

## Experiment 6

**Aim:** Streaming Analytics with Kafka and Spark: Set up a data streaming pipeline using Apache Kafka to ingest real-time data and process it using Apache Spark Streaming for real-time analytics.

***Starting Services:***

```
C:\Windows\system32\cmd.exe /c Command Prompt - C:\Windows\system32\cmd.exe /c Command Prompt
C:\Kafka\bin>cd C:\Kafka\bin
C:\Kafka\bin>. \windows\zookeeper-server-start.bat .\config\zookeeper.properties
[2024-10-17 18:54:44,986] INFO Reading configuration from: .\config\zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,001] WARN .\config\zookeeper.properties is relative. Prepend .\ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,017] WARN tmp/zookeeper is relative. Prepend .\ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,033] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,033] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,033] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,033] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,040] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DataDirCleanupManager)
[2024-10-17 18:54:45,040] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DataDirCleanupManager)
[2024-10-17 18:54:45,040] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DataDirCleanupManager)
[2024-10-17 18:54:45,040] INFO Either no config or no quorum specified in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2024-10-17 18:54:45,040] INFO Log4j 1.2 jmx support not found; jmx disabled (org.apache.zookeeper.jmx.ManagedUtil)
[2024-10-17 18:54:45,049] INFO Reading configuration from: .\config\zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,049] WARN .\config\zookeeper.properties is relative. Prepend .\ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,059] WARN tmp/zookeeper is relative. Prepend .\ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,059] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,059] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,065] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,065] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2024-10-17 18:54:45,065] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2024-10-17 18:54:45,112] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@30ee2816 (org.apache.zookeeper.server.ServerMetrics)
[2024-10-17 18:54:45,144] INFO ACL digest algorithm is: SHA1 (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2024-10-17 18:54:45,144] INFO zookeeper.DigestAuthenticationProvider.enabled = true (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2024-10-17 18:54:45,159] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2024-10-17 18:54:45,191] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,191] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,191] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,191] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,207] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,207] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,207] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,207] INFO (org.apache.zookeeper.server.ZooKeeperServer) (org.apache.zookeeper.server.ZooKeeperServer)
[2024-10-17 18:54:45,207] INFO Server environment:zookeeper-versions3.8.4-9316c2a797e166dbf4593f34dd6fc36cc436c, built on 2024-02-12 22:16 UTC (org.apache.zookeeper.server.ZooKeeperServer)
```

**Fig 1: Zookeeper Server**

```
C:\Users\rawat>cd C:/Kafka/bin
C:\Kafka\bin>. \windows\kafka-server-start.bat ..\config\server.properties
[2024-10-17 18:54:56,878] INFO Registered KafkaTypepeKafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2024-10-17 18:54:56,986] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util
1)
[2024-10-17 18:54:56,986] INFO RemoteLogManagerConfig values:
  log.local.retention.bytes = -2
  log.local.retention.ms = -2
  remote.fetch.max.wait.ms = 500
  remote.log.index.file.cache.total.size.bytes = 1073741824
  remote.log.manager.copier.thread.pool.size = 10
  remote.log.manager.copy.max.bytes.per.second = 9223372036854775807
  remote.log.manager.copy.quota.window.num = 11
  remote.log.manager.copy.quota.window.size.seconds = 1
  remote.log.manager.expiration.thread.pool.size = 10
  remote.log.manager.fetch.max.bytes.per.second = 9223372036854775807
  remote.log.manager.fetch.quota.window.num = 11
  remote.log.manager.fetch.quota.window.size.seconds = 1
  remote.log.manager.task.interval.ms = 30000
  remote.log.manager.task.retry.backoff.max.ms = 30000
  remote.log.manager.task.retry.backoff.ms = 500
  remote.log.manager.task.retry.jitter = 0.2
  remote.log.manager.thread.pool.size = 10
  remote.log.metadata.custom.metadata.max.bytes = 128
  remote.log.metadata.manager.class.name = org.apache.kafka.server.log.remote.metadata.storage.TopicBasedRemoteLogMetadataManager
  remote.log.metadata.manager.class.path = null
  remote.log.metadata.manager.impl.prefix = rlm.config.
  remote.log.metadata.manager.listener.name = null
  remote.log.reader.max.pending.tasks = 100
  remote.log.reader.threads = 10
  remote.log.storage.manager.class.name = null
  remote.log.storage.manager.class.path = null
  remote.log.storage.manager.impl.prefix = rsm.config.
  remote.log.storage.system.enable = false
  (org.apache.kafka.server.log.remote.storage.RemoteLogManagerConfig)
[2024-10-17 18:54:57.042] INFO starting (kgfka.server.KafkaServer)
```

**Fig 2: Kafka Server/Broker**

```
C:\Kafka\bin\windows>. \kafka-console-producer.bat --broker-list localhost:9092 --topic rainfall_data
>{"division": "North", "year": "2023", "jan": 2.5, "feb": 1.5, "mar": 3.0, "apr": 4.0, "may": 5.0, "jun": 6.0, "jul": 7.0, "aug": 8.0, "sep": 9.0, "oct": 10.0, "nov": 11.0, "dec": 12.0, "annual": 78.0}
>
>{"division": "East", "year": "2023", "jan": 3.2, "feb": 2.0, "mar": 4.1, "apr": 5.3, "may": 6.7, "jun": 7.2, "jul": 8.5, "aug": 9.1, "sep": 10.4, "oct": 11.6, "nov": 12.8, "dec": 13.0, "annual": 79.2}
>{"division": "West", "year": "2023", "jan": 1.5, "feb": 1.0, "mar": 2.8, "apr": 3.5, "may": 4.0, "jun": 4.5, "jul": 5.5, "aug": 6.0, "sep": 6.5, "oct": 7.0, "nov": 8.0, "dec": 8.5, "annual": 54.0}
>{"division": "South", "year": "2023", "jan": 4.0, "feb": 3.5, "mar": 5.0, "apr": 6.0, "may": 7.0, "jun": 8.0, "jul": 9.0, "aug": 10.0, "sep": 11.0, "oct": 12.0, "nov": 13.0, "dec": 14.0, "annual": 89.5}
```

**Fig 3: Sending Data to Kafka Streaming using Dictionary line by line**

```
C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 5>python kafka_producer.py
```

**Fig 4: Sending Data to Kafka Streaming using CSV File**

### Spark Code Execution:

```
C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 5>spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.2 spark_streaming.py
:: loading settings :: url = jar:file:/C:/spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: C:\Users\rawat\ivy2\cache
The jars for the packages stored in: C:\Users\rawat\ivy2\jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-c450689e-e0d6-45db-a0f3-cba56eb3e250;1.0
  confs: [default]
    found org.apache.spark#spark-sql-kafka-0-10_2.12;3.5.2 in central
    found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.2 in central
    found org.apache.kafka#kafka-clients;3.4.1 in central
    found org.lz4#lz4-java;1.8.0 in central
    found org.xerial.snappy#snappy-java;1.1.10.5 in central
    found org.slf4j#slf4j-api;2.0.7 in central
    found org.apache.hadoop#hadoop-client-runtime;3.3.4 in central
    found org.apache.hadoop#hadoop-client-api;3.3.4 in central
    found commons-logging#commons-logging;1.1.3 in central
    found com.google.code.findbugs#jsr305;3.0.0 in central
    found org.apache.commons#commons-pool2;2.11.1 in central
  :: resolution report :: resolve 708ms :: artifacts dl 110ms
  :: modules in use:
    com.google.code.findbugs#jsr305;3.0.0 from central in [default]
    commons-logging#commons-logging;1.1.3 from central in [default]
    org.apache.commons#commons-pool2;2.11.1 from central in [default]
    org.apache.hadoop#hadoop-client-api;3.3.4 from central in [default]
    org.apache.hadoop#hadoop-client-runtime;3.3.4 from central in [default]
```

**Fig 5: Execution of Spark File/Code**

```
org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.2 from central in [default]
org.lz4#lz4-java;1.8.0 from central in [default]
org.slf4j#slf4j-api;2.0.7 from central in [default]
org.xerial.snappy#snappy-java;1.1.10.5 from central in [default]
-----
|               | modules | artifacts |
|               | search|downlded|evicted|| number|downlded|
|-----|-----|-----|
| default      | 11    | 0         | 0       || 11    | 0       |
|-----|-----|-----|
retrieving :: org.apache.spark#spark-submit-parent-0dc1a2df-4075-44ca-aaca-e9eef6e2fd3d
  confs: [default]
  0 artifacts copied, 11 already retrieved (0kB/35ms)
17 19:12:00 INFO SparkContext: Running Spark version 3.5.2
17 19:12:00 INFO SparkContext: OS info Windows 11 10.0 amd64
```

**Fig 6: Output of Execution of Spark File/Code showing installation of all required modules**

```
-----
Batch: 1
-----
+-----+-----+-----+-----+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+-----+-----+-----+-----+
| NULL   | 1    | NULL          | NULL          | NULL          |
+-----+-----+-----+-----+-----+

Batch: 2
-----
+-----+-----+-----+-----+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+-----+-----+-----+-----+
| NULL   | 1    | 78.0          | 78.0          | 78.0          |
| North  | 1    | 78.0          | 78.0          | 78.0          |
+-----+-----+-----+-----+-----+

Batch: 3
-----
+-----+-----+-----+-----+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+-----+-----+-----+-----+
| NULL   | 1    | NULL          | NULL          | NULL          |
| North  | 2    | 78.0          | 78.0          | 78.0          |
+-----+-----+-----+-----+-----+

Batch: 4
-----
+-----+-----+-----+-----+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+-----+-----+-----+-----+
| NULL   | 2    | NULL          | NULL          | NULL          |
| North  | 2    | 78.0          | 78.0          | 78.0          |
+-----+-----+-----+-----+-----+
```

**Fig 7: Batch Execution of Spark Code showing batch 1-4**

```

+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| North | 2 | 78.0 | 78.0 | 78.0 |
+-----+
Batch: 5
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| North | 2 | 78.0 | 78.0 | 78.0 |
+-----+
Batch: 6
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| West | 1 | 54.0 | 54.0 | 54.0 |
| North | 2 | 78.0 | 78.0 | 78.0 |
+-----+

```

Fig 8: Batch Execution of Spark Code showing batch 5-6

```

Batch: 7
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| South | 2 | 89.5 | 89.5 | 89.5 |
| Central | 1 | 66.5 | 66.5 | 66.5 |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| West | 1 | 54.0 | 54.0 | 54.0 |
| North | 2 | 78.0 | 78.0 | 78.0 |
+-----+
Batch: 8
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| South | 2 | 89.5 | 89.5 | 89.5 |
| Central | 1 | 66.5 | 66.5 | 66.5 |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| West | 1 | 54.0 | 54.0 | 54.0 |
| North | 3 | 78.0 | 78.0 | 78.0 |
+-----+

```

Fig 9: Batch Execution of Spark Code showing batch 7-8

```

Batch: 8
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| South | 2 | 89.5 | 89.5 | 89.5 |
| Central | 1 | 66.5 | 66.5 | 66.5 |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| West | 1 | 54.0 | 54.0 | 54.0 |
| North | 3 | 78.0 | 78.0 | 78.0 |
+-----+
Batch: 9
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| NULL | 2 | NULL | NULL | NULL |
| South | 2 | 89.5 | 89.5 | 89.5 |
| ANDAMAN & NICOBAR... | 51 | 2352.1 | NaN | NaN |
| Central | 1 | 66.5 | 66.5 | 66.5 |
| East | 1 | 79.2 | 79.2 | 79.19999694824219 |
| West | 1 | 54.0 | 54.0 | 54.0 |
| North | 3 | 78.0 | 78.0 | 78.0 |
+-----+

```

Fig 10: Batch Execution of Spark Code showing batch 8-9

```

Batch: 10
+-----+
|division|count|min_annual_rainfall|max_annual_rainfall|mean_annual_rainfall|
+-----+
| VIDARBHA | 115 | 578.5 | 1606.3 | 1095.4591303286345 |
| NAGA MANI MIZO TR... | 115 | 1353.8 | 4316.2 | 2433.619123641304 |
| CHHATTISGARH | 115 | 904.6 | 1974.0 | 1371.7286875849186 |
| NULL | 2 | NULL | NULL | NULL |
| SUB HIMALAYAN WES... | 115 | 1988.2 | 3655.1 | 2752.2173955502717 |
| GANGETIC WEST BENGAL | 115 | 1015.1 | 2099.8 | 1490.4878226902174 |
| HIMACHAL PRADESH | 115 | 776.1 | 1919.2 | 1260.3452153744904 |
| BIHAR | 115 | 629.2 | 1660.4 | 1197.63390847911 |
| ORISSA | 115 | 987.0 | 1945.3 | 1458.1695694633152 |
| JAMMU & KASHMIR | 115 | 657.0 | NaN | NaN |
| ASSAM & MEGHALAYA | 115 | 1743.4 | 3403.5 | 2580.695658542799 |
| South | 2 | 89.5 | 89.5 | 89.5 |
| LAKSHADWEEP | 114 | 992.6 | NaN | NaN |
| ANDAMAN & NICOBAR... | 110 | 1849.4 | NaN | NaN |
| TAMIL NADU | 115 | 318.0 | 1365.3 | 943.7130437436311 |
| Central | 1 | 66.5 | 66.5 | 66.5 |
| NORTH INTERIOR KA... | 115 | 470.3 | 1095.6 | 717.7956508470618 |
| WEST UTTAR PRADESH | 115 | 371.9 | 1244.2 | 827.1147813879925 |
| SOUTH INTERIOR KA... | 115 | 733.3 | 1409.5 | 1040.3913027556046 |
| EAST MADHYA PRADESH | 115 | 653.8 | 1747.1 | 1205.000001061481 |
+-----+
only showing top 20 rows

```

Fig 11: Batch Execution of Spark Code showing batch 10