MK

DATA PROJECTS

amazon web services™ | EC2

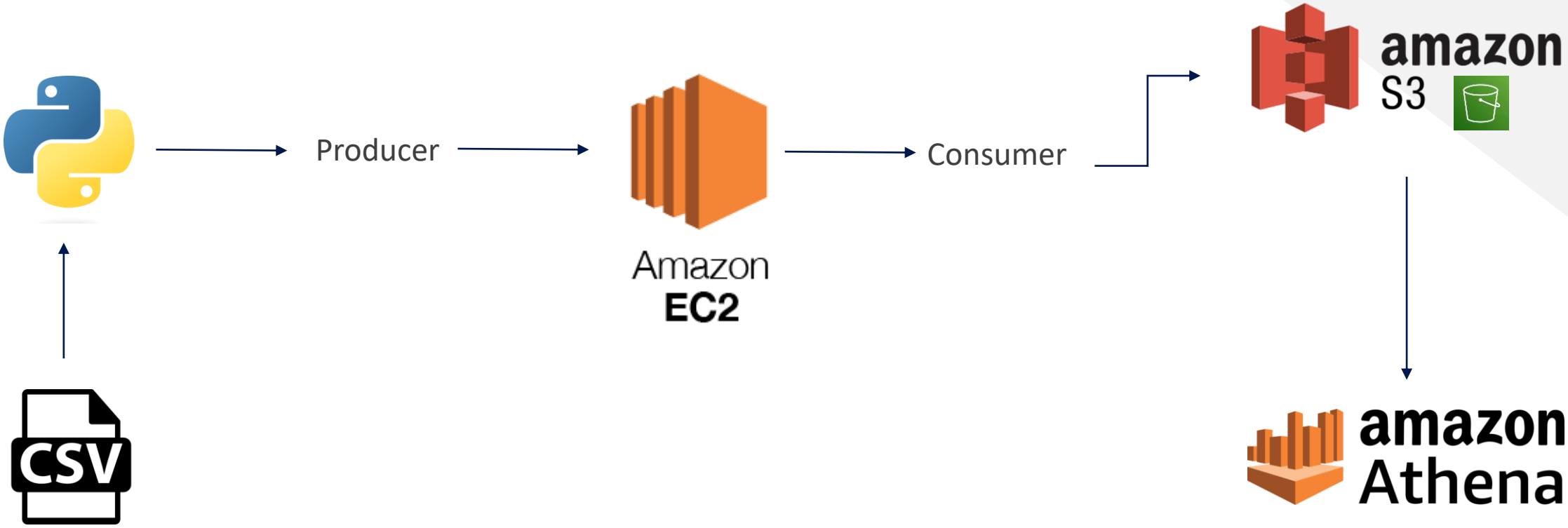
Analyse des Flux de Données de Marché avec Kafka

Du Flux Brut à la Sagesse des Données : L'Utilisation Stratégique de Kafka et d'EC2

SOMMAIRE

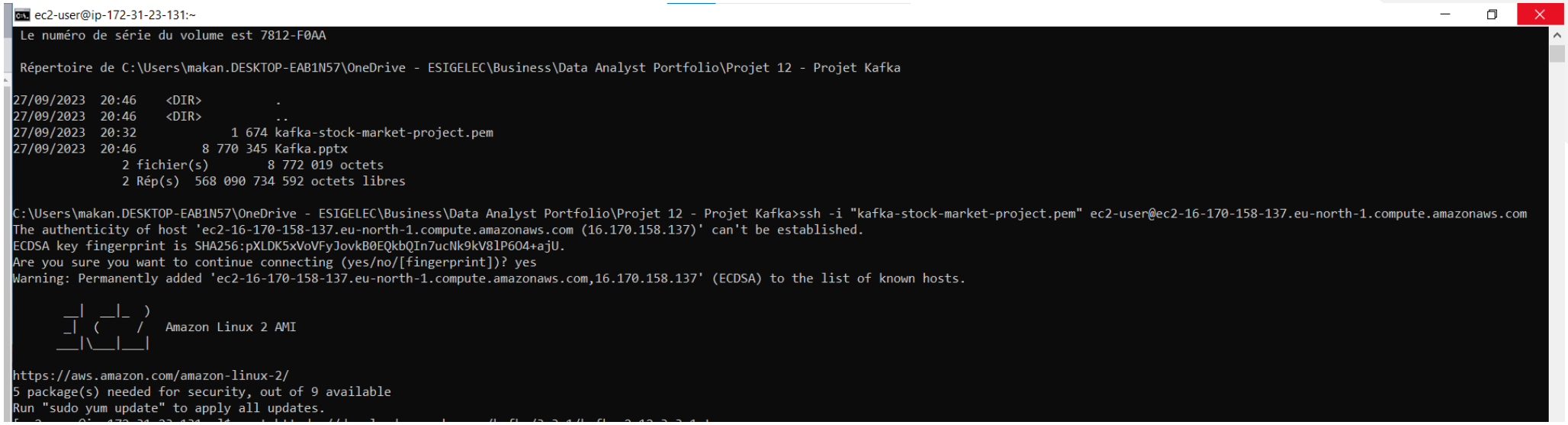
I - Mise en place de l’instance

Vue globale de l’implémentation

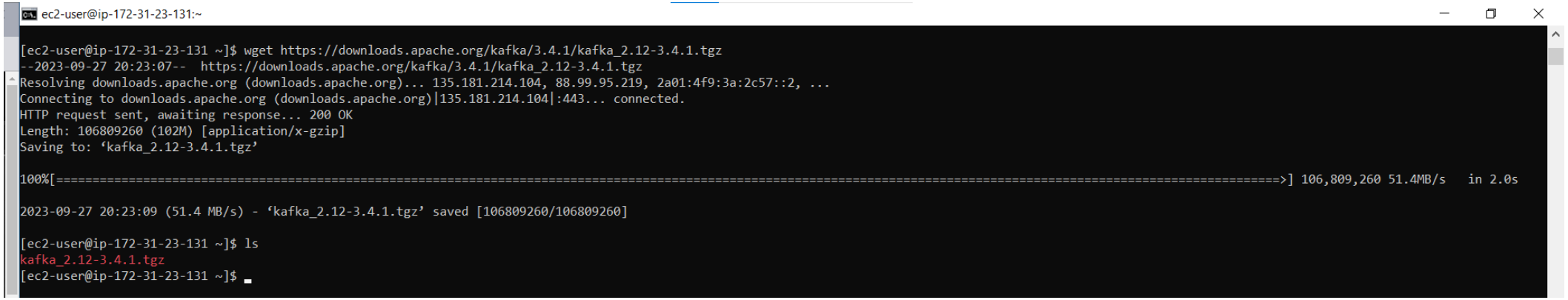


I - Mise en place de l'instance

I.1) Connexion à l'instance



I.2) Implémentation de Kafka



I - Mise en place de l'instance

I.3) Installation : Java

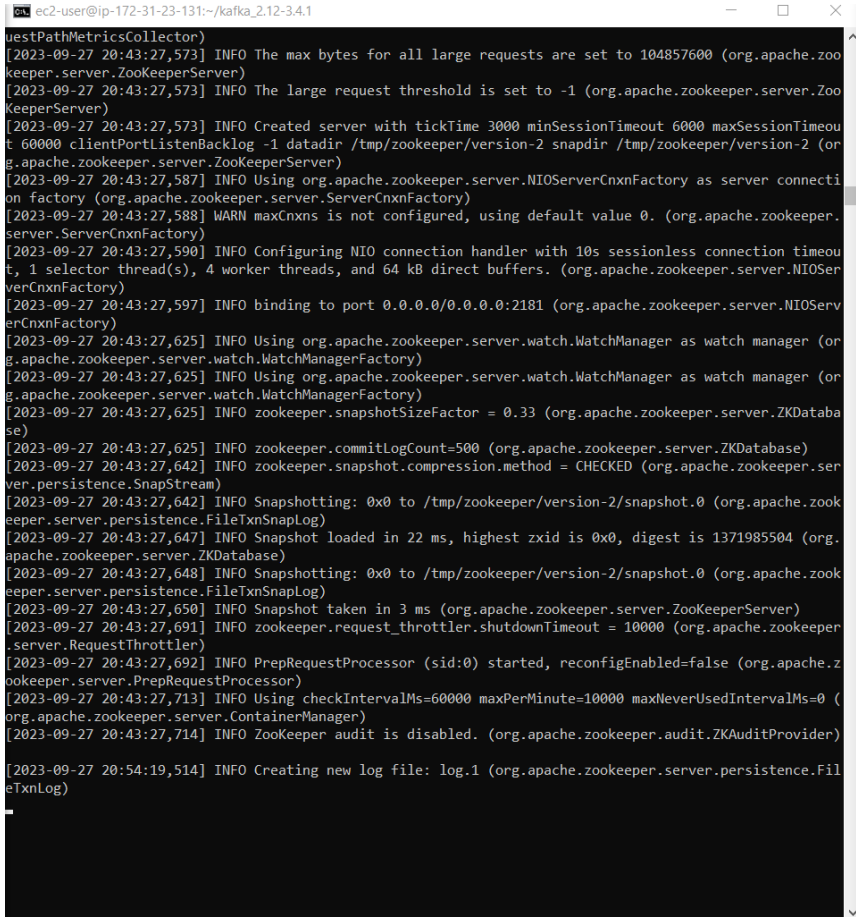
```
  Verifying  : libXfixes-5.0.3-1.amzn2.0.2.x86_64                                                                    61/64
  Verifying  : libICE-1.0.9-9.amzn2.0.2.x86_64                                                                       62/64
  Verifying  : libX11-common-1.6.7-3.amzn2.0.3.noarch                                                                63/64
  Verifying  : javapackages-tools-3.4.1-11.amzn2.noarch                                                              64/64

Installed:
  java-1.8.0-openjdk.x86_64 1:1.8.0.382.b05-1.amzn2.0.1

Dependency Installed:
  alsa-lib.x86_64 0:1.1.4.1-2.amzn2                atk.x86_64 0:2.22.0-3.amzn2.0.2                avahi-libs.x86_64 0:0.6.31-20.amzn2.0.2
```

I.4) Lancement de zookeeper

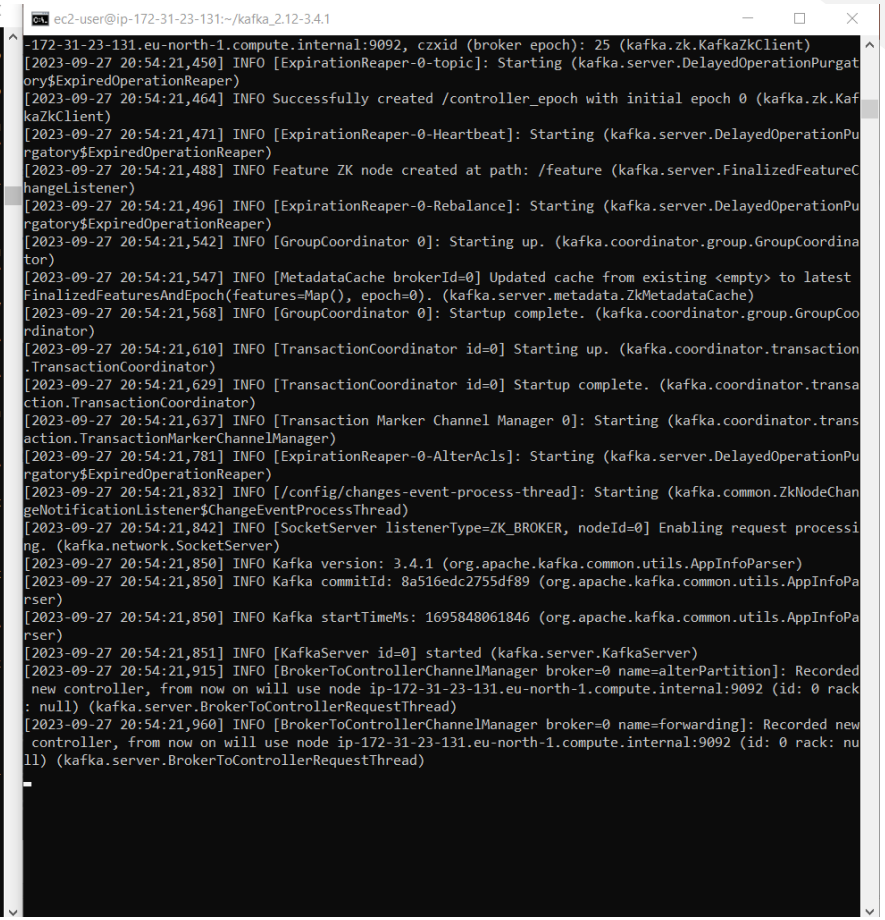
```
Sélection ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1
[2023-09-27 20:43:27,556] INFO Server environment:os.memory.total=512MB (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.enableEagerACLCheck = false (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.digest.enabled = true (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.closeSessionTxn.enabled = true (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.flushDelay=0 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.maxWriteQueuePollTime=0 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,556] INFO zookeeper.maxBatchSize=1000 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,557] INFO zookeeper.intBufferStartingSizeBytes = 1024 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,559] INFO Weighed connection throttling is disabled (org.apache.zookeeper.server.BlueThrottle)
[2023-09-27 20:43:27,561] INFO minSessionTimeout set to 6000 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,561] INFO maxSessionTimeout set to 60000 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,563] INFO Response cache size is initialized with value 400. (org.apache.zookeeper.server.ResponseCache)
[2023-09-27 20:43:27,563] INFO Response cache size is initialized with value 400. (org.apache.zookeeper.server.ResponseCache)
[2023-09-27 20:43:27,564] INFO zookeeper.pathStats.slotCapacity = 60 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,565] INFO zookeeper.pathStats.slotDuration = 15 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,565] INFO zookeeper.pathStats.maxDepth = 6 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,565] INFO zookeeper.pathStats.initialDelay = 5 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,565] INFO zookeeper.pathStats.delay = 5 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,565] INFO zookeeper.pathStats.enabled = false (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
[2023-09-27 20:43:27,573] INFO The max bytes for all large requests are set to 104857600 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,573] INFO The large request threshold is set to -1 (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,573] INFO Created server with tickTime 3000 minSessionTimeout 6000 maxSessionTimeout 60000 clientPortListenBacklog -1 datadir /tmp/zookeeper/version-2 snapdir /tmp/zookeeper/version-2 (org.a
pache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,587] INFO Using org.apache.zookeeper.server.NIOServerCnxnFactory as server connection factory (org.apache.zookeeper.server.ServerCnxnFactory)
[2023-09-27 20:43:27,588] WARN maxCnxns is not configured, using default value 0. (org.apache.zookeeper.server.ServerCnxnFactory)
[2023-09-27 20:43:27,590] INFO Configuring NIO connection handler with 10s sessionless connection timeout, 1 selector thread(s), 4 worker threads, and 64 kB direct buffers. (org.apache.zookeeper.server.NIOServer
CnxnFactory)
[2023-09-27 20:43:27,597] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.zookeeper.server.NIOServerCnxnFactory)
[2023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
[2023-09-27 20:43:27,625] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache.zookeeper.server.ZKDatabase)
[2023-09-27 20:43:27,625] INFO zookeeper.commitLogCount=500 (org.apache.zookeeper.server.ZKDatabase)
[2023-09-27 20:43:27,642] INFO zookeeper.snapshot.compression.method = CHECKED (org.apache.zookeeper.server.persistence.SnapStream)
[2023-09-27 20:43:27,642] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2023-09-27 20:43:27,647] INFO Snapshot loaded in 22 ms, highest zxid is 0x0, digest is 1371985504 (org.apache.zookeeper.server.ZKDatabase)
[2023-09-27 20:43:27,648] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2023-09-27 20:43:27,650] INFO Snapshot taken in 3 ms (org.apache.zookeeper.server.ZooKeeperServer)
[2023-09-27 20:43:27,691] INFO zookeeper.request_throttler.shutdownTimeout = 10000 (org.apache.zookeeper.server.RequestThrottler)
[2023-09-27 20:43:27,692] INFO PrepRequestProcessor (sid:0) started, reconfigEnabled=false (org.apache.zookeeper.server.PrepRequestProcessor)
[2023-09-27 20:43:27,713] INFO Using checkIntervalMs=60000 maxPerMinute=10000 maxNeverUsedIntervalMs=0 (org.apache.zookeeper.server.ContainerManager)
[2023-09-27 20:43:27,714] INFO ZooKeeper audit is disabled. (org.apache.zookeeper.audit.ZKAuditProvider)
```

I - Mise en place de l'instance

I.5) Zookeeper et Kafka



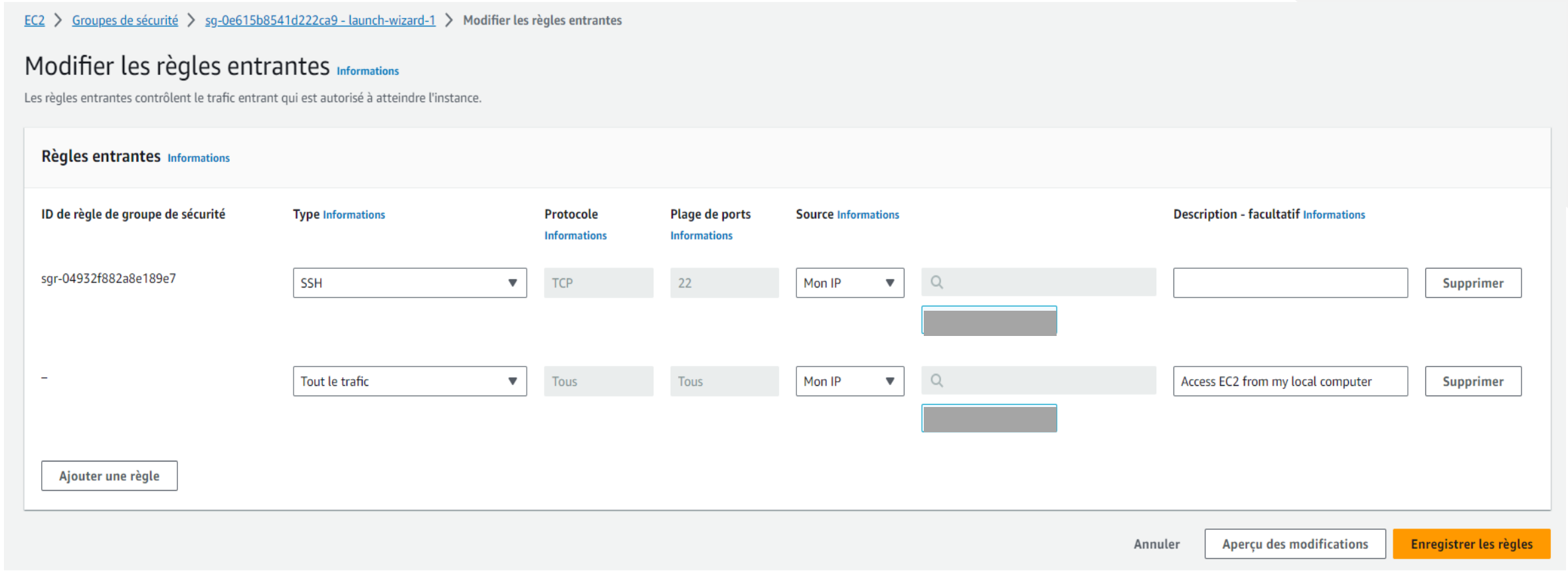
zookeeper



kafka

I - Mise en place de l’instance

I.6) Exemple de création de règles pour autoriser le Traffic

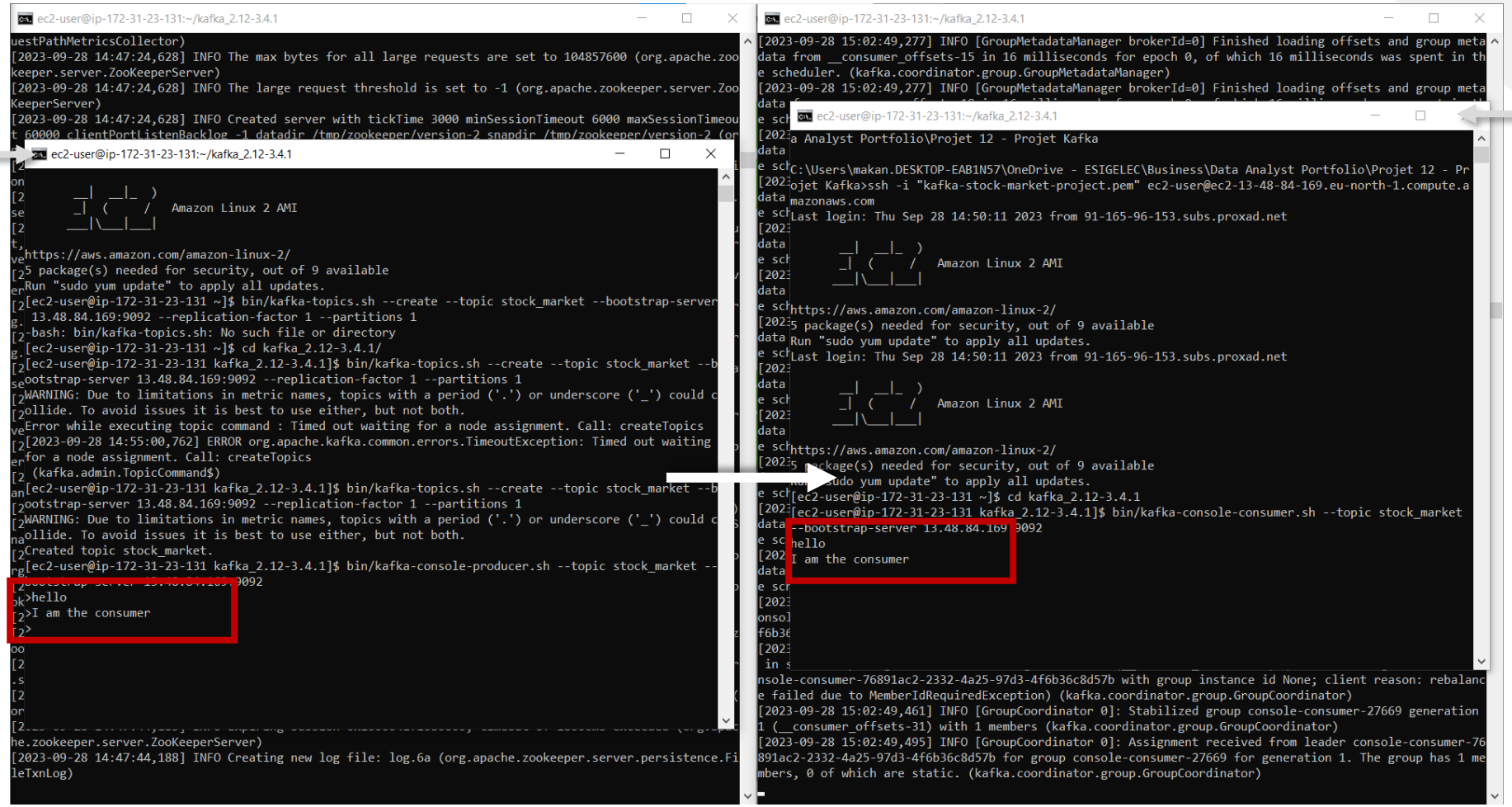


I - Mise en place de l'instance

I.7) Lancement du Producer et du consumer en CLI

Producer

Consumer



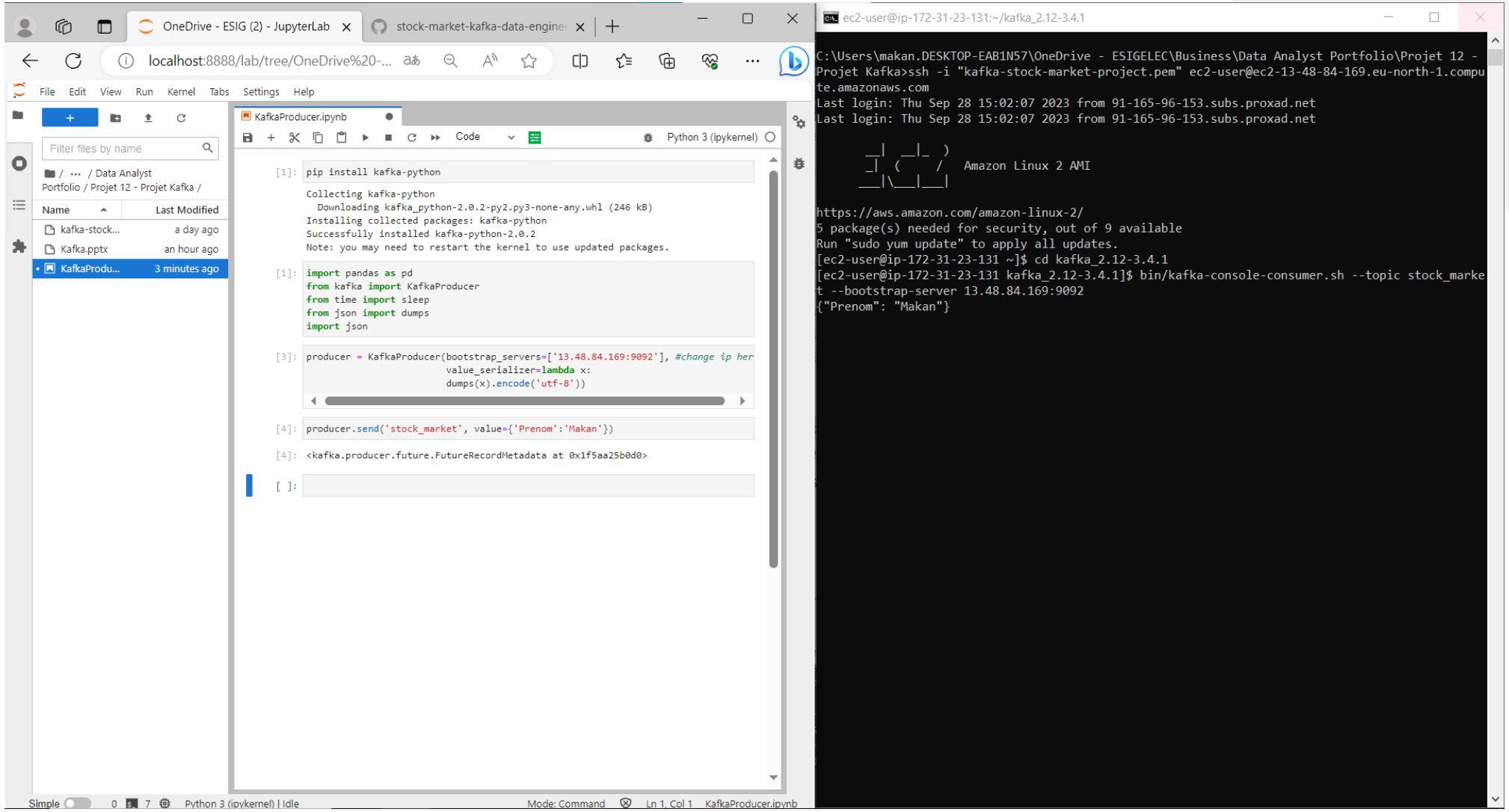
zookeeper

kafka

II - Implémentation de scripts python pour l’automatisation des processus

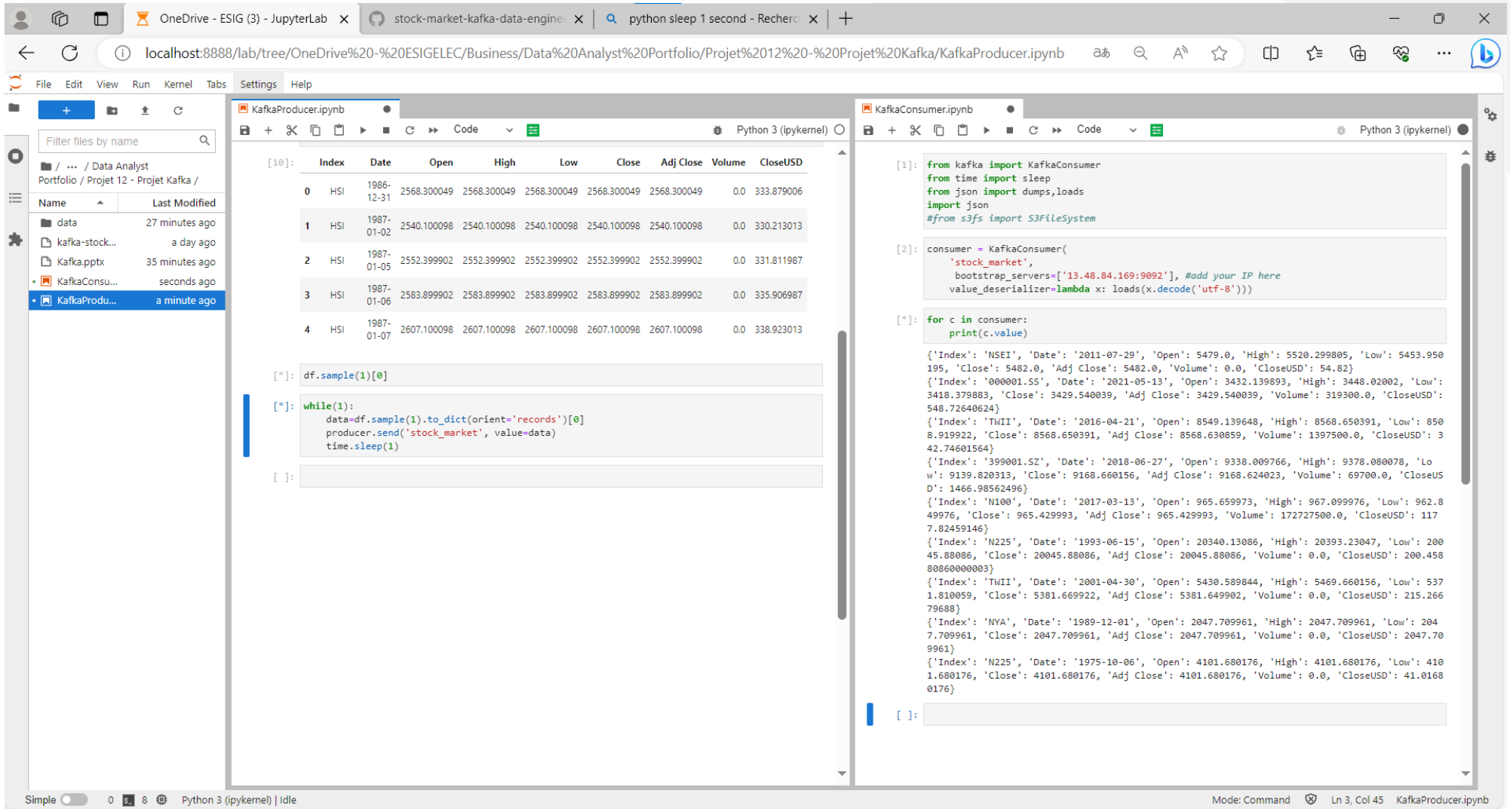
II - Implémentation de scripts python

II.1) Création d'un Producer python



II - Implémentation de scripts python

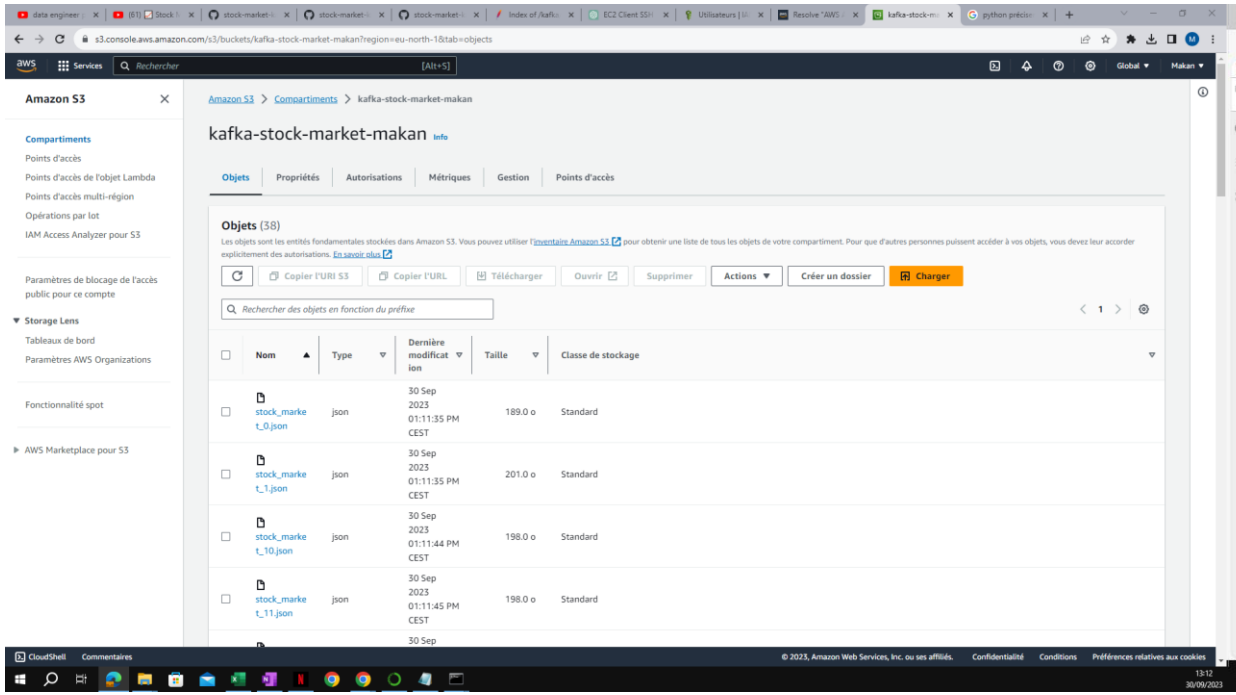
II.2) Simulation d'ingestion en temps réel et Création d'un Consumer python

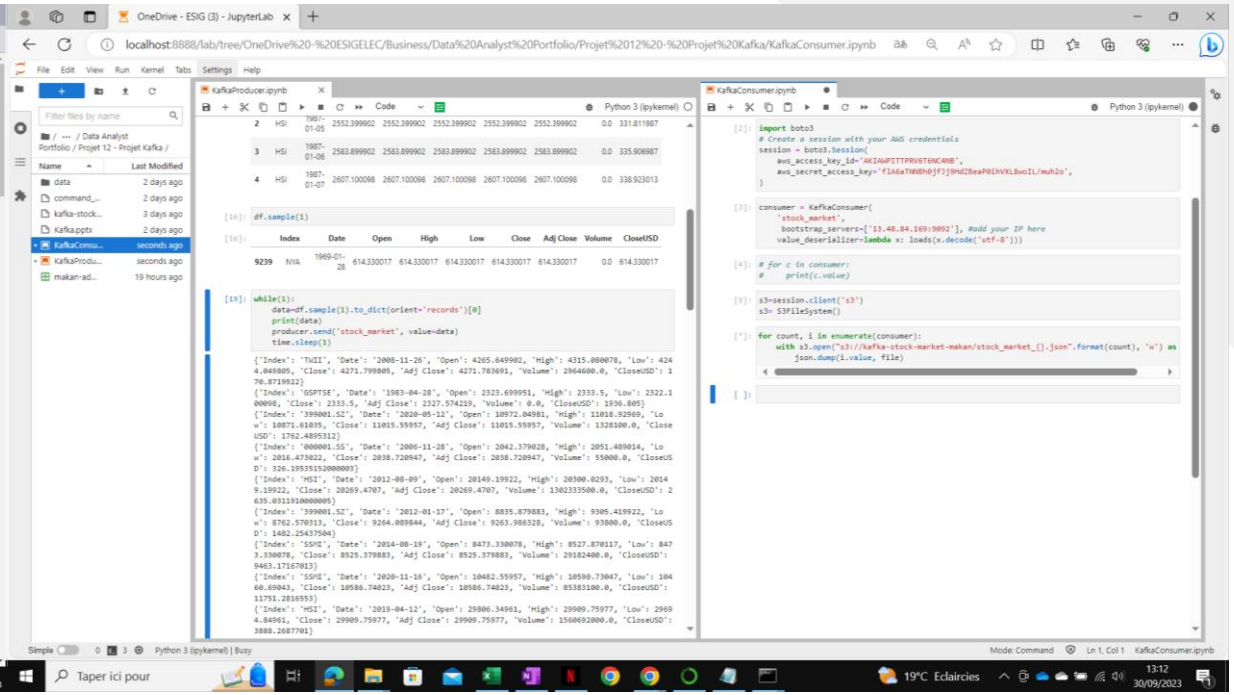


III – Accès aux données

III – Accès aux données

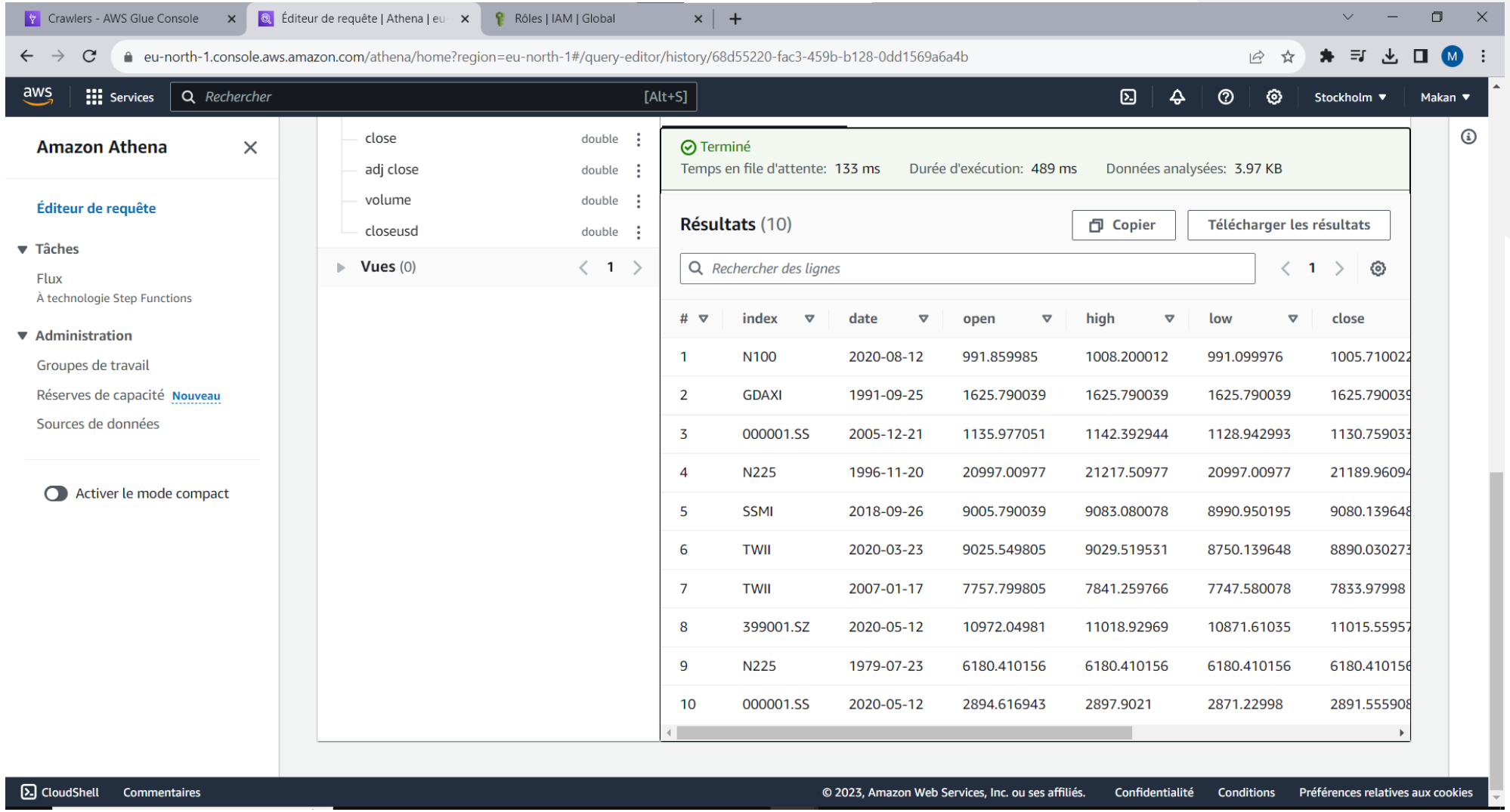
III.1) Création d’un compartiment de données (« Bucket ») pour le stockage des données issues du consumer





III – Accès aux données

III.2) Requêtes Athena pour la lecture de données



BILAN

FR

- **Objectif atteint**
  - Création réussie d'un flux de données en temps réel à partir d'un fichier CSV
  - Simulation efficace grâce à l'utilisation d'un DataFrame
- **Technologies clés**
  - Kafka, Amazon EC2, Crawler, Bucket, ZooKeeper, Producer, Consumer
- **Résultats majeurs**
  - Flux de données fonctionnel en temps réel
  - Intégration harmonieuse de l'écosystème technologique
- **Avantages significatifs**
  - Analyse de données plus réactive
  - Optimisation des ressources cloud avec Amazon EC2
  - Flexibilité dans la gestion des flux de données

BILAN

FR

- **Perspectives et Innovations**
  - Expansion des sources de données
  - Intégration future de l'apprentissage automatique
  - Développement de tableaux de bord en temps réel
- **Conclusion générale**
  - Réussite dans la réalisation d'un flux de données en temps réel
  - Adaptation aux défis technologiques
  - Prêt pour évoluer en réponse aux besoins futurs