Virtual Machine is 19 GB. Don’t use cell phone connection to download. Use wifi/adsl/fiber connection.

GİZLİLİK VE GÜVENLİK POLİTİKASI

Eğitim kapsamında paylaşılan kod parçaları, notebooklar, sunumlar, ders notları, ses ve video görüntüleri, sanal makineler, sanal makine imajları, kurulum dokümanları vb. tüm içeriğin telif hakkı Miuul’a aittir. Üçüncü kişilerle paylaşılması ve çoğaltılması kesinlikle yasaktır.

Download Linki:

https://drive.google.com/file/d/1qWPMeRpsOAKZsIQDK4xh8DuZCj5WExTJ/view?usp=sharing

User: train
Password: Ankara06
root password: Ankara06

Hardware Requirements

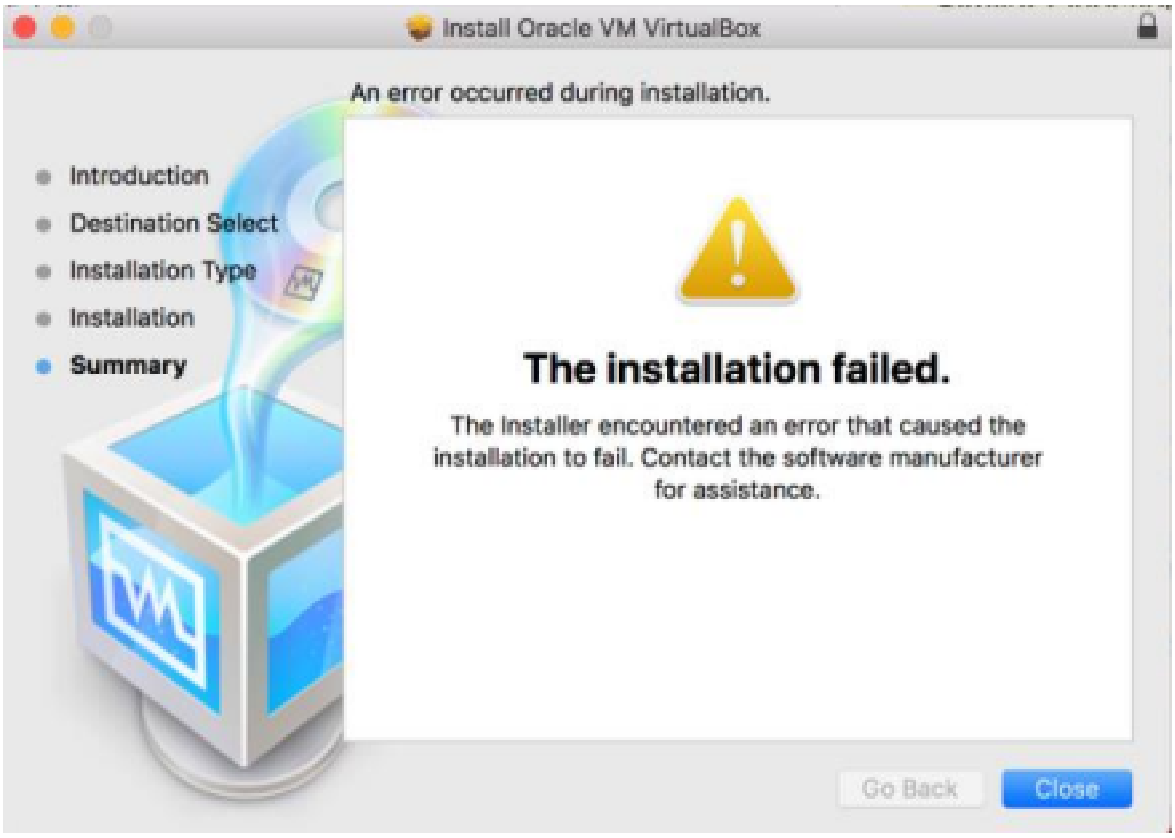
Your pc/laptop should have at least **8 GB RAM** (greater is better), **4 cpu cores** and **60 GB** free disk space. **SSD disk highly recommended**. If you use hdd give vm higher memory e.g. 10 GB.

Software Requirements

Windows 7-10 or MacOS operating system. (not M1 cpu)
VirtualBox or VmWare Player or Workstation.

Optionally recommended installations: MobaXTerm or gitbash to connect VM (Windows users only).

- Possible MacOS Virtualbox Installation Error and Solution



An error occurred during installation.

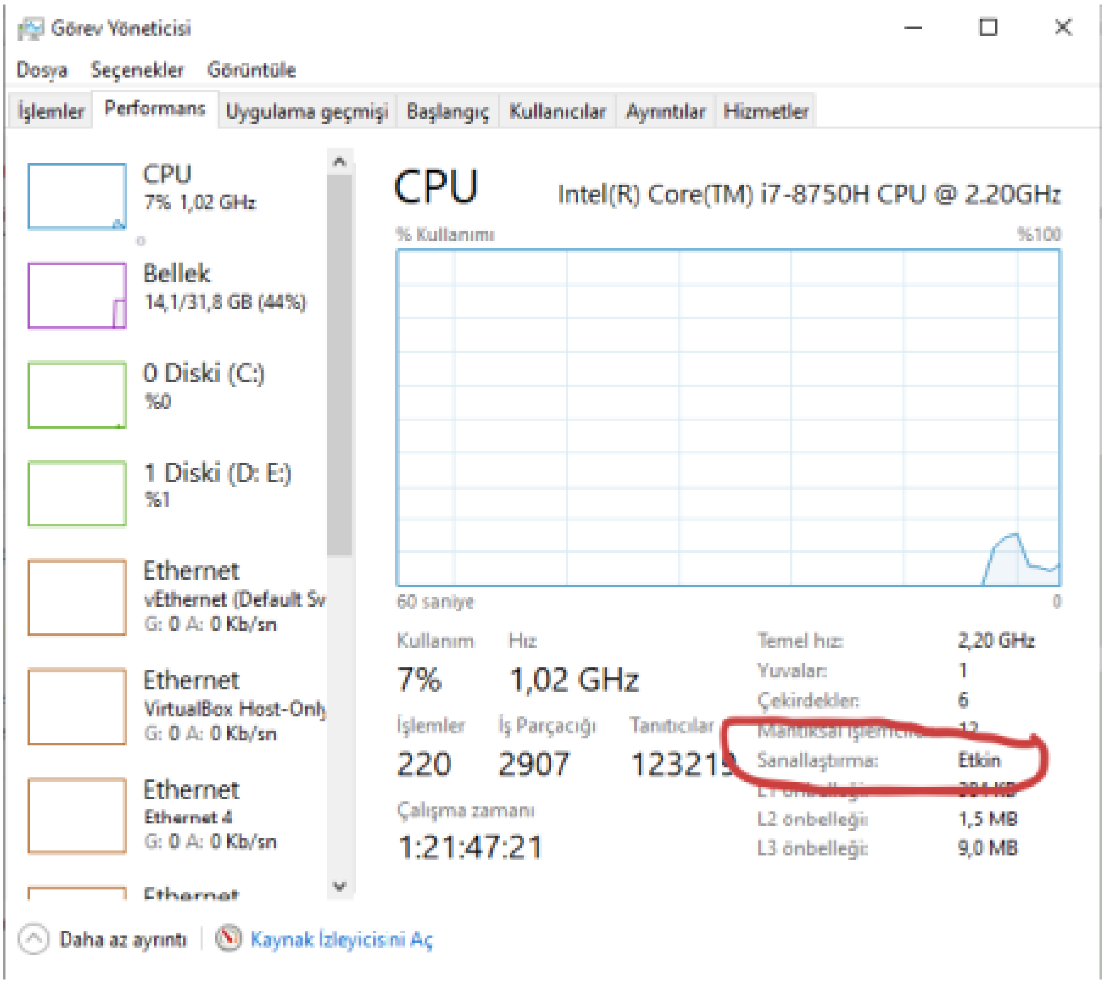
The installation failed.

The Installer encountered an error that caused the installation to fail. Contact the software manufacturer for assistance.

If you get above error, Solution#1 can help you in this post: https://medium.com/@DMeechan/fixing-the-installation-failed-virtualbox-error-on-mac-high sierra-7c421362b5b5

CPU virtualization must be enabled

CPU virtualization is enabled (Generally it is enabled, if not, you must activate from BIOS). Below you see how to check for Windows computers.



CPU
7% 1,02 GHz

Bellek
14,1/31,8 GB (44%)

0 Diski (C:)
%0

1 Diski (D: E:)
%1

Ethernet
vEthernet (Default Sv
G: 0 A: 0 Kb/sn

Ethernet
VirtualBox Host-Only
G: 0 A: 0 Kb/sn

Ethernet
Ethernet 4
G: 0 A: 0 Kb/sn

Ethernet

CPU Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz

Kullanım 7%

Hız 1,02 GHz

İşlemler 220

İş Parçacığı 2907

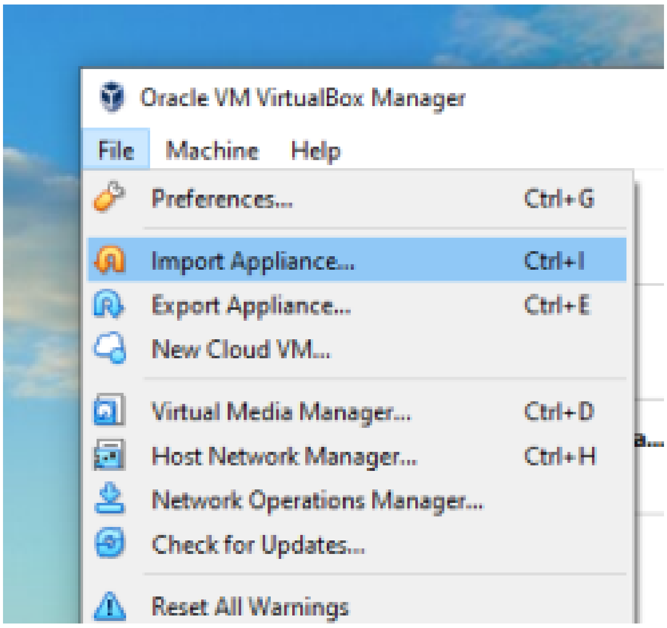
Tanıtıcılar 123219

Çalışma zamanı 1:21:47:21

Temel hız:	2,20 GHz
Yuvalar:	1
Çekirdekler:	6
Mantıksal işlemciler:	12
Sanallaştırma:	Etkin
L1 önbelleği:	384 KB
L2 önbelleği:	1,5 MB
L3 önbelleği:	9,0 MB

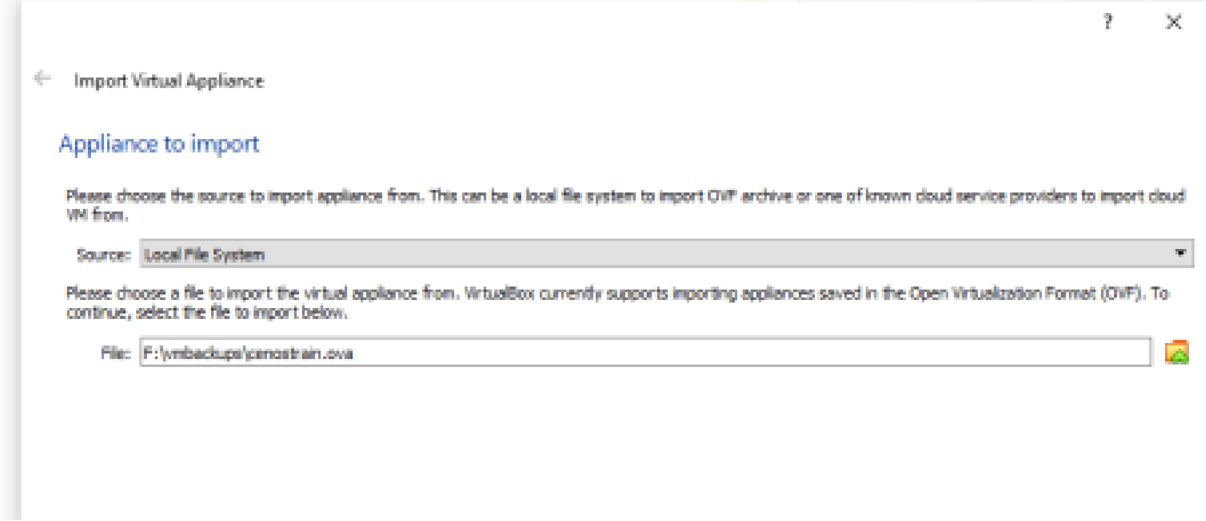
Importing vm into VirtualBox

1. Open VirtualBox and Import Appliance


Oracle VM VirtualBox Manager
File
Machine
Help
Preferences... Ctrl+G
Import Appliance... Ctrl+I
Export Appliance... Ctrl+E
New Cloud VM...
Virtual Media Manager... Ctrl+D
Host Network Manager... Ctrl+H
Network Operations Manager...
Check for Updates...
Reset All Warnings

---

2. Locate your VM



Appliance to import

Please choose the source to import appliance from. This can be a local file system to import OVF archive or one of known cloud service providers to import cloud VM from.

Source: Local File System

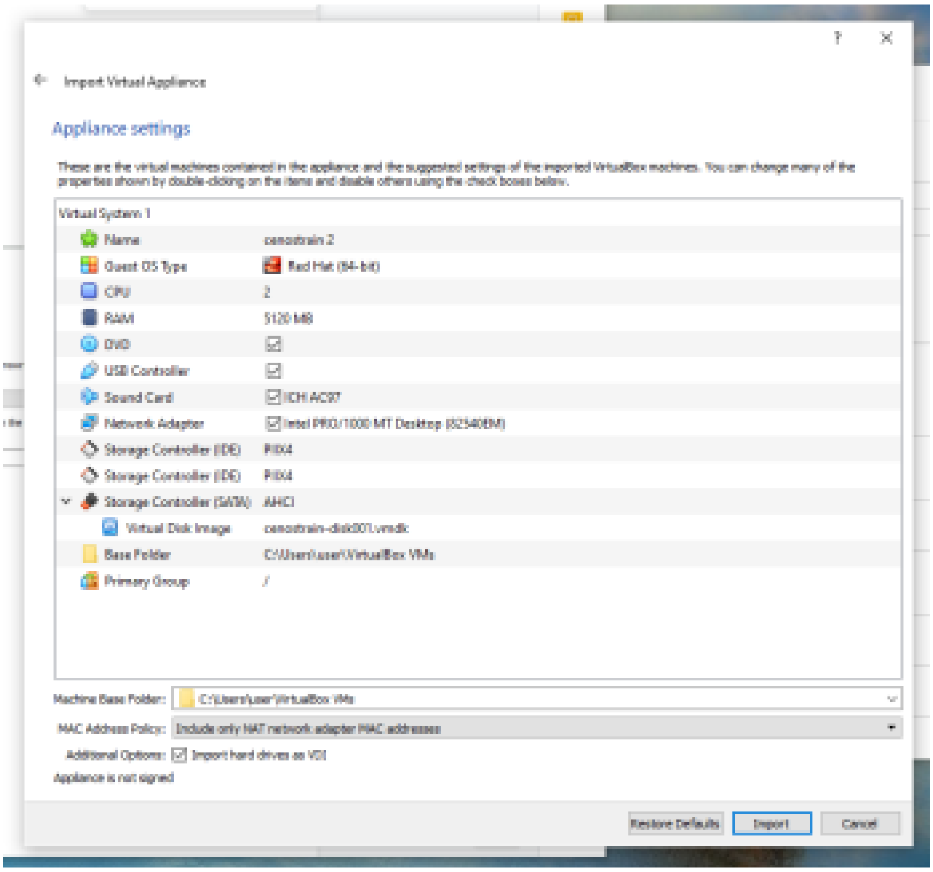
Please choose a file to import the virtual appliance from. VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

File: F:\vmbackups\cenostrain.ova

-

3. Import VM

If necessary change Machine Base Folder (must be at least 60 Gb free disk space) click Import



Import Virtual Appliance

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Virtual System 1

Name	cenostrain 2
Guest OS Type	Red Hat (64-bit)
CPU	2
RAM	5120 MB
DVD	☑
USB Controller	☑
Sound Card	☑ ICH AC97
Network Adapter	☑ Intel PRO/1000 MT Desktop (82540EM)
Storage Controller (IDE)	PIIX4
Storage Controller (IDE)	PIIX4
Storage Controller (SATA)	AHCI
Virtual Disk Image	cenostrain-disk001.vmdk
Base Folder	C:\Users\user\VirtualBox VMs
Primary Group	/

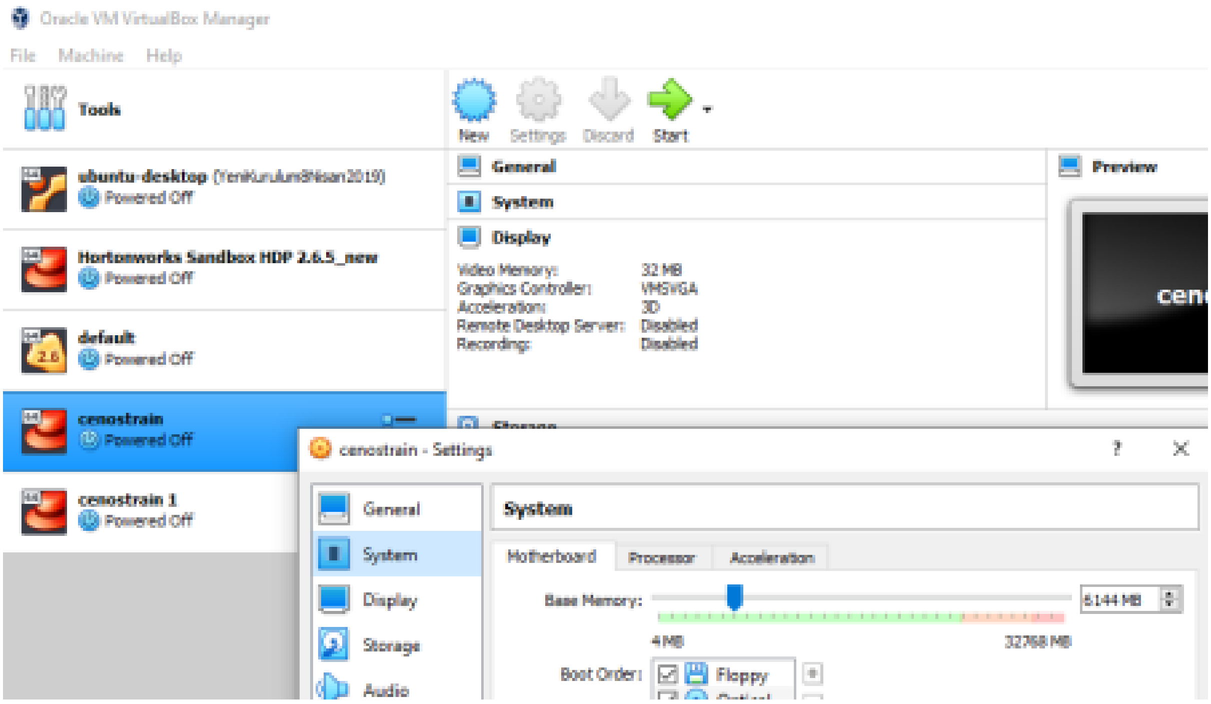
Machine Base Folder: C:\Users\user\VirtualBox VMs

MAC Address Policy: Include only NAT network adapter MAC addresses

Additional Options: ☑ Import hard drives as VDI

Appliance is not signed

Give vm higher memory and cpu core. The higher the better. But don’t forget that your host machine also needs some resources, find the balance.



Oracle VM VirtualBox Manager

ubuntu-desktop (YeniKurulumBNisan2019) Powered Off

Hortonworks Sandbox HDP 2.6.5_new Powered Off

default Powered Off

cenostrain Powered Off

cenostrain 1 Powered Off

Video Memory: 32 MB
Graphics Controller: VMSVGA
Acceleration: 3D
Remote Desktop Server: Disabled
Recording: Disabled

cenostrain - Settings

System

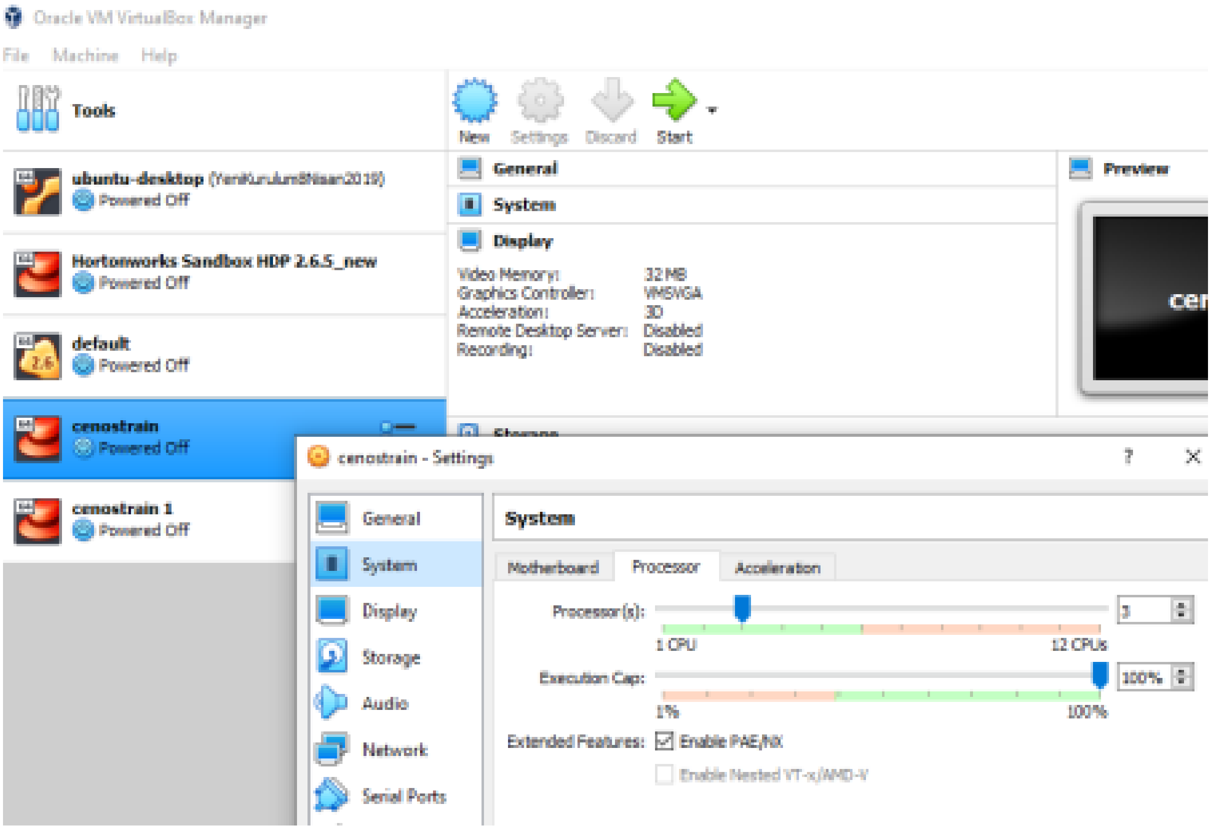
Base Memory: 6144 MB

4 MB 32768 MB

Boot Order: ☑ Floppy

-

CPU



Video Memory: 32 MB
Graphics Controller: VMSVGA
Acceleration: 3D
Remote Desktop Server: Disabled
Recording: Disabled

System

Processor(s): 3 (1 CPU – 12 CPUs)

Execution Cap: 100% (1% – 100%)

Extended Features: ☑ Enable PAE/NX

□ Enable Nested VT-x/AMD-V

Test Services

Caution!! Some commands exceed the second line. Copy both lines at the same time.

1. Test Kafka

Start Zookeeper and Kafka

```
[train@localhost ~]$ sudo systemctl start zookeeper
[train@localhost ~]$ sudo systemctl status zookeeper
      Must be running

[train@localhost ~]$ sudo systemctl start kafka
[train@localhost ~]$ sudo systemctl status kafka
      Must be running
```

Kafka topic create

```
[train@localhost ~]$ kafka-topics.sh --bootstrap-server localhost:9092 --create
-- topic test5 --partitions 3 --replication-factor 1
```

Expected output

```
Created topic test5.
```

Kafka topic list

```
[train@localhost ~]$ kafka-topics.sh --bootstrap-server localhost:9092 --list
```

Expected output

```
__consumer_offsets
test
test1
test2
```

```
test5
```

Stop Kafka

```
[train@localhost ~]$ sudo systemctl stop kafka
[train@localhost ~]$ sudo systemctl stop zookeeper
```

First start zookeeper, First stop kafka.

2. Test Hadoop, YARN and Hive

```
[train@localhost ~]$ start-all.sh
```

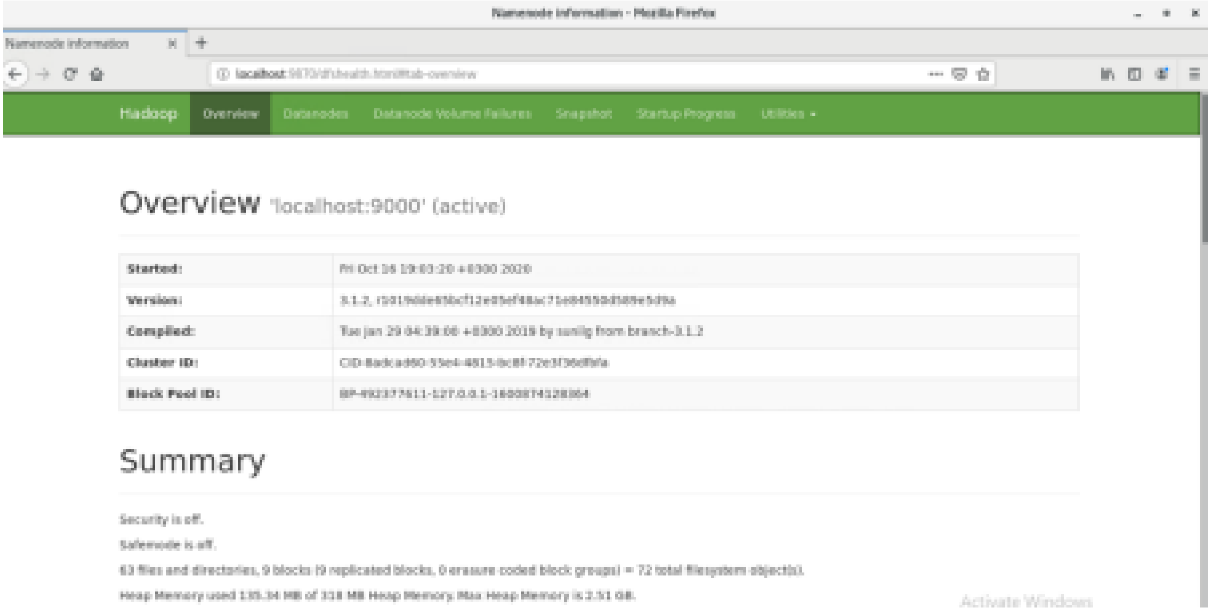
Wait till warnings end, it will take some time.

Hadoop Test

```
[train@localhost ~]$ jps
9552 SecondaryNameNode
9825 ResourceManager
10338 RunJar
9142 NameNode
10393 RunJar
10825 Jps
9276 DataNode
9950 NodeManager
```

Namenode Web UI Test

Open browser (Applications -> Firefox) and enter `http://localhost:9870/` see namenode ui



Overview 'localhost:9000' (active)

Started:	Fri Oct 16 19:03:20 +0300 2020
Version:	3.1.2, r1019dde65bcf12e05ef48ac71e84550d589e5d9a
Compiled:	Tue Jan 29 04:39:00 +0300 2019 by sunilg from branch-3.1.2
Cluster ID:	CID-8adcad60-55e4-4815-bc8f-72e3f36dfbfa
Block Pool ID:	BP-492377611-127.0.0.1-1600874128364

Summary

Security is off.

Safemode is off.

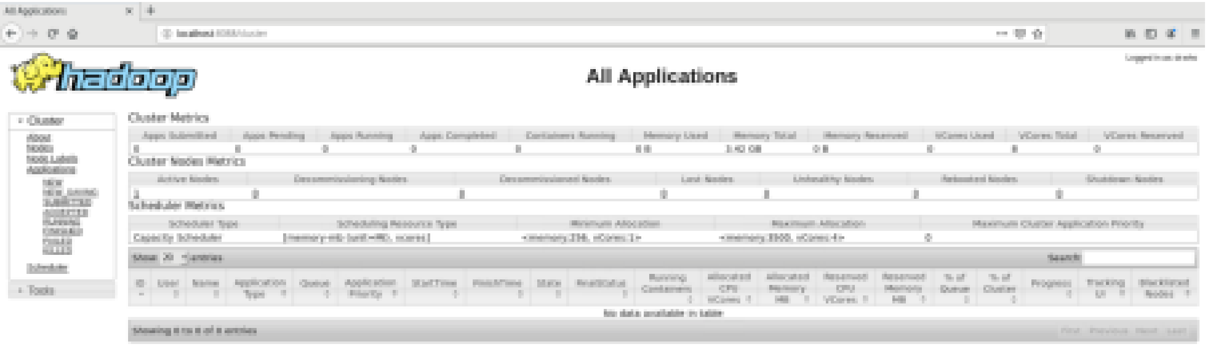
63 files and directories, 9 blocks (9 replicated blocks, 0 erasure coded block groups) = 72 total filesystem object(s).

Heap Memory used 135.34 MB of 318 MB Heap Memory. Max Heap Memory is 2.51 GB.

See: Live Nodes 1 (Decommissioned: 0, In Maintenance: 0)

YARN Test

On browser `http://localhost:8088/cluster` you should see the resource manager page



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	3.42 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:256, vCores:1>	<memory:3500, vCores:4>	0

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																			

Showing 0 to 0 of 0 entries

HDFS Write Test

```
[train@localhost ~]$ hdfs dfs -put ~/datasets/Advertising.csv
hdfs://localhost:9000/tmp
```

```
[train@localhost ~]$ hdfs dfs -ls hdfs://localhost:9000/tmp
-rw-r--r-- 1 train supergroup 4556 2021-05-09 11:54
hdfs://localhost:9000/tmp/Advertising.csv
. . .
. . .
```

. . . represents other files

3. Hive and beeline test

After starting Hadoop **wait a while (at least 30 secs to up Hive services)** then test the beeline.

Is Hive2server running?

```
pgrep -f org.apache.hive.service.server.HiveServer2
```

Expected output

```
<pid>
```

Is Hive Metastore running?

```
pgrep -f org.apache.hadoop.hive.metastore.HiveMetaStore
```

Expected output

```
<pid>
```

Beeline Test

Connect hive through beeline

```
[train@localhost ~]$ beeline -n train -u jdbc:hive2://127.0.0.1:10000
```

Stop beeline warnings

```
0: jdbc:hive2://127.0.0.1:10000> set

hive.server2.logging.operation.level=NONE;
```

Create database table insert and select

```
0: jdbc:hive2://127.0.0.1:10000> create database if not exists bookstore;
```

```
0: jdbc:hive2://127.0.0.1:10000> use bookstore;
```

```
0: jdbc:hive2://127.0.0.1:10000> create table if not exists bookstore.books(id
int,book_name string,isbn bigint,book_id bigint,price float,price_currency
string,rating_count int,author_id bigint,publisher_id bigint);
```

```
0: jdbc:hive2://127.0.0.1:10000> insert into bookstore.books values(13,"Madam
Bovary (Ciltli)",6050948752,489127179,25.115735,"TRY",5,4098249,46868),
(22,"Mai ve Siyah (Eleştirel
Basım)",9750523533,492625951,25.349610000000002,"TRY",17,3066057,63217)
, (27,"Nutuk",9759914288,9927355,11.48147,"TRY",23,9705003,46868),
 (34,"Devlet",9754734263,395307782,27.9994,"TRY",0,8978000,20709);
```

```
0: jdbc:hive2://127.0.0.1:10000> select id, book_name from
bookstore.books; OK
+-----+---------------------------------+
 | id | book_name |
+-----+---------------------------------+
 | 13 | Madam Bovary (Ciltli) |
 | 22 | Mai ve Siyah (Eleştirel Basım) |
 | 27 | Nutuk |
 | 34 | Devlet |
+-----+---------------------------------+
4 rows selected (0.369 seconds)
```

Exit beeline

```
!q
```

4. Postgresql test

Command

```
[train@localhost ~]$ systemctl status postgresql-10
```

Expected output

```
● postgresql-10.service - PostgreSQL 10 database server
 Loaded: loaded (/usr/lib/systemd/system/postgresql-10.service; enabled; vendor
preset: disabled)
 Active: active (running) since Wed 2020-09-23 16:49:49 +03; 1h 38min ago
Docs: https://www.postgresql.org/docs/10/static/
 Process: 1134 ExecStartPre=/usr/pgsql-10/bin/postgresql-10-check-db-dir ${PGDATA}
 (code=exited, status=0/SUCCESS)
 Main PID: 1156 (postmaster)
```

5. Sqoop test

```
[train@localhost ~]$ sqoop version | grep Sqoop
```

Output

```
2020-09-23 19:54:35,388 INFO sqoop.Sqoop: Running Sqoop version:
1.4.6 Sqoop 1.4.6
```

```
[train@localhost ~]$ sqoop job --list | grep -A10 'Available
jobs'
```

Output

```
2021-05-14 10:24:46,960 INFO sqoop.Sqoop: Running Sqoop version:
1.4.6 Available jobs:
```

6. Pyspark test

Local mode

```
[train@localhost ~]$ pyspark --master local
```

Output

```
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
 [GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
2021-08-14 23:28:17,984 WARN util.Utils: Your hostname, localhost.localdomain
resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface
enp0s3)
2021-08-14 23:28:17,988 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
2021-08-14 23:28:19,249 WARN util.NativeCodeLoader: Unable to load
native-hadoop library for your platform... using builtin-java classes where
applicable Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
Welcome to
  ____ __
 / __/__ ___ _____/ /__
 _\ \/ _ \/ _ `/ __/ '_/
 /__ / .__/\_,_/_/ /_/\_\ version 3.1.1
 /_/

Using Python version 3.6.8 (default, Nov 16 2020 16:55:22)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local, app id =
local-1628972904777). SparkSession available as 'spark'.
```

From http://localhost:4040 check spark web ui.

Local mode exit

```
>>> exit()
```

YARN mode

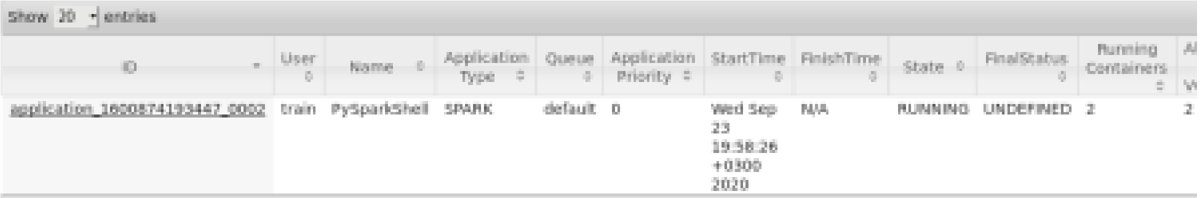
```
 [train@localhost ~]$ pyspark --master yarn
```

Output

```
Python 3.6.8 (default, Nov 16 2020, 16:55:22)
 [GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
2021-08-14 23:29:58,266 WARN util.Utils: Your hostname, localhost.localdomain
resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface
enp0s3)
2021-08-14 23:29:58,269 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to
another address
2021-08-14 23:29:59,451 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable Setting
default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
2021-08-14 23:30:06,471 WARN yarn.Client: Neither spark.yarn.jars nor
spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
  ____ __
 / __/__ ___ _____/ /__
 _\ \/ _ \/ _ `/ __/ '_/
 /__ / .__/\_,_/_/ /_/\_\ version 3.1.1
 /_/

Using Python version 3.6.8 (default, Nov 16 2020 16:55:22)
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = yarn, app id =
application_1628972365632_0002).
SparkSession available as 'spark'.
```

From http://localhost:8088/cluster/apps/RUNNING check spark have resources



Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Al
application_1600874193447_0002	train	PySparkShell	SPARK	default	0	Wed Sep 23 19:58:26 +0300 2020	N/A	RUNNING	UNDEFINED	2	2

Write with Pyspark to hive

```
>>>df=spark.read.option("header",True).option("inferSchema",True).csv("/tmp/Advert
i sing.csv")

>>> df.write.mode("overwrite").saveAsTable("advertising")

>>> spark.sql("select * from advertising").show(3)
+---+-----+-----+--------+-----+
| ID| TV|Radio|Newspaper|Sales|
+---+-----+-----+--------+-----+
| 1|230.1| 37.8| 69.2| 22.1|
| 2| 44.5| 39.3| 45.1| 10.4|
| 3| 17.2| 45.9| 69.3| 9.3|
+---+-----+-----+--------+-----+
only showing top 3 rows
```

YARN mode exit

```
>>> exit()
```

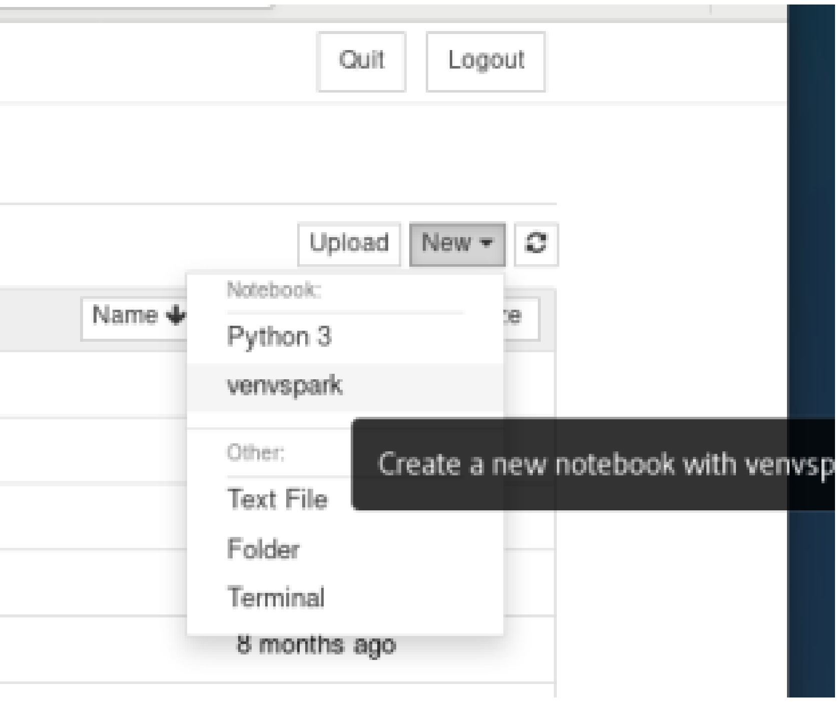
Spark python virtual environment

```
[train@localhost ~]$ source venvspark/bin/activate
```

Open Jupyter notebook

```
(venvspark) [train@localhost ~]$ jupyter notebook > notebook.log 2>&1
&
```

Jupyter will automatically open up in the browser.



Create **venvspark** notebook

Close Jupyter notebook

Save your notebook and close the browser.

On terminal stop jupyter notebook

```
 (venvspark) [train@localhost ~]$ jupyter notebook stop

Shutting down server on port 8888 ...
 [1]+ Done jupyter notebook > notebook.log 2>&1
```

Close spark virtual environment

```
 (venvspark) [train@localhost ~]$ deactivate
```

7. Stop Hadoop, YARN and Hive

```
stop-all.sh
```

8. Airflow test

```
 [train@localhost ~]$ sudo systemctl start airflow
```

```
 [train@localhost ~]$ sudo systemctl start airflow-scheduler
```

Wait 30 secs for the Airflow web server spin up. Open browser http://127.0.0.1:1502 and see airflow web ui. **username: admin, password: admin**

Stop Airflow

```
 [train@localhost ~]$ sudo systemctl stop airflow-scheduler
 [train@localhost ~]$ sudo systemctl stop airflow
```

9. Docker test

```
 [train@localhost ~]$ sudo systemctl start docker
```

```
 [train@localhost ~]$ sudo systemctl status docker
```

10. Docker-compose test

```
 [train@localhost ~]$ docker-compose version

docker-compose version 1.29.2, build 5becea4c
docker-py version: 5.0.0
CPython version: 3.7.10
OpenSSL version: OpenSSL 1.1.0l 10 Sep 2019
```

11. Minikube (Kubernetes) test

```
 [train@localhost ~]$ minikube delete
```

```
 [train@localhost ~]$ minikube start
```

```
 [train@localhost ~]$ kubectl get all
```

```
NAME TYPE CLUSTER-IP EXTERNAL-IP PORT(S) AGE service/kubernetes ClusterIP
10.96.0.1 <none> 443/TCP 10m
```

Stop minikube

```
[train@localhost ~]$ minikube stop
```

12. Jenkins Test

Start Jenkins

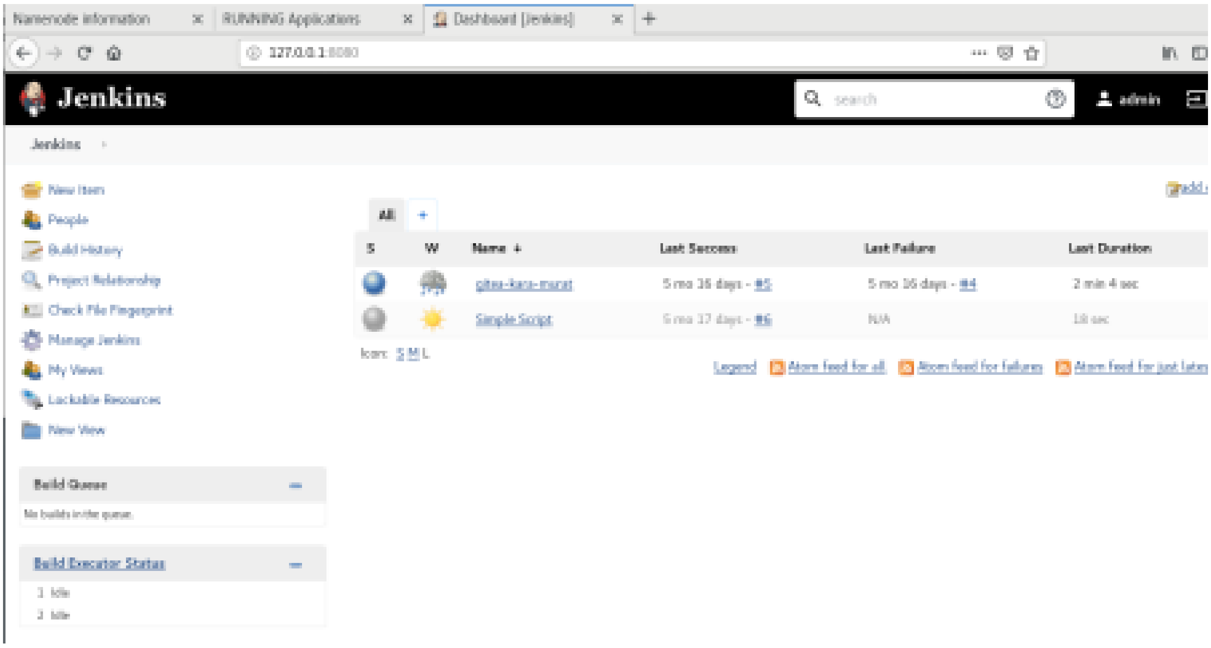
```
[train@localhost ~]$ sudo systemctl start jenkins
```

Open Jenkins UI

On browser open `localhost:8080`

username: admin, password: Ankara06

You will see this page



S	W	Name ↓	Last Success	Last Failure	Last Duration
●	☁	gitea-kara-murat	5 mo 16 days - #5	5 mo 16 days - #4	2 min 4 sec
●	☀	Simple Script	5 mo 17 days - #6	N/A	18 sec

Stop Jenkins

```
[train@localhost ~]$ sudo systemctl stop jenkins
```

13. Gitea test

Start Gitea

```
[train@localhost ~]$ sudo systemctl start gitea
```

Open Gitea UI

On browser open `localhost:3000`

Sign in with **username: jenkins, password Ankara_06**

Stop Gitea

```
[train@localhost ~]$ sudo systemctl stop gitea
```

14. Close virtual machine

Check if any docker container is running. If there are any, stop them.

stop kafka

```
[train@localhost ~]$ sudo systemctl stop kafka
```

stop zookeeper

```
[train@localhost ~]$ sudo systemctl stop zookeeper
```

stop Hadoop services

```
[train@localhost ~]$ stop-all.sh
```

Stop any other apps, services e.g. **docker, airflow, gitea** etc. like `sudo systemctl stop <xxxxxxx>`

Shutdown the machine

```
[train@localhost ~]$ sudo shutdown now
```

Congrats VM Test has finished !!!

See Troubleshooting in the following pages.

Common Errors and Fixes

1. YARN Spark Accepted State Disk usage

If your YARN apllication stuck at ACCEPTED state your disk usage probably exceeded %90. To clean it run this command and

```
rm -rf /tmp/hadoop-train/nm-local-dir/*
```

check

```
df -h /
Filesystem Size Used Avail Use% Mounted on
/dev/mapper/centos-root 38G 30G 8.0G 79% /
```

Your application will switch to RUNNING state after a while.

2. Windows Users Additional Setup

Install MobaXTerm to connect virtual machine and file transfer between host and guest machine.

3. Windows Users ssh connection to Vm

Using gitbash or putty

```
$ ssh train@localhost
```

If you get an error like below

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@ WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED! @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
```

```
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the ECDSA key sent by the remote host is
SHA256:hqpjsGjn9WsPsE+KNr87ozyUqhfyNBj7YBT0DrC+PNQ.
Please contact your system administrator.
Add correct host key in /c/Users/user/.ssh/known_hosts to get rid of this message.
Offending ECDSA key in /c/Users/user/.ssh/known_hosts:49
ECDSA host key for localhost has changed and you have requested strict checking.
Host key verification failed.
```

Run

```
$ ssh-keygen.exe -R localhost
```

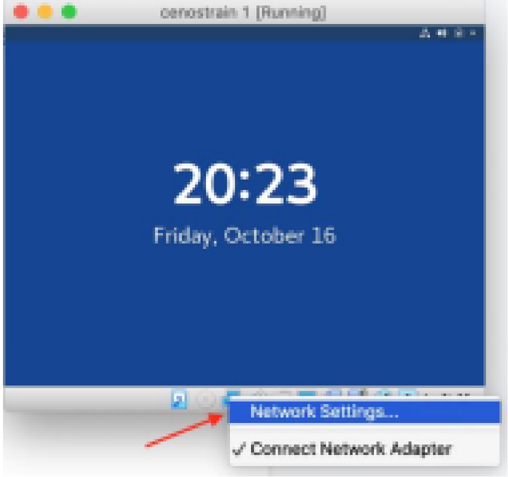
Then retry to connect.

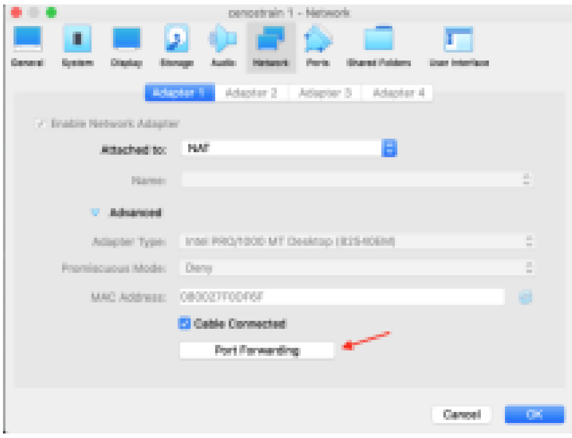
4. MacOS VM connection

MacOS users can use MacOS terminal to virtual machine.
Port forwarding over 22 is not stable on mac however it works on windows. So we need to change 22 port to 2222 to be able to make ssh connection

ssh connection settings for mac

1- Open advanced network setting and click port forwarding





✓ Enable Network Adapter

Attached to: NAT

Name:

Advanced

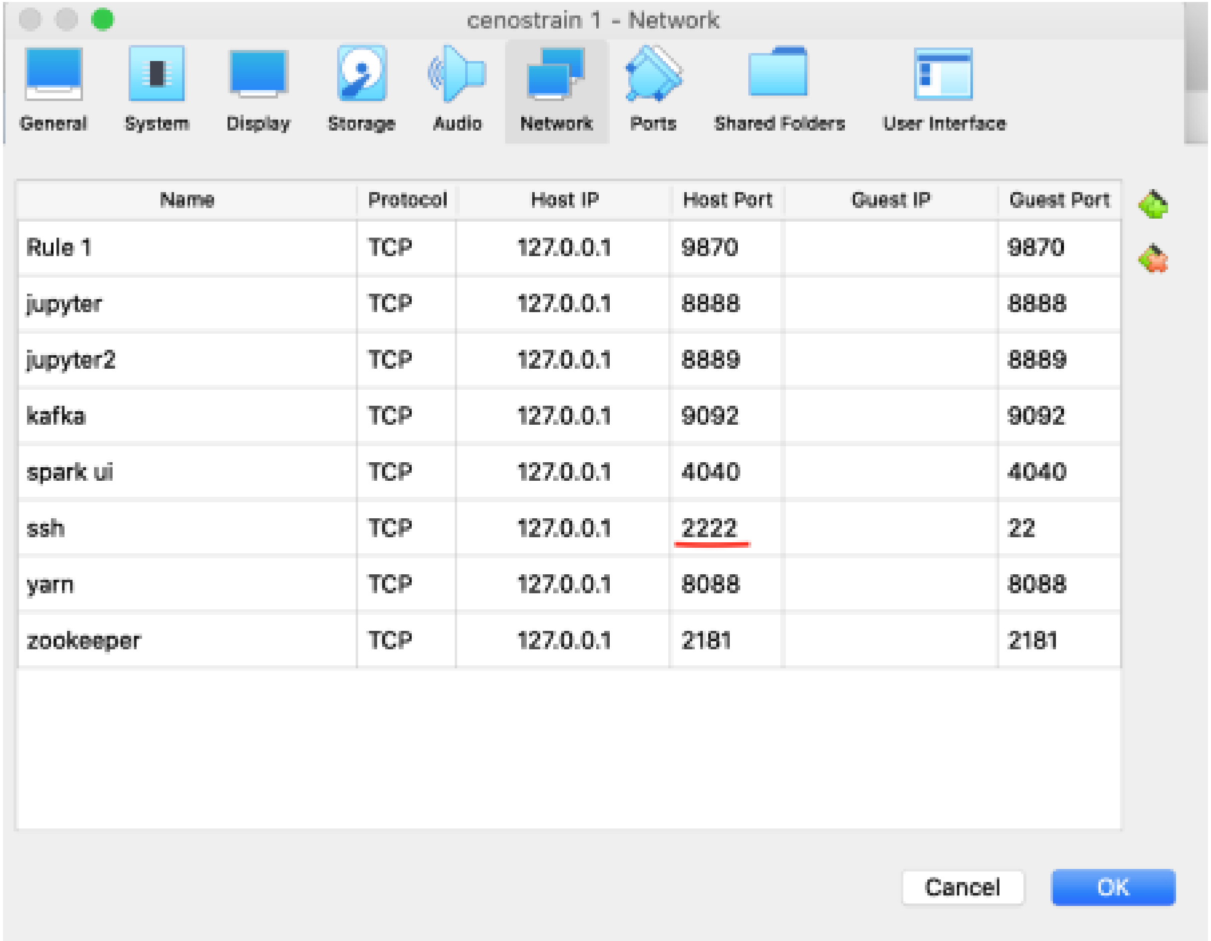
Adapter Type: Intel PRO/1000 MT Desktop (82540EM)

Promiscuous Mode: Deny

MAC Address: 080027F0DF6F

☑ Cable Connected

change 22 -> 2222



Name	Protocol	Host IP	Host Port	Guest IP	Guest Port
Rule 1	TCP	127.0.0.1	9870		9870
jupyter	TCP	127.0.0.1	8888		8888
jupyter2	TCP	127.0.0.1	8889		8889
kafka	TCP	127.0.0.1	9092		9092
spark ui	TCP	127.0.0.1	4040		4040
ssh	TCP	127.0.0.1	2222		22
yarn	TCP	127.0.0.1	8088		8088
zookeeper	TCP	127.0.0.1	2181		2181

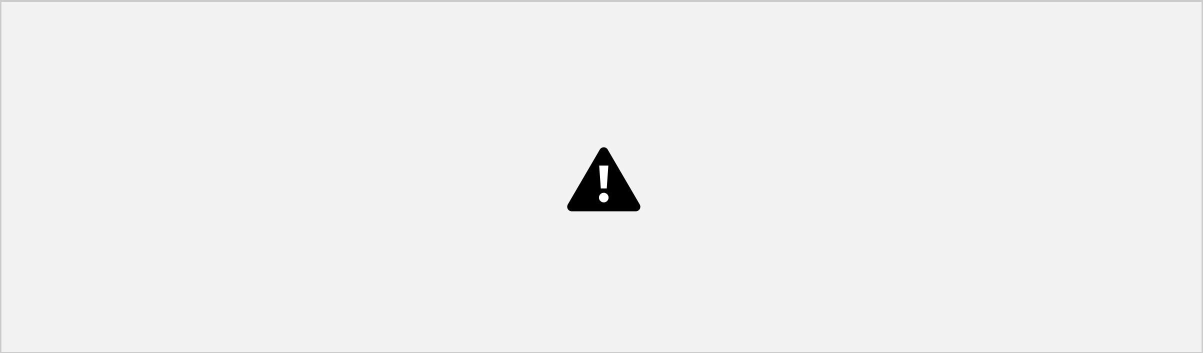
click OK

2- open mac terminal and enter ssh connection command below

```
ssh train@127.0.0.1 -p 2222
```

write yes

enter password for train user Ankara06



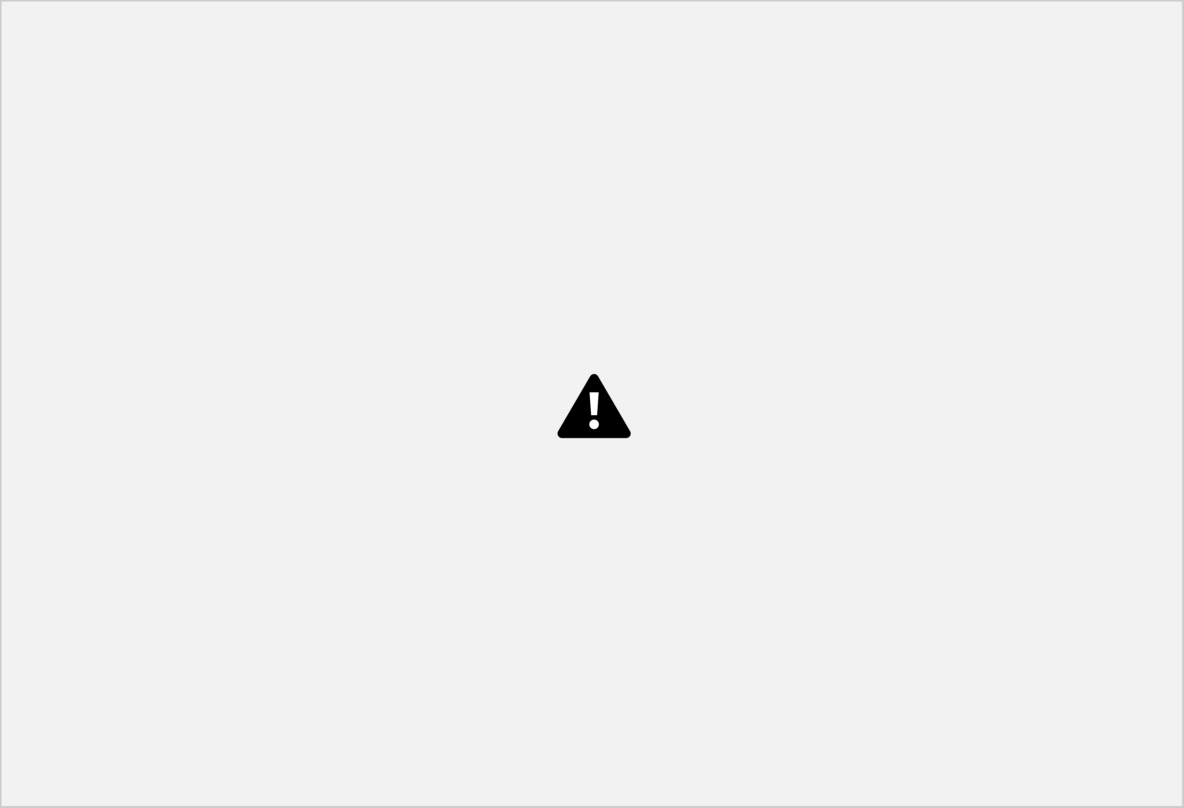
5. Jenkins password issue

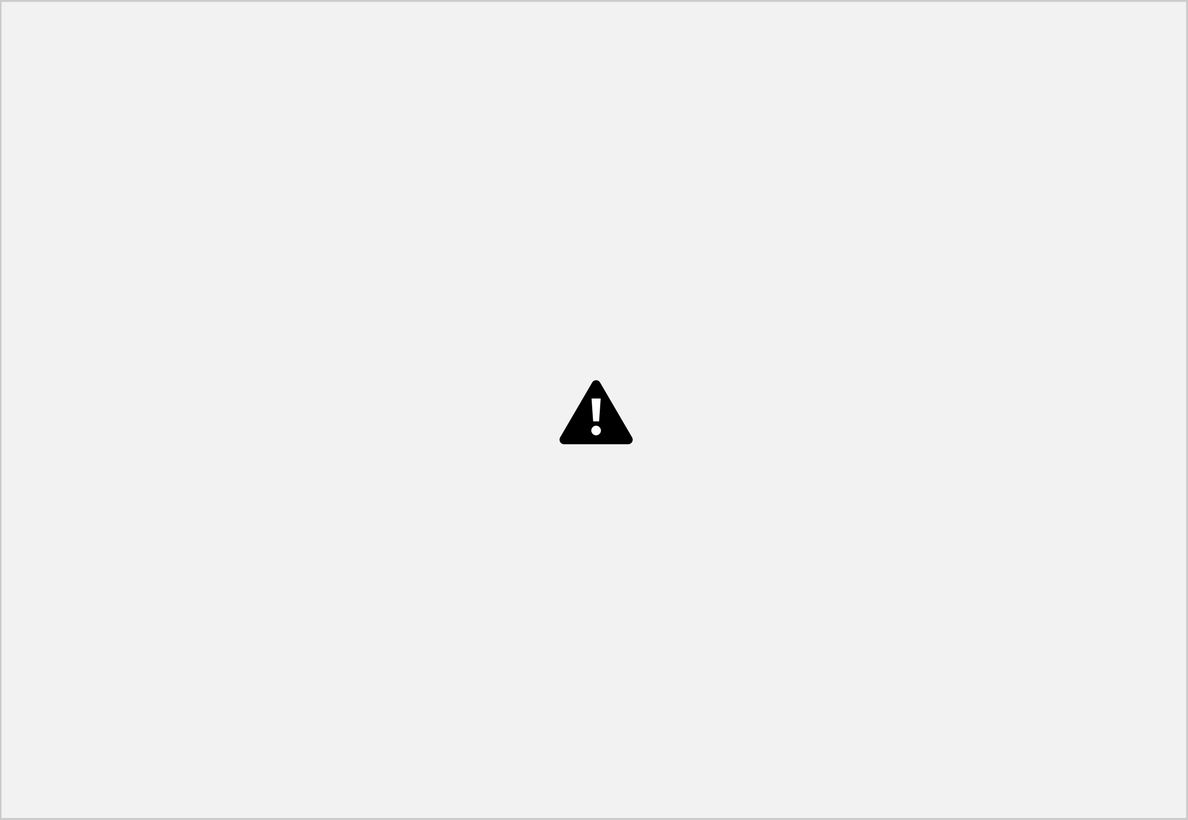
Learn Jenkins admin password

```
[train@localhost]$ sudo cat
/var/lib/jenkins/secrets/initialAdminPassword
260cd473916f4c01a0e6969857c55128
```

6. Additional Settings

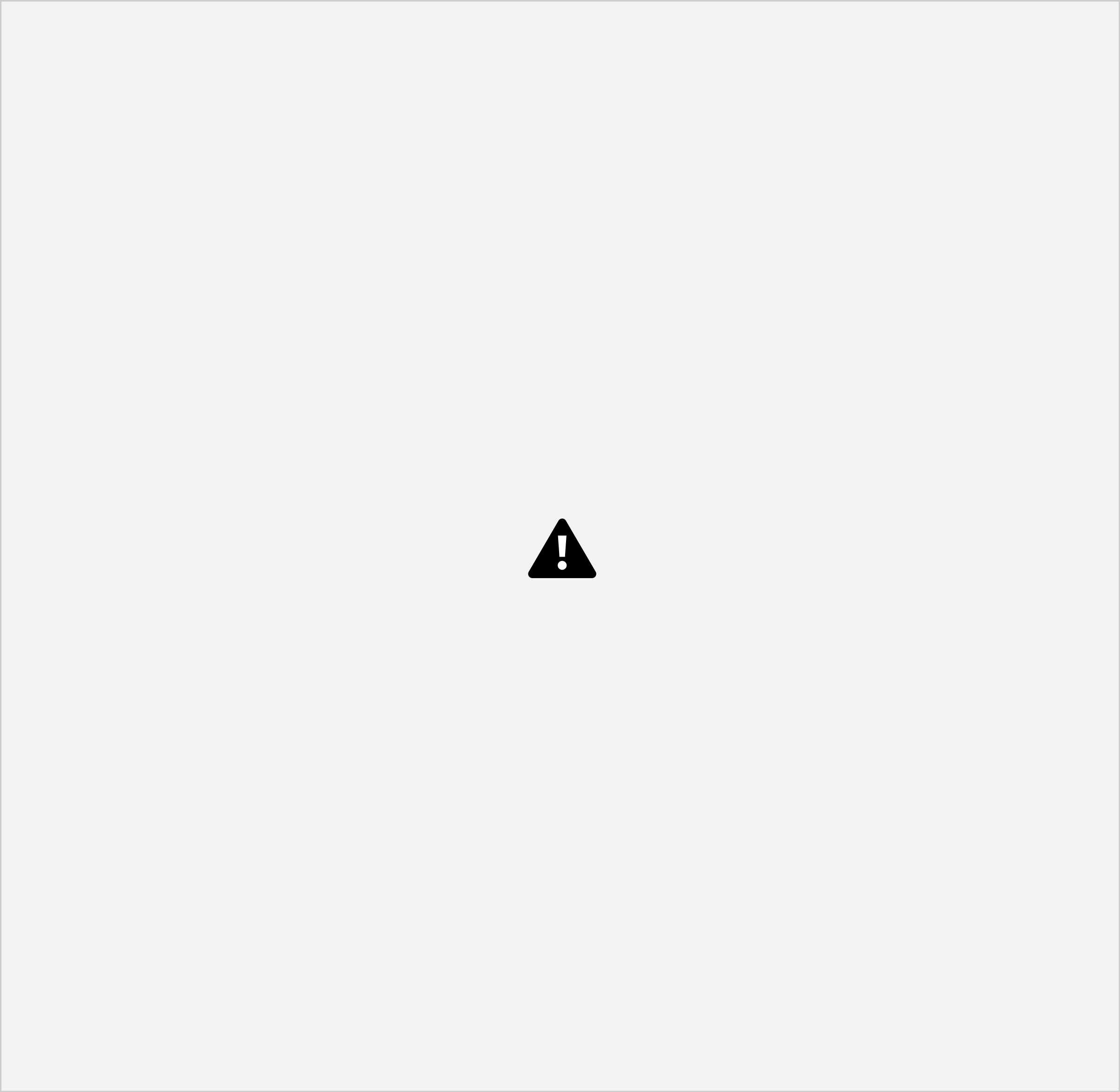
1- Disable Screen Lock.





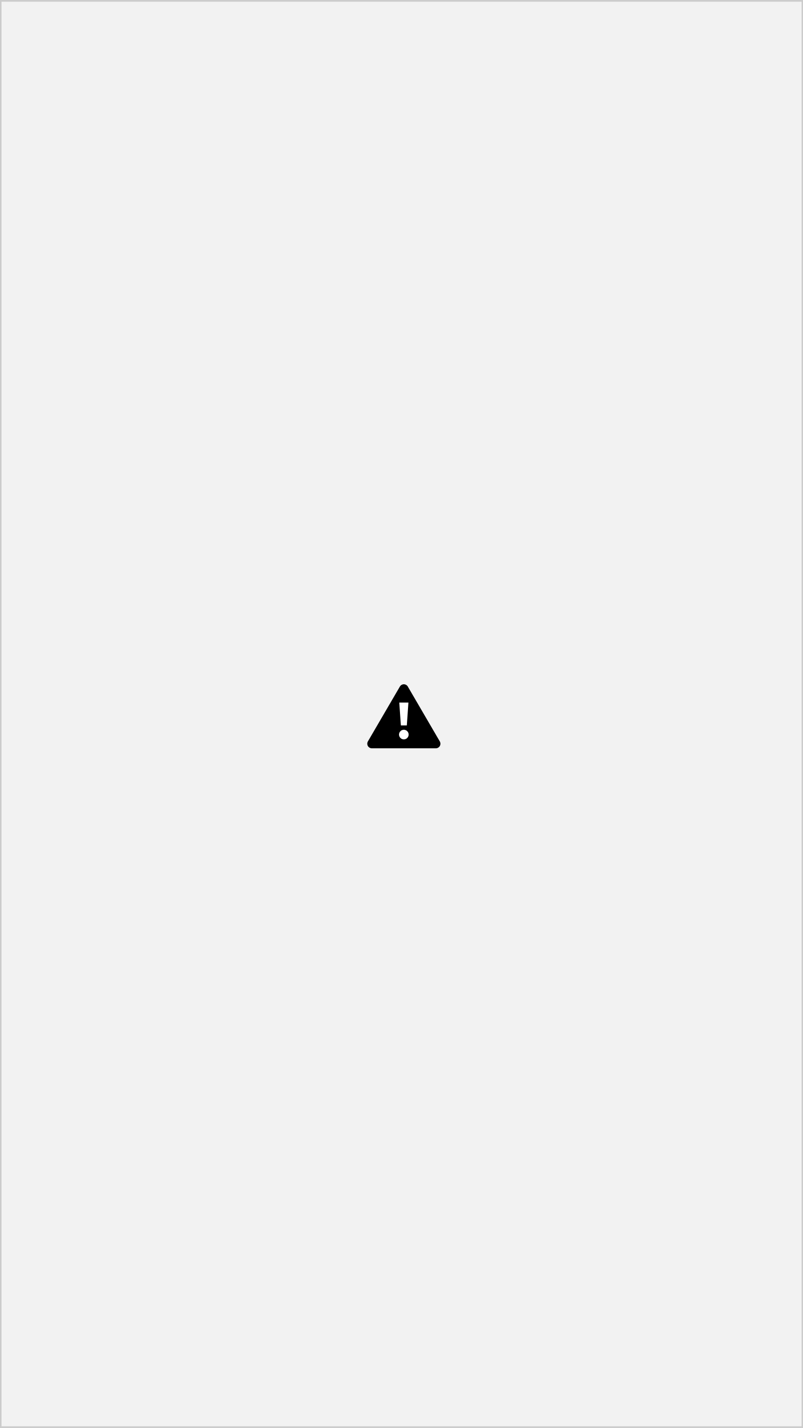
2- Language - Keyboard settings:
Default Keyboard is Turkish for this Virtual Machine. You can add additional keyboard layouts

7. windows vmware setup error



to solve this problem go to control panel and click turn windows features on or off

check windows hypervisor platform and virtual machine platform



Hyper-V Disable

If Hyper-V is enabled, the virtualbox will work slowly. To close Hyper-V completely follow the instructions on this link: https://forums.virtualbox.org/viewtopic.php?f=25&t=99390