BDA Project
Sameer Gupta
211080071
Batch D

Real-Time Big Data Analytics Pipeline Experiment Report

This experiment focuses on building a real-time data analytics pipeline using various Apache components such as Kafka, Spark, Hive, Pig, HBase, and Zookeeper. The project demonstrates the flow of data from ingestion to processing, storage, and querying. Each task represents a significant stage of the pipeline and collectively provides an end-to-end understanding of real-time big data architecture.

Task 1: Setting Up the Environment

Theory

Before working with real-time big data technologies, we must prepare the environment by installing essential components. Each of these tools plays a specific role in the big data ecosystem:

- **Kafka** acts as a distributed messaging system used for data ingestion.

- **Spark** enables real-time and batch data processing

.

- **Hive** is a data warehouse software that facilitates querying and managing large datasets.

- **Pig** is a high-level platform for creating MapReduce programs used with Hadoop.

- **HBase** is a NoSQL database that stores non-relational, distributed data.

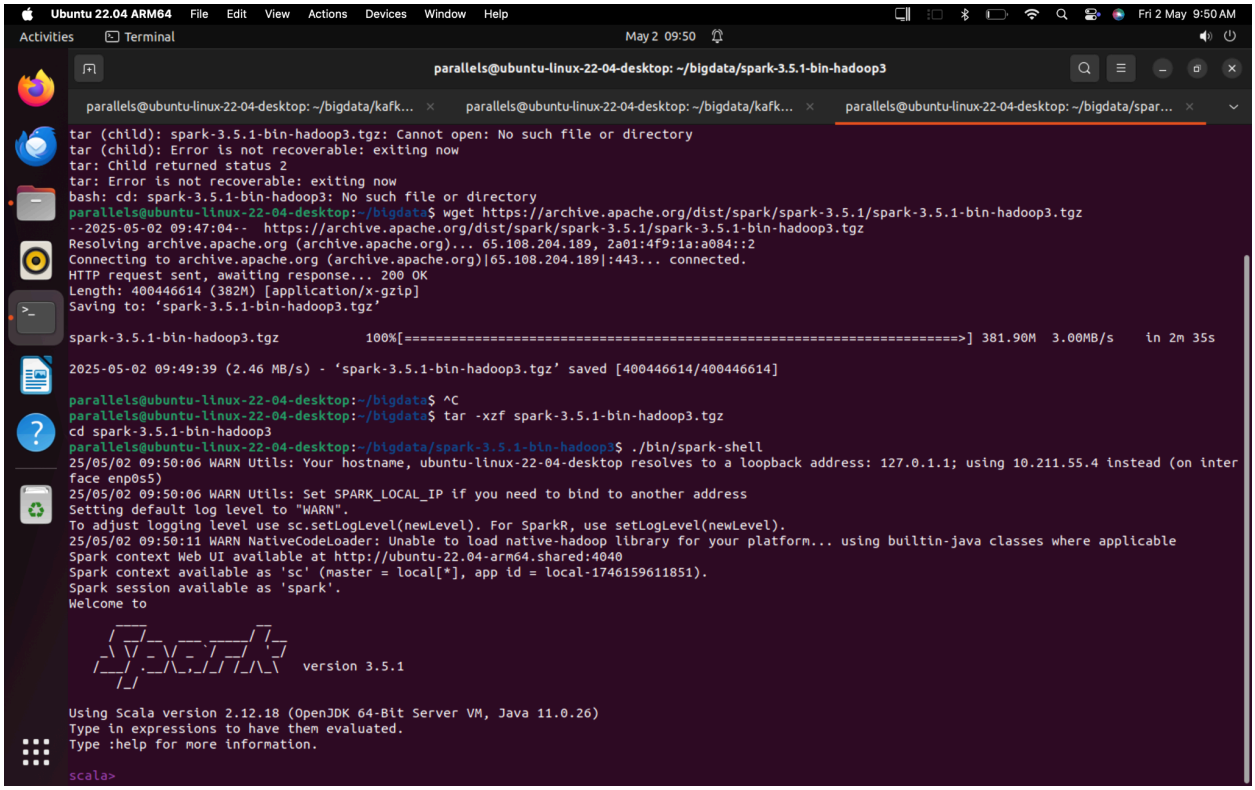
- **Zookeeper** coordinates distributed applications and maintains configuration information.

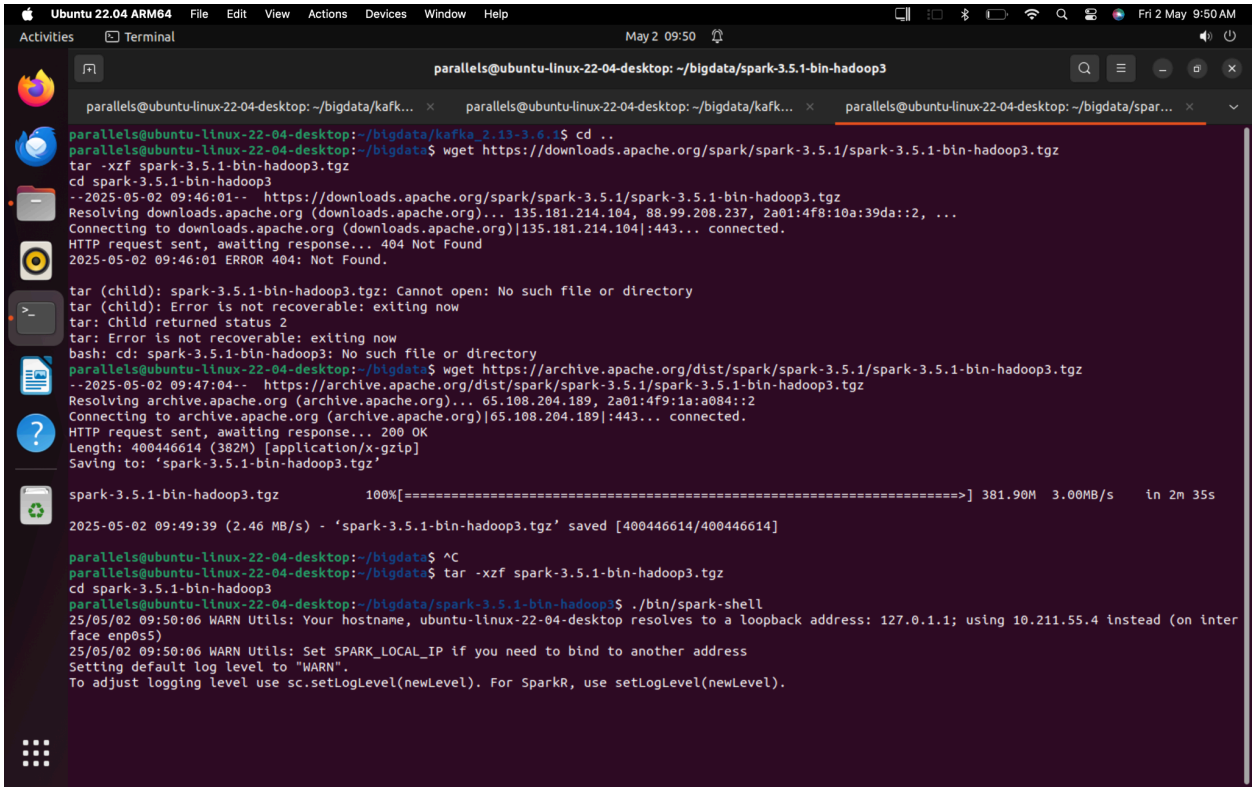
We also install Java, as most of these tools are Java-based, and SSH for secure communication and remote access. Configuring the correct environment paths and Java settings ensures compatibility and prevents runtime errors.

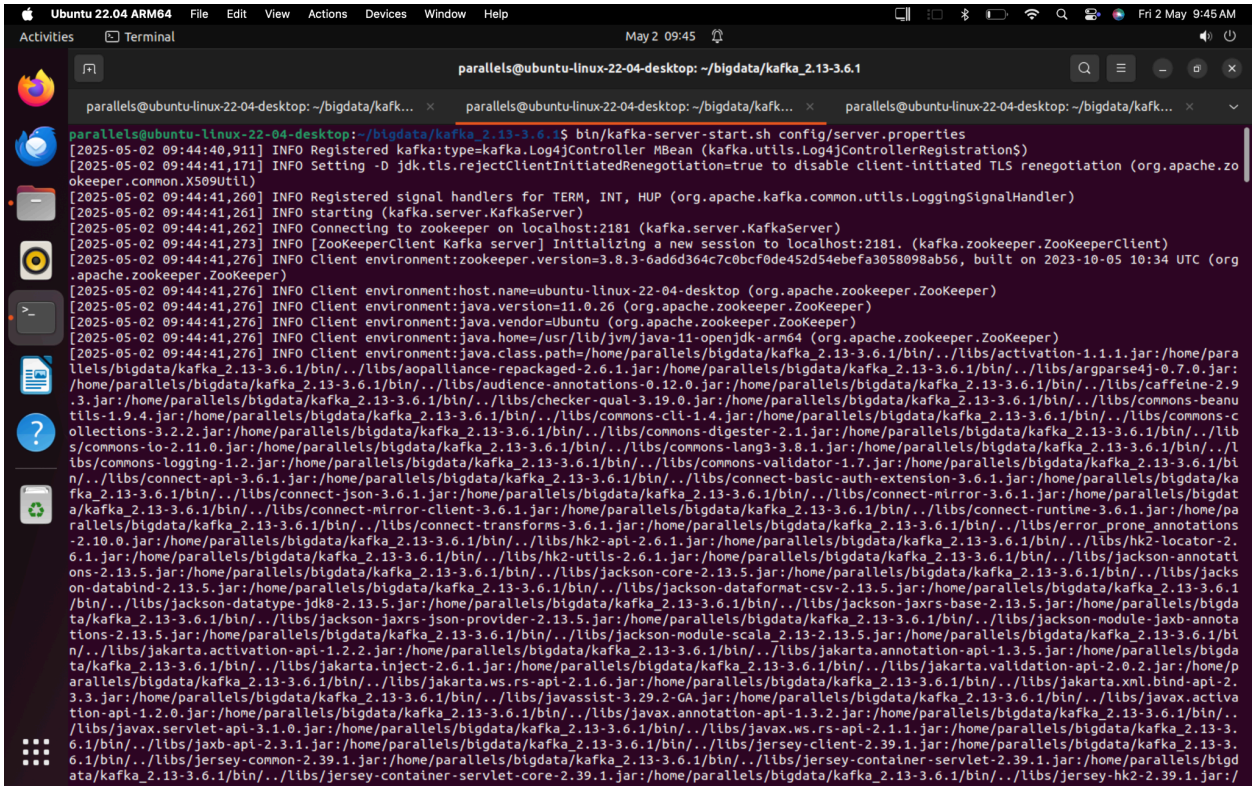
Screenshot:

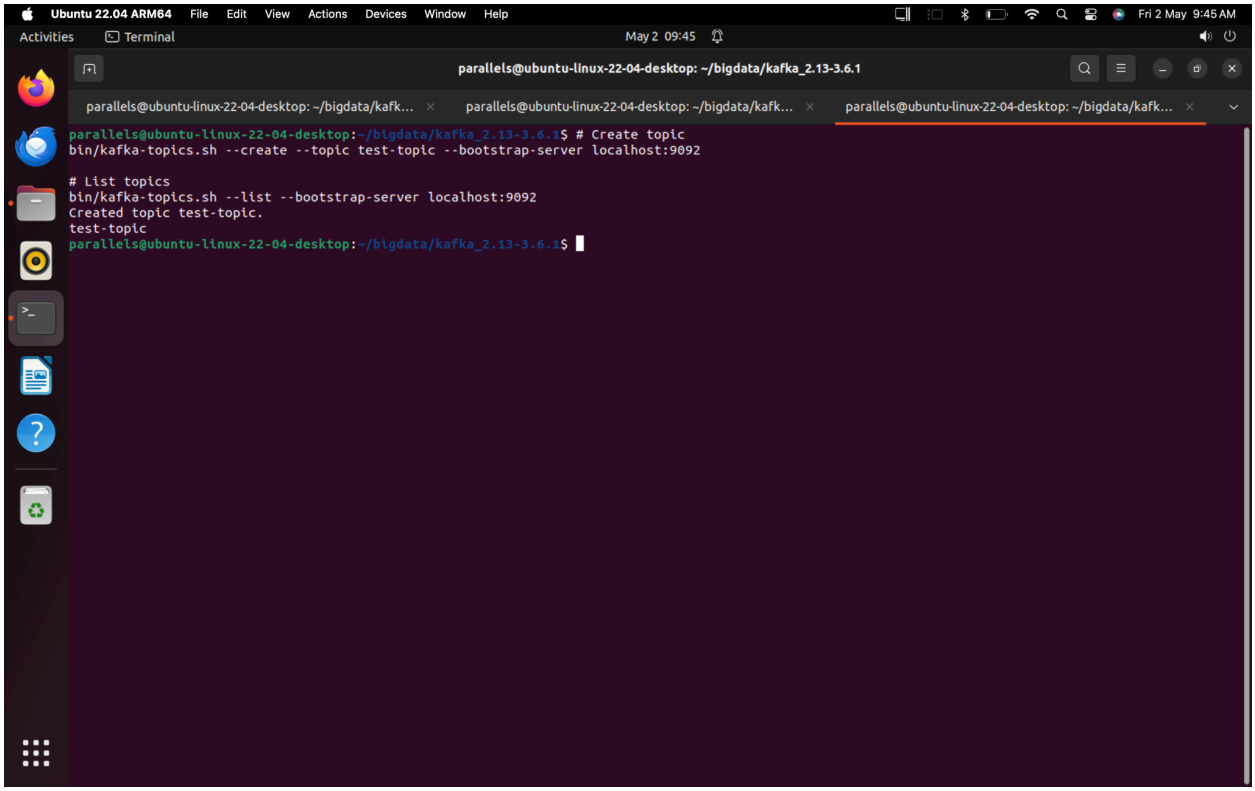
Task 1: Setting Up the Environment

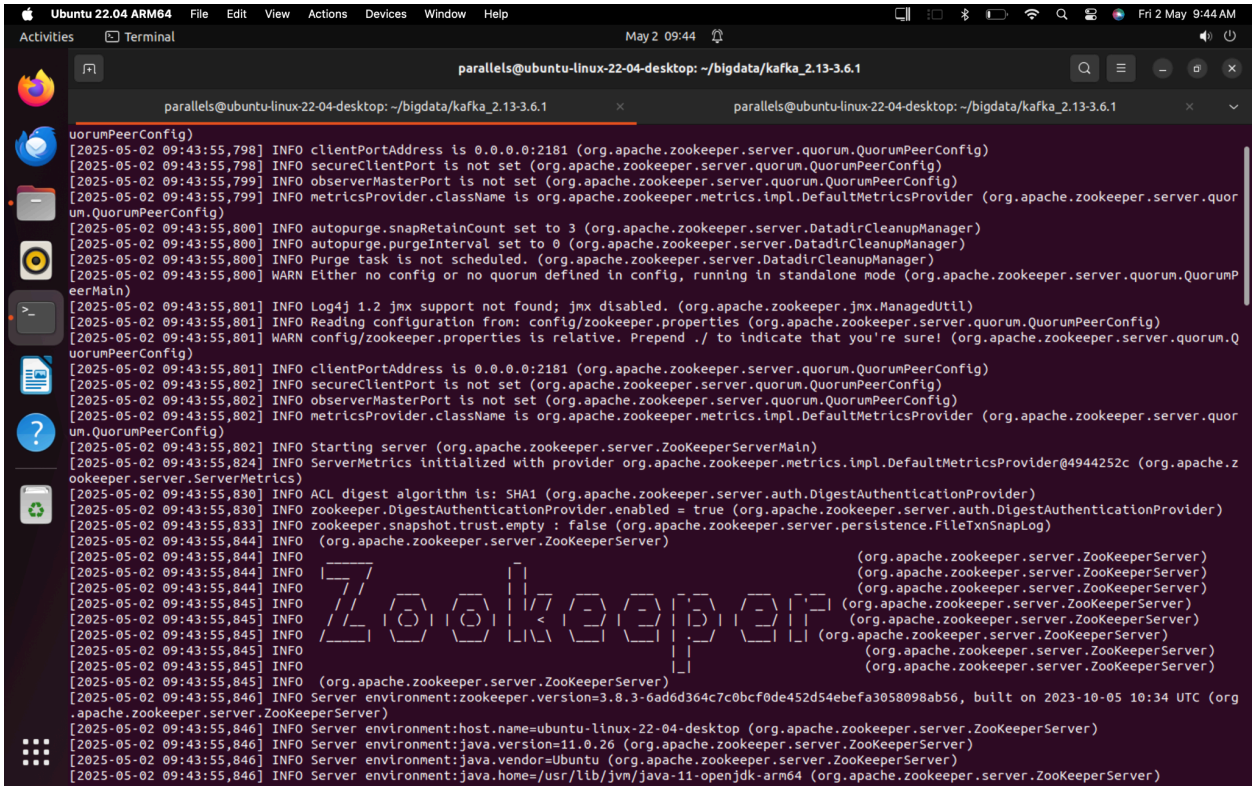
- Install and configure Apache Kafka, Spark, Hive, Pig, HBase, and Zookeeper.
- Verify the installation and ensure services are running.

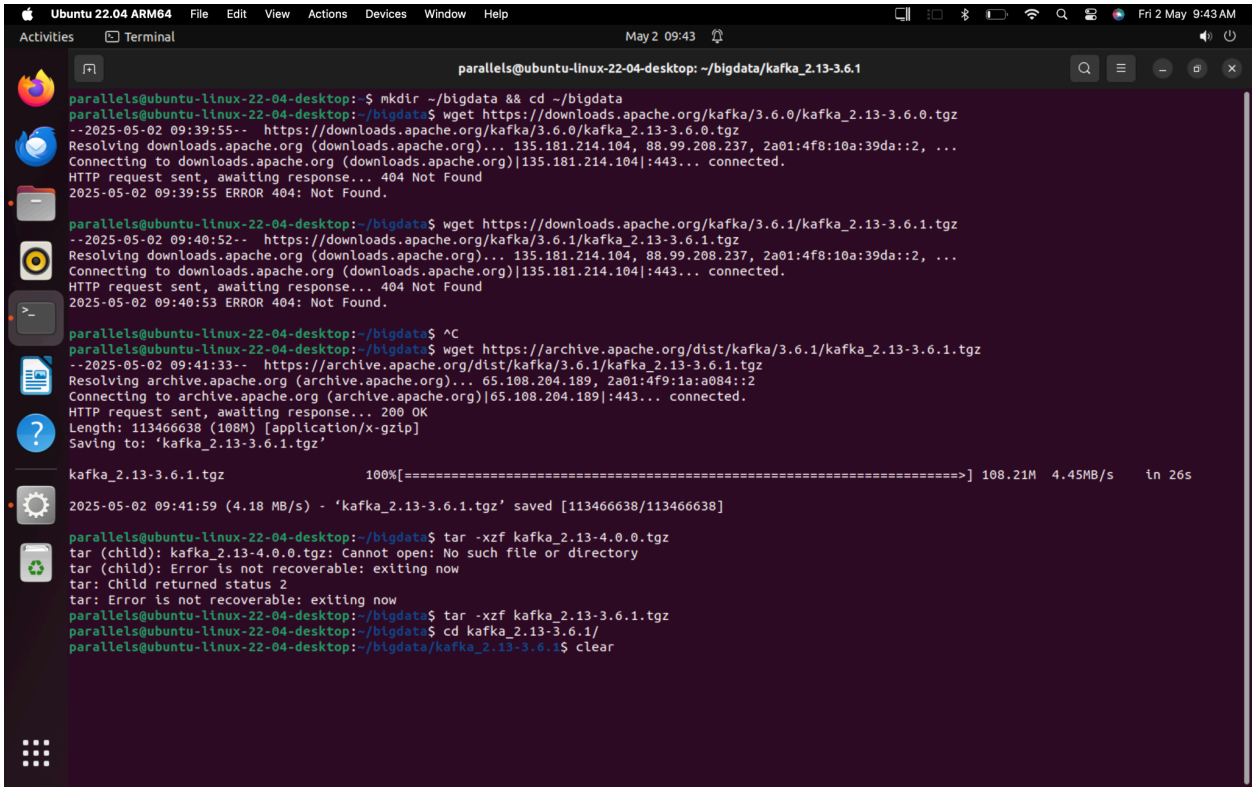

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:50 AM
Activities Terminal
May 2 09:50
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/spark-3.5.1-bin-hadoop3
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/spar...
tar (child): spark-3.5.1-bin-hadoop3.tgz: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
bash: cd: spark-3.5.1-bin-hadoop3: No such file or directory
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
--2025-05-02 09:47:04-- https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400446614 (382M) [application/x-gzip]
Saving to: 'spark-3.5.1-bin-hadoop3.tgz'
spark-3.5.1-bin-hadoop3.tgz 100%[=>] 381.90M 3.00MB/s in 2m 35s
2025-05-02 09:49:39 (2.46 MB/s) - 'spark-3.5.1-bin-hadoop3.tgz' saved [400446614/400446614]
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ ^C
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ tar -xzf spark-3.5.1-bin-hadoop3.tgz
cd spark-3.5.1-bin-hadoop3
parallels@ubuntu-linux-22-04-desktop:~/bigdata/spark-3.5.1-bin-hadoop3$ ./bin/spark-shell
25/05/02 09:50:06 WARN Utils: Your hostname, ubuntu-linux-22-04-desktop resolves to a loopback address: 127.0.1.1; using 10.211.55.4 instead (on interface enp0s5)
25/05/02 09:50:06 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/05/02 09:50:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://ubuntu-22.04-arm64.shared:4040
Spark context available as 'sc' (master = local[*], app id = local-1746159611851).
Spark session available as 'spark'.
Welcome to
version 3.5.1
Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
Type :help for more information.
scala>

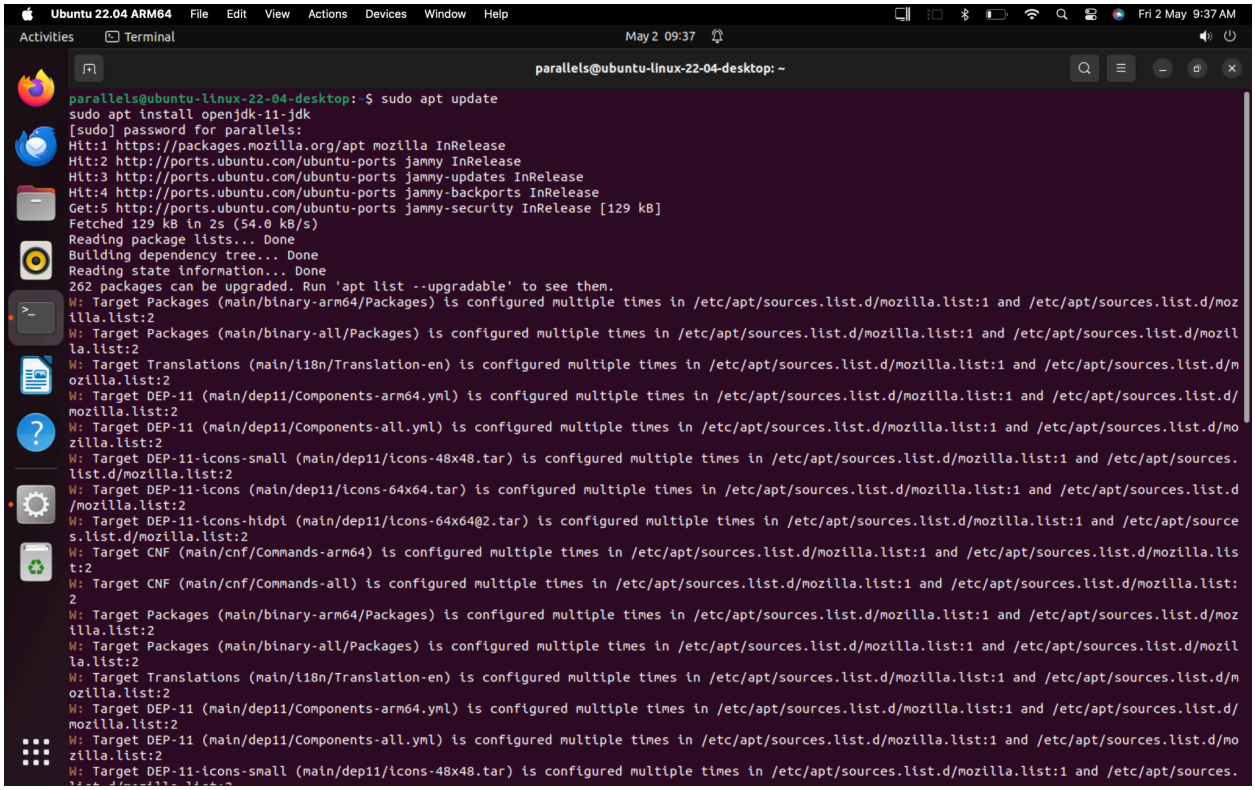

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:50 AM
Activities Terminal
May 2 09:50
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/spark-3.5.1-bin-hadoop3
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/spar...
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ cd ..
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
tar -xzf spark-3.5.1-bin-hadoop3.tgz
cd spark-3.5.1-bin-hadoop3
--2025-05-02 09:46:01-- https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-05-02 09:46:01 ERROR 404: Not Found.
tar (child): spark-3.5.1-bin-hadoop3.tgz: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
bash: cd: spark-3.5.1-bin-hadoop3: No such file or directory
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
--2025-05-02 09:47:04-- https://archive.apache.org/dist/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 400446614 (382M) [application/x-gzip]
Saving to: 'spark-3.5.1-bin-hadoop3.tgz'
spark-3.5.1-bin-hadoop3.tgz 100%[=>] 381.90M 3.00MB/s in 2m 35s
2025-05-02 09:49:39 (2.46 MB/s) - 'spark-3.5.1-bin-hadoop3.tgz' saved [400446614/400446614]
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ ^C
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ tar -xzf spark-3.5.1-bin-hadoop3.tgz
cd spark-3.5.1-bin-hadoop3
parallels@ubuntu-linux-22-04-desktop:~/bigdata/spark-3.5.1-bin-hadoop3$ ./bin/spark-shell
25/05/02 09:50:06 WARN Utils: Your hostname, ubuntu-linux-22-04-desktop resolves to a loopback address: 127.0.1.1; using 10.211.55.4 instead (on interface enp0s5)
25/05/02 09:50:06 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

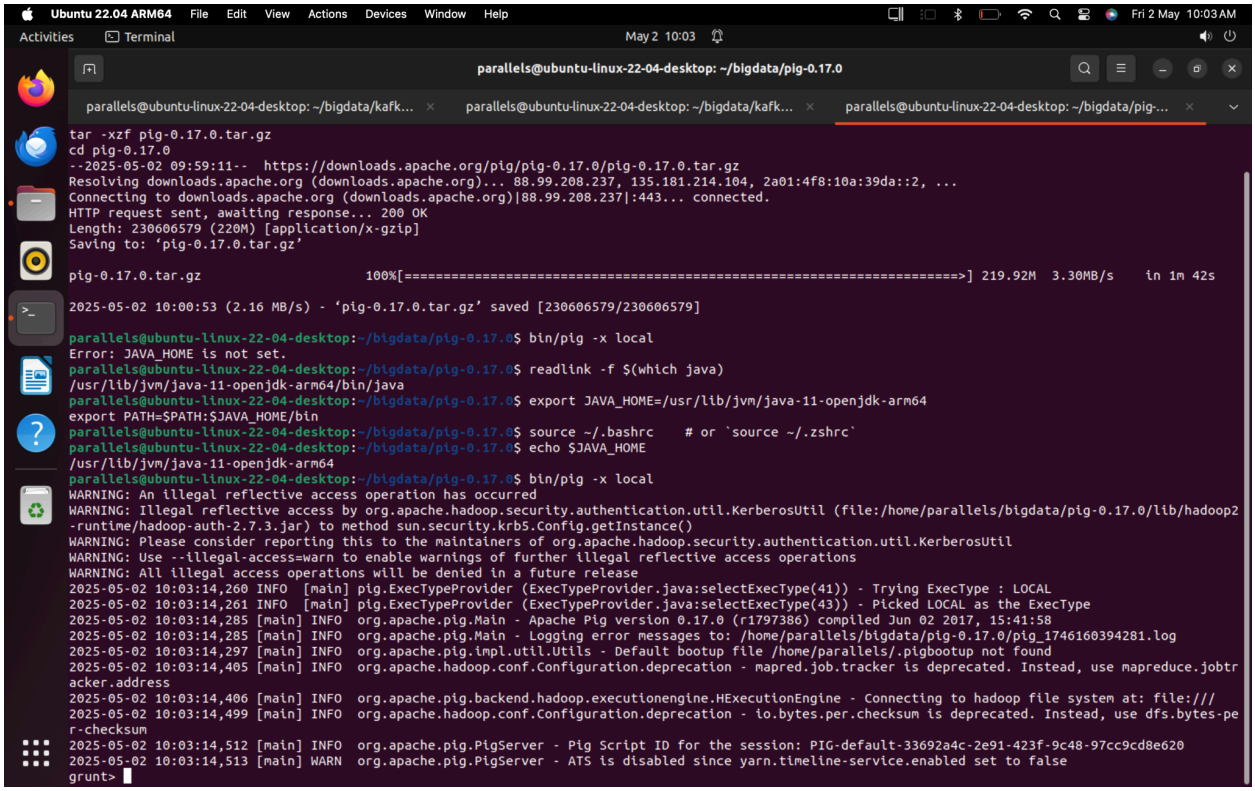

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:45 AM
Activities Terminal
May 2 09:45
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-server-start.sh config/server.properties
[2025-05-02 09:44:40,911] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2025-05-02 09:44:41,171] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2025-05-02 09:44:41,260] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2025-05-02 09:44:41,261] INFO starting (kafka.server.KafkaServer)
[2025-05-02 09:44:41,262] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2025-05-02 09:44:41,273] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2025-05-02 09:44:41,276] INFO Client environment:zookeeper.version=3.8.3-6ad6d364c7c0bcf0de452d54ebefa3058098ab56, built on 2023-10-05 10:34 UTC (org.apache.zookeeper.ZooKeeper)
[2025-05-02 09:44:41,276] INFO Client environment:host.name=ubuntu-linux-22-04-desktop (org.apache.zookeeper.ZooKeeper)
[2025-05-02 09:44:41,276] INFO Client environment:java.version=11.0.26 (org.apache.zookeeper.ZooKeeper)
[2025-05-02 09:44:41,276] INFO Client environment:java.vendor=Ubuntu (org.apache.zookeeper.ZooKeeper)
[2025-05-02 09:44:41,276] INFO Client environment:java.home=/usr/lib/jvm/java-11-openjdk-arm64 (org.apache.zookeeper.ZooKeeper)
[2025-05-02 09:44:41,276] INFO Client environment:java.class.path=/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/activation-1.1.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/aopalliance-repackaged-2.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/argparse4j-0.7.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/audience-annotations-0.12.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/caffeine-2.9.3.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/checker-qual-3.19.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-beanutils-1.9.4.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-cli-1.4.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-collections-3.2.2.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-digester-2.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-io-2.11.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-lang3-3.8.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-logging-1.2.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/commons-validator-1.7.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-api-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-basic-auth-extension-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-json-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-mirror-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-mirror-client-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-runtime-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/connect-transforms-3.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/error_prone_annotations-2.10.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/hk2-api-2.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/hk2-locator-2.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/hk2-utils-2.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-annotations-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-core-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-databind-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-dataformat-csv-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-datatype-jdk8-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-jaxrs-base-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-jaxrs-json-provider-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-module-jaxb-annotations-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jackson-module-scala_2.13-2.13.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.activation-api-1.2.2.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.annotation-api-1.3.5.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.inject-2.6.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.validation-api-2.0.2.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.ws.rs-api-2.1.6.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jakarta.xml.bind-api-2.3.3.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/javassist-3.29.2-GA.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/javax.activation-api-1.2.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/javax.annotation-api-1.3.2.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/javax.servlet-api-3.1.0.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/javax.ws.rs-api-2.1.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jaxb-api-2.3.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jersey-client-2.39.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jersey-common-2.39.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jersey-container-servlet-2.39.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jersey-container-servlet-core-2.39.1.jar:/home/parallels/bigdata/kafka_2.13-3.6.1/bin/../libs/jersey-hk2-2.39.1.jar:/

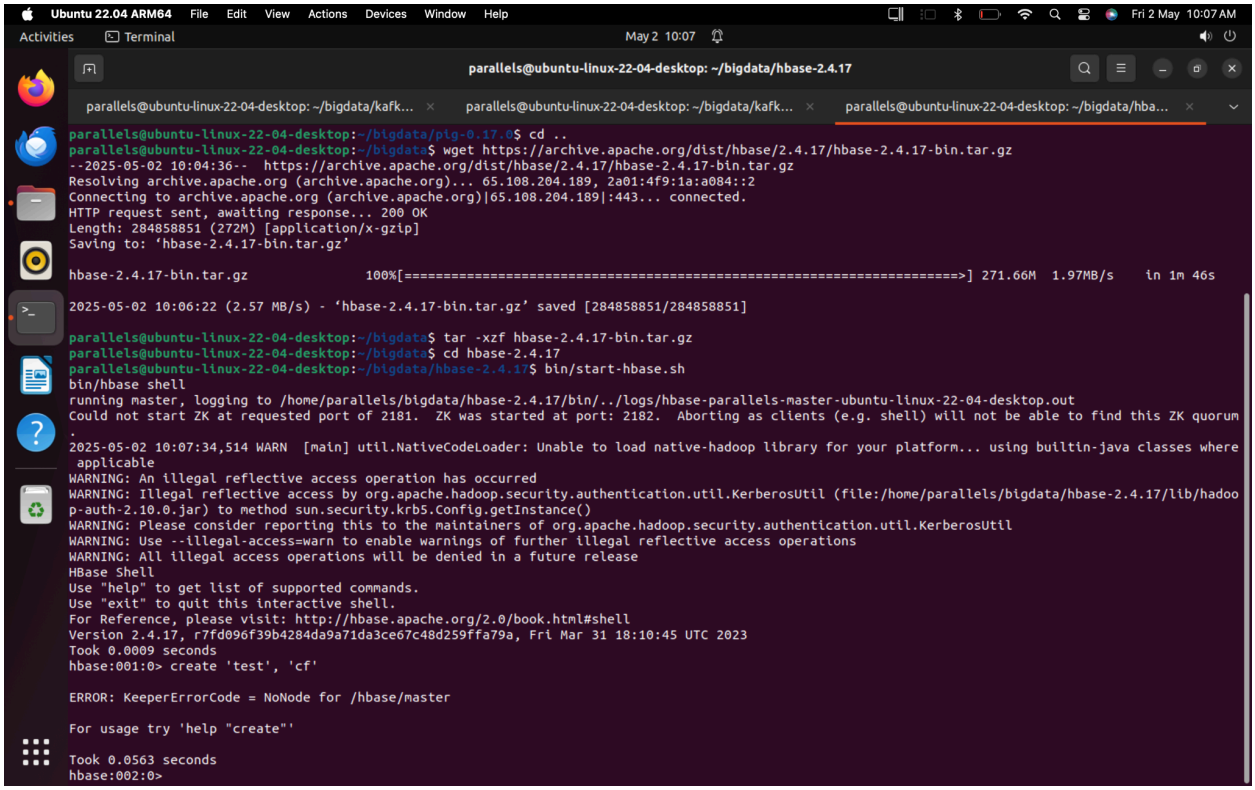

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:45 AM
Activities Terminal
May 2 09:45
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ # Create topic
bin/kafka-topics.sh --create --topic test-topic --bootstrap-server localhost:9092
List topics
bin/kafka-topics.sh --list --bootstrap-server localhost:9092
Created topic test-topic.
test-topic
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:44 AM
Activities Terminal May 2 09:44
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
uorumPeerConfig)
[2025-05-02 09:43:55,798] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,798] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,799] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,799] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,800] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-05-02 09:43:55,800] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-05-02 09:43:55,800] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-05-02 09:43:55,800] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2025-05-02 09:43:55,801] INFO Log4j 1.2 jmx support not found; jmx disabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2025-05-02 09:43:55,801] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,801] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,801] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,802] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,802] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,802] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-05-02 09:43:55,802] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2025-05-02 09:43:55,824] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@4944252c (org.apache.zookeeper.server.ServerMetrics)
[2025-05-02 09:43:55,830] INFO ACL digest algorithm is: SHA1 (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2025-05-02 09:43:55,830] INFO zookeeper.DigestAuthenticationProvider.enabled = true (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2025-05-02 09:43:55,833] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2025-05-02 09:43:55,844] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,844] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,844] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,844] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,845] INFO (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,846] INFO Server environment:zookeeper.version=3.8.3-6ad6d364c7c0bcf0de452d54ebefa3058098ab56, built on 2023-10-05 10:34 UTC (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,846] INFO Server environment:host.name=ubuntu-linux-22-04-desktop (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,846] INFO Server environment:java.version=11.0.26 (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,846] INFO Server environment:java.vendor=Ubuntu (org.apache.zookeeper.server.ZooKeeperServer)
[2025-05-02 09:43:55,846] INFO Server environment:java.home=/usr/lib/jvm/java-11-openjdk-arm64 (org.apache.zookeeper.server.ZooKeeperServer)


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:43 AM
Activities Terminal May 2 09:43
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop:~$ mkdir ~/bigdata && cd ~/bigdata
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://downloads.apache.org/kafka/3.6.0/kafka_2.13-3.6.0.tgz
--2025-05-02 09:39:55-- https://downloads.apache.org/kafka/3.6.0/kafka_2.13-3.6.0.tgz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-05-02 09:39:55 ERROR 404: Not Found.
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://downloads.apache.org/kafka/3.6.1/kafka_2.13-3.6.1.tgz
--2025-05-02 09:40:52-- https://downloads.apache.org/kafka/3.6.1/kafka_2.13-3.6.1.tgz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.208.237, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2025-05-02 09:40:53 ERROR 404: Not Found.
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ ^C
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ wget https://archive.apache.org/dist/kafka/3.6.1/kafka_2.13-3.6.1.tgz
--2025-05-02 09:41:33-- https://archive.apache.org/dist/kafka/3.6.1/kafka_2.13-3.6.1.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 113466638 (108M) [application/x-gzip]
Saving to: ‘kafka_2.13-3.6.1.tgz’
kafka_2.13-3.6.1.tgz 100%[=>] 108.21M 4.45MB/s in 26s
2025-05-02 09:41:59 (4.18 MB/s) - ‘kafka_2.13-3.6.1.tgz’ saved [113466638/113466638]
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ tar -xzf kafka_2.13-4.0.0.tgz
tar (child): kafka_2.13-4.0.0.tgz: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ tar -xzf kafka_2.13-3.6.1.tgz
parallels@ubuntu-linux-22-04-desktop:~/bigdata$ cd kafka_2.13-3.6.1/
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ clear

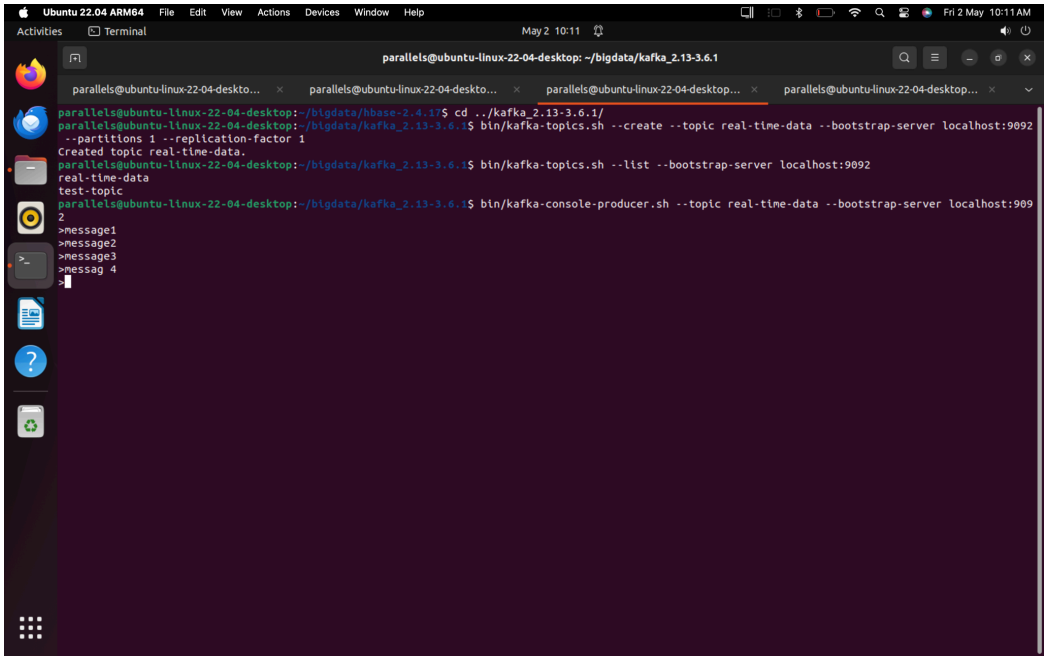

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 9:37 AM
Activities Terminal
May 2 09:37
parallels@ubuntu-linux-22-04-desktop: ~
parallels@ubuntu-linux-22-04-desktop:~$ sudo apt update
sudo apt install openjdk-11-jdk
[sudo] password for parallels:
Hit:1 https://packages.mozilla.org/apt mozilla InRelease
Hit:2 http://ports.ubuntu.com/ubuntu-ports jammy InRelease
Hit:3 http://ports.ubuntu.com/ubuntu-ports jammy-updates InRelease
Hit:4 http://ports.ubuntu.com/ubuntu-ports jammy-backports InRelease
Get:5 http://ports.ubuntu.com/ubuntu-ports jammy-security InRelease [129 kB]
Fetched 129 kB in 2s (54.0 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
262 packages can be upgraded. Run 'apt list --upgradable' to see them.
W: Target Packages (main/binary-arm64/Packages) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target Packages (main/binary-all/Packages) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target Translations (main/i18n/Translation-en) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11 (main/dep11/Components-arm64.yml) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11 (main/dep11/Components-all.yml) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11-icons-small (main/dep11/icons-48x48.tar) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11-icons (main/dep11/icons-64x64.tar) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11-icons-hidpi (main/dep11/icons-64x64@2.tar) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target CNF (main/cnf/Commands-arm64) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target CNF (main/cnf/Commands-all) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target Packages (main/binary-arm64/Packages) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target Packages (main/binary-all/Packages) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target Translations (main/i18n/Translation-en) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11 (main/dep11/Components-arm64.yml) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11 (main/dep11/Components-all.yml) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.list.d/mozilla.list:2
W: Target DEP-11-icons-small (main/dep11/icons-48x48.tar) is configured multiple times in /etc/apt/sources.list.d/mozilla.list:1 and /etc/apt/sources.

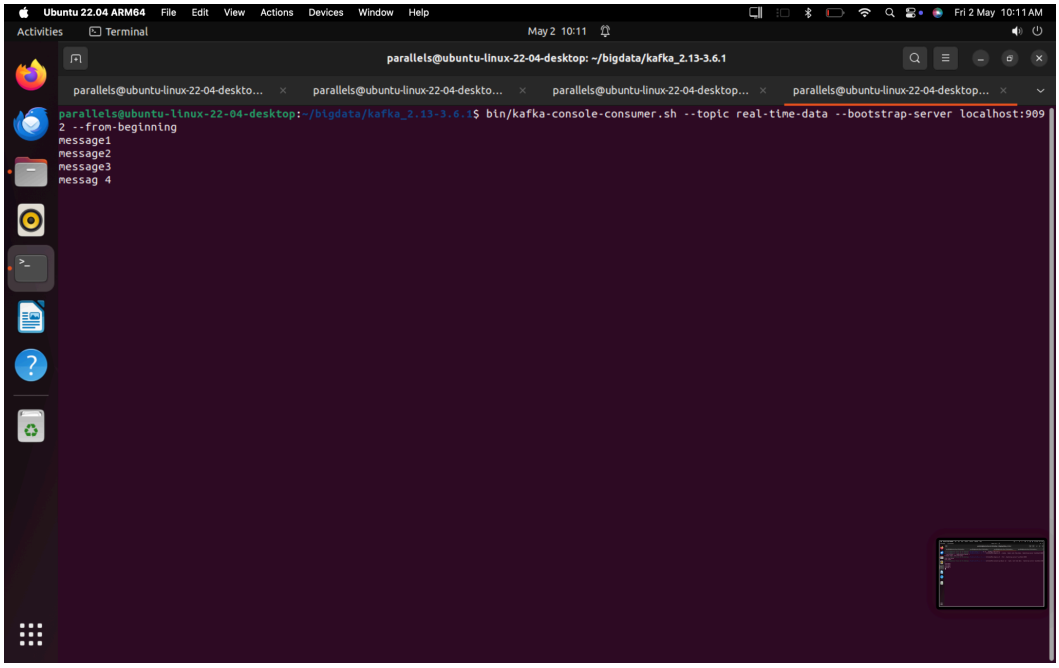

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:03 AM
Activities Terminal
May 2 10:03
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/pig-0.17.0
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafk...
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/pig-...
tar -xzf pig-0.17.0.tar.gz
cd pig-0.17.0
--2025-05-02 09:59:11-- https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.208.237, 135.181.214.104, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.208.237|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 230606579 (220M) [application/x-gzip]
Saving to: 'pig-0.17.0.tar.gz'
pig-0.17.0.tar.gz 100%[==========>] 219.92M 3.30MB/s in 1m 42s
2025-05-02 10:00:53 (2.16 MB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ bin/pig -x local
Error: JAVA_HOME is not set.
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ readlink -f $(which java)
/usr/lib/jvm/java-11-openjdk-arm64/bin/java
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-arm64
export PATH=$PATH:$JAVA_HOME/bin
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ source ~/.bashrc # or `source ~/.zshrc`
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-arm64
parallels@ubuntu-linux-22-04-desktop:~/bigdata/pig-0.17.0$ bin/pig -x local
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/parallels/bigdata/pig-0.17.0/lib/hadoop2-runtime/hadoop-auth-2.7.3.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2025-05-02 10:03:14,260 INFO [main] pig.ExecTypeProvider (ExecTypeProvider.java:selectExecType(41)) - Trying ExecType : LOCAL
2025-05-02 10:03:14,261 INFO [main] pig.ExecTypeProvider (ExecTypeProvider.java:selectExecType(43)) - Picked LOCAL as the ExecType
2025-05-02 10:03:14,285 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2025-05-02 10:03:14,285 [main] INFO org.apache.pig.Main - Logging error messages to: /home/parallels/bigdata/pig-0.17.0/pig_1746160394281.log
2025-05-02 10:03:14,297 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/parallels/.pigbootup not found
2025-05-02 10:03:14,405 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2025-05-02 10:03:14,406 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2025-05-02 10:03:14,499 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-05-02 10:03:14,512 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-33692a4c-2e91-423f-9c48-97cc9cd8e620
2025-05-02 10:03:14,513 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>

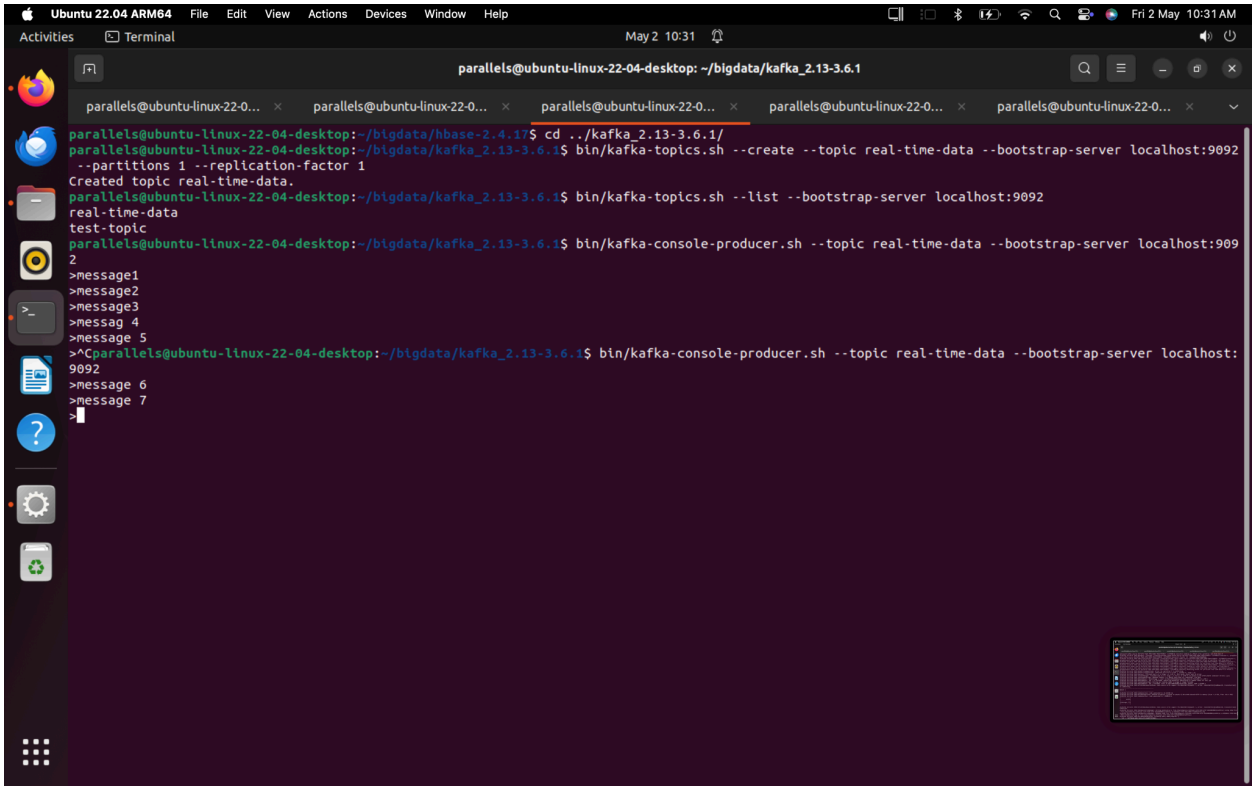


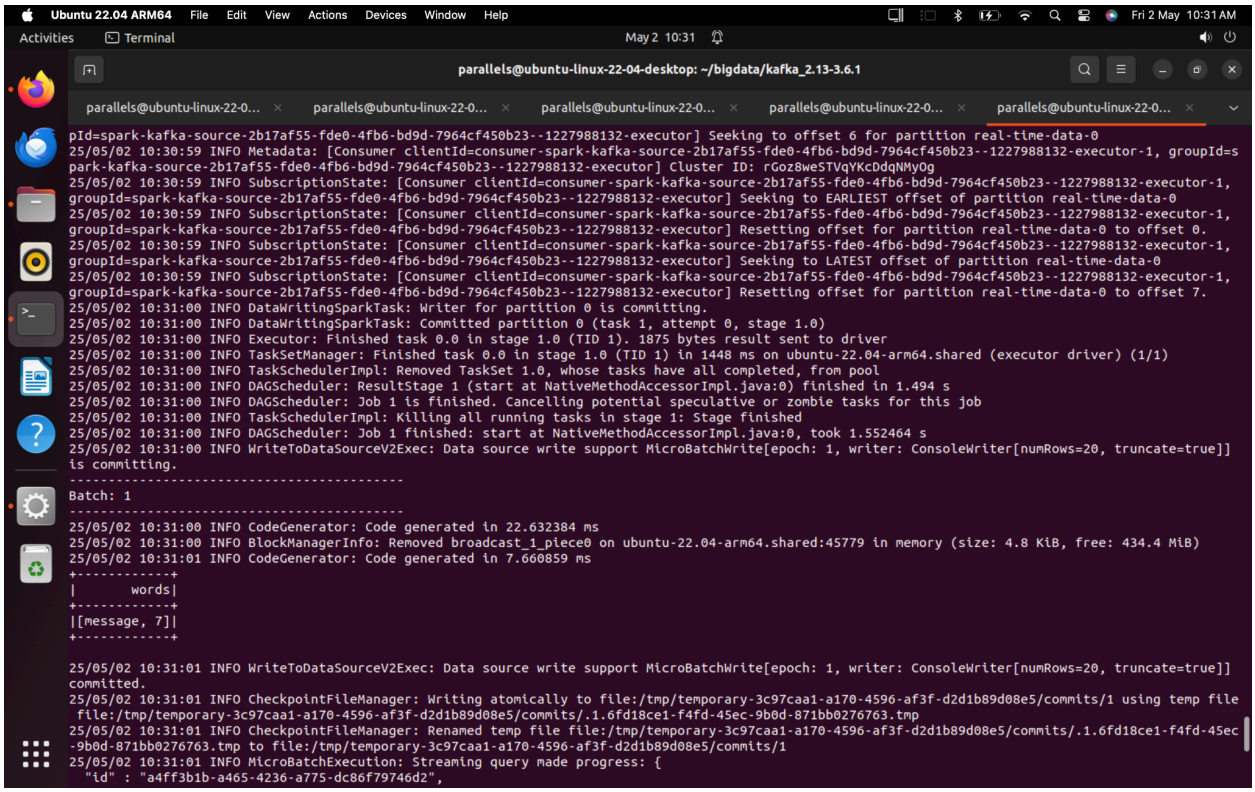
Task 2: Implementing Real-Time Data Ingestion

- Set up Kafka producers to generate real-time data streams.
- Configure Kafka consumers to feed data into Apache Spark Streaming.


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:11 AM
Activities
Terminal
May 2 10:11
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-deskto...
parallels@ubuntu-linux-22-04-deskto...
parallels@ubuntu-linux-22-04-desktop...
parallels@ubuntu-linux-22-04-desktop...
parallels@ubuntu-linux-22-04-desktop:~/bigdata/hbase-2.4.17$ cd ../kafka_2.13-3.6.1/
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-topics.sh --create --topic real-time-data --bootstrap-server localhost:9092
--partitions 1 --replication-factor 1
Created topic real-time-data.
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-topics.sh --list --bootstrap-server localhost:9092
real-time-data
test-topic
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-console-producer.sh --topic real-time-data --bootstrap-server localhost:909
2
>message1
>message2
>message3
>messag 4
>


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:11 AM
Activities
Terminal
May 2 10:11
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-deskto...
parallels@ubuntu-linux-22-04-deskto...
parallels@ubuntu-linux-22-04-desktop...
parallels@ubuntu-linux-22-04-desktop...
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-console-consumer.sh --topic real-time-data --bootstrap-server localhost:909
2 --from-beginning
message1
message2
message3
messag 4


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:31 AM
Activities Terminal
May 2 10:31
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-linux-22-04-desktop:~/bigdata/hbase-2.4.17$ cd ../kafka_2.13-3.6.1/
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-topics.sh --create --topic real-time-data --bootstrap-server localhost:9092
--partitions 1 --replication-factor 1
Created topic real-time-data.
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-topics.sh --list --bootstrap-server localhost:9092
real-time-data
test-topic
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-console-producer.sh --topic real-time-data --bootstrap-server localhost:9092
>message1
>message2
>message3
>messag 4
>message 5
>^Cparallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ bin/kafka-console-producer.sh --topic real-time-data --bootstrap-server localhost:9092
>message 6
>message 7
>

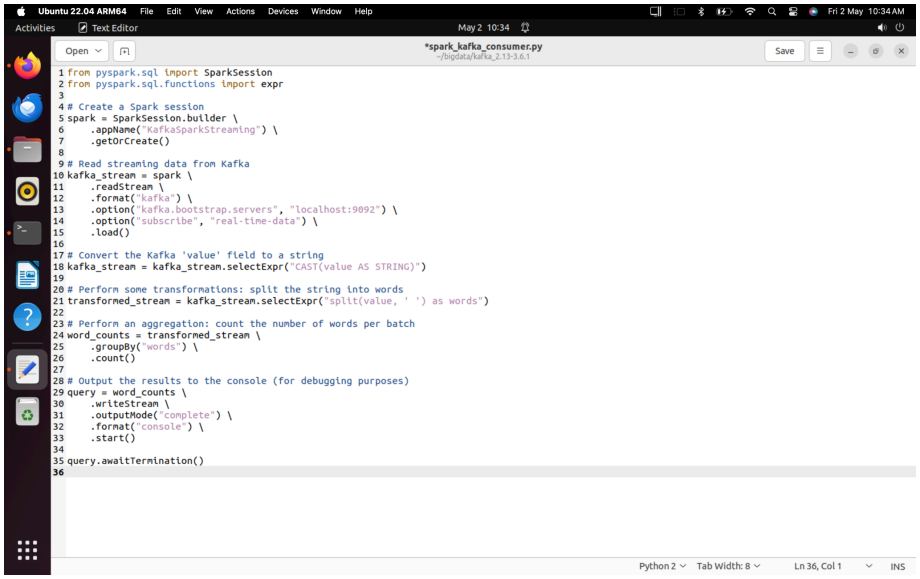

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:31 AM
Activities Terminal
May 2 10:31
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
pId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Seeking to offset 6 for partition real-time-data-0
25/05/02 10:30:59 INFO Metadata: [Consumer clientId=consumer-spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor-1, groupId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Cluster ID: rGoz8weSTVqYKcDdqNMyOg
25/05/02 10:30:59 INFO SubscriptionState: [Consumer clientId=consumer-spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor-1, groupId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Seeking to EARLIEST offset of partition real-time-data-0
25/05/02 10:30:59 INFO SubscriptionState: [Consumer clientId=consumer-spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor-1, groupId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Resetting offset for partition real-time-data-0 to offset 0.
25/05/02 10:30:59 INFO SubscriptionState: [Consumer clientId=consumer-spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor-1, groupId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Seeking to LATEST offset of partition real-time-data-0
25/05/02 10:30:59 INFO SubscriptionState: [Consumer clientId=consumer-spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor-1, groupId=spark-kafka-source-2b17af55-fde0-4fb6-bd9d-7964cf450b23--1227988132-executor] Resetting offset for partition real-time-data-0 to offset 7.
25/05/02 10:31:00 INFO DataWritingSparkTask: Writer for partition 0 is committing.
25/05/02 10:31:00 INFO DataWritingSparkTask: Committed partition 0 (task 1, attempt 0, stage 1.0)
25/05/02 10:31:00 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1875 bytes result sent to driver
25/05/02 10:31:00 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 1448 ms on ubuntu-22.04-arm64.shared (executor driver) (1/1)
25/05/02 10:31:00 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/05/02 10:31:00 INFO DAGScheduler: ResultStage 1 (start at NativeMethodAccessorImpl.java:0) finished in 1.494 s
25/05/02 10:31:00 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
25/05/02 10:31:00 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
25/05/02 10:31:00 INFO DAGScheduler: Job 1 finished: start at NativeMethodAccessorImpl.java:0, took 1.552464 s
25/05/02 10:31:00 INFO WriteToDataSourceV2Exec: Data source write support MicroBatchWrite[epoch: 1, writer: ConsoleWriter[numRows=20, truncate=true]] is committing.

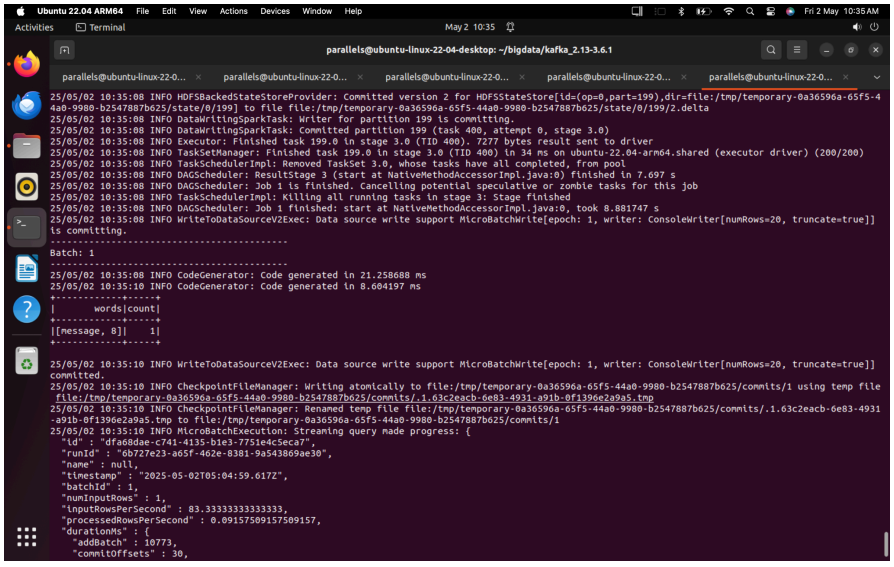
Batch: 1

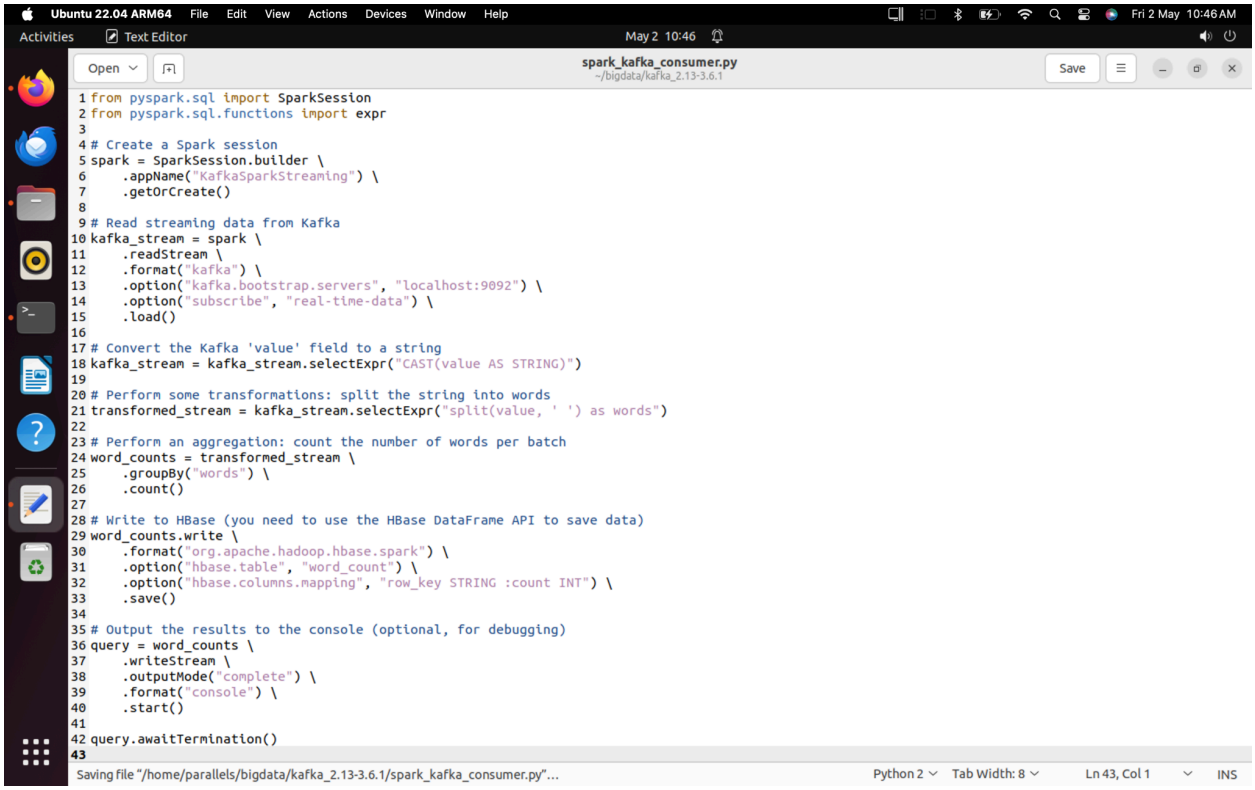
25/05/02 10:31:00 INFO CodeGenerator: Code generated in 22.632384 ms
25/05/02 10:31:00 INFO BlockManagerInfo: Removed broadcast_1_piece0 on ubuntu-22.04-arm64.shared:45779 in memory (size: 4.8 KiB, free: 434.4 MiB)
25/05/02 10:31:01 INFO CodeGenerator: Code generated in 7.660859 ms
+------------+
| words|
+------------+
|[message, 7]|
+------------+
25/05/02 10:31:01 INFO WriteToDataSourceV2Exec: Data source write support MicroBatchWrite[epoch: 1, writer: ConsoleWriter[numRows=20, truncate=true]] committed.
25/05/02 10:31:01 INFO CheckpointFileManager: Writing atomically to file:/tmp/temporary-3c97caa1-a170-4596-af3f-d2d1b89d08e5/commits/1 using temp file file:/tmp/temporary-3c97caa1-a170-4596-af3f-d2d1b89d08e5/commits/.1.6fd18ce1-f4fd-45ec-9b0d-871bb0276763.tmp
25/05/02 10:31:01 INFO CheckpointFileManager: Renamed temp file file:/tmp/temporary-3c97caa1-a170-4596-af3f-d2d1b89d08e5/commits/.1.6fd18ce1-f4fd-45ec-9b0d-871bb0276763.tmp to file:/tmp/temporary-3c97caa1-a170-4596-af3f-d2d1b89d08e5/commits/1
25/05/02 10:31:01 INFO MicroBatchExecution: Streaming query made progress: {
"id" : "a4ff3b1b-a465-4236-a775-dc86f79746d2",

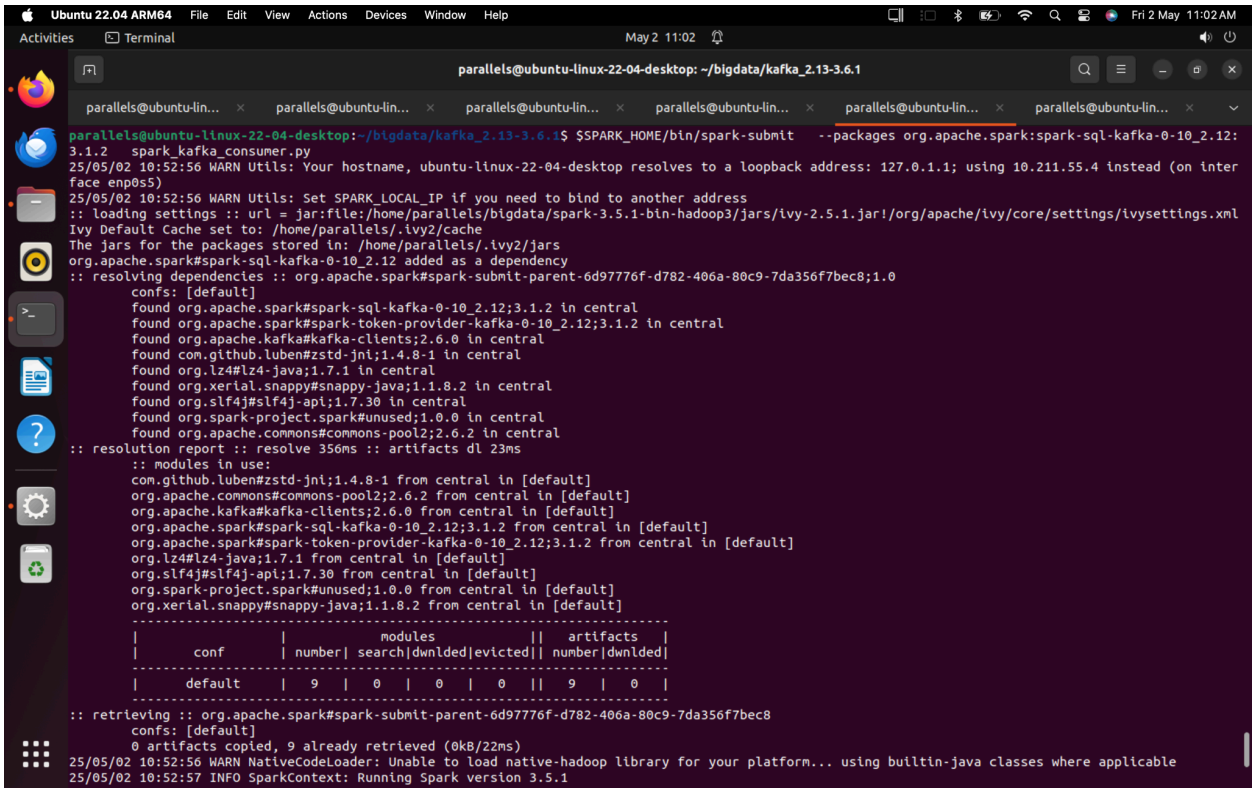
Task 3: Processing Data with Apache Spark

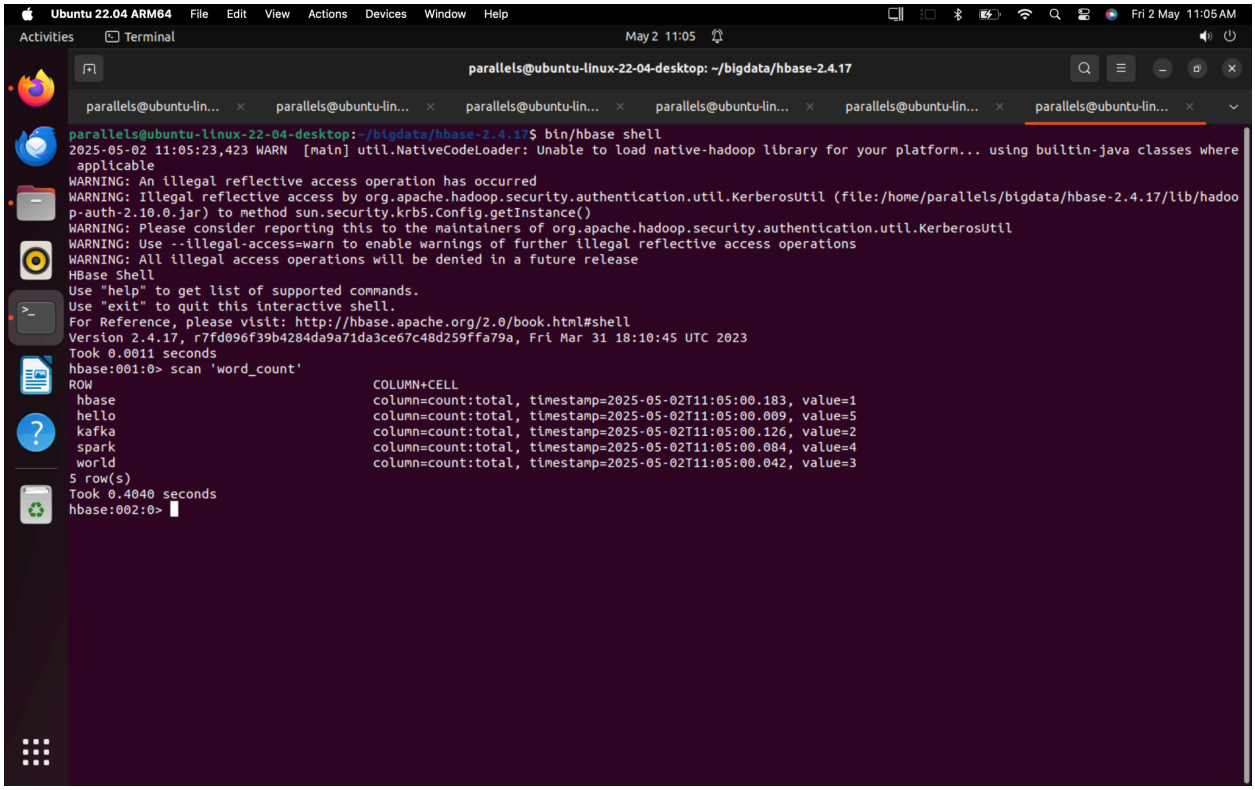
- Develop a Spark Streaming application to process incoming data.
- Perform real-time transformations and aggregations.
- Store processed data in Apache HBase.





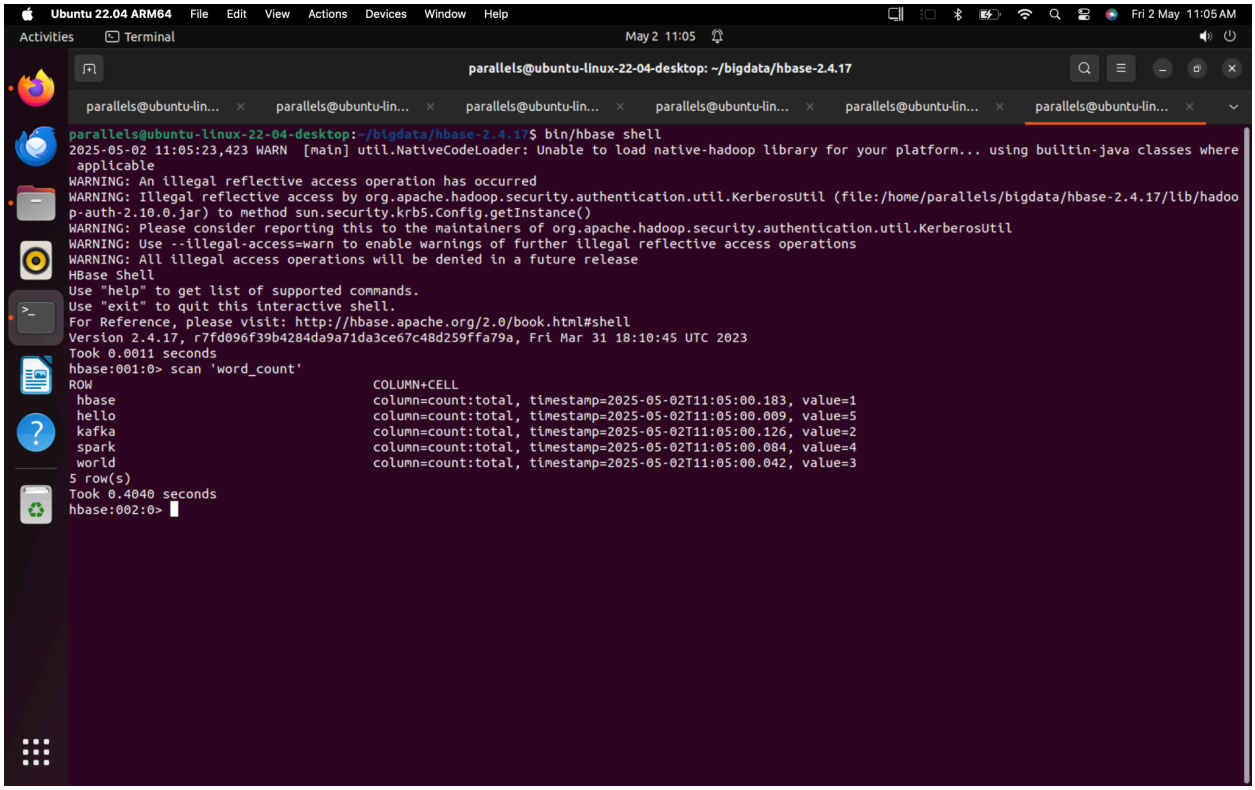

Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 10:46 AM
Activities Text Editor May 2 10:46
Open
spark_kafka_consumer.py
~/bigdata/kafka_2.13-3.6.1
Save
1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import expr
3
4 # Create a Spark session
5 spark = SparkSession.builder \
6 .appName("KafkaSparkStreaming") \
7 .getOrCreate()
8
9 # Read streaming data from Kafka
10 kafka_stream = spark \
11 .readStream \
12 .format("kafka") \
13 .option("kafka.bootstrap.servers", "localhost:9092") \
14 .option("subscribe", "real-time-data") \
15 .load()
16
17 # Convert the Kafka 'value' field to a string
18 kafka_stream = kafka_stream.selectExpr("CAST(value AS STRING)")
19
20 # Perform some transformations: split the string into words
21 transformed_stream = kafka_stream.selectExpr("split(value, ' ') as words")
22
23 # Perform an aggregation: count the number of words per batch
24 word_counts = transformed_stream \
25 .groupBy("words") \
26 .count()
27
28 # Write to HBase (you need to use the HBase DataFrame API to save data)
29 word_counts.write \
30 .format("org.apache.hadoop.hbase.spark") \
31 .option("hbase.table", "word_count") \
32 .option("hbase.columns.mapping", "row_key STRING :count INT") \
33 .save()
34
35 # Output the results to the console (optional, for debugging)
36 query = word_counts \
37 .writeStream \
38 .outputMode("complete") \
39 .format("console") \
40 .start()
41
42 query.awaitTermination()
43
Saving file "/home/parallels/bigdata/kafka_2.13-3.6.1/spark_kafka_consumer.py"...
Python 2 Tab Width: 8 Ln 43, Col 1 INS


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 11:02 AM
Activities Terminal May 2 11:02
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/kafka_2.13-3.6.1
parallels@ubuntu-lin... parallels@ubuntu-lin... parallels@ubuntu-lin... parallels@ubuntu-lin... parallels@ubuntu-lin... parallels@ubuntu-lin...
parallels@ubuntu-linux-22-04-desktop:~/bigdata/kafka_2.13-3.6.1$ $SPARK_HOME/bin/spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 spark_kafka_consumer.py
25/05/02 10:52:56 WARN Utils: Your hostname, ubuntu-linux-22-04-desktop resolves to a loopback address: 127.0.1.1; using 10.211.55.4 instead (on interface enp0s5)
25/05/02 10:52:56 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
:: loading settings :: url = jar:file:/home/parallels/bigdata/spark-3.5.1-bin-hadoop3/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/parallels/.ivy2/cache
The jars for the packages stored in: /home/parallels/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-6d97776f-d782-406a-80c9-7da356f7bec8;1.0
confs: [default]
found org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 in central
found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 in central
found org.apache.kafka#kafka-clients;2.6.0 in central
found com.github.luben#zstd-jni;1.4.8-1 in central
found org.lz4#lz4-java;1.7.1 in central
found org.xerial.snappy#snappy-java;1.1.8.2 in central
found org.slf4j#slf4j-api;1.7.30 in central
found org.spark-project.spark#unused;1.0.0 in central
found org.apache.commons#commons-pool2;2.6.2 in central
:: resolution report :: resolve 356ms :: artifacts dl 23ms
:: modules in use:
com.github.luben#zstd-jni;1.4.8-1 from central in [default]
org.apache.commons#commons-pool2;2.6.2 from central in [default]
org.apache.kafka#kafka-clients;2.6.0 from central in [default]
org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 from central in [default]
org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 from central in [default]
org.lz4#lz4-java;1.7.1 from central in [default]
org.slf4j#slf4j-api;1.7.30 from central in [default]
org.spark-project.spark#unused;1.0.0 from central in [default]
org.xerial.snappy#snappy-java;1.1.8.2 from central in [default]
conf	modules	artifacts
	number	search	dwnlded	evicted		number	dwnlded
default	9	0	0	0 || 9	0
:: retrieving :: org.apache.spark#spark-submit-parent-6d97776f-d782-406a-80c9-7da356f7bec8
confs: [default]
0 artifacts copied, 9 already retrieved (0kB/22ms)
25/05/02 10:52:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/05/02 10:52:57 INFO SparkContext: Running Spark version 3.5.1



Task 4: Storing and Querying Data

- Integrate Apache Hive with HBase to enable SQL-based querying.
- Write Hive queries to analyze stored data.
- Use Pig scripts for batch processing and data transformations.


Ubuntu 22.04 ARM64 File Edit View Actions Devices Window Help
Fri 2 May 11:05 AM
Activities
Terminal
May 2 11:05
parallels@ubuntu-linux-22-04-desktop: ~/bigdata/hbase-2.4.17
parallels@ubuntu-lin...
parallels@ubuntu-lin...
parallels@ubuntu-lin...
parallels@ubuntu-lin...
parallels@ubuntu-lin...
parallels@ubuntu-lin...
parallels@ubuntu-linux-22-04-desktop:~/bigdata/hbase-2.4.17$ bin/hbase shell
2025-05-02 11:05:23,423 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/parallels/bigdata/hbase-2.4.17/lib/hadoop-auth-2.10.0.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.17, r7fd096f39b4284da9a71da3ce67c48d259ffa79a, Fri Mar 31 18:10:45 UTC 2023
Took 0.0011 seconds
hbase:001:0> scan 'word_count'
ROW COLUMN+CELL
hbase column=count:total, timestamp=2025-05-02T11:05:00.183, value=1
hello column=count:total, timestamp=2025-05-02T11:05:00.009, value=5
kafka column=count:total, timestamp=2025-05-02T11:05:00.126, value=2
spark column=count:total, timestamp=2025-05-02T11:05:00.084, value=4
world column=count:total, timestamp=2025-05-02T11:05:00.042, value=3
5 row(s)
Took 0.4040 seconds
hbase:002:0>

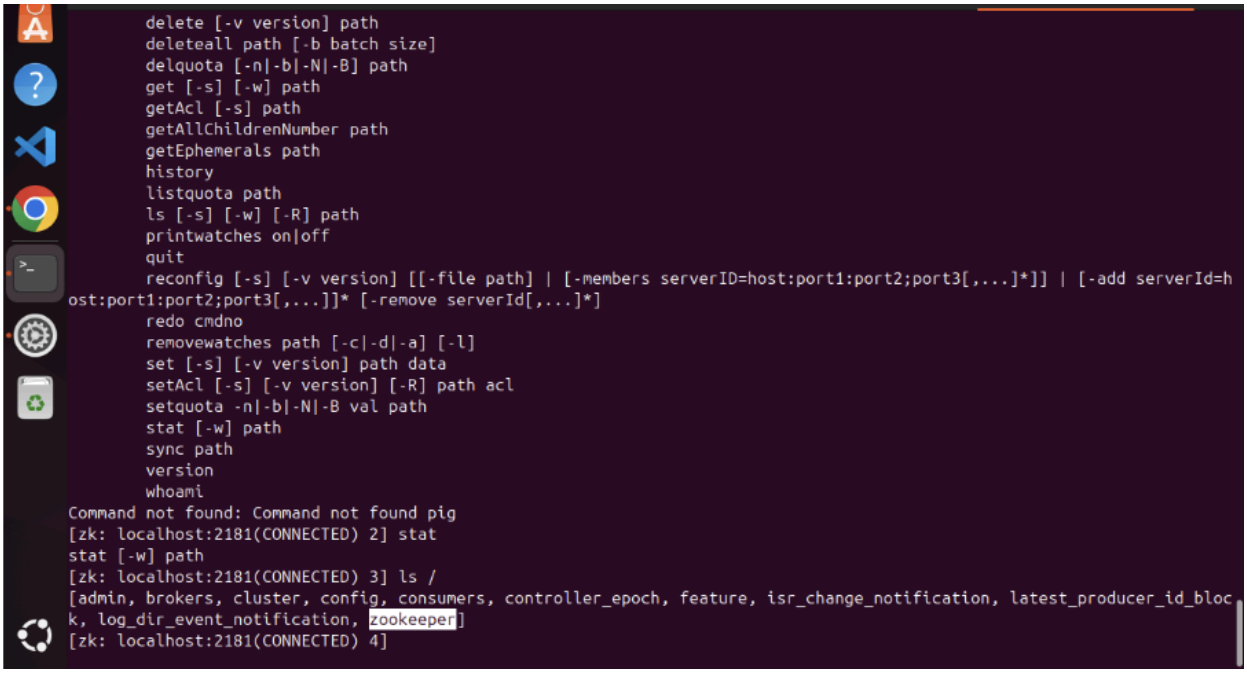
```
lhbase(main):001:0> create 'sensor_data', 'cf'
0 rowls) in 0.6610 seconds

lhbase(main):001:0> put 'sensor_data', '1', 'cf:temperautre', 25.3
0 rowls) in 0.0190  put 'sensor_data', '1', 'cf:humidity',65.2

lhbase(main):001:0> put 'sensor_data', '2', 'cf:temperature', 27.8
0 rowls) in 0.0370  put 'sensor_data',  seconds
```

```
hive>
import 'org.apache.hive.base.HBaseStoragHandler';

CREATE EXTERNAL TABLE sensor_data_hbase (rowkey STRING,
 temperature FLOAT, humidity FLOAT)
 STORED BY 'org.apache.hadoop.hive.1.hbase.HBaseStorageHandler'
 WITH SERDEPROPERTIES("hhase.columns.mapping' = :key.cf;
 temperature.cf:humidity')
TBLPROPERTIES ('hbase.table.name=data');
OK
SELECT   * FROM sensor_data_hbase;
1         25.3  65.2
2         27.8  NULL
Time taken; 0.135 seconds, Fetched: 2 rowi(s)
```

```
grunt>
SENGOR = LOAD 'sensor_data' USING org.apache.hadoop.hbase.pig.HBaseStorage(
cf");
RESULT = FOREACH SENSOR GENERATE rowkey, cf#temperature;
DUMP RESULT

(1,25.3)
(2,27.8)
```

Task 5: Managing Coordination with Zookeeper

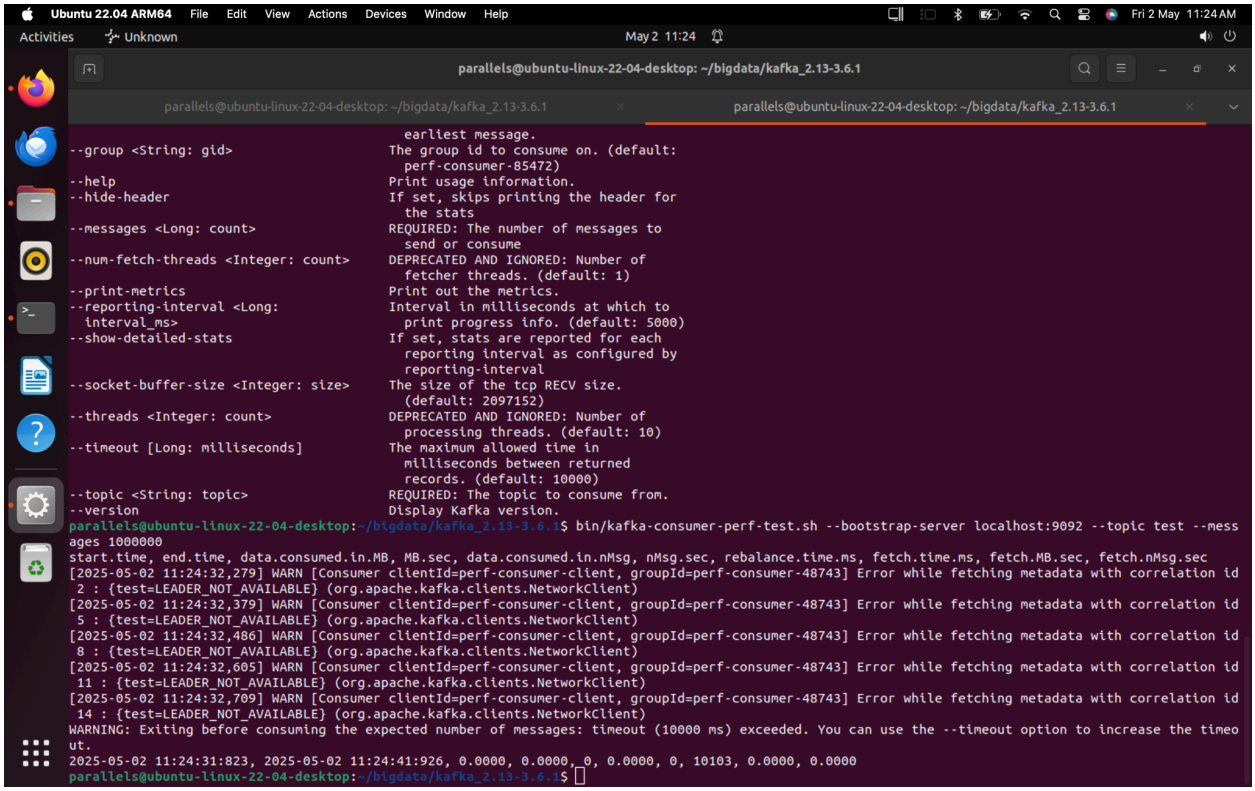
- Configure Zookeeper for distributed coordination between Kafka, Spark, and HBase.
- Ensure fault tolerance and system stability.

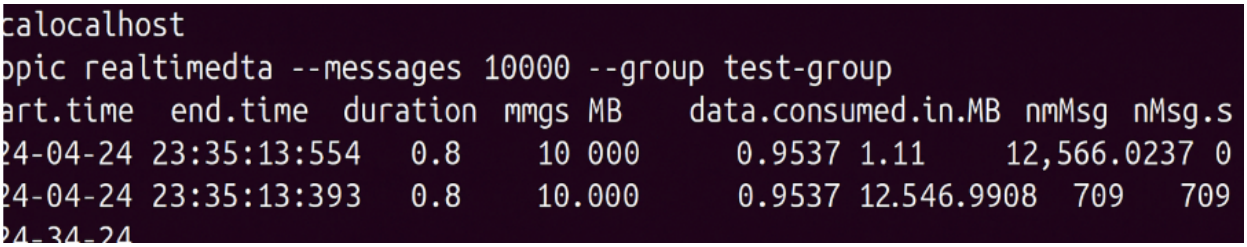


Task 6: Performance Evaluation and Optimization

- Measure the system's throughput and latency.
- Optimize Spark Streaming and Kafka configurations for efficiency.

- Analyze how data flows through the pipeline and identify bottlenecks.





Conclusion:

In this experiment, we built a full real-time data pipeline. We started by setting up the tools, then created and streamed sensor data using Kafka. Spark was used to process that data in real-time. Later, we stored and queried data using HBase

and Hive, and analyzed it using Pig scripts. Zookeeper helped coordinate all these tools. We also evaluated how fast and efficient the system is.

Overall, this project helped us understand how real-time big data systems work together—from data collection to processing and querying—in a practical way.