A: The memory consumption of hadoop's namenode?

. For a more detailed analysis of memory usage, check this link **out**: https://issues.apache.org/jira/browse/**HADOOP**-1687 You also might find this question interesting: **Hadoop** namenode memory usage … I suppose the memory consumption would depend on your HDFS setup, so depending on overall size of the HDFS and is relative to block size. From the **Hadoop** NameNode wiki: Use a good server …

answered nov 9 '12 by Pitt

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

A: Library/data structure for handling huge data

Check **out** STXXL - Standard Template Library for Extra Large Data Sets. "The core of STXXL is an implementation of the C++ standard template library STL for external memory (**out**-of-core … with existing applications, another design priority is high performance." Also, if you can dedicate several computers for the task, check **Hadoop**. Especially HBase, Hive and MapReduce. …

answered aug 17 '10 by Lior Kogan

A: pig join gets OutOfMemoryError in reducer when mapred.job.shuffle.input.buffer.percent=0.70

will spill part of the bag to disk. This allows a large job to make progress, albeit slowly, rather than crashing from "**out** of memory" errors. (This paragraph is from pig's blog) However, shuffle phase … is controlled by **Hadoop** itself, so that SpillableMemoryManager does not take effect in shuffle phase (exactly speaking, it can take effect in combine() which is used in Group By. But Join does not have …

answered aug 14 '13 by Lijie Xu

0
answers

Q: Reading 100gb xml file in hadoop

I'm coding a sort of word count in **hadoop** (for now on standalone pseudo-distributed virtual machine running on vmware player). I have a single huge 100GB xml file to read (close to 125 millions … suppose for the swap on hd as the memory run **out** ... am i right?) So the question is: which solution (configuration, data structure ...) i have left to solve the problem? …

xml file hadoop

asked jul 23 by Fujitina

A: How to use HBase and Hadoop to serve live traffic AND perform analytics? (Single cluster vs …

of memory and CPU) in high load situations. You mention, too, that heavy analytics can impact live performance, which is also true. In my cluster, we use **Hadoop** quite a bit to preprocess data for ingest … into HBase. We do things like enrichment, filtering **out** records we don't want, transforming, summarization, etc. If you are thinking you want to do something like this, I suggest sending your data …

answered jul 12 '11 by Donald Miner

3
answers

Q: Flume NG not writing to HDFS

I'm new at using Flume and **Hadoop** so I'm trying to setup the simplest (but somewhat helpful/realistic) example I can. I'm using the HortonWorks Sandbox in a VM client. After following one tutorial 12 …

hadoop hdfs flume

asked sep 3 '13 by BeanBagKing

A: Hadoop on EMR - Map Tasks Not Parallel

is that setting this value too high can cause the EC2 instances in your cluster to run **out** of memory. To set mapred.tasktracker.map.tasks.maximum, launch the cluster and specify the Configure **Hadoop** … Check if this helps you : The mapper daemons that **Hadoop** launches to process your requests to export and query data stored in DynamoDB are capped at a maximum read rate of 1 MiB per second to limit …

answered jun 10 by Mayank Agarwal

A: Hadoop Pipes: how to pass large data records to map/reduce tasks

Hadoop is not designed for records about 100MB in size. You will get OutOfMemoryError and uneven splits because some records are 1MB and some are 100MB. By Ahmdal's Law your parallelism will suffer … greatly, reducing throughput. I see two options. You can use **Hadoop** streaming to map your large files into your C++ executable as-is. Since this will send your data via stdin it will naturally …

answered oct 26 '10 by Spike Gronim

A: Running Hadoop: insufficient memory for the Java Runtime Environment to continue

This indicates you have run **out** of virtual memory, try increasing the swap space, or decreasing the heap to leave the rest of your program mroe virtual memory. a 32-bit program is limited to ~3 GB … On Windows a 32-bit program is limited to around 1.5 GB of virtual memory. As **hadoop** is a big data solution it is typically run on much bigger machines. e.g. 256 GB to 1 TB is not unusual. Given 32 GB is pretty cheap these days I would consider getting at least this much, or a lot more memory. …

answered jan 11 by Peter Lawrey

A: Hadoop namenode memory usage

) is costly as well, because it stores the collided elements in a linked list. The RPC Server needs threads to handle requests- **Hadoop** ships with his own RPC framework and you can configure … may **out** of memory (new Threads are also allocating a good chunk of stack space, maybe you need to increase this). So these are the reasons why your namenode needs so much memory. I haven't …

answered nov 2 '12 by Thomas Jungblut

A: Hadoop: Is it possible to have in memory structures in map function and aggregate them?

way. Pulling tricks like collecting data in memory in a is a minor and sometimes necessary sin, so, nothing really wrong with it. It does mean you really need to know the semantics that **Hadoop** guarantees, test well, and think about running **out** of memory if not careful. …

answered mar 1 '12 by Sean Owen

1 answer

Q: TestDFSIO fails with exitcode -1000

I setup a two node **hadoop** cluster. After having started the cluster it looks like this: machine : second machine : So, after having started the cluster I tried to perform a standard benchmark …

hadoop exit-code yarn

asked feb 18 '14 by toom

1 answer

Q: Memory-efficient distributed approach to determining unique values?

Approaches Single pass in-memory A single pass approach, storing column values in an associative array (a Java HashMap in my case). This will run **out** of memory at some point, but it's fast when … . The good news is that this approach is easy to implement in something like **hadoop**, utilising its distributed sorting step. MY QUESTION The multipass approach is painfully slow compared to a single-pass …

java sorting hadoop redis key-value

asked apr 11 by jkgeyti

A: Hadoop suitability for recursive data processing

The main strength of **Hadoop** is its ability to transparently distribute work on a large number of machines. In order to fully benefit from **Hadoop** your application has to be characterized, at least … . It seems that **out** of these 3, your application has only the last 2 characteristics (with the observation that you are trying to recursively use a scatter - gather procedure - which means a large number …

answered aug 6 '12 by Razvan

A: 1 GB Data with key and value, what kind of data structure to store them? 1TB? 1 PB?

IMHO, the choice of structure depends heavily on how much memory you have. But RAM is totally **out** of the question when you say 1TB or 1PB. When interviewers ask questions like … platform, like **Hadoop**, as Sreejith has said. In systems like **Hadoop** you use multiple systems together as a single system in order to leverage their combined power to gain better performance …

answered aug 20 '13 by Tariq

1 answer

Q: How to set heap size for EMR Master

I have a job which I am trigger from in EMR. The master triggers the mapper. Once it is done, it loads a heavweight operation in memory and then evenutualy will dump **out**. Right now, the job which … runs on the cluster fails after a few minutes because it runs **out** of heap space. By default it sets about 1000m on its master Tried the exact action below, but that did not work . The program is still …

elastic-map-reduce emr

asked aug 6 '13 by user2655578

1 answer

Q: Flexible heap space allocation to Hadoop MapReduce Mapper tasks

I'm having trouble figuring **out** the best way to configure my **Hadoop** cluster (CDH4), running MapReduce1. I'm in a situation where I need to run both mappers that require such a large amount of Java …

hadoop mapreduce cloudera

asked sep 11 '13 by Alex A.

1 answer

Q: Python MemoryError: The processed data set does not fit into 32-bit process address space

) **Hadoop** (Pydoop) (map/reduce) (The expression "Disk-based dictionaries" is used on the following webpage, but these libraries seem mostly **out** of date(?) except what is already mentioned above: https … I am trying to create a large nested dict in python, but the program runs **out** of memory (fails with MemoryError). (I am aware that 64-bit Python can use more memory that 32-bit, but am looking …

python sqlite pandas hdf5 pytables

asked dec 24 '13 by user843355

2 answers

Q: Duplicate Key Filtering

I am looking for a distributed solution to screen/filter a large volume of keys in real-time. My application generates over 100 billion records per day, and I need a way to filter duplicates **out** … **Hadoop**. Would HBase be the correct solution to use? Has anyone ever tried a partially in-memory solution like Zookeeper? …

hadoop duplicates distributed deduplication

asked nov 21 '13 by scottw

1

Q: Hadoop: Reduce-side join get stuck at map 100% reduce 100% and never finish

answer

I'm beginner with **Hadoop**, these days I'm trying to run reduce-side join example but it got stuck: Map 100% and Reduce 100% but never finishing. Progress,logs, code, sample data and configuration … this code in pseudo-mode. The join key from two files is equivalent. If I change the cite.txt file to 99 lines or lesser, It runs well while from 100 lines or above, it gets stuck like the logs shown. Please help me figure **out** the problem. I appreciate your explanation. Best regards, HaiLong …

hadoop mapreduce

asked oct 4 '12 by Hai Long

0
answers

Q: Hadoop MapReduce secondary sort: Reducer not getting called

I am trying to do a secondary sort on 4 values in my output. I referred to this tutorial. I have a 4 node cluster running **Hadoop** 2.2.0. I use Idea IntelliJ IDE for debugging locally. Following …

java sorting hadoop mapreduce

asked aug 26 by anixg33k

3
answers

Q: Pig “Max” command for pig-0.12.1 and pig-0.13.0 with Hadoop-2.4.0

I have a pig script I got from Hortonworks that works fine with pig-0.9.2.15 with **Hadoop**-1.0.3.16. But when I run it with pig-0.12.1(recompiled with -Dhadoopversion=23) or pig-0.13.0 on **Hadoop**-2.4.0 … , it won't work. It seems the following line is where the problem is. Here's the whole script. And here's the **hadoop** error info: 2014-07-29 18:03:02,957 [main] ERROR …

hadoop apache-pig

asked jul 30 by user2921752

1
answer

Q: What's the best way to do set-membership tests in hadoop?

I'm using **hadoop** to process a sequence of analytics records for my application. I want to categorise users based on which events I see in their stream and then use that information in a later stage … when iterating over the stream again. For example, suppose I want to generate data on all the users that never activate my app. I can work **out** who never activates by iterating over the stream once …

java hadoop amazon-web-services elastic-map-reduce

asked sep 16 '11 by Fasaxc

1
answer

Q: Hadoop Mapreduce Wrong result without errors - Column count check

I tried to built a **Hadoop** Mapreduce program in CentOS to check the columns on input file. File contains text only, not the XML and looks like this inside : Columns are divided by delimeter '|'. My …

java hadoop mapreduce hdfs

asked aug 19 by Pacharapol Huang

A: Java & Mysql processing massive data

The idea of **hadoop** is that it can help you to parallelize code execution. If you have only one machine I don't think that **hadoop** is suitable for you. Since you have 2 cores you may take advantage … array could be GCed, that means that you would not run **out** of memory. Queue size might be a subject for tuning based on maximum MySQL batch size, MySQL retrieval time and time spent on one batch …

answered nov 13 '13 by andrershov

A: Hadoop vs Spark

(check **out** DataBricks), but there's definitely more built on **Hadoop** than Spark. All the major **Hadoop** distributions are now on the Spark train. Spark is infinitely easier to configure and run than … There are lots of things to consider/discuss related to your question, and in some sense the answer is a matter of opinion. However I will say that Spark can be considered a replacement for **Hadoop** …

answered aug 12 by rs_atl

1
answer

Q: Pointing HiveServer2 to MiniMRCluster for Hive Testing

version of Hive and **Hadoop** (preferably, 2.0.0-cdh4.7.0) It needs to be all local. Meaning, the **Hadoop** cluster and Hive server should start on the beginning of the test, run a few queries, and teardown … for building an in-memory MapReduce cluster (I decided to use MiniMRCluster for this) Setting up both (1) and (2) above to work with each other. I was able to get (1) **out** of the way by looking at many …

jdbc hive integration-testing

asked oct 31 by Nishant Kelkar

5
answers

Q: How can I group a large dataset

col values from 1 - 10 million are considered in next run 10 million to 20 million are considered and so on. but this turned **out** to be really slow. The pig / **hadoop** solution is interesting because … the and extending them on . I am now planning to write a pig script which I am planning to run on a pseudo distributed **hadoop** machine (An Amazon EC3 High Memory Large instance). I wanted to know …

python data-structures hadoop apache-pig

asked aug 4 '10 by largescaled

A: Hadoop put performance - large file (20gb)

into packets. Last I heard **Hadoop** doesn't use DMA features **out** of the box, so these operations will be performed by your CPU rather than the NIC. Components: Memory, CPU Transmit packets to **hadoop** file …

answered oct 24 '13 by Axel Magnuson

A: How to put the files into memory using Hadoop Distributed cache?

great question. I am also trying to solve the similar issue. I don’t think **Hadoop** supports in memory cache **out** of the box. However it should not be very difficult to have another in memory cache …

answered may 10 by Saket

2 answers

Q: Are these Hadoop setup/cleanup/run times reasonable?

I've set up and am testing **out** a pseudo-distributed **Hadoop** cluster (with namenode, job tracker, and task tracker/data node all on the same machine). The box I'm running on has about 4 gigs memory, 2 … file, 60 seconds for a 100 mb file, and about 2 minutes for a 1 gig file. I also created my own Map Reduce program which cuts **out** all the logic entirely; the map and reduce functions are empty …

hadoop

asked jan 20 '11 by knt

A: MapReduce or Spark for Batch processing on Hadoop?

I'm assuming when you say **Hadoop** you mean HDFS. There are number of benefits of using Spark over **Hadoop** MR. Performance: Spark is at least as fast as **Hadoop** MR. For iterative algorithms … of the expressiveness of these functional constructs. Spark's API supports Java, Scala and Python (for most APIs). There is experimental support for R. Multiple Datastore Support: Spark supports many data stores **out** …

answered nov 2 by Soumya Simanta

1 answer

Q: How smart is the Java JVM about GC'ing during a reduce() operation on a long Scala list or S…

OK, let me see if I can explain. I have some code that wraps a Java iterator (from **Hadoop**, as it happens) in a Scala Stream, so that it potentially can be read more than once, by client code that I … the iterator will be extremely large, so that storing all the items in it will lead to **out**-of-memory errors. However, in general, the situations where the client code needs the multiple-iteration …

java scala stream garbage-collection jvm

asked sep 24 '12 by Urban Vagabond

A: Hardware requirements for Facebook Presto

Most people are running Presto on the **Hadoop** nodes they already have. At Facebook we typically run Presto on a few nodes within the **Hadoop** cluster to spread **out** the network load. Generally, I'd go …

answered nov 8 '13 by Dain Sundstrom

A: Hadoop Buffering vs Streaming

Hadoop Streaming in general refers to using custom made python or shell scripts to perform your map-reduce logic. (For example, using the Hive TRANSFORM keyword.) **Hadoop** buffering, in this context … , refers to the phase in a map-reduce job of a Hive query with a join, when records are read into the reducers, after having been sorted and grouped coming **out** of the mappers. The author is explaining …

answered jun 17 by Jerome Banks

1 answer

Q: Detailed dataflow in hadoop's mapreduce?

I am struggling a bit to understand the dataflow in mapreduce. Recently a very demanding job crashed when my disks ran **out** of memory in the reduce phase. I find it difficult to estimate how much disk … with permutation groups of words. Since identical words need to be joined the reduce function requires a temporary hash map which is always <= 3GB. Since I have 12GB of RAM and my **hadoop** daemons require 1GB …

java memory hadoop mapreduce

asked oct 21 '13 by DDW

A: WordNetSimalarity in large dataset of synsets

be buggy (better the more recent your perl) Database connections are not shared across threads In general, to get good performance **out** of Perl threads it's best to start a pool of threads and reuse … of your search/compareTo space to less than n^2, then map reduce or **Hadoop** may be a good option; otherwise, you'll just have a bunch of overhead and no use of the real scalability that **Hadoop** offers (@Thomas Jungblut). …

answered may 19 '13 by Steve P.

0 answers

Q: Configuring Hadoop Yarn CDH5 -> Stuck in runing an example job

I am configuring an **hadoop** cluster CDH5 on 3 nodes. HDFS Works. I configured YARN using Cloudera and advices from other websites I try to use the example calcultating Pi with this command on my … logs i get in mapred-mapred-historyserver-datanode1.**out** (I interupted the first Job and tried to run it a second time that's why time isn't matching) Every 3 minutes it does this... (and can keep …

hadoop configure jobs yarn

asked nov 18 by Miraculous

A: MapReduce and SQL GROUP BY

A lot of folk use MongoDB as the data storage and **Hadoop** for processing as there's connector between the two. Each MongoDB node can handle multiple **Hadoop** nodes reading into it. As a note, I'd … recommend is separating mongo and **Hadoop** nodes for memory. In case you don't have them, here's some documents for you Quick Start Scenarios (including Map-Reduce) The Interconnector Streaming One …

answered jul 6 '12 by Mark Hillick

A: 'Big dictionary' implementation in Java

it does useful for your task **out** of the box: Persistance to disk via memory mapped files (see comment by Michał Kosmulski) Lazy load (disk pages are loaded only on demand) -> fast startup If your … does the processing within a map-reduce-like framework, e. g. **Hadoop**. Strings are stored in UTF-8 form, -> ~50% memory savings if strings are mostly ASCII (as maaartinus noted) or values takes just …

answered sep 30 by leventov

A: Flexible heap space allocation to Hadoop MapReduce Mapper tasks

(PS you should use the newer name of this property with **Hadoop** 2 / CDH4: . But both should still be recognized.) The value you configure in your cluster is merely a default. It can be overridden … . Then **Hadoop** can decide how many mappers to run, and decide where to put the workers for you, and use as much of the cluster as possible per your configuration. No fussing with your own imaginary resource pool …

answered sep 14 '13 by Sean Owen

1 answer

Q: Hadoop - Reducer is waiting for Mapper inputs?

as explained in the title, when i execute my **Hadoop** Program (and debug it in local mode) the following happens: 1. All 10 csv-lines in my test data are handled correctly in the Mapper … > The bold marked lines repeat endlessly from this point. 4. Alot of open processes are active after the mapper saw every tuple: Is there any reason, why **Hadoop** expects more output from …

hadoop local reduce

asked may 23 '12 by Elmar Macek

A: Big Data Process and Analysis in R

apply to any **Hadoop**/MapReduce work. you can also get a **Hadoop** cluster via AWS/EC2. Check **out** Elastic MapReduce for an on-demand cluster, or use Whirr if you need more control over your **Hadoop** deployment. … and tweet text.) On the other hand, if your analysis is amenable to segmenting the data -- for example, you want to first group the tweets by author, date/time, etc -- you could consider using **Hadoop** …

answered dec 2 '11 by qethanm

A: Hadoop installation on Amazon cloud

or better yet use their elastic-mapreduce api which is a modified version of **hadoop**. You can run a 3 node cluster for around 00.25 cents an hour. If you really want to learn big data this is the way I went. You should check **out** their documentation here http://aws.amazon.com/documentation/elasticmapreduce/ … I have tried this personally and you will not really be able to use **hadoop** on a single micro instance due to memory restrictions. IMHO you should atleast try a medium instance to run **hadoop** …

answered jun 16 by Chris Hinshaw

1 answer

Q: Number of concurrently running mappers per node drops precipitously on Elastic MapReduce w/ …

In a related question (How to set the precise max number of concurrently running tasks per node in **Hadoop** 2.4.0 on Elastic MapReduce), I ask for formulas relating the number of concurrently running … mappers/reducers to YARN and MR2 memory parameters. It turns **out** that on Elastic MapReduce, when my cluster has between 2 and 10 c3.2xlarge nodes, variations of the formulas mentioned there work okay …

hadoop amazon-web-services amazon-ec2 elastic-map-reduce yarn

asked aug 10 by verve

4 answers

Q: Hadoop reduce become slower when there are less reduce task

I am experiencing a really weird case when I am doing some performance tuning of **hadoop**. I was running a job with large intermediate output (like InvertedIndex or WordCount without combiner … and shuffle, but it is not the case. It turns **out** that the job with 5 WAVES of reduce task is about 10% faster than the one with only one WAVE of task. And I checked the log and it turns **out** …

configuration map hadoop shuffle reduce

asked may 1 '12 by cyw

1 answer

Q: Potential tradeoffs of this use case between Cassandra and Couchbase

for real-time aggregated reports. We also plan to utilize **hadoop** directly on CassandraFS (as a replacement of HDFS - offered by datastax) to natively run Map Reduce jobs on the data residing in Cassandra … for more involved analytics. The output of the MapR jobs would be written back onto ColumnFamilies in Cassandra natively. **Hadoop** map reduce runs on a read-only replica of the main cassandra cluster …

hadoop cassandra couchbase

asked aug 1 '13 by NG Algo

A: hadoop/HDFS: Is it possible to write from several processes to the same file?

Are you able to explain what you plan to do with this file after you have created it. If you need to get it **out** of HDFS to then use it then you can let **Hadoop** M/R create separate files and then use … the downloaded parts in memory before writing them off to the output file in the correct location. You unfortunately don't have the ability to perform random output with **Hadoop** HDFS. …

answered aug 9 '12 by Chris White

A: General Method for Determining Hadoop Conf Settings on a Single Node Cluster

Time has passed and no one has tried to formulate an answer. So I will put forth some ideas in the hope that others will point **out** flaws if they exist. The most important thing in configuring … **Hadoop** is to not allow too many resources to be consumed; jobs will fail and the exceptions are not always helpful in quickly determining what went wrong. Particularly the memory resource will cause …

answered nov 18 '11 by SetJmp