from Super User 019 to All Participants:

https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/CapacityScheduler.html

from Super User 019 to All Participants:

<https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-site/FairScheduler.html>

flume;

<https://www.eduonix.com/blog/bigdata-and-hadoop/flume-installation-and-streaming-twitter-data-using-flume/>

to format the data

freeformater website

Maven

<https://acadgild.com/blog/building-a-hadoop-application-using-maven/>

<https://zoom.us/webinar/register/0584588dd535f15d7c24e00bf0acd2b8>

<https://acadgild.com/blog/building-hadoop-application-using-gradle/>

blogspot

<http://hadooped.blogspot.in/2013/09/reduce-side-joins-in-java-map-reduce.html>

matrix multiplication

<http://hadoopgeek.com/mapreduce-matrix-multiplication/>

secondary sorting

<http://codingjunkie.net/secondary-sort/>

mr unit testing

<http://codingjunkie.net/testing-hadoop-programs-with-mrunit/>

<https://acadgild.com/blog/how-to-test-hadoop-jobs-with-mrunit/>

Sampling

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Sampling>

hive

<https://parquet.apache.org/>

hive

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+WindowingAndAnalytics>

spark full doc

<http://spark.apache.org/docs/latest/graphx-programming-guide.html>

oozie

<https://www.edx.org/course/introduction-r-data-science-microsoft-dat204x-4?utm_source=sailthru&utm_medium=email&utm_campaign=studentnewsletter-actives-20170509&utm_content=Final&utm_term=NL%20Actives>

<http://oozie.apache.org/docs/4.3.0/DG_QuickStart.html>

sqoop

<https://www.digitalocean.com/community/tutorials/how-to-create-a-new-user-and-grant-permissions-in-mysql>

oozie

<https://acadgild.com/blog/oozie-job-scheduling-in-hive/>

imp blogs

<https://eldadlevy.wordpress.com/2011/02/05/hadoop-binary-files-processing-entroduced-by-image-duplicates-finder/>

<https://github.com/jt6211/hadoop-binary-analysis>

<https://github.com/endgameinc/binarypig>

<https://hortonworks.com/tutorials/>

courses for free

<https://bigdatauniversity.com/courses/>

Mandatory courses:

Spark

nosqlDB

mongoDB/couchdatabase(preferrable)

<https://training.couchbase.com/online>

data science and its related technologies and concepts

-statistics

<https://bigdatauniversity.com/courses/statistics-101/>

some data science courses

<https://www.edx.org/course/subject/data-analysis-statistics>

visualization tools

-R as a visualization tool

-Tableau

<https://www.tableau.com/learn/training>

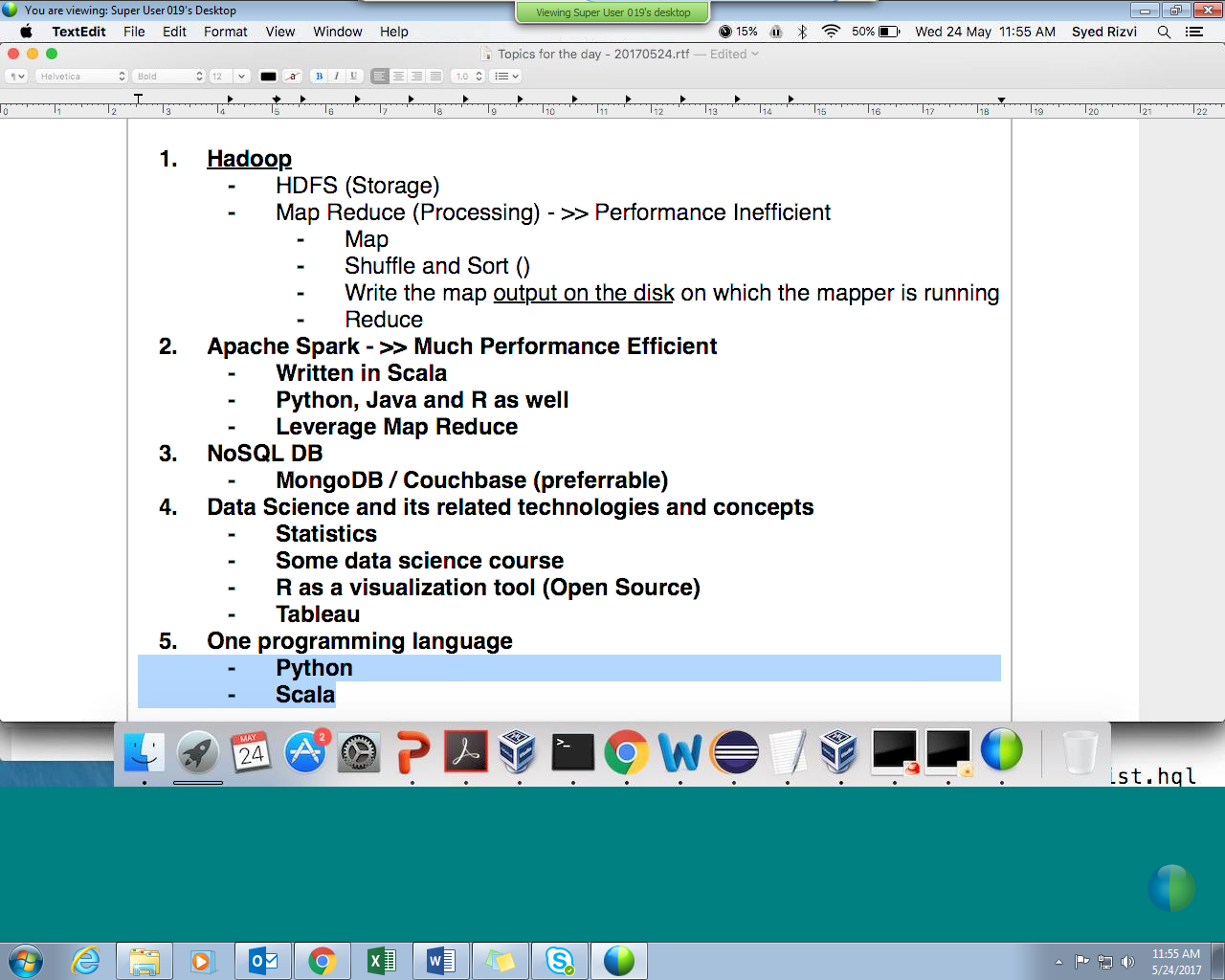
mandatatory programming language

- python

- scala

Projects

https://acadgild.com/blog/



tutorials:

<https://hortonworks.com/tutorials/>

<https://www.cloudera.com/more/training/certification/cca-spark.html>

[masroor.rizvi@gmail.com](mailto:masroor.rizvi@gmail.com)