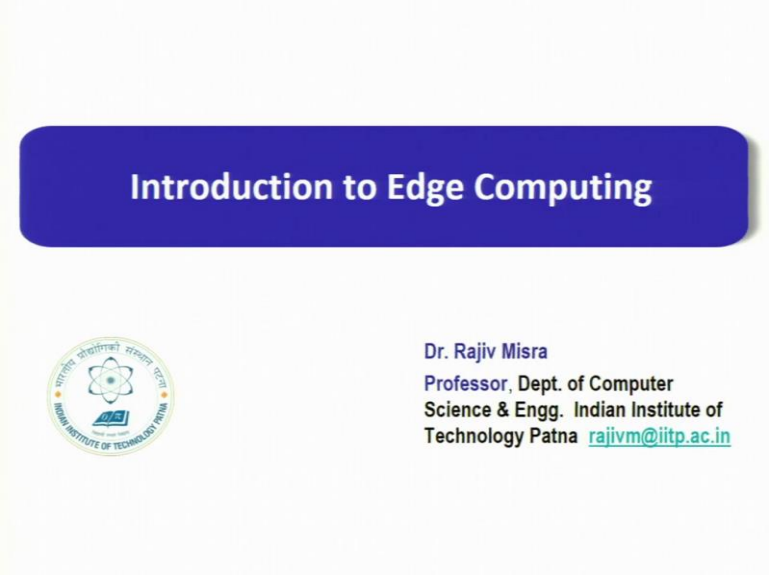
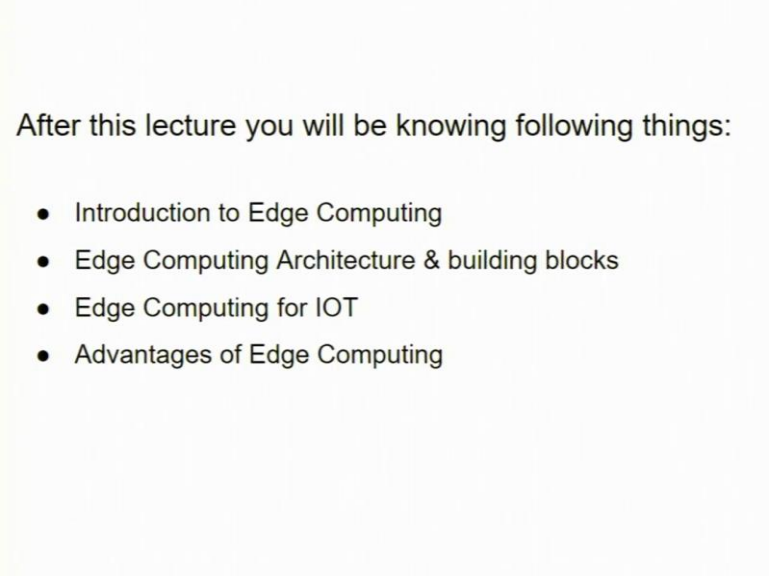
Foundation of Cloud IoT Edge ML
Professor Doctor Rajiv Mishra
Department of Computer Science and Engineering
Indian Institute of Technology, Patna
Lecture 01
Introduction to Edge Computing

(Refer Slide Time: 00:13)



Welcome to this lecture on edge computing. The title of the lecture is introduction to edge computing. Myself Doctor Rajiv Mishra from IIT Patna.

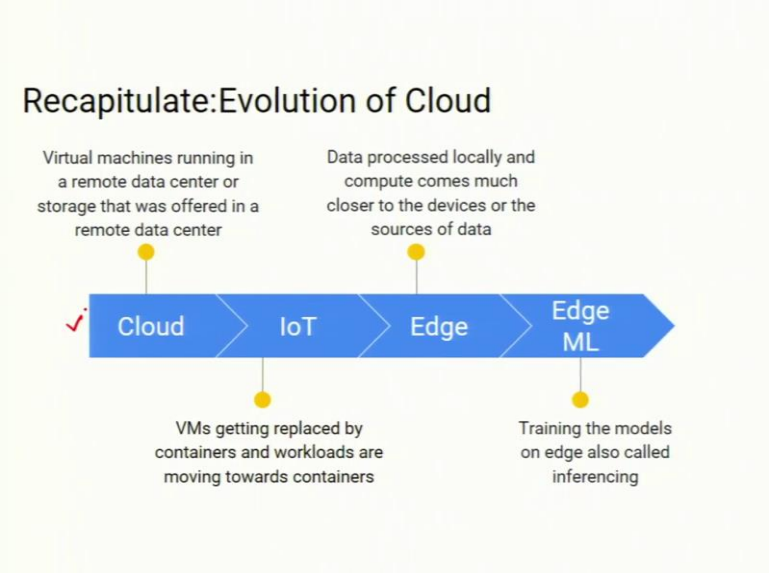
(Refer Slide Time: 00:27)



In this lecture, we are going to cover the following topics. The first one is the introduction to edge computing. The second one is about edge computing building block and architecture,

edge computing for IoT, and advantages of edge computing. So, around the edge computing, we are going to introduce you in this particular lecture, which is an essential building block in this course.

(Refer Slide Time: 00:59)

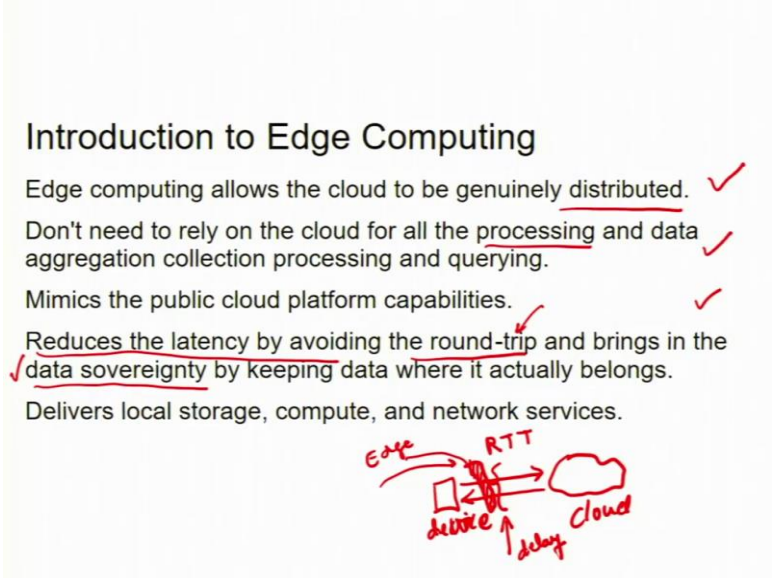


Recapitulate, what we have covered so far is the evolution of cloud. So, that means, what we have covered so far is to be briefed in here in this particular slide, that four different concepts are related with each other and which are linked with the cloud. So, we have introduced the cloud first and we have also inform you that the classical cloud provides the concept of a virtual machine, which is nothing but infrastructure as a service.

These virtual machines running on the remote data center using the technology which is called virtualization or the storage that was offered by the remote data center. Then, we have also seen that this virtual machine will get replaced by containers and more and more workloads are now moving towards the container just to allow IoT paths. Third important concept we have seen around the development of cloud towards the distributed cloud with the help of technology which is called edge computing.

And therefore, using this edge data is processed locally and the computation comes more closer to the devices or the source of data generation that is an edge and the topic of today's this particular lecture. Edge also has an advantage of applying artificial intelligence and machine learning which was the de facto standard for cloud computing for AI and ML. So, cloud is now used to train the model whereas the inferencing is to be carried on edge also called the inferencing layer. So, let us go ahead and understand about the edge computing.

(Refer Slide Time: 03:07)



So, edge computing allows the cloud to be genuinely distributed. In contrast to what we have seen that the cloud was highly centralized and all the resources are classified as or being defined as client-server architecture, where the cloud becomes the server in that architecture and the cloud is highly centralized. The edge is the technology edge computing which allows this cloud to be distributed and we will see how that particular edge will provide this kind of notion into the cloud computing.

Therefore, this edge computing do not rely on the cloud for all purposes and the data aggregation collection processing coding all can be done at the edge, therefore, leaving out the cloud. Hence, this particular notion is now taken away and with the edge computing the cloud can become the distributed where most of these functions which earlier cloud used to do now can be done at the edge that is the processing, data aggregation, collection, coding, and so on.

In this particular lecture, we are going to see how the edge computing will mimic the cloud computing functionalities. So, therefore, the other important aspect of edge computing is that it will mimic the public cloud platform capabilities. So, we are going to see more details about this and understand how these activities which were earlier done by the cloud now also can be as well done by the edge.

Now, there will be a choice for the application which one to choose with more advantage at the edge. Fourth important point about the edge computing introduction is that it reduces the

latency, how, by avoiding this round trip, when you say a round trip, you mean you have to understand that the device is in the cloud.

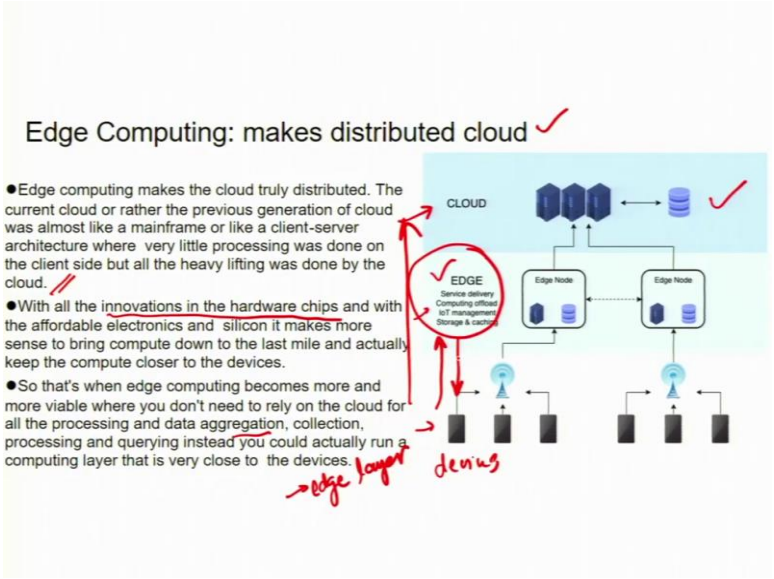
So, the device has to send the data to the cloud for doing the computation and after the computation is finished the results have to come back to the device, and this kind of delay is called a round trip time. And this particular aspect of RTT that is round trip time will introduce a sufficient amount of delay and this delay can be reduced if let us say here another layer which is called an edge is introduced.

So, edge computing will relay this particular round trip time to a very little minimal, why, because this edge is very close to the device. Therefore, it will reduce the latency and it will also avoid the round trip to the cloud, and brings into the data sovereignty by keeping the data where it belongs. Data sovereignty, you know that it is a important concept in the security. For example, many government organization or let us say hospital or a personal data.

Sometimes they do not people do not want that it is to be sent to the cloud with the fear that the personal identification or many personal data will be disclosed. Therefore, to protect this data sovereignty, you have to keep the data where it belongs that is at the source itself and edge computing will get this kind of data sovereignty applicable by not sending to the cloud yet. All the computation can be performed very close to the source that is at the location where the data is generated.

For example, we are going to discuss giving you the example in the further slides. Finally, the edge computing introduction also talks about how the edge computing will deliver the local storage, compute and network services which are earlier in a highly centralized manner the cloud used to give, so the edge will be also doing or mimics the public cloud capabilities. And that is what is about the edge computing we are going to see how these capabilities are going to be given to the edge computing.

(Refer Slide Time: 08:23)



So, let us go in more detail for each of these introductory points for the edge computing. So, first important thing is that edge computing makes the cloud truly distributed. How? So, you can see that the current cloud or other previous generation of cloud was almost like a mainframe as we have already seen. Or you can also think of like client and server architecture where very little processing was done at the client side.

But all the heavy lifting of computation and the storage was done by the cloud because cloud can do these kind of heavy jobs. But how the edge computing will make this kind of architecture which only cloud can do as the server how these functionality can be shifted away from the cloud, and yet all the computation can be performed. So, we are going to see through this particular diagram you can see there is a cloud and below the cloud, you can find an edge.

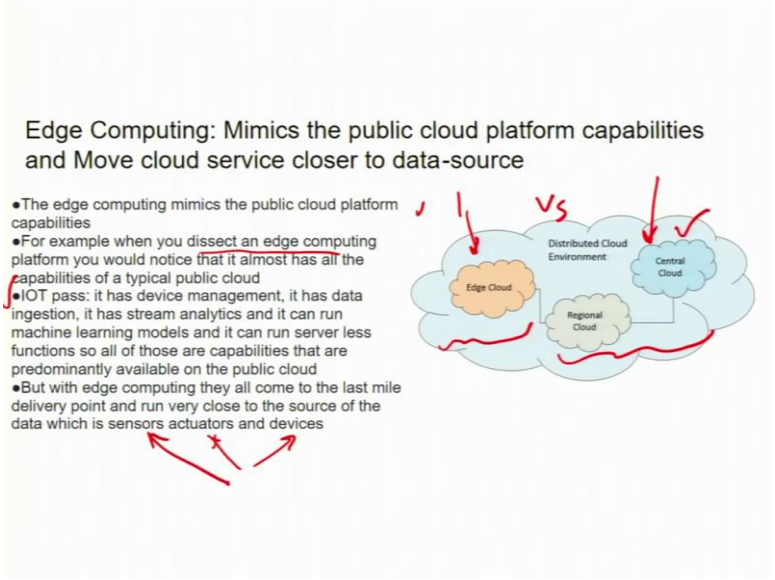
So, edge nodes also has the capabilities of storage and computation, and if it is done at the edge, then no need of sending the entire computation at the cloud. Therefore, cloud is no longer be the client-server architecture. Therefore, with the introduction of the edge and cloud, it becomes a truly distributed. Second point is that with all the innovations how that has all happened is because of the innovations in the hardware chips with the affordable electronics and silicon makes it more sense to bring the compute down to the last mile and actually keep the compute closer to the devices.

So, that means these capabilities are now possible to bring into the edge because of the innovations in the hardware and also affordable to provide large number of edge nodes. And

making this compute very close to the devices and making this particular cloud a truly distributed cloud. So, that is when the edge computing becomes more and more viable, where you do not need to rely on the cloud only for all the processing data, aggregation, collection, processing, querying instead you can also run a computing layer, which is very close to the device, which is called the edge layer.

So, you can see that if this is the device, and this is the cloud, we are earlier devices used to send all the data to the cloud. Now, they do not have to send directly to the cloud, they can send to the edge and edge can respond with all that computations back to the devices and this is quite close to the source device where the data is generated or very close to the devices. So, edge computing makes the cloud truly distributed.

(Refer Slide Time: 11:34)



Now, another aspect of edge computing, which we have mentioned previous slide is that the edge computing mimics the public cloud platform capabilities and therefore, it allows to move the cloud services closer to the data source. So, let us see how those capabilities are also brought into the edge layer, therefore becoming an edge computing. So, the edge computing mimics the cloud public cloud platform capabilities.

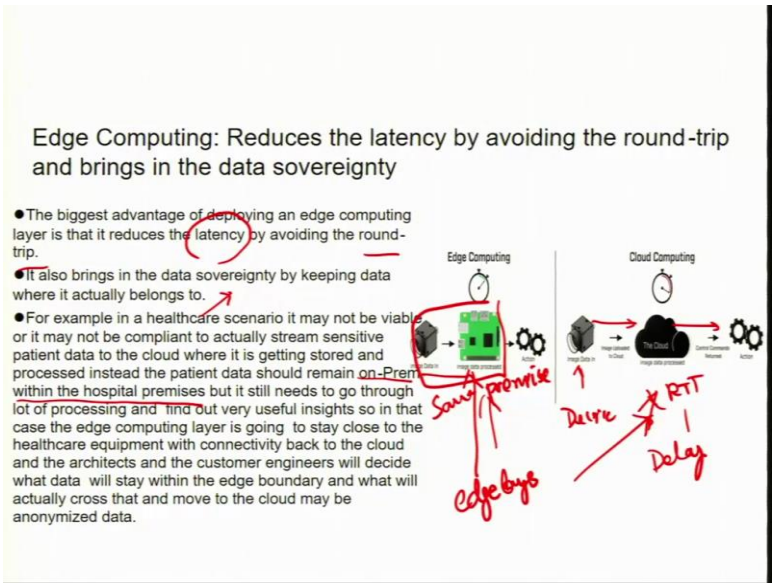
So, you can see here, this is the cloud, whether it is the central cloud or the regional cloud. And this is the capabilities which is being created very close to the source. And that is called the edge. So, if you go in details about the capabilities, which are to be required to become an edge computing, so that means when you dissect an edge computing platform you would notice that it almost has all the capabilities of a typical public cloud.

So, let us have the contrast that public cloud versus edge computing. So, what are the capabilities which public cloud has now being also capable into the edge layer and that is what edge computing we are talking about. Now, why this is needed is to support the IoT paths. So, IoT has, you know that a device management requirements earlier cloud used to do it, it has the data ingestion in the cloud earlier, it has the stream analytics in the cloud earlier to support IoT paths, and it can run the machine learning models into the cloud very well.

And it can also run the serverless functions into the cloud. So, all those are the capabilities that are predominantly available in the public cloud. But the edge computing they all come to the last mile delivery point and run very close to the source if all these capabilities are brought into the edge, when you talk about the source which is generating the data that is nothing but the devices.

So, as I told you, the devices means the sensors which generate the data and actuator which consume the data both of them are called the devices and if this computation is very close to the device, that means the capabilities of cloud if it is provided to the edge, then all the advantage and new applications is possible that is what is we have understood here by means of the edge computing.

(Refer Slide Time: 14:09)



Now, another aspect of giving up the cloud truly distributed by allowing the edge computing is to reduce the latency by avoiding the round trip and brings in the data sovereignty. So, the biggest advantage of deploying the edge computing layer is that it reduces the latency by avoiding the round trip. So, this is not a small thing.

This is a very big specification, which allows many applications which were not ready or willing to be performed into the cloud. Now with the help of this edge computing now it is possible because of lighting the round trip and reducing the latency which is now supported by various applications.

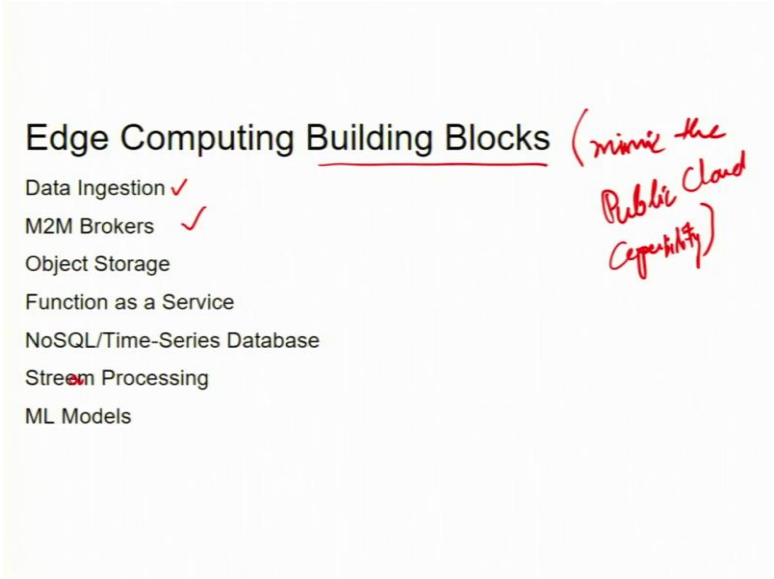
Second important point about the edge computing is that it brings you data sovereignty by keeping the data where it belongs to for example, in a healthcare situation, it will not be viable or it will not be compliant to actually stream the sensitive patient data to the cloud where it is getting stored and processed initially ropey patient data should be should remain on prem that is within the hospital reminds yet it needs to go lower processing there itself with the help of edge computing to find out very useful insights through the edge layer.

So, Edge layer is going to stay very close to the healthcare equipments with connectivity back to the cloud and the architects and the customer will engineers will decide what data will stay at the cloud at the edge boundary and what will be now actually crossing the edge and moving to the cloud that may be anonymized. So, in this particular diagram, you can see these differences that if the data which is generated by let us say that device it is sent to the to the cloud, and then the cloud will send back the response.

So, this is this particular delay is called round trip time and it will introduce the delay which may sometimes not be acceptable by the applications for example, in a medical application such a delay may sometimes take away the life of the patient. Therefore, the edge computing in contrast, if you see if the computation is possible very close to the device, maybe that in the same premise that is in the hospital.

So, the data from the devices are now being processed by the edge layer and data need not have to travel to the cloud. Therefore, the round-trip time can be avoided here in this particular case.

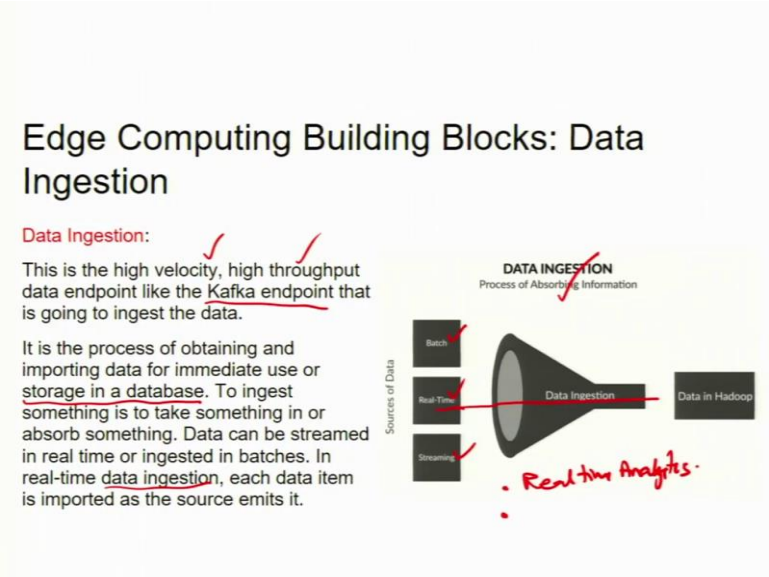
(Refer Slide Time: 17:10)



So, with this, we are going to move to the next discussion that is about the edge computing building blocks we are seeing what are the building blocks are the capabilities which are now added also in the edge, which was earlier only done at the cloud. So, the first capability is called data ingestion, machine-to-machine brokers object storage function as a service, no sequel, or the time series database is stream processing and machine learning models.

These are some of the basic building blocks, which can be performed at the edge and which was earlier the capabilities of a cloud. So, this will mimic the public cloud capabilities. Let us see one by one that how these building blocks how they are to be supported by the edge layer.

(Refer Slide Time: 18:17)



So, the first one is called data ingestion. So, data ingestion is the endpoint for getting the source data, which comes at a high velocity, high throughput data endpoint, such as Kafka. So, such kind of data ingestion, tools, and techniques or capabilities was with the cloud, but now those kinds of capabilities are also now allowed at the edge. So, this particular data ingestion will going to ingest the data which becomes the high throughput data endpoint at the edge itself.

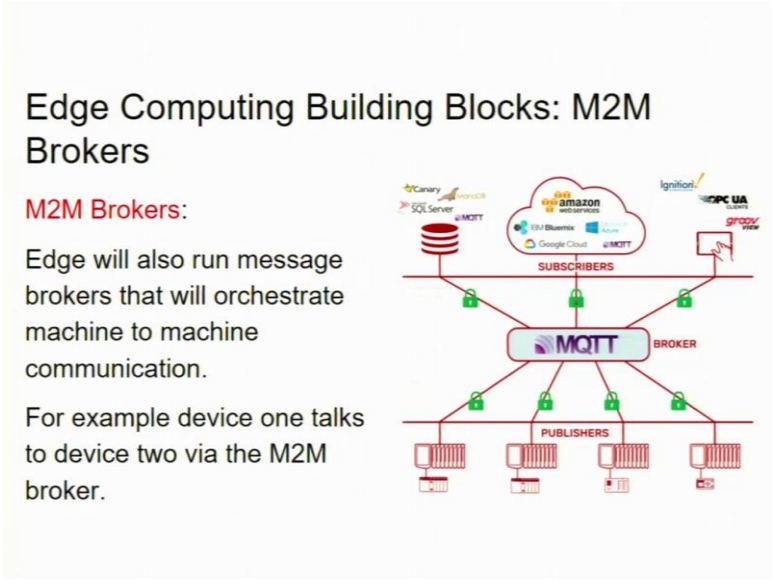
So, it is the process of obtaining and importing the data for immediate use or the storage in a database to ingest something is to take something in or absorb something and data can be streamed in real-time or ingested in batches in the real-time data ingestion, each data item is important as source emits. So, here you can see that these are the sources the first source is called the batch.

Batch means the data which is ingested will be stored somewhere into the database. The second source of the data is to be in real-time. So, data which comes as a stream it will be captured by the data ingestion has to be computed or to be processed without storage. And therefore, this streaming data becomes an essential building block for real-time analytics.

For example, the sensor data which is monitoring a particular temperature of a machine, whether it gets very hot or not. So, all the time that data is being generated and the real-time analytics has to be done it cannot be stored like a batch processing can be stored data and then it can happen later on. So, in the real-time data ingestion, each data item is imported as a source as the source emits it.

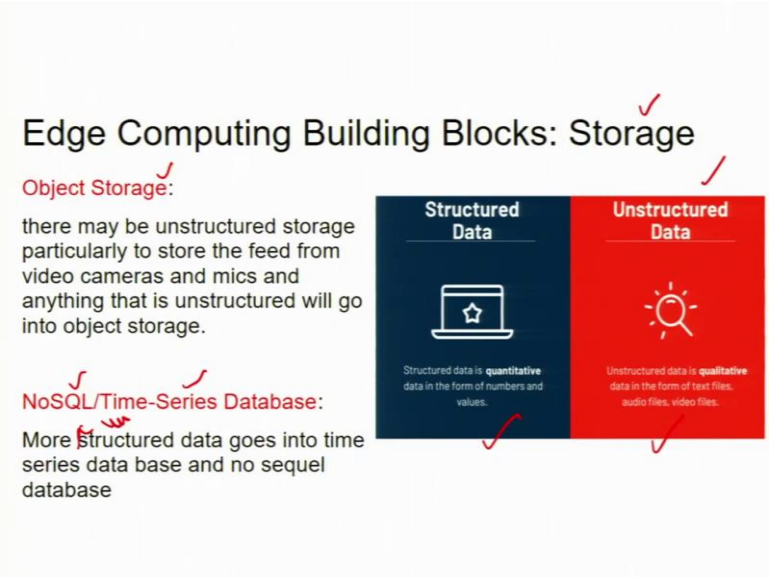
So, there are many examples, which these kinds of data ingestion, which is nothing but high-velocity data endpoints, which was done by the tool like Kafka which was only been performed by the cloud now has become the building blocks for in the edge computing. So, edge can support this kind of data ingestion technologies.

(Refer Slide Time: 20:59)



Now, another edge computing building block is in terms of machine-to-machine brokers. So, edge will also randomize Message Broker that will orchestrate the machine-to-machine communications. So, for example, the device one talks to the device to wire machine to machine broker. And now, earlier these kind of machine-to-machine brokers are supported by the cloud now, it can be other basic building block of edge computing as well.

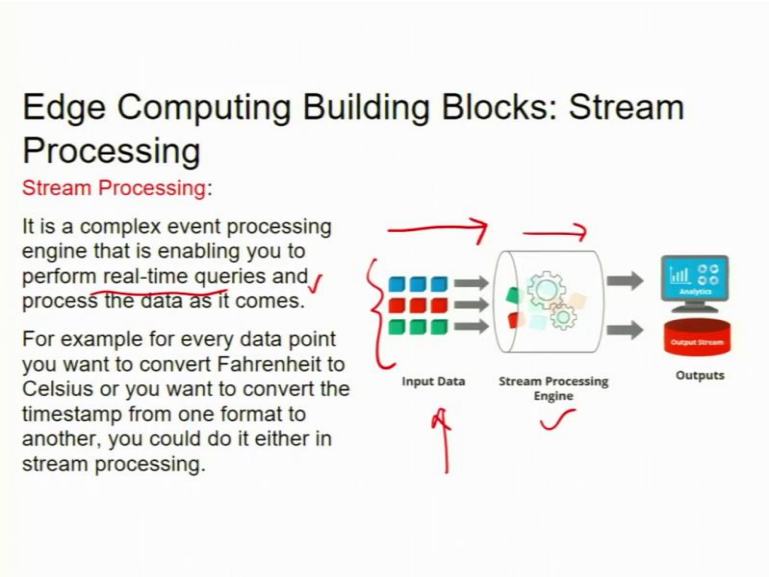
(Refer Slide Time: 21:29)



Now, another basic building block of edge computing is the storage. So, the story is sometimes the cloud is storage we also call it as object storage. So, object storage, you can see there are many types of unstructured storage particularly to store the feet from video cameras mics, and anything that is unstructured will go into the object storage. Now, this kind of object storage also can happen at the edge which was earlier only being performed by the cloud, the technologies for doing this is no sequel and the time series database.

So, more or less structured data goes into the time series database and no SQL database. So, it can also support unstructured data. So, here you can see the example that unstructured data is a qualitative data in the form of text file, audio video, and which often comes from the devices structure data is quantitative data in the form of numbers and the values. So, what type of object storage is now also possible at the edge as a building block, which was earlier being done only at the cloud.

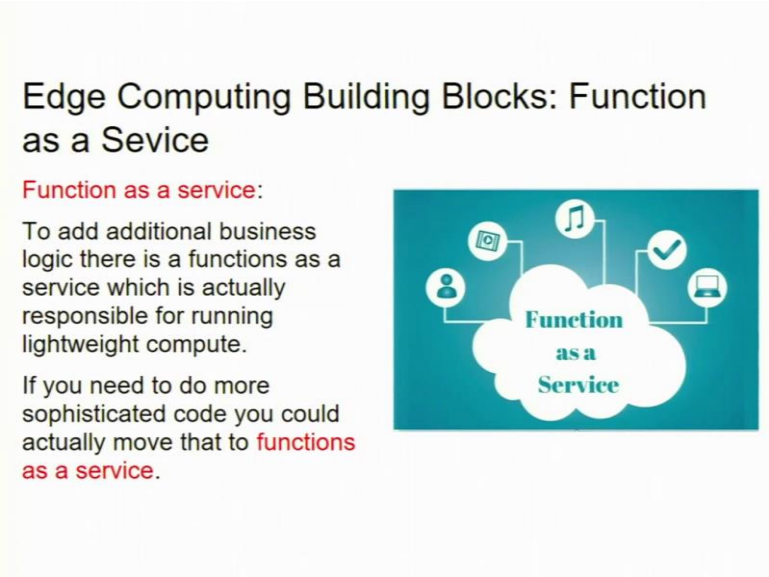
(Refer Slide Time: 22:45)



Now, finally, the edge computing building blocks as the stream processing, stream processing, you know, that is a complex event processing engine that is enabling you to perform the real-time queries and process the data as it comes. So, you can see that the data is coming in the form of a stream just like you can see a water stream. So, this particular data which requires real-time stream processing, to perform the real-time queries has to be carried out by a layer in the cloud and is a stream processing engine.

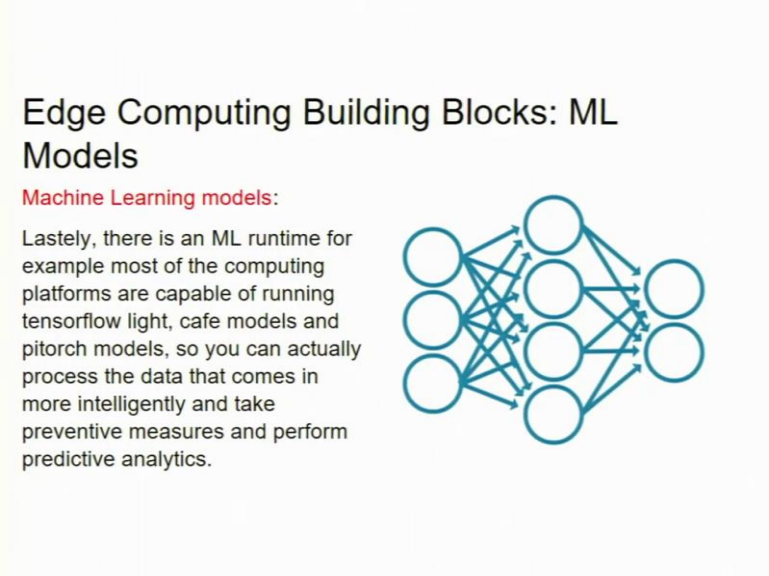
Now it can run as the capabilities have grown into the edge computing. So, therefore, these kind of techniques that is stream processing techniques are also included as a basic building blocks in the edge computing. So, for example, every data point you want to convert Fahrenheit to Celsius or you want to convert the timestamp from one format to another format, you could do it in the stream processing on the applied in the real-time.

(Refer Slide Time: 23:54)



And the edge computing building block also function as a service. So, in addition, additional to business logic that is the function as a service, which actually is responsible for running the lightweight compute, so, if you want to do more sophisticated, you could code you could actually move that to the function as a service. So, for example, that if you have a business intelligence in place, and you want to apply while the data is being computed or is being received at that time, so, therefore, function as a service can be applied as the business intelligence.

(Refer Slide Time: 24:37)

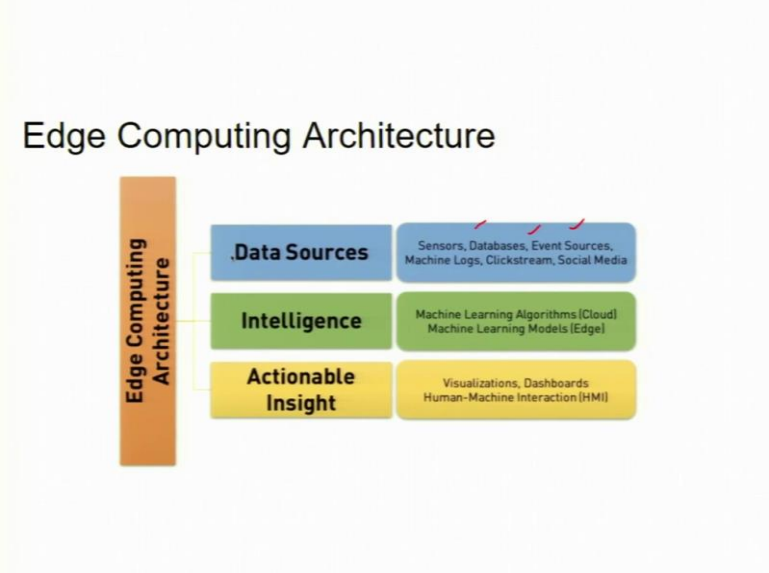


And this is also possible to be done at the edge computing unlike, which was only being carried out by the highly centralized cloud computing, which is depicted as the client-server

architecture model of computation. Now, edge computing building blocks is also supporting nowadays, the machine learning model. So, machine learning models is there is a machine learning runtime for example, most of the computing platforms are capable or run running the TensorFlow light cafe models pytorch.

So, that you can actually process the data when it comes in for more intelligently and to take the preventive measures and perform the preventive and predictive analytics that was the capabilities of the cloud earlier now, these kinds of techniques are platforms like TensorFlow, cat model, pie torch, all these models, which supports this machine learning model can also be done at the edge by supporting these capabilities.

(Refer Slide Time: 25:47)



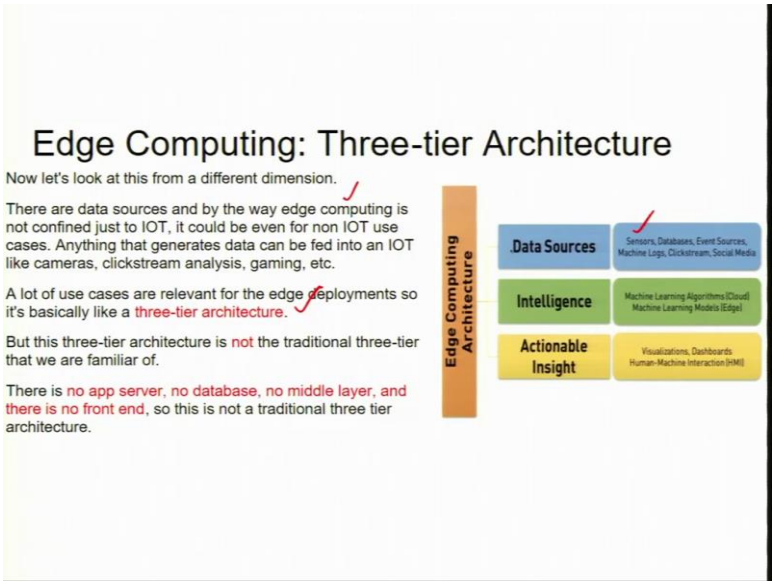
Now, coming to the next topic, which is the architecture of edge computing. Now, to understand this edge computing in more detail, that is in a systematic manner, we have now divided into three different layers of an edge-computing architecture, these layers may vary, but for our understanding of our lecture, let us understand let us assume that the edge computing architecture we can divide into three different layers.

The first layer is called the data source. So, that layer data source comprises of sensors, database events, sources, machine logs, clickstream, social media, they are all nothing but the data sources with these data sources in place into the edge computing that means edge computing supports the data ingestion from these different sources.

And once the data is ingested into the edge computing system, then the edge computing is also as we have seen supporting the algorithms like machine learning AI and building the machine learning models at the edge. So, the training if it is done at the machine learning models algorithms in the cloud, then the inferencing of that machine learning train model can be done at the edge.

And therefore, this edge can run the intelligence or artificial intelligence at very close to the source where these data sources are now generated. third layer is called actionable insights, which can be performed with the visualizations dashboards and human-machine interaction. So, it comprises of three different layers one is called data source, the second is intelligence, third is actionable insights.

(Refer Slide Time: 27:44)

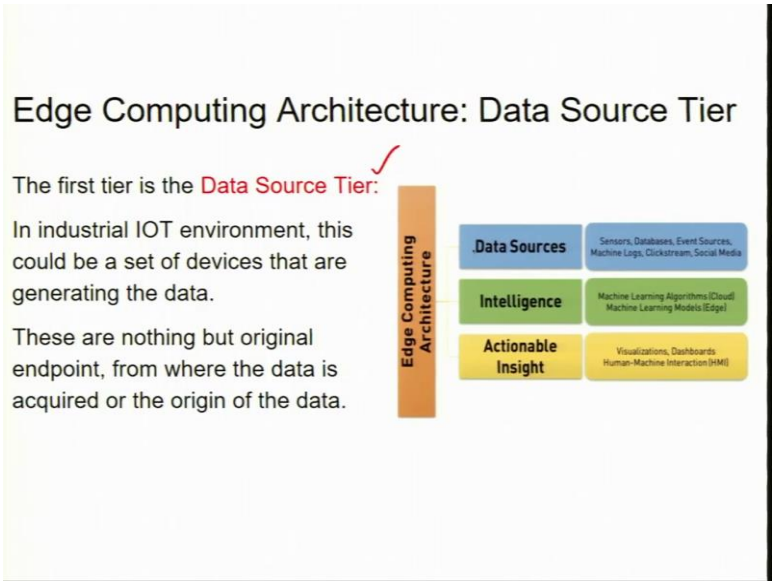


So, three-tier architecture, let us go in more detail about that. So, there are data sources. And by way of edge computing is not confined only to IoT it can be even non-IoT cases like we have given the example of sensors. So, non-IoT examples also are possible, for example, clickstream, social media logs, machine logs, and so on, generates the data and can be fed into the IoT like, such as camera or clickstream, gaming, and so on.

So, a lot of use cases are relevant for the edge deployment. So, it is basically like a three-tier architecture. But this three-tier architecture is not the traditional three-tier architecture we are familiar with. So, there is no app server, no database, no middle layer, like we have seen in the audience, the traditional three-tier architecture. So, it is not a traditional three-tier architecture, which we are talking about for the edge computing.

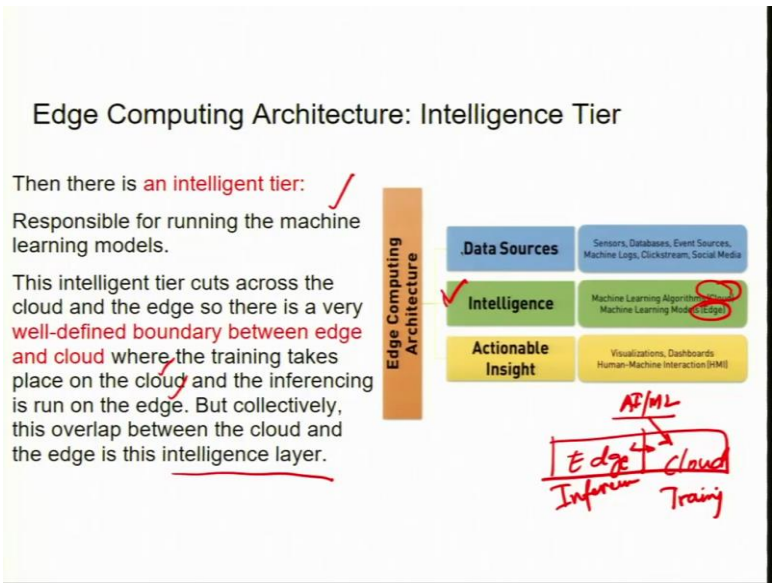
It is a new three-tier architecture for edge computing model we are discussing to make you understand the edge computing architecture.

(Refer Slide Time: 28:50)



So, first tier out of these three-tier four edge computing architecture is called Data Source tier. So, in the industrial IoT environment, this could be set of devices that are generating the data that that we know. So, they are these are nothing but the original endpoints from where the data is acquired or the origin of the data.

(Refer Slide Time: 29:14)

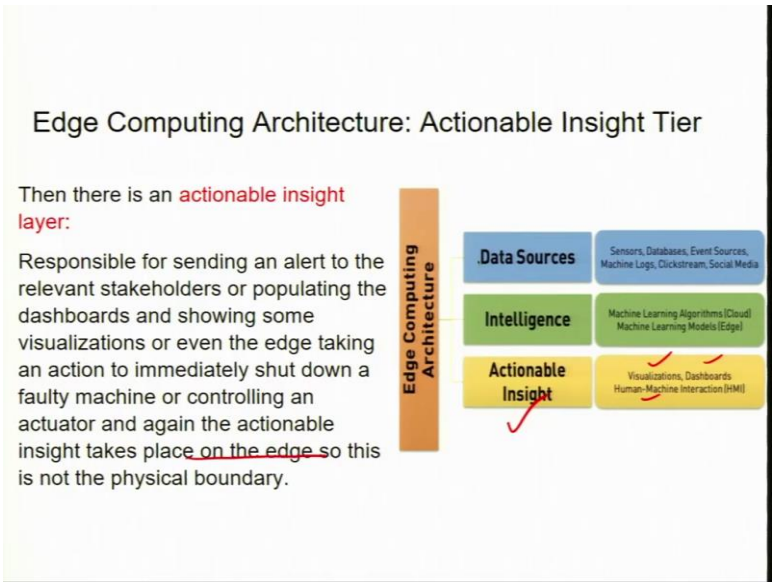


The second one is called intelligent tier, this is responsible for running the machine-learning models. So, this intelligently our cuts across the cloud and the edge. So, there is very well-

defined boundary between the cloud and the edge where the training takes place is the cloud and we are the influencing takes place is the edge, but collectively this overlap between the cloud and the edge is called the intelligence layer.

So, intelligence layer as you can see comprises of both the cloud and an edge and when the data or the AI or machine learning if you want to make it so, this division that is one part of the AI and machine learning that is called training will be performed the cloud whereas the other part of the machine learning called inferencing will be done at the edge. So, edge and cloud together join to do this intelligence computation and that is called intelligent tier.

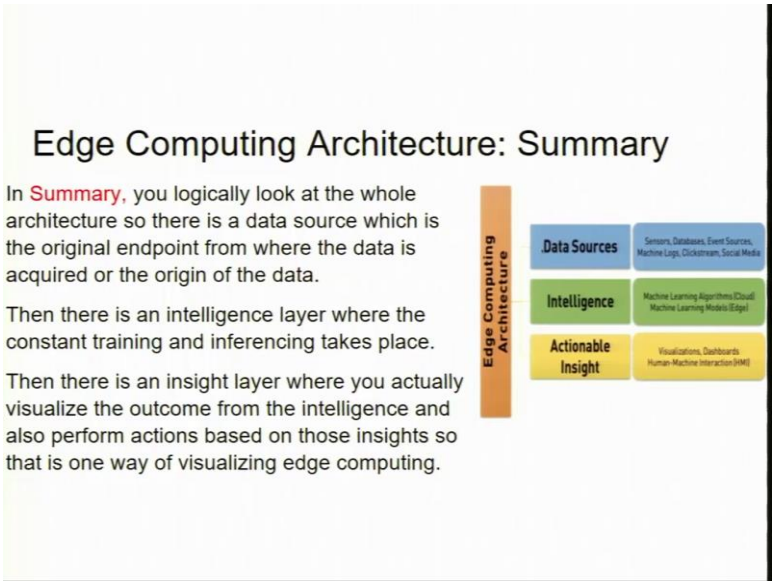
(Refer Slide Time: 30:31)



Finally, there is a tier which is called actionable insight layer responsible for sending an alert to the relevant stakeholder or populating the dashboards and showing some visualizations or even the edge taking an action or immediately shut down a faulty machine or controlling activator. And again, actionable insight takes place on the edge. So, therefore, you can see that these visualizations dashboard, and human-machine interaction, they are all supported at the edge itself.

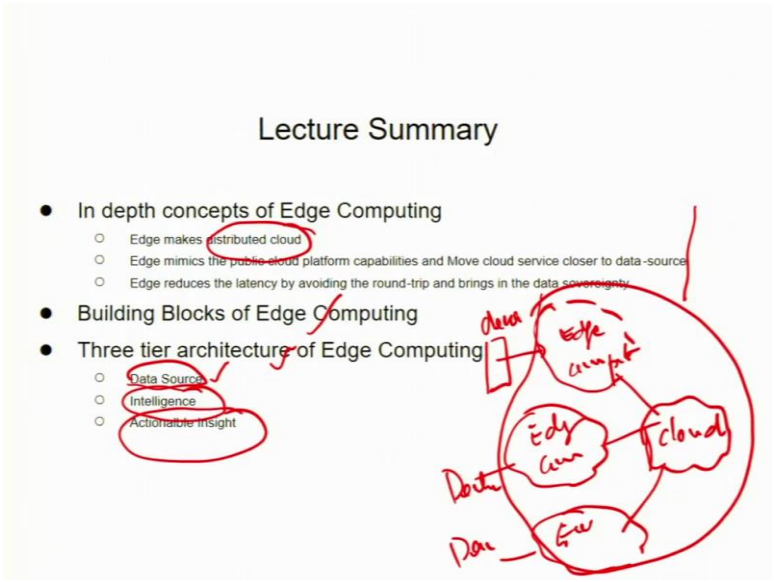
So, the data source when it is presented to the edge, so, it will generate it was capable of doing computation and generating this actionable insight. So, these three-tier architecture that is in the form of data sources, intelligence, and actionable insights, we have discussed and this forms the enough background for edge computing. In contrast with the cloud computing.

(Refer Slide Time: 31:43)



Let us summarize this architecture in summary, you will logically look at the whole architecture. So, there is a data source which is the origin original endpoint from where the data is acquired. And then there is intelligence layer, we have the constant training and inferencing takes place, they are both cloud and edge joins together in doing this intelligence computation, then there is a insight layer where you actually visualize the outcome from the intelligence, and also perform the actions based on those insights. So, that there is one way of visualizing the edge computing. So, therefore, this actionable insight can be performed at the edge itself.

(Refer Slide Time: 32:27)



So, we have presented in-depth concepts of edge computing, we have also seen that this edge computing makes the distributed cloud edge mimics the public cloud platform capabilities we have seen in this particular lecture, all these describing the details, whether it is the stream processing or it is data ingestion or it is data storage. All this is all done at the edge unlike what we have seen that only the capabilities of a cloud therefore, edge mimics the public cloud platform capabilities and move the cloud services closer to the data source.

This edge often reduces the latency by avoiding the round trip and brings in the data sovereignty. So, these different concepts you know, that provides that earlier used to be a cloud computing model, but now, you have now, different Edge locations or edge layers. There can be many edge computing. And these edge computing are very close to the data source that is the devices.

So, therefore, these capabilities, which earlier used to happen at the cloud is now given to the edge. So, this entire ecosystem that is that edge computing and the cloud computing together forms a distributed cloud. So, this particular edge computing edition will make the cloud truly distributed that is what we have shown in this particular lecture.

Then, we have also seen the basic building blocks of the edge computing and basic building blocks that means, what are the basic capabilities that edge computing has to offer to make this cloud a truly distributed that we have already understood, then we have also described the three-tier architecture of the edge computing meaning to say that the functionalities that is the basic building blocks are divided into three different tiers.

For example, when you talk about the data source, then all those capabilities which is such as data ingestion, then storage, AI, and some kind of computation becomes the necessity and the basic building blocks at the edge and this is good enough to support the data source. So, this is called the first tier that is called Data Source second tier is once the data is ingested into the edge computing system.

Then, the second important aspect is to perform the intelligence using that particular data to gain the meaningful insight to perform the intelligence, the edge computing, and the cloud computing together do the artificial intelligence and machine learning models computations. Now, with the data set or the historical data set in place, the AI and machine learning model can be trained in the cloud, these train model can be moved to the edge.

So, the edge will become capable to do the intelligence computations when a new data is presented. And this layer is called the intelligence layer or in the three-tier architecture of edge computing. Finally, the actionable insight that is what this system is being built. So, the data sources which are presented perform the intelligence computation.

And then finally, how that Intel how the actionable insight is to be presented whether it is given to the actuator to perform any action or being displayed, that is called actionable insight and that will become the third layer. So, therefore, we have given the full introduction of edge computing. In the next class, we are going to use this edge computing for an IoT pass. Thank you very much.