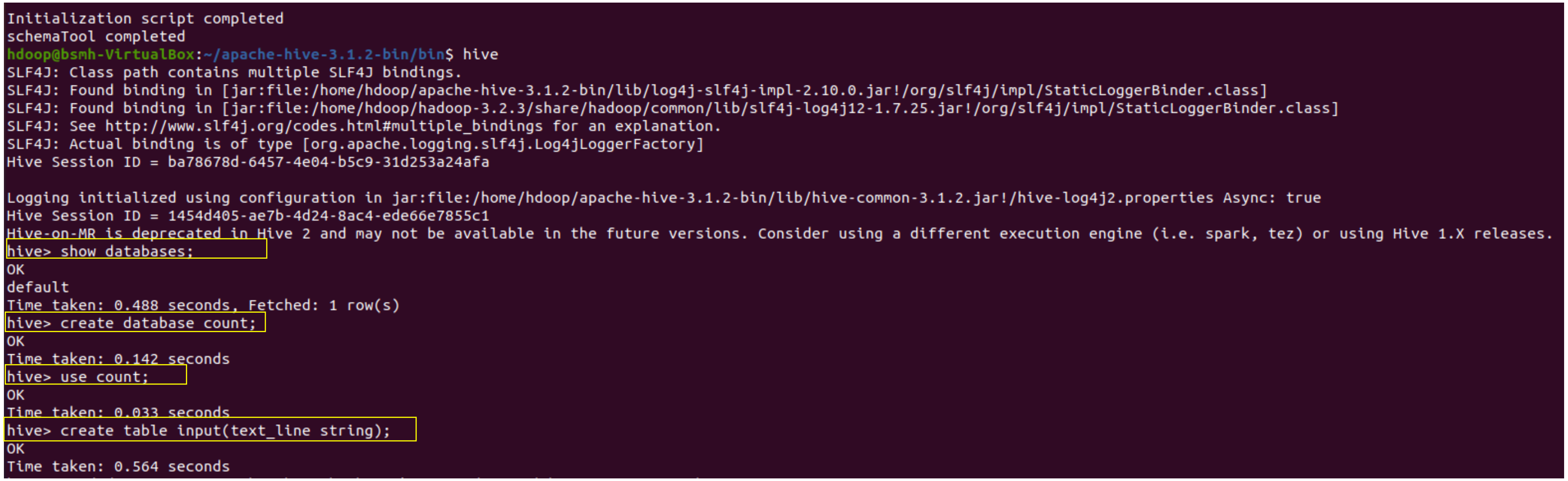
Word count Exercise:

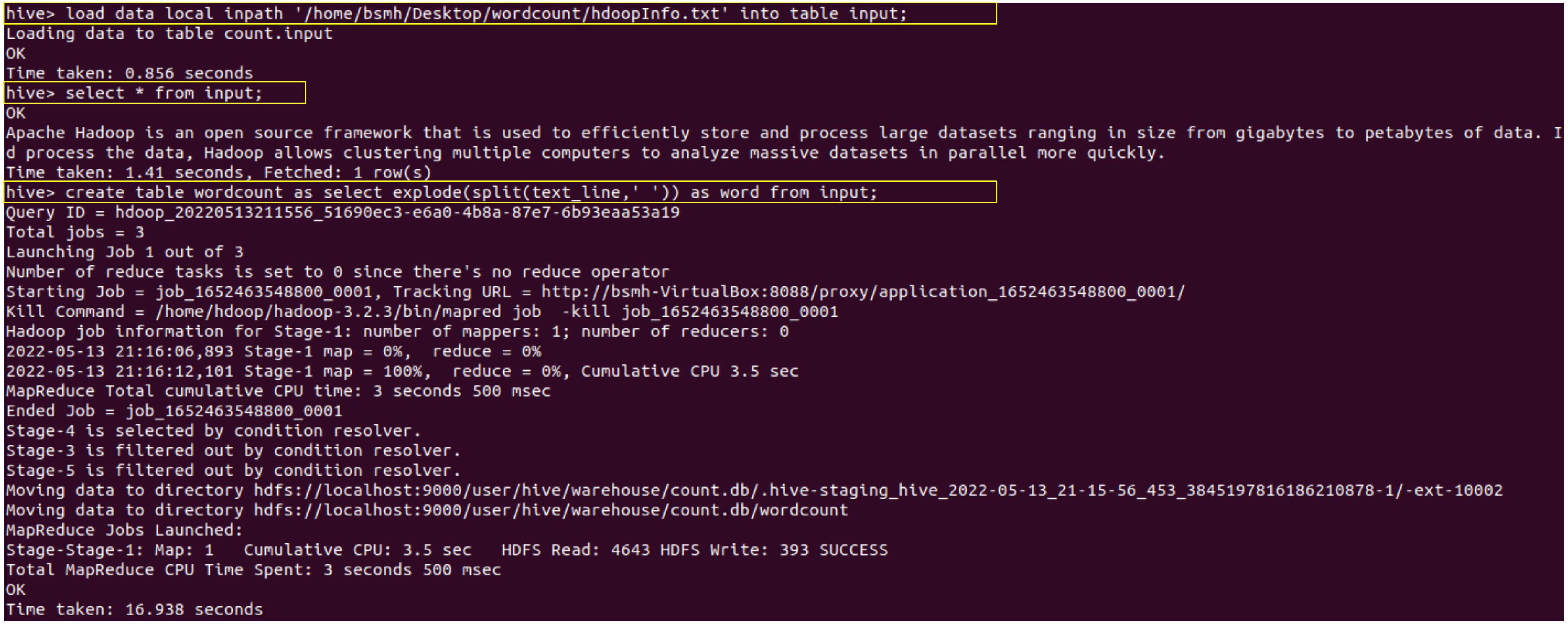
1-



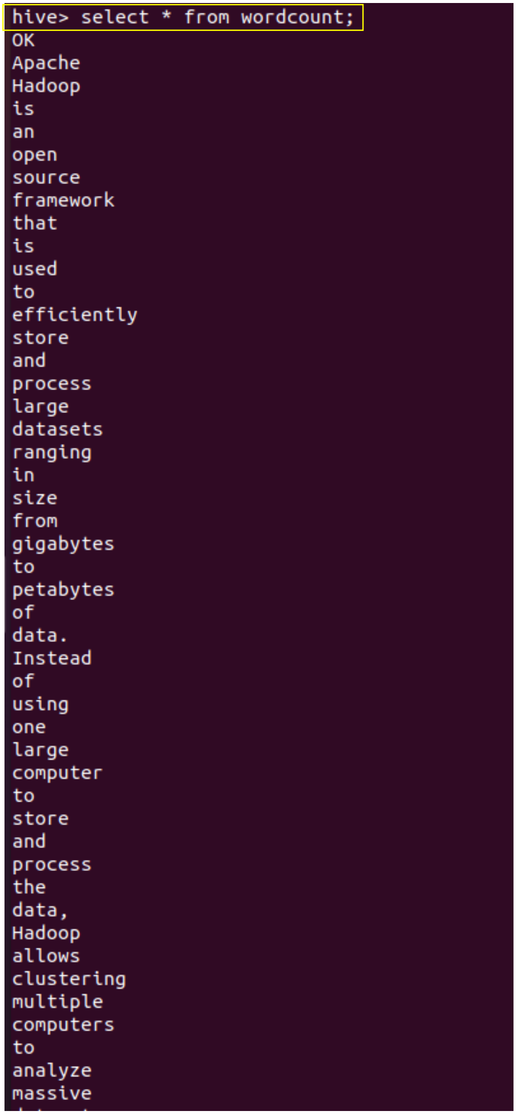
```
Initialization script completed
schemaTool completed
hdoop@bsmh-VirtualBox:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = ba78678d-6457-4e04-b5c9-31d253a24afa

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive Session ID = 1454d405-ae7b-4d24-8ac4-ede66e7855c1
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show databases;
OK
default
Time taken: 0.488 seconds, Fetched: 1 row(s)
hive> create database count;
OK
Time taken: 0.142 seconds
hive> use count;
OK
Time taken: 0.033 seconds
hive> create table input(text_line string);
OK
Time taken: 0.564 seconds
```

2-

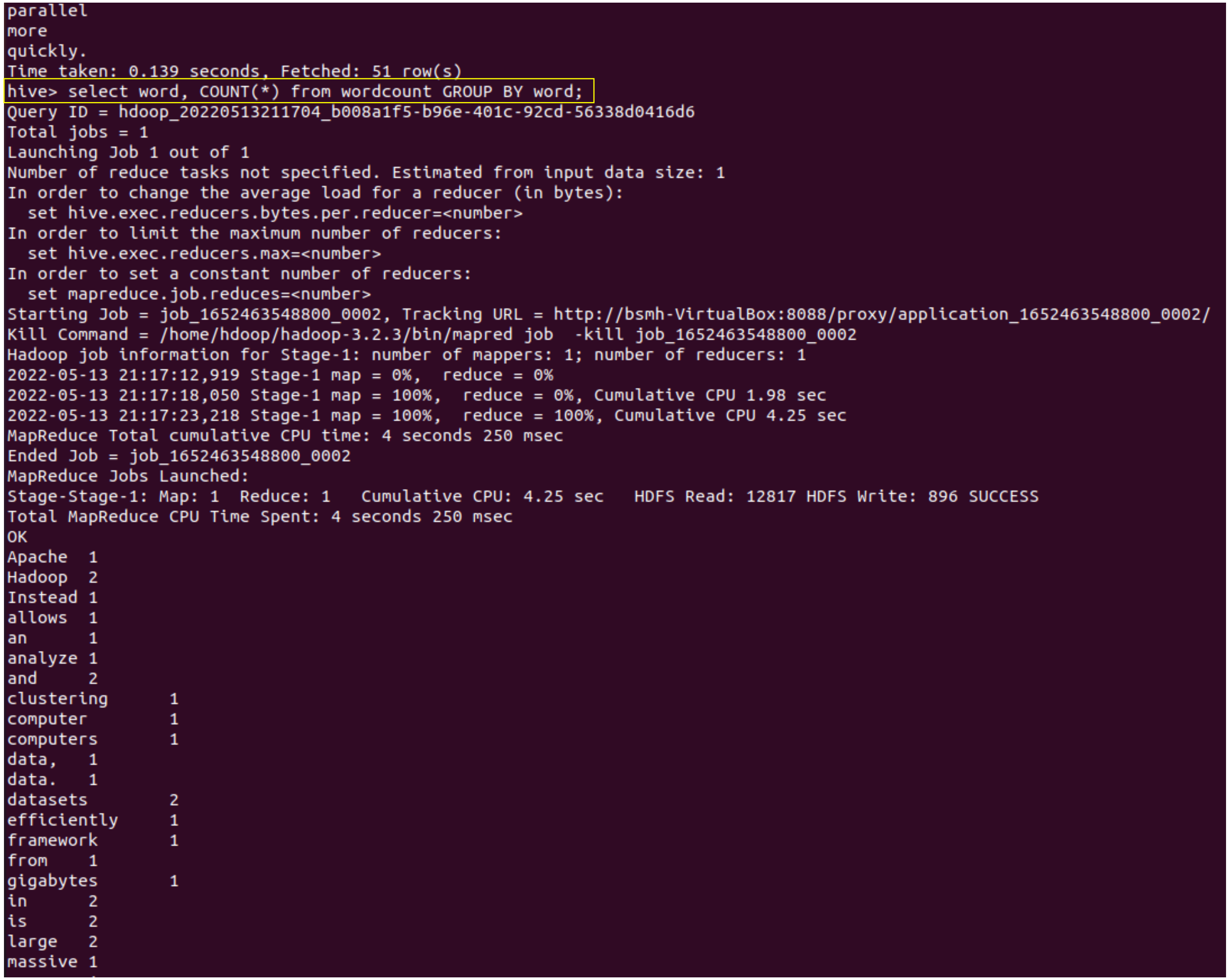


```
hive> load data local inpath '/home/bsmh/Desktop/wordcount/hdoopInfo.txt' into table input;
Loading data to table count.input
OK
Time taken: 0.856 seconds
hive> select * from input;
OK
Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. I
d process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.
Time taken: 1.41 seconds, Fetched: 1 row(s)
hive> create table wordcount as select explode(split(text_line,' ')) as word from input;
Query ID = hdoop_20220513211556_51690ec3-e6a0-4b8a-87e7-6b93eaa53a19
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1652463548800_0001, Tracking URL = http://bsmh-VirtualBox:8088/proxy/application_1652463548800_0001/
Kill Command = /home/hdoop/hadoop-3.2.3/bin/mapred job  -kill job_1652463548800_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-05-13 21:16:06,893 Stage-1 map = 0%,  reduce = 0%
2022-05-13 21:16:12,101 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.5 sec
MapReduce Total cumulative CPU time: 3 seconds 500 msec
Ended Job = job_1652463548800_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/count.db/.hive-staging_hive_2022-05-13_21-15-56_453_3845197816186210878-1/-ext-10002
Moving data to directory hdfs://localhost:9000/user/hive/warehouse/count.db/wordcount
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.5 sec   HDFS Read: 4643 HDFS Write: 393 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 500 msec
OK
Time taken: 16.938 seconds
```



```
hive> select * from wordcount;
OK
Apache
Hadoop
is
an
open
source
framework
that
is
used
to
efficiently
store
and
process
large
datasets
ranging
in
size
from
gigabytes
to
petabytes
of
data.
Instead
of
using
one
large
computer
to
store
and
process
the
data,
Hadoop
allows
clustering
multiple
computers
to
analyze
massive
```

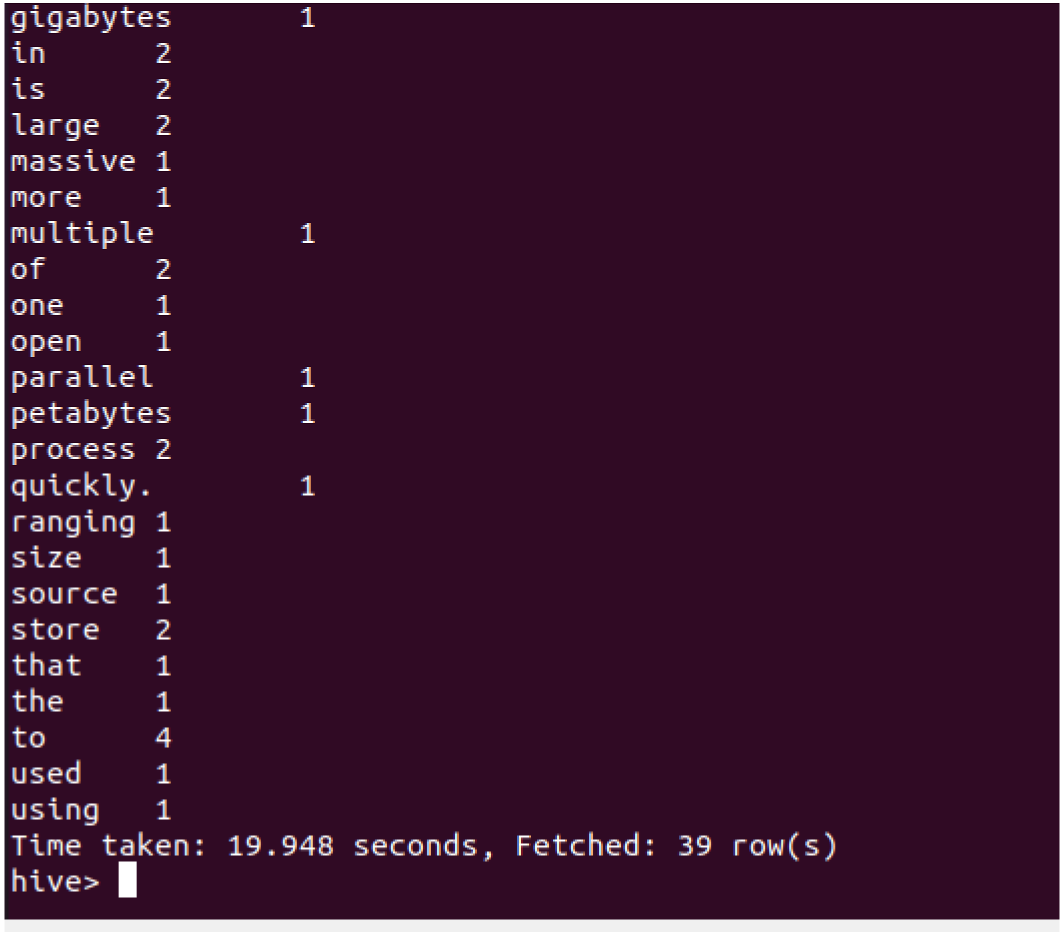
3-



```
parallel
more
quickly.
Time taken: 0.139 seconds, Fetched: 51 row(s)
hive> select word, COUNT(*) from wordcount GROUP BY word;
Query ID = hdoop_20220513211704_b008a1f5-b96e-401c-92cd-56338d0416d6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1652463548800_0002, Tracking URL = http://bsmh-VirtualBox:8088/proxy/application_1652463548800_0002/
Kill Command = /home/hdoop/hadoop-3.2.3/bin/mapred job  -kill job_1652463548800_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-05-13 21:17:12,919 Stage-1 map = 0%,  reduce = 0%
2022-05-13 21:17:18,050 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.98 sec
2022-05-13 21:17:23,218 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.25 sec
MapReduce Total cumulative CPU time: 4 seconds 250 msec
Ended Job = job_1652463548800_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.25 sec   HDFS Read: 12817 HDFS Write: 896 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 250 msec
OK
Apache  1
Hadoop  2
Instead 1
allows  1
an      1
analyze 1
and     2
clustering      1
computer        1
computers       1
data,   1
data.   1
datasets        2
efficiently     1
framework       1
from    1
gigabytes       1
in      2
is      2
large   2
massive 1
```

4-

5-



```
gigabytes       1
in      2
is      2
large   2
massive 1
more    1
multiple        1
of      2
one     1
open    1
parallel        1
petabytes       1
process 2
quickly.        1
ranging 1
size    1
source  1
store   2
that    1
the     1
to      4
used    1
using   1
Time taken: 19.948 seconds, Fetched: 39 row(s)
hive> 
```