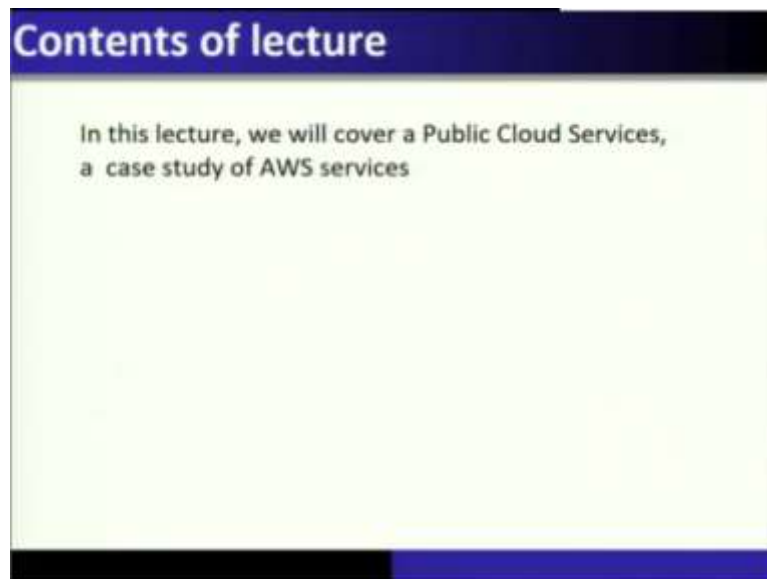


**Foundation of Cloud IoT Edge ML**  
**Professor Rajiv Misra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology Patna**  
**Lecture 11**  
**Public Cloud Services: Case Study of AWS Services**

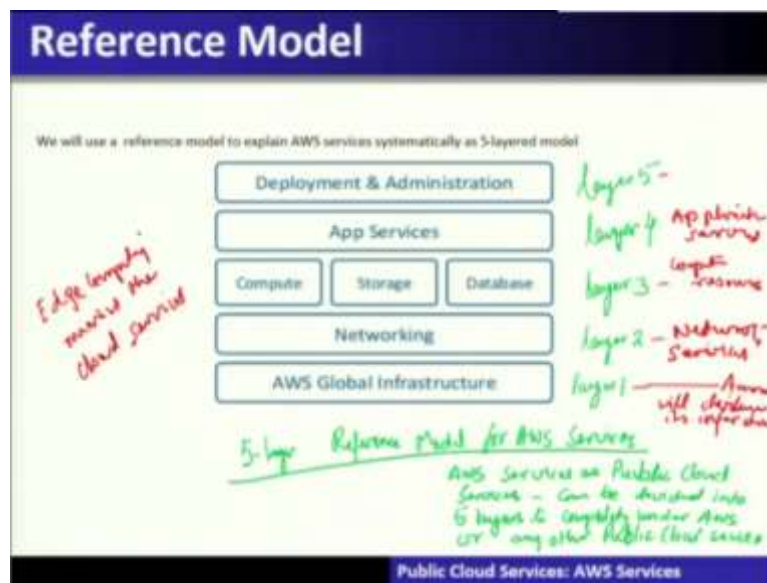
I am Doctor Rajiv Misra from Indian Institute of Technology, Patna. So, the topic of this lecture is Public Cloud Services, a case study of AWS services.

(Refer Slide Time: 0:27)



Content of this lecture. In this lecture, we will cover a Public Cloud Services, a case study of AWS services. Also, we will discuss about the particular use case of using the AWS services to build a particular website or a company, which will be the IT infrastructure, running a startup company, we will discuss at the end.

(Refer Slide Time: 0:51)



Reference model, we will follow a reference model to explain AWS services systematically. It will be five layered model, which is shown what here. So, at the layer 1, layer 2, layer 3, layer 4, layer 5. So, therefore, all the services AWS, which offers as the public cloud services can be divided into 5 layers to completely understand AWS or any other public cloud services.

So, here you can see that at the layer 1, we will talk about the AWS global infrastructure. That means how Amazon will deliver its infrastructure. So, to understand this Amazon's infrastructure for giving out these services, we will understand about AWS global infrastructure. That is how this AWS uses the global distributed infrastructure to provide these services will be described in layer 1.

Layer 2 will be how AWS network these global infrastructure and what are the different networking virtualized services which are being provided for different use cases. Third is about this networking related services will be presented in layer 2. Layer 3 will be about the compute infrastructure or compute resources, compute resources such as compute, storage and database services through AWS. Then various application services.

And finally, the deployment and the administration related services which AWS provides that we will discuss. So, by understanding this AWS services which is being offered by the public cloud services. So, similar such services are also being very quite similar to the services such as Microsoft, Azure and Google Cloud services. So, having understood one such case study of AWS, you will be able to understand how the other public cloud services are available.

And, any of them can be used to build this cloud edge architecture or reference model that is collaborative cloud edge system.

Now, as we have already pointed out that the edge computing mimics the cloud services. So, the functionality of all the services which are available, most of them which are available by the public cloud, are also available at the edge computing in some form. So, we will see that understanding the public cloud services, you will also be able to know how the edge computing will grow.

(Refer Slide Time: 5:48)

**Global Infrastructure**

**Regions (AWS)**  
An independent collection of AWS resources in a defined geography  
A solid foundation for meeting location-dependent privacy and compliance requirements

**Availability Zones**  
Designed as independent failure zones  
Physically separated within a typical metropolitan region

**Edge Locations**  
To deliver content to end users with lower latency  
A global network of edge locations  
Supports global DNS infrastructure (Route53) and CloudFront CDN

Development & Administration  
App Services  
Compute Storage Database  
Networking  
AWS Global Infrastructure

Public Cloud Services: AWS Services

**Global Infrastructure**

**Regions**  
An independent collection of AWS resources in a defined geography  
A solid foundation for meeting location-dependent privacy and compliance requirements

**Availability Zones**  
Designed as independent failure zones  
Physically separated within a typical metropolitan region

**Edge Locations**  
To deliver content to end users with lower latency  
A global network of edge locations  
Supports global DNS infrastructure (Route53) and CloudFront CDN

Storage - co-located services  
Compliant with CDN  
AWS Global Infrastructure

Public Cloud Services: AWS Services

So, let us see about the global infrastructure of AWS. So, this is the first layer, AWS global infrastructure. You can see this board map. This board map shows different blue spots at different places in the world, and they are called the regions. So, regions are the technical

term which is used by AWS to name its data centres or the collection of data centres, they call it as the regions. So, regions of AWS means that an independent collection of AWS resources in a defined geography.

And you know that these blue spots are nothing but the collection of AWS resources in these geographic location. So, this will become the solid foundation for meeting the location dependent privacy and compliance requirement. So, for example, if these regions are present in some part of the country, let us say that, near close to India, maybe in Singapore or some other place, then obviously the location-based services which requires a close proximity where the entire data is to be stored by AWS.

Maybe a user can give the preference to site which is very close to that particular nation or the country, and also that particular country which has the commitment or a trust to make the privacy and compliance requirement enforcement very strictly. So, these are some of the importance of these regions of AWS.

So, another important thing is within the regions, they are availability zones. So, these availability zones are designed as independent failure zones, meaning to say that to avoid a single point of failure, there is a possibility, there is a way that is called replication. So, the same data is replicated at more than one site, so that if one site is failed, then it can continue serving or giving that particular services called availability.

So, availability that is provided through the availability zones. So, within the regions there are availability zones. So, that means that within a particular region, for example, if a particular, data centre or the region, let us say India, so within the India at different places, these availability zones are being created. And for the example of availability zones is nearby cities, and their data centre or their, data centres are networked together using trusted and high bandwidth network. So, these availability zones are designed as independent failure zones.

So, physically they are separated with a typically a metropolitan region. And finally, there is a thing which is called the edge location, which is a part of AWS global infrastructure. So, edge location is to deliver the content to the user with a lower latency. For example, if, let us say that the region and the availability zones are very close to India, and that is the nation. But as far as if the contents which are being accessed or is created by, let us say the users in India. So, they have to be now, then an alternative way is to provide the edge locations.

So, edge locations are to deliver the content to the end users with a low latency. That means the pop location or a cache locations are called edge locations. So, this is a global network of edge locations and it supports the global DNS infrastructure. And these services are called Route 53 and Cloud Front CDNs, that is Content Delivery Network.

Let us see in this particular use case, where this global infrastructure or availability zone will be more useful is in terms of the storage. When, storage as a service is a cloud service when it is taken by the user pay as you go model, that means it uses the users want to use this service to store their files, media files and so on, or a large data set files. Now this particular con storage as a service to make this contents available to the user at a low latency, what they will do, they will now compliment this service or a storage with content delivery network.

So, content delivery network means that if the user who is sitting in India, let us say State Bank of India, is now creating a particular content and the users of that particular service is going to use it, then to make this particular services make available at a low latency, it will be complemented with the content delivery network. Content delivery networks are the cache locations created beyond this regions and availability zones, very close to the user site. They are called edge locations.

Sometimes the concept of edge computing is also involved in content delivery networks. So, content delivery network is also one of the important cloud services which is called cloud front CDNs, and it often requires DNS - Domain Name Service. So, for example, whenever user are searching a particular content, then domain name service, that is route 3 of this particular, we will see that this particular service will divert the route to a very closed, a very close points near to that particular user with the help of this cache location, this lower latency is achieved.

(Refer Slide Time: 12:09)



So, let us see that this global infrastructure often brings into important concept of a global footprint. So, to, for AWS, if you can see that it is available to make its services available. So, it has to provide the availability in terms of regions and available zones and the edge locations across the globe, that is called a global footprint.

So, for example, AWS is available today in US, Brazil, Europe, Japan, Singapore, Australia, and China. So, additional regions are created in UK, Canada, China, Ohio, and there are many more to expect in in further, period of time. And this particular global presence is giving these public cloud services to the users across the globe. Now over 1 million active customers per month, you know that require these kind of services across 190 countries to give this massive, public cloud services without any restriction on a on demand access. Therefore, it requires a global presence.

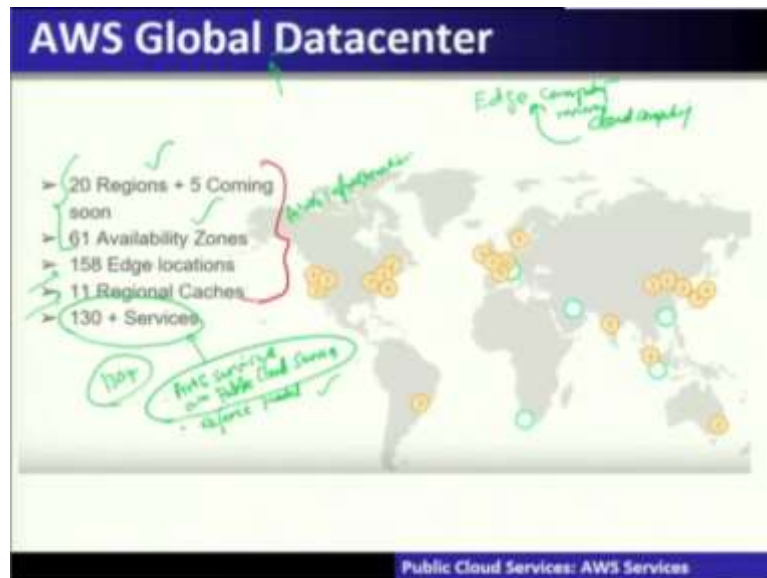
Similarly, let us see that not only 1 million active customer per month across 190 countries, but several 2,300 government agencies are also accessing through this global presence. Similarly, the 7,000 educational institution are also accessing. To support these large number of varied users, so, AWS provides its global infrastructure in form of 13 regions, which are spreaded across the globe. So, you can see that with the orange, with the orange dots. So, these are 13 different locations.

So, you can see that these are the regions which are spreaded across the globe. And you can point out that 13 different regions are there. That means 13 different locations across the globe are shown here in this word map. Besides 13 regions, there are 35 availability zones and many more will be coming soon. And there are 56 edge locations. So, you can see that

the entire globe AWS, to give the AWS services will have this kind of global footprint or presence.

So, it is not only AWS, but any other public cloud provider like Microsoft, Azure or the Google compute engine GC, all of them will also has similar kind of global infrastructure and therefore, they have the global presence.

(Refer Slide Time: 14:58)



Now, AWS Global Datacentre, if you talk about the same thing we have already explained. Let see in this particular word map that AWS global datacentres are present in 20 different regions, and 61 availability zones will be there in those regions. And besides these availability zones and regions, there are 158 different edge locations and there are 11 regional caches which are providing 130 plus services.

So, we will be looking about what these AWS services as a part of this particular lecture, what are these 130 plus services, how they are since to explain all of them, we have given here in the form of reference model. So, this reference model will explain these 130 plus services, which are packed into these different layers. And these services are provided or is being supported with the help of different regions, availability zones and edge locations and regional cache. So, all of them together will, be used to give these kind of services.

So, these are AWS services as part of Public Cloud Services. So, we are going to see, we are going to look at these kind of 130 plus, 130 plus AWS services and to understand this reference model. Now these services are provided by this AWS infrastructure and this AWS infrastructure has the global presence. So, these are all, we know that in the, in the cloud





geographic regions of an AWS is is a geographic area where AWS services are available or through which the AWS services are being provided. So, the customers, they have to now opt or they have to choose these regions for their AWS services.

We will see that in particular server of the services you have to mention, for example, storage and other services. You have to mention the regions where your data is need to be stored. So, those kind of regions you have to, other customers need to know what are the different presents in the form of regions are there, and they have to now choose these regions for their AWS services.

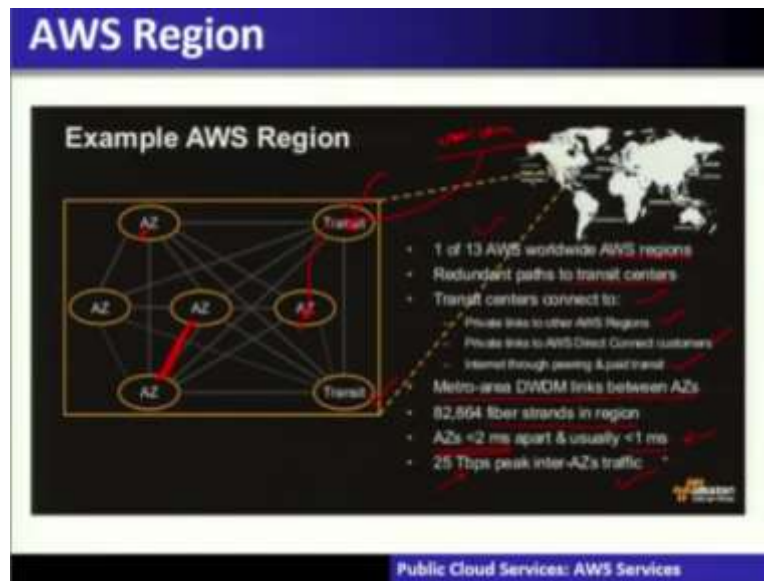
So, this importance, understanding about the regions is very important to run these services on public cloud that is AWS. Now, there are 11 regions worldwide. So, these regions are keep on changing or growing. So, you can understand here within a particular region, the configuration will look like in this manner. So, here you can see 5 different, within one region, there are 5 availability zones.

So, regions may be for example, India. So, within the India, you can see that different locations, for example, this is the Mumbai. Then let us say this is Bangalore. Let us say this is Hyderabad, let us say this is Chennai, or let us say this is again Mumbai, this is again, Bangalore. Within this particular region, let us say India, you can have multiple availability zones or many availability zones. So, this is called a region. This is one region.

So, you can see that there are 11 or 12 different regions which are being operated. And this availability zones are having complete full-fledged data centre. So, these data centres are run with the availability zones. These data centres are networked together with a high bandwidth and they are all connected in a mesh. Out of them, five, there are two other availability zones or these particular regions. These particular regions are connected with the outside world. So, these regions will have the connectivity with the outside world. And this is a failure resistant structure.

For example, if the data is stored at one place that is in one of these availability zones, it has to be replicated in more than one places. So, if let us say this availability zone has a failure, then still you keep on continuing getting the data. Therefore, this particular architecture of the distributed form of availability zones, then will make the services available all the time.

(Refer Slide Time: 22:17)



So, this is an example of again AWS. You can see that out of 13 AWS worldwide, there are different AWS regions. So, these AWS regions are or has multiple availability zones within it. And these availability zones are connected with the redundant paths. You can see that there are many ways of reaching this availability zones from here or from any other places. There are many redundant path are there.

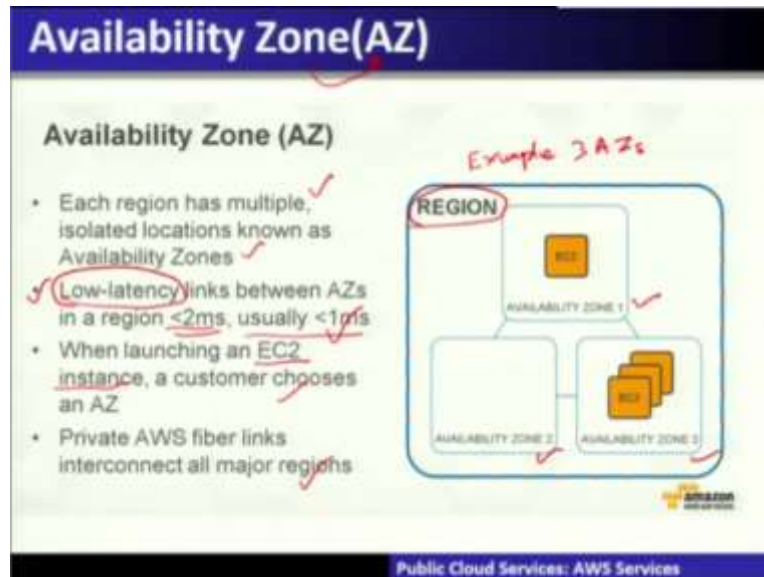
So, these transit centres are used to connect to the private links to the other AWS regions. Private links are there to AWS direct connect with the customers and internet through peering and the paid transits. So, these are the points by which whenever you type www dot google dot com. So, maybe that if that webpage is kept in one of these places, then these are the transit through the transit point. It will go inside the data centre and inside data centre, a particular server which will be serving this kind of services.

So, initiative of google dot com, maybe let say that, uber dot com or some other services when you type. So, it has to be or it will be serviced by these availability zones or the servers within it. Similarly, these particular AWS regions as the motor metropolitan area, links between this AWS regions or, availability zones, which are nothing but having the fiber strands in that region to provide this availability zones, the latency which is less than 2 milliseconds apart.

So, many number of availability zones are present within a particular region so that this particular latency should remain less than 2 milliseconds or usually less than 1 milliseconds. So, this is an important parameter. To bring this important parameter, this architecture or AWS global presence is there. So, here you can see that there are 25 terabytes per second

peak inter-AZ traffic. So, you can see that the high bandwidth links are being created across availability zones.

(Refer Slide Time: 24:42)



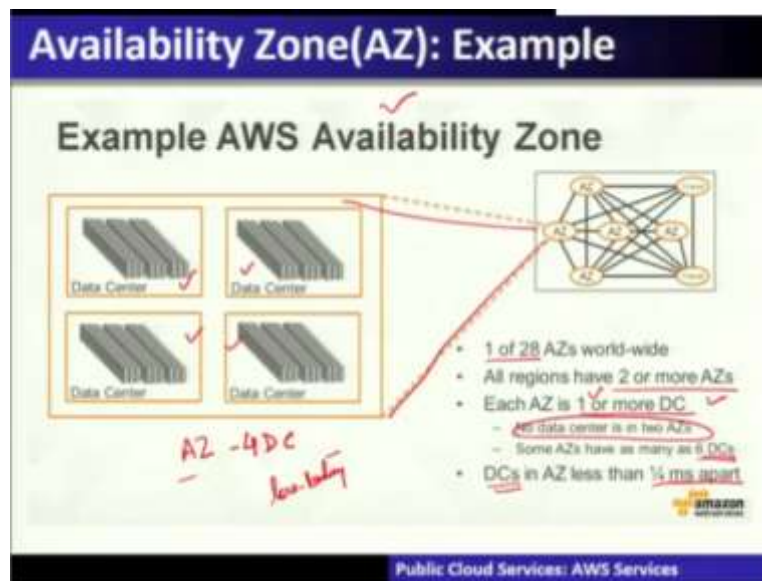
So, therefore the availability zones, if you see inside view of availability zones, so this is an example. So, we are going to see an example often of availability zones and what kind of services or AWS services these availability zones are offering to the public cloud as a service. So, here you can see that the short form of avail availability zones is AZ. To understand this availability zone, you have to now first understand about the regions. So, within the regions, these availability zones are operating.

So, availability zones are not without regions. So, under that regions, these availability zones are there. So, each region has multiple isolated locations and they are called availability zones. So, here in this particular example of a region, this region has 3 availability zones. These availability zones are there to provide low latency links between availability zone in the region, which is to provide the access, in the space, which is quite less than 2 milliseconds and usually it is less than 1 millisecond latency.

So, this particular latency is called a low latency. If the latency is less than 1 milliseconds to X for access through that particular link, then it is called a low latency. So, whenever a low latency is referred, that means we are talking about a latency less than 1 milliseconds. Now, whenever you want to launch a AWS EC2 instance, which is nothing but an infrastructure as a service, then the customer has to choose this availability zone.

So, from where this particular infrastructure as a service customer wants to choose? So, for availability zone, for example, if a infrastructure as a service is used, so customer has to specify that availability zone number. For example, whether availability zone 1, 2, or 3 in that region. So, customer has to choose. So, within that particular region availability zone, this infrastructure as a service will be hosted for that particular user. Therefore, when launching an EC2 instance, a customer has to choose, availability zones. Now, private AWS fiber links, interconnects all the major regions. So, that is being there.

(Refer Slide Time: 27:13)

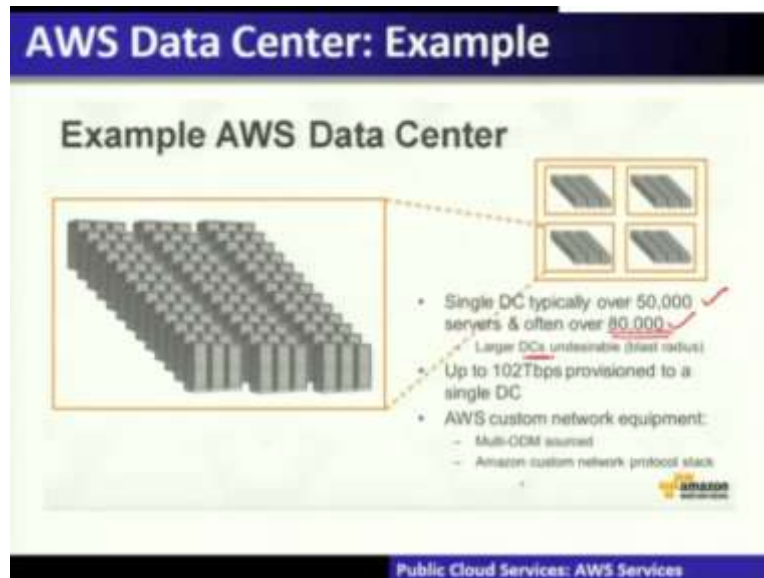


Now let us see within that availability zone, what is there? So, within our availability zone, you can see that it has more than one data centres. So, more than one data centres which are clustered together and it is called availability zone. So, example of an availability zone is shown where here that 1 out of 28 availability zones are there worldwide. And all these regions have 2 or more availability zones.

So, each availability zone has 1 or more data centre. So, here in this availability zone, it has 4 data centres. So, either each availability zones would have 1 or more data centres. So, no data centre is in two different availability zones. So, so this data centre is exclusively in this particular availability zones. So, this is very important to understand that no data centre is in two availability zones.

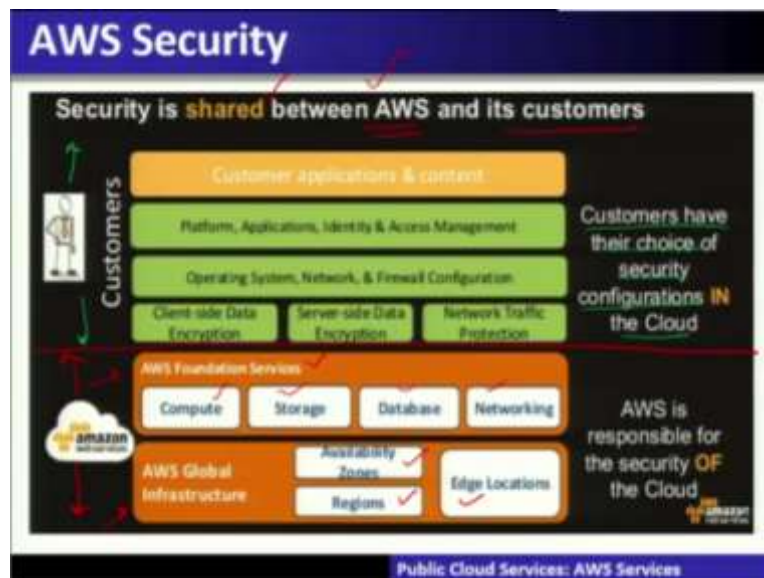
So, some availability zones have as many as 6 data centres. And these data centres in availability zones will provide less than one fourth of a millisecond. So, as I told you that it is, this particular configuration is to provide low latency services.

(Refer Slide Time: 28:21)



Now, inside the data centre, if you look, what you will find is that a single data centre has 50,000 servers and often over 80,000 servers. Therefore, you can see that inside view of a data centre is nothing but 80,000 servers are placed at one data centre in that is in AWS. So, larger data centres normally are undesirable. So, after 102 terabytes per second provision to a single data centre, AWS custom network equipments are used here in AWS data centres.

(Refer Slide Time: 28:56)



Now let us see about the cloud security or AWS security. Now the security is shared between AWS and its customers. That means, it is not a single entity's responsibility to provide the full security, therefore security is shared between AWS and its customers. So, let us see that what part of the security is the responsibility of AWS and what other is the customer's

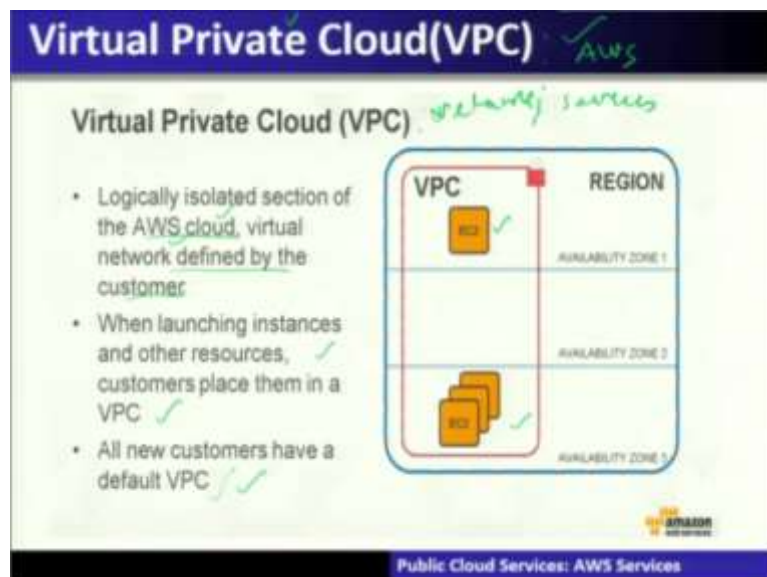
responsibility. As far as AWS responsibility of providing the security is that that is AWS responsible for the security of the cloud is in the following form.

So, AWS is responsible when it talks about AWS global infrastructure. That is, the security of availability zones, regions edge location is the responsibility of AWS. Similarly, AWS foundational services such as compute, storage, database and networking is a part of AWS services. So, whatever is shown below this line is a part of AWS. So, if you see the reference model, so this is AWS global infrastructure and AWS Foundational Services, foundation services, these are the two places where this is the responsibility of AWS.

Now, beyond this, that is the upper part is mainly the applications part. And this applications are owned and run by the customers. So, rest of these part of the security is the responsibility of the customers. So, customers have their choice of the security configurations in the cloud.

So, therefore, as far as the clients are concerned or customers are concerned, they are responsible for the client-side data encryption, server-side data encryption, network traffic production, operating system network, firewall, platform applications, identity and access management, customer application and contents are all, are the responsibilities of the customer as for the security is concerned.

(Refer Slide Time: 30:56)



So, let us see about the networking services. So, the networking services, if you see in that reference model, there is a network services. So, network services, which AWS provides is called as is called as virtual private cloud. So, virtual private cloud is the networking services. So, let us understand virtual private cloud. Virtual private cloud is a logically isolated section



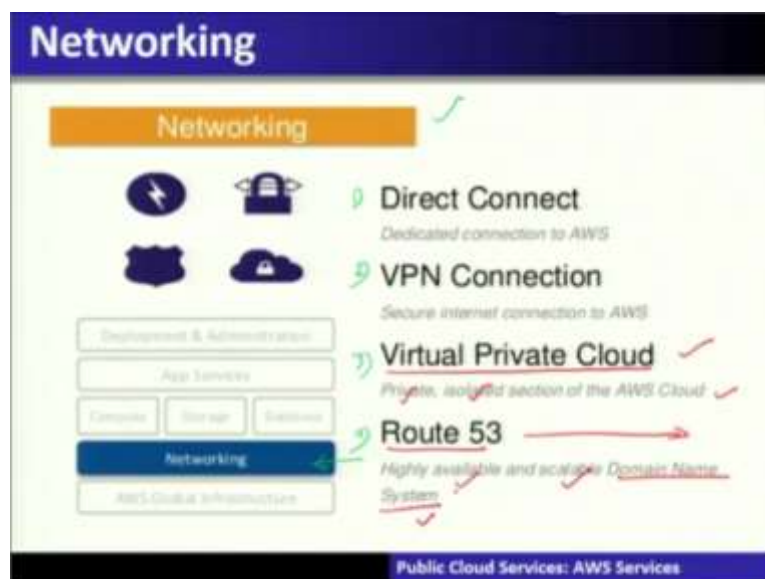
of AWS cloud. Virtual networks are defined by the customers. So, when launching the instances and other resources, customers place them in this virtual private cloud. All the new customers have the default virtual private cloud.

For example, the client requires its services in the form of infrastructure service EC2, it will place it into availability zone, let us say 1 of a particular region and rest other EC2 are placed in availability zone 3. Now, this particular different EC2 instances, EC2 if you understand it is a virtual machine. So, virtual machines, you know that if you have the virtual machines or a physical machine in your company, then they must be network together, otherwise isolated form of machines is of no use.

So, how do you create the network on virtual network, using the AWS services. So, this is called virtual private cloud. So, once you have this EC2 define, then the networking of this part is done using the virtual private cloud. So, again, I am repeating this logically isolated section of AWS cloud that is virtual networks are defined by the customers. So, customers has to define this virtual private cloud for the security of their services or applications when they are doing it.

So, when launching these instances and other resources, customers place them in a virtual private cloud and all, whenever a new customer joins its services, it will be now have a default virtual private cloud.

(Refer Slide Time: 33:08)



Again, let us go in more details about the networking part of that layer. So, you can see where the networking part of the layer is. Layer 2 talks about the networking services, networking

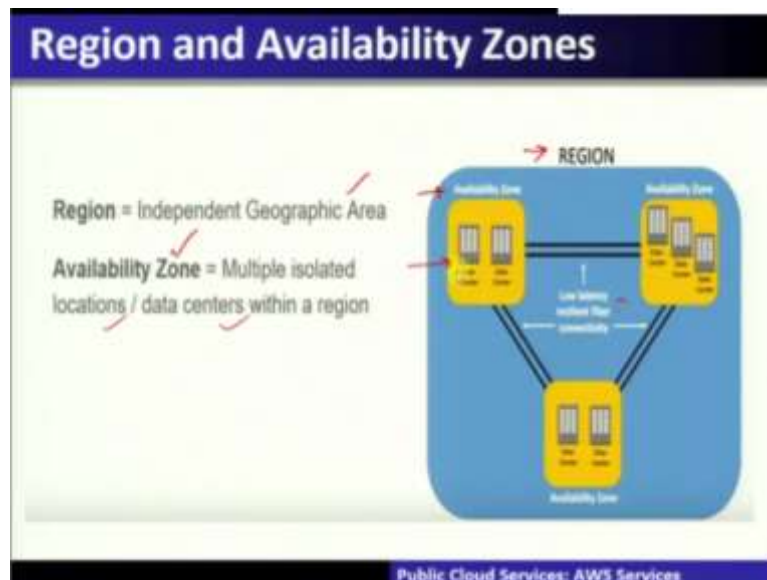
of an AWS services which it provides to the customers. So, as far as the networking services are a concerned, the important networking services which AWS provides is, direct connect VPN connection, three is the virtual private cloud. Fourth is a route 53.

Let us understand few of them, but not all of them. So, few of them means we were talking about virtual private cloud. So, virtual private cloud is a very important service. For example, local area network of a particular company is an example of a virtual private cloud on, using the cloud services. So, it provides a private isolated section of AWS cloud. Another important service is called Route 53.

So, Route 53 is a domain name system DNS on the cloud. For example, if you are running your services, your web server or your website hosted on these virtual machines or EC2 systems which are connected using virtual private cloud than to reach or to access these services, you require a DNS.

So, DNS also is a service which the AWS will provide, and this is called Route 53. So, Route 53 is a highly available scalable domain name system. So, networking services are categorized in these different forms. So, the AWS, cloud provides these networking services, which we have explained here, in the reference model.

(Refer Slide Time: 34:50)



Now, let us go above in all these. So, regions, again, let us revisit that regions and available zones together. So, we have shown you the presence of regions, which is nothing but independent geographic area. Availability zones is a multiple isolated locations, so that we have already explained and the data centres are within that particular regions.



So, you can see in this particular example that this is the region, this independent geographic area. Within that regions, there are availability zones. Within the availability zones, there are data centres. These data centres are available zones are often interconnected with a low latency resilient fiber connectivity to provide a low latency services.

(Refer Slide Time: 35:35)

**Why Availability Zones?**

**Why Availability Zones?**

Challenges with traditional asynchronous replication between distant data centers

- Committing to an SSD order 1 to 2 ms
- But Sydney to Melbourne is 20 ms
- You can't wait 20 ms to commit a transaction

Traditional failure, difficult decision:

- Failover & lose transactions, or
- Or don't failover & lose availability
- Difficult choice

AZs for no-admin failover

- Sync works when < 2 ms
- Combine with regional replication for very high availability (VHA)

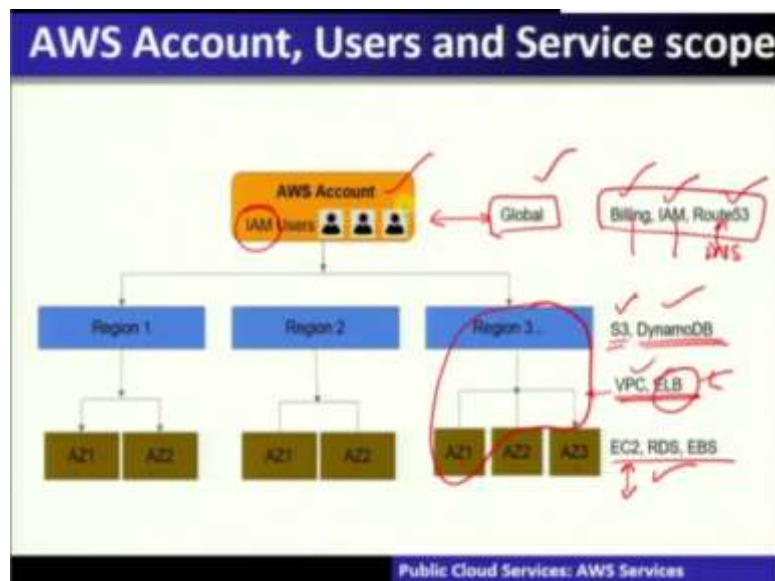
20 ms

Public Cloud Services: AWS Services

Now the question is why availability zones are there. Now, challenges with a traditional asynchronous replication between distant data centres, and committed to that particular latency of 1 to 2 milliseconds. And it is very difficult and obviously traditional things are failure prone, but it is to be, overcome from this. So, availability zones, are there for no admin failure. And this particular synchronization brings the latency less than 2 milliseconds, often combined with the regional replication for a very high.

So, this is the use case that if let us say that, otherwise, such a long distance will incur the latency to the tune of 20 milliseconds. But with the help of the AWS regions, AWS with the help of regions, availability zones and their data centre classification and networking, this particular availability zones is less than 2 milliseconds.

(Refer Slide Time: 36:34)



Now let us start with the most basic and primitive use when the user use the AWS, then it has to use these services called AWS accounts, users and services. And then we have to understand. So, AWS minimum thing which is required to become AWS user is to have this AWS account. So, for that, there is, identity and access management service IAM of AWS will provide all these services.

Now, this is responsible that means once an user account is created with the help of identity and access management to create different user account. So, this particular user account, AWS is a unique or globe. So, it is a global entity and, this will be used for billing identity and access management. And Route 53, that is the DNS service. DNS also understands this particular AWS users and identity access management also creates the user's power. And the billing also is done based on this user account.

Now this AWS account services, which is a global one. So, over that, if you go inside, the deep you will find there are different regions are there. So, there are 11, 12 regions, which we have shown in the previous, slides. So, let us see that 1, 2, 3 and so on. So, there are 13 or 14 or more than that ever-growing the regions are there. So, let us see that what are the AWS services that is 130 plus services are there out of them which services will require to specify the regions where these services are required to run, to be specified?

So, as far as, these S3 is concerned, we will make you understand what you mean by S3, the simple storage service of AWS. DynamoDB is a database services of AWS. So, if you are using a storage type of service or a database type of services, then you have to specify the regions, whether you accept to store your data. For example, you don't want your data to let

us say be stored in unfriendly countries. So, therefore, the regions or, or you have the preference to store your data in some friendly countries.

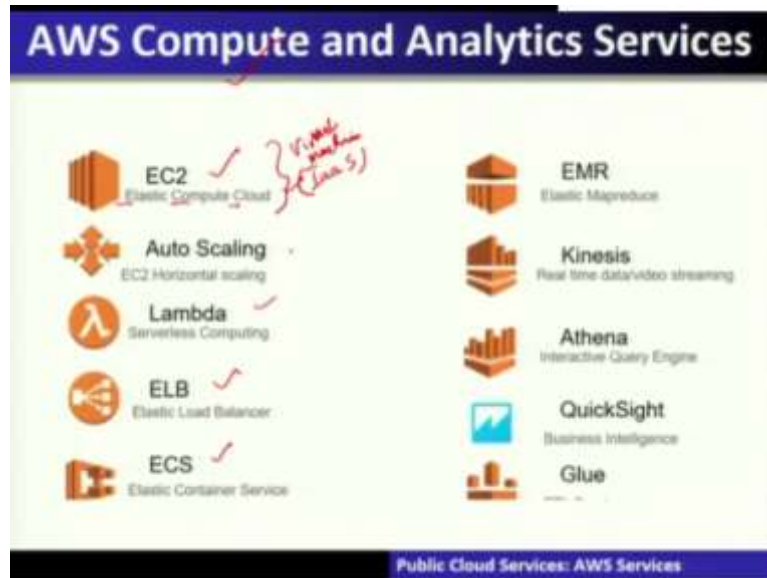
So, the regions which are based on these, geography locations, you can specify and become comfortable that your data is following a lot of regulations and that is acceptable to you as well. Now, as far as, the networking services are concerned, for example, VPC, that is Virtual Private Cloud. So, virtual private cloud is operating between regions and availability zones. So, you can see that it is in between.

So, when you go for virtual private cloud, then you have to specify both regions and availability zones together. Similarly, ELB - Elastic Load Balancer. So, elastic load balancer is also a networking service. So, networking service also you have to specify the regions and availability zones. So, these are some of the networking services. Storage services requires regions to be specified. Networking services requires regions and availability zones to be specified.

How about this type of infrastructure as a service called EC2, RDS and elastic blockage storage. So, these are all infrastructure as a service. So, infrastructure as a service. For example, you have a virtual machine. So, similar to the virtual machine here, AWS has EC2. So, EC2, to run that service, you have to specify which availability zone you are choosing where these EC2 machines will run.

Now you know that availability zones within a particular availability zones, you have one or more data centres. So, inside data centres you have 50,000 to 80,000 servers and your services EC2 services will run from these kind of servers. So, you have to specify during when you are going for e EC2 type of service, which availability zone you have to specify. So, that means user of a public cloud services need to know about AWS or the cloud providers, global infrastructure. So, that knowledge has to be there to use all these services of AWS. That is why we are explaining in more details.

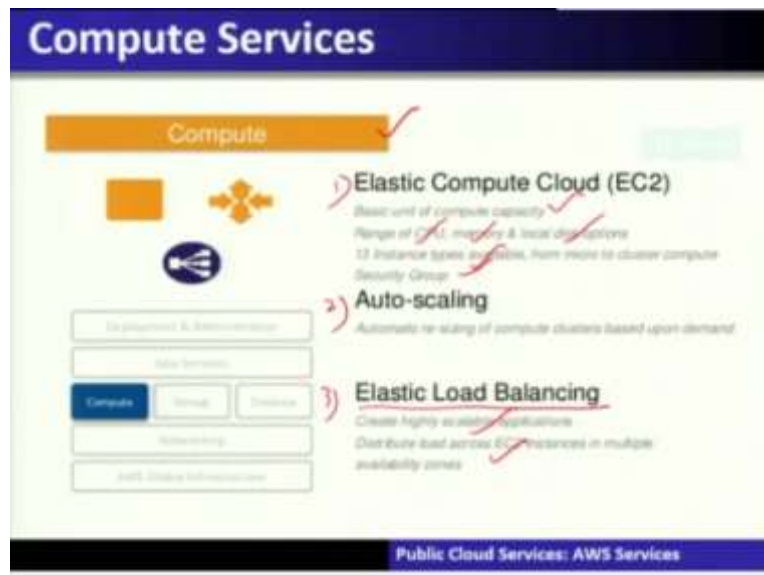
(Refer Slide Time: 40:39)



Now, AWS compute and analytics services. So, we are now going in more detail about those compute services. So, what are the different compute and analytics services are being provided by AWS out of 130 plus services, they are the following. Some of them we will explain not, but not all of them. So, very important. AWS compute services is called EC2. So, EC2, that means E and 2 times C. That means elastic compute cloud.

So, they are writing ECC, administrative ECC, they are writing EC2. So, that is a elastic compute cloud. This is nothing but a virtual machine. And this is a typical example of infrastructure as a service. Similarly, there are other type of services called ECS - Elastic Container Service, elastic load balancer, Lambda, autoscale, EMR and so on. Let us go ahead.

(Refer Slide Time: 41:33)



So, this is a compute service. So, compute services are three important compute services there. So, Elastic Compute Cloud - EC2. So, elastic compute cloud is a basic unit of compute capacity. A range of CPU memory and local disk options are there that you would choose. For example, it is very similar or it mimics to a physical machine.

So, if it mimics a physical machine, then it has the CPU, it has the memory, and it has a local DIS options. So, all things has to be being provided as a basic unit of compute capacity under EC2. There are 13 different instances types are available for e EC2 from micro to cluster compute and security groups.

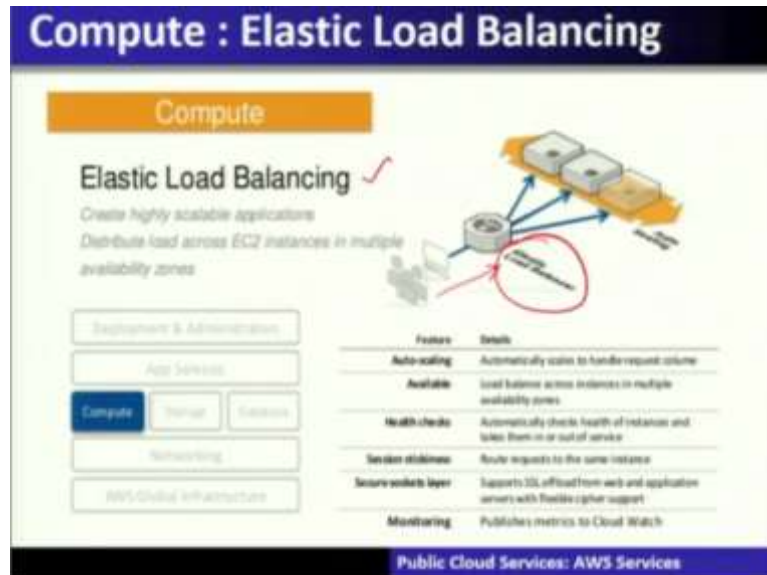
So, another example of this compute services called auto scaling automatically auto resizing of compute clusters based on their demands, so called auto scaling services there. Third is elastic load balancing is to create highly scalable applications, distribute the load across EC2 instances, multiple availability zones.

So, this is also very important service. For example, if you are running a, a startup company and many EC2 instances are there running in different availability zones. For example, to run a Uber application across the globe. So, you require to know how to use this elastic load balance service, which is a type of compute services.



scaling that is automatically resizing of the compute cluster based on the demand. So, all these are being provided under this.

(Refer Slide Time: 43:44)



**Compute : Elastic Load Balancing**

**Compute**

**Elastic Load Balancing** ✓  
Create highly scalable applications  
Distribute load across EC2 instances in multiple availability zones

Deployment & Administration  
App Services  
**Compute** Storage Database  
Networking  
AWS Global Infrastructure

**Features** **Details**

<b>Auto-scaling</b>	Automatically scales to handle request volume
<b>Available</b>	Load balance across instances in multiple availability zones
<b>Health checks</b>	Automatically checks health of instances and takes them in or out of service
<b>Session stickiness</b>	Route requests to the same instance
<b>Secure websockets layer</b>	Supports SSL offload from web and application servers with flexible cipher support
<b>Monitoring</b>	Publishes metrics to CloudWatch

Public Cloud Services: AWS Services

Elastic load balancer. For example, when, load is entered from the customer, for example, if it is www amazon dot com, which is nothing but e-commerce site, e-commerce site means multiple users are simultaneously want to buy the online shopping. Then in that case, you know that, to have this load equally balanced across different EC2 instances running across multiple availability zones. Auto is scaling, that is elastic load balancer is there. So, our elastic load balancer that is ELB, create highly scalable applications distribute load across EC2 instances in multiple availability zones.

(Refer Slide Time: 44:25)



**Storage**

**Storage**

Deployment & Administration  
App Services  
Compute **Storage** Database  
Networking  
AWS Global Infrastructure

✓ **S3 - Durable storage, any object** ✓  
99.999999999% durability of objects  
Unlimited storage of objects of any type  
Up to 5TB size per object

✓ **Elastic Block Store** ✓  
High performance block storage device  
1GB to 1TB in size  
Mount as drives to instances

Public Cloud Services: AWS Services







So, this is the typical example of simple storage service that we have explained. Another type of service also we have explained is elastic block storage, EBS. Then another important services under this is called database as a service. So, relational database as a service. So, it is a database as a service, no need to install or manage the database instances. So, they are already created. You have to pay as you go by, as you use by model.

So, DynamoDB is a important relational database, service. It is provision, throughput and no, no SQL database. Fast and predictable performance, fully distributed, full toler architecture. We are not going in detail about these databases, but, you can understand that different flavours of relational databases as a services is being provided by AWS. It is up to the users application to choose whether Oracle or MySQL or Microsoft my SQL server is used or here.

(Refer Slide Time: 46:17)



Now another important service is called application services. You can see that we have finished, layer 1, 2, 3, and we are now discussing about fourth layer in the reference model. Fourth layer has different application services. So, these application services such as, CloudFront, it is worldwide content delivery network.

For example, if you are running a company like Netflix. So, Netflix means that you are, providing the content, with a low latency. So, to provide the content with a low latency, the availability zones and the regions will not be sufficing. So, therefore, edge locations will be very much needed and beyond edge locations, this content delivery network or a caching locations are there. And that will provide using the service, which is called a CloudFront.

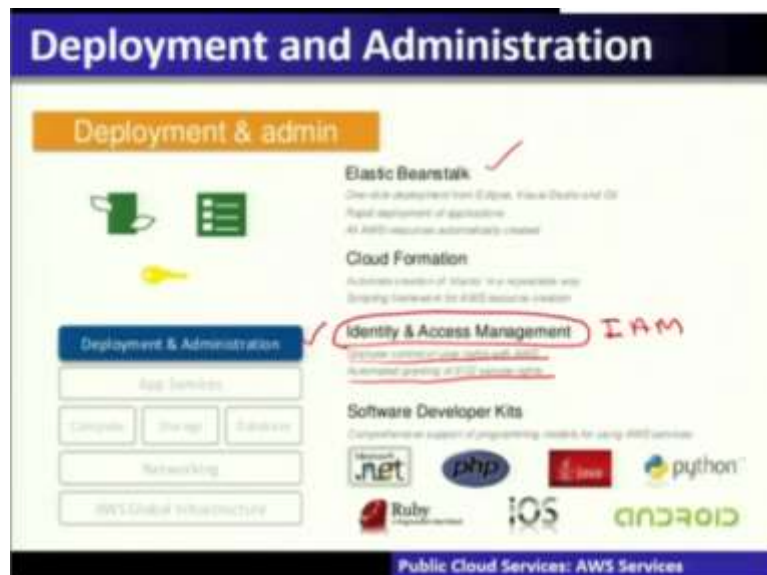
So, easily, distribute the content to the end user with a low latency and high transfer speeds and no commitment. So, a lot of the startup companies are now, are there and they are using this kind of cloud service.

(Refer Slide Time: 47:17)



So, CloudFront, we have already covered.

(Refer Slide Time: 47:20)










So, now the next type of, service is called deployment and administration. So, finally the fifth layer is deployment administration layer. In this, let us see what the different services are available. The different services are the Elastic Beanstack, cloud formation, identity and access management.

Most important is identity and access management - IAM, which I have told you that this is the global service. So, it provides the granular control of user rights with AWS automated, granting of AWS access rights. All this thing is done over by this deployment and administration services.




(Refer Slide Time: 47:59)

### AWS Storage and Database Services

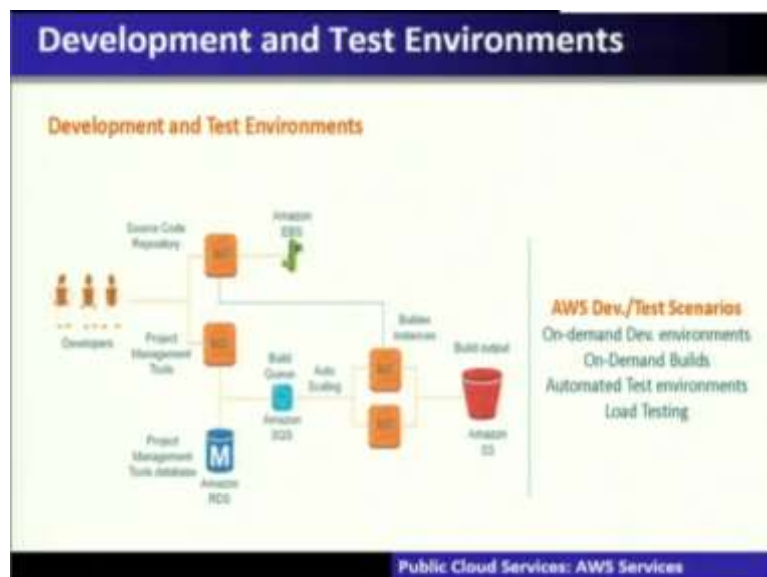
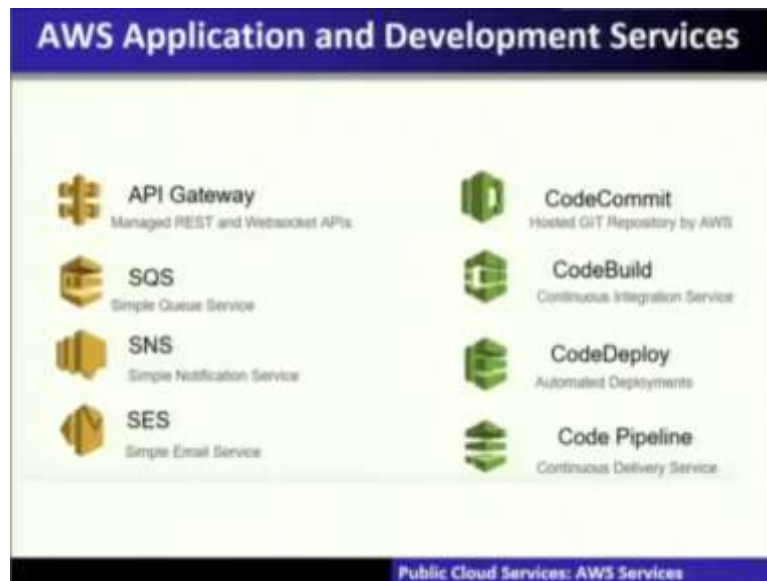
 <b>EBS</b> Elastic Block Storage	 <b>RDS</b> Relational Database Service
 <b>S3</b> Simple Storage Service	 <b>DynamoDB</b> AWS NoSQL Database
 <b>EFS</b> Elastic File System	 <b>Redshift</b> Data Warehousing
	 <b>Elasticache</b> Fast and Flexible caching

Public Cloud Services: AWS Services

### AWS Network and Management Services

 <b>VPC</b> Virtual Private Cloud	 <b>CloudWatch</b> Application & Infrastructure monitoring
 <b>Route53</b> AWS DNS Service	 <b>CloudFormation</b> Provision Infrastructure as a Code
 <b>Direct Connect</b> Dedicated Network	 <b>Elastic Beanstalk</b> Application Orchestration Service
 <b>CloudFront</b> Content Delivery Network	 <b>Opsworks</b> Infrastructure Configuration Management

Public Cloud Services: AWS Services



So, again, we are repeating some of the important public cloud services, which are being provided are storage and database services and different forms, network and management services that we have also explained. Application and deployment services, then deployment and test environment. So, that means using this AWS public cloud services, one can create its own virtual company on the network without upfront buying all these different IT infrastructure.

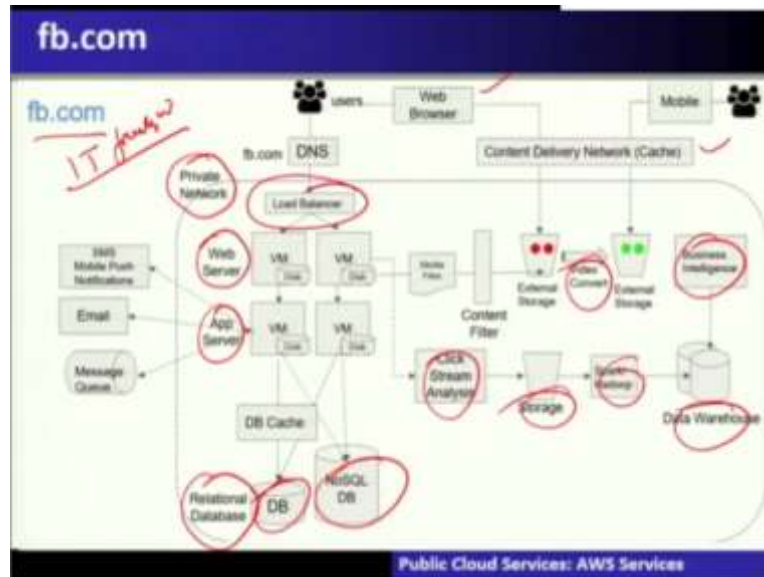






Storage and backup and archival services also can be created in the company. Disaster recovery services also can be created in the company. These companies also can create web, mobile and social app through these kind of services.

(Refer Slide Time: 48:52)



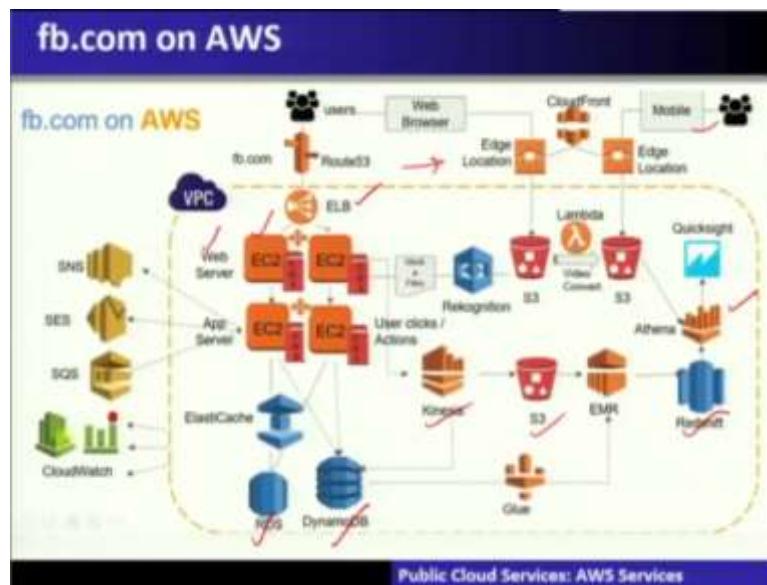
And let us now understand that a particular startup company, let us say fb dot com earlier, used to buy all these different type of IT infrastructure to run this fb dot com. So, what is this fb dot com? What this company has to do, it has to create a private network of different servers called web servers, application servers, relational database, and then it has to buy the load balancer. And then inside that it has to deal with different media files and media converters.

It requires doing a lot of business intelligence for that. It requires to buy a lot of storage and so on. That is, it has to create data warehouse, noSQL database, DBMS and so on. So, all these things is required to run this IT infrastructure to provide two types of services. One is content delivery network that is the caching to the web.

So, the users who are accessing that company through web browser, it has to be provided with a low latency content through content delivery network. If a user is accessing this application or a company using mobile phone, then also has to go through this content delivery network. But these kind of media files requires the converters for mobile phone. It is different for web browser. It is different that it has to now interface with, message queues, email and so on.

So, that means if a company is owning its own IT infrastructure, it requires a huge investment. Not only that, it requires lot of IT manpower to manage it, and that particular company will become lost in all these things besides running their actual application or business. Instead of that, lot of new startup companies are there without investing so much in the IT infrastructure. They can use the cloud services and build it.

(Refer Slide Time: 50:58)



So, this is the typical example that if the same company is running with the help of 130 plus public cloud services, it can now run without such prior investments. For example, these web servers are nothing but the EC2 instances of AWS. The load balancer is nothing but ELB service and then RDBMS, RDS and DynamoDB services will be there. And then if let us say it is a mobile phone, then the content delivery network will be provided in the form of edge locations.

This is, the data warehouse called Redshift and business intelligence is, is Athena. And, so big data is in the form of Kinesis and storage is in the form of S3 buckets. So, everything you know that is now being provided as AWS services. That is why this cloud public cloud services are very much in need and in the cloud edge type of environment, these services are very much needed. Why? Because these edge locations, they have to decide what kind of cloud services it requires, whatever is not available in the edge.

So, these type of cloud services are now shifting to run very close to the user devices called edge location. So, edge computing is becoming more and more important these days because of many advantages which we have seen. And explaining these public cloud services is very



relevant for the edge computing, so that you will see that the developments of edge computing will bring into these services to run in the in the Edge locations. Thank you.