AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

ANT303

Democratizing data for self-service analytics and ML

Greg Khairallah
Director, WW Analytics GTM Specialist
AWS

Shruti Worlikar
Senior Manager, AMER Analytics Specialist SA
AWS

Clay Martin
Senior Data Scientist
Jobcase

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Achieve business outcomes

Easy analytics for everyone

Amazon Redshift ML

Demo

Jobcase use case

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What if I can democratize data?

ACHIEVE YOUR BUSINESS OUTCOMES



Improve operational efficiency

Increase agility, run your business more efficiently, and provide a better experience for customers by moving data to the cloud



Make more informed decisions

Inform business decision making with more meaningful insights by bringing together the full picture of data across your organization



Accelerate innovation

Unlock opportunities that were either too difficult or impossible to do before by automating processes with AI and machine learning

By making 10% more data accessible, a typical Fortune 1000 company will see a ***$65 million increase in net income***[1]

[1] Dykes, "The Four Key Pillars To Fostering A Data-Driven Culture," www.forbes.com/sites/brentdykes/2019/03/28/the-four-key-pillars-to-fostering-a-data-driven-culture/?sh=45d40bee7d90

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why is making data available so hard?



Data is messy, sits in many locations, including SaaS applications and third-party data



Skills to build data pipelines is hard



Managing instance types and performance monitoring are a pain



Machine learning (ML) drives innovation, but ML skills are hard to find



Data is needed by other systems, not just reports

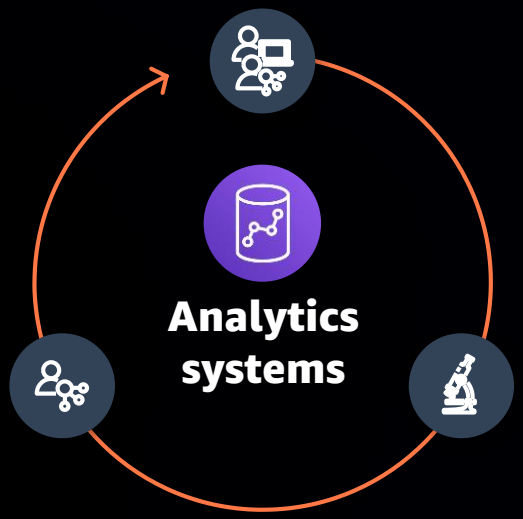


Business analysts have questions, but BI tools provide reports

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Easy analytics for everyone

FOCUS ON GETTING FROM DATA TO INSIGHTS IN SECONDS



Data engineers

Ingest and prepare data

Business leaders

Make business decisions and deliver on outcomes

Developers, analysts, data scientists

Build applications and analyze data

Automatic provisioning and scaling

Automatically provisions and scales the underlying compute resources to deliver high performance for demanding and unpredictable workloads

Visualize your data

Single, visual interface for querying data to improve productivity through one-click visual analytics, collaboration, version control, and scheduling

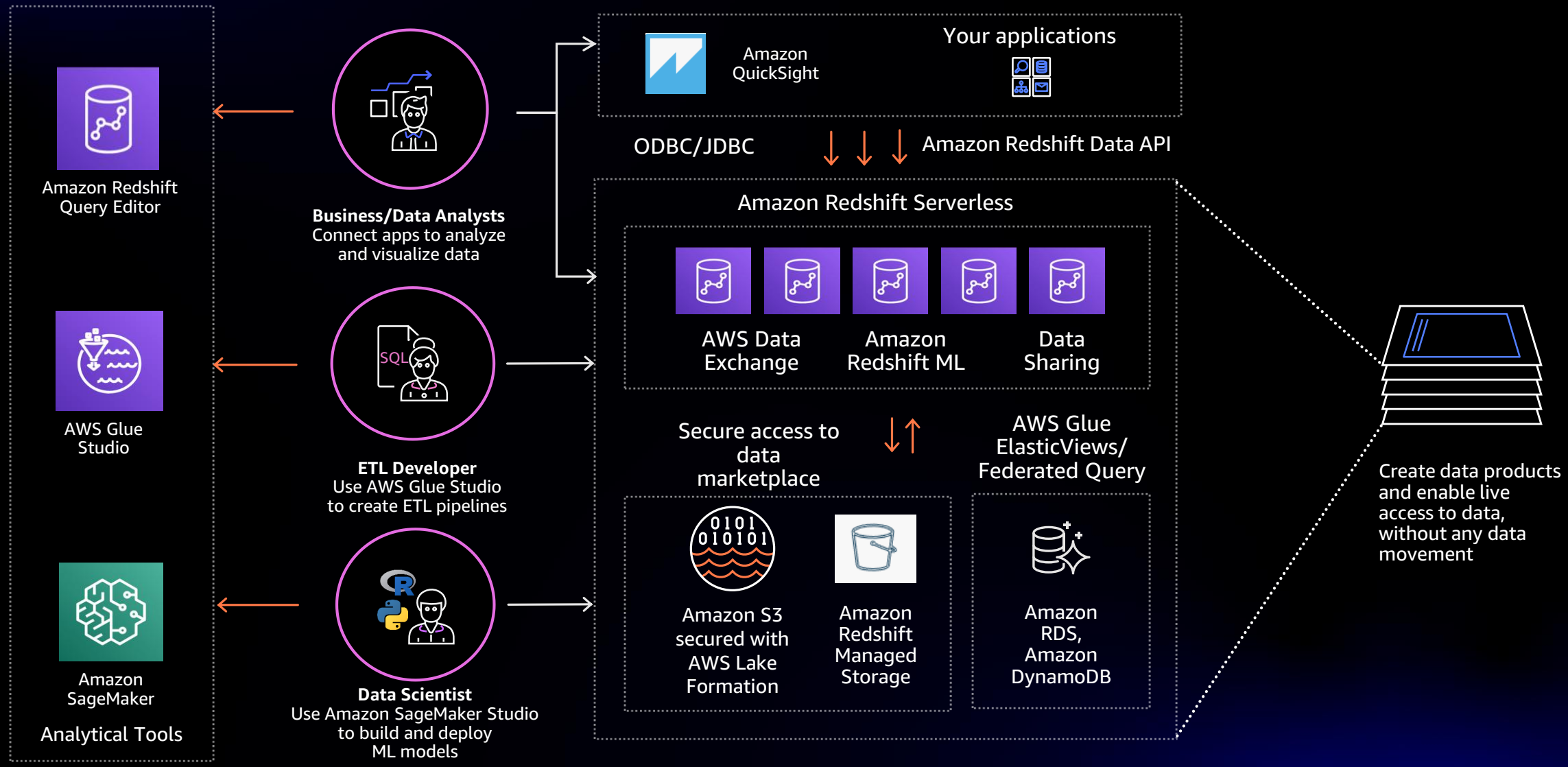
Bypass administrative tasks

Take advantage of automated provisioning, backup, patching, tuning, and monitoring in Amazon Redshift

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Democratize access to all your data



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Augment your data with machine learning

POWERED BY AMAZON QUICKSIGHT



ML predictions

Visualize and build predictive dashboards with Amazon SageMaker models



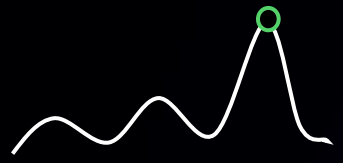
QuickSight Q

Ask questions using natural language



Autonarratives

Summarize your business metrics in plain language



Anomaly detection

Discover unexpected trends and outliers against millions of business metrics



Forecasting

Machine learning forecasting with point-and-click simplicity

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo workflow



Data Engineer (Miguel)

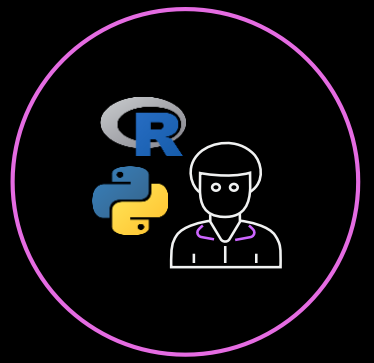
Load Sales Data
from Amazon S3 to Amazon Redshift using AWS Glue Studio



Business/Data Analyst (Jordan)

Analyze Sales Data
using Amazon Redshift Query Editor v2

Visualize Sales Data
using Amazon QuickSight Q



Data Scientist (Ravi)

Predict Sales
using Amazon Redshift ML

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Democratizing data in action

Jobcase job search content recommendation use case

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Who is Jobcase

Jobcase is an online community dedicated to empowering the world's workers

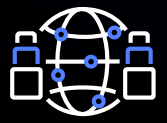
Third-largest destination for job search in the United States



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Democratizing data: Overview

jobcase



Barriers to democratizing data and ML: Data size and data/model co-location



Overcoming these barriers with Amazon Redshift and Amazon Redshift ML



Designing scalable recommender systems, or, how to build a recommender system without leaving Amazon Redshift

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Democratization at scale

jobcase

Scale can be measured in two dimensions



Data/storage scale – measured in terms of terabytes or billions of rows in a table



Organization/user scale – measured by the number of people in an organization that can unlock value in your data

Democratizing data is about the organization/user scale

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Barriers to democratizing data: Size



We store terabytes of data in highly optimized, compressed tables; these data capture 8+ years of email, arrival, job impression, and user data

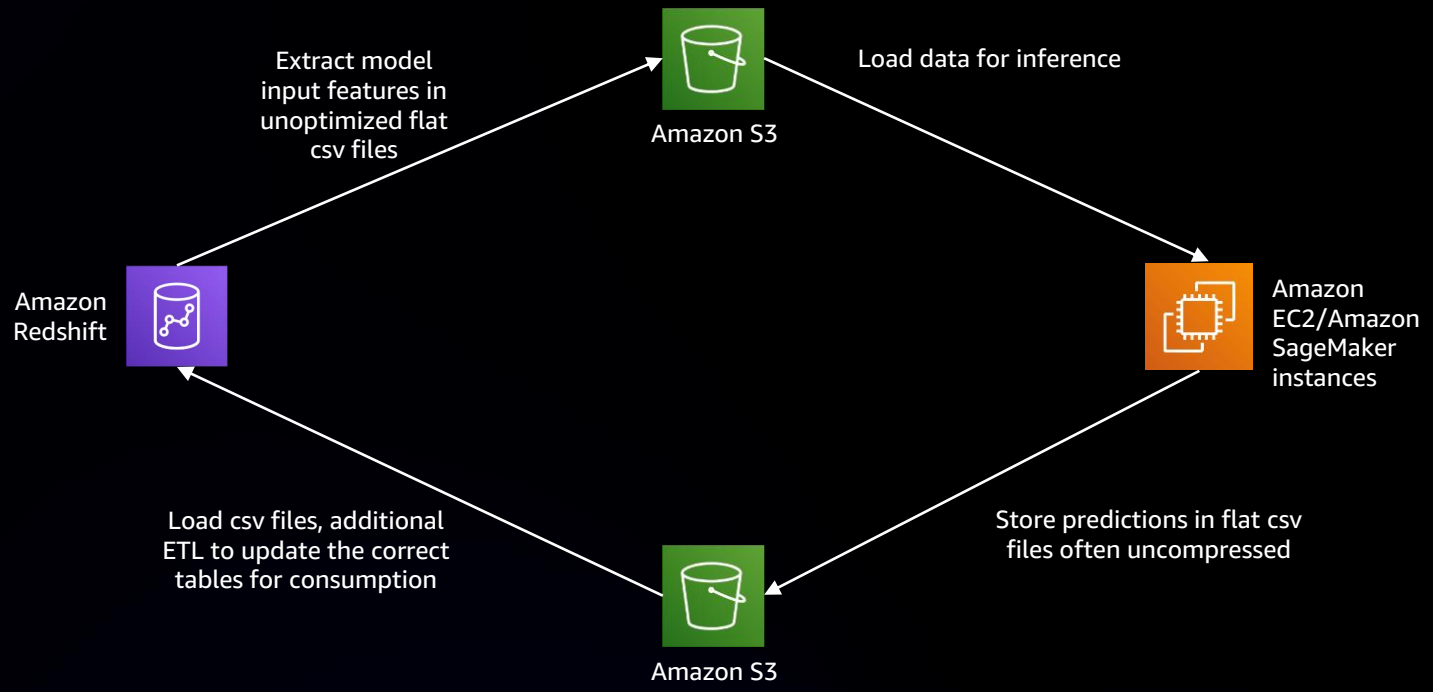
But the data structures of choice for analysts and data scientists, such as R and pandas dataframes, do not scale to these data sizes

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Barriers to democratizing ML: Data model co-location

Typical data science pipeline before Amazon Redshift ML



Old data pipeline – a slow, expensive, and error-prone process

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Solution: Bring the models to the data

Typical data science pipeline after Amazon Redshift ML

Can you spot the difference?



Amazon Redshift

Batch inference happens in place in the instances where the distributed data resides using optimized Amazon SageMaker Neo models

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Jobcase: Job recommendation



Generate job search content for over 20+ million daily active members



Content includes 20–30 million available job listings and 1+ million search terms and companies



Scoring every user-item combination is not feasible

Recommended items can be specific job listings, job search queries, companies, etc.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recommender system design: Reduce candidate sets

jobcase

Reduce the candidate set of items per user to around 300–500 via unsupervised/collaborative filtering techniques

Technique	Compute within Amazon Redshift?
Classical data mining: Jaccard distance, cosine similarity	Yes (compute efficiently using probabilistic approximate count – HyperLogLog)
Clustering algorithms	Not yet
Vector embeddings (e.g., GloVe or word2vec)	Not yet

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Example item clusters

jobcase

Item similarities calculated within Amazon Redshift, based on Jaccard distance

If you searched for …	You might also be interested in …
Data scientist	Research scientist, data analyst, business analyst
Warehouse worker	Material handler, forklift operator, warehouse associate
Registered nurse	School nurse, staff nurse, RN case manager
Front office assistant	Receptionist, mailroom clerk, file clerk, office assistant

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recommender system design: Scoring

Score reduced candidate set of 2+ billion user-item pairs with a model trained via supervised learning

Feature engineering done within Amazon Redshift



User-level features
Education, skills, years of experience, etc.



Item-level features
Skill requirement, ONET code, etc.



User-item–level features
Similarity score, previous user-item engagements, etc.

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recommender system design: Scoring

jobcase

The model class we choose depends on the key performance indicator (KPI) we wish to maximize

We score billions of user-item pairs in minutes with Amazon Redshift ML

KPI	Model class	Compute in Amazon Redshift?
Engagement rate/ clickthrough rate (CTR)	Logistic regression, feedforward neural network (MLP), XGBoost	Yes
Value per send	Linear regression, feedforward neural network, XGBoost	Yes

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Summary

jobcase

- Provides access to building and deploying ML models in production at scale to a broad set of analysts through easy-to-use SQL interface
- Running model predictions at scale, performing 2-3 billion predictions in 15 minutes
- Significant reduction in the model development cycle from weeks to days by eliminating complex data pipelines and being able to run tests of bigger cohorts
- No cost increase since Amazon Redshift ML uses existing cluster resources to do local in-database inference
- 5–10% improvement in member engagement rates across several different email template types

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next steps

Visit or view on-demand related sessions

Keynote by Swami Sivasubramanian

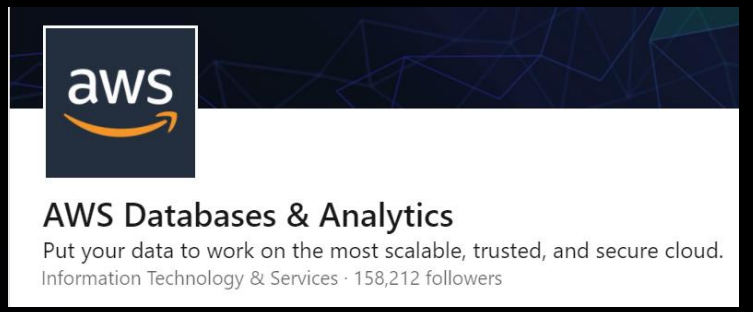
Leadership session by Rahul Pathak

ANT301: Concurrency and scalability strategies with Amazon Redshift

ANT215: Introduction to AWS Data Exchange for Amazon Redshift

Follow news on LinkedIn

linkedin.com/showcase/aws-databases/



aws.amazon.com/redshift/

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Greg Khairallah
gregkh@amazon.com

Shruti Worlikar
shrutiw@amazon.com

Clay Martin
clay@jobcase.com

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.