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

A: CAP with distributed System

HDFS has a unique central decision point, the . As such it can only fall in the CP side, since taking down the namenode takes down the entire HDFS system (no Availability). **Hadoop** does not try … . Is there any database **out** there that allows user to switch its choice accordingly between CP and AP Systems that allows you to modify both the write and the read quorums can be tuned to be either CP or AP, depending on the needs. …

answered nov 12 '13 by Remus Rusanu

A: getting started with EC2 for compute-intensive (non-web) parallel application

: Elastic Map Reduce might be what you're looking for the parallelize your work easily, etc.. If that is too limited, you could look into Cloudera. A ready-to-rumble **hadoop** distribution with support … for EC2 as well. If map-reduce is not to your liking, then you need to setup your own instance. Roughly speaking, the keypoints are as follows: You want to figure **out** a way to start EC2 instances. You …

answered dec 1 '11 by Till

A: What difficulties should I expect if I write a NoSQL db using golang but want to run Hadoop …

Ok, I'm not much of a **Hadoop** user so I'll give you some more general lessons learned about the issues you'll face: Protocol. If you're going with REST Go will be fine, but expect to find some … that: a. the Thrift implementation for Go, last I checked, was lacking and relatively slow. b. Go has great support for RPC but it might not play well with other languages. So you might want to check **out** …

answered jan 30 '14 by Not_a_Golfer

2 answers

Q: Class Not Found Exception in KMeanClustering--Mahout

it on a single node **Hadoop** cluster(CHD-4.2.1), with mahout installed on it. The mahout examples run fine on this cluster, so no issues regarding installation. I use the following command in command Promt to run … ; adding HADOOP_CONF_DIR to classpath. Running on **hadoop**, using /usr/lib/**hadoop**/bin/**hadoop** and HADOOP_CONF_DIR=/etc/**hadoop**/conf MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.7-cdh4.3.0-job.jar 13 …

hadoop cluster-analysis classnotfoundexception mahout k-means

asked jun 6 '13 by user2454360

2 answers

Q: Why the map/reduce output a empty file, with no exception reported

I implemented a simple map/reduce program to make a inverted-sort index(a part of my small search engine). Every thing worked just fine. But after I enlarged the input file size to more than 20MB(stil …

hadoop mapreduce

asked oct 27 '12 by zhoutall

A: NoSQL & AdHoc Queries - Millions of Rows

into the various NoSQL solutions **out** there. The default unit of distribution is key. Therefore you need to remember that you need to be able to split your data between your node machines effectively otherwise … loss in a prod system is a big deal) . I am using **Hadoop**, HBase, Hive, MongoDB, Solr, MySQL and even good old text files. Remember that to productionize a system using these technogies is a bit harder …

answered jul 5 '11 by NightWolf

A: Mapper and Reducer for K means algorithm in Hadoop in Java

Map level Reduce level The Job Level The job level is fairly simple, it is writing the input (Key = the class called and Value = the class called), handling the iteration with the **Hadoop** job … to know what is coming in and what is going **out** (in terms of objects). So from the job level we know that we have and as input, whereas the is currently just a dummy. For sure we want to have the same …

answered may 18 '12 by Thomas Jungblut

A: How to print frequency k words from more than 1TB of data?

quicksort and heapsort code. As a practical matter this is just about the textbook problem for which "Big Data" (**Hadoop**, and other Map/Reduce systems) were built. If the data is distributed over N nodes … structure. In general I'd think a heap would be better (possibly including dropping things from the bottom of the in-memory heap **out** into a storage heap or trie). I said "adaptive" earlier because one …

answered apr 6 '13 by Jim Dennis

2 answers

Q: Why does spark-ec2 fail with ERROR: Could not find any existing cluster?

I have recently downloaded Spark and I am attempting to access my first cluster through Spark-ec2. I used the commands: And the startup appears to run without error. However when I run: it retu …

amazon-web-services amazon-ec2 apache-spark spark

asked jan 16 '14 by LFoos24

A: Please explain MapReduce simply

paper (Dean & Ghemawat, 2004 – link here) as a solution to make computations in Big Data using a parallel approach and commodity-computer clusters. In contrast to **Hadoop**, that is written in Java … described the approach in the paper and not released its proprietary software, many open-source frameworks were created in order to implement the model. As examples one may say **Hadoop** or the limited …

answered nov 18 by Prometheus

A: Best NoSQL approach to handle 100+ million records

into the various NoSQL solutions **out** there. The default unit of distribution is key. Therefore you need to remember that you need to be able to split your data between your node machines effectively otherwise … loss in a prod system is a big deal) . I am using **Hadoop**, HBase, Hive, MongoDB, Solr, MySQL and even good old text files. Remember that to productionize a system using these technogies is a bit harder …

answered jul 4 '11 by NightWolf

3 answers

Q: Java: Can I deserialize an object without calling its constructor?

lookups in **Hadoop**? edit: So, it looks like the no-arg constructor of the least-inherited class is the one that will be called. Here's my hirearchy: abstract KDTree QuadTree extends KdTree … implements Serializable From all my experiments, I need to implement Serializable on both. Now, and Since I never see it print **out** (aside from the initial construction...), I guess it's all good. …

java serialization

asked feb 12 '13 by dranxo

1 answer

Q: MapReduce for cross-correlating datasets drawn from 100TB of data [closed]

. In our current implementation, we spawn off threads to process each region, and fetch individual elements by reading files over NFS. It turns **out** that this solution is I/O bound, and we're now … say I went with **Hadoop**. My first thought would be to get the data into HDFS as chunks, attempting to make each chunk consist of elements from the same region as best as possible. Each Map task would …

hadoop mapreduce hdfs

asked dec 2 '12 by David A Tarris

A: Nuodb and HDFS as storage

: Local Files System Amazon Web Services: Simple Storage volume (S3), Elastic Block Storage (EBS) **Hadoop** Distributed Files System (HDFS) So, to elaborate on how NuoDB works with HDFS … , to ensure that data held in cache is kept up to date with all of the changes happening within the system. Garbage Collection also kicks in and clears **out** atoms in a Least Recently Used order, when …

answered jan 29 '13 by NuoDB Support

4 answers

Q: Realistic Social Network Model with tens of millions Users. Which technologies should I be u…

. 2. Should I be using a different language all together. I know that technologies such as Lucene, **Hadoop**, etc. were created with Java, and are used for large amounts of data...But I have never used … before I can give a User a list of Friends. Sorry for the semi-long read, but I wanted to lay **out** exactly where I am so you could guide me in the right direction. Thank you to everyone that took the time to read/help me with this topic. …

java database arrays memory data-structures

asked jan 12 '11 by Eddie

1 answer

Q: Json parse with elephantbird in Pig

I can't get the following data to parse in Pig. It's what the twitter API returns after getting all tweets from a certain user. source data: (I removed some numbers to not invade on anyone's privacy …

hadoop apache-pig elephantbird

asked nov 3 by Havnar

A: What are the disadvantages of using .Rdata files compared to HDF5 or netCDF?

well with Python, Matlab, and many other systems. HDF5 is superior to Python's pickle storage in many ways - check **out** PyTables and you'll very likely see good speedups. Matlab used to have (and may … to store, how will you store it, and how will it be represented and accessed? Once you get your data conversion plan in place, you can then use tools like **Hadoop** or even basic multicore functionality …

answered oct 25 '11 by Iterator

2 answers

Q: How to dynamically scale StarCluster/qsub/EC2 to run parallel jobs across multiple nodes

of parallel jobs as and when output is written **out**? I am test running few scenarios but I would like to know if there are people who have experimented on similar scenarios. Any suggestions using **Hadoop** Plugin? http://star.mit.edu/cluster/docs/0.93.3/plugins/hadoop.html Thanks in Advance Karthick …

python hadoop qsub parallel-python starcluster

asked mar 11 '13 by user1652054

1 answer

Q: Pipe Broken exception every time when I run Mahout samples at EC2 server

I've installed mahout at bitnami AMI ami-02fb006b, (as well as several other ami's, otherwise I won't be asking the question) according to instructions provided here and here: I'm always getting st …

amazon-ec2 amazon-web-services mahout

asked aug 16 '11 by Arsen Zahray

1 answer

Q: Maven build issue - Connection to repository refused

I wish to import, change, rebuild, test and push/check-in my changes to the code available in this Github

repository Currently, I do not wish to use any IDE or plug-ins for this purpose. I installed …

maven github maven-3

asked may 24 '13 by Kaliyug Antagonist

0
answers

Q: Correlation Network Implementation

, and making sure there are no duplicate nodes, is a nightmare that I have not been able to figure **out** Boost Graph Library Intractable, arcane definitions for objects, and how to use them Documentation … already incorporated Reads easier than the BGL Not as many functions Still arcane Further down the road, I envision the graph living on a distributed network, with distributed storage (**hadoop** …

algorithm boost graph graph-algorithm minimum-spanning-tree

asked feb 22 '13 by NuclearAlchemist

3
answers

Q: Maven Building Error

when i try to run mvn clean package I am getting the following error pom.xml File Through wget http://repo.maven.apache.org it's not showing any error. Can any one help me in this. I a …

java maven maven-2 ycsb

asked may 21 '13 by Rohit

0
answers

Q: distcp not working on CD4.5

I upgraded from CDH 4.4.0-1.cdh4.4.0.p0.39 to CDH 4.5.0-1.cdh4.5.0.p0.30. I had distcp working in prior version from hdfs to s3: when upgrade I'm getting a few weird errors and warnings now: 13 …

cloudera

asked dec 12 '13 by user706794