EAS 504 : Applications of Data Science - Industry Overview

Assignment - 9

Name : Aboli Rawanhate

UB Number - 50374341

Q1. Describe the market sector or sub-space covered in this lecture.

The market sector or sub-space covered in this lecture by Arun Venkatachar and Sashi Obilisetty was Data Science in Electronic Design Automation (EDA). EDA industry is primarily concerned with chip design and the design of electronic devices such as printed integrated circuits, circuit boards, electronic devices using Synopsys and software tools, as well as determining the flow for implementation, synthesizing, and commercialization required for manufacturing these products. This business provides a way to create software programs, connect them to hardware, and map them. These software tools are used to mimic, imitate, and synthesize the design until it is ready for production. In order to develop this design, this sector is separated into three primary business sectors.

1. Software and special hardware that are used to design integrated circuits.
2. Units of chips (IP) such as USB, HDMI and memories.
3. Software development tools to check the quality and security.

He also mentioned the firm Synopsys and its working as EDA. It provides tools and ways for designing and verifying complex chips and the advanced processes and models required to manufacture the chips. To achieve high optimization, Synopsys fabricates silicon chips using TCAD and lithography equipment. Then to ensure high productivity, best-in-class digital and bespoke AMS platforms are employed in design. A wide selection of silicon-proven IP, including the greatest interfaces and embedded memory. The quickest engines and unified platforms are used to verify hardware and software. Advanced software tools are used to assess and improve application security and quality.

The aforesaid characteristics lead to increased semiconductor chip sales and scientific and technological progress. The widespread use of AI chips, growing IoTs, 5G, and the automotive industry's move toward self-driving cars have all contributed to this industry's phenomenal growth in recent years.

Q2. What data science related skills and technologies are commonly used in this sector?

- Domain expertise, data understanding, asking the proper questions and determining what sort of issues to address, methodical data analysis, and varied and unconventional thinking are all essential abilities.
- On top of ML methods, APIs such as Data access API, AutoML- High Leveled API are established. Data Cleaning, Data Transformation, and Data Normalization are all part of the Data Access APIs. Regression APIs use ML algorithms like regression, ensemble

methods, regularizations, instance-based algorithms, and so on; Classification APIs use ML algorithms like KNN, Decision Tree, Random Forest, LinearSVC, MultinomialNB; Clustering APIs use ML algorithms like Affinity Propagation, Kmeans, MiniBatchKmeans, DBScan, BayesianGaussianMixture, GaussianMixture and so on.

- Data Lakes, Data Storages, Machine Learning and Deep Learning libraries and frameworks, AI APIs, and In-House APIs are some of the data science technologies utilized in the EDA industry.
- Many reinforcement learning approaches are employed in any design where there is a requirement to increase synthesis power and analysis.
- Machine Learning, Distributed Computing, and Cluster Orchestration are some of the data science capabilities employed in the EDA business.

Q3. How are data and computing related methods used in typical workflows in this sector? Illustrate with an example

Different data and computing techniques are utilized to tackle EDA difficulties utilizing data and computer related technologies in the EDA business. Incremental algorithms, divide and conquer, parallelization, and sampling approaches are all used in ML algorithms.

We will explore Common ML Platform(CMLP) as *an example* in the EDA sector. It has key features like Agility which allows for faster development of machine learning applications, Scalability that allows for better data management, Deployability which is a dependable and error-free installation procedure. Data Collection, Data Processing, Visualization, and Collaborating APIs are the most important phases of CMLP:

- Data Collection - Data is gathered from a variety of sources, including system tools, user systems, bug tracking systems, and user compute. Data is collected using a variety of tools and libraries. CLL: Synopsys' Common Logging Library is a tool-independent key-value data model library for data gathering.
- Data Processing - Distributed data storage and object storage is used for storing data, and it is processed using a variety of ETL technologies.
- Data Visualization: Machine Learning methods are used to make important conclusions. Visualizing such findings is made easier with tools like QlikView, Tableau and Power BI.
- APIs for collaboration: This aids in the integration of hardware and software.

Q4. What are the data science related challenges one might encounter in this domain?

Data science related challenges in this domain are:

- Data from one design may not convert to another; data from one process may not translate to other processes; models developed from a small corpus of designs may not

translate effectively; customers may need to provide more data to improve model efficiency.

- If the data is missing, it must be produced from scratch, including the labels. The produced data may produce biased results, skewing the outcome.
- It is not the amount of the data that is the issue; it is the quality of the data that is the issue.
- Deep learning models are employed for superior integrated architecture. While training the model with millions of data, they might be computationally costly.
- Invest in a big-data platform strategy - Data collection and administration are critical components of analytics and machine learning; either create your own or hire a provider.
- ML models that will be utilized for training should be stable and portable. As a result, there should be a strong reliance on design and procedure.

Understanding the key focus while getting information will help you overcome these obstacles. It is necessary to assess if efficiency or quality is more essential. Features like digital and bespoke AMS platforms, software development tools, and high-quality silicon proven semiconductor chips can help achieve this.

Q5. What do you find interesting about the nature of data science opportunities in this domain?

The most interesting I found about the nature of data science opportunities in this domain is Al Chip Market. Al is quickly becoming the key driver of semiconductor growth. Following the plateauing of smartphone/mobile devices, the next "wave of innovation" will emerge. It is a significant driver in a variety of economic sectors, including robotics/autonomous agents, smart cities, and emerging areas. In the context of EDA, ML is defined.

Statistical techniques, heuristics, prediction, and other methods of learning have been used in EDA. As classical algorithms reached their limits, the usage of machine learning in this area has increased. Traditional algorithmic/heuristic techniques are being pushed to their limitations by the increasing complexity of specific EDA challenges. The large amounts of data provided by tools provide little or no utility. Users may utilize machine learning to acquire better insights about how to manage complexity increase.

To drive outcomes, ML employs statistical models created from data - ML-based computing approaches are currently being used to these challenges. In the EDA sector, machine learning algorithms provide statistical optimization at scale for prediction, classification, and/or estimation. Machine learning is a technique for automating the construction of electrical circuits and optimizing mechanical designs. For faster resolution of issues in electronic design automation (EDA) and VLSI design, such as circuit recognition, machine learning is used.

Q6. Describe some of the challenges in applying machine learning approaches to this domain
Data Quality, Data Diversity, and Investment in Big Data Platform Strategy are the three primary problems that this sector faces.

- Data Quality: It is not the amount of the data that is the issue; it is the quality of the data that is the issue.
- Data DIversity: data from one design may not convert to another; data from one process may not translate to other processes; models developed from a small corpus of designs may not translate effectively; customers may need to provide more data to improve model efficiency.
- Investment: Invest in a big-data platform strategy - Data collection and administration are critical components of analytics and machine learning; either create your own or hire a provider. Investing in a big data platform approach might be hazardous since clients may request model upgrades, which can raise or reduce model efficiency.

Q7. Describe two illustrative use cases from this domain where ML approaches have been successfully used.

VC-Static : VC-Static is a tool that improves result analysis by employing an unsupervised learning technique. In general, the results provide a lot of false positives. Clustering is used in this tool to increase debug productivity. It does analysis and prioritizes left to the user, with the user debugging issue symptoms and prioritization by violation severity/volume to provide clusters of findings. These clusters assist users in debugging clusters and root-cause scenarios in clusters. It also does Auto-prioritization by violation impact by assessing the impact of a cluster on the system.

PrimeTime ECO : A static timing analysis, such as PrimeTime ECO, is a way of verifying a design's timing performance. It looks for any potential path violations. PrimeTime is the industry's most used static timing analysis tool, and it employs supervised self-learning. A classification model is used to rapidly determine which gate or buffer should be replaced. It is made up of two primary components. - Existing flow based on PBA time and power analysis of signoff quality - ML-based training on the same block to achieve optimal power QoR. The ML-based flow has the same timing and power QoR as the old flow, but with a -6X quicker turnaround time.