Google Discussions

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

Go to Exam

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 58 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 58

Topic #: 1

[All Professional Data Engineer Questions]

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.

B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.

C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.

D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Show Suggested Answer

by [deleted] at *March 21, 2020, 11:31 a.m.*

Comments

Type your comment...

Submit

SteelWarrior Highly Voted 4 years, 7 months ago

Should go with B. Two reasons, it is a cleaner approach with single job to handle the calibration before the data is used in the pipeline. Second, doing this step in later stages can be complex and maintenance of those jobs in the future will become challenging.

upvoted 58 times

> **Yiouk** 3 years, 9 months ago
>
> B. different MR jobs execute in series, adding 1 more job makes sense in this case.
>
> upvoted 7 times

[Removed] Highly Voted 5 years, 1 month ago

Answer: A
Description: My take on this is for sensor calibration you just need to update the transform function, rather than creating a whole new mapreduce job and storing/passing the values to next job

upvoted 20 times

> **Jphix** 3 years, 11 months ago
>
> It's B. A would involving changing every single job (notice it said jobS, plural, not a single job). If that is computationally intensive, which it is, you're repeating a computationally intense process needlessly several times. SteelWarrior and YuriP are right on this one.
>
> upvoted 11 times
>
> > **mark1223jkh** 11 months, 3 weeks ago
> >
> > Why all jobs, change only the first job for calibration, right?
> >
> > upvoted 1 times

Marwan95 Most Recent 10 months ago

Selected Answer: A

I'll choose A. WHY? cause the process already takes DAYS and adding another step will increase the time more

upvoted 1 times

jin0 2 years, 2 months ago

What kinds of sensor calibrations exists? I don't understand how computation in pipeline would be expense due to calibration being omitted..?

upvoted 1 times

samdhimal 2 years, 3 months ago

B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.

This approach would ensure that sensor calibration is systematically carried out every time the ETL process runs, as the new MapReduce job would be responsible for calibrating the sensors before the data is processed by the other steps. This would ensure that all data is calibrated before being analyzed, thus avoiding the omission of the sensor calibration step in the future.
It also allows you to chain all other MapReduce jobs after this one, so that the calibrated data is used in all the downstream jobs.

upvoted 1 times

> **samdhimal** 2 years, 3 months ago
>
> Option A is not ideal, as it would be time-consuming to modify all the transformMapReduce jobs to apply sensor calibration before doing anything else, and there is a risk of introducing bugs or errors.
> Option C is not ideal, as it would rely on users to apply sensor calibration themselves, which would be inefficient and could introduce inconsistencies in the data.
> Option D is not ideal, as it would require a lot of simulation and testing to develop an algorithm that can predict the variance of data output accurately and it may not be as accurate as calibrating the sensor directly.
>
> upvoted 1 times

DipT 2 years, 4 months ago

Selected Answer: B

It is much cleaner approach

upvoted 1 times

DGames 2 years, 4 months ago

Selected Answer: B

Best approach is calibration will be separate job because if we need to tune the calibration later also it would be to maintain without worries about all other jobs.

upvoted 1 times

odacir 2 years, 4 months ago

Selected Answer: B

Should be B. My reason, this is like an Anti corruption layer, and that's a good practice,
C- , if you modify your transformMapReduce will be harder to test and debug, so it's a bad practice.
C the idea de introduce manual operation is an anti patron and has a lot of problems
D It's overkilling, a don't have sense in this scenario.

upvoted 1 times

ZIMARAKI 3 years, 3 months ago

Selected Answer: B

SteelWarrior explanation is correct :)

upvoted 3 times

lord_ryder 3 years, 3 months ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

medeis_jar 3 years, 4 months ago

Selected Answer: B

SteelWarrior explanation is correct

upvoted 1 times

hendrixlives 3 years, 4 months ago

Selected Answer: B

SteelWarrior's answer is correct

upvoted 1 times

anji007 3 years, 6 months ago

Ans: B
Adding a new job in the beginning of chain makes more sense than updating existing chain of jobs.

upvoted 1 times

sumanshu 3 years, 10 months ago

Vote for 'B' (introduce new job) over 'A', (instead of modifying existing job)

upvoted 5 times

YuriP 4 years, 9 months ago

Should be B. It's a Data Quality step which has to go right after Raw Ingest. Otherwise you repeat the same step unknown (see "job_s_" in A) number of times, possibly for no reason, therefore extending ETL time.

upvoted 5 times

[Removed] 5 years, 1 month ago

It's between A or B.
Should choose A

upvoted 4 times

EXAMTOPICS

Platform

- Home
- All Exams
- Examtopics PRO
- Training Courses