



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



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



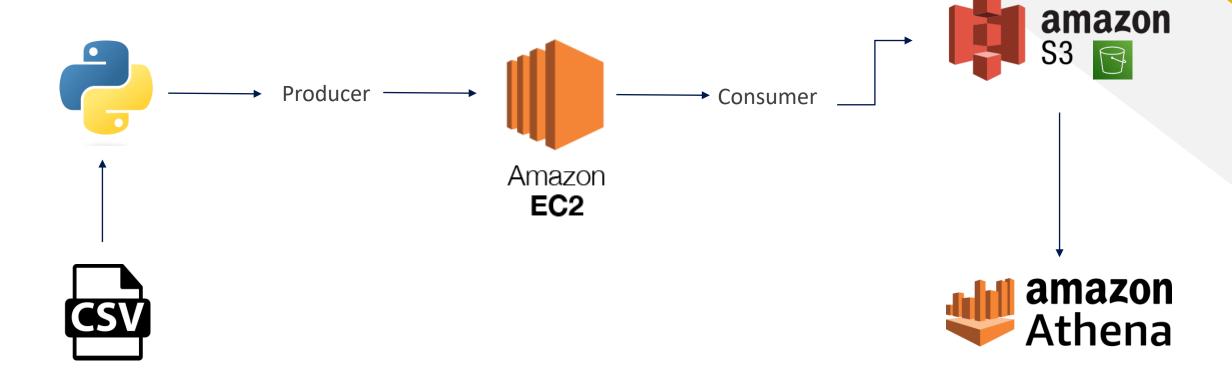
III – Accès aux données

BILAN



Vue globale de l'implémentation







I.1) Connexion à l'instance

```
ec2-user@ip-172-31-23-131:~
                                                                                                                                                                                                        Le numéro de série du volume est 7812-F0AA
 Répertoire de C:\Users\makan.DESKTOP-EAB1N57\OneDrive - ESIGELEC\Business\Data Analyst Portfolio\Projet 12 - Projet Kafka
27/09/2023 20:46 <DIR>
27/09/2023 20:46
27/09/2023 20:32
                             1 674 kafka-stock-market-project.pem
27/09/2023 20:46
                         8 770 345 Kafka.pptx
                              8 772 019 octets
              2 fichier(s)
              2 Rép(s) 568 090 734 592 octets libres
 C:\Users\makan.DESKTOP-EAB1N57\OneDrive - ESIGELEC\Business\Data Analyst Portfolio\Projet 12 - Projet Kafka>ssh -i "kafka-stock-market-project.pem" ec2-user@ec2-16-170-158-137.eu-north-1.compute.amazonaws.com
The authenticity of host 'ec2-16-170-158-137.eu-north-1.compute.amazonaws.com (16.170.158.137)' can't be established.
ECDSA key fingerprint is SHA256:pXLDK5xVoVFyJovkB0E0kb0In7ucNk9kV8lP604+ajU.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
 Warning: Permanently added 'ec2-16-170-158-137.eu-north-1.compute.amazonaws.com,16.170.158.137' (ECDSA) to the list of known hosts.
                    Amazon Linux 2 AMI
 nttps://aws.amazon.com/amazon-linux-2/
 package(s) needed for security, out of 9 available
    "sudo yum update" to apply all updates.
```

I.2) Implémentation de Kafka



I.3) Installation: Java

```
Verifying : 1:110g1VNu-g1x-1.0.1-0.1.g1c50dale5.dm2n2.0.1.X86_64

Verifying : libXfixes-5.0.3-1.amzn2.0.2.x86_64

Verifying : libX11-common-1.6.7-3.amzn2.0.3.noarch

Verifying : javapackages-tools-3.4.1-11.amzn2.noarch

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
    023-09-27 20:43:27,556] INFO Server environment:os.memory.total=512MB (org.apache.zookeeper.server.ZooKeepe
    023-09-27 20:43:27,556] INFO zookeeper.enableEagerACLCheck = false (org.apache.zookeeper.server.ZooKeeperServer)
   023-09-27 20:43:27,556] INFO zookeeper.digest.enabled = true (org.apache.zookeeper.server.ZooKeeperServer)
  023-09-27 20:43:27,556] INFO zookeeper.closeSessionTxm.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
   1923-09-27 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 20:43:27,503] INFO maxSessionTimeout set to 60000 (org.apache.zookeeper.server.ZookeeperServer)
2023-09-27 20:43:27,503] INFO mexpessionTimeout set to 50000 (org.apache.zookeeper.server.ResponseCache)
2023-09-27 20:43:27,503] INFO Response cache size is initialized with value 400. (org.apache.zookeeper.server.ResponseCache)
2023-09-27 20:43:27,503] INFO Response cache size is initialized with value 400. (org.apache.zookeeper.server.ResponseCache)
   023-09-27 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
   023-09-27 20:43:27,565] INFO zookeeper.pathStats.delay = 5 (org.apache.zookeeper.server.util.RequestPathMetricsCollector)
   023-09-27 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)
   1923-09-27 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.
  2023-09-27 20:43:27,587] INFO Using org.apache.zookeeper.server.NIOServerCnxnFactory as server connection factory (org.apache.zookeeper.server.ServerCnxnFactory)
   023-09-27 20:43:27,588] WARN maxCnxns is not configured, using default value 0. (org.apache.zookeeper.server.ServerCnxnFactory)
   023-09-27 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.NICOServe
  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)
   023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
   023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.WatchManager as watch manager (org.apache.zookeeper.server.watch.WatchManagerFactory)
   1923-09-27 20:43:27,625] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache.zookeeper.server.ZKDatabase)
   023-09-27 20:43:27,625] INFO zookeeper.commitLogCount=500 (org.apache.zookeeper.server.ZKDatabase)
   023-09-27 20:43:27,642] INFO zookeeper.snapshot.compression.method = CHECKED (org.apache.zookeeper.server.persistence.SnapStream)
    023-09-27 20:43:27,642] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
   023-09-27 20:43:27,647] INFO Snapshot loaded in 22 ms, highest zxid is 0x0, digest is 1371985504 (org.apache.zookeeper.server.ZKDatabase)
   023-09-27 20:43:27,648] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2/snapshot.0 (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
   1923-09-27 20:43:27,650] INFO Snapshot taken in 3 ms (org.apache.zookeeper.server.ZooKeeperServer)
   1923-09-27 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)
   023-09-27 20:43:27,713] INFO Using checkIntervalMs=60000 maxPerMinute=10000 maxNeverUsedIntervalMs=0 (org.apache.zookeeper.server.ContainerManager)
  1023-09-27 20:43:27,713] INFO ZooKeeper audit is disabled. (org.apache.zookeeper.audit.ZKAuditProvider)
```

MK

I.5) Zookeeper et Kafka

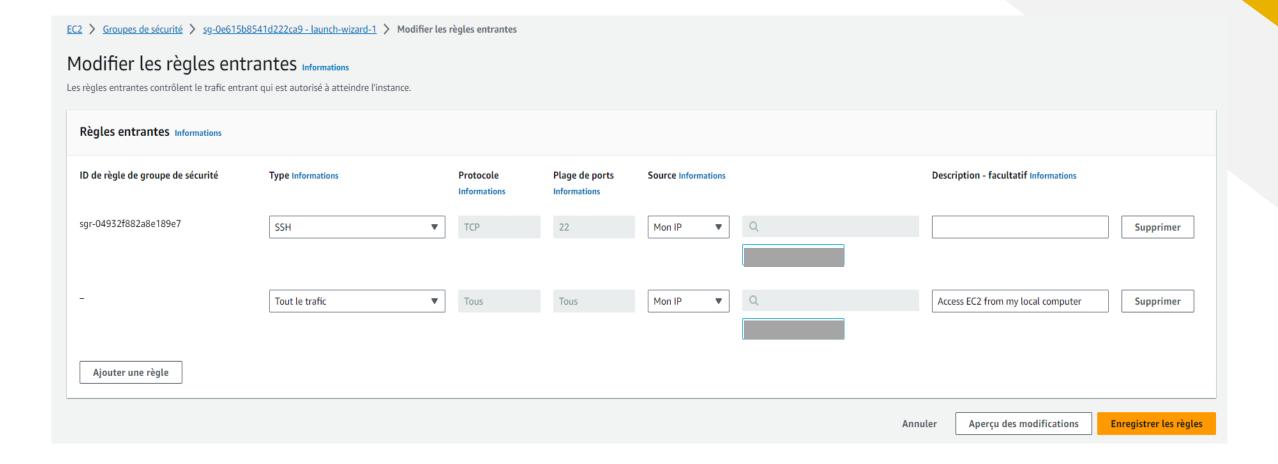
ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1	- □ ×	ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1	- 🗆 ×
uestPathMetricsCollector)		^ -172-31-23-131.eu-north-1.compute.internal:9092, czxid (broker epoch): 25 (kafka zk KafkaZkCliont)
[2023-09-27 20:43:27,573] INFO The max bytes for all large requests are set	to 104857600 (org.apache.zoo	[2023-09-27 20:54:21,450] INFO [ExpirationReaper-0-topic]: Starting (kafka.	
keeper.server.ZooKeeperServer)		ory\$ExpiredOperationReaper)	ve. ve. ibezaj edope. dezom di gae
[2023-09-27 20:43:27,573] INFO The large request threshold is set to -1 (org	.apache.zookeeper.server.Zoo	[2023-09-27 20:54:21,464] INFO Successfully created /controller_epoch with	initial epoch 0 (kafka.zk.Kaf
KeeperServer) [2023-09-27 20:43:27,573] INFO Created server with tickTime 3000 minSessionT	imagut 6000 mayEassianTimagu	kaZkClient)	
t 60000 clientPortListenBacklog -1 datadir /tmp/zookeeper/version-2 snapdir		[2023-09-27 20:54:21,471] INFO [ExpirationReaper-0-Heartbeat]: Starting (ka	fka.server.DelayedOperationPu
g.apache.zookeeper.server.ZooKeeperServer)	/ cmp/ 2000(ccpc) / vci 31011 2 (0)	rgatory\$ExpiredOperationReaper)	
[2023-09-27 20:43:27,587] INFO Using org.apache.zookeeper.server.NIOServerCn	xnFactory as server connecti	[2023-09-27 20:54:21,488] INFO Feature ZK node created at path: /feature (k	atka.server.FinalizedFeatureC
on factory (org.apache.zookeeper.server.ServerCnxnFactory)		[2023-09-27 20:54:21,496] INFO [ExpirationReaper-0-Rebalance]: Starting (ka	fka server DelavedOnerationPu
[2023-09-27 20:43:27,588] WARN maxCnxns is not configured, using default val	ue 0. (org.apache.zookeeper.	rgatory\$ExpiredOperationReaper)	ka. Ser ver ibelageaoper actom a
server.ServerCnxnFactory)		[2023-09-27 20:54:21,542] INFO [GroupCoordinator 0]: Starting up. (kafka.co	ordinator.group.GroupCoordina
[2023-09-27 20:43:27,590] INFO Configuring NIO connection handler with 10s st, 1 selector thread(s), 4 worker threads, and 64 kB direct buffers. (org.ap		tor)	
verCnxnFactory)	ache.200keeper.server.n103er	[2023-09-27 20:54:21,547] INFO [MetadataCache brokerId=0] Updated cache fro	
[2023-09-27 20:43:27,597] INFO binding to port 0.0.0.0/0.0.0:2181 (org.apa	che.zookeeper.server.NIOServ	FinalizedFeaturesAndEpoch(features=Map(), epoch=0). (kafka.server.metadata.	
erCnxnFactory)		[2023-09-27 20:54:21,568] INFO [GroupCoordinator 0]: Startup complete. (kaf	ka.coordinator.group.GroupCoo
[2023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.Watch	Manager as watch manager (or	[2023-09-27 20:54:21,610] INFO [TransactionCoordinator id=0] Starting up. (kafka coordinator transaction
g.apache.zookeeper.server.watch.WatchManagerFactory)		.TransactionCoordinator)	karka.eoor ainacor.er ansaccion
[2023-09-27 20:43:27,625] INFO Using org.apache.zookeeper.server.watch.Watch	Manager as watch manager (or	[2023-09-27 20:54:21,629] INFO [TransactionCoordinator id=0] Startup comple	te. (kafka.coordinator.transa
<pre>g.apache.zookeeper.server.watch.WatchManagerFactory) [2023-09-27 20:43:27,625] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apac</pre>	ha zaakaanan camuan 7KDataha	ction.TransactionCoordinator)	
[2023-09-27 20.43.27,023] INFO 200Keeper.Shapshot312eractor = 0.33 (org.apac	ne.zookeeper.server.zkbacaba	[2023-09-27 20:54:21,637] INFO [Transaction Marker Channel Manager 0]: Star	ting (kafka.coordinator.trans
[2023-09-27 20:43:27,625] INFO zookeeper.commitLogCount=500 (org.apache.zook	eeper.server.ZKDatabase)	action.TransactionMarkerChannelManager)	Cl
[2023-09-27 20:43:27,642] INFO zookeeper.snapshot.compression.method = CHECK		<pre>[2023-09-27 20:54:21,781] INFO [ExpirationReaper-0-AlterAcls]: Starting (ka rgatory\$ExpiredOperationReaper)</pre>	rka.server.DelayedOperationPu
ver.persistence.SnapStream)		[2023-09-27 20:54:21,832] INFO [/config/changes-event-process-thread]: Star	ting (kafka common 7kNodeChan
[2023-09-27 20:43:27,642] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2	/snapshot.0 (org.apache.zook	geNotificationListener\$ChangeEventProcessThread)	erig (narmaresimionremeacemen
eeper.server.persistence.FileTxnSnapLog)	1: 1: 4374005504 ([2023-09-27 20:54:21,842] INFO [SocketServer listenerType=ZK_BROKER, nodeId	=0] Enabling request processi
[2023-09-27 20:43:27,647] INFO Snapshot loaded in 22 ms, highest zxid is 0x0 apache.zookeeper.server.ZKDatabase)	, digest is 13/1985504 (org.	ng. (kafka.network.SocketServer)	
[2023-09-27 20:43:27,648] INFO Snapshotting: 0x0 to /tmp/zookeeper/version-2	/snapshot.0 (org.apache.zook	[2023-09-27 20:54:21,850] INFO Kafka version: 3.4.1 (org.apache.kafka.commo	
eeper.server.persistence.FileTxnSnapLog)	, snapsnocto (oi grapachic zook	[2023-09-27 20:54:21,850] INFO Kafka commitId: 8a516edc2755df89 (org.apache	.kafka.common.utils.AppInfoPa
[2023-09-27 20:43:27,650] INFO Snapshot taken in 3 ms (org.apache.zookeeper.	server.ZooKeeperServer)	rser) [2023-09-27 20:54:21,850] INFO Kafka startTimeMs: 1695848061846 (org.apache	kafka common utils AppInfoRa
<pre>[2023-09-27 20:43:27,691] INFO zookeeper.request_throttler.shutdownTimeout =</pre>	: 10000 (org.apache.zookeeper	rser)	. Karka. common. uciis. Appiniora
.server.RequestThrottler)	- 13 163 7	[2023-09-27 20:54:21,851] INFO [KafkaServer id=0] started (kafka.server.Kaf	kaServer)
[2023-09-27 20:43:27,692] INFO PrepRequestProcessor (sid:0) started, reconficokeeper.server.PrepRequestProcessor)	gEnabled=false (org.apache.z	[2023-09-27 20:54:21,915] INFO [BrokerToControllerChannelManager broker=0 n	
[2023-09-27 20:43:27,713] INFO Using checkIntervalMs=60000 maxPerMinute=1000	0 mayNeverUsedIntervalMs-0 (new controller, from now on will use node ip-172-31-23-131.eu-north-1.comp	ute.internal:9092 (id: 0 rack
org.apache.zookeeper.server.ContainerManager)	o maxilever oscalireer valids=0 (: null) (kafka.server.BrokerToControllerRequestThread)	6 11 3 8 1 1
[2023-09-27 20:43:27,714] INFO ZooKeeper audit is disabled. (org.apache.zook	eeper.audit.ZKAuditProvider)	[2023-09-27 20:54:21,960] INFO [BrokerToControllerChannelManager broker=0 n controller, from now on will use node ip-172-31-23-131.eu-north-1.compute.	
		ll) (kafka.server.BrokerToControllerRequestThread)	Internal:9092 (Id: 0 rack: Nu
[2023-09-27 20:54:19,514] INFO Creating new log file: log.1 (org.apache.zook	eeper.server.persistence.Fil	11) (Karka. Ser Ver . Broker rocorter offer hequestrin eau)	
eTxnLog)			
•			
		v ·	V

zookeeper

kafka



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



ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1

he.zookeeper.server.ZooKeeperServer)

Producer

MK

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

uestPathMetricsCollector) 2023-09-28 15:02:49,277] INFO [GroupMetadataManager brokerId=0] Finished loading offsets and group met 2023-09-28 14:47:24.628] INFO The max bytes for all large requests are set to 104857600 (org.apache.zoo data from — consumer offsets-15 in 16 milliseconds for epoch 0, of which 16 milliseconds was spent in t keeper.server.ZooKeeperServer) scheduler. (kafka.coordinator.group.GroupMetadataManager) [2023-09-28 14:47:24,628] INFO The large request threshold is set to -1 (org.apache.zookeeper.server.Zoo 2023-09-28 15:02:49,277] INFO [GroupMetadataManager brokerId=0] Finished loading offsets and group meta ec2-user@ip-172-31-23-131:~/kafka 2.12-3.4.1 [2023-09-28 14:47:24,628] INFO Created server with tickTime 3000 minSessionTimeout 6000 maxSessionTimeou 2023a Analyst Portfolio\Projet 12 - Projet Kafka **■** ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1 schc:\Users\makan.DESKTOP-EAB1N57\OneDrive - ESIGELEC\Business\Data Analyst Portfolio\Projet 12 - Po ²⁰²³ojet Kafka>ssh -i "kafka-stock-market-project.pem" ec2-user@ec2-13-48-84-169.eu-north-1.compute.a schLast login: Thu Sep 28 14:50:11 2023 from 91-165-96-153.subs.proxad.net https://aws.amazon.com/amazon-linux-2/ Amazon Linux 2 AMI 5 package(s) needed for security, out of 9 available Run "sudo yum update" to apply all updates. [ec2-user@ip-172-31-23-131 ~]\$ bin/kafka-topics.sh --create --topic stock market --bootstrap-server schhttps://aws.amazon.com/amazon-linux-2/ 13.48.84.169:9092 --replication-factor 1 --partitions 1 2023 package(s) needed for security, out of 9 available -bash: bin/kafka-topics.sh: No such file or directory ata Run "sudo yum update" to apply all updates. [ec2-user@ip-172-31-23-131 ~]\$ cd kafka 2.12-3.4.1/ schLast login: Thu Sep 28 14:50:11 2023 from 91-165-96-153.subs.proxad.net [ec2-user@ip-172-31-23-131 kafka_2.12-3.4.1]\$ bin/kafka-topics.sh --create --topic stock_market --b ootstrap-server 13.48.84.169:9092 --replication-factor 1 --partitions 1 ,WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could o Amazon Linux 2 AMI collide. To avoid issues it is best to use either, but not both. Error while executing topic command : Timed out waiting for a node assignment. Call: createTopics [2023-09-28 14:55:00,762] ERROR org.apache.kafka.common.errors.TimeoutException: Timed out waiting schhttps://aws.amazon.com/amazon-linux-2/ for a node assignment. Call: createTopics 2023_{5 reckage(s)} needed for security, out of 9 available (kafka.admin.TopicCommand\$) sudo yum update" to apply all updates. [ec2-user@ip-172-31-23-131 kafka_2.12-3.4.1]\$ bin/kafka-topics.sh --create --topic stock market sch[ec2-user@ip-172-31-23-131 ~]\$ cd kafka 2.12-3.4.1 ootstrap-server 13.48.84.169:9092 --replication-factor 1 --partitions 1 2023 <u>ec2-user@ip-172-31-23-131 kafka 2</u>.12-3.4.1]\$ bin/kafka-console-consumer.sh --topic stock_market WARNING: Due to limitations in metric names, topics with a period ('.') or underscore (' ') could --bootstrap-server 13.48.84.169 9092 ollide. To avoid issues it is best to use either, but not both. nello Created topic stock market. am the consumer [ec2-user@ip-172-31-23-131 kafka_2.12-3.4.1]\$ bin/kafka-console-producer.sh --topic stock market >I am the consumer sole-consumer-76891ac2-2332-4a25-97d3-4f6b36c8d57b with group instance id None; client reason: rebalanc failed due to MemberIdRequiredException) (kafka.coordinator.group.GroupCoordinator) 2023-09-28 15:02:49,461] INFO [GroupCoordinator 0]: Stabilized group console-consumer-27669 generation (__consumer_offsets-31) with 1 members (kafka.coordinator.group.GroupCoordinator)

ec2-user@ip-172-31-23-131:~/kafka_2.12-3.4.1

Consumer

[2023-09-28 14:47:44,188] INFO Creating new log file: log.6a (org.apache.zookeeper.server.persistence.Fi

bers, 0 of which are static. (kafka.coordinator.group.GroupCoordinator)

[2023-09-28 15:02:49,495] INFO [GroupCoordinator 0]: Assignment received from leader console-consumer-7

391ac2-2332-4a25-97d3-4f6b36c8d57b for group console-consumer-27669 for generation 1. The group has 1 me

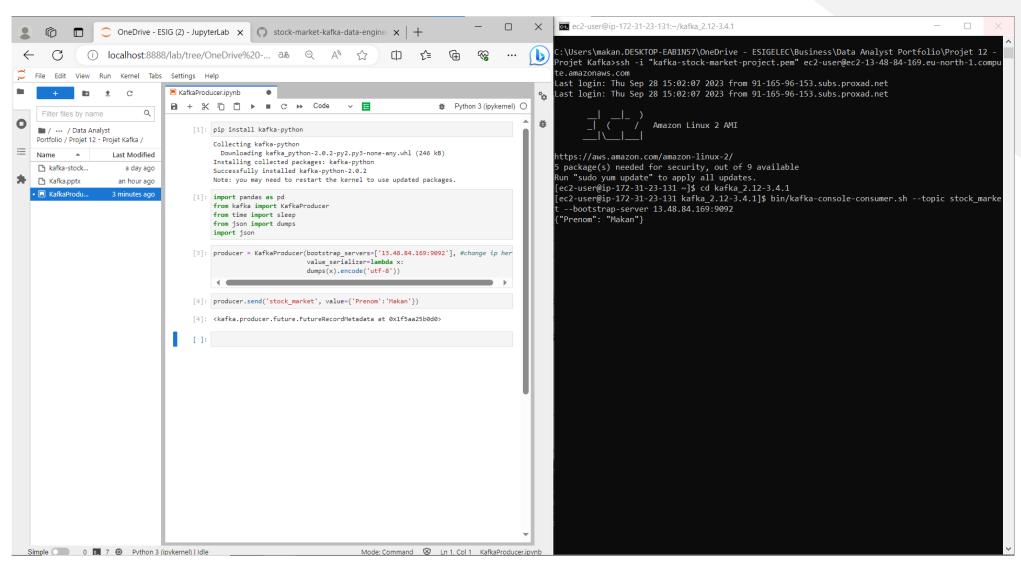


II - Implémentation de scripts python pour l'automatisation des processus

II - Implémentation de scripts python

MK

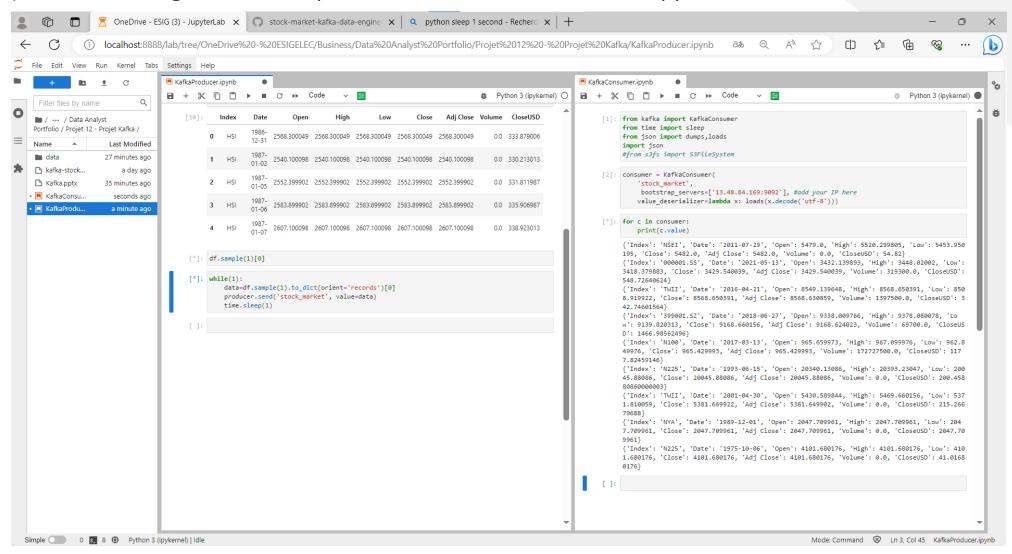
II.1) Création d'un Producer python



II - Implémentation de scripts python

MK

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



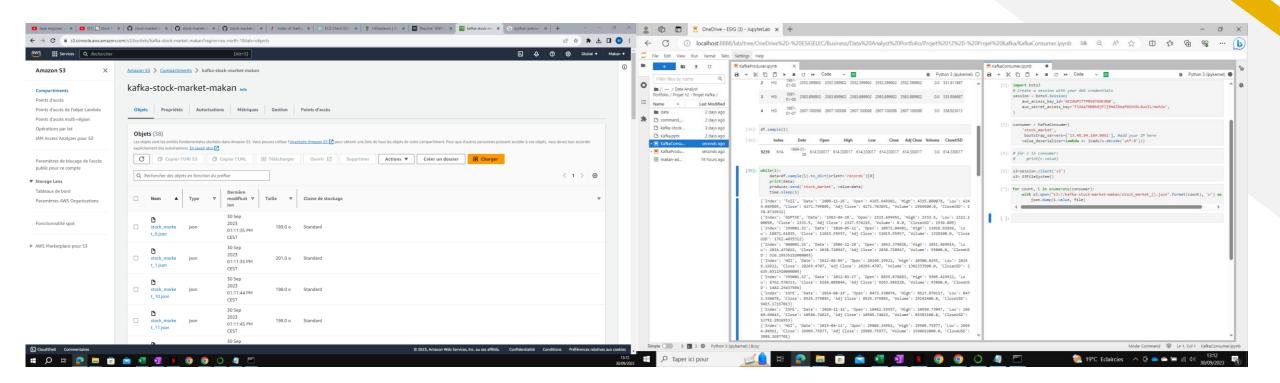


III – Accès aux données

III – Accès aux données



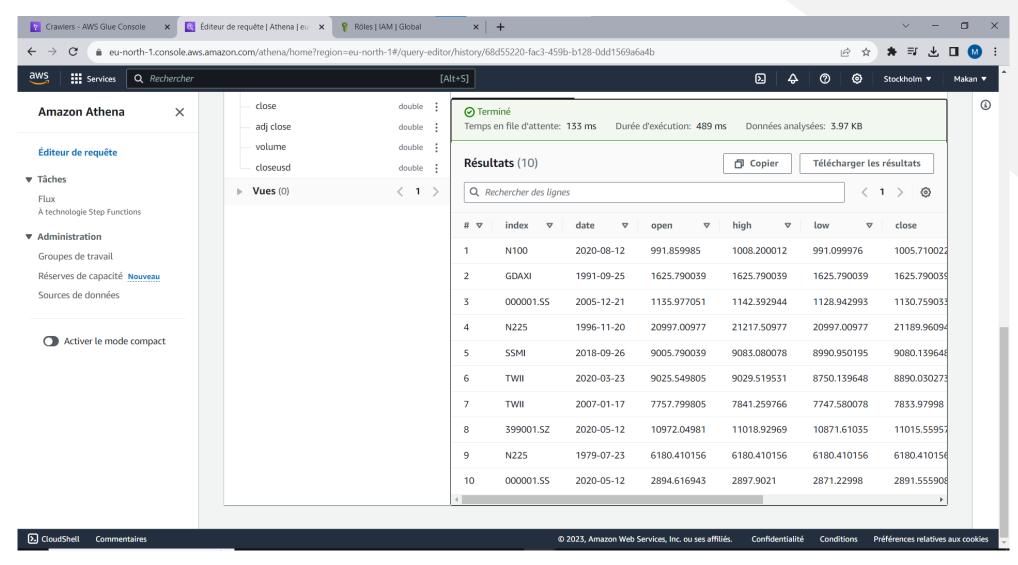
III.1) Création d'un compartiment de données (« Bucket ») pour le stockage des données issues du consumer



III – Accès aux données

MK

III.2) Requêtes Athena pour la lecture de données



BILAN

FR

16

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