AWS SUMMIT ONLINE



T R A 0 4

Mastering your data journey one step at a time

Rada Stanic

Principal Solutions Architect Amazon Web Services





Current data challenges

Introducing data lakes

Security and governance patterns

AWS Lake Formation – making things easy

Data lake design patterns

Phased architecture build

Current data challenges



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Decision making used to...

... revolve around the Enterprise **Data Warehouse** (in the 90s – 00s)



Enterprise data warehouse

Data no longer fits



There is more data than people think

Data is more diverse

Wider range of workloads



There are more people accessing data

That want to analyse it in different ways

And there are more rules around data use

Introducing data lakes



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The data lake is the new information hub

A centralised repository that enables you to secure, discover, share, and analyse structured and unstructured data at any scale

The concept of a data lake

- All data in one place, a single source of truth
- Handles structured/semi-structured/unstructured/raw data
- Supports fast ingestion and consumption
- Schema on read
- Designed for low-cost storage
- Decouples storage and compute
- Supports protection and security rules

Building a data lake on AWS



Options for structured data ingestion



Amazon Simple Storage Service (Amazon S3) (raw data)



The core of a data lake

Versatile compute layers



Amazon Athena



Amazon EMR

Data lake

Data and metadata



Amazon S3



AWS Glue Data Catalog



Amazon Redshift Spectrum

Tiered approach



Amazon S3

Tier 1 data lake: Ingestion

- Single source of truth for raw data
- Organised by Ingestion Time
- Use least transformations
- Use lifecycle policies to Amazon S3 IA or Amazon S3 Glacier

Tiered approach



Amazon S3

Tier 2 data lake: Analytics

- Use columnar formats Parquet/ORC
- Organised into partitions
- Organised by Event Time
- Combine to larger partitions over time
- Optimised for analytics

Tiered approach



Amazon S3

Tier 3 data lake: Analytics

- Domain-level data mart
- Organised by use cases
- Optimised for specialised analysis



Security and governance patterns



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data storage security

Implement access control in a multi-team environment

- Fine-grained
- Coarse-grained

Secure and segregated access to

- Amazon S3
- Amazon EMR clusters
- Amazon Redshift clusters
- Serverless analytics services and other tools used in the pipeline

Encrypt data assets

Control access to data – fine grained ACL

"Fine-grained" data and resource ownership

- Teams share S3 buckets and \bullet clusters
- Access control is complex to set up and maintain
- Common in a "shared services" \bullet architecture



Control data access - coarse grained

Prefer "coarse-grained" data and resource ownership

- Teams own entire S3 buckets and clusters
- Ownership segregated by AWS accounts
- Access control easier to setup and maintain
- Suitable for **autonomous teams**



Amazon EMR Clusters

Control access to data

Configure Amazon S3 permissions

- Implement your access control matrix using IAM policies
- Use S3 bucket policies for easy cross-account data sharing
- Limit role-based access from an Amazon EMR cluster's EC2 instance profile
- Authorise access from other tools such as Amazon Redshift using IAM roles



Amazon S3

Block public access to Amazon S3

Amazon S3 provides four settings

- BlockPublicAcls
- IgnorePublicAcls
- BlockPublicPolicy
- RestrictPublicBuckets

But what is "public"?

- Public object (or bucket) ACL → grants permission to members of the predefined AllUsers or AuthenticatedUsers groups
- Public bucket policy
 → wild cards in Principal and Condition elements

Encryption - data-at-rest and in-motion



- Amazon S3 offers multiple forms of • encryption
 - Server-side and client-side encryption
 - Encryption with keys managed by S3 or **AWS Key Management Service**
 - Encryption with keys that customers manage
- Encrypts data in transit when replicating across regions
- Data movement services can use the same AWS KMS service

AWS Lake Formation



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation

AWS Lake Formation Dashboard Show ove Overview Data lake lifecycle stages and activities Stage 1 Stage 2 Stage 3 Stage 4 ρ A A Monitor and Audit Ingest and Register Security and Control Collaborate and Use Automatically ingest, clean, organize, Define access controls that provide Review data access history Search and discover using catalog and catalog structured and semithe right data to the right users, multiple use cases from a sil metadata. All access is checked structured data, including logs and groups, and roles. Database, table, against policy, so your data is and download data access data from relational databases into and column permissions offer protected even if tools change or granular security. your data lake. new data arrives. Create data lake Table permissions Search data catalog Monitor activity User permissions Add metadata Download logs Import data

erview 🔵	
across	
ingle place audit logs.	
У	
s	

Typical steps of building a data lake



Secure once - access in multiple ways





2. User tries to access data via one of the services

Security permissions in AWS Lake Formation

Control data access with simple grant and revoke permissions

Specify permissions on tables and columns rather than on buckets and objects

Easily view policies granted to a particular user

Audit all data access at one place

Tables (2)		Actions v Use	rawler 🗹 🛛 Add t	able	
Q Filter by tags and attributes or sea	rch by keyword	View data 🗹 Edit	< 1 >	۲	
Name 🔻 Databa	se v Location v	Сору	Last updated	Ŧ	
reviews sales	S3://datalake/sales/reviews/	Delete	23 October 2018 6:1	5 AM	
orders sales	S3://datalake/sales/orders/	Grant	23 October 2018 6:1	7 AM	
		Verify permissions			
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran	nd roles for the selected tab	le.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran	nd roles for the selected tab	le.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran Q. User, group, or role name Johnd X. salesgrp X. analyst X. Permissions Info Select the specific access permissions to grant	nd roles for the selected tab	te.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran Q User, group, or role name johnd X salesgrp X analyst X Permissions Info Select the specific access permissions to grant G Grant all Full permissions as well as the ability to grant a	nd roles for the selected tab	te.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran Q. User, group, or role name johnd X salesgrp X analyst X Permissions Info Select the specific access permissions to grant G Grant all Full permissions as well as the ability to grant a Specific permissions	nd roles for the selected tab	le.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to gran Q. User, group, or role name johnd X salesgrp X analyst X Permissions Info Select the specific access permissions to grant Grant all Full permissions as well as the ability to grant a Specific permissions Select all Create Select	nd roles for the selected tab t access permissions nd revoke other users. t 2 Insert 2 Alter	le.		
	Grant access permissions to specific users, groups, a IAM user, group, and roles Info Add one or more IAM users, groups, or roles to grant Q. User, group, or role name [johnd X] salesgrp X] analyst X Permissions Info Select the specific access permissions to grant Grant all Full permissions as well as the ability to grant a Specific permissions Select all I Create I Select Columns – optional Info Add one or more columns governed by the access permissions to granted by the access permissi	nd roles for the selected tab t access permissions nd revoke other users. t Insert I Alter ermissions	le.		

Contemporary Location S3://datalake/sales/reviews/ S3://datalake/sales/orders/ Grant permissions to table orders Grant access permissions to specific users, groups, and AM user, group, and roles Info	Edit Copy Delete Grant Revoke Verify permission	Last updated 23 October 201 23 October 201 s ed table.	8 6:15 AM 8 6:17 AM	×
 Location S3://datalake/sales/reviews/ S3://datalake/sales/orders/ Grant permissions to table orders Grant access permissions to specific users, groups, and AM user, group, and roles Info	Copy Delete Grant Revoke Verify permission	s Last updated 23 October 201 23 October 201 23 October 201	▼ 8 6:15 AM 8 6:17 AM	×
S3://datalake/sales/reviews/ S3://datalake/sales/orders/ Grant permissions to table orders Grant access permissions to specific users, groups, and AM user, group, and roles Info	Delete Grant Revoke Verify permission nd roles for the select	23 October 201 23 October 201 s ed table.	8 6:15 AM 8 6:17 AM	×
S3://datalake/sales/orders/ Grant permissions to table orders Grant access permissions to specific users, groups, and AM user, group, and roles Info	Grant Revoke Verify permission and roles for the select	23 October 201 s ed table.	8 6:17 AM	×
Grant permissions to table orders Grant access permissions to specific users, groups, an AM user, group, and roles Info	Verify permission	s ed table.		×
Grant permissions to table orders Grant access permissions to specific users, groups, and AM user, group, and roles Info	nd roles for the select	ed table.		×
Q User, group, or role name				
johnd X salesgrp X analyst X Permissions Info Select the specific access permissions to grant				
Grant all Full permissions as well as the ability to grant and	nd revoke other users.			
Specific permissions				
🗌 Select all 🛛 🗹 Create 🗹 Select	t 🗹 Insert 🗹 /	Alter 🗌 Drop 🗌 Del	ete	
Columns – optional Info	ermissions			
Q Column name - wildcards allowed				
Columns – optional Info Add one or more columns governed by the access pe Q Column name – wildcards allowed	ermissions			

Security permissions in Lake Formation

Control data access with simple grant and revoke permissions

Specify permissions on tables and columns rather than on buckets and objects

Easily view permissions granted to a particular user

Audit all data access in one place

Column name	Data t
marketplace	string
customer_id	bigint
review_id	string
product_id	string
product_parent	bigint
product_title	string
star_rating	string
helpful_votes	bigint
total_votes	bigint
vine	string
verified_purchase	string
review_headline	string
review_body	string
review_date	string
product_category	string



User 1



Security deep dive



AWS Lake Formation



Granular permissions control for users



Grant permissions	Add one or more columns to include.
Grant access permissions to specific users and roles.	Choose columns
IAM users and roles Add one or more IAM users or roles.	useridentity X string string eventsource X string string
Choose IAM principals to add	sourceipaddress X string
datalake_user_redshift 🗙 User	Table permissions Choose the specific access permissions to grant.
Database Add one or more databases.	 ✓ Select all ✓ Alter ✓ Insert ✓ Delete ✓ Select
Choose databases	Grant all Enabling this permission grants full access to the specified restill logging access requests based on the individual permissi above. It is typically used for debugging. Specific individual permission access to the specific individual pe
Table - optional Add one or more tables.	Grantable permissions Choose the specific permissions that may be granted to othe
Choose tables 🔹	Select all Create table Alter
amazoncloudtrail_cloudtrail X Column - optional Grant permissions to:	Grant all Enabling this permission grants full access to the specified re still logging access requests based on the individual permissi above. It is typically used for debugging. Specific individual p above will take effect when Grant all is later removed.
The include columns 🔹	



sources while on settings ermissions set

ers.

Drop

sources while on settings permissions set

Cancel



Data lake design patterns



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Log analytics, IoT sensor data



8-8

Amazon Redshift Data Warehouse





Reporting

Log analytics, IoT sensor data



) %

Amazon Redshift Data Warehouse







Reporting



Machine Learning - batch training pipeline

Amazon SageMaker Endpoint

Machine Learning - predictions on streaming data

Presto/Spark on EMR

Amazon Redshift Data Warehouse

Machine Learning - predictions on streaming data

Presto/Spark on EMR

Amazon Redshift Data Warehouse

Phased architecture build

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Bringing it all together ...

Conclusion

Building modern data architecture is a phased journey

Data lake should be designed as a self-service platform

Leverage tools to automate (AWS Lake Formation)

Leverage design patterns to deliver key use cases

Use security and governance controls

Thank you!

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

