

SPRING 2024
DSE I2450 - 3GG
BIG Data &
Scalable Computation

Week 3
Spark + PySpark



Dall-E 3: "inside a big data system, use your imagination"

If you burned all of
the data created in
just 1 days onto DVDs,
you could stack them
on top of each other
and reach the moon -
twice.
(Marr, 2014 !!)



Margaret Hamilton, NASA code (1969)



Plan for today:

- Review of MapReduce, finish demo
- Group work: RDDs + Spark
- Share back, questions
- Demo: Spark on Databricks
- Upcoming schedule: async weeks!
Assignments, housekeeping


Week 4, Feb. 28:

- No class, **make (min.) 1 update post by Friday**
- **Programming assignment shared by Feb. 26th**
- Assignment = Databricks notebook, w/instructions
- Work on programming assignment + research
- **Choosing topics + pairs**


Week 5, March 6th:

- No class, **make (min.) 1 update post by Friday**
- **Work on programming assignment + bibliography**

March 8th by 11:59pm:

- **Programming assignment due!** Codecademy complete 
- Will be an upload form to submit

Week 6, March 13th:

- Sync class resumes, on zoom
- **Pairs, bibliography + abstract due,** by 11:59pm 
- I will give feedback ASAP after so you can start reading