**Assignment 02: Create, Filter, and Clean an RDD**

**Display all the airports based in New York using PySpark**

**DESCRIPTION**

**Problem:**

Use the given dataset to count and display all the airports based in New York using PySpark.

Perform the following steps:

View all the airports listed in the dataset

View only the first 10 records

Filter the data for all airports located in New York

Clean up the dataset, if required

You can download the additional resources from the link shared -

<https://www.dropbox.com/sh/99zzg0rn5ekhir5/AACfmgTPUfJI8GuG9LCe2PLSa?dl=0>

**Assessment:**

(Note: For code refer the attached notebook in pdf format)

**1: Perform the preparatory activities.**

* Verify the Spark Context using ‘sc’ command.



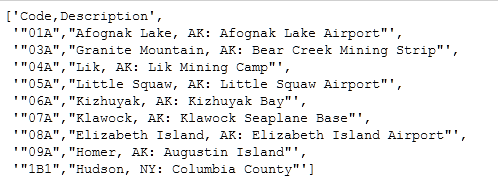
* Import the given airport dataset using sc.textfile ().

**2: View all the airports listed in the dataset.**

* Use collect () function, which returns all elements of the dataset as an array.
* Use count (), which returns the number of elements present in the dataset.
* The number of elements in dataset is 6389.

**3: View only the first 10 records.**

* Use datafile.take (10)**,** which returns number of elements (n) as specified by the number in parenthesis.

****

**4: Filter the data for all airports located in New York.**

* Use New York state (NY) code to filter the required data.

****

* + Verify the total number of airports present in the filtered list.
  + The total number of airports in the filtered list is 81.
  + Use collect (), View all the details of the airports in the filtered list.

**5: Clean up the dataset (Hint: Identify and use the common factor among all the airports that are located in New York).**

* Update the filter condition to get the required output.
* Verify the total number of airports present in the new filtered list.
* The total number of airports present in the new filtered list is 70.
* View the contents of the new filtered list using collect ().