1 answer

Q: Storm-YARN : Application container fails to launch

below issues. In Distributed **Hadoop** mode I'm getting the below error [1] while launching the YARN application. In **Hadoop** (local mode, with 1 box only) Yarn is spawnning the nimbus server and storm-ui … you help me **out** in understanding the reason of this container failure? There are no errors/info present in application logs. [1] YARN container fails to launch with below error on running. …

hadoop storm yarn

asked nov 6 by mukh007

A: Apache Spark vs. Apache Storm

Spark a unique form of fault tolerance based on lineage information. If you are interested in, for example, executing a **Hadoop** MapReduce job much faster, Spark is a great option (although memory … makes tradeoffs that are worth knowing. I would suggest checking **out** these links. Edit: discovered this today: http://xinhstechblog.blogspot.com/2014/06/storm-vs-spark-streaming-side-by-side.html …

answered jun 9 by plambre

3 answers

Q: 5.6 GB not enough for Cloudera?

I am running Cloudera **Hadoop** on my laptop and Oracle VirtualBox VM. I have given 5.6 GB **out** of mine 8 and six from eight cores as well. And still I am not able to keep it up and running. Even without … and the Hive went down again. What could be reason for more-or-less all **Hadoop** services going down if the VM starts to swap? I don't have enough reputation to post the picture to here, but when Hive went …

hadoop out-of-memory development-environment cloudera

asked jul 18 by Timo Riikonen

3 answers

Q: Reaching a saturation point when loading data; how can I control Java GC generations?

to the point where I find **out** if I have enough memory to load all of them, Java slows down asymptotically, and jvisualvm tells me that this is because nearly all of the CPU time is spent in garbage … that, however, the data-loading grinds to a halt because the garbage collector is apparently checking the same objects over and over (left plot). It must be expecting them to go **out** of scope, but I'm …

java garbage-collection permgen

asked sep 25 '13 by Jim Pivarski

2 answers

Q: Hive not enforcing bucketing

I am going through the Hive tutorial in the O'Reilly **Hadoop** book by Tom White. I am trying to make a bucketed table, but I can't get Hive to create the buckets. I can create the table and load … the data into it, but all of the data is then stored in one file. I am running a pseudo-distributed **Hadoop** cluster. I'm using **Hadoop** 1.2.1 and Hive 0.10.0 with a MySql metastore. The data (shown below …

hadoop hive

asked oct 4 '13 by Katrina

4 answers

Q: algebraic error when running “aggregate” function on dataset

I'm learning **hadoop**/pig/hive through running through tutorials on hortonworks.com I have indeed tried to find a link to the tutorial, but unfortunately it only ships with the ISA image …

java hadoop hdfs apache-pig piglatin

asked jun 14 '13 by Yuck

1 answer

Q: Container is running beyond physical memory for larger files

I have a small **hadoop** (2.5.1) cluster where I have the following configuration (concerning memory limits) mapred-site.xml: yarn-site.xml: And a map streaming task with python (without … a reducer) where I just read lines from a file and select specific fields to print **out** (I keep one of the fields as a key and the rest one big string). Each line holds quite a big of an array so the default …

python hadoop memory-management mapreduce hadoop-streaming

asked nov 27 by user1676389

A: graph database for Erlang with good query/traversal capability?

is a distributed graph database, which I believe is based on **Hadoop**. Titan is designed to scale-**out** and can offer an interesting approach for massive paralelism, with some overhead. There are use cases where this is more appropriate, such as similar to Google Pregel use cases. … and is based on RB-Tree algorithms that are very well performing for various graph use-cases. What most of the graph databases have in common is that they scale up only. They can scale-**out**, although …

answered jan 10 '14 by gextra

A: Block Replication Limits in HDFS

. The values in indicate their defaults. Some description of this is available at https://hadoop.apache.org/docs/current/**hadoop**-project-dist/**hadoop**-hdfs/hdfs-default.xml You can perhaps try to increase … to max **out** the re-replication workload such that it affects regular cluster work, as recovery of 1/3 replicas may be of lesser priority than missing job/query SLAs due to lack of network resources …

answered feb 15 '14 by Harsh J

A: Compare in-memory cluster computing systems

by using client-side manual sharding, they cannot be scaled **out** in a cluster (a Redis cluster implementation is on-going though). Spark is a system to expedite large scale analytics jobs … data management. But Redis (and memcached) play in the same ballpark as the other OLTP NoSQL stores, while Spark is rather similar to an **Hadoop** map/reduce system. Redis is good at running numerous …

answered may 22 '13 by Didier Spezia

A: Is this a viable MapReduce use case or even possible to execute?

You certainly need a scalable, parallel-izable solution and **hadoop** can be that. You just have to massage your solution a bit so it would fit into the **hadoop** world. First, You'll need to make … across multiple servers. These objects (call it class ModelApply) would hold the results of a specific model-to-widget application and can be processed in the usual way with **hadoop** to repost on best …

answered oct 2 '12 by Chris Gerken

A: Store data locally long term

It is hard to answer you question. Because i don't know about ur data volume and type. But i can tell you now. If u are thinking about file for that, it may have scale **out** issue, if u scale **out** … python server to # of box. So u may need a shared storage. In that case u have to think about shared storage file system like glusterFS or **Hadoop**. (glusterFS is more eaisier). But the access time …

answered jan 20 '14 by Terry Cho

2 answers

Q: What is the best way to mine text on large files (1 GB+) in Python?

I have a handful of text files, ranging from 1 to 5 GBs. Content are simple unique one-liners. I would like to: Problem: Memory runs **out**. IDE can't cope. Even when using generators … . Question: What is the best approach to work with such large files? Batching? Map/reduce? **Hadoop**? Using database instead of Python? What I don't want is to write a function to find …

python-3.x large-files

asked jan 21 by user1552294

A: How do I output whole files from a map job?

SequenceFile ruin your application as you will end up with corrupted output PDF files. Try this **out** and see for yourself. You have lots of input files...and this is where **hadoop** sucks. (read … fop. I dont think this is a **hadoop** use case in first place...becoz whatever you want to do can be done by a batch script. How big are your xml files ? How many xml files do you have to process ? EDIT …

answered apr 11 '12 by Tejas Patil

A: what are the disadvantages of mapreduce?

swapping, that's why I made the transition. The Namenode keeps track of the metadata of all files in your distributed file system. I am reading a **hadoop** book (**Hadoop** in action) and it mentioned that Yahoo … . This list is definitely not complete, just a few first remarks. Obviously you have to keep in mind that it is geared towards Big Data, and that's where it will perform at its best. There are plenty of other distribution frameworks **out** there with their own characteristics. …

answered sep 3 '13 by DDW

A: how can i handle billion records effectivily

will be the quickest to setup and test. However, if you anticipate map-reduce type analytic needs, then HBase is more tightly integrated with the **Hadoop** ecosystem, and should work **out** well. Performance-wise, they are both neck to neck, so take your pick. …

answered sep 27 '13 by Nikhil

A: getting close to real-time with hadoop

You need to provide a lot more information about the goals and challenges of your system to get good advice. Perhaps **Hadoop** is not what you need, and you just require some distributed systems foo … ://hadoop.apache.org/hbase/ It could be that you just need some help with managing replication and sharding of data. Check **out** Gizzard, a middleware to do just that: http://github.com/twitter/gizzard Processing …

answered may 24 '10 by SquareCog

A: BigData analysis choose technology stack

+Pig would be good. I'm not saying Cassandra is bad, but Hbase was developed ground up to be used with **Hadoop**. 4- There are lot of 'cool' things **out** there, but you are better off keeping the number low … Cloudera Impala on the same Hive tables(when is important). Impala uses same Hive metadata. So you don't have to worry about that. 3- If you are planning to work on **Hadoop** platform then HDFS+Hive+HBase …

answered jun 25 '13 by Tariq

A: deciding between subprocess, multiprocessing and thread in Python?

descriptors, or buffering space becoming available on some other descriptors or sockets (writable), or some exceptional conditions (TCP **out**-of-band PUSH'd packets, for example), or a TIMEOUT. Thus … as evocative of the programming model ... since your approach to the problem must be, in some sense, "twisted" inside **out**. Rather than conceiving of your program as a series of operations on input data …

answered apr 16 '13 by Jim Dennis

2 answers

Q: Does Cassandra 0.7.2 have performance issues with get_range_slices?

to the record's last update date so that I only get the latest record for each key. Later phases need to read everything back **out** of Cassandra, perform some processing on the records, and add the records back … to a different column family using various other keys, so that the records can be grouped. I accomplished this batch reading by using Cassandra.Client.describe_ring() to figure **out** which machine …

performance cassandra batch-processing

asked mar 21 '11 by Adam Brown

A: High Level Java Optimization

The word count problem is one of the most widely covered problems in the Big Data world; it's kind of the Hello World of frameworks like **Hadoop**. You can find ample information throughout the web … . Now, as you get even more data, you want to bring in map-reduce frameworks like **hadoop** to do the word counting on clusters of machines. Now, I've heard when you get into obscenely large datasets …

answered aug 13 '11 by Ray Toal

A: Computing user similarity using mahout mapreduce

this time, too. 100k users and 5 attributes is not really "**hadoop** size big data" yet. You may end up paying more for the **Hadoop** overhead than you get **out** as opposed to a fast low-level implementations … during or before generation). Secondly, note that Mahout builds on the **Hadoop** platform, but doesn't solve everything with just MapReduce. A lot of the **Hadoop** things do not do just "map+reduce …

answered sep 7 '12 by Anony-Mousse

0 answers

Q: hive job stuck at map=100%, reduce 0%

I'm running hive-0.12.0 on **hadoop**-2.2.0. After submitting the query: I get the following errors in the logs: And then the last line repeats every second or so ad infinitum. If I look … at container logs I see: I've searched for the Exit code 143, but most the stuff **out** there refers to memory issue and I have memory set pretty large (following the advice of Container is running beyond …

hadoop hive

asked jul 7 by harschware

2 answers

Q: ERROR org.apache.sqoop.tool.ExportTool - Error during export: Export job failed

a deprecated API. Note: Recompile with -Xlint:deprecation for details. 9516 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-**hadoop** library for your platform... using builtin … ! BUILD SUCCESS Total time: 30.419s Finished at: Fri Aug 23 15:28:03 IST 2013 Final Memory: 14M/113M After, we checked the mysql table contain only 100 records **out** of 1600 records. Same program, we …

mysql hadoop sqoop

asked aug 23 '13 by Manish V

1 answer

Q: IO Exception while running K-means clustering using mahout and hadoop jar [closed]

I am trying to run a clustering program using Mahout.Following is my java code which I am using But when i run it ,it starts executing normally but at the end gives me an error.. Following is stack …

java hadoop cluster-analysis mahout k-means

asked may 30 '13 by The good the bad the ugly

A: NLTK in production environment?

: adapt your algorithms to MapReduce and run them on MongoDB/**Hadoop**/Google MapReduce/... There are different places to host such solutions (Amazon, Google, Rackspace, ...) The second, "roll your own … for a scaled infrastructure: rent processors, memory, disk space, ... per hour, just enough to do the load test and bail **out**. That way, you don't need to buy equipment. more risky: do a load test …

answered may 12 '13 by pvoosten

A: Hadoop's Map-side join implements Hash join?

key, and then project (flatten) back **out** on the pairs. Because of this, when generating a frequently-read dataset it's often a good idea to do a total sort in the last pass. Zebra and other databases may also give you total-sorted input for (almost) free. …

answered jun 3 '10 by mrflip

A: What is faster- Java or C# (or good old C)?

think you should consider two options: MapReduce Your problem sounds like a good match for something like **Hadoop**, which is designed for very data-intensive jobs. **Hadoop** has scaled to 10,000 nodes … , you could consider adapting your app to use MPI. This would also allow you to scale **out** to (potentially thousands) of cores. MPI is the de-facto standard for computationally intensive, distributed …

answered apr 8 '09 by tgamblin

A: Database scalability concerns

. Maybe you can summarize the data with some sort of a process and store it in much smaller tables... in which case good indices should keep you **out** of trouble. Maybe you need to fall back to something … **Hadoop**-based, like Storm or Impala or Spark. Elastic search might come in handy too, depending, or Redis/memcache. It really all depends on (a) what data you will be storing (b) what queries you …

answered jun 9 by evanv

A: Hadoop data flow efficiency for “customers who bought x also bought y”

could even use an LRU Map to keep the most frequently observed pairs in the map, and write **out** those 'expired' entries when they are forced **out**. For an example adapted for the Word Count example … , see http://www.wikidoop.com/wiki/**Hadoop**/MapReduce/Mapper#Map_Aggregation Of course, if you have a huge product set, or random product pairings, this isn't going to save you that much. You also need …

answered mar 21 '12 by Chris White

A: Java VS RAM: How much data Java can handle without 'OutOfMemory'?

should probably rethink your design. Moreover, scaling **out** a 'big data' system becomes cost effective only if you can use convenience hardware (cheap low end servers) where the available resources … above, most Java based (and not) big data technologies most likely never need to load huge amounts of data in memory. For example **Hadoop** processes individual rows of data, dumping the result back …

answered feb 19 '14 by Giovanni Botta

1 answer

Q: Could not load org.apache.hadoop.util.ShutdownHookManager when shutdown tomcat server

could not load **hadoop** shutdownHookManager exception. I am sure the **hadoop**-common (contains ShutDownManager) is in tomcat classpath. Can anyone help me **out**? Exception I got: Oct 14, 2013 5:57:54 PM … " java.lang.NoClassDefFoundError: org/apache/**hadoop**/util/ShutdownHookManager$2 at org.apache.hadoop.util.ShutdownHookManager.getShutdownHooksInOrder(ShutdownHookManager.java: …

java tomcat hadoop shutdown-hook

asked oct 15 '13 by Terminal User

A: MongoDB vs. Cassandra vs. MySQL for real-time advertising platform

for usage patterns. I don't feel that the information **out** there is that dated because the concepts at play are very fundamental. There may be new NoSQL databases and fixes to existing ones, but your … data **out** is pretty quick too, and you have a lot of flexibility with data format changes. The tradeoff is that you can't use SQL (a benefit for some) so getting reports **out** may be trickier …

answered may 28 '11 by Brian Lyttle

3 answers

Q: Limit CPU / Stack for Java method call?

I am using an NLP library (Stanford NER) that throws OOM errors for rare input documents. I plan to eventually isolate these documents and figure **out** what about them causes the errors … , but this is hard to do (I'm running in **Hadoop**, so I just know the error occurs 17% through split 379/500 or something like that). As an interim solution, I'd like to be able to apply a CPU and memory limit …

java nlp stanford-nlp

asked jul 4 '09 by Kevin Peterson

1 answer

Q: What Database for extensive logfile analysis?

. Which kind of database would fit this task? Is it worth to setup a single machine instance with **hadoop**+hbase... or is this all a bit over-sized? What database would you choose to do high … -performance logfile analysis? EDIT: Maybe **out** of my question it is not clear that we cannot spend money for cloud services or new hardware. The Question is if there are benefits in using noSQL approaches …

sql database nosql analysis logfile

asked jan 9 '13 by six86

A: Machine Learning & Big Data

. One of the most well known framework is the MapReduce abstraction, available through Apache **Hadoop**. **Hadoop** can be run on 10 thousands nodes cluster, probably much more than you will ever need. If you … do not own the hardware, you can "rent" the use of a **Hadoop** cluster, for example through Amazon MapReduce. Unfortunately, the MapReduce abstraction is not suited to all Machine Learning computations …

answered may 3 '13 by Matthieu Durut

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

in memory on a single machine, then why are you using **Hadoop**? Honestly, relying on a dense ordered primary key is one of the first things that gets thrown **out** in a NoSQL DB as it assumes a single …

answered may 9 by b4hand

0 answers

Q: Map task fails. Status: FAILED. Error: INSTANCE. Exception thrown-java.io.IOException: Job …

I have written a code to expand shortened url in mapreduce. But the map task fails. It throws exception java.io.IOException. I am running this in pseudo distributed mode in cloudera **hadoop** vm. Below is the error message and my code. Please help me. My code: …

java hadoop mapreduce mapper

asked jul 29 by ananya choudhury

A: kmeans with big data

-means is one of those algorithms that's "embarassingly paralellizable." If you rent **out** server space, you could run this on a **hadoop** cluster and that should help a lot. What are you trying to accomplish … the number of clusters, that will help a lot. Are you sure you need all 512 dimensions? If you can trim **out** or combine some of those dimensions that could also help. Have you tried running PCA …

answered aug 4 '13 by David Marx

A: machine-learning, artificial-intelligence and computational-linguistics

learning, then you're on the wrong path... try a simpler problem. Which database(s)? Update: Depends on the size of your corpus: if it's ginormous, then I would go with **hadoop** (since you … is really big, then I would definitely use a robust combination like mahout/**hadoop**. Both of them are built exactly for that purpose and you would have a really hard time "duplicating" all …

answered apr 23 '11 by Lirik

A: Get a value from RichPipe

on top of **Hadoop**) and then read the result from disk in order to use it in your application. So the code

will be something like: Then you'll need some higher level code to run the job and read … the data back into memory to get at the result. Rather than write **out** all the code to do this (it's pretty straightforward), why don't you give some more context about what your use case is and what you …

answered jul 17 by samthebest

A: data block size in HDFS, why 64MB?

the metadata in memory, which in turn brings other advantages that we will discuss in Section 2.6.1. Finally, I should point **out** that the current default size in Apache **Hadoop** is is 128 MB. …

answered oct 21 '13 by cabad

A: Which NoSQL database should I use for logging?

. These daemons are supposed to be launched at every application nodes, and takes the logs from app processes. They buffer the logs and asynchronously writes **out** the data to other systems like MongoDB … MongoDB's problem is it starts slowing down when the data volume exceeds the memory size. At that point, you can switch to other solutions like Apache **Hadoop** or Cassandra. If you have a distributed logging …

answered nov 17 '12 by Kazuki Ohta

A: What is a better approach of storing and querying a big dataset of meteorological data

be important). There is a package (SciHadoop) that combines **Hadoop** and HDF5. HDF5 makes it relatively easy to do **out** core computation (i.e. if the data is too big to fit into memory). PyTables …

answered jun 4 '13 by Ümit

1 answer

Q: Where should computations take place for complex algorithms

but they rely on **Hadoop** for scaling. Is there another way **out**? is there a Mahout or Machine Learning equivalent for ruby and if so how where does the computation take place? … Background: I'm a software engineering student and I was checking **out** several algorithms for recommendation systems. One of these algorithms, a collaborative filtering has a lot of loops int …

ruby-on-rails hadoop scalability mahout

asked apr 16 '13 by fernandohur

A: Which NoSQL Database for Mostly Writing

? Or possibly a big data Map/Reduce system like **Hadoop** (I know **Hadoop** is written in Java) If C is key requirement, maybe you want to look at Tokyo/Kyoto Cabinet? EDIT: more details MongoDB does not support … will no longer fit into RAM and your performance will start to drop off dramatically. (this is well-documented under MongoDB) So it's going to be really important to figure **out** which queries you want to run. …

answered apr 5 '12 by Gates VP

A: Moving from file logging to database logging

. These daemons are supposed to be launched at every application nodes, and takes the logs from app processes. They buffer the logs and asynchronously writes **out** the data to other systems like MongoDB … problem is it starts slowing down when the data volume exceeds the memory size. At that point, you can switch to other solutions like Apache **Hadoop** or Cassandra. If you have a distributed logging …

answered nov 17 '12 by Kazuki Ohta

A: MongoDB Aggregation as slow as MapReduce?

I will place an answer basically summing up my comments. I cannot speak for other techs like **Hadoop** since I have not yet had the pleasure of finding time to use them but I can speak for MongoDB … database. There is no easy way to do this in real-time in line to your own application. Map reduce could be a way **out** if you didn't need to return the results immediately but since I am guessing you …

answered dec 27 '12 by Sammaye

A: Practical example for each type of database (real cases)

unhappy when I bought a product and they said later they were **out** of stock. I did not want a compensated transaction. I wanted my item! • to scale then NoSQL or SQL can work. Look for systems … that support scale-**out**, partitioning, live addition and removal of machines, load balancing, automatic sharding and rebalancing, and fault tolerance. • to always be able to write to a database because you …

answered aug 18 '13 by loops

A: Large Data: Storage and Query

. If all your queries are date-based, you might put the data for each month on different physical servers, for instance. I'd start with an RDBMS, create a test data set to work **out** if it meets your … scalability needs by running and tuning sample queries. Tune the hardware, and add more if you can afford to. I don't think you will get much benefit from **Hadoop** - you're not doing much processing, you're …

answered jan 18 '13 by Neville K