AWS re:Invent

AMZ304

Prime Video: Processing analytics at petabyte scale

Parwinder Singh
Sr. Data Engineer
Amazon.com – Prime Video

Karthik Kumar Odapally
Sr. Solutions Architect
Amazon Web Services

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

A brief history

AWS’ification

Metric governance

Implementation examples

- Clickstream & playback

Conclusion

A brief history

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AMAZON ORIGINAL

The Marvelous Mrs. Maisel




AMAZON ORIGINAL
TOM CLANCY'S
JACK
RYAN



PREMIER LEAGUE LIVE 3-5 DEC MUN v TOT LIV v EVE BUR v MCI prime video

amazon.co.uk Try Prime

Prime Video

Hello, Sign in Account & Lists

Returns & Orders

Try Prime

0 Basket

Deliver to United States

Cyber Monday Sale | Christmas Shop | Vouchers | AmazonBasics | Best Sellers | Gift Ideas | New Releases | Gift Cards | Customer Service | Free Delivery | Shopper Toolkit | Sell

Prime Video | Amazon Originals | TV Shows | Movies | Kids | Your Watchlist | Your Video Library | Settings | Getting Started | Help

PREMIER LEAGUE LIVE ON PRIME VIDEO

prime video Premier League



AWS'ification

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

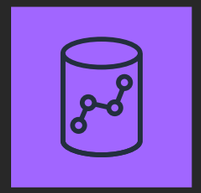
10,000-ft view

prime video



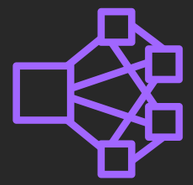
Amazon Elastic Compute Cloud (Amazon EC2)

Heavy lifting and light use



Amazon Redshift

Scalable data warehousing



Amazon EMR

Spark on EMR



Amazon Simple Storage Service (Amazon S3)

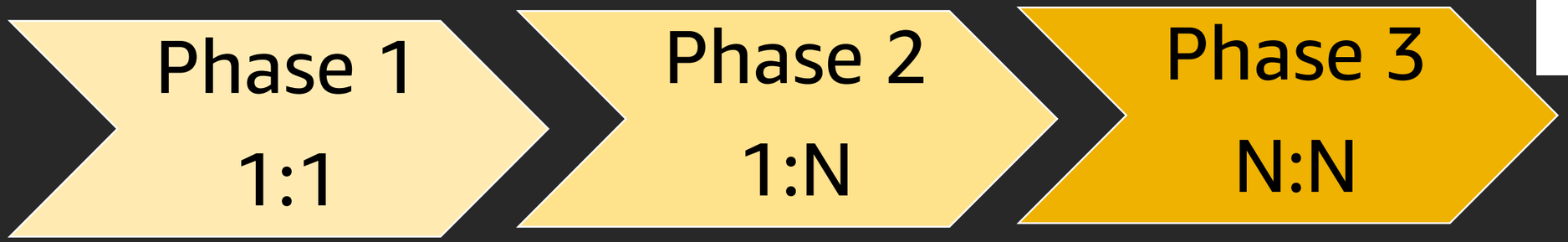
Data lake; intermediary storage



AWS Lambda

No operational overhead

Team history



Phase 1: A centralized team

Phase 2: A failed governance attempt

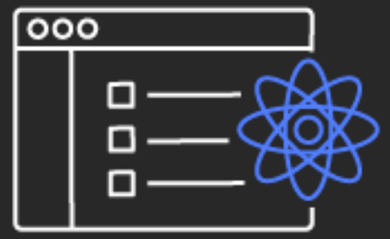
Phase 3: Needs mechanism

Mechanisms

prime video



Metric governance



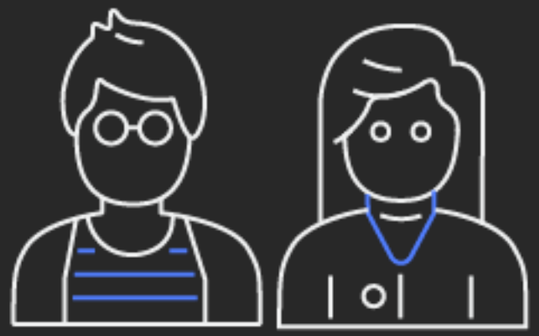
Business logic encapsulated data platform



Role-based access control for data lake

Metric governance

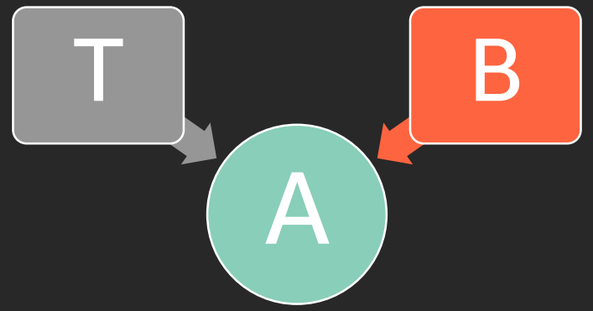
prime video



Scheduled meetings



Document



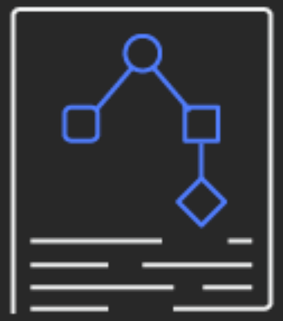
Approval chains

Business logic encapsulated data platform

prime video



Data primitives



Apply business logic

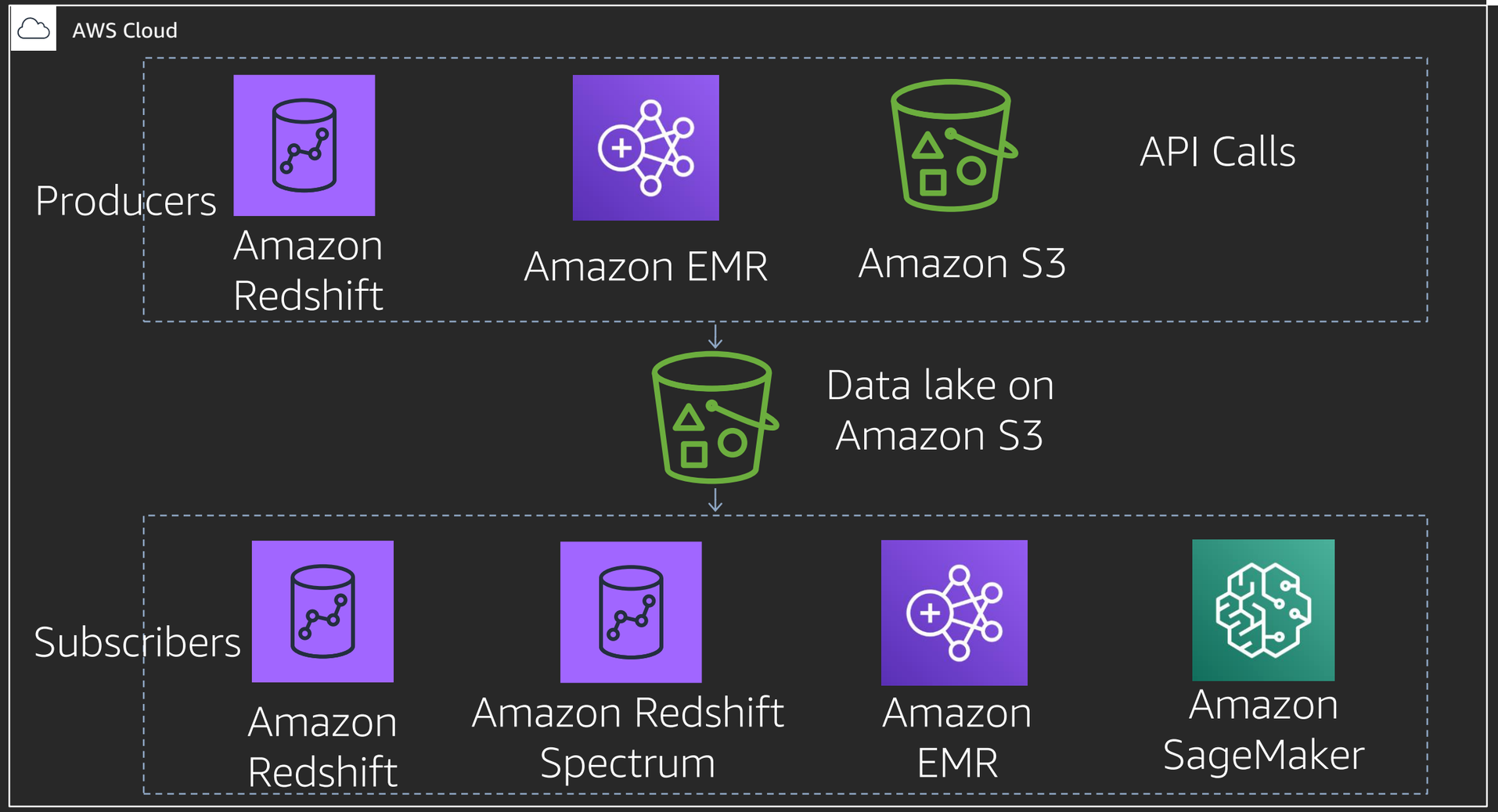


Data lake

Tables Views

Role-based access control for data lake

prime video



Data lake

prime video

Implementation

- Encryption
  - In transit
  - At rest
- Web UI (non-tech) and API (tech)
- Integrated user communication
- Easy data discoverability



Data lake

prime video

Features

- Role-based access control
- Backed by S3, based on partitions
  - Replace partition
  - Update partition
- Managed schema changes
  - In place
  - Replace

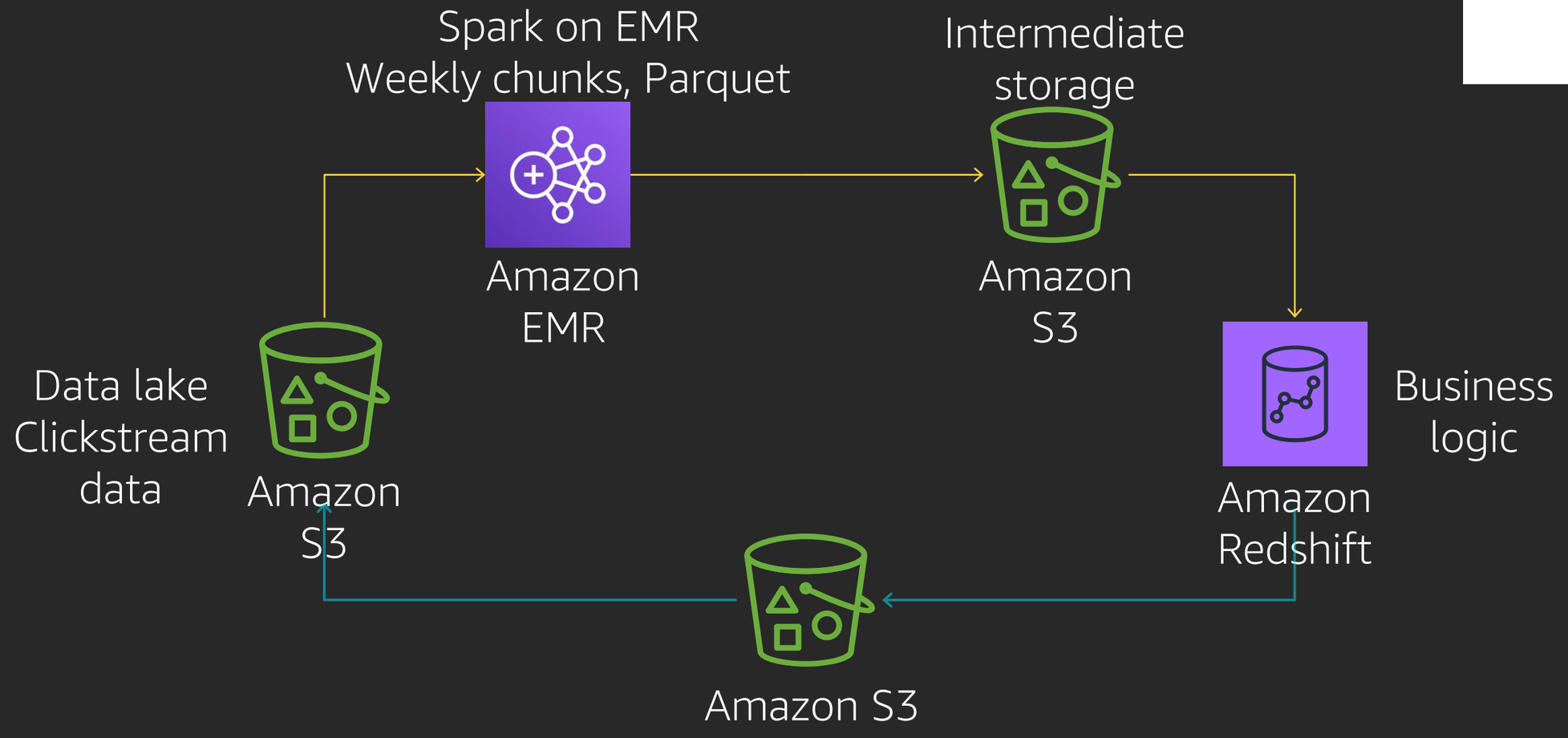


Implementation examples: Clickstream

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.


Business Challenges

Clickstream data




Technical Challenges

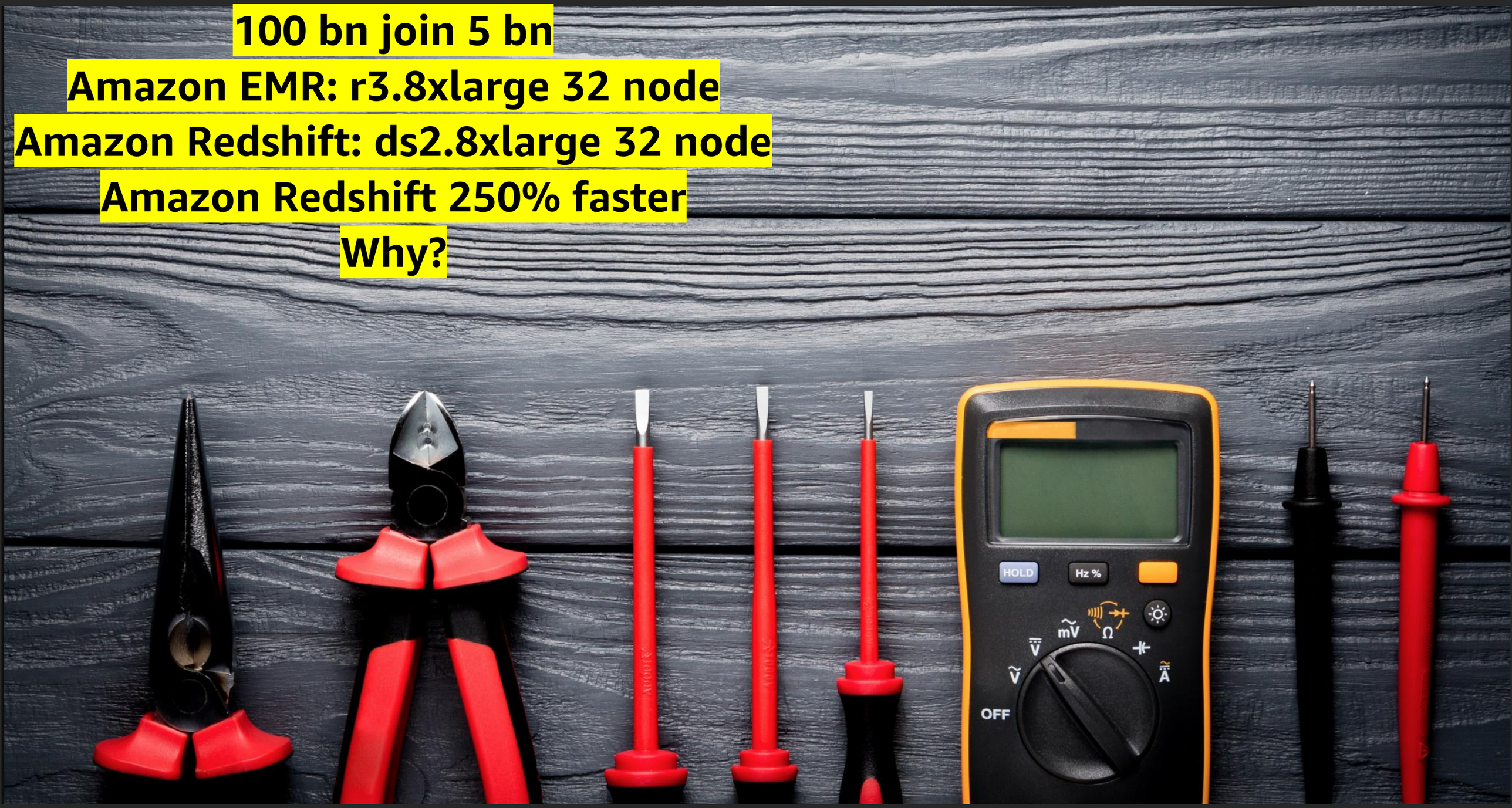
100 bn join 5 bn

Amazon EMR: r3.8xlarge 32 node

Amazon Redshift: ds2.8xlarge 32 node

Amazon Redshift 250% faster

Why?



Amazon Redshift best practices

prime video

- Distribution key
  - High cardinality
  - Same column (and data type) for tables usually joined (e.g., customer_id)
- Sort key
  - Commonly used for filtering, joining, or both
  - Don’t use an interleaved sort key on monotonically increasing attributes, i.e., timestamps, identity columns

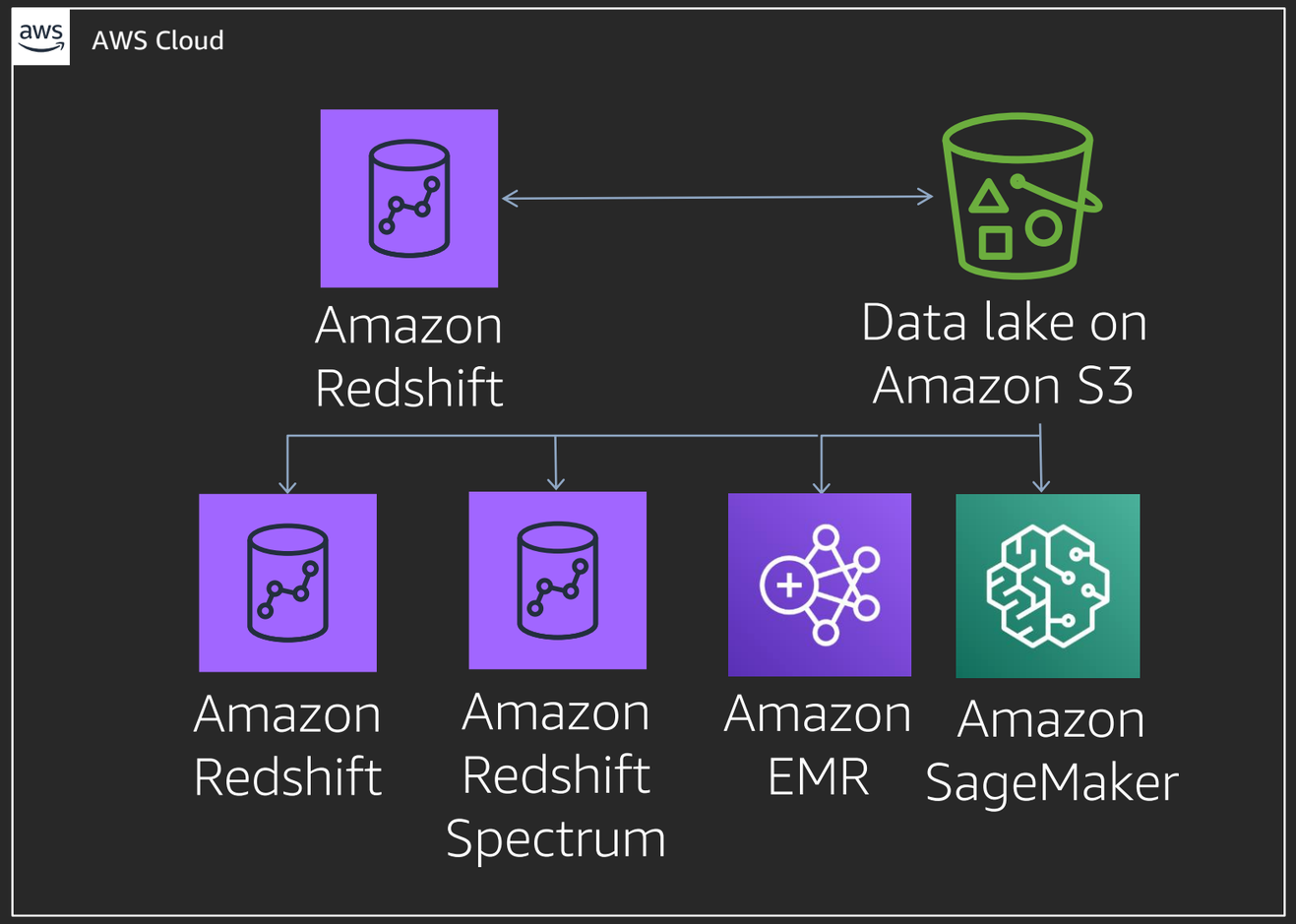


Implementation examples: Playback

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

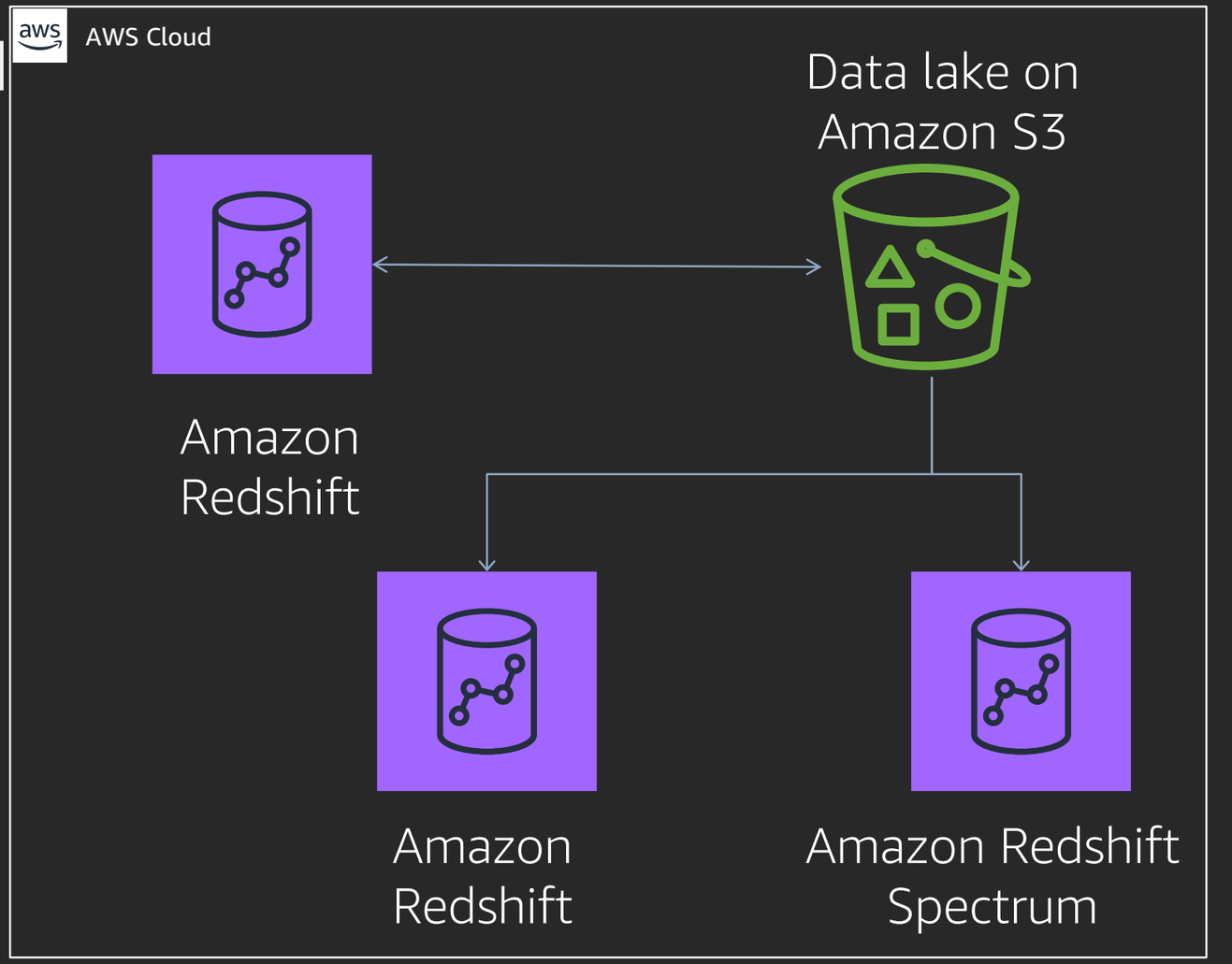
Playback: Lifeblood of Prime Video

prime video



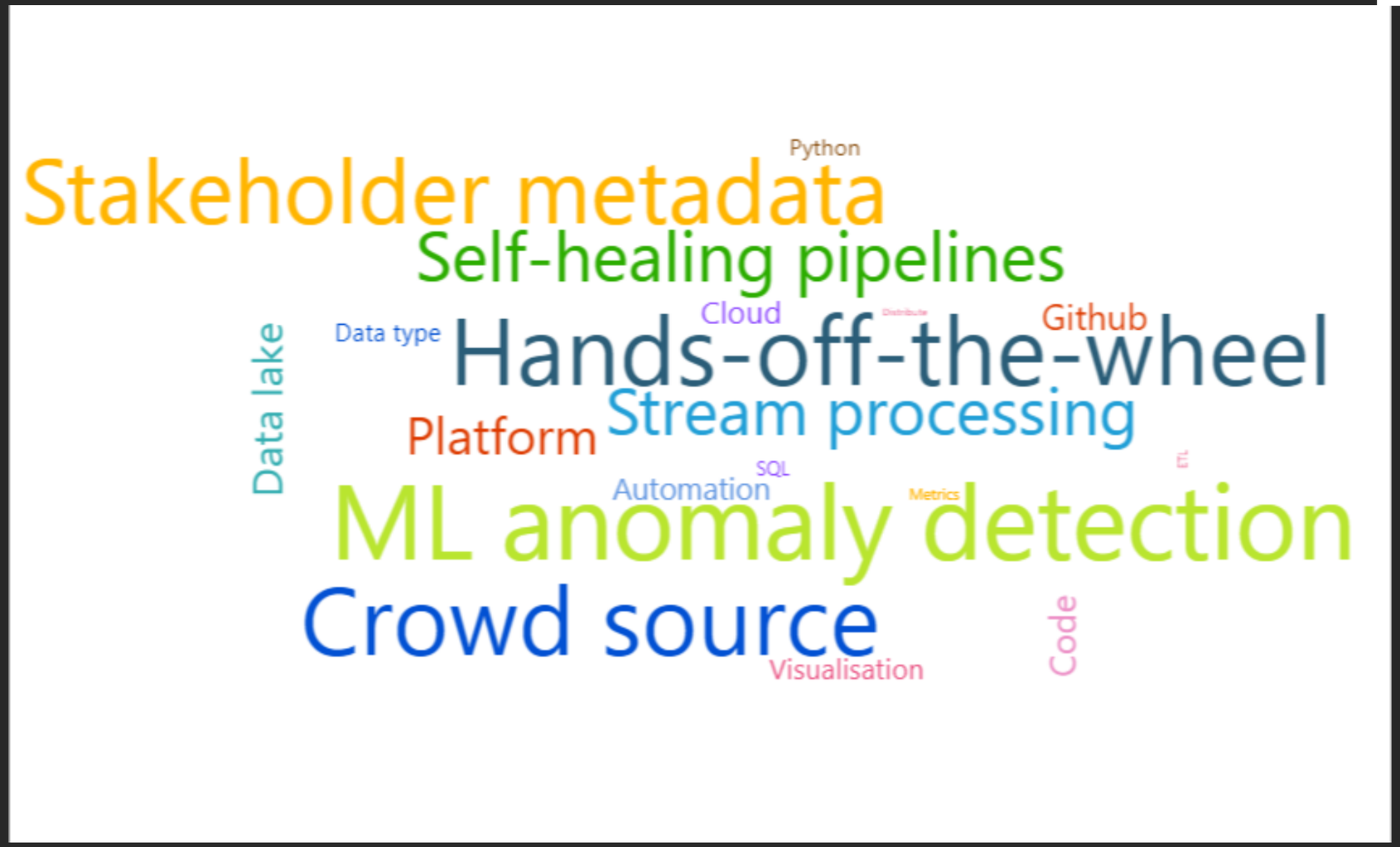
Playback data view

- Instant response to business need
  - Change in business logic
  - Bad data
  - Event
  - Row/column-level security (currently, not natively supported in Amazon Redshift)

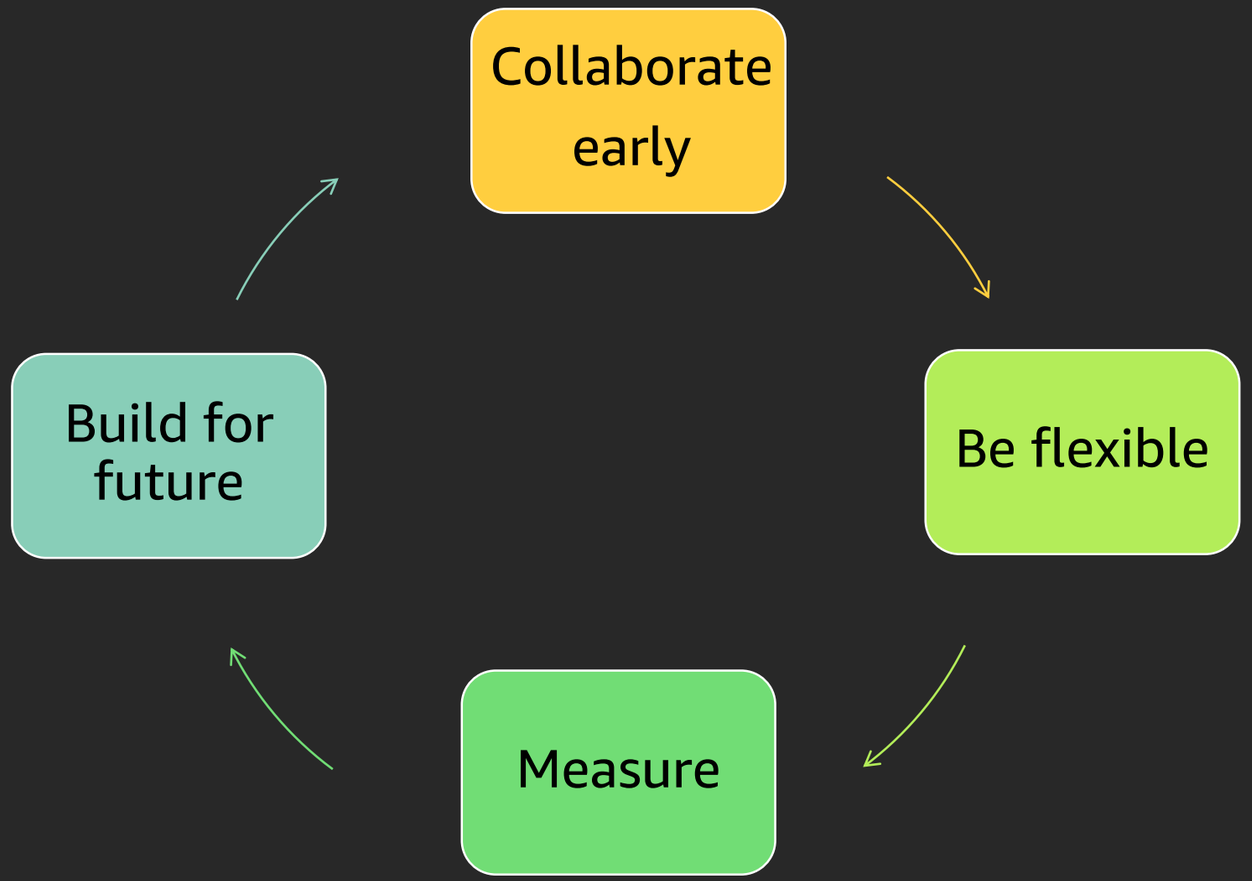


Back to the future

prime video



Conclusion



“If we can partition, we can scale”

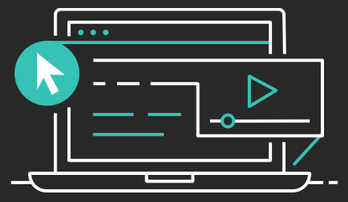
Key takeaways



- Data lake
  - Managed schema changes
  - Views on data lake
- Architecture evolution
  - Understand constraints of the in-house tech
  - Perform data-driven PoC
- Centralized pool of EMR clusters

Learn analytics with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills



New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, Amazon Redshift, Amazon Athena, and AWS Lake Formation



Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs



Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws training and certification

Thank you!

Parwinder Singh
parwinde@amazon.com

Karthik Kumar Odapally
karthio@amazon.com

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey in the mobile app.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.