**Azure**

Install and use Hortonworks on Azure

<https://bigdata-etl.com/how-to-install-hortonworks-in-azure/>

<https://bigdata-etl.com/how-to-install-hortonworks-sandbox-with-data-platform-in-microsoft-azure-part-2/>

<https://www.youtube.com/watch?v=VefG02MA0Ik&t=519s>

**Connecting and loading data to HIVE:**

**hive jdbc driver:**

jdbc:hive2://sandbox-hdp.hortonworks.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2

CREATE TABLE CustomerOrders

(CustomerName VARCHAR(30),

OrdDate DATE,

Category CHAR(1),

ProductName VARCHAR(30),

Quantity INT,

Price DECIMAL(5,2)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

tblproperties ("skip.header.line.count"="1")

;

LOAD DATA INPATH '/demo/data/Customer/Customer-Orders.csv'

OVERWRITE INTO TABLE customerorders;

**Azure account**

[K\_rasuri@outlook.com](mailto:K_rasuri@outlook.com)

Submit.27

**HortonWorks sandbox**

hdpuser

Generate.123

root

Admin123

**Ambari connection**

http://hwsandbox.eastus.cloudapp.azure.com:8888

maria\_dev

maria\_dev

**Connecting to Hortonworks VM from Azure CLI**

ssh [hdpuser@40.112.53.25](mailto:hdpuser@40.112.53.25)

ssh –p 2222 root@localhost

beeline

!connect jdbc:hive2://sandbox-hdp.hortonworks.com:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2;

**HDInsight Cluster:**

admin, Generate.123

sshuser, Generate.123

**Exercises:**

* PySpark commands from spark shell

<https://spark.apache.org/docs/2.4.0/quick-start.html>

* Build a program from Visual studio code. Deploy it to the cluster and run t
* Run an actual program on 6 node cluster

Selecting the right VM for your HDInsight cluster

<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-selecting-vm-size>

Execute spark and Hive queries from Visual Studio Code

<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-for-vscode>

Create HDInsight cluster from scripts

<https://docs.microsoft.com/en-us/azure/hdinsight/azure-cli-samples>

<https://github.com/Azure-Samples/hdinsight-dotnet-sdk-samples>

<https://docs.microsoft.com/en-us/python/api/overview/azure/hdinsight?view=azure-python>

Debugging failed Spark job in IntelliJ

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-intellij-tool-failure-debug>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-intellij-tool-plugin-debug-jobs-remotely>

Build spark application using IntelliJ, deploy and run it on HDInsight cluster

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-create-standalone-application>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-intellij-tool-plugin>

Manage Cluster resources

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-resource-manager>

Monitor spark applications using history server

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-azure-spark-history-server>

<https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-log-management>

Improve spark jobs performance

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-improve-performance-iocache>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-perf>

<https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-settings>

How Apache Spark works

<https://www.dezyre.com/article/apache-spark-architecture-explained-in-detail/338>

Data Modeling CosmosDB

<https://docs.microsoft.com/en-us/azure/cosmos-db/how-to-model-partition-example>

<https://docs.microsoft.com/en-us/azure/cosmos-db/modeling-data#hybrid-data-models>

Difference between NO SQL and Relational databases

<https://docs.microsoft.com/en-us/azure/cosmos-db/relational-nosql>