DataXper Data Platfrom Sandbox

Student Guide

dataXper

Think Big Value

Legal Notice

This guide is for anyone who is following our course on Big Data and Data Engineering at Scale, and is using our DDP sandbox, and has heard about Apache Hadoop and Apache Spark ecosystem tools and is curious to learn more but keeps getting lost in advanced documentation sites around these tools communities.

We feel you; we hear you and we want to say:

Look no further! Give this a read and we look forward to meeting you in our virtual classrooms in the future!

“The expert at anything was once a beginner.” - Helen Hayes

DDP Student Guide: 1.5
Current DDP version: 3.2.0.2409
Author: Khaled Tannir
Last Published: 2024-10-01

Table of contents

About the Sandbox

The dataXper Data Platform Sandbox is a straightforward, pre-configured, learning environment that contains the latest developments from Apache Hadoop, Spark and many other tools including many sample files and datasets. It allows you to learn and explore DDP on your own.

The DDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Apache Kafka, Trino, Nifi and Superset…

The DDP Sandbox is delivered as a virtual appliance. The virtual appliance (indicated by an .ovf or .ova extension in the filename) runs in the context of a virtual machine (VM), a piece of software that appears to be an application to the underlying (host) operating system (OS), but that looks like a bare machine, including CPU, storage, network adapters, and so forth, to the operating system and applications that run on it.

To use the dataXper Data Platform sandbox, VirtualBox is the only one supported virtual machine application and needs to be installed on your host machine.

The sandbox is an isolated environment where researchers, students and/or data analysts can test and experiment before making changes to production. The sandbox contains some practical tutorials for who want to learn and practice Hadoop / Spark ecosystems and/or implementing Proof of Concepts.

Introduction

The virtual machine is a pre-configured environment for your workshops and hands-on exercises. In preparing a virtual machine (VM) for your workshop, we focused on creating an environment that is fully equipped and ready for immediate use. The virtual machine was meticulously set up with all the necessary tools, libraries, and configurations pre-installed, ensuring that participants can jump straight into the hands-on exercises without any delays.

Virtual Machine Preparation:

1. **Software Installation:**
    - We installed all required software, including Apache Hadoop, Spark, Hive, Trino, Zeppelin, Superset, Cassandra and many other tools and libraries related to the Hadoop and Spark ecosystems. Also, we installed MySQL Server, the Scala compiler SBT, as well as relevant Python libraries.
    - Ensured that each tool is configured correctly, with dependencies properly managed to avoid conflicts or issues during use.
2. **Environment Configuration:**
    - Configured system paths, environment variables, and settings to optimize performance and ensure that all tools work harmoniously together.
    - Set up HDFS (Hadoop Distributed File System) within the VM to allow for seamless data storage and retrieval.
3. **Sample Data and Tutorials:**
    - Pre-loaded the virtual machine with sample datasets, including MySQL sample databases.
    - Included detailed tutorials and documentation within the VM, guiding users on how to use each tool effectively and how to navigate the environment.
4. **User-Friendly Experience:**
    - The virtual machine is designed to be user-friendly, with a familiar operating system interface that most participants will find easy to navigate.
    - We ensured that all tools are accessible from a central location, with shortcuts and scripts provided for common tasks.

Benefits of Using a Virtual Machine:

We decided to offer a virtual machine to our students rather than a Docker script because it comes pre-installed and configured with all the necessary tools. This approach provides a seamless experience, allowing you to dive straight into learning and practicing without the hassle of setup. As explained in the previous section, the virtual machine includes tutorials, sample datasets, MySQL sample databases, a Scala compiler, and essential Python libraries, ensuring that you have everything you need at your fingertips. This setup allows you to focus entirely on honing your skills, exploring

the tools, and working on your projects, free from the distractions of installation and configuration issues.

While Docker offers flexibility and portability, the virtual machine approach was chosen for this course because it provides a more comprehensive and accessible environment. The VM eliminates the need for participants to manage the intricacies of containerization, allowing them to focus solely on learning and using the tools provided. Additionally, the VM environment is closer to a real-world setup, where all tools and services are fully integrated and configured, providing a more realistic and practical experience.

By using the virtual machine, students benefit from a ready-to-use environment where they can start practicing immediately, with everything they need already set up and functioning, minimizing any potential technical barriers. This approach ultimately supports a more effective learning experience, as users can fully immerse themselves in the hands-on exercises without the overhead of managing software installation and environment configuration.

Getting Started with the DDP Sandbox

To get started with the virtual machine provided for the course, follow these steps to ensure you can make the most of the pre-configured environment:

Step 1: Download and Install Virtualization Software

If you haven't already, you'll need to install a virtualization software that can run the virtual machine. We recommend using **VirtualBox** which is free to download and use. The **Part 1 Chapter** will guide you through the instructions to install and configure VirtualBox.

Step 2: Download the Virtual Machine Image

We provided you a link to download the virtual machine image file (usually in `.ova or .ovf` format). This file contains the entire pre-configured environment. You need to download the file to a location on your computer with sufficient storage space.

Step 3: Import the Virtual Machine

After downloading the DDP sandbox you need to import it into VirtualBox and configure it. The **Part 2 Chapter** will guide you to correctly import and configure your DDP sandbox.

Step 4: Start the Virtual Machine

Once the virtual machine has been imported, you'll see it listed in your virtualization software. You will learn how to start and stop it in **Part 3 Chapter**.

Step 5: Explore the Pre-Configured Environment

You'll find shortcuts on the main screen of the sandbox for easy access to the main tools like Apache Hadoop, Saprk, Hive, Trino, Zeppelin, Nifi and Superset. Simply click these icons to launch the respective tools.

Inside the **/home/training/Data** directory, you'll find folders containing sample datasets, such as the weather data, sales, employees, as well the data for the sandbox tutorials to help you get started with each tool.

You can access the sandbox from the external tools to run commands, manage directories and execute Apache Kafka scripts.

Step 6: Practice and Experiment

Hands-On Tutorials: The virtual machine includes tutorials designed to help you practice key concepts. Follow theses tutorials to perform data transformations, query data, and visualize results.

Data Analysis: We use Apache Zeppelin to run Shell, Spark, MySQL, Cassandra, etc … codes. This code is provided as pre-built notebooks that contain example queries and visualizations using Hive and Trino as well Spark and Hadoop. You can modify these notebooks or create your own to analyze any other dataset.

Exploration: Feel free to explore the tools further, experimenting with different configurations, queries, and data analyses.

Step 7: Saving Your Work

Persistent Storage: All changes you make in the virtual machine will be saved within the virtual environment. If you want to back up your work or transfer it to another machine, you can create a snapshot, export your virtual machine or copy files from the VM to your host system.

Step 8: Shutting Down the Virtual Machine

When you're done, you can safely shut down the virtual machine by going to the VM menu and selecting **Shut Down**.

This virtual machine is designed to provide a hassle-free environment where you can focus on learning and experimenting with Big Data tools without worrying about installation and configuration issues. Enjoy exploring the capabilities of the tools and the datasets provided and make the most of your hands-on experience!

Part 1

Preparing the Virtual Environment

This part aims to get you to install, configure and run VirtualBox Manager on your host machine. But before starting to install VirtualBox and importing the DDP Sandbox you need to ensure that your host machine satisfies the requirements.

Prerequisites

- Hardware: A 64-bit machine with a multi-core CPU that supports virtualization. Please investigate your laptop / desktop machine's documentation to verify if you need to enable virtualization and / or hyperthreading.
- Ensure that you can **dedicate** the following resources to the sandbox. Your host machine should have at least **32+ Gb** of physical RAM and **8+ VCPUs** to run the VM.
- The DDP Sandbox requires the following dedicated resources (mandatory).

20 GB	
8+ CPUS	

- To connect to the sandbox instance from a SSH client, you need to install/use an SSH client installed on your machine (e.g `MobaXterm`, PuTTY). We recommend using MobaXterm (not available for MacOS). A full installation guide is provided in the sandbox.

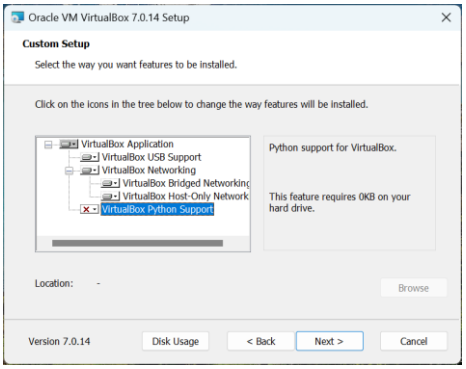
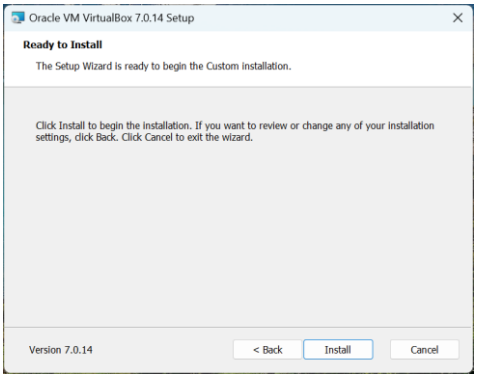
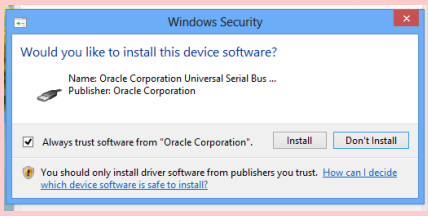
1. Installing VirtualBox Manager

Follow the instructions in this section to install and configure VirtualBox Manager on your machine (Administrator privileges are required).

Go to the link below to download VirtualBox Manager. Choose the version related to your operating system and download the binary related to your host machine. In this guide we will be using the version for Windows operating system hosts.

https://www.virtualbox.org/wiki/Downloads

After downloading, run the executable to begin installing the software. When you start the installation, follow the installation wizard steps:

Click **Next** to begin VirtualBox installation.	
Uncheck the VirtualBox Python Support and click **Next.**	
During the installation wizard, you'll get a Network Interfaces warning: Click **Yes** to proceed	
Click on **Install** to proceed.	
If prompted with a message to install (Trust) Oracle Universal Serial Bus, Click **Install** to continue and continue with the wizard until you're done.	

Note: If you're in the process of downloading or copying files and data, this will interrupt your network connection briefly. So, maybe pause the process or wait until you're done before installing VirtualBox.	
Click on Finish to close the installation wizard.	

After that, VirtualBox should be installed. When you're done installing, you are ready to install VirtualBox Extension pack.

2. Installing VirtualBox Extension Pack.

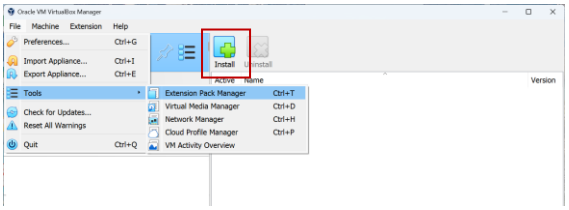
The extension pack extends the functionality of VirtualBox base packages. It provides the following enhancements to VirtualBox:

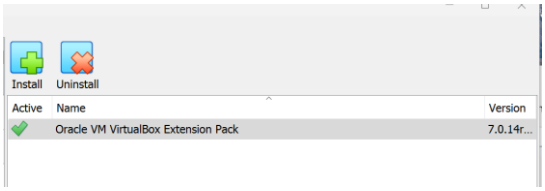
- Virtual USB 2.0 (EHCI) device.
- Virtual USB 3.0 (xHCI) device.
- VirtualBox Remote Desktop Protocol (VRDP) support.
- Host webcam pass-through.
- Intel PXE boot ROM.
- Experimental support for PCI pass-through on Linux hosts.
- Disk image encryption with AES algorithm.

To install the extensions pack, go back to VirtualBox's download page:

VirtualBox Extension Pack

Download and save the current pack for **all supported platforms.** After downloading and saving the extensions pack, follow these steps to install it.

Open VirtualBox Manager and go to: **File → Tools → Extension Pack Manager** Click on **Install.** Then find the downloaded extension pack.	

Click on **Install** Then Scroll down and Click on **I Agree**	
The VirtualBox Extension Pack is now installed	

When you're done, VirtualBox software will be ready to be used and install guest(s) operating systems.

Part 2

Installing and Configuring the Sandbox

This second part of the guide aims to get you to install, configure the Virtual Machine (DDP sandbox). This virtual machine includes a fully Hadoop / Spark Single-Node Cluster and all the tools needed in the course and making it quick and easy to get started with Data at Scale course workshops. Please follow the instructions in this section to download, install and configure the DDP sandbox.

Prerequisites:

- VirtualBox Manager + Extensions pack already installed.
- Having enough resources to run the VM.

Host Machine	Dedicated to the sandbox
32+ GB RAM	20 GB RAM
20+ VCPUs	10+ VCPUs
100+ GB Free disk space	100 GB disk space

1. Downloading the DDP Sandbox

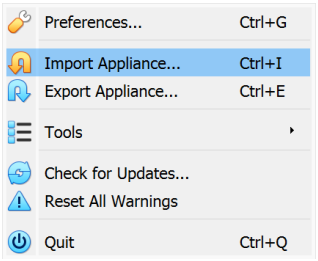
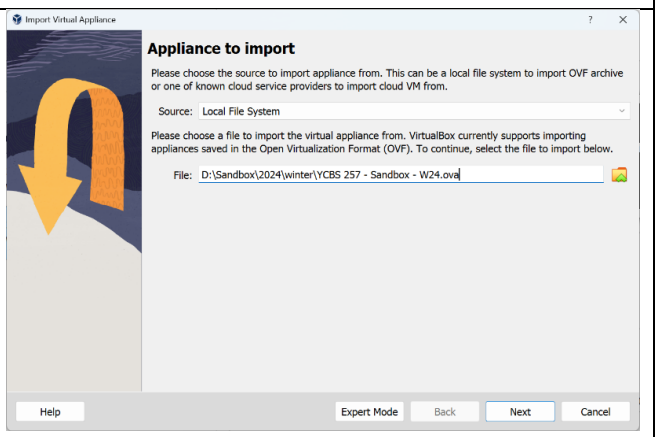
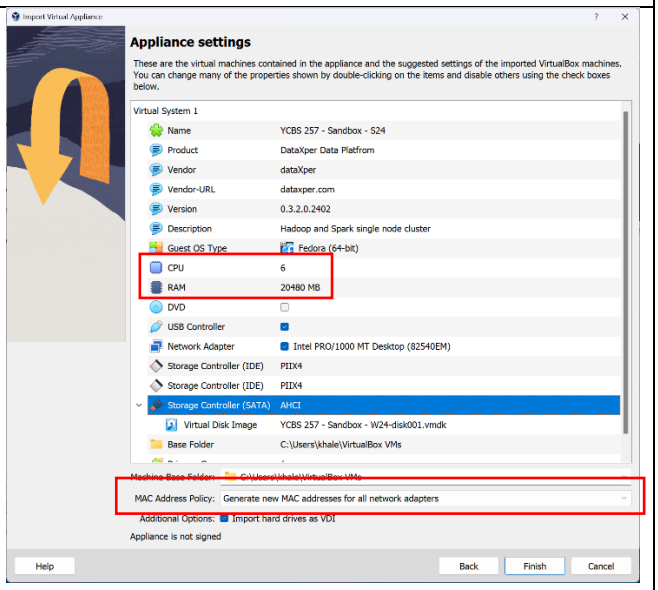
Step 1. **Downloading the sandbox (.ova file) for VirtualBox.**

a) You can download the VM (~22 Gb) from this link:

DDP Sandbox - F24

2. Importing into VirtualBox

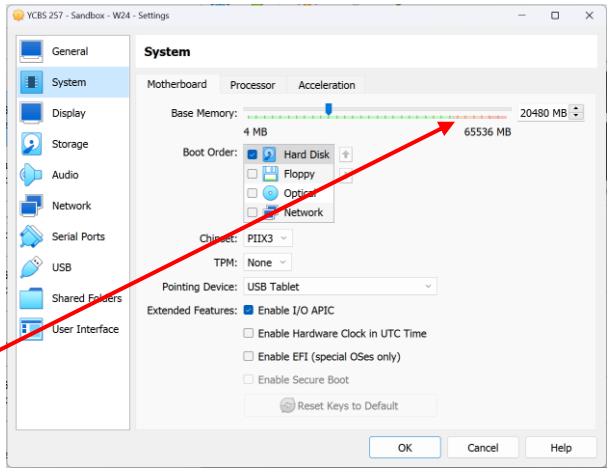
Step 2. **Importing the sandbox in VirtualBox.**

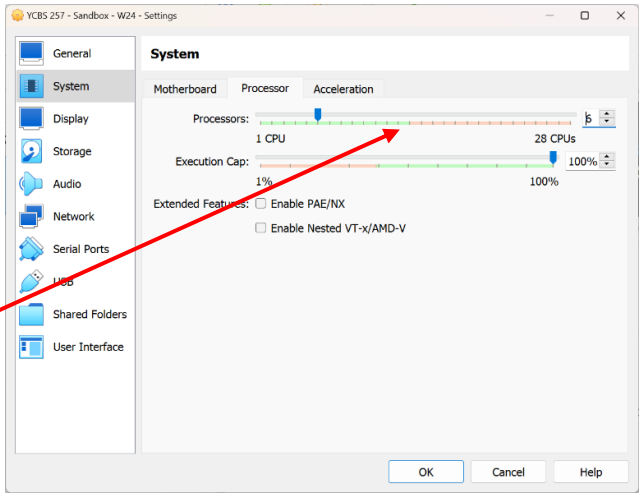
<table>
<tr>
<td>a) Launch VirtualBox.
b) Click on File – Import Appliance</td>
<td> 
</td>
</tr>
<tr>
<td>c) Browse the location where you downloaded the VM.</td>
<td> 
</td>
</tr>
<tr>
<td>d) You might need to adjust the allocated resources based on your machine capacity.
Default values:
6-8 CPUs, 20480 MB RAM
e) Choose to ‘Generate new MAC address for all network adapters’.
f) Click Finish.</td>
<td> 
</td>
</tr>
</table>

3. Configuring The DDP Sandbox

Please Check the VM allocated resources: Memory amount and VCPUs cores.

<table>
<tr>
<td>a) From Virtualbox Manager select the VM.
b) Click on Settings -> System to Adjust Base Memory and Processor CPUs.
c) Set the Base Memory as: 20480 Gb. (Minimal value: 20480 Gb)

When settings your Base Memory resources DO NOT EXCEED the green area.</td>
<td> 
</td>
</tr>
<tr>
<td>d) Set the Processors as: 6 VCPUs Cores.
Minimal value: 6 CPUs
Default Value: 8 CPUs

When settings your VCPUs DO NOT EXCEED the green area.</td>
<td> 
</td>
</tr>
</table>

4. Taking Initial Snapshot of the Sandbox

Before starting your VM, you need to take a Snapshot. (Very Important)

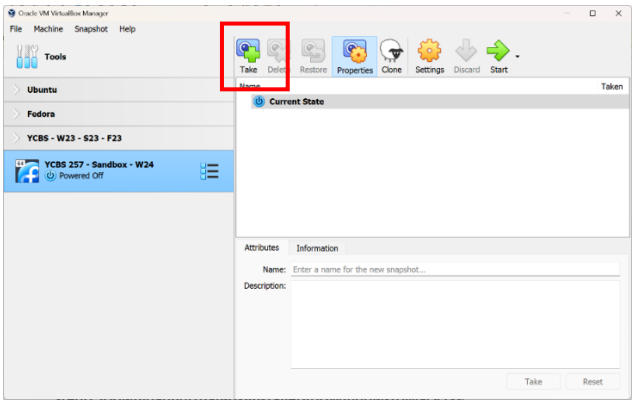
With snapshots, you can save a particular state of a virtual machine for later use. At any later time, you can revert to that state, even though you may have changed the VM considerably since then. A snapshot of a virtual machine is thus like a machine in Saved state, but there can be many of them, and these saved states are preserved.

Taking a snapshot makes a copy of the machine's current state, to which you can revert at any given time later.

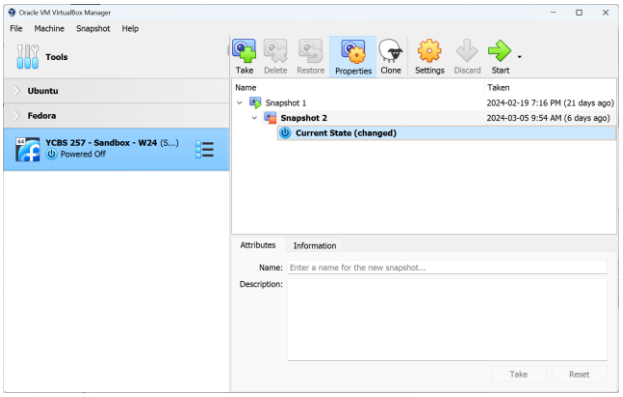
<table>
<tr>
<td>a. Select the VM name in the Oracle VM VirtualBox main window.

b. Click the List icon next to the machine name.

c. Select Snapshots. The snapshots window is shown.

d. Click the Take icon.</td>
<td> 
</td>
</tr>
<tr>
<td>To see the snapshots of a virtual machine:

• Click on the machine name in VirtualBox Manager.
• Then click the List icon next to the machine name, and select Snapshots.

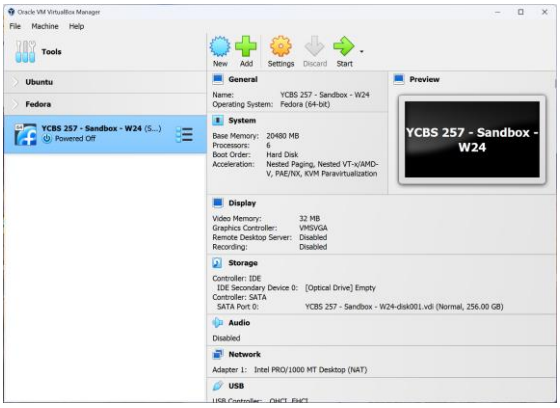
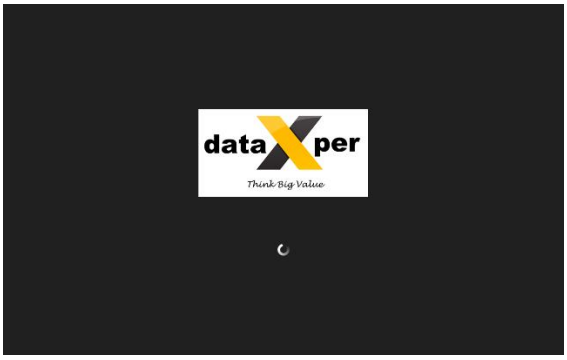
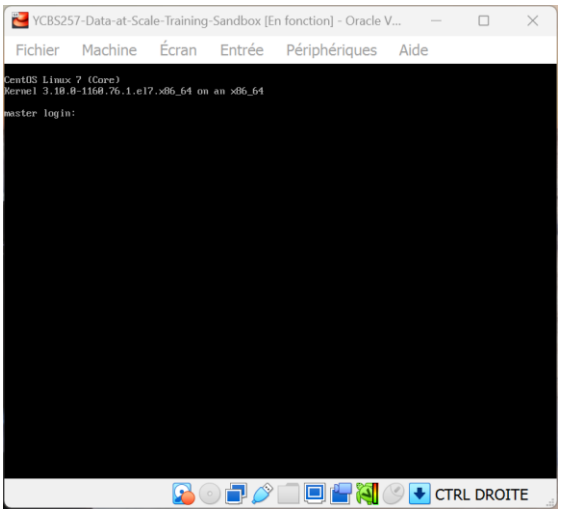
Until you take a snapshot of the machine, the list of snapshots will be empty except for the Current State item, which represents the “now” point in the lifetime of the virtual machine.</td>
<td> 
</td>
</tr>
</table>

Part 3

Working with the DDP Sandbox

In this part you will learn how to manage (start, stop and save the state of the sandbox).

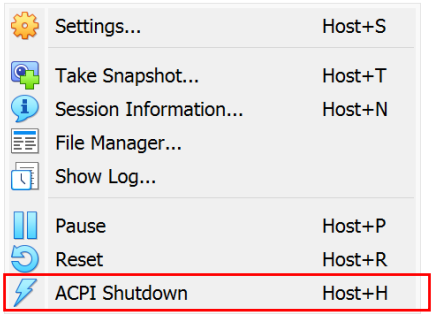
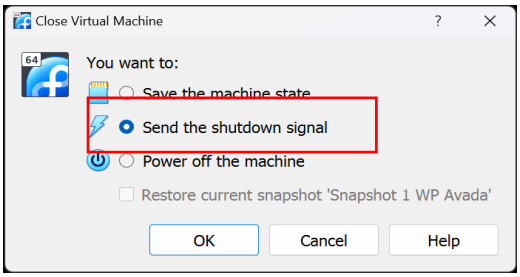
1. Starting the sandbox.

a) Select the sandbox from VirtualBox Manager on the left panel. b) Click the **Start** (green arrow) icon.	
The Sandbox is loading	
A similar window should open after finishing loading and the VM starts running. **From here, there is nothing else to do.** You can minimize this window. You will interact with your VM from your **browser** and your **SSH client**.	

Please allow 5 to 8 minutes before all the Hadoop / Spark Services are up and running.

2. Shutting down the sandbox.

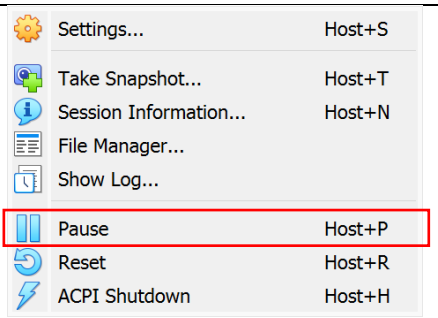
Once you have finished using your VM, you can shutdown it by sending the **ACPI** shutdown signal.

a) Switch the sandbox running window. b) Go to **Machine -> ACPI Shutdown.**	
You can also choose the Send the shutdown signal when you close the VM widows.	

Note: *It is important to always shut down the VM properly. Otherwise, invalid, or corrupted blocks might appear in HDFS which might make the sandbox unusable*.

3. Pausing the Sandbox

You can pause the VM if you need to interrupt your work for a short time.

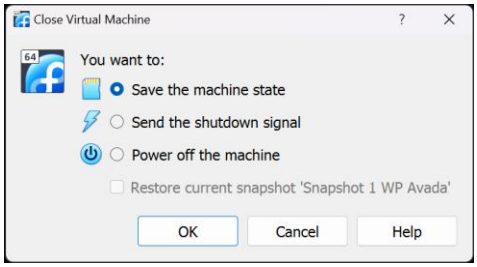
c) Switch the sandbox running window. d) Go to **Machine -> Pause.**	

The pause command temporarily stops the execution of the VM. When paused, the VM's state is not permanently changed.

The VM window appears gray, and the title bar of the window indicates that the VM is currently Paused.

4. Saving the Sandbox State

While working with your VM and if you need to interrupt or pause your work and come back later, you can then save the state of the VM and restore it back later.



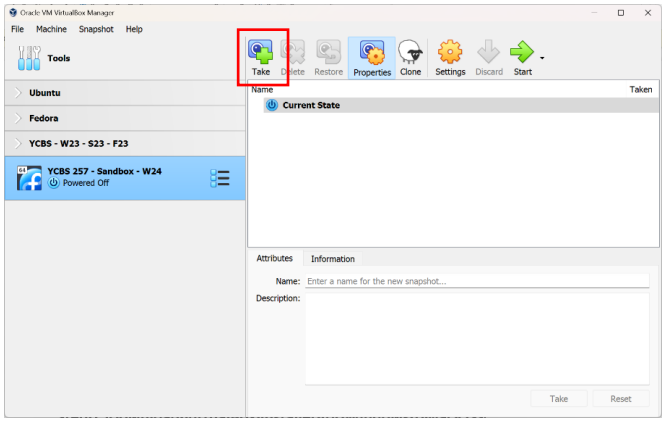
5. Take a Snapshot

With snapshots, you can save a particular state of a virtual machine for later use. At any later time, you can revert to that state, even though you may have changed the VM considerably since then. A snapshot of a virtual machine is thus like a machine in Saved state, but there can be many of them, and these saved states are preserved.

Taking a snapshot makes a copy of the machine's current state, to which you can go back at any given time later.

a) Select the VM name in the VirtualBox main window, click the **List** icon next to the machine name and select **Snapshots**. The snapshots window is shown.
Do one of the following:

b) Click the **Take** icon.
OR
b) Right-click on the **Current State** item in the list and select **Take**.

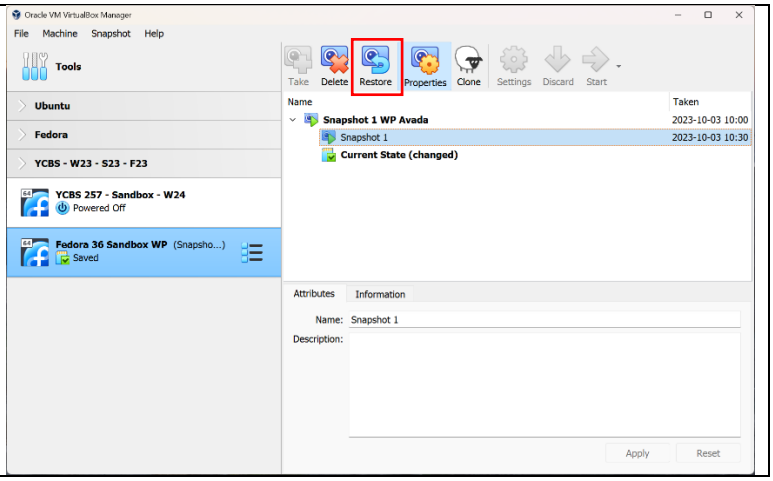


6. Restore a Snapshot

By restoring a snapshot, you go back or forward in time. The current state of the machine **is lost**, and the machine is restored to the exact state it was in when the snapshot was taken.

To Restore a snapshot

a) Power off your VM
b) From the list of snapshots, right-click on any snapshot you have taken
c) Select **Restore**.



Note: Restoring a snapshot will affect the virtual hard drives that are connected to your VM, as the entire state of the virtual hard drive will be reverted as well. This means also that all files that have been created since the snapshot and all other file changes *will be lost*.

Part 4

Exploring the DDP Sandbox

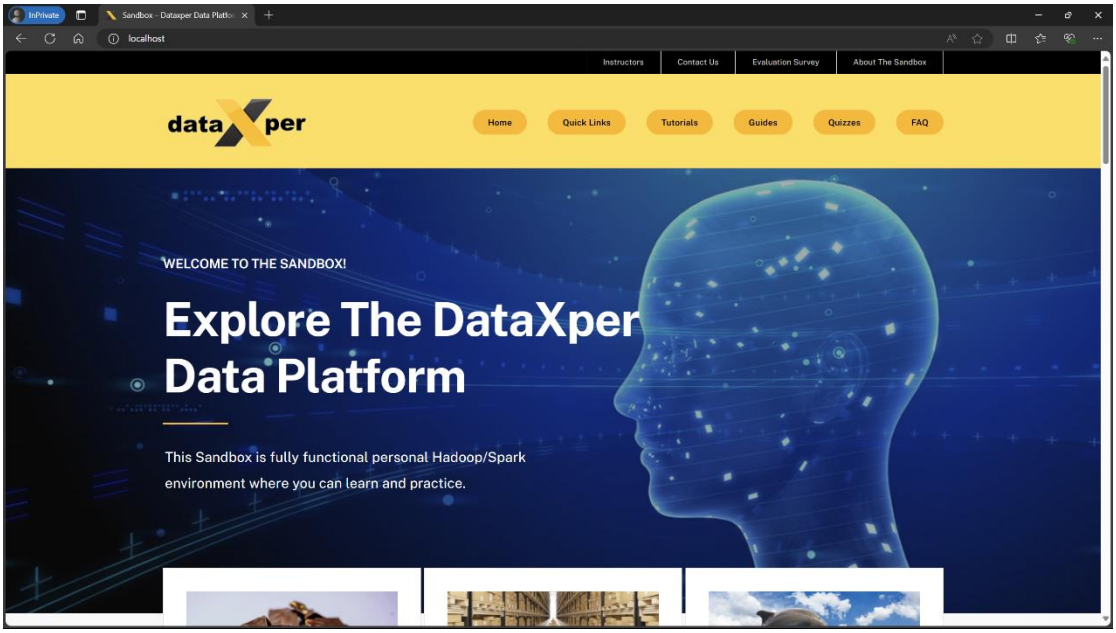
This part aims to get you to be familiar with the sandbox and help you to explore it. This virtual machine includes many **Tutorials**, integration **Guides** and **Quizzes** to help you to be more comfortable with the Hadoop and Spark concepts covered in the course. Also, there is a few tutorials to help you getting some knowledge about the SQL language, Scala language, Linux and many others.

1. Sandbox Home Page.

Once your VM is up and running, open your browser and enter the following URL to show the VM Home page.

http://localhost

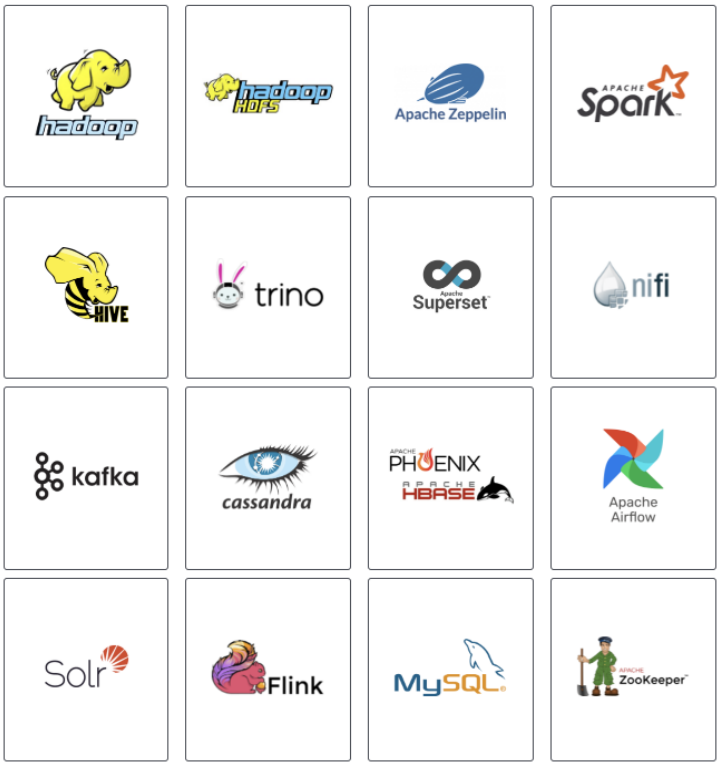
Note: In case you get a security warning, you can just continue to the local site and ignore the warning.



2. Sandbox Quick Links Page.

The **Quick Links** page give you an access to all the tools installed in the sandbox.

- Go over any box with your mouse and it will flip to show you the links and credentials (when needed).
- Clicking on the link will open the corresponding web page for the tool in a new tab in your browser.

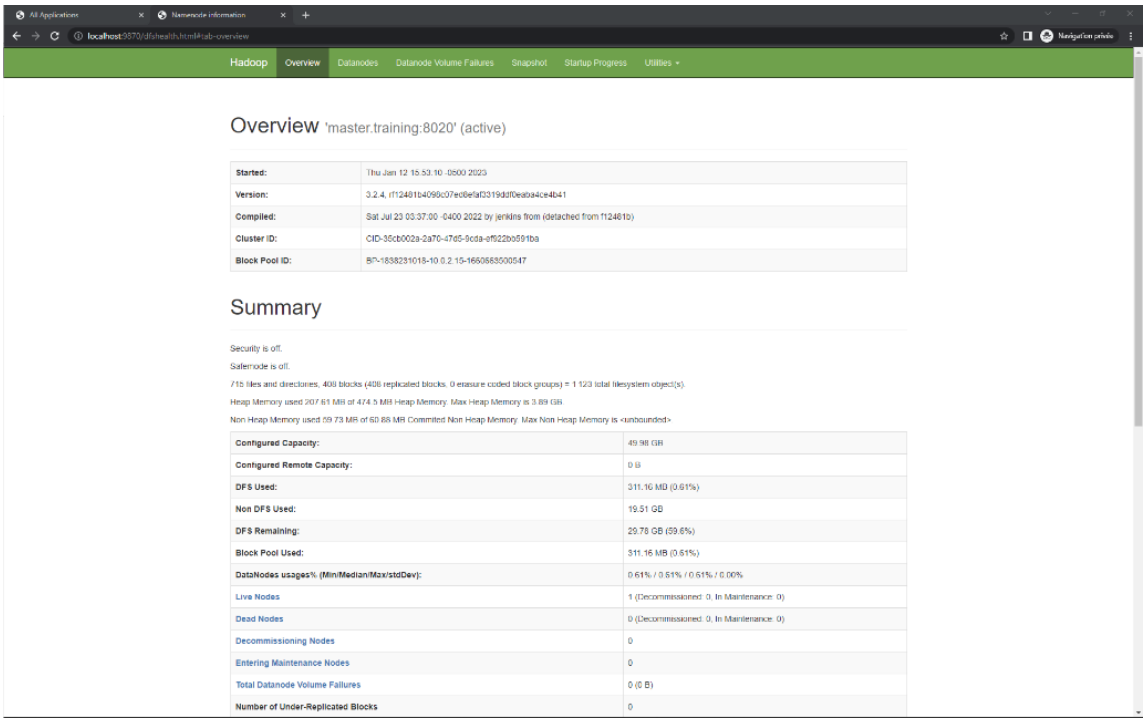


Try This by Your Self:

Open Hadoop NameNode Page and start exploring the NameNode UI.

• Go over the Hadoop image. • Click on NameNode UI url: http://localhost:9870 • To keep the flip box open, click on the mango area. It will flip back to the initial state once you click anywhere on the page.	

Your browser will open a new tab, and you should see a screen like the screen below.



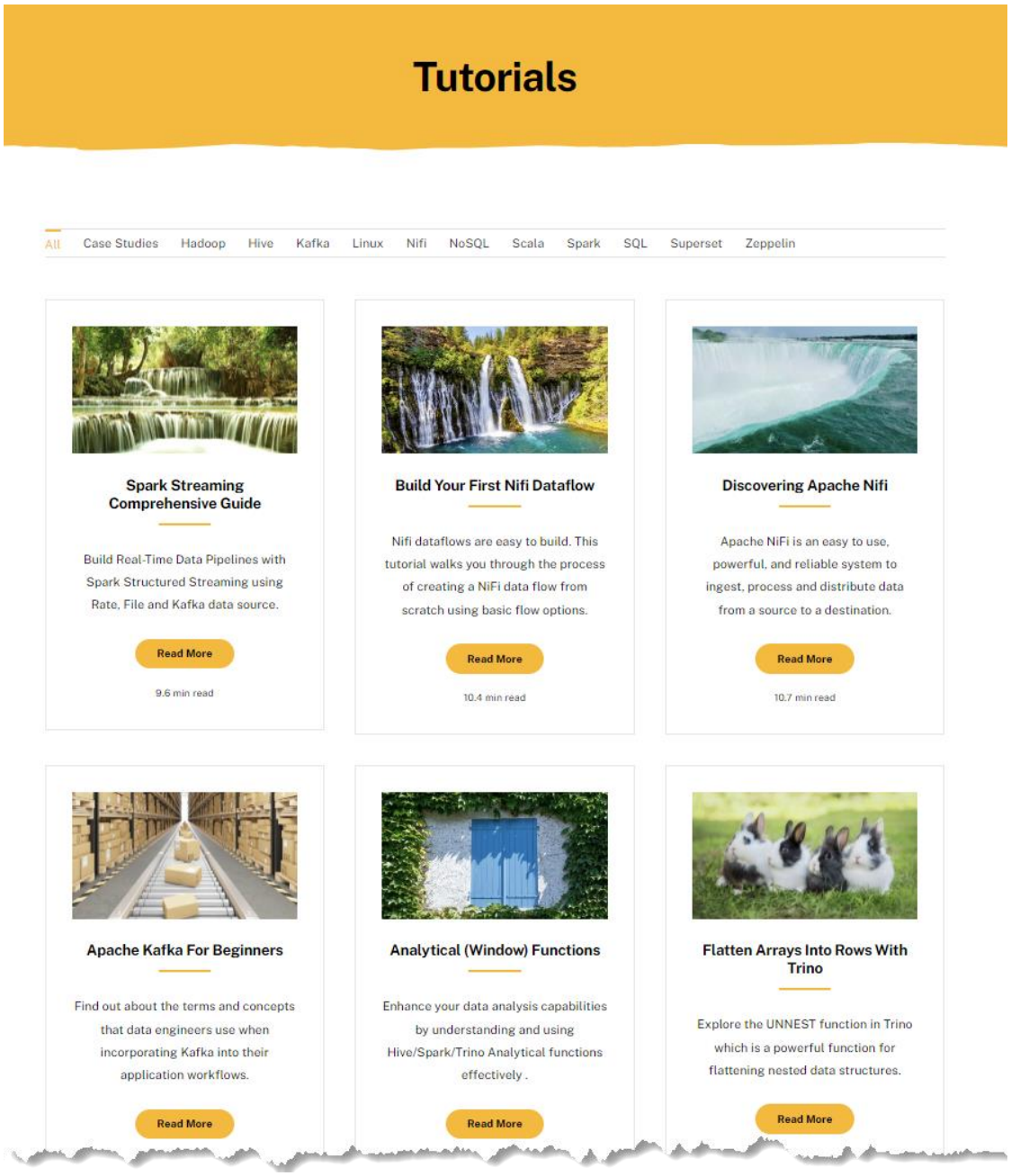
Notice that the `NameNode UI` contains a lot of information about the cluster.

The Overview page shows the version of Hadoop and other details such the storage capacity in the Summary section.

3. Tutorials Page.

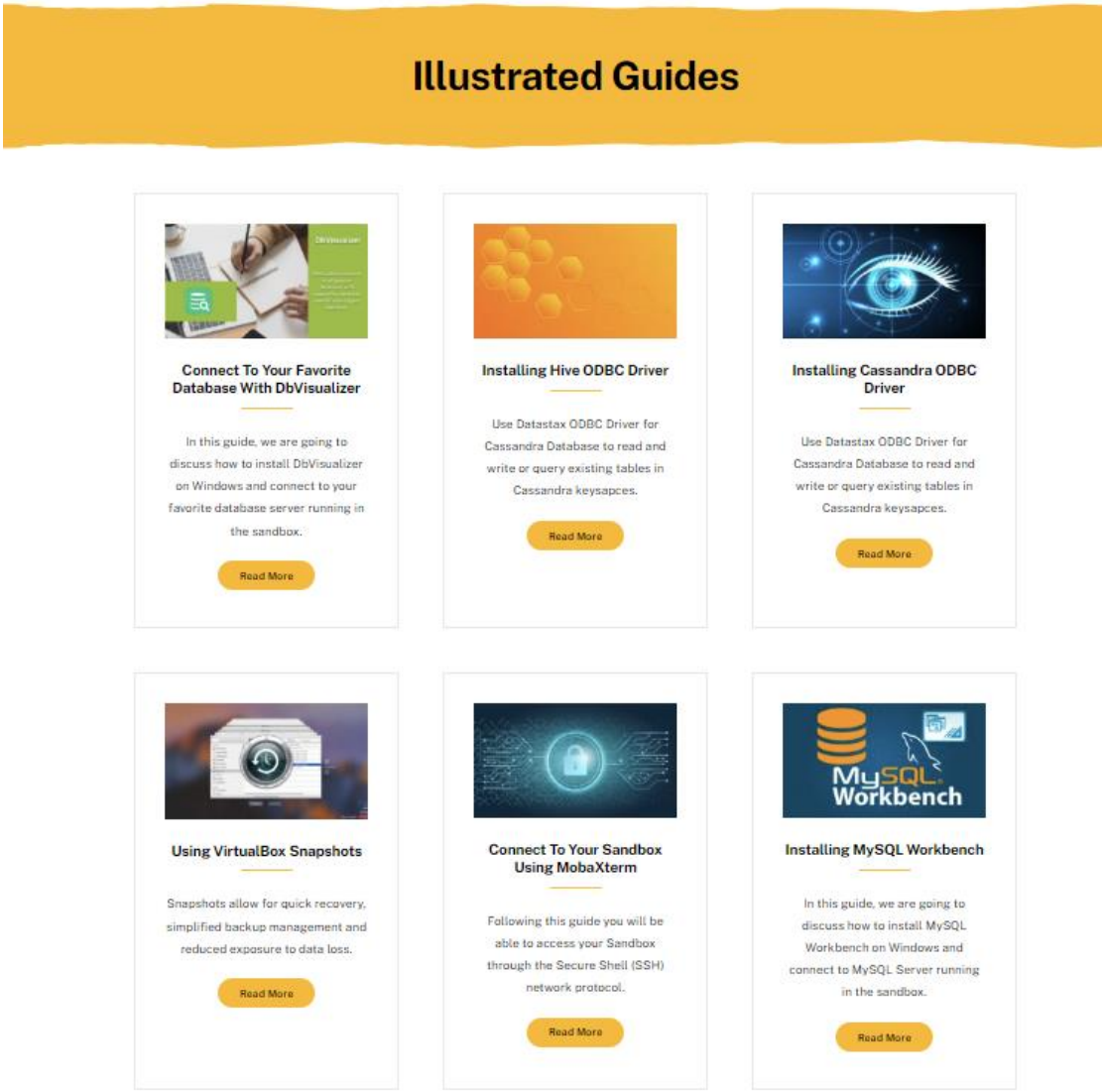
The **Tutorials** page contains more than 40 tutorials about Hadoop and Spark and many other tools such as Scala, Hive, Trino and MySQL, etc…

Also, there are two Case Studies to get more practice with data engineering using Hadoop, Spark and Hive.



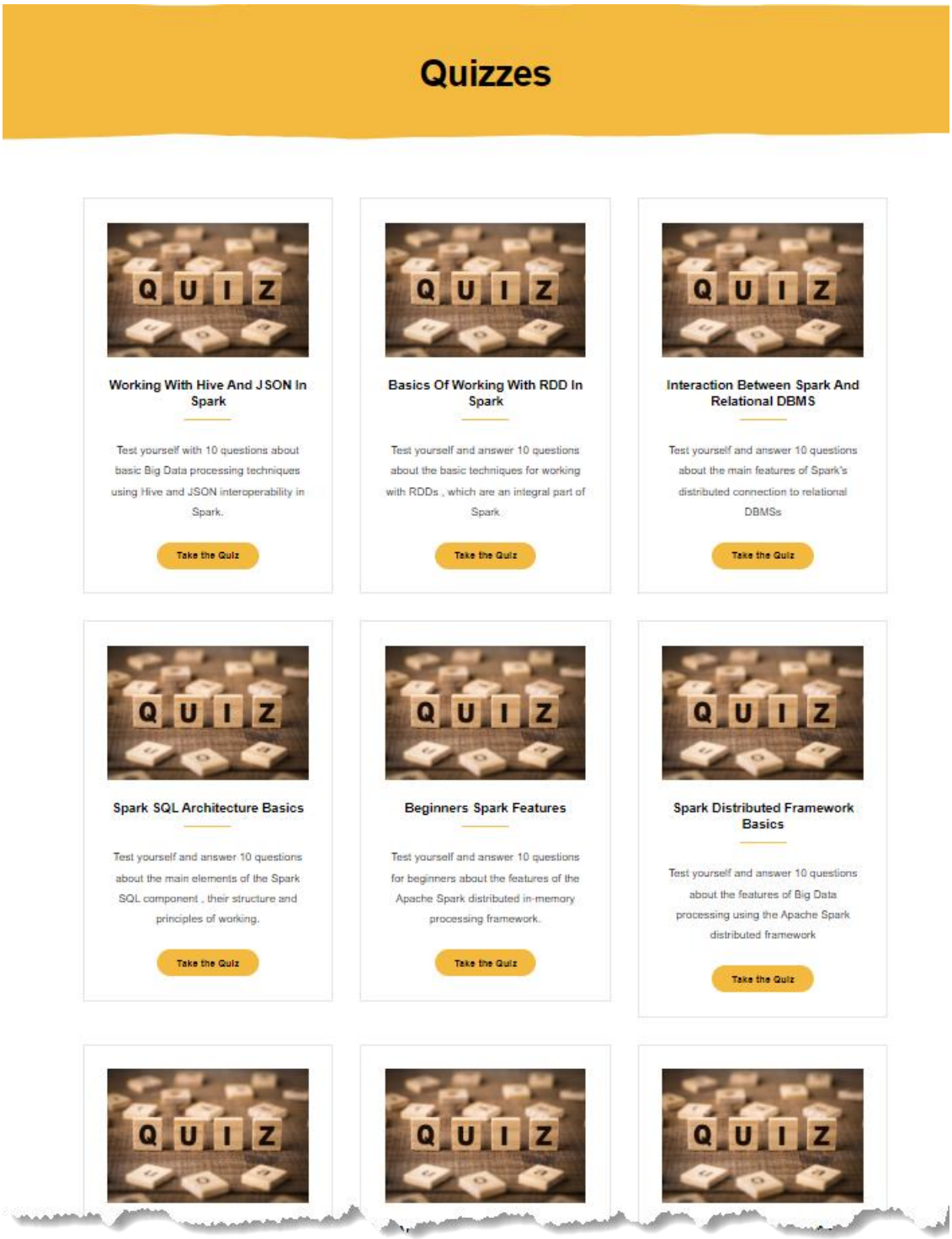
4. Illustrated Guides Page.

The Illustrated Guides page contains instructions about how to connect externals tools to the sandbox such as ODBC drivers and databases management and visualization tools.



5. Quizzes Page.

The Quizzes page contains 10+ quizzes to test your knowledge about the tools you are learning. These quizzes are not graded, and you can take them as many times you want and at any time.



Part 5

Connecting Using SSH

SSH, or Secure Socket Shell, is a protocol which allows you to connect securely to a remote computer or a server by using a text-based interface.

An SSH client is a program that allows establishing secure and authenticated SSH connections to SSH servers. SSH client software is available for major enterprise environment operating systems, such as Unix variations, Microsoft Windows and Mac OS.

In this part, we are going to discuss how to configure MobaXterm SSH client to connect to the sandbox where to find the sample data files.

1. SSH Clients for Windows

There are several SSH clients available for Windows. Some of the most popular are:

MobaXterm (recommended) is an SSH client for Windows. It provides all the important remote network tools (SSH, X11, RDP, VNC, FTP, MOSH, ...) and Unix commands (bash, ls, cat, sed, grep, awk, rsync, ...) to Windows desktop, in an installer or single portable exe file which works out of the box. The Home Edition can be downloaded for free.

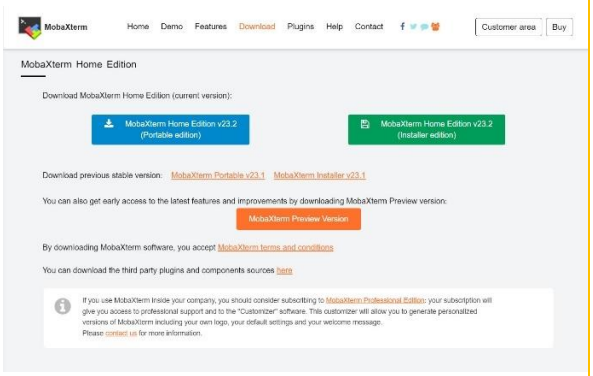
PuTTY is a free client for the SSH and telnet protocols.

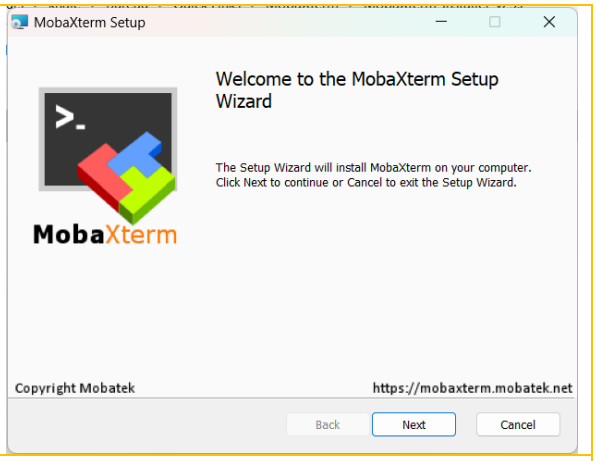
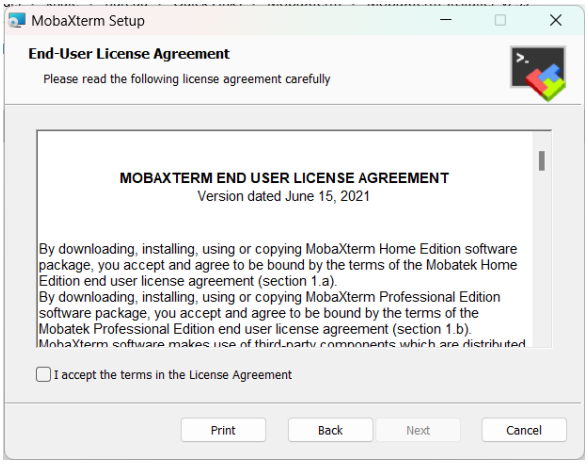
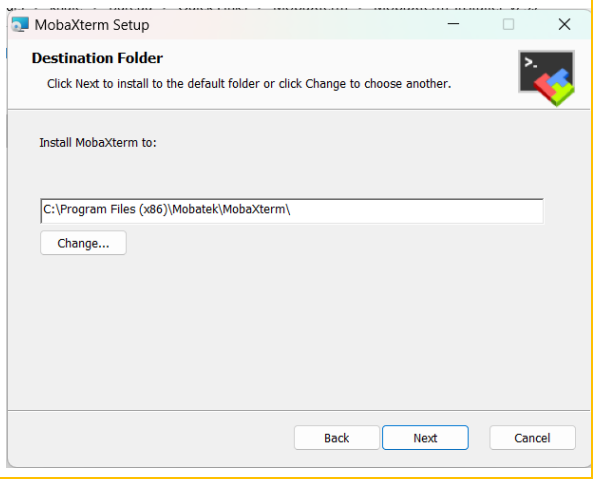
Chrome SSH extension - The Google Chrome browser can be turned into an SSH client with an extension available in the Chrome Web Store. Chrome SSH (beta) offers a basic SSH protocol capability.

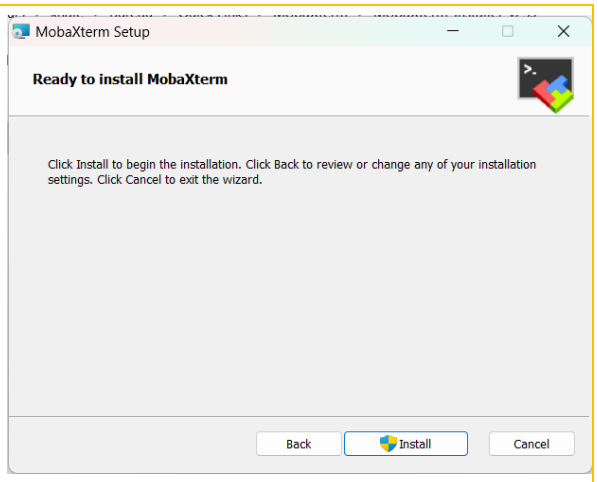
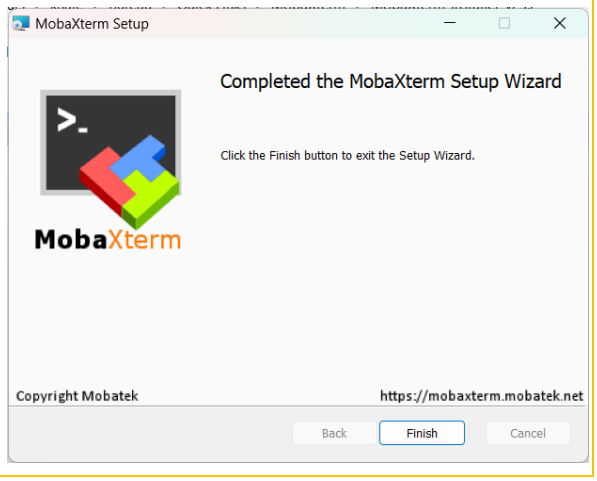
In this part you will be downloading and configuring MobaXterm Home Edition to connect to the sandbox.

2. Download and Install MobaXterm

The software that we recommend is called **MobaXterm**, and this can be obtained from https://mobaxterm.mobatek.net/download-home-edition.html. Download the Installer edition zip archive, open it, and run the MSI installer found there. Then follow the Setup Wizard as usual. This will typically install an icon on your desktop (as well as an entry in the Start Menu).

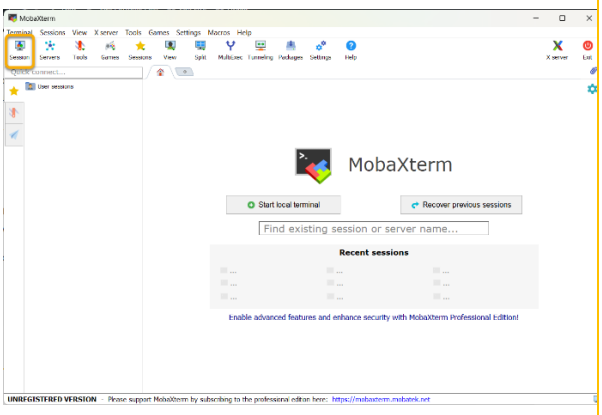


2. After the download, open the installer. • It will ask for permission; when it does, click **Yes**. The installer will then open. • Once you start the program, the window shown below will appear. Click on **Next**	
3. Accept the license terms and Click on **Next.**	
4. Click on **Next** again	

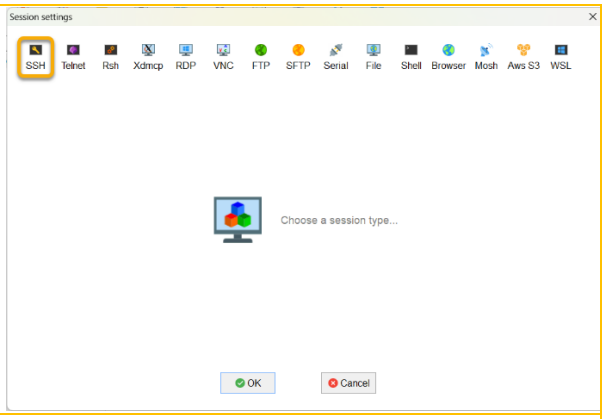
5. Click on **Install**	
6. When the installation is done, Click on **Finish.**	

3. Configure MobaXterm

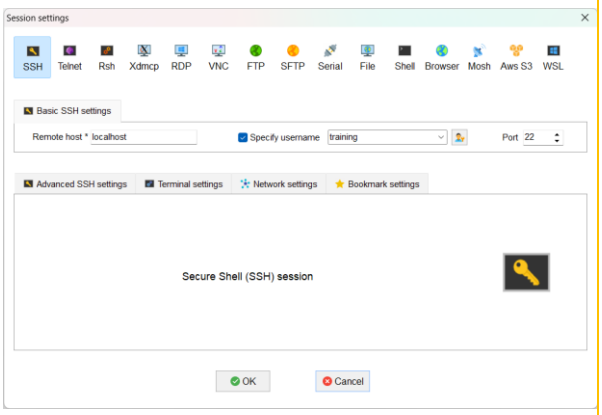
The following steps illustrate how to configure MobaXterm to connect to the sandbox and can use it to move, copy, paste, and delete files to and from your sandbox.

7. Once MobaXterm is installed, open it and the following screen will appear. • Click on the **Session** button in the toolbar to create a new session setting.	

<table>
<tr>
<td>8. a new window will appear.

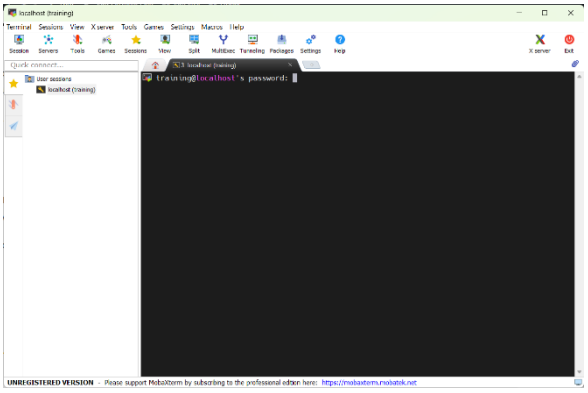
• Click on the SSH button in the toolbar to create a new session setting.</td>
<td> 
</td>
</tr>
<tr>
<td>9. The Session Settings window will open. Set the connection parameters as following:

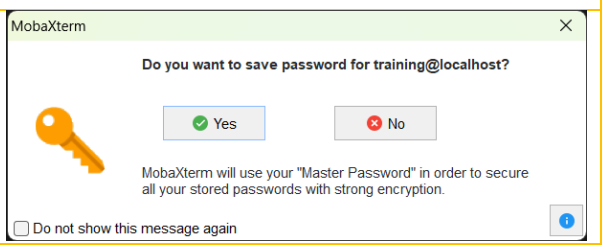
• Remote host: localhost
• Check 'Specify username': training
• Port: 22

Then click on "OK" button to validate the settings and start the SSH session.</td>
<td> 
</td>
</tr>
<tr>
<td>10. MobaXterm will connect you to the sandbox and you will be asked to enter the training's password.

• Password: training

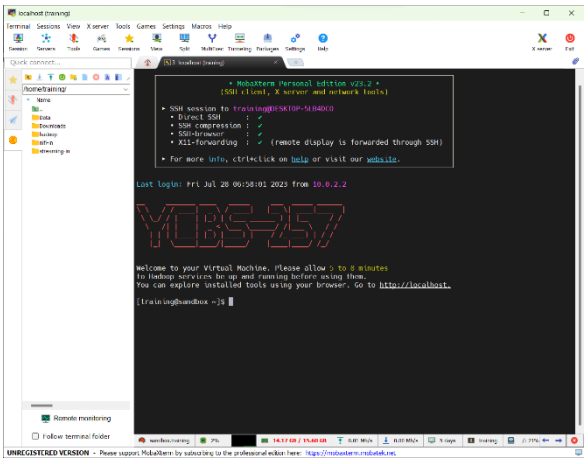
Enter the password and hit Enter.

Note: Nothing will appear on the screen while typing the password.</td>
<td> 
</td>
</tr>
<tr>
<td>11. You will be prompted to store the password in the local vault.

• Click on: Yes</td>
<td> 
</td>
</tr>
</table>

12. You are now logged in.

- The panel on the left (white background in the image at right) shows the files and directories and can be used for drag-and-drop file transfers (upload or download).
- The panel on the right (black background) is a terminal session, which can be used to enter commands.

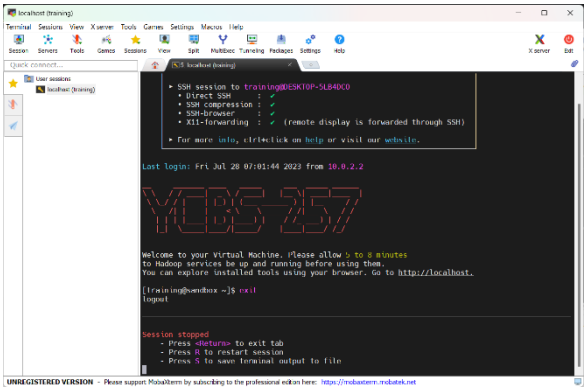


4. Close / Re-Open SSH Session

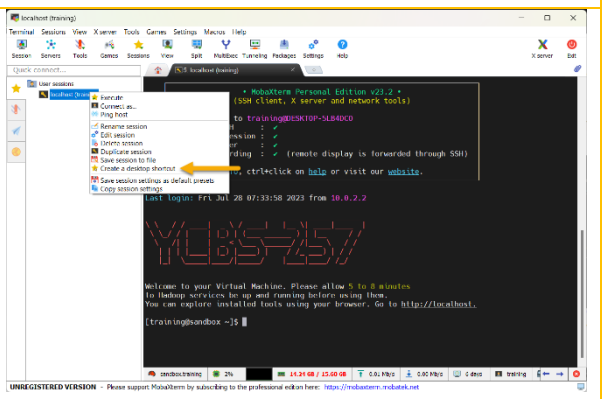
Closing an open SSH session and re-opening it is very easy. The following illustrations show how to do it.

To disconnect from the sandbox type **exit** or close the terminal tab.

You will find on the left sidebar (in the Sessions tab) a shortcut to the session you just setup. From now on, when you open MobaXterm, you can just double click that shortcut, and you will start a SSH session on the sandbox (same that you used in previous steps).

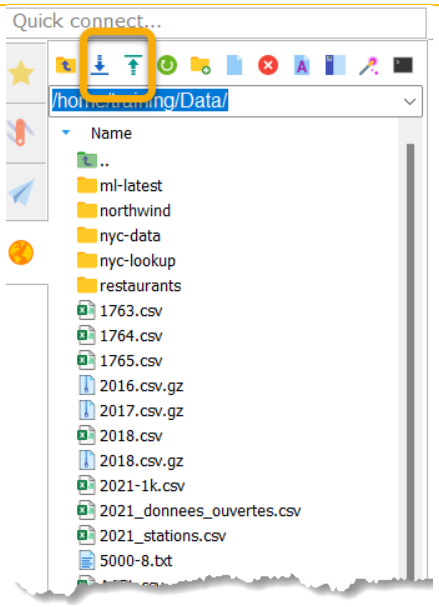


To create a direct shortcut on your desktop (optional), right click on the saved session name and choose *Create a desktop shortcut* (see image). An icon will appear on your Desktop that will start MobaXterm and open a session in the sandbox.



5. Copying files to and from the sandbox

Once you've connected to the sandbox, you will see on the left sidebar (in the *Sftp* tab) a file browser on the sandbox.
You can simply drag and drop files from your computer to that panel and they will be copied to the sandbox. Or you can drag and drop files from the sandbox to your computer.
Alternatively, you can use the file tools located at the top of the file browser (Up and Down arrows).
Remember to **always** press the Refresh folder button **after** you copied something or created/removed a file or folder on the sandbox.



Note: *If the Explorer Panel doesn't appear after opening the SSH session, please check the Troubleshooting section to fix it.*

This guide is also available in the sandbox.

Goto this link to open it.

http://localhost/guides/connect-to-your-sandbox-using-mobaxterm/

6. Accessing the Sample Data files

All the sample data files used in the training course (and more 😊) can be found in your sandbox.
Just SSH the sandbox, go to your home directory and navigate to the `Data` directory:

```
$ cd /home/training/Data
```

7. Connecting External tools

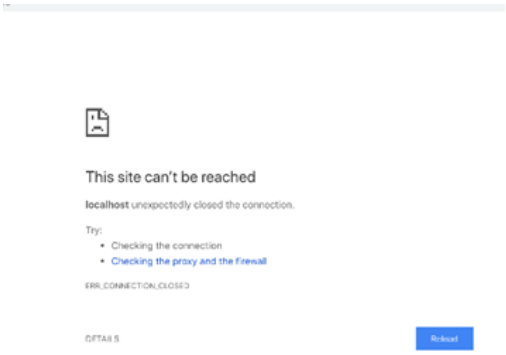
The sandbox supports connecting some external tools. To connect an external tool to your sandbox, you need to first check the Guides page to know about the tools you can connect to the sandbox and then follow the guide step by step.

Appendix

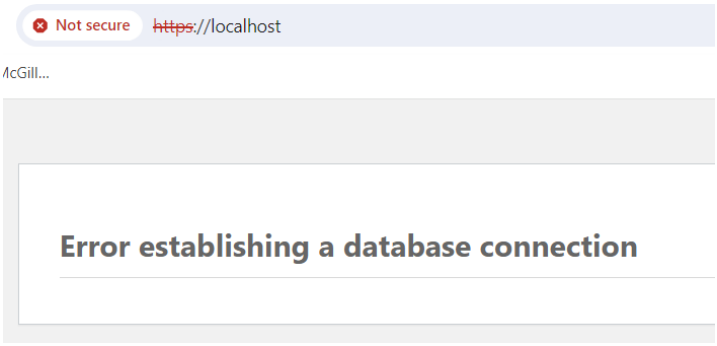
Troubleshooting

You might be facing issues launching some tools or disconnections, many factors might impact the sandbox and causing these interruptions (memory, network, etc..). If you are facing the same issues this troubleshooting guide might help you solve the problem(s).

This section contains troubleshooting information to help you identify, isolate, and resolve component and system issues.

Problem
I am not able to access the localhost web page. 
Solution • Check the VM required resources. You should allocate at least: 20 GB RAM, 6 CPUS and 50+ GB disk space. • Verify that the sandbox is loaded and is not paused. • Verify that VirtualBox Network Interface is connected to your host machine and is not blocked by a local firewall.

Problem
The Sandbox Localhost welcome page is not showing. Instead, you got the following error.

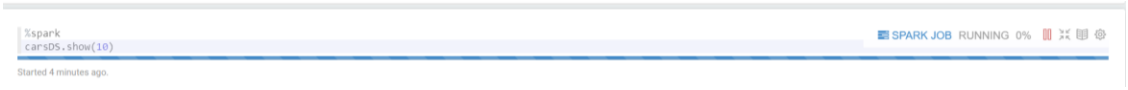


Solution

The sandbox is still **loading**. Please wait a few minutes and retry.

Problem

Spark task doesn't start (Pending) or doesn't finish (Running):



Solution

This happens when Spark is waiting for resources allocation from Yarn. You need to free the Yarn/Spark queue: (no more memory room available). You can try one of the following solutions:

1- From Zeppelin:
   - Stop the current paragraph.
   - Run in a new paragraph:

     %spark
     sc.stop

     Then re-run your zeppelin note.

2- Restart the Sandbox

Problem

Spark task cannot run. I am getting a similar error message:

```
java.lang.IllegalStateException: LiveListenerBus is stopped.
  at org.apache.spark.scheduler.LiveListenerBus.addToQueue(LiveListenerBus.scala:97)
  at org.apache.spark.scheduler.LiveListenerBus.addToStatusQueue(LiveListenerBus.scala:80)
  at org.apache.spark.sql.internal.SharedState.<init>(SharedState.scala:115)
  at org.apache.spark.sql.SparkSession.$anonfun$sharedState$1(SparkSession.scala:139)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.sql.SparkSession.sharedState$lzycompute(SparkSession.scala:139)
  at org.apache.spark.sql.SparkSession.sharedState(SparkSession.scala:138)
  at org.apache.spark.sql.SparkSession.$anonfun$sessionState$2(SparkSession.scala:158)
  at scala.Option.getOrElse(Option.scala:189)
  at org.apache.spark.sql.SparkSession.sessionState$lzycompute(SparkSession.scala:156)
  at org.apache.spark.sql.SparkSession.sessionState(SparkSession.scala:153)
  at org.apache.spark.sql.DataFrameReader.<init>(DataFrameReader.scala:732)
  at org.apache.spark.sql.SparkSession.read(SparkSession.scala:658)
  ... 46 elided
```

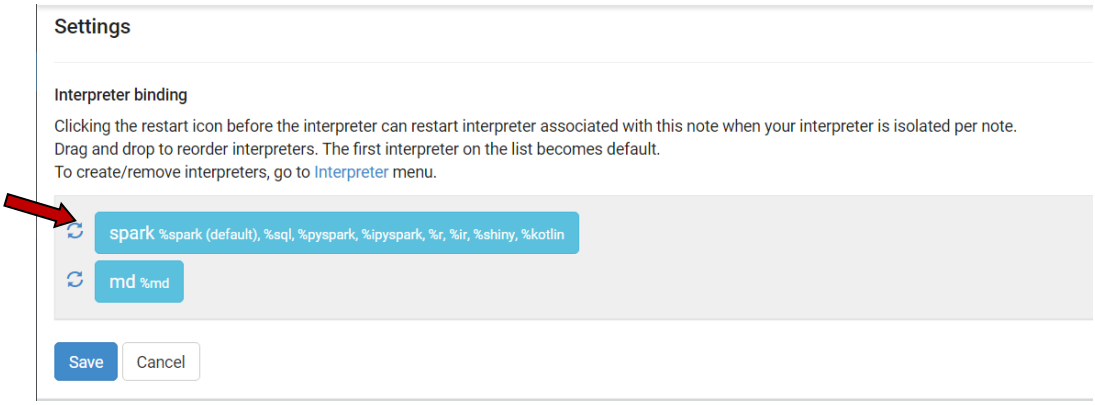
Solution

This message is raised when you try to run a Spark code, and the Spark context has been stopped (either manually or after a long time of inactivity).

You need to restart the Zeppelin Spark Interpreter.

From Zeppelin:

- Navigate to your note.
- Click on “Interpreter binding”
- Select the Spark interpreter by clicking on the interpreter’s name.
- Click on the double arrow to restart the interpreter.
- Click on Save when finish.
- Then re-run your zeppelin note.



Problem

When opening my SSH client I see a similar message:

```
[root@sandbox ~]#
Message from syslogd@sandbox at May  2 16:01:44 ...
 kernel:watchdog: BUG: soft lockup - CPU#3 stuck for 573s! [swapper/3:0]
```

Solution

This message appears when you are not using the VM and your host machine goes into standby mode.

Just hit ENTER and you will get the prompt message.

Problem

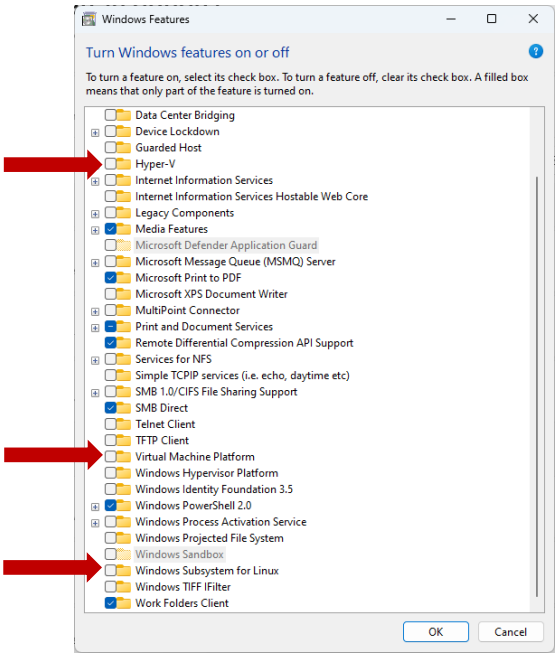
I am running on MS Windows host machine and my Sandbox is running very slow.

Solution

This slowness appears when some MS Windows features are installed, and you need to remove them (requires admin privileges).

Open the MS Windows features dialog box and check the following features if any are installed. If it is the case, you need to uninstall them: *(restart your Windows machine after uninstalling these features)*

1. Hyper-V
2. Virtual Machine Platform
3. Windows Subsystem for Linux

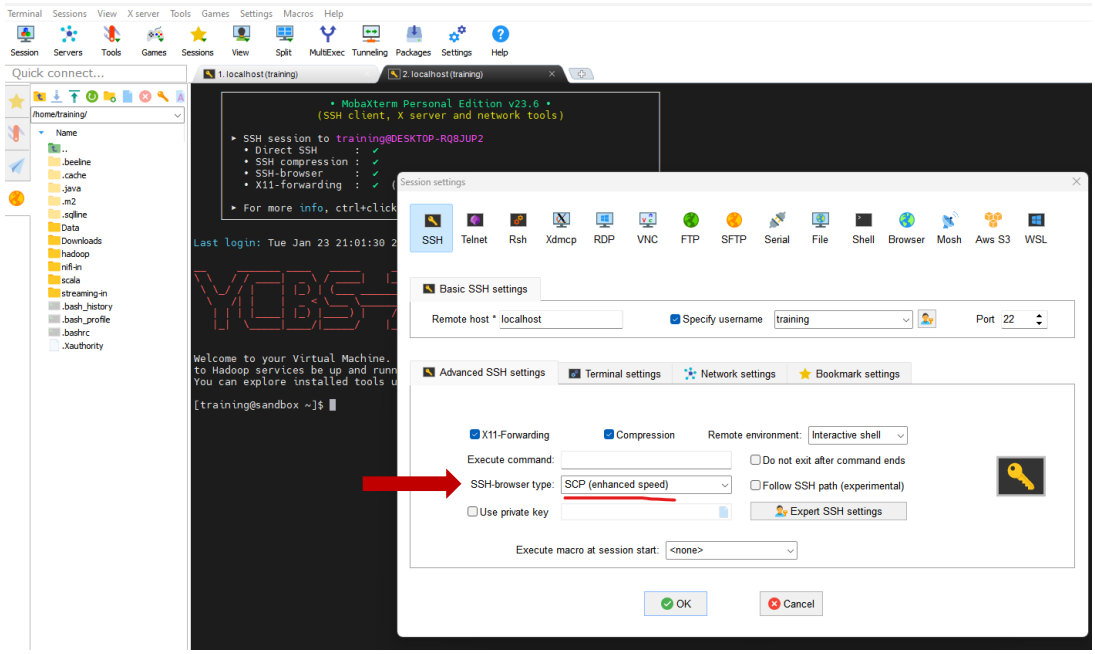


Problem

The File Explorer Panel does not appear when I open my SSH session.

Solution

This seems to be a problem in MobaXterm that appears randomly. To fix this behavior, edit your SSH session setting, Advanced SSH Settings and select the SSH-browser type as in the following screenshot.



Sandbox Installed Tools

Tool	Version	Description
airflow	2.8.3	Airflow, is an open-source tool and framework that lets you programmatically author, schedule, and monitor your data pipelines using Python.
cassandra	4.1.1	Cassandra is a distributed NoSQL database that delivers continuous availability, high performance, & linear scalability.
flink	1.15.3	Flink is an open-source, unified stream-processing and batch-processing framework.
hadoop	3.3.4	Hadoop is an open source, Java-based software platform that manages data processing and storage for big data applications.
hbase	2.4.13	HBase is an open-source, NoSQL, distributed big data store. It enables random, strictly consistent, real-time access to petabytes of data.
hive	3.1.3	Hive is open-source data warehouse software designed to read, write, and manage large datasets extracted from the Apache Hadoop Distributed File.
kafka	2.8.1	Kafka is an open-source distributed streaming system used for stream processing, real-time data pipelines, and data integration at scale.
livy	0.7.1	Livy is an open-source RESTful web service that enables data scientists and developers to easily interact with Spark clusters over a remote interface.
mysql	8.0.32	MySQL is free and open-source, widely used, relational database management system (RDBMS).
nifi	1.20.0	NiFi is a software project from the Apache Software Foundation designed to automate the flow of data between software systems.
nifi registry	1.20.0	Registry is a subproject of Apache NiFi. It is a complementary application that provides a central location for storage and management of shared resources across one or more instances of NiFi.
phoenix	5.1.2	Phoenix is an open source, massively parallel, relational database engine supporting OLTP for Hadoop using Apache HBase as its backing store.

solr	8.11.2	Solr is an open-source platform that enables near real-time indexing, database integration, dynamic clustering, full-text search, monitorable logging, rich document parsing, and more.
spark	3.2.1	Spark is an open-source analytics engine used for big data workloads that can handle both batches as well as real-time analytics.
superset	3.1.0	Superset is an open-source software application for data exploration and data visualization able to handle data at petabyte scale (big data).
tez	0.10.1	Tez is an open-source data processing framework built on top of Apache Hadoop YARN. It was designed to optimize complex directed acyclic graph (DAG) of MapReduce tasks and accelerate data processing.
trino	4.1.1	Trino is an open-source distributed SQL query engine designed to query large data sets distributed over one or more heterogeneous data sources.
zeppelin	0.10.1	Zeppelin is an open-source web-based platform for data processing, analytics, and visualization. It provides an interactive notebook that allows data scientists, analysts, and developers to collaborate efficiently. It supports various programming languages and data sources such as SQL, Python, R, and more.
zookeeper	3.5.9	Zookeeper is an open-source project providing a centralized configuration service and naming registry for large distributed systems.

Happy Learning!!