* Load
* Filter
* Viualize
* Large real-world dataset
* Cloud-based distributed computing environment
* Hadoop
* Spark
* S3
* Google Ngram
  + Google books
  + 4%
  + 1800-2000
* Host
* S3 bucket
* Amazon S3 open data registry
* CSV
* Public S3 bucket
* Big data analytics workflow
  + Public bucket
  + HDFS
  + Spark
  + Head node
  + Personal bucket
  + Local machine
* Step
  + AWS console
  + EMR cluster
    - Hadoop
    - Hive
    - Spark
    - Jupyterhub
  + Head node: SSH
  + HDFS
    - distcp
  + Read
    - Jupyterhub
      * EMR
    - Pyspark
      * Terminal
  + Schema
  + Row
  + Query: sparksql
  + Row
  + Write
  + File
  + S3 bucket
  + Pandas dataframe
    - Command line
  + Frequency
* Requirement
  + Screen shot
  + Command line
  + Jupyter notebook