Week Two Lecture Videos

Indiana University Bloomington: Big Data Applications

Naimesh Chaudhari: naichaud@iu.edu

Assignment 2 – September 14, 2019

1. Abstract

In this paper, I provide a brief introduction about the information presented in the Week 1 lecture videos. I also provide additional information on a subtopic that I am interested in from the videos and finally I provide ideas that could improve some of the existing work that has been done.

2. Summary

2.1 2019 Hype Cycles

In these lectures the focus was primarily on how technologies evolve over the years. I will primarily focus on the 2019 trends as they are the most recent. Generally, a hype cycle has a few stages: innovation triggers, peak of inflated expectations, trough of disillusionment, slope of enlightenment and plateau of productivity. The technologies in 2019 that I am interested in are AI PaaS and autonomous driving.

2.2 Clouds/Big Data Applications

These lectures focused on big data & big data deluge turning into the deep learning deluge. Data is growing at a rapid rate over the years. By 2023 we are expected to reached close to 163 zettabytes. Most of this comes from data replication. Cost of storage is getting cheaper. GE is one of the larger companies that has focuses on intelligent machines via the industrial internet of things. They collect a tremendous amount of data. We also leaned about how cyberinfrastructure defers from non-cyberinfrastructure. There isn't much difference other then its on the internet and is cheaper. Finally, we learned about how Google, Facebook, and Microsoft are focused on being machine learning first companies.

2.3 Jobs

The discussion for this lecture was focused on jobs, more specifically jobs related to data science, clouds and computer science, and computer engineering. There is a concern that robots will take over in the future, but in reality, these trends have been happening since 1920. Jobs change over years. China leads the market in cloud jobs. There will be a lack of people in the future with talent in the big data field. Analytics and data science jobs have grown rapidly over the years. There are no unicorns in big data employees.

2.4 Industry, Technology, & Industry Trends

These lectures discuss how computing cost, bandwidth, & storage costs have gone down over the years. The tech disruption driver are cheaper computer powers + increased storage capacity. Increase usage of social media, smart phones, parcel volume, and online advertising over the years. China has the largest amount of mobile user. Basically, cloud in inevitable.

2.5 Digital Disruption & Transformations

Due to private shippers' postal services have reported a loss. The evolution of commers has transitioned from stores -> more stores -> malls -> e-commerce. E-commerce has caused malls and stores to collapse. Media, sporting goods, and hobby goods and still primarily bought in a retail environment. Internet sales are growing but retail sales have plateaued. Reduction on households with major channels like ESPN.

2.6 Research Models

We learn about the 4th paradigm, how to go from theory to data driven science. We learn about the petabyte age. Now the quest of knowledge begins with massive data. The 4 paradigms in order are theory, experiment or observations, simulation of theory or model, and data drive. Generally, more data beats better algorithms and displayed by the Netflix challenge.

2.7 Computing Models

The computing model primarily states clouds are it. Python adoption has grown rapidly over the last few years. Java & C are most popular and have been reducing over the year but of late are rebounding. We need lots of computers to support the computing model. From 2008-2012 cloud computing and big data have been dominant in the garner priority matrix. Cloud service continues to grow with Amazon being completely dominant. Going to cloud allows for little to no infrastructure and maintenance but requires customers to rely on vendor security. Each datacenter is about 11.5 times the size of a football field. This allows for cheaper storage, cost, and network. Virtualization made many things convenient.

2.8 Data science Pipeline

The data science pipeline is going from data to information to knowledge to wisdom to decisions (DIKW). In the wisdom & decision area community acceptance is important. Google maps data comes from traditional maps and satellites. Information is presented by basic google maps. Knowledge is created through optimized routes. Decision are made via taking those routes. In the hype cycle for data science the peak of inflated expectations is in augmented analytics, auto ML, citizen data scientist, and decision management. Leaders is the platform are RapidMiner, KNIME, SAS, TIBCO Software, & Alteryx.

2.9 Physics

We learn about the discovery of the Higgs Particle with large hadron colliders. LHC produces close to 15 petabytes of data per year. Runs underground in 2 major circles. Projects are ATLAS, ALUCE, & CHS. The LHC computing grid has close to 200,000 cores. Data is analyzed to get insight.

2.10 Recommender Systems

Netflix uses recommender systems for their users. Recommender systems are created to give personalized matching of products. Amazon does buy recommendations, YouTube does video recommendations., and Pandora does music recommendations. They solve a matching problem. Idea is map users in a space with their items and find other similar users. Distance measures such as the Pearson coefficient and Cosine Measure can be used. KNN algorithm does a good job of finding neighbors. Clustering can be done through LSA or LDA process.

2.11 Web Search & Informatica

Within web search we get all the words and word positions that are on a page and process it using TF-IDF to quantify importance of word matching apply a rank

and recommend the top ranked pages. Yahoo uses a multi objective optimization system to make recommendation to users on their page.

2.12 Cloud Applications

We learn about the science of clouds and internet of things. Major providers have large scale super computers. It is projected that there will be 24-75 billion devices on the internet of things. In the internet of things devices communicate to each other. Drones are used in agriculture, sports and disaster relief. QR codes are used to scan and be scanned. The peak hype for internet of things was in 2016. GE has the Predix platform for its internet of things. The IOT platform market is projected to grow substantially in the next few years. The primary messaging standard is HTTP.

2.13 Parallel Computing & Map Reduce

There are a tremendous cloud offering from man companies in the big data field. The apache big data stack has close to 21 layers. Google implementation of some of these layers took many years. Parallel computing is composed of HPC & Clouds that use MapReduce. MapReduce take a set of key value pairs and reduces them at different layers to get a final list of values.

2.14 Online Education & Conclusion

The idea of big data has transitioned to Cloud, Edge, & Deep Learning. Clouds are here to stay. Data Analytics & employment opportunities in the field continue to grow. Data science & distance education continue to grow. Data science is composed of domain expertise, programming skills, research, & statistics and data mining. There are a broad range of topics in the data science field. Online education has a large adoption than residential.

3. Interested Subtopic

The subtopic I am interested in lies under the recommender systems branch. I would like to research this further on how other companies have built recommender system using deep learning. I will review the systems that is developed by ike. Intuition engineering states, "*iki is your personal career and professional consultant powered by machine learning and artificial intelligence related technologies. After a user has defined the areas of his professional interests, iki creates daily content feed developing knowledge and chosen skills with the help of the recommender system with the deep learning architecture described below*."

Ike captures six parameters from its user to build the recommender system, user interest and skills, user like disliked content, user watched content, user SVD (Singular Values Decomposition) profile, vectorized content, and, content SVD profile. They used the glove language model which was developed by Stanford. Their vector contained 300 dimensions. They calculated the tf-idf weights and calculated the weighted sum for ranking. As each user enters their personal bio of skills, they can compare to other users who have had similar skills and recommend skills that the original user has not explored yet.

Intuition engineering states, "*The goal of iki recommender system is not only to provide the content on corresponding topic but also to rank content elements by quality and expertise level to fit the resulting feed to each user's level of expertise."* They used multiple layers to make the data denser for the recommender system.

They measured their deep learning solutions to the commonly used methods of creating recommender systems (SVD, NNMF, et al.). They specifically went towards the deep learning method because not only did it outperform the common methods it was also much faster. Due to

their large amount of user base, they were concerned with model performance on typical methods. Due to the deep learning framework being able to run on GPU they were able to get good recommendations with fast performance. They also did the dense approach because in real world user can enter a wide variety of data, to be able to make that dense gives them a way to process it efficiently.

This is another example of how big data and deep learning are helping developers create new offering to consumers. I am extremely excited about this field and can't wait to learn more.

4. Ideas that can improves existing work

One of the major concerns that many companies have with moving to cloud is the available security. Business have proprietary data that they need to protect. Putting this data on the cloud means they must trust the third-party provider on their security. This loss of control seems to be a big factor for business that choose not to move to the cloud. I believe having more education in this would allow businesses to finally make the move. I am confident with both Amazon and Microsoft to safeguard our data, but I do wish they could be more transparent about the security levels. The other things between the two larges providers is that it is easy to get on the cloud but after one is in the cloud, if one wants to pull out, one must pay a plethora of fees to do so. This gives business the feeling that their data is held hostage. I believe the providers should be more transparent about this and come up with a strategy that allows its customers to make changes or moves between providers quickly without additional fees.

Second in todays world we must buy into one of three largest providers if we do move to the cloud. The technology is evolving quickly, and different providers provide different niche services. In the future it would be very useful if the three largest providers could collaborate and allow for cross platform services. This ideally would be the most beneficial to the customer and they can pick and choose the service that want implemented.

Third as mentioned in the data science pipeline video there are not that many providers that support the entire data science pipeline. The leader in this is RapidMiner, I would like to see additional vendors approach this space and develop fantastic tools. In the corporate world if a citizen data scientist is armed with the proper tools, he would be able to add much value to his department. Although open source software like R and Python have fantastic packages to develop ML models, not all users are good at the programing side of this. Having tools in the Auto ML space or tool that require minimum coding would allow novices to enter the field and really start exploring the capabilities of machine learning.

Finally, as learned from the videos there are countless technologies out there in the Big Data space. At times it is difficult to understand which of these technologies are really needed for the corporation and which are not. Also, this space is evolving rapidly and there is a feeling that by the time you decide on an infrastructure, there is already some new and better out there. I would like to see vendors demystify this.

5. References

1 Intuition Engineering. (2019). *Recommender systems with deep learning architectures*. [online] *Available at: https://towardsdatascience.com/recommender-systems-with-deep-learning-architectures-1adf4eb0f7a6 [Accessed 15 Sep. 2019].*