Best practices and lessons learnt from Running NiFi at Renault

Kamelia Benchekroun- Big Data Architect
Renault

Abdelkrim Hadjidj - Solution Engineer
Hortonworks

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

Agenda

- The NiFi journey at at Renault
- Best practices for running NiFi in production
- Lessons learnt at Renault
- Questions & answers

GROUPE RENAULT

HORTONWORKS

The NiFi journey at Renault

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

About Us

The Datalake Squad:

- A passionate Team who work really hard to deliver solutions and services and empower Data Initiatives at Renault

Datalake Platform Metrics:

450 Connected Users

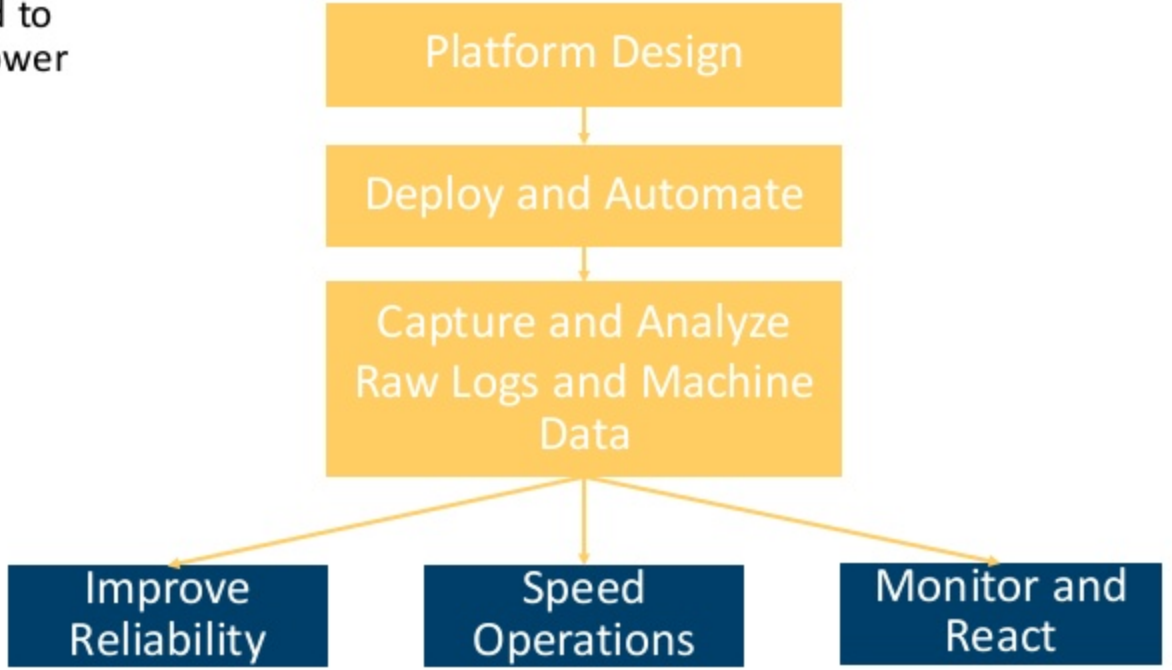
30 Data Initiatives and use cases

8000 Daily Queries

3500 Service Requests

300TB Efficient Storage

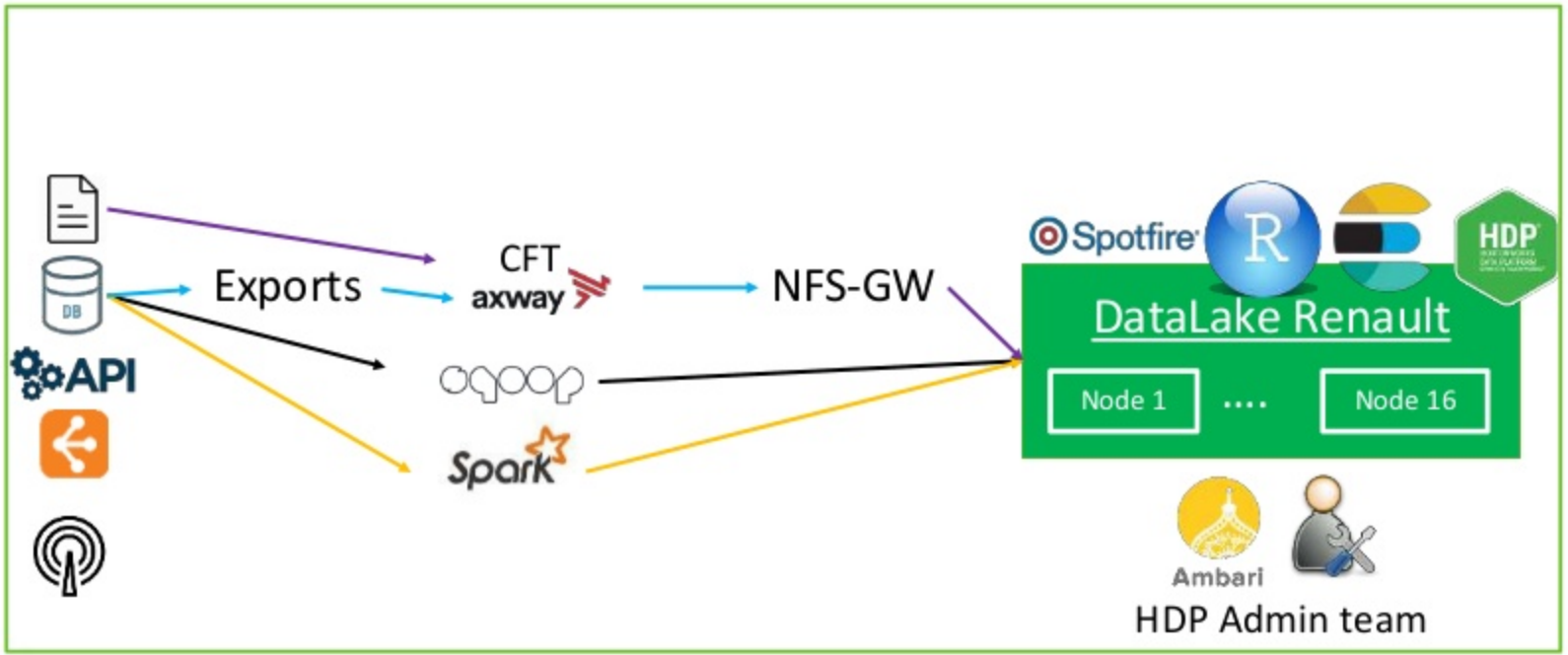
Datalake Squad Activities



GROUPE RENAULT

HORTONWORKS

Data Lake at Renault – the beginning of the story



June 2016

- 10 Users
- Requests by emails, 1 Admin
- Manual provisioning

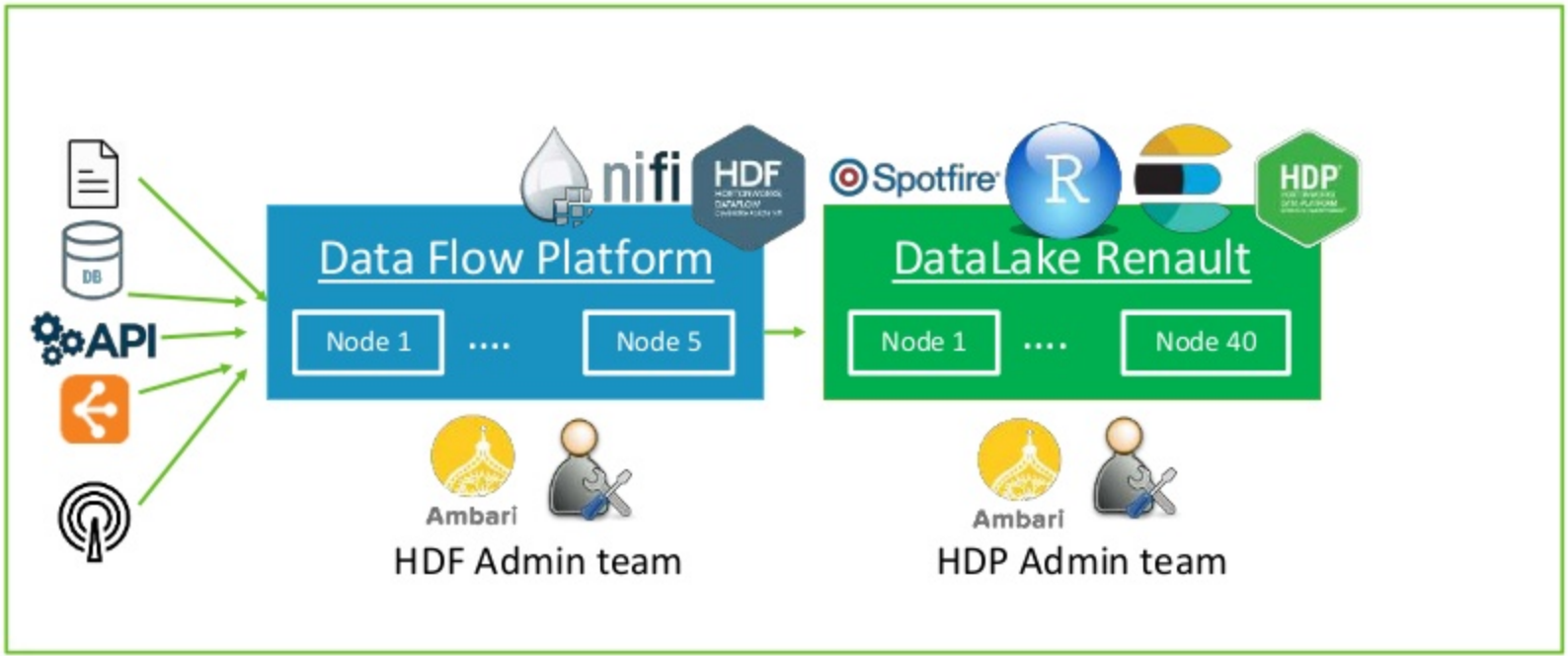
GROUPE RENAULT

HORTONWORKS

Big Data Ingestion

GROUPE RENAULT

HDF came in (Q1 2017)

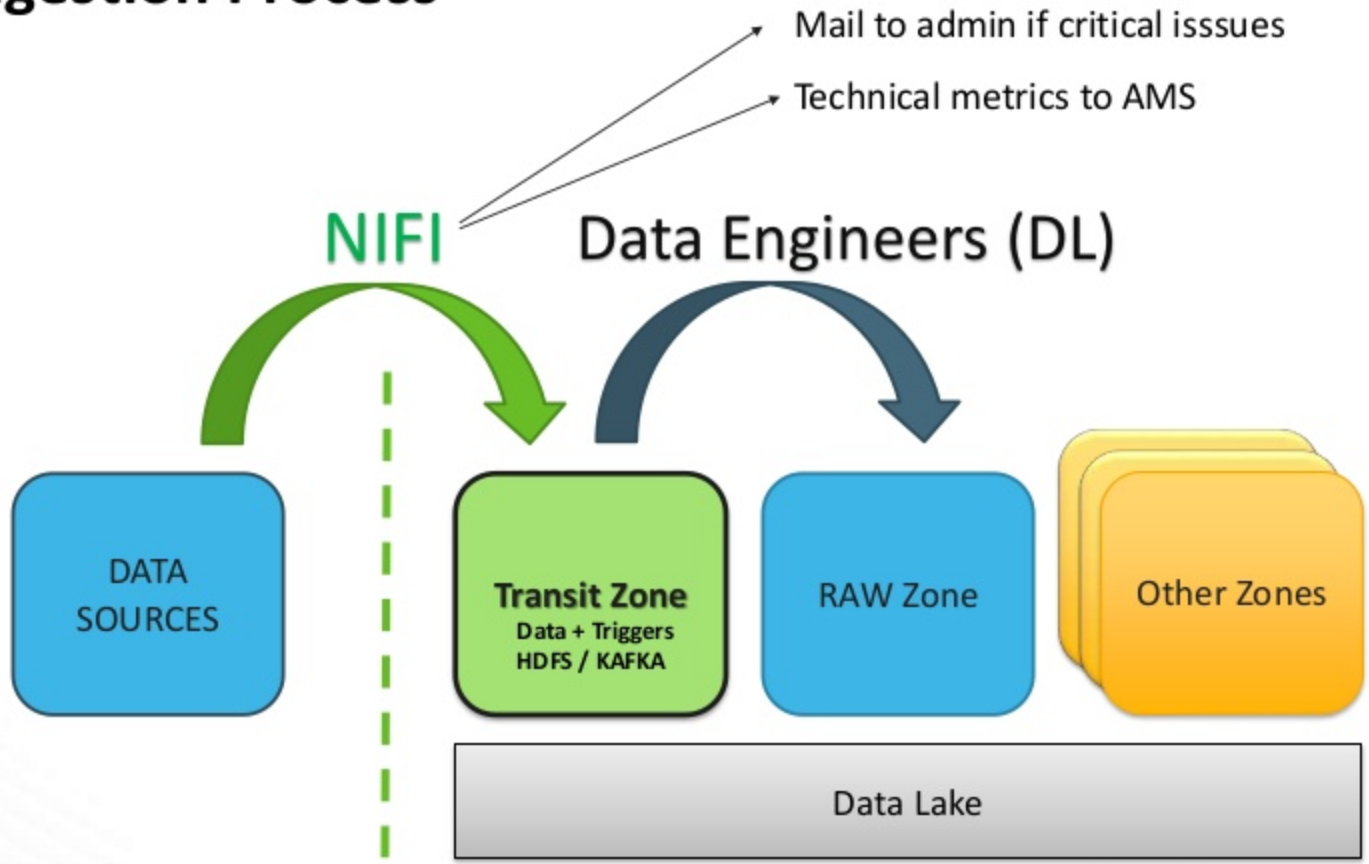


2016
- 10 Users
- Requests by emails, 1 Admin
- Manual provisioning

2017
- 200 users
- Jira portal, Admins/devops
- NiFi, Automation (Jenkins for HDP)

Today
- 300 – 500 active users
- 7000 query per day
- 40 data source in production

Data Ingestion Process

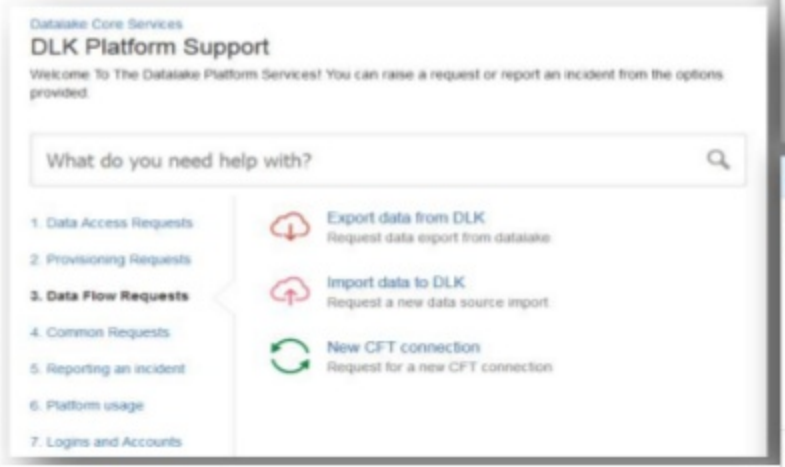


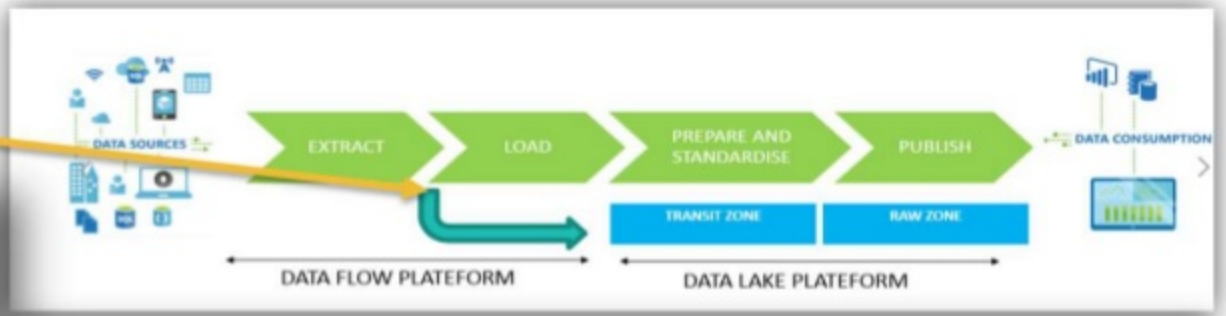
GROUPE RENAULT

HORTONWORKS

Data ingestion workflow

NIFI

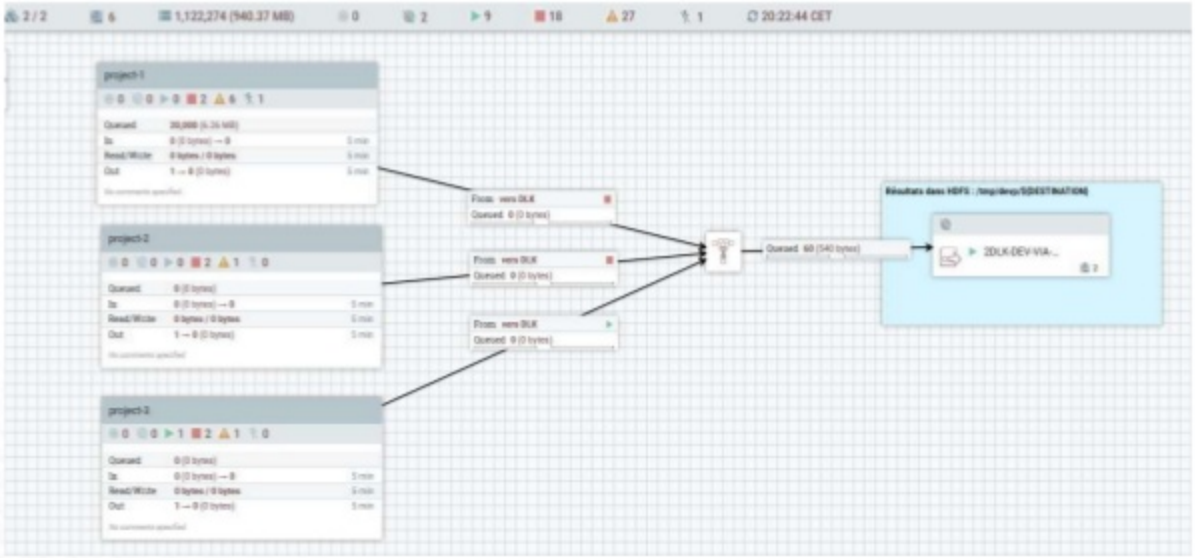




What	How ?	Who ?	When ?
Extract and Load Data Sources	By submitting a request to the Big Data Squad : [redacted] /servicedesk/customer/portal /1/create/18 So that Extraction can be implemented in the Data Flow Platform (DFP)	The Data Flow is scheduled by the DFP Admin at Squad Team	Once the data sources prerequisite are ready. A One shot data offload is launched, Then, when the customer is satisfied with the result. The Frequently import is set.
Publish Data Sets in the Raw Zone	By releasing the Data Pipelines into Production	The Data Engineer is responsible for the Go Live The Datalake Admin is responsible for Auditing Data Format and Usage	Once the Business or Data Analyst confirm Data accuracy
Prepare and Standardize	By developping Data Preparation and Standardization Pipelines. Using HQL queries and Spark applications	The Data Engineer is responsible for developping Data Pipelines	Once Data is loaded by the Data Flow Platform

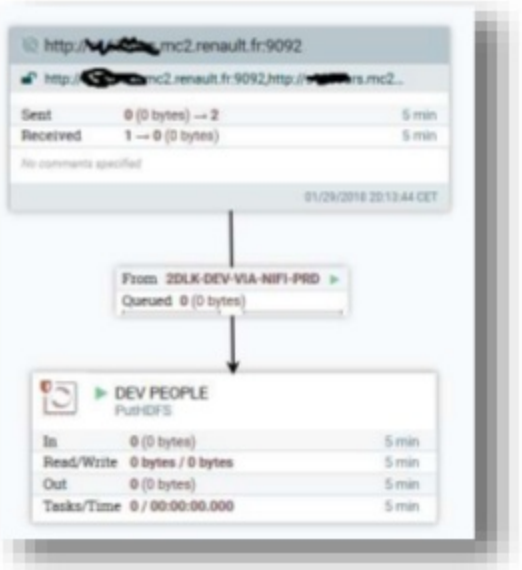
Dev to prod testing

Dev



S2S

Prod



/tmp/devp/${DESTINATION}

GROUPE RENAULT

HORTONWORKS

NiFi Value for Renault

90% of data ingestion since2016

One Platform for all data sources : Files, DBs, API, Brokers, etc. **Offload CFT**

100 active data flows, + 2000 processors in production,.

Accelerate time to insights, use case development and improve monitoring and governance

Also used **export data** from Data Lake to other systems/sites/Cloud

Enables **new use cases** connected plants, package tracking, IoT

GROUPE RENAULT

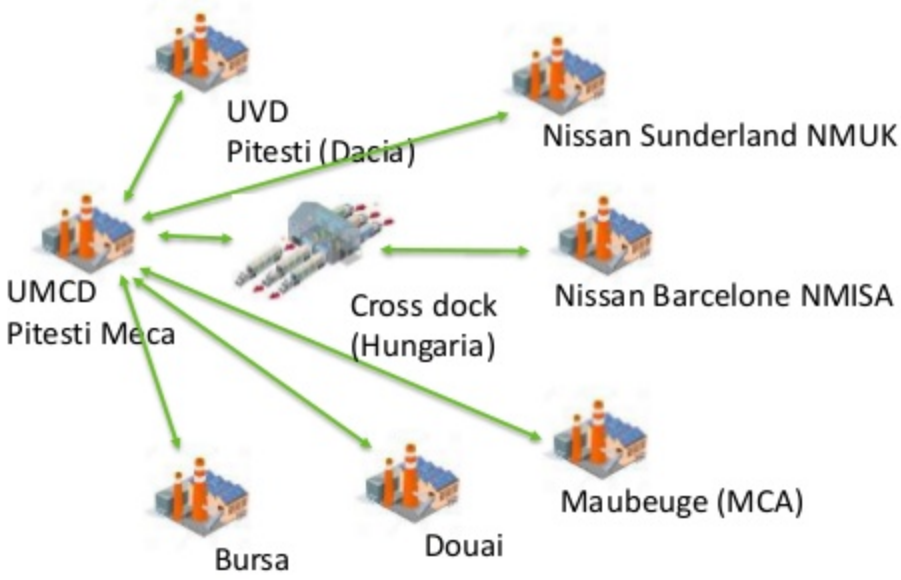
HORTONWORKS

Packaging Traceability

GROUPE RENAULT

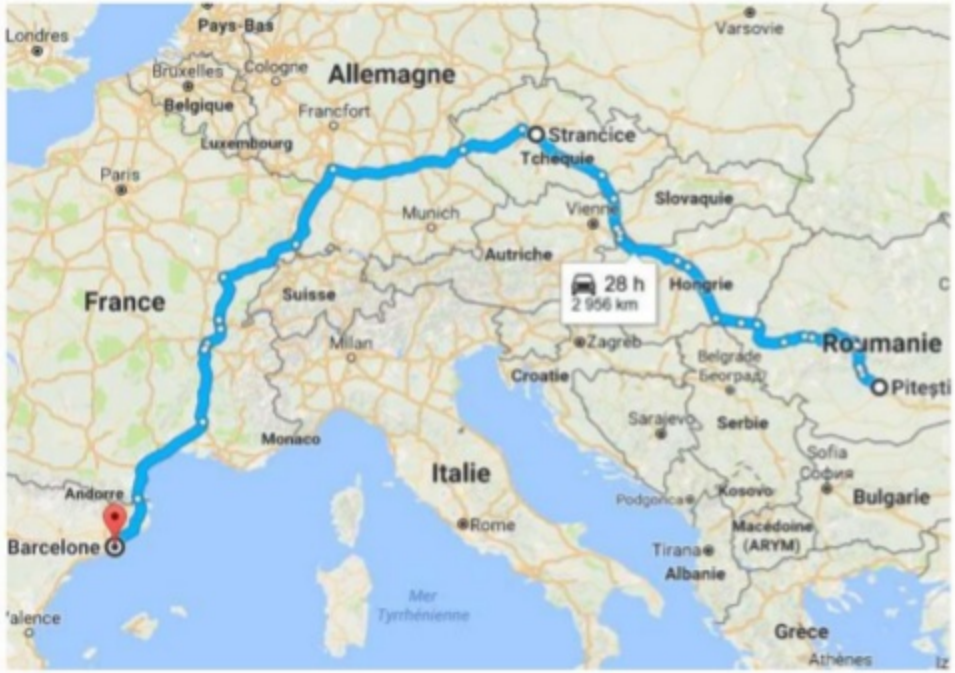
Packaging Traceability

- POC: 2500 Packages running in the Loop
  - 1 package lost = 800 € (= IPhone)
- If the solution is generalized : 600k package
- 2016 Status
  - 400 K€ packaging re-investment due to packaging losses
- 2017 Status
  - 100 K€ cardboard in January
- Test Expectations:
  - Reduce cardboard costs and packaging losses
  - Test LoRa technology in an industrial context and Renault activities
  - Validate operational added Value of LoRa technology



Example flow

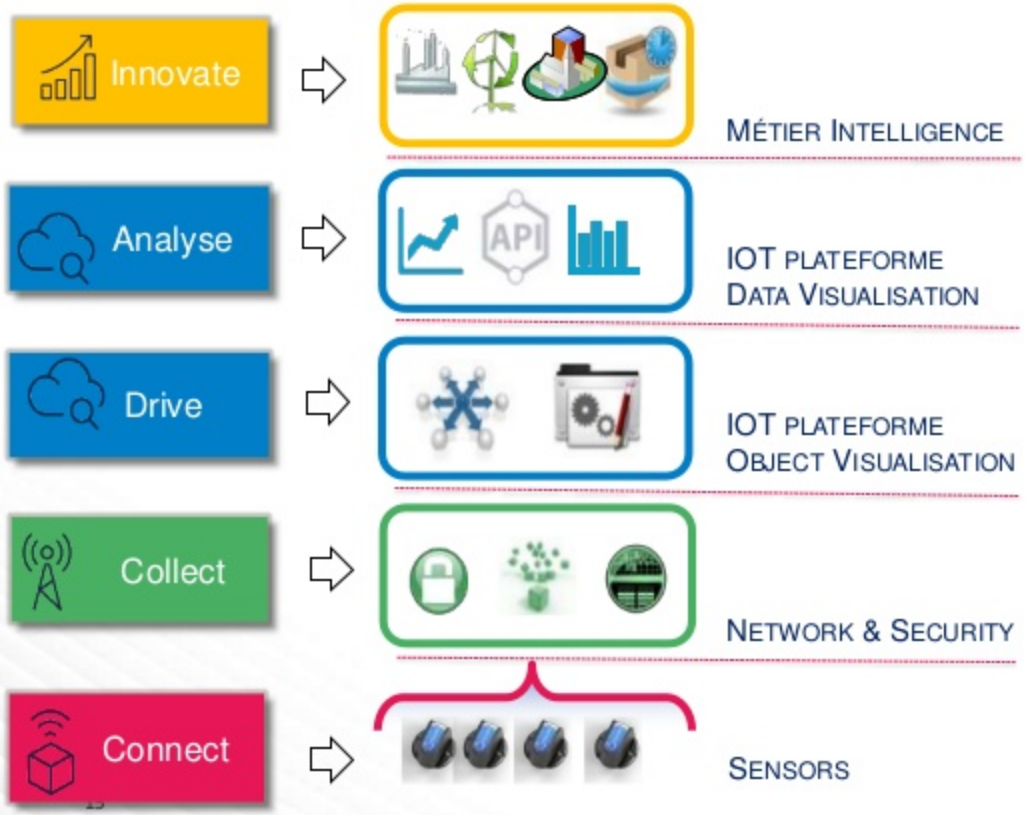
- Pitesti -> Barcelona (full)
  - 3 transports/week
  - 1 truck each time
  - Lead Time : 6 days
- NMUK -> Pitesti (Empty)
  - once a week
  - 1 truck each time
  - Lead Time : 6 days
  - Cross Dock in Strančice (Tchek Rep.)

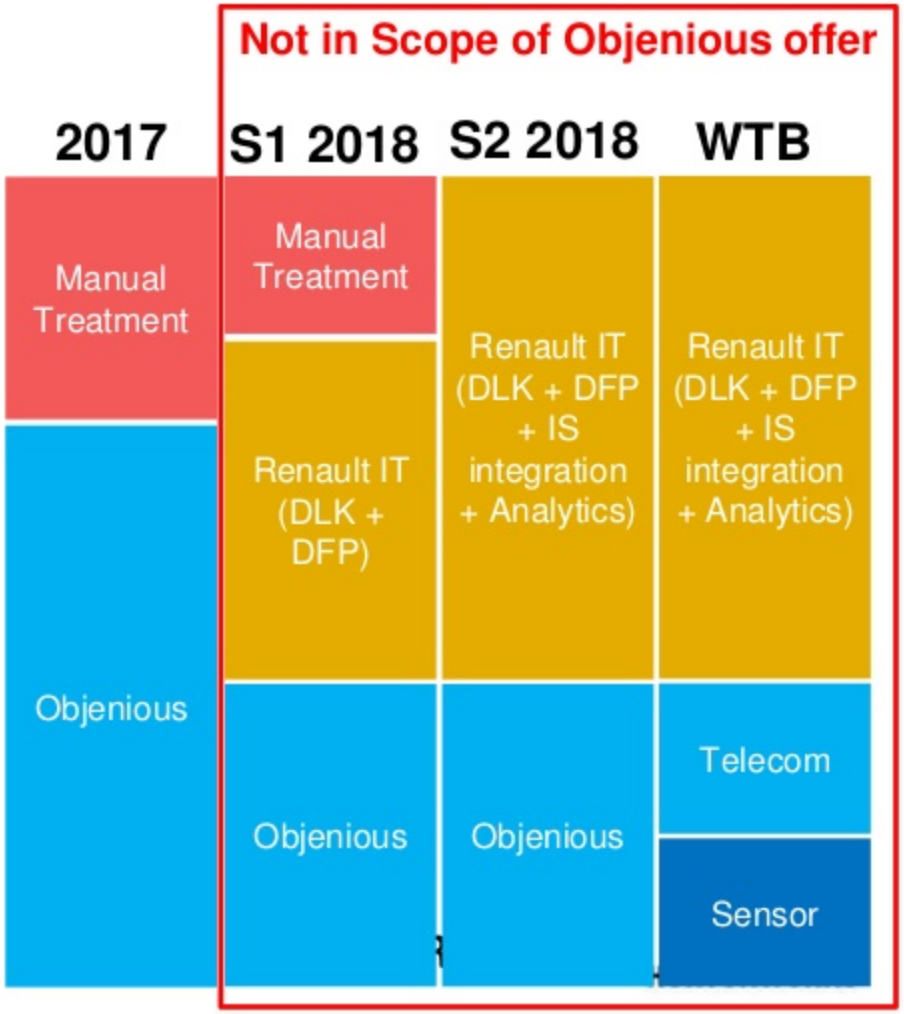


GROUPE RENAULT

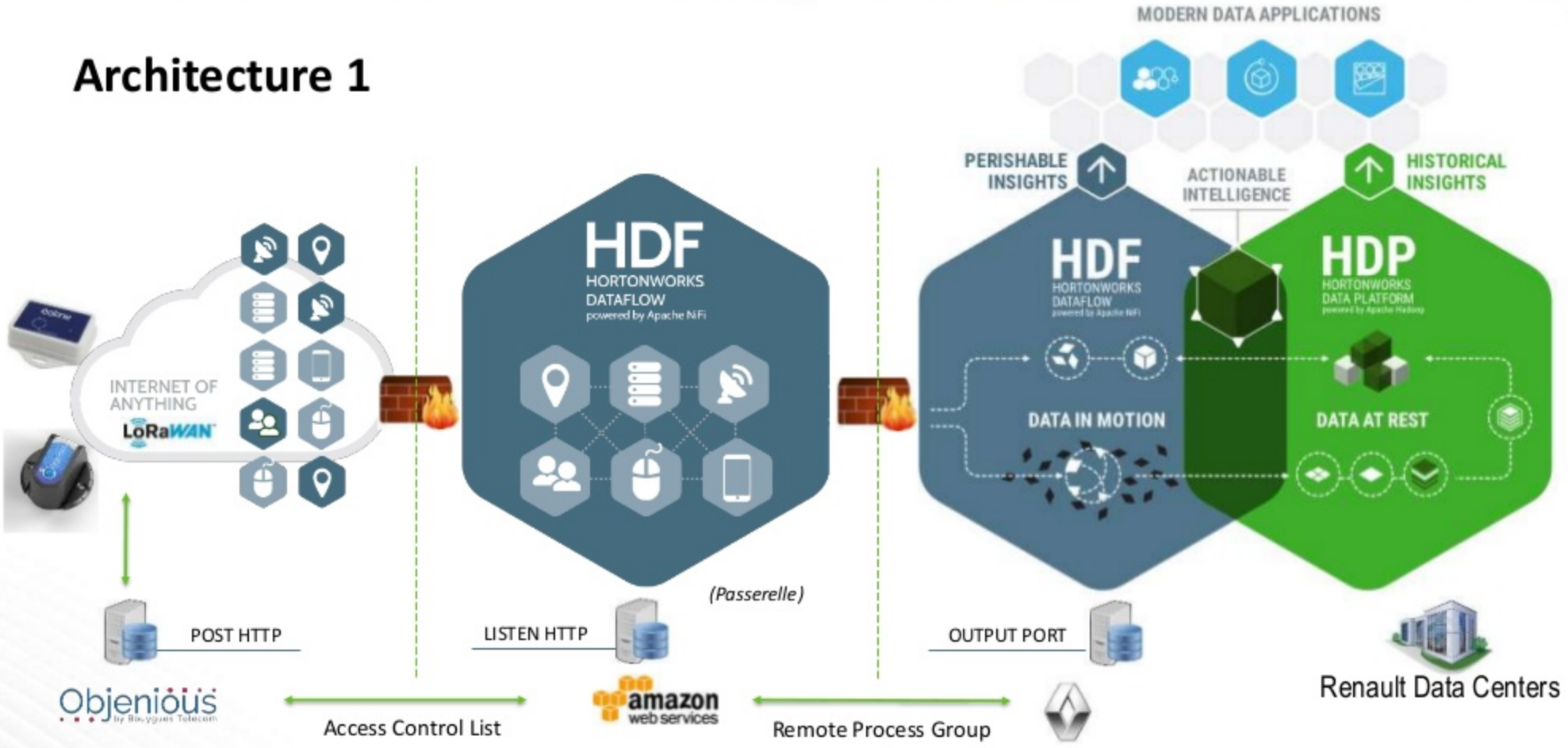
HORTONWORKS

Use case scope





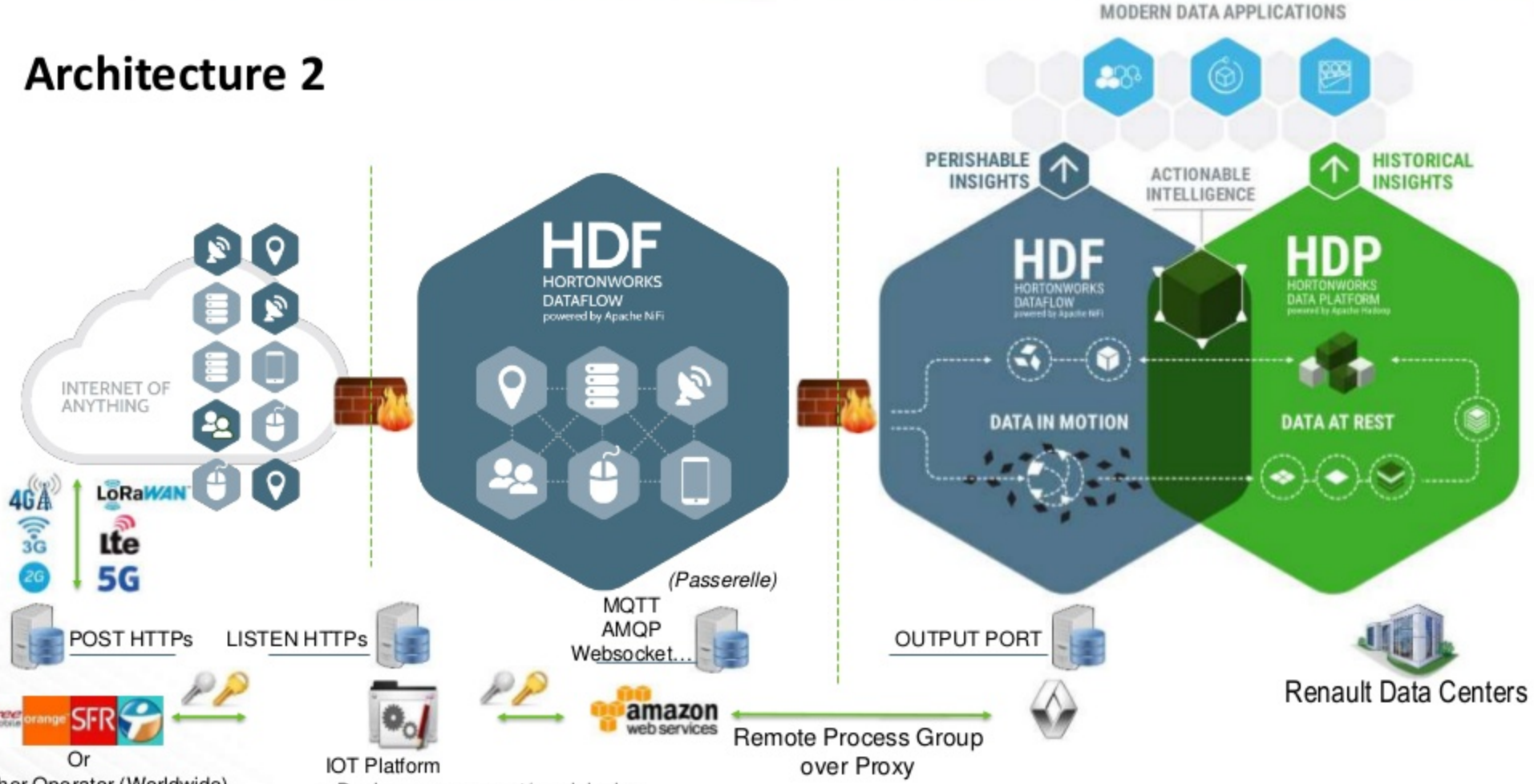
Architecture 1



GROUPE RENAULT

HORTONWORKS

Architecture 2

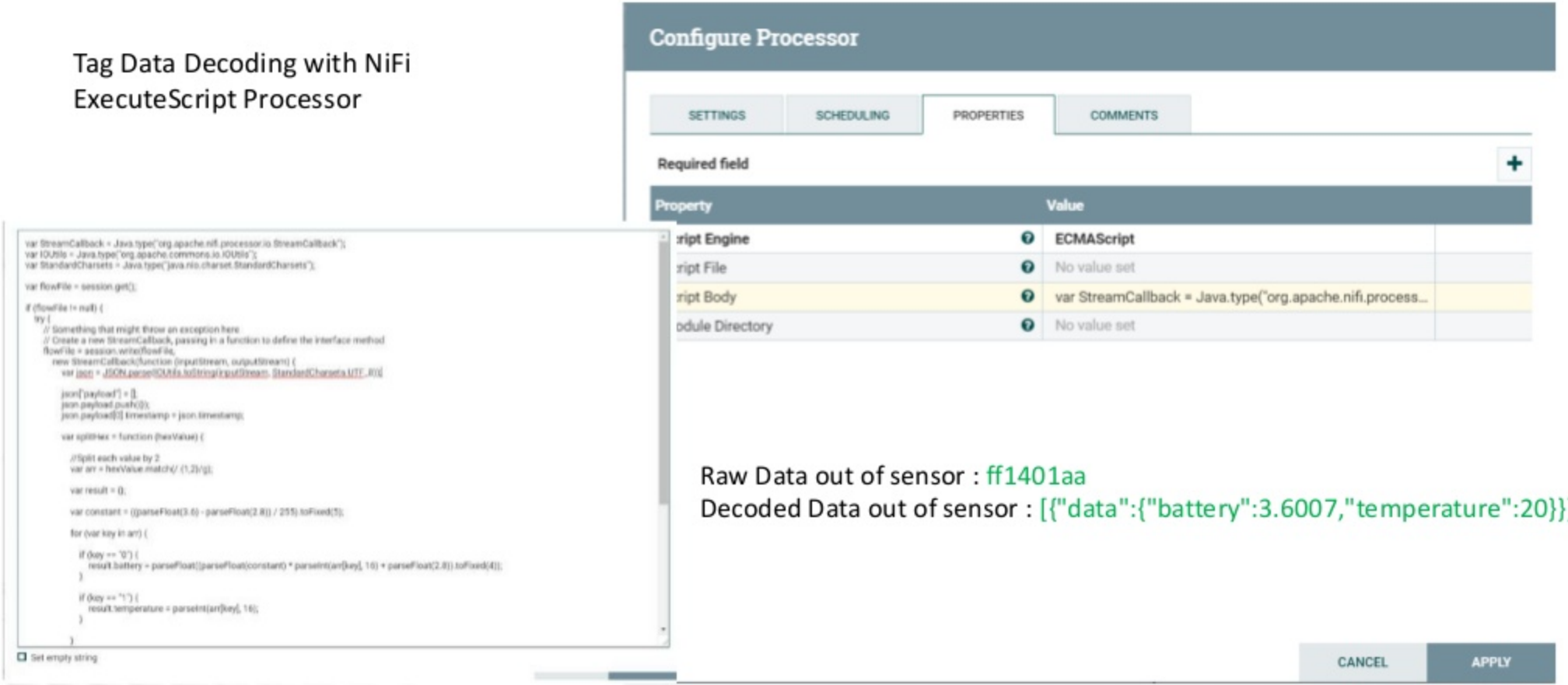


GROUPE RENAULT

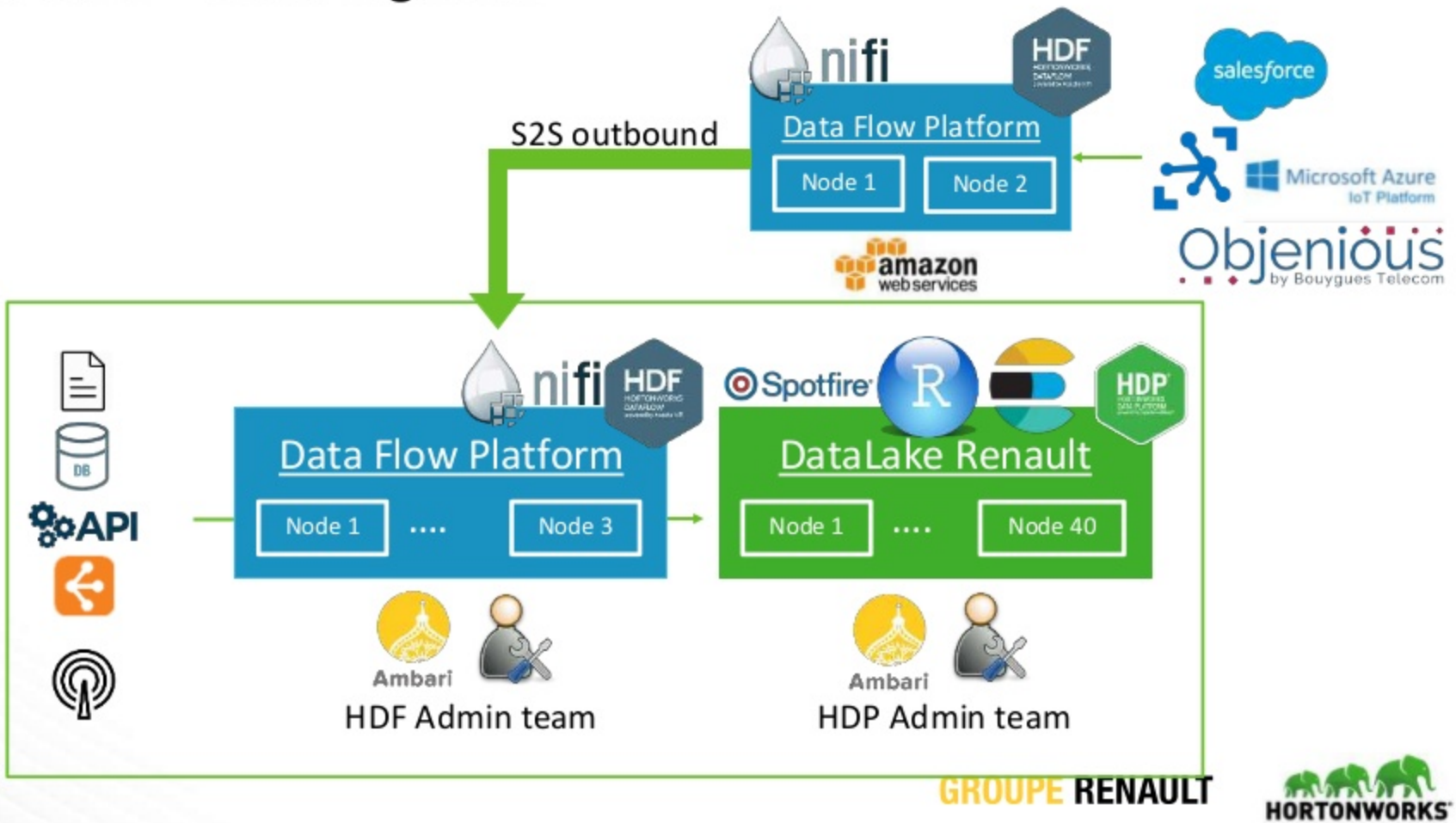
HORTONWORKS

Reduce cloud communication cost by 50%

Tag Data Decoding with NiFi
ExecuteScript Processor



Architecture + Cloud ingestion

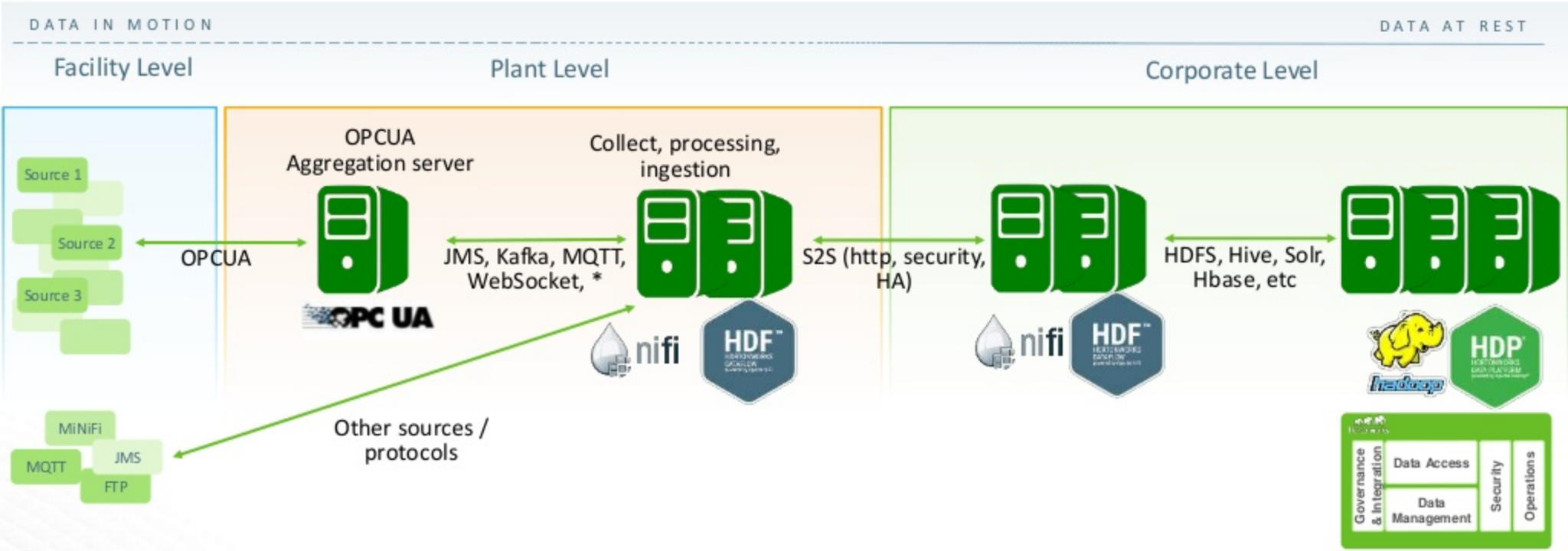


Connected plants

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

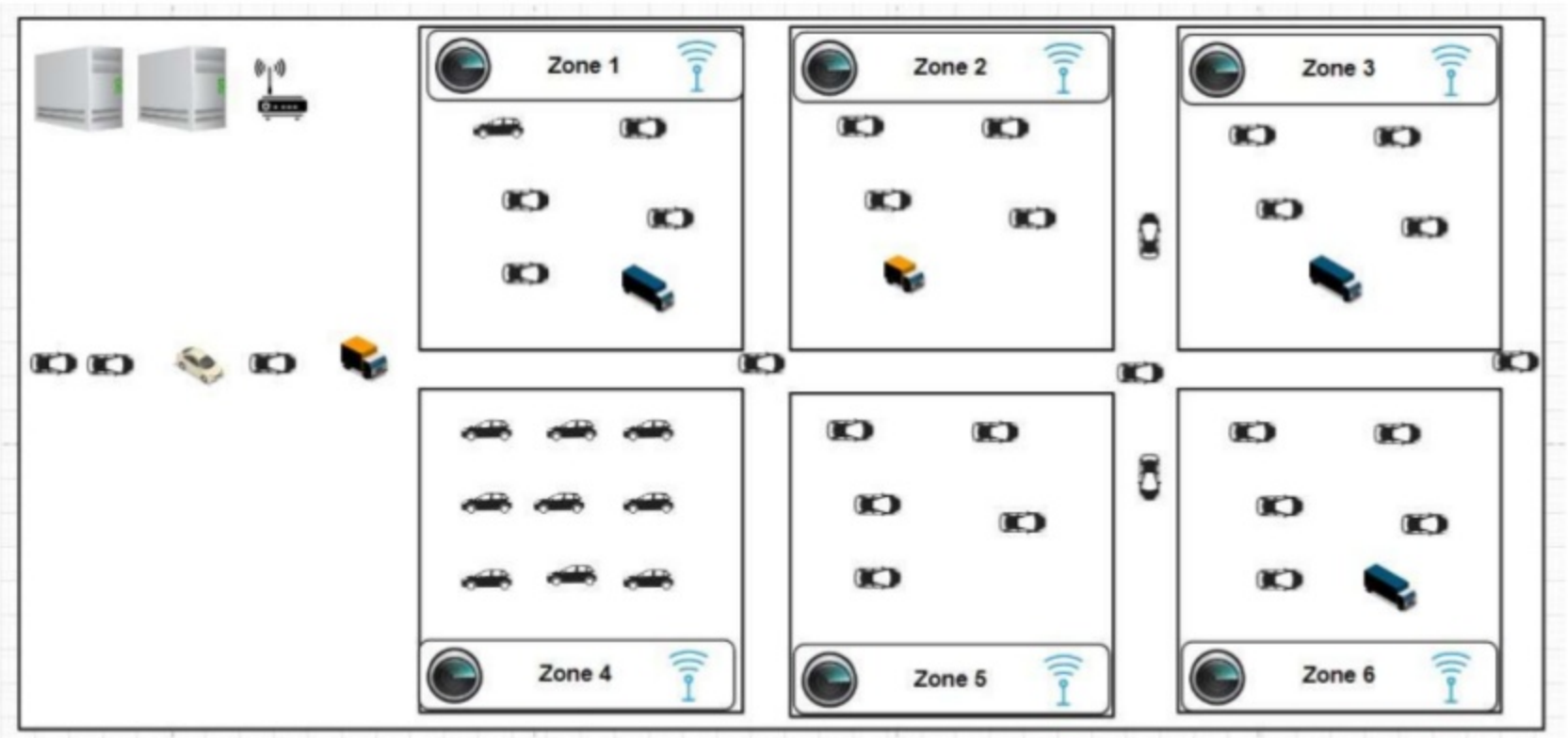
Connected plants



GROUPE RENAULT

HORTONWORKS

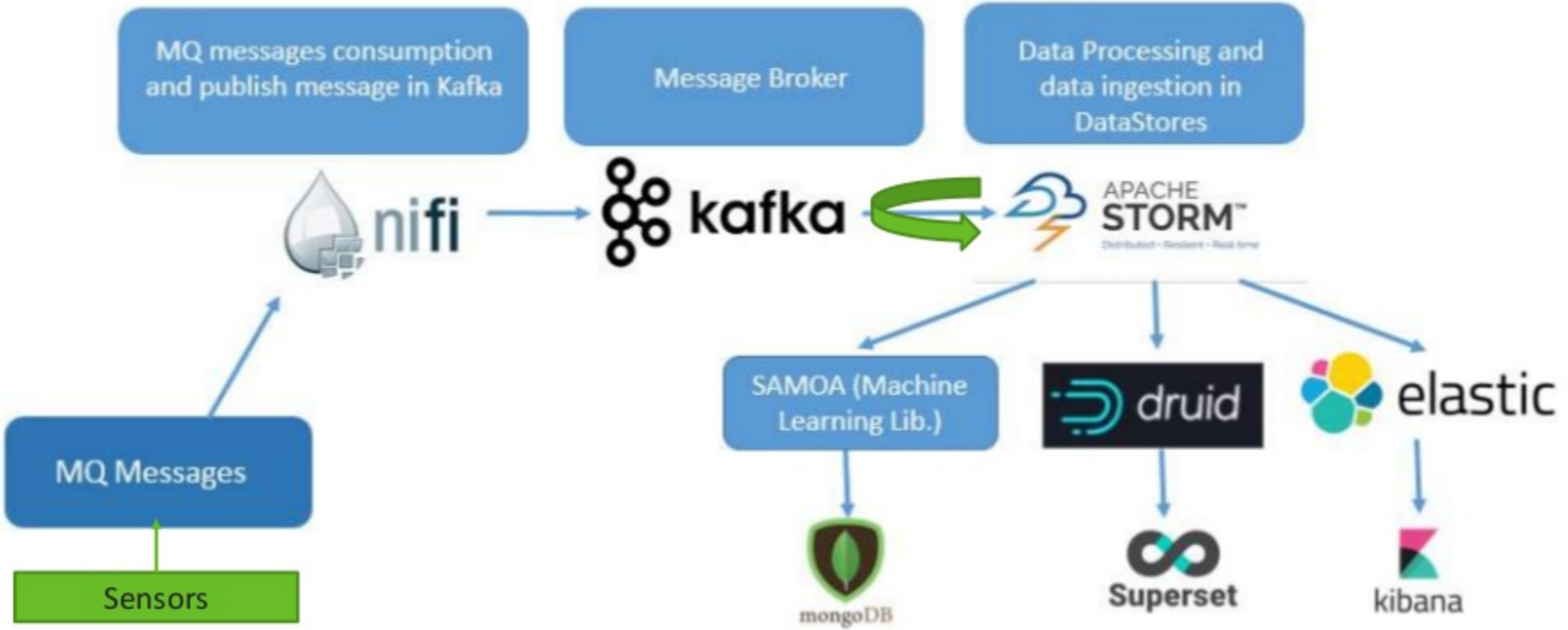
Overview of Plant Assembly Factory UC



GROUPE RENAULT

HORTONWORKS

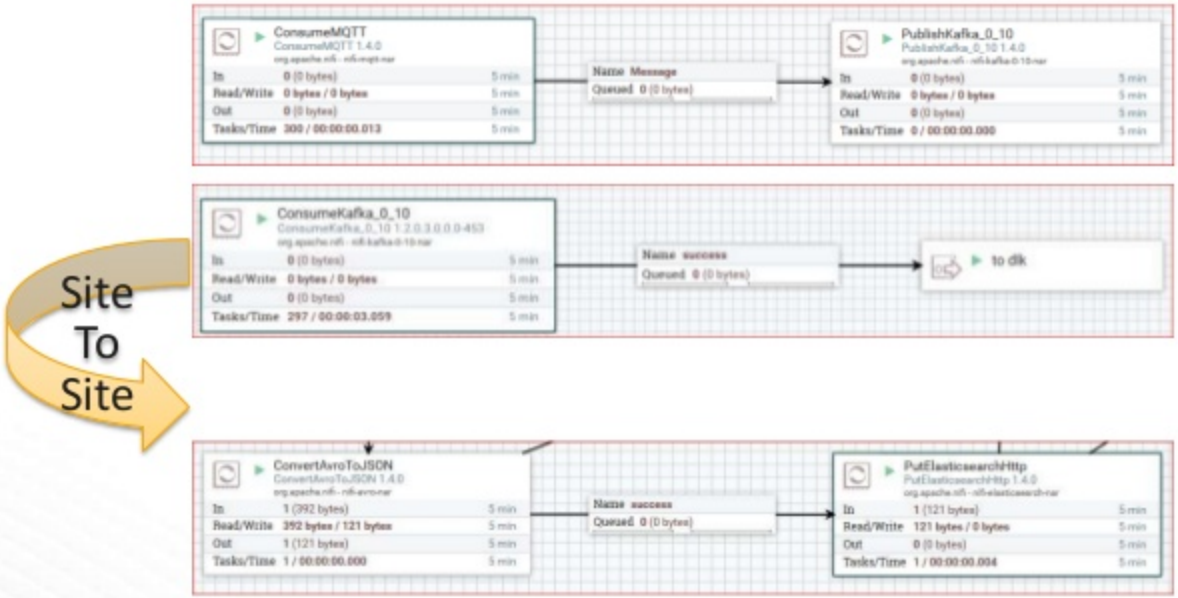
Full HDF use case (NiFi, Kafka and Storm)



GROUPE RENAULT

HORTONWORKS

Connected plants



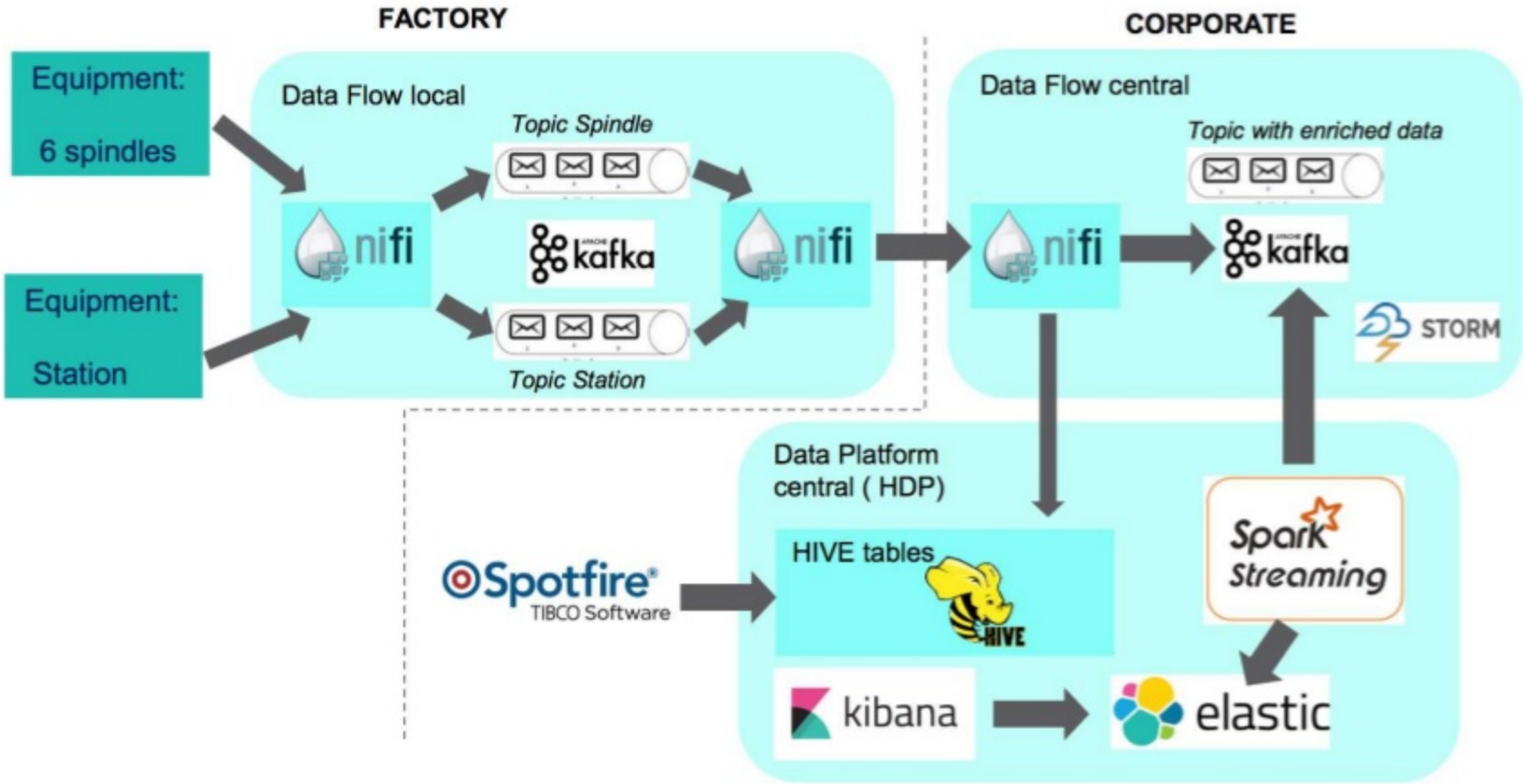
Demo with MQTTool from Hand to Data Lake !

Handy, MQTT Broker, NiFi ConsumeMQTT to Kafka, NIFI consume Kafka to ElasticSearch and Hive…

GROUPE RENAULT

HORTONWORKS

Screwing tools valladolid data analysis



Architecture + Connected plants

nifi
HDF
Data Flow Platform
Node 1
Node 2
Connected plants
S2S
S2S outbound
nifi
HDF
Data Flow Platform
Node 1
Node 2
salesforce
Microsoft Azure
IoT Platform
amazon web services
Objenious
by Bouygues Telecom
DB
API
nifi
HDF
Data Flow Platform
Node 1
....
Node 3
Spotfire
R
HDP
DataLake Renault
Node 1
....
Node 40
Ambari
HDF Admin team
Ambari
HDP Admin team
GROUPE RENAULT
HORTONWORKS

Best practices for running NiFi in production

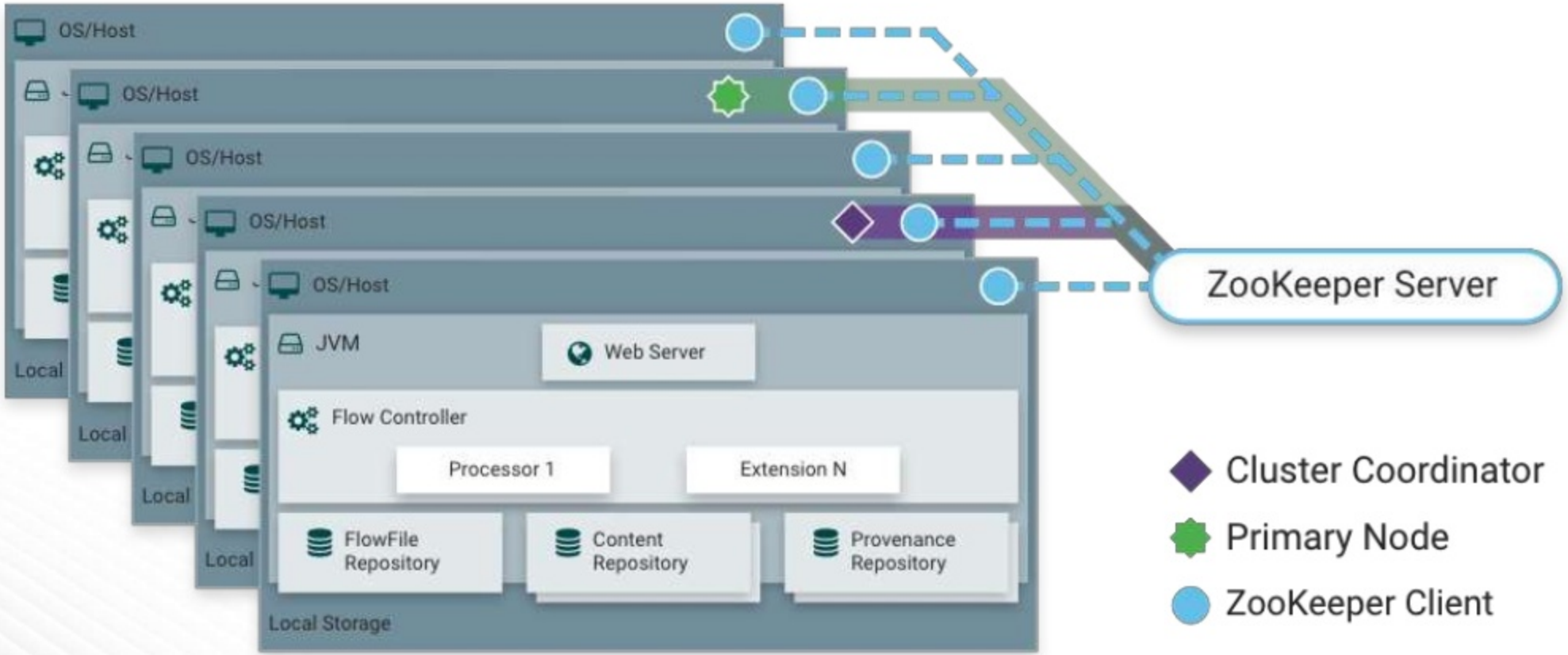
GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

Architecture & sizing

GROUPE RENAULT

Typical NiFi Production Cluster Logical View



GROUPE RENAULT

HORTONWORKS

Sizing considerations

- There are baselines for NiFi sizing but NiFi is like a “programming language” !!
- NiFi resources usage depends on used processors but it’s always IO intensive
  - Use different disks / volumes for the three repositories : flow file, content & provenance
  - For content repository, it’s recommended to have multiple mount points
  - For content and provenance repositories, SSD can provide the best performances
  - Heap sizing depends on the use case and used processors
  - Depending on the workload, we can scale vertically or horizontally
- Default settings are only for getting started. Tune based on your use case.
  - Thread pool size : Maximum Timer Driven Thread Count
  - Timeouts: nifi.cluster.node.connection/read.timeout, nifi.zookeeper.connect/session.timeout
  - Pluggable modules: WAL Provenance Repository

Development & deployment

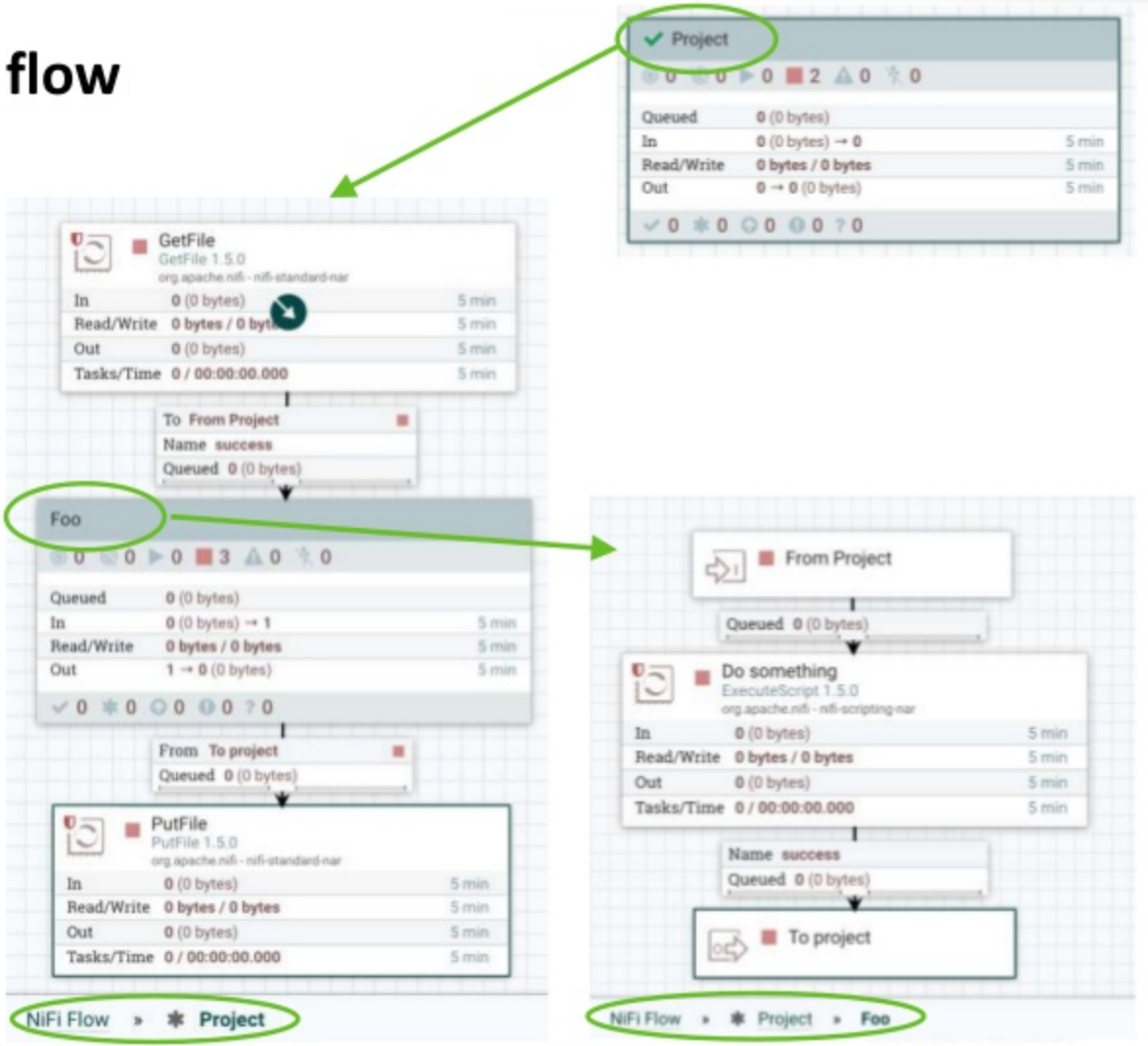
GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

Let's consider a simple data flow

```
Public class Project {

  private int foo(int x) {
        //do something with x -> y
        return y;
  }
  public static void main(String [] args) {
      //some code
      try {
           BufferedReader bufferRead = new
           BufferedReader ...;
           String data =bufferRead.readLine(
           data = foo(data);
           System.out.println(data);
      } catch ...
  }
}
```



Flow development = software development

Design > Code > Test > Deploy > Monitor > Evolve

Programing language

- Integrated Development Environment
- Algorithm, code, instructions
- Functions (arguments, results)
- Libraries
- …

Apache NiFi

- Apache NiFi UI
- Flows, processors, funnels 
- Process groups (input/out ports)   
- Templates 
- …

GROUPE RENAULT

HORTONWORKS

General guideline

Principle

- Use separate environments
- No flows at root level: use a PG per department, BL, project, etc
- Break your flow into process groups
- Use a naming convention, use comments (labels, comments)
- Use variable when possible
- Organize your projects into three PGs: ingestion, test & monitoring

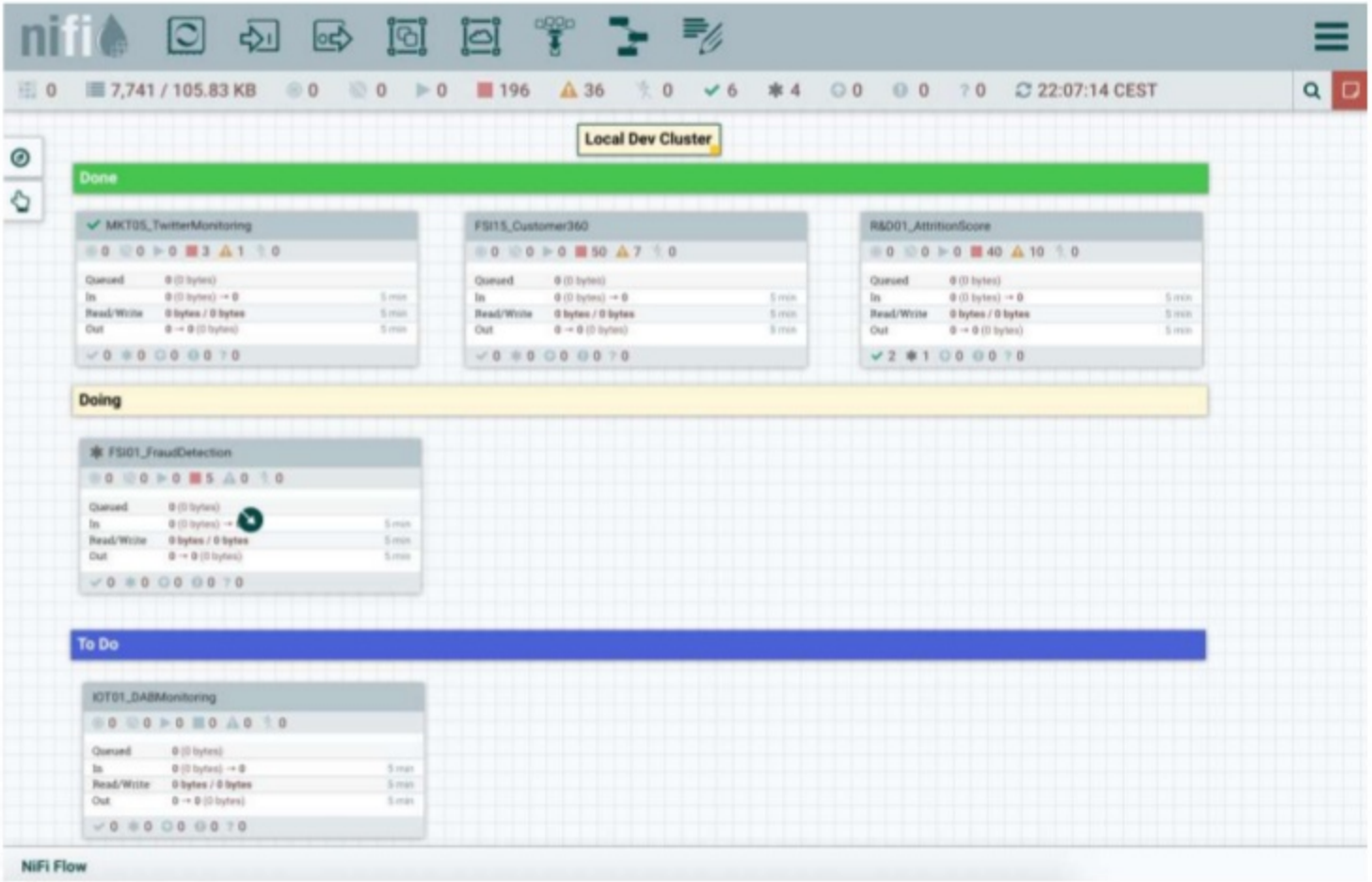
Benefits

- Performance, security and SLA
- Easy to secure, everything in NiFi is attached to a PG, heritage
- Easy to test, update & version
- Very useful for development/monitoring. Ideally use unique names
- Promotion (dev > test > prod), update
- NiFi can generate data for TDD (Test Driven Dev), can collect/parse logs for BAM (Business App Monitoring)

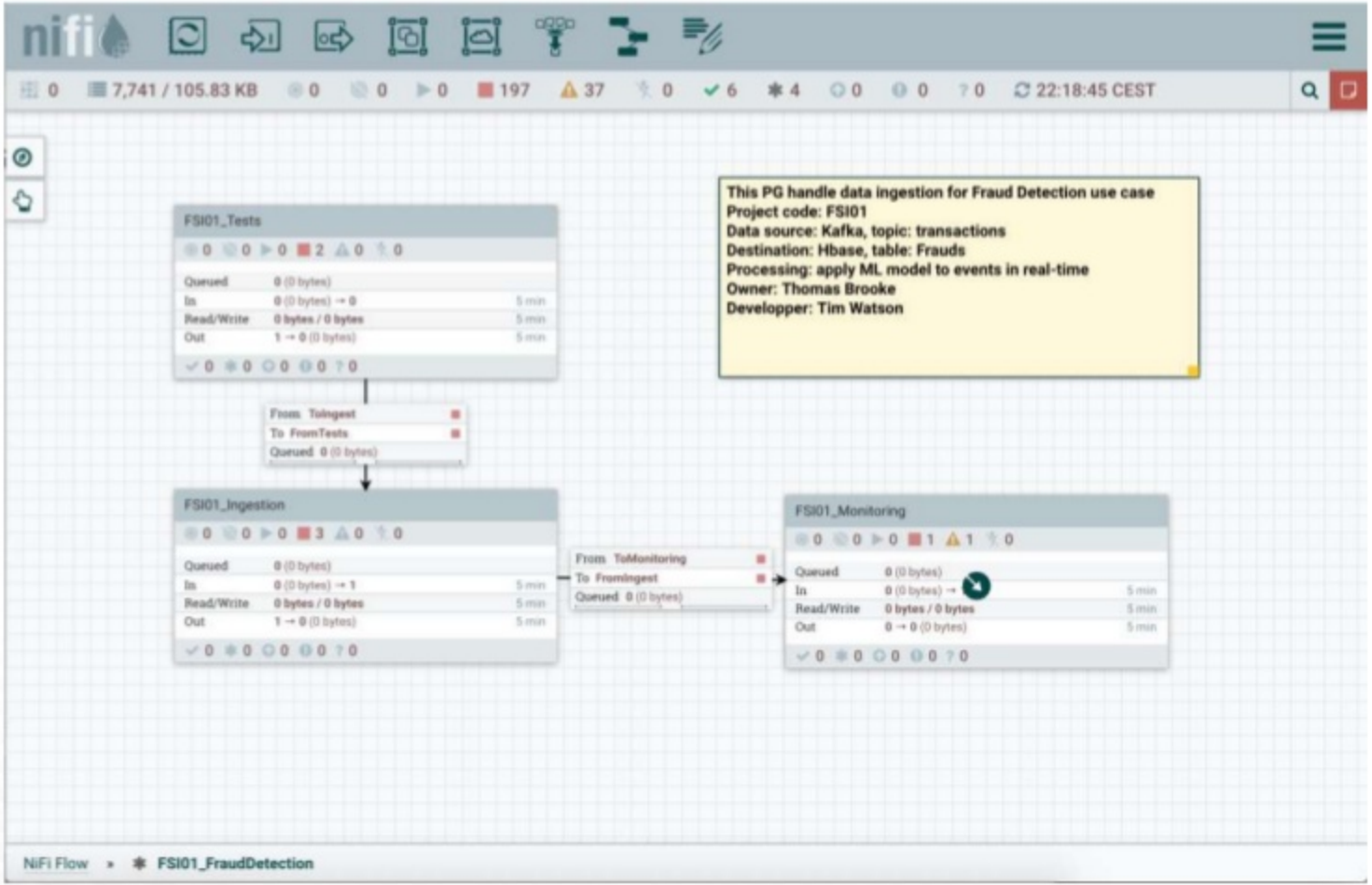
GROUPE RENAULT

HORTONWORKS

NiFi organization example



NiFi organization example



Know & use NiFi Design Patterns

Fan IN/OUT
List & Fetch

Attributes promotion
Extract, Update Attribute

Throttling
ControlRate, expiration

Funneling
RPG, RouteOnAttribute

Error loops
Relations, Counters

etc

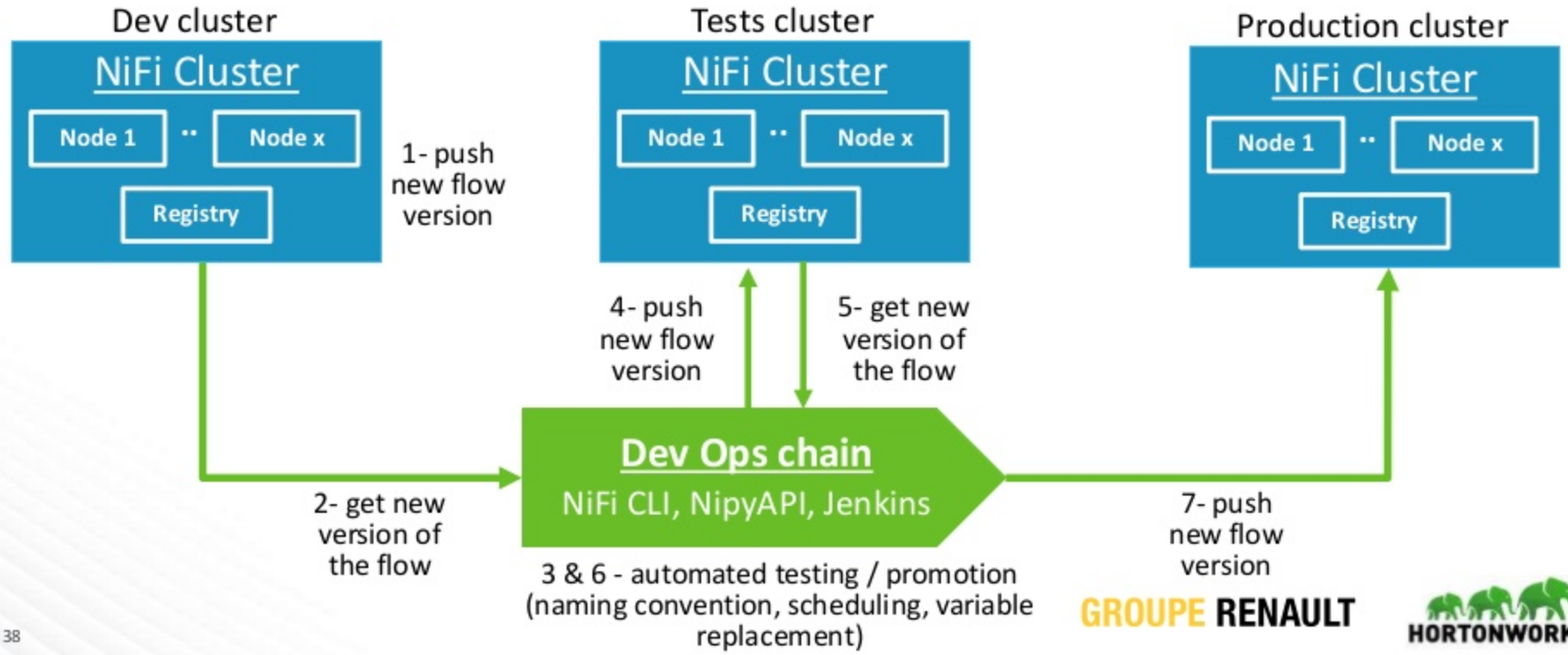
FDLC: Flow Development LifeCycle

APACHE NIFI registry

Forget Duplicating Local Changes: Apache NiFi and the Flow Development Lifecycle (FDLC)

TECHNICAL ● Data Processing and Warehousing

Thursday, April 19
4:00 PM - 4:40 PM
Room II



GROUPE RENAULT

HORTONWORKS

Monitoring

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

NiFi Monitoring

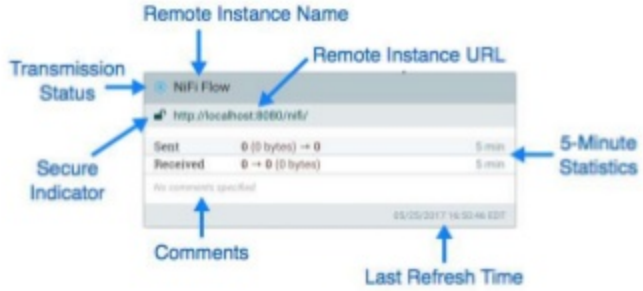
Service monitoring

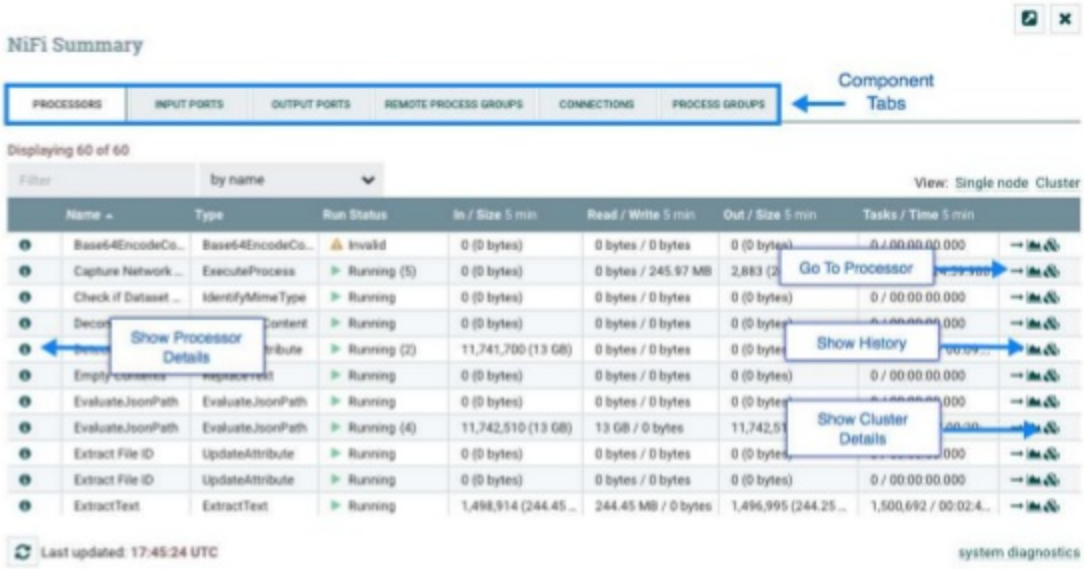
- Is NiFi service running correctly?
- Monitor global system metrics such as threads, JVM, disk, etc
- Monitor global flow metrics such as number of flow files sent, received or queued, processors stopped, etc
- Solutions
  - NiFi UI
  - Reporting tasks
  - Ambari
  - Grafana

Applications (Flow) monitoring

- Are a particular flow running correctly?
- Monitor per application (flow, PG, processor) metrics such as number of flow files, data size, queues, back pressure, etc
- Solution
  - S2S Reporting tasks
  - Custom flow developments (integrate monitoring and reporting in the application logic)

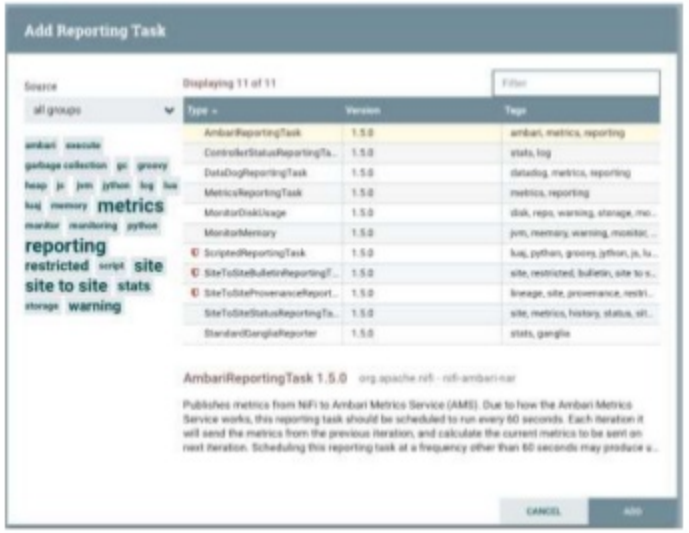
NiFi UI for service monitoring





Tools available for service monitoring

- Bootstrap notifier: send notification when the NiFi starts, stops or died unexpectedly
  - Email/HTTP notification services
- Use reporting tasks: export metrics to your monitoring solution
  - AmbariReportingTask (global, process group)
  - MonitorDiskUsage (Flowfile, content, provenance repositories)
  - MonitorMemory
- Also, monitor inactivity
  - NiFi has a built-in MonitorActivity processor
  - To be used with the S2SBulletinReportingTask
  - You can use InvokeHTTP to call the reporting Rest API



GROUPE RENAULT

HORTONWORKS

How to achieve granular monitoring?

Monitoring Driven Development (MDD)

- Integrate your monitoring logic in the flow design
  - Count data (lines, tables, events, etc)
  - Extract business metadata from data (table names, project name, source directory, etc)
- Handle different types of errors (connection, format, schema, etc)
- Send extracted KPI to brokers, dashboards, API, files, etc

S2S reporting tasks

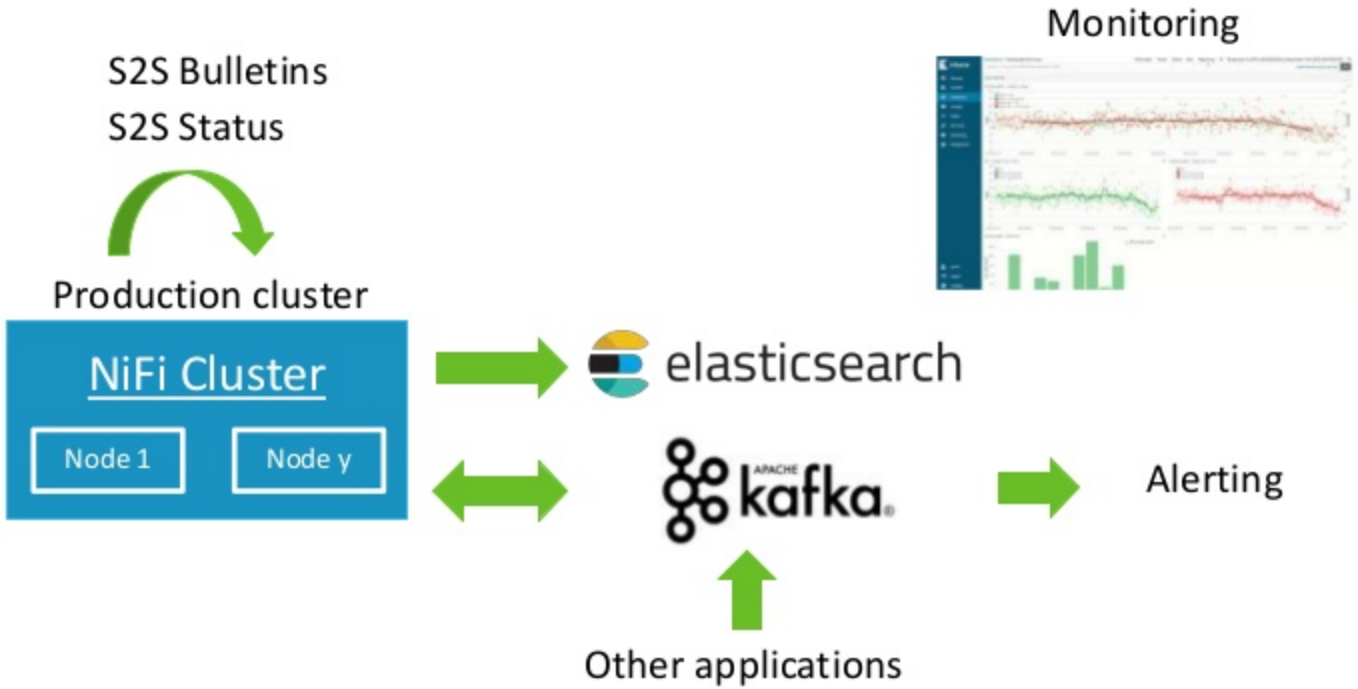
- NIFInception: NiFi can export metrics to to another cluster or to itself
- Bulletins provide information on errors
- Status provide metrics on usage
- These reports become data and we can use the power of NiFi to extract our KPI
- **Naming convention is key**

Monitoring Driven Development

GROUPE RENAULT

NiFInception architecture

- Generate status and bulletins
- Send them through S2S to a local input port
- Ingest bulletin and status reports
- Parse reports (JSON) and extract useful KPI
- Send KPI to a dashboard tool (AMS, Elastic, etc)
- Use a broker for alerting (Kafka)
- Can ingest logs from other systems or application

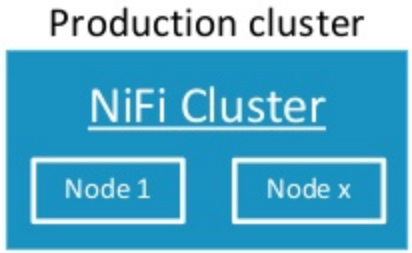


GROUPE RENAULT

HORTONWORKS

NiFInception architecture 2

- Ingest data
- Generate status and bulletins
- Send them through S2S to a remote cluster



S2S Bulletins
S2S Status

Monitoring

- Ingest bulletin and status reports
- Parse reports (JSON) and extract useful KPI
- Send KPI to a dashboard tool (AMS, Elastic, etc)
- Use a broker for alerting (Kafka)
- Can ingest logs from other systems or application

elasticsearch

APACHE kafka

Alerting

Other applications

GROUPE RENAULT

HORTONWORKS

Lessons learnt at Renault

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA

Recommendations from day to day life

- High availability means several weeks to validate before Go live
- Use Ranger authorizations instead of NIFI build in ACL management
- Too much « if then do else »: do the data flow and do not offload other processing solutions
- S2S : Minimize the number of RPG and self-RPG. Use attributes and routing.
- Put hive via Two redundant processors
- Discuss with DBA to better handle impacts (Views VS Tables, Open Sockets ..etc).
- Backup flowfile.xml.gz (users.xml et autorizations.xml) using NiFi itself.

Recommendations from day to day life

- Success flag can be done trough counters instead of InvokHTTP
- In case of CIFS, do not forget to use AUTO MOUNT for NFS client side on NiFi Servers.
- Check ‘0’ size before transmitting into HDFS (especially for IoT use case).
- Configure a TIMER for when we use FAILURE redirections to avoid back-pressure scenario.
- Build separate clusters for separate use cases (Real Time with SSD, Batch, etc …)
- CA PKI very useful for internal communications, no need to wait Security teams answers but need security skills (SSL) .

Questions

GROUPE RENAULT

HORTONWORKS
POWERING THE FUTURE OF DATA