTTP Proxy User

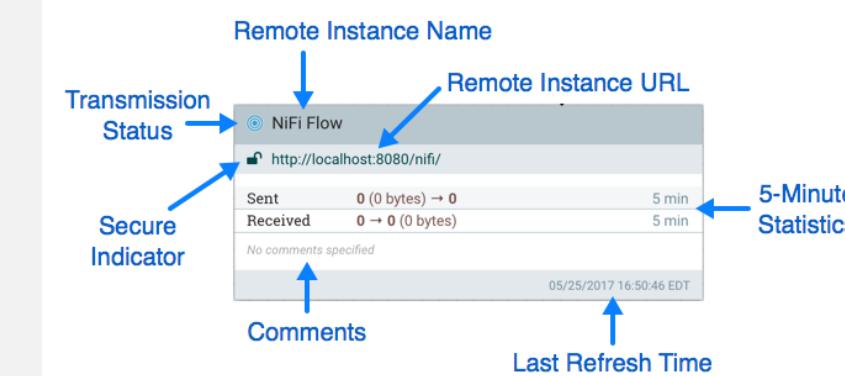
HTTP Proxy Password

Communications Timeout **30 sec**

Yield Duration **10 sec**

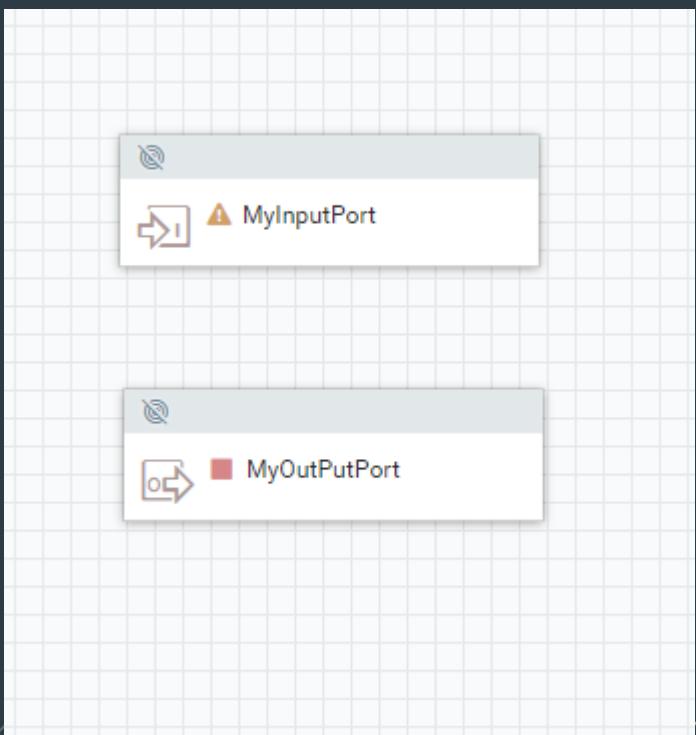
CANCEL ADD

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

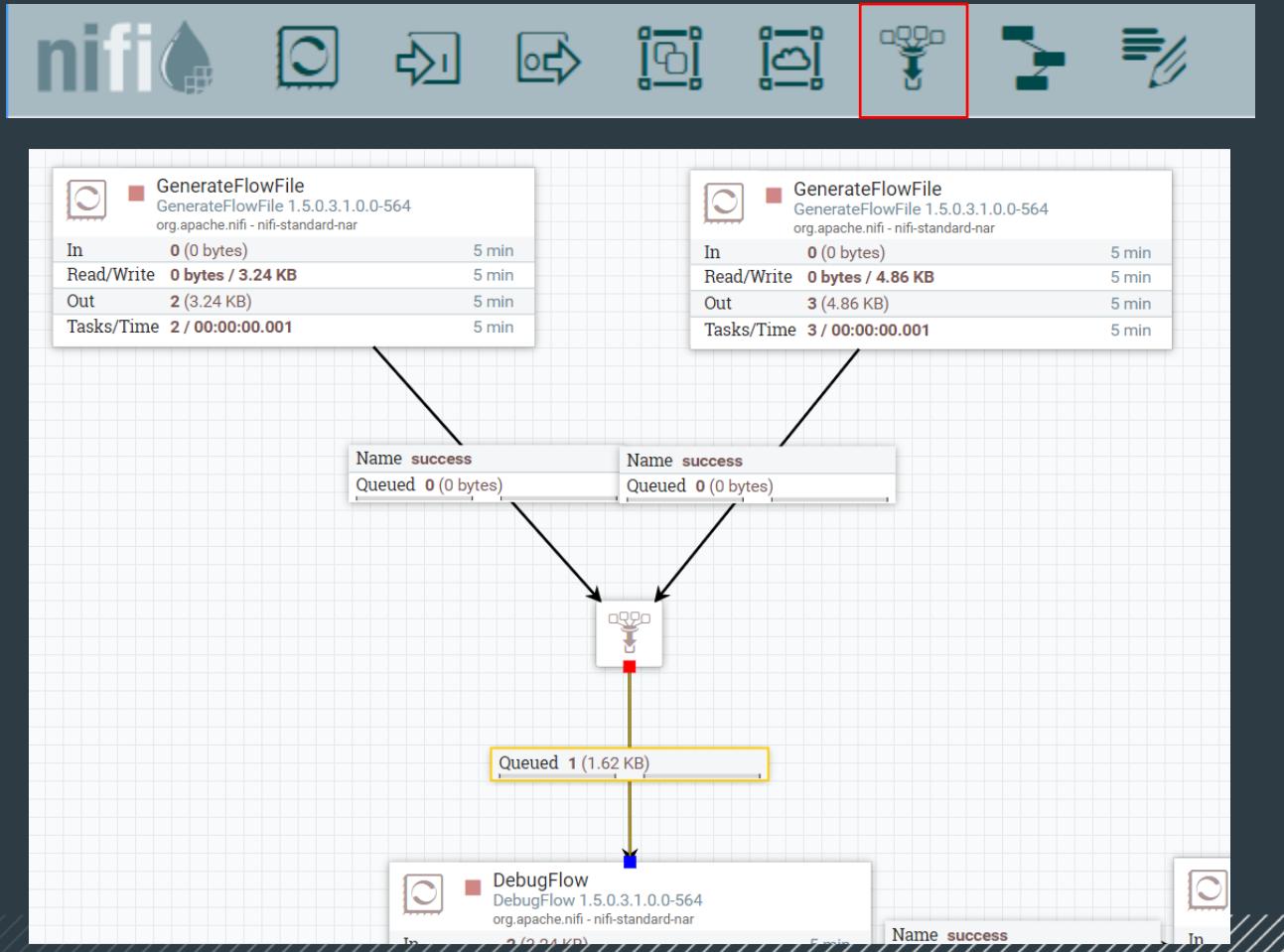
为处理器组提供了数据接入的接口

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

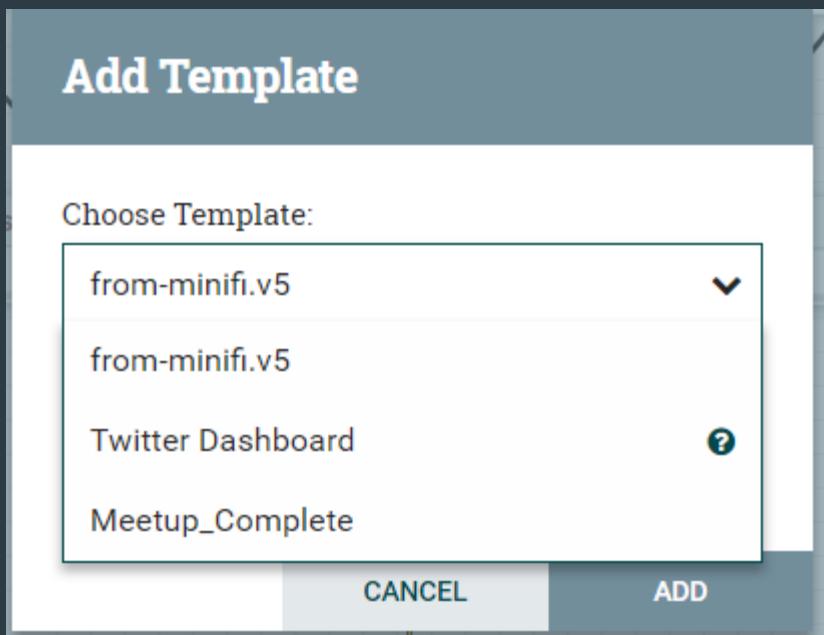


汇聚器

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

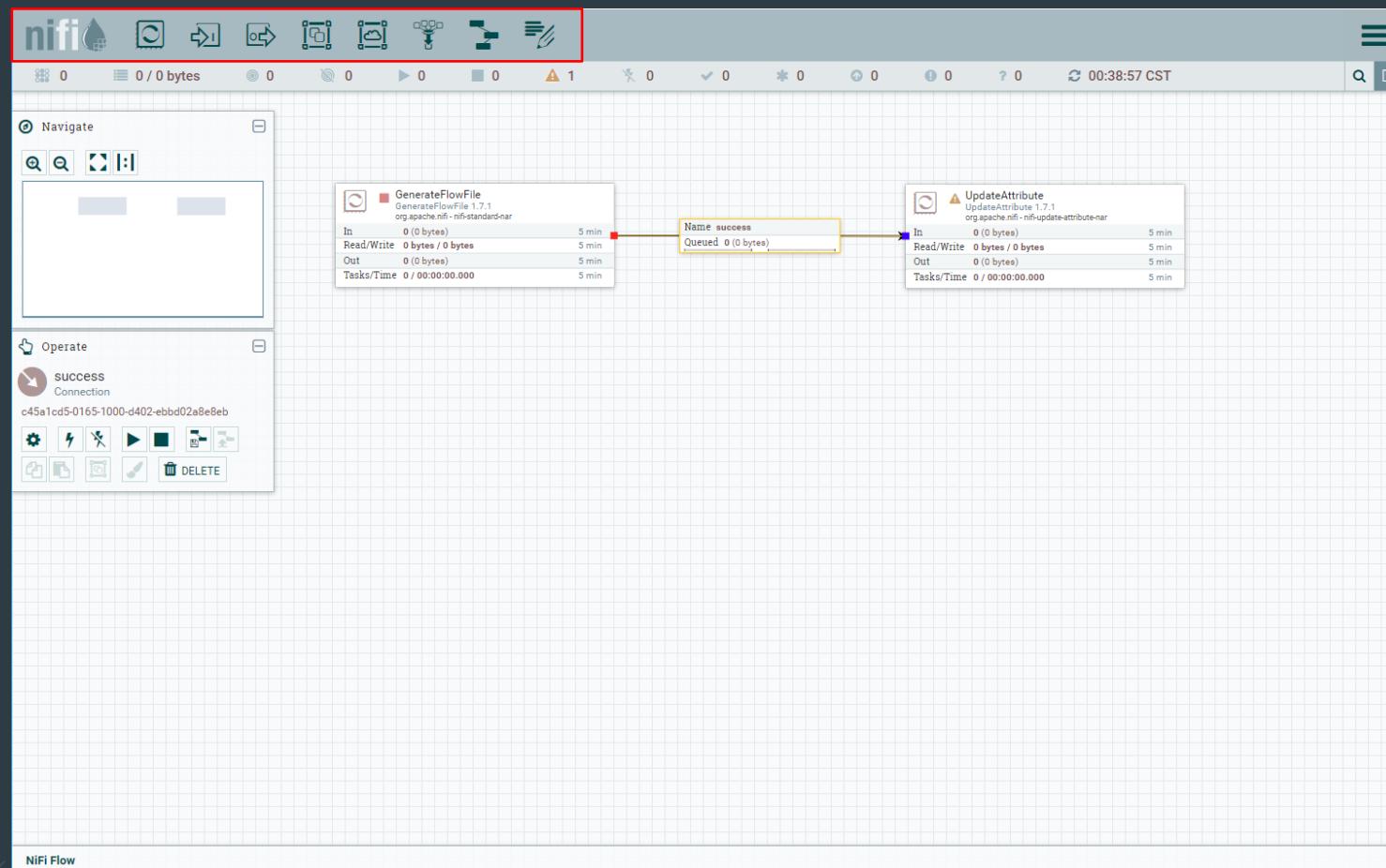


模板

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

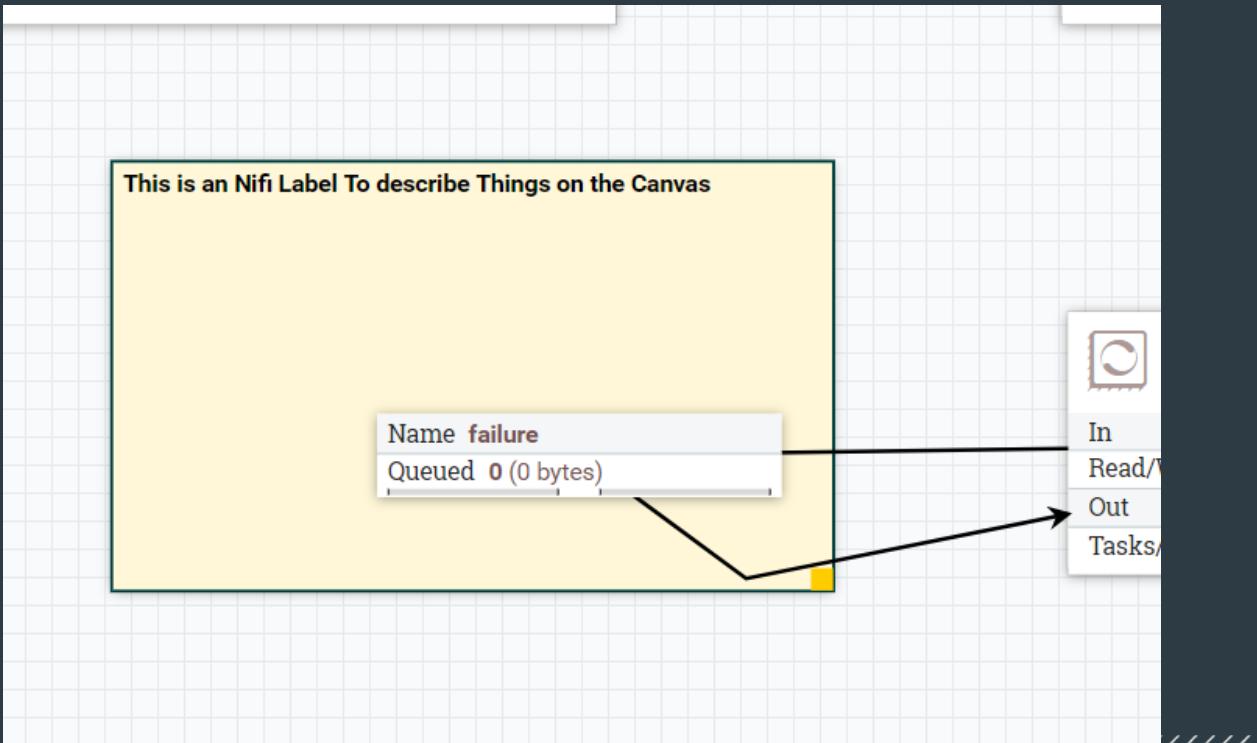
常用工具栏



常用工具栏

Apache NIFI UI 介绍

常用工具栏 – 便利贴

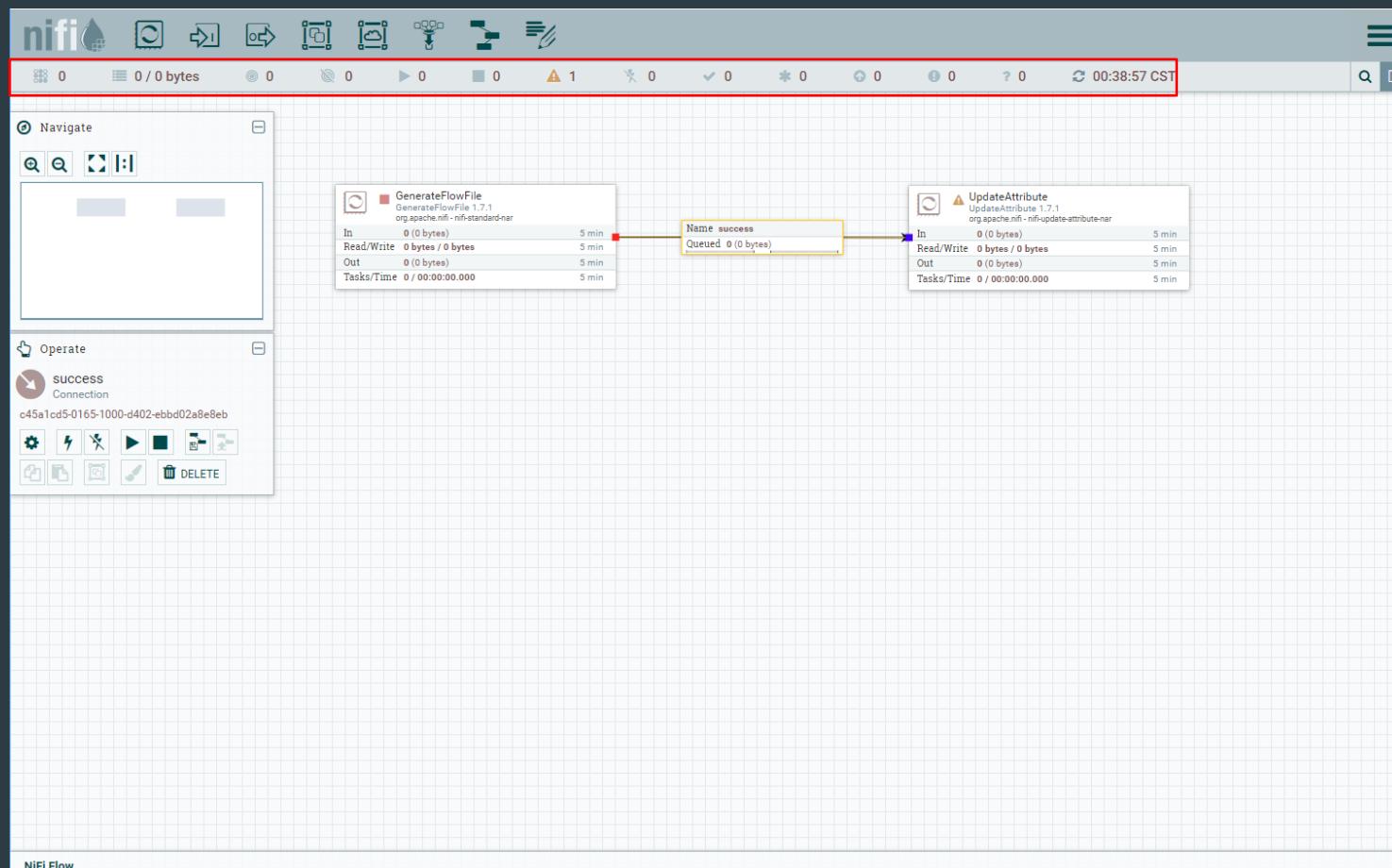


便利贴

在画布上对Dataflow做额外说明

Apache NIFI UI 介绍

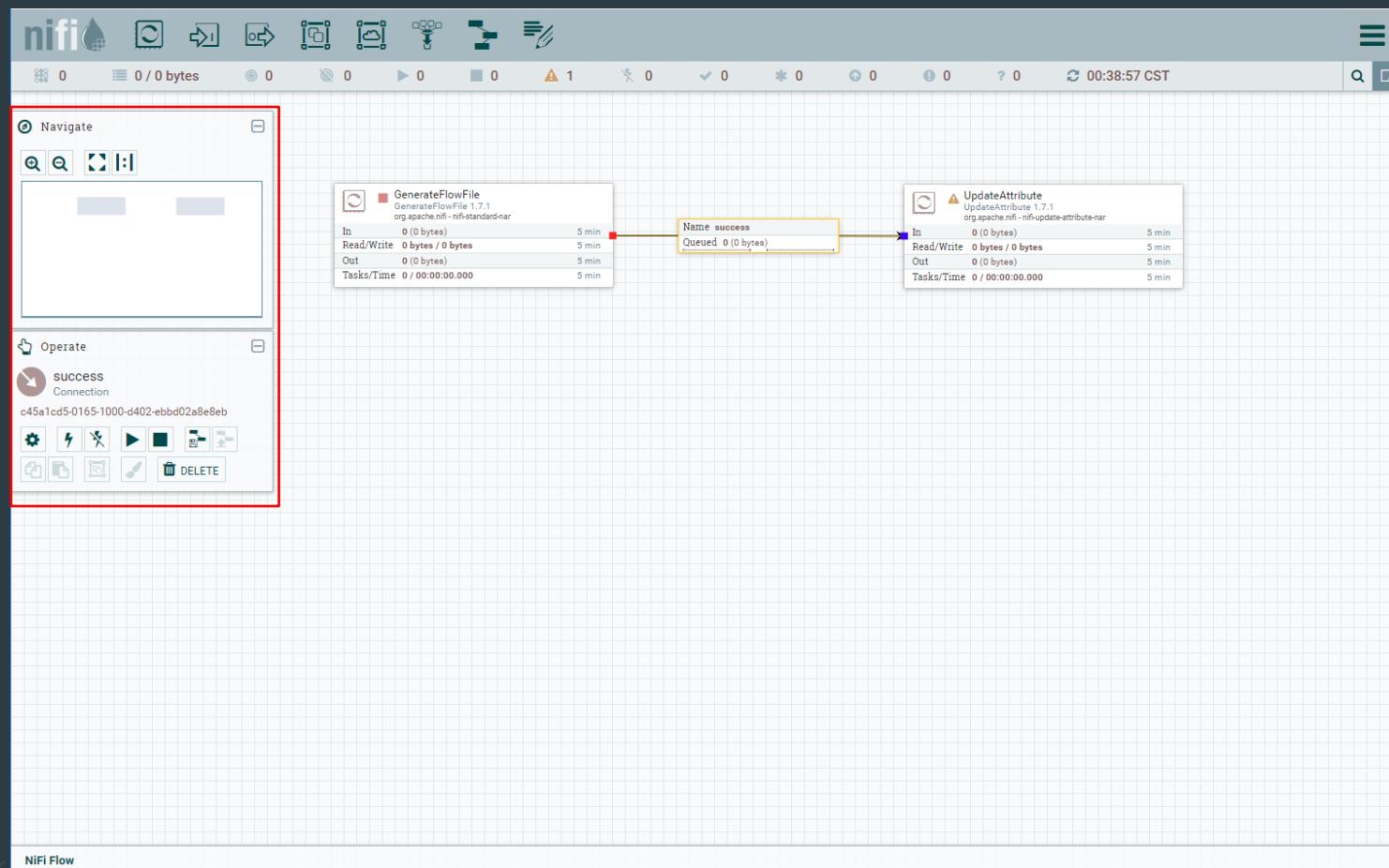
统计信息及状态栏



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

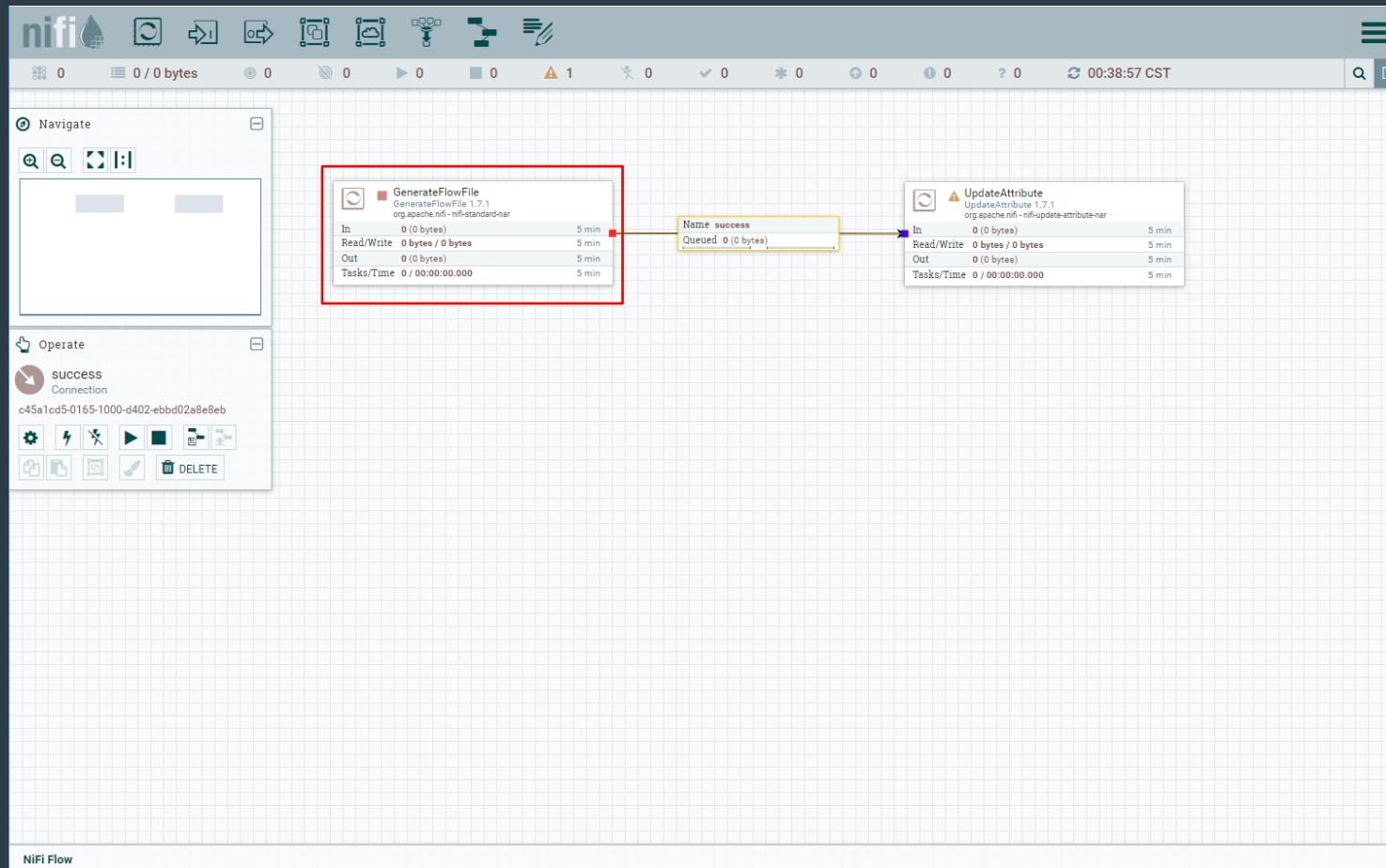
快速定位画布上有多少个处理器。

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



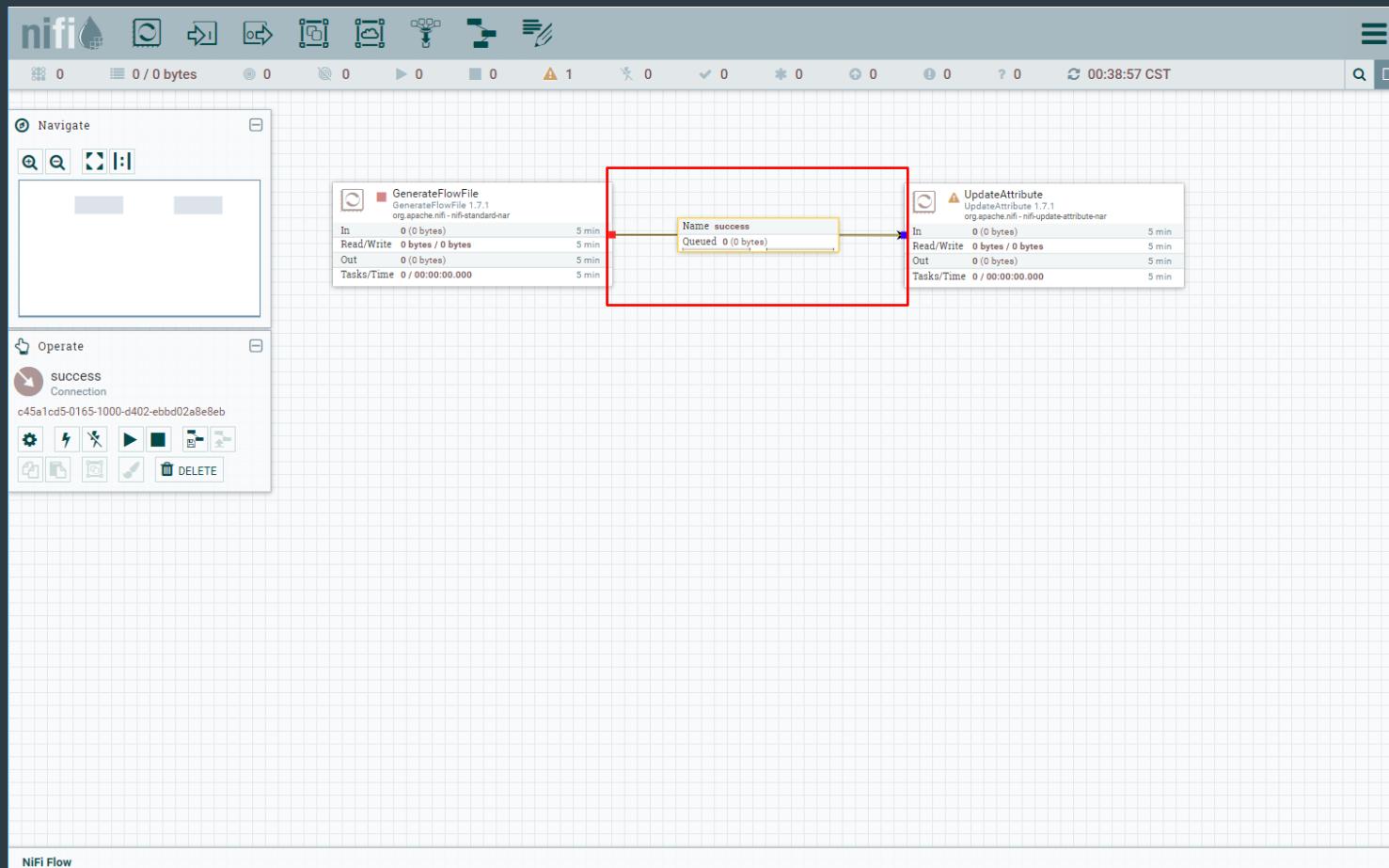
处理器

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

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Name: GenerateFlowFile Enabled: Automatically Terminate Relationships: success

Id: aa1efbb7-0165-1000-617f-27b1e434de25

Type: GenerateFlowFile 1.7.1

Bundle: org.apache.nifi - nifi-standard-nar

Penalty Duration: 30 sec

Yield Duration: 1 sec

Bulletin Level:

CANCEL APPLY

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

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Name: GenerateFlowFile Enabled

Id: aa1efbb7-0165-1000-617f-27b1e434de25

Type: GenerateFlowFile 1.7.1

Bundle: org.apache.nifi - nifi-standard-nar

Penalty Duration: 30 sec

Yield Duration: 1 sec

Bulletin Level:

Automatically Terminate Relationships success

CANCEL APPLY

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

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Scheduling Strategy: Timer driven

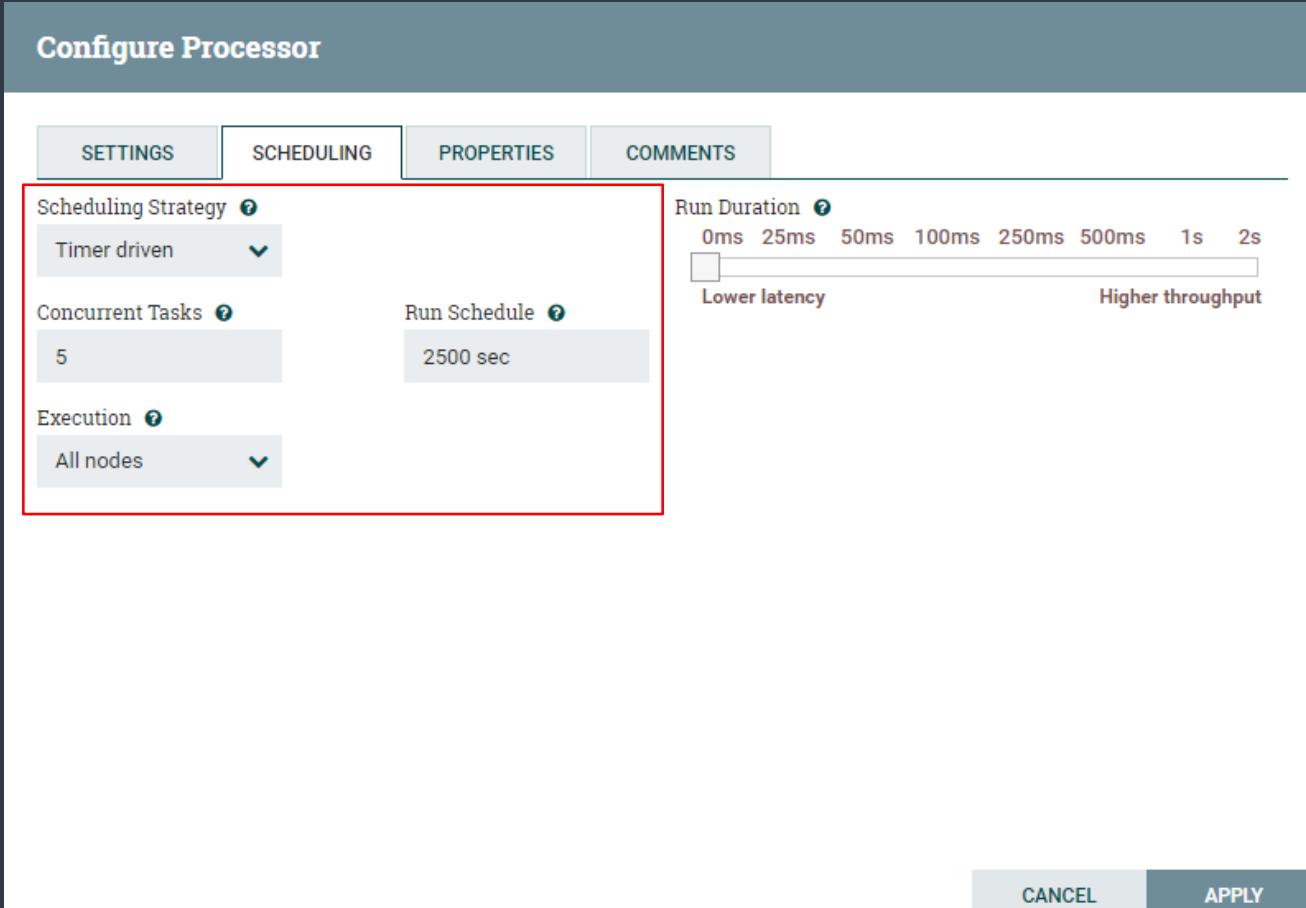
Concurrent Tasks: 5

Execution: All nodes

Run Duration: 2500 sec

Run Duration slider: 0ms to 2s (Lower latency to Higher throughput)

CANCEL APPLY



执行调度器计划

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

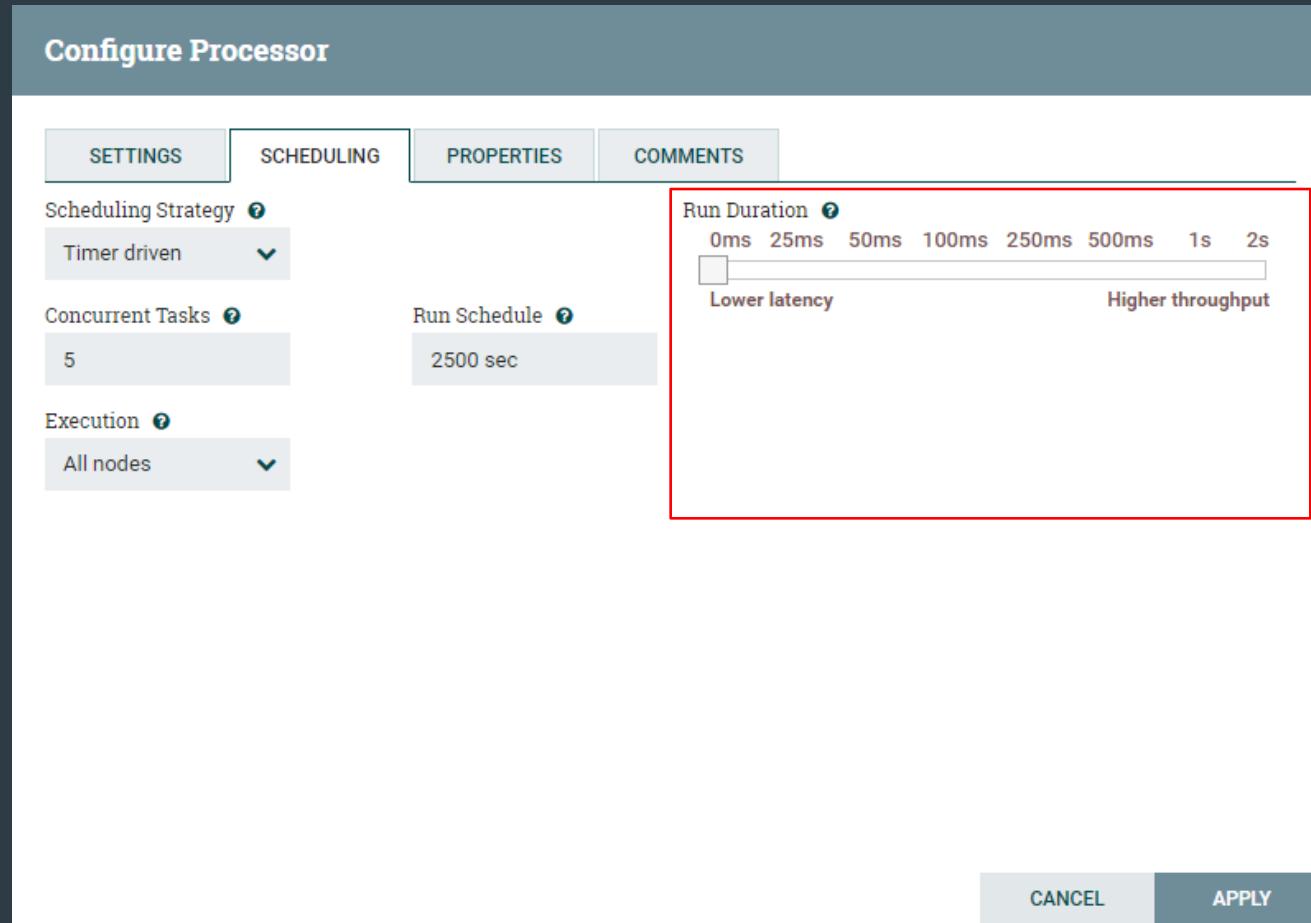
ConcurrentTaks: 并发线程

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

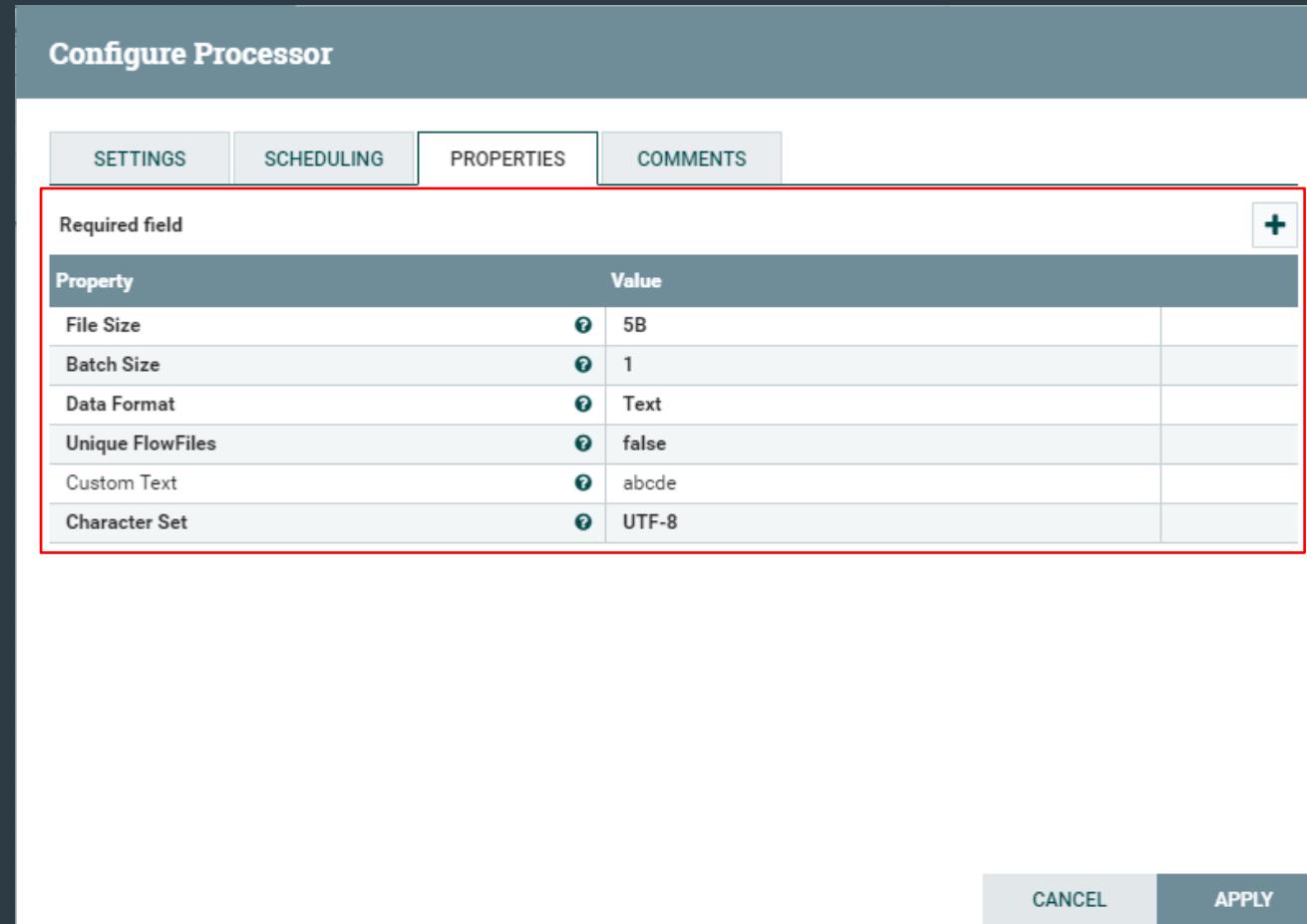
Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

Property	Value
File Size	5B
Batch Size	1
Data Format	Text
Unique FlowFiles	false
Custom Text	abcde
Character Set	UTF-8

CANCEL APPLY



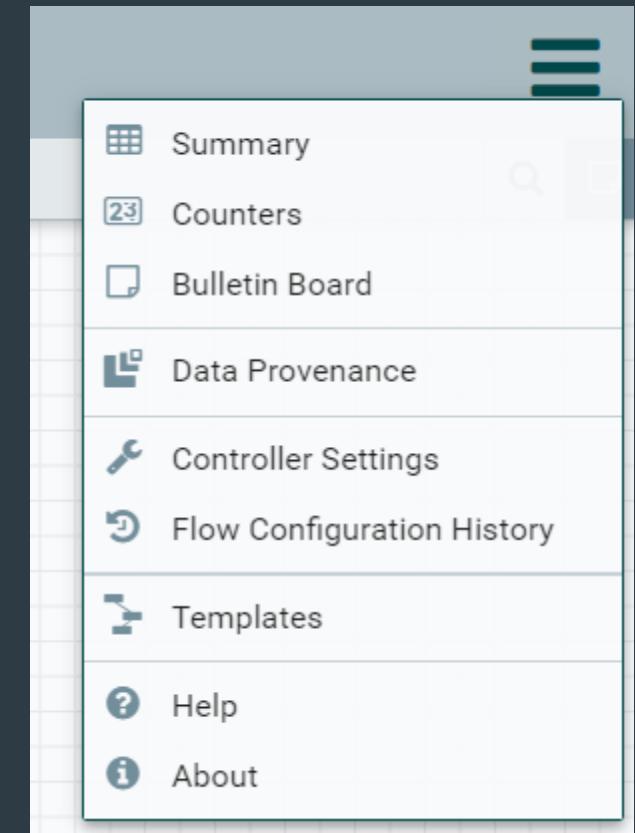
处理器相关配置

- 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)

NiFi Summary

PROCESSORS	INPUT PORTS	OUTPUT PORTS	REMOTE PROCESS GROUPS	CONNECTIONS	PROCESS GROUPS			
Displaying 9 of 9								
Filter		by name						
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)

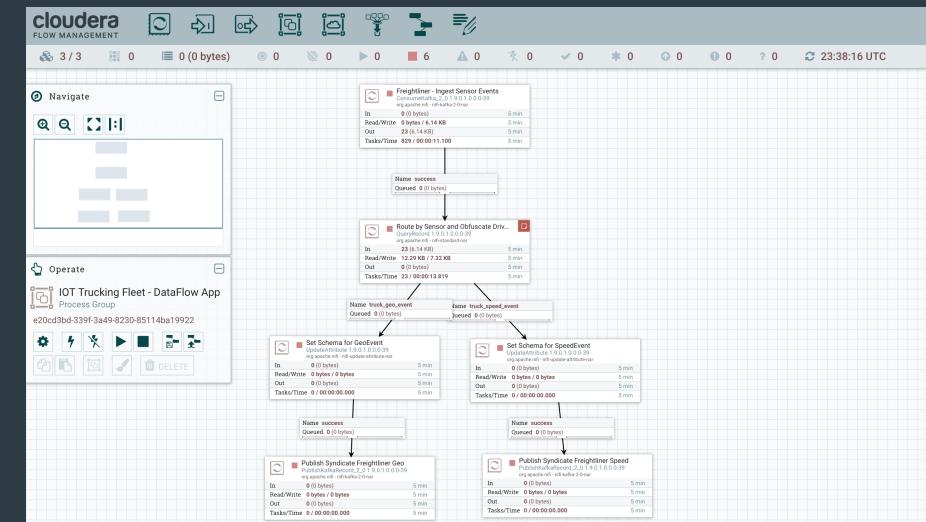
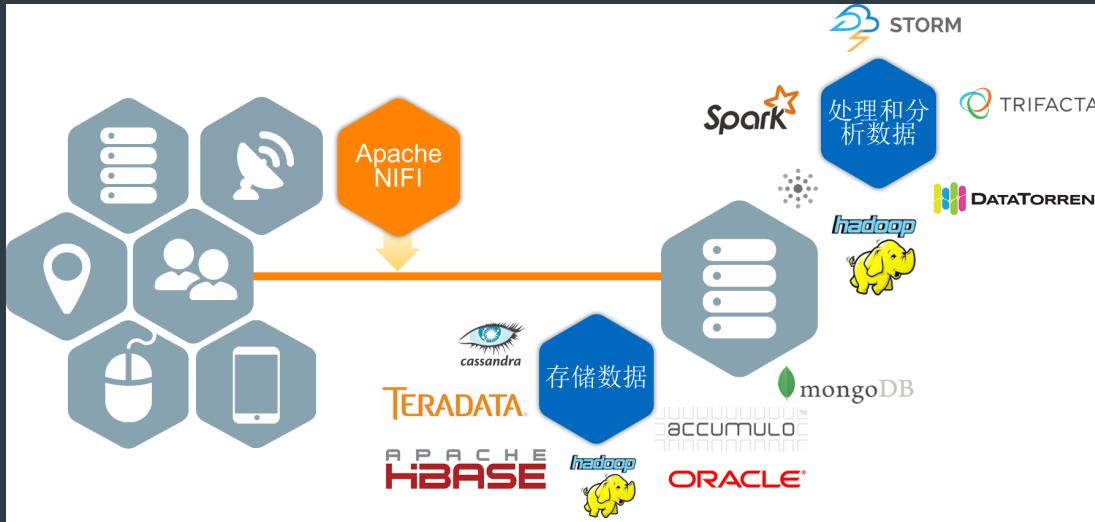
The screenshot shows the Apache NiFi Developer's Guide page. The left sidebar contains a navigation menu with sections like Toolkit Guide, Developer (Rest API, Developer Guide, Apache NiFi In Depth), and Processors (AttributeRollingWindow, AttributesToCSV, AttributesToJson, Base64EncodeContent, CalculateRecordStats, CaptureChangeMySQL, CompareFuzzyHash, CompressContent, ConnectWebSocket, ConsumeAMQP, ConsumeAzureEventHub, ConsumeEWS, ConsumeGCPubSub, ConsumeIMAP, ConsumeJMS, ConsumeKafka, ConsumeKafka_0_10). The main content area features the title "NiFi Developer's Guide" and the subtitle "Apache NiFi Team – dev@nifi.apache.org". Below the subtitle is a "Table of Contents" section with links to Introduction, NiFi Components, Processor API, Supporting API, FlowFile, ProcessSession, ProcessContext, PropertyDescriptor, Validator, ValidationContext, PropertyValue, Relationship, StateManager, ProcessorInitializationContext, ComponentLog, and Flowfile API.

The screenshot shows the documentation for the ConnectWebSocket processor. At the top, it says "- NiFi Documentation 1.9.2 ConnectWebSocket 1.9.2". Below that is a "Description" section stating: "Acts as a WebSocket client endpoint to interact with a remote WebSocket server. FlowFiles are transferred to downstream relationships according to received message types as WebSocket client configured with this processor receives messages from remote WebSocket server." There is also a "Tags" section listing "subscribe, WebSocket, consume, listen". A "Properties" section follows, with a table showing properties like Name (WebSocket Client ControllerService), Default Value (Controller Service API: WebSocketClientService), Allowable Values (Implementation: jettyWebSocketClient), and Description (A WebSocket CLIENT Controller Service which can connect to a WebSocket server). Below this is a "Relationships" section with a table showing Name (WebSocket Client Id) and Description (The client ID to identify WebSocket session. It should be unique within the WebSocket Client Controller Service. Otherwise, it throws WebSocketConfigurationException when it gets started).

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>

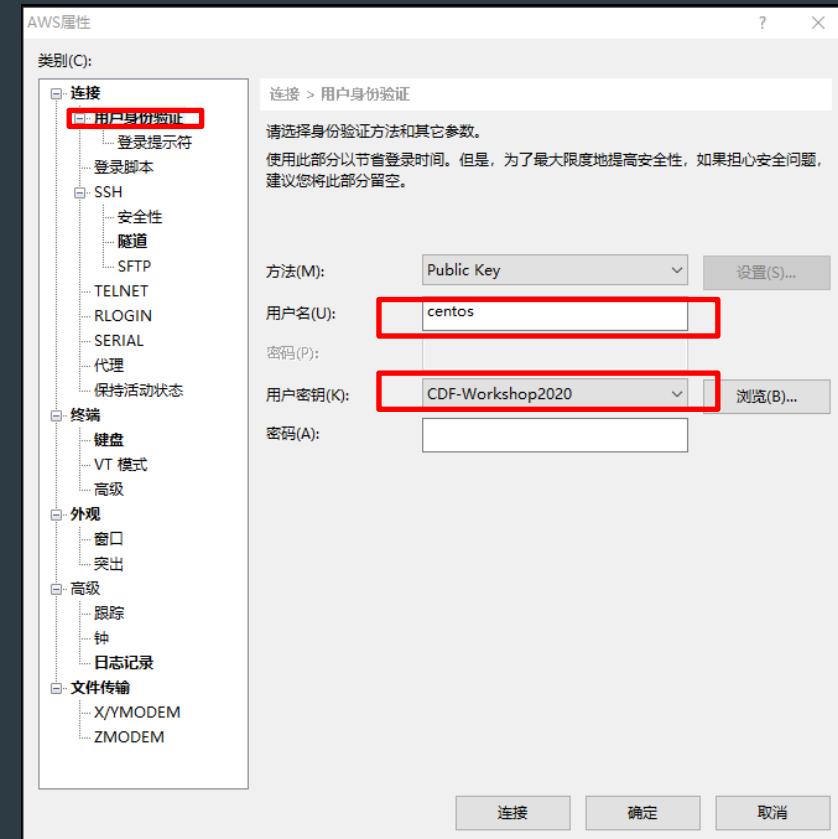
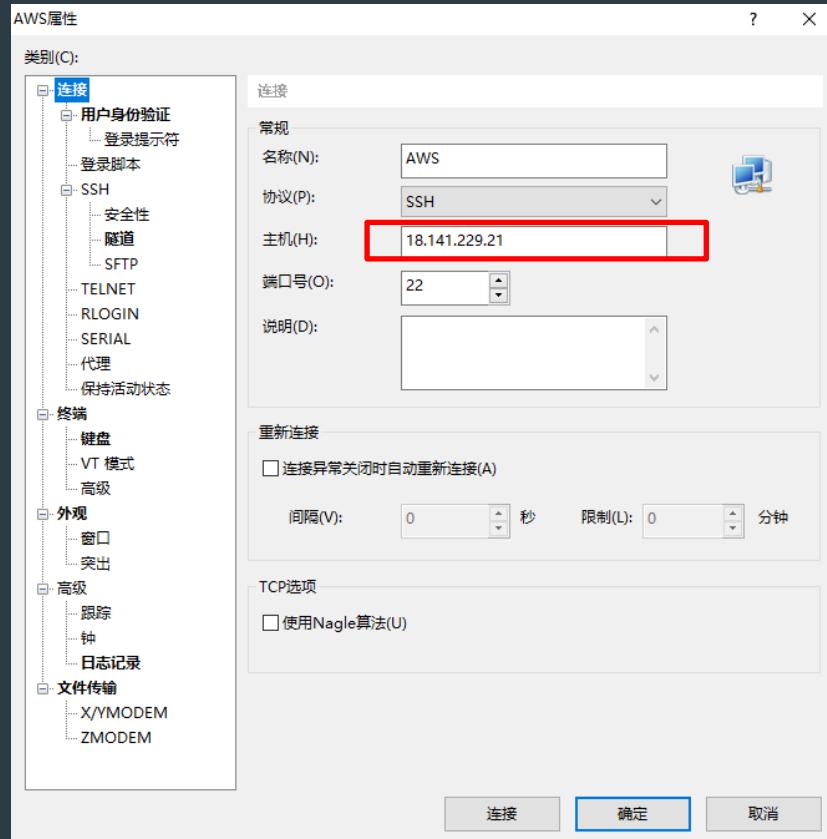
LAB 1

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
 - Mac, Linux用户- 下载PEM KEY文件 - chmod 400
 - Windows用户- 下载PPK KEY文件

LAB 1

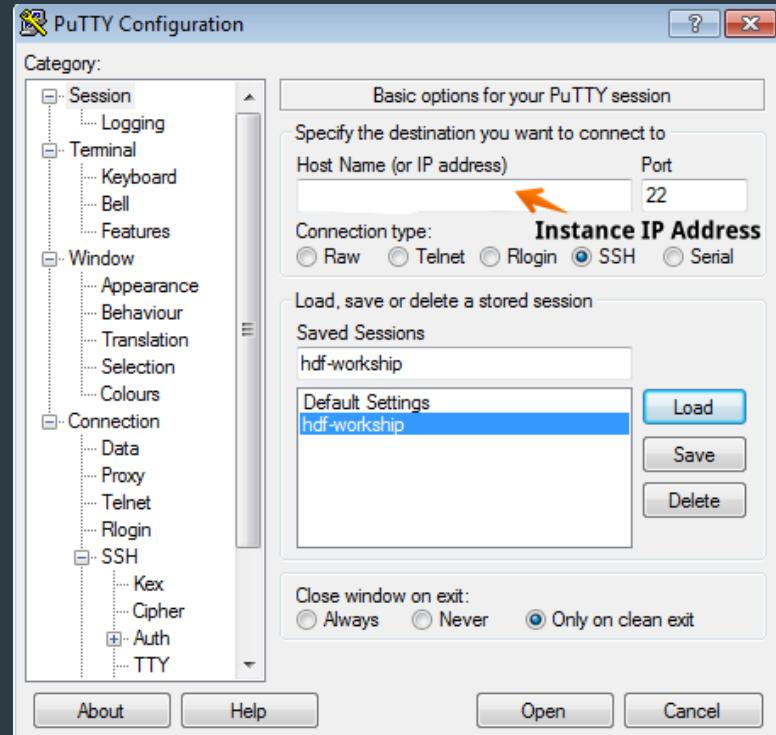
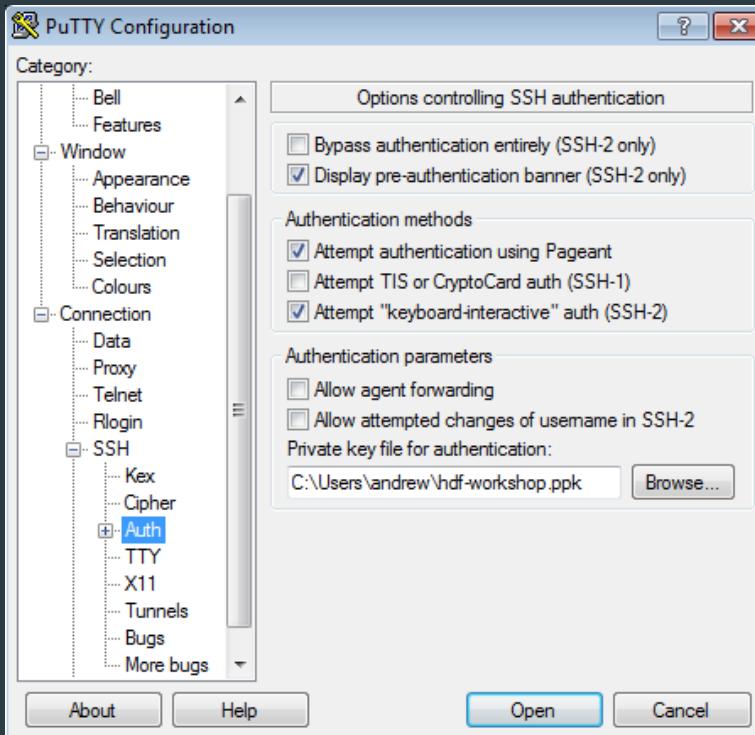
登录你的实验环境 - Xshell



动手实战

Lab 1

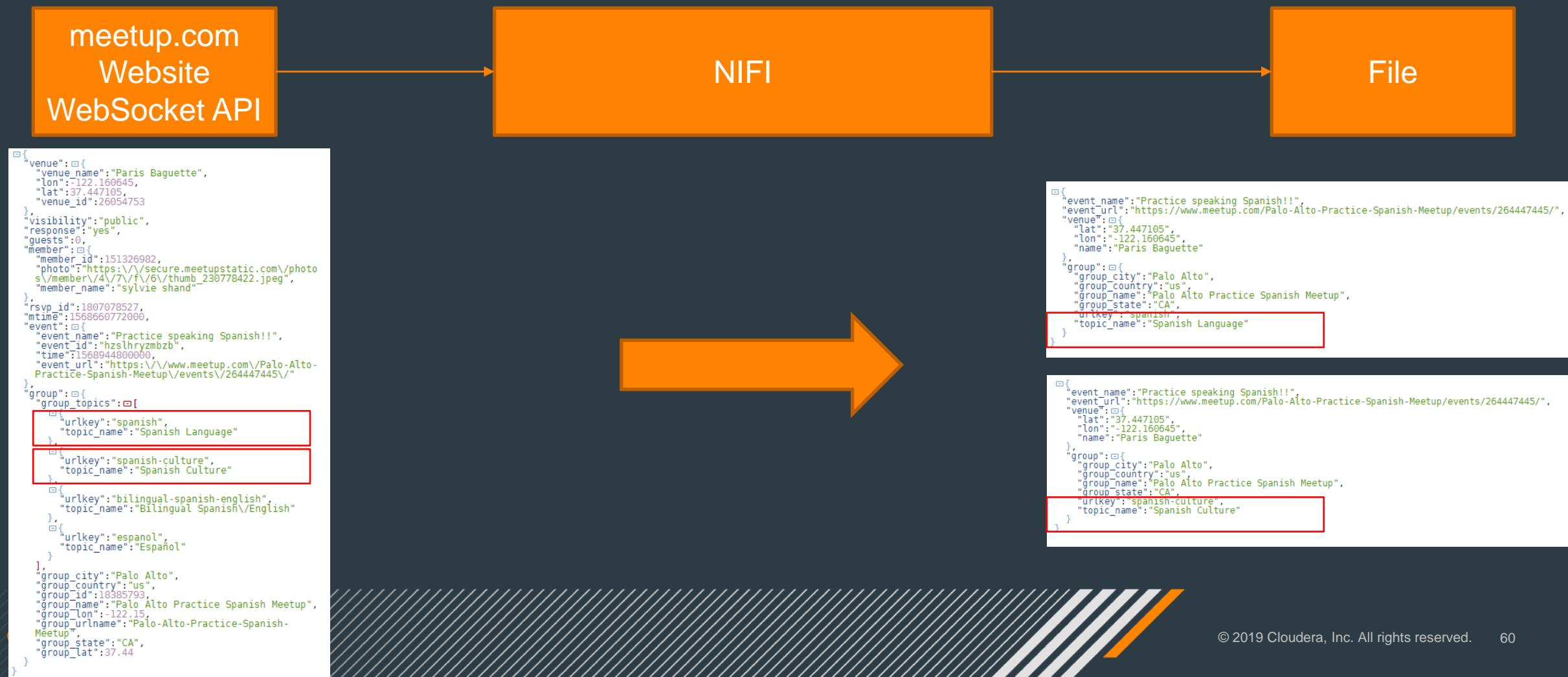
- Windows - 进入OS环境



- 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



Lab 2

使用NIFI构建第一个数据流

- 添加一个ConnectWebSocket 处理器

Source	Displaying 3 of 251		
all groups	Type	Version	Tags
	ConnectWebSocket	1.5.0.3.1.0.0-564	subscribe, consume, listen, We...
	ListenWebSocket	1.5.0.3.1.0.0-564	subscribe, consume, listen, We...
	PutWebSocket	1.5.0.3.1.0.0-564	publish, send, WebSocket

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

The screenshot shows two panels of the Apache NiFi web interface.

Left Panel: Add Controller Service

- Required field:** Property
- WebSocket Client ControllerService:** JettyWebSocketClient 1.9.2 from org.apache.nifi - nifi-websocket-services-api-nar
- Compatible Controller Services:** JettyWebSocketClient 1.9.2
- Controller Service Name:** JettyWebSocketClient 1.9.2
- Bundle:** org.apache.nifi - nifi-websocket-services-jetty-nar
- Tags:** client, Jetty, WebSocket
- Description:** (empty)

Buttons at the bottom: CANCEL, CREATE

Bottom Buttons: CANCEL, APPLY

Right Panel: Configure Processor

- Required field:** WebSocket Client ControllerService: JettyWebSocketClient (highlighted with a red box)
- WebSocket Client Id:** No value set

Buttons at the top: SETTINGS, SCHEDULING, PROPERTIES, COMMENTS

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

NiFi Flow Configuration

Name	Type	Bundle	State	Scope	Action
CSVReader	CSVReader 1.9.2	org.apache.nifi - nifi-rec...	Enabled	NiFi Flow	
CSVRecordSetWriter	CSVRecordSetWriter 1...	org.apache.nifi - nifi-rec...	Enabled	NiFi Flow	
JettyWebSocketClient	JettyWebSocketClient 1...	org.apache.nifi - nifi-we...	Invalid	NiFi Flow	

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

Controller Service Details

SETTINGS	PROPERTIES	COMMENTS
Required field		
Property	Value	
Input Buffer Size	?	4 kb
Max Text Message Size	?	64 kb
Max Binary Message Size	?	64 kb
WebSocket URI	?	ws://stream.meetup.com/2/rsvps
SSL Context Service	?	No value set
Connection Timeout	?	3 sec
Session Maintenance Interval	?	10 sec

Lab 2

使用NIFI构建第一个数据流

- 启用Jetty客户端服务

NiFi Flow Configuration

Name	Type	Bundle	State	Scope	Action
CSVReader	CSVReader	CSVReader 1.9.2	Enabled	NiFi Flow	
CSVRecordSetWriter	CSVRecordSetWriter	CSVRecordSetWriter 1....	Enabled	NiFi Flow	
JettyWebSocketClient	JettyWebSocketClient	JettyWebSocketClient 1...	Disabled	NiFi Flow	

Lab 2

使用NIFI构建第一个数据流

- 添加Update Attribute 处理器

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

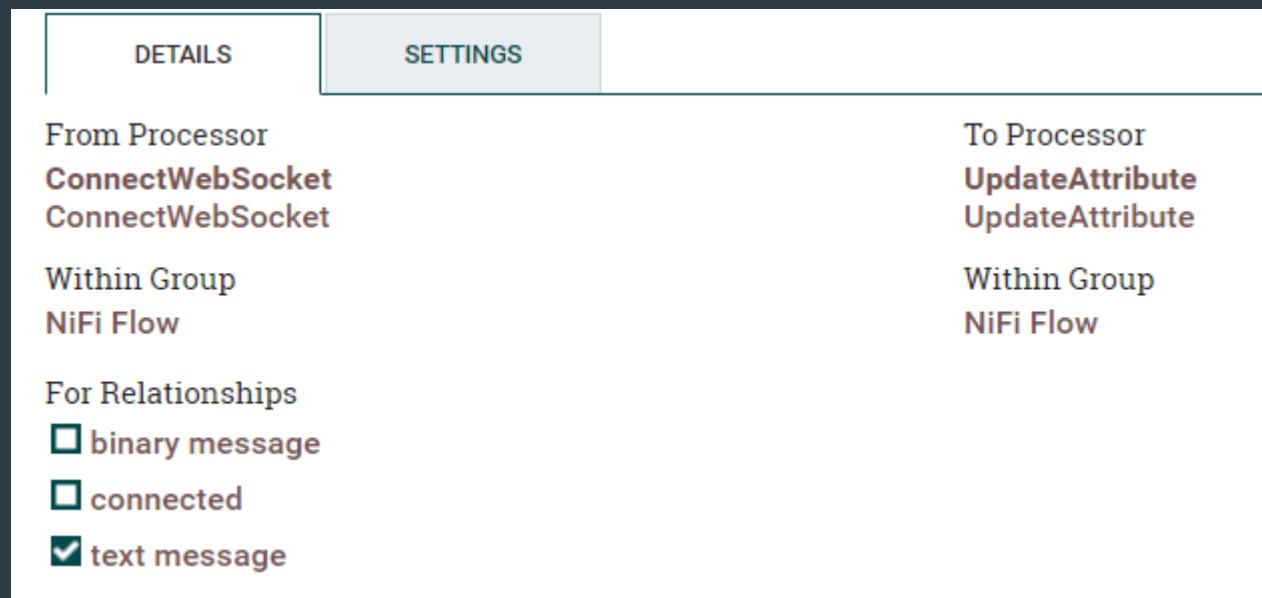
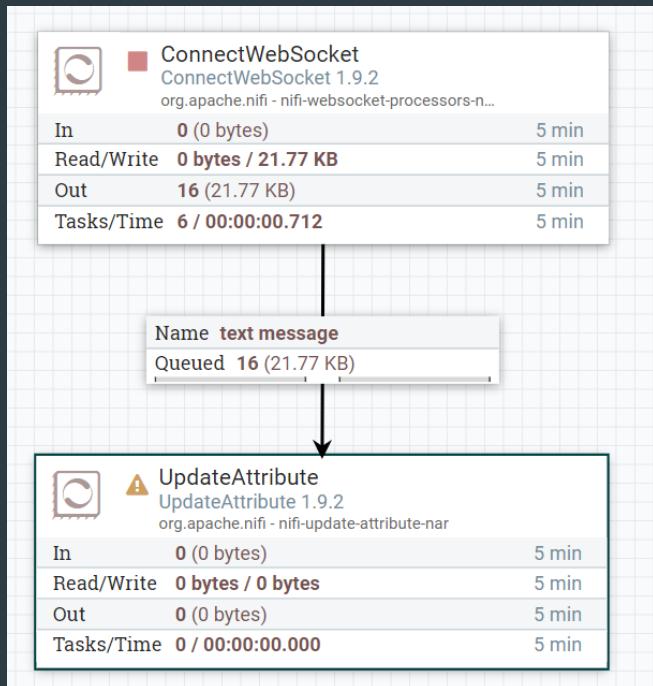
Property	Value	
Delete Attributes Expression	?	No value set
Store State	?	Do not store state
Stateful Variables Initial Value	?	No value set
mime.type	?	application/json

+

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

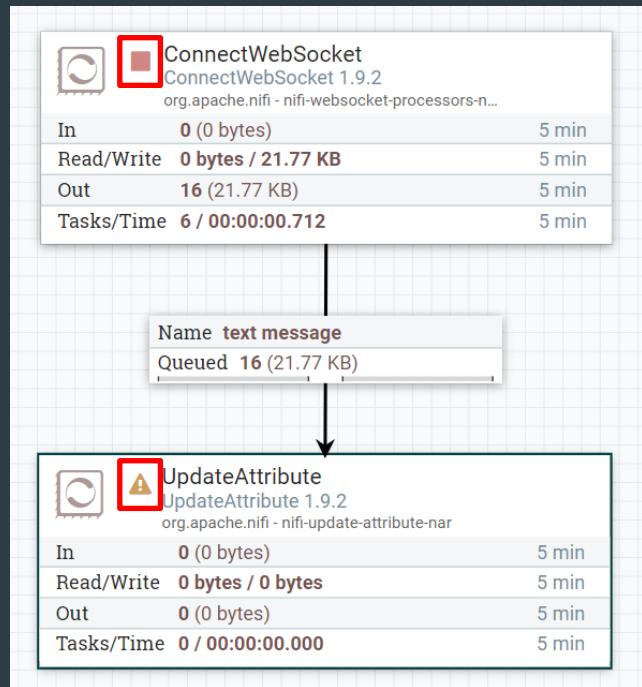
Configure Processor

SETTINGS	SCHEDULING	PROPERTIES	COMMENTS
Name ConnectWebSocket	<input checked="" type="checkbox"/> Enabled	Automatically Terminate Relationships ?	
Id 3b71cc12-016d-1000-720b-dd103950401a		<input checked="" type="checkbox"/> binary message The WebSocket binary message output	
Type ConnectWebSocket 1.9.2		<input checked="" type="checkbox"/> connected The WebSocket session is established	
		<input type="checkbox"/> text message The WebSocket text message output	

Lab 2

使用NIFI构建第一个数据流

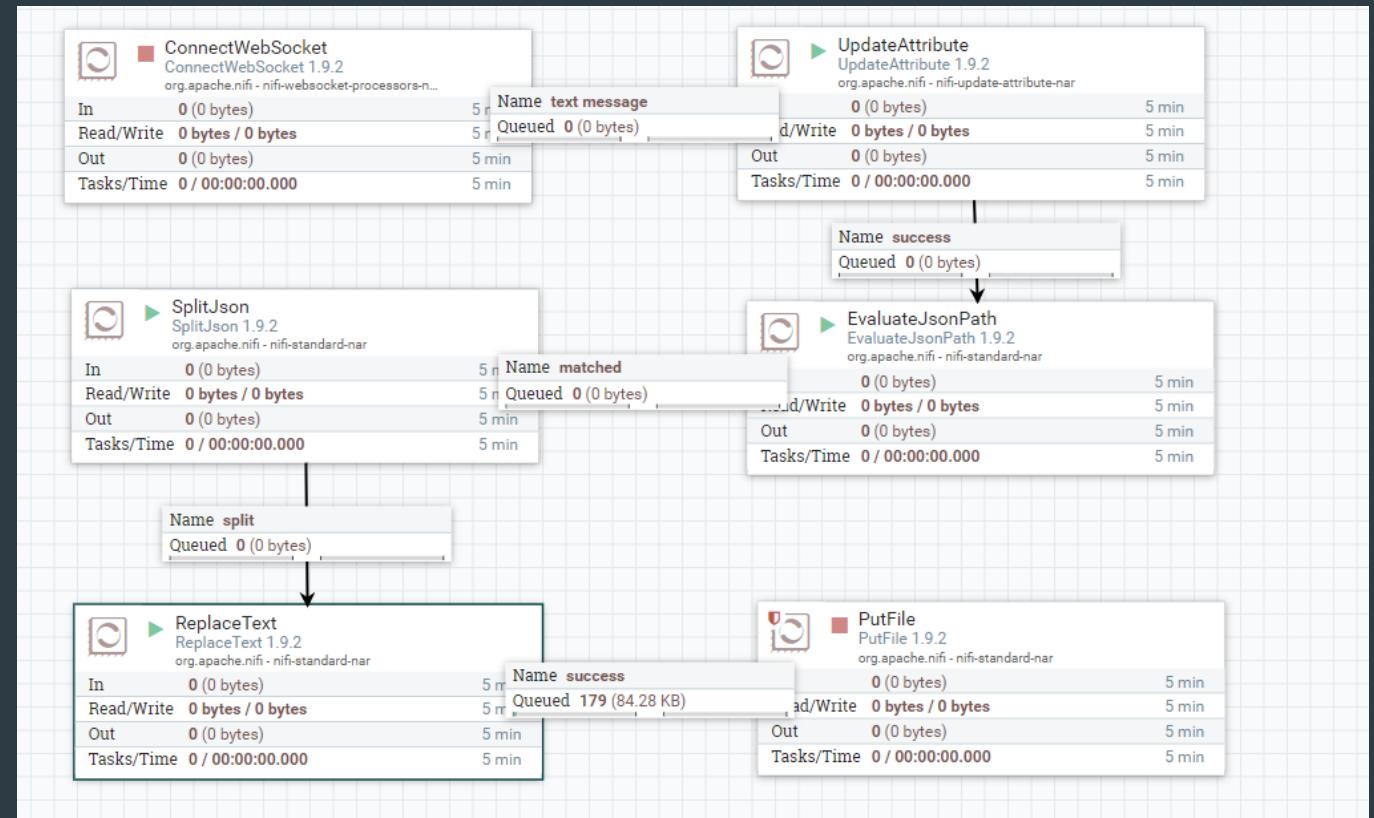
- 当ConnectWebSocket为红色方块时代表处理器基本配置已满足，可运行



Lab 2

使用NIFI构建第一个数据流

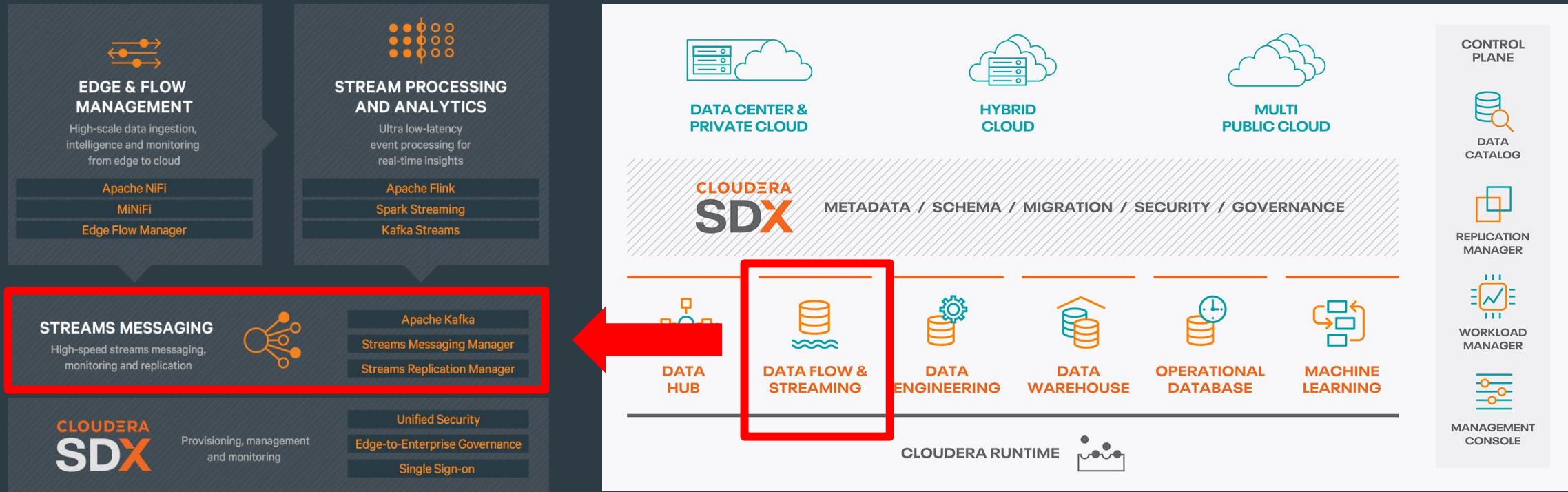
- 完成后续的数据流构建



Apache Kafka 介绍

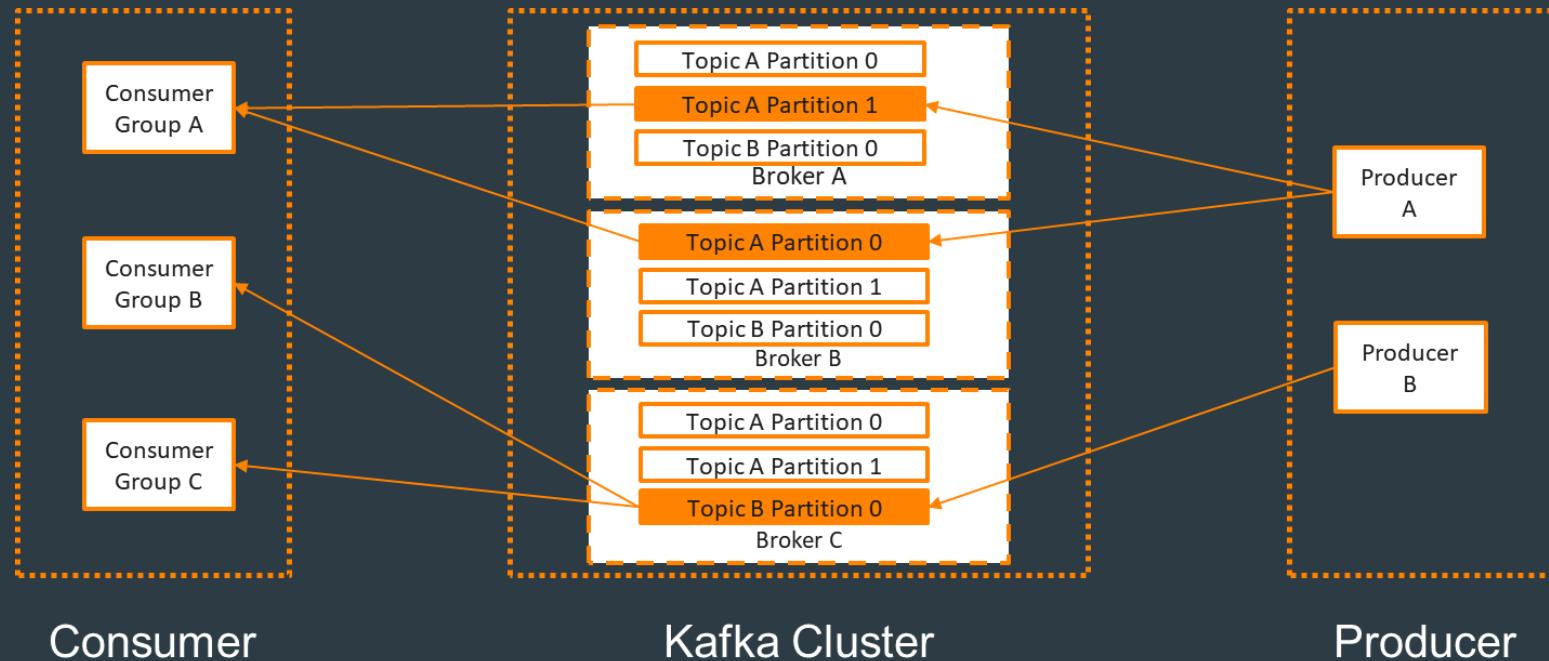
Cloudera DataFlow (CDF)

- 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

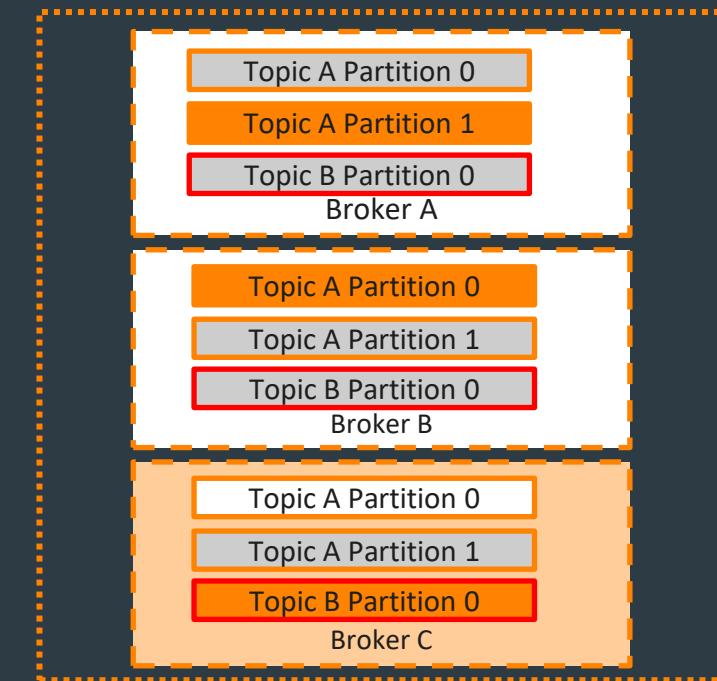
Apache Kafka介绍

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

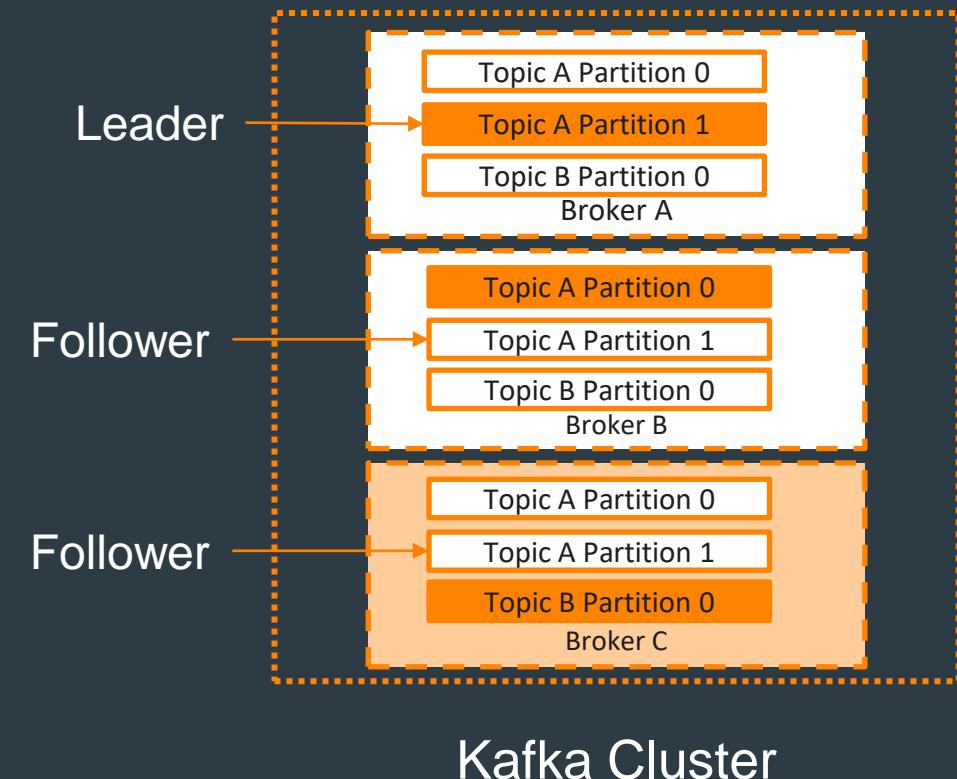
Topic B 存在Partition 0



Apache Kafka介绍

Apache Kafka – 核心概念

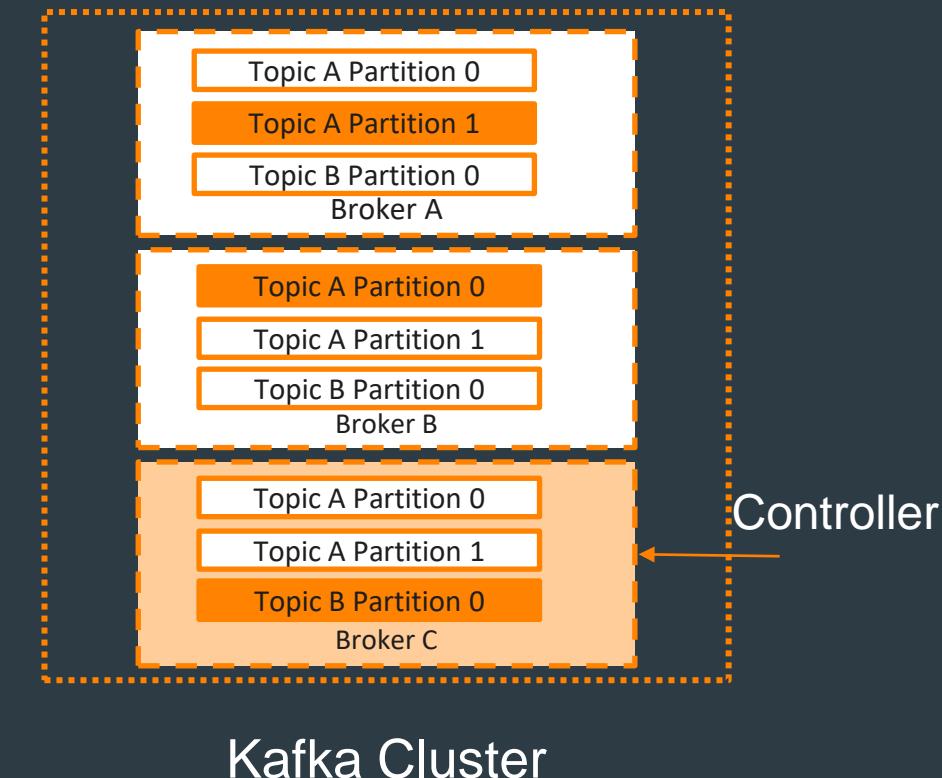
- Broker
 - Partition Replica
 - Partition Leader 和 Follower
 - ISR (In-Sync-Replica)



Apache Kafka介绍

Apache Kafka – 核心概念

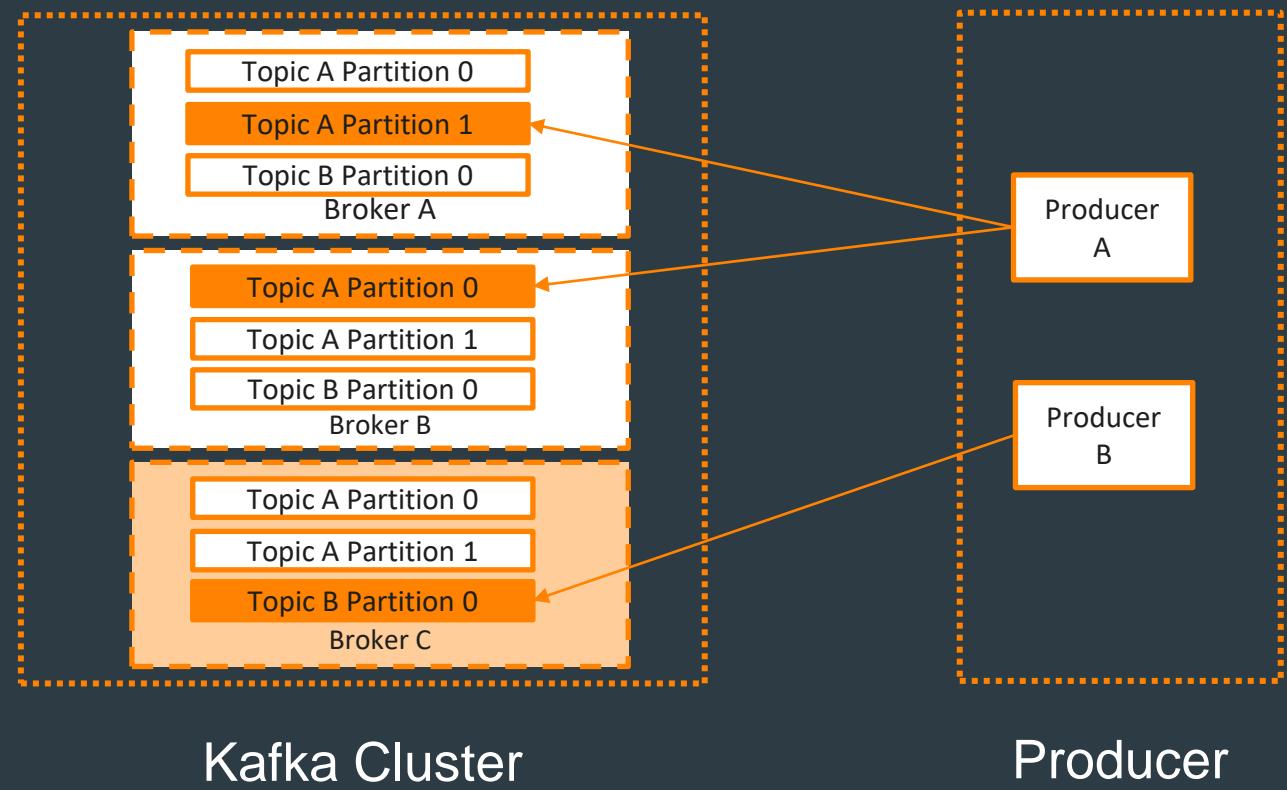
- Broker
 - Controller
 - 每一个Kafka 集群有且只有一个Controller
 - 负责监控其他Broker的活动
 - 维护Topic的CRUD
 - 选举新的Partition Leader
 - 同步新的Partition Leader至Broker



Apache Kafka介绍

Apache Kafka – 核心概念

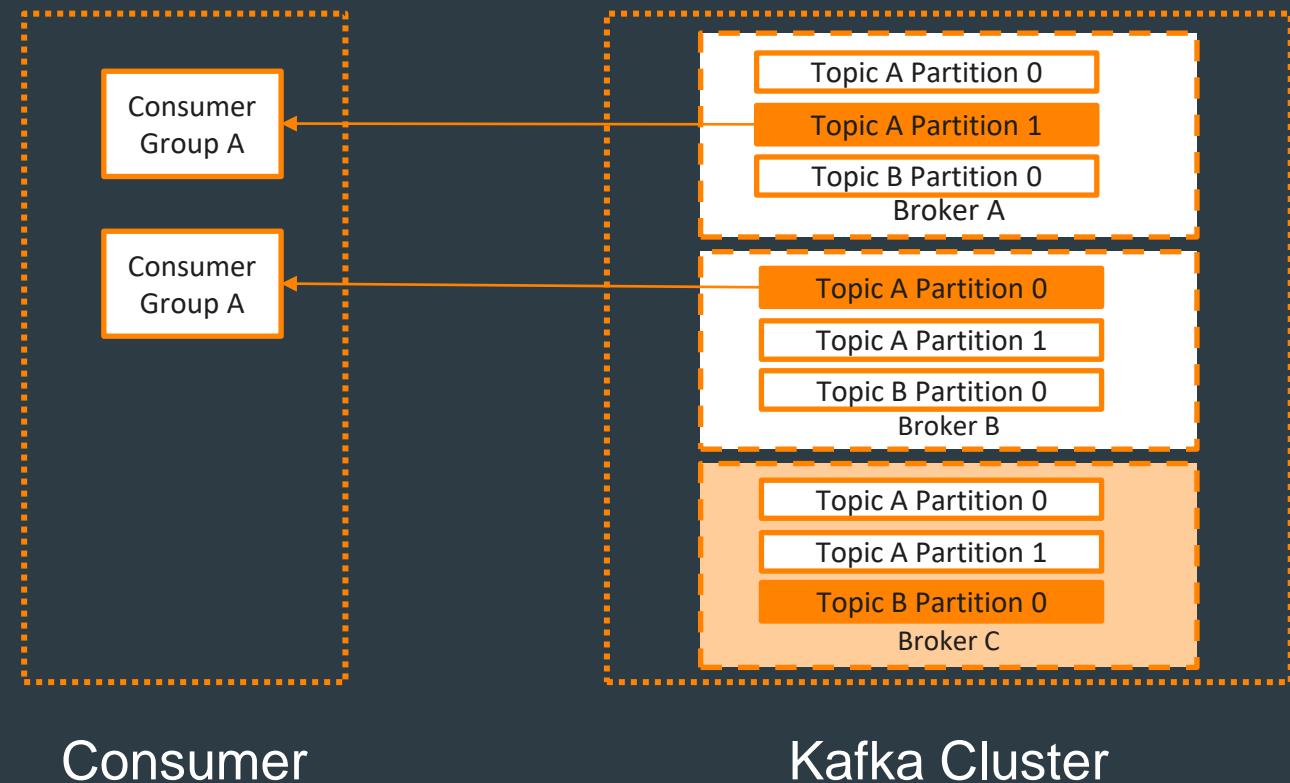
- Producer (客户端)
 - Partitioner
 - RoundRobin
 - Key
 - Custom
 - ACKs
 - Transaction



Apache Kafka介绍

Apache Kafka – 核心概念

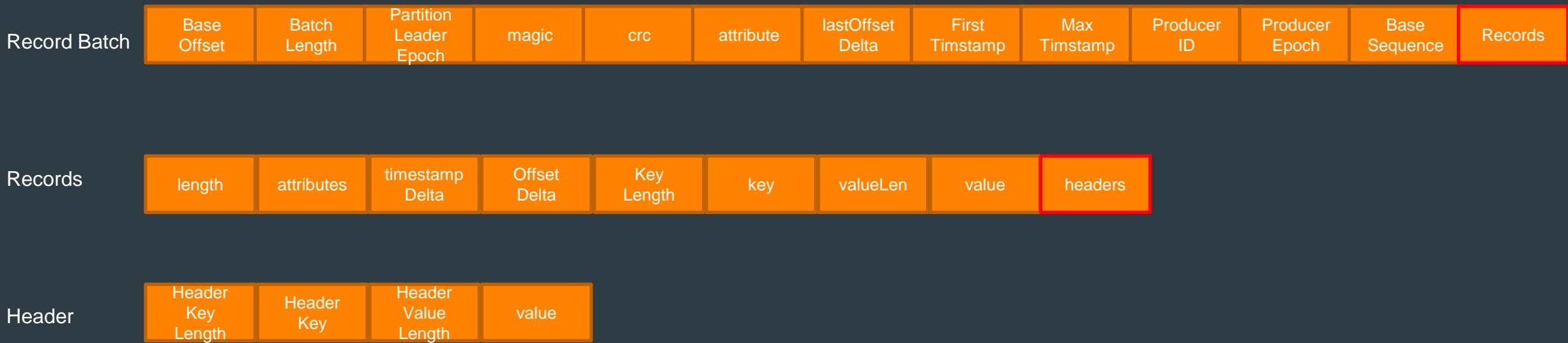
- Consumer (客户端)
 - Consumer Group
 - Consumer Rebalance
 - Consumer Lag
 - Read Committed



Apache Kafka介绍

Apache Kafka – 核心概念

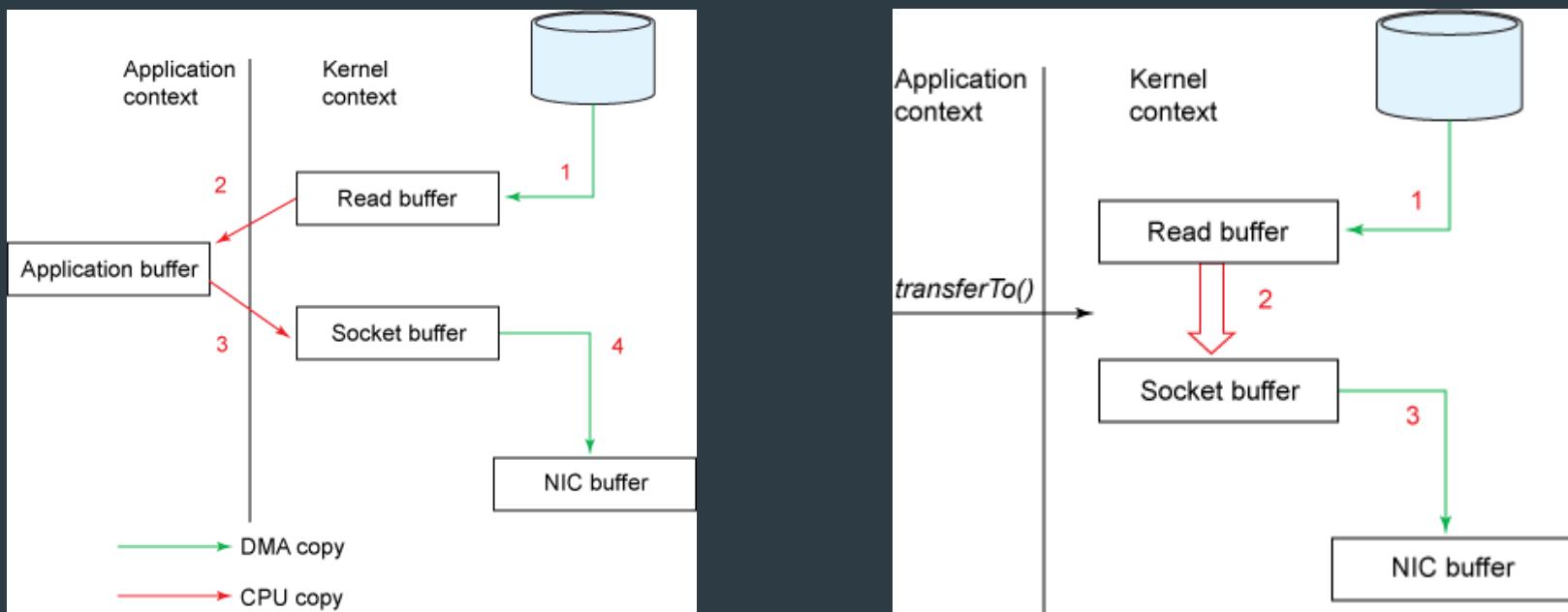
- Broker
 - Message



Apache Kafka介绍

Apache Kafka – 核心概念

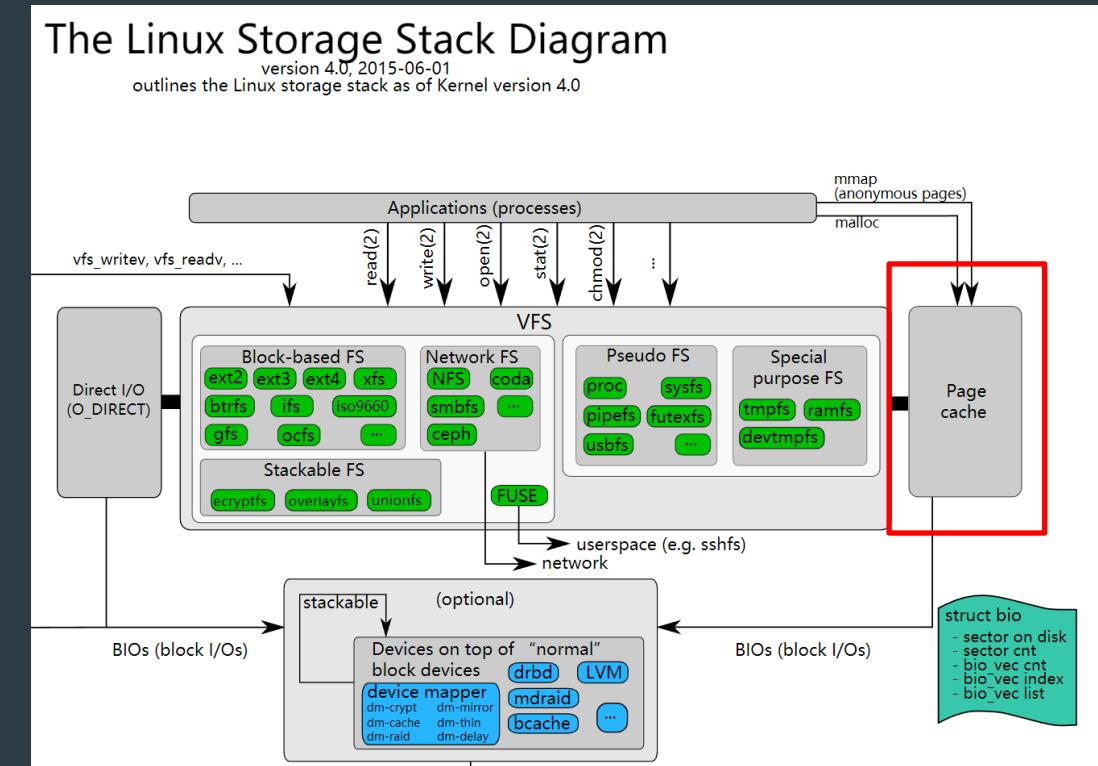
- Kafka 为什么这么快
 - Zero Copy
 - Page Cache



Apache Kafka介绍

Apache Kafka – 核心概念

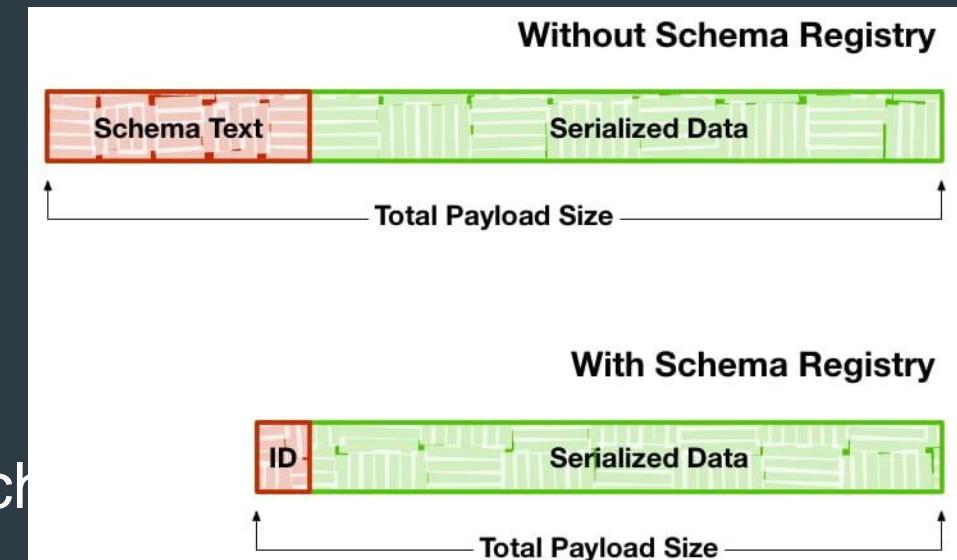
- Kafka 为什么这么快
 - Zero Copy
 - Page Cache



Schema Registry

Schema Registry概念

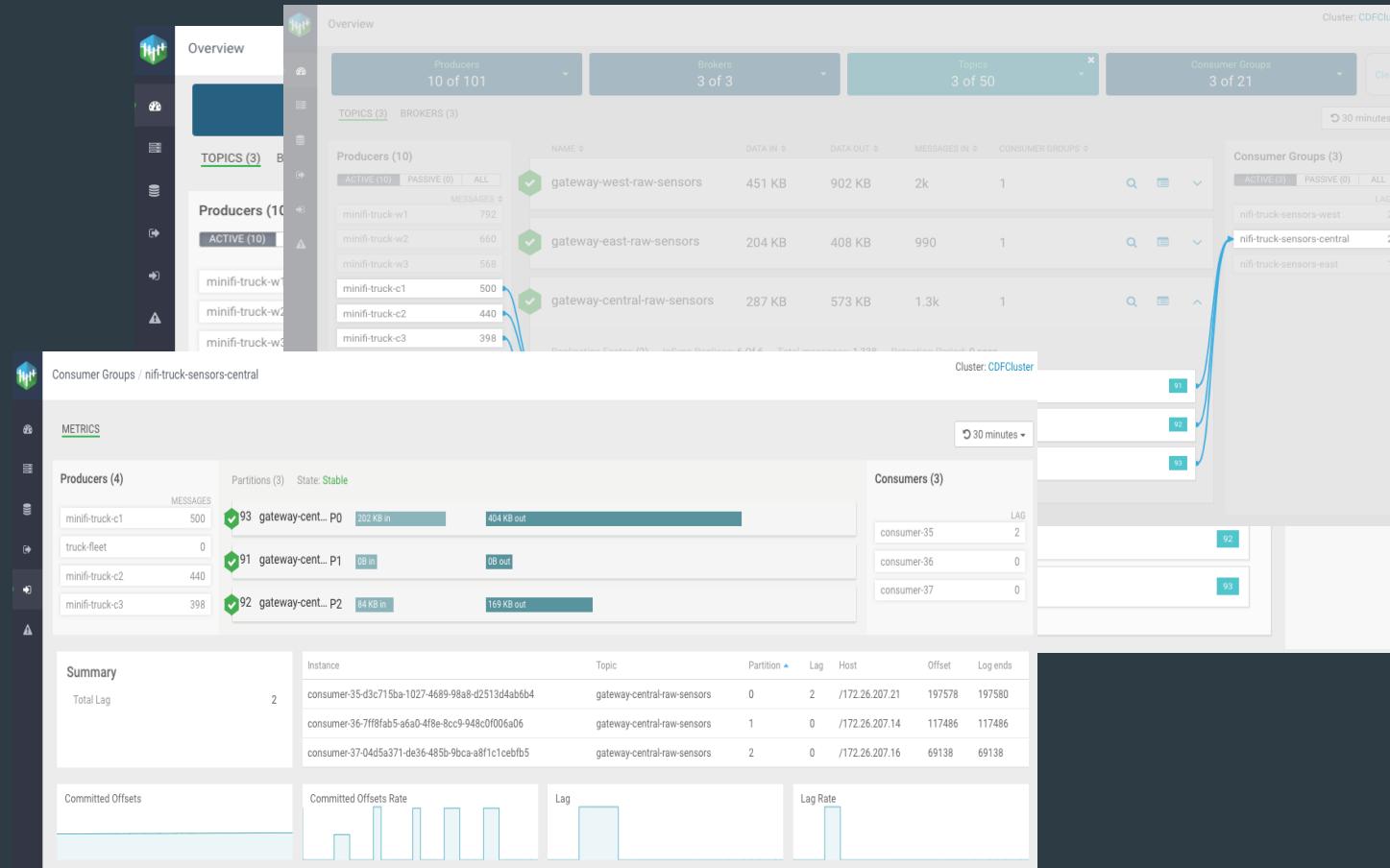
- 一个公共的Schema服务，用于保存或读取Schema
 - 1. 更好的元数据管理
 - 提供了可重用的Schema
 - 提供了Schema 版本管理
 - 2 运维效率
 - 避免了在消息本体中添加Schema
 - 消费者和生产者不需要完全统一固定Schema
 - 数据校验 Schema Validation



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

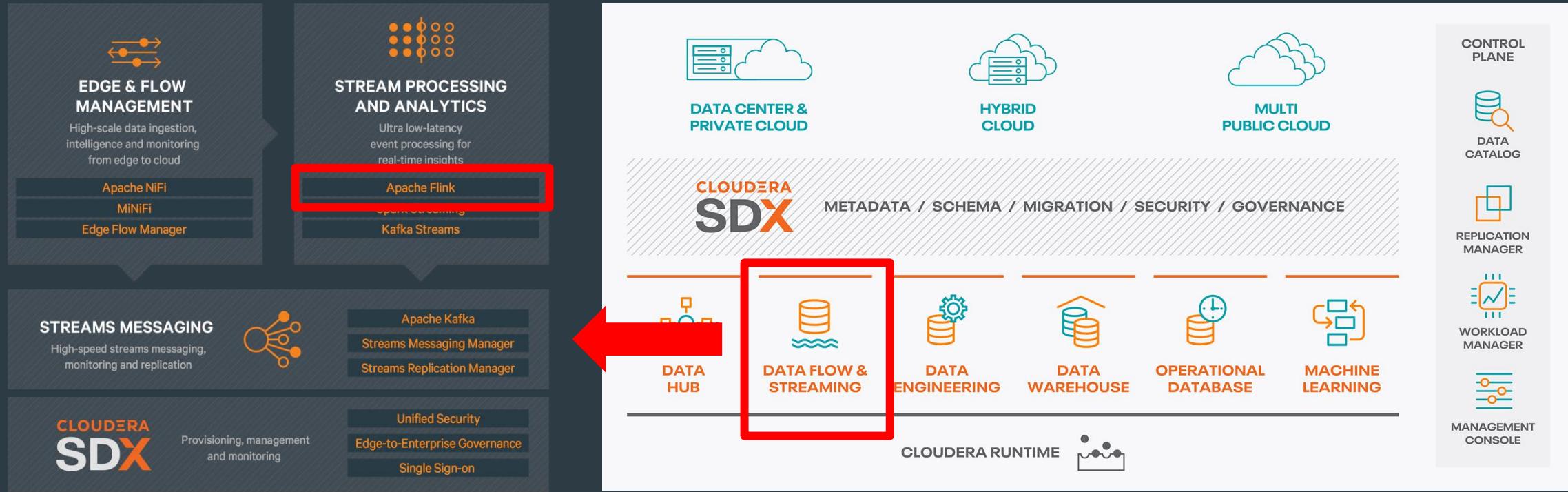
Lab 3 & 4

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

Apache Flink

Apache Flink 项目介绍

- Cloudera CSA



Apache Flink

目录

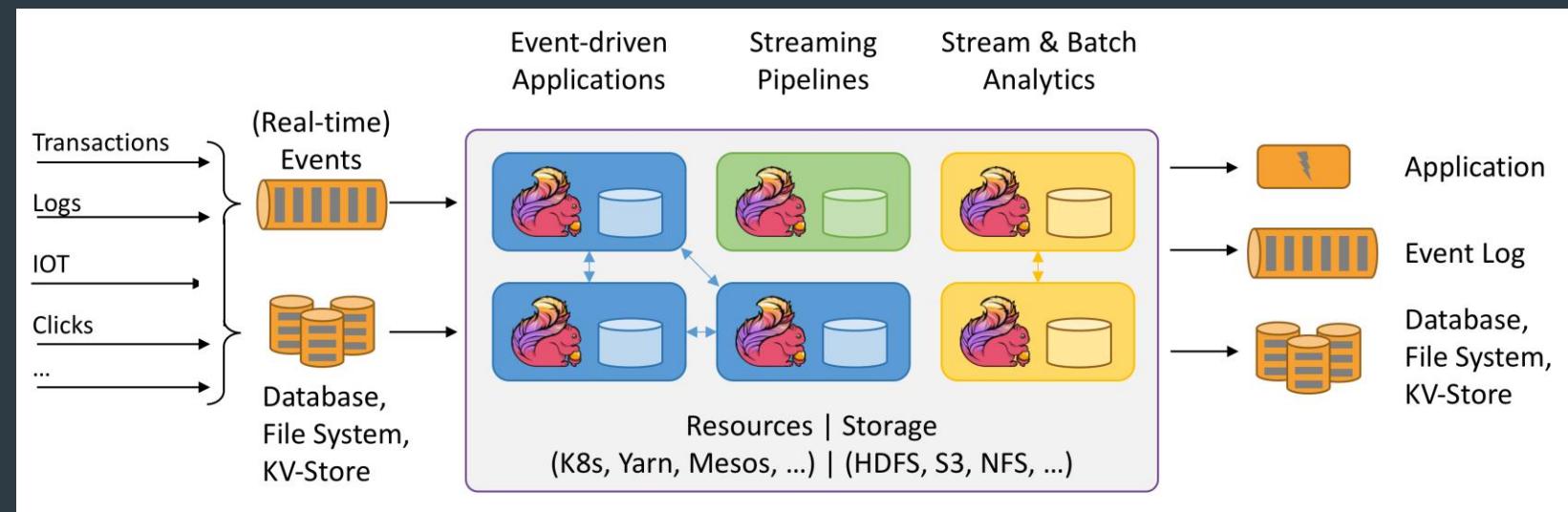
- Apache Flink 项目介绍
- Apache Flink 基本能力介绍
- Apache Flink 适用场景介绍
- Cloudera CSA 模块介绍
- Apache Flink 未来展望

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

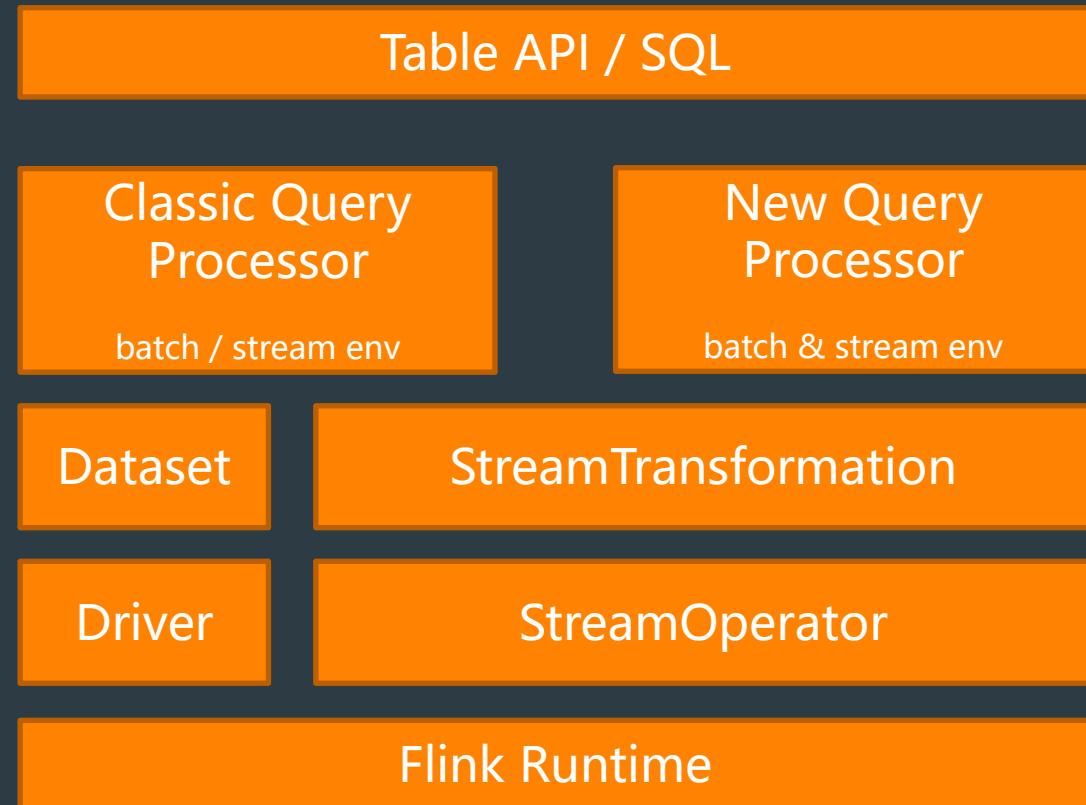
- 低延时
- 高吞吐
- 有状态
- 分布式



Apache Flink

Apache Flink 项目介绍

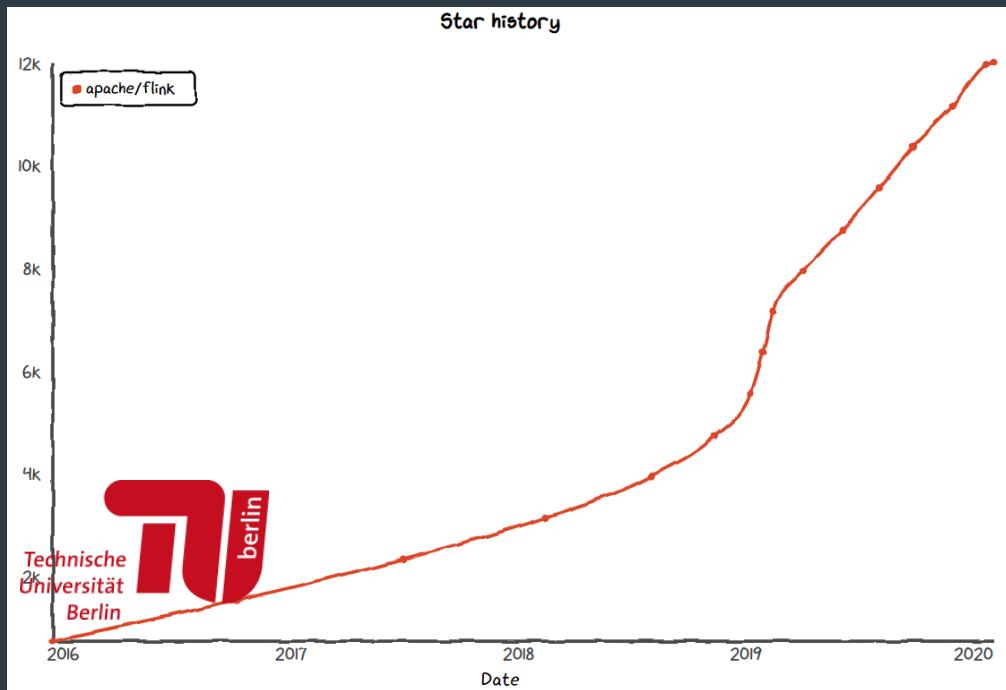
- Flink API
 - Stateful Stream Processing
 - DataStream/ DataSet API
 - Table API/ SQL



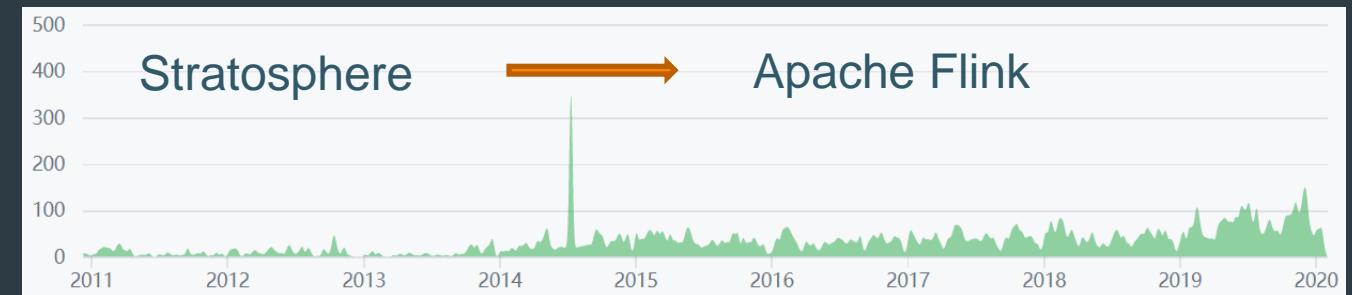
Apache Flink

Apache Flink 项目介绍

- 项目背景



CLOUDERA



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/powerby.html>

Apache Flink

目录

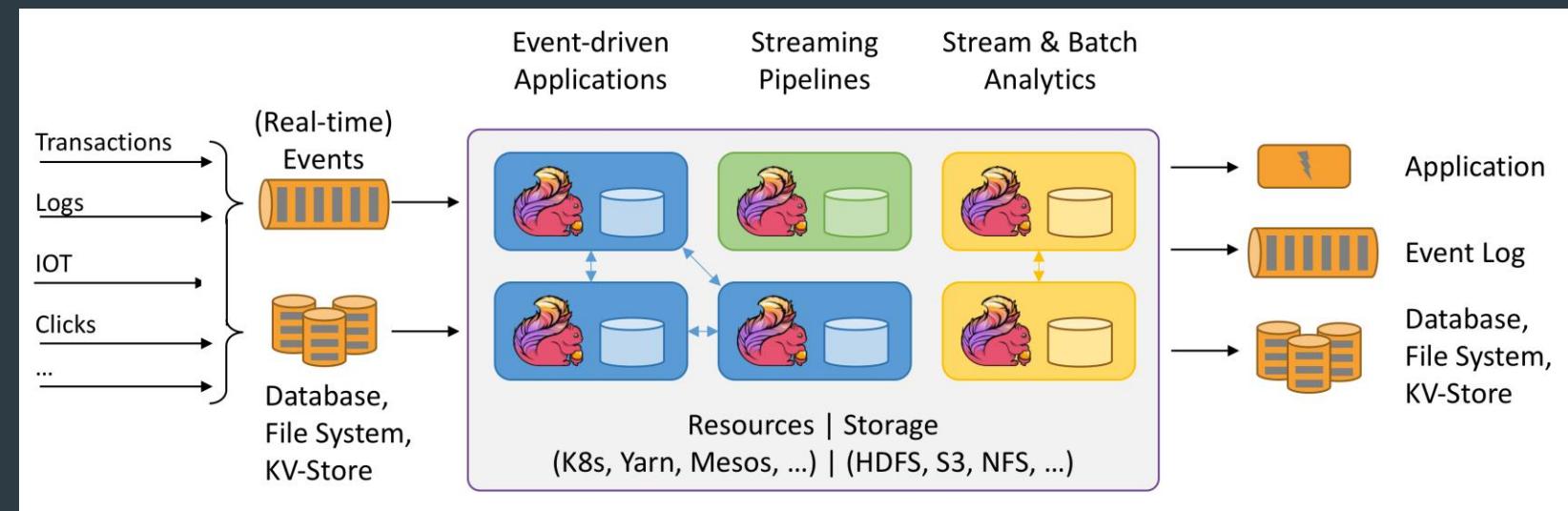
- Apache Flink 项目介绍
- Apache Flink 基本能力介绍
- Apache Flink 适用场景介绍
- Cloudera CSA 模块介绍
- Apache Flink 未来展望

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

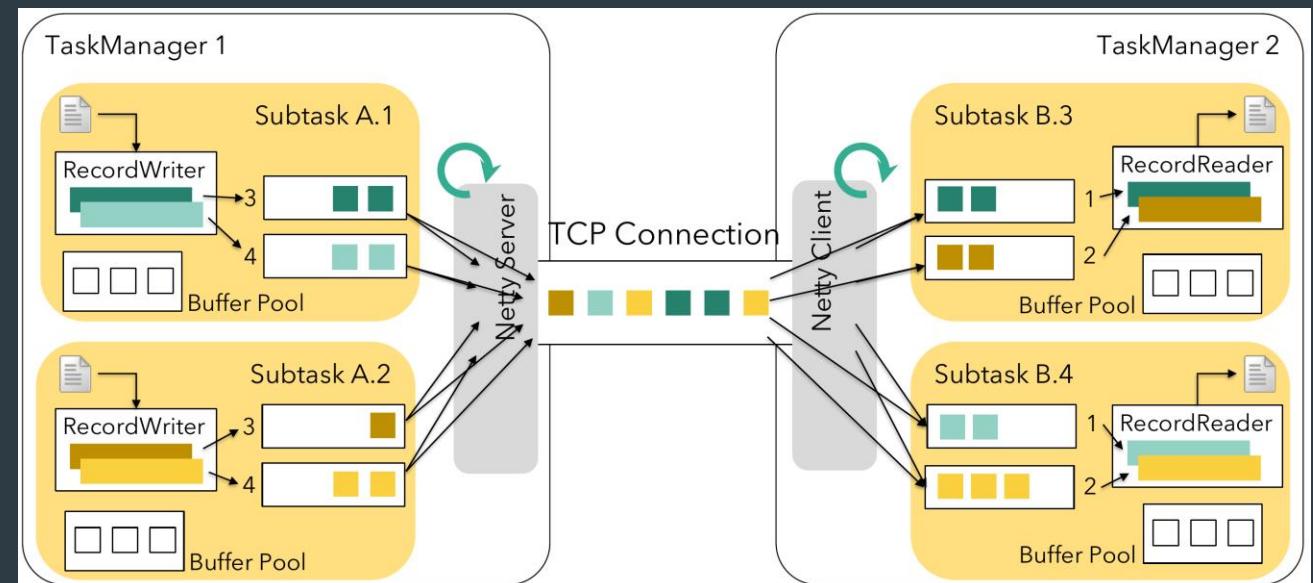
- 低延时
- 高吞吐
- 有状态
- 分布式



Apache Flink

What is Apache Flink

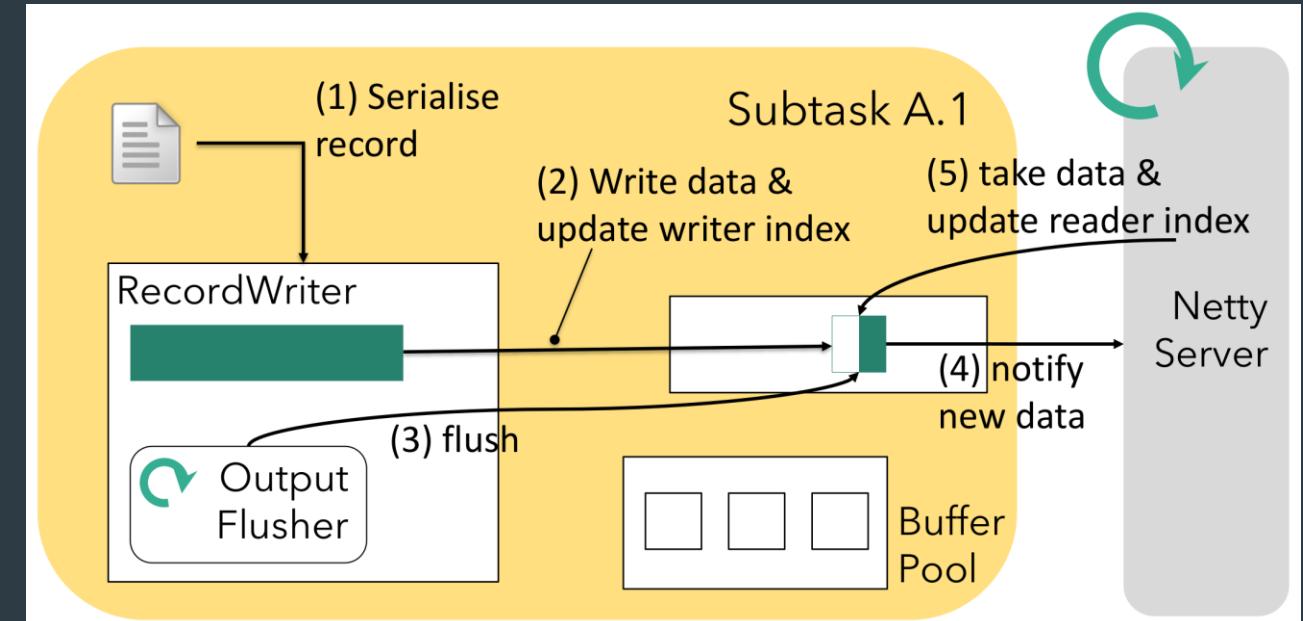
- 高吞吐
 - 每一个Task的subtask在TaskManager之间共享一个TCP链接



Apache Flink

What is Apache Flink

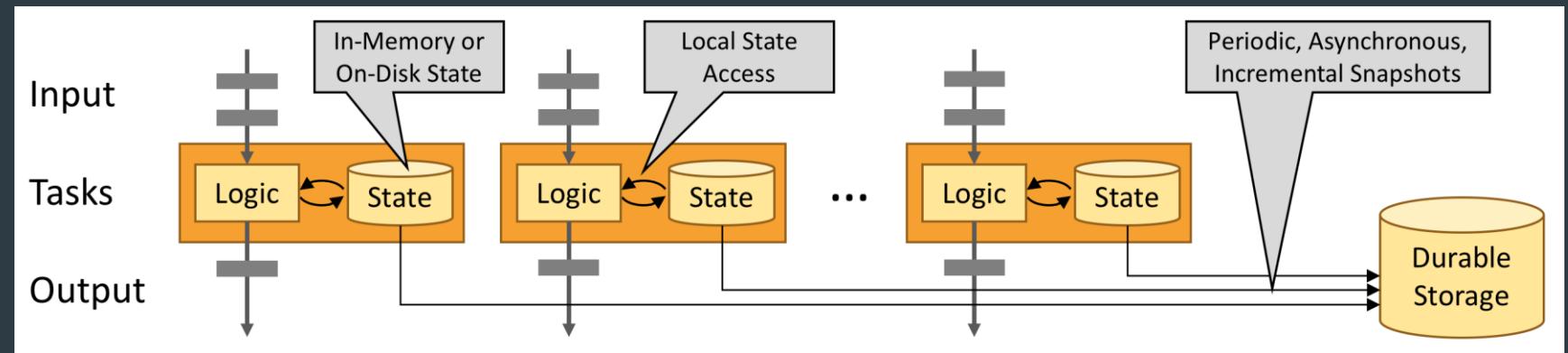
- 低延时
 - 可配置的缓冲区定时推送
 - setBufferTimeout = -1,0 ,x ms



Apache Flink

What is Apache Flink

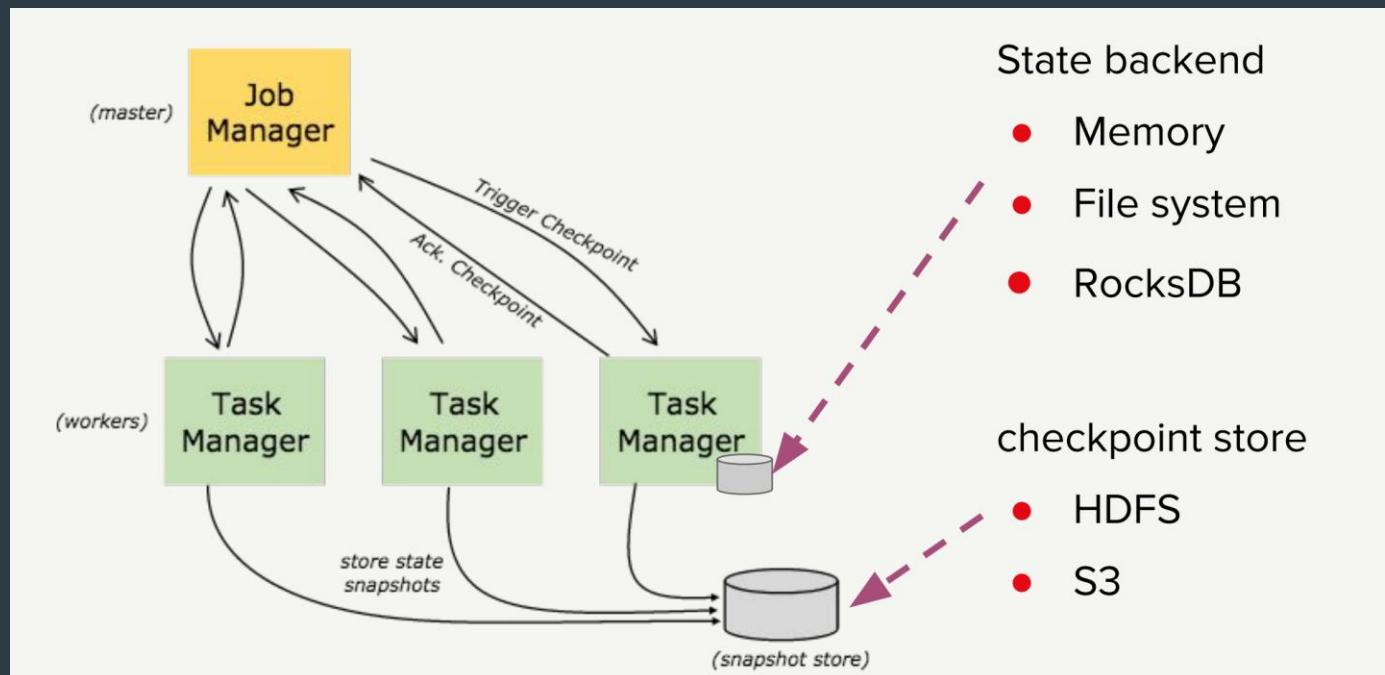
- 有状态



Apache Flink

What is Apache Flink

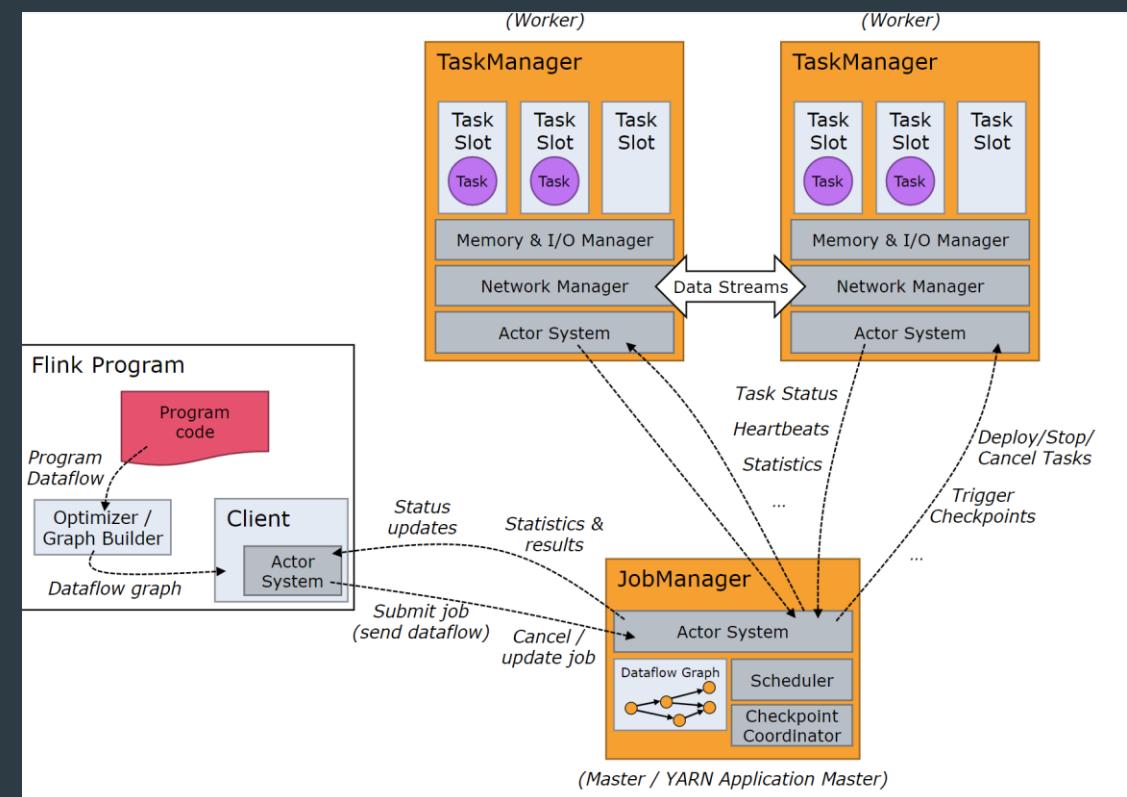
- 有状态
 - 内存
 - 文件
 - Rocks DB



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

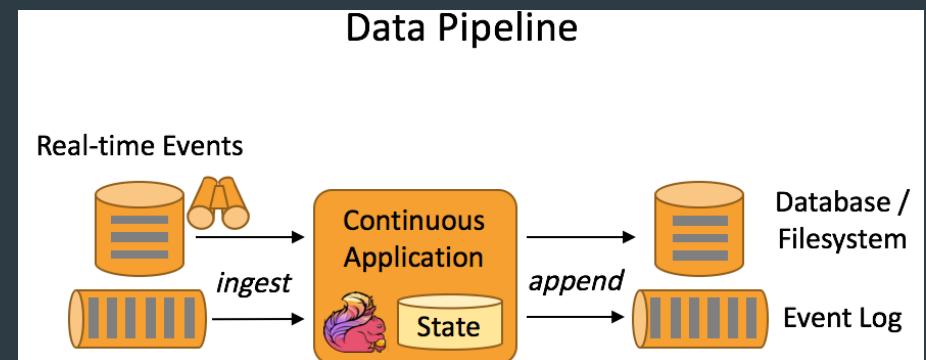
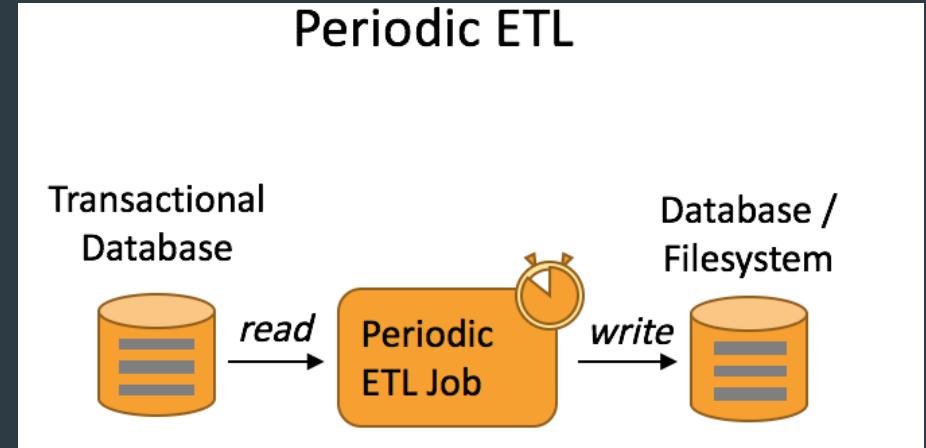
- 适用场景
 - 流式ETL
 - 实时数据分析
 - 事件驱动型的应用



Apache Flink

适用场景 – 流式ETL

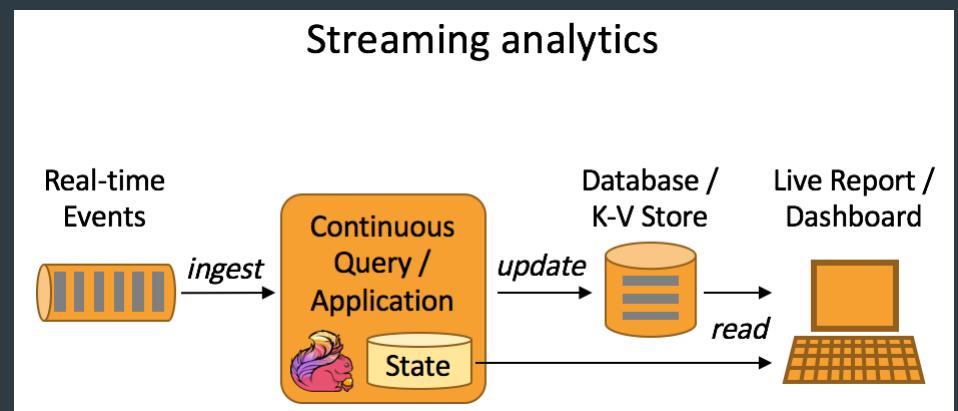
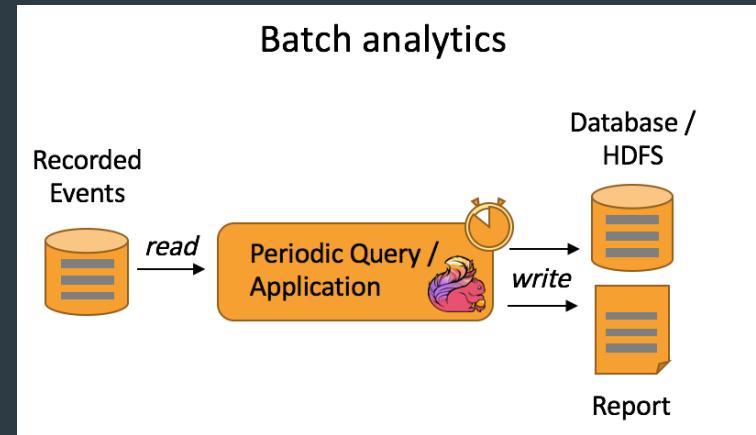
- 传统的ETL
 - 通过周期性的调用ETL脚本完成批处理的作业
- 流式ETL
 - 低延时
 - 持续的数据转换



Apache Flink

适用场景 – 实时数据分析

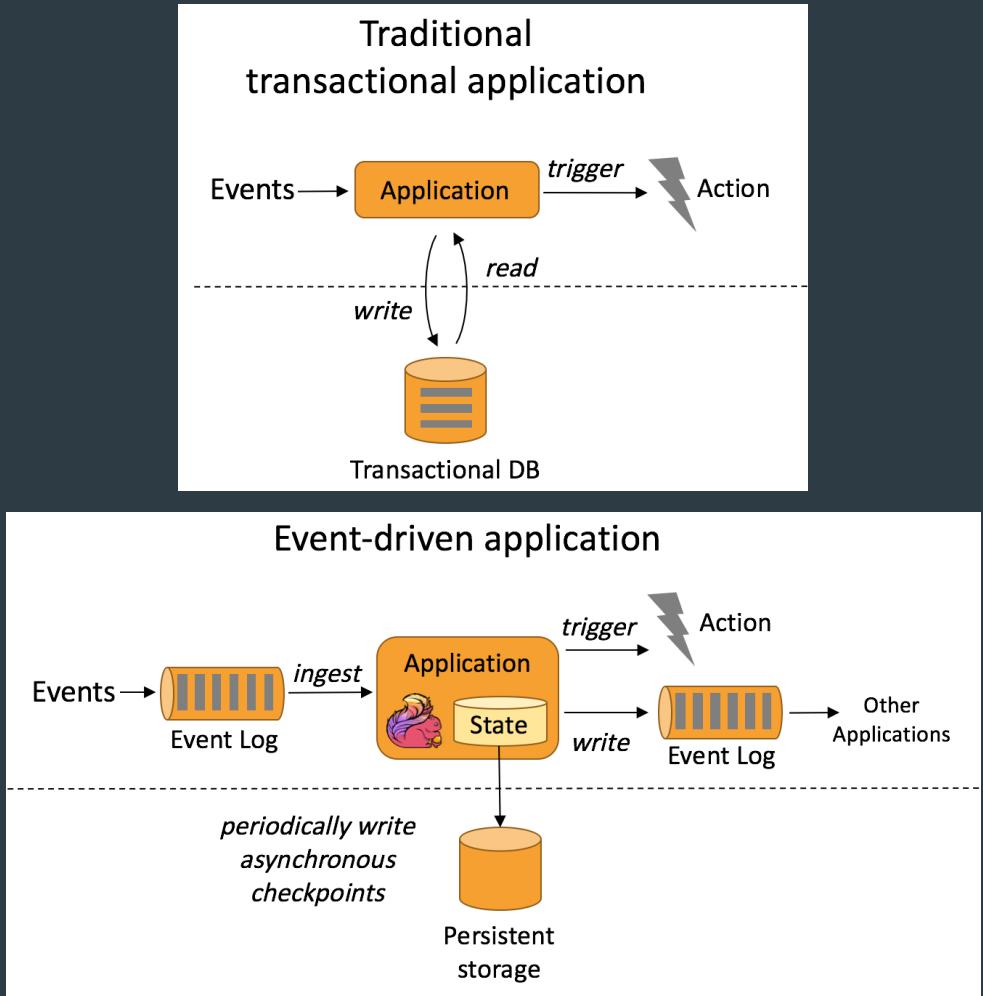
- 批处理性的分析
 - ad-hoc 查询
 - 查询的变化远大于数据的变化
 - 交互性的分析和预分析
- 实时数据分析
 - 数据的变化远大于查询的变化
 - 实时/低延迟的查询结果
 - 不需要Lambda架构的系统设计



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

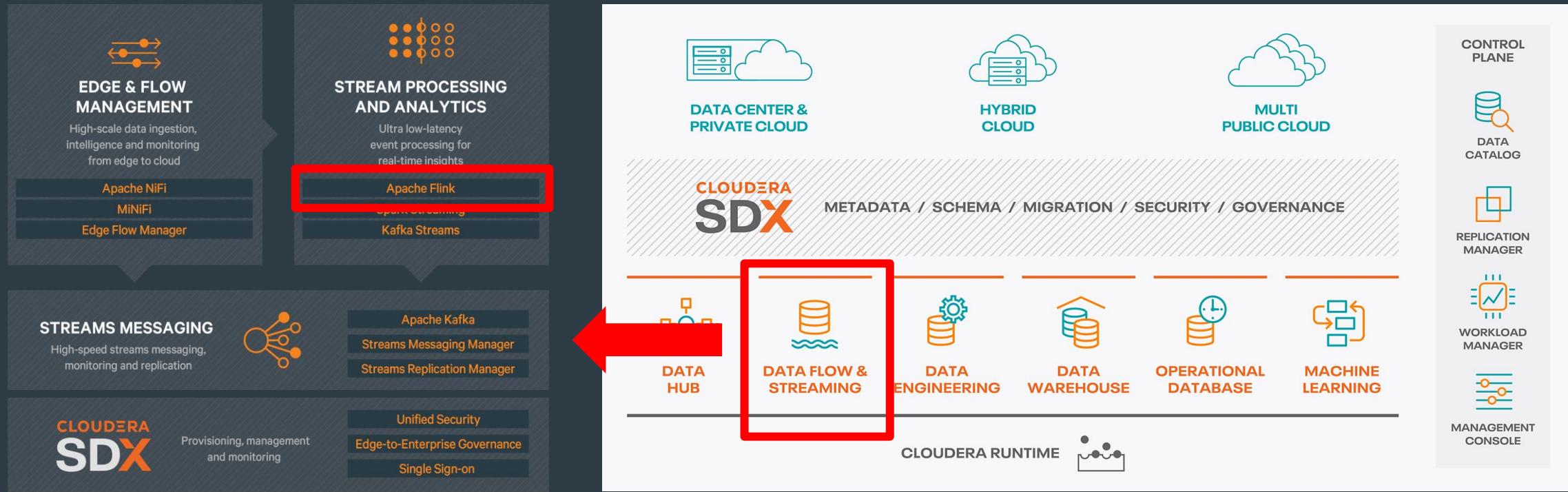
目录

- Cloudera CSA 模块介绍

Apache Flink

Cloudera CSA 模块介绍

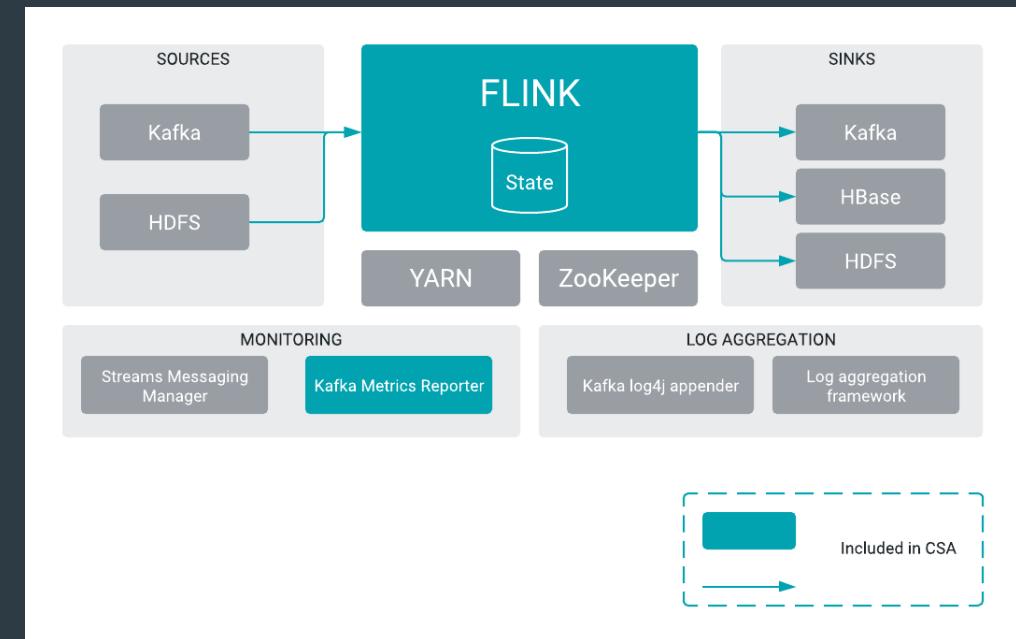
- Cloudera CSA 模块介绍



Apache Flink

Cloudera CSA 模块介绍

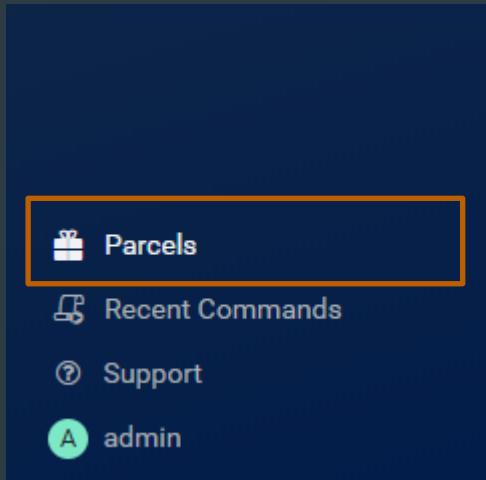
- Cloudera CSA 模块介绍
 - StableAPI: Flink Datastream API
 - Flink Kafka Connector
 - Flink HDFS Connector
 - Flink HBASE Connector
- Flink Kafka Metrics Reports
- Cloudera Schema Registry



Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 模块下载



The screenshot shows the 'Parcels' page in Cloudera Manager. The left sidebar has 'Parcels' selected. The main content area shows a table for 'CDH703' with a 'FLINK' row highlighted by an orange box. Below the table, a list of service types is shown, with 'Flink' also highlighted by an orange box.

Parcel Name	Version
ACCUMULO	1.9.2-1.ACUMULO6.1.0.p.908695
Cloudera Runtime	1.7.2-5.5.0.ACUMULO5.5.0.p.08
FLINK	7.0.3-1.cdh7.0.3.p.0.1635019
FLINK	1.9.1-csa1.1.0.0-cdh7.0.3.0-79-1751706

Select the type of service you want to add.

Service Type	Description
Atlas	Apache Atlas provides a set of metadata management and governance services that enable you to find, organize, and manage data assets.
Data Analytics Studio	Data Analytics Studio is the one stop shop for Apache Hive warehousing. Query, optimize and administrate your data with this powerful interface.
Flink	Apache Flink is a framework and distributed processing engine for stateful computations over unbounded and bounded data streams.
HBase	Apache HBase is a highly scalable, highly resilient NoSQL OLTP database that enables applications to leverage big data.
HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data.
Hive	Apache Hive is a SQL based data warehouse system. In CDH 6 and earlier, this service includes Hive Metastore and HiveServer2. In Cloudera Runtime, part of the Hive on Tez service.

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

The screenshot shows the Cloudera Manager interface for the History Server. At the top, there's a green checkmark icon and a small icon of a person with a crown, followed by the text "History Server". Below this is a "Actions" dropdown menu. The main area displays a table with two rows: one for the History Server (status green, role type History Server, state Started, hostname cdpdc703.cloudera.com) and one for the Gateway (status grey, role type Gateway, state N/A, hostname cdpdc703.cloudera.com). At the bottom, there's a navigation bar with tabs: Status (which is blue and underlined), Configuration, Processes, Commands, Charts Library, Audits, Log Files, Stacks Logs, History Server Web UI (which is highlighted with an orange box), and Quick Links.

Apache Flink

下载/安装/使用 Flink

- Cloudera CSA 使用

The screenshot shows the Apache Flink Dashboard interface. On the left, there is a dark sidebar with the following navigation options:

- Overview (selected)
- Jobs
 - Running Jobs
 - Completed Jobs
- Task Managers
- Job Manager
- Submit New Job

The main content area has two sections: "Running Job List" and "Completed Job List".

Running Job List: This section is currently empty, displaying a "No Data" message.

Job Name	Start Time	Duration	End Time	Tasks	Status
----------	------------	----------	----------	-------	--------

Completed Job List: This section shows one completed job: "Streaming WordCount".

Job Name	Start Time	Duration	End Time	Tasks	Status
Streaming WordCount	2020-01-16 01:37:11	7s	2020-01-16 01:37:18	3	FINISHED

At the top right of the dashboard, there is a status bar with the following information: Version: 1.9.1-csa1.1.0.0 Commit: 3ef6b4a @ 04.01.2020 @ 19:34:15 UTC | Message: 0 | [Logout](#).

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>${flink.version}</version>
</dependency>
<dependency>
<groupId>org.apache.flink</groupId>
<artifactId>flink-avro-cloudera-registry</artifactId>
<version>1.9.1-csa1.1.0.</version>
</dependency>
<dependency>
<groupId>org.apache.avro</groupId>
<artifactId>avro-compiler</artifactId>
<version>${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" );
}
```

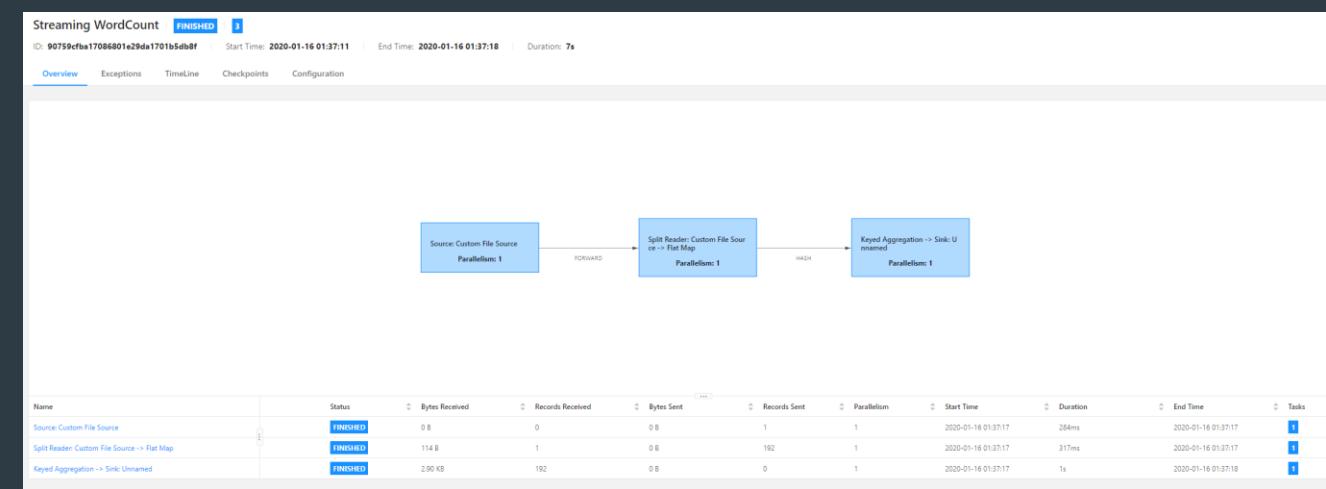
Apache Flink

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

The screenshot shows a list of tasks or components for a streaming job. The tasks listed include:
- (Kafka,1)
- (Deserialization,1)
- (Schema,1)
- (Cloudera,1)
- (Registry,1)
- (Kafka,1)
- (Deserialization,1)
- (Schema,1)
- (Cloudera,1)
- (Registry,1)
- (Kafka,1)
- (Serialization,1)
- (Schema,1)
- (aaa,1)
- (cbd,1)
- (aaa,1)
- (Cloudera,1)
- (kafka,1)
- (Kafka,1)

02:35:39,034 INFO org.apache.kafka.clients.consumer.internals.AbstractCoordinator



Apache Flink

Apache Flink 未来展望

- Apache Flink 未来展望

PAAS on CDP Public Cloud

- 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

Ranger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

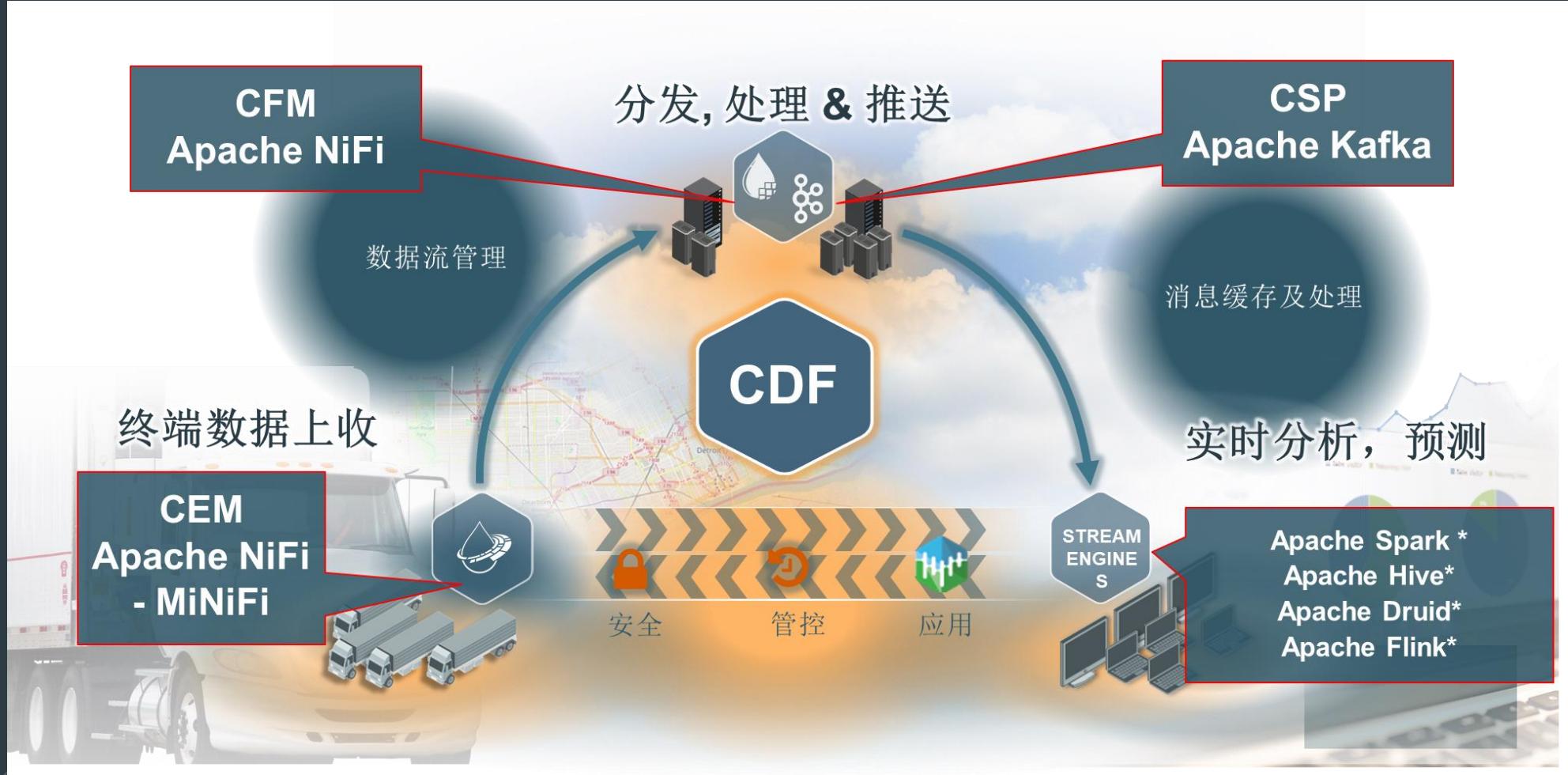
Atlas集成

- 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

Flink部署管理

- 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)

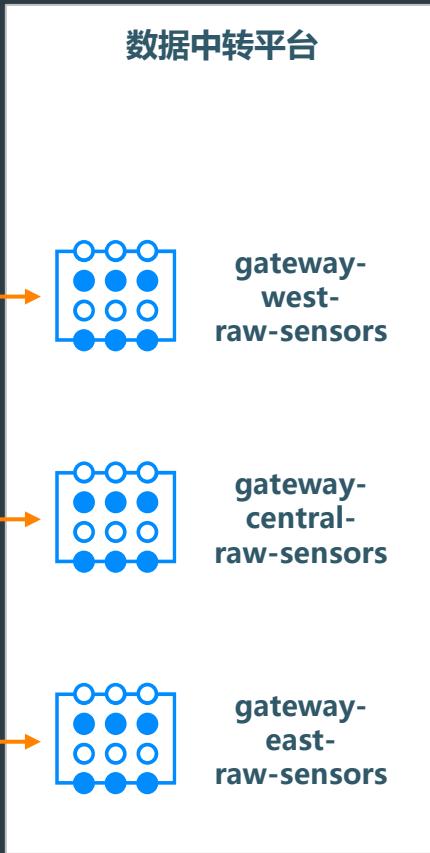


CDF IN Action 整体架构

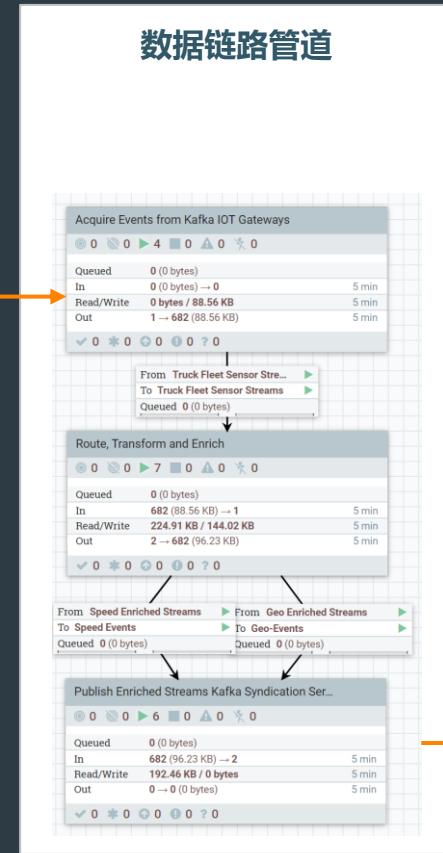
MiNiFi



Apache Kafka



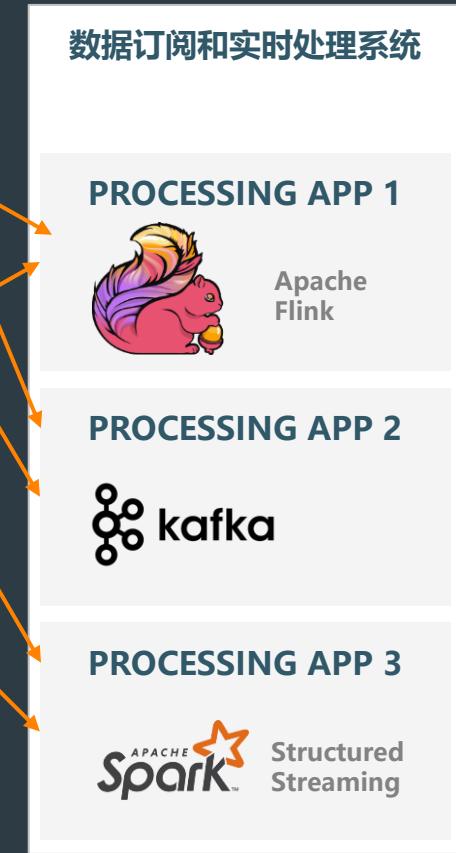
Apache NiFi



Apache Kafka



Apache Flink



CLOUDERA DATA PLATFORM

全球第一个企业数据云

HYBRID &
MULTI-CLOUD

SECURITY &
GOVERNANCE

ANALYTICS
EDGE TO AI

OPEN
DISTRIBUTION



DATA CENTER &
PRIVATE CLOUD



HYBRID
CLOUD



MULTI
PUBLIC CLOUD

CLOUDERA
SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE



DATA
HUB



DATA FLOW &
STREAMING



DATA
ENGINEERING



DATA
WAREHOUSE



OPERATIONAL
DATABASE



MACHINE
LEARNING

CLOUDERA RUNTIME



CONTROL
PLANE



DATA
CATALOG



REPLICATION
MANAGER



WORKLOAD
MANAGER



MANAGEMENT
CONSOLE

Apache Flink

Q & A

