UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

Data Stream Processing

MSDS 682 (2 units)
Fall 2023
Lecture

(v1.30)

Instructor Information

Jeremy W. Gu
wgu9@usfca.edu
Office Hours
8:45-9:30am on Wed
Office Location: Virtual Zoom

Course Description

This class are to equips students with the skills necessary to process continuous data streams at scale and in real-time.

Students will gain hands-on experience with Apache Kafka and other modern data engineering tools. We focus on blending foundational knowledge with hands-on skills, often using real-world examples to anchor theoretical concepts. The primary programming language for the course is Python.

Prerequisite Details

To ensure your success in this course, the following knowledge prerequisites are essential:

- **Statistics:** Grasp of concepts like mean, variance, data visualization, tabular data, and pandas. Relevant courses include MSDS 504 - Review Probability and Stats and MSDS 593 - EDA and Visualization.
- **Python:** Familiarity with Python Classes and Objects. While proficiency with the Mac command line is beneficial, there will be no focus on Java or Scala in the JVM ecosystem.
- **SQL and Data Pipelines:** Knowledge derived from courses such as MSDS 681 - Data Lakehouse is ideal.

- **Machine Learning:** Familiarity with scikit-learn, supervised and unsupervised learning, evaluation of model performance, and feature selection. Courses like MSDS 621 - Intro to Machine Learning, MSDS 630 - Advanced Machine Learning, and MSDS 680 - Machine Learning Operations are recommended, though not mandatory.
- **Business Setting:** This course requires project presentations and reports. Effective and professional communication in both written and spoken English is crucial. Business writing should be concise, straightforward, and to the point.

Course Learning Outcomes

Learning Objectives

This applied, hands-on course provides students with the essential skills for processing and analyzing real-time data streams using modern data engineering tools and technologies. The goal is to equip students with foundational knowledge paired with practical skills in core data streaming tools and technologies like Kafka to prepare them for real-world data engineering and data science roles.

The course focuses on core concepts like the Kafka ecosystem while providing practical skills development through hands-on projects. Students will gain experience with real-time data ingestion using Kafka, conducting streaming analytics, and extracting insights. Optional materials will expose students to additional technologies like Faust and Spark Structured Streaming for building and evaluating streaming applications.

Fundamentals

• **Kafka Ecosystem**: Grasp the components of data streaming systems and delve into the Kafka ecosystem, recognizing the challenges each solution addresses. Kafka remains the primary focus.

• **Confluent Kafka** and **Confluent Cloud**: Use the Confluent Kafka Python library for tasks like topic management, production, and consumption. Demonstrations will be held using Confluent Cloud.

• **Real-time Data Streaming with Apache Kafka**: Students will undertake a course project for hands-on experience, ingesting real-time data using Apache Kafka, conducting live analytics, and extracting insights from streaming console reports.

Additional Skills

• **Faust Stream Processing**: Familiarize with the Faust Stream Processing Python library to craft real-time stream-based applications.

• **Additional Tools**: Time permitting, other tools will be introduced. For instance, students will explore the integration of Apache Spark Structured Streaming with Apache Kafka and understand the reports

generated by the Structured Streaming console.

Assignments

There will be three Assignments and one Final Project. It's crucial to adhere to deadlines. No submissions will be accepted past the due date.

Policy on Collaboration and Academic Integrity

- **Group Discussions & Collaborations:** We encourage effective group discussions and collaborations. Discussing concepts, techniques, and general approaches to problems with classmates can be beneficial. However, the work you submit must be your own.
- **Originality of Work:** All submitted assignments, code, and project reports must be your own original work. Copying and pasting code or text from classmates or any other source is strictly prohibited.
- **Using AI Services (e.g., ChatGPT):** Utilizing ChatGPT or other LLM models to aid in understanding course materials is permissible. However, always write in your own words and code independently. If you incorporate insights or specific code from any AI Services, clearly indicate which part(s) were influenced or sourced from them.
- **Online Sources & References:** If you utilize or are inspired by online sources, always provide proper citation and reference. Respect the intellectual property of others.
- **Cheating and plagiarism:** We have a zero-tolerance stance on cheating and plagiarism. Engaging in any form of academic dishonesty will result in severe penalties, which may include failure in the course.

Grading Breakdown and Grading Policies

Grading Breakdown

- Attendance and Professionalism: 5%
    - Regular attendance and active participation are expected.
- Individual Assignment: 30%
    - Assignments (10% each) will assess individual understanding and application of course material.
- Final Project: 45%
    - Project Proposal 5%
    - Written Report 20%

    - Final Presentation 20%
- Midterm Exam: 20%
- No Final Exam.

This class is a standard, graded course with letter grades **A - F**. Each grade reflects the quality and understanding demonstrated by the student, as follows:

A (90-100): Exceptional understanding and application. Demonstrates depth of knowledge and skill and indicates readiness to apply concepts in a professional setting.

- A+: 96-100

- A: 93-95

- A-: 90-92

B (80-89): Competent understanding and application of course material, representing the expected level of competence in a business setting.

- B+: 87-89

- B: 83-86

- B-: 80-82

C (70-79): Basic understanding, with room for improvement in application and depth, indicating achievements lower than the expected competence in the subject.

- C+: 77-79

- C: 73-76

- C-: 70-72

F (Below 70): Limited understanding and application of course material, representing an unacceptably low level of knowledge and understanding of the subject matter.

The expected class average is 85+ (Letter grade of B or above), with a normal distribution around this mean.

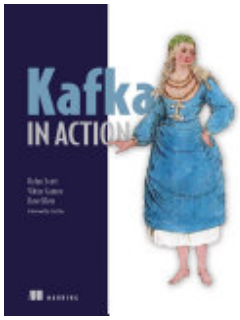
Beyond Grades

While the grading system in this course is a measure of academic performance, it can also shed light on the challenges one might face in the professional realm and the manner in which these are addressed. For example, when faced with tight deadlines, how does one ensure the quality of work? How does collaboration occur within a team, especially when compromises are necessary to meet

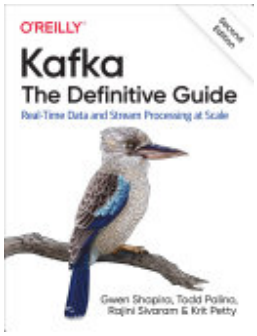
delivery dates amidst differing opinions? How well do you explain data science concepts to people with no data background, and how effective is communication with peers and superiors? How are colleagues persuaded when introducing innovative ideas?

In reality, certain missteps or attitude problems can have severe consequences, while standout performances can bring about many opportunities in your career. These are aspects indirectly reflected in your grades. If you earn an 'A+', there's a high probability that you will be a superstar in the workplace. Earning a 'B' is also commendable, as in a professional setting, it typically signifies "Meeting Expectations", meaning that you are performing your job. Conversely, if your grades are low, you might face significant risks of criticism from both managers and peers during performance evaluations. I hope every student can treat this course as a practical experience of the working world.

Texts and Supplies



Kafka in Action
ISBN: 9781617295232
Authors: Dylan Scott, Viktor Gamov, Dave Klein
Publisher: Simon and Schuster
Publication Date: 2022-02-15
Required or recommended?: Recommended reading as a reference book.



Kafka: The Definitive Guide
ISBN: 9781492043034
Authors: Gwen Shapira, Todd Palino, Rajini Sivaram, Krit Petty
Publisher: "O'Reilly Media, Inc."
Publication Date: 2021-11-05
Edition: 2nd Edition
Required or recommended?: Recommended. Not Required.

Course Schedule

Tentative Schedule (Each lecture: 5:30 - 7:20pm PST)

- Lecture #1: Oct 20, 2023 (F) - San Francisco-101 Howard 156
- Lecture #2: Oct 24, 2023 (T) - San Francisco-101 Howard 157 (Tentative Room)
- Lecture #3: Oct 27, 2023 (F) - San Francisco-101 Howard 156

- **Assignment #1: Due by 11:59pm on 10/28/2023**
- Lecture #4: Oct 31, 2023 (T) - San Francisco-101 Howard 156
- Lecture #5: Nov 03, 2023 (F) - San Francisco-101 Howard 156
- **Assignment #2: Due by 11:59pm on 11/4/2023**
- Lecture #6: Nov 07, 2023 (T) - San Francisco-101 Howard 156
- Lecture #7: Nov 10, 2023 (F) - San Francisco-101 Howard 529 | **60-min Midterm Exam**
- Lecture #8: Nov 14, 2023 (T) - San Francisco-101 Howard 156
- **Final Project Proposal: Due by 11:59pm on 11/13/2023**
- Lecture #9: Nov 17, 2023 (F) - San Francisco-101 Howard 156
- Lecture #10: Nov 21, 2023 (T) - San Francisco-101 Howard 156
- **Assignment #3: Due by 11:59pm on 11/22/2023**
- Lecture #11: Nov 24, 2023 (F) - San Francisco-101 Howard 156 | **Thanksgiving: No Class**
- Lecture #12: Nov 28, 2023 (T) - San Francisco-101 Howard 156
- Lecture #13: Dec 01, 2023 (F) - San Francisco-101 Howard 156
- **Final Project Written Report and Code: Due by 11:59pm on 12/3/2023**
- **Final Project Presentation Deck: Must be submitted before the final lecture 5:30pm on 12/5/2023**
- Lecture #14: Dec 05, 2023 (T) - San Francisco-101 Howard 157 (Tentative Room)
    - Note: This lecture is allocated for the Final Project Presentation. Each student has a 10-minute presentation slot.

Program Learning Outcomes

Please refer to "Course Learning Outcomes" Section.

Attendance Policy

Mandatory attendance for every lecture.

Use of Laptops: Please keep your laptops closed unless instructed otherwise, specifically during demo or exercise sessions. This is to ensure focus and participation during lectures.

Absence Due to Illness: If you are unable to attend a lecture due to sickness or any other unavoidable circumstance, please notify me in advance.

No Distractions: Mobile phones and other electronic devices should be kept silent and should not be used during class time.

University Policies

Credit-hour Policy

One unit of credit in lecture, seminar, and discussion work approximates one hour of direct faculty instruction (or 50 minutes plus a break) and a minimum of two hours of out-of-class student work per week through one 15-week semester. For further details, see USF's Credit Hour Policy.

The below resources and additional information can be found in the Student Life Resource Toolkit.

Students with Disabilities

If you are a student with a disability or disabling condition, or if you think you may have a disability, contact USF Student Disability Services (SDS) within the first week of class, or immediately upon onset of disability, to speak with a disability specialist.
If you are determined eligible for reasonable accommodations, please meet with your disability specialist so they can arrange to have your accommodation letter sent to me, and we will discuss your needs for this course. For more information, please visit Student Disability Services, email: sds@usfca.edu, or call (415) 422-2613.

Behavioral Expectations

The University of San Francisco is committed to providing an environment consistent with the academic nature and core values of the institution in which students can participate in learning as a humanizing, social activity rather than a competitive exercise to support the full, integral development of each person and all persons, with the belief that no individual or group may rightfully prosper at the expense of others.
It is important you know what is expected of you academically and behaviorally through the published course syllabus, the Student Conduct Code, and other instructions provided by the instructor. Therefore, you are expected to uphold the following:

- Behave in accordance with the Student Conduct Code and other USF policies
- Refrain from disrupting the ability of fellow students to learn or the instructor's ability to teach. Examples of disruption include:
    - Cell phone or computer use that significantly or repeatedly distracts others
    - Coming to class late, leaving early, or excessively physically relocating oneself in the classroom

    - Speaking frequently without being called on
    - Yelling, cursing, or engaging in aggressive behavior
- When interacting online, communicate in a respectful fashion. This includes, but is not limited to:
    - Refraining from name calling, profanity, or typing in all capital letters
    - Sending multiple emails with one sentence
    - Avoiding rants or discussing non-relevant topics

Open discussion and disagreement are encouraged when done respectfully and in the spirit of academic discourse. There are also a variety of behaviors that, while not against a specific University policy, may create disruption in this course. Students whose behavior is disruptive or who fail to comply with the instructor may be dismissed from the class for the remainder of the class period and may need to meet with the instructor or Dean prior to returning to the next class period. If necessary, referrals may also be made to the Student Conduct process for violations of the Student Conduct Code.

Academic Integrity

As a Jesuit institution committed to Cura Personalis—the care and education of the whole person — USF has an obligation to embody and foster the values of honesty and integrity. All members of the USF academic community are responsible for maintaining the standards of honesty and integrity. The honor code applies to all students (undergraduate and graduate) in the College of Arts and Sciences, the School of Education, the School of Management, and the School of Nursing and Health Professions. Faculty and students in the School of Law should review their own honor code for policies and procedures. Students enrolled in distance learning (online courses) are subject to these policies as well as supplemental policies set forth by their program. All students should review and familiarize themselves with the honor code, prohibited conduct, and procedures.

Counseling and Psychological Services

Many college students experience mental health struggles. Counseling and Psychological Services (CAPS) is a great source of support for issues such as anxiety, loneliness, struggles with relationships, stress, identity development, racial/cultural concerns, and mild depression. However, CAPS does not prescribe medication and does not have a psychiatrist on staff, so students with more severe mental health concerns will be referred off-campus for treatment.

Counseling and Psychological Services (CAPS) offers remote individual and group teletherapy to students residing within California. Students seeking services are scheduled for a 15-20-minute phone triage to assess immediate risk, identify treatment needs, and provide initial recommendations. These may include a crisis intake session, brief, intermittent individual teletherapy (every 2-3 weeks), single session teletherapy, weekly individual teletherapy via UWill,

weekly group therapy, or referrals to off-campus providers. There are no fees for services. To make an appointment, students must call 415.422.6352 or request an appointment via the CAPS. CAPS does not accept walk-in appointments.

If you are concerned about a student and would like someone to follow up, please contact the Dean of Students Office at 415.422.5330. If you know someone who is an immediate risk of harming themselves or others please contact Public Safety at 415.422.2911 in San Francisco, or out of state dial 911, or call the National Suicide & Crisis Lifeline by dialing 988. In addition, CAPS All Hours line can be reached 24/7 by calling 855.531.076. All students are encouraged to check out CAPS and access our extensive online resources, podcasts, mental health apps, videos, self-care strategies, and more.

Title IX

The Title IX Office seeks to stop, remedy, and prevent occurrences of sex and gender-based discrimination, sexual harassment, and sexual violence. The University has a Policy on Nondiscrimination based on Sex and Gender, Sexual Harassment and Sexual Misconduct. If you have experienced any of these behaviors, we encourage you to report the incident. If you report these behaviors to any staff or faculty member, they must notify the USF Title IX Coordinator.

Students who wish to report any sexual misconduct should use the online mandatory reporting form, or contact the Title IX Office directly. Other reporting options are available by visiting the Title IX website. The Title IX Office is located in Lone Mountain Room 145.

As an employee at USF, and your Professor, I am a mandatory reporter, meaning I have to share any instances of sexual harassment or sexual violence shared with me or that become known to me. I will have to share this information, including names and any details known, to the Title IX Office to connect you with resources. If you would like more information about the resources available, you can ask me at any time this semester. You do not need to tell me why you are asking to get help for a friend, another student, or yourself.

Confidential Resources for Reporting Sexual Misconduct

- Students may speak to someone confidentially which will not generate a report to the Title IX Office by contacting Counseling and Psychological Services at (415) 422-6352 during M-F 9-4pm, or speaking to a clergy member in University Ministry at (415) 422-4463.
- If you need to speak to a mental health clinician immediately, please **call the CAPS 24/7 All Hours Line at 855-531-0761** (available daily, including weekends and holidays, and accepts international calls), Public Safety (415-422-2911), 911, the Suicide Hotline (dial 988), or go to your nearest emergency room
- For off-campus resources, and local Bay Area organizations, view this webpage.

Learning, Writing, and Speaking Centers

The Learning, Writing, and Speaking Centers (LWSC) at USF provide individualized support to assist students in better understanding course material and to aid them on their path to success. Services are free and include tutoring, collaborative peer support services, academic skills coaching, writing, and speaking support. Services are available in-person and on Zoom. LWSC staff can be reached Monday through Thursday between 8:00am-8:00pm and Friday between 8:00am-5:00pm at LWSC@usfca.edu or through the chat box on our myUSF webpage or by phone at (415) 422-6713. To make an appointment for subject tutoring, academic skills coaching, the writing center, or the speaking center, students should visit the Student Appointment Dashboard.

Communication

All course communications, like all other USF communications, will be sent to your USF official email address. You are therefore strongly encouraged to monitor that email account.

Gleeson Library

Looking for help with a research paper or project? Set up a consultation with a Librarian or get 24/7 research help online.

Additional USF Resources

USF Food Pantry

The USF Food Pantry is an intermediate, short-term solution for any registered USF student to receive food and toiletry resources. Students are invited to stop by the pantry located on the first floor of Gleeson Library in the Atrium, and take the items that they need. Items are available on a first-come, first-serve basis until our supply is depleted. You will be asked to check-in via QR code before entering the pantry. For more information and the current schedule, visit the USF food pantry website. If you have further questions, please contact the Pantry Coordinator at usfpantry@usfca.edu or 415-422-4099 (during business hours Monday thru Friday from 9:00am - 5:00pm). You can find out about additional food security resources through the USF food insecurity resource page and the CalFresh resources site.