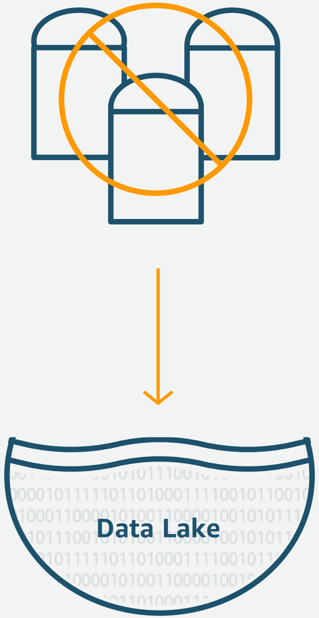
Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation

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

Visualization & Machine Learning

- Dashboards
- Predictive Analytics

Analytics

- Data Warehousing
- Big Data Processing
- Serverless Data processing
- Interactive Query
- Operational Analytics
- Real time Analytics

Data Lake Infrastructure & Management

- Infrastructure
- Security & Management
- Data Catalog & ETL

Data Movement

Migration & Streaming Services

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 NEW

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

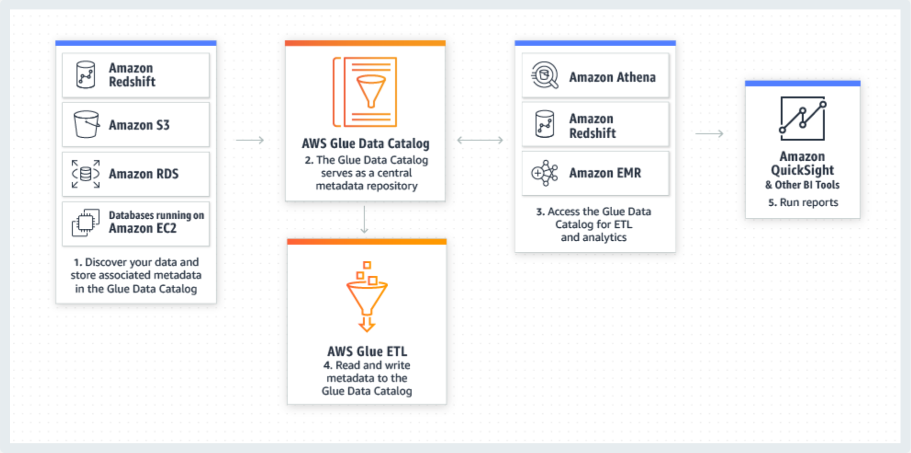
with AWS Glue

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


Amazon Redshift
Amazon S3
Amazon RDS
Databases running on Amazon EC2
1. Discover your data and store associated metadata in the Glue Data Catalog
AWS Glue Data Catalog
2. The Glue Data Catalog serves as a central metadata repository
AWS Glue ETL
4. Read and write metadata to the Glue Data Catalog
Amazon Athena
Amazon Redshift
Amazon EMR
3. Access the Glue Data Catalog for ETL and analytics
Amazon QuickSight
& Other BI Tools
5. Run reports

BEESWAX

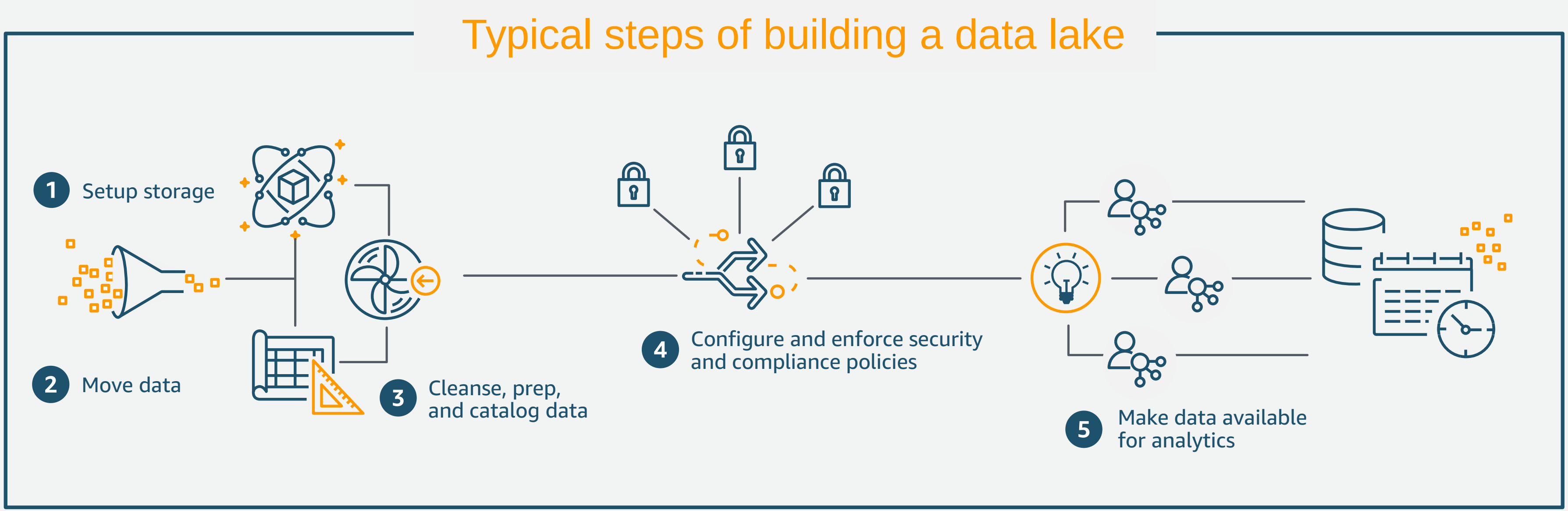
“Beeswax uses Amazon S3 and AWS Glue Data Catalog to build a highly reliable data lake that is fully managed by AWS. Our platform leverages the AWS Glue Data Catalog integration with Amazon EMR in Hive and SparkSQL applications to deliver reporting and optimization features to our customers.”

—Ram Kumar Rengaswamy, CTO, Beeswax

Data lake infrastructure & management

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

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

Enforce security policies across multiple services

Enforce security policies across multiple services

Gain and manage new insights

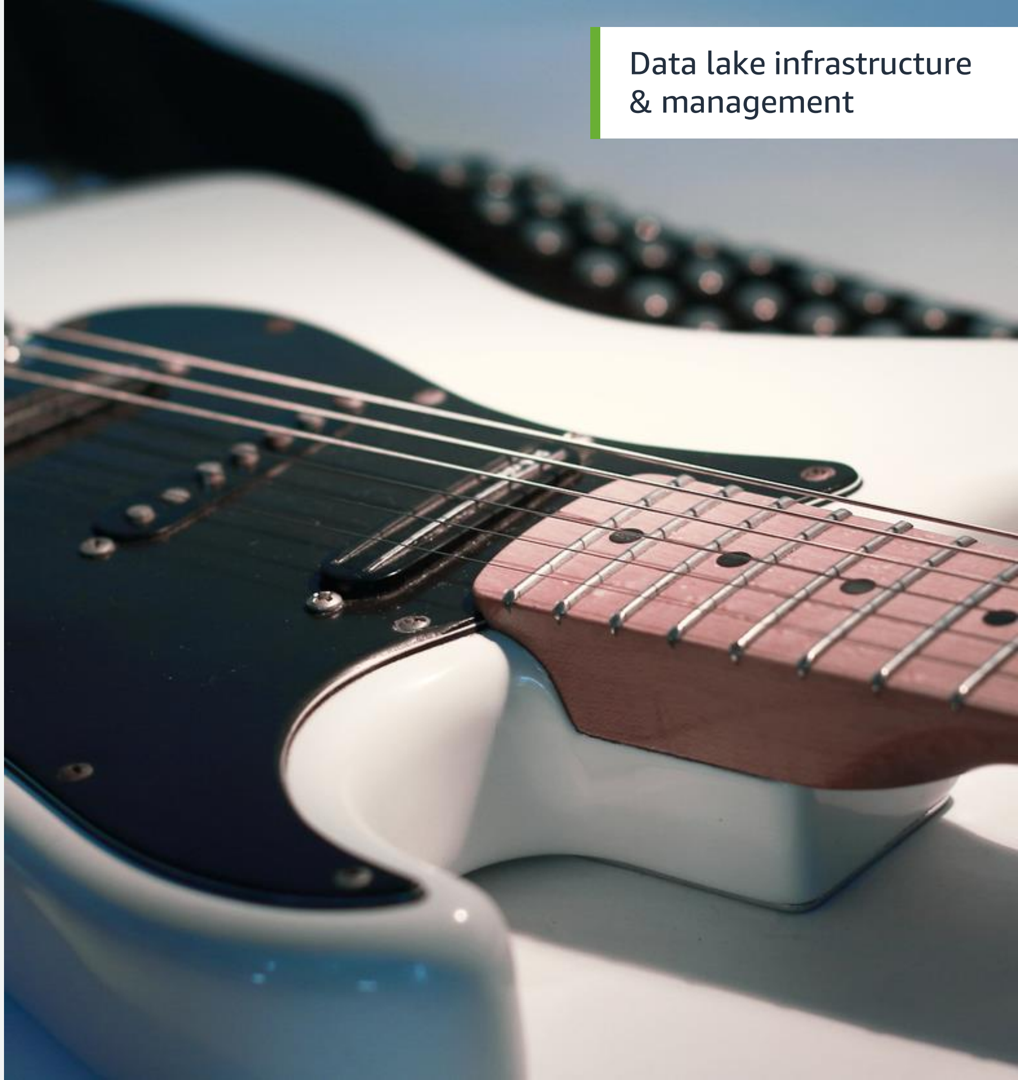
Empower analyst and data scientist to gain and manage new insights

Data lake infrastructure & management

Fender DIGITAL

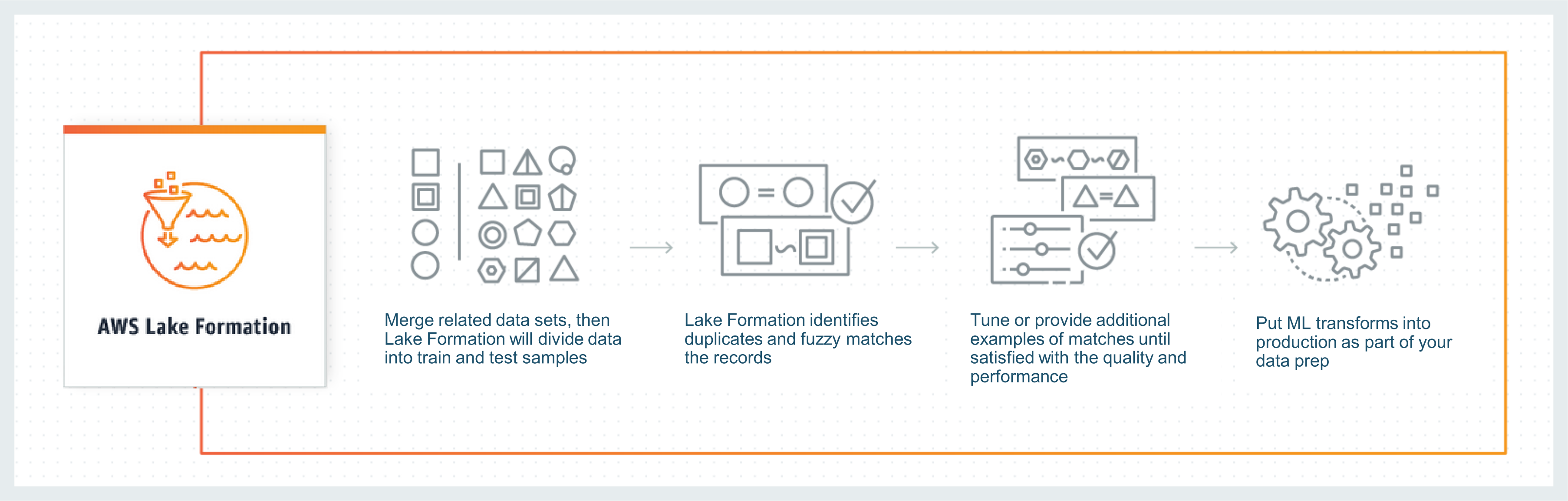
"With an enterprise-ready option like Lake Formation, we will be able to spend more time deriving value from our data rather than doing the heavy lifting involved in manually setting up and managing our data lake."

—Joshua Couch, VP Engineering at Fender Digital



Using the ‘FindMatches’ ML Transform

Data integration and deduplication with FindMatches



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

“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)

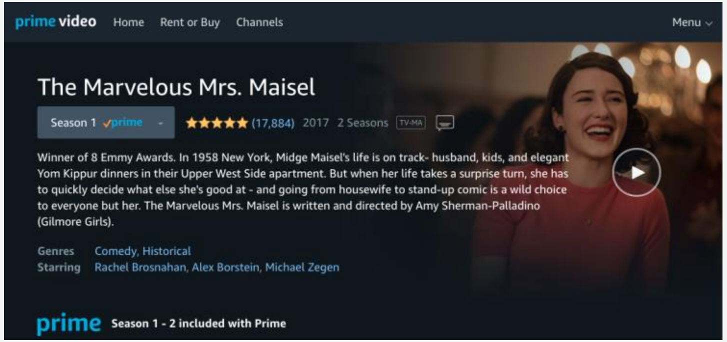
Deduplication

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…



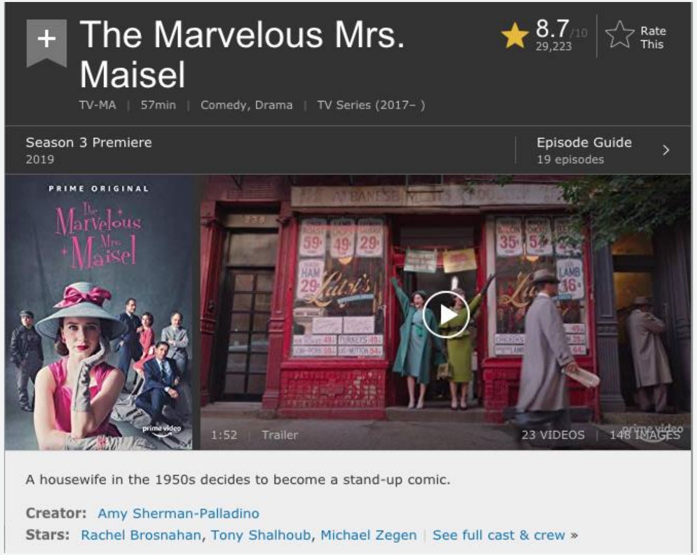
The Marvelous Mrs. Maisel

Season 1 prime ★★★★★ (17,884) 2017 2 Seasons TV-MA

Winner of 8 Emmy Awards. In 1958 New York, Midge Maisel's life is on track- husband, kids, and elegant Yom Kippur dinners in their Upper West Side apartment. But when her life takes a surprise turn, she has to quickly decide what else she's good at - and going from housewife to stand-up comic is a wild choice to everyone but her. The Marvelous Mrs. Maisel is written and directed by Amy Sherman-Palladino (Gilmore Girls).

Genres Comedy, Historical
Starring Rachel Brosnahan, Alex Borstein, Michael Zegen

prime Season 1 - 2 included with Prime



The Marvelous Mrs. Maisel

★ 8.7/10 29,223 ☆ Rate This

TV-MA | 57min | Comedy, Drama | TV Series (2017–)

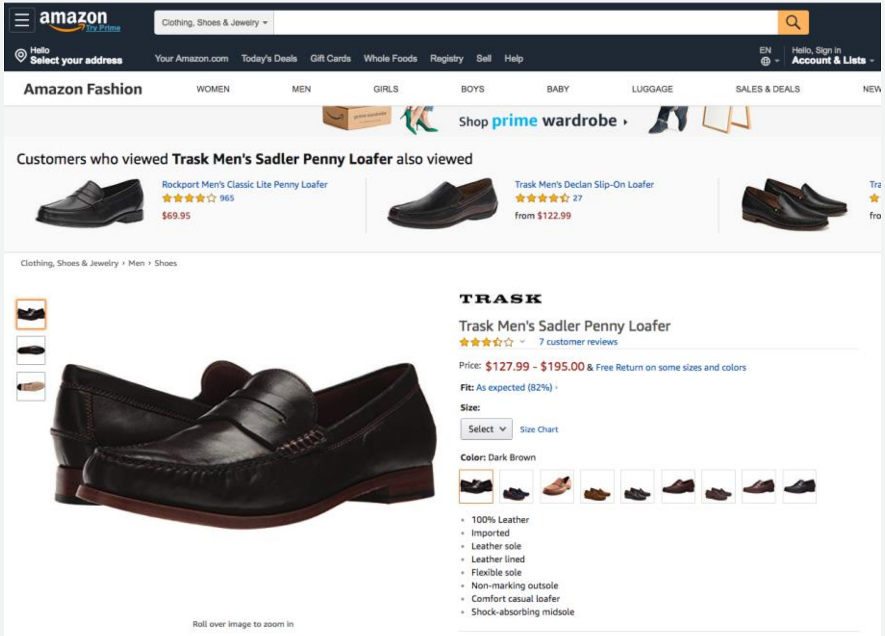
Season 3 Premiere
2019

Episode Guide
19 episodes

A housewife in the 1950s decides to become a stand-up comic.

Creator: Amy Sherman-Palladino
Stars: Rachel Brosnahan, Tony Shalhoub, Michael Zegen | See full cast & crew »

Data integration in products



Customers who viewed **Trask Men's Sadler Penny Loafer** also viewed

Rockport Men's Classic Lite Penny Loafer
965
$69.95

Trask Men's Declan Slip-On Loafer
27
from $122.99

Clothing, Shoes & Jewelry › Men › Shoes

TRASK

Trask Men's Sadler Penny Loafer
7 customer reviews

Price: $127.99 - $195.00 & Free Return on some sizes and colors

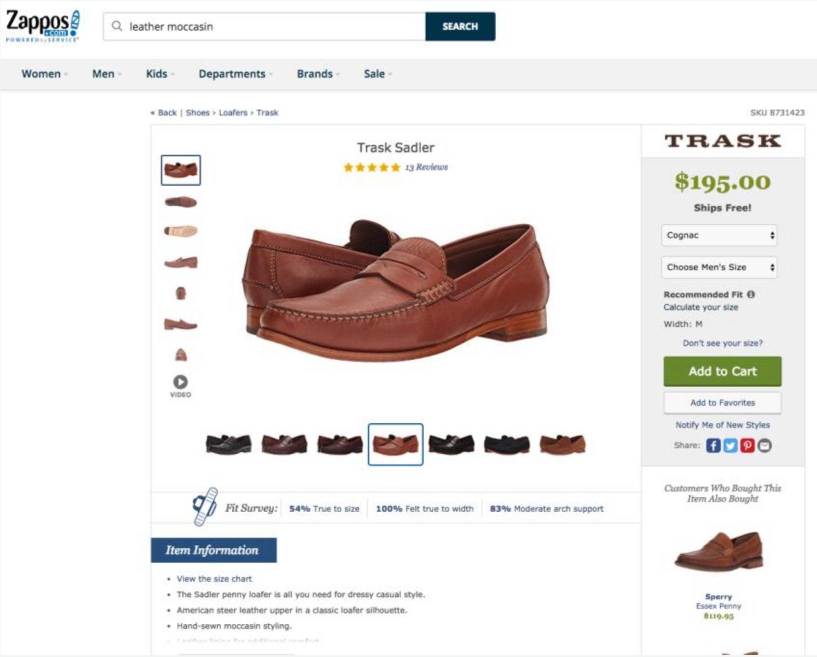
Fit: As expected (82%) ›

Size: Select Size Chart

Color: Dark Brown

- 100% Leather
- Imported
- Leather sole
- Leather lined
- Flexible sole
- Non-marking outsole
- Comfort casual loafer
- Shock-absorbing midsole

Roll over image to zoom in



SKU 8731423

Trask Sadler
13 Reviews

TRASK

$195.00

Ships Free!

Cognac

Choose Men's Size

Recommended Fit
Calculate your size
Width: M
Don't see your size?

Fit Survey: **54%** True to size **100%** Felt true to width **83%** Moderate arch support

Item Information

- View the size chart
- The Sadler penny loafer is all you need for dressy casual style.
- American steer leather upper in a classic loafer silhouette.
- Hand-sewn moccasin styling.

Customers Who Bought This Item Also Bought

Sperry
Essex Penny
$119.95

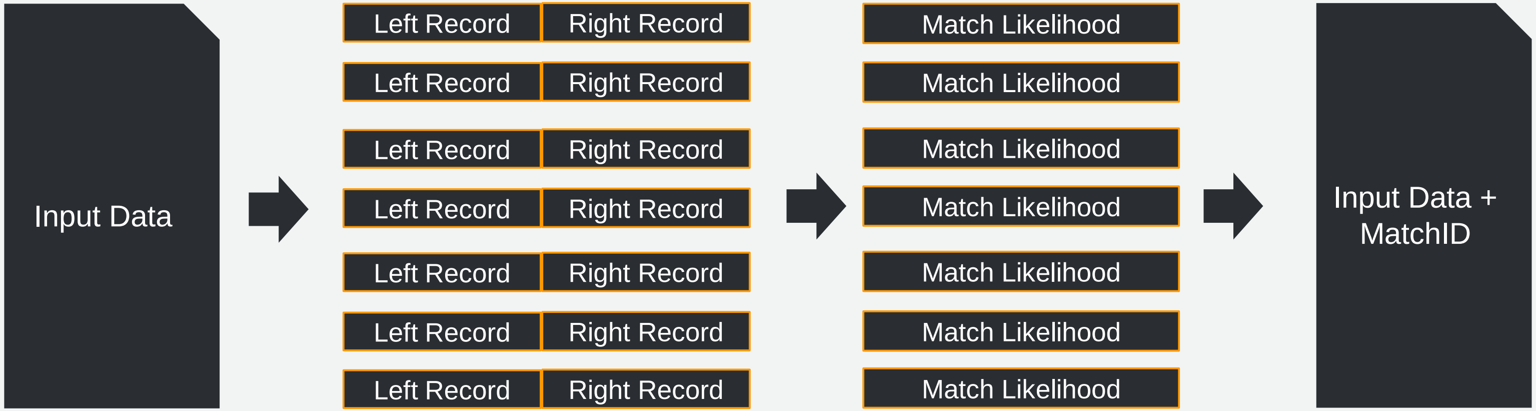
Data integration for People

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