Google Word Datasets

<http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>

RHIPE is interested in comparisons

<https://groups.google.com/forum/#!topic/rhipe/EpLL8EWK_pE>

Single file

enwiki-latest-pages-articles.xml #48GB

RHIPE

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 | Time #4 | Block |
| 1 | 1539 | 385 | 4mins, 15sec | 5mins, 18sec | 5mins, 18sec | 6mins, 27sec | 32 |
| 2 | 770 | 192 | 5mins, 51sec | 5mins, 6sec | 4mins, 50sec | 4mins, 34sec | 64 |
| 3 | 513 | 142 | 6mins, 5sec | 5mins, 42sec | 5mins, 51sec | 6mins, 0sec | 96 |
| 4 | 385 | 96 | 6mins, 15sec | 6mins, 3sec | 5mins, 29sec | 6mins, 30sec | 128 |

RHADOOP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |
| 1 | 1539 | 385 |  |  |  |
| 2 | 770 | 192 |  |  |  |
| 3 | 513 | 142 | 33mins, 14sec | 37mins, 46sec |  |
| 4 | 385 | 96 |  |  |  |

Streaming Py

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |  |  |  |
| 1 | 1539 | 385 | 5mins, 38sec | 5mins, 39sec | 5mins, 39sec | 5mins, 54sec | 5mins, 35sec | 5mins, 54sec |
| 2 | 770 | 192 | 6mins, 21sec | 6mins, 42sec | 6mins, 24sec |  |  |  |
| 3 | 513 | 142 | 6mins, 17sec |  |  |  |  |  |
| 4 | 385 | 96 |  |  |  |  |  |  |

Streaming R

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |
| 1 | 1539 | 385 | FAILS | FAILS | FAILS |
| 2 | 770 | 192 |  |  |  |
| 3 | 513 | 142 |  |  |  |
| 4 | 385 | 96 |  |  |  |

Java

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 | Time #4 | Block |
| 1 |  |  |  |  |  |  | 32 |
| 2 |  |  |  |  |  |  | 64 |
| 3 |  |  |  |  |  |  | 96 |
| 4 |  |  |  |  |  |  | 128 |

Single file reading

936MB source /user/dotcz12/data/googletxt/googlebooks-eng-all-5gram-20090715-0.csv

RHIPE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 | Block |
| 1 | 118 | 28 | 1mins, 14sec | 1mins, 13sec | 1mins, 11sec | 8 |
| 2 | 59 | 14 | 1mins, 34sec | 1mins, 32sec | 1mins, 27sec | 16 |
| 3 | 15 | 4 | 3mins, 4sec | 3mins, 11sec | 2mins, 55sec | 64 |
| 4 | 10 | 3 | 3mins, 24sec | 3mins, 27sec | 3mins, 40sec | 96 |
| 5 | 8 | 2 | 4mins, 55sec | 4mins, 41sec | 5mins, 34sec | 128 |

RHADOOP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |
| 1 | 118 | 28 | 2mins, 59sec | 2mins, 55sec | 3mins, 3sec |
| 2 | 59 | 14 | 3mins, 59sec | 3mins, 54sec | 3mins, 44sec |
| 3 | 15 | 4 | 12mins, 51sec | 12mins, 32sec | 12mins, 51sec |
| 4 | 10 | 3 | 16mins, 31sec | 18mins, 41sec | 17mins, 17sec |
| 5 | 8 | 2 | 22mins, 49sec | 25mins, 52sec | 23mins, 0sec |

Directory reading

20\*937MB ~ 18GB

source /user/dotcz12/data/googletxt/

RHIPE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 | Block |
| 1 |  |  |  |  |  | 16 |
| 2 |  |  |  |  |  | 64 |
| 3 |  |  |  |  |  | 96 |
| 4 |  |  |  |  |  | 128 |

RHADOOP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |

Directory reading (ZIP)

20\*112MB ~ 2.5GB

source /user/dotcz12/data/googlezip/

RHIPE

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |  | Block |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |  |

RHADOOP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Mappers | Reducers | Time #1 | Time #2 | Time #3 |
| 1 | 770 | 192 |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |

[dotcz12@login TACCPROJECT]$ ./tests.sh rhipe 8 128 96

Loading required package: methods

Loading required package: codetools

Loading required package: rJava

------------------------------------------------

| Please call rhinit() else RHIPE will not run |

------------------------------------------------

Rhipe: Using Rhipe.jar file

Initializing Rhipe v0.75.0

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/usr/lib/hadoop/client/slf4j-log4j12.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12.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.slf4j.impl.Log4jLoggerFactory]

2015-03-31 04:51:18,634 WARN [main][NativeCodeLoader] Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Initializing mapfile caches

Deleted 1 file

Saving 1 parameter to /tmp/rhipe-temp-params-00389d114c68d30e97a484733fe39454 (use rhclean to delete all temp files)

java.io.IOException: java.io.EOFException

at org.godhuli.rhipe.RHMR.readParametersFromR(RHMR.java:437)

at org.godhuli.rhipe.RHMR.init(RHMR.java:124)

at org.godhuli.rhipe.RHMR.run(RHMR.java:174)

at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)

at org.godhuli.rhipe.RHMR.fmain(RHMR.java:97)

at org.godhuli.rhipe.PersonalServer.rhex(PersonalServer.java:147)

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)

at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:606)

at RJavaTools.invokeMethod(RJavaTools.java:386)

Caused by: java.io.EOFException

at java.io.DataInputStream.readFully(DataInputStream.java:197)

at org.godhuli.rhipe.RHMR.readParametersFromR(RHMR.java:432)

... 10 more

java.lang.RuntimeException: java.io.IOException: java.io.EOFException

at org.godhuli.rhipe.RHMR.init(RHMR.java:136)

at org.godhuli.rhipe.RHMR.run(RHMR.java:174)

at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)

at org.godhuli.rhipe.RHMR.fmain(RHMR.java:97)

at org.godhuli.rhipe.PersonalServer.rhex(PersonalServer.java:147)

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)

at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)

at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:606)

at RJavaTools.invokeMethod(RJavaTools.java:386)

Caused by: java.io.IOException: java.io.EOFException

at org.godhuli.rhipe.RHMR.readParametersFromR(RHMR.java:437)

at org.godhuli.rhipe.RHMR.init(RHMR.java:124)

... 9 more

Caused by: java.io.EOFException

at java.io.DataInputStream.readFully(DataInputStream.java:197)

at org.godhuli.rhipe.RHMR.readParametersFromR(RHMR.java:432)

... 10 more

Error in .jcall("RJavaTools", "Ljava/lang/Object;", "invokeMethod", cl, :

java.lang.RuntimeException: java.io.IOException: java.io.EOFException

Calls: rhwatch ... tryCatch -> tryCatchList -> tryCatchOne -> <Anonymous>

In addition: Warning messages:

1: In writeBin(as.integer(nchar(x)), conffile, size = 4, endian = "big") :

problem writing to connection

2: In writeBin(charToRaw(x), conffile, endian = "big") :

problem writing to connection

Execution halted

Just keep retrying….. It’ll work after x many tries… ? The cluster is strained for resources.