BDiF

Big Data in Finance

"Big Data in Finance" 2016 course outline

Friday 21st January 2016, Revision B

1. Rationale behind the course

Practicing Quants build useful tools. Big Data skills are not just useful in themselves, they are now a requirement of many jobs. Which is to say that you need Big Data and data science skills not just to be a Big Data practitioner, but you now need those same skills as part of other jobs, such as being a financial Quant. As companies become more and more data driven, Big Data and data science skills are not only more in demand, it's also the case that having good data skills are essential for pretty much any quantitative job in finance.

Good financial quants these days require many skills, and they have to be very good at each of them: knowledgeable about finance and markets, a mathematician, an ace programmer, and a data scientist. The Big Data in finance course will help with the latter two. If you work diligently and complete the course, I can pretty much guarantee that your programming skills will have gone up several notches, and you will be able to take on any dataset no matter how big and ugly!

Moreover, these skills are in high demand in the market, so you can take them to the bank (pun intended).

2. New for 2016

Each year the "Big Data in Finance" course has continued to evolve based on student feedback, advances in technology, and the jobs/skills landscape.

New for this year:

- A TA (Teaching Assistant) to help students with assignments and learning to get the most out of the course.
- New state-of-the-art hardware (an SGI shared-memory machine with 12TB RAM [1], 384 processors and GPU acceleration). Shared-memory architectures are particularly good at solving graphs.
- Graphs (of the mathematical vertex and edge kind, not pretty pictures!) over Big Data.
- More real-world financial Big Data (bigger and more interesting data sets).

- Real-time streaming Big Data using Apache Spark, rather than batch oriented Hadoop (we're dropping Hadoop at this point because the market is shifting to Spark).
- In-memory databases, such as Kx.
- Advanced C++14 for ultra-high performance and running at massive scale.
- Using the "Blockchain" (the technology underling Bitcoin and other crypto-currencies) as a platform. I've covered the Blockchain in previous years, but I'm giving it more emphasis because its rapid rise in mainstream finance.
- Data visualization.

3. Course outline

The course follows an 4 step methodology for solving Big Data problems. Any Big Data problem. Here are the steps:

1. (D1) Data Exploration ("know thy data").
2. (D2) Data Programming ("put the right data 'plumbing' in place").
3. (D3) Data Analysis ("seek-out value in your data").
4. (D4) Data Insights ("profit from your data").

This methodology permeates all that we do in class:

- Class 1: Big Data foundations. (D1)
    - Data science. An overview from a Big Data perspective.
    - Information theory (and its relevance to data). Data representation.
    - Mathematical modeling of Big Data.
    - Data characteristics. Data quality.
    - Big Data algorithms and tools.
- Class 2: Big Data software development masterclass. (D2)
    - Professional development tools and the toolchain (your toolchain). On being a professional and productive programmer.
    - Writing programs that work—that give the correct results.
    - Writing programs that are robust—that don't fail ungraciously.
    - Writing programs that scale (massively)—work at all data sizes.
    - Writing programs that are easy to debug—that you can fix easily.
    - Step-by-step setup of remote access to your user account at the CUNY HPC (high performance computing) center. Getting your first programs working on "big iron" (supercomputers).
- Class 3: Running and solving Big Data problems at massive scale. (D2)
    - Thinking in parallel. Parallel programming.
    - Multithreading. CPU & GPU programming.
    - Distributed programming.
    - Parallel I/O.
- Class 4: In memory databases and Big Data. (D3)
    - In-memory databases.
    - NoSQL databases.
    - Graph databases.

  - The Kx in-memory database.
  - Graphs and Big Data.
- Class 5: Streaming Big Data. (D3)
  - Real-time Big Data.
  - Apache Spark.
- Class 6: Big Data visualization. (D4)
  - Introduction to R.
  - Using R for Big Data.
  - Data visualization with R.
  - Example: The Blockchain.
- Class 7: Final exam.
- Assignments: I'm still thinking through the details of the assignments—most of which are new for this year—so the following list may change, but it does give a sense of what you will be doing in terms of assignments.
  - Assignment 1: Distributed/parallel analysis of financial Big Data using C++ and MPI.
  - Assignment 2: Kx in-memory database and complex event processing and analysis on a large shared-memory machine.
  - Assignment 3: Credit risk and real-time enterprise CVA using streaming Big Data with Apache Spark.

4. **Course workload and grading**

This is a challenging course. But the effort is worth it. (I myself do very well as a Big Data consultant using the same skills as I teach!) And many of the skills you learn in my class are transferrable to other areas of quantitative finance, such as high-performance C++14, parallel programming, solving problems at scale, and mathematical modeling (of data).

The course will be graded as follows:

- Assignments 1 to 3, each 15%. In terms of assignments, if you spend more than 8 hours per week (on average) over the first six weeks of the course working on them, something is wrong. Come see me and I (or the TA) will help. The assignments are really important because that's where you learn the most; if you do the assignments well, you will learn an awful LOT. All assignments are individual assignments, because group assignments have been a problem in the past.
- Final exam 25%. Three hours.
- Class participation 5%.
- MFE forum participation 5%.
- Professionalism (how you present your work and work with others, timeliness of submitting work and attending class, clarity of communication, and so on) 10%.
- Instructor discretion (points for flashes of brilliance) 10%. Yes, some past students have gotten the full 10% !
- You will lose points for the following:

- Disrupting class (e.g. making or taking phone calls in class): -5%
- Missing an individually scheduled tutorial: -5%

5. **Course pre-requisites**

This is a hands-on course, so you had better be ready to roll up your sleeves and get your hands dirty!

So, what do you need to know before you can begin the course? (By the way, some of these prerequisites are covered in pre-course lectures before the course proper begins.)

The "Big Data in Finance" course is a fast-paced hands-on course that covers a lot of material in just seven weeks. Prerequisites have been kept to a minimum, but understanding the concepts and tools below will make the course both easier and more enjoyable:

- ***Programming***: Some experience with programming. Ideally in C/C++ (just the basic language and syntax), but experience in any language will do. Familiarity with some dynamic language, such as Python, would be a plus, but is not essential.
- ***Mathematics***: linear algebra, matrices, calculus, and basic statistical concepts. All at the high school level or above.
- ***Command line & scripting***: Familiarity working at the command line, whether on Linux, Windows or Mac.
- ***Data analysis software***: Some familiarity working with data and graphing data in Excel, R, or other software package.
- ***Technical writing***: An ability to express yourself succinctly, and clearly, in written English; and you must be comfortable reading documentation of various kinds: course notes, assignment specifications, academic papers and articles, software specifications (APIs) and other technical documentation.

Note: The course outline is still being finalized and some parts of the course may be subject to change.