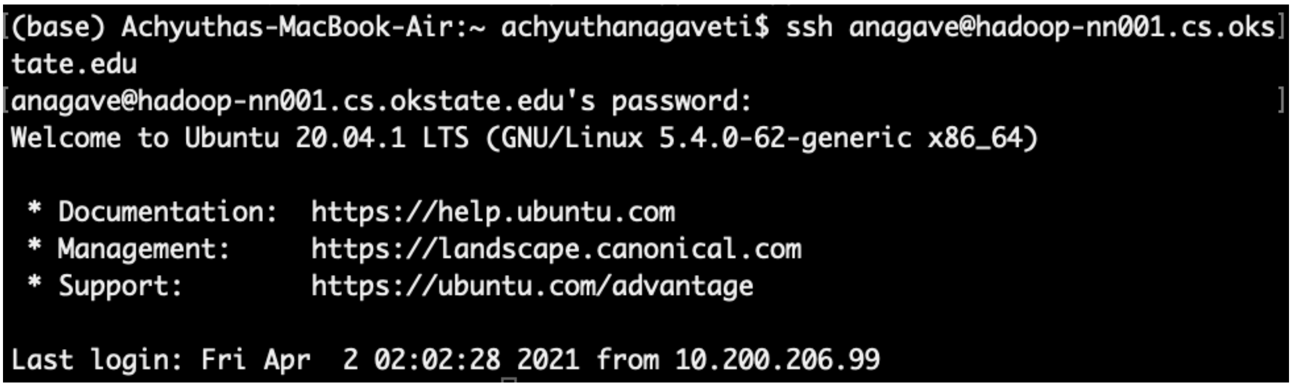
Task-3: Write a Spark job **using SQL** to count number of pressure (Pressure9am) readings at 9am for Launceston (location).

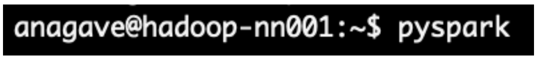
Step 1: Login to HADOOP CLUSTER and Navigate to Pyspark using pyspark command



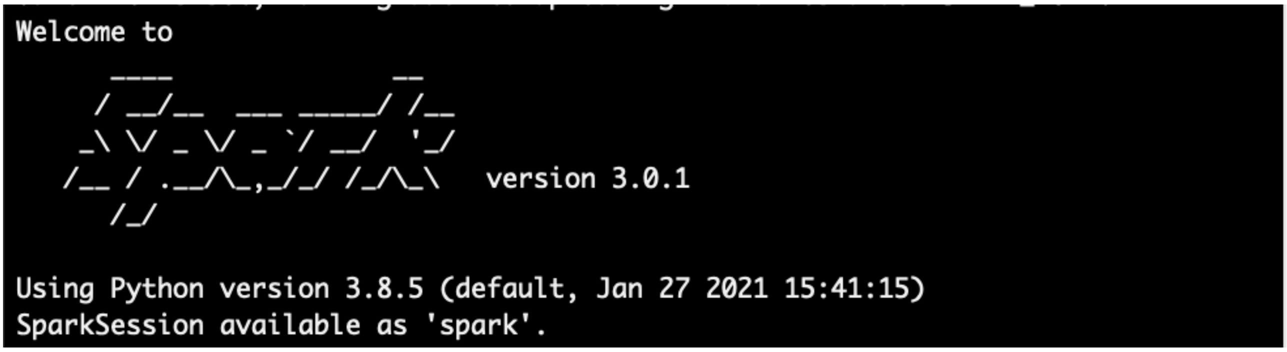
```
[(base) Achyuthas-MacBook-Air:~ achyuthanagaveti$ ssh anagave@hadoop-nn001.cs.oks]
tate.edu
[anagave@hadoop-nn001.cs.okstate.edu's password:                                 ]
Welcome to Ubuntu 20.04.1 LTS (GNU/Linux 5.4.0-62-generic x86_64)

 * 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'.
```

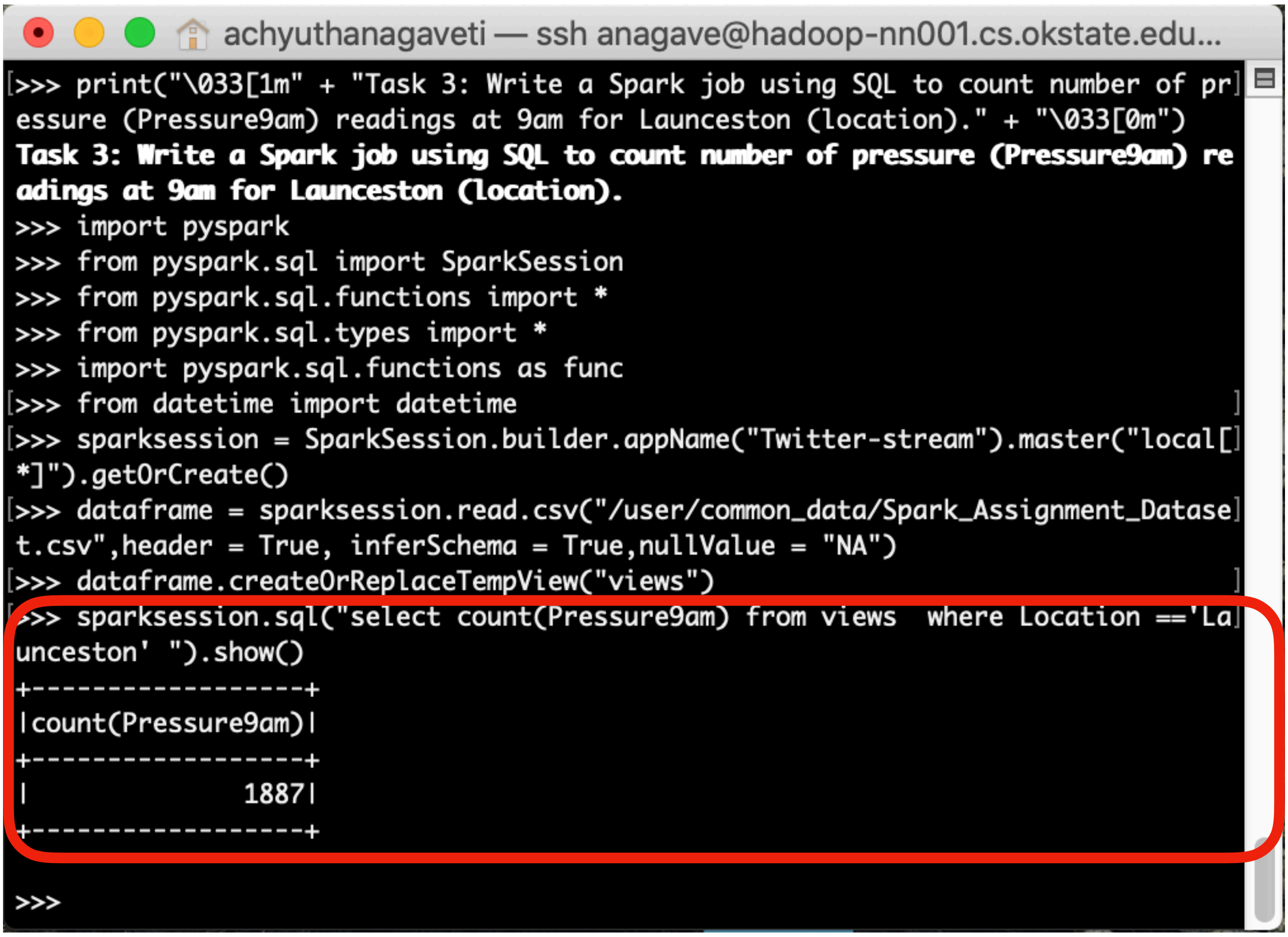
Step 2: CSV file is located in /user/common_data/Spark_Assignment_Dataset.csv path.

Step 3: Import required packages such as SparkSession, pys-ark.sql.functions, pys-ark.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: Create views for data frame and use sql query to obtain output. Results are shown below(Highlighted within red box).



achyuthanagaveti — ssh anagave@hadoop-nn001.cs.okstate.edu...

```
>>> print("\033[1m" + "Task 3: Write a Spark job using SQL to count number of pr
essure (Pressure9am) readings at 9am for Launceston (location)." + "\033[0m")
Task 3: Write a Spark job using SQL to count number of pressure (Pressure9am) re
adings at 9am 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")
>>> dataframe.createOrReplaceTempView("views")
>>> sparksession.sql("select count(Pressure9am) from views  where Location =='La
unceston' ").show()
+------------------+
|count(Pressure9am)|
+------------------+
|              1887|
+------------------+

>>>
```