



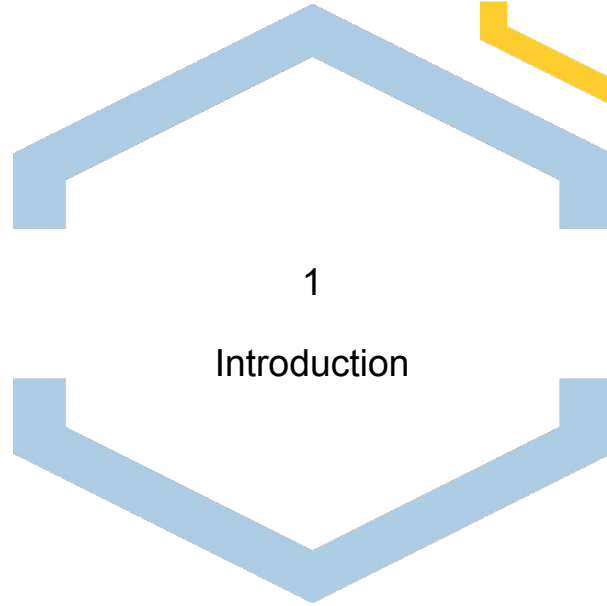
Raúl Reguillo Carmona

Apache NiFi

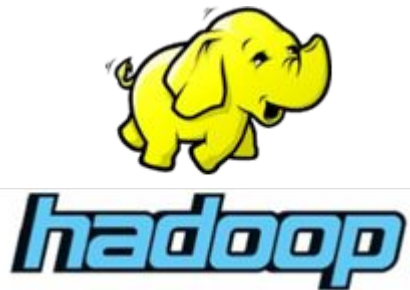
A General
Purpose Tool for
Big Data

Table of Contents

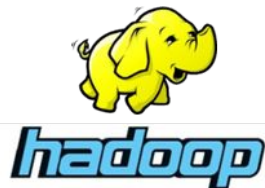
1. Introduction
2. Install and setup
3. Study cases
 - a. Tweet feeding with NiFi
 - b. Tweet transformation with NiFi + Microservices
 - c. Tweet visualization with NiFi and ELK
 - d. Use case: sentiment analysis with NiFi + IBM Watson Tone Analyzer
 - e. Other use cases with NiFi and Big Data tools
4. Conclusions



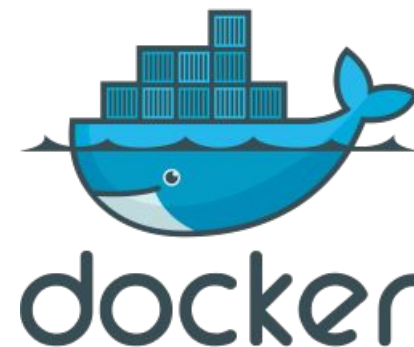
- What NiFi is...
 - General purpose Big Data Tool
 - Basic operations (extract, transform and load)
 - Based on processors and flowfiles
 - Drag and drop
 - **Integration**

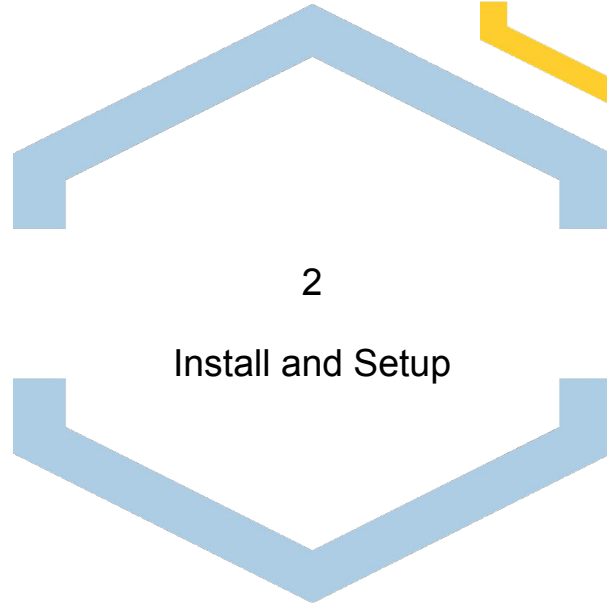


- Features
 - Clear interface
 - Changes at runtime
 - API
 - Kerberos/SSL/SSH/HTTPS/Multitenant friendly
 - Lots of processors
 - but also extensible
 - Expression language
 - dynamic configurations



- About this PoC
 - NiFi properties
 - Some use cases
 - Integration between services
 - NiFi as *man in the middle*
 - Use along with
 - Docker
 - ELK stack
 - Custom microservices



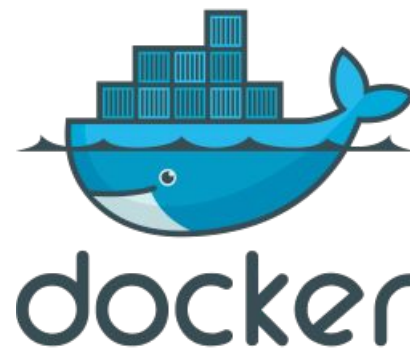


2

Install and Setup



- PoC
 - Dockerized NiFi
 - Dockerized ELK Stack
 - Dockerized Microservice
- Easy to use and test!

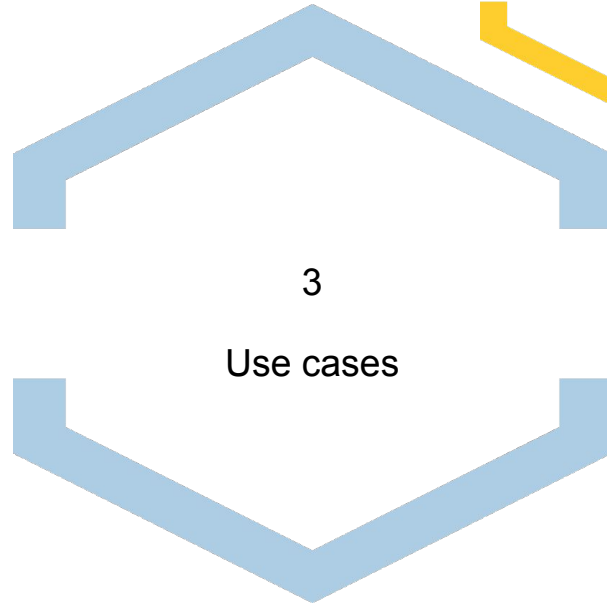


2. Install and Setup

```
git clone <repo>
cd <repo>/submodules/ms.semtweet/ && mvn clean package
cd <repo>/submodules/ && docker-compose up
```

1. NiFi
 - <http://localhost:9090/nifi/>
2. semtweet (microservice)
3. ElasticSearch
4. Kibana
 - <http://localhost:5601/>

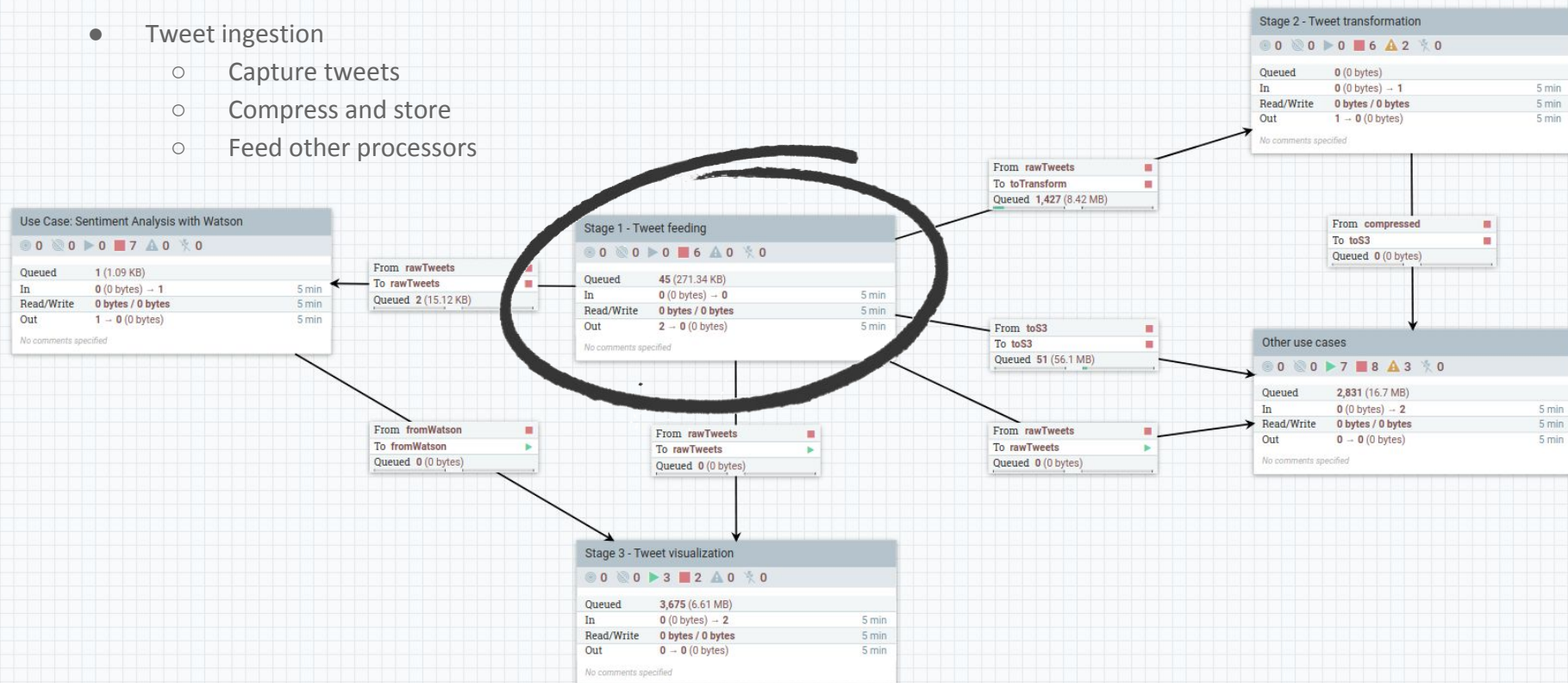




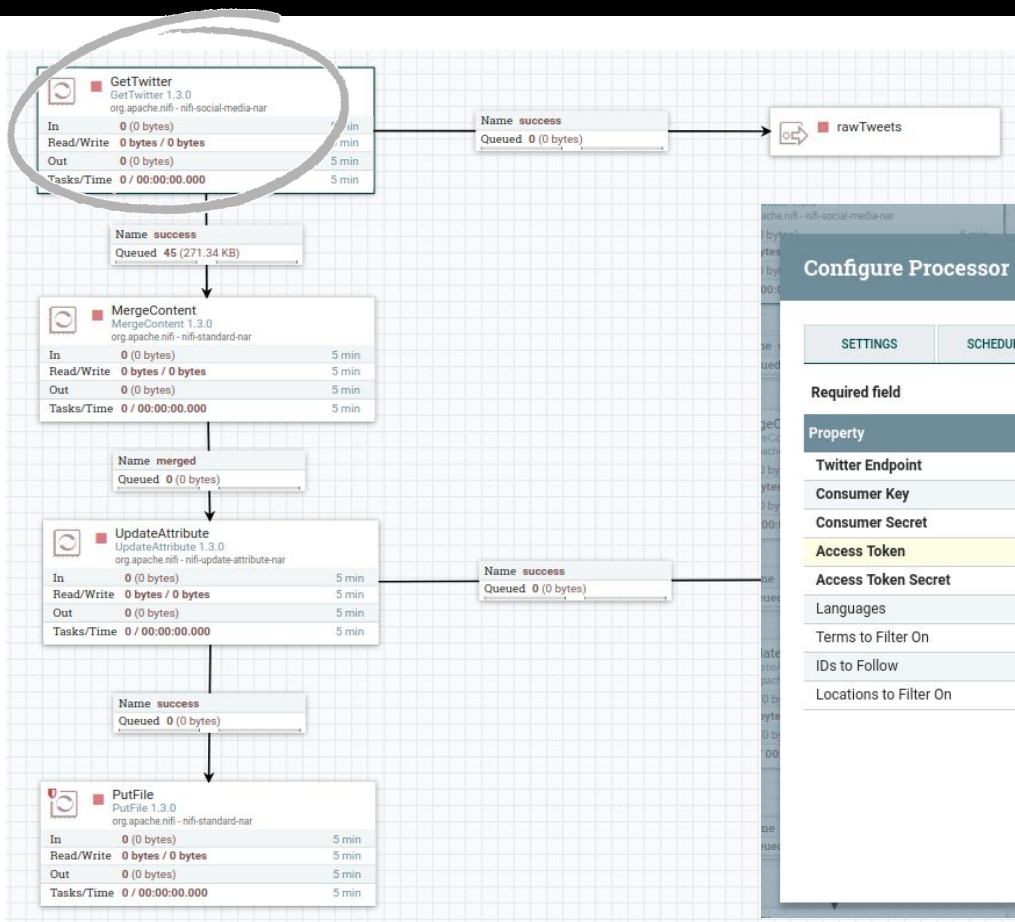
3. General topology



- Tweet ingestion
 - Capture tweets
 - Compress and store
 - Feed other processors



3. General topology - a) Tweet feeding



Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field

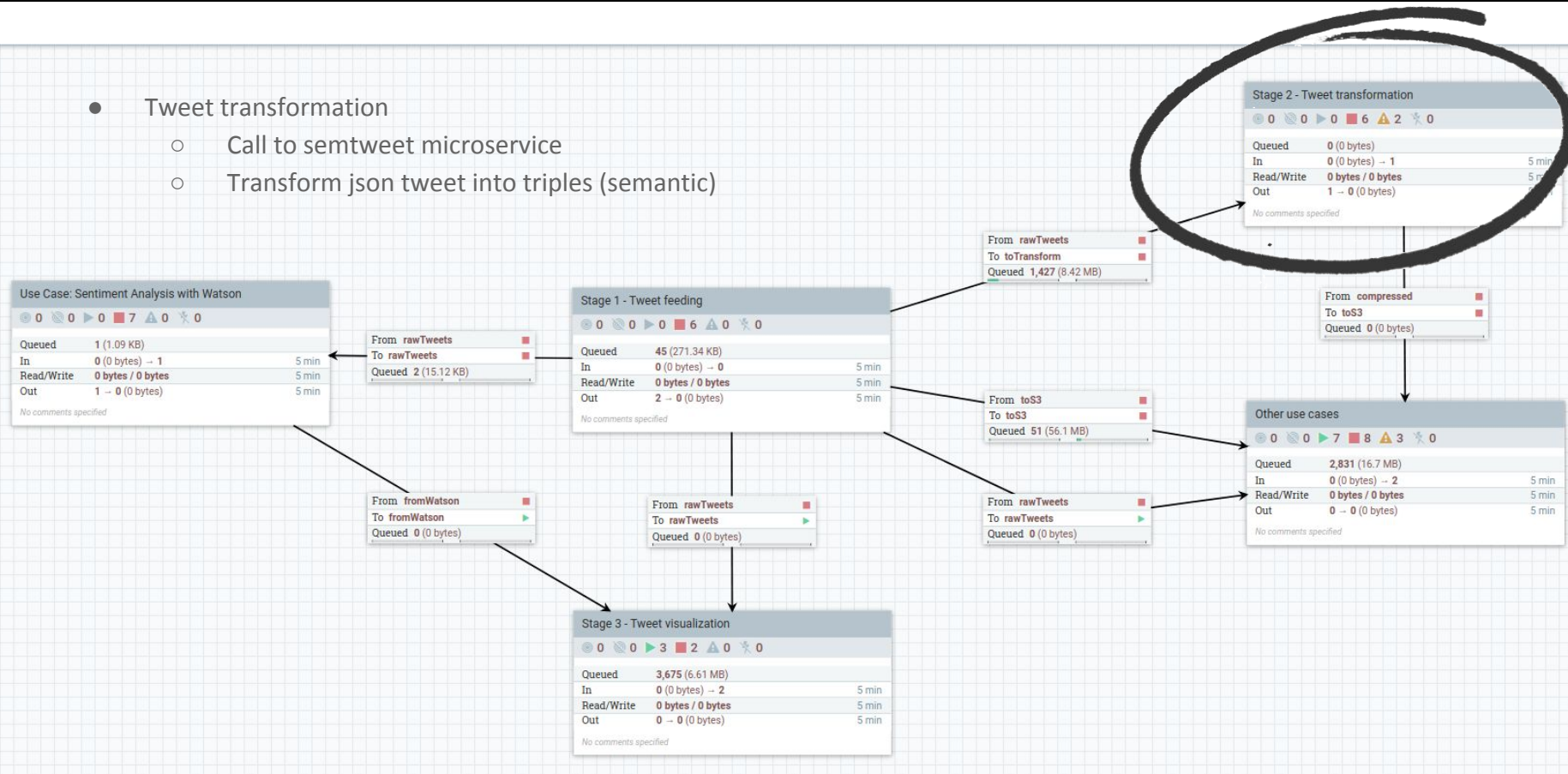
Property	Value
Twitter Endpoint	Filter Endpoint
Consumer Key	GD7Fr0uuxxDhiiO1U33zn0XHUD
Consumer Secret	Sensitive value set
Access Token	15067432-suRPk1Vc9gk0yBe8mgGCQDisH2Bz13CB2TdG...
Access Token Secret	Sensitive value set
Languages	en
Terms to Filter On	fintech, blockchain, disruptive
IDs to Follow	No value set
Locations to Filter On	No value set

CANCEL APPLY

3. General topology

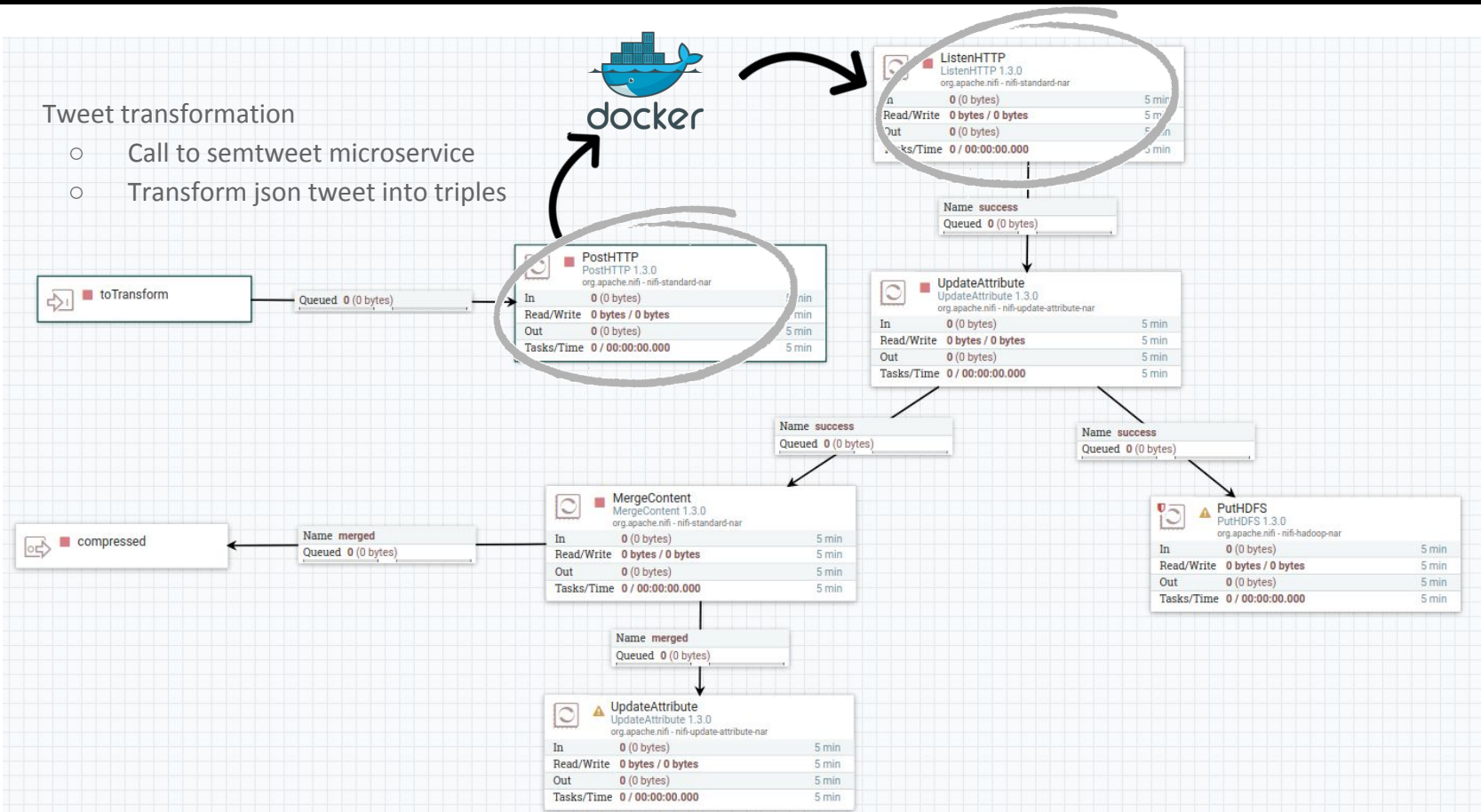


- Tweet transformation
 - Call to senttweet microservice
 - Transform json tweet into triples (semantic)



3. General topology - b) Tweet transformation

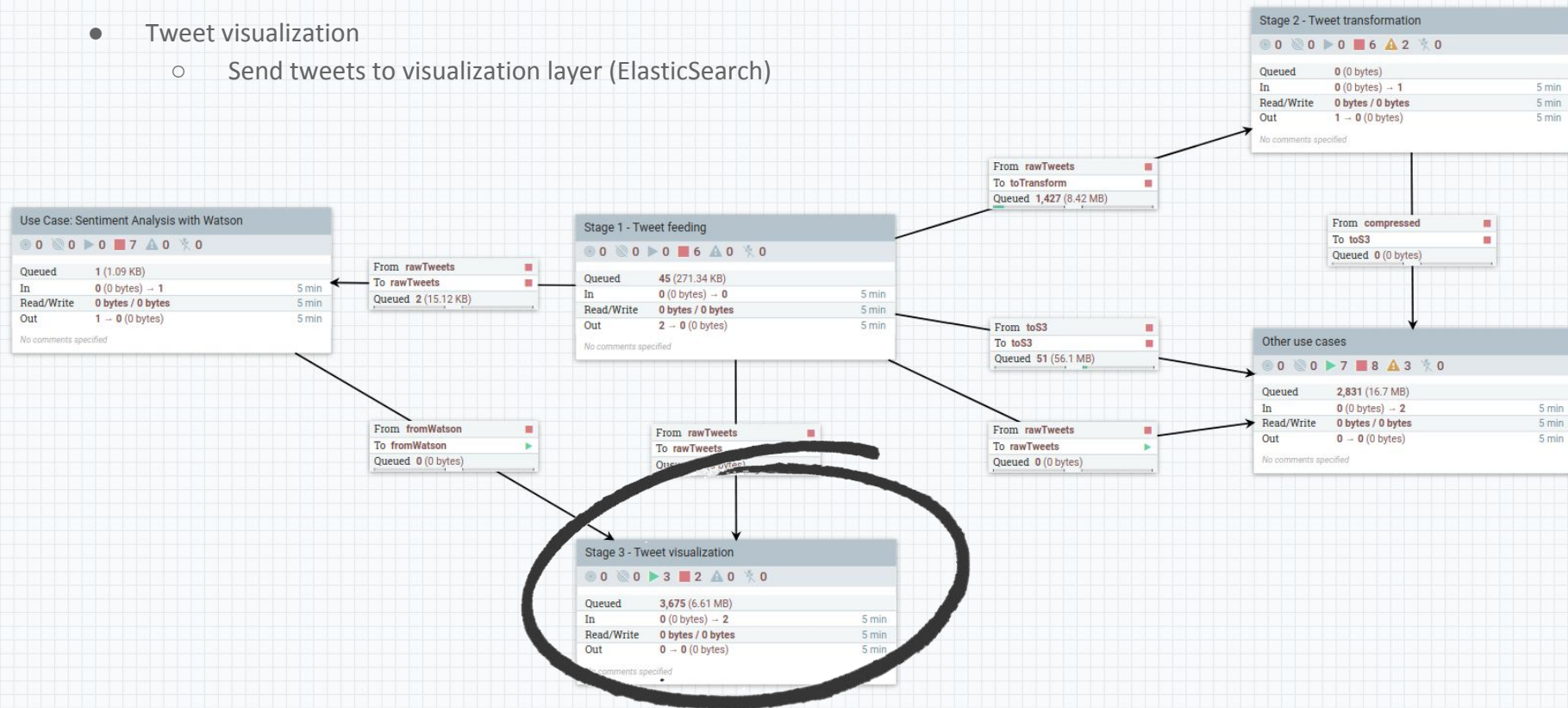
- Tweet transformation
 - Call to semtweet microservice
 - Transform json tweet into triples



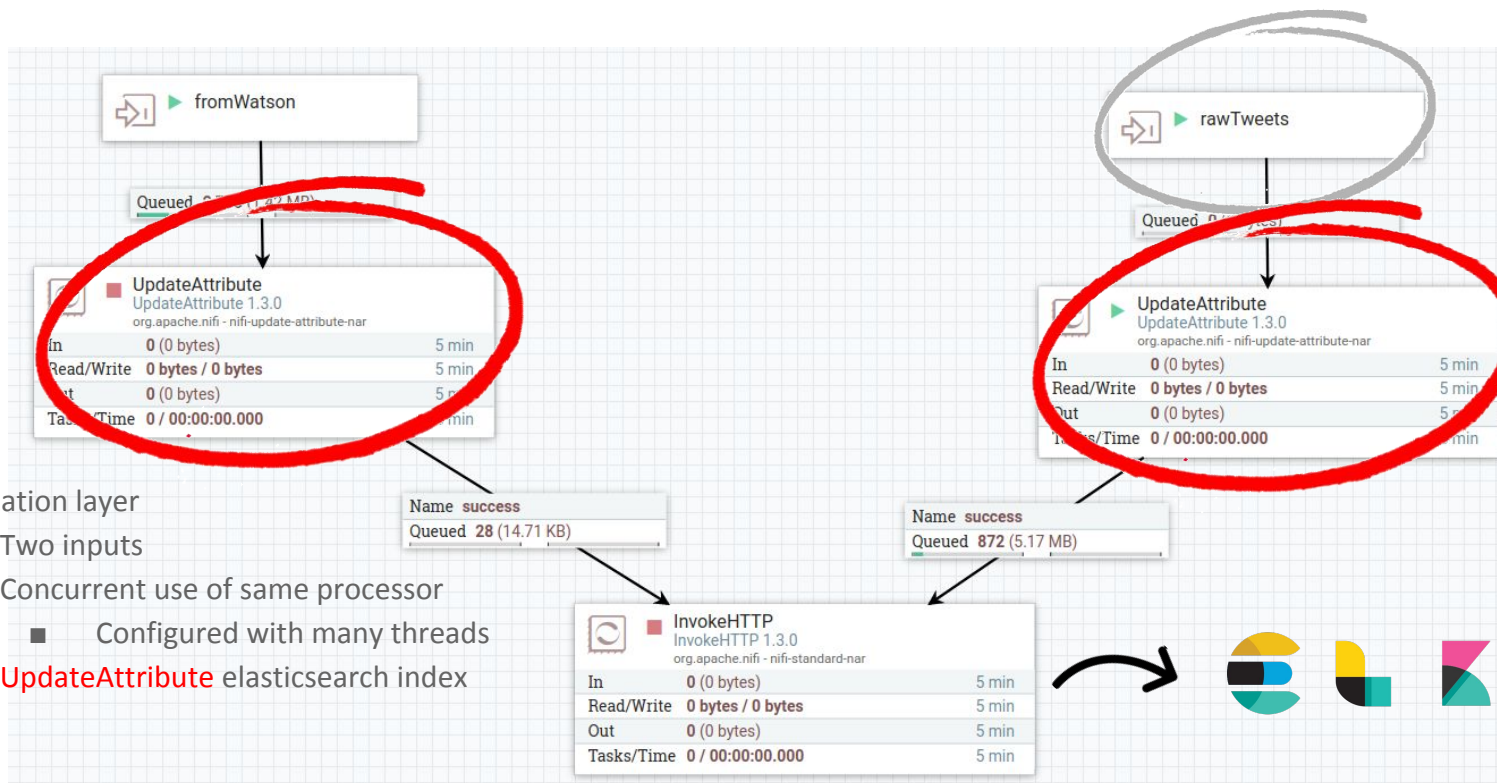
3. General topology



- Tweet visualization
 - Send tweets to visualization layer (ElasticSearch)



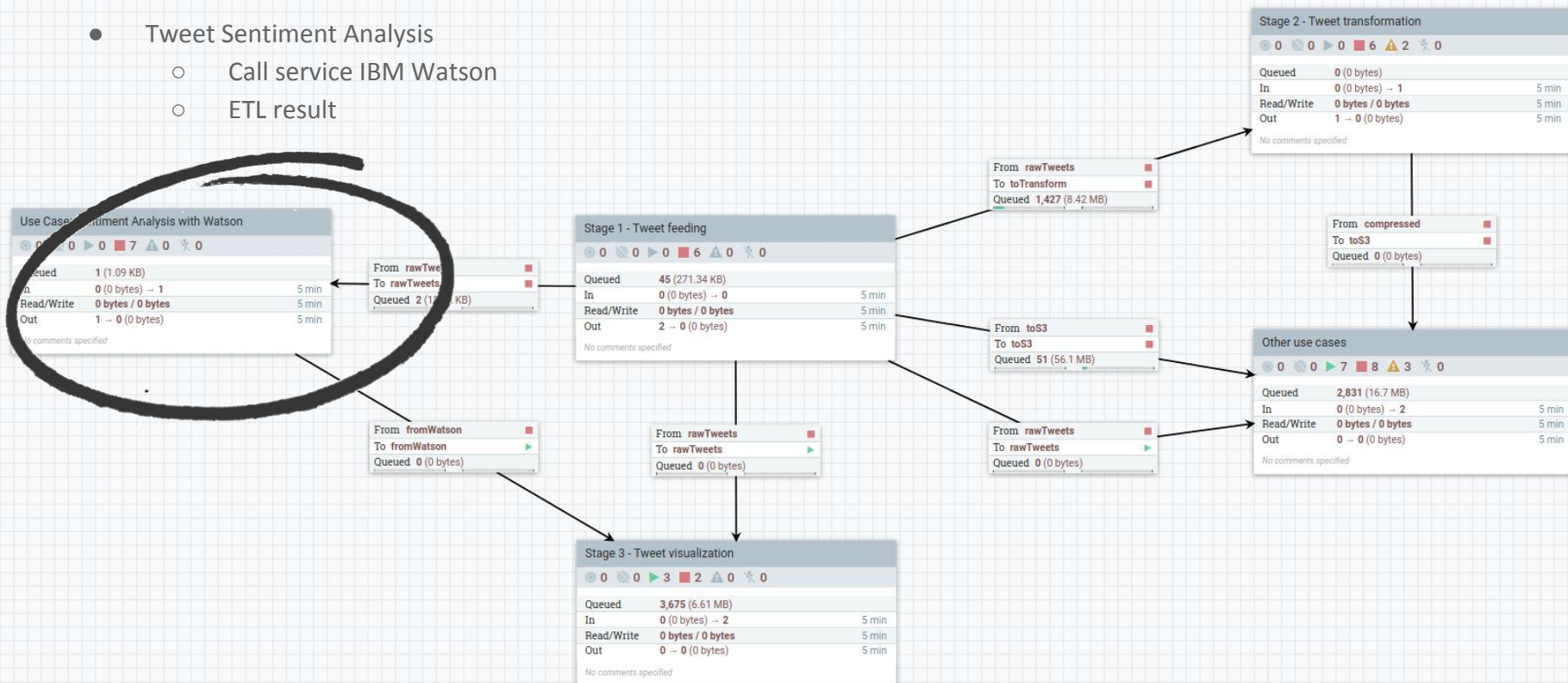
3. General topology - c) Tweet visualization



- Visualization layer
 - Two inputs
 - Concurrent use of same processor
 - Configured with many threads
 - UpdateAttribute elasticsearch index

3. General topology

- Tweet Sentiment Analysis
 - Call service IBM Watson
 - ETL result



3. General topology - d) Watson Sentiment Analysis



- Invoke HTTP (if you can *curl* it, you can use NiFi)
- Extract info from returned json

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

Property	Value
Destination	flowfile-attribute
Return Type	auto-detect
Path Not Found Behavior	ignore
Null Value Representation	empty string
agreeableness	\$.document_tone.tone_categories[2].tones[3].score
analytical	\$.document_tone.tone_categories[1].tones[0].score
anger	\$.document_tone.tone_categories[0].tones[0].score
confident	\$.document_tone.tone_categories[1].tones[1].score
conscientiousness	\$.document_tone.tone_categories[2].tones[1].score
disgust	\$.document_tone.tone_categories[0].tones[1].score
emotional_range	\$.document_tone.tone_categories[2].tones[4].score
extraversion	\$.document_tone.tone_categories[2].tones[2].score
fear	\$.document_tone.tone_categories[0].tones[2].score
inv	\$.document_tone.tone_categories[1].tones[3].score

CANCEL APPLY

Configure Processor

SETTINGS SCHEDULING PROPERTIES COMMENTS

Required field +

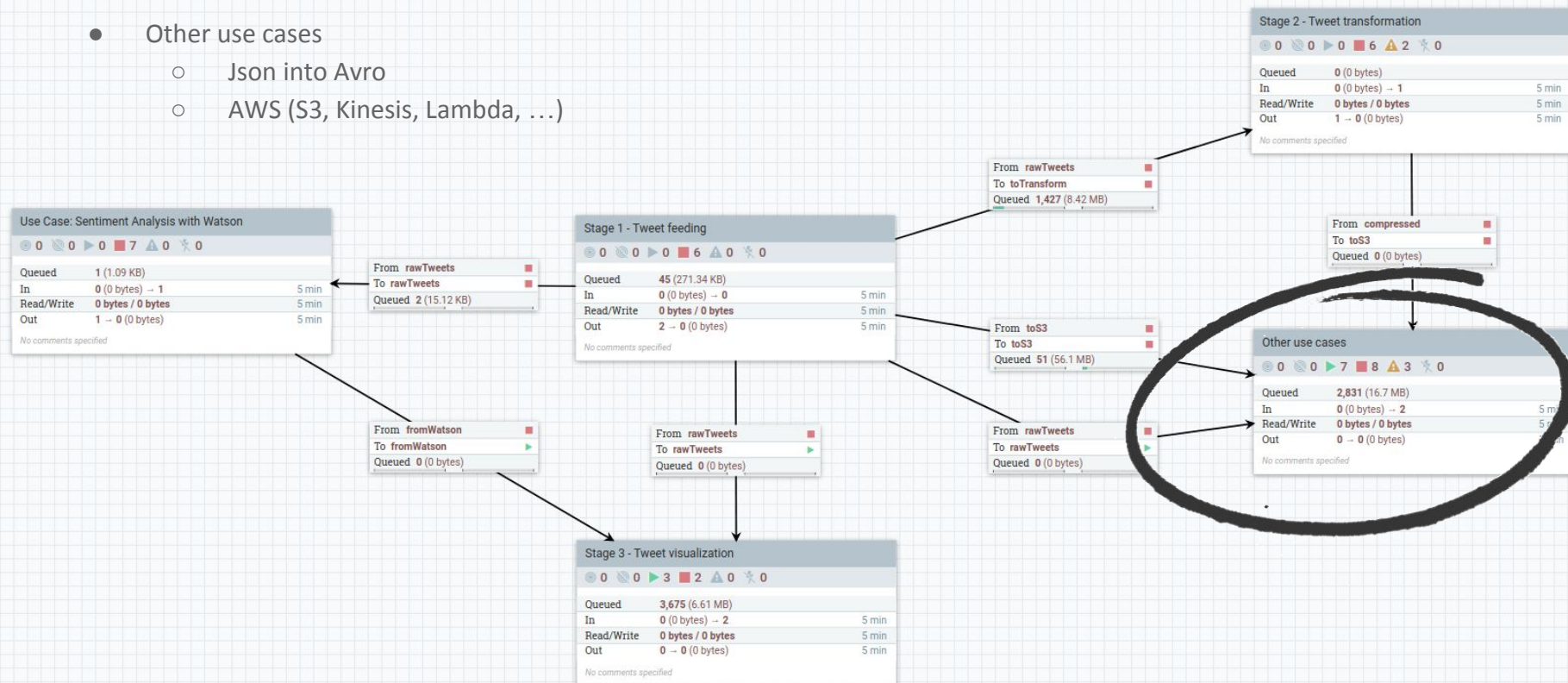
Property	Value
HTTP Method	GET
Remote URL	https://gateway.watsonplatform.net/tone-analyzer/api/v3...
SSL Context Service	No value set
Connection Timeout	5 secs
Read Timeout	15 secs
Include Date Header	True
Follow Redirects	True
Attributes to Send	No value set
Basic Authentication Username	d111dd6d-263e-4b0f-acba-f8ea7abc269f
Basic Authentication Password	Sensitive value set
Proxy Host	No value set
Proxy Port	No value set
Proxy Username	No value set
Proxy Password	No value set

CANCEL APPLY

3. General topology



- Other use cases
 - Json into Avro
 - AWS (S3, Kinesis, Lambda, ...)

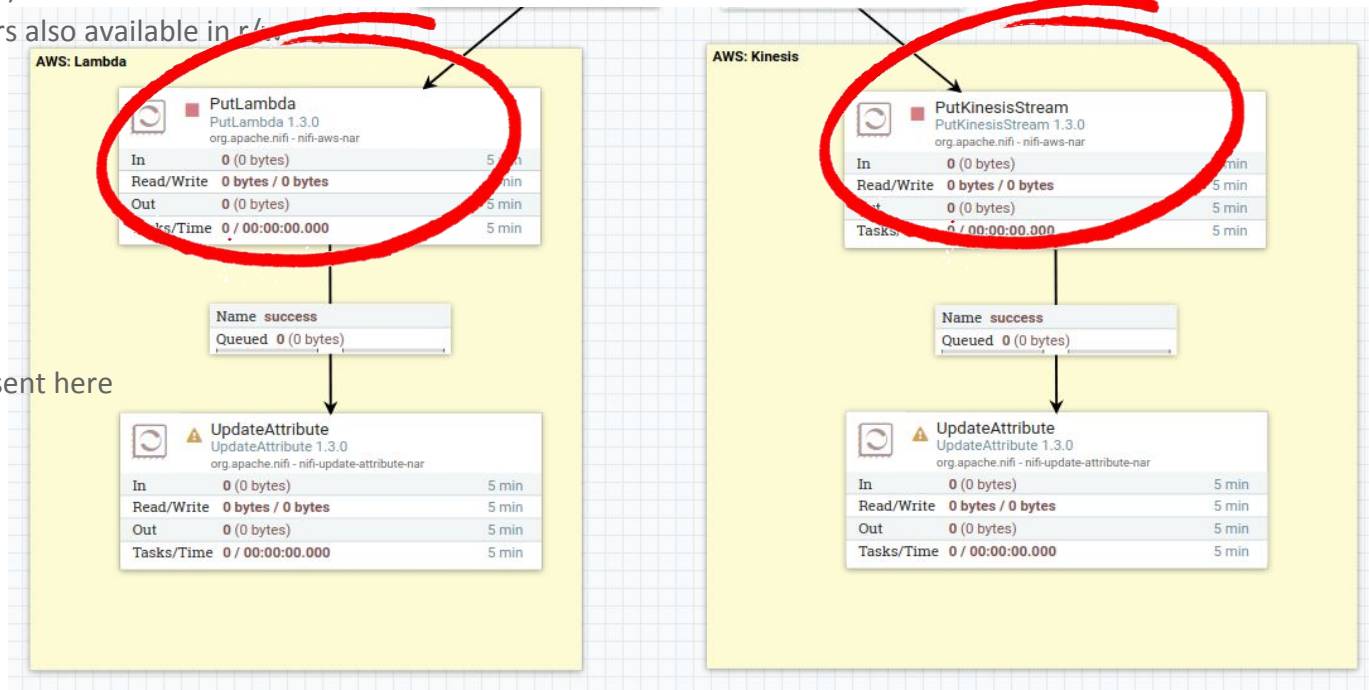


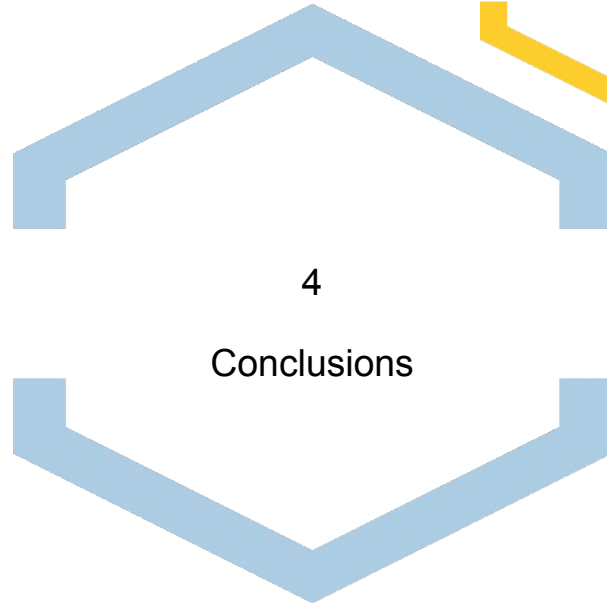
3. General topology - e) Other use cases with Big Data tools



- Used with
 - AWS Lambda: counting number of items in the json
 - AWS Kinesis (just insert)
 - Kafka processors also available in r/...
 - AWS S3
 - List
 - Fetch
 - Put
 - HDFS friendly
 - Avro/Orc/Parquet/...

- Other tested features not present here
 - Flume integration
 - MongoDB
 - Cassandra
 - HBase
 - SFTP
 - ...



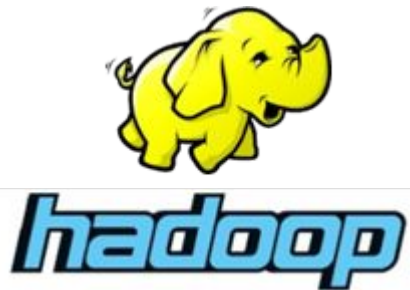


4

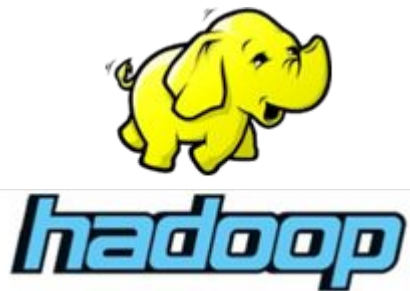
Conclusions

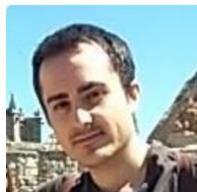


- What NiFi is...
 - General purpose Big Data Tool
 - **Integration**
 - **Versatile**
 - **Easy to use**



- **Beware!**
 - Generates so many dependencies
 - Tends to bring everything
 - **Addictive**





Raúl Reguillo Carmona

Software Engineer. Jack of all trades.

raul.reguillo@beeva.com | [@Einath_Kelaven](https://twitter.com/Einath_Kelaven)



hablemos@beeva.com

www.beeva.com