

DSC 232R: Big Data Analytics using Spark

Introductory Discussion Session: Week 1 | Winter 2026



Meet Your Instructional Team



Leadership & Support

Instructor: Edwin Solares (esolares@ucsd.edu)

Course Designers: Prof. Yoav Freund and Laura Griffin

TA: Nishanth Ramesha

Office hours and live sessions are held weekly via Zoom.
Links are available on the Canvas dashboard.

| Course Objectives & Goals

Engineering Big Data

Learn to program Spark using PySpark and identify computational bottlenecks in large-scale data analysis frameworks like Spark and XGBoost.

Analysis & Statistics

Apply methods from statistics and machine learning to analyze massive datasets, perform PCA on weather data, and visualize statistical summaries.

| The Learning Ecosystem



Canvas

Your main hub for syllabus, modules, quizzes, and grade tracking.



Discord

Primary communication channel for real-time discussion (No Code).



Piazza

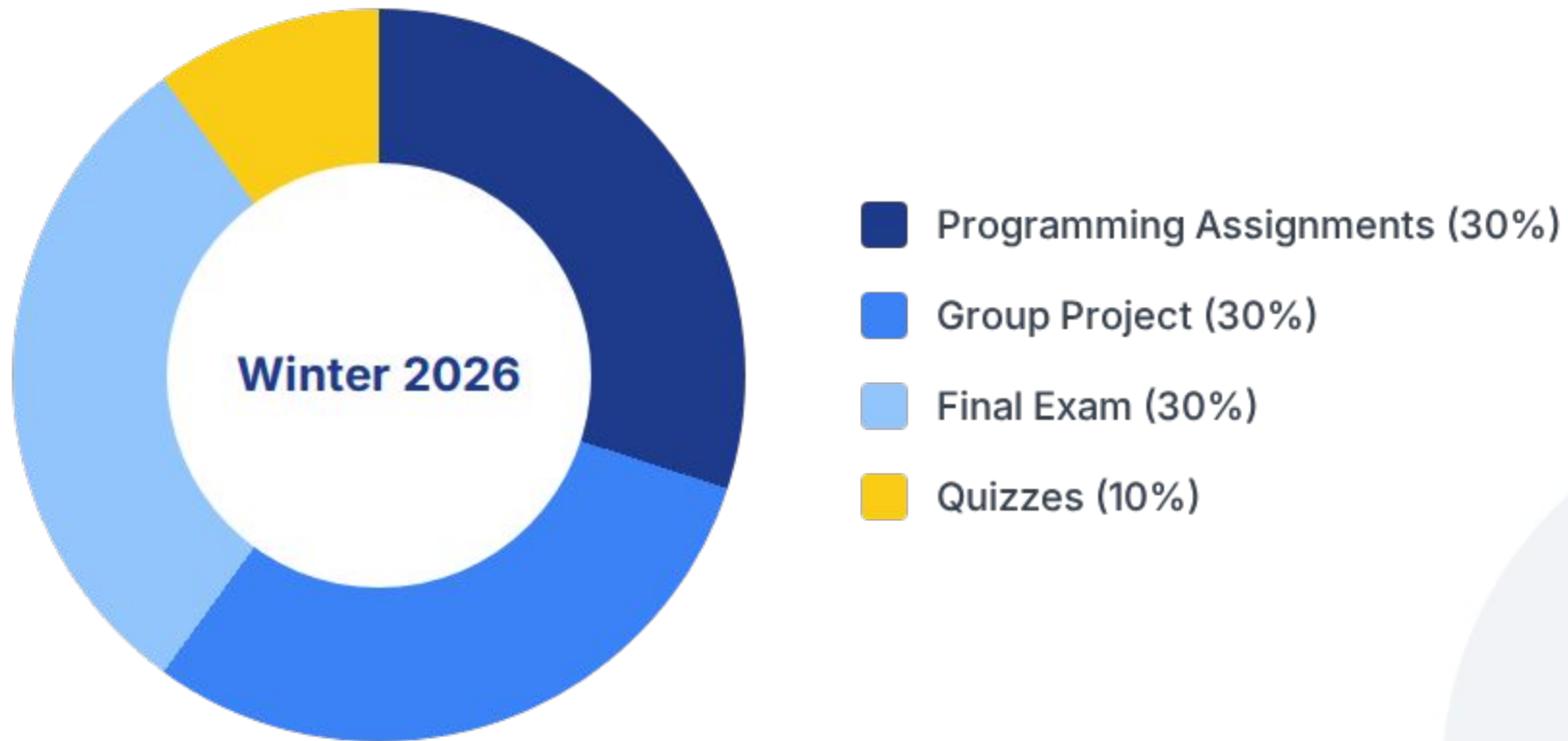
Post code snippets, debug bugs, and interact with the instructional team.



Vocareum

Jupyter-based LTI integration for all programming assignments.

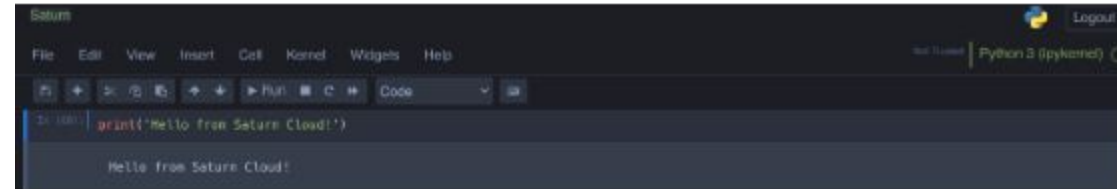
| Grading & Evaluation



Vocareum: Jupyter Environment

The course uses a Jupyter-based environment on Vocareum. Key technical highlights include:

- **Environment Setup:** Using `os` and `sys` to set PySpark variables.
- **Spark Config:** Setting application names and initialized nodes.
- **Local Workers:** Default setup uses 4 local cores to handle parallel jobs.
- **Auto-Graders:** Built-in cells that provide immediate feedback (Max submissions number preset).



Rules of Engagement





Discord vs. Piazza

Discord: Best for quick questions, networking, and group finding. Strictly NO code snippets here to prevent academic integrity issues.

Piazza: Use for technical bugs. Wrap code in triple backticks (``) for legibility. TA will monitor this for troubleshooting.`



| Weekly & Major Assessments

-  **Quizzes:** Released Mondays 12:00 AM, due Sundays 11:59 PM. Three attempts allowed; highest score counts.
-  **4 Programming Assignments:** Notebooks delivered via Vocareum. Rolling release schedule throughout the 10 weeks.
-  **Group Project:** Groups of 4 max, 4 critical Milestones. Check Canvas for due dates and more information.
-  **Final Exam:** An exam during the finals week, hosted via Vocareum.

| The Group Project Competition

4

Max Students per Team

Kaggle Competition

Work with public datasets to derive significant statistical insights.

The Prize: The top 3 projects as voted by your peers will receive extra credit toward the final grade.

Resources: Use the SDSC (San Diego Supercomputer Center) cluster for high-performance computing needs.

| Integrity & Netiquette

"Academic Integrity is expected of everyone at UC San Diego. This means you must be honest, fair, responsible, respectful, and trustworthy in your actions."

UCSD Course Syllabus Policy

! No insulting language

! No SHOUTING (All Caps)

! No code sharing on Discord

Questions?

Welcome to DSC 232R. We're excited to see your progress!

Next Steps: Join Discord | Fill When2Meet | Finish Quiz 1 & Week 1 Modules