BDA Labs — Command Crib Sheet (Exam Quick■Skim)

Labs 1–5: Install Docker • Linux & Scripting • Docker Images/Containers/Volumes/Compose • Hadoop Single■Node • MapReduce Detailed

Lab 1 — Install Docker on Ubuntu

System prep & repo:

- sudo apt update && sudo apt upgrade -y — refresh & update packages
- sudo apt install apt-transport-https ca-certificates curl software-properties-common — add HTTPS repo tooling
- curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg — trust Docker packages
- echo "deb [arch=$(dpkg --print-architecture) signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee /etc/apt/sources.list.d/docker.list >/dev/null — add repo
- sudo apt update

Install & verify:

- sudo apt install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin — engine + CLI + runtime + Buildx + Compose v2
- sudo systemctl enable --now docker — enable + start daemon
- sudo docker run hello-world — pull & test
- sudo usermod -aG docker $USER && newgrp docker — run docker without sudo

Lab 2 — Linux Commands & Scripting (inside safe container)

Start disposable container (persist to host):

- sudo docker run -it --rm -v ~/lab_out:/root ubuntu:24.04 bash — interactive shell with mounted output
- apt update && apt install -y file tar gzip findutils coreutils — basic tools

Navigation & inspection:

- pwd | ls | cd | which | file /bin/ls — paths, list, move, find path, inspect type
- test -x /bin/sh && echo Executable || echo Not — quick exec perm check

Permissions & counts:

- ls -l /bin/ls > /root/ls_perm.txt; printf "\n..." >> file — save + append notes
- ls -1 /bin | wc -l > /root/bin_count.txt — count entries

Find + sort + head (very exam■able):

- find /usr -type f -printf '%s %p\n' 2>/dev/null | sort -nr | head -n 3 > /root/top3_usr.txt — top 3 largest files

Mini script (archive top N):

- chmod +x /root/top3_archive.sh; /root/top3_archive.sh /usr /root 3 — creates /root/top_3_files.tar.gz

Ownership & chmod:

- useradd -m labuser; touch my_file.txt; chown labuser my_file.txt; chmod 600 my_file.txt — owner change & 600 perms

Processes:

- ps aux; apt -y install psmisc; pstree -p — list processes & tree
- sleep 3600 &; ps aux | grep sleep; kill — background + kill

Exit:

- exit — leave container; files stay in ~/lab_out

Lab 3 — Docker Images, Containers, Volumes, Compose

Images:

- docker pull nginx | redis | mongodb/mongodb-community-server:7.0.2-ubi8 — fetch image
- docker images; docker rmi ; docker image prune [-a]; docker system prune [-a] — list/remove/prune
- docker tag nginx:latest nginx:22sep — add tag

Build & run:

- docker build -t justasample:v1 . — build from Dockerfile
- docker run -p 8080:80 justasample:v1 — map host:container ports
- docker tag justasample:v1 /sampleapp:v1; docker push /sampleapp:v1 — publish

Containers (lifecycle & introspect):

- docker run -d --name myredis -p 6379:6379 redis — detached service
- docker ps [-a]; docker logs ; docker stop/start ; docker rm -f — manage & logs
- docker cp :/path ./ | docker cp ./file :/path — copy in/out
- docker exec -it /bin/bash — shell into container
- docker info — daemon environment

Storage:

- docker volume create myvol | ls | inspect | rm | prune — manage volumes
- docker run -d --mount source=myvol,target=/app/data busybox sleep 3600 — named volume
- docker run -d -v /host/dir:/container/dir busybox sleep 3600 — bind mount

Compose:

- docker compose up -d | down | up --build -d | logs -f | ps | up -d --scale svc=N — orchestration cheats
- Minimal:

```
version: "3.9"
services:
web:
image: nginx:latest
ports: ["8080:80"]
```

Lab 4 — Hadoop Single■Node (Phase■1/2)

Bring up:
- docker compose up -d; docker compose ps; docker compose logs --tail 50 — start & verify services

HDFS basics:
- hdfs dfs -mkdir -p /user/root/input — make dirs
- hdfs dfs -copyFromLocal /tmp/words.txt /user/root/input — upload data

Run MapReduce (WordCount):
- hadoop jar /path/hadoop-mapreduce-examples-2.7.1.jar org.apache.hadoop.examples.WordCount /user/root/input /user/root/output_wc — run job
- hdfs dfs -cat /user/root/output_wc/part-r-00000 | head — view results

Daemons/JVM:
- jps — confirm NN/DN/RM/NM/HS

Cleanup:
- docker compose down — stop (keep data)
- docker compose down -v — stop + wipe HDFS volumes

Lab 5 — MapReduce Detailed (multi■DN, datasets, monitoring, snapshots)

Reset & up:
- docker compose down; docker rm -f ; sudo rm -rf ./data — clean
- docker compose up -d — NN + 3DN + YARN + HS

Data staging:
- docker cp words.txt namenode:/tmp/; hdfs dfs -mkdir -p /user/inputdata/set{0,1,2,3}
- hdfs dfs -put /tmp/bigdata/words_setX.txt /user/inputdata/setX/ — upload sets

Run WordCount (each set):
- hdfs dfs -rm -r -f /user/output_wc_setX — clean old
- hadoop jar /tmp/hadoop-examples.jar wordcount /user/inputdata/setX /user/output_wc_setX — run
- hdfs dfs -get /user/output_wc_setX/part-r-00000 /tmp/part-r-00000_setX; docker cp namenode:/tmp/part-r-00000_setX ./ — fetch output

Admin/health:
- hdfs dfs -du -h /user/inputdata | hdfs dfs -count -h /user/inputdata — sizes & counts
- hdfs dfsadmin -report — cluster capacity & DN health
- hdfs fsck / -files -blocks -locations — file→block→replica mapping

YARN:
- yarn application -list -appStates ALL — list apps
- yarn application -status — details

Web UIs:
- NN 9870 | RM 8088 | NM 8042 | HS 8188 — confirm runs

Snapshots & replication:
- hdfs dfs -createSnapshot /user/inputdata snap_lab — snapshot
- hdfs dfs -ls /user/inputdata/.snapshot — list
- hdfs dfs -setrep -w 3 /user/inputdata/set3/words_set3.txt — set replicas=3 (waits)
- hdfs fsck /user/inputdata/set3/words_set3.txt -files -blocks — verify block/replicas
- hdfs dfs -deleteSnapshot /user/inputdata snap_lab — remove snapshot

Stop:
- docker compose down; sudo rm -rf ./data — stop & (optionally) wipe

Glossary — What Sir May Ask You to Explain (1■liners)

docker pull/build/run/ps/logs/exec/cp/stop/rm/prune/tag — image lifecycle & container management
docker volume vs bind mount — managed persistent store vs host path
docker compose up/down/--build/logs/scale — multi■container orchestration verbs
HDFS dfs vs dfsadmin vs fsck — file ops vs cluster report vs integrity/placement
YARN application list/status — job listing & details
Data locality — run map tasks where blocks live to cut network I/O
Combiner — local pre■aggregation (idempotent) to reduce shuffle bytes
Speculative execution — duplicate slow tasks to shave tail latency
Default block size — typically 128 MB; align input splits to blocks
Snapshot — point■in■time HDFS directory version (fast, metadata■based)