(ECBS5211) Data Engineering 3: Batch Jobs and APIs

Academic year: 2024

Background and overall aim

You will learn what components usually make up an automated, maintainable and scalable R-based data architecture, how to use that in production, and what other technologies you should become familiar with for different use-cases. The class starts with deploying real-world examples of batch business rules implemented in R onto cloud-computing resources running in AWS (Amazon Web Services), then will look into stream-processing with R, creating dashboards, and implementing API endpoints. We will also cover some related data processing and database technologies outside of R: how they can be used for solving a variety of problems and how to connect to these from R.

Course prerequisites

Introduction to R

Laptop with Internet access and an SSH client is required in the class and for the take-home assignments as well. SSH installation steps will be shared on GitHub. AWS login information is also required.

Learning activities and teaching methods

By the end of the course you will have a better understanding on the building blocks of an R-based production data infrastructure in the cloud, including automated reports, dashboards, stream-processing and service integrations.

Other outcomes. The course will also help develop skills in the following areas:
Critical Thinking - The importance of security and logging in application design.
Technology Skills - Cloud infrastructure (Amazon Web Services), job scheduler (Jenkins), APIs, stream processing.

Assessment

20% quizzes (at the beginning of the classes) and 80% final project (creating an R-based data pipeline on AWS).

Grading Policy
Students shall not miss more than 1 day of classes, failing to do so will yield an administrative fail grade. To pass, students will need to get at least 50% of the quizzes AND at least 50% of the final project.
Grading will be based on the total score out of 100, in line with CEU’s standard grading guidelines

Course contents

1. Installing and using R and RStudio on cloud servers.
2. Scheduling R scripts, database connections.
3. Stream processing with R.
4. Implementing API endpoints in R.

Reading List:
Class materials will be available on GitHub.
Databases: The CEU Library boasts a range of databases covering financial and company data, market and industry reports, global news and more. For a full list of databases visit the CEU Library.
• Refinitiv (Thomson Reuters) Eikon for Students + Datastream/Thomson ONE

- Eikon: Platform used by finance practitioners including market traders to monitor and analyze financial information. Information, analytics and news on all major financial markets including real-time pricing data, financial research, global financial news and commentary, financial estimates, fundamentals analysis, visual analysis through charting. Import/export from Excel.
- Datastream: Range of economic, securities and company financial data. Excel add-in.
- Thomson ONE: Global overviews on 55,000 public companies, one million private companies. Reuters News, ownership, deals, private equity, key ratios, company filings, officers and directors. Investext analyst reports, active and historical research from 1,600 independent research firms, brokerages, investment banks.

• Standard & Poor's Capital IQ
- Web and Excel-based platform combining deep global company information, credit ratings and research, and market research with powerful tools for risk assessments. Real-time and historical information on markets, industries, companies, transactions and people. Tearsheet data.

• Lexis Nexis Academic
- Global database of news, business, legal and other sources. Full text of 350 newspapers, 300 magazines and journals, 600 newsletters. Wire services including Associated Press, Business Wire and PR Newswire. Company financial information, market research, industry reports.

Contact details

Gergely Daroczi has a PhD in Sociology, 15 years of experience with R, founder of the Hungarian R meetup and main organizer of R conferences, authored a book on R and maintains a dozen of R packages, lived and worked in Hungary and USA at market research, fintech, adtech and healthtech companies as a data scientist and engineer both in individual contributor and management roles