

Case Study 3 — Real-Time Data Lakehouse Architecture

By Kafeel Kamran

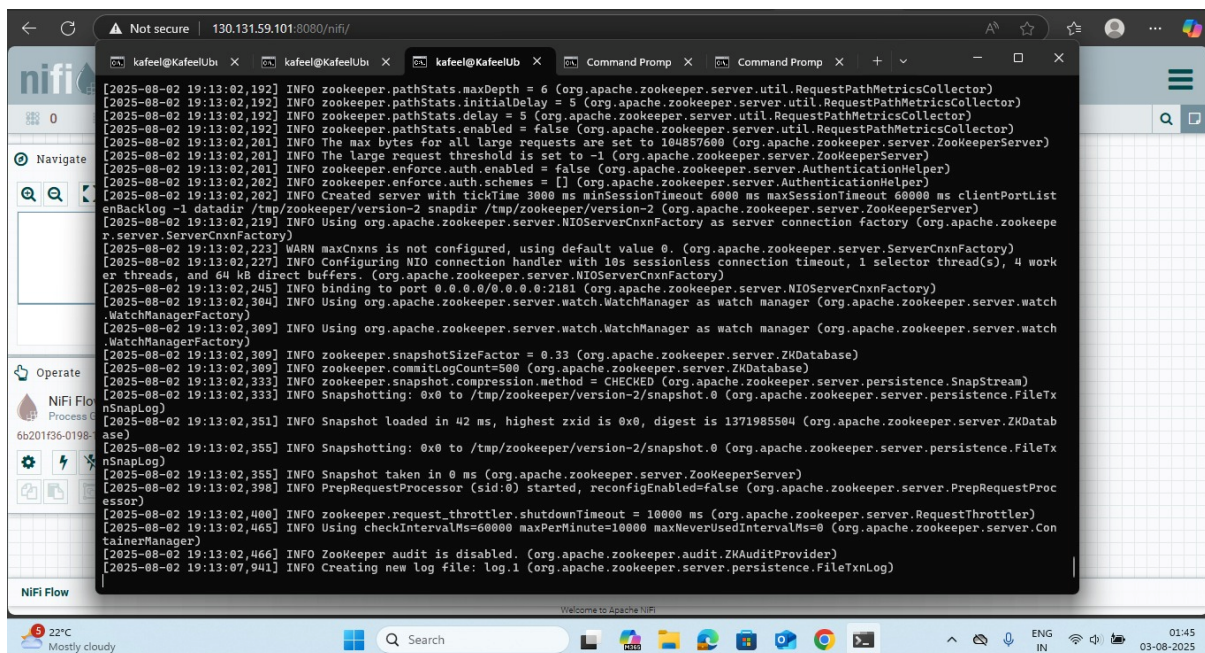
Data: hw_200.csv,

Sales.csv (<https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset>)

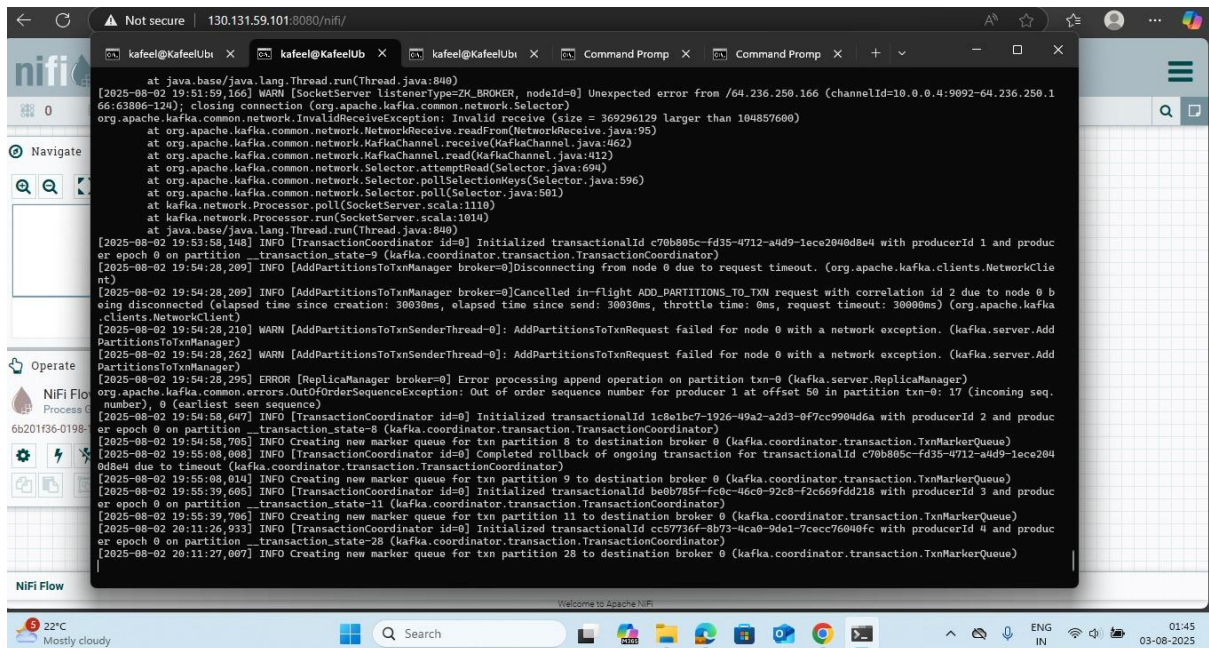
Docs for reference

Basic Environment setup (Nifi + Kafka)

- *Start and run zookeeper*

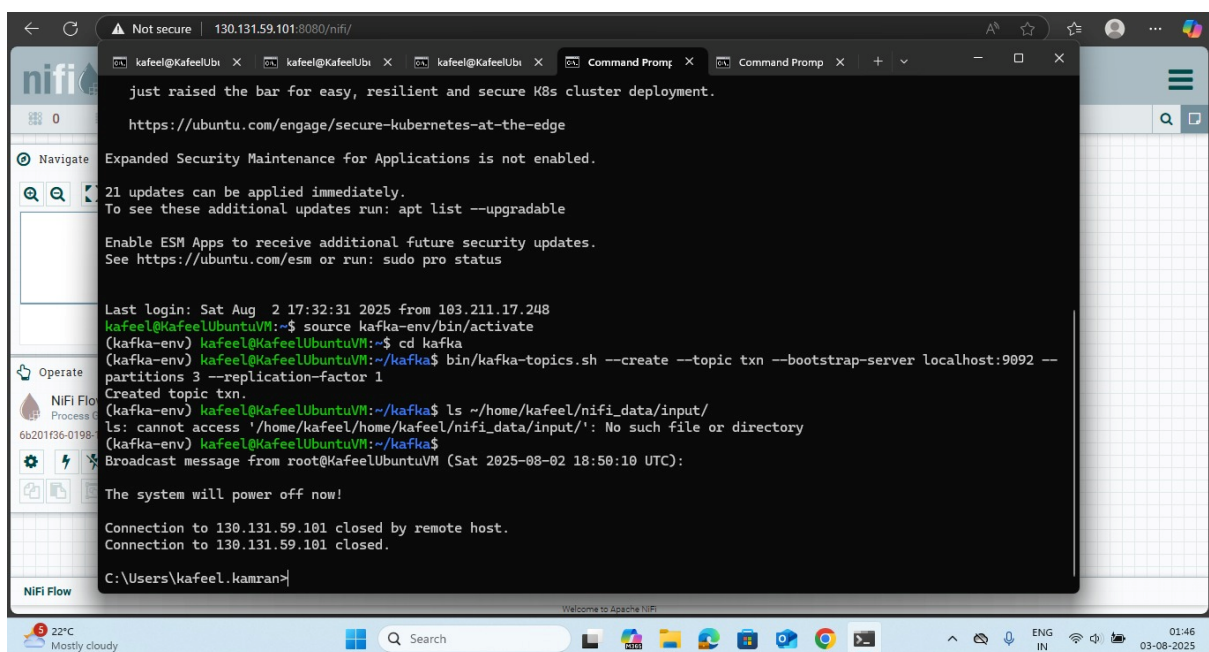


- *Start and run Kafka server*



```
at java.base/java.lang.Thread.run(Thread.java:848)
[2025-08-02 19:51:59,166] WARN [SocketServer listenerType=ZK_BROKER, nodeId=0] Unexpected error from /64.236.250.166 (channelId=10.0.0.4:9092-64.236.250.166:63806-124); closing connection (org.apache.kafka.common.network.Selector)
org.apache.kafka.common.network.InvalidReceiveException: Invalid receive (size = 369296129 larger than 104857600)
    at org.apache.kafka.common.network.NetworkReceive.readFrom(NetworkReceive.java:95)
    at org.apache.kafka.common.network.KafkaChannel.receive(KafkaChannel.java:462)
    at org.apache.kafka.common.network.KafkaChannel.read(KafkaChannel.java:412)
    at org.apache.kafka.common.network.Selector.attemptRead(Selector.java:694)
    at org.apache.kafka.common.network.Selector.pollSelectionKeys(Selector.java:596)
    at org.apache.kafka.common.network.Selector.poll(Selector.java:501)
    at kafka.network.Processor.poll(SocketServer.scala:1110)
    at kafka.network.Processor.run(SocketServer.scala:1014)
    at java.base/java.lang.Thread.run(Thread.java:848)
[2025-08-02 19:53:58,148] INFO [TransactionCoordinator id=0] Initialized transactionalId c70b805c-fd35-4712-a4d9-1ece2040d8e4 with producerId 1 and producer epoch 0 on partition _transaction_state-9 (kafka.coordinator.transaction.TransactionCoordinator)
[2025-08-02 19:54:28,209] INFO [AddPartitionsToTxnManager broker=0] Disconnecting from node 0 due to request timeout. (org.apache.kafka.clients.NetworkClient)
[2025-08-02 19:54:28,209] INFO [AddPartitionsToTxnManager broker=0] Cancelled in-flight ADD_PARTITIONS_TO_TXN request with correlation id 2 due to node 0 being disconnected (elapsed time since creation: 30030ms, elapsed time since send: 30030ms, throttle time: 0ms, request timeout: 30000ms) (org.apache.kafka.clients.NetworkClient)
[2025-08-02 19:54:28,210] WARN [AddPartitionsToTxnSenderThread-0] AddPartitionsToTxnRequest failed for node 0 with a network exception. (kafka.server.AddPartitionsToTxnManager)
[2025-08-02 19:54:28,262] WARN [AddPartitionsToTxnSenderThread-0] AddPartitionsToTxnRequest failed for node 0 with a network exception. (kafka.server.AddPartitionsToTxnManager)
[2025-08-02 19:54:28,295] ERROR [ReplicaManager broker=0] Error processing append operation on partition txn=0 (kafka.server.ReplicaManager)
org.apache.kafka.common.errors.OutOfOrderSequenceException: Out of order sequence number for producer 1 at offset 50 in partition txn=0: 17 (incoming sequence number) 0 (earliest seen sequence)
[2025-08-02 19:54:58,647] INFO [TransactionCoordinator id=0] Initialized transactionalId 1c8e1bc7-1926-49a2-a2d3-0f7cc9904d6a with producerId 2 and producer epoch 0 on partition _transaction_state-8 (kafka.coordinator.transaction.TransactionCoordinator)
[2025-08-02 19:54:58,705] INFO [TransactionCoordinator id=0] Creating new marker queue for txn partition 8 to destination broker 0 (kafka.coordinator.transaction.TxnMarkerQueue)
[2025-08-02 19:55:08,608] INFO [TransactionCoordinator id=0] Completed rollback of ongoing transaction for transactionalId c70b805c-fd35-4712-a4d9-1ece2040d8e4 due to timeout (kafka.coordinator.transaction.TransactionCoordinator)
[2025-08-02 19:55:08,614] INFO [TransactionCoordinator id=0] Creating new marker queue for txn partition 9 to destination broker 0 (kafka.coordinator.transaction.TxnMarkerQueue)
[2025-08-02 19:55:39,605] INFO [TransactionCoordinator id=0] Initialized transactionalId be0b785f-fc0c-92c8-f2c669fd218 with producerId 3 and producer epoch 0 on partition _transaction_state-11 (kafka.coordinator.transaction.TransactionCoordinator)
[2025-08-02 19:55:39,706] INFO [TransactionCoordinator id=0] Creating new marker queue for txn partition 11 to destination broker 0 (kafka.coordinator.transaction.TxnMarkerQueue)
[2025-08-02 20:11:26,933] INFO [TransactionCoordinator id=0] Initialized transactionalId cc57736f-8b73-4ca0-9de1-7ccc76040fc with producerId 4 and producer epoch 0 on partition _transaction_state-28 (kafka.coordinator.transaction.TransactionCoordinator)
[2025-08-02 20:11:27,007] INFO [TransactionCoordinator id=0] Creating new marker queue for txn partition 28 to destination broker 0 (kafka.coordinator.transaction.TxnMarkerQueue)
```

- *Create a topic 'txn' from VM*



```
just raised the bar for easy, resilient and secure K8s cluster deployment.
https://ubuntu.com/engage/secure-kubernetes-at-the-edge

Expanded Security Maintenance for Applications is not enabled.
21 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

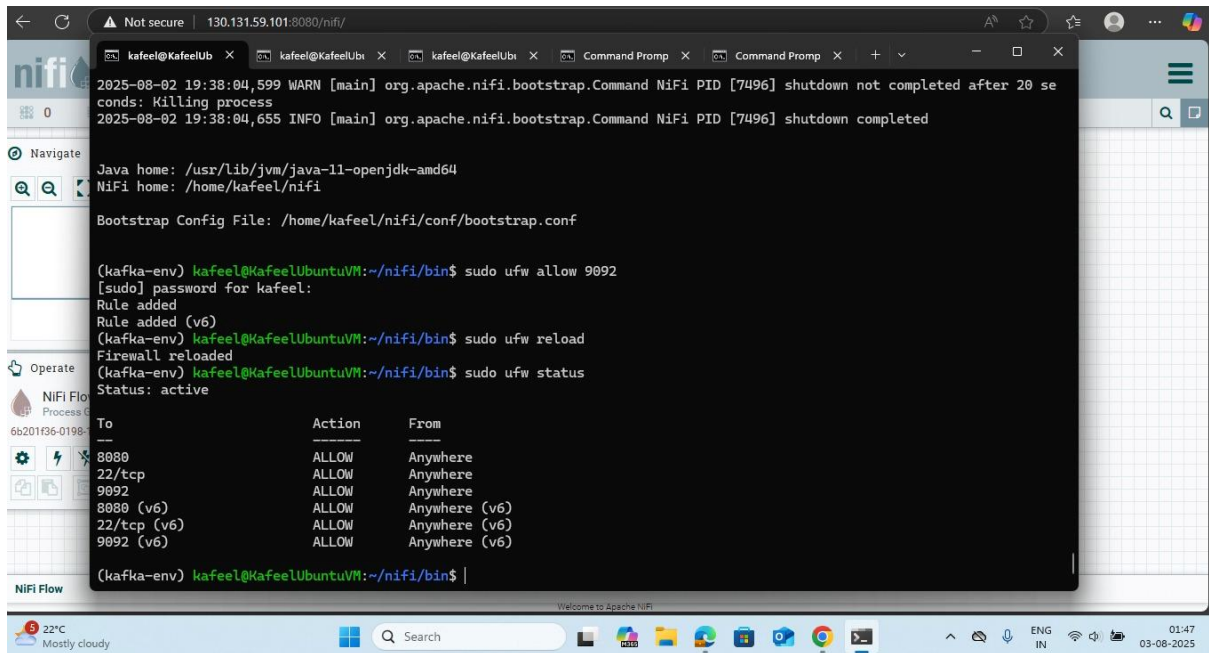
Last login: Sat Aug  2 17:32:31 2025 from 103.211.17.248
kafeel@KafeelUbuntuVM:~$ source kafka-env/bin/activate
(kafeel-env) kafeel@KafeelUbuntuVM:~$ cd kafka
(kafeel-env) kafeel@KafeelUbuntuVM:~/kafka$ bin/kafka-topics.sh --create --topic txn --bootstrap-server localhost:9092 --partitions 3 --replication-factor 1
Created topic txn.
(kafeel-env) kafeel@KafeelUbuntuVM:~/kafka$ ls ~/home/kafeel/nifi_data/input/
ls: cannot access '/home/kafeel/home/kafeel/nifi_data/input/': No such file or directory
(kafeel-env) kafeel@KafeelUbuntuVM:~/kafka$
Broadcast message from root@KafeelUbuntuVM (Sat 2025-08-02 18:50:10 UTC):

The system will power off now!

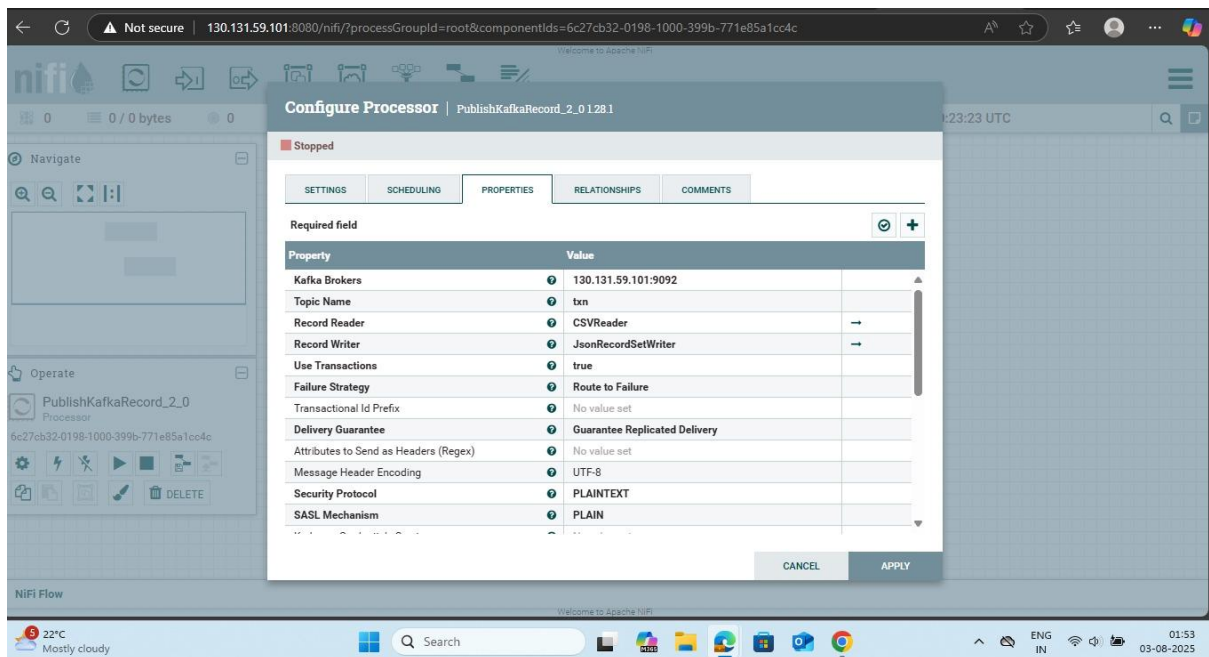
Connection to 130.131.59.101 closed by remote host.
Connection to 130.131.59.101 closed.

C:\Users\kafeel.kamran>
```

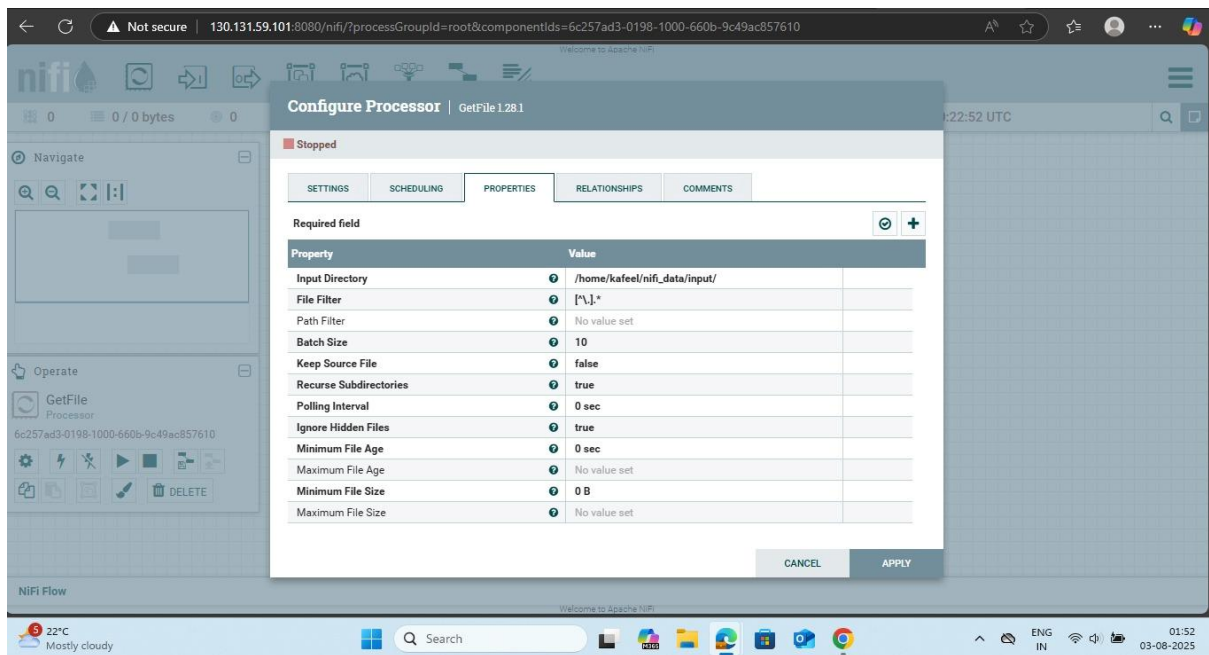
- *Configure and run Nifi*



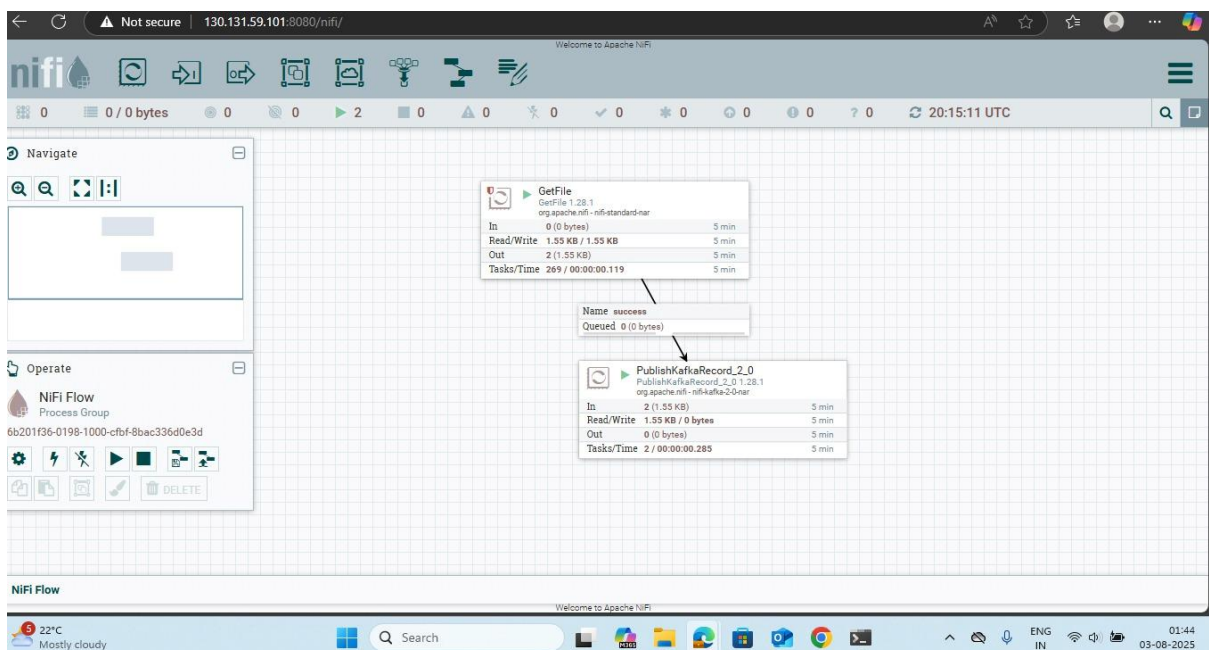
- *KafkaPublish_2.0 Config*



- *Getfile (input dir config)*



- *Nifi Ingestor Processor (Getfile --> PublishKafka_2.0)*



- *Storage Account*

The screenshot displays the Microsoft Azure portal interface for a storage account named 'datastoragekk1'. The left sidebar shows the navigation menu with 'Storage browser' selected. The main pane shows the 'Blob containers' view, listing several containers: 'slogs', 'casestudy', 'deltatable', and 'delta_output'. The 'delta_output' container is selected, showing a list of blobs. The table below lists the blobs with their names, last modified dates, access tiers, blob types, and sizes.

Name	Last modified	Access tier	Blob type	Size
[.]				
._delta_log				
._delta_log	03/08/2025, 01:41:28	Hot (Inferred)	Block blob	0
part-00000...	03/08/2025, 01:41:28	Hot (Inferred)	Block blob	1.72 KiB
part-00000...	03/08/2025, 01:40:21	Hot (Inferred)	Block blob	1.72 KiB