Using the ADA cluster

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Connecting to the cluster: Within EPFL

```
ssh <your GASPAR id>@iccluster028.iccluster.epfl.ch
```

Connecting to the cluster: Outside EPFL

1. Set up your EPFL VPN.
2. Connect to the VPN.
3. Now you’re seen as being within the EPFL network, and as such, you can use the instructions in the previous slide.

Once connected: authorisation

In order to be able to do practically *anything*, you need to initialise *Kerberos*.

You can do it by typing the following:

```
kinit
```

This will result in your being prompted to enter your password, and once you do, you’re good to go!

The Hadoop File System

- This file system is separate from the regular file system.
- In order to access it, use: hadoop fs -<your command>
- Examples:
  - `hadoop fs -cat filename.txt | less`
    - Relative addressing (will be done from your home dir)
  - `hadoop fs -ls`
  - `hadoop fs -cat hdfs:/some/dir/somefile.txt | head -20`
    - Absolute addressing

How do I submit a job?

Using spark-submit!

Example:

```
spark-submit --master yarn --deploy-mode client
--driver-memory 4G --num-executors 5 --executor-memory 4G
--executor-cores 5 some_script.py script_arg1 script_arg2
```

The arguments

- --master: **Must be set to `yarn`.**
  - **Never** use local mode for this argument.
- --deploy-mode: **Must be set to `client` or `cluster`.**
- --driver-memory: The memory allocated to your driver node (the master node that acts as the supervisor of all executors).
- --executor-memory: The memory allocated to each executor.
- --num-executors: Number of executors allocated to your job.
- --executor-cores: Number of cores per executor (hint: the executors are virtual). Set to 4 or 5 for best performance.

What NOT to do

Do not allocate large amounts of driver and executor memory to yourself unless you have already submitted the job with smaller amounts and you have run into memory problems. Start with lower amounts of memory, and only increase if absolutely necessary.

If we see a job using disproportionate amounts of memory (relative to the size of the dataset you're using), we may kill it.

In short: be mindful of the other users of the cluster, i.e. your classmates!

Advanced arguments

- `--conf spark.yarn.executor.memoryOverhead=1024`
  - The memoryOverhead off-heap memory allocated for overheads to each executor. Amount is in megabytes. Only change it if you get an error that tells you to consider boosting this parameter.
- `--py-files dependencies.zip`
  - If you have some other Python files on which your script depends, zip them into a file and ship them to the server using this argument.

Other FAQs

- Q: Can I use a Jupyter notebook on the cluster?
  - A: No, the risk of someone forgetting to close their notebook and blocking half the class is too high.
- Q: Can I use the pyspark console?
  - A: No, for the same reason as above.
- Q: How much memory is reasonable to use in spark-submit?
  - A: Depends on your dataset. For most, the example given in these slides should be enough.

Good luck!