

🔗 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 158 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 158

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You need to deploy additional dependencies to all nodes of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?

- A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master
- B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet
- C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter
- D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role

[Show Suggested Answer](#)

by [rickywck](#) at March 18, 2020, 1:46 a.m.

### Comments

Type your comment...

[Submit](#)

🗨️ [\[Removed\]](#) [Highly Voted](#) 👍 4 years, 7 months ago  
Correct: C




If you create a Dataproc cluster with internal IP addresses only, attempts to access the Internet in an initialization action will fail unless you have configured routes to direct the traffic through a NAT or a VPN gateway. Without access to the Internet, you can enable Private Google Access, and place job dependencies in Cloud Storage; cluster nodes can download the dependencies from Cloud Storage from internal IPs.

   upvoted 39 times

  **AzureDP900** 1 year, 10 months ago

Thank you for detailed explanation. C is right

   upvoted 1 times

  **rickywck** Highly Voted  4 years, 7 months ago

Should be C:

<https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>

   upvoted 12 times

  **b3e59c2** Most Recent  4 months ago

**Selected Answer: C**


A: Incorrect, Proxy allows for connection to a Cloud SQL instance, which unless you have the dependencies stored there (doesn't seem viable or smart), would achieve nothing.

B: Incorrect, will allow for a connection to the internet to be made for installing the dependencies, however this goes against the companies security policies so should not be considered.

C: Correct, only one that makes sense and is best practise (see <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions>)

D: Incorrect, provides access to a shared VPC network, however doesn't necessarily provide a way to access the dependencies. And even if it did, would go against company security policy.

   upvoted 1 times

  **gcpdataeng** 7 months, 2 weeks ago

**Selected Answer: C**

c looks good

   upvoted 1 times

  **barnac1es** 1 year, 1 month ago

**Selected Answer: C**


Security Compliance: This option aligns with your company's security policies, which prohibit public Internet access from Cloud Dataproc nodes. Placing the dependencies in a Cloud Storage bucket within your VPC security perimeter ensures that the data remains within your private network.

VPC Security: By placing the dependencies within your VPC security perimeter, you maintain control over network access and can restrict access to the necessary nodes only.

Dataproc Initialization Action: You can use a custom initialization action or script to fetch and install the dependencies from the secure Cloud Storage bucket to the Dataproc cluster nodes during startup.

By copying the dependencies to a secure Cloud Storage bucket and using an initialization action to install them on the Dataproc nodes, you can meet your security requirements while providing the necessary dependencies to your cluster.


   upvoted 3 times

  **knith66** 1 year, 2 months ago

**Selected Answer: C**

C is correct

   upvoted 1 times

  **charline** 1 year, 7 months ago

**Selected Answer: C**

C seems good

   upvoted 1 times

  **musumusu** 1 year, 8 months ago

Answer C,

It needs practical experience to understand this question. You create cluster with some package/software i.e dependencies such as python packages that you store in .zip file, then you save a jar file to run the cluster as an application such as you need java while running spark session. and some config yaml file.

These dependencies you can save in bucket and can use to configure cluster from external window , sdk or api. without going into UI.

Then you need to use VPC. to access these files

When you need to use VPC to access those resources

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **zellck** 1 year, 11 months ago

**Selected Answer: C**

C is the answer.

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#and\\_vpc-sc\\_networks](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#and_vpc-sc_networks)

With VPC Service Controls, administrators can define a security perimeter around resources of Google-managed services to control communication to and between those services.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **DataEngineer\_WideOps** 2 years, 3 months ago

Without access to the internet, you can enable Private Google Access and place job dependencies in Cloud Storage; cluster nodes can download the dependencies from Cloud Storage from internal IPs.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **medeis\_jar** 2 years, 10 months ago

**Selected Answer: C**

[https://cloud.google.com/dataproc/docs/concepts/configuring-](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only)

[clusters/network#create\\_a\\_cloud\\_dataproc\\_cluster\\_with\\_internal\\_ip\\_address\\_only](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only)

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **Prabusankar** 2 years, 10 months ago

When creating a Dataproc cluster, you can specify initialization actions in executables or scripts that Dataproc will run on all nodes in your Dataproc cluster immediately after the cluster is set up. Initialization actions often set up job dependencies, such as installing Python packages, so that jobs can be submitted to the cluster without having to install dependencies when the jobs are run

👍 ↩ 🚩 upvoted 3 times

🗄️ 👤 **JG123** 2 years, 11 months ago

Correct: C

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **clouditis** 4 years, 2 months ago

c it is!

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **Rajokkiyam** 4 years, 7 months ago

Should be C

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **[Removed]** 4 years, 7 months ago

Should be C

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **jvg637** 4 years, 7 months ago

I think the correct answer might be C instead, due to [https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create\\_a\\_cloud\\_dataproc\\_cluster\\_with\\_internal\\_ip\\_address\\_only](https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/network#create_a_cloud_dataproc_cluster_with_internal_ip_address_only)

👍 ↩ 🚩 upvoted 4 times



## Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

