Introduction to AWS Lake Formation

Prajakta Damle, Principal Product Manager, AWS Lake Formation & AWS Glue
Agenda

• Why did we build AWS Lake Formation?
• What is AWS Lake Formation?
• How does AWS Lake Formation help you?
There is more data than people think.

Data grows >10x every 5 years. Data platforms need to live for 15 years and scale 1,000x.

* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don’t Focus on Big Data, Focus on the Data That’s Big, April 2017.
There are more people accessing data

And more requirements for making data available

Data Scientists  Business Users
Analysts  Applications

Secure  Real time
Flexible  Scalable
A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.
Why data lakes?

Data Lakes provide:

- Relational and non-relational data
- Scale-out to EBs
- Diverse set of analytics and machine learning tools
- Work on data without any data movement
- Designed for low cost storage and analytics
Any analytic workload, any scale, at the lowest possible cost
More data lakes & analytics on AWS than anywhere else
Typical steps of building a data lake

1. Setup storage
2. Move data
3. Cleanse, prep, and catalog data
4. Configure and enforce security and compliance policies
5. Make data available for analytics
Building data lakes can still take months
Data preparation accounts for ~80% of the work

- Building training sets
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Refining algorithms
- Other
Sample of steps required:

- Configure access from analytics services

Rinse and repeat for other:
- data sets, users, and end-services

And more:
- manage and monitor ETL jobs
- update metadata catalog as data changes
- update policies across services as users and permissions change
- manually maintain cleansing scripts
- create audit processes for compliance

... Manual | Error-prone | Time consuming
AWS Lake Formation

Build a secure data lake in days

- Identify, ingest, clean, and transform data
- Enforce security policies across multiple services
- Gain and manage new insights
How it works

Overview
Data lake lifecycle stages and activities

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingest and Register</td>
<td>Security and Control</td>
<td>Collaborate and Use</td>
<td>Monitor and Audit</td>
</tr>
<tr>
<td>Create data lake</td>
<td>Table permissions</td>
<td>Search data catalog</td>
<td>Monitor activity</td>
</tr>
<tr>
<td>Import data</td>
<td>User permissions</td>
<td>Add metadata</td>
<td>Download logs</td>
</tr>
</tbody>
</table>

Ingest and Register
Automatically ingest, clean, organize, and catalog structured and semi-structured data, including logs and data from relational databases into your data lake.

Security and Control
Define access controls that provide the right data to the right users, groups, and roles. Database, table, and column permissions offer granular security.

Collaborate and Use
Search and discover using catalog metadata. All access is checked against policy, so your data is protected even if tools change or new data arrives.

Monitor and Audit
Review data access history across multiple use cases from a single place and download data access audit logs.
Register existing data or import new

Amazon S3 forms the storage layer for Lake Formation

Register existing S3 buckets that contain your data

Ask Lake Formation to create required S3 buckets and import data into them

Data is stored in your account. You have direct access to it. No lock-in.
Easily load data to your data lake

Blueprints

one-shot

incremental

DBs

logs

Amazon RDS

MySQL

Oracle

PostgreSQL

Amazon Aurora

Amazon Kinesis Data Firehose

AWS CloudTrail

Elastic Load Balancing

Amazon CloudFront

Lake Formation

Data import

Access Control

Data Catalog

Crawlers

ML-based data prep

Data Lake Storage

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
With blueprints

You

1. Point us to the source
2. Tell us the location to load to in your data lake
3. Specify how often you want to load the data

Blueprints

1. Discover the source table(s) schema
2. Automatically convert to the target data format
3. Automatically partition the data based on the partitioning schema
4. Keep track of data that was already processed
5. You can customize any of the above
Blueprints build on AWS Glue

---

**Blueprints**

- **Monitoring**
  - AWS Glue Jobs
  - AWS Glue Crawlers

- **Workflow**
  - AWS Glue Data Catalog
  - Connections, Databases, Tables
Easily de-duplicate your data with ML transforms

1. Merge related data sets, then Lake Formation will divide data into train and test samples.
2. Lake Formation identifies duplicates and fuzzy matches the records.
3. Tune or provide additional examples of matches until satisfied with the quality and performance.
4. Put ML transforms into production as part of your data prep.
Fuzzy de-duplication – under the hood

Naïve: look at all pairs, $N^2$ – state-of-the-art:

- Blocking
- Score pairs
- Partition
# Fuzzy de-duplication – Innovations

<table>
<thead>
<tr>
<th>Intersection Dynamic Blocking (VLDB 2008)</th>
<th>SuperPart</th>
</tr>
</thead>
<tbody>
<tr>
<td>parallelizable &amp; performant blocks on dynamic mix of columns</td>
<td>partitions based on customer-provided ground-truth</td>
</tr>
<tr>
<td>400M+ rows</td>
<td>gives confidence of grouping</td>
</tr>
<tr>
<td>7.5B+ candidate pairs</td>
<td>effective without tuning knobs</td>
</tr>
<tr>
<td>2.5 hours</td>
<td></td>
</tr>
</tbody>
</table>
Secure once, access in multiple ways

1. Set up user access in Lake Formation

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

3. Service sends user credentials to Lake Formation

4. Lake Formation returns temporary credentials allowing data access

Admin

Access Control

Data Catalog

Amazon S3

Data Lake Storage

Lake Formation

Amazon Athena

Amazon QuickSight

Amazon Redshift

AWS Glue

Amazon EMR

Amazon SageMaker

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
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 policies granted to a particular user

Audit all data access at one place
Security permissions in Lake Formation

Search and view permissions granted to a user, role, or group in one place

Verify permissions granted to a user

Easily revoke policies for a user
Grant table and column-level permissions

<table>
<thead>
<tr>
<th>Column name</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>marketplace</td>
<td>string</td>
</tr>
<tr>
<td>customer_id</td>
<td>bigint</td>
</tr>
<tr>
<td>review_id</td>
<td>string</td>
</tr>
<tr>
<td>product_id</td>
<td>string</td>
</tr>
<tr>
<td>product_parent</td>
<td>bigint</td>
</tr>
<tr>
<td>product_title</td>
<td>string</td>
</tr>
<tr>
<td>star_rating</td>
<td>string</td>
</tr>
<tr>
<td>helpful_votes</td>
<td>bigint</td>
</tr>
<tr>
<td>total_votes</td>
<td>bigint</td>
</tr>
<tr>
<td>vine</td>
<td>string</td>
</tr>
<tr>
<td>verified_purchase</td>
<td>string</td>
</tr>
<tr>
<td>review_headline</td>
<td>string</td>
</tr>
<tr>
<td>review_body</td>
<td>string</td>
</tr>
<tr>
<td>review_date</td>
<td>string</td>
</tr>
<tr>
<td>product_category</td>
<td>string</td>
</tr>
</tbody>
</table>
Security – deep dive

1. query T

User

2. request access

3. AWS Lake Formation

End-services retrieve underlying data directly from S3

IAM users, roles
Active Directory

Principals can be IAM users, roles Active Directory users via federation

4. request objs comprising T

5. return objs of T

Amazon S3

© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
Search and collaborate across multiple users

Text-based, faceted search across all metadata

Add attributes like Data owners, stewards, and other as table properties

Add data sensitivity level, column definitions, and others as column properties

Text-based search and filtering

Query data in Amazon Athena
Audit and monitor in real time

See detailed alerts in the console

Download audit logs for further analytics

Data ingest and catalog notifications also published to Amazon CloudWatch events
Example: a data lake in 3 easy steps

1. Use blueprints to ingest data
2. Grant permissions to securely share data
3. Query the data (Amazon Athena)
Step 1: Blueprints to ingest data
Monitor the import

<table>
<thead>
<tr>
<th>Load status</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run state</td>
<td>SUCCESSFUL</td>
</tr>
<tr>
<td>Imported table count</td>
<td>12</td>
</tr>
</tbody>
</table>
Imported data as table in the data lake
Step 2: Grant permissions to securely share data

[Diagram of the AWS Lake Formation UI showing the process of granting permissions to a table.]
Step 3: Run query in Amazon Athena
AWS Lake Formation Pricing

No additional charges - Only pay for the underlying services used.
Customer interest

“We are very excited about the launch of AWS Lake Formation, which provides a central point of control to easily load, clean, secure, and catalog data from thousands of clients to our AWS-based data lake, dramatically reducing our operational load. … Additionally, AWS Lake Formation will be HIPAA compliant from day one …”

Aaron Symanski, CTO, Change Healthcare

“I can’t wait for my team to get our hands on AWS Lake Formation. With an enterprise-ready option like Lake Formation, we will be able to spend more time deriving value from our data rather than doing the heavy lifting involved in manually setting up and managing our data lake.”

Joshua Couch, VP Engineering, Fender Digital
Thank you!

Join the preview: https://pages.awscloud.com/lake-formation-preview.html
lakeformation-pm@amazon.com