AWS re:Invent

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

AIM411-R2

Improve ML productivity through better feature engineering

David Ping
Senior Manager, AI/ML Solutions Architecture
AWS

Farooq Sabir
Senior AI/ML Specialist Solutions Architect
AWS

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

Agenda

Challenges with feature engineering

Amazon SageMaker Feature Store architecture

SageMaker Feature Store APIs

Extending SageMaker Feature Store functionality

Best practices

Demo

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

Where to find data?



DS in marketing

Discovering data sources and getting access are hard!



Data lake



Data warehouse



Transactions



Customers



Products



Graph



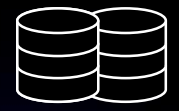
Amazon S3



Documents



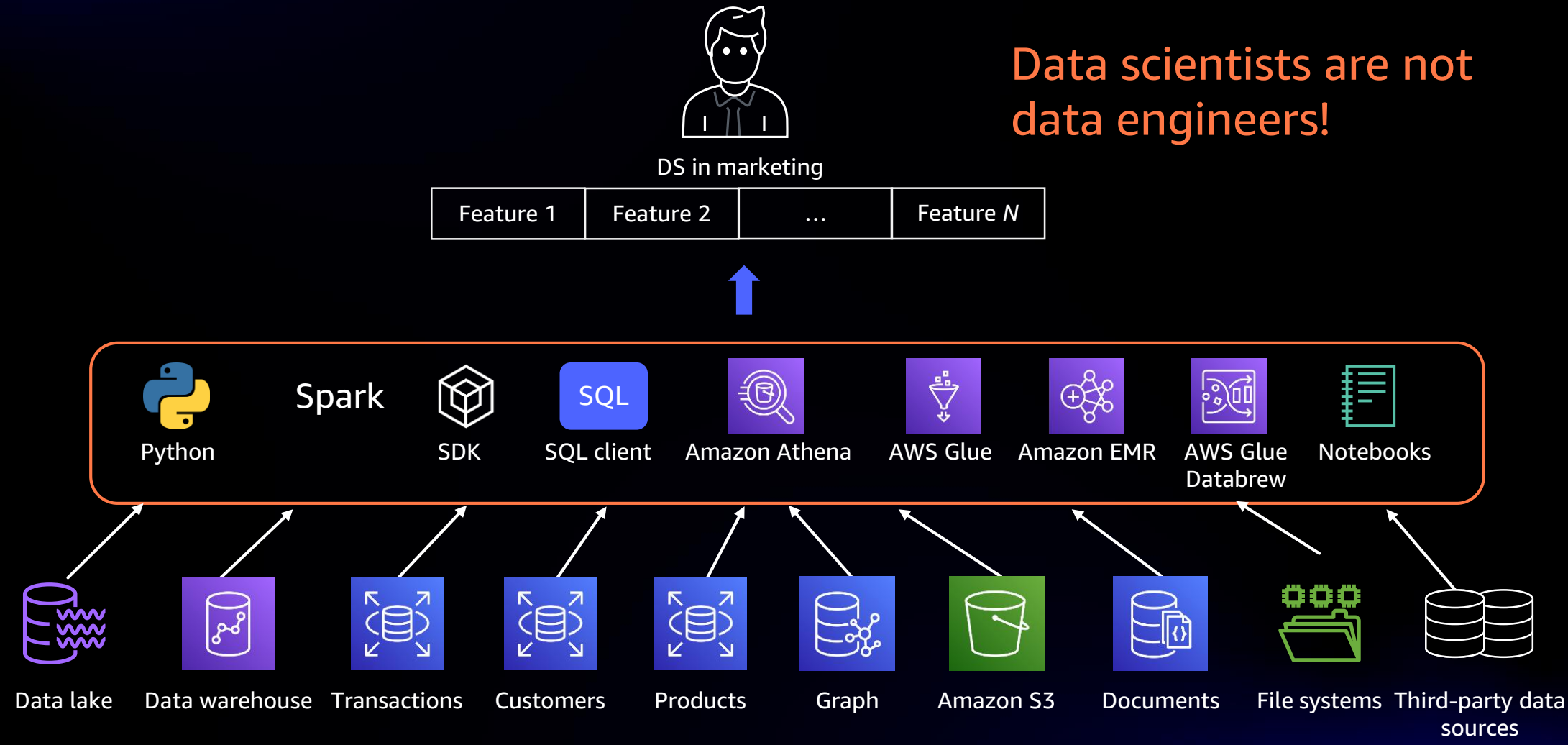
File systems



Third-party data sources

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

What tools to use?

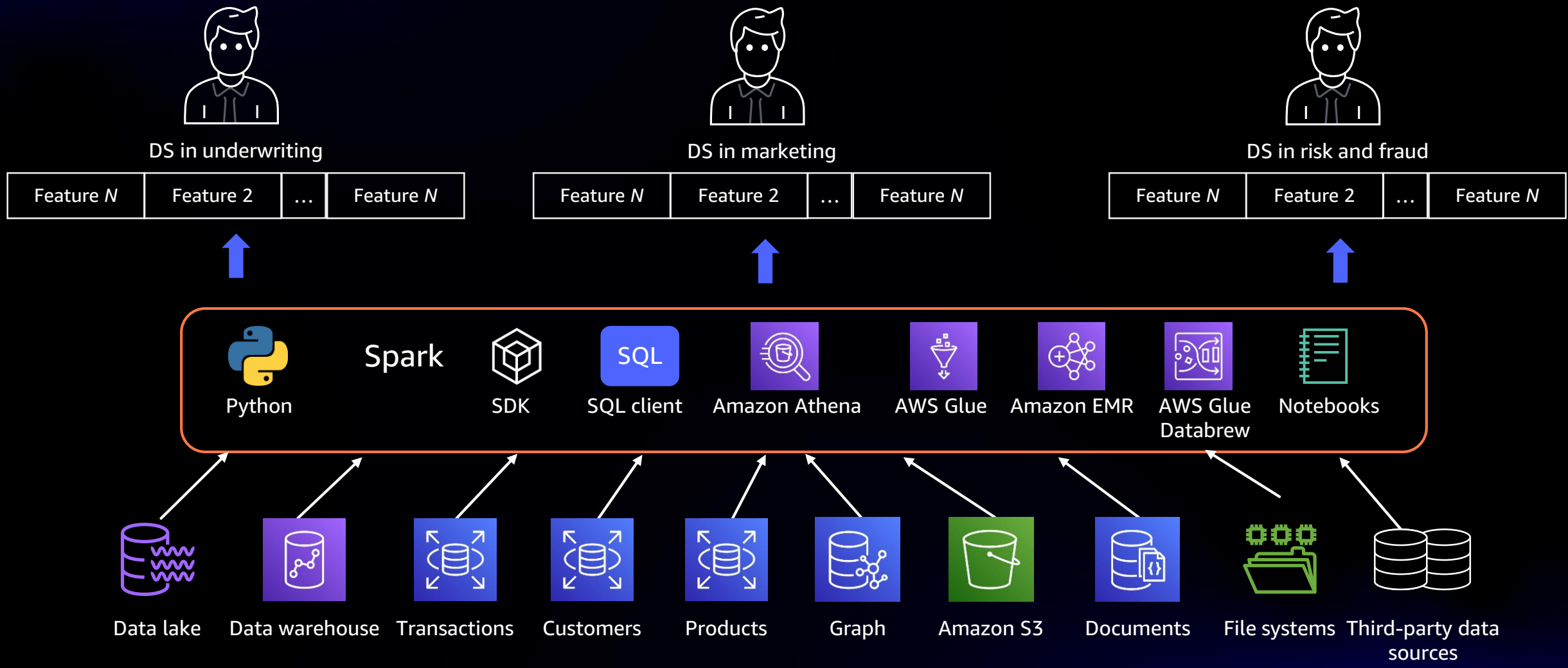


Data scientists are not data engineers!

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

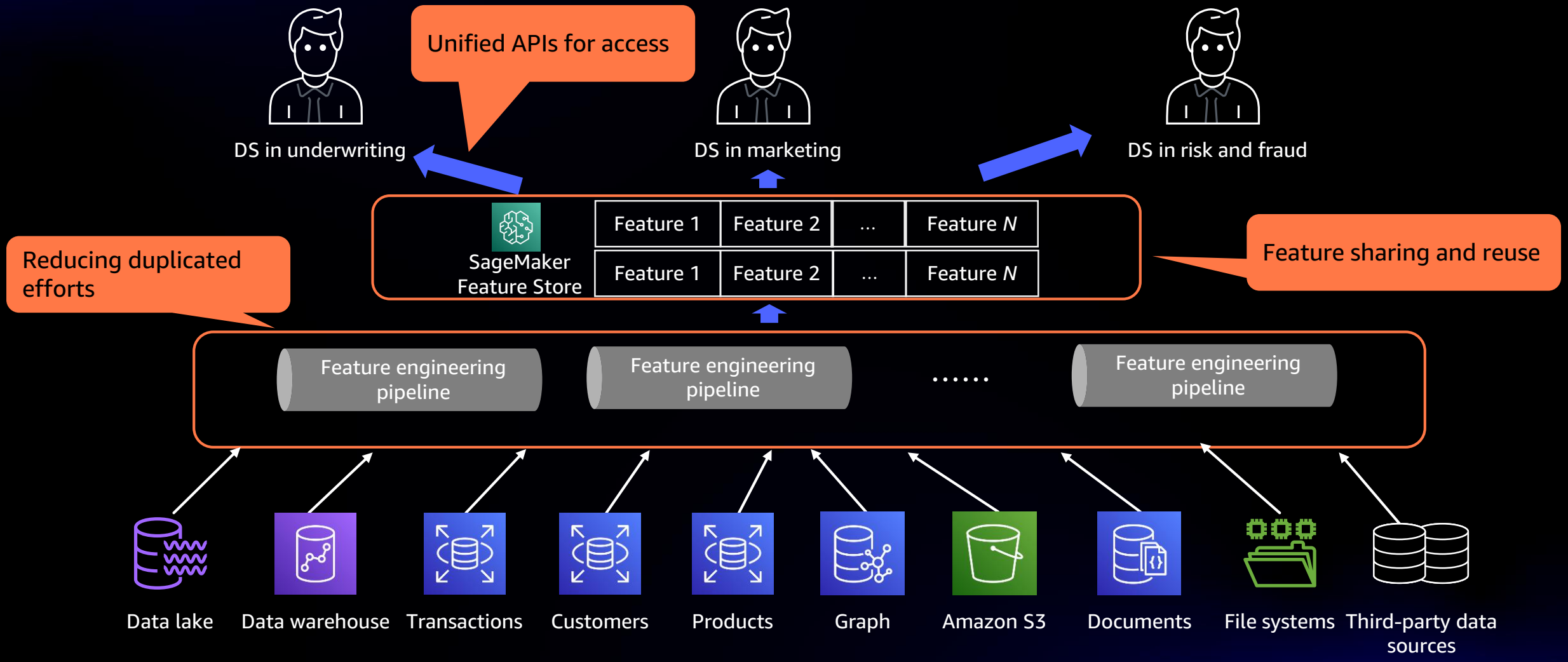
How to share?

Hard to share features across teams!



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

Feature store as the solution

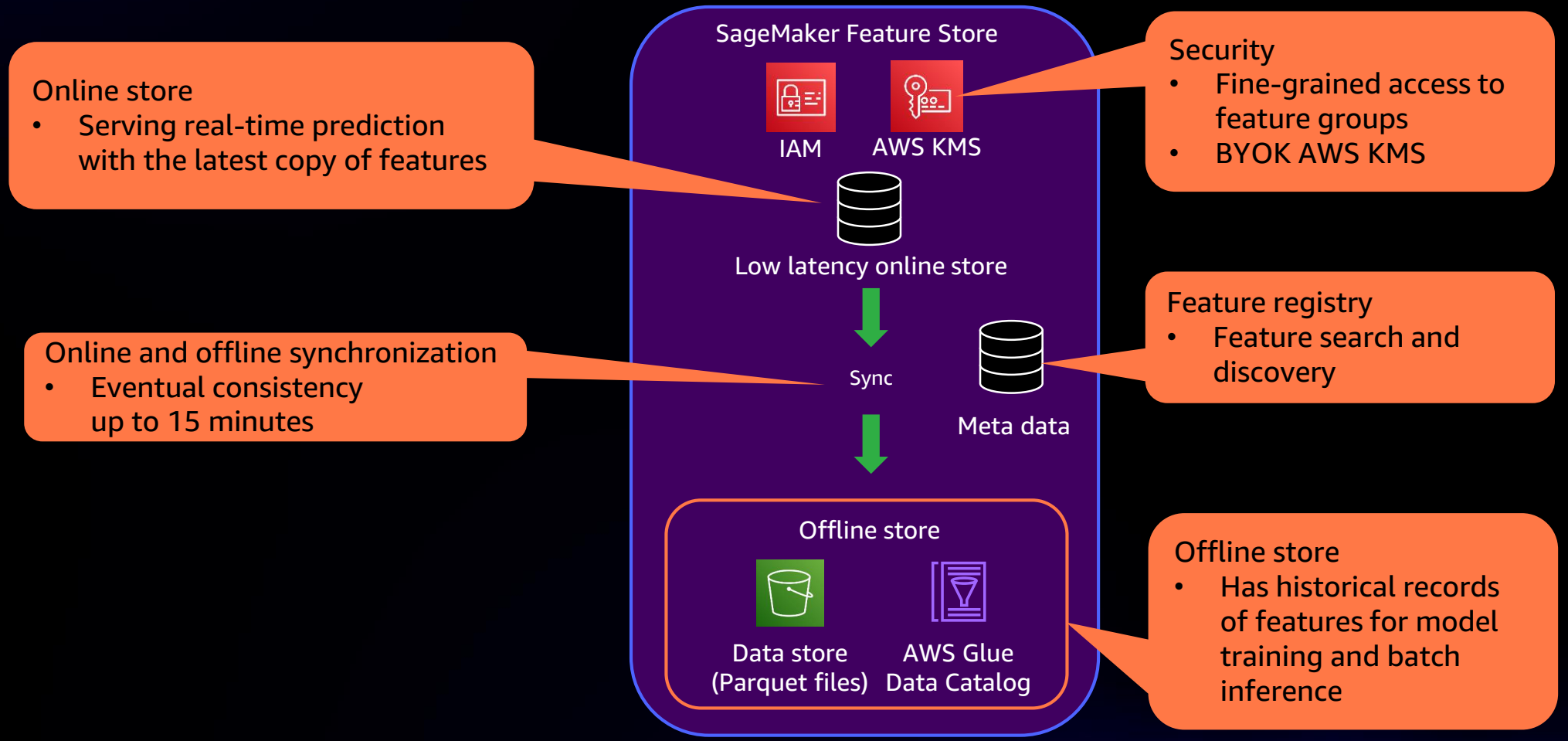


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture

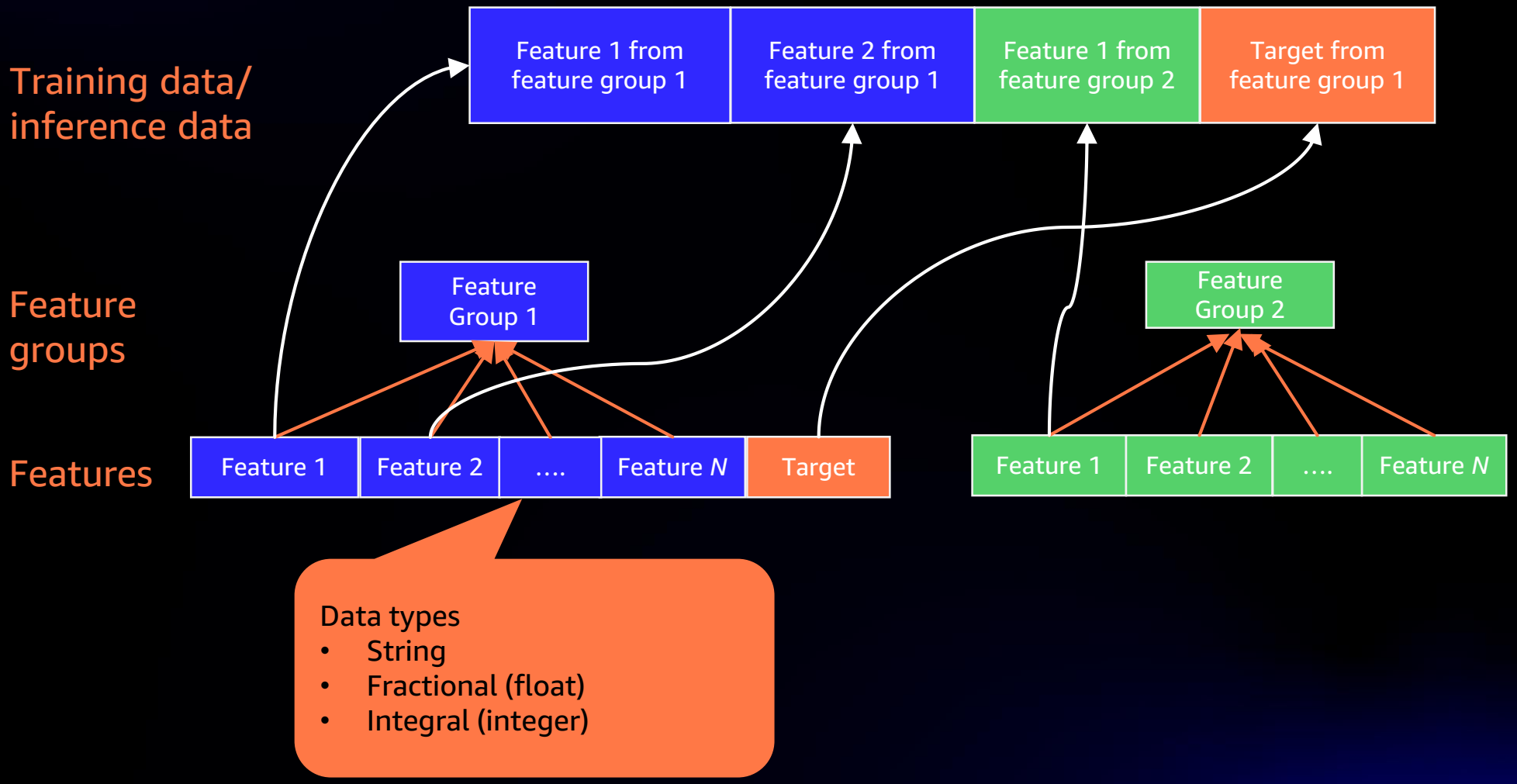
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Feature Store architecture



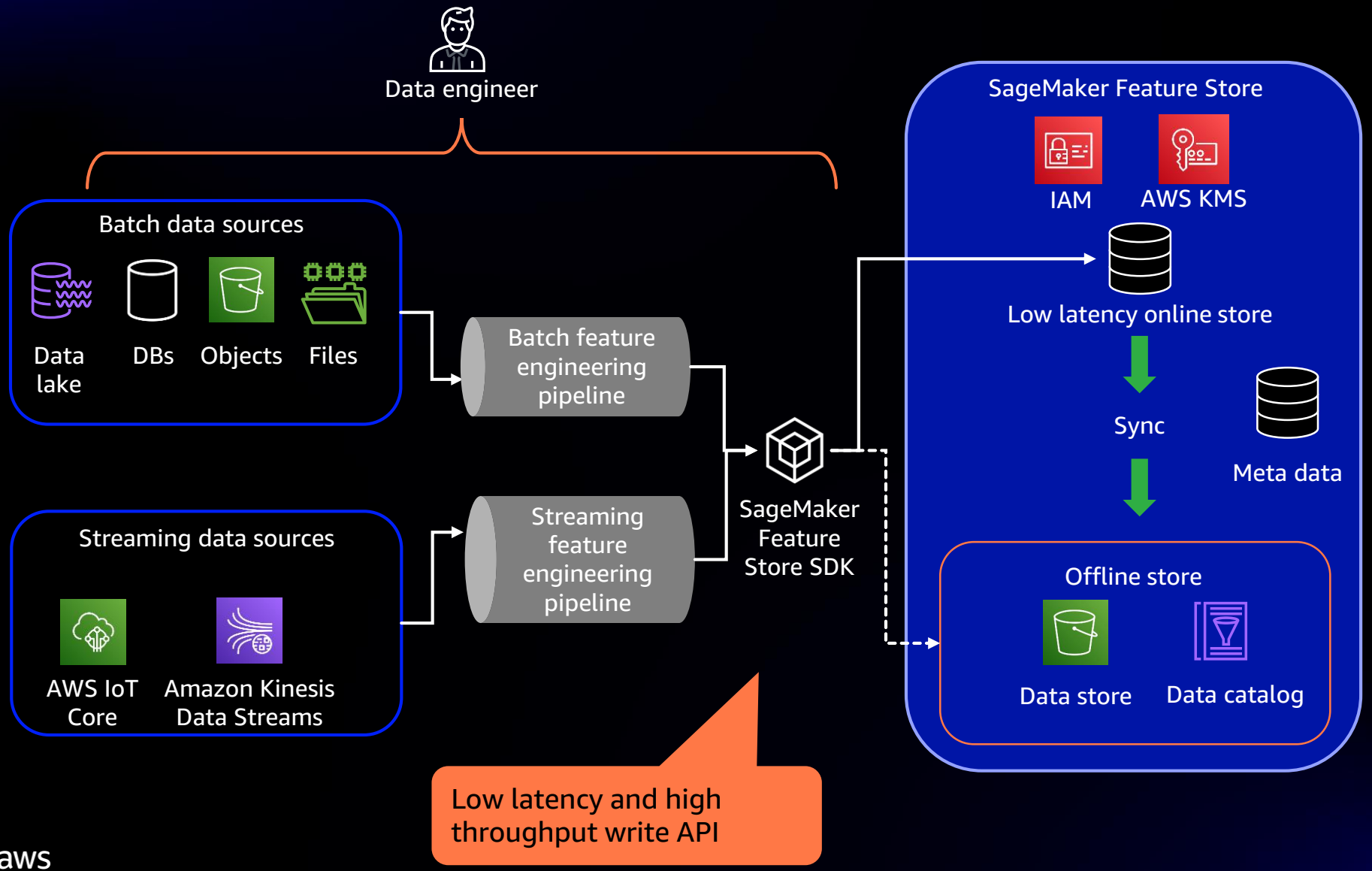
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How features are organized



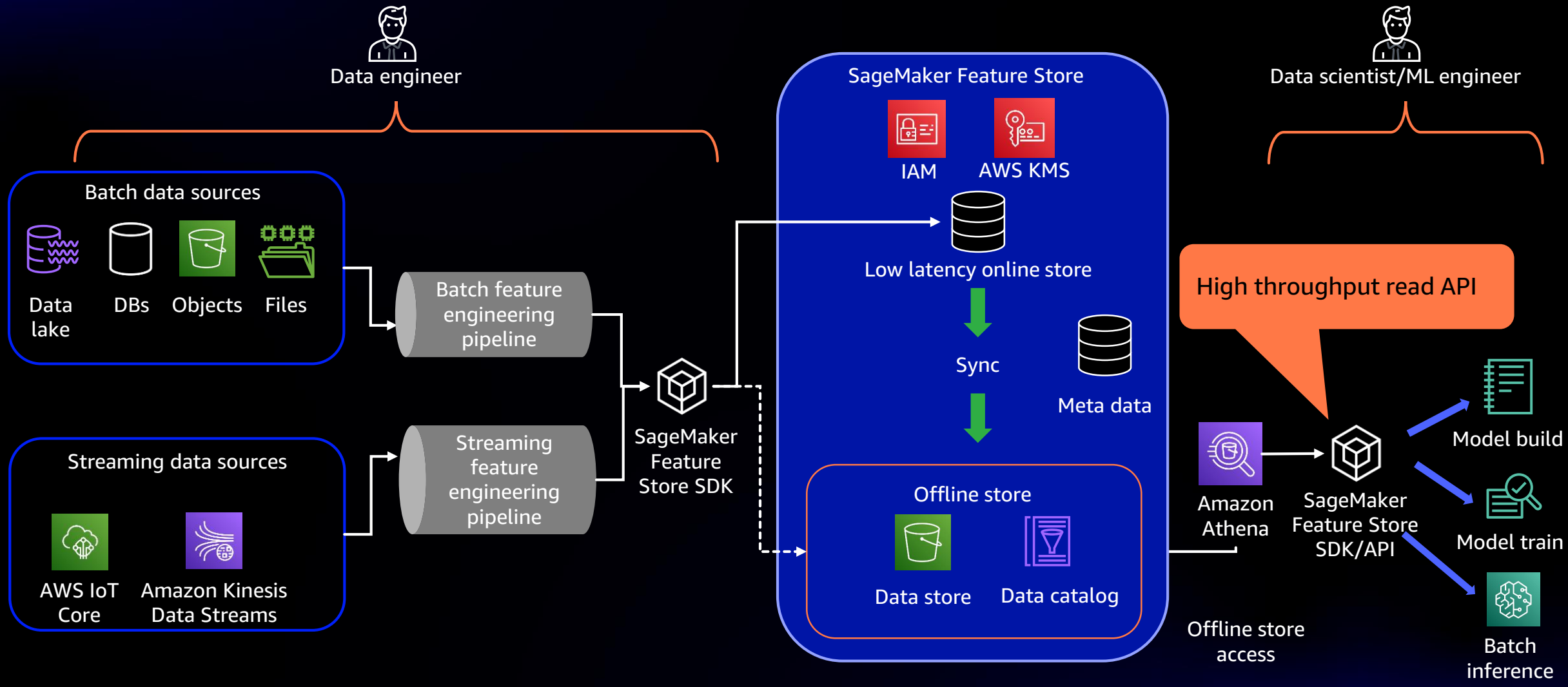
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Ingestion architecture



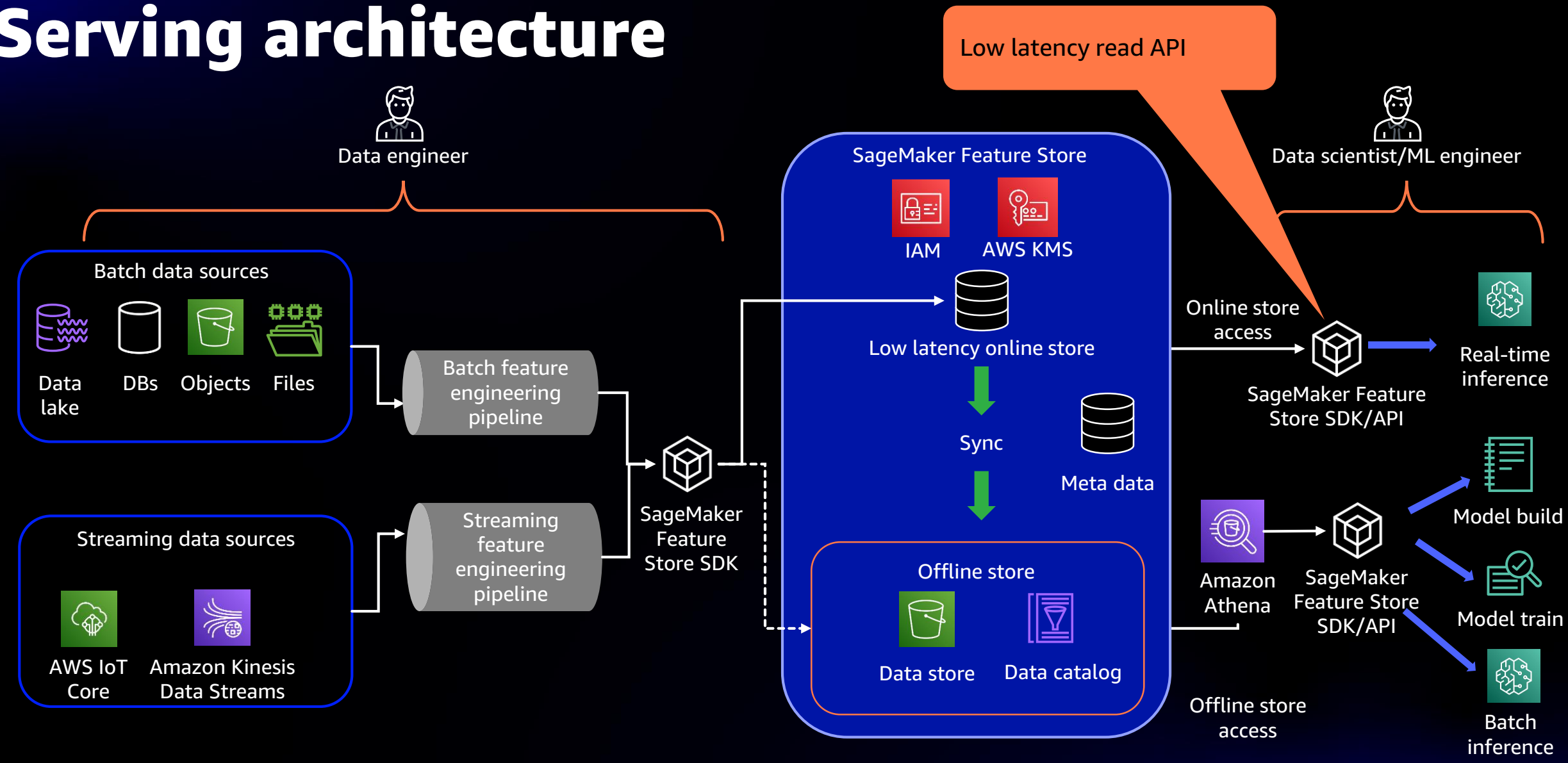
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Serving architecture



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

Serving architecture



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

APIs

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

Ingestion API

Python API

```
fg = FeatureGroup (
        name='fg-name',
        sagemaker_session=fs_session )

fg.ingest (
        data_frame=df,
        max_processes=20,
        max_worker=4 )
```

Target feature group

Pandas DataFrame for data to be ingested

Number of processes to create, each with max_worker threads

Number of threads to be created to work on different partitions of the data_frame

PySpark API

```
Spark_ingestor.ingest_data (
        input_data_frame=spark_df,
        feature_group_arn=feature_group,
        direct_offline_store=True )
```

Spark DataFrame for data to be ingested

Target feature group

Ingest data to offline Amazon S3 directly when true

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

Online store retrieval API

Get features for a single record from a single feature group

```
sm_fs.get_record(
        FeatureGroupName='fg-name',
        RecordIdenifierValueAsString='REC-001',
        FeatureName=['fname-1', 'fname-2'])
```

Record ID in the feature group for which to retrieve the features

Name of features for which to return the values

Get features for multiple records from multiple feature groups

```
sm_fs.batch_get_records(
        Identifier=[
        {FeatureGroupName': 'fg-name-1',
        'RecordIdentifiersValueAsString' :
        ['REC-001', 'REC-002', 'REC-003'],
        'FeatureNames' : ['fname-1' , 'fname-2']}
        {FeatureGroupName': 'fg-name-2',
        'RecordIdentifiersValueAsString' :
         ['REC-004', 'REC-005'],
        'FeatureNames' : ['fname-3', 'fname-4']}
        ]}
```

First feature group

Second feature group

aws

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

Offline store retrieval API

```
query = “SELECT  feature_1, feature 2,
            …
            feature_n  FROM table_name”

fg = FeatureGroup(
            name=‘fg-name’,
            sagemaker_session=fs_session)

fg.AthenaQuery.run (query_string = query,
            output_location = output_location
            )
```

SQL query string for selecting features from tables in the offline data store

Run query using Amazon Athena as the query engine

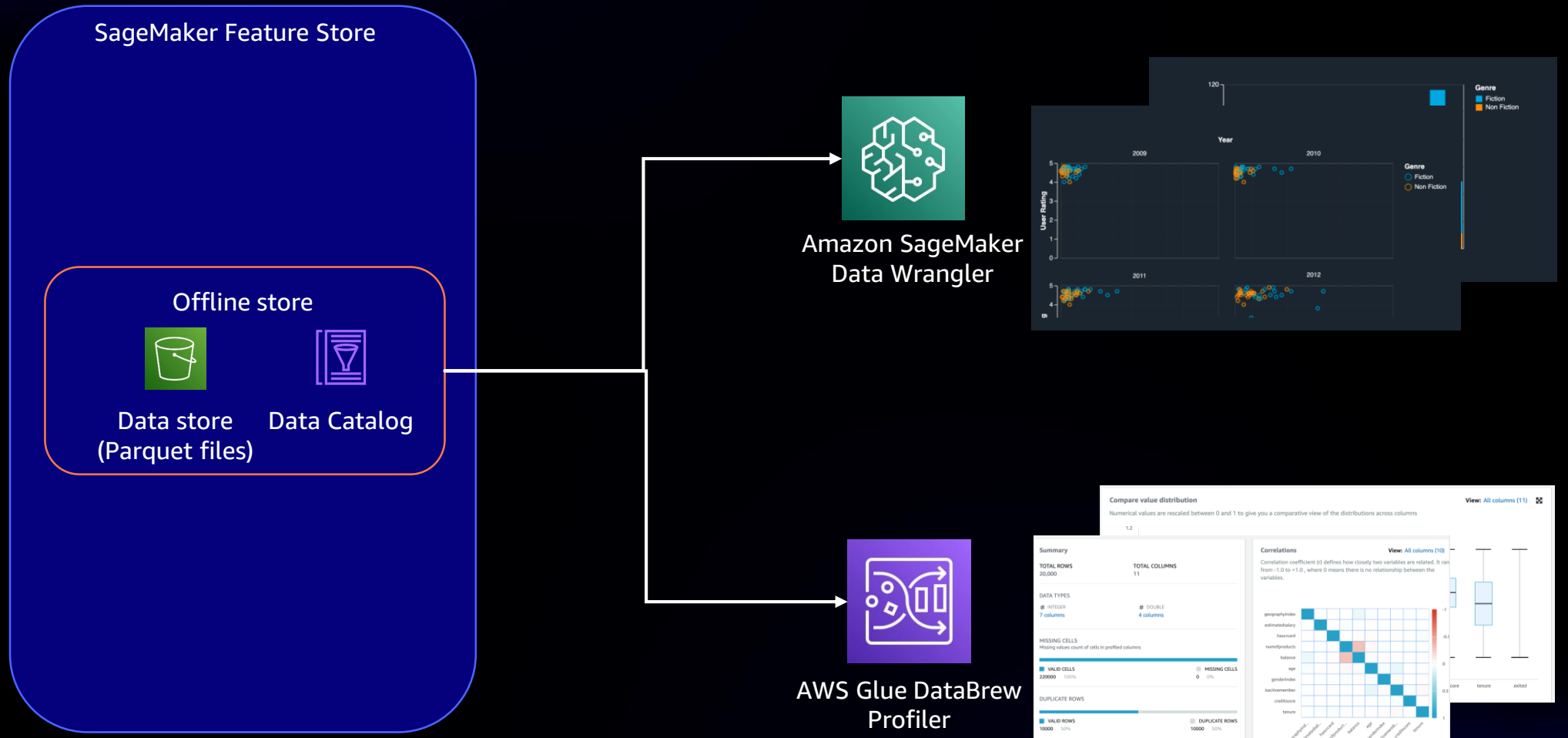
Amazon S3 location where the query results will be saved

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

Extending feature management functionality

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

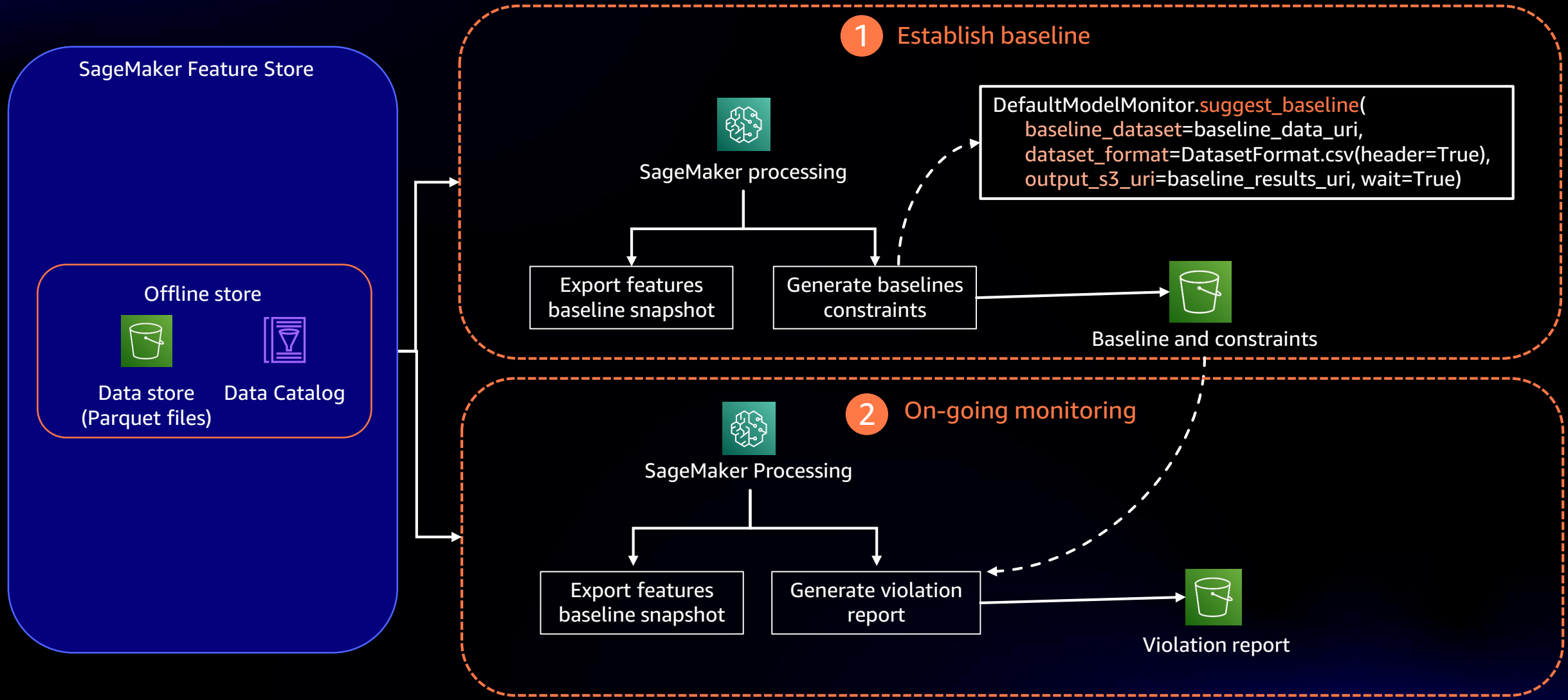
Feature statistic and visualization



- Statistics
- Data quality
- Correlation
- Outlier
- Skew
- ….

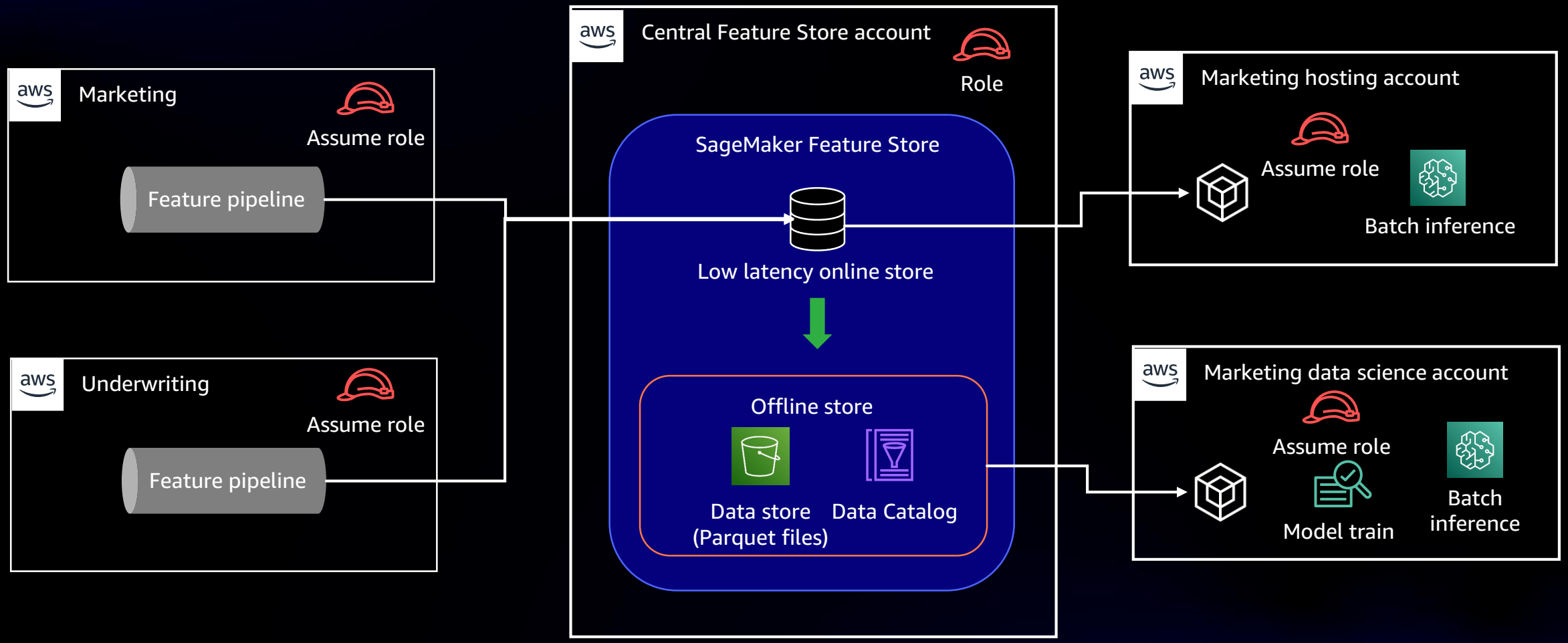
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Feature quality and monitoring



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

Cross-account access



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

SageMaker Feature Store best practices

- Design feature group schema based on entity types instead of specific model requirements
- Use SageMaker Data Wrangler, Pipeline, and Feature Store for quick feature engineering, pipeline development, and deployment
- Use parallel ingestion processes to maximize ingestion throughput
- Abstract feature lookup in a common feature lookup layer instead of the client application layer for inferences
- Enrich feature groups with metadata to facilitate feature discovery

aws

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

Thank you!

David Ping
dyping@amazon.com

Farooq Sabir
farsabir@amazon.com

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