**Project#5: Big Data with PySpark using Anaconda & Jupyter notebook.**

Do not wait till the last week to run the programs.

**Introduction:**

1. Download Anaconda Python: <https://www.anaconda.com/products/distribution>
2. Download Apache Spark™ <https://spark.apache.org/downloads>

* Choose a Spark release: 3.3.0 (Jun 16 2022)
* Choose a package type:  Pre-built for Apache Hadoop 3.3 and later
* Download Spark: [spark-3.3.0-bin-hadoop3.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-3.3.0/spark-3.3.0-bin-hadoop3.tgz)
* Double-click the file to install Spark

For MAC/Linux: Set environment variables after “conda initialize” line:

# <<< conda initialize <<<

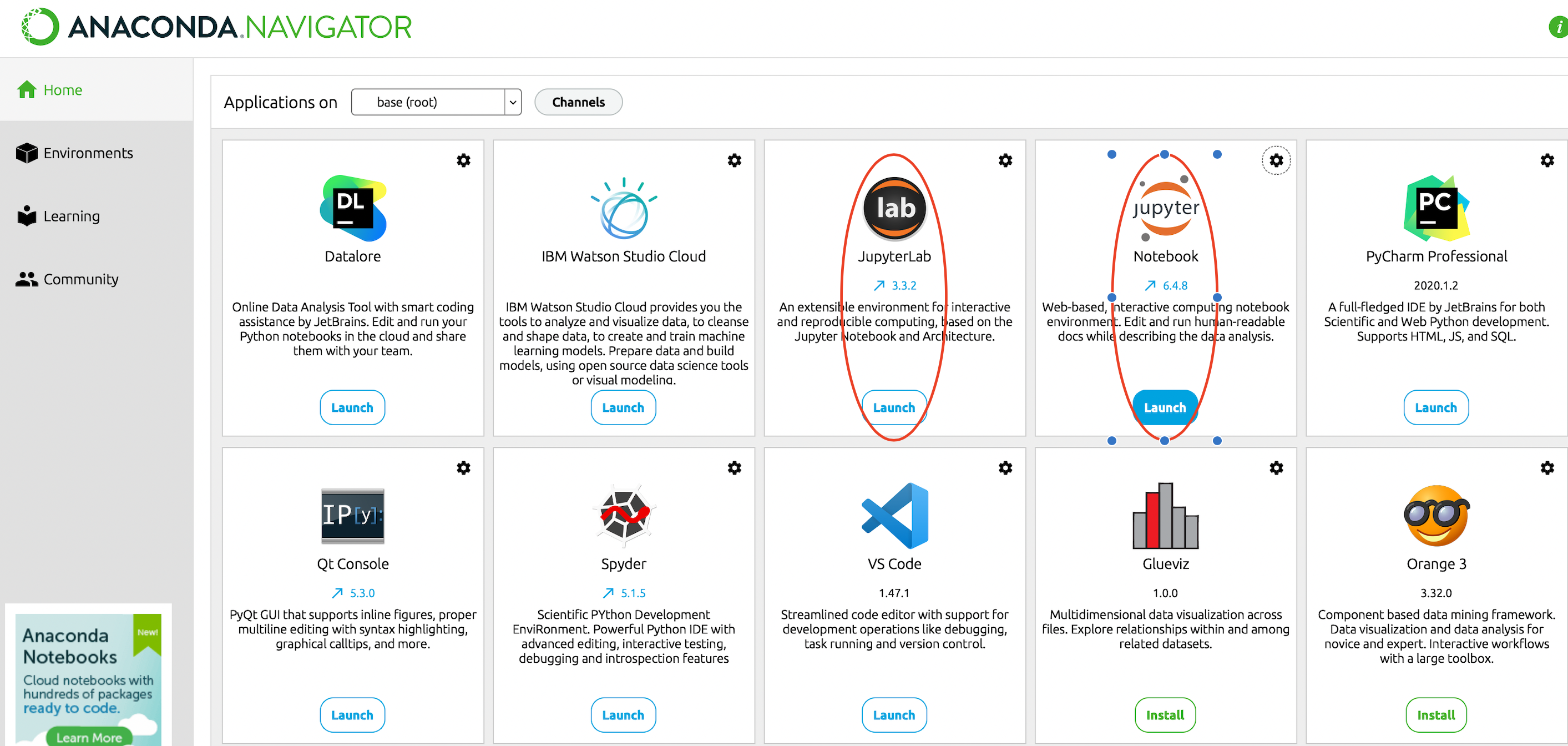
export SPARK\_HOME=/Users/r188694/spark/spark-3.3.0-bin-hadoop3

export HADOOP\_HOME=$SPARK\_HOME

* Install PySpark Python libraries: pip install pyspark

1. Launch Jupyter Notebook

* Open Anaconda-Navigator App
* Launch Jupyter Notebook (or Jupyter Lab)



1. Get Weather data from: [NCEI Bulk Download (CSV)](https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/) (https://www.ncei.noaa.gov/data/global-summary-of-the-day/access/)
2. Download the Weather data sets listed below from 2010 to 2022. Copy each year data is in the respective directory (i*.e., 2010 directory have data for the year 2010*)
3. 2010/01052099999.csv
4. 2010/99407099999.csv
5. 2011/01008099999.csv
6. 2011/01046099999.csv
7. 2012/01023099999.csv
8. 2012/01044099999.csv
9. 2013/01001499999.csv
10. 2013/01008099999.csv
11. 2014/01008099999.csv
12. 2014/01023099999.csv
13. 2015/01008099999.csv
14. 2015/01025099999.csv
15. 2016/01008099999.csv
16. 2016/01023199999.csv
17. 2017/01008099999.csv
18. 2017/01023099999.csv
19. 2018/01008099999.csv
20. 2018/01025099999.csv
21. 2019/01008099999.csv
22. 2019/01023099999.csv
23. 2020/01008099999.csv
24. 2020/01023099999.csv
25. 2021/01062099999.csv
26. 2021/01065099999.csv
27. 2022/01241099999.csv
28. 2022/02095099999.csv
29. Additional data information can be found at <https://www.ncei.noaa.gov/data/global-summary-of-the-day/doc/readme.pdf>
30. Use spark to analyze the above-mentioned data.
31. Example commands used in the class: pyspark.ipynb
32. For additional learning materials on Spark: <https://sparkbyexamples.com/pyspark-tutorial/>
33. Ignore the missing values for the dataset (Details on missing value can be found in readme.pdf).

**Tasks:**

1. **Find the hottest day (column MAX) for each year, and provide the corresponding station code, station name and the date (columns STATION, NAME, DATE).**
2. **Find the coldest day (column MIN) for the month of January across all years (2010 - 2022) , and provide the corresponding station code, station name and the date (columns STATION, NAME, DATE)..**
3. **Maximum and Minimum precipitation (column PRCP ) for the year 2015,** **and provide the corresponding station code, station name and the date (columns STATION, NAME, DATE).**
4. **Count percentage missing values for wind gust (column GUST) for the year 2019.**
5. **Find the mean, median, mode and standard deviation of the Temperature (column TEMP) for each month for the year 2020.**

**Submission:**

1. **Your result.txt file with all the above information**
2. **Your spark code/command**