A: Kmeans on hadoop

huge pain, for the reason that each iteration involved disk IO, for both input and **out** put. Besides that, lauching an iteration (a MR job) cost tens of seconds in a **Hadoop** cluster. Later I tried **out** … Spark, which is a MR like framework that can work perfectly upon **Hadoop**. It uses the memory of all commodity computers in a cluster, to cache iteration invariants instead of reading and writing back them to the disk repeatedly. You may want to check it **out** :-) …

answered aug 14 '13 by MesPost

1
answer

Q: Hadoop bringing code to the data is this the only way to achieve scale?

. It means you have to figure **out** how to port your code and install it along with any dependencies onto the server itself. This is a lot of infrastructural cost! And it's also difficult to do inside … a (Platform as a service) Paas cloud. An example using **Hadoop** streaming for instance. You gotta make sure your binary is compiled against the target server kernel. eg. Linux vs. windows etc. You also gotta …

.net architecture hadoop

asked apr 9 '12 by Alwyn

1
answer

Q: HDFS Block related questions

I have 2 questions that will help me understand how HDFS works in the context of blocks. 1. You use the **hadoop** fs -put command to write a 300 MB file using and HDFS block size of 64 MB. Just after … this command has finished writing 200 MB of this file, what would another user see when trying to access this file? A. They would see **Hadoop** throw an ConcurrentFileAccessException when they try …

hadoop hdfs

asked jan 14 by theexplorer

0
answers

Q: Understanding of merging in Hadoop on reduce side

I have problem with understanding of files merging process on reduce side in **Hadoop** as it is described in "**Hadoop**: The Definitive Guide" (Tom White). Citing it: When all the map outputs have been … copied, the reduce task moves into the sort phase (which should properly be called the merge phase, as the sorting was carried **out** on the map side), which merges the map outputs, maintaining …

hadoop mapreduce

asked sep 24 by Anton Kulev

A: Algorithm for efficient diffing of huge files

not aware of any **out**-of-the box command line tools using the compression technique. The "fs-c" open source project contains most of the code that is necessary. However, fs-c itself tries only to measure the redundancies and the analzye files in-memory or using a **Hadoop** cluster. …

answered jan 8 '10 by dmeister

A: configure() not be called when running on hadoop cluster but can be called on Eclipse, Distr…

Problem solved, it turns **out** that I made two mistake: 1) I added a System.out.println() in the beginning of configure(), but it didn't show up it turns **out** that mapreduce can't use System.out.println … () in mapreduce phases, if we want to see the it, we need to check our log, for details thanks to Where does **hadoop** mapreduce framework send my System.out.print() statements ? (stdout) 2) my real …

answered apr 15 '13 by user2070763

A: sublist for partitioning data set

. A really common tool for this kind of work is **Hadoop**. You'd copy the data into HDFS, run a map-reduce job (or more than one job) on the data and then copy the data **out** of HDFS when you're done. A simpler …

answered jul 8 '12 by Harold L

A: Require Details about edge node in cluster

+1 with the Dell explanation. In my opinion, edge nodes in a **Hadoop** cluster are typically nodes that are responsible for running the client-side operations of a **Hadoop** cluster. Typically edge-nodes … are kept separate from the nodes that contain **Hadoop** services such as HDFS, MapReduce, etc, mainly to keep computing resources separate. For smaller clusters only having a few nodes, it's common …

answered dec 5 '13 by Anthony R.

A: Any one implementing or working on impala from cloudera?

a **Hadoop** cluster, and the most recent versions of "MapReduce" or programs that run on this "YARN" cluster. There's a lot to explain here and you're probably basically familiar with it, so I'll move … complicated jobs **out** of sequences of MapReduce jobs. Sqoop: A tool for importing / exporting data between HDFS and relational databases. It does this by compiling the job to perform into a MapReduce job …

answered jan 7 '13 by user1955635

A: Recomendations for PHP & MySQL Statistics Reporting for a WebApp

. Retrieving results might be a little trickier (time for lists). Redis is also largely memory-based (though dev builds include "virtual memory" support), so it's fast, but you could run **out** of room … solid, you could use **Hadoop** or another technology to pull from the end of the list (RPOP) and archive the results into a more permanent format (XML pushed to Amazon S3, for instance). You could …

answered oct 20 '10 by mattbasta

A: Working with Hadoop with two datasets

Hive for this. If Hive is ruled **out**, the steps mentioned in this link should help you get started for large data sets Configure Map Side join for multiple mappers in **Hadoop** Map/Reduce … Distributed Cache is good for small data set and may consider to cache in memory. If the data is large, distributed cache is not an option, as you mentioned. **Hadoop** framework allows to sort on large …

answered apr 25 '13 by Niranjan Sarvi

A: How to schedule hundreds of thousands of tasks?

need to run a loop, popping off tasks **out** of the AVL tree until the next task's timestamp is in the future. Then you could sleep for a few seconds and start checking again. Whenever a task runs, you … you have 4GB+ RAM and several hard drives in RAID 0+1 on your master server. Finally, if you really want to get complicated, **Hadoop** might work too: http://hadoop.apache.org/ …

answered mar 16 '10 by PsychoDad

A: Making a Twitter-like timeline with MongoDB

array of statuses. You will use more storage which with mongoDB mean more memory.... At Facebook they are using **Hadoop** with HBase for the main storage then they have huge arrays of servers with lots … solution if you follow 3users, you need to grab all their feeds then sort them, then render them... [Edit] Like Shekhar point **out** int the comment, Mongo has a document Limit. You need to create a status …

answered dec 9 '11 by zzarbi

A: How to cheaply process large amounts of data (local setup or cloud)?

then you probably want something like **Hadoop** or Storm to manage the compute nodes. I don't know how feasible it is to go through 4TB of data in a matter of seconds but again that really depends on the kind … is the time your project will be running. Or to put it another way, how long will you get some use **out** of your specialty hardware. If it's a project that is supposed to serve the EU parliament …

answered nov 13 '12 by Matti Lyra

A: Work on samples instead of the entire data base

and just process the web logs in **Hadoop** for analytics. The problem is that having an OLTP system for your 'Products' and another OLAP system for your analytical insight is problematic. Sure, you can set … , short time, you may get some mileage **out** of an appropriate index on . You may do some pre-aggregation and staging, with batched updates and some delay. But there is no silver bullet, easy answer. …

answered jan 10 '14 by Remus Rusanu

0
answers

Q: mahout seqdirectory mapreduce just get one map task

. note as mahout-833 I am tracing SequenceFileFromDirectory.java from Mahout 0.9 source code, try to figure **out** why. Here is my configuration info: job config file, **hadoop** config file(mapred-site.xml, **hadoop**-env.sh) … I want to have more map task to parallel Mahout seqdirectory job, but every time I try, it still produce just one map task, please help me!! **Hadoop** version: 1.2.1 Mahout version: 0.8/0.9 (tested …

hadoop configuration mapreduce mahout

asked jul 4 by user1412306

1
answer

Q: What are good/established database backing strategies for logic programming in clojure

to be found deal with facts created in memory. A really nice library **out** there is Cascalog. It sits on top of **hadoop** and abstracts away the map-reduce with a nice and intuitive query language inspired … by datalog. But Cascalog is **hadoop** only and I wonder if there are similar clojure solutions which go well with other types of databases, be it nosql or relational (deductive databases). …

database clojure logic-programming

asked apr 4 '12 by nansen

A: Hadoop block size and file size issue?

One block is consumed by **Hadoop**. That does not mean that storage capacity will be consumed in an equivalent manner. The output while browsing the HDFS from web looks like this: You see that each … of blocks available dependent on the capacity of the HDFS. You are wasting blocks as you will run **out** of them before utilizing all the actual storage capacity. Remember that Unix filsystem also has …

answered jul 6 '12 by pyfunc

A: handling large scale dataset

need to do analysis or transformations on the data. This is a deep topic with a lot of different approaches and challenges. Map/reduce frameworks like CouchDB and **Hadoop** have recently become popular … tools for research and application in this area. Data that's too big for a single database instance. This can be a problem of disk size (ran **out** of space) or performance (memory cache keeps getting …

answered jun 27 '10 by Owen S.

0
answers

Q: host key verification failed - Bash

: The catch I'm seeing is the fact that the directory has as owner. I need to format **hadoop** and then start the cluster with . When I execute those three commands from the vm's shell, it works … really don't know what's going on, and that's the only missing piece to finish my **hadoop** cluster! What can I do to fix it? I'm executing from the terminal, to a vm. …

linux bash shell hadoop amazon-web-services

asked may 9 by philippe

A: Processes exceeding thread stack size limit on RedHat Enterprise Linux 6?

Turns **out** that RHEL6 2.11 have changed the thread model such that each thread where possible gets allocated its own thread pool, so on a larger system you may see it grabbing up to the 64MB. On 64 … /linux_glibc_2_10_rhel_6_malloc_may_show_excessive_virtual_memory_usage?lang=en glibc bug malloc uses excessive memory for multi-threaded applications http://sourceware.org/bugzilla/show_bug.cgi?id=11261 Apache **hadoop** have fixed the problem by setting MALLOC_ARENA_MAX https://issues.apache.org/jira/browse/**HADOOP**-7154 …

answered nov 5 '12 by Rory

2 answers

Q: Best way to process line at a time data from hdfs file from within CPython (without using st…

I would like to use CPython in a **hadoop** streaming job that needs access to supplementary information from a line-oriented file kept in a **hadoop** file system. By "supplementary" I mean that this file … is in addition to the information delivered via stdin. The supplementary file is large enough that I can't just slurp it into memory and parse **out** the end-of-line characters. Is there a particularly elegant way (or library) to process this file one line at a time? Thanks, SetJmp …

python hadoop line hdfs

asked dec 19 '11 by SetJmp

1 answer

Q: Deserializing Flume Events in C# via Avro

, and there's a Microsoft **Hadoop** library that has deserialization libraries. I created a local object to deserialize to: and try deserializing with this: then I get this error: I'm almost … certainly going about all this in the wrong way, and I'm sure people are going to be telling me not to use C#, but I've pretty much run **out** of sources on Google, so if anyone else has actually done this and point me in the right direction, I'd be very grateful Toby …

c# flume avro

asked nov 13 '13 by TobyEvans

A: Big data File: Read and Create structured file

to write to a table where the primary key is auto incremented, with a field called "key" (or "left") and a field called "value" (or "right"). Then SELECT **out** of the table which was the MIN(key) and MAX(key …). However, IMHO this question is not correctly tagged as "big data": in order to require distributed computations on **Hadoop** or similar, your input data should be much more than what you can hold …

answered apr 9 by logc

9 answers

Q: Redis administration panel [closed]

to logs, trends on memory usage, etc. would be nice but not necessary. I'm running Redis on a **Hadoop** cluster, in which I enjoy having pages for the JobTracker, NameNode, Ganglia, etc. There are a few … **out** there, but at first glance they don't seem ready for prime time. http://www.servicestack.net/mythz_blog/?p=381 http://code.google.com/p/redis-admin/ …

redis

asked oct 10 '11 by Donald Miner

1 answer

Q: Scalable real time item based mahout recommender with precomputed item similarities using it…

I have the following setup: boolean data: (userid, itemid) **hadoop** based mahout itemSimilarityJob with following arguements: --similarityClassname Similarity_Loglikelihood … is stands for and how it works. If I set this to 10/20/50, will I be able to achieve what stated above. Also is there way to accomplish this via the api? I am using a rescorer for filtering **out** …

hadoop machine-learning mahout

asked aug 31 '12 by gk99

A: Oozie error while trying to execute “bin/mkdistro.sh -DskipTests”

Me too faced the same probelm. Try this installtion steps it is worked for me change the versions in below steps depends on what version you need.

answered jul 4 by MarHserus

A: How to use apache mahout results in jsp?

though. If the data is too large, you have to go the **hadoop**/db route. check **out** the in-memory item and user based recommenders here: https://github.com/apache/mahout/tree/trunk/mrlegacy/src/main/java …

answered may 15 by pferrel

A: good (noSQL?) database for physical measurements

is an in-memory database. It's extremely fast and efficient at getting data in and **out** of RAM. It does have the ability to use disk for storage, but it's not terribly good at it. It's great … only one node is necessary to perform operations within a row. **Hadoop** plugs right into Cassandra, with "native drivers" for MapReduce, Pig, and Hive, so it could potentially be used to aggregate …

answered nov 3 '11 by rbranson

A: Getting started with massive data

Hadoop is great, but can be a pain in the ass to set up. This is by far the best article I've read on **Hadoop** setup. I strongly recommend it: http://www.michael-noll.com/wiki … , which may or may not be helpful. I'm not a math guy but it seems most math calculations are very parallelizable, with little need of threads sharing memory. Either way, you might want to check **out** …

answered aug 10 '10 by Jieren

A: Datameer for Real Time Querying

into **Hadoop** is much easier. No static schema creation, no ETL, etc. Just use a wizard to download data from your database, log files, social media, etc. Designing analytics or making changes is a lot faster … If you have real time requirements you should not pull data directly **out** of Datameer, Hive, Impala, etc.. Columnar storages make some processing faster but will still not be low latency. But you can use …

answered jan 17 '14 by Peter Voss

A: Using java + php in windows

Here is your problem: I'm unaware of any sort of object called "Java" on PHP that you'll get **out** of the box by simply installing Java. I think you're trying to execute the Java Bridge integration … a compiled language. Java support structure is massive. There is tons of tooling and libraries for Java that just aren't available to other stacks. Things like Lucene, Spring, **Hadoop**, Tomcat, PDF …

answered jan 14 '12 by chubbsondubs

A: is it feasible to use hadoop over amazon emr to process > 10TB of input?

the rest of the cluster is fine)? If none of the above, there probably is a **Hadoop** resources that is getting maxed **out**, and needs to be configured differently. Since you didn't mention any particular … it wrong". Usually the problem is you are getting tied up in the sort/spill/merge process, with nodes maxing **out** disk IO in some fashion. **Hadoop** has a number of tuning parameters that start to get wacky …

answered nov 1 '12 by tphyahoo

A: OperationTimeoutException Cassandra cluster AWS / EMR

So, what happens if you set your timeouts to -1 instead? Personally, I would dig into the astyanax code and try to figure **out** how to disable the timeouts. Run your stuff again and it should keep … .maven.org/maven2/org.apache.cassandra/cassandra-all/1.2.2/org/apache/cassandra/**hadoop**/ColumnFamilyRecordReader.java#ColumnFamilyRecordReader.0 in which we can at least see this is getting a key …

answered apr 16 '13 by Dean Hiller

A: Database design or architecture suitable for storing logs, real time reporting and utilized …

for single-person usage. If it's not simply a matter of adding up numbers but analyzing complex relationships (graph like analysis) on large volumes, you're **out** of luck. Old solutions don't scale well … solution. Personally, I'd go with postgres (backend) + pentaho and qlikview (both front-end) with kettle for traditional ETL and **hadoop** or custom code to precalculate results for more complicated analysis. In postgres split up your data in an operational store and a DWH. …

answered aug 7 '13 by r.m

A: Hadoop Yarn Container Does Not Allocate Enough Space

in the **Hadoop** mailing list had the same problem and in their case, it turned **out** they had a memory leak in their code. …

answered dec 27 '13 by cabad

A: Streaming data and Hadoop? (not Hadoop Streaming)

. Worth checking it **out** M3 : Stream Processing on Main-Memory MapReduce. I could not find the source code or API of this M3 anywhere, if somebody found it please share the link here. Also, **Hadoop** Online … As you know the main issues with **Hadoop** for usage in stream mining are the fact that first, it uses HFDS which is a disk and disk operations bring latency that will result in missing data in stream …

answered sep 26 by Ambodi

0
answers

Q: get error when I use java api to connect Hbase(standalone mode without Hadoop)

Hbase version: 0.98.0 when I tried to use java api to connect to Hbase and do some 'CRUD' operation, I got the following error: I think it was the last one where I got stuck. But I don't know why. …

java eclipse api hadoop hbase

asked mar 28 by Rickie Lau

A: Using Amazon MapReduce/Hadoop for Image Processing

: starting with a set of files, performing the same action to each of the files, and then writing **out** the new file's output as it's own file. Guess why? :D **Hadoop** is not the right thing for this task … There are several problems with your task. **Hadoop** does not natively process images as you've seen. But you can export all the file names and paths as a textfile and call some Map function on it. So …

answered oct 19 '11 by Thomas Jungblut

2
answers

Q: Hbase daemon crashes at start

I am trying to run Hbase 0.96.1.1 for **Hadoop** 2 on a Mac book air. When I run ./start-hbase.sh, starting master, logging to..... but it crashes right after. It seems that iface is an network … interface on Linux system. Does that mean this version can not be run on Mac? Edited: I tested hbase version 0.98 also. Same issue. The only version that is working is hbase 0.94 but it is not compatible with **hadoop** 2. …

osx hbase

asked mar 29 by user1996536

A: Searching for regex patterns on a 30GB xml dataset. Making use of 16gb of memory

First, try to find **out** what's slowing you down. How much faster is the parser when you parse from memory? Does using a with a large size help? Is it easy to split up the XML file? In general … , shuffling through 30 GiB of any kind of data will take some time, since you have to load it from the hard drive first, so you are always limited by the speed of this. Can you distribute the load to several machines, maybe by using something like **Hadoop**? …

answered sep 21 '08 by Torsten Marek

0
answers

Q: How to process huge data sets using mapreduce

Could some one please help me to find an approach for below scenario. I am trying to explore **hadoop** and **hadoop** related tools. I want to develop a small mapreduce application where it should read … input feed file from hdfs and prepare some data, read data from hive and prepare some data and compare these two data sets to find **out** the data accuracy. What is the best way to do this? How can we do …

hadoop mapreduce hive hdfs bigdata

asked apr 2 by Ram

1
answer

Q: RHadoop Job failing on Single Node Ubuntu cluster

I am posting a similar question a second time because I believe I now have a far more precise view of the problem. Environment : **Hadoop** 2.2.0 running as a Single Node Cluster on an Ubuntu 14.04 … laptop machine. RStudio version 0.98.507, R version 3.0.2 (2013-09-25), Java Version 1.7.0_55 Any R (or Python) program works perfectly with the **Hadoop** Streaming utility located at /usr/local/hadoop220 …

r hadoop rhadoop

asked may 10 by Calcutta

A: Hadoop one Map and multiple Reduce

and multiple "different" reduce function makes no sense, it shows that you are just using map to pass **out** data to different machines to do different things. you dont require **hadoop** or any other special architecture for that. …

answered feb 28 '10 by none

2
answers

Q: Mahout on Elastic MapReduce: Java Heap Space

. Based on previous questions here and elsewhere, I've cranked up every memory knob I can find: conf/**hadoop**-env.sh: setting all the heap spaces there up to 1.5GB on small instances and even 4GB …

hadoop heap mahout emr

asked apr 29 '12 by David M.

A: Using ElasticSearch with Hadoop Map Reduce

You might be able to make this work. One possible way would be to have each **hadoop** node run an embedded routing only elastic search node. That should make querying a bit more efficient since … the node will figure **out** which nodes need to be contacted for each query and utilize the efficient internal protocol to do so. You can scale horizontally by adding more es data nodes. The only downside …

answered jul 18 '13 by Jilles van Gurp

1
answer

Q: Why the RDD is not persisted in memory for every iteration in spark?

I use the spark for machine learning application. The spark and **hadoop** share the same computer clusters with **out** any resource manger such as yarn. We can run **hadoop** job while running spark task … make the application so slowly. Then, my question is why the rdd was not persisted in the memory when there was enough free memory? because of the **hadoop** jobs? I add the following jvm parameters …

scala apache-spark

asked jul 24 by Tim

2
answers

Q: Hadoop program runs well with “java -jar” but not with “hadoop jar”

I have coded a MapReduce program that connects to two HBase databases. I have written it on Eclipse and I have exported it with the "Runnable Jar" option (with all the libraries). It runs well with th …

java eclipse hadoop jar hbase

asked sep 10 by brunneis

0
answers

Q: Permission denied error while running embedded pig in Java on Hadoop

Last week I used the user "root" to start the Hadoop's dfs & mapreduce and run the Embedded-Pig Java code. It was running fine. This week I'd like to perform the same task by using a non-root user: ch …

java permissions hadoop apache-pig

asked apr 17 '12 by C.c. Huang

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 spawning 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

---

prev 1 2 3 4 5 6 next

15 30 50 per page