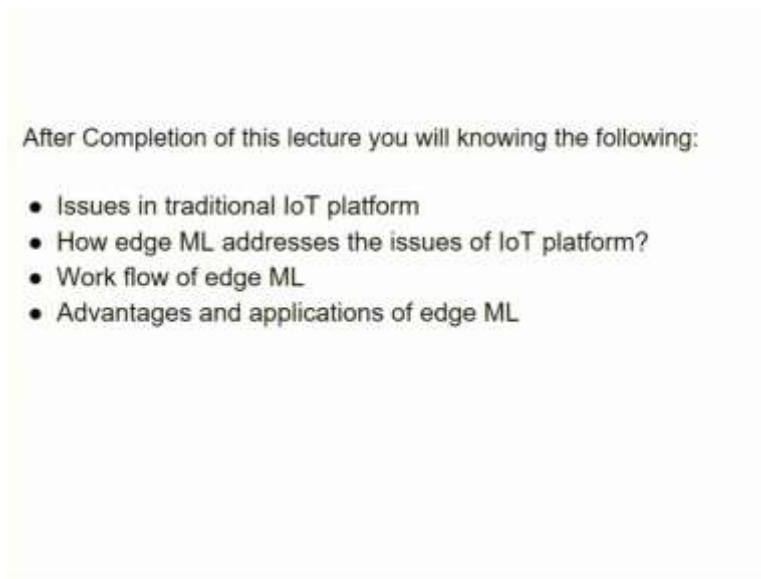


Foundation of Cloud IoT Edge ML
Professor Rajiv Misra
Department of Computer Science and Engineering
Indian Institute of Technology, Patna
Lecture 05
Enabling Intelligence at Edge layer for IoT

My self Dr. Rajiv Misra from IIT Patna. The topic of today's lecture is enabling intelligence at the edge layer for Internet of Things.

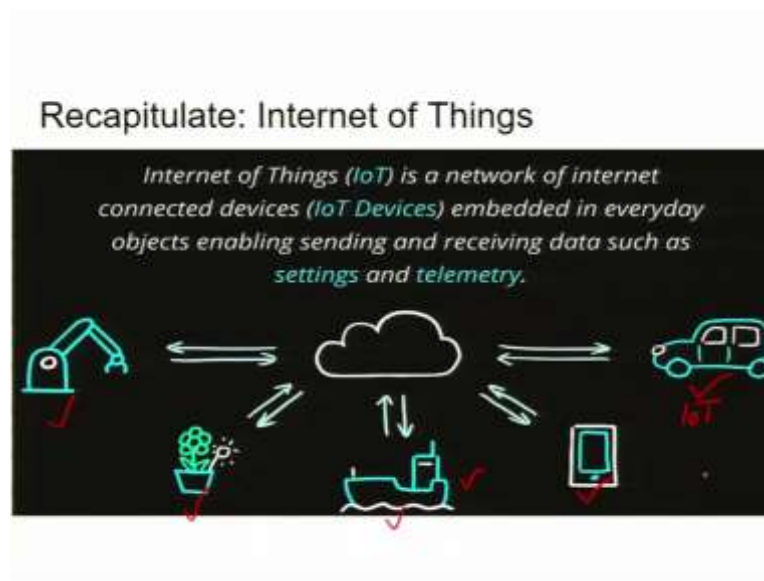
(Refer Slide Time: 0:26)



So, in this lecture, we will be covering the following topics. First thing is that you will be understand to know the issues in the traditional IoT platform, then you will also come to know how the edge machine learning addresses these issues which arises in IoT platform that is in the traditional form of IoT, then, we will introduce you to the workflow of edge machine learning the advantages and the application of edge machine learning.

So, these are the use cases or these are the topics which will enable you to understand the use of machine learning for IoT. So, that is called an edge machine learning for Internet of Things.

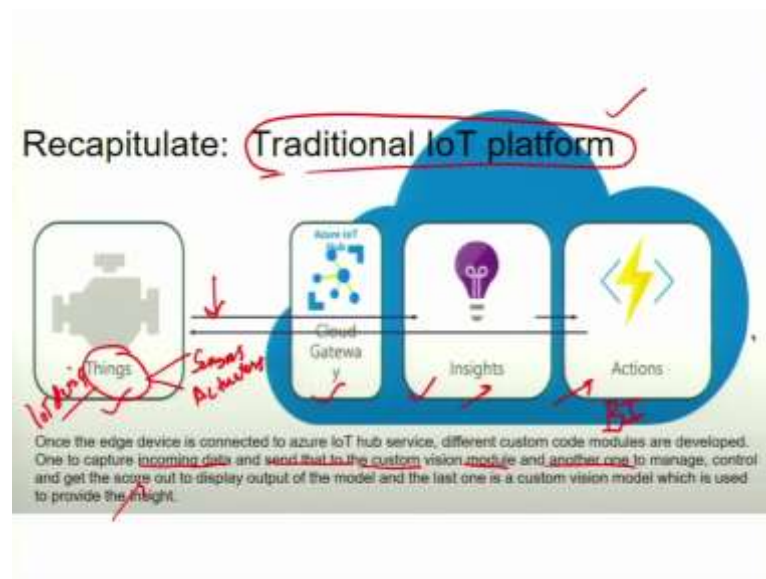
(Refer Slide Time: 1:16)



Let us recapitulate about the Internet of Things. So, Internet of Things is a network of internet connected devices that is called IoT devices, which embedded in everyday objects this enables sending and receiving of the data to the cloud, such as the settings configurations and telemetry. So, look at this particular use case or the example that these are the different things which you can see over hear that is whether it is a car or a vehicle, you have a mobile phone or you are moving on some river over the ship or you are having a garden or a flower or a robot, all these are different devices.

Which are embedded with the internet connected devices, they are called IoT devices. So, Internet of Things is the network of internet connected devices, which are embedded in everyday objects that we have seen, which enables sending and receiving of the data to the cloud, such as the settings and telemetry etc.

(Refer Slide Time: 2:34)



So, we will also recapitulate about the traditional IoT platform where these are the things or Internet of Things, which we have now discussed that is the network of internet connected devices, whether it is the sensors or actuators these are the specific things which we are now interested as the IoT devices. So, in the traditional form of the platform or in a classical IoT system, earlier days, you can see that these things are Internet of Things, they were directly connected for computations to the cloud.

So, you can see this kind of connections is happening to this particular cloud this is an example of a cloud and inside the cloud that we have already seen in an IoT platform, the data which is sent from the IoT devices will be used for three different use cases. The first is that after ingesting the data with the help of a cloud gateway, then the next step would be that this particular data is now processed and so, after processing, it will gain the insight into the data which is brought by the sensors, which are embedded in the everyday objects.

So, these insights will be now given to the business intelligence for making the decisions or the actions based on the policies. So, these are the specific use cases whereby these things or Internet of Things are directly communicating to the cloud, they are dependent upon the cloud. So, that is what is written let us understand it once the device that is that is the IoT device is connected to this cloud platform or an IoT platform. Then it will be able to use different custom code or modules which are there which are running inside the cloud.

So, that is it will be in the form of capturing or ingesting their incoming data. And then now it will be sent for doing the Insight using the custom build modules. And then it will be now

will be used another module to manage control. And to get the source or the details out and display on the dashboards, these are some of the actions of the reports and these all things are covered under the insight. So, this computation is performed directly by the cloud in the traditional IoT platform system.

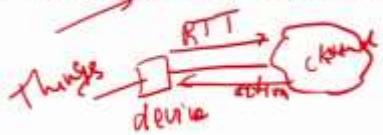
(Refer Slide Time: 5:20)

Recapitulate: Limitation of traditional IoT platform

Poor internet connection, when the internet is down the system fails. For example, if a smart fire alarm system just detect fires when internet connection is up, then it fails in performing its task.

Data gravity, IoT devices create lots of data that demand more way to find the insights locally on a device than shipping all of the data to the cloud. For example, a smart doorbell, you don't want to stream video to the cloud 24-7 just to identify faces for the two minutes that someone is in front of your door, you would rather do that, locally on the smart doorbells.

Real time responses, as opposed to near real time responses that you cannot get by sending data to the cloud finding insights and then sending the actions back down.



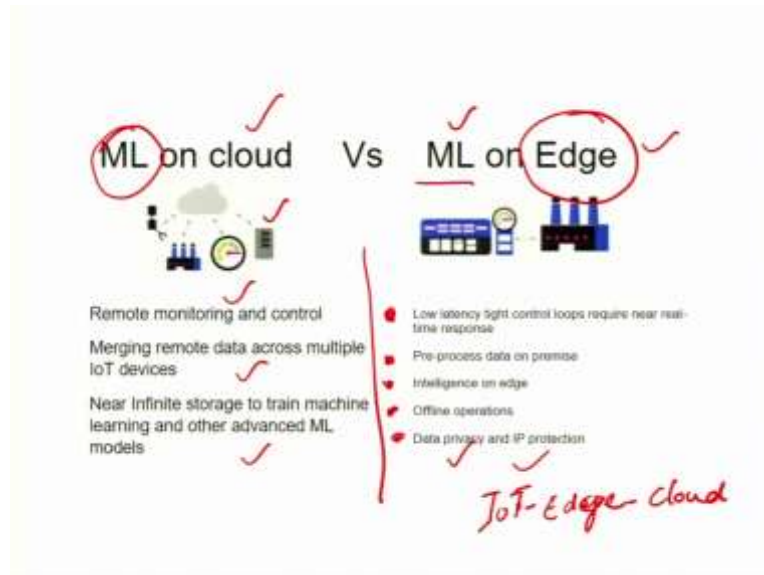
Now, there are see what is this limitation, although it has the advantages, but it has some of the limitations, that limitation comes in the form of a poor internet connectivity. So, if the internet outages are there, then this particular cloud connectivity will fail. Hence, the insight or the actions which are dependent upon the cloud computation will no longer be accessible. So, it is not that all the time, internet connectivity is there, for IoT devices, then you have another aspect that is lots of data has to move from because it is not only one or two IoT devices, it is hundreds and 1000s of IoT devices.

Which are embedded into the daily objects. So, these devices are now connected to the cloud. So, so much of data is to be shipped to the cloud for getting the insight. So, this is a issue, where too much of data is being swamped into the cloud, and they dependent upon the cloud services all the time and also the internet connectivity, this is another problem.

Third problem is that which is happening in the traditional IoT system, IoT platform is called real time responses, as opposed to the near real time responses, we cannot get by sending the data to the cloud to find the insight and sending back the actions. So, therefore, this particular delay or the latency is due to the round trip time from the device or you can also call as the things to the cloud.

So, sending the data are getting the insight and then actions are to be now taken or being communicated by the cloud is taking lot of time and some of the applications cannot wait for that longer duration. Therefore, not many applications are supported in by the traditional IoT platforms.

(Refer Slide Time: 7:35)



So, this particular drawback is due to the sending the IoT data that is the things to the cloud. And most of these insights is happening with the help of machine learning. So, to perform the machine learning into the cloud, the data has to be sent back to the cloud and then cloud has to report about the actions. So, these kinds of situations are happening in the remote control or a remote monitoring type of situations.

That means if you want to do a remote monitoring or control of an industrial situation or industry or a specific machine, using this IoT or other device, which sends the information and waits for actuators or waiting for the response out of this insight or merging the remote data across multiple IoT devices, that is aggregation or near infinity storage to train the machine learning models and other advanced machine learning models.

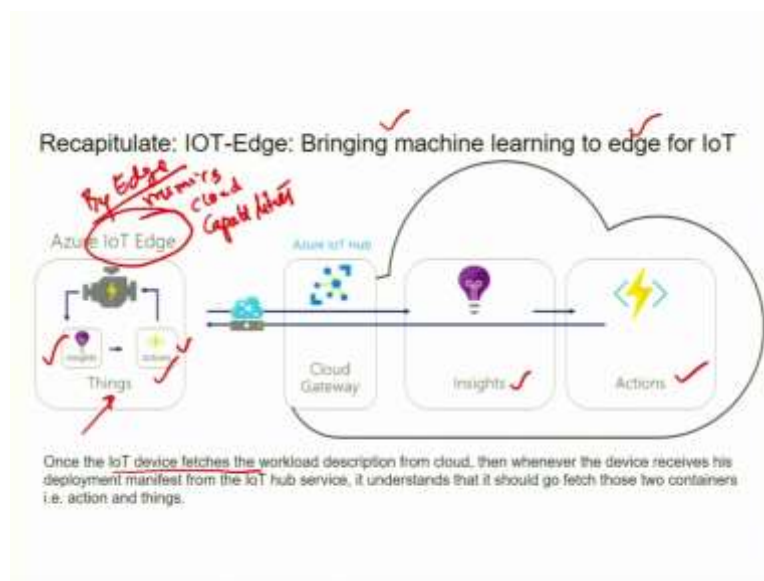
So, this particular data not from one sensors, but So, many sensors are required to be aggregated and this data is to be stored somewhere into the cloud and then the machine learning algorithm can be performed. So, we have seen the drawbacks of this particular model. So, what is the way out is performing the machine learning very close to the device that is called machine learning on the edge.

So, what are you gaining? So, you are gaining the following advantages that it will be a low latency tight control loop near real time response, if the machine learning is being performed on the edge that is very close to the source not on the cloud. Also, the sensor data will be now preprocessed on premise need not have to send everything to the cloud.

Third important thing is that you can do the intelligence to get the insight a preprocessing of machine learning is to be performed at the edge. And if let us say that most of the task is completed need not have to go to the cloud for this purpose. Similarly, if there is no connectivity between the edge and the cloud, that means that edge can also work offline to support the offline operation that is also possible, if let us say machine learning is being done at the edge.

Finally, the data need not have to send directly to the cloud all the time therefore data privacy and IP protection, all that is being met. So, therefore, nowadays, you know that doing machine learning on the edge will add a layer of edge that is called for IoT that is called an IoT Edge, is to be added in between this particular IoT and Cloud.

(Refer Slide Time: 10:38)



So, this is the model which we will discuss also, we have so, far briefly mentioned in the previous class. So, what we are going to see today is that, bringing the machine learning to the edge for IoT, so, therefore, you can see that earlier slide things was only having that the devices, but now, the things is now populated with the insights which earlier also will be done only by the cloud and taking the action also can be performed at the edge.

So, this kind of edge is now adding up the capabilities which or it mimics the capabilities by edge you can see here that these are the capabilities, which earlier used to be performed at the cloud, which is now enabled to be performed at the edge. So, if you are doing the machine learning to get the insights so machine learning inferencing will be performed at the edge for doing the IoT.

So, here once the IoT device fetches the workload description from the cloud, then whenever the device receives this deployment manifestation from the IoT service, it will understand that. So, these are more details about how this insight and action you can run in the edge as well, we are going to discuss this in more details about this development.

(Refer Slide Time: 12:18)

The slide features a title 'Enabling Intelligence at Edge layer for IOT' with a red oval around it. Below the title is a paragraph: 'To manage the increasing amount of data that is generated by the devices, sensors, most of the business logic is now applied at the edge instead of the cloud to achieve low-latency and have faster response time for IOT devices using Machine learning at edge.' A red bracket underlines this paragraph. Below it, a red arrow points to the text 'Edge layer is delivering three essential capabilities'. This is followed by a numbered list: 1. local data processing, 2. filtered data transfer to the cloud and, 3. faster decision-making. A red line underlines the second and third items of the list.

Now, enabling intelligence at the edge layer, this is the topic of today's lecture after clearing up the background of simple IoT platform. So, enabling the intelligence at the edge layer for IoT is the topic of this particular lecture, the idea is to manage the increasing amount of data that is being generated by the devices that is the sensors and the actuators, which is to be sent to the cloud for doing this kind of business intelligence or performing the business logic.

Now, you can do the same kind of business logic at the edge instead of the cloud. Therefore, if data need not have to send to the cloud, you will achieve a low latency that is the delay will be not there and also have a faster response time for IoT devices, if you perform the machine learning at the edge. So, we are now here that the theme of this particular session or a lecture is enabling the intelligence at the edge layer for Internet of Things.

Now, let us understand by how are you going to enable the intelligence at the edge layer for Internet of Things. So, for that, we have to understand what the edge layer has to deliver. That is how the edge layer is going to mimic the capabilities of a cloud services. So, edge layer will be delivering the three essential capabilities. One is that it will enable the local data processing at the edge.

Second thing, this particular capability of edge layer will be to filter the data transfer to the cloud that is it can do most of the job at the edge but whatever it has to be supported only by the cloud only that will data will be filtered and then transferred to the cloud. Third important thing is that the business intelligence or the logic itself runs over the edge therefore, it will enable the faster decision making. So, let us see all these three different essential capabilities, which are to be enabled at the edge layer, how that is all being supported.

(Refer Slide Time: 14:47)

Enabling Intelligence at Edge layer for IOT

- Local data processing:**
 - In order to deal with increasing amount of data generated by sensors, most of the business logic is now deployed at the edge layer instead of cloud to ensure low-latency and faster response time.
 - Only a subset of the data generated by sensors is sent to the cloud after aggregating and filtering the data at the edge.
- Filtered data transfer to cloud:**
 - This Edge Computing approach significantly saves the bandwidth and cloud storage.
- Faster decision-making:**
 - AI has enabled new capabilities for edge computing. Since most of the decision-making is now taking advantage of artificial intelligence, the edge layer is becoming the perfect destination for deploying machine-learning models trained in the cloud.

So, the first thing is supporting the local data processing is a important capability of enabling intelligence at the edge layer for IoT. So, in order to deal with the increasing amount of data which is generated by the sensors, most of these business logics are now deployed at the edge layer in a state of the cloud to ensure that low latency and a faster response time.

So, that means, these business logic after the insight of that particular data which is generated at the edge layer will be now then fed with the business logic which is also running at the edge layer. And this both together will enable the local data processing only the subset of the data which is generated by the sensors need to be sent to the cloud after aggregating and filtering of the data at the edge.

Now, second important capabilities is to filter the data transferred to the cloud. So, this particular edge computation that is called edge computing approach significantly saves both the bandwidth and the cloud storage. Third important thing is that, this will enable that faster decision making. So, you can see that this AI enabled new capabilities for the edge computing will enable the decision making and the advantage is in the form of doing the artificial intelligence at the edge layer.

So, doing the artificial intelligence at the edge layer is the perfect place for deploying machine learning model which is trained in the cloud for doing this inferencing that is taking the business decisions or taking the faster decision.

(Refer Slide Time: 16:28)

Performance vs Cost trade-off in IOT-Edge Platforms

ML implementation on edge heavily depends on specialized processors that complement the CPU. There's no conventional CPU can increase the speed of training ML model.

To bridge the gap between the cloud and edge, innovations in chip designs offers purpose-built accelerator that speed up model inferencing significantly. Chip manufacturers such as Qualcomm, NVIDIA and ARM have launched specialized chips that speed up the execution of ML-enabled applications.

These modern processors GPUs assist the CPU of the edge devices by taking over the complex mathematical calculations needed for running deep learning models, accelerate the inference process.

This result in faster prediction, detection and classification of data ingested to the edge layer.

The solutions like Microsoft Azure IoT Edge runtime, and the Qualcomm Neural Processing SDK for ML, makes it possible to take models trained in the cloud and run hardware-accelerated inference at the intelligent edge.

Handwritten notes:
- Edge GPU
- Following way to make independent edge = GPU - Accelerators are now available for edge layer
- AI models which are trained in cloud with help of can do inference at edge

Now, let us see more detail about this kind of IoT Edge platform. So, there are two things one is called performance the other is called cost. So, performance versus cost trade off, let us see how it reflects in IoT Edge platform. So, you can see that the machine learning implementation on the edge heavily depends on some specialized processors, that you may be knowing that these processors these are often very much needed in the cloud that complements the CPU.

That is no conventional CPU can increase the speed of training the machine learning algorithm. So, how are you going to get these specialized processors enabled at the edge. So, to bridge this gap between the cloud and the edge these innovations in the chip design offers a purpose built in accelerators that speed up the model inferencing at the edge significantly.

So, you can see a lot of chip manufacturers such as Qualcomm NVIDIA, ARM, they have launched a specialized chip that speeds up the execution of machine learning enabled applications and they have also reduced the form factor so, that they can also run at the edge as well. So, the edge now can have the GPUs also. So, they are in a small form and also they are cheaper and this particular innovation has led to this kind of edge computing or edge layer.

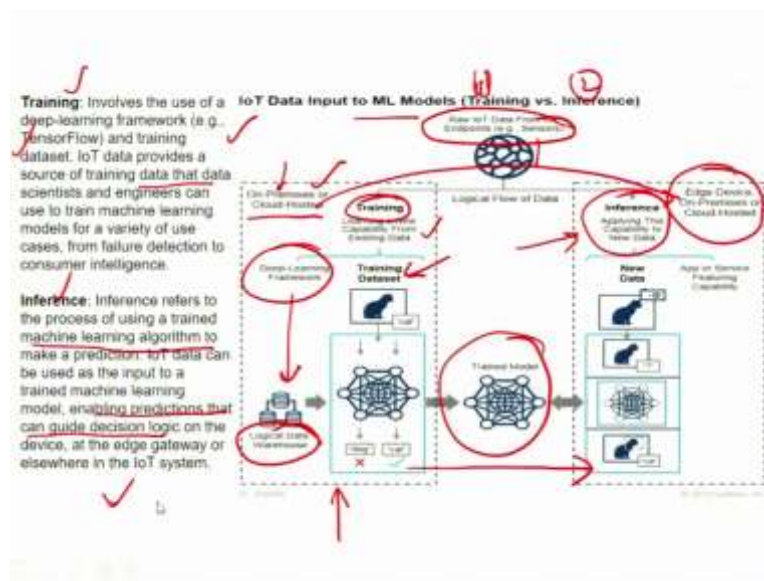
So, therefore, these modern processors which are, which is called GPUs are will assist the CPUs for the edge devices by taking over the complex mathematical calculations which are needed for running the machine learning models such as deep learning and therefore, these accelerators will accelerate the inference process. So, this result in a faster prediction, detection and classification of the data ingested at the edge itself.

So, the solutions like Microsoft Azure IoT Edge runtime and Qualcomm neural processing SDKs for making the machine learning possible to take these models trained at the cloud, but they are now running with the hardware accelerated inferencing at intelligent edge. So, how are you going to make this intelligent edge is the following way. First thing is that these GPUs they are also called the accelerators are now available for the edge layer and this is due to the innovations in the chip design.

So, once this accelerator are available at the edge, so, that machine learning model or AI model which are trained in the cloud can do the inferencing with the help of accelerators like GPUs at the edge. So, this is what we are going to see that due to the reduction in the cost of these specialized processors, which the chip designers innovation has brought into the use at the edge layer also.

So, the important chip manufacturers, which manufactures GPUs, which will run in complement to the CPUs add at the edge layer is used to do this run the inferencing which of the AI models which are trained at the cloud. This will enable this particular predictions, faster predictions, object detections, image classification of the data which is ingested at the edge layer by the sensor nodes.

(Refer Slide Time: 21:20)



Now, this is therefore, moving forward with the IoT data and performing these machine learning on this IoT data. Now, as far as machine learning process is a two process two stage process one is called training, the other is called inferencing. Now, training requires huge data set is a lot of compute power therefore, the training it itself confined at the cloud or it is at on premise servers. So, you can see here in this particular figure, that the training part of a machine learning data which is of a machine learning on a data which is now generated by these particular sensors or an IoT data need to be trained at the cloud that is called the training.

So, training means learning new capabilities from the, from these kind of IoT data. So, this has to happen at the cloud or on premise why because it needs more resources. So, this particular training requires a specific deep learning algorithm or a framework and the data set called training data set and this you know, that these two things requires a huge storage in the form of the data warehouse, this storage will feed this data set into the deep learning framework and this will generate a trained model.

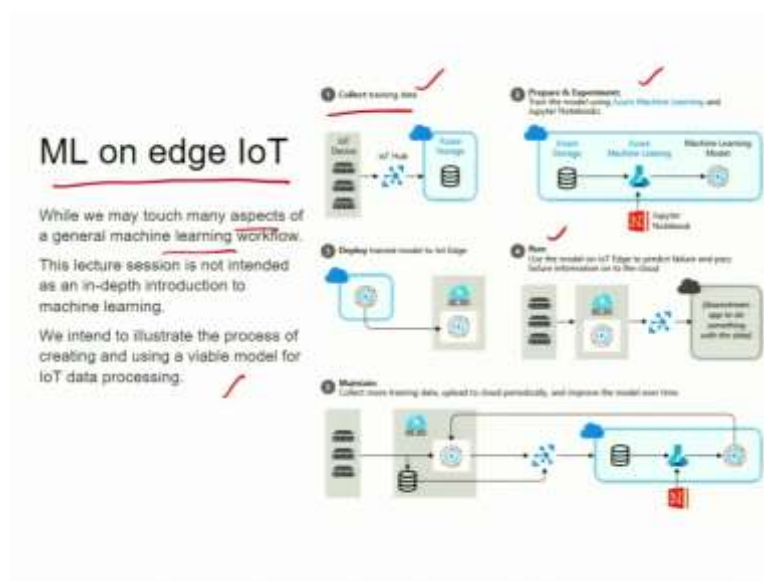
Now, this trained model, if it is deployed or shipped to the edge, now, then edge device now can work in conjunction with the cloud. So, therefore, once this trained model is deployed and run on the edge, then the other part of the machine learning can happen at the edge itself and that is called the inferencing. So, when a new data comes for prediction or the inferencing that can happen on the using the trained model or the edge and this CPUs along with complementing the CPUs, the GPUs are also there with the edge and therefore, inferencing can happen.

So, this particular capabilities are now possible with the innovation in the new kind of chip manufacturers and a new kind of technologies which are now possible, which has made this possible to run the GPUs at the edge layer. So, let us see that these kinds of discussions, which I have described in the form with explained with the help of this diagram, let us see that the training which involves the use of deep learning framework, and this is offering the tensor flow based deep learning model or a framework which we are talking about, is used to train the data set.

So, IoT provides the source of training the data that the data scientists and engineers uses to train the machine learning models for a variety of use cases, such as failure detection to the consumer intelligence. So, here the data scientist requires is required here to after the IoT generates the data, it has to build into a data set, which is to be fed to a deep learning framework that is a tensor flow based model for training and running that particular business intelligence from the inside.

So, for that you require the inferencing So, inference refers to the process of using the trained machine learning model algorithm to make the prediction. So, IoT data, then if it is presented to this train machine learning model, it will enable the predictions and that particular prediction can guide the decision logic at the device itself at the edge gateway or elsewhere in the system makes this inferencing very faster and usable for the modern applications.

(Refer Slide Time: 26:07)



So, going beyond understanding how that is possible to run the machine learning on an edge IoT. So, this is the topic of this particular slide. So, here you can see that machine learning on

edge IoT, how that is all being done. So, to do this machine learning on an edge for IoT, we have to touch many aspects of a general machine learning workflow.

So, this particular slide or this lecture is not intended to present an in depth understanding of machine learning algorithms, but nevertheless, we will be able to make the understanding here are the people who do not know much background of machine learning to understand how the machine learning can be implemented over the edge in this lecture session. So, we intend to illustrate the process of creating and using a model for IoT data processing with the help of machine learning that is what we are now intended here.

So, for this we have divided this particular process flow of doing machine learning on the edge in five different steps. First is to collect the training data, second is to prepare and do the experiments. Third is to deploy the trained model on IoT Edge. Fourth one is to run the model and get the data. Fifth one is to maintain that is to collect more and more training data upload to the cloud for retraining and improve the model over the time that is learning and relearning and so on.

(Refer Slide Time: 27:54)

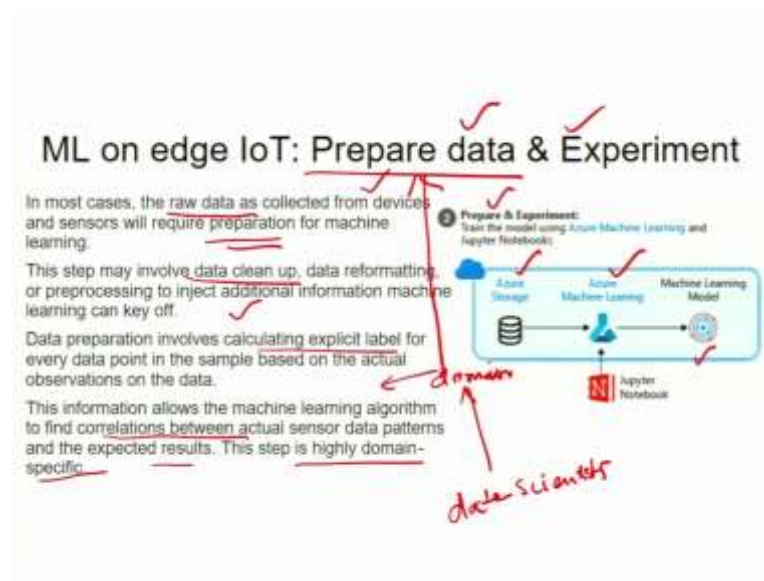


So, let us see all these steps in a more details. So, the first step is to collect the training data in this particular method. So, the process will begin by collecting the training data and you know the data we mean that it is the data which comes from the sensors. So, in some cases that data has already been collected that is called historical data. And it is available in the database or in some data files or in other cases especially for the IoT scenarios the data need

to be collected from IoT devices and sensors and now, then it will be stored in the cloud and then make it usable for doing the machine learning.

So, the first step is to collect the training data, this particular figure is explained explaining that this IoT device that is the sensor data will now send this particular source data, sourced that is data ingested into the Azure cloud storage and will be stored in some databases. This is the first step to perform the machine learning on the edge for IoT.

(Refer Slide Time: 29:20)



The second step is to prepare this particular data and it can be usable for doing the machine learning experiments. So, this raw data as it is collected from the devices and the sensors need to require, need to be prepared. So, that it can be used to build, it can be usable for the machine learning algorithms. So, it requires the data preparation. So, that is called raw data. raw data is not usable by the machine learning. If it is once it is collected from the devices requires a step.

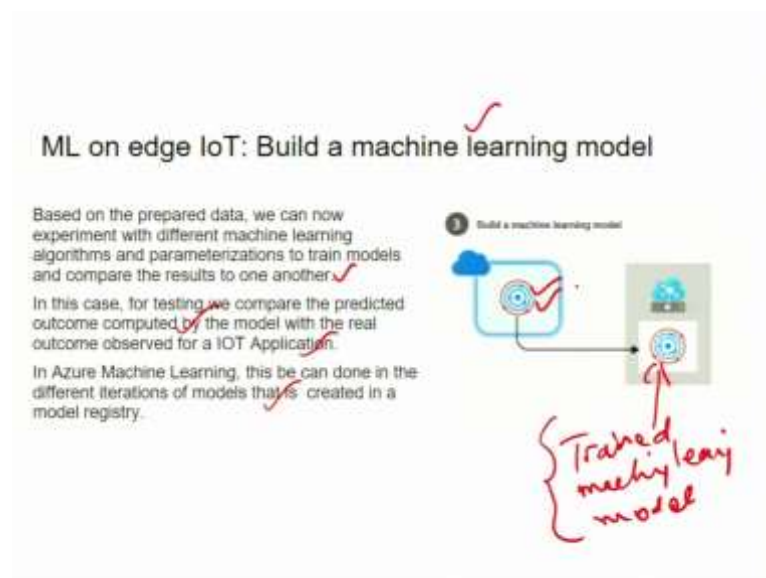
So, this step is very much essential to prepare the data So, that it can be used for the machine learning. So, this step may involve data cleanup, data reformatting or a preprocessing to inject additional information machine learning can use it. So, that is called a data preparation involves all these explicit label for every data point in the sample based on the actual observations. So, it requires the domain expert to prepare this particular data, it often needs the domain expert and the data scientist.

So, here comes the role of a data scientist in preparing the data which is now usable for the machine learning algorithm. So, this information allows the machine learning algorithm to

find once it is done, then this particular data set will be usable by the machine learning algorithm to find various correlations between the actual sensor data pattern and the expected results and this is, this step is highly domain specific, so, data scientists are needed.

So, let us see in this particular figure. So, what do you mean by preparing and doing the experiment is to train the model using Azure machine learning. So, here you can see that the data which is stored will be now made into the form which is usable by some machine learning and so, this is all happening into the cloud. So, so, using a particular machine learning algorithm, the machine learning model is being generated or is called machine learning model.

(Refer Slide Time: 31:52)



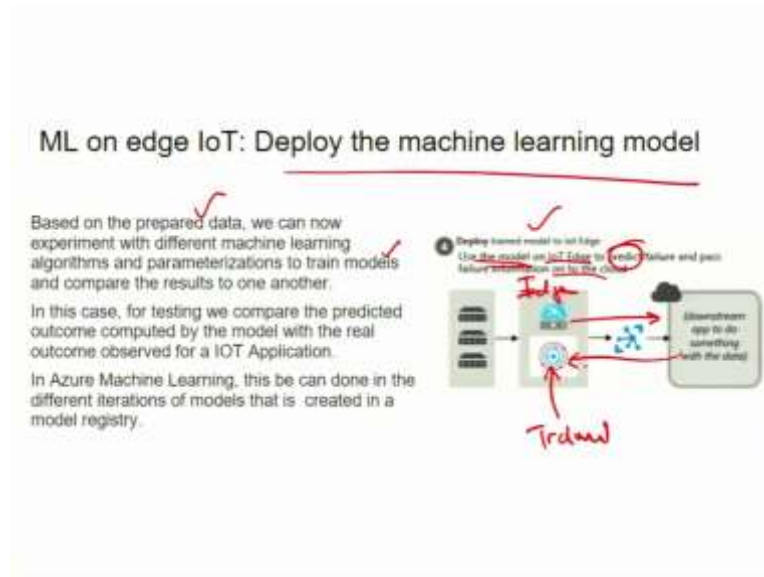
Now, the next step is to build a machine learning model is to build a machine learning model is based on the data which is prepared in the previous step. And you can now use the, you can now do the experiments with different machine learning algorithm and the parameter to train the model and compare the result to one another. So, you can see which machine learning algorithm is doing the best job for running the business logic.

So, in this case for testing, we have to compare the predicted outcome computed by different models with a real outcome for an observed, observed for an IoT applications. So, Azure machine learning this can be done in a different iterations of the model that is created in the model registry, and here this is shown over here that the model.

Which is trained in the previous step is built with the help of a machine learning algorithm and the prepared data set and this will now generate this trained model. Trained machine

learning model this icon is showing that it is the trained machine learning model which is generated at the cloud.

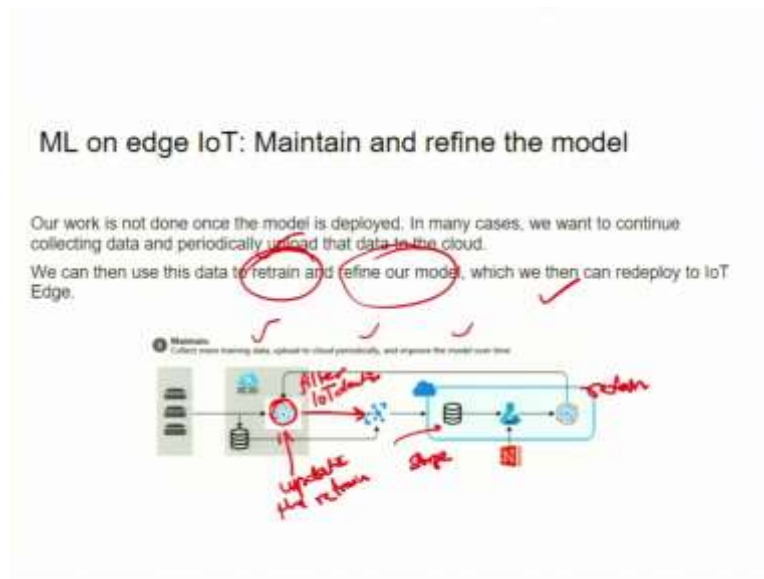
(Refer Slide Time: 33:14)



Now, once the machine learning model is available, then you can deploy the machine learning model at the edge. So, based on the prepared data, you have now parameterized and got the trained model for and this particular model now can be deployed on an IoT Edge. So, use this particular model, so, that it can be now running at the IoT edge to predict various kind of actions are the events which is very much needed for the business logic to take the decisions.

So, that is what is shown here in this particular side, this is the trained model and often this particular trained model is running now, at the edge layer. So, this is, if it is the edge layer. So, what it will do, it will do the inferencing in most of the cases, but in some of the cases where the data which is not having the good accurate result and need the machine learning model to be retrained that particular data will be sent back to the cloud for retraining and this rebuilding this particular trained model on the machine learning.

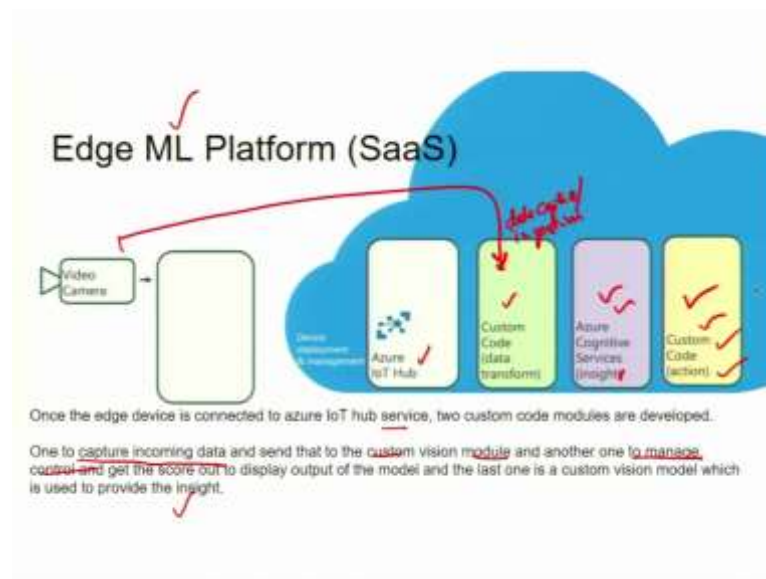
(Refer Slide Time: 34:33)



So, maintain and refine the model. So, that our work is not done once the model is deployed in many cases, you have to continue collecting the data and periodically upload it to the cloud and you can retrain and refine your model. So, this is an iterative process. So, when we can redeploy to the IoT edge, and the entire ecosystem is shown over here that this particular data after now going through this particular inferencing from this particular model.

And the edge decides that this data has to be sent after filter, this IoT data need to be sent to the cloud and this cloud and it will be aggregated here in the storage and then it will be used to retrain the model. So, retrain model is again updated so, that is what is written over here you have to keep on refining the model at the edge and this is the way so, therefore, you have to collect more training data upload to the cloud periodically and improve the model over the time and this is all done intelligently with the filter at the edge.

(Refer Slide Time: 35:57)



So, now, we are going to see how software as a service, which is being provided by different cloud provider, how this edge machine learning platform is being supported by them. So, here you can see in this workflow that this particular cloud, how it is going to offer this software as a service for edge machine learning platforms. So, one such example here is from Microsoft Azure IoT Hub, which is running in the cloud.

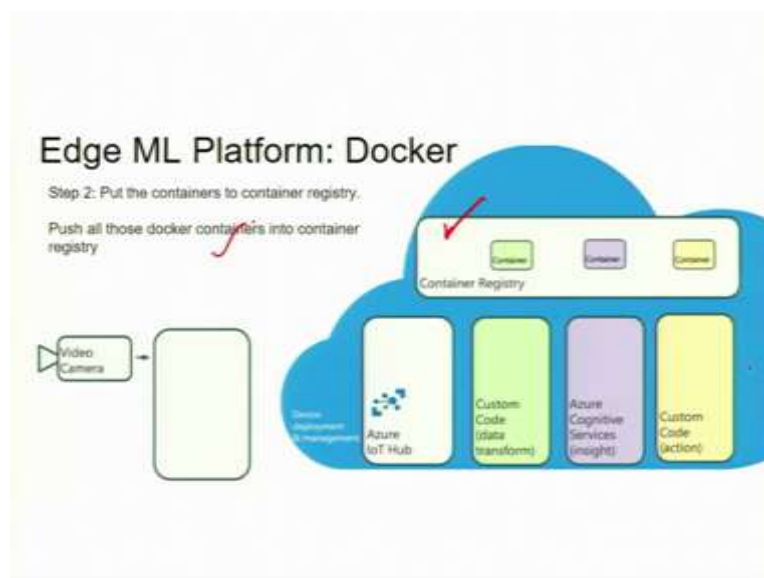
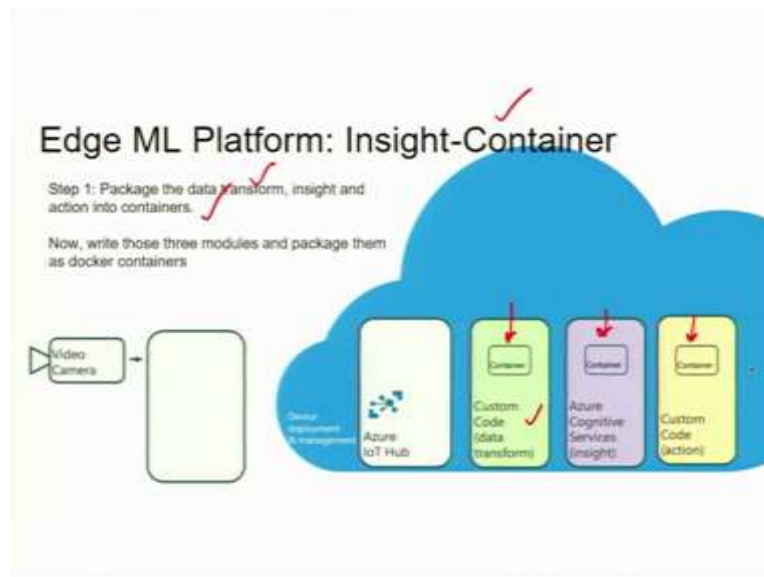
So, some of the modules which are running inside as a part of Azure IoT Hub, they are called custom code which will perform the data transformation. Then second one is called Azure cognitive services, which is used for getting the insight, third one is called custom code which is used for the action. So, let us see that once the edge device is connected to the Azure IoT Hub service, now, these custom modules are being developed one to capture the incoming data that is called here.

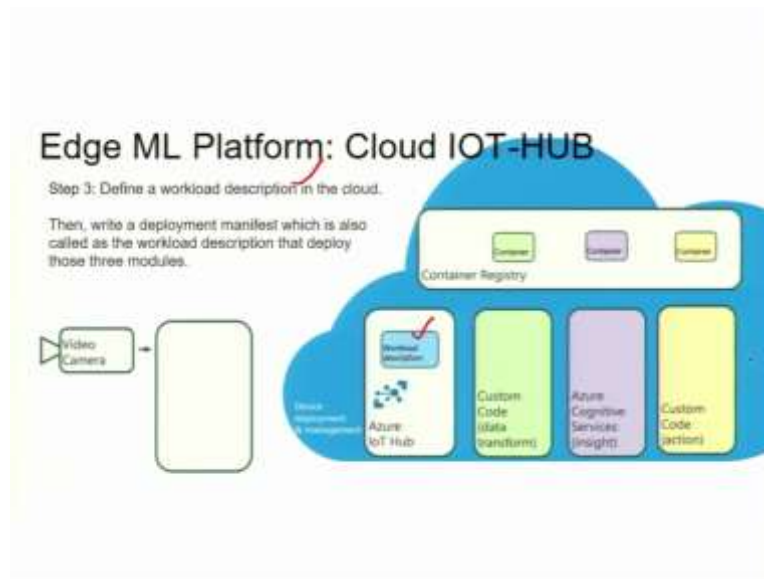
So, incoming data from these particular devices will be captured it is also called data ingestion. If the volume is more, if the volume is less than using telemetry MQTT protocol is used, if the volume is more than then this data will be ingested using some of the open source like Kafka. So, so, the first module will capture the incoming data and will send to the custom modules. So, these are the custom build module to take the decisions or to take the actions for that it has to use the insight for that it uses various other services.

Now, there are many other services which are used to manage and control to get the score out of the display of this particular model and we are going to see this in more details. So, the entire system of doing this particular data ingestion insight, and the action which is

happening in the cloud is being offered the software as a service to for this doing this edge ML platform. Let us see in more detail how this particular services which are running in the cloud using the software as a service platform, it can support this edge ML.

(Refer Slide Time: 38:48)

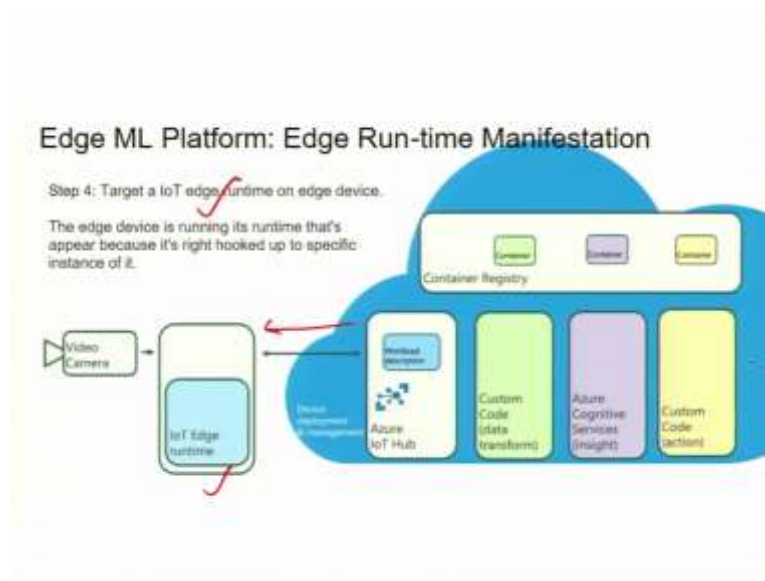




So, for this the first step is the containers. So, these particular modules which are running inside the cloud, now offer containerization in the form of containers, so, the data transformation is packaged or is containerized. Similarly, the insight is also containerized and the action is also containerized. So, therefore, this is that first step is that it will package the data transfer insight and actions into the containers. This is called containerization.

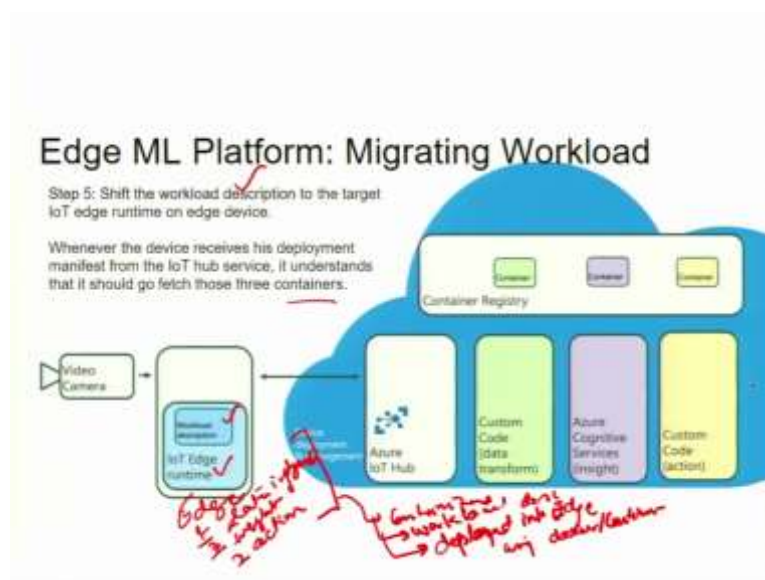
Now, then, you have to write these modules and package them into the Docker containers. So, let us see, this is the next step. So, if you package it into the Docker container, so, it will require to be registered as a container registry. So, all these containers are now registered. So, what it says is to push all those Docker containers into the container registry and having done this then this will define the workload description into the cloud, then write a deployment manifest, which is also called the workload description that deploy these three modules. So these three models are now packed into the workload description into the cloud into the Azure IoT Hub.

(Refer Slide Time: 40:22)



Now, this workload description, which is containerized, using Docker container, and is available with the Azure IoT Hub, in the software as a service SaaS model will now make this available to the edge layer. In the form of IoT Edge runtime, will use this workload description. So, the target IoT Edge runtime on the edge device, So, the edge device is running its runtime that is appeared because its right hook to the specific instance of it.

(Refer Slide Time: 41:00)

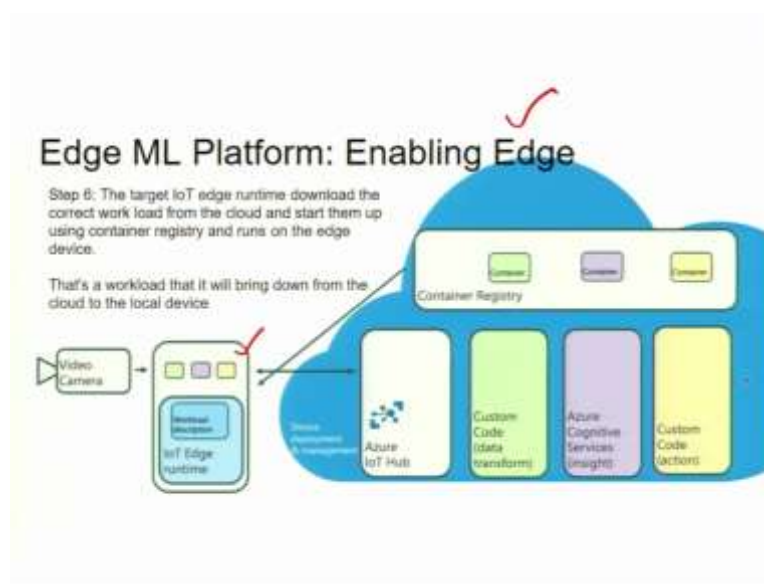


So, you see that the IoT Edge runtime will receive this workload description in the form of Docker and container will facilitate this migration or movement or deployment into the edge. So, therefore, this particular module. One is data ingestion. Second is insight. Third one is the actions they are all containerized. And is available as the workload description. And then

using the Docker container, it will be now migrated or is deployed into edge using Docker container.

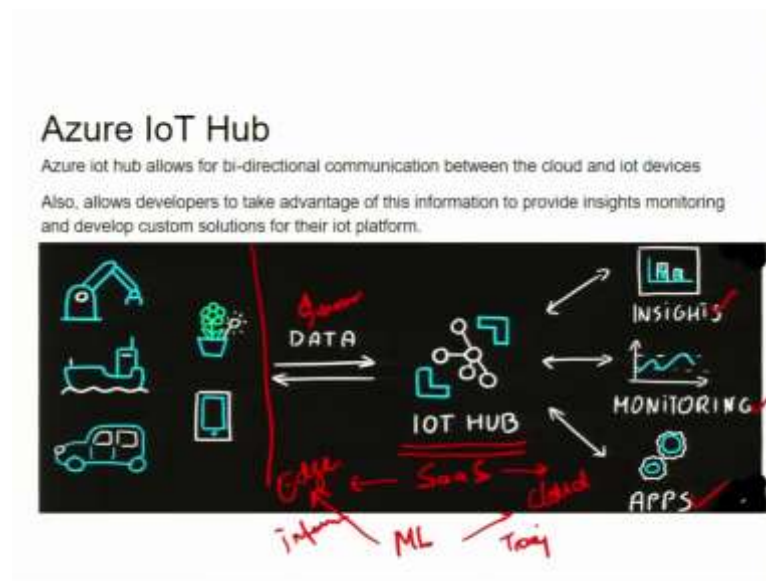
So, therefore, this workload description will bring into all these capabilities, which this cloud is now happening. So, therefore, shift the workload description to the target IoT runtime on the edge device. So, this is the edge device. So, whenever the device receives this deployment manifest from the IoT Hub service, it understood that it should go and fetch these three containers. So, these three containers described in the workload description from the container registry will be will be deployed.

(Refer Slide Time: 42:42)



So, that you can see here, enabling the edge with these capabilities. So, the target edge runtime, download the correct workloads from the cloud, and then start using the Container Registry and runs it on. So, all these containers will now start running these three different capabilities that is data ingestion, insight and action which earlier cloud used to do to the local device.

(Refer Slide Time: 43:13)



Now, let us see the entire picture that is being defined here in Azure IoT Hub. So, Azure IoT Hub allows bi-directional communication between the cloud and IoT devices, it also allows the developer to take advantage of the information to provide insight, monitoring, and develop custom solutions to the IoT platform. So, here it is quite explained that these IoT devices, which are equipped with sensors and actuators, will send the data that is called sensor data to the Azure IoT Hub. And Azure IoT Hub will perform the insight, monitoring, and use different applications.

Now, this Azure IoT Hub will use software as a service and therefore will enable the computation to be done very close to the devices that is called Edge, and the rest of the things will be performed by the cloud. So, the machine learning model training will be performed at the cloud, whereas the inferencing will be performed at the edge, and this particular capability will allow the data to be transmitted back and forth between the devices and this particular software as a service that is whether it is edge computing or a cloud computing.

(Refer Slide Time: 44:50)

Azure IoT Hub: key characteristics

- Managed service for bi-directional communication: it is a managed service for bi-directional communication between the cloud and IoT devices.
- Platform as a service (Paas): it's a platform as a service offering in Azure for IoT development.
- Highly secure, scalable and reliable: it's a highly secure, scalable and reliable service for IoT devices.
- Integrates with lots of Azure services: perfectly integrates with a lot of Azure services.
- Programmable SDK for popular languages: you do not need to learn any new language to take advantage of IoT Hub for their development purposes.
- Multiple protocols: it supports for multiple common standards on the market when it comes to communication protocols.

MQTT, HTTP, etc

So, therefore, let us understand the key characteristics of such software as a service IoT Edge platform which is being provided Under the software as a service which is called Azure IoT Hub, so, Azure IoT Hub is a software as a service platform and which enables the edge layer and a cloud model to facilitate, so, therefore, this particular software as a service that is called Azure IoT Hub or you can say edge IoT or IoT Edge is used to manage the services for bidirectional communication.

So, it is managed service for by directional communication between the cloud and the IoT devices. So, it is a Platform as a Service. So, it is a platform this we are talking about this as the edge IoT Edge platform. So, it is a platform as a service offering in the Azure for IoT development, it is highly secure, scalable and reliable. And it integrates with a lot of Azure services programmable SDKs for popular languages are available and it also supports multiple protocols such as MQTT, HTTP, HTML, sorry, HTTP, etc.

(Refer Slide Time: 46:33)

IoT-Edge: key characteristics

The **Camera Capture Module** handles scanning items using a camera. It then calls the Image Classification module to identify the item, a call is then made to the "Text to Speech" module to convert item label to speech, and the name of the item scanned is played on the attached speaker.

The **Image Classification Module** runs a Tensorflow machine learning model that has been trained with images of fruit. It handles classifying the scanned items.

The **Text to Speech Module** converts the name of the item scanned from text to speech using Azure Speech Services.

A USB Camera is used to capture images of items to be bought.

A Speaker for text to speech playback.

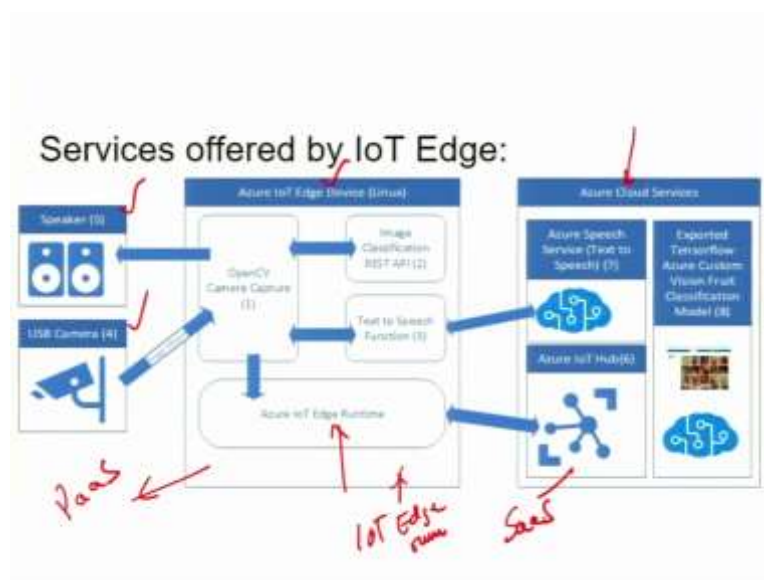
Azure IoT Hub (Free tier) is used for managing, deploying, and reporting Azure IoT Edge devices running the solution.

Azure Speech Services (free tier) is used to generate very natural speech telling the shopper what they have just scanned.

Azure Custom Vision service was used to build the fruit model used for image classification.

So, let us see more characteristics of such IoT Edge platform, which runs as a platform as a service. So, there will be a camera module, which handles the scanning if the camera is a device as an IoT device, then it will also use another module which is called Text to Speech, where the microphone is the device as the IoT device. Similarly, the IoT Azure IoT Hub used to manage and deploy using these Azure IoT Hub edge devices.

(Refer Slide Time: 47:15)

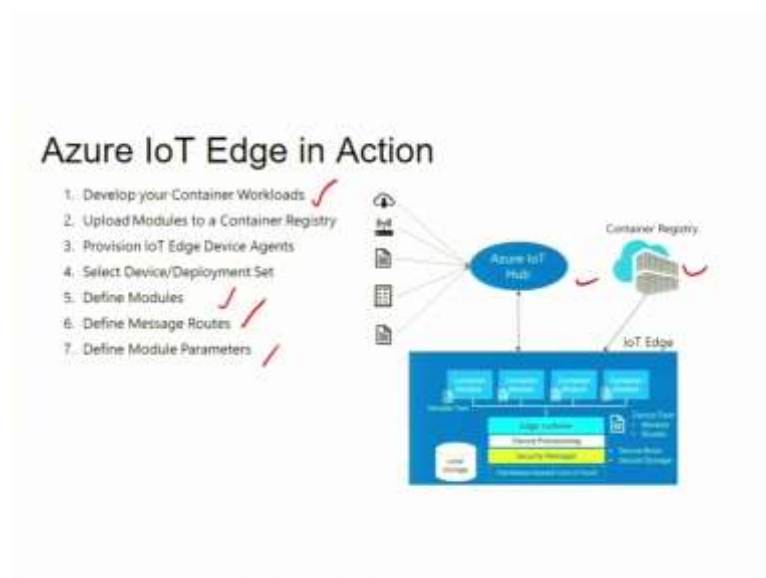


So, there are various services which are being offered by this particular type of platform as a service that is called IoT Edge platform, you can have the microphone, you can have the camera as the devices. And therefore this the edge devices will have the support of data

ingestion, and performing the insight and the action for these kinds of devices with the help of Azure IoT Edge runtime.

Similarly, this Azure IoT runtime is being supported by the cloud that is Azure cloud services that is software as a service for this particular edge device that is Azure IoT Edge runtime will use this Azure cloud as a Software as a Service. So, this entire thing, this edge will be offered as a platform as a service whereas this Azure cloud services for this IoT Edge runtime will be offered as a Software as a Service.

(Refer Slide Time: 48:23)



So, Azure IoT Edge in action, you can think of that you can use this kind of software as a service to develop your container workloads, upload your modules in the container registry, do the provisioning of these device agents, select the device deployments, define the modules and then define the message routes and define the model parameters that is what is shown over here. So, all these things we have already explained, and these features are available to the developer.

(Refer Slide Time: 48:59)

Azure IoT edge: Functionalities

- **Target workload at the correct type of device**
 - Once the workload description sent down to the edge, the run time will download the correct work load from the cloud and start them up and running.
- **Create workload which can include high value ML**
 - This results in the custom code, machine learning modal, and business logic all running locally independent of cloud connection and also all of those values of edge analytics.
- **Run those workload locally, in disconnected manner**
 - The runtime is smart enough to detect if the workload is trying to send messages to the cloud while it doesn't have internet connection, the runtime will catch those messages and sync them with the cloud once the internet is up.
- **Monitor the health of the workloads**
 - Azure IoT edge ensures that the work loads continue to run and report status sent back to the cloud. Reporting the status back to the cloud allows to understand if there is any issues issues in the deployment and take preventive actions.

So, Azure IoT Edge, other functionalities is that it will target the workload to the correct type of device. It will create the workload which can include high value machine learning and run those workloads locally even in the disconnected manner and monitor the health of the workloads.

(Refer Slide Time: 49:17)

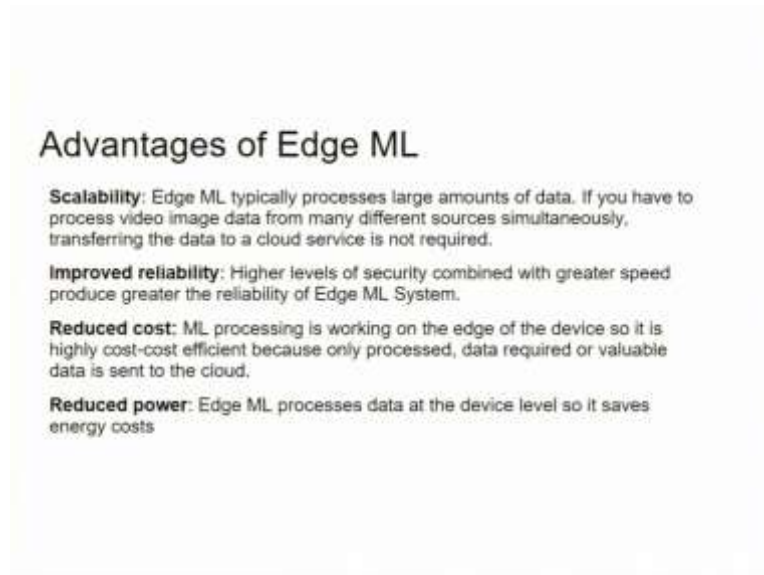
Advantages of Edge ML

- **Reduced latency:** Transfer of data back & forth from the cloud takes time. Edge ML reduces latency by processing data locally (at the device level).
- **Real-time analytics:** Real-time analytics is a major advantage of Edge Computing. Edge ML brings high-performance computing capabilities to the edge, where sensors and IoT devices are located.
- **Higher speeds:** Data is processed locally which significantly improves processing speed as compared to cloud computing
- **Reduced bandwidth requirement:** Edge ML processes the data locally on the device itself, reducing the cost of internet bandwidth and cloud storage.
- **Improved data security:** Edge ML systems perform the majority of data processing locally i.e. on the edge device itself. This greatly reduces the amount of data that is sent to the cloud and other external locations.

So, that advantages of edge machine learning is that the first advantage is it facilitates the reduced latency that is the transfer of data back and forth from the cloud takes a lot of time and therefore, this introduced introduction of edge layer, which uses the machine learning computation locally at the device level reduces the latency. Second important advantage is the real time analytics can happen using the advantage of edge computing.

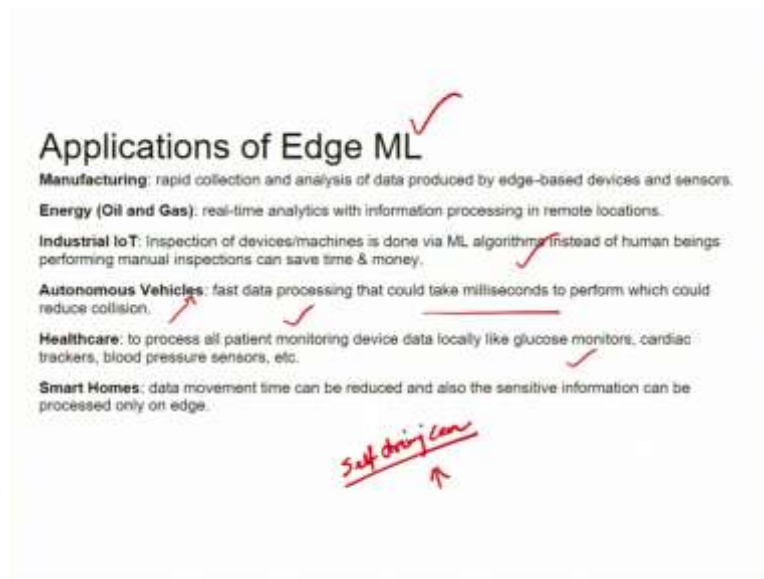
Third is the high speed so the data is processed locally, and need not have to send to the cloud. So, therefore, a higher speed of inferencing or insight is possible here. and it will also reduce the bandwidth requirement because not all data need to be sent to the cloud, it will also improve the data security.

(Refer Slide Time: 50:10)



Other advantages are scalability, improved reliability, reduced cost and reduced power most important out of this is that first that it is scalable. So, this particular platform or the services will be going to support large number of devices. Therefore, Edge ML typically processes large amount of data and if you have the big data source like video image is coming from various sources simultaneously transferring to the cloud is now not required. So, it will be also having the improved reliability higher level of security with a greater speed will bring into the reliability reduced cost is also there and reduce power is there.

(Refer Slide Time: 50:55)

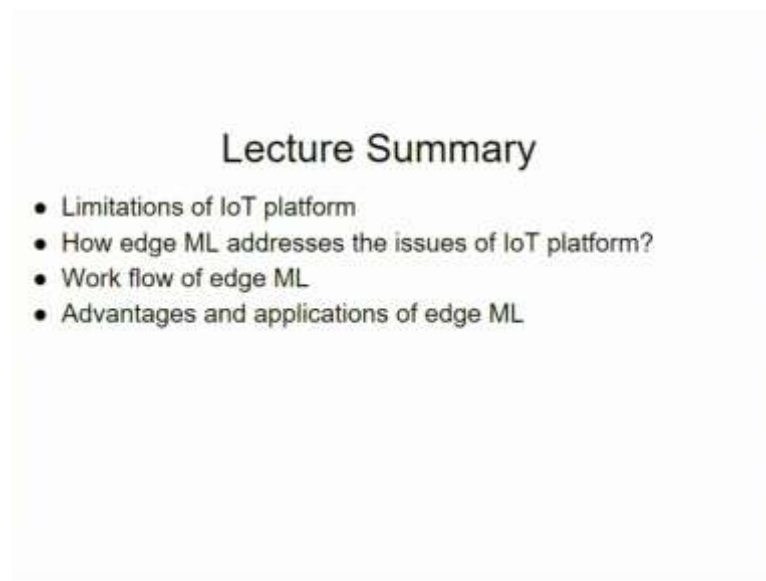


There are various applications that are awaiting for this kind of IoT Edge and machine learning platform in the manufacturing, in the Energy sector, in the Industrial IoT Autonomous vehicle, Health care and Smart homes. So, autonomous vehicle is the perfect example. What do you mean by autonomous vehicle is the driverless car self-driving cars, self-driving car you know that driver is not there.

So, the using camera these particular devices used to sense the environment and it has to take the decisions whether to move forward or to stop or it works, what is speed and this computation has to be performed within the milliseconds which cannot earlier possible with the help of cloud. So, with the help of this edge ML this autonomous vehicle is a new kind of applications is possible.

Similarly, the healthcare the patient monitoring device has to now deliver the prescription or the decision for the patients whether it requires immediate attention of medication similarly, smart homes and all these are the new kinds of similarly, the industrial application. So, monitoring of the devices or the machine and the inspection with the human is not possible in some other situation to know then when the machine is going to stop functioning unless it is put on the repairs.

(Refer Slide Time: 52:40)



Lecture Summary

- Limitations of IoT platform
- How edge ML addresses the issues of IoT platform?
- Work flow of edge ML
- Advantages and applications of edge ML

So, let us summaries this entire lecture that we have brought into the limitation of an IoT platform, IoT platform, classical IoT platform comprises of IoT and the cloud. So, IoT data is directly sent to the cloud that has a lot of limitations. Therefore, the edge machine learning is possible. Now, we have seen with the advent of lot of innovation in the chip manufacturers this particular Edge ML is a possibility and therefore, this limitation of IoT platform is possible is overcome from this introduction of the Edge ML.

We have seen how the edge machine learning, how the machine learning at the edge is possible. The entire workflow we have discussed here, we have also seen the advantages and the applications of Edge ML. So, with this, we conclude this lecture. Thank you.