A: cloudera hadoop mapreduce job GC overhead limit exceeded error

The Mahout jobs are very memory intensive. I don't know whether the mappers or reducers are the culprits, but, either way you will have to tell **Hadoop** to give them more RAM. "GC Overhead Limit … Exceeded" is just a way of saying "**out** of memory" -- means the JVM gave up trying to reclaim the last 0.01% of available RAM. How you set this is indeed a little complex, because there are several …

answered oct 18 '13 by Sean Owen

2 answers

Q: Hadoop MapReduce: Read a file and use it as input to filter other files

out of the other files in the folder. How can I achieve this? By the way, I have a running **hadoop** mapreduce application which takes as input a path to a folder, does the processing and writes **out** … I would like to write a **hadoop** application which takes as input a file and an input folder which contains several files. The single file contains keys whose records need to be selected and extracted …

java hadoop mapreduce

asked jun 19 '12 by Bob

A: Hadoop Balancer fails with - IOException: Couldn't set up IO streams (LeaseRenewer Warning)

Once the number of threads created by **Hadoop** RPC reaches the (ulimit -u) on the node's number of processes, java will report it as an **out**-of-memory error. Try increasing the maximum number of processes allowed, i.e. your ulimit -u value. …

answered may 29 by David Kjerrumgaard

1 answer

Q: Why do I have multple instances of Java running after installing HDnsight?

I've installed HDInsight on a desktop computer to learn to work with Hive. When I restarted my computer and logged into my profile everything was moving slow. I've figured **out** that **Hadoop** is running multiple instances of Java.exe. Is there a way to limit the amount of memory that **Hadoop** uses for this? …

hadoop hive hdinsight azul-zulu

asked dec 3 by Tremvelope_Filter

A: Hadoop Processing time in clustered and standalone system

running a cluster in virtual-machines will only slow down your map-reduce (because of the overhead from running the virtual-os and multiple **hadoop** instances) especially if you run **out** of memory …

answered feb 21 '13 by Josh

1 answer

Q: How to see hadoop's heap use?

I am doing a school work to analyze the use of heap in **hadoop**. It involves running two versions of a mapreduce program to calculate the median of the length of forum comments: the first one … to handle the data. The purpose is to use both programs to process data of different sizes and watch how the memory usage goes up faster in the first one (until it eventually runs **out** of memory). My …

hadoop mapreduce heap

asked jun 22 '13 by user2510940

0 answers

Q: Hadoop MapReduce vs MPI (vs Spark vs Mahout vs Mesos) - When to use one over the other?

I am new to parallel computing and just starting to try **out** MPI and **Hadoop**+MapReduce on Amazon AWS. But I am confused about when to use one over the other. For example, one common rule of thumb … on MPI (MR-MPI) which does not provide fault tolerance but seems to be more efficient on some benchmarks than MapReduce on **Hadoop**, and seems to handle big data using **out**-of-core memory. Conversely …

hadoop parallel-processing mapreduce mpi

asked jan 6 by crackjack

1 answer

Q: Hadoop component is not starting

I'm new to **hadoop** well I've followed micheal install (http://www.michael-noll.com/tutorials/running-**hadoop**-on-ubuntu-linux-single-node-cluster/). When I run the command /usr/local/**hadoop**/bin/start …

hadoop

asked sep 20 '13 by mezgani

2 answers

Q: Hadoop nodes die (crash) after a while

): Another machine showed this error, with **hadoop** dfs running, while no job was running: and Here is another screenshot (**out** of which I cannot make any sense): Here is the log of a crashed … I have a **hadoop** cluster of 16 (ubuntu 12.04 server) nodes (1 master and 15 slaves). They are connected through a private network and the master also has a public IP (it belongs to two networks). When …

networking ubuntu hadoop cluster-computing

asked jan 31 '14 by vefthym

0 answers

Q: Hadoop 2.2.0 stuck in Accepted state in Jobtracker

So I am trying to get a simple 1 node **hadoop** setup running (as in 1 manager and 1 worker), Virtual Cluster has been working just fine on either machine but when I try to change it to a real cluster …

hadoop cloud hadoop2

asked jul 17 by user3846838

0 answers

Q: Computing simple moving average using Map Reduce in MongoDB

, why is it that **Hadoop** Reduce phase doesn't crash **out** of memory, since it has to deal with at least several TBs of mapped data. … I stumbled upon this article:

http://blog.cloudera.com/blog/2011/04/simple-moving-average-secondary-sort-and-mapreduce-part-3/ which mentions how to calculate moving average using **Hadoop** …

mongodb hadoop mapreduce

asked may 16 '13 by P.Prasad

A: How could I use mahout with hadoop in Eclipse?

with the **Hadoop** cluster). Access by typing and it will print a help screen. Once you have run the **hadoop** job on the cluster you will need to write code to lookup the recs for a specific user **out** … Your example is not **Hadoop** code. The Mahout recommenders come in non-**hadoop** "in-memory" versions, as you've used in your example, and **Hadoop** versions. The **Hadoop** version has a very different API …

answered jun 30 by pferrel

0 answers

Q: Hadoop YARN reducer/shuffle stuck

is mostly 100% idle. The job can run successfully on **Hadoop** 1.2.1. I checked the log messages from resourcemanager and found **out** that since map finished, no more container was allocated so there's … I was migrating from **Hadoop** 1 to **Hadoop** 2 YARN. Source code were recompiled using MRV2 jars and didn't have any compatibility issue. When I was trying to run the job under YARN, map worked fine …

hadoop shuffle reduce yarn

asked nov 26 by Chandler Lee

A: Hadoop on Windows

I use **Hadoop** natively on Windows as a virtual 2-node cluster running on one machine. It runs inside Cygwin (so no VM). Works well to try **Hadoop out** and I still use it to test new code in small before … putting it on the cluster. You basically get every bit of functionality as with a full cluster. Getting it to work can be a bit tricky though. I used the following short guide: Stanford **Hadoop** …

answered oct 16 '12 by Jeroen Vuurens

A: How many Mapreduce Jobs can be run simultaneously

of running some **hadoop** code, which requires some amount of memory, so eventually you would run **out** of memory on your machine. You might also have to configure job queues cleverly in order to run a ton at the same time. Now, what is possible is a very different question than what is a good idea... …

answered oct 30 '13 by Joe K

1 answer

Q: Cassandra setInputSplitSize is not working properly

I am using **Hadoop** + Cassandra. I use setInputSplitSize(1000) to not overload mappers (and receive **out** of heap memory) as default it is 64K. All together I have only 2M lines to process. Actually … %. When I check the log, I found 40K-64K rows processed. It is not crashing or giving **out** of memory, but these 2-3 tasks begin in the middle of processing and continue for 2-3 hours after all other have …

java hadoop mapreduce cassandra

asked aug 11 '11 by Anton

1 answer

Q: Using a Lucene Index as an Input For Hadoop

I am trying to build an adjacency list **out** of a corpus. I am thinking of using Map-Reduce because in-memory solutions have proven to be extremely expensive. The sequence of jobs that I think … inverted index --- I want to use a Lucene index which seems rather easy to generate. However, I am not really clear how I could take the Lucene index and generate pairs that Map in **Hadoop** can use? Could some one clarify how one goes about doing that? …

lucene hadoop

asked jun 17 '11 by dvk

0 answers

Q: Why Hadoop Job slows down after some time and after some run?

I have a **Hadoop** cluster of 2 nodes running latest stable version(**hadoop**-0.20.203.0). I do some memory intensive work in my map function.My job run properly for first few runs. But now it slows down … and terminates after some time. I tried cleaning **out** the mapred.local.dir nad mapred.tmp.dir but not helped. Sometimes it says "GC overhead". Sometimes it shows "just killed". …

garbage-collection hadoop mapreduce

asked jan 16 '12 by samarth

2 answers

Q: Oracle R Enterprise (ORE) KMeans Package

I have a task to run K-Means clustering algorithm on a SAS server but ran **out** of memory. The dataset is 500G, i know i can sample it down to fit into memory, but if I want to run the model … of available Oracle R packages? Will I also run into memory issue if I run the kmeans algorithm (R CRAN package) in Oracle R Enterprise? Is there any R clustering package available in BDA that is written to run on distributed **Hadoop** clusters? Thanks …

oracle r hadoop

asked may 31 '12 by user1319866

2 answers

Q: Spring hadoop Mapper configuration

I'm using **Hadoop** 1.2.1 and Spring **Hadoop** 1.0.2 I wanted to check the Spring autowiring in a **Hadoop** Mapper. I wrote this configuration file: Then I created this Mapper } As you can see …

spring hadoop

asked dec 2 '13 by Angelo Immediata

A: Removing Duplicate Words Across Multiple and Large Dictionary Files

On that that scale of 300GB+, you may want to consider using **Hadoop** or some other scalable store - otherwise, you will have to deal with memory issues through your own coding. You can try other … , more direct methods (UNIX scripting, small C/C++ programs, etc...), but you will likely run **out** of memory unless you have a ton of duplicate words in your data. Addendum Just came across memcached …

answered jan 31 '13 by kfmfe04

1 answer

Q: Is the input to a Hadoop reduce function complete with regards to its key?

see in White's book the discussion about "the shuffle" and am tempted to wonder if when you come **out** of merging and the input to a reducer is sorted by key, if all the data for a key … a vertical (?) partition where the values for a particular key are in different files. Said another way, the columns for a complete record each come from different files. Does **Hadoop** re-assemble that? ...at least for a single key at a time. …

hadoop mapreduce

asked nov 21 '11 by Chris

0 answers

Q: Cannot Understand the TOP command output on Hadoop Datanode

so confused and wondering if some experienced **hadoop** admin could help me understand if my cluster is working fine. Why there is only 1 task running **out** of 897 while the other 896 sleeping …

linux hadoop hdfs cloudera metrics

asked jun 1 by B.Mr.W.

1 answer

Q: Unexpected output from the mapper. It adds a number before the output

So I am giving my Mapper an input from another MapReduce job. Here I had done some partitioning of my input so that the reducer iterable doesn't go **out** of memory (This is just a test program). So … was to limit the size of value iterable in Reducer. where for words **hadoop** and fs we would get output as 5 and 5 . Here I've limited the reducer values to 3 by partitioning the mapper somehow but I'm …

java hadoop

asked sep 17 by Akshay Hazari

A: Hadoop map step returns to big dataset

If this was an actual issue **hadoop** would be crippled. Data is stored in the Reducer Nodes local directory. **Hadoop** by design is meant to take on big data problems 100G text files are nothing. **Hadoop** … is ready to take terabytes of data and run jobs on it. But if you are running **out** of virtual memory in your actual method calls(like for some reason you trying to load 100g into a data structure …

answered feb 5 '14 by Dan Ciborowski

1 answer

Q: Suspending hadoop nodes temporarily - background hadoop cluster

I wonder if it is possible to install a "background" **hadoop** cluster. I mean, after all it is meant to be able to deal with nodes being unavailable or slow sometimes. So assuming some university has … the local disks are not used a lot. Sounds like a good idea to me to use the systems as **Hadoop** cluster in their idle time. The simplest setup would be of course to have a cron job start the cluster …

hadoop parallel-processing cluster-computing preemption yarn

asked sep 25 '12 by Anony-Mousse

2 answers

Q: Hadoop Map Reduce wordcount Shuffle Error: Exceeded MAX_FAILED_UNIQUE_FETCHES; bailing-out

I have installed and configure **hadoop** as single node using manul from following site. http://www.michael-noll.com/tutorials/running-**hadoop**-on-ubuntu-linux-single-node-cluster/#running-a-mapreduce-job I have compiled wordcount example and run it but it takes long time and generate Any Clue? …

java hadoop

asked jul 3 '12 by Suvendu

A: Why would an AWS EMR job with c3.8xlarge servers have serious lag vs the same job with cc2.8…

problem. c3.8xlarge have fewer memory allocated for mapreduce tasks (the default configuration). Check here for verification if you use **Hadoop** 2.0 or here if you use **Hadoop** 1.0 In case you use … **Hadoop** 1.0 as you can see in the link provided, the number of mappers and reducers is much higher (by default) for c3.8xlarge instances. This means that less memory is allocated for each map an reduce …

answered nov 24 by Memos

1 answer

Q: Hadoop: intermediate merge failed

I'm running into a strange issue. When I run my **Hadoop** job over a large dataset (>1TB compressed text files), several of the reduce tasks fail, with stacktraces like these: Not all of my … an implementation of a secondary sort. We are using the CDH3 **Hadoop** distribution. Here is my custom implementation: The method is very simple, and I can't see any problems in this class …

hadoop mapreduce cloudera

asked apr 7 '11 by ajduff574

0 answers

Q: Connection Issue for Presto with Hive :Read timed out Exception

I am using presto 0.89 with cdh5 **Hadoop**. For Hive using mysql for metastore. Here is my Configuration: hive.properties Here is config.properties Error message: …

java hadoop hive cloudera-cdh prestodb

asked jan 9 by amitp

A: How to consume terabytes of data using a Java RESTful client

machines (so you don't run **out** of memory) and then apply an aggregating function across the data to get the result you're looking for before storing the final result in your results database. Off the top … of my head I'd recommend looking into Cassandra or HDFS for the distributed data grid (NoSQL cluster) then **Hadoop** for creating jobs to query/aggregate/manipulate that data. I hope that helps. …

answered dec 16 '11 by simonlord

1 answer

Q: Distributed local clustering coefficient algorithm (MapReduce/Hadoop)

). I tried to tune my **hadoop** platform and the code however the results were unsatisfactory (to say the least). No I have turned my attention to actually change/improve the algorithm. Below is my current … algorithm (pseudo code) Little bit more details about the code. For directed graphs neighbour data is restricted to node ID and **OUT** edges destination IDs (to decrease the data size), for undirected …

performance algorithm graph hadoop mapreduce

asked jun 10 '12 by alien01

0 answers

Q: Is data aggregated from all data nodes into the client during a HIVE query?

I was playing around with **Hadoop** and HIVE and I wanted the following queries to be addressed: I understand that **Hadoop** is a distributed File System. So, a file is spread across various datanodes … is fetching a million records, or 10 concurrent users are writing 10 queries fetching 10000 records, does the client node over loaded/**out** of memory? Awaiting your response. Regards, Jones …

hadoop hive hdfs

asked aug 14 by Jones

A: How to “break” HBase

Store keys with large values (Megabyes) in a loop so that you run **out** of diskspace Mismatch **Hadoop** version with the one shipped in Set allowed number of opened files on the OS to a very low number …

answered nov 23 '12 by Ravindranath Akila

A: Sorting using Map-Reduce - Possible approach

for **Hadoop** use) **Out** of the two possible approaches you mention, I think only one would work within the **Hadoop** infrastructure. Num 2, Since **Hadoop** leverages many nodes to do one job, sorting becomes … between time and space and in **Hadoop** we tend to think in terms of space rather than time unless you use products that are optimized for time (TeraData has the capability of putting Databases in memory …

answered jun 4 '13 by Engineiro

A: Weird error in Hadoop reducer

You're being caught **out** by a efficiency mechanism employed by **Hadoop** - Object reuse. Your calls to is returning the same object reference each time, all **Hadoop** is doing behind the scenes … is replaced the contents of that same object with the underlying bytes (deserialized using the method). To avoid this you'll need to create deep copies of the object returned from - **Hadoop** actually has …

answered dec 20 '13 by Chris White

0 answers

Q: Hadoop 2.2.0 on Windows: Job is Successful but just outputs the input file

I downloaded the source code for **Hadoop** 2.2.0 and all of the relevant dependencies and built it/installed it/configured it. Note that I built it using MinGW from Git rather than Cygwin … , if that matters. I am running a job but the only output that I get is the input file. Logging from my mapper or reducer is not working, either - in fact, I don't even know how to get **Hadoop** to tell me …

java windows hadoop mapreduce

asked feb 4 '14 by blspeiser

A: Spark fails on big shuffle jobs with java.io.IOException: Filesystem closed

failed). If it is a large shuffle, it might be an **out**-of-memory error which cause executor failure which then caused the **Hadoop** Filesystem to be closed in their shutdown hook. So, the RecordReaders … As of September 1st 2014, this is an "open improvement" in Spark. Please see https://issues.apache.org/jira/browse/SPARK-3052. As syrza pointed **out** in the given link, the shutdown hooks are likely …

answered sep 2 by Niketan

5 answers

Q: Comparing using Map Reduce(Cloudera Hadoop 0.20.2) two text files of size of almost 3GB

I'm trying to do the following in **hadoop** map/reduce(written in java, linux kernel OS) Text files 'rules-1' and 'rules-2' (total 3GB in size) contains some rules, each rule are separated by endline …) with the rules inside 'rules-1' and 'rules-2' Problem is, if I pull **out** each line of rules-1 and rules-2 files to a static arraylist only once, so that each mapper can share the same arraylint …

memory hadoop mapreduce compare overflow

asked apr 9 '11 by SSaikia_JtheRocker

4 answers

Q: What's the best way to count unique visitors with Hadoop?

hey all, just getting started on **hadoop** and curious what the best way in mapreduce would be to count unique visitors if your logfiles looked like this... and for each site you wanted to find **out** …

python hadoop mapreduce

asked may 21 '10 by James

A: How to run large Mahout fuzzy kmeans clustering without running out of memory?

mapper by default. When you subtract **out** all the JVM overhead, room for splits and combining, etc, you probably don't have a whole lot left. You set **Hadoop** params in a bootstrap action. Choose … Yes you're running **out** of memory. As far as I know, that "memory intensive workload" bootstrap action is long since deprecated, so may do nothing. See the note on that page. A should use 384MB per …

answered apr 20 '13 by Sean Owen

A: Kmeans with an huge array

may have to figure **out** a distributed solution, such as Mahout on **Hadoop**. Another option is that you may want to take sample **out** of all data somehow, and do clustering on the sample, if it is acceptable to

your requirement. …

answered aug 14 '13 by MesPost

A: Cloud shared memory management in openstack

been very difficult to get good performance **out** of a distributed shared memory system because of a problem called false sharing (I have no idea how much this affects ScaleMP, I've never used it). You … can also buy a more expensive system with specialized hardware for supporting distributed shared memory, like an SGI UV. You're probably best off modifying your code to take advantage of something like **Hadoop** or MPI. …

answered nov 11 '12 by Lorin Hochstein

A: Pig: Hadoop jobs Fail

Check your logs, increase the verbosity level if needed, but probably you're facing and **Out** of Mem error. Check this answer on how to change Pig logging. To change the memory in **Hadoop** change …

answered dec 17 by Paulo Fidalgo

0
answers

Q: Scala or Java analogues of PyTables & numexpr

over **out** of memory data. In particular the libraries should optimize the movement of data between disk, ram, cache & cpu. Sparse matrices should be handled efficiently. For more background on what … for building something like I need in Scala if what I am looking for does not already exist. **Hadoop**, MapReduce, graph databases & most of the popular nosql stores are not what I am looking for. Though what I am looking for is technically a kind of nosql store. …

java scala bigdata pytables numexpr

asked nov 15 '12 by Daniel Mahler

0
answers

Q: Hadoop - Single Node Standalone setup grep example crash

As a beginner trying **out hadoop**, I am referring to this article for setting up a single node stand alone **hadoop** instance. I believe my path variables have all been set up fine but when i try to run …

java hadoop

asked oct 31 '13 by Ajay Nair

3
answers

Q: Parallel processing options in Python

the problem down into steps and hopefully take advantage of the scalable parallel processing I'd get by using Amazon Web Services. But a friend of mine pointed **out** the fact that **Hadoop** is really … I recently created a python script that performed some natural language processing tasks and worked quite well in solving my problem. But it took 9 hours. I first investigated using **hadoop** to break …

python amazon-web-services parallel-processing

asked oct 6 '11 by Trindaz

A: Hadoop cluster. 2 Fast, 4 Medium, 8 slower machines?

In a nutshell, you want to max **out** the number of processor cores and disks. You can sacrifice reliability and quality, but don't get the cheapest hardware **out** there, as you will have too many … at, but maxing **out** cores and drives versus buying more boxes is generally a good choice - less power costs, easier to administer, and faster for some operations. More drives means more simultaneous disk …

answered jun 24 '09 by Colin Evans

2
answers

Q: How to export a large table (100M+ rows) to a text file?

fields, etc.) and store it int a big text file, for later processing with **Hadoop**. So far, I tried two things: Using Python, I browse the table by chunks (typically 10'000 records at a time) using … the full table with this. Using the command-line tool, I tried to output the result of my query in form to a text file directly. Because of the size, it ran **out** of memory and crashed. I am currently …

python mysql database hadoop export

asked jan 18 '13 by Wookai

0
answers

Q: How to make a matrix multiplication faster and managable?

) values. By multiplying them, I am trying to grab total commons for each row value. I applied MATLAB to this end but it was not possible to multiply them due to **out** of memory issue. So I partitioned … to apply **Hadoop** in java and do the process within java? Or is there any other suggestion? Thanks in advance. …

java matlab hadoop matrix-multiplication

asked dec 10 by user3198674