Assignment #5

1. **[Hadoop] (35)**

   The attached CSV file contains hourly normal recordings for temperature and dew point temperature at Asheville Regional Airport, NC, USA. The unit of measurement is tenth of a degree Fahrenheit. So, 344 is 34.4 F.

   Write a program using Hadoop to compute and output daily average measurements for temperature and dew point temperature. The daily average measurements should include measurements for 24-hour period, for example from 20100101 00:00 (2010, January 1st, 00:00) to 20100101 23:00 (2010, January 1st, 23:00). Output the result to text file(s) in the format shown below - the columns are date and the combined result (separated by comma) of daily temperature and daily dew point temperature:

   ```
   20100101    377.04, 285.58
   20100102    378.67, 286.92
   ....        ...., ....
   ```

   You may write the application in Java, C/C++ or Python language. Provide both source code and compiled code, if applicable, for your program, as well as the output file.

2. **[Spark RDDs] (35)**

   Consider the CSV file containing hourly normal recordings of temperature and dew point temperature at Asheville Regional Airport, NC, USA.

   Write a program using Spark RDDs (not DataFrames) to compute and output daily average measures for temperature and dew point temperature. The daily average measurements should include measurements for 24-hour period, for example from 20100101 00:00

(2010, January 1st, 00:00) to 20100101 23:00 (2010, January 1st, 23:00). Output the result to text or CSV file(s) in the format shown below - the columns are date and the combined result (separated by comma) of daily temperature and daily dew point temperature:

```
20100101        377.04, 285.58
20100102        378.67, 286.92
....            ...., ....
```

Write you application in Java, Scala or Python. Provide source code and compiled code, if applicable, for your program as well as the output data file.

3. **[PySpark DataFrames] (30)**

Consider the CSV file containing hourly normal recordings of temperature and dew point temperature at Asheville Regional Airport, NC, USA.

Write a program using PySpark and its DataFrame APIs to compute daily average measures for temperature and dew point temperature. The daily average measurements should include measurements for 24-hour period, for example from 20100101 00:00 (2010, January 1st, 00:00) to 20100101 23:00 (2010, January 1st, 23:00). Output the result to text or CSV file(s) in the format shown below - the columns are date and the combined result (separated by comma) of daily temperature and daily dew point temperature:

```
20100101        377.04, 285.58
20100102        378.67, 286.92
....            ...., ....
```

Write you application in Python. Provide the source code for your program as well as the output data file.