

Homework 2: MapReduce Concepts & Spark Fundamentals

Points: 20 | **Due:** Sunday, February 8, 2026 @ 11pm Pacific

Author: Richard Young, Ph.D. | UNLV Lee Business School

Compute: CPU (free tier is fine)

Learning Objectives

1. **Set up** Apache Spark on Google Colab
 2. **Understand** how Spark partitions data for parallel processing
 3. **Measure and compare** processing performance with different configurations
 4. **Apply** K-Means clustering and interpret business results
 5. **Explain** distributed computing through your own diagram
-

Grading

Component	Points	What We're Looking For
Spark Setup	3	Create a Spark session and explain configuration
Data Partitioning	5	Demonstrate understanding of data distribution
Performance Experiment	5	Analyze why speedup isn't linear
K-Means Clustering	5	Apply clustering and interpret results
Diagram	2	Clear diagram explaining distributed processing
Total	20	

Instructions

1. Open MIS769_HW2_MapReduce_Spark.ipynb in Google Colab
 2. Complete Part 1: Install Spark and create a session
 3. Complete Part 2: Explore data partitioning
 4. Complete Part 3: Run performance experiment with 1, 2, and 4 cores
 5. Complete Part 4: Apply K-Means to Netflix data and name your clusters
 6. Complete Part 5: Draw your own diagram of how Spark processes data
-

Questions to Answer

- **Q1:** What does `local[2]` mean? What would `local[4]` do differently?
 - **Q2:** Why doesn't 4 cores give exactly 4x speedup?
 - **Q3:** What characterizes each cluster? Give them descriptive names.
-

Submission

Upload to Canvas: - Your completed .ipynb notebook with all cells executed