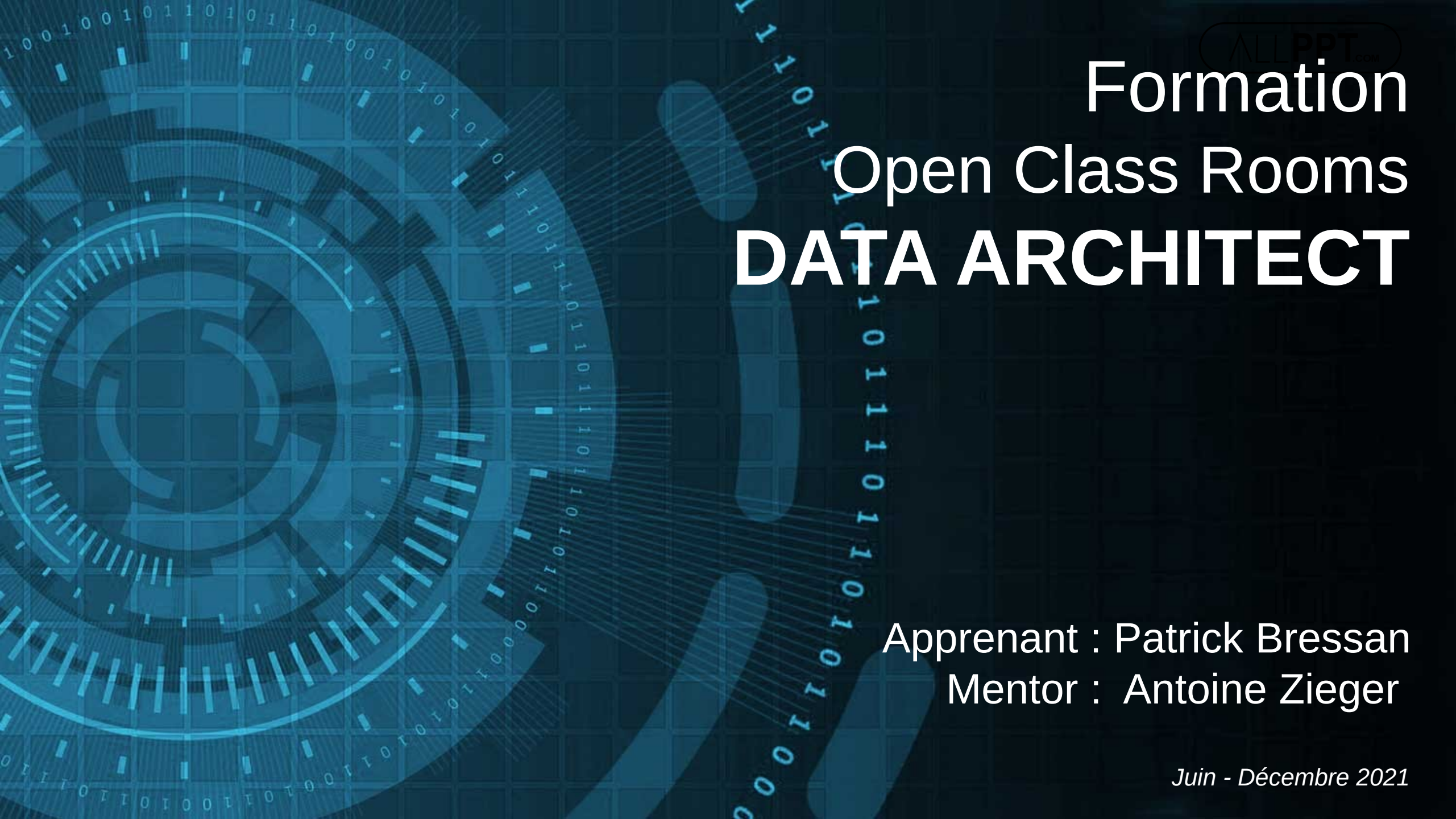


Formation
Open Class Rooms
DATA ARCHITECT

Apprenant : Patrick Bressan
Mentor : Antoine Zieger

Juin - Décembre 2021

Projet 5

Développez une architecture Big Data complète

Résumé du contexte, présentation des enjeux (1/3)

Votre mission

Vous êtes un data architect chez Twitter et votre rôle est de déployer une solution complète d'analyse de données pour créer un top 10 des sujets les plus tendance. Vous allez devoir mettre en place :

1. la collecte des données,
2. leur stockage dans des structures adaptées,
3. leur traitement au coup par coup et en temps réel,
4. des solutions pour améliorer la robustesse de l'architecture globale et de chacun de ses composants.

Résumé du contexte, présentation des enjeux (2/3)

Vous devrez créer un outil qui permet de lister les dix hashtags les plus fréquents pour chaque heure. Vous devez par exemple pouvoir répondre aux questions suivantes :

- Quels sont les mots-clés les plus tendance depuis une heure ?
- Quels étaient les mots-clés les plus tendance lundi dernier entre 7h et 8h ?
- J'ai fait une erreur dans mon algorithme d'identification des hashtags ; peut-on recalculer les hashtags les plus fréquents utilisés depuis lundi dernier à 7h ?

Pour ce faire, vous devrez mettre en place à la fois un outil de calcul de statistiques en batch, à partir de données stockées sur HDFS, et en temps réel, à l'aide d'un pipeline de traitement de données.

Résumé du contexte, présentation des enjeux (3/3)

1. Les données temps-réel seront transmises à un cluster Kafka pour à la fois être stockées dans un système de fichiers distribués (HDFS) et transférées à un pipeline Storm.
2. Batch layer : Les données présentes dans HDFS seront analysées périodiquement et par lot (“batch”) par des jobs MapReduce (Hadoop ou Spark).
3. View layer : Le résultat de ces jobs MapReduce sera stocké dans MongoDb.
4. Speed layer : Les données traitées par le pipeline Storm/Kafka seront stockées dans MongoDb. Les données rendues inutiles par l’exécution des tâches de la batch layer devront être effacées au moment opportun.
   Un simple script en ligne de commande suffira pour visualiser les tendances Twitter à une heure donnée.

Progression d’apprentissage et de réalisation

#	PROGRESSION
1	Déploiement local de dockers kafka, storm, mongodb, HDFS
2	Prise en main de kafka-connect et confluent schema-registry
3	Déploiement local de l’intégralité de l’infrastructure via docker-compose
4	Développement du « kafka producer » en python (Twitter API)
5	Configuration de kafka-connect pour synchronisation avec HDFS
6	Prise en main de MongoDB et spécification des collections
7	Développement de KafkaSpout et StormBolt et intégration avec MongoDB
8	Développement du script d’aggrégation des #hashtags heure par heure
9	Développement du script d’interrogation des #hashtags
10	Déploiement de l’infrastrucure sur AWS

Schéma d’architecture (1/4) || API to Kafka

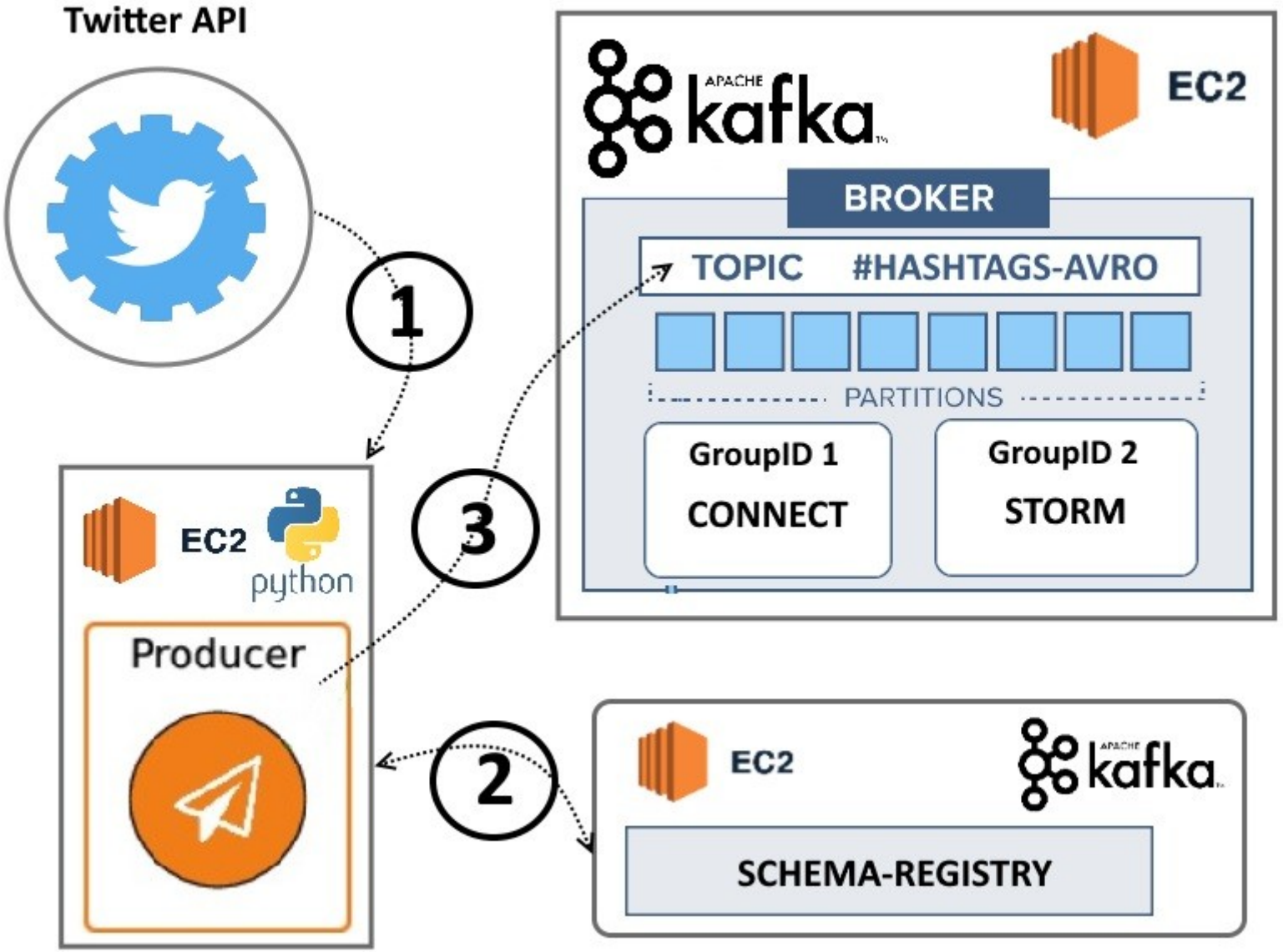


Schéma d’architecture (2/4) || API to HDFS

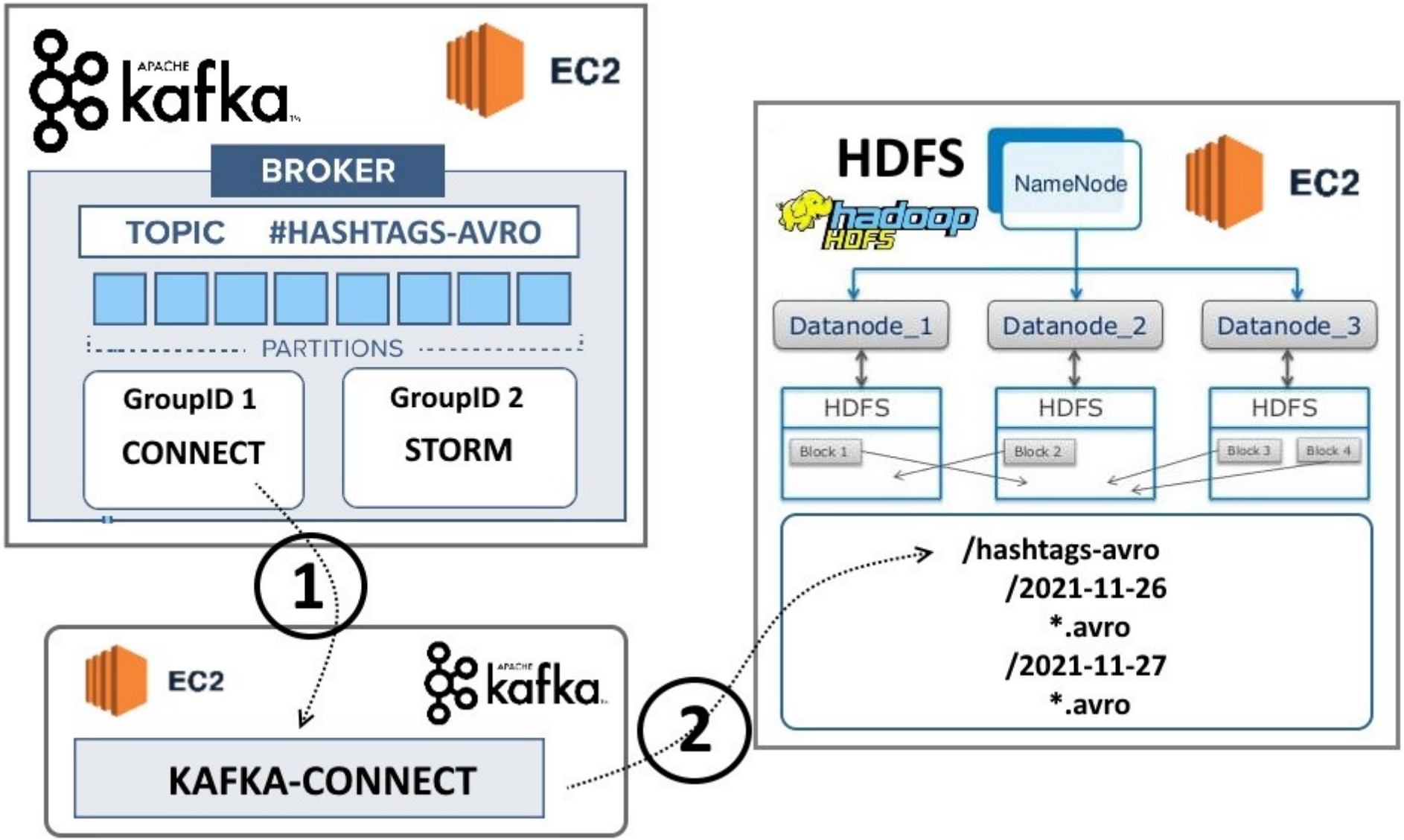


Schéma d’architecture (3/4) || Kafka to MongoDB

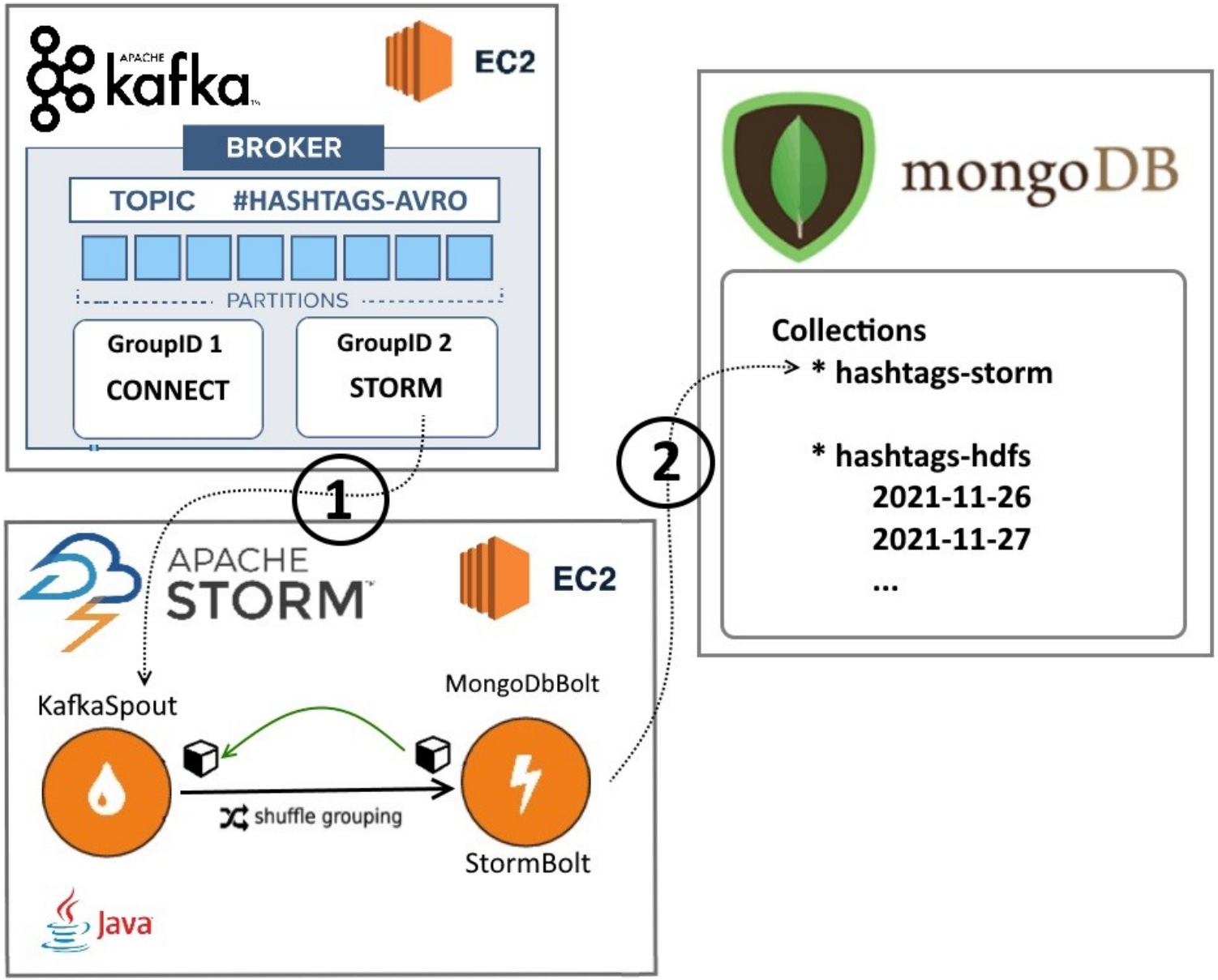
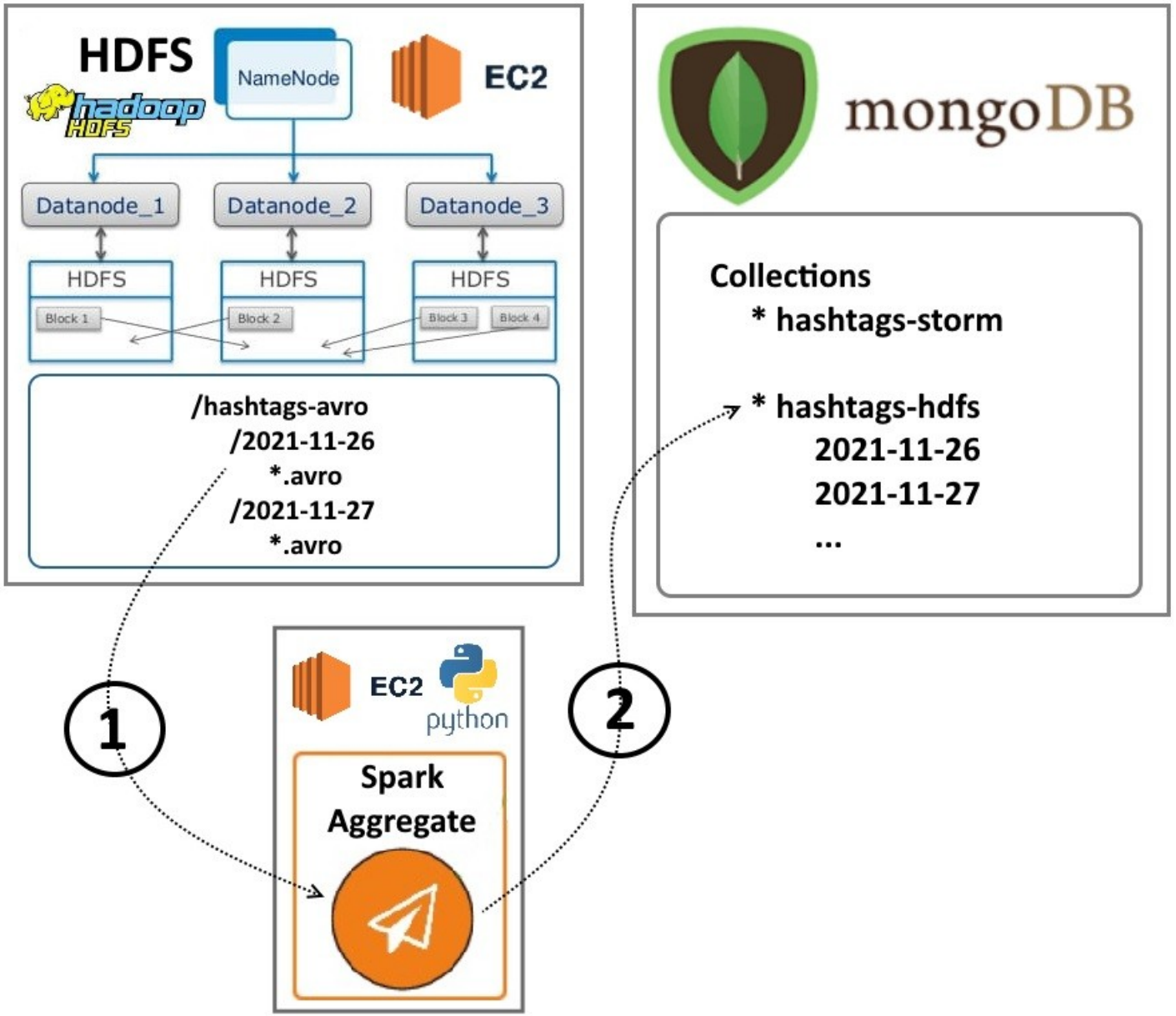
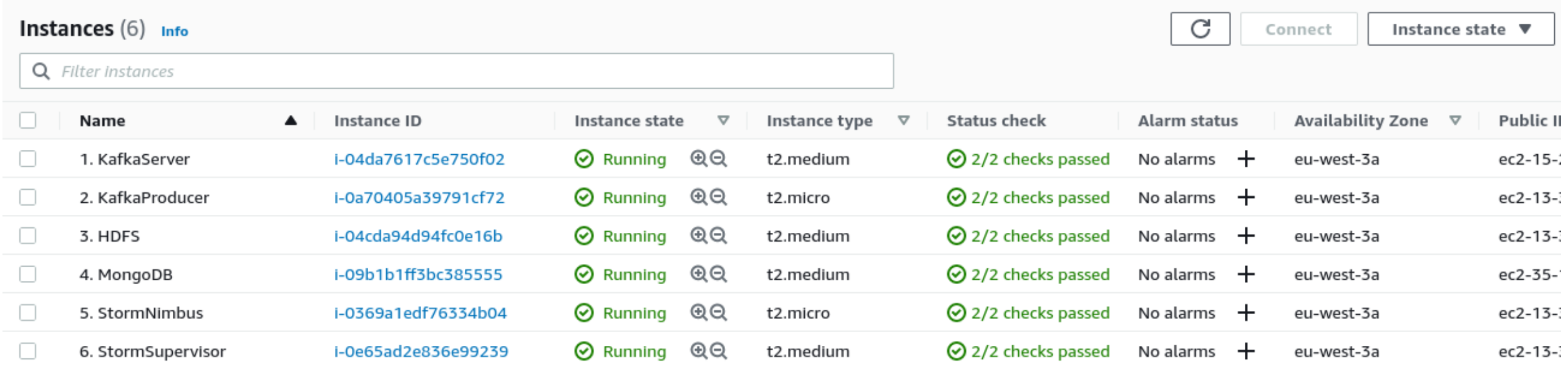


Schéma d’architecture (4/4) || HDFS to MongoDB



AWS (EC2)



Composants d'Architecture	Nombre
ZooKeeper	1
Kafka Brokers	3
Kafka Topics : hashtags-avro || Partition 3 / Replication Factor 3	2
Kafka Producer (Python), Kafka-Connect (HDFS Sink Tasks)	1, 3
HDFS Master Node, HDFS Secondary Master Node, HDFS Data Nodes	1, 1, 3
Schema Registry	1
MongoDB, Shards, ReplicaSet	1, 0, 0
Storm Nimbus, UI, Supervisor (Workers)	1, 1, 3

Kafka Python Producer (2/2)

```
#! /usr/bin/env python

from confluent_kafka import avro
from confluent_kafka.avro import AvroProducer
```

```
    schema = avro.loads(avro_schema)

    avro_producer = AvroProducer({
        'bootstrap.servers': ["172.31.2.27:9092", "172.31.2.27:9093", "172.31.2.27:9094"],
        'on_delivery': delivery_report,
        'schema.registry.url': 'http://localhost:8081'
        }, default_value_schema=schema
    )
```

```
                    avro_producer.produce(
                        topic='hashtags-avro',
                        value=msg,
                        value_schema=schema)
```

Kafka Python Producer (1/2)

```
ubuntu@ip-172-31-13-207:~/logs$ tail -20f twitter-api_v3.log
b'{"unique_id": "26485952e0e6f3d41da80613d33bff6ab79509ebfa72a36f1c527c8f9b6fa1c1", "timestamp": 1637935588, "date": "2021-11-26", "hour": "14", "hashtag": "quantique"}'
b'{"unique_id": "34e66597cb83e1f9d8daae0a140713b8dd05efffe730d3210792b7fde7d869a2", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "CryptoNews"}'
b'{"unique_id": "fa67e695eb524098dcd16639592bf67ef3778c54a7a88d120a16cfb0c929f539", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "BSC"}'
b'{"unique_id": "131d3e681c0bb57520ad153b10cbe4de555ce440a8d3dd78845cb1fe0b79045e", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "BSCGem"}'
b'{"unique_id": "c25e5f6c907bce083bce809565d754cf2fa5858ddfea30a464bd98bed0e97389", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "altcoin"}'
b'{"unique_id": "5f844556fc58059a1041df8fdf4124abf4165cca8ac3d223107e4a1a4ebc1d3c", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "ArtificialIntelligence"}'
b'{"unique_id": "cf69e28491d0f184856e46793db27d99275a569b8c03ce4218c63ceb94992fe5", "timestamp": 1637935598, "date": "2021-11-26", "hour": "14", "hashtag": "ai"}'
b'{"unique_id": "94172b9dd85df0ecebddcb98a8026ede0a664a02a19ec0f31413275558eed628", "timestamp": 1637935602, "date": "2021-11-26", "hour": "14", "hashtag": "STM"}'
b'{"unique_id": "2731e645a6c92bc93dae73ed7c2cc995f2643f9cc069fcc08c175f437037d0ef", "timestamp": 1637935613, "date": "2021-11-26", "hour": "14", "hashtag": "Hulot"}'
b'{"unique_id": "09b1b0e50d2d187bdf0be6da32804a9d6b0654660414ff95794e64086ca05aff", "timestamp": 1637935613, "date": "2021-11-26", "hour": "14", "hashtag": "Macron"}'
b'{"unique_id": "1191f3bdd8d7677d7586362b19c0a025dff9f0aa7050188d266a7806c98a5421", "timestamp": 1637935630, "date": "2021-11-26", "hour": "14", "hashtag": "R\\u00e9alisations"}'
b'{"unique_id": "a9228bbe043615543dd54f8fdcca7cbeab8c6203e145630e12a50a5fba8387bc", "timestamp": 1637935630, "date": "2021-11-26", "hour": "14", "hashtag": "Dictature_des_G\\u00e9n\\u
b'{"unique_id": "790ac587345c886dd401531cd209fa11e6be18d6745fb5232bc945ddfdd1b655", "timestamp": 1637935630, "date": "2021-11-26", "hour": "14", "hashtag": "Terrorisme"}'
b'{"unique_id": "37eebe43a37d1092d1db383beb0ef59ca0417cd7bebb44efeb9402c55fb886c4", "timestamp": 1637935645, "date": "2021-11-26", "hour": "14", "hashtag": "Lustre"}'
b'{"unique_id": "509cf8a54763eb31dd2f5043d91af9ab49c3ed7edcd996d50fcd801f6f2205b4", "timestamp": 1637935648, "date": "2021-11-26", "hour": "14", "hashtag": "DEGAGE"}'
b'{"unique_id": "65c87e22d6774ed9501d6118d1b66457280b9edf0f279f71fb0314e42a1f773b", "timestamp": 1637935650, "date": "2021-11-26", "hour": "14", "hashtag": "Concours"}'
b'{"unique_id": "a7abd513f9f110828ca92bd0a5e9f93f9c590520a9e00ed4ef77a725df4e8aca", "timestamp": 1637935650, "date": "2021-11-26", "hour": "14", "hashtag": "BlackFriday"}'
b'{"unique_id": "2cb4d991b88713866c876c9ebe0a2891eb54ffb470458950d87a1340abe3d16b", "timestamp": 1637935653, "date": "2021-11-26", "hour": "14", "hashtag": "GDRAGON"}'
b'{"unique_id": "4c7583e8916aebb4cc35a0b6eb890f28f9d4bd2d86826bb80fb20fe970930768", "timestamp": 1637935662, "date": "2021-11-26", "hour": "14", "hashtag": "GenerationMelenchon"}'
b'{"unique_id": "5fd08995037da19a0dc2899d477c4421b55094eefc6fde713266ea34e98f47bc", "timestamp": 1637935663, "date": "2021-11-26", "hour": "14", "hashtag": "Arknights"}'
```

AVRO SCHEMA

```
avro_schema = """
{
    "namespace": "ocr.p5",
    "name": "hashtags",
    "type": "record",
    "fields" : [
        { "name" : "unique_id", "type" : "string" },
        { "name" : "timestamp", "type" : "int", "logicalType": "date" },
        { "name" : "date", "type" : "string" },
        { "name" : "hour", "type" : "string" },
        { "name" : "hashtag", "type" : "string" }
    ]
}
"""
```

KAFKA to HDFS Connector Config

```
{
    "name":"hdfs-sink","config":{
        "connector.class":"io.confluent.connect.hdfs.HdfsSinkConnector",
        "tasks.max":3,
        "topics":"hashtags-avro",
        "hdfs.url":"hdfs://namenode:9000",
        "flush.size":30,
        "logs.dir":"/logs/",
        "topics.dir":"/",
        "format.class":"io.confluent.connect.hdfs.avro.AvroFormat",
        "partitioner.class":"io.confluent.connect.storage.partitioner.FieldPartitioner",
        "partition.field.name":"date"
    }
}
```

HDFS Storage

```
hduser_@ip-172-31-0-46:~$ hdfs dfs -ls /hashtags-avro/
2021-11-26 14:27:18,783 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
Found 4 items
drwxr-xr-x   - ubuntu supergroup          0 2021-11-19 14:43 /hashtags-avro/date=2021-11-19
drwxr-xr-x   - ubuntu supergroup          0 2021-11-22 15:18 /hashtags-avro/date=2021-11-22
drwxr-xr-x   - ubuntu supergroup          0 2021-11-24 16:23 /hashtags-avro/date=2021-11-24
drwxr-xr-x   - ubuntu supergroup          0 2021-11-26 14:26 /hashtags-avro/date=2021-11-26
hduser_@ip-172-31-0-46:~$
```

```
hduser_@ip-172-31-0-46:~$ hdfs dfs -ls /hashtags-avro/date=2021-11-26
2021-11-26 14:24:15,322 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes
Found 46 items
-rw-r--r--   3 ubuntu supergroup       3228 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000000+0000000029.avro
-rw-r--r--   3 ubuntu supergroup       3235 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000030+0000000059.avro
-rw-r--r--   3 ubuntu supergroup       3253 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000060+0000000089.avro
-rw-r--r--   3 ubuntu supergroup       3244 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000090+0000000119.avro
-rw-r--r--   3 ubuntu supergroup       3282 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000120+0000000149.avro
-rw-r--r--   3 ubuntu supergroup       3167 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000150+0000000179.avro
-rw-r--r--   3 ubuntu supergroup       3232 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000180+0000000209.avro
-rw-r--r--   3 ubuntu supergroup       3249 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000210+0000000239.avro
-rw-r--r--   3 ubuntu supergroup       3258 2021-11-26 13:07 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000240+0000000269.avro
-rw-r--r--   3 ubuntu supergroup       3226 2021-11-26 13:10 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000270+0000000299.avro
-rw-r--r--   3 ubuntu supergroup       3241 2021-11-26 13:10 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000300+0000000329.avro
-rw-r--r--   3 ubuntu supergroup       3257 2021-11-26 13:12 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000330+0000000359.avro
-rw-r--r--   3 ubuntu supergroup       3247 2021-11-26 13:14 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000360+0000000389.avro
-rw-r--r--   3 ubuntu supergroup       3235 2021-11-26 13:16 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000390+0000000419.avro
-rw-r--r--   3 ubuntu supergroup       3238 2021-11-26 13:18 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000420+0000000449.avro
-rw-r--r--   3 ubuntu supergroup       3268 2021-11-26 13:21 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000450+0000000479.avro
-rw-r--r--   3 ubuntu supergroup       3217 2021-11-26 13:23 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000480+0000000509.avro
-rw-r--r--   3 ubuntu supergroup       3216 2021-11-26 13:24 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000510+0000000539.avro
-rw-r--r--   3 ubuntu supergroup       3245 2021-11-26 13:26 /hashtags-avro/date=2021-11-26/hashtags-avro+0+0000000540+0000000569.avro
```

Topologie Storm dans StormUI

Storm UI

Search twitter-1-1638048257: Search Search Archived Logs:

Topology summary

Name	Id	Owner	Status	Uptime	Num workers	Num executors	Num tasks	Replication count
twitter	twitter-1-1638048257	ubuntu	ACTIVE	4m 27s	1	3	3	1

Topology actions

Activate | Deactivate | Rebalance | Kill | Debug | Stop Debug | Change Log Level

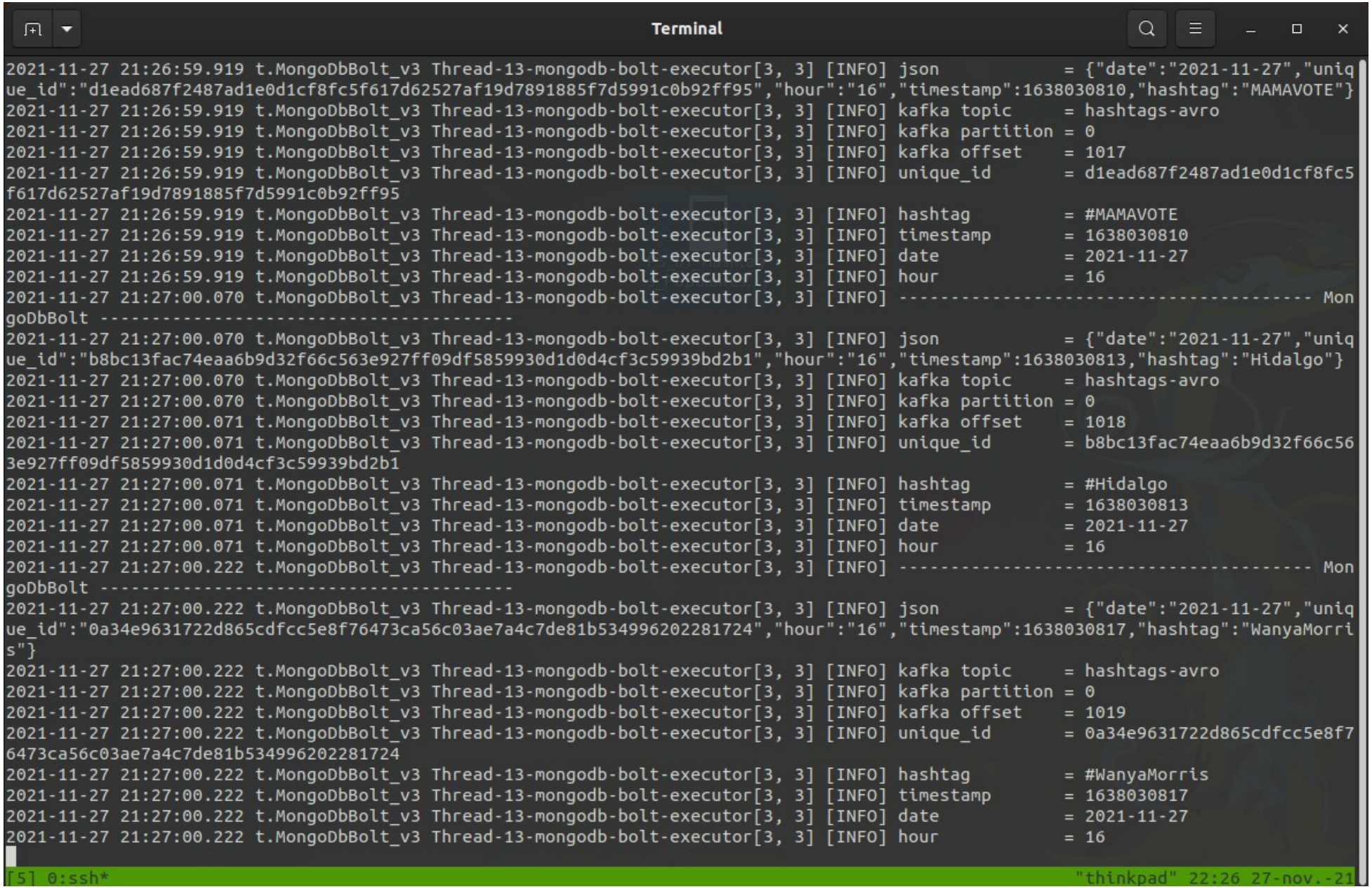
Topology stats

Window	Emitted	Transferred	Complete laten
10m 0s	5580	4000	0
3h 0m 0s	5580	4000	0
1d 0h 0m 0s	5580	4000	0
All time	5580	4000	0

Spouts (All time)

Id	Executors	Tasks	Emitted	Transferred	Complete latency (ms)
kafka-spout	1	1	4000	4000	0

Topologie Storm en Action



MongoDB Storm Collection

```
db.createCollection( "twitter.hashtags_storm",
{
   validator: { $jsonSchema: {
      bsonType: "object",
      required: [ "unique_id", "hashtag", "date", "hour" ],
      properties: {
         unique_id: {  bsonType: "string",  description: "must be a string and is required"     },
         hashtag:   {  bsonType : "string", description: "must be a string and is required"     },
         timestamp: {  bsonType : "long",   description: "must be a long and is required"       },
         date:      {  bsonType : "string", description: "must be a string and is required"     },
         hour:      {  bsonType : "string", description: "must be a string and is required"     }
      }
   } }
} )

db.twitter.hashtags_storm.createIndex( { "date": 1 } )
db.twitter.hashtags_storm.createIndex( { "hour": 1 } )
db.twitter.hashtags_storm.createIndex( { "hashtag": 1 } )
db.twitter.hashtags_storm.createIndex( { "timestamp": 1 }, { "expireAfterSeconds": 18000 } )
```

MongoDB HDFS Collection

```
db.createCollection( "twitter.hashtags_hdfs",
{
   validator: { $jsonSchema: {
      bsonType: "object",
      required: [ "date", "hour", "rank", "hashtag", "count" ],
      properties: {
         date:    { bsonType: "string",  description: "must be a string and is required"    },
         hour:    { bsonType: "string",  description: "must be a string and is required"    },
         rank:    { bsonType: "int",     description: "must be an integer and is required"  },
         hashtag: { bsonType: "string",  description: "must be a string and is required"    },
         count:   { bsonType : "int",    description: "must be an integer and is required"  }
      }
   } }
} )

db.twitter.hashtags_hdfs.createIndex( { "date": 1 } )
db.twitter.hashtags_hdfs.createIndex( { "hour": 1 } )
```

Façon d’interroger les tendances twitter

STORM

```
print('  > STORM PATH')

records = collection.aggregate([
    {"$match": {"date": args.date, "hour": args.hour}},
    {"$group": {"_id": "$hashtag", "count": {"$sum": 1}}},
    {"$sort": {"count": -1}},
    {"$limit": 10}
])
```

HDFS

```
print('  > HDFS PATH')

records = collection.find({"date": args.date, "hour": args.hour}).sort("rank", 1).limit(10)
```


Démo de l’application

Script d’Aggrégation des données HDFS

```
********************************************************************************************
[[ STARTING OpenClassRooms PROJET N°5 | CONTEXT = local ]]

No command arguments provided.
  > Processing last 3 days (_DEFAULT_PROCESSING_DAYS).
  > --from_date = 2021-11-25   --to_date = 2021-11-27

Starting to aggregate HDFS avro files for date : 2021-11-25
    hdfs://172.31.0.46:54310/hashtags-avro/date=2021-11-25/*.avro
    Not found

Starting to aggregate HDFS avro files for date : 2021-11-26
    hdfs://172.31.0.46:54310/hashtags-avro/date=2021-11-26/*.avro
  > 2021-11-26:00H      --->     Update MongoDB: No data
  > 2021-11-26:01H      --->     Update MongoDB: No data
  > 2021-11-26:02H      --->     Update MongoDB: No data
  > 2021-11-26:03H      --->     Update MongoDB: No data
  > 2021-11-26:04H      --->     Update MongoDB: No data
  > 2021-11-26:05H      --->     Update MongoDB: No data
  > 2021-11-26:06H      --->     Update MongoDB: No data
  > 2021-11-26:07H      --->     Update MongoDB: No data
  > 2021-11-26:08H      --->     Update MongoDB: No data
  > 2021-11-26:09H      --->     Update MongoDB: No data
  > 2021-11-26:10H      --->     Update MongoDB: No data
  > 2021-11-26:11H      --->     Update MongoDB: No data
  > 2021-11-26:12H      --->     Update MongoDB: Done
  > 2021-11-26:13H      --->     Update MongoDB: Done
  > 2021-11-26:14H      --->     Update MongoDB: Done
  > 2021-11-26:15H      --->     Update MongoDB: Done
  > 2021-11-26:16H      --->     Update MongoDB: Done
  > 2021-11-26:17H      --->     Update MongoDB: Done
  > 2021-11-26:18H      --->     Update MongoDB: Done
  > 2021-11-26:19H      --->     Update MongoDB: Done
  > 2021-11-26:20H      --->     Update MongoDB: Done
  > 2021-11-26:21H      --->     Update MongoDB: Done
  > 2021-11-26:22H      --->     Update MongoDB: Done
  > 2021-11-26:23H      --->     Update MongoDB: Done

Starting to aggregate HDFS avro files for date : 2021-11-27
    hdfs://172.31.0.46:54310/hashtags-avro/date=2021-11-27/*.avro
```

Script d’interrogation des #hashtags tendances

```
********************************************************************************
[[ STARTING OpenClassRooms PROJET N°5 | CONTEXT = local ]]

Command arguments provided.
  > --date = 2021-11-26   --hour = 15
  > HDFS PATH

--------------------------------------------------------------------------
RANK     HASHTAG                          COUNT      DATE            HOUR
--------------------------------------------------------------------------
1        #BlackFriday                     54         2021-11-26      15
2        #ENHYPEN                         44         2021-11-26      15
3        #ENniversary_D4                  41         2021-11-26      15
4        #TeufeurSBlackFriday             27         2021-11-26      15
5        #PetitPapaTopAchat               26         2021-11-26      15
6        #Lot2                            24         2021-11-26      15
7        #Concours                        13         2021-11-26      15
8        #MAMAVOTE                        13         2021-11-26      15
9        #Bitcoin                         9          2021-11-26      15
10       #MAMA2021                        7          2021-11-26      15
--------------------------------------------------------------------------

ubuntu@ip-172-31-2-199:~$
```


MERCI