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 78 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 78

Topic #: 1

[All Professional Data Engineer Questions]

You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

A. PigLatin using Pig

B. HiveQL using Hive

C. Java using MapReduce

D. Python using MapReduce

Show Suggested Answer

by [deleted] at *March 21, 2020, 6:11 p.m.*

Comments

Type your comment...

Submit

[Removed] Highly Voted 5 years, 1 month ago

Answer: A
Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

upvoted 23 times

[Removed] Highly Voted 5 years, 1 month ago

Should be A

upvoted 15 times

oliiivier Most Recent 1 week, 1 day ago

Selected Answer: C

Réponse : C
Explication :
La bonne réponse ici est C. Java using MapReduce.
Explication rapide :
Tu parles de besoin de "checkpointing" et "splitting pipelines".
PigLatin (Pig) et HiveQL (Hive) sont des langages déclaratifs (de plus haut niveau), pas faits pour un contrôle précis sur la manière dont les jobs se séquencent, checkpointent ou s'articulent.
MapReduce en Java te donne un contrôle total sur :
L'ordonnancement précis.
Le checkpointing manuel.
Le découpage et le chaînage des étapes.
Python avec MapReduce est possible mais beaucoup moins natif ; Hadoop est conçu principalement pour Java MapReduce, donc en Python ce serait plus compliqué, plus fragile et moins performant.
Résumé rapide pour que tu t'en souviennes : Besoin de contrôle précis sur la logique ETL complexe = MapReduce en Java.

upvoted 1 times

Parandhaman_Margan 1 month, 3 weeks ago

Selected Answer: C

MapReduce in Java (C) allows more control for checkpointing and splitting.

upvoted 1 times

desertlotus1211 1 month, 3 weeks ago

Selected Answer: C

Writing your pipeline in Java using MapReduce allows you to implement these custom controls and fine-tune the execution, ensuring robust and manageable ETL processes on your Hadoop cluster

upvoted 1 times

grshankar9 3 months, 2 weeks ago

Selected Answer: A

Pig Latin supports both splitting pipelines and checkpointing, allowing users to create complex data processing workflows with the ability to restart from specific points in the pipeline if necessary.

upvoted 1 times

SamuelTsch 6 months, 2 weeks ago

Selected Answer: A

I would go to A.
C, D are similar. So both are excluded. B, Hive is actually a data warehouse system. I don't use Apache Pig. But, BCD are wrong. Then A should be correct.

upvoted 1 times

AnonymousPanda 1 year, 8 months ago

Selected Answer: A

A as others have said

upvoted 1 times

Oleksandr0501 2 years ago

Selected Answer: C

Comment content is too short

upvoted 2 times

juliobs 2 years, 1 month ago

Selected Answer: A

PigLatin is the correct answer, however... the last release was 6 years ago and has lots of bugs.

upvoted 2 times

musumusu 2 years, 2 months ago

This answer depends which language you are comfortable with.
Hadoop is your framework, where mapReduce is your Native programming model in JAVA, which is designed to scale, parallel processing, restart pipeline from any checkpoint etc. , So if you are comfortable with JAVA, you can customize your checkpoint at lowlevel in better way. otherwise, choose PIG which is another programming concept run over JAVA but then you need to learn this also, if not choose python as it can be deployed with hadoop because hadoop has been making

updates for python clients regularly.
Option C: is the best one.

upvoted 7 times

samdhimal 2 years, 3 months ago

C. Java using MapReduce or D. Python using MapReduce

Apache Hadoop is a distributed computing framework that allows you to process large datasets using the MapReduce programming model. There are several options for writing ETL pipelines to run on a Hadoop cluster, but the most common are using Java or Python with the MapReduce programming model.

upvoted 4 times

samdhimal 2 years, 3 months ago

A. PigLatin using Pig is a high-level data flow language that is used to create ETL pipelines. Pig is built on top of Hadoop, and it allows you to write scripts in PigLatin, a SQL-like language that is used to process data in Hadoop. Pig is a simpler option than MapReduce but it lacks some capabilities like the control over low-level data manipulation operations.

B. HiveQL using Hive is a SQL-like language for querying and managing large datasets stored in Hadoop's distributed file system. Hive is built on top of Hadoop and it provides an SQL-like interface for querying data stored in Hadoop. Hive is more suitable for querying and managing large datasets stored in Hadoop than for ETL pipelines.

Both Java and Python using MapReduce provide low-level control over data manipulation operations, and they allow you to write custom mapper and reducer functions that can be used to process data in a Hadoop cluster. The choice between Java and Python will depend on the development team's expertise and preference.

upvoted 3 times

cetanx 1 year, 11 months ago

It has to be C
because while Pig can be used to simplify the writing of complex data transformation tasks and can store intermediate results, it doesn't provide the detailed control over checkpointing and pipeline splitting in the way that is typically implied by those terms.

also, while one can write MapReduce jobs in languages other than Java (like Python) using Hadoop Streaming or other similar APIs, it may not be as efficient or as seamless as using Java due to the JVM-native nature of Hadoop.

upvoted 2 times

Koushik25sep 2 years, 7 months ago

Selected Answer: A

Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

upvoted 1 times

BigDataBB 3 years, 2 months ago

Why not D?

upvoted 1 times

rbeeraka 3 years, 3 months ago

Selected Answer: A

PigLatin supports checkpoints

upvoted 1 times

davidqianwen 3 years, 3 months ago

Selected Answer: A

Answer: A

upvoted 1 times

maddy5835 3 years, 6 months ago

Pig is just a scripting language, how pig can be used in creation of pipelines, should be answer from c & D

upvoted 3 times

Load full discussion...

Platform

- Home
- All Exams
- Examtopics PRO
- Training Courses