

MSDS 694

Distributed Computing

DIANE WOODBRIDGE, PH.D



Your Team

Three to Five people per team.

- Everyone needs to work !! (We are going to survey.)



UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

Your Team

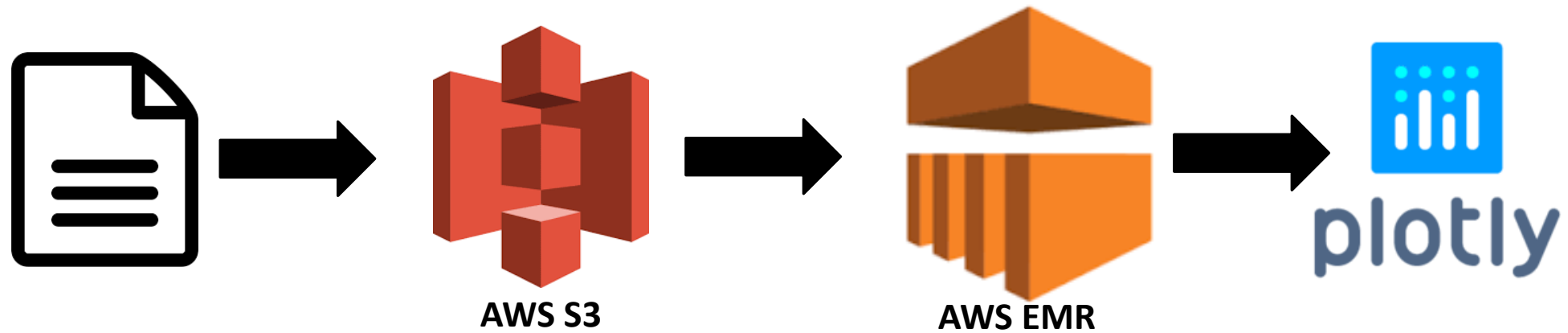
Three to Five people per team.

- Everyone needs to join Group under People (on Canvas)

The screenshot shows the Canvas LMS interface for the course MSDS-694-02. The left sidebar contains navigation links: Home, Modules, Assignments, Grades, People, Pages, Quizzes, Echo360 ALP, and a search bar. The main content area is titled 'Groups' and shows a list of six 'FINAL PROJECT GROUP' entries. Each entry has '0 students' and a lock icon. A red box highlights the group list, and a red 'X' marks the '+ Group' button.

Group Name	Students	Status
FINAL PROJECT GROUP 1 FINAL PROJECT GROUP	0 students	Locked
FINAL PROJECT GROUP 2 FINAL PROJECT GROUP	0 students	Locked
FINAL PROJECT GROUP 3 FINAL PROJECT GROUP	0 students	Locked
FINAL PROJECT GROUP 4 FINAL PROJECT GROUP	0 students	Locked
FINAL PROJECT GROUP 5 FINAL PROJECT GROUP	0 students	Locked
FINAL PROJECT GROUP 6 FINAL PROJECT GROUP	0 students	Locked

Group Project



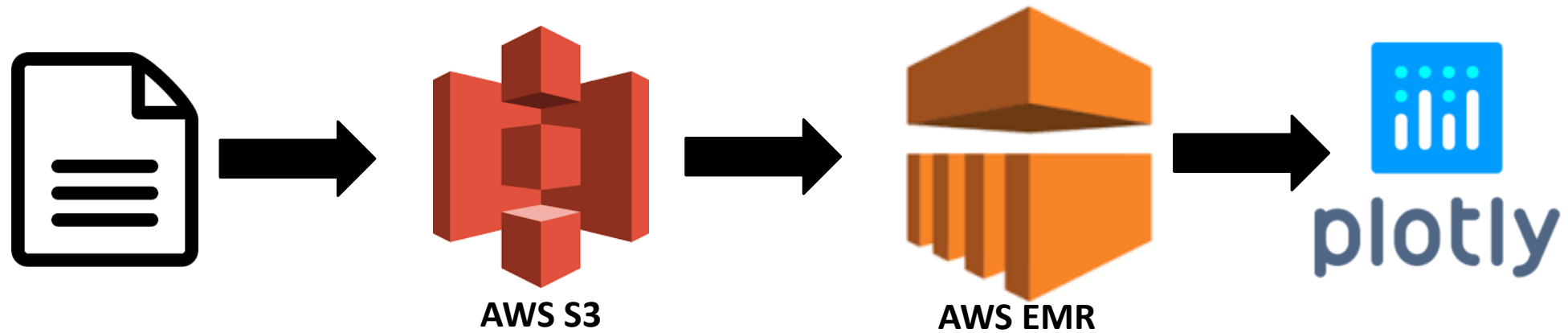
Step 1. Choose Data Sets

Step 2. Load to S3

Step 3. Apply data preprocessing to retrieve interesting stats/trends.

Step 4. Visualize the outcome.

Group Project



Step 1. Choose Data Sets

Step 2. Load to S3

Step 3. Apply data preprocessing to retrieve interesting stats/trends.

Step 4. Visualize the outcome.

Which Data?



Some Data Example

Awesome Public Data : <https://github.com/awesomedata/awesome-public-datasets>

Kaggle Data Sets : <https://www.kaggle.com/datasets>

UC Irvine Machine Learning Repository : <https://archive.ics.uci.edu/ml/index.php>

Data.gov : <https://www.data.gov/>

Registry of Open Data on AWS : <https://registry.opendata.aws/>

Requirements

Submit a 1-page data description.

1. Each group member proposes one or two topics (3pt).

- List Student Name, Data titles, List data sources (URLs), Size, Reasons why you chose, Possible analytic goals.
- If you are planning to collect your own data (web crawling, smartphone application, IoT application, etc.), please describe your specific plans and timeline.
- Data should be at least 2GB and over 1M records.

2. Each group chooses one data set (2pt).

- Describe reasons why you chose.

Things to Be Considered

Novel data analytics goals

- Data fusion from multiple data sources
- Develop(or Apply) and validate novel algorithms
- Compare results of different algorithms
- Compare results from different machine specs- Costs/Speed



Questions?