

Mini project1 : Building a Data Ingestion Pipeline

1.Set Up the Environment

I have already these frameworks installed in my VM



2. Data Migration with Apache Sqoop and HDFS:

1. Create a local MySQL RDBMS database and populate it with sample data.

Connecting to MySQL:

1. Log in to MySQL using the following command:

```
mysql --user=student --password=student labs
```

2. Creating a New Database in MariaDB:

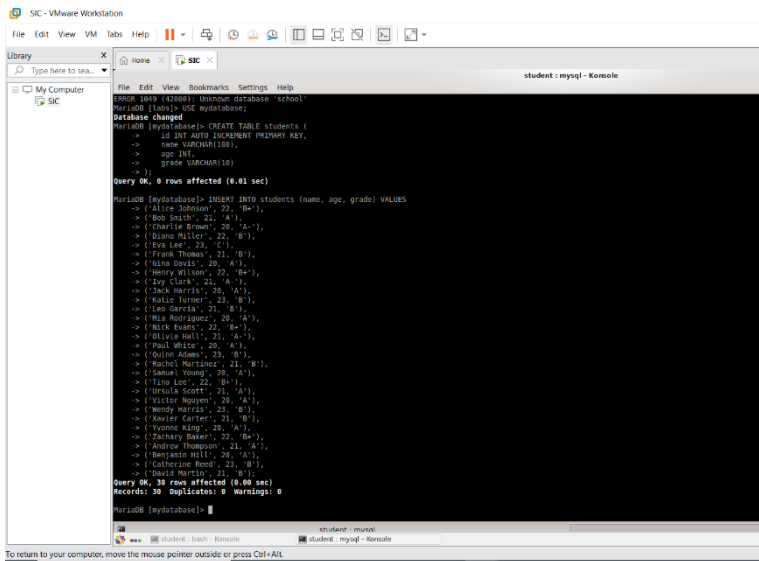
```
CREATE DATABASE mydatabase;
```

3. Creating a Table and Adding Data:

```
USE mydatabase;

CREATE TABLE students (
  id INT AUTO_INCREMENT PRIMARY KEY,
  name VARCHAR(100),
  age INT,
  city VARCHAR(100)
);

INSERT INTO students (name, age, city) VALUES
('John Doe', 25, 'New York'),
('Jane Smith', 22, 'Los Angeles'),
('Ahmed Ali', 28, 'Cairo');
```



```
student: mysql - Konsole
File Edit View Bookmarks Settings Help
[root@localhost ~]# mysql -uuserstudent -passwordstudent labs
MariaDB [labs]> use mydatabase;
Database changed
MariaDB [mydatabase]> CREATE TABLE students (
  id INT AUTO INCREMENT PRIMARY KEY,
  name VARCHAR(100),
  age INT,
  grade VARCHAR(10)
);
Query OK, 0 rows affected (0.01 sec)

MariaDB [mydatabase]> INSERT INTO students (name, age, grade) VALUES
-> ('Alice Johnson', 22, 'B+'),
-> ('Bob Smith', 21, 'A'),
-> ('Charlie Brown', 20, 'A-'),
-> ('Diana Miller', 22, 'B'),
-> ('Evo Lee', 23, 'C'),
-> ('Frank Thomas', 21, 'B'),
-> ('Gina Davis', 20, 'A'),
-> ('Henry Wilson', 22, 'B+'),
-> ('Ivy Clark', 21, 'A'),
-> ('Jack Harris', 20, 'A'),
-> ('Katie Turner', 23, 'B'),
-> ('Leo Garcia', 21, 'B'),
-> ('Mia Rodriguez', 20, 'A'),
-> ('Nick Evans', 22, 'B+'),
-> ('Olivia Hall', 21, 'A-'),
-> ('Paul White', 20, 'A'),
-> ('Quinn Adams', 23, 'B'),
-> ('Rachel Martinez', 21, 'B'),
-> ('Samuel Young', 20, 'A'),
-> ('Tina Lee', 22, 'B+'),
-> ('Ursula Scott', 21, 'A'),
-> ('Victor Nguyen', 20, 'A'),
-> ('Wendy Harris', 23, 'B'),
-> ('Xavier Carter', 21, 'B'),
-> ('Yvonne King', 20, 'A'),
-> ('Zachary Baker', 22, 'B+'),
-> ('Andrew Thompson', 21, 'A'),
-> ('Benjamin Hill', 20, 'A'),
-> ('Catherine Reed', 23, 'B'),
-> ('David Martin', 21, 'B');
Query OK, 30 rows affected (0.00 sec)
Records: 30 Duplicates: 0 Warnings: 0

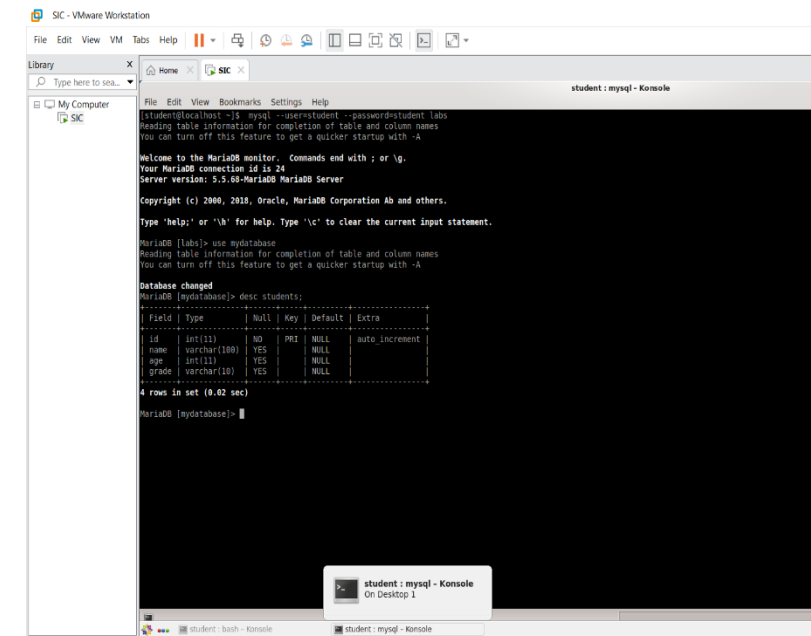
MariaDB [mydatabase]>
```

4. Viewing Data in the Table (SELECT Statement):

```
SELECT * FROM students;
```

5. Describing the Table

```
DESC students;
```



```
student: mysql - Konsole
File Edit View Bookmarks Settings Help
[student@localhost ~]# mysql -uuserstudent -passwordstudent labs
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MariaDB connection id is 24
Server version: 5.5.68-MariaDB MariaDB Server

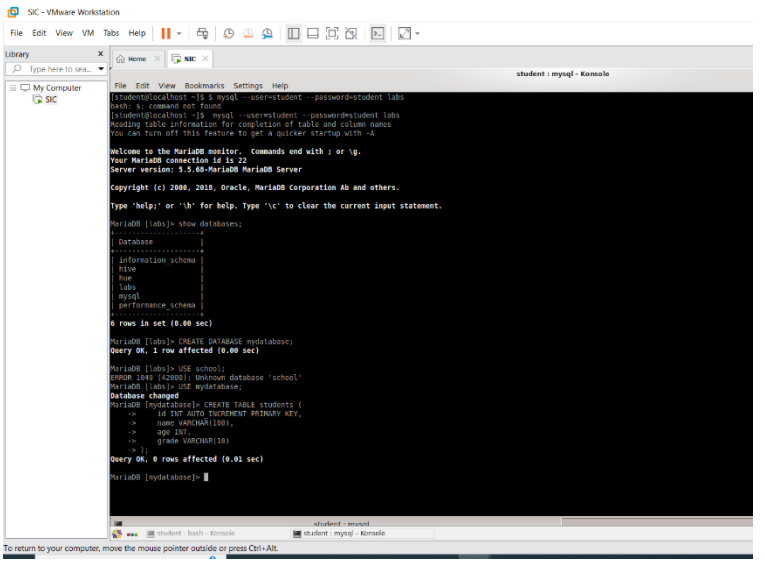
Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [labs]> use mydatabase;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MariaDB [mydatabase]> desc students;
+----+-----+-----+-----+-----+
| id  | Type | Null | Key | Default | Extra |
+----+-----+-----+-----+-----+
| id  | int(11) | NO   | PRI | NULL    | auto_increment |
| name | varchar(100) | YES  |     | NULL    |               |
| age  | int(11) | YES  |     | NULL    |               |
| grade | varchar(10) | YES  |     | NULL    |               |
+----+-----+-----+-----+-----+
4 rows in set (0.02 sec)

MariaDB [mydatabase]>
```



```
student: mysql - Konsole
File Edit View Bookmarks Settings Help
[student@localhost ~]# mysql -uuserstudent -passwordstudent labs
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MariaDB connection id is 22
Server version: 5.5.68-MariaDB MariaDB Server

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [labs]> show databases;
+-----+
| Database |
+-----+
| information_schema |
| mysql |
| performance_schema |
+-----+
0 rows in set (0.00 sec)

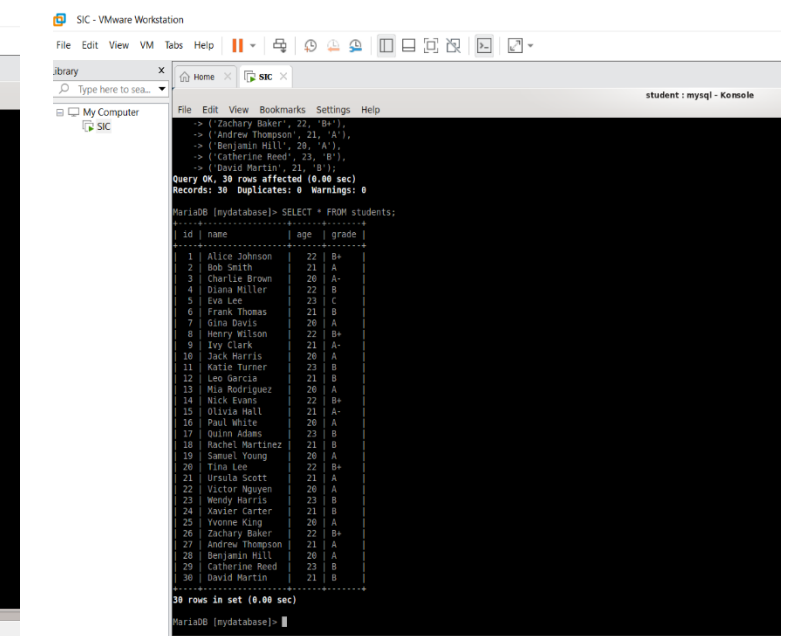
MariaDB [labs]> CREATE DATABASE mydatabase;
Query OK, 1 row affected (0.00 sec)

MariaDB [labs]> USE school;
ERROR 1006 (42000): Unknown database 'school'

MariaDB [labs]> USE mydatabase;
Database changed
MariaDB [mydatabase]> CREATE TABLE students (
  id INT AUTO INCREMENT PRIMARY KEY,
  name VARCHAR(100),
  age INT,
  grade VARCHAR(10)
);
Query OK, 0 rows affected (0.01 sec)

MariaDB [mydatabase]>
```

To return to your computer, move the mouse pointer outside or press Ctrl+Alt.



```
student: mysql - Konsole
File Edit View Bookmarks Settings Help
[student@localhost ~]# mysql -uuserstudent -passwordstudent labs
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MariaDB connection id is 24
Server version: 5.5.68-MariaDB MariaDB Server

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [labs]> use mydatabase;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MariaDB [mydatabase]> desc students;
+----+-----+-----+-----+-----+
| id  | Type | Null | Key | Default | Extra |
+----+-----+-----+-----+-----+
| id  | int(11) | NO   | PRI | NULL    | auto_increment |
| name | varchar(100) | YES  |     | NULL    |               |
| age  | int(11) | YES  |     | NULL    |               |
| grade | varchar(10) | YES  |     | NULL    |               |
+----+-----+-----+-----+-----+
4 rows in set (0.02 sec)

MariaDB [mydatabase]> SELECT * FROM students;
+----+-----+-----+-----+
| id | name          | age | grade |
+----+-----+-----+-----+
| 1  | Alice Johnson | 22  | B+    |
| 2  | Bob Smith     | 21  | A      |
| 3  | Charlie Brown | 20  | A-     |
| 4  | Diana Miller  | 22  | B      |
| 5  | Evo Lee       | 23  | C      |
| 6  | Frank Thomas | 21  | B      |
| 7  | Gina Davis    | 20  | A      |
| 8  | Henry Wilson  | 22  | B+     |
| 9  | Ivy Clark     | 21  | A      |
| 10 | Jack Harris   | 20  | A      |
| 11 | Katie Turner  | 23  | B      |
| 12 | Leo Garcia    | 21  | B      |
| 13 | Mia Rodriguez | 20  | A      |
| 14 | Nick Evans    | 22  | B+     |
| 15 | Olivia Hall   | 21  | A-     |
| 16 | Paul White    | 20  | A      |
| 17 | Quinn Adams   | 23  | B      |
| 18 | Rachel Martinez | 21  | B      |
| 19 | Samuel Young  | 20  | A      |
| 20 | Tina Lee      | 22  | B+     |
| 21 | Ursula Scott  | 21  | A      |
| 22 | Victor Nguyen | 20  | A      |
| 23 | Wendy Harris  | 23  | B      |
| 24 | Xavier Carter | 21  | B      |
| 25 | Yvonne King   | 20  | A      |
| 26 | Zachary Baker | 22  | B+     |
| 27 | Andrew Thompson | 21  | A      |
| 28 | Benjamin Hill | 20  | A      |
| 29 | Catherine Reed | 23  | B      |
| 30 | David Martin  | 21  | B      |
+----+-----+-----+-----+
30 rows in set (0.00 sec)

MariaDB [mydatabase]>
```

2. Writing Sqoop commands to import data from the local database to HDFS

```
$ sqoop import-connect jdbc:mysql://localhost/mydatabase --username student --password student --table students
```

```

2024-08-07 03:43:06,947 INFO mapreduce.Job: map 0% reduce 0%
2024-08-07 03:43:14,104 INFO mapreduce.Job: map 50% reduce 0%
2024-08-07 03:43:16,110 INFO mapreduce.Job: map 100% reduce 0%
2024-08-07 03:43:16,152 INFO mapreduce.Job: Job job_1722968506208_0002 completed successfully
2024-08-07 03:43:16,152 INFO mapreduce.Job: Counters: 33
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=112940
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=390
HDFS: Number of bytes written=617
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=4
Other Local map tasks=4
Total time spent by all maps in occupied slots (ms)=19245
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=19245
Total vcore-seconds taken by all map tasks=19245
Total megabyte-seconds taken by all map tasks=1970680
Map-Reduce Framework
Map input records=30
Map output records=30
Input split bytes=390
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=349
CPU time spent (ms)=3720
Physical memory (bytes) snapshot=471849684
Virtual memory (bytes) snapshot=1232194560
Total committed heap usage (bytes)=250371712
Peak Map Physical memory (bytes)=243470336
Peak Map Virtual memory (bytes)=2014705604
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=617
2024-08-07 03:43:16,197 INFO mapreduce.ImportJobBase: Transferred 617 bytes
2024-08-07 03:43:16,200 INFO mapreduce.ImportJobBase: Retrieved 30 records

```

```

student@localhost ~$ sqoop import --connect jdbc:mysql://localhost/mydatabase --username student --password student --table students --fields-terminated-by ';' --target-dir /dwh/_students
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set HCAT HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-08-07 03:42:47,450 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-08-07 03:42:47,496 INFO tool.BaseSqoopTool: Setting your password on the command line is insecure. Consider using -P instead.
2024-08-07 03:42:47,583 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-08-07 03:42:47,583 INFO tool.CodeGenTool: Beginning code generation
2024-08-07 03:42:48,235 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 03:42:48,314 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 03:42:48,329 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/98d57c3c6ba31aa40a7483174c4/students.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-08-07 03:42:49,561 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/98d57c3c6ba31aa40a7483174c4/students.jar
2024-08-07 03:42:49,571 INFO manager.MySQLManager: It looks like you are importing from mysql.
2024-08-07 03:42:49,571 INFO manager.MySQLManager: This transfer can be faster! Use the --direct
2024-08-07 03:42:49,571 INFO manager.MySQLManager: option to exercise a MySQL specific fast path.
2024-08-07 03:42:49,571 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-08-07 03:42:49,574 INFO mapreduce.ImportJobBase: Beginning import of students
2024-08-07 03:42:49,575 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-07 03:42:49,640 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-08-07 03:42:50,110 INFO client.DefaultHadoopWALinuxProxyProvider: Connecting to ResourceManager at jv8.6.8:8032
2024-08-07 03:42:50,381 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/staging/job_1722968506208_0002
2024-08-07 03:42:54,804 INFO db.DatabaseMetaData: using read committed transaction isolation
2024-08-07 03:42:54,804 INFO db.DatabaseMetaData: RowidRangeQuery: SELECT MAX(id), MAX(id) FROM 'students'
2024-08-07 03:42:54,804 INFO mapreduce.JobSubmitter: Split size: 7; Num splits: 4 from: 1 to: 30
2024-08-07 03:42:54,806 INFO mapreduce.JobSubmitter: number of splits=4
2024-08-07 03:42:55,302 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1722968506208_0002
2024-08-07 03:42:55,582 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-07 03:42:56,680 INFO conf.Configuration: resource.types not found.
2024-08-07 03:42:56,866 INFO resource.ResourceMgrUtil: Unable to find 'resource.types.xml'.
2024-08-07 03:42:56,764 INFO impl.YarnClientImpl: Submitted application application_1722968506208_0002
2024-08-07 03:42:56,830 INFO mapreduce.Job: The url to track the job: http://localhost:8080/proxy/application_1722968506208_0002/
2024-08-07 03:42:56,831 INFO mapreduce.Job: Running job: job_1722968506208_0002
2024-08-07 03:43:06,941 INFO mapreduce.Job: Job job_1722968506208_0002 running in uber mode : false
2024-08-07 03:43:16,110 INFO mapreduce.Job: map 100% reduce 0%
2024-08-07 03:43:16,213 INFO mapreduce.Job: Job job_1722968506208_0002 completed successfully
2024-08-07 03:43:16,152 INFO mapreduce.Job: Counters: 33
File System Counters

```

- Making target directory in Hadoop HDFS

```
$ hdfs dfs -Mkdir /dwh_
```

- Import database to hdfs target directory

```

$ sqoop import \
--connect jdbc:mysql://localhost/mydatabase \
--username student \
--password student \
--table students \
--fields-terminated-by ';' \
--target-dir /dwh/_students

```

```

student@localhost ~$ sqoop import --connect jdbc:mysql://localhost/mydatabase --username student --password student --table students --fields-terminated-by ';' --target-dir /dwh/_students
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! HCatalog jobs will fail.
Please set HCAT HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-08-07 04:50:19,944 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-08-07 04:50:19,976 INFO tool.BaseSqoopTool: Setting your password on the command line is insecure. Consider using -P instead.
2024-08-07 04:50:20,870 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-08-07 04:50:20,880 INFO tool.CodeGenTool: Beginning code generation
2024-08-07 04:50:20,294 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 04:50:20,427 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 04:50:20,434 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/a6b3ad8bde8c41daf7f249f2a398c8a/students.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-08-07 04:50:21,599 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student/compile/a6b3ad8bde8c41daf7f249f2a398c8a/students.jar
2024-08-07 04:50:21,609 INFO manager.MySQLManager: It looks like you are importing from mysql.
2024-08-07 04:50:21,609 INFO manager.MySQLManager: This transfer can be faster! Use the --direct
2024-08-07 04:50:21,609 INFO manager.MySQLManager: option to exercise a MySQL specific fast path.
2024-08-07 04:50:21,609 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
2024-08-07 04:50:21,613 INFO mapreduce.ImportJobBase: Beginning import of students
2024-08-07 04:50:21,614 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-07 04:50:21,721 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-08-07 04:50:22,255 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-08-07 04:50:22,296 INFO client.DefaultHadoopWALinuxProxyProvider: Connecting to ResourceManager at jv8.6.8:8032
2024-08-07 04:50:22,918 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/staging/job_1722968506208_0003

```

```
student: hash - Konsole <2>
File Edit View Bookmarks Settings Help
2024-08-07 04:59:51.111 INFO mapreduce.job: map 50% reduce 0%
2024-08-07 04:59:52.129 INFO mapreduce.job: map 100% reduce 0%
2024-08-07 04:59:52.211 INFO mapreduce.job: Job job_1722968586208_0003 completed successfully
2024-08-07 04:59:52.211 INFO mapreduce.job: Counter: 13
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=1129908
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=908
  HDFS: Number of bytes written=417
  HDFS: Number of read operations=24
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=0
  HDFS: Number of bytes read.errored=0
  HDFS: Number of bytes read.errored.coded=0
Job Counters
  Launched map tasks=4
  Other local map tasks=4
  Total time spent by all maps in occupied slots (ms)=46322
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=46322
  Total vcore-millisecspend taken by all map tasks=46322
  Total megabyte-millisecspend taken by all map tasks=47433728
Map-Reduce Framework
  Map input records=30
  Map output records=30
  Input split bytes=308
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=425
  CPU time spent (ms)=790
  Physical memory (bytes) snapshot=726740996
  Virtual memory (bytes) snapshot=1129649804
  Total committed heap usage (bytes)=67864352
  Peak map physical memory (bytes)=2480172
  Peak Map Virtual memory (bytes)=281353804
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=1
2024-08-07 04:59:52.215 INFO mapreduce.Job: Transfered 317 bytes (2.29.9516 seconds (20.5724 bytes/sec)
[student@localhost ~]$
```

- checking whether it was imported or not

hdfs dfs -ls /dwh/_students

```
[student@localhost ~]$ hdfs dfs -ls /dwh/_students
Found 5 items
-rw-r--r-- 1 student supergroup 0 2024-08-07 04:59 /dwh/_students/_SUCCESS
-rw-r--r-- 1 student supergroup 155 2024-08-07 04:59 /dwh/_students/part-m-00000
-rw-r--r-- 1 student supergroup 141 2024-08-07 04:59 /dwh/_students/part-m-00001
-rw-r--r-- 1 student supergroup 145 2024-08-07 04:59 /dwh/_students/part-m-00002
-rw-r--r-- 1 student supergroup 176 2024-08-07 04:59 /dwh/_students/part-m-00003
[student@localhost ~]$
```

- open any part and show the data

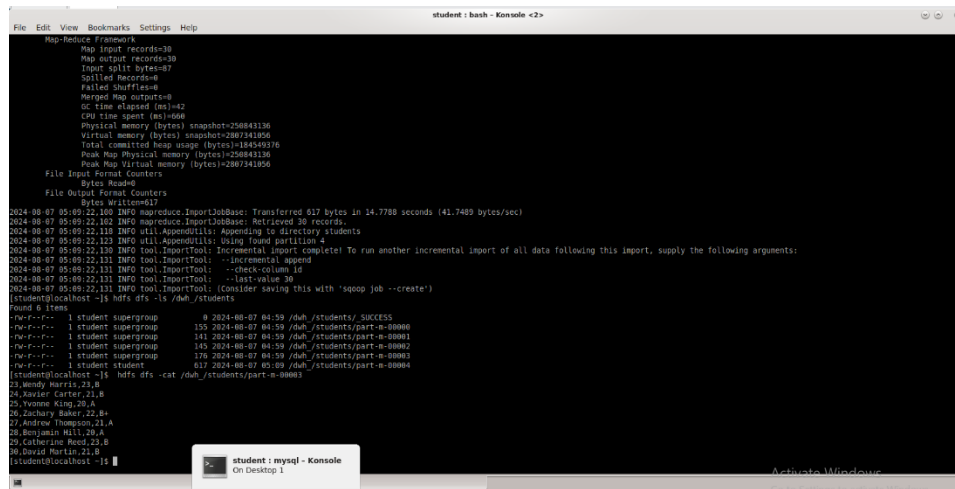
hdfs dfs -cat /dwh/_students/part-m-00001

```
[student@localhost ~]$ hdfs dfs -cat /dwh/_students/part-m-00001
9,Ivy Clark,21,A-
10,Jack Harris,20,A
11,Katie Turner,23,B
12,Leo Garcia,21,B
13,Mia Rodriguez,20,A
14,Nick Evans,22,B+
15,Olivia Hall,21,A-
[student@localhost ~]$
```

- Set up incremental import with Sqoop to handle new data entries.

```
[student@localhost ~]$ sqoop import \
  --connect jdbc:mysql://localhost/mydatabase \
  --username student \
  --password student \
  --table students \
  --target-dir /dwh/_students \
  --incremental append \
  --check-column id \
  --last-value 0 \
  --h 1
Warning: /usr/local/sqoop/sqoop-1.4.7/./hcatalog does not exist! Hcatalog jobs will fail.
Please set HCATALOG_HOME to the root of your Hcatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./accumulo does not exist! Accumulo imports will fail.
Please set ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.7/./zookeeper does not exist! Accumulo imports will fail.
Please set ZOOKEEPER_HOME to the root of your Zookeeper installation.
2024-08-07 05:09:04.746 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
2024-08-07 05:09:04.772 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
2024-08-07 05:09:04.865 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
2024-08-07 05:09:05.177 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 05:09:05.215 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM 'students' AS t LIMIT 1
2024-08-07 05:09:05.215 INFO err.ConfigurationManager: HADOOP_MAPRED_HOME is /home/hadoop/hadoop
Note: /tmp/sqoop-student/compile/6a3f6822dc52165939f6379d97e147f/students.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
2024-08-07 05:09:06.803 INFO err.ConfigurationManager: Writing jar file: /tmp/sqoop-student/compile/6a3f6822dc52165939f6379d97e147f/students.jar
2024-08-07 05:09:06.812 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(id) FROM 'students'
2024-08-07 05:09:06.817 INFO tool.ImportTool: Incremental import based on column 'id'
2024-08-07 05:09:06.817 INFO tool.ImportTool: Lower bound value: 0
2024-08-07 05:09:06.817 INFO tool.ImportTool: Upper bound value: 30
2024-08-07 05:09:06.817 WARN manager.MySQLManager: It looks like you are importing from mysql.
2024-08-07 05:09:06.817 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
2024-08-07 05:09:06.817 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
2024-08-07 05:09:06.817 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertNull (mysql)
2024-08-07 05:09:06.820 INFO mapreduce.ImportJobBase: Beginning import of students
2024-08-07 05:09:06.821 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-08-07 05:09:06.899 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
2024-08-07 05:09:07.312 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
2024-08-07 05:09:07.415 INFO client.DefaultHadoopMailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-08-07 05:09:07.852 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/job_1722968586208_0004
2024-08-07 05:09:11.264 INFO dh.DistributedFormat: Using read committed transaction isolation
2024-08-07 05:09:11.296 INFO mapreduce.JobSubmitter: number of splits:1
2024-08-07 05:09:11.447 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1722968586208_0004
2024-08-07 05:09:11.447 INFO mapreduce.JobSubmitter: Executing with tokens: []
[student@localhost ~]$
```

3. Real-Time Data Ingestion with Apache Flume and Apache Kafka:



```
File Edit View Bookmarks Settings Help
student: bash - Konsole «2»

Map-Reduce Framework
Map input records=30
Map output records=30
Input split bytes=47
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=42
CPU time spent (ms)=606
Physical memory (bytes) snapshot=258843136
Virtual memory (bytes) snapshot=280741956
Total committed heap usage (bytes)=184549376
Peak Map Physical memory (bytes)=258843136
Peak Map Virtual memory (bytes)=280741956
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
2024-08-07 05:09:22.109 INFO mapreduce.ImportJobBase: Transferred 617 bytes in 14.7788 seconds (41.7409 bytes/sec)
2024-08-07 05:09:22.162 INFO mapreduce.ImportJobBase: Retrieved 30 records.
2024-08-07 05:09:22.112 INFO util.AppendUtils: Appending to directory:students
2024-08-07 05:09:22.123 INFO util.AppendUtils: Using found partition 4
2024-08-07 05:09:22.136 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:
2024-08-07 05:09:22.131 INFO tool.ImportTool: --incremental append
2024-08-07 05:09:22.131 INFO tool.ImportTool: --check-column id
2024-08-07 05:09:22.131 INFO tool.ImportTool: --last-value 30
2024-08-07 05:09:22.131 INFO tool.ImportTool: (Consider saving this with 'sqoop job --create')
student@localhost ~$ hdfs dfs -ls /dwh/_students
Found 6 items
rw-r--r-- 1 student supergroup 0 2024-08-07 04:59 /dwh/_students/_SUCCESS
rw-r--r-- 1 student supergroup 350 2024-08-07 04:59 /dwh/_students/part-m-00000
rw-r--r-- 1 student supergroup 141 2024-08-07 04:59 /dwh/_students/part-m-00001
rw-r--r-- 1 student supergroup 140 2024-08-07 04:59 /dwh/_students/part-m-00002
rw-r--r-- 1 student supergroup 176 2024-08-07 04:59 /dwh/_students/part-m-00003
rw-r--r-- 1 student student 617 2024-08-07 05:09 /dwh/_students/part-m-00004
student@localhost ~$ hdfs dfs -cat /dwh/_students/part-m-00001
23,Wendy Harris,21.8
24,Xavier Carter,21.8
25,Yvonne King,20.4
26,Zachary Baker,22.8
27,Andrew Thompson,21.4
28,Benjamin Hill,20.4
29,Catherine Wood,23.8
30,David Martin,21.8
student@localhost ~$
```

setting up flume to send data to kafka topic

- first i have to configure the agent that is responsible for getting the data source

```
#conf file

agent1.sources = streaming-txt-source

agent1.sinks = kafka-sink logger-sink

agent1.channels = memory-channel

agent1.sources.streaming-txt-source.type = spooldir

agent1.sources.streaming-txt-source.spoolDir = /home/student/Documents/spool

agent1.sinks.kafka-sink.type = org.apache.flume.sink.kafka.KafkaSink

agent1.sinks.kafka-sink.topic = stream_text

agent1.sinks.kafka-sink.brokerList = localhost:9092

agent1.sinks.kafka-sink.batchSize = 5

agent1.channels.memory-channel.type = memory

agent1.channels.memory-channel.capacity = 10000

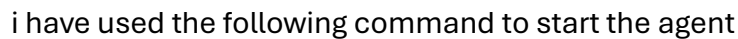
agent1.channels.memory-channel.transactionCapacity = 100

agent1.sinks.logger-sink.type = logger

agent1.sources.streaming-txt-source.channels = memory-channel

agent1.sinks.kafka-sink.channel = memory-channel

agent1.sinks.logger-sink.channel = memory-channel
```

[illegible]

start the zookepr service and Kafka service as well

```
sudo systemctl stop kafka
```



```
sudo systemctl stop zookeeper
```

```
sudo systemctl start zookeeper
```

```
sudo systemctl status zookeeper
```

```
sudo systemctl start kafka
```

```
sudo systemctl status kafka
```

```
student: bash - Konsole
File Edit View Bookmarks Settings Help
[student@localhost ~]$ sudo stop hbase.sh
no hbase master found
[student@localhost ~]$
[student@localhost ~]$ sudo systemctl stop kafka
[student@localhost ~]$ sudo systemctl stop zookeeper
[student@localhost ~]$ sudo systemctl start zookeeper
[student@localhost ~]$ sudo systemctl status zookeeper
● zookeeper.service
   Loaded: loaded (/etc/systemd/system/zookeeper.service; disabled; vendor preset: disabled)
   Active: active (running) since Wed 2024-08-07 22:33:08 KST; 17s ago
     Main PID: 14000 (java)
    CGroup: /system.slice/zookeeper.service
            └─14000 java -Xms32M -Xmx32M -server -XX:+UseG1GC -XX:MaxGCPauseMillis=20 -XX:InitiatingHeapOccupancyPercent=35 -XX:+ExplicitGCInvokesConcurrent -XX:MaxInlineLevel=15 -Djava.awt.headless=true -X...

Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,416] INFO maxSessionTimeout set to 60000 (org.apache.zookeeper.server.ZooKeeperServer)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,416] INFO Created server with tickTime 3000 minSessionTimeout 6000 maxSessionTimeout 60000 dataDir /home...eperServer)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,426] INFO Using org.apache.zookeeper.server.NIOServerCnxnFactory as server connection factory (org.apac...nFactory)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,429] INFO Configuring NIO connection handler with 10s sessionless connection timeout, 1 selector thread...nFactory)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,433] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.zookeeper.server.NIOServerCnxnFactory)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,461] INFO zookeeper.snapshotSizeFactor = 0.33 (org.apache.zookeeper.server.ZNDatabase)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,472] INFO Reading snapshot /home/hadoop/hadoopdata/zookeeper/zookeeper-0/version-2/snapshot.2c1 (org.ap...e.FileSnap)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,499] INFO Snapshotting: 0x2e9 to /home/hadoop/hadoopdata/zookeeper/zookeeper-0/version-2/snapshot.2e9 (...:FileSnapLog)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,522] INFO PrepRequestProcessor (sid:0) started, reconfigEnabled=false (org.apache.zookeeper.server.Prepro...tProcessor)
Aug 07 22:33:09 localhost.localdomain zookeeper-server-start.sh[14000]: [2024-08-07 22:33:09,523] INFO Using checkIntervalMs=60000 maxPerMinute=10000 (org.apache.zookeeper.server.ContainerManager)
Hint: Some lines were ellipsized, use -l to show in full.
[student@localhost ~]$ sudo systemctl start kafka
[student@localhost ~]$ sudo systemctl status kafka
● kafka.service
   Loaded: loaded (/etc/systemd/system/kafka.service; disabled; vendor preset: disabled)
   Active: active (running) since Wed 2024-08-07 22:33:35 KST; 10s ago
     Main PID: 14393 (sh)
    CGroup: /system.slice/kafka.service
            └─14393 /bin/sh -c /home/kafka/kafka/bin/kafka-server-start.sh /home/kafka/kafka/config/server.properties > /home/kafka/kafka/kafka.log 2>&1
              └─14395 java -Xms1G -Xmx1G -server -XX:+UseG1GC -XX:MaxGCPauseMillis=20 -XX:InitiatingHeapOccupancyPercent=35 -XX:+ExplicitGCInvokesConcurrent -XX:MaxInlineLevel=15 -Djava.awt.headless=true -Xlogg...

Aug 07 22:33:35 localhost.localdomain systemd[1]: Started kafka.service.
[student@localhost ~]$
```

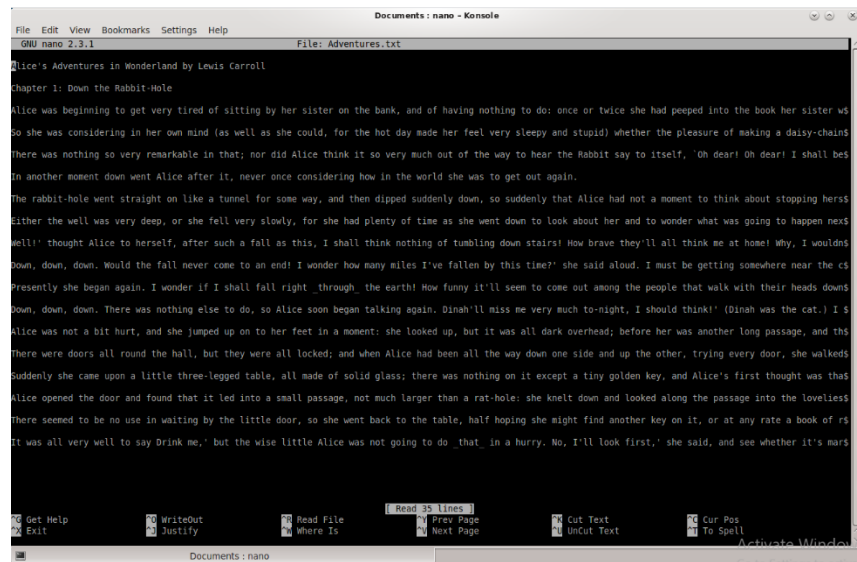
Create destination directory to collect streaming text

```
[student@localhost Documents]$ mkdir spool
[student@localhost Documents]$ ls
spool  spooldir.conf  spool_stream.py
[student@localhost Documents]$
```

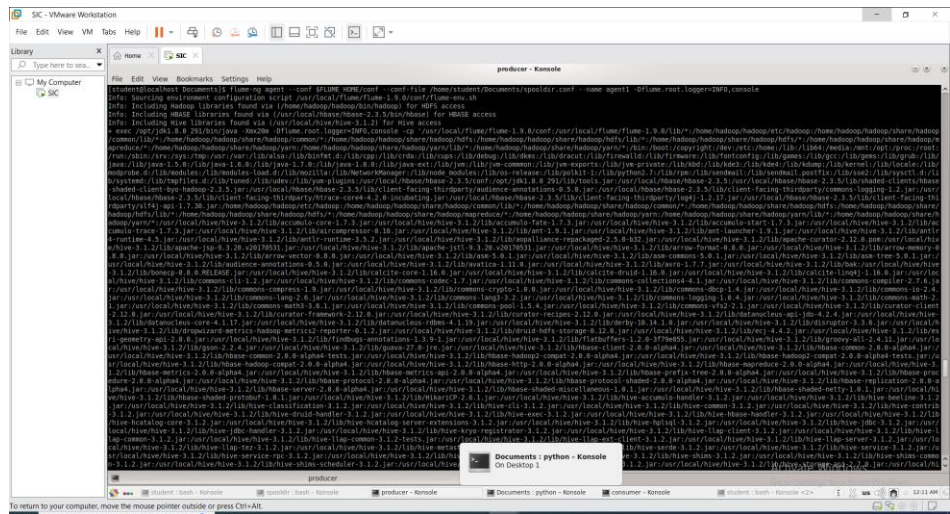
Let's create Kafka topic and name it **stream_text**

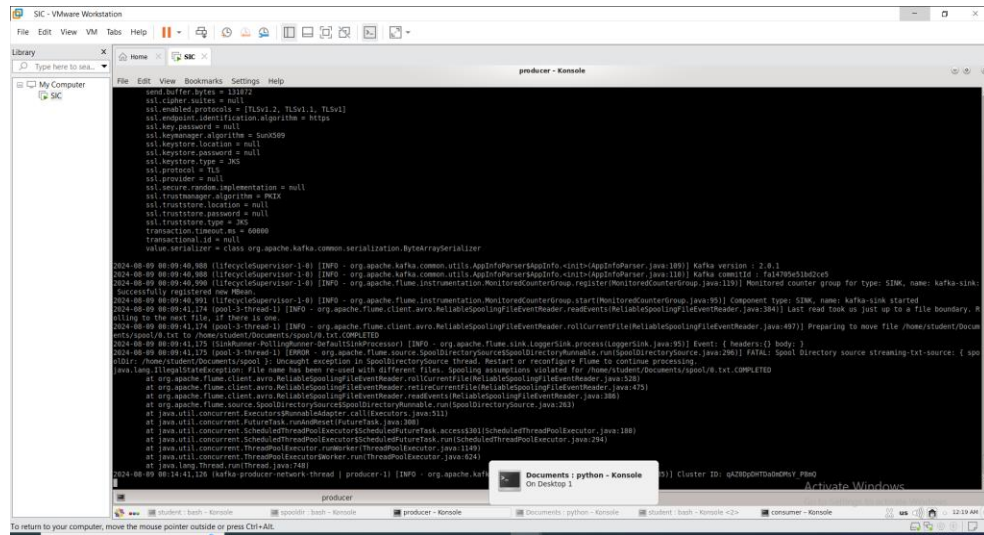
```
Documents: bash - Konsole
File Edit View Bookmarks Settings Help
[student@localhost Documents]$ kafka-topics --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic stream_text
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid issues it is best to use either, but not both.
Created topic stream_text.
[student@localhost Documents]$
```

This is the data that I will send it to kafka

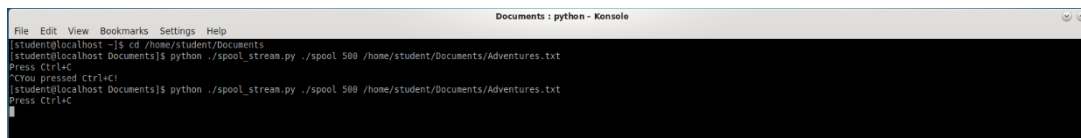


In another terminal We will navigate to /home/Documents/ to run producer





In another terminal We will navigate to /home/Documents to run python script to send our streaming data



In a new terminal I called it consumer, we will create Kafka consumer to receive messages from the server.

With this command

kafka-console-consumer --bootstrap-server localhost:9092 --topic stream_text --from-beginning

this is the data after run the pipeline

