Akash Tanwani A20448831

CSP554—Big Data Technologies

Assignment #9

Exercise 1) 5 points

Read the article "Real-time stream processing for Big Data" available on the blackboard in the 'Articles' section and then answer the following questions:

a) (1.25 points) What is the Kappa architecture and how does it differ from the lambda architecture?
   - The basic idea of the Kappa architecture is to not periodically recompute all data in the batch layer, but to do all computation in the stream processing system alone and only perform recomputation when the business logic changes by replaying historical data.
     The advantage of Kappa architecture over Lambda architecture is its simplicity. With Lambda, we would need to maintain two different processes and possibly two different set of teams which can put pressure on small budget projects. In Kappa, there's only one level of process and one set of teams so it's cheaper to maintain. Also, from end-user perspective, with Kappa there's only one plug-in required to read the data while in Lambda there are two different views for batch and real-time data results.

b) (1.25 points) What are the advantages and drawbacks of pure streaming versus micro-batch real-time processing systems?
   - Purely stream-oriented systems such as Storm and Samza provide very low latency(advantage) and relatively high per-item cost(drawback), while batch-oriented systems achieve unparalleled resource-efficiency at the expense of latency that is prohibitively high for real-time applications.
     The space between these two extremes is vast and some systems like Storm Trident and Spark Streaming employ micro-batching strategies to trade latency against throughput: Trident groups tuples into batches to relax the one-at-a-time processing model in favour of increased throughput as an advantage, whereas Spark Streaming restricts batch size in a native batch processor to reduce latency as a drawback.

c) (1.25 points) In few sentences describe the data processing pipeline in Storm.
   - A data pipeline or application in Storm is called a topology and is a directed graph that represents data flow as directed edges between nodes which again represent the individual processing steps: The nodes that ingest data and thus initiate the data flow in the topology are called spouts and emit tuples to the nodes downstream which are called bolts and do processing, write data to external storage and may send tuples further downstream themselves. Storm comes with several groupings that control data flow between nodes, e.g. for shuffling or hash-partitioning a stream of tuples by some attribute value, but also allows arbitrary custom groupings.

d) (1.25 points) How does Spark streaming shift the Spark batch processing approach to work on real-time data streams?
   - Spark Streaming shifts Spark's batch-processing approach towards real-time requirements by chunking the stream of incoming data items into small batches, transforming them into RDDs and processing them as usual. It further takes care of data flow and distribution automatically. Data is ingested and transformed into a sequence of RDDs which is called DStream (discretised stream) before processing through workers. All RDDs in a DStream are processed in order, whereas data items inside an RDD are processed in parallel without any ordering guarantees.

Exercise 2) 5 points extra credit

Follow the document “Instructions for setting up a VM with Kafka” included with this assignment and execute the demo code. Provide enough screen shots to indicate you have completed the document through section 4. Then remember to terminate your VM.

KT1

```
MINGW64:/c/Users/Akash Tanwani

Akash@Akash MINGW64 ~
$ scp -i /h/Illinois_tech_masters/Third_Semester/bigData/Assignment2/emr-key-pair.pem /h/Illinois_tech_masters/Third_Semester/bigD
ata/Assignment9/kafka_2.12-2.3.0.tgz ubuntu@ec2-54-144-142-18.compute-1.amazonaws.com:/home/ubuntu
kafka_2.12-2.3.0.tgz                                                                          100%   55MB   2.0MB/s   00:26

Akash@Akash MINGW64 ~
$ |
```

KT2

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
kafka_2.12-2.3.0/libs/javax.servlet-api-3.1.0.jar
kafka_2.12-2.3.0/libs/jetty-http-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-io-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-continuation-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/jetty-util-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/log4j-1.2.17.jar
kafka_2.12-2.3.0/libs/hk2-api-2.5.0.jar
kafka_2.12-2.3.0/libs/hk2-utils-2.5.0.jar
kafka_2.12-2.3.0/libs/jakarta.inject-2.5.0.jar
kafka_2.12-2.3.0/libs/jakarta.annotation-api-1.3.4.jar
kafka_2.12-2.3.0/libs/osgi-resource-locator-1.0.1.jar
kafka_2.12-2.3.0/libs/validation-api-2.0.1.Final.jar
kafka_2.12-2.3.0/libs/aopalliance-repackaged-2.5.0.jar
kafka_2.12-2.3.0/libs/javassist-3.22.0-CR2.jar
kafka_2.12-2.3.0/libs/connect-api-2.3.0.jar
kafka_2.12-2.3.0/libs/javax.ws.rs-api-2.1.1.jar
kafka_2.12-2.3.0/libs/connect-runtime-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-json-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-transforms-2.3.0.jar
kafka_2.12-2.3.0/libs/jetty-client-9.4.18.v20190429.jar
kafka_2.12-2.3.0/libs/reflections-0.9.11.jar
kafka_2.12-2.3.0/libs/maven-artifact-3.6.1.jar
kafka_2.12-2.3.0/libs/guava-20.0.jar
kafka_2.12-2.3.0/libs/plexus-utils-3.2.0.jar
kafka_2.12-2.3.0/libs/commons-lang3-3.8.1.jar
kafka_2.12-2.3.0/libs/connect-file-2.3.0.jar
kafka_2.12-2.3.0/libs/connect-basic-auth-extension-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-2.3.0.jar
kafka_2.12-2.3.0/libs/rocksdbjni-5.18.3.jar
kafka_2.12-2.3.0/libs/kafka-streams-scala_2.12-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-test-utils-2.3.0.jar
kafka_2.12-2.3.0/libs/kafka-streams-examples-2.3.0.jar
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ |
```

KT2

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
erationReaper)
[2020-10-26 03:55:57,034] INFO [ExpirationReaper-0-Rebalance]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredOp
erationReaper)
[2020-10-26 03:55:57,042] INFO Successfully created /controller_epoch with initial epoch 0 (kafka.zk.KafkaZkClient)
[2020-10-26 03:55:57,058] INFO [GroupCoordinator 0]: Starting up. (kafka.coordinator.group.GroupCoordinator)
[2020-10-26 03:55:57,059] INFO [GroupCoordinator 0]: Startup complete. (kafka.coordinator.group.GroupCoordinator)
[2020-10-26 03:55:57,067] INFO [GroupMetadataManager brokerId=0] Removed 0 expired offsets in 9 milliseconds. (kafka.coor
dinator.group.GroupMetadataManager)
[2020-10-26 03:55:57,084] INFO [ProducerId Manager 0]: Acquired new producerId block (brokerId:0,blockStartProducerId:0,b
lockEndProducerId:999) by writing to Zk with path version 1 (kafka.coordinator.transaction.ProducerIdManager)
[2020-10-26 03:55:57,111] INFO [TransactionCoordinator id=0] Starting up. (kafka.coordinator.transaction.TransactionCoord
inator)
[2020-10-26 03:55:57,113] INFO [Transaction Marker Channel Manager 0]: Starting (kafka.coordinator.transaction.Transactio
nMarkerChannelManager)
[2020-10-26 03:55:57,113] INFO [TransactionCoordinator id=0] Startup complete. (kafka.coordinator.transaction.Transaction
Coordinator)
[2020-10-26 03:55:57,151] INFO [/config/changes-event-process-thread]: Starting (kafka.common.ZkNodeChangeNotificationLis
tener$ChangeEventProcessThread)
[2020-10-26 03:55:57,180] INFO [SocketServer brokerId=0] Started data-plane processors for 1 acceptors (kafka.network.Soc
ketServer)
[2020-10-26 03:55:57,203] INFO Kafka version: 2.3.0 (org.apache.kafka.common.utils.AppInfoParser)
[2020-10-26 03:55:57,204] INFO Kafka commitId: fc1aaa116b661c8a (org.apache.kafka.common.utils.AppInfoParser)
[2020-10-26 03:55:57,204] INFO Kafka startTimeMs: 1603684557181 (org.apache.kafka.common.utils.AppInfoParser)
[2020-10-26 03:55:57,205] INFO Got user-level KeeperException when processing sessionid:0x100002b3a580000 type:multi cxid
:0x38 zxid:0x1c txntype:-1 reqpath:n/a aborting remaining multi ops. Error Path:/admin/preferred_replica_election Error:K
eeperErrorCode = NoNode for /admin/preferred_replica_election (org.apache.zookeeper.server.PrepRequestProcessor)
[2020-10-26 03:55:57,206] INFO [KafkaServer id=0] started (kafka.server.KafkaServer)
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ |
```

KT3

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
Akash@Akash MINGW64 ~
$ ssh -i "/h/Illinois_tech_masters/Third_Semester/bigData/Assignment2/emr-key-pair.pem" ubuntu@ec2-54-144-142-18.compute-1.amazonaws.co
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 5.3.0-1035-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Mon Oct 26 03:46:01 UTC 2020

  System load:  0.0               Processes:           117
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.53.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:26:06 2020 from 208.59.159.174
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partition
s 1 --topic test
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ |
```

KT3

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Mon Oct 26 03:46:01 UTC 2020

  System load:  0.0               Processes:           117
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.53.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:26:06 2020 from 208.59.159.174
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partition
s 1 --topic test
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-topics.sh --list --bootstrap-server localhost:9092
test
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-console-producer.sh --broker-list localhost:9092 --topic test
>Hello This is Akash Tanwani
>This is the first time I am using Kagka
>In the above sentence it's Kafka instead of kagka
>
```

KT4

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  System information as of Mon Oct 26 03:51:09 UTC 2020

  System load:  0.0               Processes:           119
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.53.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:46:02 2020 from 208.59.159.174
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --f
rom-beginning
Hello This is Akash Tanwani
This is the first time I am using Kagka
In the above sentence it's Kafka instead of kagka
```

KT3

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
[2020-10-26 04:15:41,273] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
[2020-10-26 04:15:51,274] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
[2020-10-26 04:15:51,274] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
[2020-10-26 04:16:01,275] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
[2020-10-26 04:16:01,275] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa

[2020-10-26 04:16:11,275] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
[2020-10-26 04:16:11,276] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
[2020-10-26 04:16:21,276] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
[2020-10-26 04:16:21,276] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
[2020-10-26 04:16:31,277] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
[2020-10-26 04:16:31,277] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
[2020-10-26 04:16:41,277] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
rSourceTask:398)
[2020-10-26 04:16:41,278] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
che.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:16:51,278] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
rSourceTask:398)
[2020-10-26 04:16:51,278] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
che.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:17:01,279] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
rSourceTask:398)
[2020-10-26 04:17:01,279] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
che.kafka.connect.runtime.WorkerSourceTask:415)
[2020-10-26 04:17:11,279] INFO WorkerSourceTask{id=local-file-source-0} Committing offsets (org.apache.kafka.connect.runtime.Worke
rSourceTask:398)
[2020-10-26 04:17:11,280] INFO WorkerSourceTask{id=local-file-source-0} flushing 0 outstanding messages for offset commit (org.apa
che.kafka.connect.runtime.WorkerSourceTask:415)
```

KT4

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0

  System information as of Mon Oct 26 03:51:09 UTC 2020

  System load:  0.0               Processes:           119
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.53.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:46:02 2020 from 208.59.159.174
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --f
rom-beginning
Hello This is Akash Tanwani
This is the first time I am using Kagka
In the above sentence it's Kafka instead of kagka
^CProcessed a total of 3 messages
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$
```

KT4

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
  Usage of /:   7.0% of 30.96GB   Users logged in:     1
  Memory usage: 1%                IP address for ens3: 172.31.53.0
  Swap usage:   0%


46 packages can be updated.
38 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Mon Oct 26 03:46:02 2020 from 208.59.159.174
ubuntu@ip-172-31-53-0:~$ export PATH=/home/ubuntu/kafka_2.12-2.3.0/bin:$PATH
ubuntu@ip-172-31-53-0:~$ cd /home/ubuntu/kafka_2.12-2.3.0
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test --f
rom-beginning
Hello This is Akash Tanwani
This is the first time I am using Kagka
In the above sentence it's Kafka instead of kagka
^CProcessed a total of 3 messages
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic connect-
test --from-beginning
{"schema":{"type":"string","optional":false},"payload":"foo"}
{"schema":{"type":"string","optional":false},"payload":"bar"}
```

KT4

```
ubuntu@ip-172-31-53-0: ~/kafka_2.12-2.3.0
bar
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$ more test.sink.txt
foo
bar
Another line
Another line
ubuntu@ip-172-31-53-0:~/kafka_2.12-2.3.0$
```