#SFexp6

Real-time Machine Learning Analytics Using Structured Streaming and Kinesis Firehose

Caryl Yuhas (@ckred)
Myles Baker (@mydpy)
June 6th, 2017

databricks

About Databricks

TEAM

Started Spark project (now Apache Spark) at UC Berkeley in 2009

MISSION

Making Big Data Simple

PRODUCT

Unified Analytics Platform

Impact of Real-Time Analytics

- Capturing customer interactions, user behavior, and sensor readings is rapidly increasing
- Businesses need to respond immediately to new information as it arrives
- Real-time analytics is at the core of next-generation IT systems

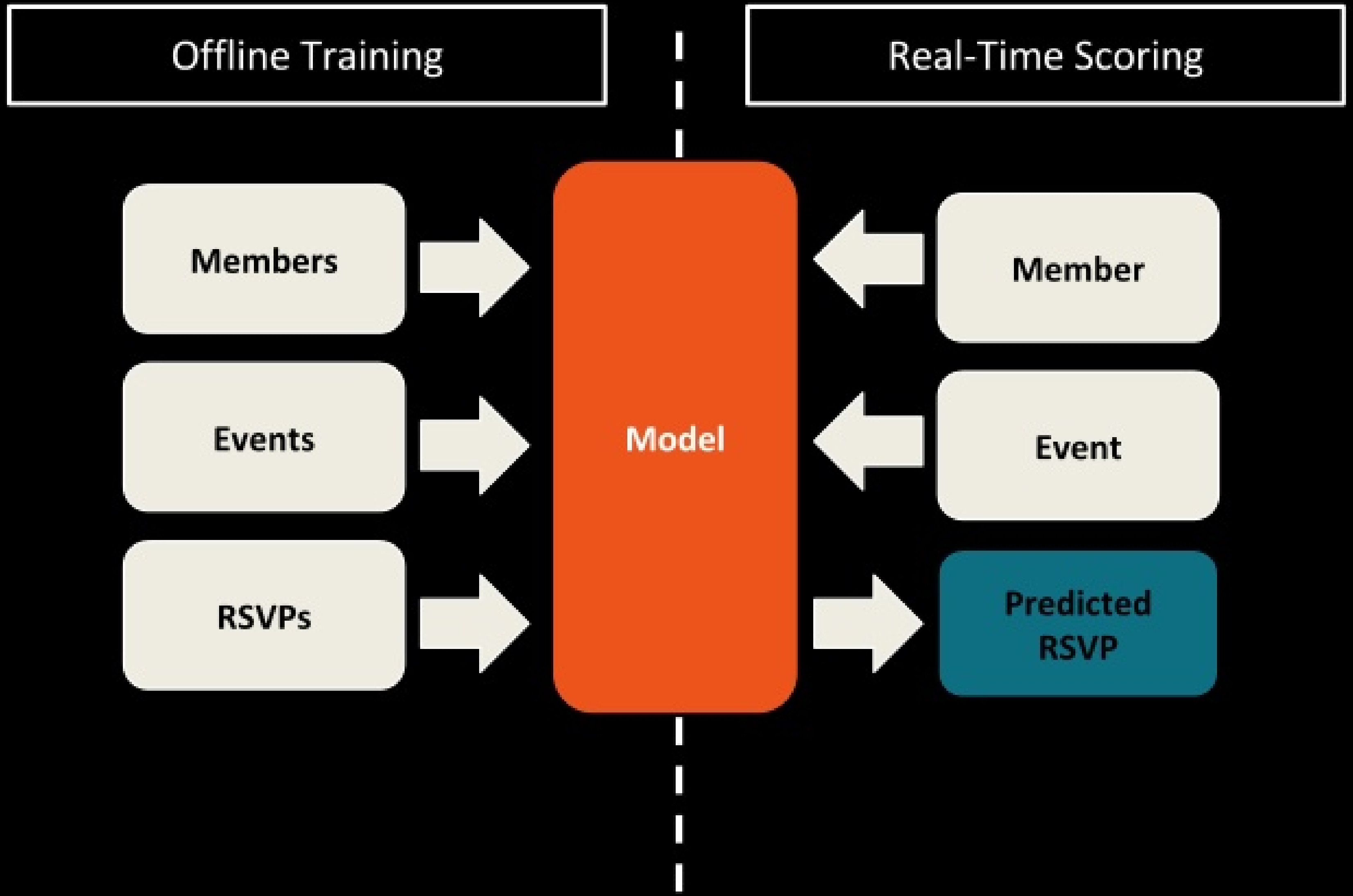
Challenges Building a Solution

- Performant, scalable real-time analytics requires connecting multiple tools
- Streaming data comes with all of the problems of static data with added complexity
- Machine learning models need to be trained on historical data and scored with real-time data

The Meetup Streaming API

- Can we explore Meetup data in real-time?
- Can we predict RSVPs for new Meetups using streaming data from the Meetup API?
  - Members
  - Events
  - RSVPs

A Data Model for Training and Scoring



Offline Training

Real-Time Scoring

Members

Events

RSVPs

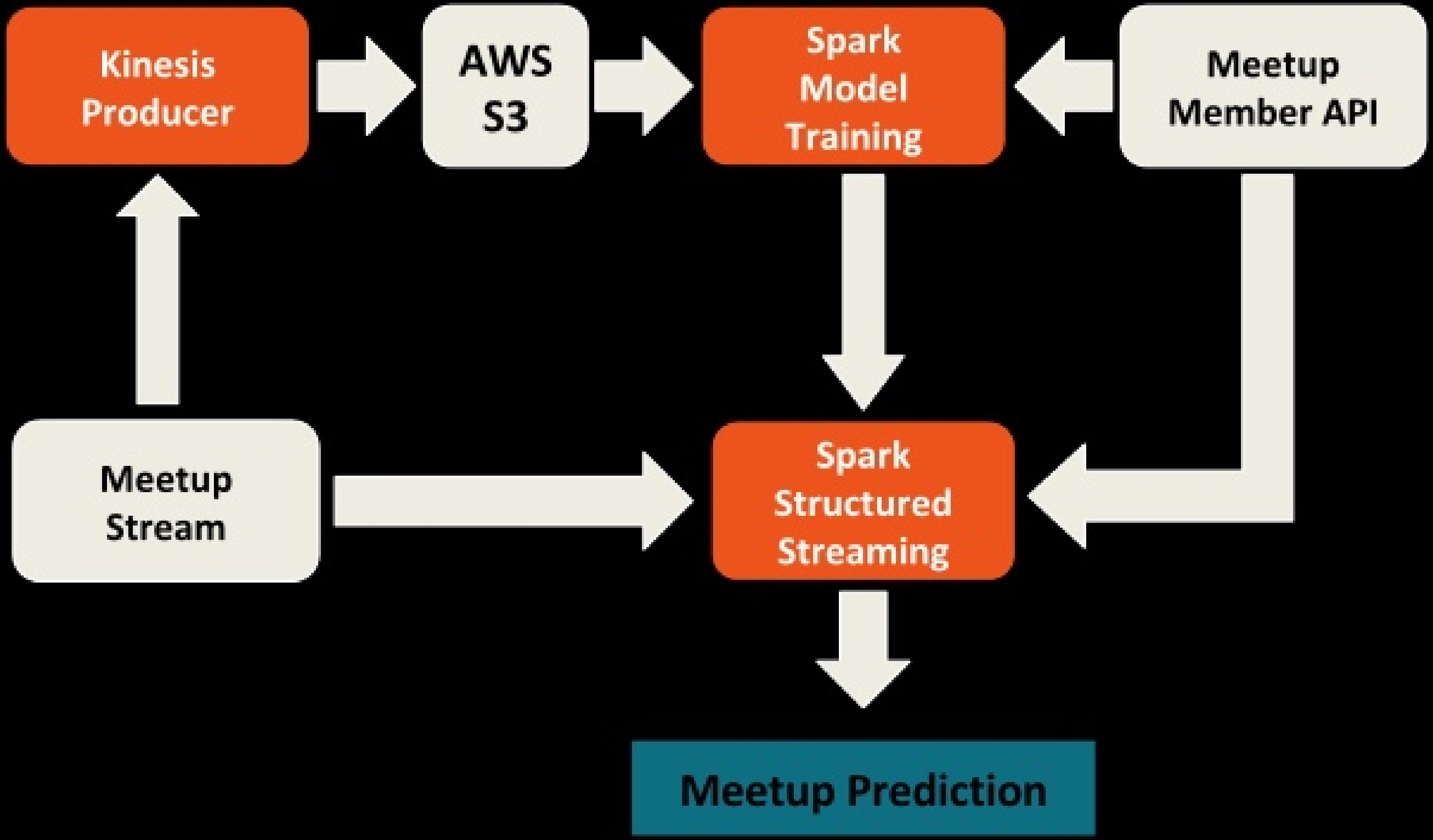
Model

Member

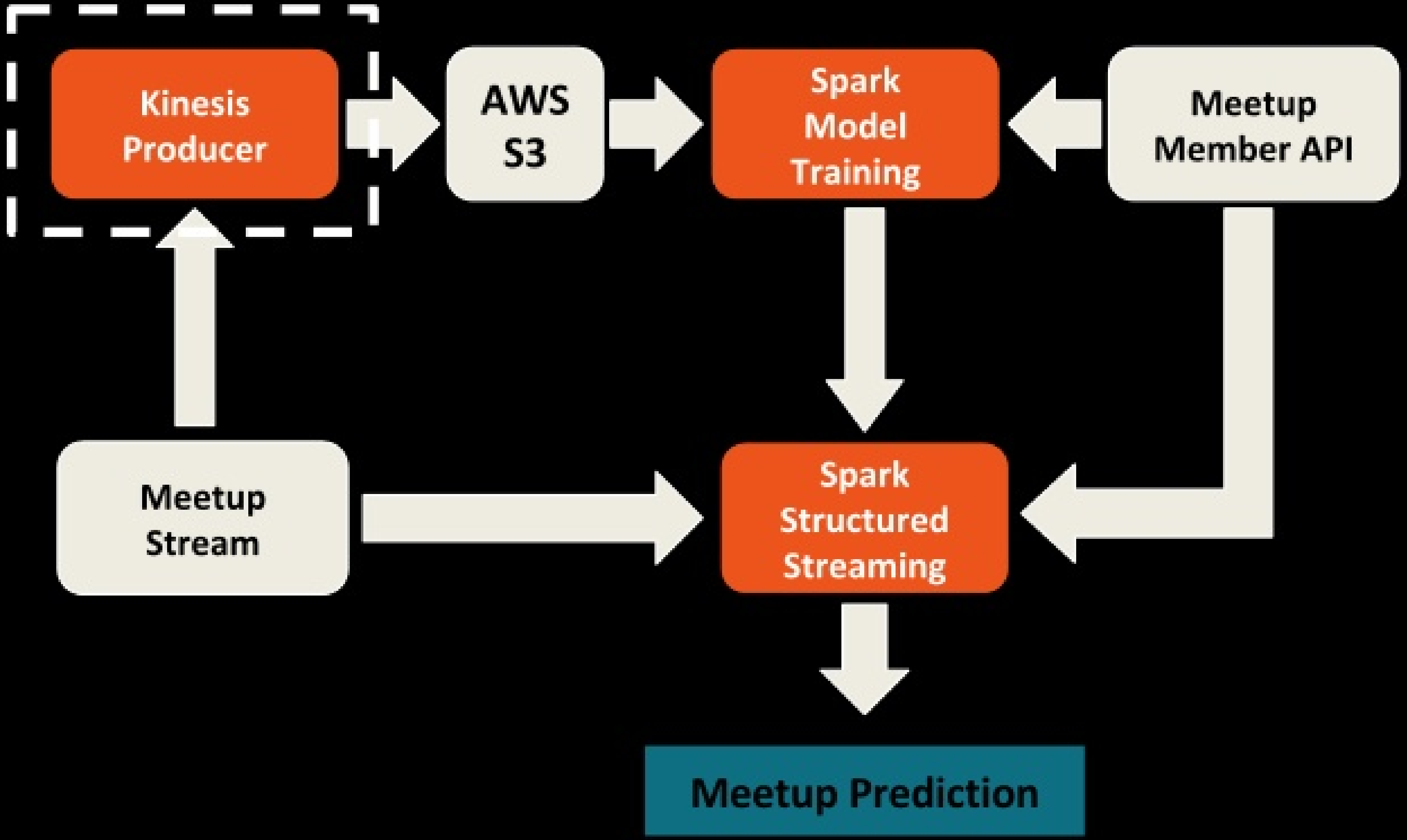
Event

Predicted RSVP

Component Integration and Serving



Component Integration and Serving



Producing the Kinesis Firehose Stream

```
requests.get(apiURL, stream = True)
kinesis = boto3.client('firehose')
```

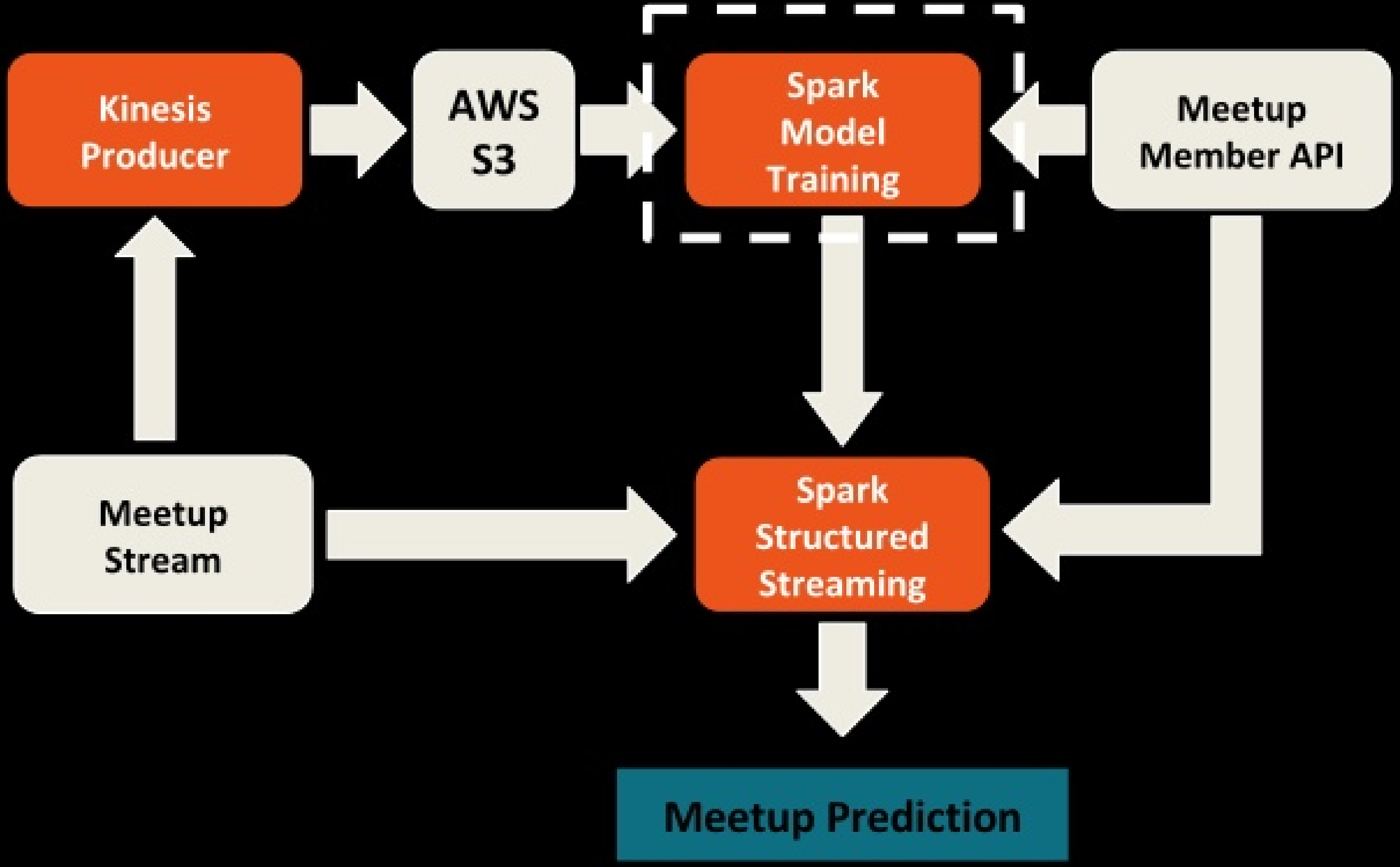
requests.get() makes a request to the Meetup API, keeping the stream open

boto3.client creates a firehose kinesis client

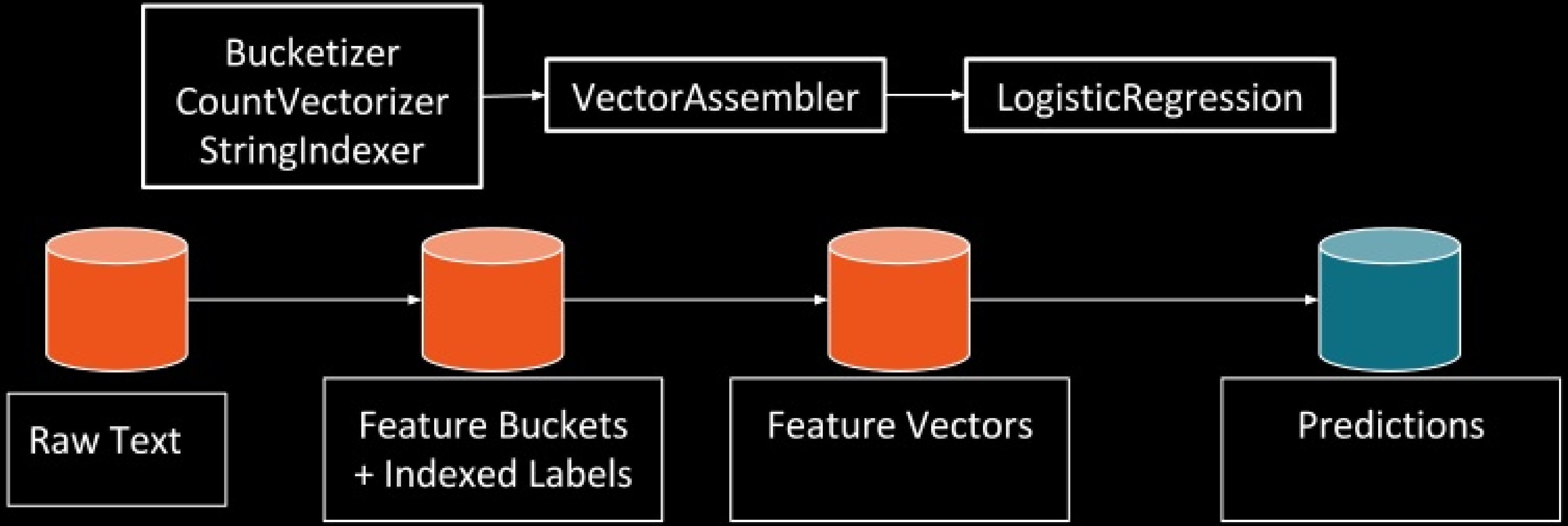
```
kinesis.put_record_batch(
     DeliveryStreamName='meetup',
     Records=rsvps)
```

kinesis.put_record_batch() writes the records streamed to S3 using the Kinesis Firehose delivery stream ‘meetup’

Component Integration and Serving



Our Meetup ML Pipeline



Bucketizer
CountVectorizer
StringIndexer

VectorAssembler

LogisticRegression

Raw Text

Feature Buckets
+ Indexed Labels

Feature Vectors

Predictions

Create an ML Pipeline

```
val pipeline = new Pipeline()
  .setStages(Array(
    transformers,
    estimators,
    models))
```

A **Pipeline** allows us to simply chain a series of transformations and estimators

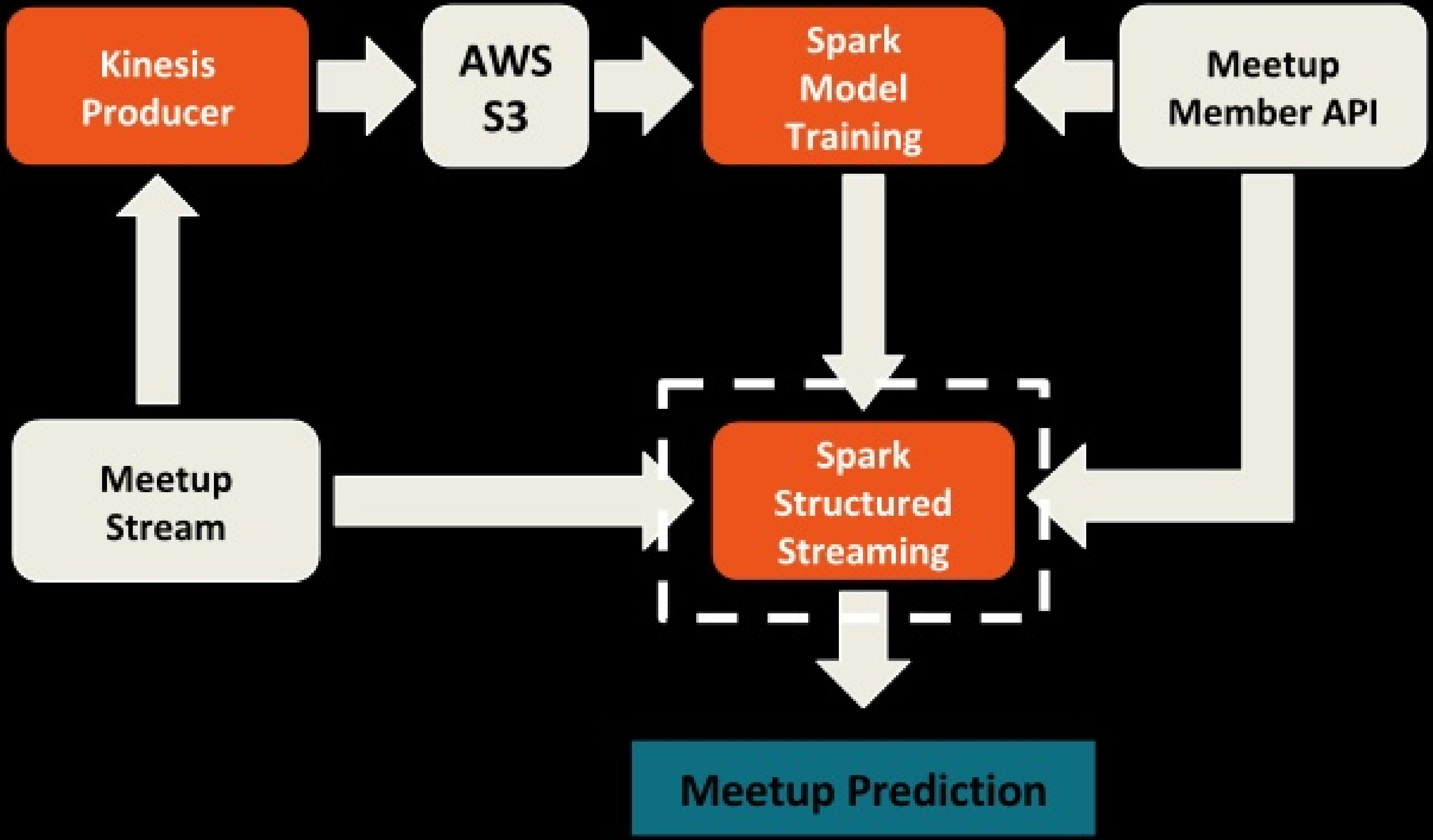
```
val model = pipeline.fit(meetup)
```

Fit a **model** based on the pipeline

```
model.write.overwrite().save(...)
```

Save the model to disk for scoring

Component Integration and Serving



Scoring the Model in Real-time

```
val model = PipelineModel.load(...)
```

Load the **trained** model

```
val events = spark.readStream
                  .parquet(...)
```

Stream meetup event data

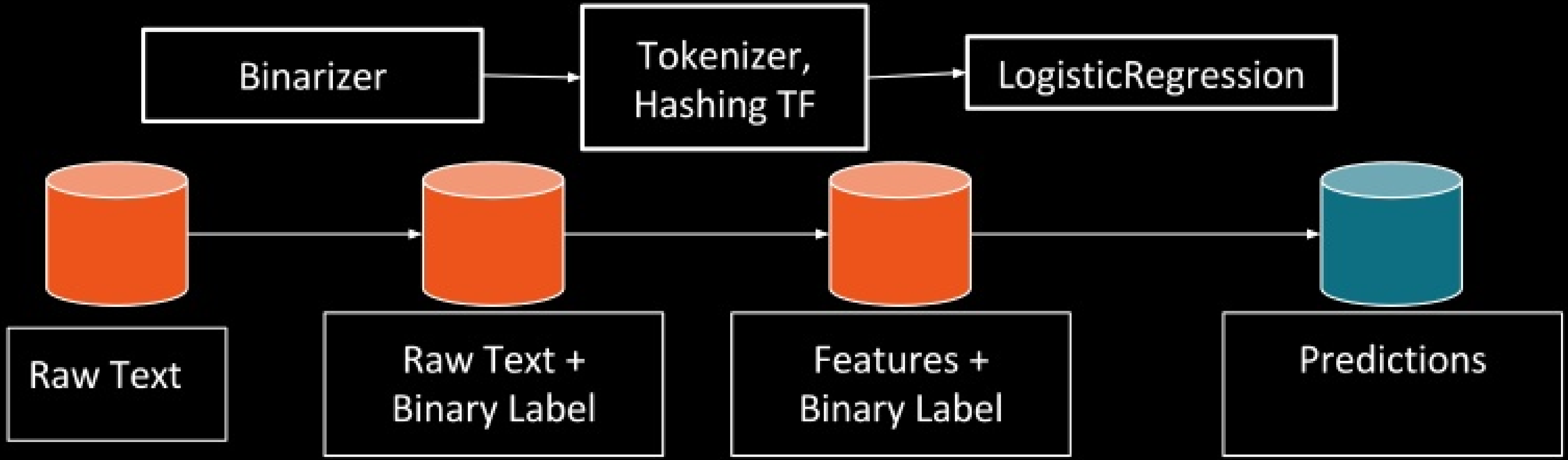
```
model.transform(members)
```

Score the model

ML Limitations in Structured Streaming

- Structured streaming does not support operations needed by ML methods
  - count, collect, round, aggregate*, etc.
- Many models, transformers, and estimators are not supported
  - K-Means, SVM, CountVectorizer, VectorAssembler, StringIndexer, etc.

Our Streaming Meetup ML Pipeline



Binarizer

Tokenizer, Hashing TF

LogisticRegression

Raw Text

Raw Text + Binary Label

Features + Binary Label

Predictions

Alternative Scoring: Model Export

1) Fit ML model in Databricks using Spark MLlib.

2) Export model (as JSON files) in Databricks

```
val lrModel = new LogisticRegression().fit(myData)
ModelExporter.export(lrModel, "s3a:/...")
```

3) Deploy model in external system

```
import com.databricks.ml.local.ModelImport
val lrModel = ModelImport.import("s3a:/...")
val jsonInput = json(...)
val jsonOutput = lrModel.transform(jsonInput)
```

Try Apache Spark in Databricks!

UNIFIED ANALYTICS PLATFORM

- Collaborative cloud environment
- Free version (community edition)

DATABRICKS RUNTIME 3.0

- Apache Spark - optimized for the cloud
- Caching and optimization layer - DBIO
- Enterprise security - DBES

Try for free today.

databricks.com

#SFexp6

mydpy/ss-2017-structured-streaming

Thank you

caryl@databricks.com
mbaker@databricks.com

databricks

Common Questions

- Can I consume data from Kinesis directly with Structured Streaming?
- Does MLlib support streaming data frames?
- Why did you use Boto3 to produce the Kinesis stream?
- How well does this scale? Can we test volume?