# Open Sourcing Education for Data Engineering and Data Science

David E Drummond
Data Engineering

Insight Data Engineering
Palo Alto, CA

***Abstract*—The fields of Data Engineering and Data Science have emerged in recent years as an exciting intersection between leading academic research and industry leaders, and include diverse, cutting-edge topics like distributed systems, machine learning, and artificial intelligence. While these topics are interesting in their own right, perhaps the most compelling aspect of these fields is how the tools, which are in widespread use, have been developed by an Open Source community. Individuals across several universities and companies have worked together in a distributed fashion to build and improve the leading generation of data technologies. These Open Source principles have enabled the latest industry-adopted tools to constantly evolve at an incredible rate. One way for educators to keep pace with these developments and maintain advanced curriculum is to adopt the same collaborative principles used in open source. At the Insight Data Fellowship, we've used this open source model to provide immediate feedback and drive our curriculum forward, while fostering a culture of independence and curiosity. This session will show Engineering educators how to use open source principles and tools to develop their own curriculum.**

***Keywords—open source; data science***

## I. A Novel Approach

In addition to learning about the fields of Data Engineering and Data Science, attendees will experiment with a new approach of applying open source principles to educational curriculum. This style combines the constant feedback or Agile software development with the distributed, collaborative, and scalable nature of Open Source. These techniques, which have been widely adopted in the tech industry, have only recently begun influencing the educational field and represent a singular approach to improving curriculum in emerging fields.

## II. Learning by Building

Participants will learn in a hands-on way to use the same open source data engineering technologies adopted in industry. At the same time, we will work together as a group to improve on the accompanying curriculum using distributed version control technologies. This will provide first hand experience of how our Fellows have quickly and effectively learned advanced material, and also show attendees how to incorporate these open source principles in their own curriculum.

## III. Agenda

The session will be broken into two parts - the first half will be a brief review of the fields of Data Engineering and Data Science, best practices in the open source community, and a discussion of how Insight has leveraged these open source materials to develop our educational curriculum. Specifically, we will discuss the recent advance and adoption of one of the most popular open source platforms, Apache Spark. This will include an overview of Spark's history as it transitioned from academia to industry as an open source project. Additionally, we will very briefly review the underlying technology used in Spark, and how our Fellows quickly learn to use tools like this at the Insight Fellowship. This review will ensure that all participants are familiar with the latest developments in these fields, and will be able to apply them in the interactive session without any prior knowledge of specific data technologies.

The second half will be an interactive session where participants develop open source curriculum for setting up a distributed cloud computing cluster (fully provided by the Speaker). Once this is set up, each participant will work through an example data science algorithm using this platform. All the materials and resources required for this exercise will be available through the distributed version control platform, Github, and participants will actively improve this material throughout the process. This example will serve as a vignette of how a group of students can efficiently learn and improve advanced material in a scalable and collaborative way.

## IV. Expected Outcome

Participants will come out of the session with a greater familiarity of the emerging fields of Data Engineering and Data Science. Attendees will also get hands-on experience developing curriculum using Open Source tools and principles, which can be directly applied to their own content. Finally, all materials used during the session will be available for further use and development, both as an example of data technologies and as open source educational material.