TECHNOLOGY TRANSFER PRESENTS

RUSSELL JURNEY

AGILE DATA SCIENCE 2.0

FULL-STACK ANALYTICS APPLICATION DEVELOPMENT WITH KAFKA AND SPARK

ONLINE LIVE STREAMING

OCTOBER 24-28, 2022

DUE TO TIME ZONES, THIS CLASS WILL TAKE PLACE IN THE AFTERNOON
FROM 2 PM TO 6 PM ITALIAN TIME

info@technologytransfer.it
www.technologytransfer.it

ABOUT THIS SEMINAR

Agile Data Science 2.0 covers the theory and practice of an agile development methodology created to enable analytics application development. Students will learn the theory and application of Agile Data Science, a development methodology in which a Data Scientist uses agile methods and a lightweight stack to perform full-stack analytics application development. Students will learn how to define, implement and use a Big Data full stack, and how to roll their own Big Data applications from the ground up. This will enable them to effectively present their findings as applications, helping them make change within technological organizations. Students will emerge from this course with skills and a technological template from which to derive their own applications using their own datasets.

AUDIENCE

- Data Scientists interested in learning "Big Data" application development
- Programmers interested in building full-stack Big Data applications
- Data Scientists interested in applying Agile methods to Data Science
- Practicing Statisticians who want to learn to build entire applications
- Entry level Data Scientists who want to learn how to craft full-stack applications

Background requirements:

- Basic understanding of imperative, C-like languages
- Some experience with Python
- Some exposure to Javascript
- Some exposure to HTML/CSS
- Some experience working with data (SQL or Excel counts)
- Basic Linux/bash experience, or expertise in running FOSS on Windows or a Linux VM (Windows will not be supported, so you need to be capable of setup yourself)

Required materials and preparation:

- Students must read Agile Data Science 2.0, Chapters 1-4, before the course begins
- Students must follow the setup instructions in Agile Data Science 2.0, chapter 2, to set up their development environment
- Students must download the book's source code at http://github.com/rjurney/Agile_Data_Code_2 .
- Students are recommended to have an Amazon Web Services account to launch the development environment (optional but recommended to fully participate)
- Students can run a script from the book's GitHub repo to initialize their system with all the example datasets

Course Supplies: Acquiring The Book

You can acquire the book, Agile Data Science 2.0 from Amazon, Wordery (free shipping to Italy) or eBooks. You should do so in plenty of time to do the course reading, or as soon as you sign up for the course.

Outline

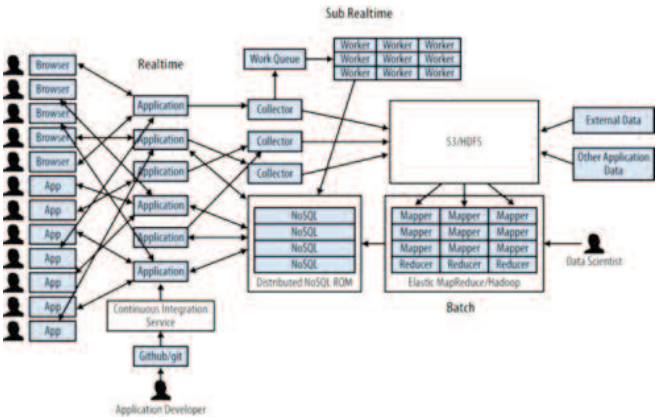
DAY 1

1. Lecture: Agile Data Science

A lecture will present the analytics development methodology outlined in Agile Data Science 2.0. This will focus on how to think with agility while doing Data Science.

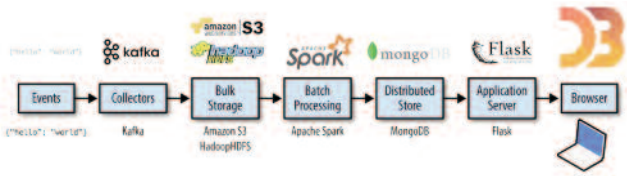
2. Introducing the Analytics Stack

The introductory unit will begin to establish a theoretical background for operating on data at scale. The terms Big Data and Data Science will be defined, analyzed and given a historical context. During the lecture we will introduce the software and system that form the backbone for performing Data Science in the field.
This stack is pictured below:



3. Demo: Walking Through our Full Stack

Students were instructed that course preparation required that they work through Agile Data Science 2.0 Chapter 2, Setup. We will review that setup, and then walk the class through the the operation of the stack, from the ground up.
We'll begin by looking at how we download data programmatically, and then inspect the files we download. We'll walk through using the Kafka console producer and consumer to send messages, and we'll store these messages in a file. Then we'll show how to upload the file to S3, and we'll view it in the AWS S3 interface. Next we'll process this file in PySpark, demonstrating a few APIs. Then we'll store the data from PySpark into MongoDB, followed by querying the data in MongoDB. Next we'll use Jupyter Notebooks to query MongoDB from Python, and we'll demonstrate a simple Flask application. Next we'll create a very simple Jinja2 HTML template which will load d3.js and fetch our data from Flask, rendering a simple chart.



At this point we will have interactively demonstrated the entire stack, and will have linked students to documentation for each step they can refer to as they work at each level.

4. Exercise: Data Processing in PySpark

Students will be presented with very simple API documentation with examples for 3-4 PySpark operations, and a simple dataset. They will be assigned a list of things to compute using these APIs. After ten minutes, I will ask a few students to present their answers and will cover correct answers to the exercises.

5. Exercise: Querying Data in MongoDB

Students will be presented with very simple API documentation with examples for 3-4 MongoDB operations, and a simple dataset. They will be assigned a list of things to compute using these APIs. After ten minutes, I will ask users to present their answers and will cover a correct answer to the exercises.

Next, I will present the students with an example of using PyMongo to query MongoDB from Python, and their assignment will be to write code to do the

same for a different dataset.

6. Exercise: Creating a Web Service

After a brief lecture around a demonstration of a Web Service from the example code, students will be given the assignment to create a slightly different Web Service around a different dataset. After fifteen minutes, I will demonstrate my own version of this Web Service, and then we'll discuss common problems and compare different implementations of the Web Service.

DAY 2

7. Demo: Hacking Charts in d3.js

We will start from a demo dataset and a corresponding Web Service providing that data in the javascript console of a browser. I will walk through the process of choosing which chart to use to visualize the dataset and then searching for examples of that chart type implemented in d3.js. I will show students the process of adapting existing d3.js examples to a different dataset. I will take them through, step by step, the process of altering only a few lines to make an example chart work for a different dataset.

8. Exercise: Hacking Charts in d3.js

Students will be supplied with example code of a Web Service that gives a blank page with d3.js and the example dataset already loaded. Students will be guided through deciding which chart to use and finding an example. Students will then be challenged to alter the example to fit the dataset in the exercise. At the end of the session I will review the correct answer, and we will address common pitfalls and problems that arose.

9. Lecture/Demo: Predictive Modeling in PySpark

A lecture/demonstration will cover the theory behind predictive modeling along with a corresponding implementation in PySpark/Spark MLlib.

10. Predictive Modeling in PySpark

Students will be assigned with building and testing a predictive model using a different dataset. Every 5 minutes or so I will review another portion of the answer so that stragglers can catch up.

11. Deploying Spark Predictive Models

A lecture will review how Kafka interacts with PySpark to deploy predictive models in real-time. We will cover the back-end operation of real-time predictive models using Kafka and PySpark.

DAY 3

12. Exercise: Deploying Spark Predictive Models

Students will be challenged with deploying an existing predictive model. They will use the Kafka console producer to generate requests for the model, and will retrieve predictions from MongoDB.

13. Demo: Predictions on the Web

I will show students how to create a Web Service to create prediction requests, and the corresponding front-end javascript code to submit a request to the Web Service and then poll for and display the result. I will use a notebook to walk through the back-end code and the console of a web browser to walk students through the front-end code.

14. Exercise: Predictions on the Web

A script in the GitHub repo will initialize the back-end of a predictive service. Students will then be challenged with implementing the front end of the service using the example from the previous step as a guide.

15. Build large knowledge graphs with GraphFrames

Students will transform the tabular data we have worked with into a graph via Spark's GraphFrames library and learn how to do network science and perfect business knowledge graphs for business applications.

16. Discussion: Lessons Learned

A discussion will cover what students have learned, areas that are still unclear, pitfalls they found and solutions they discovered, what interested them about the material and how they plan to use it in the future.

17. Lecture: Wrapping Up

A lecture will walk through what the students have learned and the examples they have worked and will drive home the theory and application of full-stack Web development.

EXPECTED OUTCOMES

Define what someone will gain by taking this course: what they will know and be able to do by the end of it?

1. Participants will understand
    - How to define “full-stacks” of Big Data tools
    - How to apply Agile methods to Data Science
    - Python/Flask Web development
    - Exploratory data analysis against Big Data

2. Participants will be able to
    - Use full-stacks of Big Data tools
    - Work with some of the most popular Big Data tools: Python, Spark, Kafka, Elasticsearch, MongoDB
    - Build full-stack analytics applications
    - Build visualizations in d3.js
    - Build and deploy complete predictive analytics applications and systems
    - Build Web applications using Python/Flask
    - Explore Big Data interactively

COMMON MISUNDERSTANDINGS

- Big data tools are only for Big Data
- Big Data is too hard for me!
- It takes an entire team to build an analytics application
- Agile software methods apply directly to Data Science
- Agile doesn’t work for Data Science
- Full-stack applications are beyond my reach!

LEARNING ACTIVITIES

Assignments prior to the live scheduled meetings online
Inspired by How to Flip a Class.

- Read Chapter 1, Theory, of Agile Data Science 2.0
- Work through Chapter 2, Setup, of Agile Data Science 2.0
- Read (the short) Chapter 3, Data
- Work through Chapter 4, Collecting and Displaying Records

This work prepares us to work our way through visualization, reports, interactive application development, and real-time predictive analytics.

INFORMATION

PARTICIPATION FEE

€ 1400

The fee includes all seminar documentation.

SEMINAR TIMETABLE

2.00 pm - 6.00 pm (Italian time)

HOW TO REGISTER

You must send the registration form with the receipt of the payment to:
info@technologytransfer.it

TECHNOLOGY TRANSFER S.r.l.
Piazza Cavour, 3 - 00193 Rome (Italy)

PAYMENT

Wire transfer to:
Technology Transfer S.r.l.
Banca: Cariparma
Agenzia 1 di Roma
IBAN Code:
IT 03 W 06230 03202 000057031348
BIC/SWIFT: CRPPIT2P546

GENERAL CONDITIONS

DISCOUNT

The participants who will register 30 days before the seminar are entitled to a 5% discount.

If a company registers 5 participants to the same seminar, it will pay only for 4.

Those who benefit of this discount are not entitled to other discounts for the same seminar.

CANCELLATION POLICY

A full refund is given for any cancellation received more than 15 days before the seminar starts. Cancellations less than 15 days prior the event are liable for 50% of the fee. Cancellations less than one week prior to the event date will be liable for the full fee.

CANCELLATION LIABILITY

In the case of cancellation of an event for any reason, Technology Transfer's liability is limited to the return of the registration fee only.

RUSSELL JURNEY
AGILE DATA SCIENCE 2.0

October 24-28, 2022

Registration fee:
€ 1400

If registered participants are unable to attend, or in case of cancellation of the seminar, the general conditions mentioned before are applicable.

first name .....

surname .....

job title .....

organisation .....

address .....

postcode .....

city .....

country .....

telephone .....

fax .....

e-mail .....

Stamp and signature

Send your registration form with the receipt of the payment to:
Technology Transfer S.r.l.
Piazza Cavour, 3 - 00193 Rome (Italy)
Tel. +39-06-6832227 - Fax +39-06-6871102
info@technologytransfer.it
www.technologytransfer.it

Speaker

Russell Jurney is principal consultant at Data Syndrome, a product analytics consultancy dedicated to advancing the adoption of the development methodology Agile Data Science, as outlined in the book Agile Data Science 2.0. He has worked as a Data Scientist building data products for over a decade, starting in interactive web visualization and then segwaying towards data products, Machine Learning and Artificial Intelligence at companies such as Ning, LinkedIn and Hortonworks. He is a self taught visualization software engineer, data engineer, data scientist, writer and most recently, he is becoming a teacher. In addition to applied work building analytics products, Data Syndrome offers live and video training courses.