

Guide for HW2

Useful Resources for HW2:

- <http://www.evontech.com/what-we-are-saying/entry/apache-hadoop-vs-apache-spark-two-popular-big-data-frameworks-compared.html>
- <https://spark.apache.org/docs/2.1.1/ml-guide.html>
- <http://data-flair.training/blogs/apache-spark-rdd-vs-dataframe-vs-dataset/>

Some Datasets you can check :

- <https://www.kaggle.com/datasets>
- <https://archive.ics.uci.edu/ml/index.php>
- https://en.wikipedia.org/wiki/Wikipedia:Database_download

Personally, I would recommend spark/pyspark as your primary big data framework.