Week1

INTRODUCTION

Hi Everyone,

I am Anthony Pagan from New York. I have been working in backend infrastructure for 15+ years mostly focused on 3 tiered internal and vendor-based applications. I have some scripting experience using powershell and some introductory knowledge of R and Python.

I am interested in Data analytics as a tool to build processes that can uncover hidden system behaviors in our current environment. Outside of my current position, I am also interested in learning how to use data analytics as a social science tool.

I look forward to the weeks ahead and hope to extend my Python knowledge in relation to data analytics.

QOTW

Week1

I don’t have a lot of experience in Python. Most of my exposure has been through MOOC training programs like Udemy, Coursera etc. I have also been exposed to R programming through our program.

From some reading I learned Python has been available some 1991 while R was available as of 1995. Python had a bit of a head start, but R is catching up. They are both used in Data analytics. R may be easier to use for some users, while others mays say Python has less of a learning curve. Some say Python to provide better performance than R. Both are open source.

Do experienced data analyst prefer one over the other, or are both language a necessity for various data analytics projects ?

Week2

I agree with the other responses, it’s too early to tell if Julia is a keeper. Although Julia appears to address the issue of speed, especially for complex calculations and large data, it needs some more years to work out bugs and kinks. As mentioned in Agile Scientific article below, Julia is compiled at runtime, which explains its speed, unlike Python, which is an interpreted language.

<https://agilescientific.com/blog/2014/9/4/julia-in-a-nutshell.html>

Julia has some promise, but not ready for primetime quite yet.

Week 3

I have used applications as simple as NotePad ++ for basic scripts, Powershell ISE for powershell scripting, R Studio for R and IDLE for Python. I have tinkered with Visual Code, which can be used for several languages, but it has taken me a bit to get used to.  I have begun using Spyder this semester. Spyder has a similar layout as Powershell and R Studio. I now prefer Spyder to IDLE for Python because of its familiarity.

Experienced developers would probably prefer an IDE that can handle different languages so they can do all their coding in one place. As a novice coder, I would also prefer an IDE that supports multiple languages if it also had the ability to execute test code in one place, with auto-complete and debugging tools. NotePad ++ supports multiple languages, but the code has to be tested outside of the application, so it is not a real IDE.  Visual Code may suffice for what I am looking, it just requires a little more time to explore and learn.

Week 4

I first heard about blockchain when the value of Bitcoins and cryptocurrency exploded a few years ago just before it came crashing down. My understanding is that it is a shared security concept where there are an unknown number of checkers with decentralized controls. It has a chance to disrupt markets and change the way we do business and exchange information. My opinion is it is an ideal solution for data science when working with confidential data. Some of the blockchain mining ideas of using high powered processors and taking advantage of grid computing for high throughput, falls in line with what is needed to work with large datasets in data science. I haven’t really dug too deep into blockchain, but its approaches to data and processing is worth deeper look.

Week 5

I have used SSIS for Extracting, Transforming and loading data. It is a useful tool for anyone that is used to working with MS SQL. I have started to use Orchestrator for Microsoft. It is one of many tools from Microsoft System Center. It is used for automation and the GUI is similar to SSIS. It combines automation with data manipulation. I have also used powershell to extract and transform data from application that have powershell modules as well as PowerBI and Qlikview.

Week 6

I usually have an optimistic outlook on life. In this response I’ll take a pessimistic and slightly morbid view.

I remember seeing a documentary on civilizations awhile back. There was a part on the translation of writings on the wall from 1000’s years ago. As they read it, I realized the translation could have been placed in any era of human civilization and still be the same.  Although technology has advanced rapidly since those writings on the wall from a millennia ago, our traits and behaviors as a civilization has remained the same.

I think AI can will be the next phase of technological evolution as it produces more efficient processes and allow us to live longer. At the same time, it can be the next phase of human devolution. As we continue to depend on technology, it will not only automate processes, but make decisions for us.

As an example, let’s say you brought a red car. Did you make the decision to buy the red car or did AI analyze your social media, your social connections, words you text, emailed and commands you relayed to Siri Echo. Then based on the verbs and nouns you used, your current age, marital status, race, religion and other attributes AI calculates with a 99.9999 probability that you have just enough in your bank account to buy the red car you never wanted on the first Monday in June if it sends enough emails with words red and car hidden in the wording 1 week prior.  Did you make that decision? Sure you did, at least you think you did.

In a way, AI could be a good thing. If it is built with good intentions, maybe it will make enough decisions for us to move our civilization past the writings on the wall from 1000 years ago and into the next phase of human development.

Week 7

Information is the key to everything and it is everywhere.  Devices are getting smaller and data can be embedded into a circuit board as tiny as a fingertip. With that comes an exponential growth of data.  Someone, somewhere has to make sense of all if it.

Tools are what make Data Science what it is today. The math and data analytic have always existed, but the advanced tools are what bring everything together and transform the mathematical data analytics into data science.

Just like any other business, Education industry is in the business of making a profit and staying relevant. Data analytics is a need that many businesses are trying to fill and the Education industry is trying to businesses fill that need.

I for one prefer the online course because it gives me more flexibility. With all of the collaboration tools available, any interaction you would normally have in an in-person classroom can happen online. The online interaction is sufficient for me. There are plenty of additional free courses online to fill any gaps that are missed in online classes. I use the MIT Open courseware and Khan academy for any of the Math and Statistical Gaps and Coursera or Udemy for coding.

Week8

Data Science is a game changer when you see how it may have impacted the last election. The days of using static demographic data such as age, location and ethnicity no longer give a competitive advantage. Using data science on social media analyzes changing dynamic living data and results in understanding individual behaviors, perceptions and influences. There should be additional disclosure by social media companies so the public is aware how their data and activity is being analyzed in-place and which organizations the results are being sold to.

Week 9

When you go from MB/GB to TB/PT datasets the big concerns are performance and cost. Larger datasets require higher IO on disks, more memory for in memory processing and high throughput networks. Standard drives would have lower IO throughput than Solid State Drives. NAS/SAN drives can be configured for higher throughput, but may not give same IO as SSDs. All of these factors affect read and writes, and will alter the speed of processing and reading data. As a result, application performance and reporting tools can be affected. High performance needs result in higher costs. There are some alternatives like cloud computing that can take advantage of grid computing, parallel processing and can separate services from OS needs. These cloud options can help reduce costs. Examples include Data as a service, Platform as Service, Software as a service etc…

Week10

I don’t have experience managing a Data Science team, but do manage a small team of Infrastructure specialist. Depending on the management level you may eventually go from a tactical mindset to a more strategic one with team-oriented objectives. I agree with the article on bringing in team members that are a good fit for the collective group. Adding member from different background with skills that complement one another strengthens a team’s ability to build creative solutions.

Week12

If and when we get to the point of technological singularity, jobs may be the least of our problems. It will be the point where technology surpasses human intelligence and the point of no return. In the book Home Deus by Yavil Noah Harari, he talks about how algorithms will, and in some case, have already removed humans from decision making. He believes that at some point humans would either merge with technology or be replaced by it. I think it’s a bit extreme, but if you step back to look at big picture, the best approach is to embrace the technology, find ways to use advances for the greater good, instill good values onto the next generation so they can continue the inevitable progression of technological advancement.

Week13

In the NY Tri-State area, the salaries for any field will be higher than average. With the cost of living added, it probably evens out in other locations. Like any other career, providing well rounded value to your employer increases your value. Knowing data, statistics and programming is a big part of it, but not all of it. Good communication skills and some business understanding should go a long way on the initial salary as well as continued increases as the years go by.

I would expect the salaries to continue increase in 5 years. Information is key to any business and having a team or an individual that knows how to extract, transform and prepare the data and communicate ways it can help the business is priceless.

This Forbes article predicts that Data Science salaries will increase 28% by 2020:

<https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#4483f8b7e3bd>