

**Task-2:** Write a Spark job **without using SQL** to determine the average pressure at 9am (Pressure9am) for Launceston (location).

---

**Step 1:** Login to HADOOP CLUSTER and Navigate to Pyspark using pyspark command

```
(base) Achyuthas-MacBook-Air:~ achyuthanagaveti$ ssh anagave@hadoop-nn001.cs.okstate.edu
anagave@hadoop-nn001.cs.okstate.edu's password:
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-62-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

Last login: Fri Apr  2 02:02:28 2021 from 10.200.206.99
```

```
anagave@hadoop-nn001:~$ pyspark
```

```
Welcome to
      ____
     /  _ \   _ __   _
    /  _ \  / __ \  / _ \
   /  _ \ /  __ \|  __/
  /  _ \|  _  \| |  | |
 /  _ \|  \  __/| |__| |
/  _ \|   \____|_____|
/  _ \|
/  _ \|

version 3.0.1

Using Python version 3.8.5 (default, Jan 27 2021 15:41:15)
SparkSession available as 'spark'.
```

**Step 2:** CSV file is located in /user/common\_data/Spark\_Assignment\_Dataset.csv path.

---

**Step 3:** Import required packages such as SparkSession, pyspark.sql.functions, pyspark.sql.types, date time.

---

**Step 4:** Create spark session variable to assign sparksession and dataframe variable to create a data frame from csv

---

**Step 5:** Apply filter on the data frame to show the Average pressure, and print the same. Results are shown below(Highlighted in the red box)

```
achyuthanagaveti — ssh anagave@hadoop-nn001.cs.okstate.edu...
>>> print("\033[1m" + "Task 2: Write a Spark job without using SQL to determine
the average pressure at 9am (Pressure9am) for Launceston (location)" + "\033[0m"
)
Task 2: Write a Spark job without using SQL to determine the average pressure at
9am (Pressure9am) for Launceston (location)
>>> import pyspark
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql.functions import *
>>> from pyspark.sql.types import *
>>> import pyspark.sql.functions as func
[>>> from datetime import datetime
[>>> sparksession = SparkSession.builder.appName("Twitter-stream").master("local[
*]").getOrCreate()
[>>> dataframe = sparksession.read.csv("/user/common_data/Spark_Assignment_Datase
t.csv",header = True, inferSchema = True,nullValue = "NA")
[>>> Avg_Pressure_9am = dataframe.filter("Location == 'Launceston']").agg(func.avg
(func.col("Pressure9am")))
[>>> print(" Average Pressure at 9 am where location is Launceston is \n ")
Average Pressure at 9 am where location is Launceston is

[>>> Avg_Pressure_9am.show(vertical = True)
-RECORD 0-----
avg(Pressure9am) | 1015.6792792792787
```