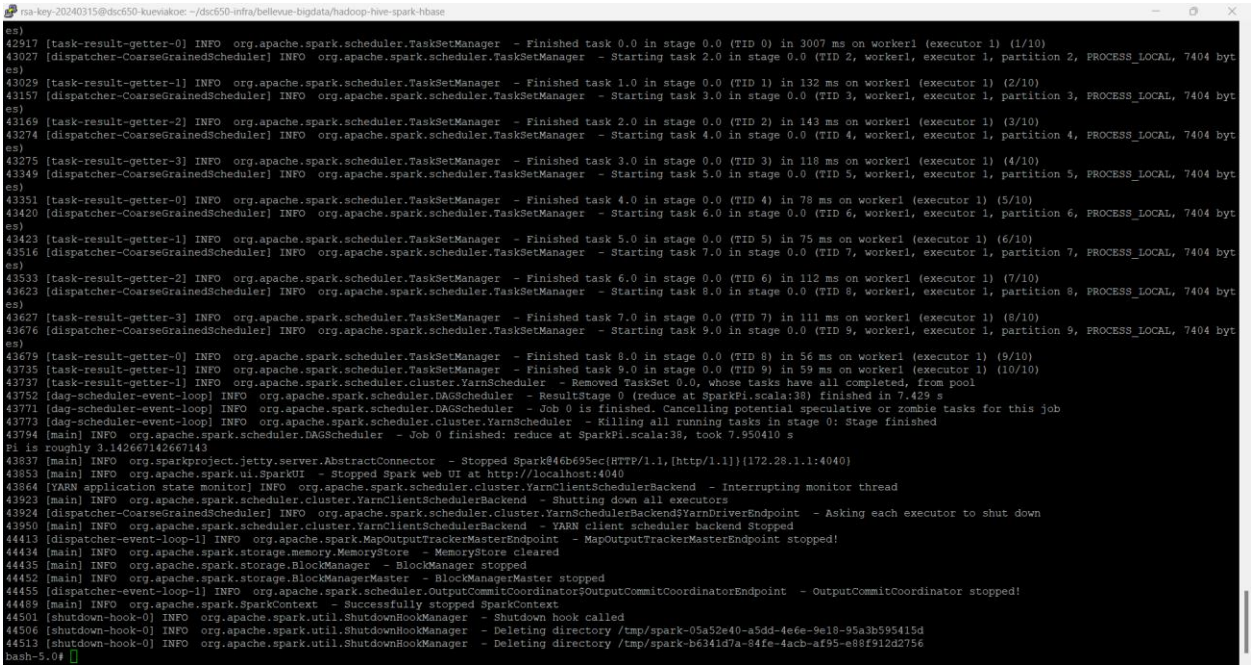
DSC 650

Week 4 Assignment – Introduction to Apache Spark using Scala and PySpark

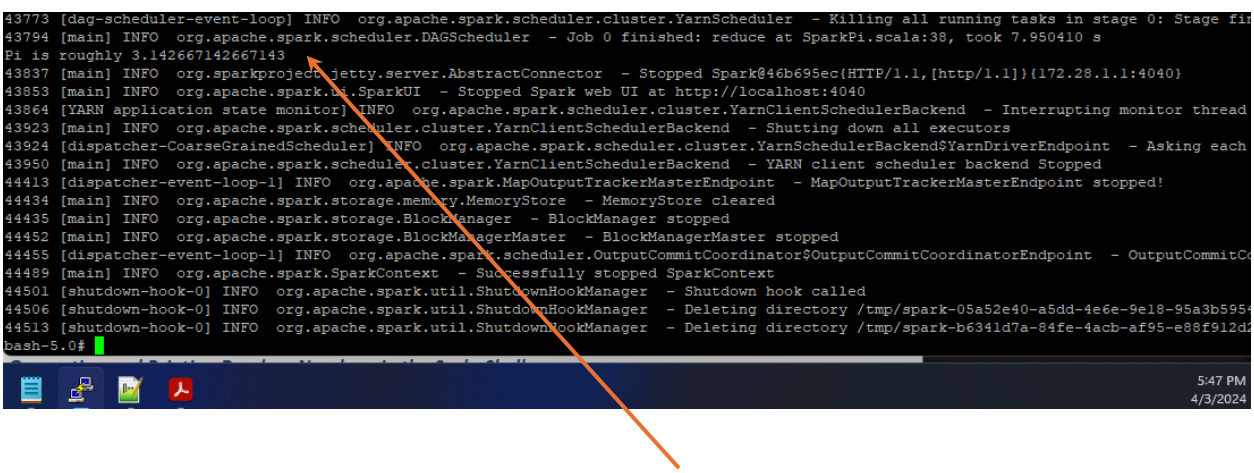
Eyram Kueviakoe

April 03, 2024

Screenshot of the SparkPi output



```
rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase
es)
42917 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 0.0 in stage 0.0 (TID 0) in 3007 ms on worker1 (executor 1) (1/10)
43027 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 2.0 in stage 0.0 (TID 2, worker1, executor 1, partition 2, PROCESS_LOCAL, 7404 byt
es)
43029 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 1.0 in stage 0.0 (TID 1) in 132 ms on worker1 (executor 1) (2/10)
43157 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 3.0 in stage 0.0 (TID 3, worker1, executor 1, partition 3, PROCESS_LOCAL, 7404 byt
es)
43169 [task-result-getter-2] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 2.0 in stage 0.0 (TID 2) in 143 ms on worker1 (executor 1) (3/10)
43274 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 4.0 in stage 0.0 (TID 4, worker1, executor 1, partition 4, PROCESS_LOCAL, 7404 byt
es)
43275 [task-result-getter-3] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 3.0 in stage 0.0 (TID 3) in 118 ms on worker1 (executor 1) (4/10)
43349 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 5.0 in stage 0.0 (TID 5, worker1, executor 1, partition 5, PROCESS_LOCAL, 7404 byt
es)
43351 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 4.0 in stage 0.0 (TID 4) in 78 ms on worker1 (executor 1) (5/10)
43420 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 6.0 in stage 0.0 (TID 6, worker1, executor 1, partition 6, PROCESS_LOCAL, 7404 byt
es)
43423 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 5.0 in stage 0.0 (TID 5) in 75 ms on worker1 (executor 1) (6/10)
43516 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 7.0 in stage 0.0 (TID 7, worker1, executor 1, partition 7, PROCESS_LOCAL, 7404 byt
es)
43533 [task-result-getter-2] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 6.0 in stage 0.0 (TID 6) in 112 ms on worker1 (executor 1) (7/10)
43623 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 8.0 in stage 0.0 (TID 8, worker1, executor 1, partition 8, PROCESS_LOCAL, 7404 byt
es)
43627 [task-result-getter-3] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 7.0 in stage 0.0 (TID 7) in 111 ms on worker1 (executor 1) (8/10)
43676 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.TaskSetManager  - Starting task 9.0 in stage 0.0 (TID 9, worker1, executor 1, partition 9, PROCESS_LOCAL, 7404 byt
es)
43679 [task-result-getter-0] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 8.0 in stage 0.0 (TID 8) in 56 ms on worker1 (executor 1) (9/10)
43735 [task-result-getter-1] INFO  org.apache.spark.scheduler.TaskSetManager  - Finished task 9.0 in stage 0.0 (TID 9) in 59 ms on worker1 (executor 1) (10/10)
43737 [task-result-getter-1] INFO  org.apache.spark.scheduler.cluster.YarnScheduler  - Removed TaskSet 0.0, whose tasks have all completed, from pool
43752 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.DAGScheduler  - ResultStage 0 (reduce at SparkPi.scala:38) finished in 7.429 s
43771 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.DAGScheduler  - Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
43773 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.cluster.YarnScheduler  - Killing all running tasks in stage 0: Stage finished
43794 [main] INFO  org.apache.spark.scheduler.DAGScheduler  - Job 0 finished: reduce at SparkPi.scala:38, took 7.950410 s
Pi is roughly 3.142667142667143
43837 [main] INFO  org.sparkproject.jetty.server.AbstractConnector  - Stopped Spark@46b695ec{HTTP/1.1,[http/1.1]}{172.28.1.1:4040}
43853 [main] INFO  org.apache.spark.ui.SparkUI  - Stopped Spark web UI at http://localhost:4040
43864 [YARN application state monitor] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Interrupting monitor thread
43923 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Shutting down all executors
43924 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnDriverEndpoint  - Asking each executor to shut down
43950 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - YARN client scheduler backend Stopped
44413 [dispatcher-event-loop-1] INFO  org.apache.spark.MapOutputTrackerMasterEndpoint  - MapOutputTrackerMasterEndpoint stopped!
44434 [main] INFO  org.apache.spark.storage.memory.MemoryStore  - MemoryStore cleared
44435 [main] INFO  org.apache.spark.storage.BlockManager  - BlockManager stopped
44452 [main] INFO  org.apache.spark.storage.BlockManagerMaster  - BlockManagerMaster stopped
44455 [dispatcher-event-loop-1] INFO  org.apache.spark.scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint  - OutputCommitCoordinator stopped!
44489 [main] INFO  org.apache.spark.SparkContext  - Successfully stopped SparkContext
44501 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Shutdown hook called
44506 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-05a52e40-a5dd-4e6e-9e18-95a3b595415d
44513 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-b6341d7a-84fe-4acb-af95-e88f912d2756
bash-5.0#
```



```
43773 [dag-scheduler-event-loop] INFO  org.apache.spark.scheduler.cluster.YarnScheduler  - Killing all running tasks in stage 0: Stage fi
43794 [main] INFO  org.apache.spark.scheduler.DAGScheduler  - Job 0 finished: reduce at SparkPi.scala:38, took 7.950410 s
Pi is roughly 3.142667142667143
43837 [main] INFO  org.sparkproject.jetty.server.AbstractConnector  - Stopped Spark@46b695ec{HTTP/1.1,[http/1.1]}{172.28.1.1:4040}
43853 [main] INFO  org.apache.spark.ui.SparkUI  - Stopped Spark web UI at http://localhost:4040
43864 [YARN application state monitor] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Interrupting monitor thread
43923 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - Shutting down all executors
43924 [dispatcher-CoarseGrainedScheduler] INFO  org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnDriverEndpoint  - Asking each
43950 [main] INFO  org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend  - YARN client scheduler backend Stopped
44413 [dispatcher-event-loop-1] INFO  org.apache.spark.MapOutputTrackerMasterEndpoint  - MapOutputTrackerMasterEndpoint stopped!
44434 [main] INFO  org.apache.spark.storage.memory.MemoryStore  - MemoryStore cleared
44435 [main] INFO  org.apache.spark.storage.BlockManager  - BlockManager stopped
44452 [main] INFO  org.apache.spark.storage.BlockManagerMaster  - BlockManagerMaster stopped
44455 [dispatcher-event-loop-1] INFO  org.apache.spark.scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint  - OutputCommitC
44489 [main] INFO  org.apache.spark.SparkContext  - Successfully stopped SparkContext
44501 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Shutdown hook called
44506 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-05a52e40-a5dd-4e6e-9e18-95a3b595
44513 [shutdown-hook-0] INFO  org.apache.spark.util.ShutdownHookManager  - Deleting directory /tmp/spark-b6341d7a-84fe-4acb-af95-e88f912d
bash-5.0#
5:47 PM
4/3/2024
```

Screenshot of the 100 generated random numbers

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
scala> numbersRDD.take(100).foreach(println)
871
563
696
304
373
28
539
154
543
188
204
905
207
235
408
518
848
892
104
426
384
706
882
305
616
679
471
90
426
584
457
928
718
868
35
283
25
251
325
599
898
307
538
23
826
710
744
98
```

Transformation 1: Convert the sentence into uppercase.

Our first transformation will be to convert each sentence into uppercase.

The command we will execute is:

val transformedSentences = sentencesRDD.map(sentence => sentence.toUpperCase())

transformedSentences.take(100).foreach(println)

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
scala> val transformedSentences = sentencesRDD.map(sentence => sentence.toUpperCase())
transformedSentences: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at map at <console>:25

scala> transformedSentences.take(100).foreach(println)
DATE GRAPE CHERRY HONEYDEW.
APPLE BANANA.
APPLE CHERRY DATE HONEYDEW.
GRAPE APPLE DATE ELDERBERRY.
HONEYDEW APPLE BANANA.
DATE CHERRY GRAPE.
GRAPE BANANA ELDERBERRY FIG DATE CHERRY.
FIG GRAPE APPLE.
APPLE DATE ELDERBERRY FIG HONEYDEW.
ELDERBERRY FIG DATE.
ELDERBERRY DATE BANANA.
ELDERBERRY APPLE.
DATE ELDERBERRY CHERRY.
HONEYDEW APPLE ELDERBERRY CHERRY DATE.
DATE BANANA APPLE ELDERBERRY GRAPE CHERRY.
GRAPE BANANA APPLE.
ELDERBERRY APPLE DATE.
GRAPE DATE BANANA FIG.
HONEYDEW.
CHERRY HONEYDEW FIG ELDERBERRY.
FIG BANANA GRAPE CHERRY ELDERBERRY DATE.
ELDERBERRY DATE APPLE CHERRY.
CHERRY HONEYDEW APPLE.
APPLE HONEYDEW ELDERBERRY.
GRAPE HONEYDEW CHERRY DATE ELDERBERRY APPLE.
HONEYDEW ELDERBERRY CHERRY.
FIG ELDERBERRY.
ELDERBERRY CHERRY GRAPE HONEYDEW FIG.
HONEYDEW FIG GRAPE ELDERBERRY BANANA CHERRY.
FIG.
HONEYDEW DATE CHERRY APPLE FIG.
ELDERBERRY CHERRY GRAPE BANANA DATE HONEYDEW.
CHERRY ELDERBERRY DATE HONEYDEW.
DATE APPLE GRAPE HONEYDEW CHERRY.
HONEYDEW FIG APPLE BANANA GRAPE ELDERBERRY.
APPLE.
ELDERBERRY BANANA HONEYDEW FIG DATE.
APPLE.
FIG DATE ELDERBERRY APPLE BANANA CHERRY.
ELDERBERRY APPLE.
BANANA CHERRY HONEYDEW DATE.
BANANA CHERRY.
APPLE BANANA GRAPE CHERRY.
GRAPE BANANA.
HONEYDEW ELDERBERRY BANANA.
```

Transformation 2: Sort the words in the sentence alphabetically

For the second transformation, we will sort the words in the sentence alphabetically

The code we will execute is:

val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").sorted.mkString(" "))

transformedSentences.take(100).foreach(println)

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
scala> val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").sorted.mkString(" "))
transformedSentences: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[8] at map at <console>:25

scala> transformedSentences.take(100).foreach(println)
cherry date grape honeydew.
apple banana.
apple cherry date honeydew.
apple date elderberry. grape
apple banana. honeydew
cherry date grape.
banana cherry. date elderberry fig grape
apple. fig grape
apple date elderberry fig honeydew.
date. elderberry fig
banana. date elderberry
apple. elderberry
cherry. date elderberry
apple cherry date. elderberry honeydew
apple banana cherry. date elderberry grape
apple. banana grape
apple date. elderberry
banana date fig. grape
honeydew.
cherry elderberry. fig honeydew
banana cherry date. elderberry fig grape
apple cherry. date elderberry
apple. cherry honeydew
apple elderberry. honeydew
apple. cherry date elderberry grape honeydew
cherry. elderberry honeydew
elderberry. fig
cherry elderberry fig. grape honeydew
banana cherry. elderberry fig grape honeydew
fig.
apple cherry date fig. honeydew
banana cherry date elderberry grape honeydew.
cherry date elderberry honeydew.
apple cherry. date grape honeydew
apple banana elderberry. fig grape honeydew
apple.
banana date. elderberry fig honeydew
apple.
apple banana cherry. date elderberry fig
apple. elderberry
banana cherry date. honeydew
```

Transformation 3: Join the words in the sentence with * as a delimiter

For the 3rd transformation, we will use * to join the words in the sentence

The code we will execute is:

val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").mkString("*"))

transformedSentences.take(100).foreach(println)

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
scala> val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").mkString("*"))
transformedSentences: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at map at <console>:25

scala> transformedSentences.take(100).foreach(println)
date*grape*cherry*honeydew.
apple*banana.
apple*cherry*date*honeydew.
grape*apple*date*elderberry.
honeydew*apple*banana.
date*cherry*grape.
grape*banana*elderberry*fig*date*cherry.
fig*grape*apple.
apple*date*elderberry*fig*honeydew.
elderberry*fig*date.
elderberry*date*banana.
elderberry*apple.
date*elderberry*cherry.
honeydew*apple*elderberry*cherry*date.
date*banana*apple*elderberry*grape*cherry.
grape*banana*apple.
elderberry*apple*date.
grape*date*banana*fig.
honeydew.
cherry*honeydew*fig*elderberry.
fig*banana*grape*cherry*elderberry*date.
elderberry*date*apple*cherry.
cherry*honeydew*apple.
apple*honeydew*elderberry.
grape*honeydew*cherry*date*elderberry*apple.
honeydew*elderberry*cherry.
fig*elderberry.
elderberry*cherry*grape*honeydew*fig.
honeydew*fig*grape*elderberry*banana*cherry.
fig.
honeydew*date*cherry*apple*fig.
elderberry*cherry*grape*banana*date*honeydew.
cherry*elderberry*date*honeydew.
date*apple*grape*honeydew*cherry.
honeydew*fig*apple*banana*grape*elderberry.
apple.
elderberry*banana*honeydew*fig*date.
apple.
fig*date*elderberry*apple*banana*cherry.
elderberry*apple.
banana*cherry*honeydew*date.
banana*cherry.
apple*banana*grape*cherry.
grape*banana.
honeydew*elderberry*banana.
```

Transformation 4: We will replace all vowels by '-'

For the 4th transformation, we will replace all vowels in the sentence with a dash.

The code we will execute is:

val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").map(_.replaceAll("[aeiouAEIOU]", "-")).mkString(" "))

transformedSentences.take(100).foreach(println)

rsa-key-20240315@dsc650-kueviakoe: ~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase

```
scala> val transformedSentences = sentencesRDD.map(sentence => sentence.split(" ").map(_.replaceAll("[aeiouAEIOU]", "-")).mkString(" "))
transformedSentences: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[9] at map at <console>:25

scala> transformedSentences.take(100).foreach(println)
d-t- gr-p- ch-rry h-n-yd-w.
-ppl- b-n-n-.
-ppl- ch-rry d-t- h-n-yd-w.
gr-p- -ppl- d-t- -ld-rb-rry.
h-n-yd-w -ppl- b-n-n-.
d-t- ch-rry gr-p-.
gr-p- b-n-n- -ld-rb-rry f-g d-t- ch-rry.
f-g gr-p- -ppl-.
-ppl- d-t- -ld-rb-rry f-g h-n-yd-w.
-ld-rb-rry f-g d-t-.
-ld-rb-rry d-t- b-n-n-.
-ld-rb-rry -ppl-.
d-t- -ld-rb-rry ch-rry.
h-n-yd-w -ppl- -ld-rb-rry ch-rry d-t-.
d-t- b-n-n- -ppl- -ld-rb-rry gr-p- ch-rry.
gr-p- b-n-n- -ppl-.
-ld-rb-rry -ppl- d-t-.
gr-p- d-t- b-n-n- f-g.
h-n-yd-w.
ch-rry h-n-yd-w f-g -ld-rb-rry.
f-g b-n-n- gr-p- ch-rry -ld-rb-rry d-t-.
-ld-rb-rry d-t- -ppl- ch-rry.
ch-rry h-n-yd-w -ppl-.
-ppl- h-n-yd-w -ld-rb-rry.
gr-p- h-n-yd-w ch-rry d-t- -ld-rb-rry -ppl-.
h-n-yd-w -ld-rb-rry ch-rry.
f-g -ld-rb-rry.
-ld-rb-rry ch-rry gr-p- h-n-yd-w f-g.
h-n-yd-w f-g gr-p- -ld-rb-rry b-n-n- ch-rry.
f-g.
h-n-yd-w d-t- ch-rry -ppl- f-g.
-ld-rb-rry ch-rry gr-p- b-n-n- d-t- h-n-yd-w.
ch-rry -ld-rb-rry d-t- h-n-yd-w.
d-t- -ppl- gr-p- h-n-yd-w ch-rry.
h-n-yd-w f-g -ppl- b-n-n- gr-p- -ld-rb-rry.
-ppl-.
-ld-rb-rry b-n-n- h-n-yd-w f-g d-t-.
-ppl-.
f-g d-t- -ld-rb-rry -ppl- b-n-n- ch-rry.
-ld-rb-rry -ppl-.
b-n-n- ch-rry h-n-yd-w d-t-.
b-n-n- ch-rry.
-ppl- b-n-n- gr-p- ch-rry.
gr-p- b-n-n-.
h-n-yd-w -ld-rb-rry b-n-n-.
```