EMR cluster master node configuration for the project:

Cluster: 5.30.1
Hadoop distribution: 2.8.5
Spark 2.4.5
Hive 2.3.6, Hue 4.6.0
HBase 1.4.13
Sqoop 1.4.7
HCatalog 2.3.6
Kafka 2.3.0 (Not included in cluster, need to download separately)

YARN Configuration:

[{"classification":"yarn-site", "properties":{"yarn.nodemanager.resource.memory-mb":"10240", "yarn.scheduler.maximum-allocation-mb":"8192"}, "configurations":[]}]

Login as User - **hadoop**

---

echo -e "installation of required kafka and other packages\n"

cd

wget https://archive.apache.org/dist/kafka/2.3.0/kafka_2.12-2.3.0.tgz

tar xzf kafka_2.12-2.3.0.tgz

wget https://de-mysql-connector.s3.amazonaws.com/mysql-connector-java-8.0.25.tar.gz

tar xzf mysql-connector-java-8.0.25.tar.gz

sudo cp ./mysql-connector-java-8.0.25/mysql-connector-java-8.0.25.jar /usr/lib/sqoop/lib/

sudo pip3 install kafka-python

sudo pip3 install mysql-connector

sudo pip3 install boto3

echo -e "import required libraries from hive and hbase to access through spark"

sudo cp /usr/lib/hive/lib/hive-hbase-handler* /usr/lib/hbase/lib/

sudo cp /usr/lib/hive/lib/hive-hbase-handler* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/hbase-common* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/hbase-client* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/hbase-server* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/hbase-protocol* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/htrace-core* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/hbase-metrics* /usr/lib/spark/jars/

sudo cp /usr/lib/hbase/lib/metrics-core-2.2.0* /usr/lib/spark/jars/

echo -e "stored rds db credential in hdfs for sqoop job \n"

touch rdsdbpassfile | echo -n "STUDENT123" > rdsdbpassfile

hadoop fs -put rdsdbpassfile /user/hadoop/

hadoop fs -chmod 400 /user/hadoop/rdsdbpassfile

hadoop fs -ls /user/hadoop/rdsdbpassfile

Starting of Kafka Server on default cluster master node

/home/hadoop/kafka_2.12-2.3.0/bin/kafka-server-start.sh /home/hadoop/kafka_2.12-2.3.0/config/server.properties

Creation of Kafka topics

/home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic Patients-Vital-Info

/home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic Health-Alert-Messages

/home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --list --bootstrap-server localhost:9092

---

installation of packages

```
[hadoop@ip-172-31-62-240 ~]$
[hadoop@ip-172-31-62-240 ~]$ cd
[hadoop@ip-172-31-62-240 ~]$ wget https://archive.apache.org/dist/kafka/2.3.0/kafka_2.12-2.3.0.tgz
--2023-01-15 13:48:28--  https://archive.apache.org/dist/kafka/2.3.0/kafka_2.12-2.3.0.tgz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 57215197 (55M) [application/x-gzip]
Saving to: 'kafka_2.12-2.3.0.tgz'

100%[=============================================================================================================>

2023-01-15 13:48:31 (22.2 MB/s) - 'kafka_2.12-2.3.0.tgz' saved [57215197/57215197]

[hadoop@ip-172-31-62-240 ~]$ tar xzf kafka_2.12-2.3.0.tgz
[hadoop@ip-172-31-62-240 ~]$ sudo pip3 install kafka-python
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting kafka-python
  Downloading https://files.pythonhosted.org/packages/75/68/dcb0db055309f680ab2931a3eeb22d865604b638acf8c914bedf4c1a0c
    100% |                                | 256kB 4.7MB/s
Installing collected packages: kafka-python
Successfully installed kafka-python-2.0.2
[hadoop@ip-172-31-62-240 ~]$ sudo pip3 install mysql-connector
WARNING: Running pip install with root privileges is generally not a good idea. Try `pip3 install --user` instead.
Collecting mysql-connector
  Downloading https://files.pythonhosted.org/packages/28/04/e40098f3730e75bbe36a338926f566ea803550a34fb50535499f4fc478
    100% |                                | 11.9MB 107kB/s
Installing collected packages: mysql-connector
  Running setup.py install for mysql-connector ... done
Successfully installed mysql-connector-2.2.9
[hadoop@ip-172-31-62-240 ~]$
```

Starting of Kafka Server

```
[2023-02-02 15:16:30,797] INFO Stat of the created znode at /brokers/ids/0 is: 97,97,1675350990779,1675350990779,1,0,0,72057621008875526,2
 (kafka.zk.KafkaZkClient)
[2023-02-02 15:16:30,798] INFO Registered broker 0 at path /brokers/ids/0 with addresses: ArrayBuffer(EndPoint(ip-172-31-58-127.ec2.intern
(broker epoch): 97 (kafka.zk.KafkaZkClient)
[2023-02-02 15:16:30,800] WARN No meta.properties file under dir /tmp/kafka-logs/meta.properties (kafka.server.BrokerMetadataCheckpoint)
[2023-02-02 15:16:30,874] INFO [ExpirationReaper-0-topic]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2023-02-02 15:16:30,879] INFO [ExpirationReaper-0-Heartbeat]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2023-02-02 15:16:30,880] INFO [ExpirationReaper-0-Rebalance]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2023-02-02 15:16:30,889] INFO Successfully created /controller_epoch with initial epoch 0 (kafka.zk.KafkaZkClient)
[2023-02-02 15:16:30,905] INFO [GroupCoordinator 0]: Starting up. (kafka.coordinator.group.GroupCoordinator)
[2023-02-02 15:16:30,907] INFO [GroupCoordinator 0]: Startup complete. (kafka.coordinator.group.GroupCoordinator)
[2023-02-02 15:16:30,912] INFO [GroupMetadataManager brokerId=0] Removed 0 expired offsets in 6 milliseconds. (kafka.coordinator.group.Gr
[2023-02-02 15:16:30,920] INFO [ProducerId Manager 0]: Acquired new producerId block (brokerId:0,blockStartProducerId:0,blockEndProducerId
oordinator.transaction.ProducerIdManager)
[2023-02-02 15:16:30,955] INFO [TransactionCoordinator id=0] Starting up. (kafka.coordinator.transaction.TransactionCoordinator)
[2023-02-02 15:16:30,957] INFO [Transaction Marker Channel Manager 0]: Starting (kafka.coordinator.transaction.TransactionMarkerChannelMa
[2023-02-02 15:16:30,958] INFO [TransactionCoordinator id=0] Startup complete. (kafka.coordinator.transaction.TransactionCoordinator)
[2023-02-02 15:16:31,038] INFO [/config/changes-event-process-thread]: Starting (kafka.common.ZkNodeChangeNotificationListener$ChangeEvent
[2023-02-02 15:16:31,087] INFO [SocketServer brokerId=0] Started data-plane processors for 1 acceptors (kafka.network.SocketServer)
[2023-02-02 15:16:31,099] INFO Kafka version: 2.3.0 (org.apache.kafka.common.utils.AppInfoParser)
[2023-02-02 15:16:31,099] INFO Kafka commitId: fc1aaa116b661c8a (org.apache.kafka.common.utils.AppInfoParser)
[2023-02-02 15:16:31,099] INFO Kafka startTimeMs: 1675350991088 (org.apache.kafka.common.utils.AppInfoParser)
[2023-02-02 15:16:31,120] INFO [KafkaServer id=0] started (kafka.server.KafkaServer)
```

Kafka topic creation and listing of topics

```
hadoop@ip-172-31-58-161:~
[hadoop@ip-172-31-58-161 ~]$ /home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --parti
[hadoop@ip-172-31-58-161 ~]$
[hadoop@ip-172-31-58-161 ~]$ /home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --parti
[hadoop@ip-172-31-58-161 ~]$
[hadoop@ip-172-31-58-161 ~]$ /home/hadoop/kafka_2.12-2.3.0/bin/kafka-topics.sh --list --bootstrap-server localhost:9092
Health-Alert-Messages
Patients-Vital-Info
[hadoop@ip-172-31-58-161 ~]$
```