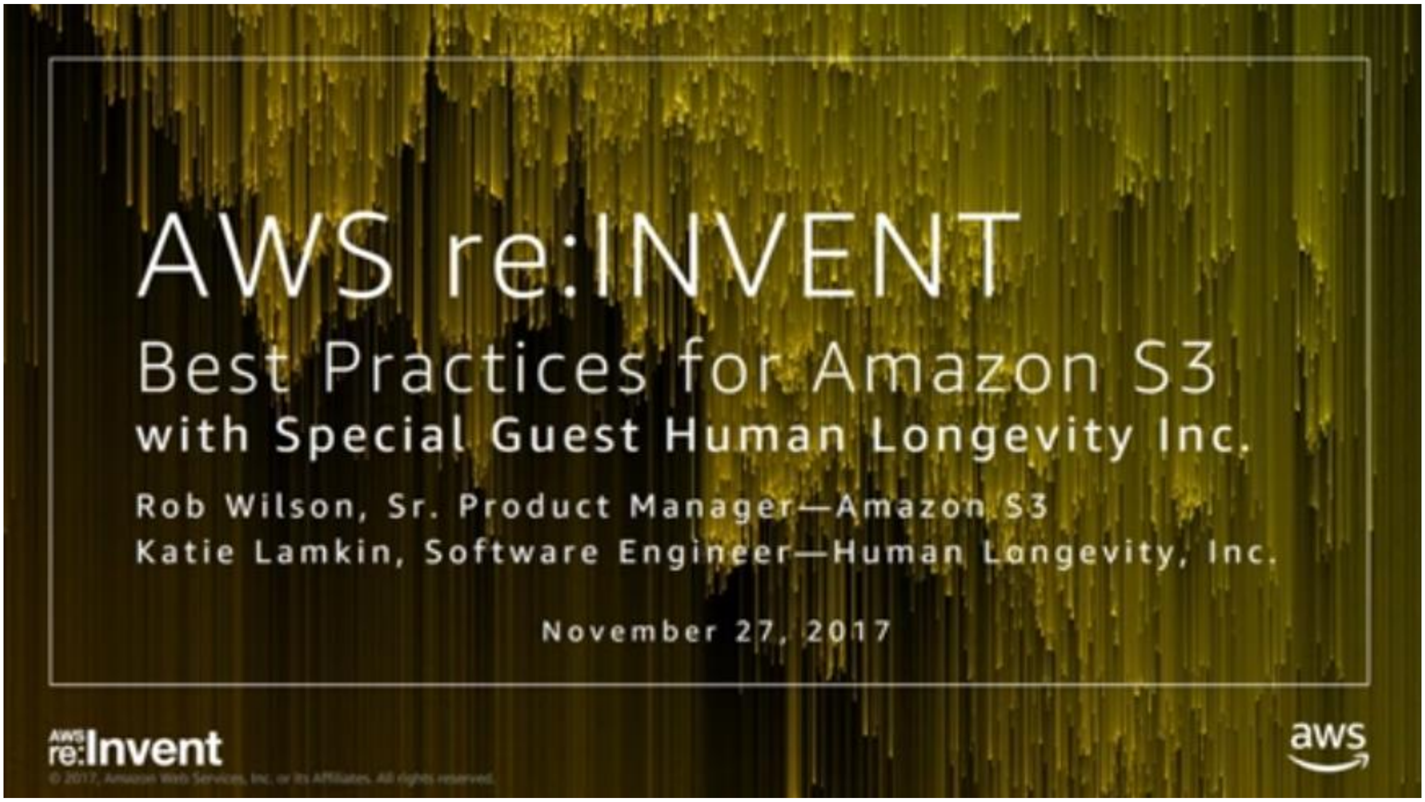


AWS re:INVENT

Best Practices for Amazon S3

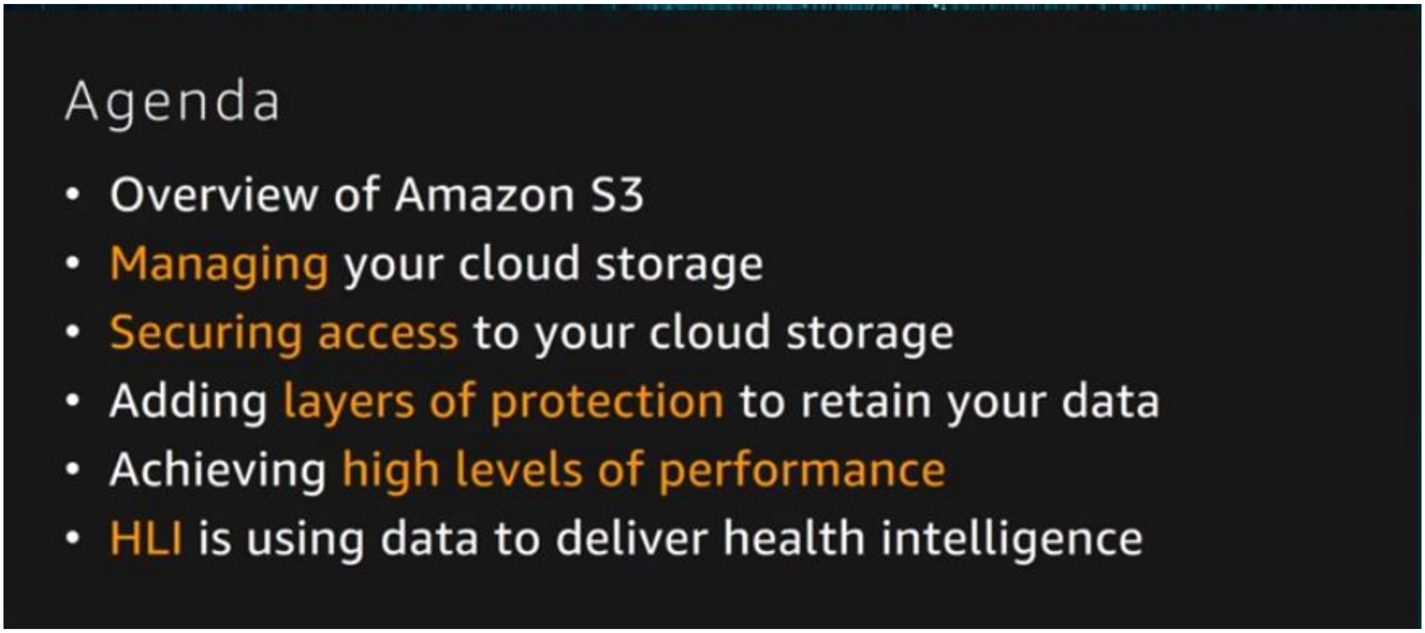
with Special Guest Human Longevity Inc.

Rob Wilson, Sr. Product Manager—Amazon S3
Katie Lamkin, Software Engineer—Human Longevity, Inc.

November 27, 2017

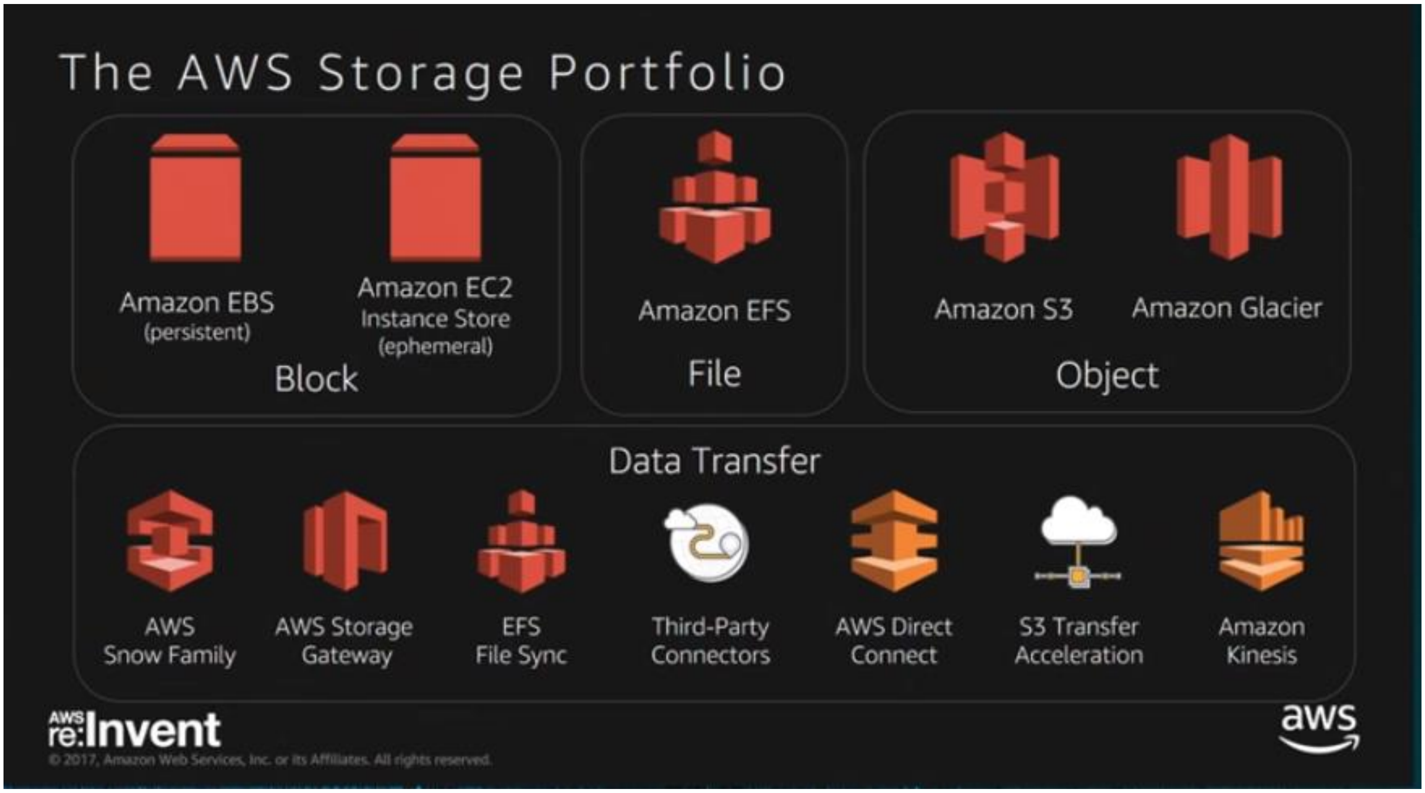
re:Invent
aws

This is a company that is using data to deliver health intelligence.



Agenda

- Overview of Amazon S3
- Managing your cloud storage
- Securing access to your cloud storage
- Adding layers of protection to retain your data
- Achieving high levels of performance
- HLI is using data to deliver health intelligence



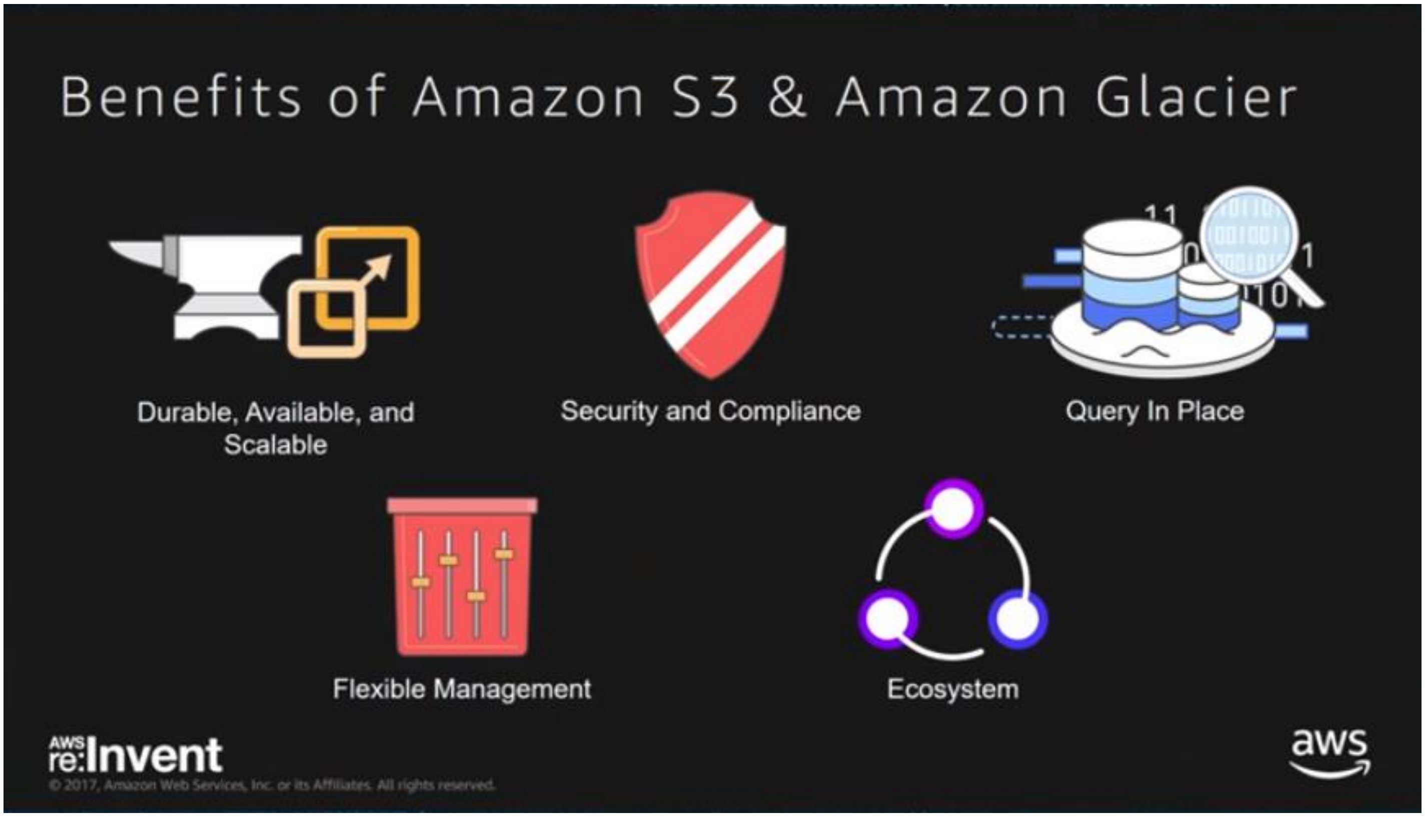
The AWS Storage Portfolio

Block
- Amazon EBS (persistent)
- Amazon EC2 Instance Store (ephemeral)

File
- Amazon EFS

Object
- Amazon S3
- Amazon Glacier

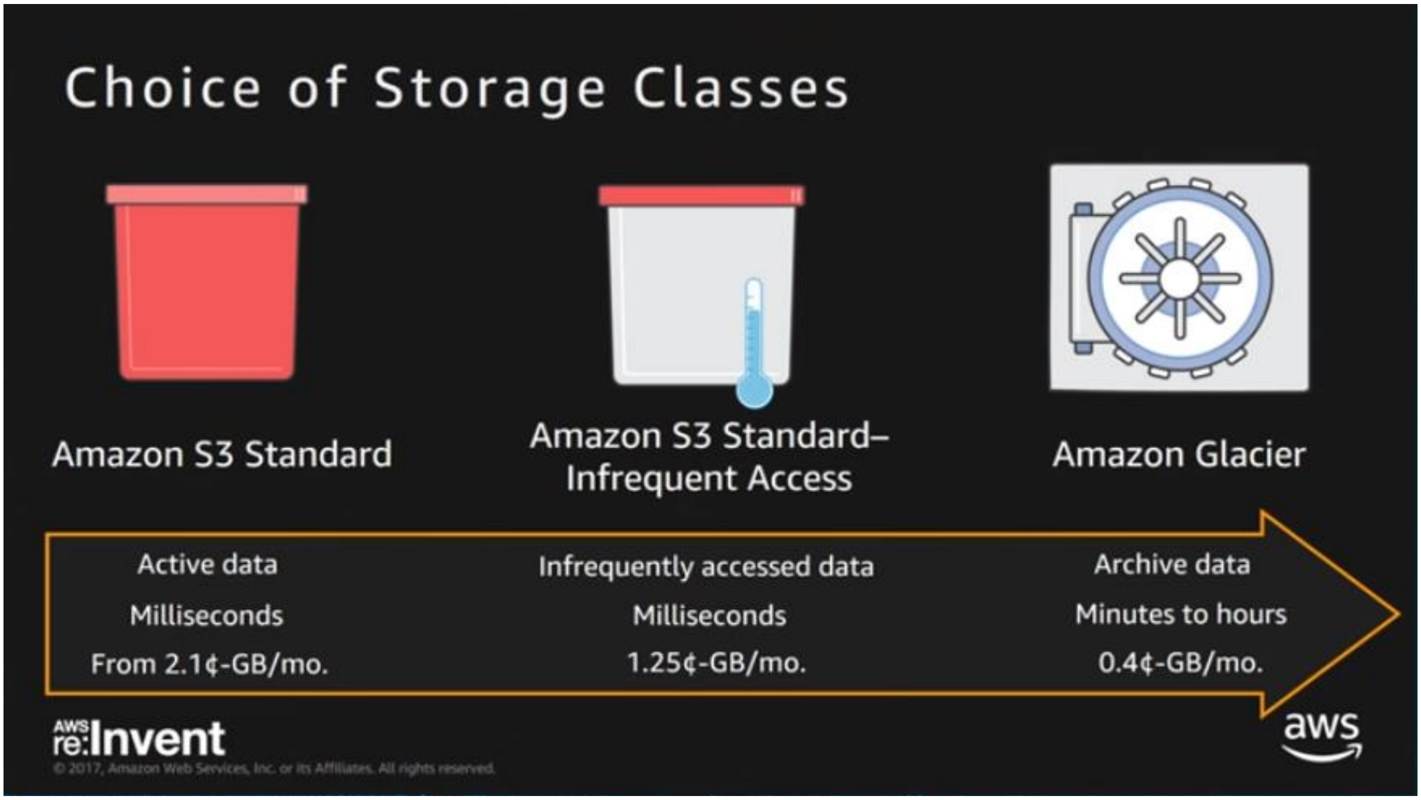
Data Transfer
- AWS Snow Family
- AWS Storage Gateway
- EFS File Sync
- Third-Party Connectors
- AWS Direct Connect
- S3 Transfer Acceleration
- Amazon Kinesis



Benefits of Amazon S3 & Amazon Glacier

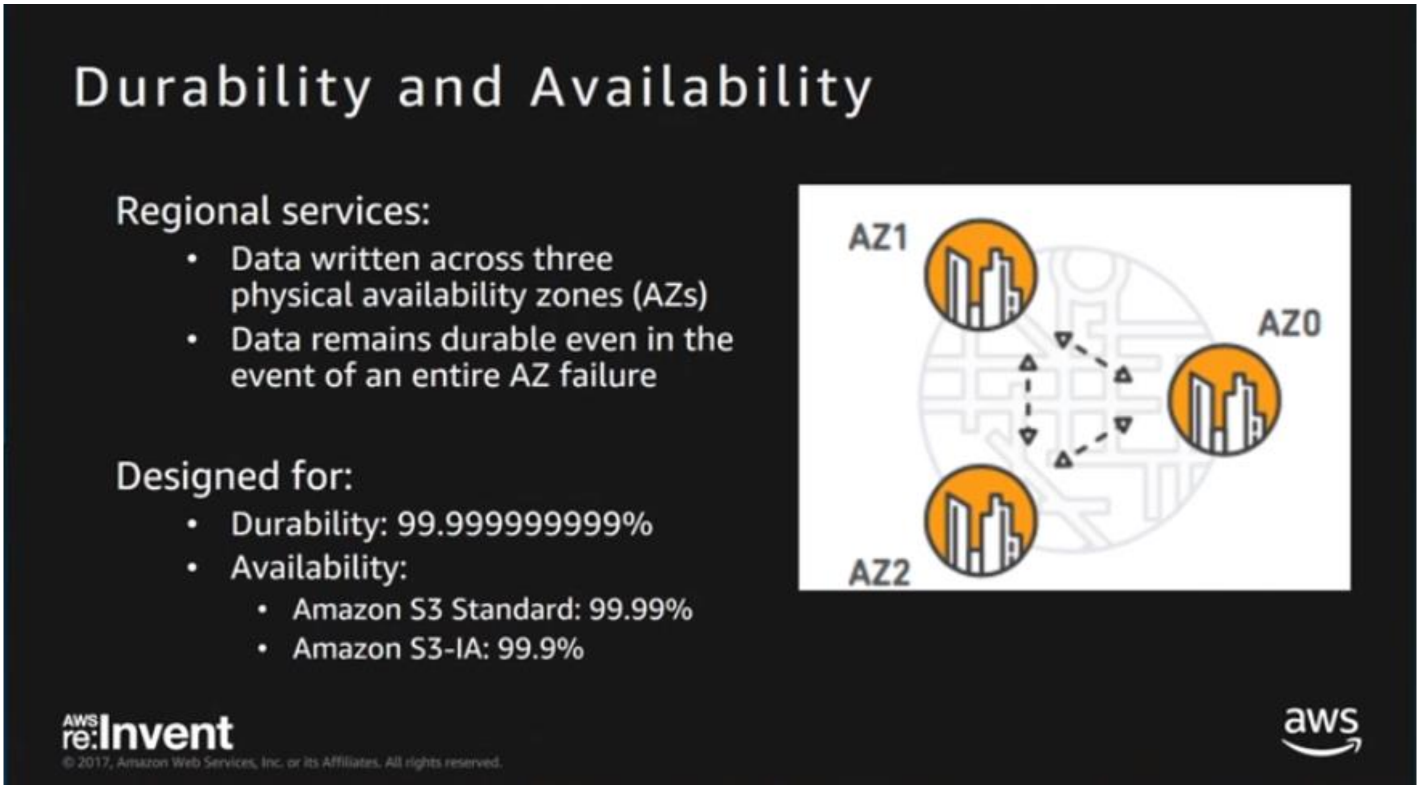
- Durable, Available, and Scalable
- Security and Compliance
- Query In Place
- Flexible Management
- Ecosystem

Amazon S3 is designed to be a durable object store with 11 nines of availability, you can also scale to Petabytes and Exabytes using S3. You have 3 different ways to encrypt your objects in S3, you can use server side managed by you the customer where you provide the key, or use the AWS KMS service to encrypt your objects. We also have compliance offerings like Federand, HIPAA, or PCI compliance using S3 storage. For Query In Place, we have Amazon Athena and Amazon Redshift Spectrum that both allow you to analyze data using SQL across S3 and taking in data from your data warehouse without requiring you to extract or load that data in any other service. We also have ***Flexible Mgt tools*** for better storage visibility, monitor, set alarms. Our Ecosystem consists of 3rd party services and in-house services.



Choice of Storage Classes

Amazon S3 Standard	Amazon S3 Standard–Infrequent Access	Amazon Glacier
Active data	Infrequently accessed data	Archive data
Milliseconds	Milliseconds	Minutes to hours
From 2.1¢-GB/mo.	1.25¢-GB/mo.	0.4¢-GB/mo.



Durability and Availability

Regional services:

- Data written across three physical availability zones (AZs)
- Data remains durable even in the event of an entire AZ failure

Designed for:

- Durability: 99.999999999%
- Availability:
  - Amazon S3 Standard: 99.99%
  - Amazon S3-IA: 99.9%

Storage Management

Object Tags

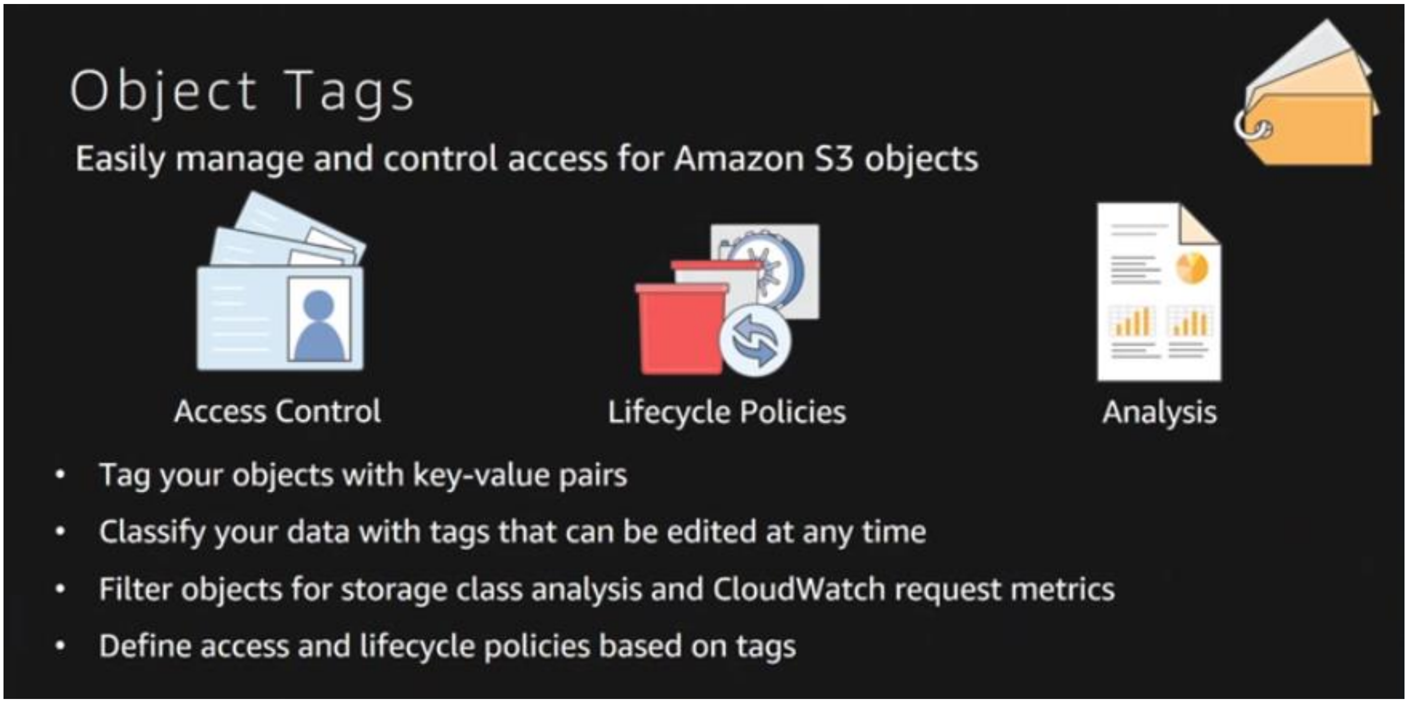
Lifecycle Management

Storage Class Analysis

Amazon S3 Inventory

Storage Management

- Cross-Region Replication
- Lifecycle Policies
- Event Notifications
- Object Tags
- Amazon CloudWatch Request Metrics
- Amazon S3 Inventory
- Storage Class Analysis
- AWS CloudTrail Data Events



Object Tags

Easily manage and control access for Amazon S3 objects

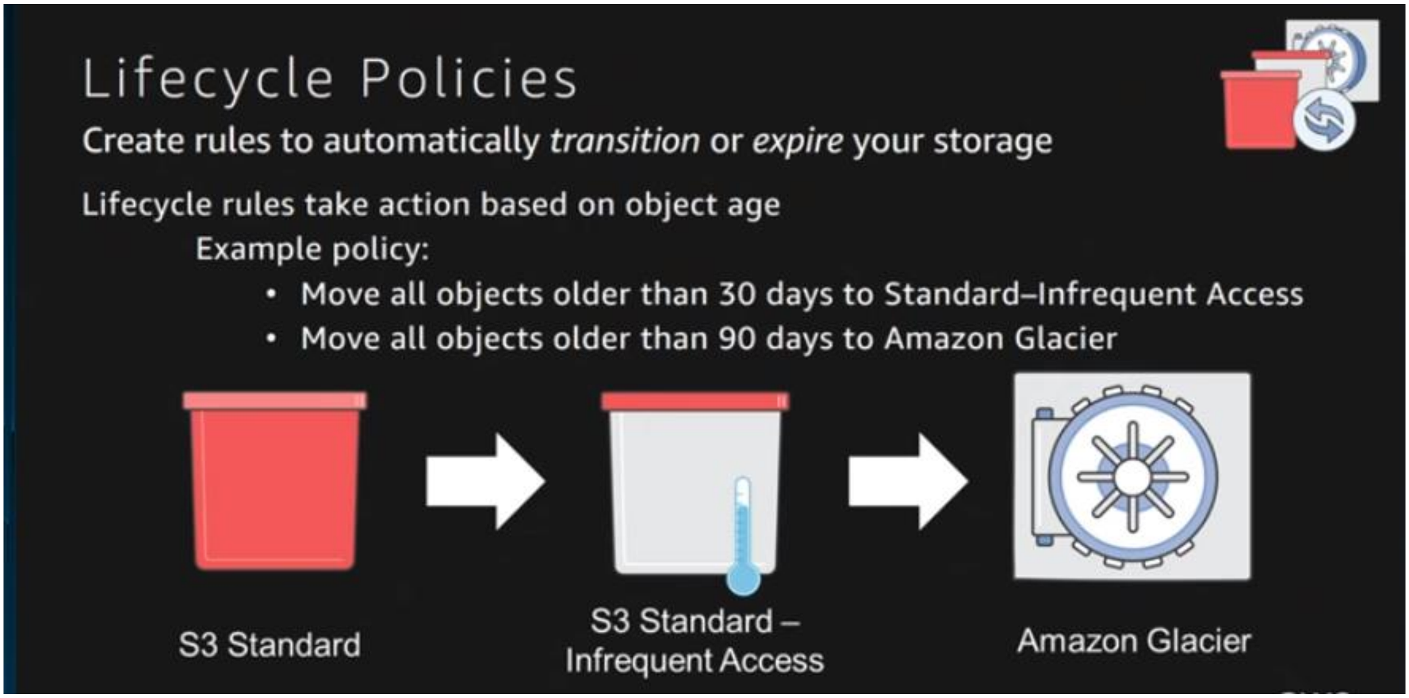
Access Control

Lifecycle Policies

Analysis

- Tag your objects with key-value pairs
- Classify your data with tags that can be edited at any time
- Filter objects for storage class analysis and CloudWatch request metrics
- Define access and lifecycle policies based on tags

You can add up to 10 tags per object stored in S3, this can be used for several analytics use cases like changing the tags to the S3 objects to control access over time without changing the data. Bucket policies allows you to set access policies and IAM User Roles policies based on the ‘what tagged objects should this user have access to?’. You can write Lifecycle policies (to transition or expire storage) per bucket or prefix level depending on how granular you want to control your lifecycle policies. You can now also use tags for this as well by attaching tags to objects with a specific lifecycle policy so that only those tagged objects will be affected.



Lifecycle Policies

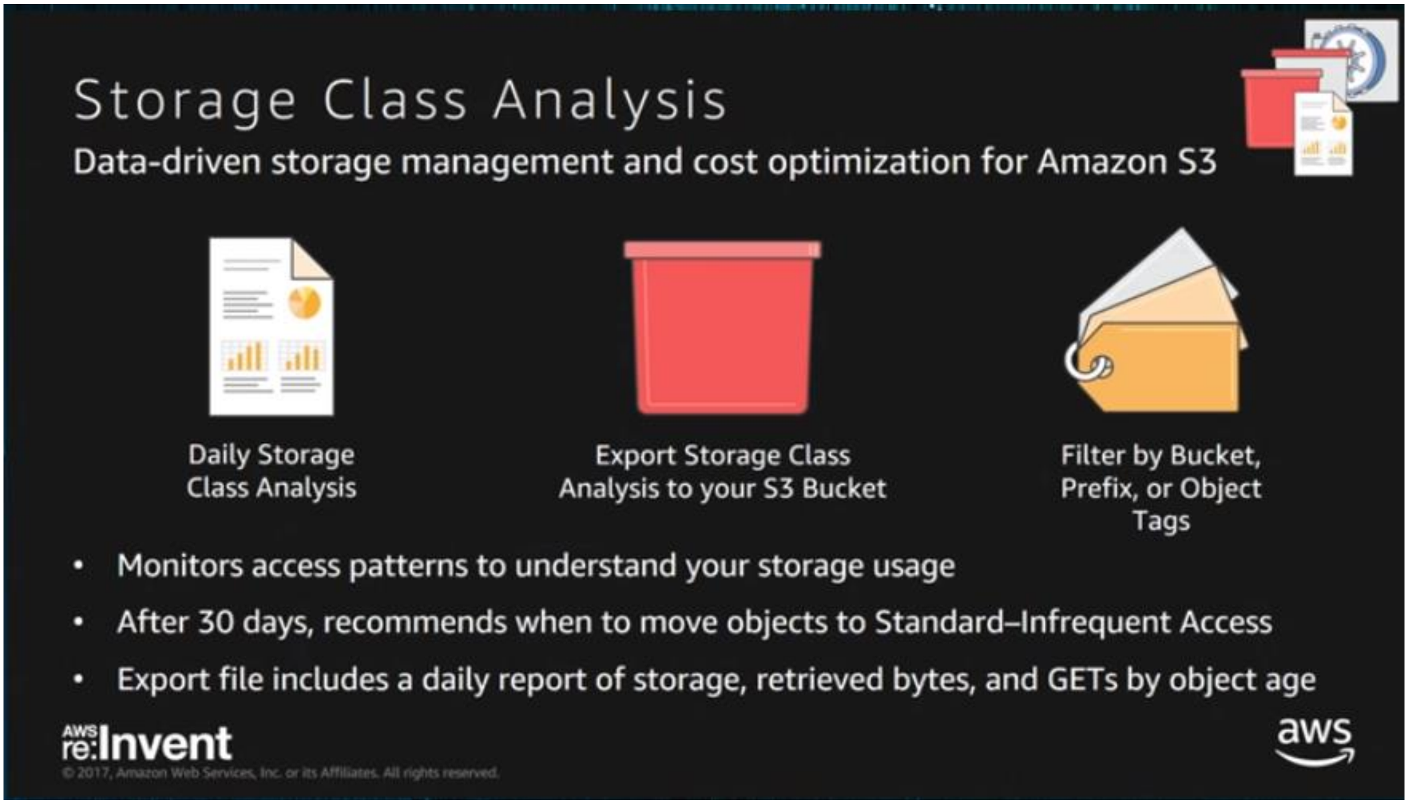
Create rules to automatically *transition* or *expire* your storage

Lifecycle rules take action based on object age

Example policy:

- Move all objects older than 30 days to Standard–Infrequent Access
- Move all objects older than 90 days to Amazon Glacier

AlertLogic is using tags on S3 objects in a very clever way, they have a number of N customers stored in a single S3 bucket. They can have 950 customers worth of data stored in a single S3 bucket, they then use tags to keep track of which customer data is a particular object and when was that data originally created. They used the ***creationDate*** and ***lastSeenDate*** as tags on objects and have policies that move the data to long term storage options when it hasn’t been accessed in say ‘3 months’.



Storage Class Analysis

Data-driven storage management and cost optimization for Amazon S3

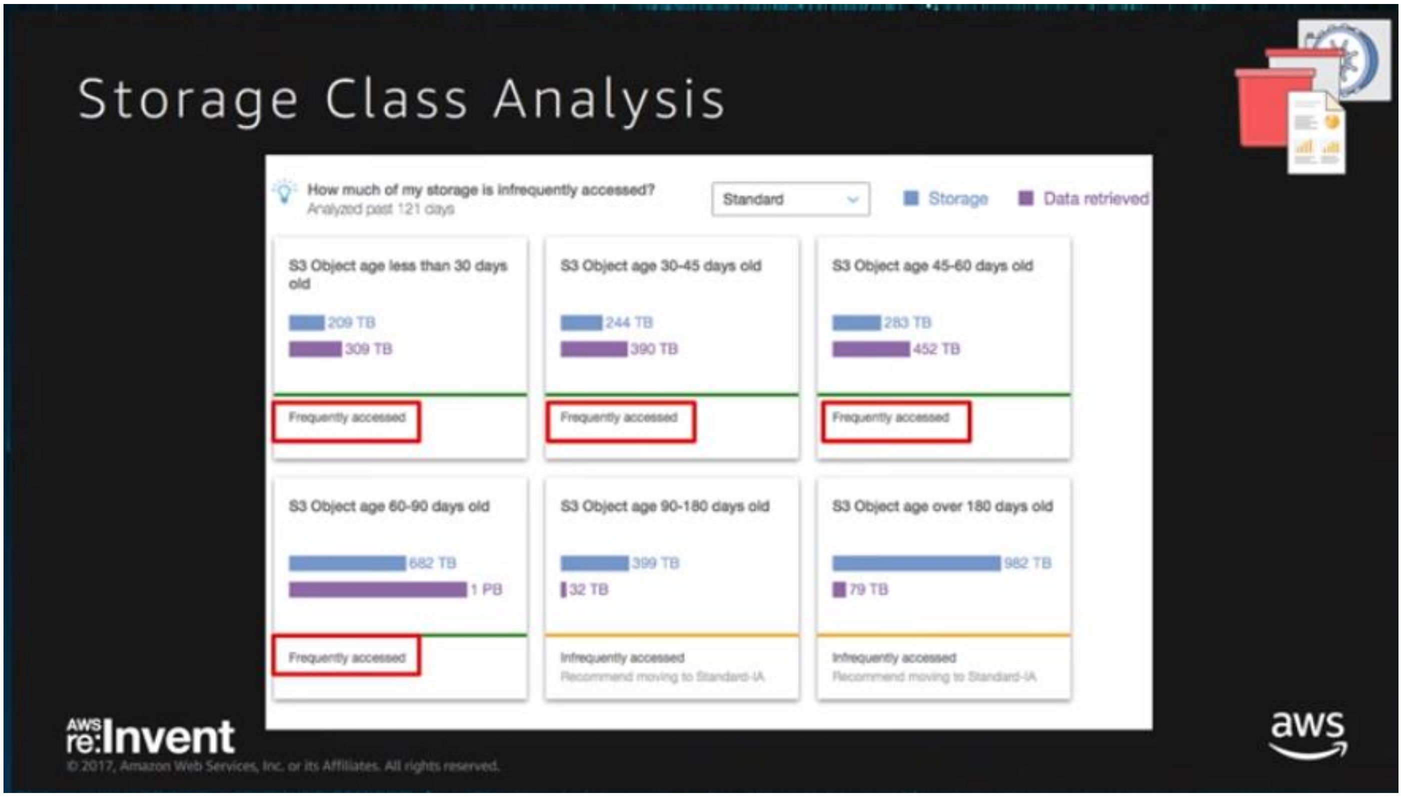
Daily Storage Class Analysis

Export Storage Class Analysis to your S3 Bucket

Filter by Bucket, Prefix, or Object Tags

- Monitors access patterns to understand your storage usage
- After 30 days, recommends when to move objects to Standard–Infrequent Access
- Export file includes a daily report of storage, retrieved bytes, and GETs by object age

How do we know that 30 days is the number of days to move to infrequent access and 90 days is the number of days to move to longer term storage like Athena for S3 objects?



Storage Class Analysis

How much of my storage is infrequently accessed?
Analyzed past 121 days

Standard

■ Storage ■ Data retrieved

S3 Object age less than 30 days old
209 TB
309 TB
Frequently accessed

S3 Object age 30-45 days old
244 TB
390 TB
Frequently accessed

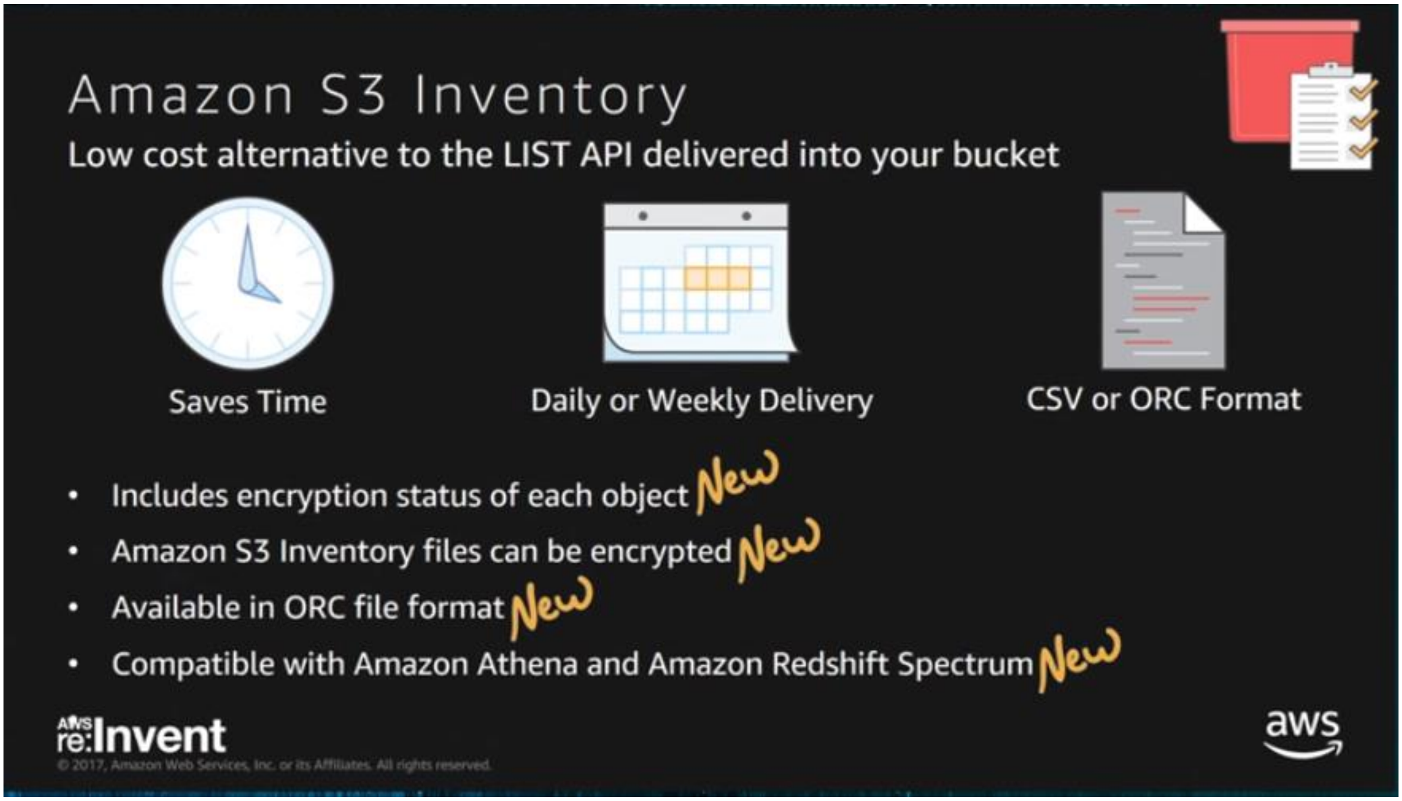
S3 Object age 45-60 days old
283 TB
452 TB
Frequently accessed

S3 Object age 60-90 days old
682 TB
1 PB
Frequently accessed

S3 Object age 90-180 days old
399 TB
32 TB
Infrequently accessed
Recommend moving to Standard-IA

S3 Object age over 180 days old
982 TB
79 TB
Infrequently accessed
Recommend moving to Standard-IA

This is what you will see in the S3 console when you set this up.



Amazon S3 Inventory

Low cost alternative to the LIST API delivered into your bucket

Saves Time

Daily or Weekly Delivery

CSV or ORC Format

- Includes encryption status of each object New
- Amazon S3 Inventory files can be encrypted New
- Available in ORC file format New
- Compatible with Amazon Athena and Amazon Redshift Spectrum New

This is a way to list out all the objects in an S3 bucket or an S3 prefix. This is a feature that you configure to have a list of S3 buckets and their associated metadata delivered to you, you can then do advanced analytics on the list.

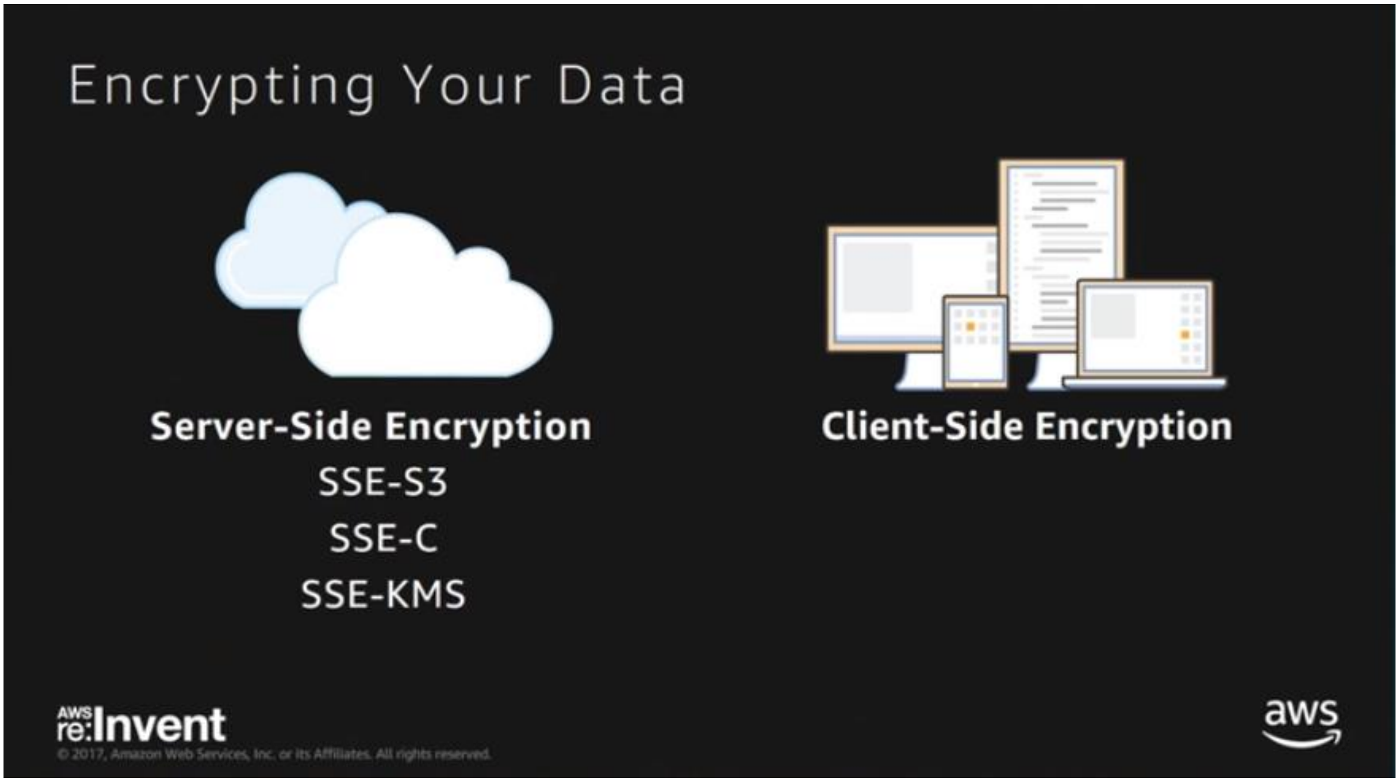
Security

Data Encryption

Access Controls

AWS CloudTrail Data Events

Amazon Macie



Encrypting Your Data

Server-Side Encryption

SSE-S3

SSE-C

SSE-KMS

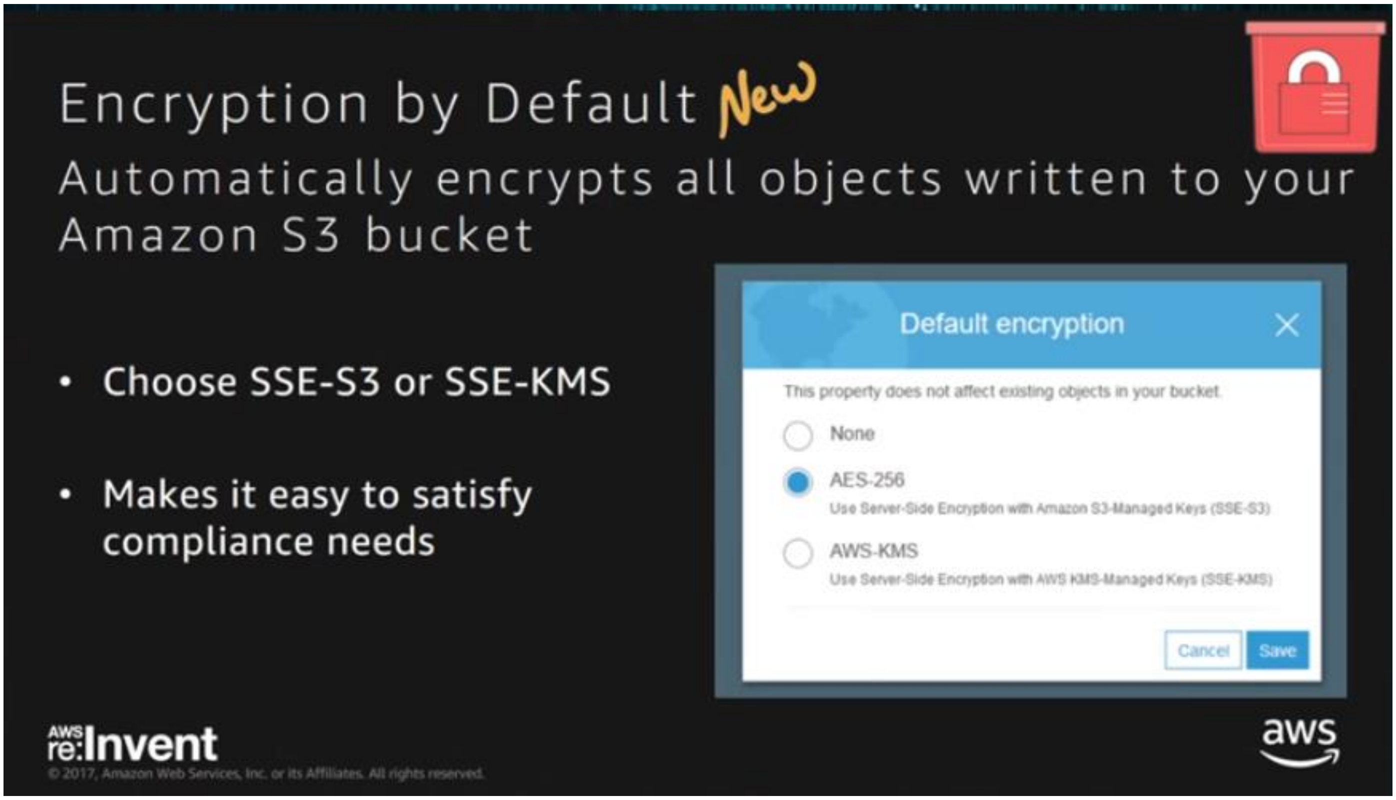
Client-Side Encryption

Client-Side encryption is when you encrypt the data object before uploading it into S3, but you can also let AWS do the data object encryption for you on the ***Server-Side*** before storing it in S3.

SSE-S3 means that AWS manages the data and master keys, you just specify whether you want the data encrypted when they are uploaded. This is encryption at REST so keep in mind that when an authorized request comes in to read your object, AWS is going to decrypt the object and then deliver it to the requesting party.

SSE-C is server-side encryption customer, it refers to you providing and managing the customer key, so with your request for retrieval you will have to specify the encryption key and AWS will use the key that you provide to encrypt or decrypt the object depending on if it is a PUT or a GET.

SSE-KMS is the Key Management Service, with KMS there is a data key used in S3 to encrypt your data and there is a master key that is retained in the KMS service. Customers need approval and authorization before they can read the objects from S3, they also need the permissions from the KMS service before they can get access to the master key they need to decrypt the object when they get it.



Encryption by Default New

Automatically encrypts all objects written to your Amazon S3 bucket

- Choose SSE-S3 or SSE-KMS
- Makes it easy to satisfy compliance needs

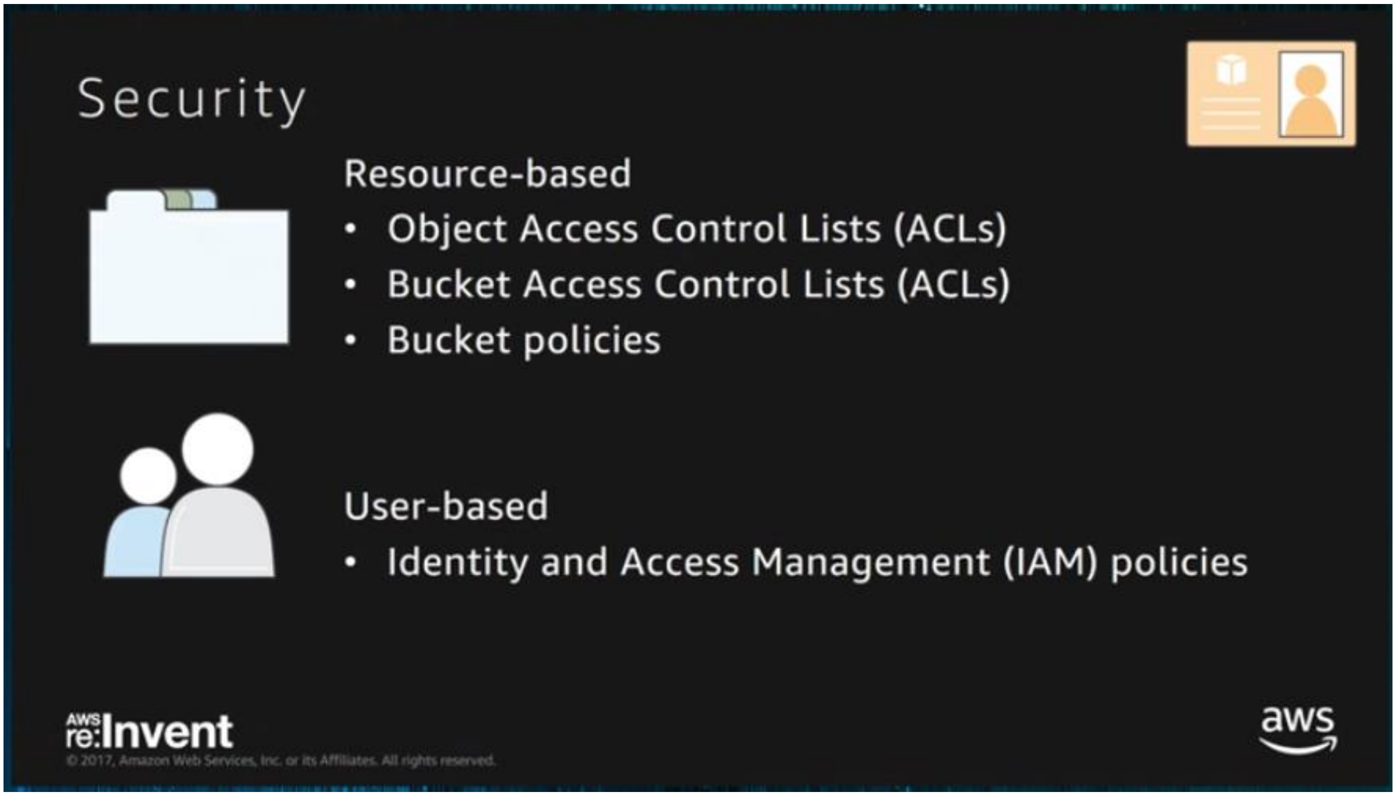
Default encryption

This property does not affect existing objects in your bucket.

○ None

● AES-256
Use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3)

○ AWS-KMS
Use Server-Side Encryption with AWS KMS-Managed Keys (SSE-KMS)



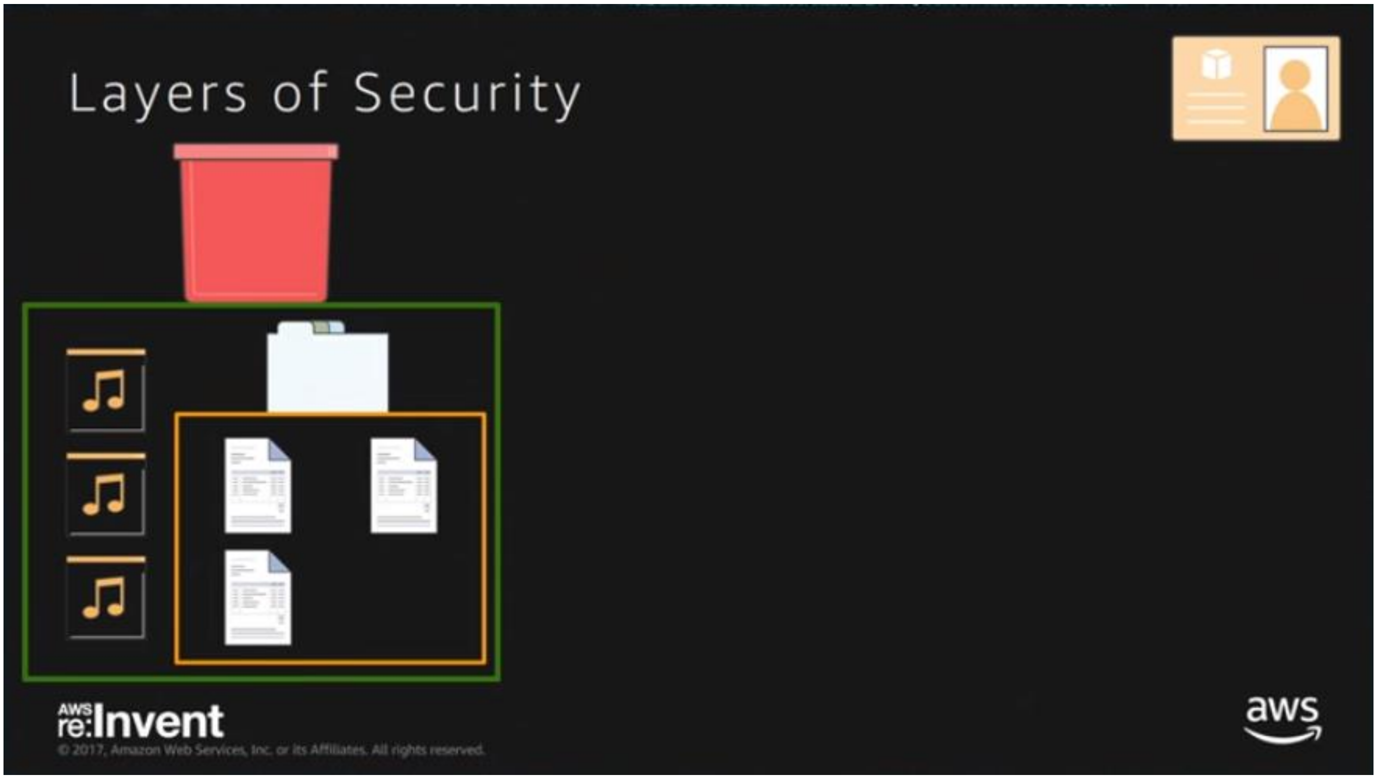
Security

Resource-based
- Object Access Control Lists (ACLs)
- Bucket Access Control Lists (ACLs)
- Bucket policies

User-based
- Identity and Access Management (IAM) policies

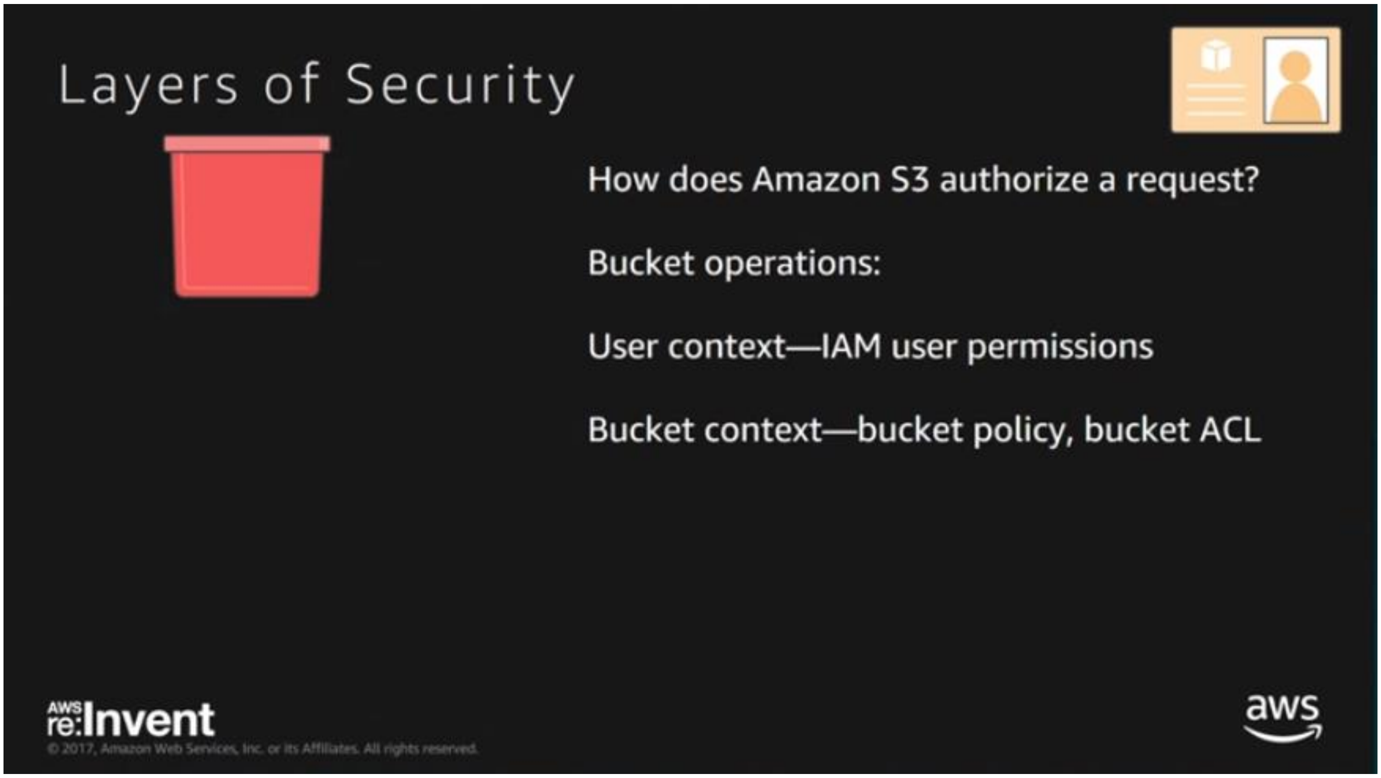
AWS re:Invent
aws

Above are the few different options and levels that you can manage for doing security on your S3 buckets. Let us now see how to use these options and how a request to your S3 bucket is authorized. ***Bucket ACLs*** are best recommended to be used for a S3 bucket logging group like for log delivery from Amazon S3 to write those logs.



Layers of Security

When you add a prefix to your object key name like a '/' for organizing your folder, it will look like a folder as above.



Layers of Security

How does Amazon S3 authorize a request?

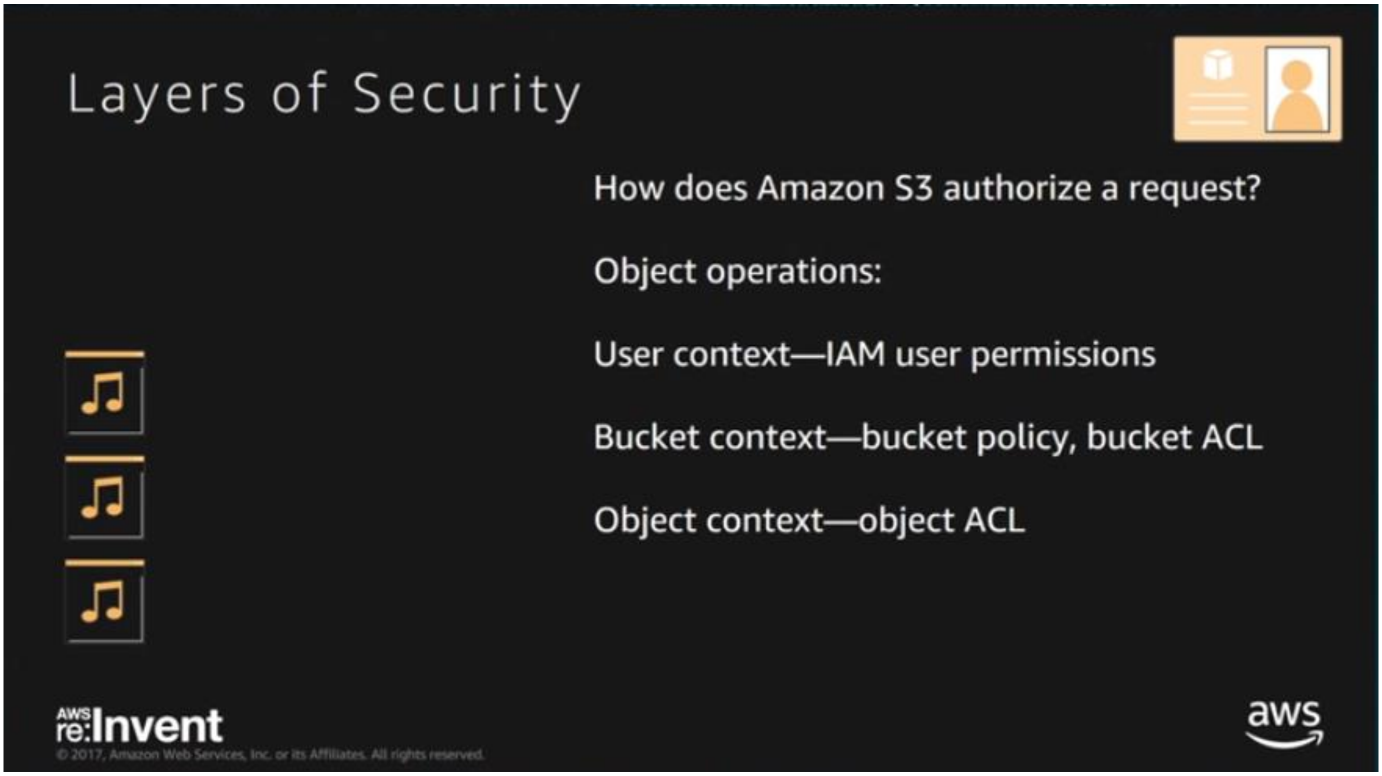
Bucket operations:

User context—IAM user permissions

Bucket context—bucket policy, bucket ACL

AWS re:Invent
aws

If the IAM Role trying to access your S3 bucket does not have access to S3 or does not have access to your account, then the request is denied immediately. If an account is trying to access your S3 bucket, then we will look at the Bucket context's policy and bucket ACL to determine access.



Layers of Security

How does Amazon S3 authorize a request?

Object operations:

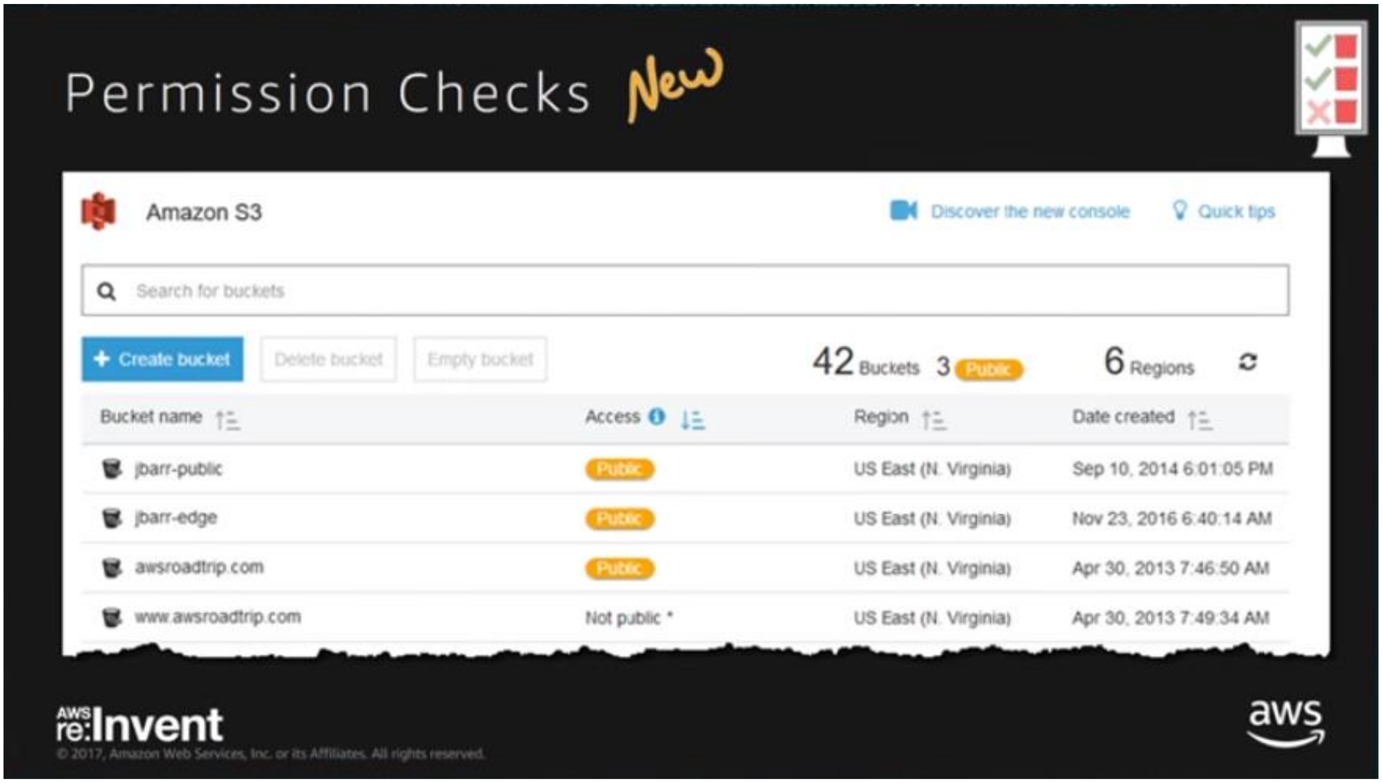
User context—IAM user permissions

Bucket context—bucket policy, bucket ACL

Object context—object ACL

re:Invent
aws

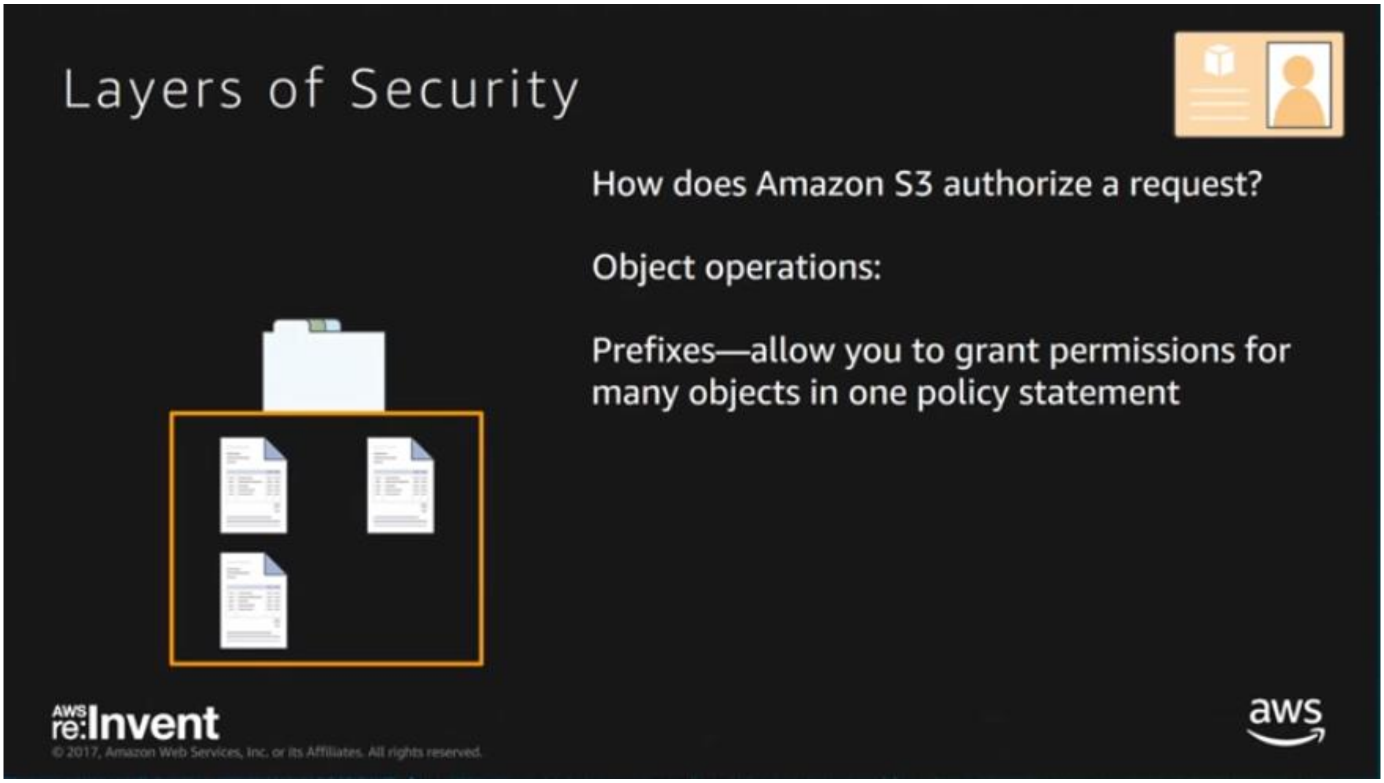
There is an additional layer when using the Object level, we will look at the object context and the object ACL.



Permission Checks New

42 Buckets 3 Public 6 Regions

Bucket name	Access	Region	Date created
jbarr-public	Public	US East (N. Virginia)	Sep 10, 2014 6:01:05 PM
jbarr-edge	Public	US East (N. Virginia)	Nov 23, 2016 6:40:14 AM
awsroadtrip.com	Public	US East (N. Virginia)	Apr 30, 2013 7:46:50 AM
www.awsroadtrip.com	Not public *	US East (N. Virginia)	Apr 30, 2013 7:49:34 AM



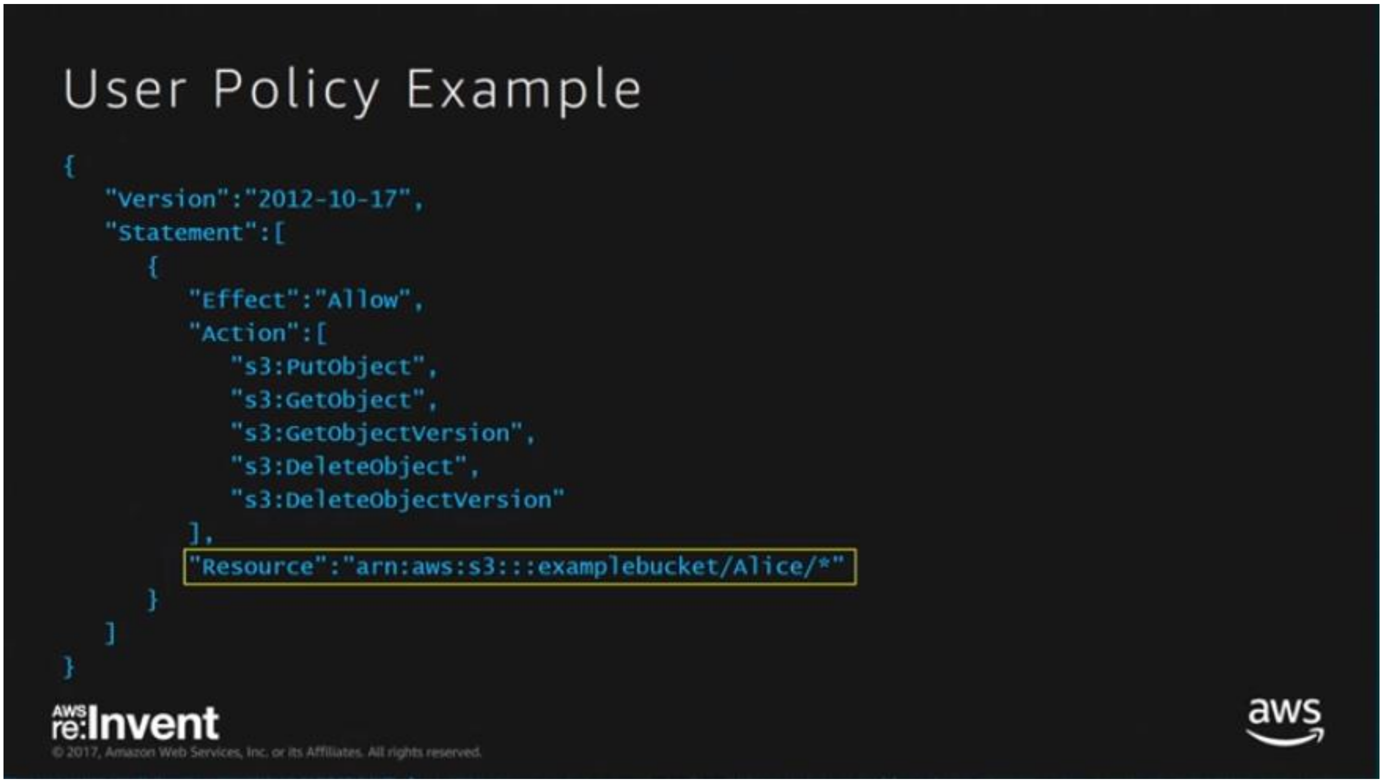
Layers of Security

How does Amazon S3 authorize a request?

Object operations:

Prefixes—allow you to grant permissions for many objects in one policy statement

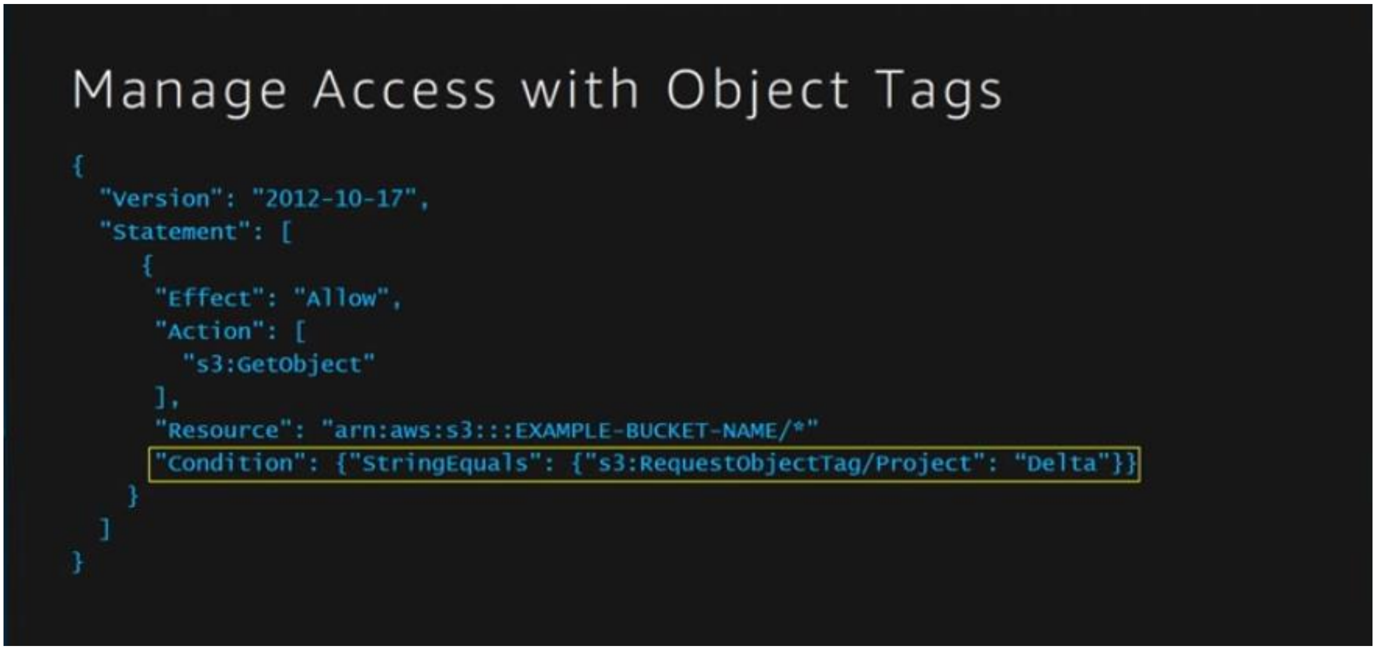
When you add a prefix to an object, it gets stored in a folder with other objects that have that prefix as above. We can then create and write S3 bucket policies for the folder as a whole and then for individual subfolders if needed.



User Policy Example

```
{
   "Version":"2012-10-17",
   "Statement":[
      {
         "Effect":"Allow",
         "Action":[
            "s3:PutObject",
            "s3:GetObject",
            "s3:GetObjectVersion",
            "s3:DeleteObject",
            "s3:DeleteObjectVersion"
         ],
         "Resource":"arn:aws:s3:::examplebucket/Alice/*"
      }
   ]
}
```

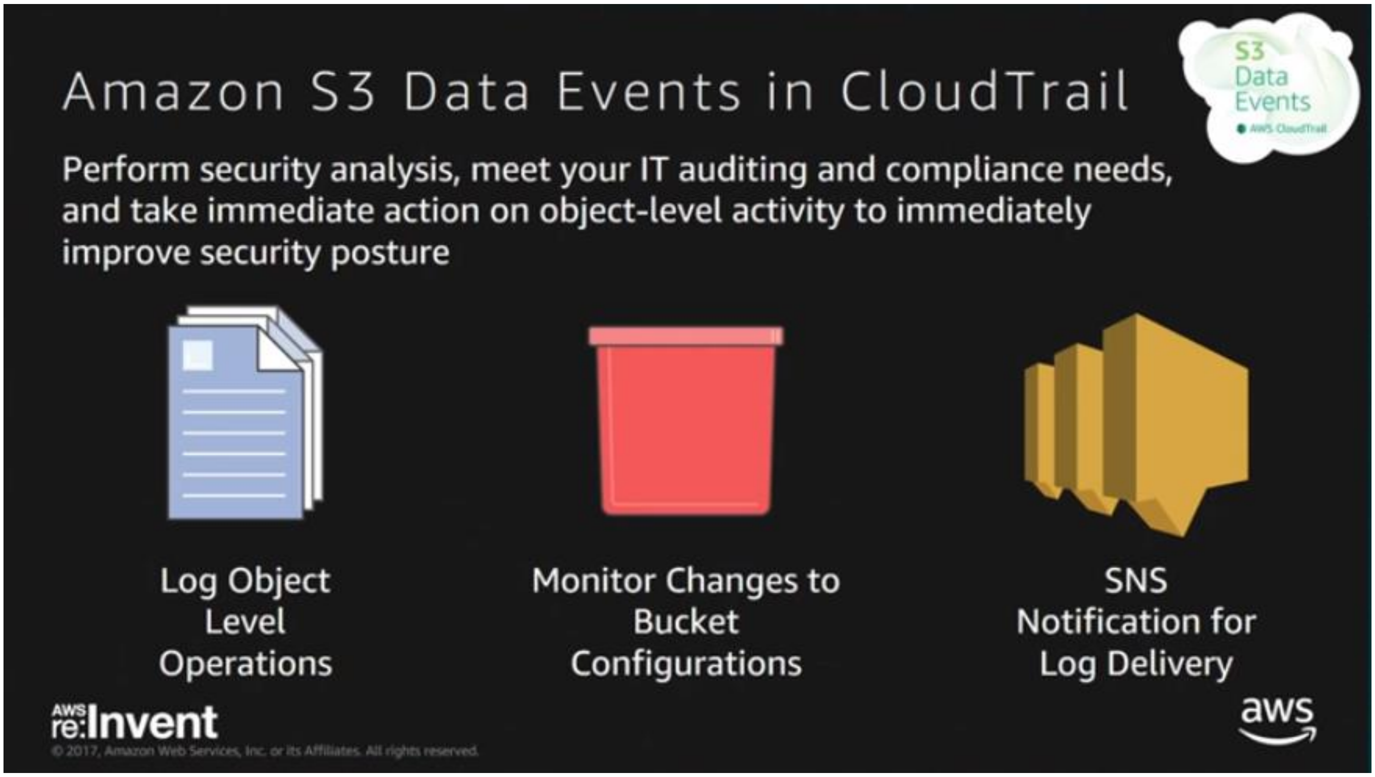
To manage a user at the ***Prefix level***, in this case the prefix ***examplebucket/***Alice. This User policy will grant PUT, GET, DELETE permissions to a particular user called Alice. The S3 Resource allowed is for the objects within the folder in examplebucket/Alice/*, the user Alice can now interact with that particular subset of S3 objects.



Manage Access with Object Tags

```
{
  "Version": "2012-10-17",
  "Statement": [
     {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject"
      ],
      "Resource": "arn:aws:s3:::EXAMPLE-BUCKET-NAME/*"
      "Condition": {"StringEquals": {"s3:RequestObjectTag/Project": "Delta"}}
    }
  ]
}
```

Next, we will talk about ***Object Tags*** on your S3 objects level. Here we are adding a ***Condition*** attribute to the policy, we want a user to be able to get all the Objects that are associated with the Project called ***Delta***. Project delta might start very small and grow over time to a very large project, we just continue to tag all the S3 objects associated with the project so that permitted users can see and use them. When Project Delta ends, we can go ahead and just simply delete all the tags on the objects or even delete all the objects themselves in addition to all the user permissions associated to Project Delta.



Amazon S3 Data Events in CloudTrail

Perform security analysis, meet your IT auditing and compliance needs, and take immediate action on object-level activity to immediately improve security posture

Log Object Level Operations

Monitor Changes to Bucket Configurations

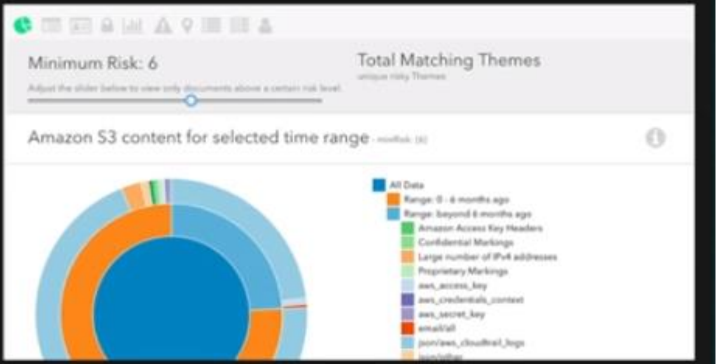
SNS Notification for Log Delivery

Data Events in CloudTrail can be turned on to start looking at object level activity and recording the GET, PUT, DELETE events to the buckets. You can automate alerts using this approach. We can set up lambda functions to revert unwanted events back to their intended state.

Amazon Macie New

A security service that uses machine learning to automatically discover, classify, and protect sensitive data in AWS

- Recognizes sensitive data
- Continuously monitors data access
- Provides dashboards and alerts



re:Invent

aws

Data Protection

Cross-Region Replication

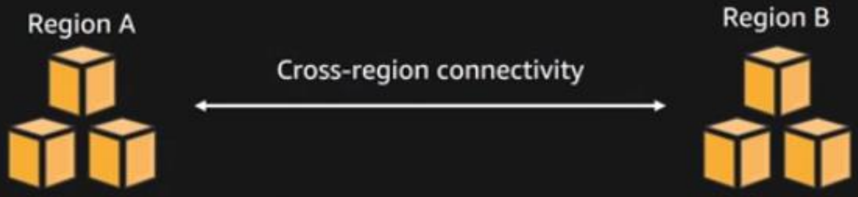
Versioning

Multi-Factor Authentication

Cross-Region Replication

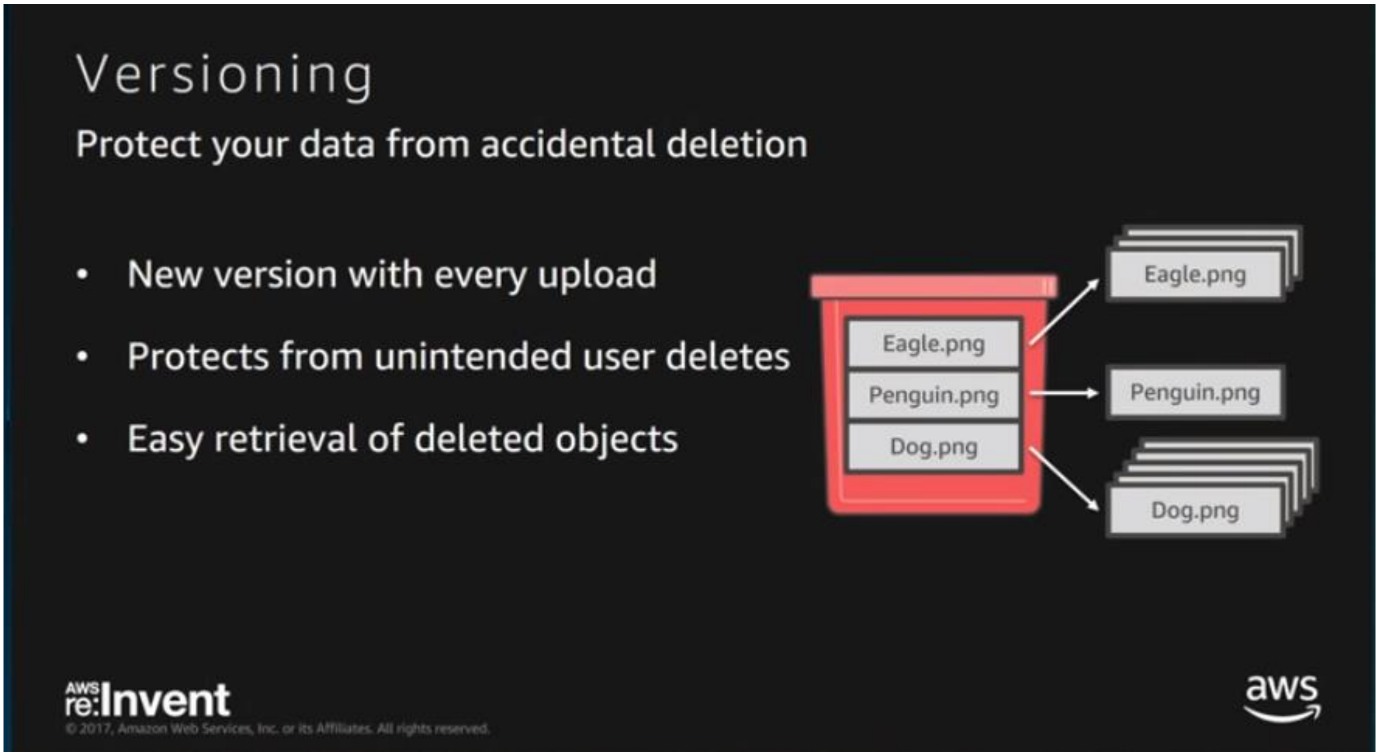
Automatically replicate data to any other AWS Region

- Replicate by object, bucket, or prefix
- Support for SSE-KMS encrypted objects New
- Ownership overwrite New
  - Change the object owner in the destination region



re:Invent

aws



Versioning

Protect your data from accidental deletion

- New version with every upload
- Protects from unintended user deletes
- Easy retrieval of deleted objects

Versioning protects your data from accidental deletion, by creating new versions of your object every time you overwrite a particular object key like the version number. You can then go back and use the particular object key name you want to get that particular version.

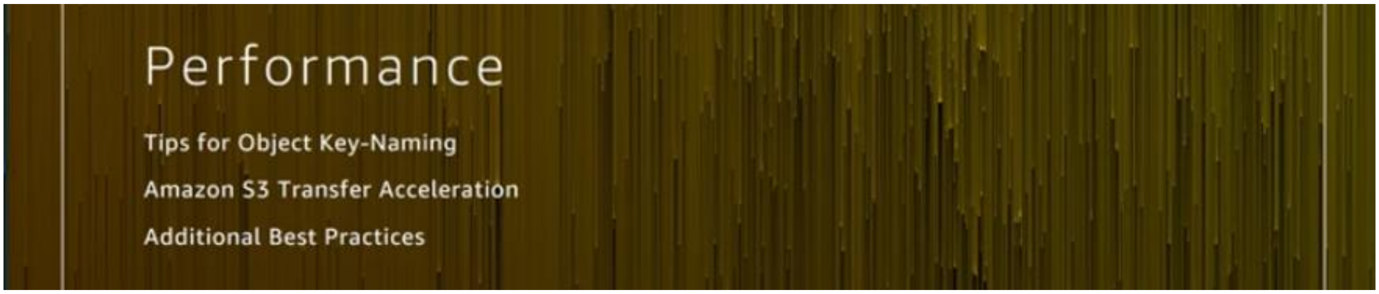
Multi-Factor Authentication (MFA)

Adds another layer of protection for deleting objects or changing the versioning state of the bucket

Requires two forms of authentication:

Your security credentials

Unique code from an approved authentication device

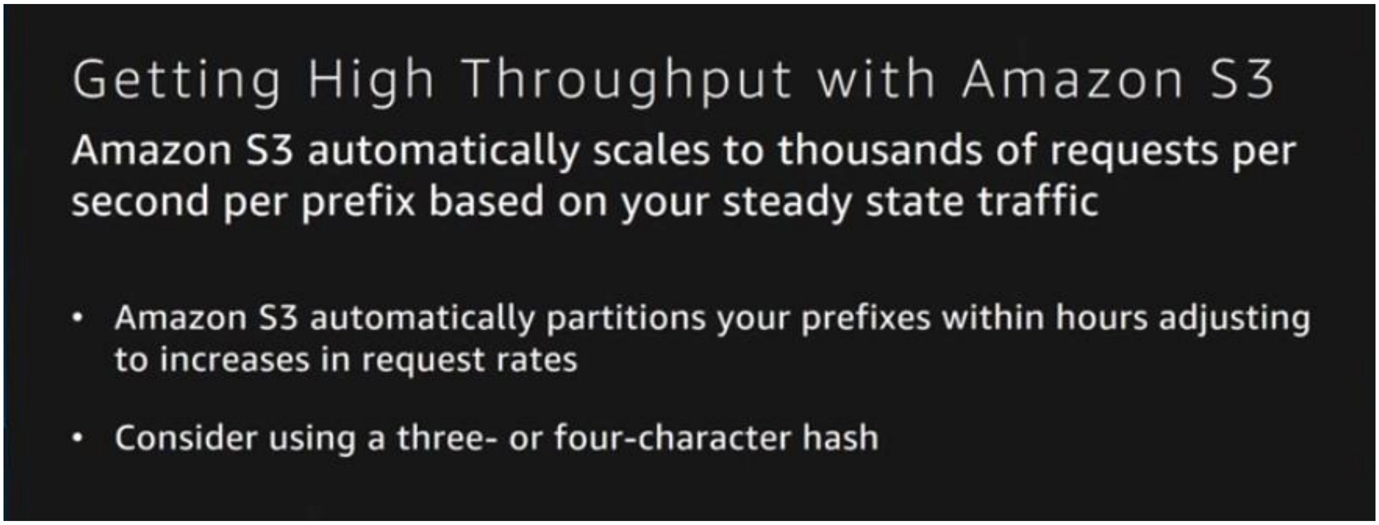


Performance

Tips for Object Key-Naming

Amazon S3 Transfer Acceleration

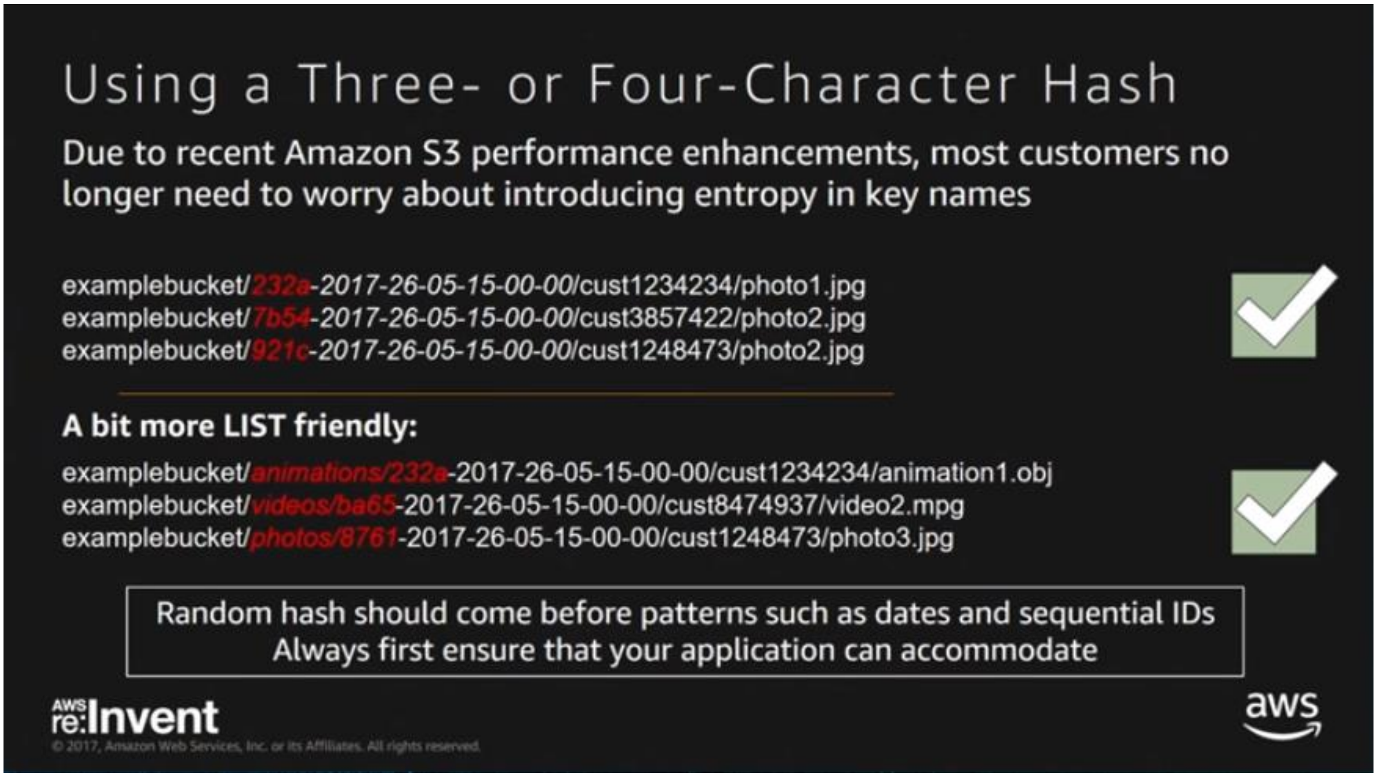
Additional Best Practices



Getting High Throughput with Amazon S3

Amazon S3 automatically scales to thousands of requests per second per prefix based on your steady state traffic

- Amazon S3 automatically partitions your prefixes within hours adjusting to increases in request rates
- Consider using a three- or four-character hash



Using a Three- or Four-Character Hash

Due to recent Amazon S3 performance enhancements, most customers no longer need to worry about introducing entropy in key names

examplebucket/*232a-2017-26-05-15-00-00*/cust1234234/photo1.jpg
examplebucket/*7b54-2017-26-05-15-00-00*/cust3857422/photo2.jpg
examplebucket/*921c-2017-26-05-15-00-00*/cust1248473/photo2.jpg ✓

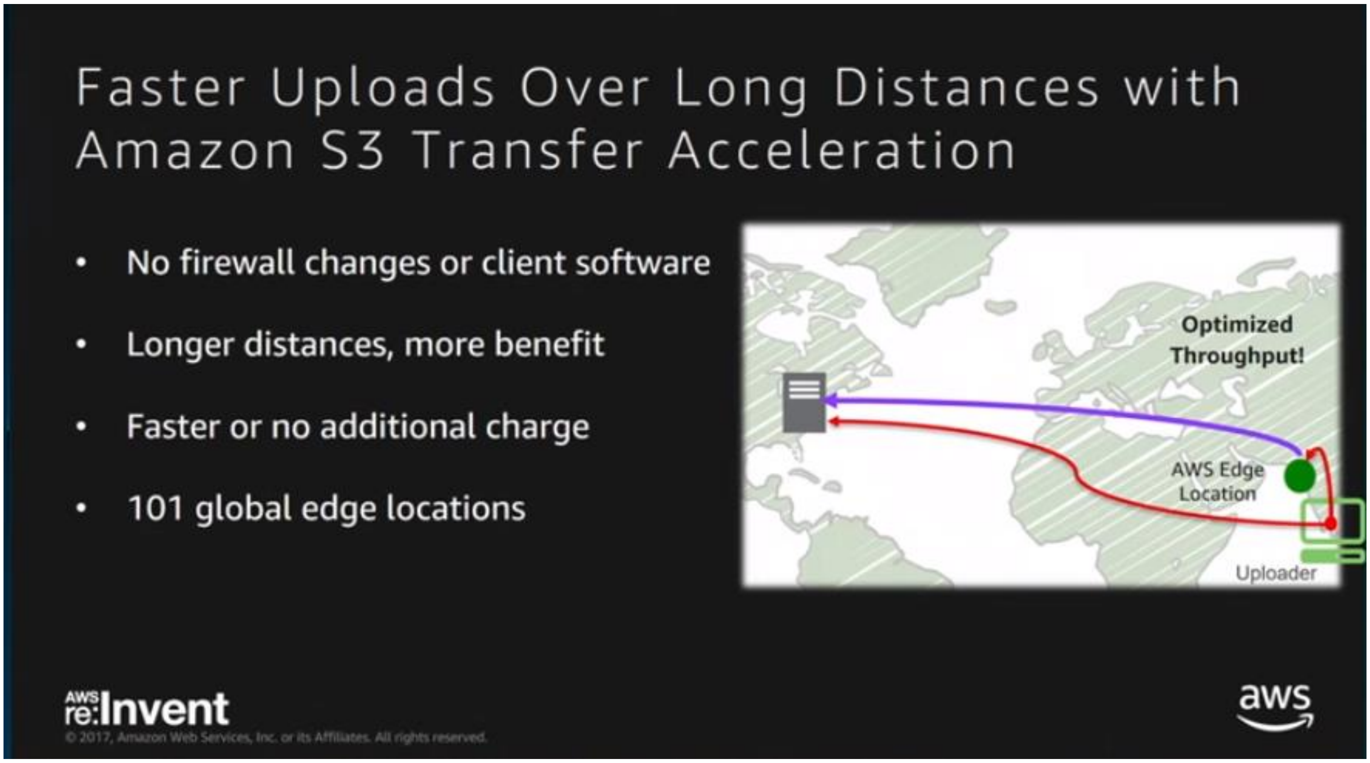
A bit more LIST friendly:

examplebucket/*animations/232a*-2017-26-05-15-00-00/cust1234234/animation1.obj
examplebucket/*videos/ba65*-2017-26-05-15-00-00/cust8474937/video2.mpg
examplebucket/*photos/8761*-2017-26-05-15-00-00/cust1248473/photo3.jpg ✓

Random hash should come before patterns such as dates and sequential IDs
Always first ensure that your application can accommodate

AWS re:Invent

This is how you can add the hash to your buckets. Object keys are stored in an index in an AWS region, and if you are constantly writing the same object keys repeatedly like for a year 2017 or a day 12, this will place all those objects close together within the same partition or index. So, if you are also doing a lot of reads on those objects, you might start to see some performance slowdowns. You can make this better by putting a ***hash*** in front of your object key names as above to add some randomness and spread the data around better. You can also add prefixes like the ***/animations/<hash>*** for specific buckets that might get a high traffic.



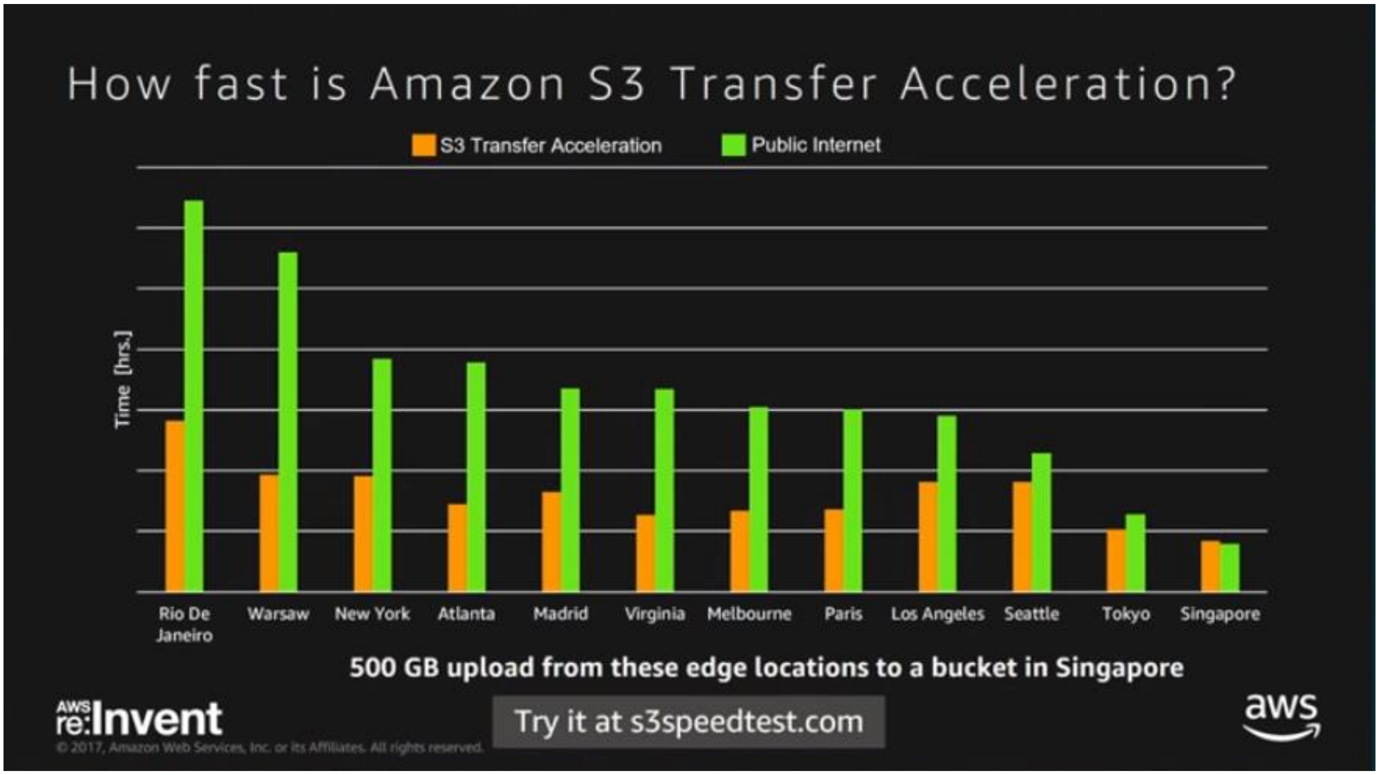
Faster Uploads Over Long Distances with Amazon S3 Transfer Acceleration

- No firewall changes or client software
- Longer distances, more benefit
- Faster or no additional charge
- 101 global edge locations

re:Invent

aws

You can specify that you want to use this feature for your S3 bucket



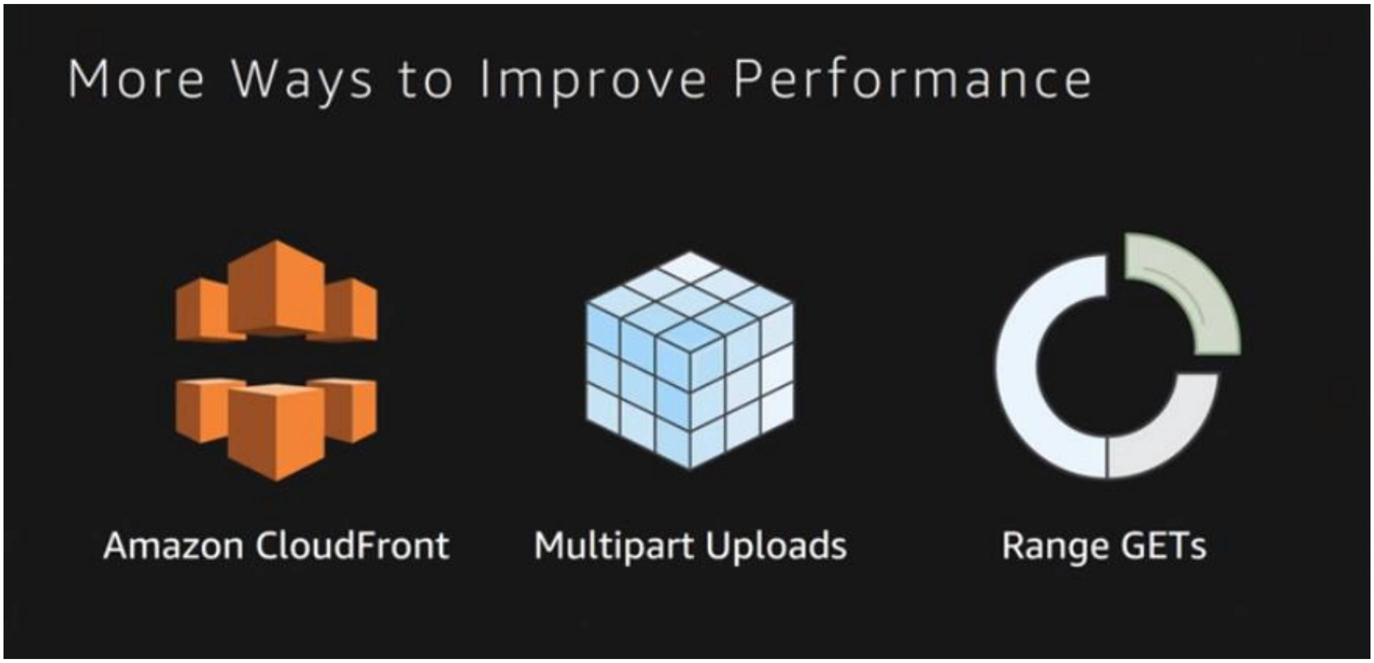
How fast is Amazon S3 Transfer Acceleration?

500 GB upload from these edge locations to a bucket in Singapore

Try it at s3speedtest.com

re:Invent

aws



More Ways to Improve Performance

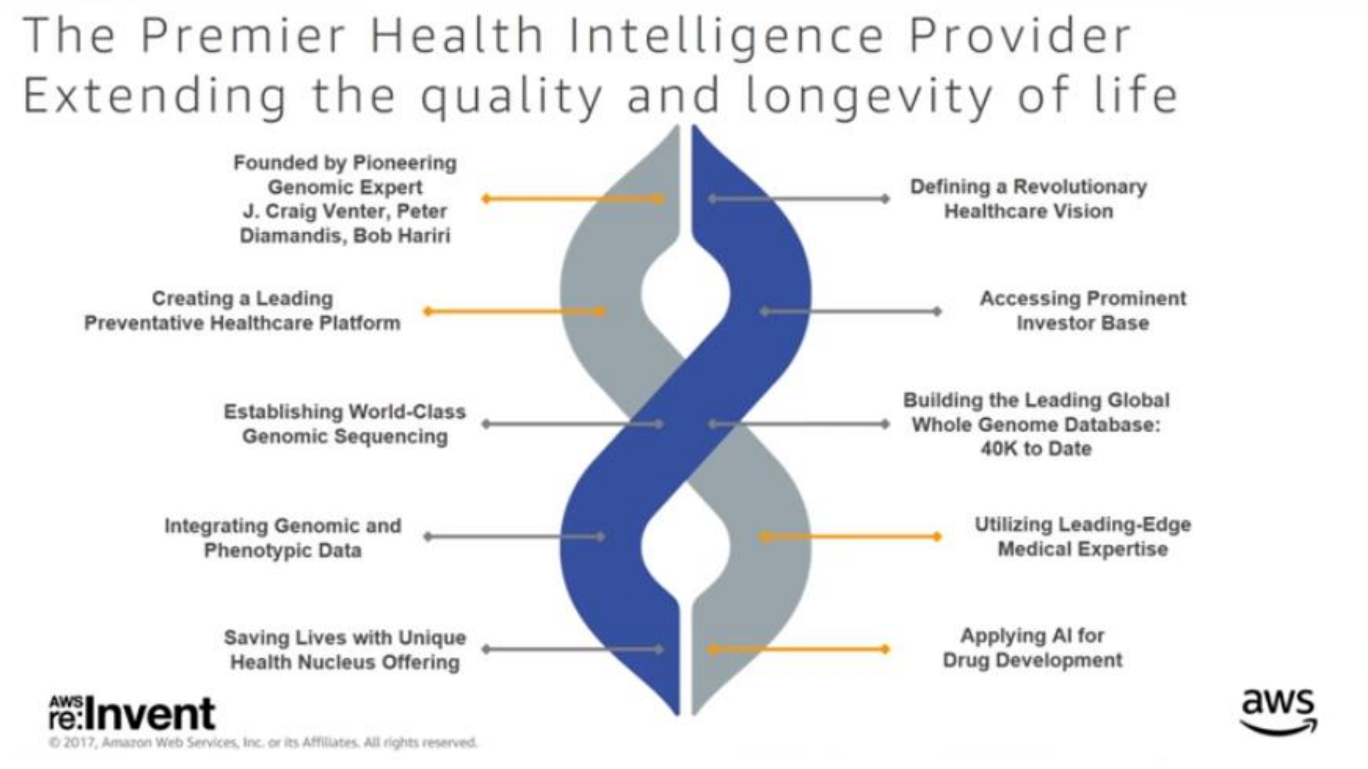
Amazon CloudFront

Multipart Uploads

Range GETs

Human Longevity Inc.

Katie Lamkin, Software Engineer



The Premier Health Intelligence Provider Extending the quality and longevity of life

Founded by Pioneering Genomic Expert J. Craig Venter, Peter Diamandis, Bob Hariri

Defining a Revolutionary Healthcare Vision

Creating a Leading Preventative Healthcare Platform

Accessing Prominent Investor Base

Establishing World-Class Genomic Sequencing

Building the Leading Global Whole Genome Database: 40K to Date

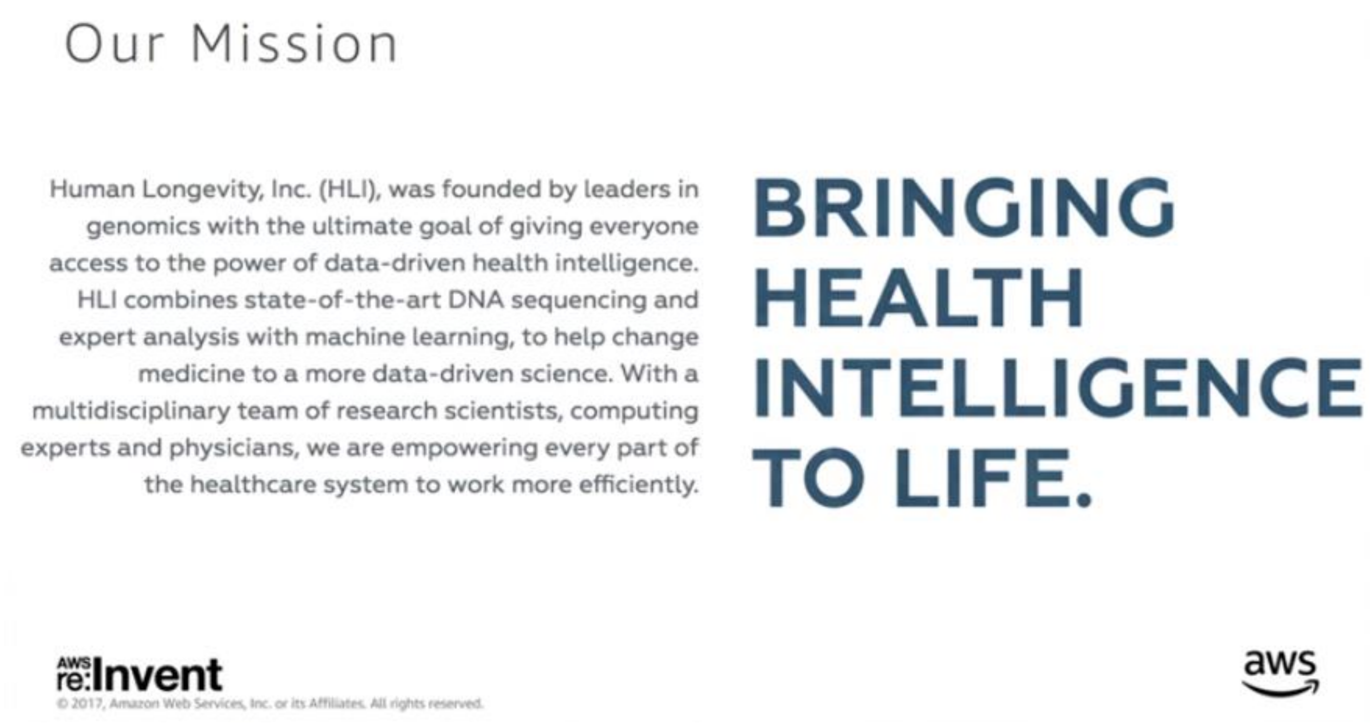
Integrating Genomic and Phenotypic Data

Utilizing Leading-Edge Medical Expertise

Saving Lives with Unique Health Nucleus Offering

Applying AI for Drug Development

Human longevity is a human genomics company founded by Craig Venter in 2013, it gathers quality genomic data as well as Phenotypic data at petabyte scale. They perform a whole genomic sequencing in order to make discoveries in part of the genome in addition with phenotypic data like a full body MRI, brain scan, CT scans, height weight, etc. This provides preventive healthcare.



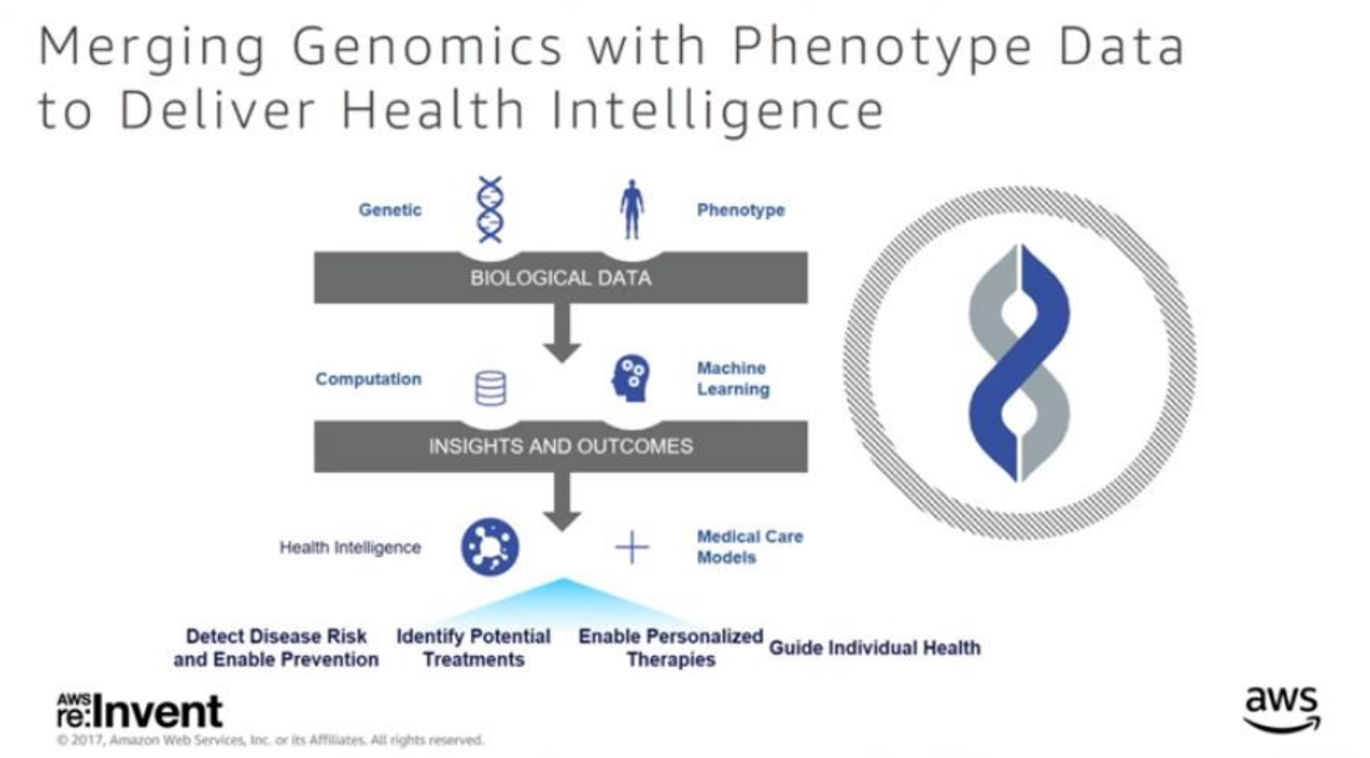
Our Mission

Human Longevity, Inc. (HLI), was founded by leaders in genomics with the ultimate goal of giving everyone access to the power of data-driven health intelligence. HLI combines state-of-the-art DNA sequencing and expert analysis with machine learning, to help change medicine to a more data-driven science. With a multidisciplinary team of research scientists, computing experts and physicians, we are empowering every part of the healthcare system to work more efficiently.

BRINGING HEALTH INTELLIGENCE TO LIFE.

re:Invent

aws



Merging Genomics with Phenotype Data to Deliver Health Intelligence

Genetic

Phenotype

BIOLOGICAL DATA

Computation

Machine Learning

INSIGHTS AND OUTCOMES

Health Intelligence

Medical Care Models

Detect Disease Risk and Enable Prevention

Identify Potential Treatments

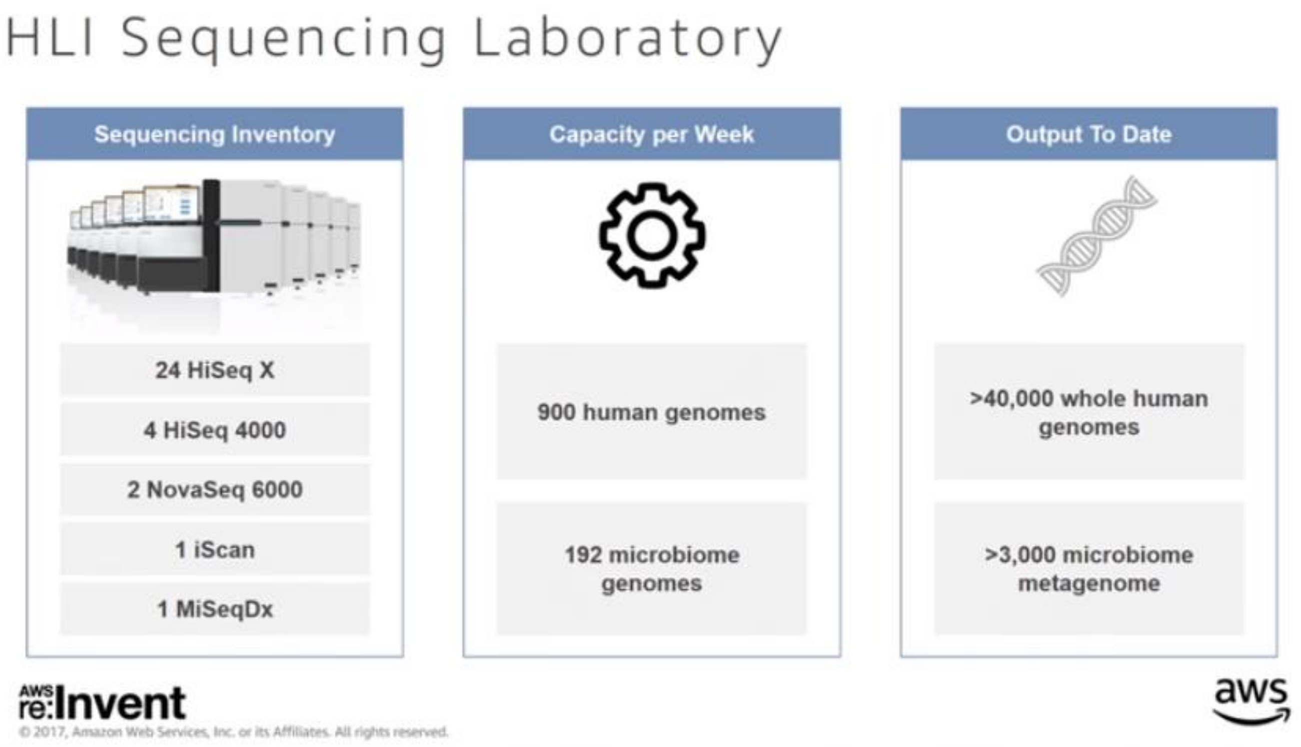
Enable Personalized Therapies

Guide Individual Health

re:Invent

aws

We then sequence their genome and do analysis on it that are given back to them in their ‘health intelligence’ report containing potential cancer or neurological risks, ancestry, etc.



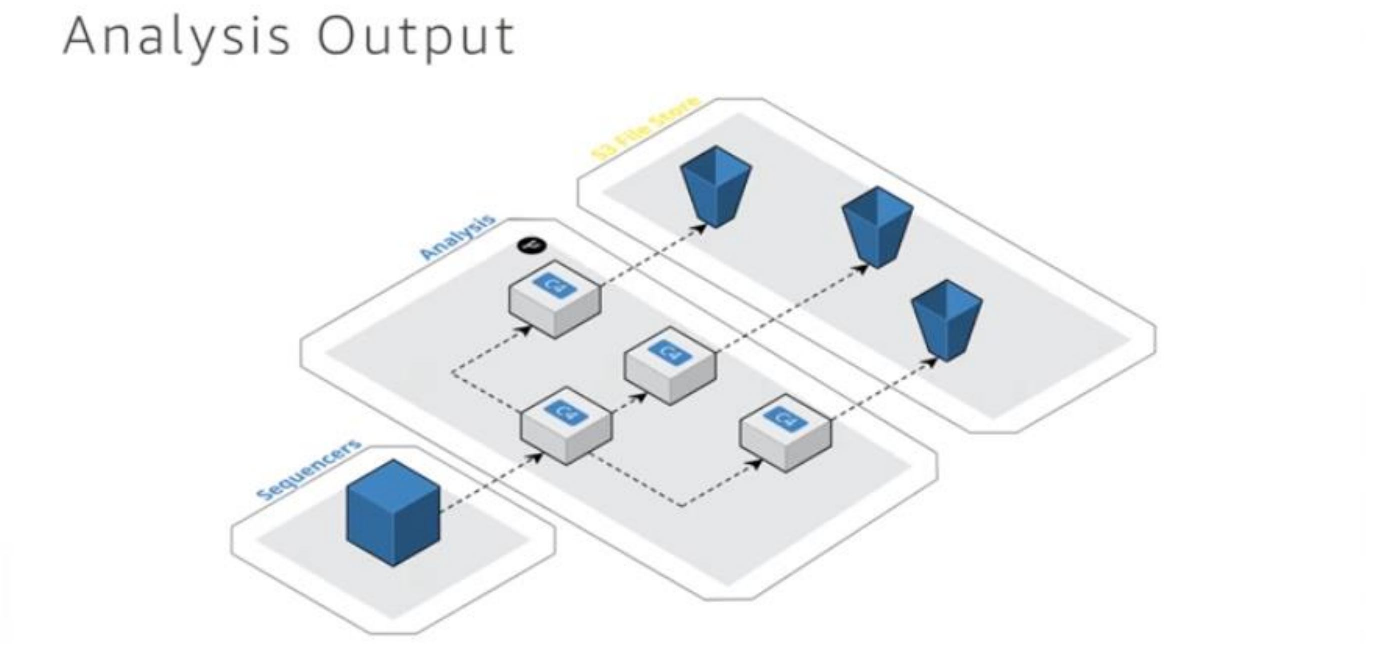
HLI Sequencing Laboratory

Sequencing Inventory	Capacity per Week	Output To Date
24 HiSeq X 4 HiSeq 4000	900 human genomes	>40,000 whole human genomes
2 NovaSeq 6000 1 iScan 1 MiSeqDx	192 microbiome genomes	>3,000 microbiome metagenome

AWS re:Invent

aws

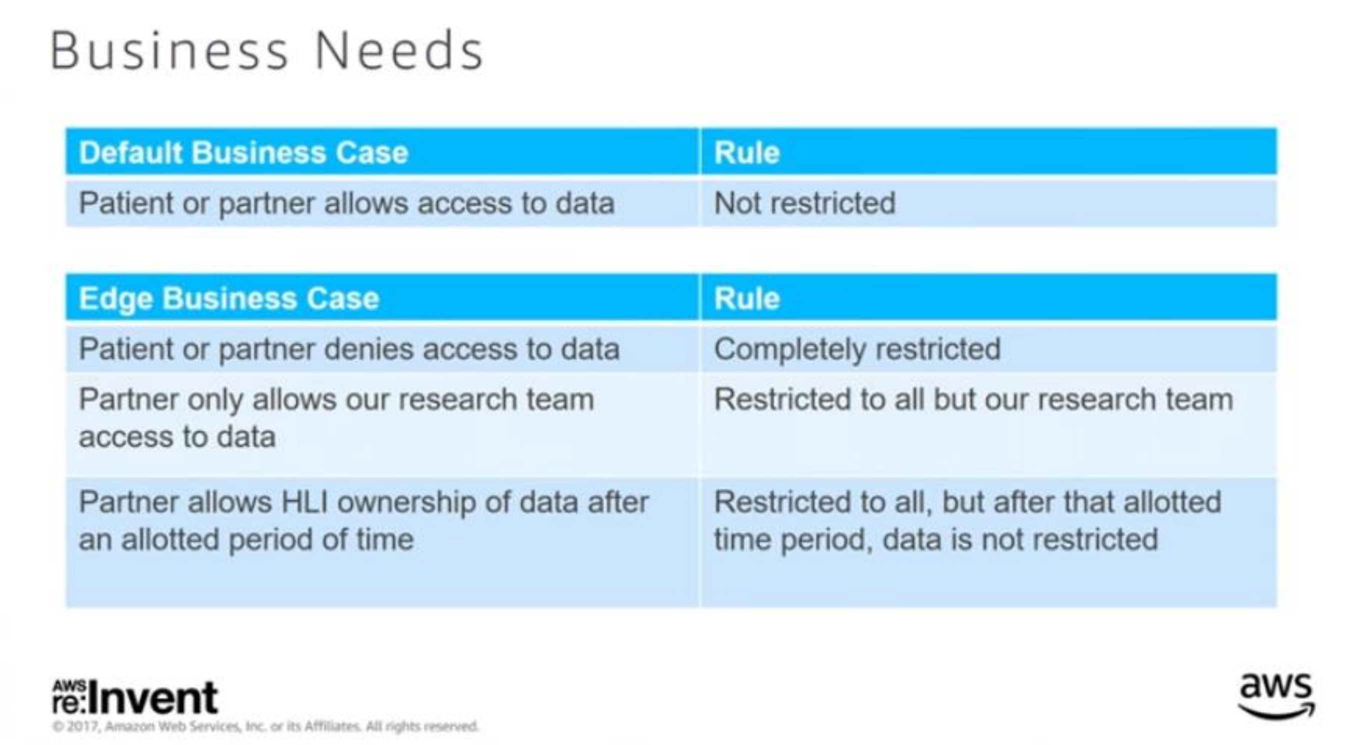
Microbiomes are the bacteria in the gut that also get sequenced. This adds up to about 6Pb of data in S3 and 10Pb in Glacier.



Analysis Output

To get the data, we put the samples into our sequencer and it dumps the data in our local aspire storage that will upload that data into S3. Those files can be up to 400,000 BCL files that all get compressed into a file that is about 600GB in size. When this 600GB file gets dropped into S3, it triggers our primary analysis platform that takes those 400,000 BCL files and turns them into the overall genome file which is about 200GB. After the primary analysis is done, a lambda gets triggered checks what is the next workflow that needs to be run like a secondary analysis step that then gets triggered. After all required workflow gets completed, all the analysis data again gets dumped into S3.

We then have professionals that map the phenomic data into specific ontologies and the result of their work again gets dumped into S3 as well. S3 is the merging ground for both genomic and phenotypic data.

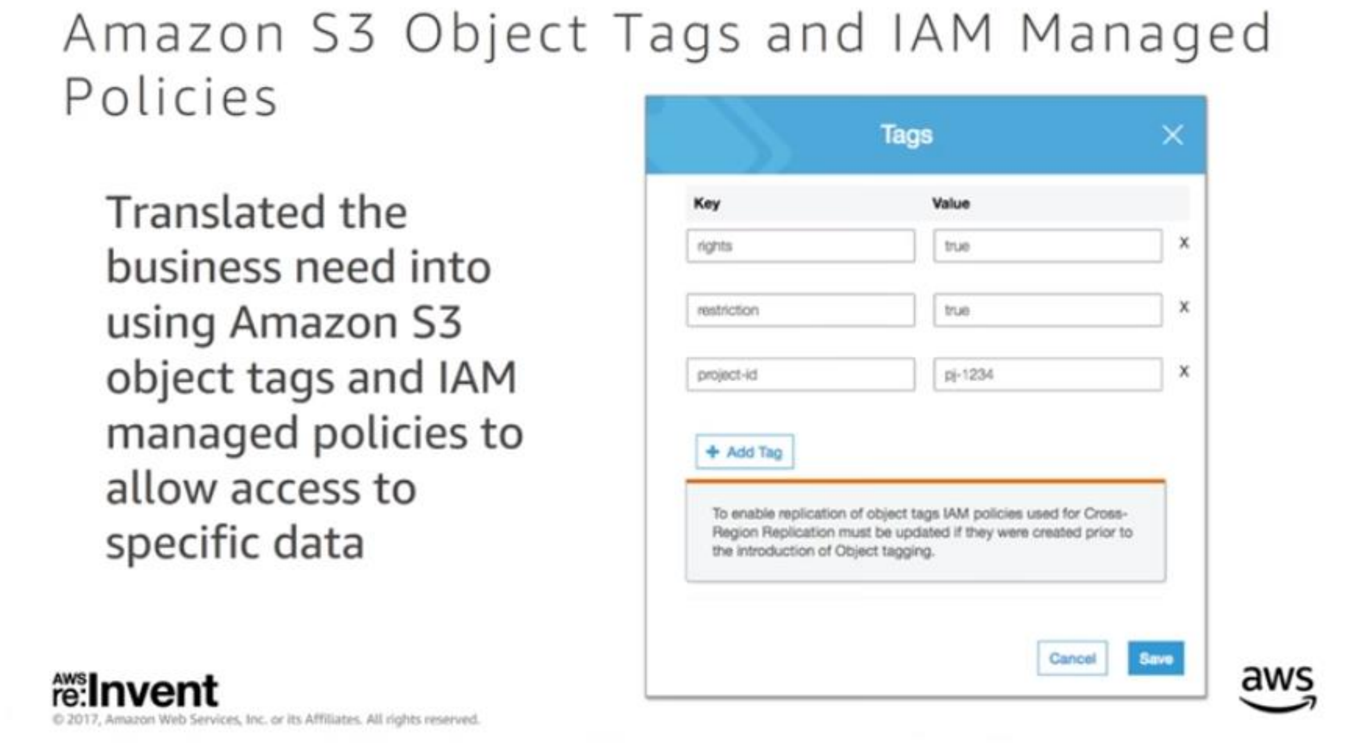


Business Needs

Default Business Case	Rule
Patient or partner allows access to data	Not restricted

Edge Business Case	Rule
Patient or partner denies access to data	Completely restricted
Partner only allows our research team access to data	Restricted to all but our research team
Partner allows HLI ownership of data after an allotted period of time	Restricted to all, but after that allotted time period, data is not restricted

There are 4 main business needs to securing our data.



Amazon S3 Object Tags and IAM Managed Policies

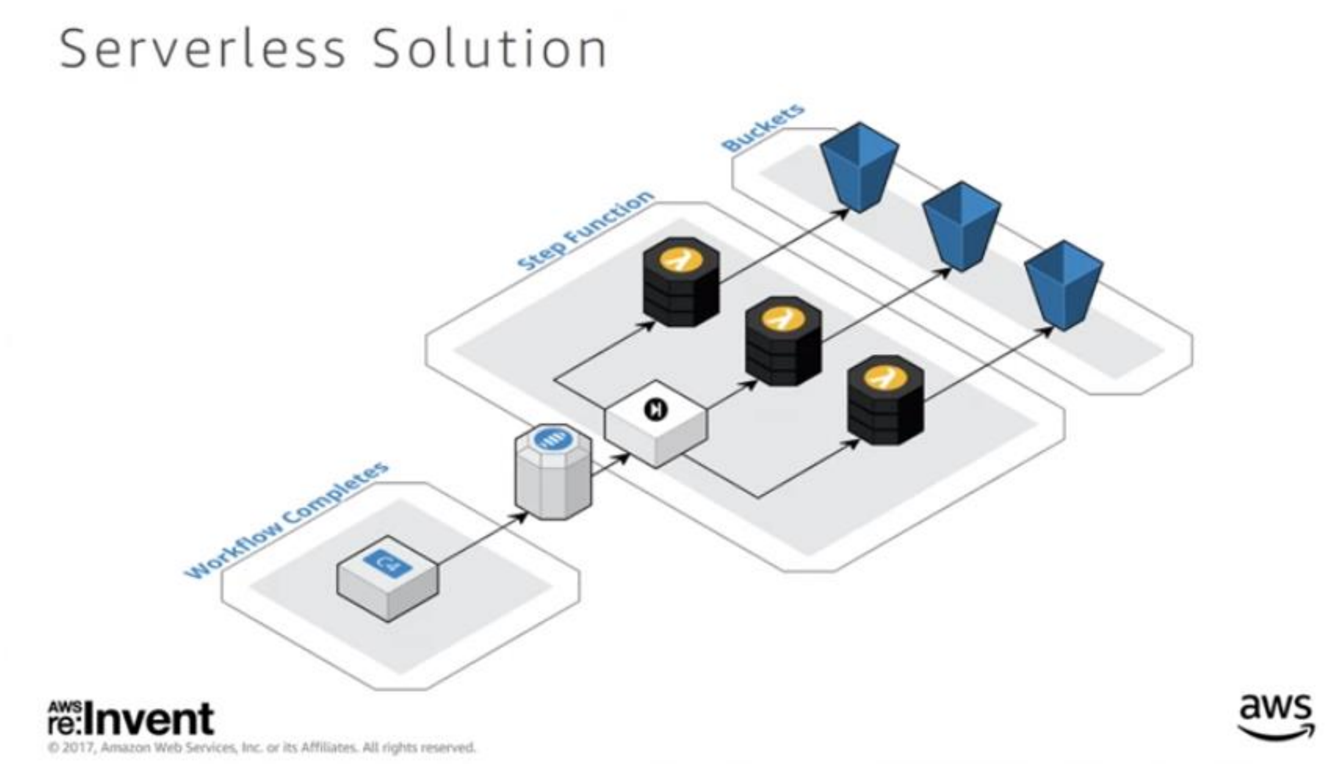
Translated the business need into using Amazon S3 object tags and IAM managed policies to allow access to specific data

Tags

Key	Value
rights	true
restriction	true
project-id	pj-1234

To enable replication of object tags IAM policies used for Cross-Region Replication must be updated if they were created prior to the introduction of Object tagging.

The 3 tags we use on the S3 buckets for access management that we use are ***rights***, ***restriction***, ***project-id***.

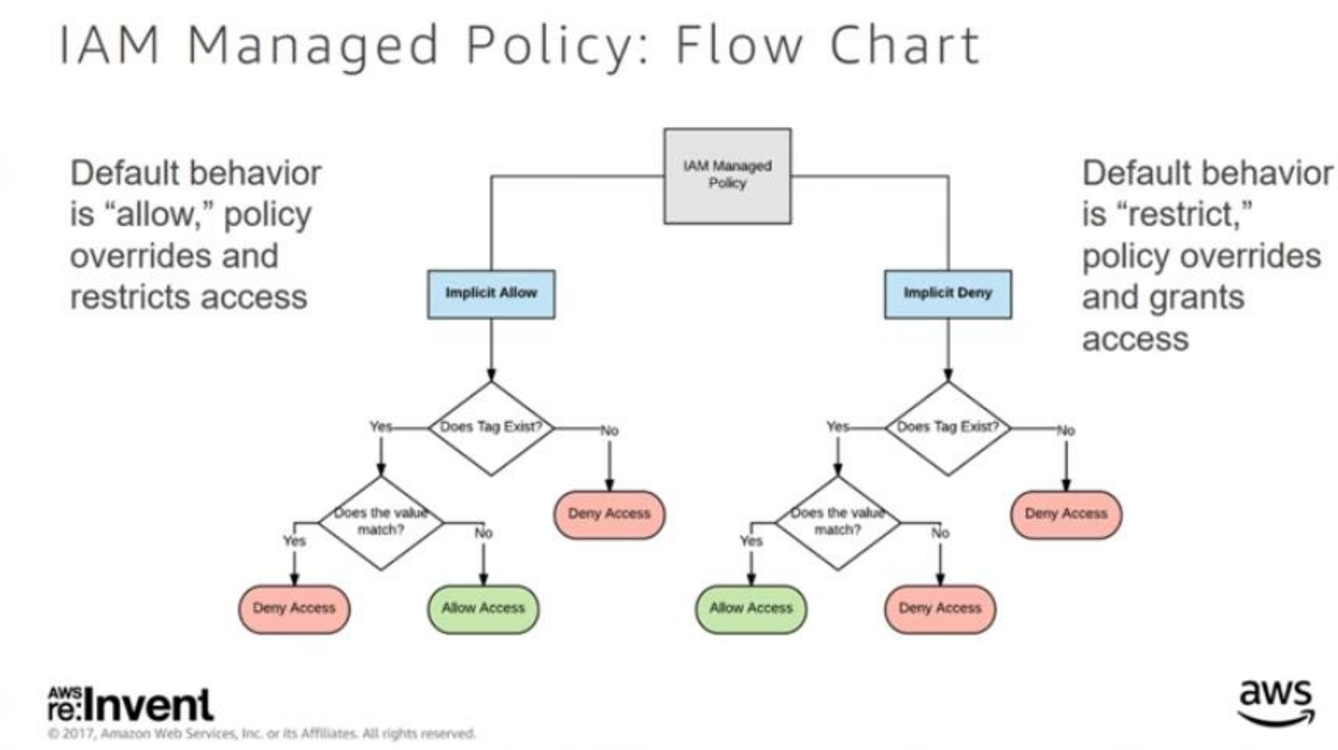


Serverless Solution

AWS re:Invent

aws

We needed an automatic solution for tagging the millions of files as appropriate, we used a serverless solution for doing this tagging as above. Once a workflow completes and has uploaded its files to S3, an SMS message gets sent off and that triggers a Step Function that has a series controlled parallel lambdas that gets triggered to go to their respective S3 buckets and tag all the files. Each of the individual lambdas will hit our own API to gather the list of files that they need to tag and what to tag them with, then they will go out and get the files, unpack them, and tag all the files in S3.



IAM Managed Policy: Flow Chart

Default behavior is "allow," policy overrides and restricts access

Default behavior is "restrict," policy overrides and grants access

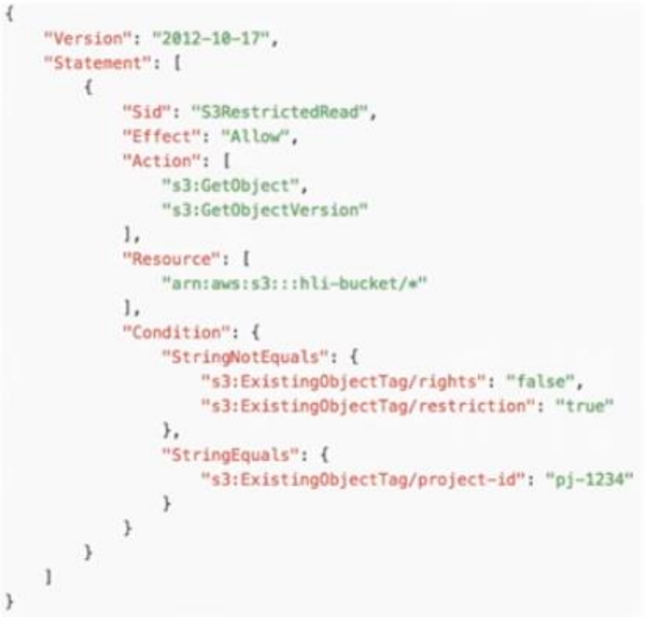
AWS re:Invent

aws

We have 2 different types of IAM policies, Implicit Allow and Implicit Deny.

IAM Managed Policy: Implicit Allow

- Deny read access:
  - Rights tag = false AND Restriction tag = true
- Override rule:
  - Project-id tag = pj-1234

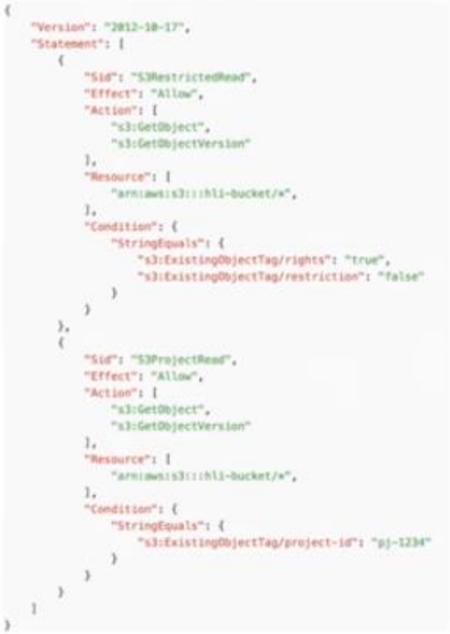


```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "S3RestrictedRead",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectVersion"
            ],
            "Resource": [
                "arn:aws:s3:::hli-bucket/*"
            ],
            "Condition": {
                "StringNotEquals": {
                    "s3:ExistingObjectTag/rights": "false",
                    "s3:ExistingObjectTag/restriction": "true"
                },
                "StringEquals": {
                    "s3:ExistingObjectTag/project-id": "pj-1234"
                }
            }
        }
    ]
}
```

re:Invent

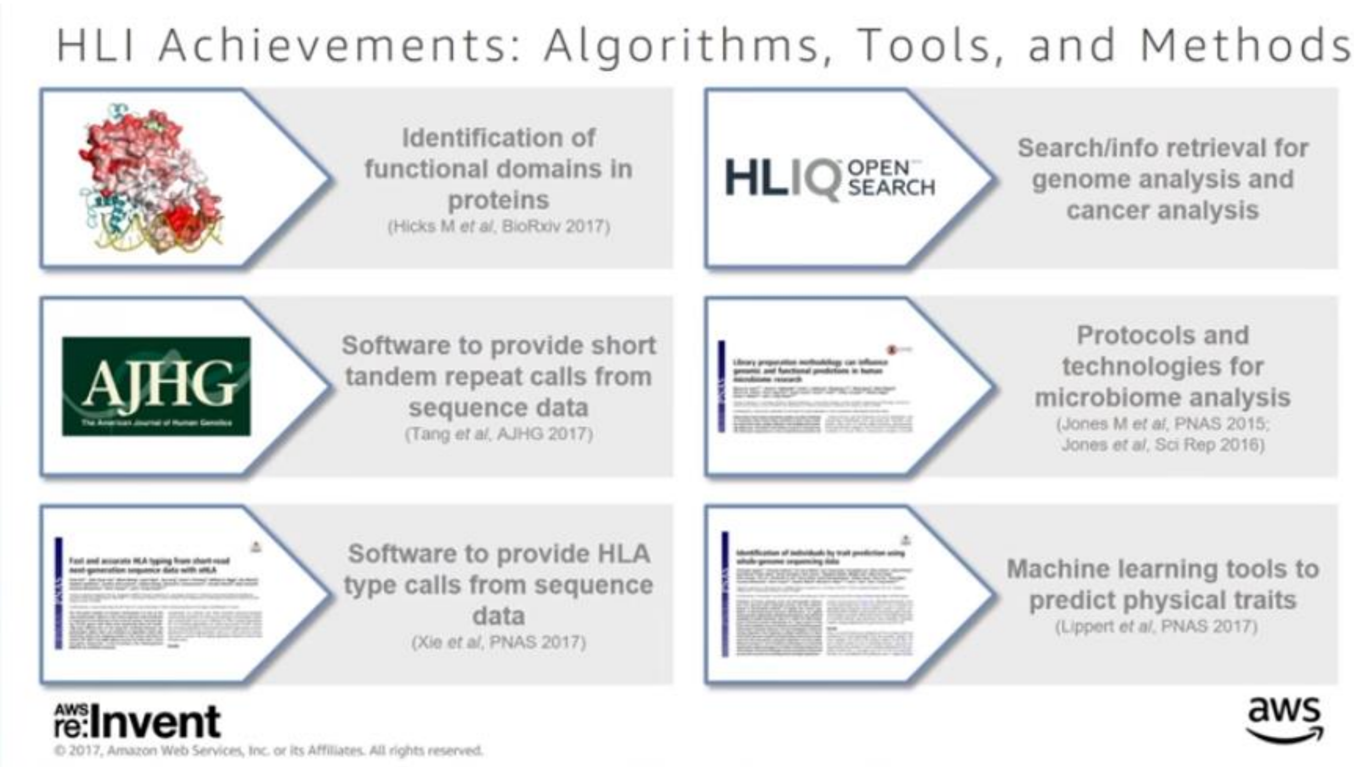
IAM Managed Policy: Implicit Deny

- Allow read access:
  - Rights tag = true AND Restriction tag = false
- Override rule:
  - Project-id tag = pj-1234



```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "S3RestrictedRead",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectVersion"
            ],
            "Resource": [
                "arn:aws:s3:::hli-bucket/*",
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ExistingObjectTag/rights": "true",
                    "s3:ExistingObjectTag/restriction": "false"
                }
            }
        },
        {
            "Sid": "S3ProjectRead",
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:GetObjectVersion"
            ],
            "Resource": [
                "arn:aws:s3:::hli-bucket/*",
            ],
            "Condition": {
                "StringEquals": {
                    "s3:ExistingObjectTag/project-id": "pj-1234"
                }
            }
        }
    ]
}
```

re:Invent



HLI Achievements: Algorithms, Tools, and Methods

Identification of functional domains in proteins
(Hicks M *et al*, BioRxiv 2017)

Search/info retrieval for genome analysis and cancer analysis

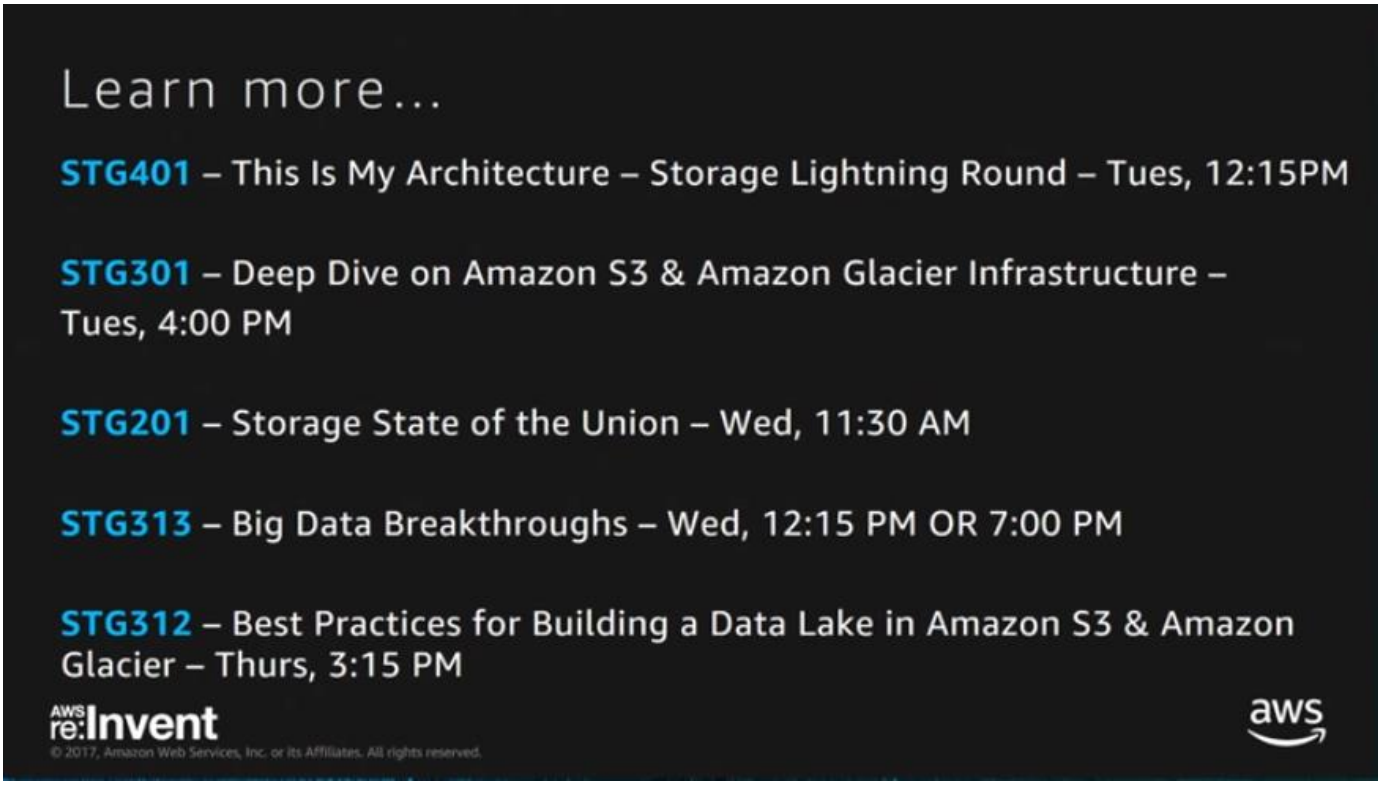
Software to provide short tandem repeat calls from sequence data
(Tang *et al*, AJHG 2017)

Protocols and technologies for microbiome analysis
(Jones M *et al*, PNAS 2015; Jones *et al*, Sci Rep 2016)

Software to provide HLA type calls from sequence data
(Xie *et al*, PNAS 2017)

Machine learning tools to predict physical traits
(Lippert *et al*, PNAS 2017)

We have done a lot with the data so far, we have created an open search tool that aggregates all the files that we have been granted open access to and built a search tool on top of them to allow easy insights and querying both the genomic and phenotypic data. we can also do queries based on demographics, age, genome variants to build graphs and insights.



Learn more…

STG401 – This Is My Architecture – Storage Lightning Round – Tues, 12:15PM

STG301 – Deep Dive on Amazon S3 & Amazon Glacier Infrastructure – Tues, 4:00 PM

STG201 – Storage State of the Union – Wed, 11:30 AM

STG313 – Big Data Breakthroughs – Wed, 12:15 PM OR 7:00 PM

STG312 – Best Practices for Building a Data Lake in Amazon S3 & Amazon Glacier – Thurs, 3:15 PM

AWS re:Invent

Thank you!

AWS re:Invent

aws