

# Cloud Computing and Big Data Analytics

## 2022 Fall

### Homework 5:

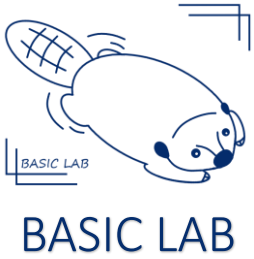
### Recommendation System with PySpark

---

TA：曾偉倫

Email: [wlt seng. ee06@nycu. edu. tw](mailto:wlt seng. ee06@nycu. edu. tw)

# Outline



- Introduction
- Problem Description
- Dataset
- Grading Policy
- Submission
  - Kaggle Competition
  - E3
- Timeline









# Introduction

- Recommendation system makes e-commerce platform better.

Your recently viewed items and featured recommendations




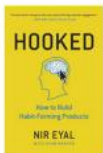




Sponsored products related to this search [What's this?](#)

Page 1 of 3

							
All-new Echo Show (2nd Gen) + Ring Video Doorbell 2- Charcoal 1 offer from \$428.99	AmazonBasics Microwave, Small, 0.7 Cu. Ft, 700W, Works with Alexa ★★★★☆ 1,375 \$59.99 ✓prime	Echo Look   Hands-Free Camera and Style Assistant with Alexa—includes Style Check to... ★★★★☆ 413 \$99.99 ✓prime	Sonos Beam - Smart TV Sound Bar with Amazon Alexa Built-in - Black ★★★★☆ 474 \$399.00 ✓prime	Echo Wall Clock - see timers at a glance - requires compatible Echo device ★★★★☆ 1,231 \$29.99 ✓prime	Echo Spot Adjustable Stand - Black ★★★★☆ 933 \$19.99 ✓prime	AHASTYLE Wall Mount Hanger Holder ABS for New Dot 3rd Generation Smart Home Speakers... ★★★★☆ 12 \$10.99 ✓prime	Angel Statue Crafted Stand Holder for Amazon Echo Dot 3rd Generation, Alexa Smart... ★★★★☆ 57 \$25.99 ✓prime

Explore more from across the store

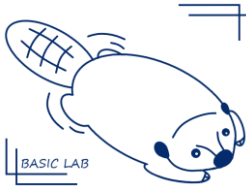
Page 1 of 6

							
Actionable Gamification: Beyond Points, Badges, ... Yu-kai Chou	The Model Thinker: What You Need to Know to ... Scott E. Page	Don't Make Me Think, Revisited: A Common ... Steve Krug	Hooked: How to Build Habit-Forming Products Nir Eyal	Microservices Patterns: With examples in Java Chris Richardson	Solving Product Design Exercises: Questions & ... Artiom Dashinsky	100 Things Every Designer Needs to Know About ... Susan Weinschenk	Infinity Jonathan Hickman ★★★★☆ 182

# Problem Description

- Given : user-item pairs
- Objective: Predict ratings
- Limitation:
  - Build your recommendation system with PySpark framework.
  - PySpark ML-lib or Other Model-based methods

# Dataset

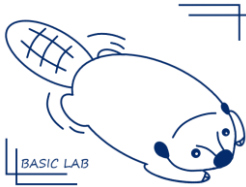


BASIC LAB

- File List
  - train\_student.csv
    - item,user,rating
  - train\_meta.csv
    - user,item,rating with additional with additional review content
  - test\_public.csv (50% of testing set)
    - item,user,rating
  - test\_private.csv (50% of testing set)
    - item,user
  - test\_all.csv
    - item,user



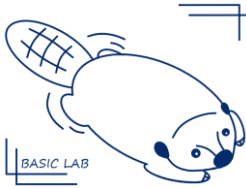
# Grading Policy



BASIC LAB

- Kaggle Competition Rank
  - Over baseline:
    - Top 20% : 100 pts
    - 20% ~ 30%: 97 pts
    - 30% ~ 40%: 95 pts
    - 40% ~ 50%: 93 pts
    - 50% ~ 60%: 91 pts
    - 60% ~ 70%: 89 pts
    - 70% ~ 80%: 87 pts
    - 80% ~ 90%: 85 pts
    - 90% ~ 100%: 83 pts(> baseline)
  - Valid submission, but  $\leq$  baseline benchmark: 75 pts
  - Invalid submission, but submit code & readme to E3: 50 pts  
**(Student must write down what you have done and tell TA what Algorithm / Method you plan to implement. )**

# Submission



BASIC LAB

- Kaggle
  - Register with invitation link: [<URL>](#)
  - Team name: <student\_ID> .
  - .csv file, containing two columns
    - Column name: U\_I,rating (I is capital "i")
    - U\_I: user-item pairs, unique identifier (string)
    - rating: predicted rating (double)
  - Make sure your total line number in submission file is 1 (column name) +83799 (U\_I,rating pair).
  - 10 submissions per day

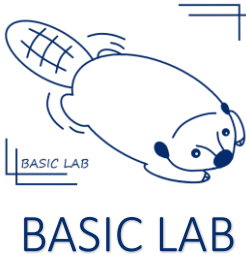
```
U_I = str(user) + "_" + str(item)
user = 12345
item = CXGZE
```

```
U_I = 12345_CXGZE
```

Example submission:

```
U_I,rating
12345CXGZE,5
12345CXGZK,1
...
```

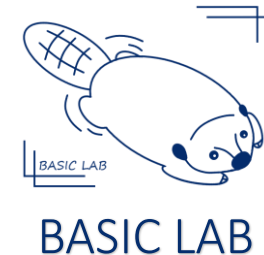
# Submission



- E3
  - <studet\_ID>.zip : Source code (in one zip file)
  - <studet\_ID>.pdf : Readme File
    - 1 page A4
    - Tell TA how to execute your code
    - Briefly explain your method
    - Reference



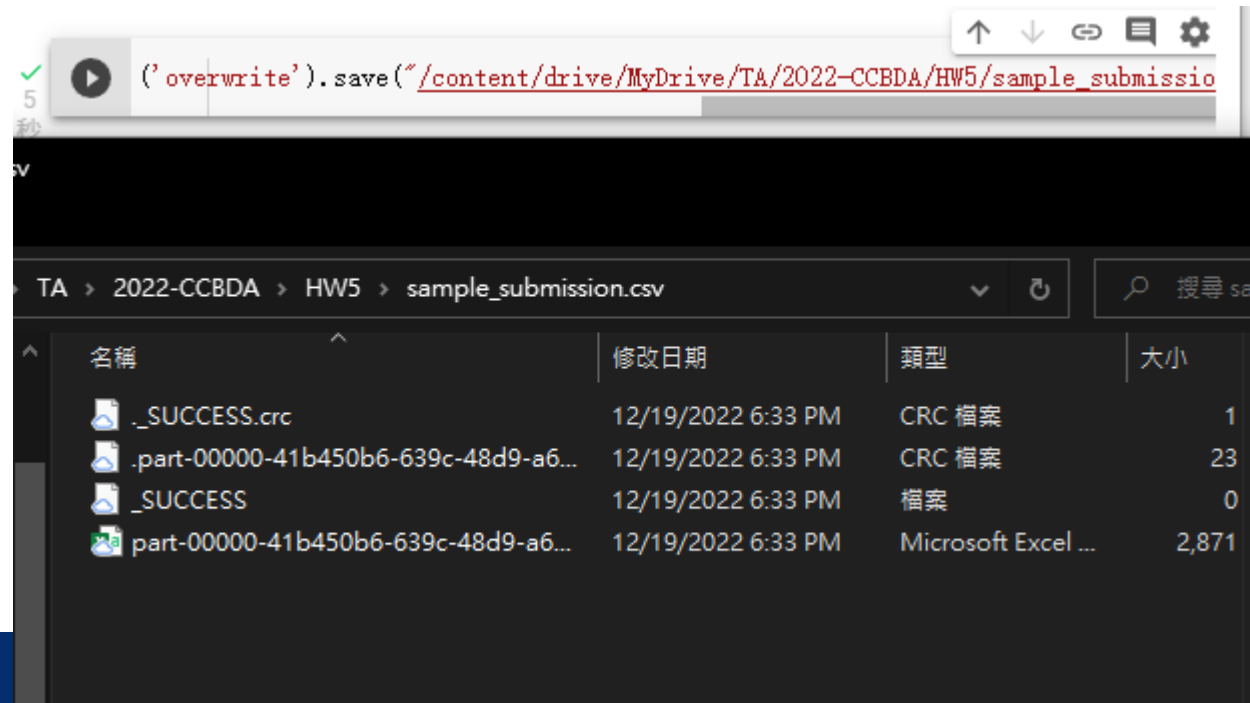
# Timeline



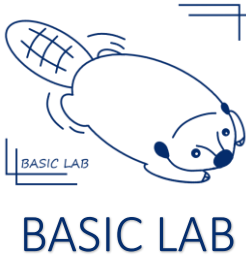
- 12/13: HW5 release
- 12/21-12/26: Kaggle Competition Submission (10 times per day)
- 12/30 : E3 Submission Deadline (before 23:59:59)
- Any problems? Contact TA:
  - TA : 曾偉倫
  - Email: [wtseng.ee06@nycu.edu.tw](mailto:wtseng.ee06@nycu.edu.tw)
  - TA Hour needs to make an appointment

# Appendix

- PySpark will output csv file in the following screenshot.
  - Please select the csv file in your output path.



# Appendix



- Useful example project:
  - [Building a Recommendation System with Spark ML and Elasticsearch | by Lijo Abraham | Towards Data Science](#)
  - [PySpark Recommender System with ALS | Towards Data Science](#)

