Access the lab environment:

1. Open https://cloudthat.learnondemand.net
2. Log in using the Microsoft account.
3. Click in Az-400 class
4. Click Launch button under Hands-on Lab to launch virtual machine

Note: You have to perform below step inside the virtual machine launched the previous step.

Apache Spark Setup in Azure Virtual Machine

Tasks:

1. Create an Azure Virtual Machine
2. Download and Install Python, Java, Apache Spark
3. Setup Environment Variable
4. Start Apache Spark in the shell

1. Create an Azure Virtual Machine
    a. Enter virtual machines in the search.
    b. Under Services, select Virtual machines.
    c. In the Virtual machines page, select Create and then Azure virtual machine. The Create a virtual machine page opens.
    d. Under Instance details,
        i. Virtual Machine Name: myVM
        ii. Region: Central India
        iii. Availability options: No Infrastructure dependency required
        iv. Security Type: Standard
        v. Image: Pro Windows 10 Pro, version 22H2 - x64 Gen2
        vi. Leave the other defaults.
    e. Under Administrator account,
        i. Username: vmadmin
        ii. Password: demo!pass123
        iii. Confirm Password: demo!pass123
    f. Under Inbound port rules, choose Allow selected ports and then select RDP (3389).

g. Leave the remaining defaults and then select the **Review + create** button at the bottom of the page.
h. After validation runs, select the **Create** button at the bottom of the page.
i. After deployment is complete, select **Go to resource**

Connect to virtual machine

a. Create a remote desktop connection to the virtual machine.
b. On the overview page for your virtual machine, select the Connect > Connect.
c. In the **Connect with RDP**, keep the default options to connect by IP address, over port 3389, and click **Download RDP file**.
d. Open the downloaded RDP file and **click Connect** when prompted.

2. Download and Install Python, Java, Apache Spark, Git Bash
   a. Python: 3.10
   https://www.python.org/ftp/python/3.10.0/python-3.10.0-amd64.exe

   Do Custom Installation
   Location: c:\python
   create a 'python' folder in c drive

   b. Java: 17.0.10
   https://download.oracle.com/java/17/archive/jdk-17.0.10_windows-x64_bin.msi

   Install in the default location

   c. Apache Spark: 3.5.0
   https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

   You need to extract the downloaded file
   Note: Install Winrar: https://www.rarlab.com/rar/winrar-x64-700b4.exe

   Create a new directory in C drive named it **spark**
   Extract download file into **spark** directory

   d. Download and setup **winutils** for Hadoop

   - **winutils.exe**: https://github.com/cdarlint/winutils/blob/master/hadoop-3.3.5/bin/winutils.exe

- **hadoop.dll**: https://github.com/cdarlint/winutils/blob/master/hadoop-3.3.5/bin/hadoop.dll

create a new **winutils** directory in C drive and inside it create another directory **bin**
Move the downloaded (**winutils.exe, hadoop.dll**) files into **bin** folder

e. Instal Git bash
https://github.com/git-for-windows/git/releases/download/v2.43.0.windows.1/Git-2.43.0-64-bit.exe

Go with the default installation

f. Install VS Code
https://code.visualstudio.com/Download

click windows x64

After VS code install open it, then go to the extension tab and add the following extension

- Jupyter
- Python

Note: when you run pyspark or python it may ask you to install additional package (install it).

3. Setup Environment Variable
   a. **Edit System "PATH" Environment Variable** and add following paths:
      i. C:\python
      ii. C:\python\Scripts
      iii. C:\winutils\bin
      iv. C:\spark\spark-3.5.0-bin-hadoop3\bin
      v. C:\Program Files\Java\jdk-17\bin

Note: Please make sure you have installed Python, Java, Winutils, and Spark as mentioned above otherwise you need to change the Environment Variable Path to those locations

b. **Add new environment variable**s

Name	**: value**
SPARK_HOME	: C:\spark\spark-3.5.0-bin-hadoop3
JAVA_HOME	: C:\Program Files\Java\jdk-17
HADOOP_HOME:	: C:\winutils
PYSPARK_PYTHON	: C:\python\python.exe

c. Install PySpark

open terminal and run:
pip install pyspark

Note: If you are getting an error: pip is not recognized as Enteral or External command. Then you have not setup your environment variable correctly.