CDP Data Developer (3001) Certification Guide

(by AMRIT CHHETRI, RCS()

CREATED FOR : CDP-3001 Certification Training from Rosefinch,Siliguri

Contents

CDP Data Developer Exam (CDP 3001) Details:

CDP Data Developer is latest Developer Certification from Cloudera and short details are:

1. Exam Number: CDP-3001
2. Number of questions: 65
3. Duration: 90 minutes
4. Pass Score: unpublished
   We do not publish exam pass scores. Candidates should not be trying to achieve any particular score. Rather they should be aiming for the highest score possible.
5. Delivery: online, proctored
6. Please review the system requirements to enable online, proctored testing through QuestionMark
7. Allowed resources: none.
8. Official URL: https://www.cloudera.com/about/training/certification/cdp-datadev-exam-cdp-3001.html

1. Exam Portal(Purchase and Registration): https://support.questionmark.com/user/register
2. FREE Voucher: https://www.cloudera.com/about/training.html

Training Steps:

These are recommended steps for the preparation of Cloudera Developer CDP-3001 Certification:

9. Register Account : https://sso.cloudera.com/register.html
10. Purchase Exam -directly or from Training Center:
    https://education.cloudera.com/store/2737716-cdp-3001-cdp-data-developer-exam

CDP Private Cloud Installation:

1. **CDP Cluster Installations- for CDP Cluster:**
   The requirements for the installations of ***CDP Private Cloud Base(Cluster)*** are given below and they are based on **https://supportmatrix.cloudera.com/** **.** There is requirement of 2 Virtual Instances each with 16 GB RAM and 250 GB HDD CDP Private Cloud Base (Cluster) installation.

   A. Requirements Details –with Ubuntu 20.04 LTS:

   a. System Matrices:

Peripherals/ Infrastructure	**Peripherals/ Infrastructure**	**Specifications**	**URL/Remarks**
Operating System	Ubuntu	20.04 LTS	https://www.releases.ubuntu.com/ 20.04/
RAM	16 GB		
Storage/HDD	250 GB		
CPU	X86-64 Architecture		
Network Bandwidth	3-5 MBPS	Unlimited	
SSH User			cdproot with root privileges

a. Number of Instances : 1
b. Access Type : Root
c. Number of Days : Till Training completes

B. Requirements Details –with Ubuntu 20.04 LTS:

b. System Matrices:

Peripherals/ Infrastructure	**Peripherals/Infrastructure**	**Specifications**	**URL**
Operating System	Centos	20.04 LTS	http://isoredirect.centos.org/centos/8-stream/isos/x86_64/
RAM	16 GB Tuned: 64 GB		
Storage/HDD	250 GB		
CPU	X86-64 Architecture		
Network Bandwidth	3-5 MBPS	Unlimited	
SSH User			cdproot with root privileges

Cloudera Manager 7 Installation-Steps

2. Register an account: https://sso.cloudera.com/onboarding.html?stepId=pocForm
3. Get installer for CDP Installer (60 Days Trial) https://www.cloudera.com/downloads/cdp-private-cloud-trial/cdp-private-cloud-base-trial.html
4. Check for supportable Linux OS(Ubuntu 20.04 or Centos) .Supporting OS : https://supportmatrix.cloudera.com/, and get Ubuntu 20.04, https://releases.ubuntu.com/20.04/ubuntu-20.04.4-desktop-amd64.iso
5. Request Trial at https://www.cloudera.com/downloads/cdp-private-cloud-trial/cdp-private-cloud-base-trial.html
6. Install Cloudera Manager with steps below(they are mentioned in the URL above too):

    a. Download Cloudera Manager Installer and install:
    b. $ wget https://archive.cloudera.com/cm7/7.4.4/cloudera-manager-installer.bin
    c. $ chmod u+x cloudera-manager-installer.bin
    d. $ sudo ./cloudera-manager-installer.bin

7. Once installer completes, create SSH user(cdproot) with Root/Admin Privileges

    Option-1:

    ```
    sudo adduser cdproot
    sudo usermod -aG sudo cd root~~cdproot~~
    ```

    Option-2:

    ```
    sudo adduser  cdproot sudo
    Checks: groups cdproot
    ```

8. Install and start OpenSSH-Server to run at 22

    ```
    # sudo apt-get install openssh-server
    # sudo systemctl enable ssh
    # sudo systemctl start ssh
    ```

9. SSH Login and Root Permission:

    a. SSH Login:

    ```
    sudo nano /etc/ssh/sshd_config
                    or
    sudo nano /etc/ssh/sshd_config
    Add This : PermitRootLogin yes
    ```

    b. Root User:

    ```
    sudo gedit /etc/sudoer   and allow NOPASSWRD logins
    ```

10. Start SCM Server sudo service cloudera-scm-server start

    Other options:

    ```
    sudo service cloudera-scm-server start
    sudo service cloudera-scm-server stop
    sudo service cloudera-scm-server restart
    ```

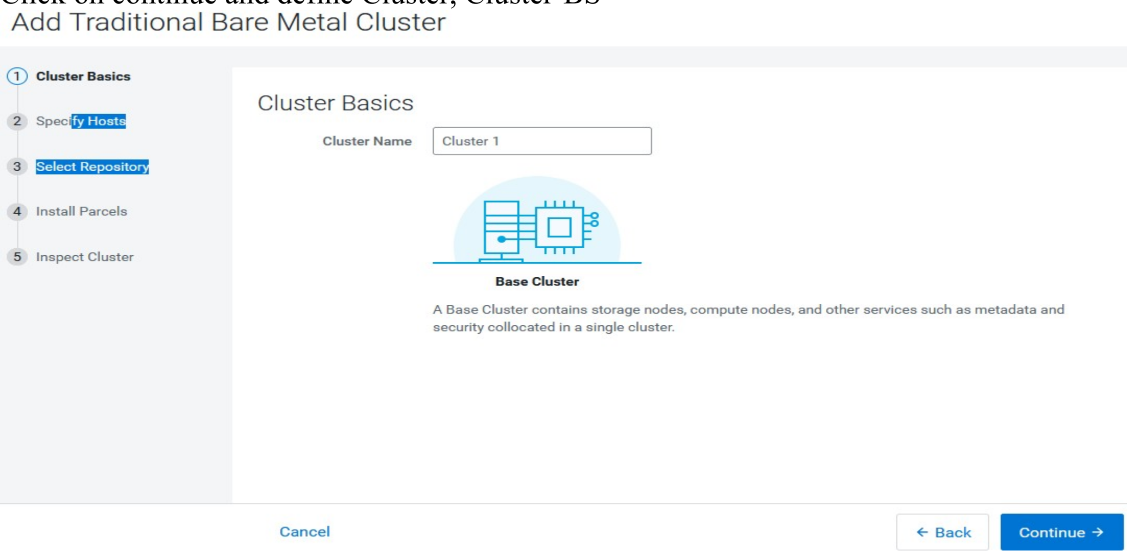
11. Access Cloudera Manager at https://<ip>:7180, username/password admin/admin/
12. Follow on-screen details to install and configure all parcels/packages
13. **Configure and verify Parcel URL Configuration before processing the installation**

    (Works in most scenarios)...

Cloudera Data Platform installation using Cloudera~~Clouder~~ Manager 7:

Now, follow the steps below to complete installation of CDP :

1. Login to Cloudera Manager 7 with given URL http://54.255.20.131:7180 and credentials
2. Click on continue and define Cluster, Cluster-BS



Add Traditional Bare Metal Cluster

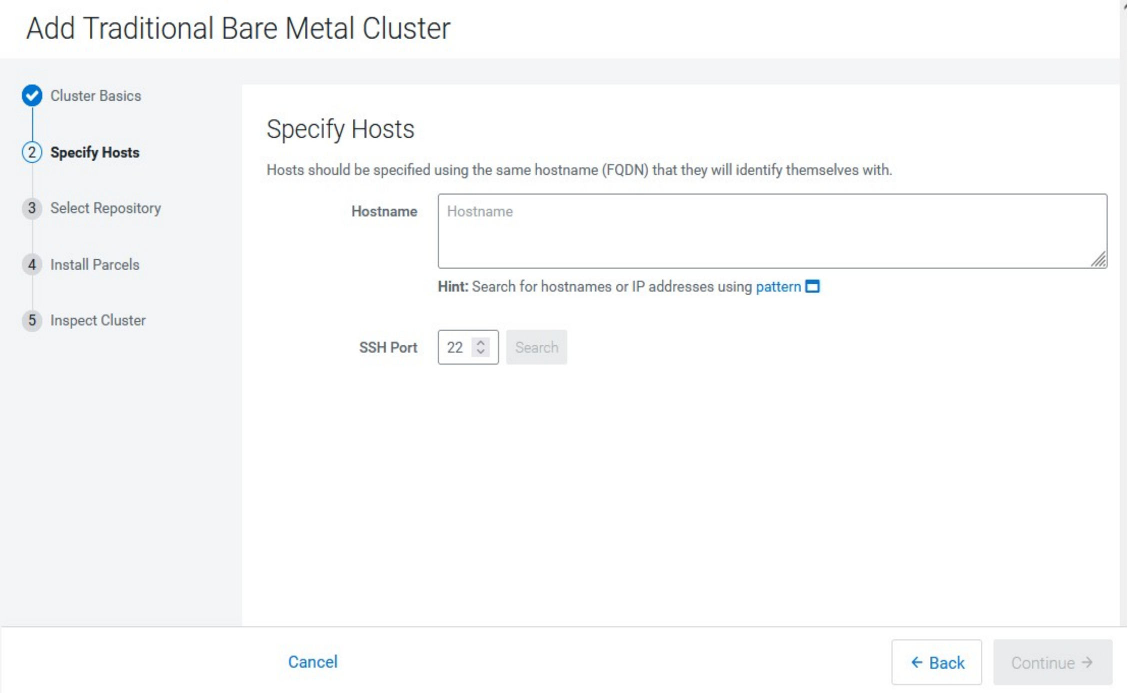
Cluster Basics

Cluster Name: Cluster 1

Base Cluster

A Base Cluster contains storage nodes, compute nodes, and other services such as metadata and security collocated in a single cluster.

3. Click on continue and search the server system by Hostname



Add Traditional Bare Metal Cluster

Specify Hosts

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Hostname: Hostname

Hint: Search for hostnames or IP addresses using pattern

SSH Port: 22

4. Before clicking Continue create another user with root permission/privileges
   Option-1:
   sudo adduser cdproot
   sudo usermod -aG sudo cd root
   Option-2:
   sudo adduser cdproot sudo
   Checks: groups cdproot

5. Permit RootLogin

```
sudo nano /etc/ssh/sshd_config
Add This : PermitRootLogin yes
sudo gedit /etc/sudoer   and allow NOPASSWRD logins
```

6. Click next and select “System JDK” and continue, need start installing all packages.

-------- Installed worked well in 2 VMs , have to update the contents here ....

CDP Developer Workstation for CDP 3001 Certification:

1. **CDP Data Developer Certification –for Participant's Environment**
   a. System Matrices:

Peripherals/ Infrastructure	**Peripherals/ Infrastructure**	**Specifications**	**URL**
Operating System	Windows 10	Professional, 64-Bits	-
RAM	4-12 GB		-
Storage/HDD	80 GB	-	-
CPU	X86-64 Architecture		-
Network Bandwidth/Internet	3-5 MBPS	Unlimited	-
Development SDK	Java	11	https://www.oracle.com/in/java/techn archive-downloads.html
	Scala		https://www.scala-lang.org/downloac
	Python	3.10	https://www.python.org/downloads/r
Development IDE	PyCharm Edu (for Python)		https://www.jetbrains.com/edu-produ #section=pycharm-edu
	IntelliJ Edu(Scala and Java)		https://www.jetbrains.com/edu-produ #section=idea-Scala
	Jupyter Notebook		https://jupyter.org/install
Putty	Smart Putty	3.0	https://sysprogs.com/SmarTTY/
Notebook	Notepad++	8.4.4	https://notepad-plus-plus.org/
MLOps Tools	Jenkins	3.65	https://www.jenkins.io/download/
	Jenkins X/GitOps		https://jenkins-x.io/
Codes Management	Git	2.37	https://git-scm.com/downloads
	Anaconda		https://www.anaconda.com/

b. Number Of Instance: 1 /Participant(Accessible Physically)

Topics of CDP Data Developer Exam (CDP 3001)

https://www.cloudera.com/about/training/certification/cdp-datadev-exam-cdp-3001.html

CDP 3001 Topics:

3. Connect and move data between systems
4. Build and manage a data warehouse
5. Build, schedule, execute, and monitor data pipelines
6. Clean and serve data to the end-users
7. Perform data quality checks
8. Debug data issues reported by end-users
9. Data backup and disaster recovery

Studying CDP 3001 Certification:

1. **Connect and move data between systems**
   a. Tools: Apache NiFi DataFlows,
   b. Moving: In and Out of Snowflake, Ozone using **NiFi**
   c. References:
      i. Moving Data: https://docs.cloudera.com/cfm/2.1.4/howto-exchanging-data.html
      ii. Clouder Products : https://www.cloudera.com/products/pricing.html

2. **Build and manage a data warehouse**
   a. Tools:
   b. References
      i. DW Fundamentals : https://www.cloudera.com/about/events/webinars/modern-dw-fundamentals.html
      ii. Designing of Enterprise Data Science Warehouse
      iii. ***Designing(on AWS, Azure):*** https://docs.cloudera.com/data-warehouse/cloud/index.html

   Handson on Dataware is must! CC free registration ….

3. **Build, schedule, execute, and monitor data pipelines:**
   a. Tools:
   b. References:
      i. Data Pipelines: https://docs.cloudera.com/cdsw/1.9.2/jobs-pipelines/topics/cdsw-jobs-pipelines.html

        ii. Scheduling: https://docs.cloudera.com/data-engineering/1.3.4/manage-jobs/topics/cde-schedule-job.html

4. **Clean and serve data to the end-users**
    a. **References:**
        i. Cleaning Old Data: https://docs.cloudera.com/runtime/7.2.6/troubleshooting-hue/topics/hue-cleanup-old-data-to-improve-performance.html
        ii.

5. **Perform data quality checks**
    a. References:
        i. Monitoring : https://docs.cloudera.com/cdp-private-cloud-base/7.1.4/monitoring-and-diagnostics/cm-monitoring-and-diagnostics.pdf

6. **Debug data issues reported by end-users**
    a. References:
        i. Debugging Issues with Models : https://docs.cloudera.com/machine-learning/1.0/models/topics/ml-models-debug.html
            -

7. **Data backup and disaster recovery**
    a. References:
        i. HBase Backup Strategies: https://docs.cloudera.com/cdp-private-cloud-base/7.1.3/hbase-backup-dr/topics/hbase-backup-dr-strategies.html
        ii. Backup Strategies: https://docs.cloudera.com/cdsw/1.9.2/cluster-management/topics/cdsw-bdr.html
        iii. Backup : https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_bdr_howto_hdfs.html#backup_hdfs_from_prod_to_backup

8. **DevOps and Clouder Security**
    a. Tools: Git/GitHub, Jenkins, Anisible
    b. Jenkins in MLOps - MLOps Guide | Jenkins X - Cloud Native CI/CD Built On Kubernetes (jenkins-x.io)
    c. Jenkin X for MlOps Pipelines - Jenkins X - Cloud Native CI/CD Built On Kubernetes (jenkins-x.io)
    d. References:
        i. Kafka Security : https://developer.confluent.io/learn-kafka/security/intro/

Self-Study Modules/Resources

Self-Study Modules/Resources- By Cloudera:
Related Training – Student should complete themselves on Cloudera OnDemand Training Library

1. https://www.cloudera.com/about/training.html

Self-Study Modules/Resources- By Instructor:

1. **Kafka Security** : https://developer.confluent.io/learn-kafka/security/intro/
   https://confluent.cloud/environments/env-57yrng/add-cluster

Everything news, others covers CCA 175 but CDP-3001 is completely new with MLops using Jenkin X, Cluster Management!!!

Books – No available

1. E-Books
   a. Book: Cloudera Data Platform Fundamentals and Concepts (hadoopexam.com)
   b. CDP Private Cloud Base: Cloudera Data Platform Private Cloud Base with IBM Spectrum Scale [Book] (oreilly.com)

CDP Coding Spark:

Systems Requirements for the Participants:

1. **CDP Access** or own installation (if anyone can manage own)

[2.] Pycharm Edu and Python - Python Coding~~Codding~~ for **PySpark**

2.[3.] Intellij IDE Edu- Java and Scala Coding for ***Spark Coding with Scala***

3.[4.] Firefox and Chrome Browsers –
   a. CDP Data Developer: https://www.cloudera.com/about/training/certification/cdp-datadev-exam-cdp-3001.html
   b. FAQ: https://www.cloudera.com/about/training/certification/cdp-faqs.html

4.[5.] **Recommended Resources Access:**
   a. Supporting Cloudera Courses: https://education.cloudera.com/store?utf8=%E2%9C%93&st=free&commit=

5.[6.] **Remote Access to CDP for PySpark Shell:**
   a. Create Linux Users on CDP Server , example CDP3001-001
   b. Getting remote session on Smart Putty,
      https://sysprogs.com/SmarTTY/download/
   c. Starting Spark shell , $ spark-shell or spark2-shell
   d. Checks with Sample Code:

```
Scala>val log = sc.textFile("/tmp/logs")
scala>print(log)
```

Note: Practical Labs on Spark(Python) on Shell, Jupyter Notebook and PyCharm, Scala on Shell, Intellij and Jupyter Notebook and all necessary labs during the sessions, started making them!

https://docs.cloudera.com/data-warehouse/1.1/managing-warehouses/topics/dw-private-cloud-create-database-on-ozone.html