AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

STG353-NEW

Introducing Amazon S3 Metadata

Paul Meighan

Director, Product Management,
Amazon S3 & Amazon S3 Glacier
AWS

Seth Markle

Senior Principal Engineer,
AWS Storage
AWS

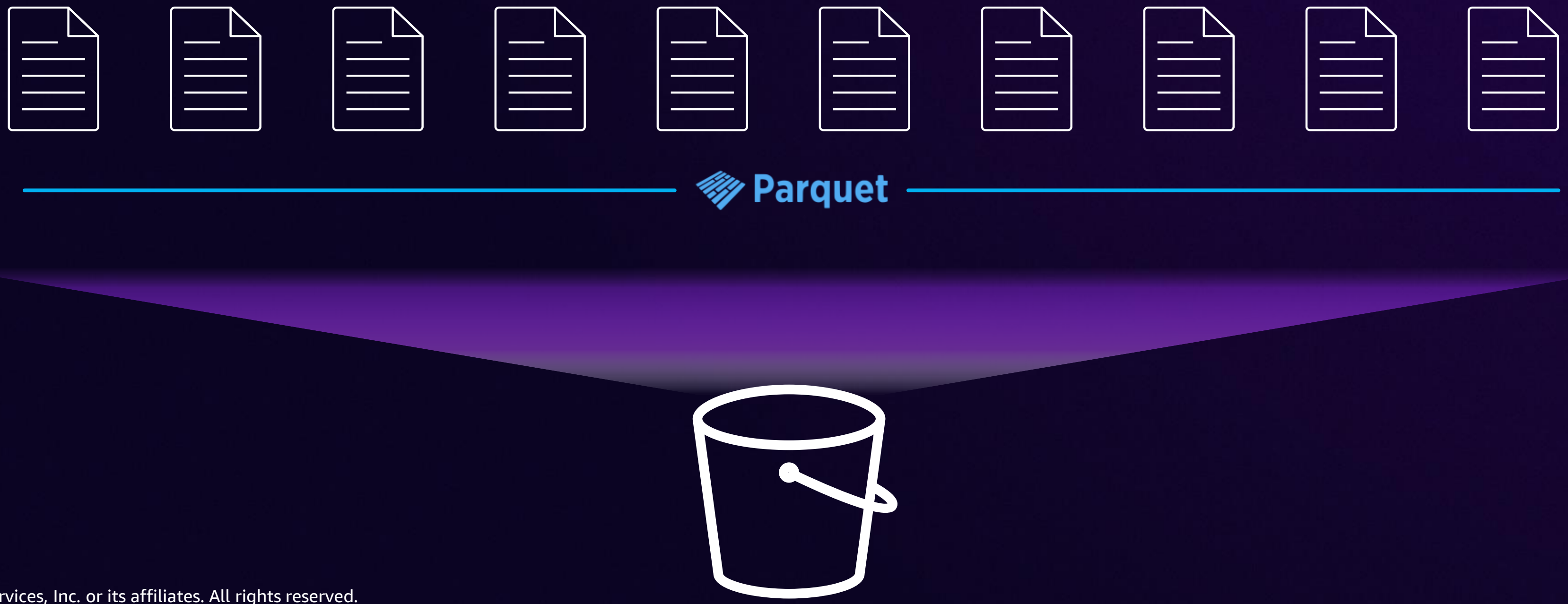
Structured data in Amazon S3

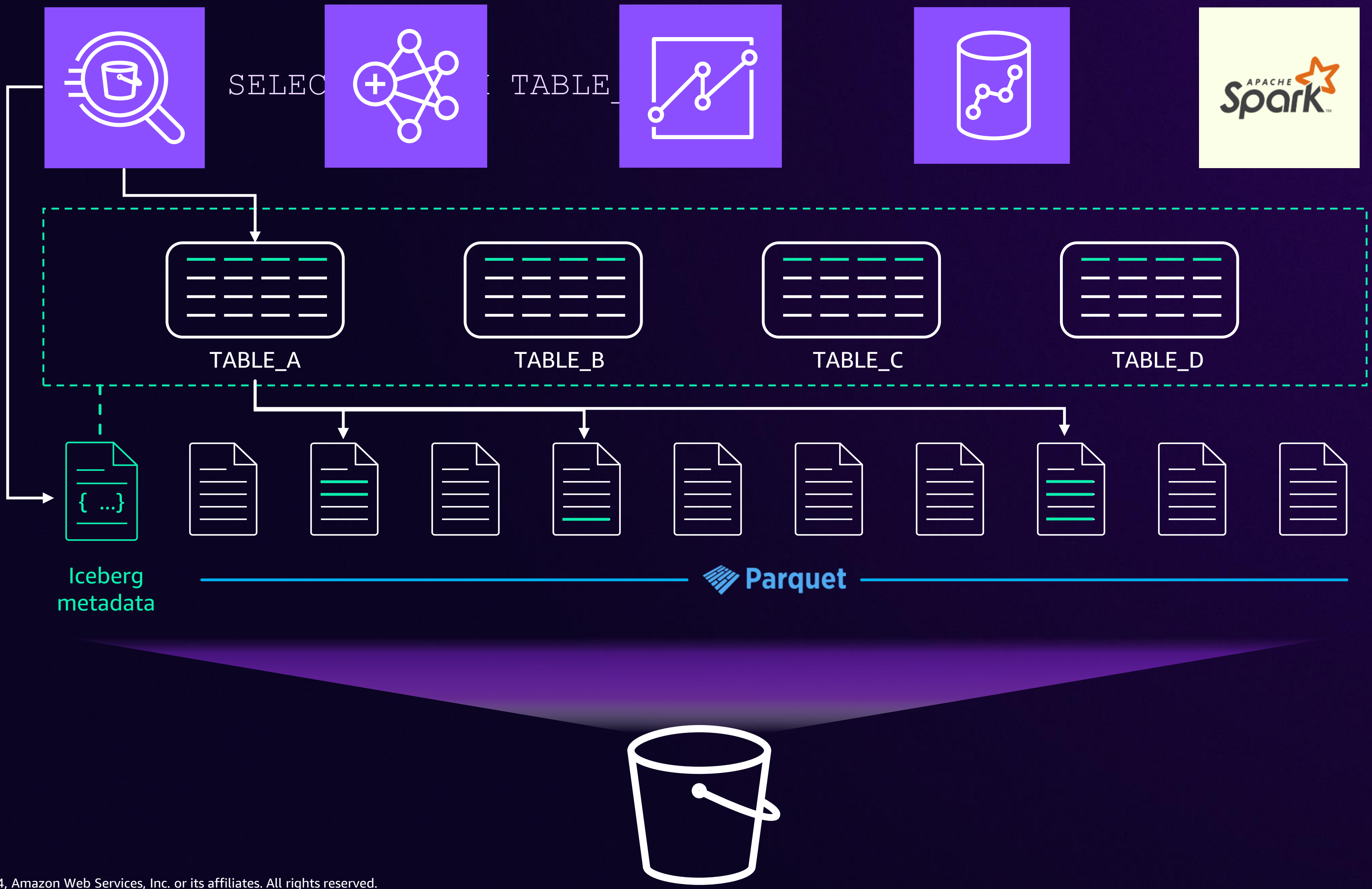
Today, Amazon S3 is also a *tabular* data store

Exabytes of Parquet data stored

15 million+ requests per second

Serve hundreds of petabytes every day




SELEC
TABLE_
Apache Spark
TABLE_A
TABLE_B
TABLE_C
TABLE_D
{ ...}
Iceberg metadata
Parquet

GA Dec. 3, 2024

Amazon S3 Tables

Fully managed Apache Iceberg tables in Amazon S3

Unstructured data in Amazon S3

Customer problems to solve

Finding the right data to complete a task takes longer and longer

Metadata stores are difficult to build, live outside of storage

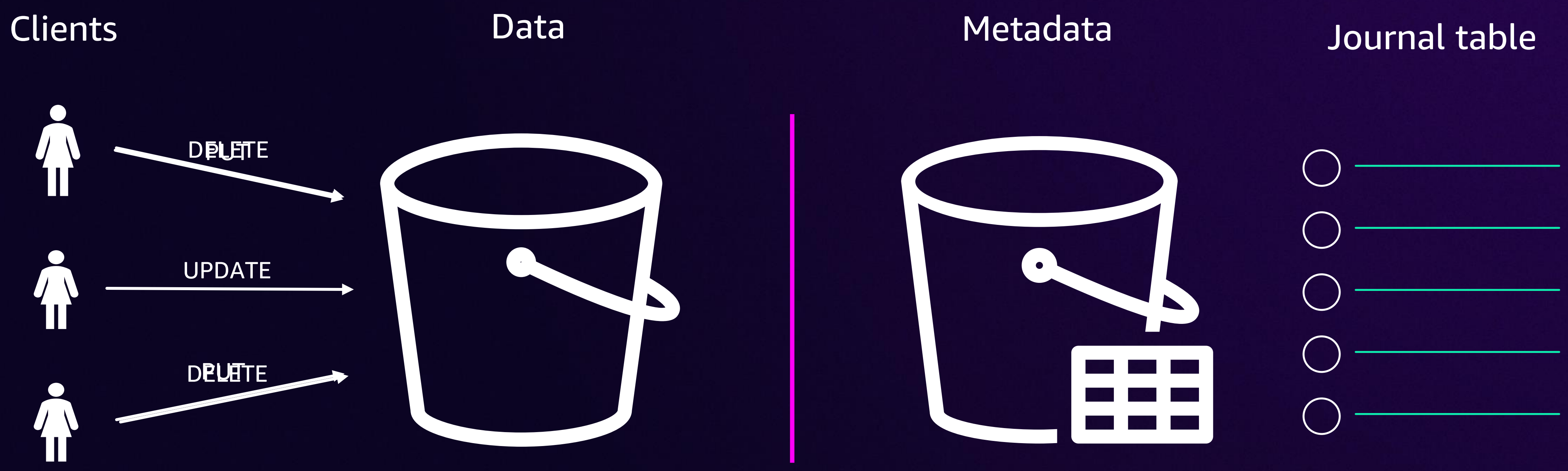
Metadata is only as useful as it is current

PREVIEW

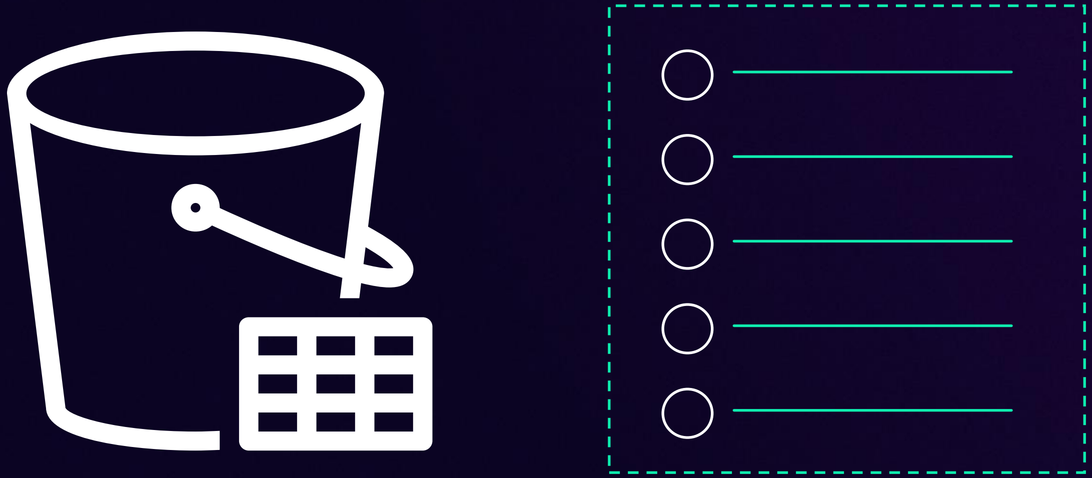
PREVIEW Dec. 3, 2024

Amazon S3 Metadata

Automatic metadata generation, accessible with simple SQL semantics


Clients
Data
Metadata
Journal table
DELETE
PUT
UPDATE
DELETE
PUT

Journal table



AWS-managed table

Read-only for non-AWS principals

It’s a “system table” for your bucket

Metadata	Data Type
bucket	String
key	String
sequence_number	String
record_type	String
record_timestamp	Timestamp
version_id	String
is_delete_marker	Boolean
size	Long
last_modified_date	Timestamp
e_tag	String
storage_class	String
is_multipart	Boolean
encryption_status	String
is_bucket_key_enabled	Boolean
kms_key_arn	String
checksum_algorithm	String
object_tags	Map <String, String>
user_metadata	Map <String, String>
requester	String
source_ip_address	String
request_id	String

21 system metadata fields automatically recorded for all new objects in your bucket

Object size
Storage class
Create date
Client information
Encryption information
…

Use cases

Find objects to process

```
// List the objects in my bucket (without using the LIST API)
SELECT bucket, key, storage_class, user_metadata
FROM aws_s3_metadata.my_metadata_table
WHERE key LIKE '{SOME_PREFIX}%’;
```

Understand and track data lineage

```
// Who is deleting data? What are they deleting?
SELECT DISTINCT bucket, key, record_type, record_timestamp, requester, source_ip_address
FROM aws_s3_metadata.my_metadata_table
WHERE record_type = 'DELETE’;
```

Learn about storage usage

```
// How much storage are my tagged objects using?
SELECT object_tags[‘MyTag'] as mytag, storage_class, count(*) as count,
sum(size) / 1024 / 1024 as usage
FROM aws_s3_metadata.my_metadata_table
GROUP BY tags['Module'], storage_class;
```

Custom metadata

Object tags

Mutable
Dedicated APIs
Dedicated permissions
Paid Amazon S3 feature

User-defined

Immutable
Only added on PUT
No cost

Joined tables

Mutable
Fully customized
Detached from objects
Cost of table storage

Video generated via Amazon Bedrock is automatically annotated with custom metadata

Amazon Bedrock

In Amazon Bedrock’s PUT request:

```
x-amz-meta-content-source: AmazonBedrock
x-amz-meta-content-model-id: arn:aws:::model/xyz-v1
```

Journal table

Query from Amazon S3 Metadata:

```
SELECT bucket, key, user_defined_metadata, row_type
FROM aws_s3_metadata.my_metadata_table
WHERE user_defined_metadata[‘content-source’] = AmazonBedrock
```

Thank you!

Please complete the session survey in the mobile app

Paul Meighan
@paulmeighan.bsky.social

Seth Markle