Applied Accelerated Artificial Intelligence
Prof. Satyajit Das
Department of Computer Science and Engineering
Indian Institute of Technology, Palakkad

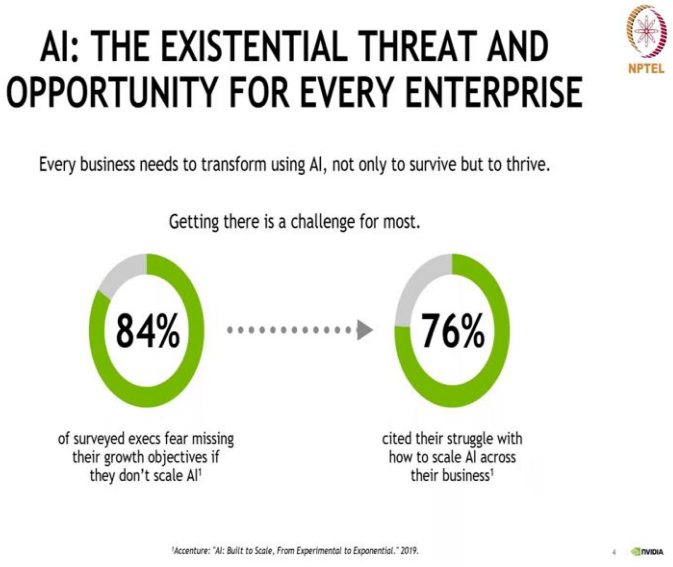
Lecture - 44
Accelerated Data Analytics Part 1

Welcome everyone to the next lecture. We have been covering different topics and different scenarios which are prevalent in the space of deep learning, artificial intelligence having its relevance in the industry. Today we will start a new a series of lecture targeted primarily towards isolated data analytics stack and one of the key framework or the SDK Software Development Kit that we are going to use is called as Rapids.

(Refer Slide Time: 01:01)

So, let us get started. So, let us first start with the motivation itself as usual.

(Refer Slide Time: 01:06)



So, there was a survey which was done by Accenture and primarily the survey was targeted towards from a point of view of not the end users, but targeted towards enterprise segment and also the scale at which the management or the higher level segment would like to basically see if they want to use AI if they do not want to use AI and there were many sets of questions which were asked.

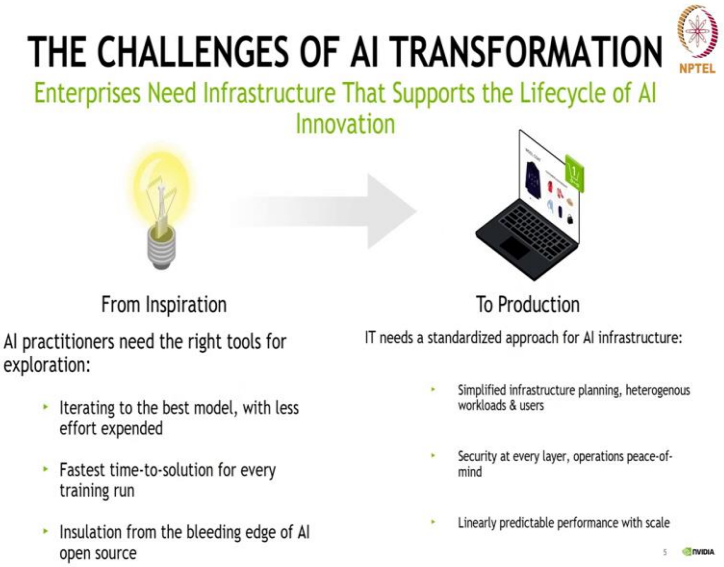
So, to the executives when there was a question which was asked with respect to AI, 84 percentage of the executives they feared missing their growth objective if they do not scale to AI which means they all wanted to be part of this journey of artificial intelligence, which is a very very critical point because if you look at artificial intelligence it is not about coming out with algorithms or anything.

It is a different mindset its a different business flow for which it requires a top to bottom approach where there are the higher executives who should be looking at how to integrate this new way of programming or a new way of providing service using artificial intelligence. And most of them were able to relate to that there is a necessity of using AI for their own enterprise business that they have been working on.

But out of that almost 76 percent of them also sighted their struggle with how to scale AI across their business. So, the problem is not just about using AI, but also using it effectively and scaling it within the organization in such a way that it becomes integral

part of the thinking process itself and that is something which is very very difficult to achieve in any organization.

(Refer Slide Time: 03:23)



The challenge of AI is not about adopting a new technology. The challenge of AI is about adopting the lifecycle of AI innovation which is from inspiration to production. AI or deep learning became really relevant at this point of time and data everything became really important because it was being applied for a real world use case.

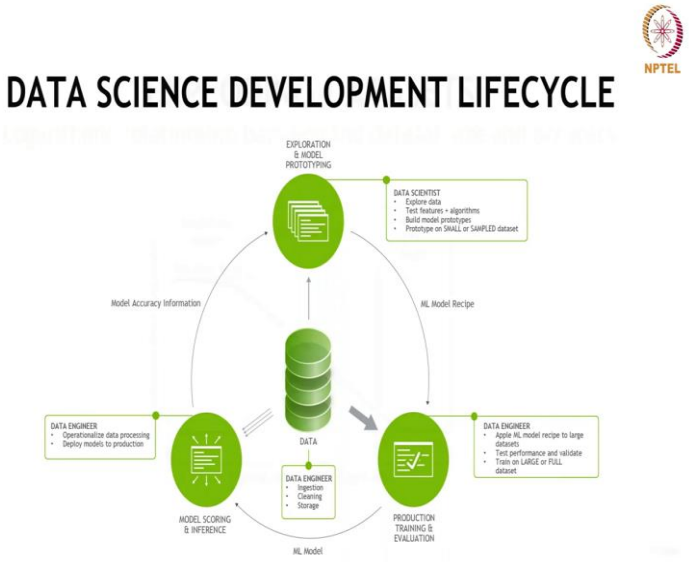
AI practitioners basically want right set of tools for exploration, they want to try out the best model and they want to spend as less effort as possible. The time to solution should be very very fast when you are trying to train it right. And we should be always at the bleeding edge of AI which means, it is also looking at if there is a possibility of having an open source initiative also.

Now, to bring it to production if you see it moves from a practitioner basket to the ID and integrating it into the AI infrastructure itself. There should be a simplified infrastructure planning heterogeneous workloads and user.

AI can bring in different services the workloads or the kind of segment that you like to have is quite varied you would require a large amount of security and also you need scaling predictability, which means that you should be able to scale with respect to the number of users or the cases that you are going to have. So, the challenge of AI

transformation is not just about as we said about the technology, but it is much more than that.

(Refer Slide Time: 05:12)



You have been hearing a lot about different kinds of AI job roles or the data science job roles which have been there out in the market.

This is the development cycle of a typical data science, the first key player in this is the data scientist. The data scientist is responsible for exploration and model prototyping basically takes the data explores different kinds of data tries to add new features or tries to explore different features and different kinds of algorithms. In the end the idea is to build a model, which is at the stage of prototype.

The prototyping is generally done on small and sample data set. The output of this activity is basically a model or ML model, which is given to the next user which is the data engineer that is the next role. The data engineers role is to take this ML model recipe and apply it to a larger data set. The reason why we have differentiation between the data scientist and data engineer here is primarily because of the aspect of not just looking at the algorithmic side of it.

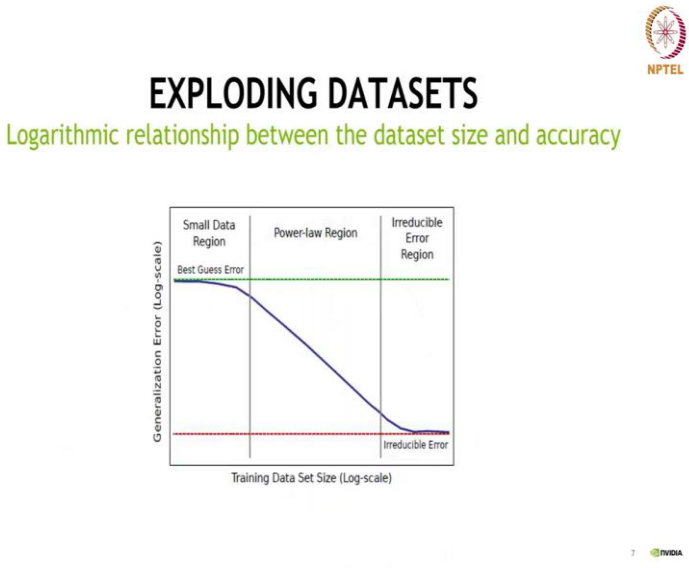
But also needs to apply it to a larger scale to be relevant and we will talk about that in a bit as well. The idea is not just to talk about accuracy there, but also to test the performance and validate it and the training happens generally on the overall data set.

So, the responsibility of data engineers is to make your model production ready. The output of that is not just a model recipe but the actual ML model. This ML model is provided and is again taken by the data engineer to make it operational in nature and deploy it. Deployment is as critical as creating the model itself because if it does not suffice the necessity of the end user and if it does not match the performance criteria's that model may not be of any use.

Finally, once you deploy and when you have done the production training and evaluation with respect to larger data set there are chances that you might end up having certain accuracy related problems because as you saw previously also when you want to deploy it in the real world you might have to reduce the overall precision to improve the overall speed and there are other kinds of techniques which are available to do that which may affect the accuracy.

So, it again goes back to the data scientist this is the overall cycle of a data science if in an enterprise segment the data which is at the centre of all of these things needs to go through different set of operations ingestion cleaning storage. So, that data is at the centre of it and that is why the name also which is given to different nodes which is data scientists and data engineers.

(Refer Slide Time: 08:50)



We have been seeing this particular trend with respect to increasing. The data set over time the amount of data sets have increased. The data set when they increase it is also

proven that when the data set increases and if you train it on larger data sets the amount of error also reduces or obviously, after certain duration there is nothing more which can be done which is in this region of irreducible error. But the relationship of data set size to accuracy has been well established in different papers.

(Refer Slide Time: 09:35)

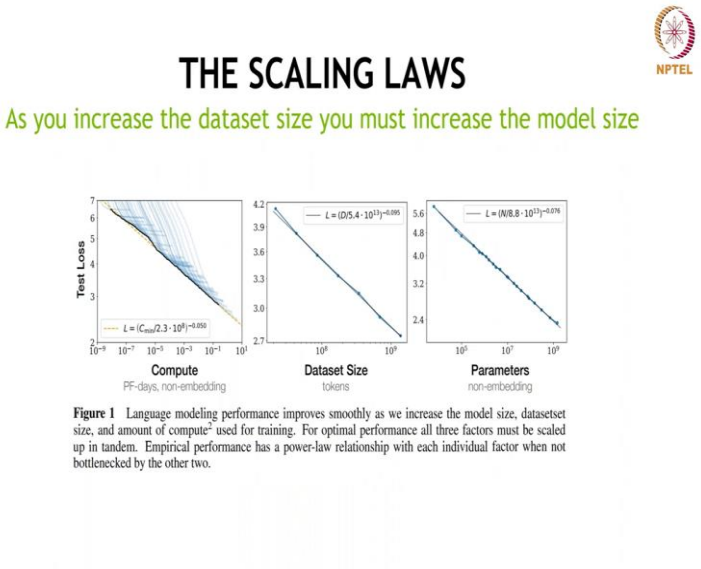
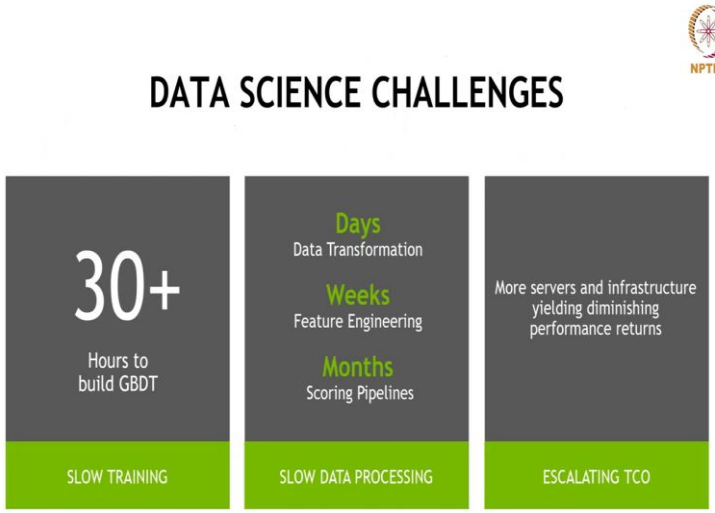


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

It is also established that when we increase the data set size, you require much more many more complicated models resulting into more necessity in the computation because when your model size increases you also end up increasing the amount of computation that is required.

So, there is a linear relationship there is a relationship which is directly proportional to the amount of computation which is required to learn all, the pattern from the data also.

(Refer Slide Time: 10:13)



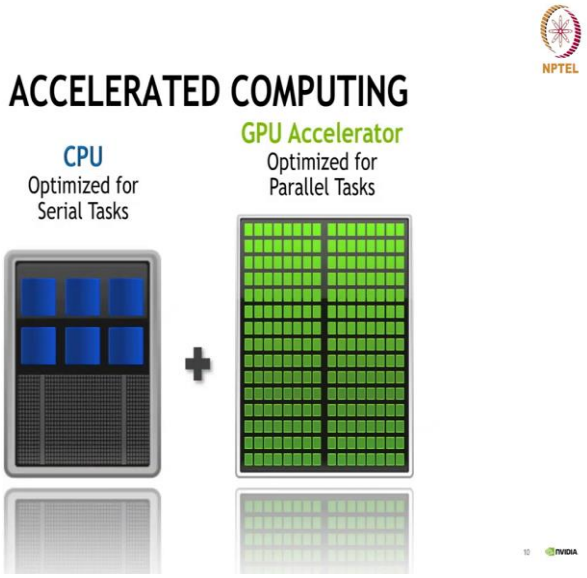
So, just to summarize what we have talked about from a data science challenge perspective. There are three kinds of challenges or it can be put into this particular definition.

So, it may require a lot of time to basically build your model which is falling into the regime of slow training. It may require days for data transformation may require weeks for feature engineering, it may require months for scoring the pipelines. Slow data processing is a very very big problem which can result into not being at the bleeding edge of AI more servers and infrastructure yielding diminishing performance returns.

So, it is not just about increasing the number of servers or increasing the number of computation, but it can result into diminishing returns as well. So, return on investment with respect to infrastructure is another aspect that most of the executives are also concerned about.

So, in order to enable the data scientists and the data engineers in order to improve the training and improve the data processing speed but at the same time making sure that the return on investment is being met is the challenge of from transformation from an idea to a actual deployable product.

(Refer Slide Time: 11:45).



One of the solutions which have been brought in which we have spoke about in various previous lecture is the usage of GPU along with a CPU.

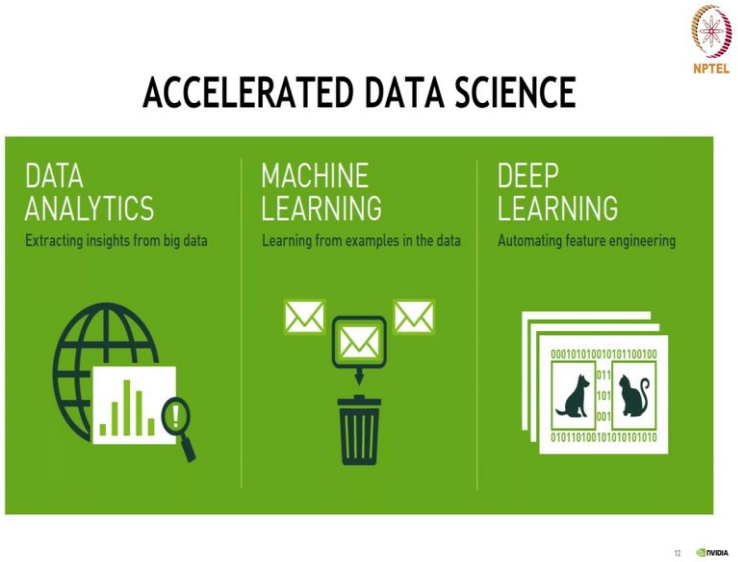
GPU Graphics Processing Units are fundamentally a parallel computing architectures initially used for graphics processing, which is also a parallel computing job can be applied to artificial intelligence data science and data analytics to speed up various of this task which we just talked about some time back and also making sure that they return on investment is also met.

(Refer Slide Time: 12:32)

The good thing about GPU computing is that not anyone needs to start from scratch. Over time there have been a lot of development in the GPU ecosystem, a complete stack has been built. One of the critical things for AI enterprise segment is the necessity of having a stable software stack it is not just about the hardware or about a product which is just made open source and there is no support attached to them.

GPU computing has seen various changes and there are more than 80 plus SDKs which are present which solve this problem. If we talk about adding data analytics we are particularly going to look at one of them in this particular lecture which is rapids. In the future lectures you would be also going through certain additional SDKs, which is like deep stream which is used for the intelligent video analytics space which is the next set of lectures that you are going to see. So, today we are going to concentrate on SDK which is called as rapids.

(Refer Slide Time: 13:51)

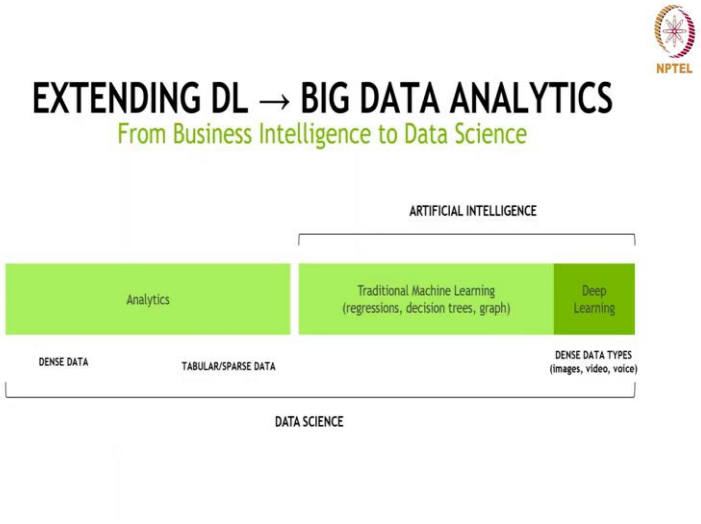


It is also important to understand different terminologies which has been there. Data analytics primarily refers to extracting insights from big data and it is a larger regime and fundamentally it has been there for a very very long time you have seen various frameworks coming there. It includes almost everything which is machine learning deep learning and all are a smaller subset of what which is the transformation of data into information or in from data to insight.

Machine learning is one of the ways to gather this insights from big data and one of the examples that we have seen in the previous lectures series is the usage of different SDKs which are very popular especially in the CPU domain like sklearn and all. Deep learning is another subset of machine learning which kind of automates the process of feature engineering in the machine learning space.

It has its own relevance, it is primarily used for unstructured data while traditional machine learning is used for structured data deep learning is popular in the unstructured data space like images or NLP Natural Language Processing and it is required and it because it automates the feature engineering part of it finds its relevance in various domains.

(Refer Slide Time: 15:30)

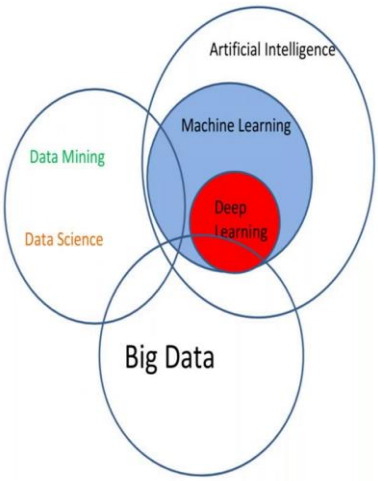


But if I wanted to say extending DL to Big data analytics or from business intelligence to data science, if suppose this was your overall gambit of things which is the overall artificial intelligence space you can see the data analytics part of it, then there is traditional machine learning which includes different algorithms like regression decision trees clustering algorithms graphs and in the end there is also data types which are of type images video and voice which is your deep learning space.

So, what we are going to concentrate today a lot is towards this particular space which is the traditional machine learning part of it.

But the data science in general constitutes of all of these parts and we have concentrated on all of these parts over this two sessions that we are doing along with you. So, today we focus primarily on these two parts.

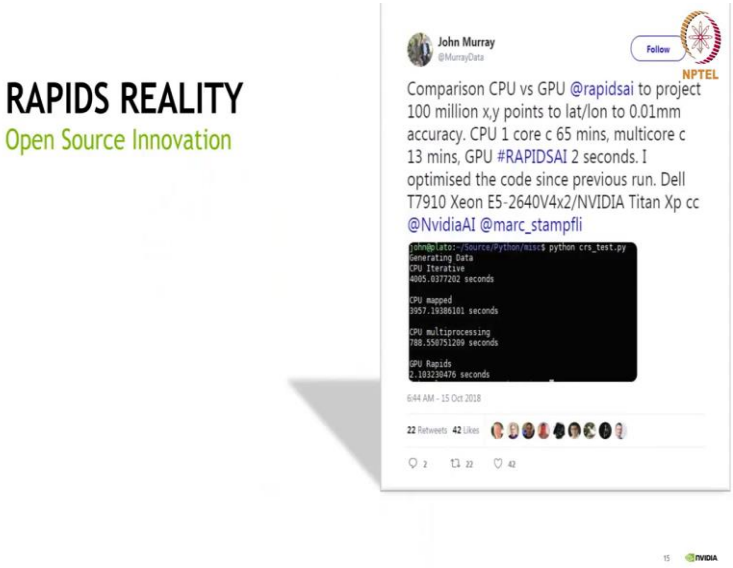
(Refer Slide Time: 16:35)



As I said some time back also. So, there have been different technologies which have evolved over time you have artificial intelligence, which has been defined by different levels, you have a subset of it which is machine learning and then there is a further subset of it called deep learning.

Big data has its overlap with the deep learning space or with the artificial intelligence space same goes with the data science and the data mining part. What we need to understand is in this crossover what is the right set of technology based on the domain that you work on or is there a crossover between them or you need to work with multiple of them for deploying for a real world use cases.

(Refer Slide Time: 17:31)

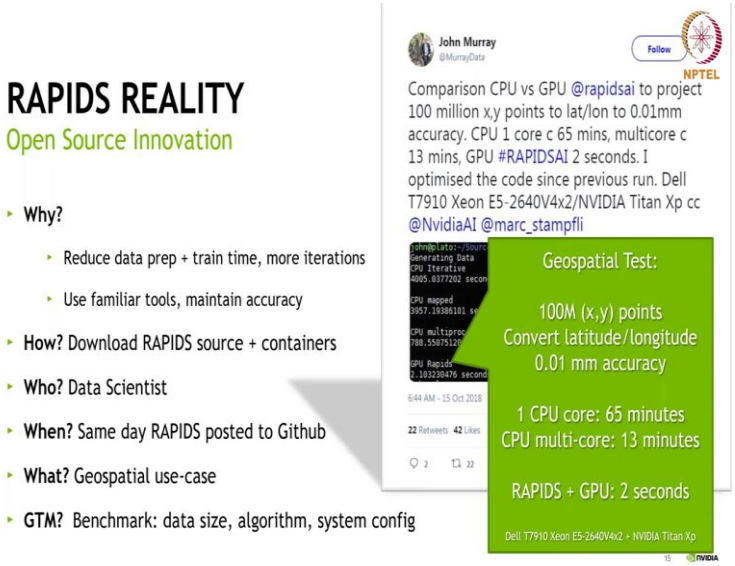


So, let us get started with what is rapids. First of all one thing that needs to be known about rapids which we talked sometime back is that, it is a open source innovation platform. You can see here on some of the comments which are there on this open source community which is the comparison of CPU versus GPU.

The project name itself is rapids AI and if you were to train it for 100 million parameters for a 1 CPU code it took around 65 minutes for training same thing when converted to a multi core took around 13 minutes and when we reached the part of GPU the minutes further reduced to 2 seconds.

So, that is the kind of speedups and these are the tweets or these are the kind of appreciations being made by the users.

(Refer Slide Time: 18:27)



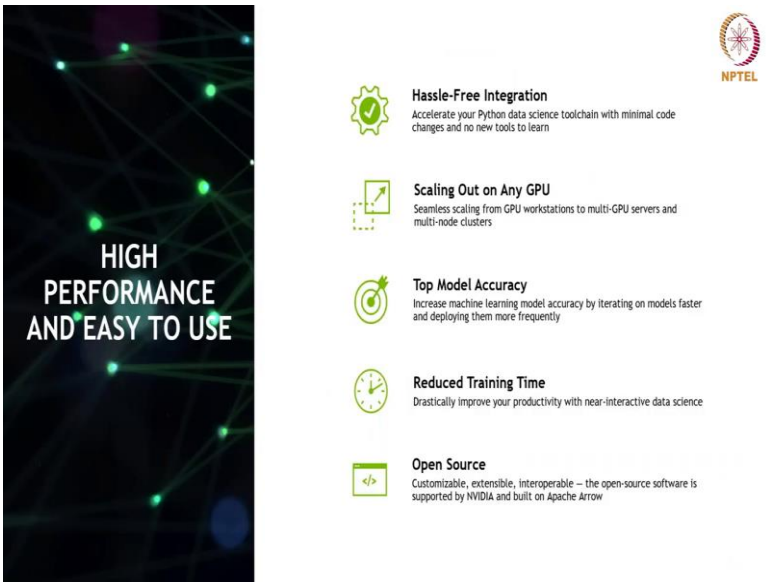
You can see here its exactly the same thing for geospatial test data all of these results are kind of mentioned. So, the rapids innovation is required because we want to reduce the time for data preparation training time and you want to do more number of iterations.

We also want to be consistent or we want to be want to use a particular SDK which is very familiar to my previous generation SDKs, which is CPU based SDKs that we have been using, but at the same time it should also maintain accuracy. So, I want to use or have my older tools which have been the traditional codes running on the CPU I do not want to invest a lot of time into getting familiar with the new tool.

So, it should look something similar like my older tools, but at the same time also maintain accuracy. How? Its very easy because rapids is a open source project you can download it and will see through quanta you can download it in form of containers and all and for whom it is?

It is primarily for data scientists and data engineers and since it is open source you can basically post it I mean you can download it anytime from GitHub in fact, you can contribute it back to the GitHub as well.

(Refer Slide Time: 19:56)



So, the reason why it is very very easy to use is because first of all you are using your traditional Python environment.

So, it has a complete Python environment data science tool chain and there is very minimal code, which is required to be changed because the APIs have been designed in a way that they look very similar to the ones that you have been using in the traditional CPU environment. The good thing about rapids is that it does not just make it parallel on a single GPU, but you can take the same application and scale them across multiple GPUs across the whole cluster.

So, it has support for multi GPU multi node kind of environment to enable the data scientists and data engineers to train much more faster. It provides the same level of accuracy that you would have seen in a traditional CPU based environment it also reduced a lot of training time it drastically reduces it.

Last, but not the least it has very I would say one of the most best open source license it is customizable, it is extensible and it is interoperable with the other frameworks if you want to use which have been traditionally used in your original enterprise segment.