**AWS S3 integration with Spark local environment via AWS CLI**

Step – 1: Install Python latest package.

A computer screen with white text

Description automatically generated

Step – 2: Install Java 8 or 11 and set JAVA\_HOME and path in environment variables

A screenshot of a computer program

Description automatically generated

Step – 3: Download and extract hadoop 3.2.2 package from this link and set HADOOP\_HOME and path in environment variables

[https://github.com/cdarlint/winutils](https://github.com/cdarlint/winutils%20)

Step – 4 : Download apache spark latest and extract it in the same folder as of Hadoop downloaded previously and set SPARK\_HOME and path in environment variables.

For extract : tar zxvf <Hadoop-file>A screenshot of a computer

Description automatically generated

Step – 5 : Check spark is running or not by running pyspark in cmd

A computer screen with white text

Description automatically generated

Step – 6 : Now, install aws cli in our local system

Cmd : msiexec.exe /i <https://awscli.amazonaws.com/AWSCLIV2.msi>

Step – 7 : Now, configure the aws with your root accesskey, secretkey and region

A screenshot of a computer program

Description automatically generated

Step – 8 : check if your aws account is connected to your local system or not

A computer screen shot of a computer program

Description automatically generated

Step – 9: Now, login to your aws console and create an s3 bucket. While creating give it publicly accessible and ACL enabled.

A screenshot of a computer

Description automatically generated

Step – 10 : Then create another folder inside S3 (optional)

A screenshot of a computer

Description automatically generated

Step – 11 : Now, download a csv file and upload the file into s3 bucket, into the folder you created

A screenshot of a computer

Description automatically generated

Step – 12: Now, we just check whether the uploaded s3 object is reflecting in our local terminal or not.

Cmd: aws s3 ls: s3//<s3-bucket-name>/<folder-name>/

A computer screen shot of a computer program

Description automatically generated

Then copy that file from s3 bucket to local environment.

cmd: aws s3 cp <s3-bucket-file-location> <local-path-location>

A screenshot of a computer program

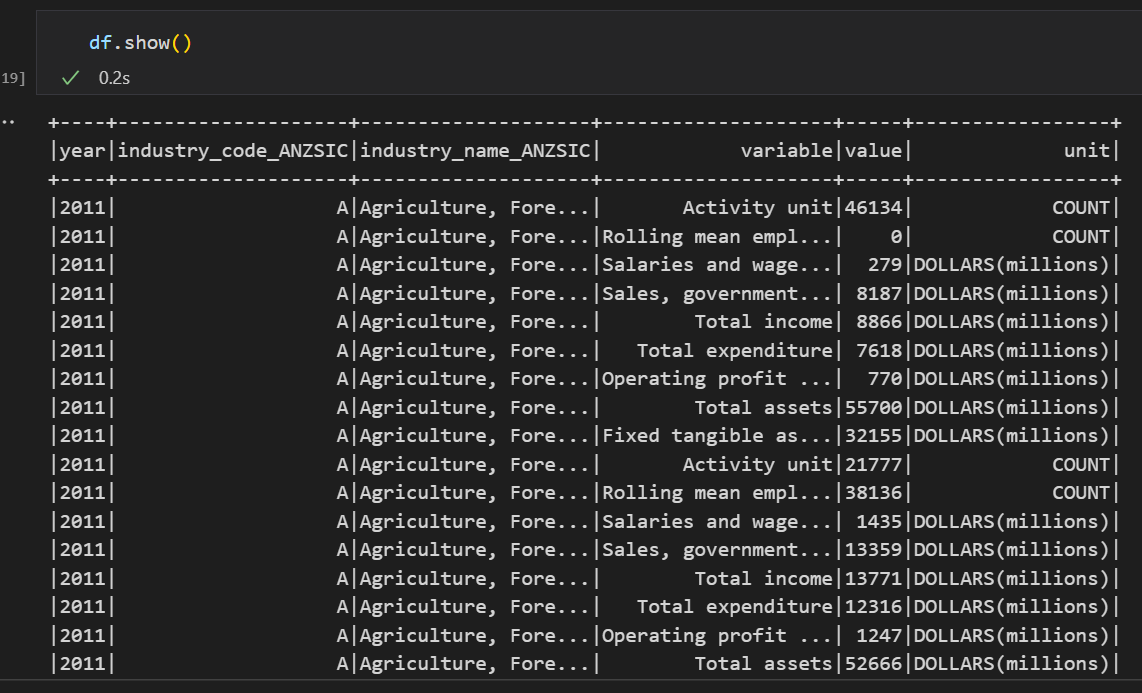
Description automatically generated

Step – 13 : Now, using that csv file create a spark dataframe

A screenshot of a computer program

Description automatically generated

Step – 14 : Now, show the data in the dataframe just now created



Step – 15: Now, remove unnecessary columns from the data frame A screenshot of a computer program

Description automatically generated

Step – 16 : Preprocess the dataframe (modify the data in the dataframe)

A black screen with white text

Description automatically generated

Step – 17: Now, ad we have changed the column type from string to integer. We will check the datatype whether it is changed or not.

A screenshot of a computer program

Description automatically generated

Step – 18 : Replace all the null values present in the value column with mean value of the column

A screen shot of a computer code

Description automatically generated

Step – 19: We have preprocessed the data frame, now we will export this data frame into a csv file

A black screen with text on it

Description automatically generated

Step – 20: Now, check whether, the csv file is created or not

A screenshot of a computer

Description automatically generated

Step – 21: Now, we will copy this modified csv file from our local system to aws s3 bucket

A screenshot of a computer

Description automatically generated

Step – 22: Now, check whether the file is uploaded to s3 bucket.

A screenshot of a computer

Description automatically generated

Step – 23: Check also in our local system, whether the newly uploaded file is reflecting

A screenshot of a computer program

Description automatically generated

Hereby, I have successfully connected our local machine with aws account, then I retrieved the .csv file present in the s3 bucket to our local system with aws cli and then created a spark data frame from the csv file and did data preprocessing. Then I have exported this data frame to a csv file. Finally, I have uploaded the modified .csv file to aws s3 bucket.