# Azure HDInsight notes

* Azure HDInsight training: <https://learn.microsoft.com/en-us/training/paths/build-oss-analytical-solutions-az-hdinsight/>

# Spark and Hadoop materials

* Submitting Spark jobs:
  + <https://www.youtube.com/watch?v=o_pne3aLW2w>
  + <https://www.youtube.com/watch?v=L3VuPnBQBCM>
  + <https://learn.microsoft.com/en-us/azure/hdinsight/spark/spark-best-practices>
  + <https://stackoverflow.com/questions/53344285/is-there-a-way-to-submit-spark-job-on-different-server-running-master>

# Multi node hadoop set up

**The best materials:**

* <https://www.youtube.com/watch?v=_iP2Em-5Abw&t=371s> - multi node hadoop set up without docker
* <https://www.cloudduggu.com/hadoop/installation-multi-node-cluster/> - multi node hadoop set up without docker
* <https://www.confessionsofadataguy.com/create-your-very-own-apache-spark-hadoop-cluster-then-do-something-with-it/> - Setup of both Hadoop and Spark. there is no info about on which nodes (servers) execute which commands in terminal for setting up hadoop. But the part for Spark seems fine.

**Official documentation:**

* Hadoop official documentation about cluster setup: [hadoop.apache.org](https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-common/ClusterSetup.html)
* Hadoop official HDFS guide: [hadoop.apache.org](https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html#Related_Documentation)

**Other materials:**

* <https://freedium.cfd/https://blog.det.life/developing-multi-nodes-hadoop-spark-cluster-and-airflow-in-docker-compose-part-1-10331e1e71b3> - medium article about multi node hadoop and spark set up with docker. It creates multiple data nodes as separate docker containers but everything is running on a single machine using docker compose.
* <https://freedium.cfd/https://medium.com/@rubenafo/some-tips-to-run-a-multi-node-hadoop-in-docker-9c7012dd4e26> - hadoop multi node set up. There is a missing piece about generating ssh keys. Here different data nodes are created in a separate docker containers and all of them are running on a single machine. Docker compose is not used so some additional task related to networks in Docker are required.
* <https://www.linkedin.com/pulse/setup-multi-node-hadoop-cluster-using-docker-komal-suthar/> - hadoop multi node set up. Some actions are done with docker, some manually.

# Multi node spark set up

**The best materials:**

* <https://www.cloudduggu.com/spark/installation-multi-node/> - spark multi node set up with hadoop.
* <https://www.confessionsofadataguy.com/create-your-very-own-apache-spark-hadoop-cluster-then-do-something-with-it/> - Setup of both Hadoop and Spark. there is no info about on which nodes (servers) execute which commands in terminal for setting up hadoop. But the part for Spark seems fine.

**Other materials:**

* <https://www.youtube.com/watch?v=f_XsaYcETnI> – spark multi node set up on hadoop
* <https://www.youtube.com/watch?v=-5TSKMXAygc> – spark multi node set up. I am not sure if it is using hadoop.
* <https://data-flair.training/blogs/install-apache-spark-multi-node-cluster/> - spark multi node set up on hadoop

# Hadoop and Spark theory

* + <https://www.youtube.com/watch?v=N6TmDNexxGI&list=PLGhXxbu7qYooyn_aWk1DqpIF1CjBzaSUn&index=2> - Hadoop
  + <https://www.youtube.com/watch?v=Tyg1FVNq40g&list=PLGhXxbu7qYooyn_aWk1DqpIF1CjBzaSUn&index=3> – Hadoop and Spark (9h video)
  + <https://www.youtube.com/watch?v=rsOSrEbK7sU&list=PLLa_h7BriLH0FzTY5aBFpH-vciOiEf4Br> - Spark

# Spark multinode setup

In this repo: <https://github.com/cluster-apps-on-docker/spark-standalone-cluster-on-docker> the docker compose file is creating a spark master and workers.

I can run that on a single machine but I can also take from that docker compose file only the section for a master and run it on one VM.

Then I can take from docker compose the section for one worker and run it on another VM, and do the same for other workers.

# Hadoop multinode setup

We can use this repo: <https://github.com/big-data-europe/docker-hadoop> and do the same as for Spark multinode setup.

That is to take from the docker compose sections for namenode, resourcemanager, nodemanager and historyserver and run that on one VM, and then take from docker compose the datanode section and run it on another VM.

# Kubernetes

With using Kubernetes and Helm Chart it might be easier to set up a multinode hadoop and spark. It can help with networking and communication between VMs.