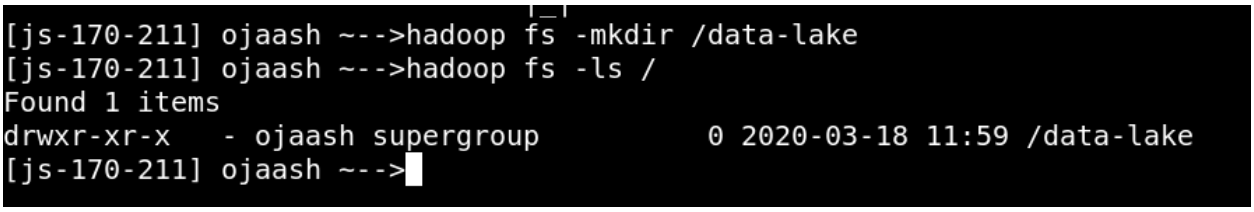
INTRODUCTION TO HDFS:

HDFS is a storage system of Hadoop framework that is used to store files across a cluster of nodes by converting the file to multiple blocks of data and storing them with redundant copying mechanisms. Redundancy is introduced in the filing system because the nodes are usually hosted on hardware terminals which can go down at any moment, hence there must be a backup for all the data chunks that are created. HDFS works particularly well for very large files. The file sizes can be in the range of thousands of Gb to Tb's. HDFS works on the principle of allowing the computing power to go the data rather than getting the data to the computation machine when the size of the data is too large. Another major advantage of HDFS is that it can work on multiple machines supporting multiple software without any compatibility issues.

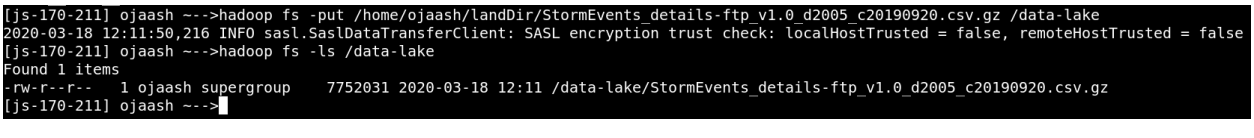
PRACTISING APACHE SPARK AND THE GIVEN QUERIES:

I used the project_utilities python file given in the mini project section of the canvas file section to read the data and to store it in the landing directory. I downloaded the data for the years 2004 and 2005 using the python script and then created a data lake directory for storing the data. I copied the file for the 2005 weather report and put it in the data lake hdfs directory on local host. Then I checked the number of records that were loaded to the data lake directory and decide to view the first few rows of the data to check for the column headings. View distinct event types helps us to view all the possible events that have occurred at least once in the data set. Then I saw how to subset the data only by choosing a few features out of all the available features. The next step that I took was to filter the data by the event type to find out the count and the occurrences of tornado.

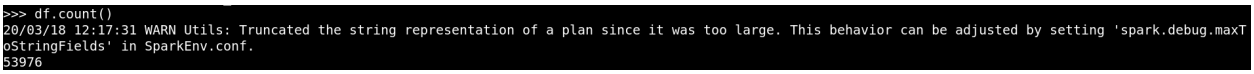
Creating a Data Lake folder:



Transferring the file to the Data-Lake directory:



Checking the number of records loaded:



Viewing the first 3 rows of the data:

```
>>> df.show(3)
+--------------+----------+----------+-------------+--------+--------+----------+--------+---------+----------+----+----------+----------+-------+-------+------
-----+---+------------------+-----------+------------------+----------------+----------------+-------------+---------------+---------------+------------+----
------------+---------+-----------------+------------+--------+-----------+------------+---------------+------------------+------------------+--------------
---+-----------+------------+--------------+----------+-----------+------------+----------+----------+-------+-------+--------------------+--------------------+
-----------+
|BEGIN_YEARMONTH|BEGIN_DAY|BEGIN_TIME|END_YEARMONTH|END_DAY|END_TIME|EPISODE_ID|EVENT_ID|   STATE|STATE_FIPS|YEAR|MONTH_NAME|EVENT_TYPE|CZ_TYPE|CZ_FIPS|     C
Z_NAME|WFO|   BEGIN_DATE_TIME|CZ_TIMEZONE|     END_DATE_TIME|INJURIES_DIRECT|INJURIES_INDIRECT|DEATHS_DIRECT|DEATHS_INDIRECT|DAMAGE_PROPERTY|DAMAGE_CROPS|
       SOURCE|MAGNITUDE|MAGNITUDE_TYPE|FLOOD_CAUSE|CATEGORY|TOR_F_SCALE|TOR_LENGTH|TOR_WIDTH|TOR_OTHER_WFO|TOR_OTHER_CZ_STATE|TOR_OTHER_CZ_FIPS|TOR_OTHER_CZ_N
AME|BEGIN_RANGE|BEGIN_AZIMUTH|BEGIN_LOCATION|END_RANGE|END_AZIMUTH|END_LOCATION|BEGIN_LAT|BEGIN_LON|END_LAT|END_LON|   EPISODE_NARRATIVE|     EVENT_NARRATIVE|
DATA_SOURCE|
+--------------+----------+----------+-------------+--------+--------+----------+--------+---------+----------+----+----------+----------+-------+-------+------
-----+---+------------------+-----------+------------------+----------------+----------------+-------------+---------------+---------------+------------+----
------------+---------+-----------------+------------+--------+-----------+------------+---------------+------------------+------------------+--------------
---+-----------+------------+--------------+----------+-----------+------------+----------+----------+-------+-------+--------------------+--------------------+
-----------+
|         200501|        4|      1800|       200501|      5|     500|   1183593| 5433896|NEBRASKA|        31|2005|   January|Heavy Snow|      Z|     54|     K
IMBALL|CYS|04-JAN-05 18:00:00|        MST|05-JAN-05 05:00:00|              0|                0|            0|              0|           null|        null|EMER
GENCY MANAGER|     null|          null|       null|    null|       null|      null|     null|         null|              null|             null|            n
ull|       null|         null|          null|     null|       null|        null|     null|     null|   null|   null|                null|                null|
        PDS|
|         200501|        4|      1800|       200501|      5|     500|   1183593| 5433897|NEBRASKA|        31|2005|   January|Heavy Snow|      Z|     19|SCOTTS
 BLUFF|CYS|04-JAN-05 18:00:00|        MST|05-JAN-05 05:00:00|              0|                0|            0|              0|           null|        null|EMER
GENCY MANAGER|     null|          null|       null|    null|       null|      null|     null|         null|              null|             null|            n
ull|       null|         null|          null|     null|       null|        null|     null|     null|   null|   null|                null|A winter storm sy...|
        PDS|
|         200501|        1|       500|       200501|      1|    1000|   1183594| 5433898| MONTANA|        30|2005|   January|Heavy Snow|      Z|     32|
CUSTER|BYZ|01-JAN-05 05:00:00|        MST|01-JAN-05 10:00:00|              0|                0|            0|              0|           null|        null|  TR
AINED SPOTTER|     null|          null|       null|    null|       null|      null|     null|         null|              null|             null|            n
ull|       null|         null|          null|     null|       null|        null|     null|     null|   null|   null|New Year's Eve an...|                null|
        PDS|
+--------------+----------+----------+-------------+--------+--------+----------+--------+---------+----------+----+----------+----------+-------+-------+------
-----+---+------------------+-----------+------------------+----------------+----------------+-------------+---------------+---------------+------------+----
```

Viewing distinct event types:

```
>>> df.select("EVENT_TYPE").distinct().show()
+----------------+
|      EVENT_TYPE|
+----------------+
|    Winter Storm|
|Storm Surge/Tide|
|    Volcanic Ash|
|Marine High Wind|
|       Avalanche|
|       Dense Fog|
| Cold/Wind Chill|
|       High Surf|
|  Tropical Storm|
|           Sleet|
|       High Wind|
|       Lightning|
|   Coastal Flood|
|      Dust Devil|
|  Winter Weather|
|        Wildfire|
|     Debris Flow|
|         Tornado|
|    Frost/Freeze|
|      Dust Storm|
+----------------+
only showing top 20 rows
```

Events by Year, Month Name and Locations

```
>>> df.select("YEAR", "MONTH_NAME", "STATE", "EVENT_TYPE").show()
+----+----------+---------+------------+
|YEAR|MONTH_NAME|    STATE|  EVENT_TYPE|
+----+----------+---------+------------+
|2005|   January| NEBRASKA|  Heavy Snow|
|2005|   January| NEBRASKA|  Heavy Snow|
|2005|   January|  MONTANA|  Heavy Snow|
|2005|   January|  MONTANA|  Heavy Snow|
|2005|   January|  MONTANA|  Heavy Snow|
|2005|   January|WISCONSIN|Winter Storm|
|2005|   January| VIRGINIA|   High Wind|
|2005|   January| VIRGINIA|   High Wind|
|2005|   January| VIRGINIA|   High Wind|
|2005|   January| VIRGINIA|   High Wind|
|2005|   January|WISCONSIN|Winter Storm|
|2005|   January| VIRGINIA|   Ice Storm|
|2005|     March| ILLINOIS|        Hail|
|2005|   January| VIRGINIA|   Ice Storm|
|2005|   January|     IOWA|   High Wind|
|2005|   January|     IOWA|   High Wind|
|2005|   January|    TEXAS|        Hail|
|2005|   January| NEBRASKA|  Heavy Snow|
|2005|   January| VIRGINIA|   Ice Storm|
|2005|   January| VIRGINIA|   Ice Storm|
+----+----------+---------+------------+
only showing top 20 rows
```

Tornado events filtered

```
>>> df_filtered.count()
1343
>>> df_filtered.show(2)
+---------------+---------+----------+-------------+-------+--------+----------+--------+--------------+----------+----+----------+----------+-------+-------+
---------+---+------------------+-----------+------------------+---------------+-----------------+-------------+---------------+----------------+------------+-
---------------+---------+---------------+-----------+--------+----------+------------+---------+-----------+------------------+----------------+---------------
-----+-----------+-------------+-------------+--------+-----------+------------+---------+---------+-------+-------+--------------------+--------------------
-+-----------+
|BEGIN_YEARMONTH|BEGIN_DAY|BEGIN_TIME|END_YEARMONTH|END_DAY|END_TIME|EPISODE_ID|EVENT_ID|         STATE|STATE_FIPS|YEAR|MONTH_NAME|EVENT_TYPE|CZ_TYPE|CZ_FIPS|
  CZ_NAME|WFO|   BEGIN_DATE_TIME|CZ_TIMEZONE|     END_DATE_TIME|INJURIES_DIRECT|INJURIES_INDIRECT|DEATHS_DIRECT|DEATHS_INDIRECT|DAMAGE_PROPERTY|DAMAGE_CROPS|
         SOURCE|MAGNITUDE|MAGNITUDE_TYPE|FLOOD_CAUSE|CATEGORY|TOR_F_SCALE|TOR_LENGTH|TOR_WIDTH|TOR_OTHER_WFO|TOR_OTHER_CZ_STATE|TOR_OTHER_CZ_FIPS|TOR_OTHER_CZ
 NAME|BEGIN_RANGE|BEGIN_AZIMUTH|BEGIN_LOCATION|END_RANGE|END_AZIMUTH|END_LOCATION|BEGIN_LAT|BEGIN_LON|END_LAT|END_LON|   EPISODE_NARRATIVE|     EVENT_NARRATIV
E|DATA_SOURCE|
+---------------+---------+----------+-------------+-------+--------+----------+--------+--------------+----------+----+----------+----------+-------+-------+
---------+---+------------------+-----------+------------------+---------------+-----------------+-------------+---------------+----------------+------------+-
---------------+---------+---------------+-----------+--------+----------+------------+---------+-----------+------------------+----------------+---------------
-----+-----------+-------------+-------------+--------+-----------+------------+---------+---------+-------+-------+--------------------+--------------------
-+-----------+
|         200501|       13|      1852|       200501|     13|    1853|   1182447| 5434191|SOUTH CAROLINA|        45|2005|   January|   Tornado|      C|     65|
MCCORMICK|CAE|13-JAN-05 18:52:00|        EST|13-JAN-05 18:53:00|              0|                0|            0|              0|               0|           0|N
WS STORM SURVEY|     null|          null|       null|    null|        F0|        .2|       40|         null|              null|             null|
 null|          2|           NE|    PARKSVILLE|        2|         NE|  PARKSVILLE|    33.78|   -82.20|  33.78| -82.20|An approaching fr...|Storm Survey foun..
.|        PDS|
|         200501|       12|      2056|       200501|     12|    2057|   1183929| 5435571|      MISSOURI|        29|2005|   January|   Tornado|      C|    153|
    OZARK|SGF|12-JAN-05 20:56:00|        CST|12-JAN-05 20:57:00|              0|                0|            0|              0|             50K|          0K|N
WS STORM SURVEY|     null|          null|       null|    null|        F0|         1|      100|         null|              null|             null|
 null|       null|         null|      TECUMSEH|     null|       null|    TECUMSEH|    36.58|   -92.28|  36.58| -92.28|                null|A brief tornado t..
.|        PDS|
+---------------+---------+----------+-------------+-------+--------+----------+--------+--------------+----------+----+----------+----------+-------+-------+
---------+---+------------------+-----------+------------------+---------------+-----------------+-------------+---------------+----------------+------------+-
---------------+---------+---------------+-----------+--------+----------+------------+---------+-----------+------------------+----------------+---------------
-----+-----------+-------------+-------------+--------+-----------+------------+---------+---------+-------+-------+--------------------+--------------------
-+-----------+
only showing top 2 rows
```

IMPLEMENTING A CUSTOM QUERY:

Question: Use the filter () function to count how many events of each type were reported by each source (SOURCE) in Georgia for your selected year. Write commands separately first and then chain them together.

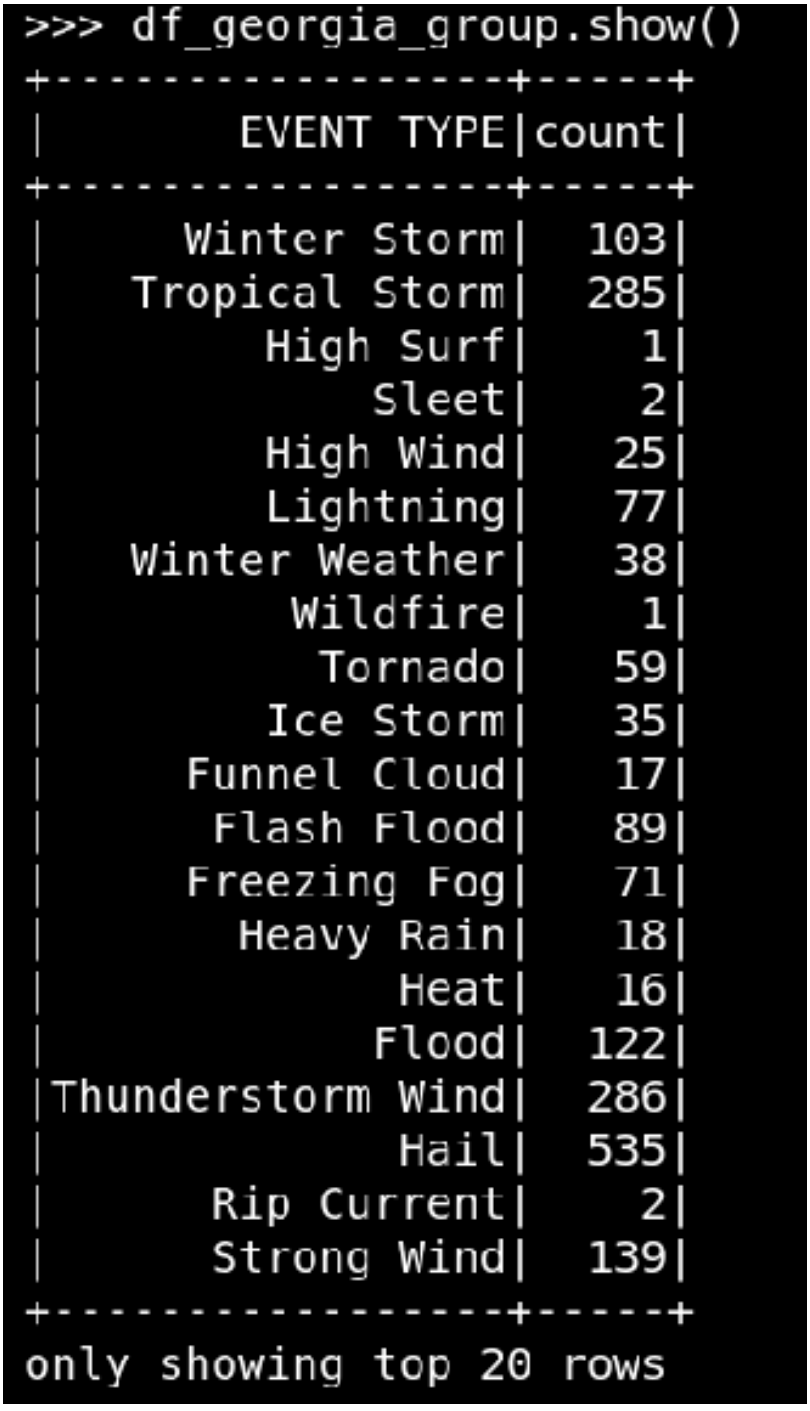
Here I first decide to write the command to filter the data by the State name as Georgia. Then I decided to have a look at the number of records that exist for Georgia using the count () function.

```
>>> df_georgia_filter = df.filter(df.STATE == "GEORGIA")
>>> df_georgia_filter.count()
2141
>>>
```

The next step was grouping the subset data for the Georgia state using the event types in order to get a count of the events that took place.

```
>>> df_georgia_group = df_georgia_filter.groupby(df.EVENT_TYPE).count()
>>> df_georgia_group.count()
21
```

Then I used the show function to display the data that I had created using the groupby command.



```
>>> df_georgia_group.show()
+------------------+-----+
|        EVENT TYPE|count|
+------------------+-----+
|      Winter Storm|  103|
|    Tropical Storm|  285|
|         High Surf|    1|
|             Sleet|    2|
|         High Wind|   25|
|         Lightning|   77|
|    Winter Weather|   38|
|          Wildfire|    1|
|           Tornado|   59|
|         Ice Storm|   35|
|       Funnel Cloud|   17|
|        Flash Flood|   89|
|       Freezing Fog|   71|
|         Heavy Rain|   18|
|               Heat|   16|
|              Flood|  122|
|Thunderstorm Wind|  286|
|               Hail|  535|
|       Rip Current|    2|
|       Strong Wind|  139|
+------------------+-----+
only showing top 20 rows
```

Finally, I combined the above syntax into a single statement and decided to print the data in ordering by the event types feature.

```
>>> df_georgia_agg = df.filter(df.STATE == "GEORGIA").groupby(df.MONTH_NAME, df.EVENT_TYPE).count().orderBy(df.EVENT_TYPE)
>>> df_georgia_agg.count()
87
>>> df_georgia_agg.show()
+----------+------------+-----+
|MONTH NAME|  EVENT TYPE|count|
+----------+------------+-----+
|     March| Flash Flood|   22|
|      July| Flash Flood|   45|
|   October| Flash Flood|    8|
|      June| Flash Flood|    7|
|    August| Flash Flood|    7|
|      July|       Flood|   57|
|      June|       Flood|    9|
|    August|       Flood|    8|
|     March|       Flood|   39|
|   October|       Flood|    5|
|  February|       Flood|    1|
|     April|       Flood|    3|
|  December|Freezing Fog|   71|
|    August|Funnel Cloud|    8|
|       May|Funnel Cloud|    1|
|  December|Funnel Cloud|    8|
|  February|        Hail|   88|
|  November|        Hail|    1|
|   October|        Hail|    9|
|     April|        Hail|  101|
+----------+------------+-----+
only showing top 20 rows
```

The implementation of the above queries was not that difficult. I could easily figure out the steps to be followed to subset the data and to apply the filters and groupby as well as orderBy commands. I f we want we could have ordered the data to be displayed by the month name or the count as well. We can easily us the above code to study the data for different states as well as for different features and to get the proper conclusions.

CONCLUSION:

This assignment helps us to understand the different intricate features of HDFS which can be exploited to make the big data file storage, handling and analytics easier than any other framework. The Hadoop hdfs file systems can be accessed and queried very easily using the aggregate, groupby, orderby, filter, show, count and many more complex commands. The data can be subset and queried to gte the results as desired as well as the data can be sent to any other machine for the visualization application and storage.