STREAM ANALYTICS

Bachelor in Data and Business Analytics BDBA SEP-2023
SA-DBA.3.M.A

Area Data Science

Number of sessions: 30

Academic year: 23-24

Degree course: THIRD

Number of credits: 6.0

Semester: 2º

Category: COMPULSORY

Language: English

Professor: **EDUARDO RODRÍGUEZ LORENZO**

E-mail: erodriguezl@faculty.ie.edu

EDUARDO RODRÍGUEZ LORENZO

Eduardo Rodriguez Lorenzo is Senior Manager at NETSCOUT and Adjunct Professor at IE School of Science and Technology. He is a technologist specializing in Telecommunication Networks, Cybersecurity, Software Architecture, Data Engineering and Analytics.

He studied at UPM (Universidad Politécnica de Madrid), King's College London and London University.

At NETSCOUT, he leads a global team of Data and Network Engineers with a strong focus on Network Service Assurance, Cybersecurity, Data Engineering and Analytics.

He has gained broad international experience delivering high-value Consulting Services (Customer Experience & Customer Journeys, Business Intelligence, Service Assurance, Data Monetization, Process Engineering...) and Data-driven Solutions (Cloud & Backend Architecture, Data Feeds, Database, Dashboard, Interaction & Visualisation Design) to global Enterprises and Communication Service Providers. He has played an active role in the launch, measurement and optimisation of Mobile Networks for various top international Telcos.

He is a member of the Spanish Charter of Telecommunications Engineers (COIT) where he is an active member of the Telecommunications Policy and Regulation Group and the Digital Transformation Group.

His main interests include Disruptive Technologies, Data Engineering Architectures, Networks, Distributed Systems and Graph technology.

He joined IE University in 2020.

IE | LinkedIn | Twitter

Office Hours

Office hours will be on request. Please contact at:

erodriguezl@faculty.ie.edu

SUBJECT DESCRIPTION

Organisations today connect to and use any number of data sources, some at rest stored in databases and exposed via APIs; others in motion and delivered as real time streams. Moreover, organisations also produce data at an unprecedented rate, data which needs to be published and delivered effectively to an unknown number of consumers.

To realise their business value, Data Streams must be collected, filtered, clean, aggregated and finally presented to southbound systems and consumers. Data processing can happen predictably in a schedule, as in Batch Processing, or constantly and in real time, as in Stream Processing.

Moreover, network and server infrastructure is required to be reliable, robust and scalable to prevent downtime and ensure supply and demand of data are correctly met.

In this course we will look into the building blocks that make data processing at scale possible, from network protocols and data serialization to messaging paradigms such as queues and brokers. Students will also learn advanced SQL techniques required to handle batches and moving data windows. We will also learn how to leverage the power of Cloud Computing to implement such solutions.

We will discuss and obtain hands-on experience with some of the key technologies in the data streaming stack, namely Kafka, Spark Streaming, Azure Event Hub, Apache Airflow, to name a few.

Students will have the opportunity to lay out and present their own Data Architectures to solve Data Streaming Use Cases, acting as CTOs in the making.

LEARNING OBJECTIVES

1. Understand the place of Streams in the Data Engineering Landscapce, its applications and key Use Cases
2. Understand the basics of Distributed Systems, Communication Networks and Protocols and the OSI Model
3. Design tradeoffs and key abstraction models for Streaming Data Processing systems
4. Understand Serialisation and most common formats used in industry: CSV, JSON, XML, AVRO
5. Learn Streaming Models and Techniques for Data Processing, including Public/Subscribe paradigms, Queues, Message Brokers
6. Design Analytical Pipelines based on data streams, using Windowing Techniques, Watermarks & Streaming SQL
7. Provide the basics of modern Design and Architecture of Distributed Systems

TEACHING METHODOLOGY

IE University teaching method is defined by its collaborative, active, and applied nature. Students actively participate in the whole process to build their knowledge and sharpen their skills. Professor's main role is to lead and guide students to achieve the learning objectives of the course. This is done by engaging in a diverse range of teaching techniques and different types of learning activities such as the following:

Learning Activity	Weighting	Estimated time a student should dedicate to prepare for and participate in
Lectures	26.67 %	40.0 hours
Discussions	10.0 %	15.0 hours
Exercises in class, Asynchronous sessions, Field Work	30.0 %	45.0 hours
Group work	16.67 %	25.0 hours
Individual studying	16.67 %	25.0 hours
TOTAL	100.0 %	150.0 hours

PROGRAM

WELCOME TO STREAM ANALYTICS!

SESSION 1 (LIVE IN-PERSON)

Introductions.

Expectations and what to expect.

Understand the place of Streams in the Data Engineering Landscapce, its applications and key Use Cases. Batch Processing vs Stream Processing at a glance.

Student Environment Setup.

DISTRIBUTED SYSTEMS AND COMPUTER NETWORKS

SESSION 2 (LIVE IN-PERSON)

Article: A Protocol for Packet Network Intercommunication (IEEE Trans on Comms, Vol Com-22, No 5 May 1974) (ced)

- OSI Reference Model. Protocols. TCP/IP Reference Model.
- Layer 2.
- Internet Protocol (IP). Addressing Concepts.
- IP version 4 Header.
- Transport Layer. TCP. 3-Way TCP Handshake. TCP Header. Common Layer 4 Ports.
- ICMP, Ping and Traceroute

- HTTP and DNS
- IPv6

SESSION 3 (LIVE IN-PERSON)

Linux Networking Practice

Practical use of Unix Pipes, Netcat, Nmap, Netstat, Curl, Wget... and other networking tools useful in Data Science projects.

RELIABILITY, SCALABILITY AND MAINTAINABILITY OF DATA SYSTEMS

SESSION 4 (LIVE IN-PERSON)

Basic Queueing Theory and System Dimensioning

SESSION 5 (LIVE IN-PERSON)

Reliability: Hardware Faults. Software Errors. Human Errors
Scalability: Describing Load, Describing Performance
Maintainability: Operability, Simplicity, Evolvability

DATA SERIALIZATION AND ENCODING

SESSION 6 (LIVE IN-PERSON)

Encoding and transmission of messages. Schema Evolution. Data Encoding formats. Dataflow modes.

SESSION 7 (LIVE IN-PERSON)

Data Encoding Lab: JSON, AVRO, HL7, ISO8583

STREAMING PROCESSING

SESSION 8 (LIVE IN-PERSON)

STREAMING FUNDAMENTALS

Article: The Dataflow Model (Proceedings of the VLDB Endowment, Vol. 8, No. 12)

Event Time Domain vs Processing Time Domain. Triggers. Watermarks. Accumulation.

Windowing techniques: Tumbling, Hopping, Sliding, Session. Advanced Windowing.

SESSION 9 (LIVE IN-PERSON)

STREAMING TECHNOLOGY

Streaming Tools and Frameworks. Spark Streaming Introduction and Setup.

SESSION 10 (LIVE IN-PERSON)

Spark Streaming Lab

Article: Discretized Streams: Fault-Tolerant Streaming Computation at Scale (SOSP'13, Nov. 3–6, 2013, Farmington, Pennsylvania, USA) (ced)

A Jupyter-led assignment to demonstrate *pyspark.streaming* module functionality

SESSION 11 (LIVE IN-PERSON)

Streams and Tables

Article: Spark SQL: Relational Data Processing in Spark (SIGMOD'15, May 31–June 4, 2015, Melbourne, Victoria, Australia) (ced)

Multimedia Documentation: Spark Structured Streaming Programming Guide (spark.apache)

Relational Algebra, Unwindowed Joins, Windowed Joins, Stream-Stream joins, Stream-Table joins, and Table-Table joins.

Spark Structured Streaming.

SESSION 12 (LIVE IN-PERSON)

Spark Structured Streaming Lab

Sample Spark Notebooks to demonstrate Spark Structured Streaming Programming model.

SESSION 13 (LIVE IN-PERSON)

RECAP, Q&A SESSION

We will review topics covered so far and do practice exercises towards the Midterm Exam.

SESSION 14 (LIVE IN-PERSON)

MID-TERM EXAM

MESSAGING

SESSION 15 (LIVE IN-PERSON)

Messaging Patterns:

- Pairwise,
- Client/Server,
- Push-Pull,
- Queues.
- Publish/Subscribe paradigm

SESSION 16 (LIVE IN-PERSON)

Messaging Paradigms Lab

Using ZeroMQ to demonstrate all messaging paradigms we learnt in the previous session.

SESSION 17 (LIVE IN-PERSON)

Azure Queues Lab

Learn about Azure Storage Queues service, its API, SDK and Use Cases.

Implement a basic streaming solution using Azure Storage Queues.

SESSION 18 (LIVE IN-PERSON)

Kafka Concepts

Terminology, Use Cases, Core APIs, Zookeeper, Topics, Brokers, Partitions, Producers, Consumers. Local Setup

SESSION 19 (LIVE IN-PERSON)

Event Streaming Use Case: Financial Transactions and ISO8583

SESSION 20 (LIVE IN-PERSON)

Kafka Lab

Hands-on Kafka Lab on a relevant topic of choice involving streaming data sources.

SESSION 21 (LIVE IN-PERSON)

Azure Event Hubs Concepts

Conceptual mapping with Kafka and key differences, API usage, integration with other Azure components. Use Case: sending IoT telemetry to Event Hub.

SESSION 22 (LIVE IN-PERSON)

Azure Event Hubs Lab

Setting up an Azure Event Hub and practice publishing to and consuming messages from it.

SESSION 23 (LIVE IN-PERSON)

Azure Streaming Analytics Concepts

Features. Input Types. Using Reference Data. The U-SQL dialect to write Streaming Queries. Available Outputs.

Lab: Simulating Sensor Data for Streaming Processing.

SESSION 24 (LIVE IN-PERSON)

Azure Streaming Analytics Lab

A Lab integrating some of the Azure Technologies covered so far: EventHub, Stream Analytics, Blob Storage and PowerBI.

SESSION 25 (LIVE IN-PERSON)

Advanced SQL Applications in Streaming Business Applications

SQL Window Functions and Streaming extensions to SQL standard. SQL Common Table Expressions (CTE).

SESSION 26 (LIVE IN-PERSON)

Advanced SQL Lab

Using SQLITE as the underlying Database, demonstrate Business Uses of Advanced SQL constructs.

ARCHITECTING MODERN STREAMING DATA SYSTEMS

SESSION 27 (LIVE IN-PERSON)

Team presentations of Group Project

List of sample topics, though students can propose their own:

- Contact Tracing
- Fraud Detection
- Predictive Maintenance
- Smart City topics: car telemetry, environmental sensors, etc

Student group will conduct a live presentation and upload accompanying materials to Campus (code, Notebooks, Slides...).

SESSION 28 (LIVE IN-PERSON)

Article: Mariam Kiran, Inder Monga: Lambda architecture for cost-effective batch and speed big data processing (Conference Paper · October 2015) (ced)

Working Paper: Martin Feick, Niko Kleer, and Marek Kohn: Fundamentals of Real-Time Data Processing Architectures Lambda and Kappa (SKILL 2018, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 1)

Lambda and Kappa architectures.

Distributed Architecture Patterns.

Job Scheduling. Synchronization. Cron Jobs. Apache Airflow & Azure Batch

FINALS

SESSION 29 (LIVE IN-PERSON)

RECAP, Q&A SESSION

We will review topics covered and practice topics for the Final Exam.

SESSION 30 (LIVE IN-PERSON)

FINAL EXAM

EVALUATION CRITERIA

Throughout this course, you will be asked to read material related to the sessions, participate in live and online discussions, complete individual assignments and labs, and participate in a group project.

Specifically, grading will be based on the following criteria.

criteria	percentage	Learning Objectives	Comments
Final Exam	20 %		
Midterm Exam	20 %		
Labs and Individual Practice	30 %		
Group Assignment	20 %		
Class Participation	10 %		

RE-SIT / RE-TAKE POLICY

Midterm Exam

There will be a midterm exam covering the topics:

- Distributed Systems and Computer Networks
- Reliability, Scalability and Maintainability of Data Systems
- Data Serialization and Encoding
- Streaming Proccessing and Pipelines

Final Exam

At the end of the course you will have to pass an individual exam which may contain questions about any of the topics covered during this course.

Labs and Individual Practice

During the course I will propose multiple labs, practices and exercises on these and other topics:

1. Networking Lab
2. Data Encoding Lab
3. Spark Streaming Lab
4. ZeroMQ Scenarios
5. Kafka Lab
6. Event Hub Lab
7. SQL Window Functions Lab
8. Azure Stream Analytics Lab
9. Streaming Architectures

I will guide the development of all Labs and Assignments, and make sure the level of difficulty is adequate to each and every student. You should be able to complete each exercise during the time allocated to each session (synchronous or asynchronous), however there may be optional parts that might require additional time from you. For asynchronous assignments, there will be a channel (typically via the forum) to ask questions to other students and the professor.

Group Assignment

Group assignments will be developed by teams of 4 or 5 students, who will have to work on a real life Use Case involving Stream Analytics. The team will play the role of a Chief Technology Officer who puts together the best solution taking into account technical, business and environmental constraints.

A list of topics will be proposed by the professor, but students are free to propose their own subjects.

Class Participation

You are expected to attend every class and participate in the discussions and class activities. This includes optional exercises, voluntary participation on the whiteboard, discussion board (forum) activity, class attendance, and active participation in in-class discussions, with the goal of ensuring a continued learning process, good teamworking, and ability to apply class concepts in real-world problems. Participation is based on the quality, rather on the quantity, of your contributions.

Late Assignments/Presentation

If you should be late in submitting an assignment, be sure to inform me in advance of the reason to try and accommodate you. Any unjustified late submission will not be graded.

Minimum passing grade

A minimum passing grade applies in final exams (3.5). If your score is lower than this minimum you will have to retake, irrespective of your overall course grade. Also keep in mind that the overall passing course grade is 5.0.

BIBLIOGRAPHY

Recommended

- Tyler Akidau, Slava Chernyak, Reuven Lax. *Streaming Systems.* 2018. O'Reilly Media, Inc.. ISBN 9781491983874 (Digital)

Streaming data is a big deal in big data these days. As more and more businesses seek to tame the massive unbounded data sets that pervade our world, streaming systems have finally reached a level of maturity sufficient for mainstream adoption. With this practical guide, data engineers, data scientists, and developers will learn how to work with streaming data in a conceptual and platform-agnostic way.

- Jay Kreps. (2014). *I Heart Logs.* O'Reilly Media, Inc.. ISBN 9781491909386 (Digital)

Why a book about logs? That’s easy: the humble log is an abstraction that lies at the heart of many systems, from NoSQL databases to cryptocurrencies. Even though most engineers don’t think much about them, this short book shows you why logs are worthy of your attention.

- Piethein Strengholt. (2020). *Data Management at Scale.* O'Reilly Media, Inc.. ISBN 9781492054788 (Digital)

As data management and integration continue to evolve rapidly, storing all your data in one place, such as a data warehouse, is no longer scalable. In the very near future, data will need to be distributed and available for several technological solutions. With this practical book, you’ll learn how to migrate your enterprise from a complex and tightly coupled data landscape to a more flexible architecture ready for the modern world of data consumption.

- Martin Kleppmann. (2017). *Designing Data-Intensive Applications.* O'Reilly Media, Inc.. ISBN 9781449373320 (Digital)

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords?

BEHAVIOR RULES

Please, check the University's Code of Conduct here. The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy here. The Program Director may provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code here. The Program Director may provide further indications.