DE project

Data exploration

- [ ] Explore million song dataset
- [ ] Identify Genre information

Preprocessing and data reduction

- [ ] data cleaning - remove missing value
  remove duplicate
- [ ] feature extraction - find most relevant feature
- [ ] data reduction

Data processing solution

- [ ] Distributed data storage (HDFS or cloud based storage)
- [ ] Processing framework
  Apache SPARK or ? Deploy own cluster
- [ ] parallel processing

ML for Genre classification

- [ ] Model selection (e.g. Knn, SVM or ?)
- [ ] training validation
- [ ] parameter tuning

Computational experiment

Dataset chosen
Analysis group
metadata

Report