Course project guidance

Romit Maulik, IST & ICDS

Email: rmaulik@psu.edu

PennState

Course project

Basic guidelines:

1. Propose a **big-data** project that leverages PySpark functionality (regression, classification, clustering, etc). Look at the syllabus.
2. Pick a problem that you **cannot solve on your laptop**. However, note that the work directory allows for 100GB storage at most.
3. Execute project on Roar Collab for varying numbers of data partitions and workers and profile the scaling of the project (in cluster mode).
4. Note - *purely jupyter-server based submissions will not be enough to get a competitive score.*
5. Use the weekly reports! These are your own highway markers.

Miniproject grading

Your final product (50% of entire project grade) will be graded by the instruction team and by your fellow classmates:

1. Peer-review from your own group members.
2. Your final presentations will be graded by other teams (in addition to the TAs and myself)
3. You must provide a CREDIT statement in your submission (see elsevier CREDIT statement) that I will compare with intra-group peer review. This ties into your division of labor.

Why are we doing this? Big-data projects in the wild are very collaborative!

Where to find data?

- Kaggle
- Google dataset search
- Data.gov
- UCI Machine Learning Repository
- https://www.earthdata.nasa.gov/
- Global Health Observatory Data Repository
- FBI Crime Data Explorer
- BFI film industry statistics
- NYC Taxi data
- https://dataverse.harvard.edu/

Miniproject proposal: Examples

- Pattern/sentiment analysis in documents/social media.
- Machine learning for classification or regression.
- Time-series analyses (Real-time Stream Processing)
- Recommendation systems.

In most cases - you *will* need to go beyond the material taught in this course. Please start looking at MLlib functionality ahead of time (I can help you if you have questions).

Use your instruction team to guide you on your projects.