Kafka + Spark Streaming + PySpark

FATEMA NAGORI 19635

TABLE OF CONTENT

Introduction

Design

Implementation

Test

Enhancement Ideas

Conclusion

References

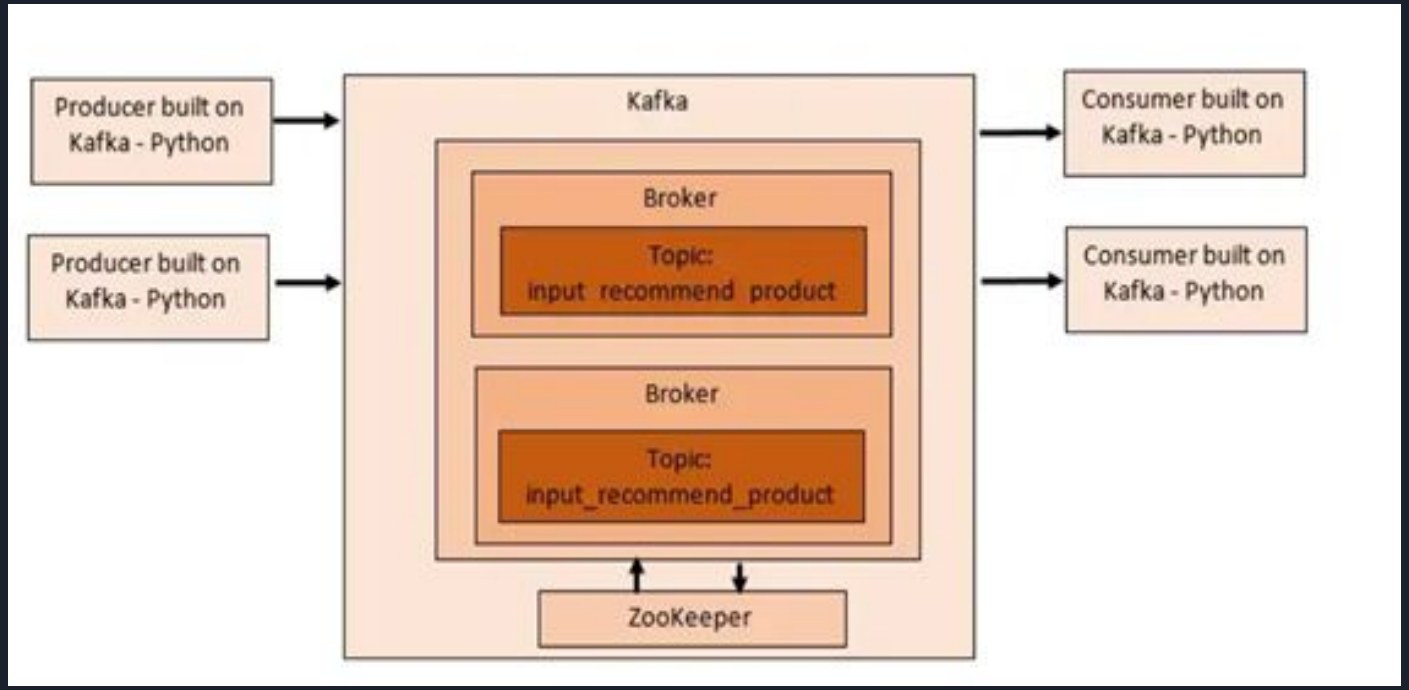
INTRODUCTION

Real-time data ingesting is a common problem in real-time analytics, because in a platform such as e-commerce, active users in a given time and the number of events created by each active user are many. Hence, recommendations (i.e., predictions) for each event or groups of events are expected to be near real-time.

The primary concerns are, *How we will [consume, produce, and process] these events efficiently?*

Apache Kafka addresses the first two problems stated above. It is a distributed streaming platform, which helps to build real-time streaming data pipelines.

KAFKA ECOSYSTEM



Step 1: Study the basic concepts about Kafka

QuickStart — Apache Kafka + Kafka-Python

1. The latest version of Kafka binary distribution is available at https://kafka.apache.org/downloads

kafka

GET STARTED DOCS POWERED BY COMMUNITY APACHE DOWNLOAD KAFKA

Otherwise any version should work (2.13 is recommended).

Kafka 3.2.1 fixes 13 issues since the 3.2.0 release. For more information, please read the detailed Release Notes.

3.2.0

- Released May 17, 2022
- Release Notes
- Source download: kafka-3.2.0-src.tgz (asc, sha512)
- Binary downloads:
  - Scala 2.12 - kafka_2.12-3.2.0.tgz (asc, sha512)
  - Scala 2.13 - kafka_2.13-3.2.0.tgz (asc, sha512)

We build for multiple versions of Scala. This only matters if you are using Scala and you want a version built for the same Scala version you use. Otherwise any version should work (2.13 is recommended).

2. Starting Zookeeper. Unzip it, get into the folders AND cd into it

3. Starting Kafka Brokers Create another terminal, do not close zookeeper

```
[2022-12-01 21:38:24,304] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,309] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerC
onfig)
[2022-12-01 21:38:24,314] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,320] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,320] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,320] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumP
eerConfig)
[2022-12-01 21:38:24,323] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-12-01 21:38:24,323] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-12-01 21:38:24,323] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2022-12-01 21:38:24,323] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2022-12-01 21:38:24,326] INFO Log4j 1.2 jmx support not found; jmx disabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2022-12-01 21:38:24,326] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,327] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerC
onfig)
[2022-12-01 21:38:24,327] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,328] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,328] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2022-12-01 21:38:24,328] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumP
eerConfig)
[2022-12-01 21:38:24,328] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2022-12-01 21:38:24,352] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@4b168fa9 (org.apache.zookeeper.s
erver.ServerMetrics)
[2022-12-01 21:38:24,361] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2022-12-01 21:38:24,381] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,382] INFO   ______                   _                                                  (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,382] INFO  |___  /                  | |                                                 (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,382] INFO     / /    ___    ___    | |  __   __   __   _ __   __ _   __ _   __ _ __   (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,382] INFO    / /    / _ \  / _ \   | |/ /  / _ \  / _ \ | '_ \  / _ \ | '__| (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,382] INFO   / /__  | (_) | | (_) |  |   <  |  __/ |  __/ | |_) |  __/ | |     (org.apache.zookeeper.server.ZooKeeperServer)
[2022-12-01 21:38:24,383] INFO  /_____|  \___/   \___/  |_|\_\  \___|  \___| | .__/   \___| |_| (org.apache.zookeeper.server.ZooKeeperServer)
```

4.Creating Kafka Topics. Create another terminal, do not close zookeeper and kafka brokers

bin/kafka-topics.sh --create --topic input_recommend_product --zookeeper localhost:2181 --partitions 3 --replication-factor 1

```
fatemanagori@DESKTOP-GLUEUFI: /mnt/c/kafka_2.12-3.3.1
-server localhost:9092
Created topic quickstart-events.
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$ bin/kafka-topics.sh --bootstrap-server localhost:9092 -list
input_recommend_product
quickstart-events
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$ bin/kafka-topics.sh --describe --topic quickstart-events --bootstr
ap-server localhost:9092
Topic: quickstart-events        TopicId: 5gxOI6C4RDS1tK_spuijXw PartitionCount: 1       ReplicationFactor: 1    Configs:

        Topic: quickstart-events        Partition: 0    Leader: 0       Replicas: 0     Isr: 0
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$
```

bin/kafka-topics.sh --create --topic input_recommend_product --bootstrap-server localhost:9092 --partitions 3 --replication-factor 1

```
fatemanagori@DESKTOP-GLUEUFI: /mnt/c/kafka_2.12-3.3.1
fatemanagori@DESKTOP-GLUEUFI:~$ cd /mnt/c/kafka_2.12-3.3.1
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$ bin/kafka-topics.sh --create --topic quickstart-events --bootstrap
-server localhost:9092
Created topic quickstart-events.
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$ bin/kafka-topics.sh --bootstrap-server localhost:9092 -list
input_recommend_product
quickstart-events
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_2.12-3.3.1$
```

5 Creating Producer and Consumer using Kafka-python : Create producer.py:

Code:

```
from kafka import KafkaProducer

producer = KafkaProducer(bootstrap_servers='localhost:9092')

producer.send('input_recommend_product', b'(1, Main Menu), (2, Phone) ,
(3, Smart Phone), (4, iPhone)')
producer.close()
```

pip3 install msgpack

pip3 install kafka-python

```
fatemanagori@DESKTOP-GLUEUFI: ~
root @ /dev/tty3: init[8026]

No VM guests are running outdated hypervisor (qemu) binaries on this host.
fatemanagori@DESKTOP-GLUEUFI:~$ pip3 install msgpack
Defaulting to user installation because normal site-packages is not writeable
Collecting msgpack
  Downloading msgpack-1.0.4-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (316 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 KB 3.8 MB/s eta 0:00:00
Installing collected packages: msgpack
Successfully installed msgpack-1.0.4
fatemanagori@DESKTOP-GLUEUFI:~$ pip3 install kafka-python
Defaulting to user installation because normal site-packages is not writeable
Collecting kafka-python
  Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 246.5/246.5 KB 3.6 MB/s eta 0:00:00
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
fatemanagori@DESKTOP-GLUEUFI:~$
```

5.2 Create comsumer.py

Code

```
from kafka import KafkaConsumer

consumer = KafkaConsumer('input_recommend_product',
bootstrap_servers=['localhost:9092'])
for msg in consumer:
    print(msg)
```

5.3 Run comsumer.py first (you can run it in your IDE)

5.4 Create another terminal, run the producer.py

5.5 Go to the consumer terminal, you can see the result

```
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_prj$ ls
consumer.py  producer.py
fatemanagori@DESKTOP-GLUEUFI:/mnt/c/kafka_prj$ python3 consumer.py
ConsumerRecord(topic='input_recommend_product', partition=2, offset=0, timestamp=1669973195741, timestamp_type=0, key=No
ne, value=b'(1, Main Menu), (2, Phone) , (3, Smart Phone), (4, iPhone)', headers=[], checksum=None, serialized_key_size=
-1, serialized_value_size=58, serialized_header_size=-1)
```

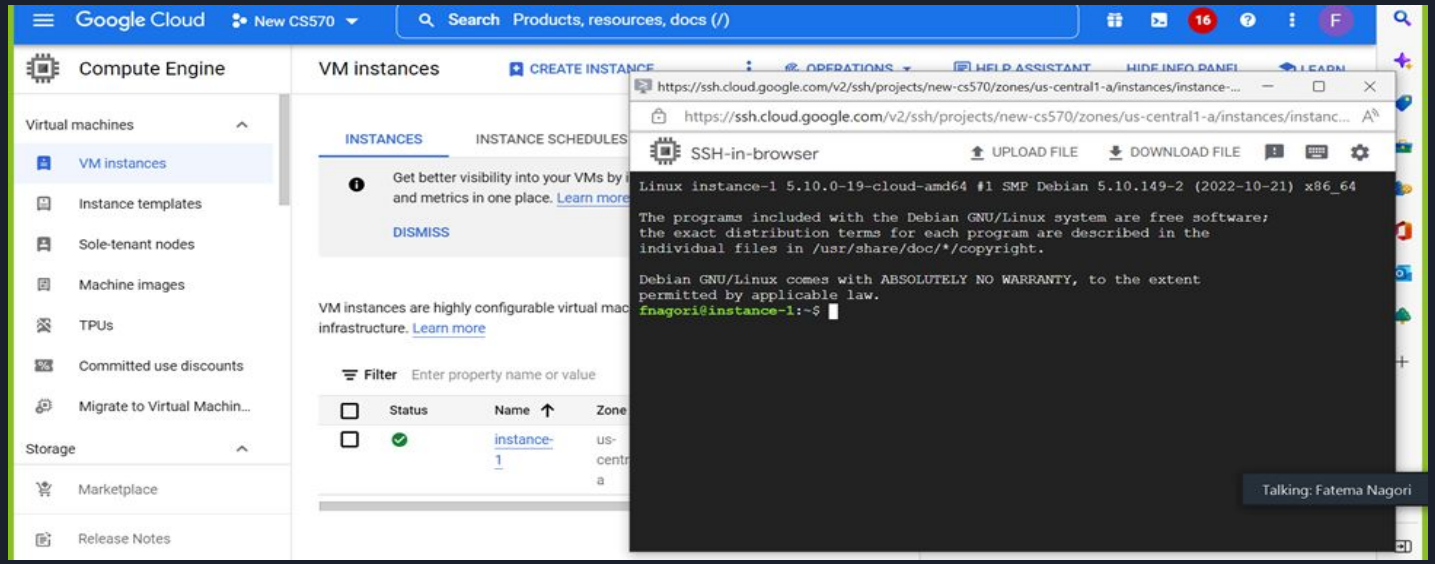
Step 2: Study the basic concepts about Spark Streaming

- Spark Streaming basic concepts AT GCP

2.1 Start a Project

2.2 Create instance at GCP

2.3 Connect to SSH



Installing Spark which is availabe at https://spark.apache.org/downloads.html

download the package and unpack it

$ wget https://dlcdn.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz

$ tar -xvf spark-3.3.1-bin-hadoop3.tgz

```
permitted by applicable law.
fnagori@instance-1:~$ wget https://dlcdn.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
-bash: wget: command not found
fnagori@instance-1:~$ wget https://dlcdn.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
-bash: wget: command not found
fnagori@instance-1:~$ sudo apt-get install wget
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
  wget
0 upgraded, 1 newly installed, 0 to remove and 0 not upgraded.
Need to get 964 kB of archives.
After this operation, 3559 kB of additional disk space will be used.
Get:1 http://deb.debian.org/debian bullseye/main amd64 wget amd64 1.21-1+deb11u1 [964 kB]
Fetched 964 kB in 0s (10.3 MB/s)
Selecting previously unselected package wget.
(Reading database ... 54255 files and directories currently installed.)
Preparing to unpack .../wget_1.21-1+deb11u1_amd64.deb ...
Unpacking wget (1.21-1+deb11u1) ...
Setting up wget (1.21-1+deb11u1) ...
Processing triggers for man-db (2.9.4-2) ...
fnagori@instance-1:~$ wget https://dlcdn.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
--2022-12-05 19:48:45--  https://dlcdn.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 299350810 (285M) [application/x-gzip]
Saving to: 'spark-3.3.1-bin-hadoop3.tgz'

spark-3.3.1-bin-hadoop3.tgz         100%[===================================================================>] 285.48M   127MB/s    in 2.2s

2022-12-05 19:48:47 (127 MB/s) - 'spark-3.3.1-bin-hadoop3.tgz' saved [299350810/299350810]

fnagori@instance-1:~$ tar -xvf spark-3.3.1-bin-hadoop3.tgz
```

Talking:

create a soft link

$ ln -s /home/fnagori/spark-3.3.1-bin-hadoop3 /home/fnagori/spark

set spark related environment varibales

export SPARK_HOME=/home/fnagori/spark

export PATH=$SPARK_HOME/bin:$PATH

export PATH=$SPARK_HOME/sbin:$PATH

```
spark-3.3.1-bin-hadoop3/sbin/start-thriftserver.sh
spark-3.3.1-bin-hadoop3/sbin/start-worker.sh
spark-3.3.1-bin-hadoop3/sbin/start-workers.sh
spark-3.3.1-bin-hadoop3/sbin/stop-all.sh
spark-3.3.1-bin-hadoop3/sbin/stop-history-server.sh
spark-3.3.1-bin-hadoop3/sbin/stop-master.sh
spark-3.3.1-bin-hadoop3/sbin/stop-mesos-dispatcher.sh
spark-3.3.1-bin-hadoop3/sbin/stop-mesos-shuffle-service.sh
spark-3.3.1-bin-hadoop3/sbin/stop-slave.sh
spark-3.3.1-bin-hadoop3/sbin/stop-slaves.sh
spark-3.3.1-bin-hadoop3/sbin/stop-thriftserver.sh
spark-3.3.1-bin-hadoop3/sbin/stop-worker.sh
spark-3.3.1-bin-hadoop3/sbin/stop-workers.sh
spark-3.3.1-bin-hadoop3/sbin/workers.sh
spark-3.3.1-bin-hadoop3/yarn/
spark-3.3.1-bin-hadoop3/yarn/spark-3.3.1-yarn-shuffle.jar
fnagori@instance-1:~$  ln -s /home/fnagori/spark-3.3.1-bin-hadoop3 /home/fnagori/spark
fnagori@instance-1:~$ ls
spark  spark-3.3.1-bin-hadoop3  spark-3.3.1-bin-hadoop3.tgz
fnagori@instance-1:~$ vi ~/.bashrc
fnagori@instance-1:~$ export SPARK_HOME=/home/fnagori/spark
fnagori@instance-1:~$ export PATH=$SPARK_HOME/bin:$PATH
fnagori@instance-1:~$ export PATH=$SPARK_HOME/sbin:$PATH
fnagori@instance-1:~$ echo $PATH
/home/fnagori/spark/sbin:/home/fnagori/spark/bin:/usr/local/bin:/usr/bin:/bin:/usr/local/games:/usr/games
```

Install JAVA

Install java8:

$ sudo apt update

$ sudo apt-get install openjdk-8-jdk

$ update-alternatives --list java

$ export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre

```
fnagori@instance-1:~$ echo $PATH
/home/fnagori/spark/sbin:/home/fnagori/spark/bin:/usr/local/bin:/usr/bin:/bin:/usr/local/games:/usr/games
fnagori@instance-1:~$ pyspark
JAVA_HOME is not set
fnagori@instance-1:~$ jave --version
-bash: jave: command not found
fnagori@instance-1:~$ java --version
-bash: java: command not found
fnagori@instance-1:~$ sudo apt-get install openjdk-8-jdk
```

```
fnagori@instance-2:~$ java -version
openjdk version "1.8.0_352"
OpenJDK Runtime Environment (build 1.8.0_352-8u352-ga-1~20.04-b08)
OpenJDK 64-Bit Server VM (build 25.352-b08, mixed mode)
fnagori@instance-2:~$
```

Start pyspark

```
fnagori@instance-2:~$ pyspark
Python 3.8.10 (default, Jun 22 2022, 20:18:18)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/12/05 20:40:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.3.1
      /_/

Using Python version 3.8.10 (default, Jun 22 2022 20:18:18)
Spark context Web UI available at http://instance-2.us-central1-a.c.new-cs570.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1670272839200).
SparkSession available as 'spark'.
>>>
```

$ start-master.sh

```
fnagori@instance-2:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /home/fnagori/spark/logs/spark-fnagori-org.apache.spark.deploy.master.Master-1-instance-2.out
fnagori@instance-2:~$
```

In your browser, paste and go the link http://34.172.96.149:8080

Not secure | 34.172.96.149:8080

Spark 3.3.1

Spark Master at spark://instance-2.us-central1-a.c.new-cs570.internal:7077

URL: spark://instance-2.us-central1-a.c.new-cs570.internal:7077
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (0)

Worker Id	Address	State	Cores	Memory	Resources

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

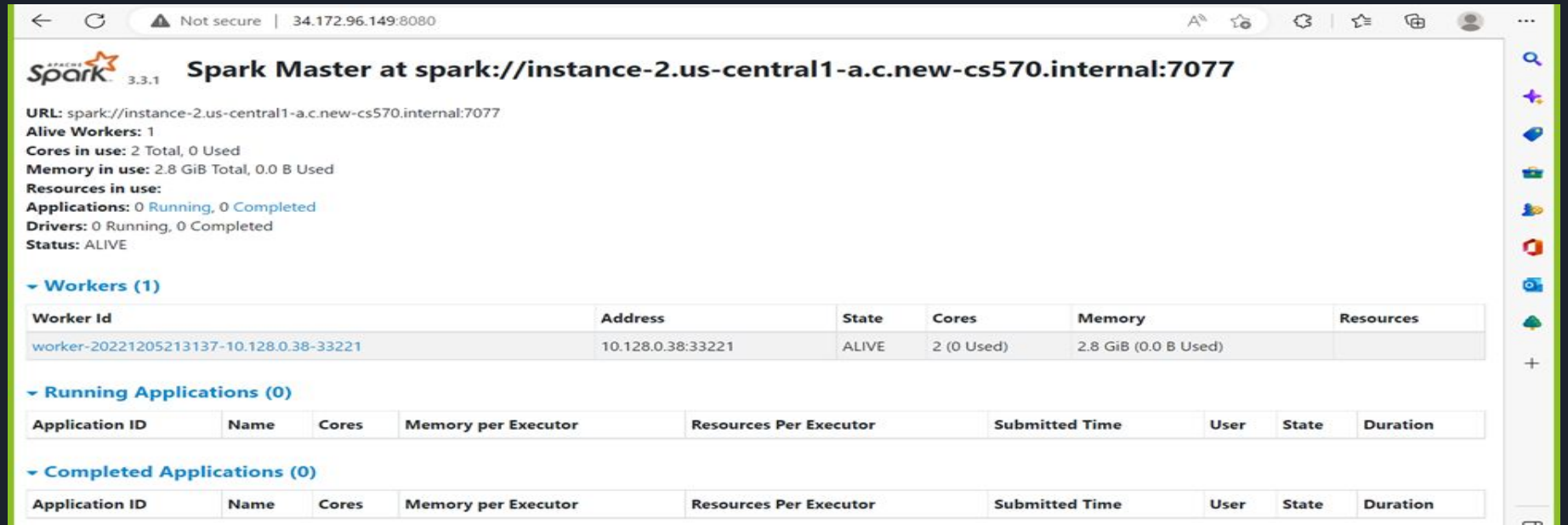
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Starting worker

$ start-slave.sh spark://34.172.96.149:7077

```
no org.apache.spark.deploy.worker.Worker to stop
fnagori@instance-2:~$ start-slave.sh spark://34.172.96.149:7077
This script is deprecated, use start-worker.sh
starting org.apache.spark.deploy.worker.Worker, logging to /home/fnagori/spark/logs/spark-fnagori-org.apache.spa
rk.deploy.worker.Worker-1-instance-2.out
fnagori@instance-2:~$
```

In the browser (http:///34.172.96.149:8080), you can see one alive worker as bellow:



Run Spark Streaming Word Count example

Open a terminal 1:

$ nc -lk 9999

```
fnagori@instance-2:~$ nc -lk 9999
hello world
how are you doing today
hello world
how are you doing today
sunshine sunrise
my name is alpha
my name is beta
how r u
```

Open another terminal 2:

```
$ ./bin/spark-submit examples/src/main/python/streaming/network_wordcount.py localhost 9999
```

```
22/12/05 21:57:26 INFO ShuffleBlockFetcherIterator: Getting 1 (88.0 B) non-empty blocks including 1 (88.0 B) loc
al and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/12/05 21:57:26 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
22/12/05 21:57:26 INFO PythonRunner: Times: total = 44, boot = -46, init = 90, finish = 0
22/12/05 21:57:26 INFO PythonRunner: Times: total = 47, boot = -40, init = 87, finish = 0
22/12/05 21:57:26 INFO Executor: Finished task 0.0 in stage 48.0 (TID 26). 1663 bytes result sent to driver
22/12/05 21:57:26 INFO TaskSetManager: Finished task 0.0 in stage 48.0 (TID 26) in 69 ms on instance-2.us-centra
l1-a.c.new-cs570.internal (executor driver) (1/1)
22/12/05 21:57:26 INFO TaskSchedulerImpl: Removed TaskSet 48.0, whose tasks have all completed, from pool
22/12/05 21:57:26 INFO DAGScheduler: ResultStage 48 (runJob at PythonRDD.scala:166) finished in 0.078 s
22/12/05 21:57:26 INFO DAGScheduler: Job 24 is finished. Cancelling potential speculative or zombie tasks for th
is job
22/12/05 21:57:26 INFO TaskSchedulerImpl: Killing all running tasks in stage 48: Stage finished
22/12/05 21:57:26 INFO DAGScheduler: Job 24 finished: runJob at PythonRDD.scala:166, took 0.082682 s
-------------------------------------------
Time: 2022-12-05 21:57:26
-------------------------------------------
('name', 1)
('is', 1)
('my', 1)
('beta', 1)

22/12/05 21:57:26 INFO JobScheduler: Finished job streaming job 1670277446000 ms.0 from job set of time 16702774
46000 ms
22/12/05 21:57:26 INFO JobScheduler: Total delay: 0.375 s for time 1670277446000 ms (execution: 0.348 s)
22/12/05 21:57:26 INFO PythonRDD: Removing RDD 86 from persistence list
22/12/05 21:57:26 INFO BlockManager: Removing RDD 86
22/12/05 21:57:26 INFO BlockRDD: Removing RDD 81 from persistence list
22/12/05 21:57:26 INFO BlockManager: Removing RDD 81
22/12/05 21:57:26 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[81] at socketTextStream at NativeMeth
odAccessorImpl.java:0 of time 1670277446000 ms
22/12/05 21:57:26 INFO ReceivedBlockTracker: Deleting batches: 1670277444000 ms
22/12/05 21:57:26 INFO InputInfoTracker: remove old batch metadata: 1670277444000 ms
22/12/05 21:57:27 INFO JobScheduler: Starting job streaming job 1670277447000 ms.0 from job set of time 16702774
47000 ms
/name
```

Run Networking WordCount example in python sucessfully:

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
22/12/05 21:57:31 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 5 ms
22/12/05 21:57:31 INFO PythonRunner: Times: total = 45, boot = -51, init = 96, finish = 0
22/12/05 21:57:31 INFO PythonRunner: Times: total = 44, boot = -45, init = 89, finish = 0
22/12/05 21:57:31 INFO Executor: Finished task 0.0 in stage 68.0 (TID 37). 1661 bytes result sent to driver
22/12/05 21:57:31 INFO TaskSetManager: Finished task 0.0 in stage 68.0 (TID 37) in 69 ms on instance-2.us-centra
l1-a.c.new-cs570.internal (executor driver) (1/1)
22/12/05 21:57:31 INFO TaskSchedulerImpl: Removed TaskSet 68.0, whose tasks have all completed, from pool
22/12/05 21:57:31 INFO DAGScheduler: ResultStage 68 (runJob at PythonRDD.scala:166) finished in 0.080 s
22/12/05 21:57:31 INFO DAGScheduler: Job 34 is finished. Cancelling potential speculative or zombie tasks for th
is job
22/12/05 21:57:31 INFO TaskSchedulerImpl: Killing all running tasks in stage 68: Stage finished
22/12/05 21:57:31 INFO DAGScheduler: Job 34 finished: runJob at PythonRDD.scala:166, took 0.085247 s
-------------------------------------------
Time: 2022-12-05 21:57:31
-------------------------------------------
('r', 1)
('how', 1)
('u', 1)

22/12/05 21:57:31 INFO JobScheduler: Finished job streaming job 1670277451000 ms.0 from job set of time 16702774
51000 ms
22/12/05 21:57:31 INFO JobScheduler: Total delay: 0.309 s for time 1670277451000 ms (execution: 0.284 s)
22/12/05 21:57:31 INFO PythonRDD: Removing RDD 126 from persistence list
22/12/05 21:57:31 INFO BlockManager: Removing RDD 126
22/12/05 21:57:31 INFO BlockRDD: Removing RDD 121 from persistence list
22/12/05 21:57:31 INFO BlockManager: Removing RDD 121
22/12/05 21:57:31 INFO SocketInputDStream: Removing blocks of RDD BlockRDD[121] at socketTextStream at NativeMet
hodAccessorImpl.java:0 of time 1670277451000 ms
22/12/05 21:57:31 INFO ReceivedBlockTracker: Deleting batches: 1670277449000 ms
22/12/05 21:57:31 INFO InputInfoTracker: remove old batch metadata: 1670277449000 ms
22/12/05 21:57:32 INFO JobScheduler: Added jobs for time 1670277452000 ms
22/12/05 21:57:32 INFO JobScheduler: Starting job streaming job 1670277452000 ms.0 from job set of time 16702774
52000 ms
22/12/05 21:57:32 INFO SparkContext: Starting job: runJob at PythonRDD.scala:166
@@@
/how
```

Step 3: Following the procedure on this web page Connecting the Dots (Python, Spark, and Kafka)

1 Downlaod kafka which is available at https://kafka.apache.org/downloads

$ wget https://downloads.apache.org/kafka/3.3.1/kafka_2.12-3.3.1.tgz

$ tar -xvf kafka_2.12-3.3.1.tgz

```
fnagori@instance-2:~$ wget https://downloads.apache.org/kafka/3.3.1/kafka_2.12-3.3.1.tgz
--2022-12-06 06:57:34--  https://downloads.apache.org/kafka/3.3.1/kafka_2.12-3.3.1.tgz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 135.181.214.104, 2a01:4f9:3a:2c57::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 105092106 (100M) [application/x-gzip]
Saving to: 'kafka_2.12-3.3.1.tgz'

kafka_2.12-3.3.1.tgz   100%[===================>] 100.22M  25.2MB/s    in 4.7s

2022-12-06 06:57:40 (21.4 MB/s) - 'kafka_2.12-3.3.1.tgz' saved [105092106/105092106]

fnagori@instance-2:~$  tar -xvf kafka_2.12-3.3.1.tgz
kafka_2.12-3.3.1/
kafka_2.12-3.3.1/LICENSE
kafka_2.12-3.3.1/NOTICE
kafka_2.12-3.3.1/bin/
kafka_2.12-3.3.1/bin/kafka-console-consumer.sh
kafka_2.12-3.3.1/bin/kafka-log-dirs.sh
kafka_2.12-3.3.1/bin/kafka-producer-perf-test.sh
kafka_2.12-3.3.1/bin/kafka-console-producer.sh
kafka_2.12-3.3.1/bin/kafka-streams-application-reset.sh
kafka_2.12-3.3.1/bin/kafka-configs.sh
kafka_2.12-3.3.1/bin/kafka-get-offsets.sh
kafka_2.12-3.3.1/bin/kafka-metadata-quorum.sh
kafka_2.12-3.3.1/bin/kafka-server-start.sh
kafka_2.12-3.3.1/bin/zookeeper-server-start.sh
kafka_2.12-3.3.1/bin/kafka-broker-api-versions.sh
kafka_2.12-3.3.1/bin/windows/
kafka_2.12-3.3.1/bin/windows/kafka-get-offsets.bat
```

Part Three: Event Processing on Apache Spark (PySpark)

Setup Spark

$ pip3 install msgpack

$ pip3 install kafka-python

if pip3 command not found,

$ sudo apt install python3-pip

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
Setting up zlib1g-dev:amd64 (1:1.2.11.dfsg-2ubuntu1.5) ...
Setting up cpp (4:9.3.0-1ubuntu2) ...
Setting up gcc-9 (9.4.0-1ubuntu1~20.04.1) ...
Setting up libpython3-dev:amd64 (3.8.2-0ubuntu2) ...
Setting up libstdc++-9-dev:amd64 (9.4.0-1ubuntu1~20.04.1) ...
Setting up gcc (4:9.3.0-1ubuntu2) ...
Setting up g++-9 (9.4.0-1ubuntu1~20.04.1) ...
Setting up python3.8-dev (3.8.10-0ubuntu1~20.04.5) ...
Setting up g++ (4:9.3.0-1ubuntu2) ...
update-alternatives: using /usr/bin/g++ to provide /usr/bin/c++ (c+
 in auto mode
Setting up build-essential (12.8ubuntu1.1) ...
Setting up python3-dev (3.8.2-0ubuntu2) ...
Processing triggers for man-db (2.9.1-1) ...
Processing triggers for libc-bin (2.31-0ubuntu9.9) ...
fnagori@instance-2:~$  pip3 install msgpack
Collecting msgpack
  Downloading msgpack-1.0.4-cp38-cp38-manylinux_2_17_x86_64.manylin
2014_x86_64.whl (322 kB)
     |[redacted]| 322 kB 4.8 MB/s
Installing collected packages: msgpack
Successfully installed msgpack-1.0.4
fnagori@instance-2:~$  pip3 install kafka-python
Collecting kafka-python
  Downloading kafka_python-2.0.2-py2.py3-none-any.whl (246 kB)
     |[redacted]| 246 kB 4.9 MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
fnagori@instance-2:~$
```

$ wget https://repo1.maven.org/maven2/org/apache/spark/spark-streaming-kafka-0-8-assembly_2.11/2.3.2/spark-streaming-kafka-0-8-assembly_2.11-2.3.2.jar

Create and Submit the park Application

Create /home/xwu/pyspark_script/spark_processor.py

$ vi pyspark_script/spark_processor.py

```
import sys
from pyspark import SparkConf, SparkContext
from pyspark.streaming import StreamingContext
#from pyspark.streaming.kafka import KafkaUtils

#processing each micro batch
def process_events(event):
    return (event[0], Counter(event[1].split(" ")).most_common(3))

#push the processed event to Kafka
def push_back_to_kafka(processed_events):
    list_of_processed_events = processed_events.collect()
    producer.send('output_event', value = str(list_of_processed_events))

#create SC with the specified configuration
def spark_context_creator():
    conf = SparkConf()
    #The master URL to connect and set name for our app
    conf.setMaster("spark://34.121.70.117:7077").setAppName("ConnectingDotsSparkKafkaStreaming")
    sc = None
    try:
        sc.stop()
        sc = SparkContext(conf=conf)
    except:
        sc = SparkContext(conf=conf)
    return sc
```

Launch spark application

Open a terminal 1:

$ start-master.sh

```
stopping org.apache.spark.deploy.master.Master
fnagori@instance-2:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /home/fnagori/spark/logs/spark-fnagori-org.apache.spark.deploy.master.Master-1-instance-2.out
fnagori@instance-2:~$
```

Open another terminal 2:

$ ./spark/bin/spark-submit --jars myrun/spark-streaming-kafka-0-10_2.12-3.3.1.jar --master spark://34.70.211.224:7077 --deploy-mode client myrun/spark_processor.py

```
fnagori@instance-2:~$ ./spark/bin/spark-submit --jars spark-streaming-kafka-0-10_2.12-3.3.1.jar --master spark://34.121.70.117:7077 --deploy-mode client spark_p
rocessor.py
22/12/08 20:55:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/12/08 20:55:16 INFO SparkContext: Running Spark version 3.3.1
22/12/08 20:55:16 INFO ResourceUtils: ==============================================================
22/12/08 20:55:16 INFO ResourceUtils: No custom resources configured for spark.driver.
22/12/08 20:55:16 INFO ResourceUtils: ==============================================================
22/12/08 20:55:16 INFO SparkContext: Submitted application: ConnectingDotsSparkKafkaStreaming
22/12/08 20:55:17 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory ->
name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
22/12/08 20:55:17 INFO ResourceProfile: Limiting resource is cpu
22/12/08 20:55:17 INFO ResourceProfileManager: Added ResourceProfile id: 0
22/12/08 20:55:17 INFO SecurityManager: Changing view acls to: fnagori
22/12/08 20:55:17 INFO SecurityManager: Changing modify acls to: fnagori
22/12/08 20:55:17 INFO SecurityManager: Changing view acls groups to:
22/12/08 20:55:17 INFO SecurityManager: Changing modify acls groups to:
22/12/08 20:55:17 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(fnagori); groups with view
 permissions: Set(); users  with modify permissions: Set(fnagori); groups with modify permissions: Set()
22/12/08 20:55:17 INFO Utils: Successfully started service 'sparkDriver' on port 46659.
22/12/08 20:55:17 INFO SparkEnv: Registering MapOutputTracker
22/12/08 20:55:17 INFO SparkEnv: Registering BlockManagerMaster
```

Connect to the master successfully, but failed due to KafkaUtils not defined.

```
22/12/08 20:56:18 INFO NettyBlockTransferService: Server created on instance-2.us-central1-a.c.new-cs570.internal:39257
22/12/08 20:56:18 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/12/08 20:56:18 INFO SparkUI: Stopped Spark web UI at http://instance-2.us-central1-a.c.new-cs570.internal:4040
22/12/08 20:56:18 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, instance-2.us-central1-a.c.new-cs570.internal, 39257, None)
22/12/08 20:56:18 INFO BlockManagerMasterEndpoint: Registering block manager instance-2.us-central1-a.c.new-cs570.internal:39257 with 366.3 MiB RAM, BlockManage
rId(driver, instance-2.us-central1-a.c.new-cs570.internal, 39257, None)
22/12/08 20:56:18 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, instance-2.us-central1-a.c.new-cs570.internal, 39257, None)
22/12/08 20:56:18 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, instance-2.us-central1-a.c.new-cs570.internal, 39257, None)
22/12/08 20:56:18 INFO StandaloneSchedulerBackend: Shutting down all executors
22/12/08 20:56:18 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
22/12/08 20:56:18 WARN StandaloneAppClient$ClientEndpoint: Drop UnregisterApplication(null) because has not yet connected to master
22/12/08 20:56:18 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/12/08 20:56:18 INFO MemoryStore: MemoryStore cleared
22/12/08 20:56:18 INFO BlockManager: BlockManager stopped
22/12/08 20:56:18 INFO BlockManagerMaster: BlockManagerMaster stopped
22/12/08 20:56:18 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/12/08 20:56:18 INFO SparkContext: Successfully stopped SparkContext
```

```
    kafkaStream = KafkaUtils.createStream(ssc, 'localhost:2181', 'test-consumer-group', {'input_event':1})
NameError: name 'KafkaUtils' is not defined
22/12/04 11:01:11 INFO SparkContext: Invoking stop() from shutdown hook
22/12/04 11:01:11 INFO SparkUI: Stopped Spark web UI at http://vm-1.us-central1-a.c.evocative-lodge-362700.internal:4040
22/12/04 11:01:11 INFO StandaloneSchedulerBackend: Shutting down all executors
22/12/04 11:01:11 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
22/12/04 11:01:11 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/12/04 11:01:11 INFO MemoryStore: MemoryStore cleared
```

Enhancement Ideas

Apache Kafka is a **Popular Open-Source Distributed Stream Data Ingesting & Processing Platform**. Providing an end-to-end solution to its users, Kafka can efficiently read & write streams of events in real-time with constant import/export of your data from other data systems.

Its Reliability & Durability allows you to store streams of data securely for as long as you want. With its **Best-in-Class performance, Low latency, Fault Tolerance, and High Throughput**, Kafka can handle & process thousands of messages per second in Real-Time.

Conclusion

1) Python, Spark, and Kafka are important frameworks in a data scientist's daily activities.

2) This article helps data scientists to perform their experiments in Python while deploying the final model in a scalable production environment.

References

USE GCP

GCP common task

QuickStart — Apache Kafka + Kafka-Python

Spark Streaming basic concepts

Connecting the Dots (Python, Spark, and Kafka)


thank
you