Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation

Nikki Rouda, Tim Jones, AWS

March 19th, 2019
Cloud data lakes are the future

Customers want:

- To move to a single store; i.e., a data lake in the cloud
- To store data securely in standard formats
- To grow to any scale, with low costs
- To analyze their data in a variety of ways
- To democratize data access and analysis
AWS analytics services
Broadest and deepest portfolio, purpose-built for builders
AWS analytics services
Broadest and deepest portfolio, purpose-built for builders

Visualization & Machine Learning
- QuickSight
- SageMaker
- Comprehend
- Lex
- Polly
- Rekognition
- Translate
- Transcribe
- Deep Learning AMIs
+ 10 more

Analytics
- Redshift
- EMR (Apache Spark & Hadoop)
- AWS Glue (Apache Spark & Python)
- Athena
- Elasticsearch Service
- Kinesis Data Analytics

Data Lake Infrastructure & Management
- S3/Glacier
- Lake Formation
- AWS Glue

Data Movement
- Database Migration Service
- Snowball
- Snowmobile
- Kinesis Data Firehose
- Kinesis Data Streams
- Managed Streaming for Kafka
Set up a catalog, ETL, and data prep with AWS Glue

Serverless provisioning, configuration, and scaling to run your ETL jobs on Apache Spark

Pay only for the resources used for jobs

Crawl your data sources, identify data formats and suggest schemas and transformations

Automates the effort in building, maintaining and running ETL jobs
“Beeswax uses Amazon S3 and AWS Glue Data Catalog to build a highly reliable data lake that is fully managed by AWS. Our platform leverages the AWS Glue Data Catalog integration with Amazon EMR in Hive and SparkSQL applications to deliver reporting and optimization features to our customers.”

—Ram Kumar Rengaswamy, CTO, Beeswax
Challenges to making a secure data lake

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
Build a secure data lake in days with AWS Lake Formation

Move, store, catalog, and clean your data faster

Enforce security policies across multiple services

Gain and manage new insights

Move, store, catalog, and clean your data faster with Machine Learning

Enforce security policies across multiple services

Empower analyst and data scientist to gain and manage new insights
“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 at Fender Digital
Using the ‘FindMatches’ ML Transform
Data integration and deduplication with FindMatches

AWS Lake Formation

Merge related data sets, then Lake Formation will divide data into train and test samples

Lake Formation identifies duplicates and fuzzy matches the records

Tune or provide additional examples of matches until satisfied with the quality and performance

Put ML transforms into production as part of your data prep

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential
“FindMatches” ML Transform Target Problems

Data Integration

Finding the relationships between multiple datasets, even when those datasets do not share an identifier (or when their identifier is unreliable)

Deduplication

Transforming a dataset that has multiple rows referring to the *same actual thing* into a dataset where no two rows refer to the *same actual thing*
Some examples
Illustrations of types of problems this technology has been applied to.
Data integration in movies...
Data integration in products
Data integration for People

Name: Tim Jones  
DOB: 1/1/1979  
Zip Code: V6T 1Z4  
Hobbies: Guitar, Reading, Computers  
Allergies: Amoxicillin

Name: Timothy Z. Jones  
DOB: 1/16/1979  
Zip Code: 98101  
Hobbies: Woodworking, Audiobooks, Computers  
Allergies: Peanuts, Amoxicillin
Demo
Candidate Generation, Pair Comparison, Clustering

Strongly influenced by:
- cost/accuracy slider
- match labels
- precision/recall slider
Thank you!

Sign up for the Lake Formation Preview: https://pages.awscloud.com/lake-formation-preview.html

Questions and Use case details? Send an email to lakeformation-pm@amazon.com