Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 42 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 42
Topic #: 1
[All Professional Data Engineer Questions]

Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.

B. Rewrite the job in Apache Spark.

C. Increase the size of the Hadoop cluster.

D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Show Suggested Answer

by jvg637 at *March 15, 2020, 1:43 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 4 years, 7 months ago

I would say B since Apache Spark is faster than Hadoop/Pig/MapReduce

upvoted 36 times

Trocinek 7 months, 2 weeks ago

But it requires much more memory causing it more expensive, which is not what we're aiming for here..

upvoted 1 times

ler_mp Highly Voted 1 year, 10 months ago

Wow, a question that does not recommend to use Google product

upvoted 19 times

axantroff Most Recent 11 months, 2 weeks ago

Selected Answer: B

Just a regular Spark. B

upvoted 1 times

DataFrame 11 months, 3 weeks ago

C. I think it should be C because intent of asking question is to realize the problem of on-prem auto-scaling not the optimization that we achieve using spark in-memory features. Its GCP exam they want to highlight if hadoop cluster commodity hard doesn't increase when data increases then it can create problem unlike GCP. Hence migrate to GCP.

upvoted 1 times

itsmynickname 1 year, 3 months ago

None. Being a GCP exam, it must be either Dataflow or BigQuery :D

upvoted 11 times

KHAN0007 1 year, 6 months ago

I would like to take a moment to thank you all guys
You guys are awesome!!!

upvoted 5 times

Whoswho 1 year, 10 months ago

looks like he's trying to spark the company up.

upvoted 8 times

itsmynickname 1 year, 3 months ago

It seems he's not well paid.

upvoted 2 times

Krish6488 1 year, 10 months ago

Selected Answer: B

Both Pig & Spark requires rewriting the code so its an additional overhead, but as an architect I would think about a long lasting solution. Resizing Hadoop cluster can resolve the problem statement for the workloads at that point in time but not on longer run. So Spark is the right choice, although its a cost to start with, it will certainly be a long lasting solution

upvoted 4 times

Mamta072 2 years, 4 months ago

Ans is B . Apache spark.

upvoted 2 times

alecuba16 2 years, 6 months ago

Selected Answer: B

SPARK > hadoop, pig, hive

upvoted 4 times

kped21 2 years, 8 months ago

B - Apache Spark

upvoted 1 times

luamail 1 year, 7 months ago

https://www.ibm.com/cloud/blog/hadoop-vs-spark

upvoted 2 times

kped21 2 years, 9 months ago

B Spark for optimization and processing.

upvoted 1 times

sraakesh95 2 years, 9 months ago

Selected Answer: B

B: Spark is suitable for the given operation is much more powerful

upvoted 1 times

medeis_jar 2 years, 10 months ago

Selected Answer: B

as explained by pr2web

upvoted 1 times

pr2web 2 years, 10 months ago

Selected Answer: B

Ans B:
Spark is a 100 times faster and utilizes memory, instead of Hadoop Mapreduce's two-stage paradigm.

upvoted 1 times

MaxNRG 2 years, 11 months ago

B as Spark can improve the performance as it performs lazy in-memory execution.
Spark is important because it does part of its pipeline processing in memory rather than copying from disk. For some applications, this makes Spark extremely fast.

upvoted 1 times

MaxNRG 2 years, 11 months ago

With a Spark pipeline, you have two different kinds of operations, transforms and actions. Spark builds its pipeline used an abstraction called a directed graph. Each transform builds additional nodes into the graph but spark doesn't execute the pipeline until it sees an action.
Spark waits until it has the whole story, all the information. This allows Spark to choose the best way to distribute the work and run the pipeline. The process of waiting on transforms and executing on actions is called, lazy execution. For a transformation, the input is an RDD and the output is an RDD. When Spark sees a transformation, it registers it in the directed graph and then it waits. An action triggers Spark to process the pipeline, the output is usually a result format, such as a text file, rather than an RDD.

upvoted 1 times

MaxNRG 2 years, 11 months ago

Option A is wrong as Pig is wrapper and would initiate Map Reduce jobs
Option C is wrong as it would increase the cost.
Option D is wrong Hive is wrapper and would initiate Map Reduce jobs. Also, reducing the size would reduce performance.

upvoted 3 times

kastuarr 2 years ago

Wont Option B increase the cost ? Cost of re-writing the job in Spark + Cost of additional memory ?

upvoted 1 times

anji007 3 years ago

Ans: B
Spark performs better than MapReduce due to in memory processing.

upvoted 2 times

Load full discussion...