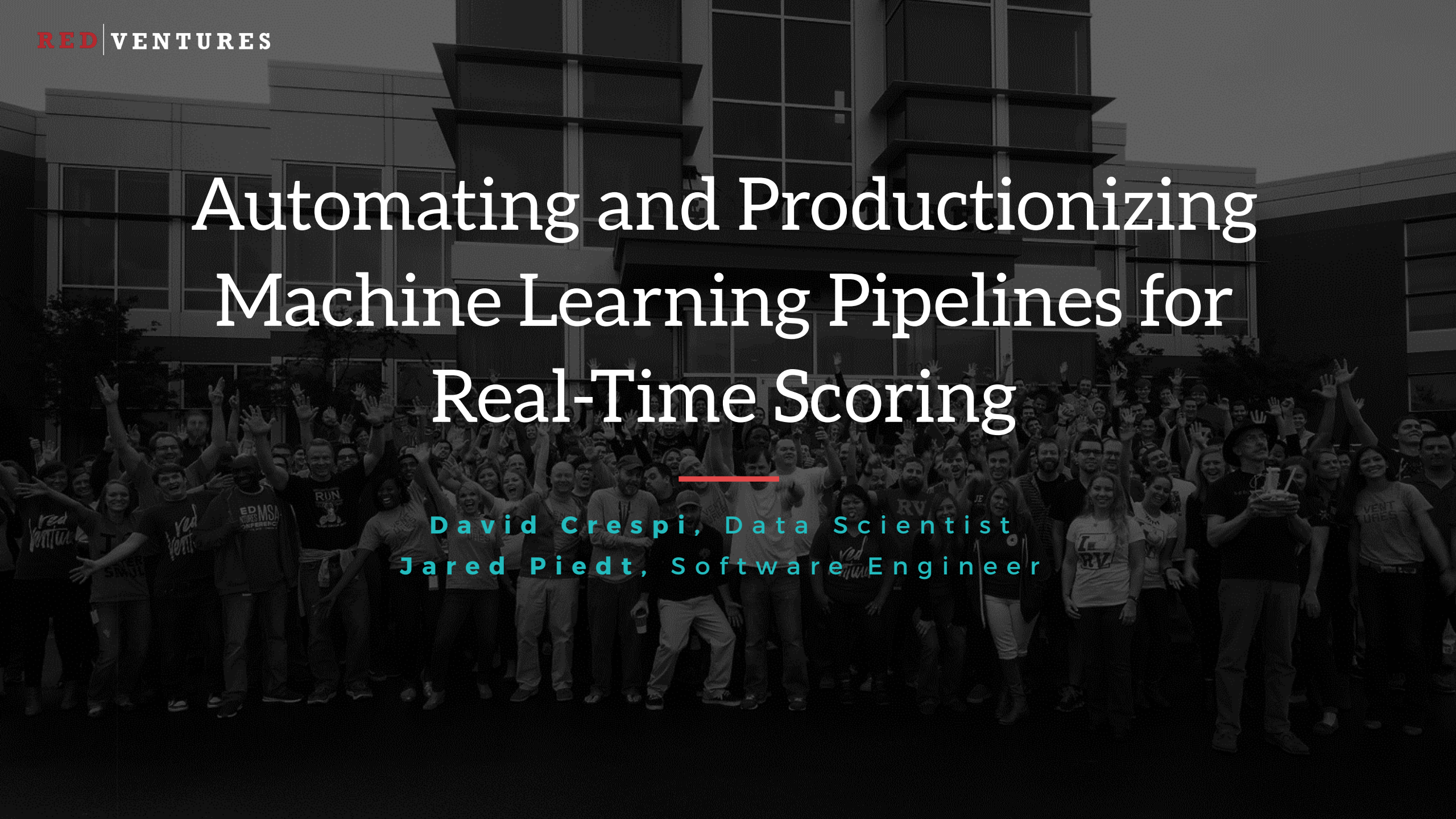


RED VENTURES

Automating and Productionizing Machine Learning Pipelines for Real-Time Scoring

David Crespi, Data Scientist
Jared Piedt, Software Engineer

What we do

- Digital consumer choice platform
- Connect online customers with products and services across high-growth industries:
  - Home services
  - Financial services
  - Healthcare

Bankrate®

the POINTS GUY

MYMOVE™

creditcards.com

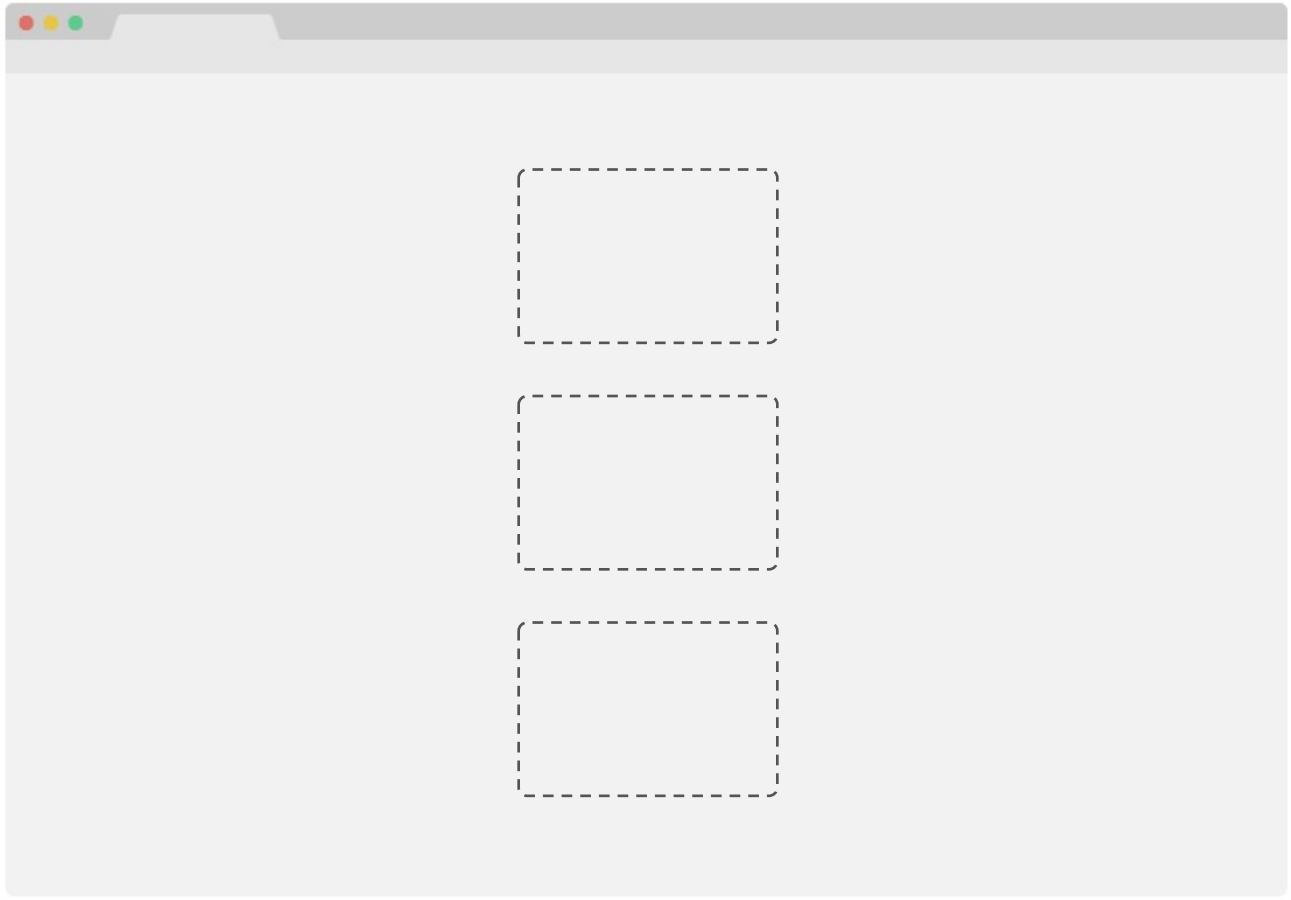
The Simple Dollar

Reviews.com

allconnect®
A really smart move.

How does data science add value?

Product recommendation



Real-Time Predictions

Requirements
1 Speed
2 Consistency

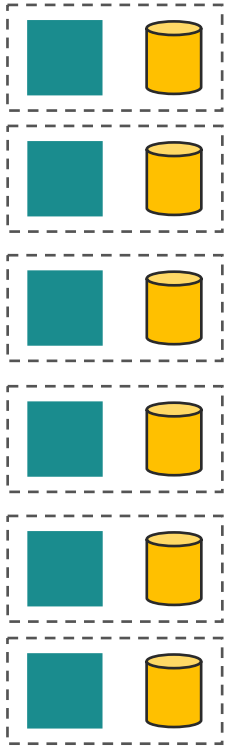
Data Science Process

1. Data Collection
2. Machine Learning Pipelines
3. Model Deployment

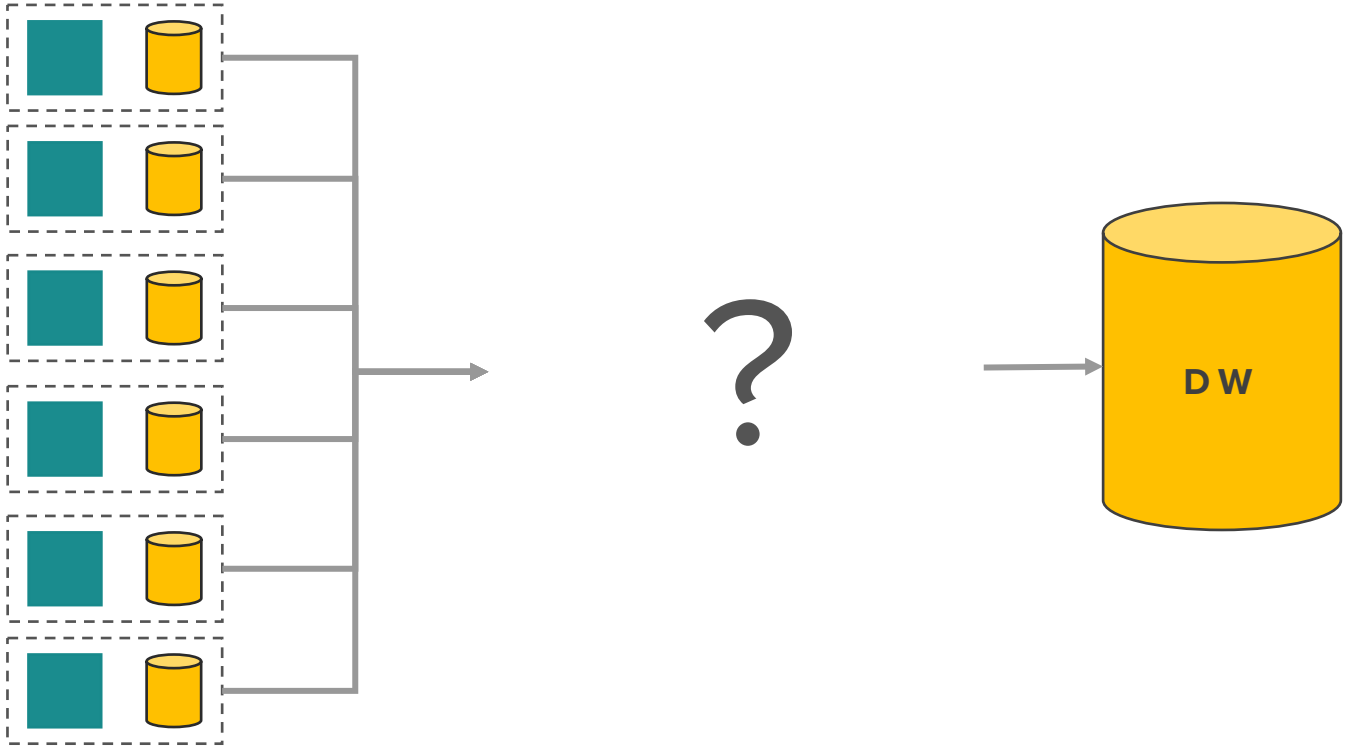
Data Science Process

1. Data Collection
2. Machine Learning Pipelines
3. Model Deployment

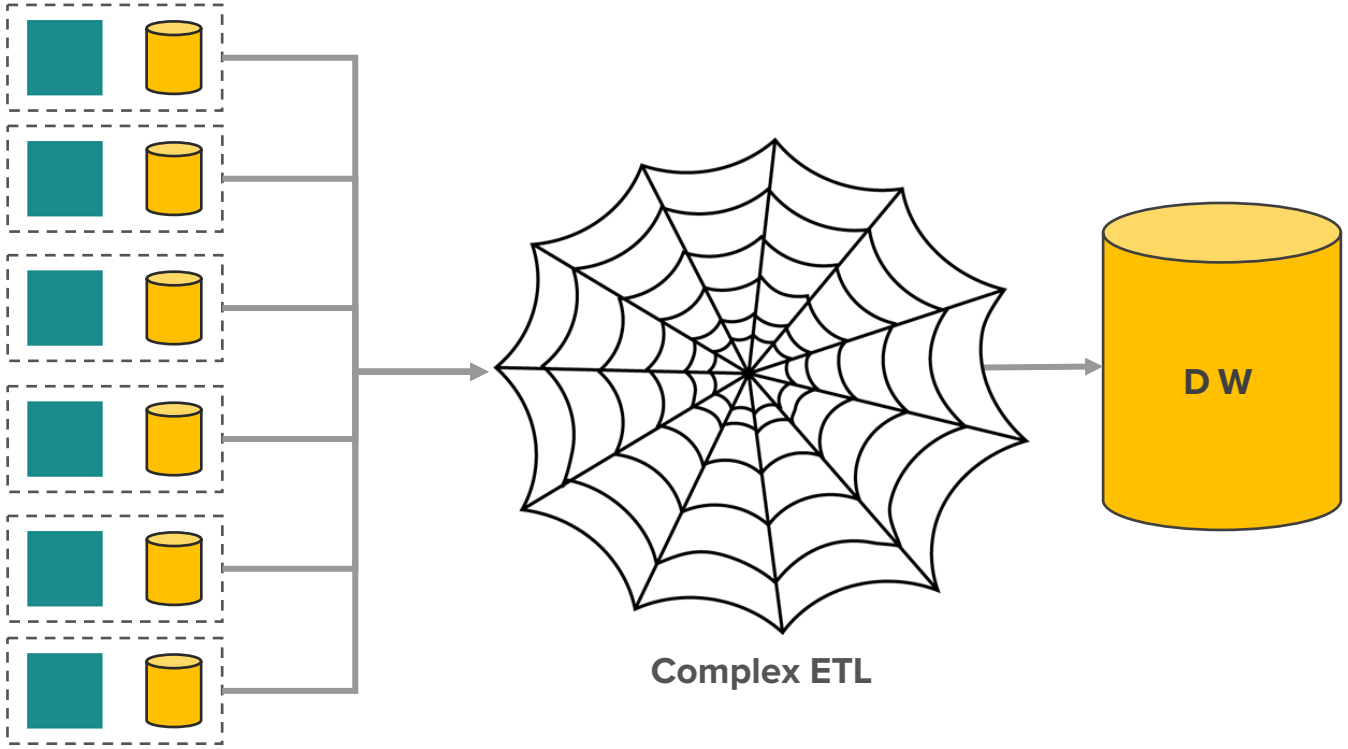
Old Data Architecture



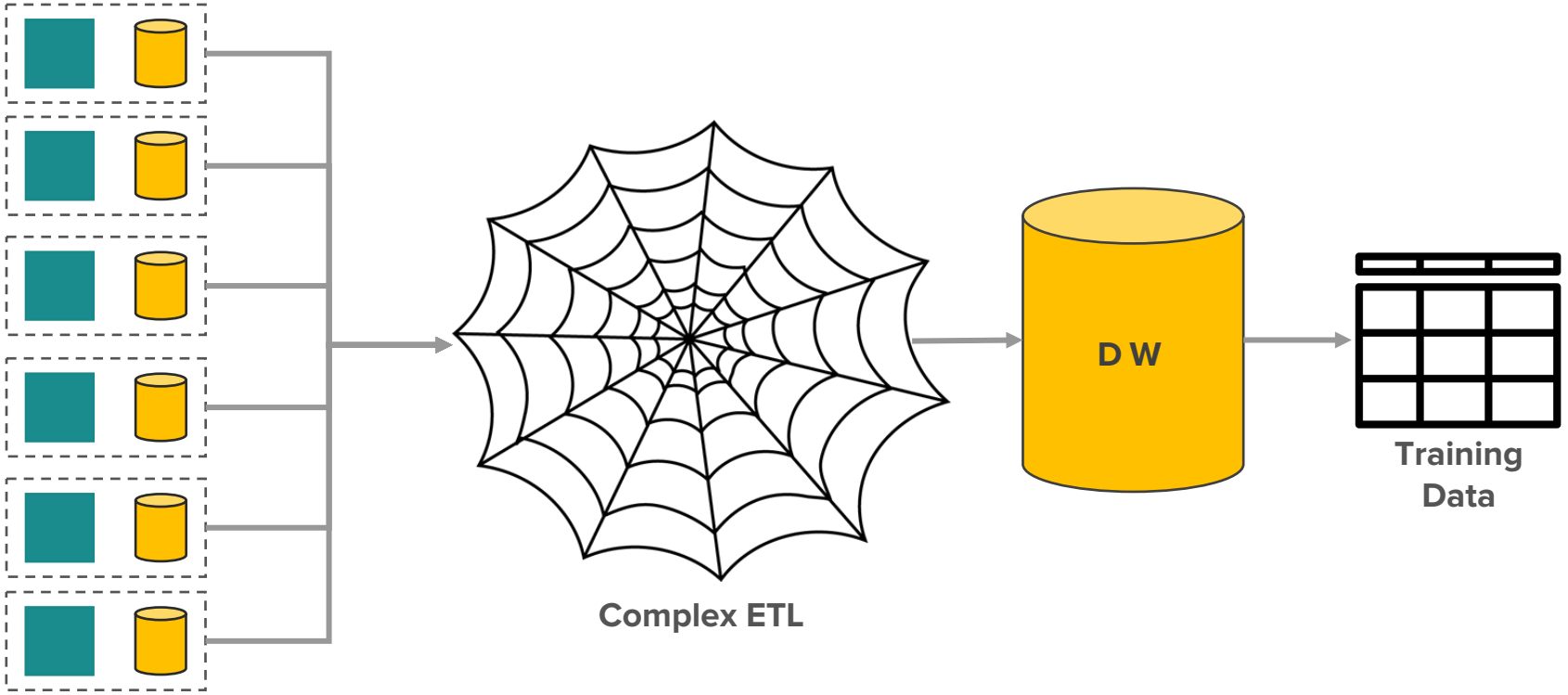
Old Data Architecture



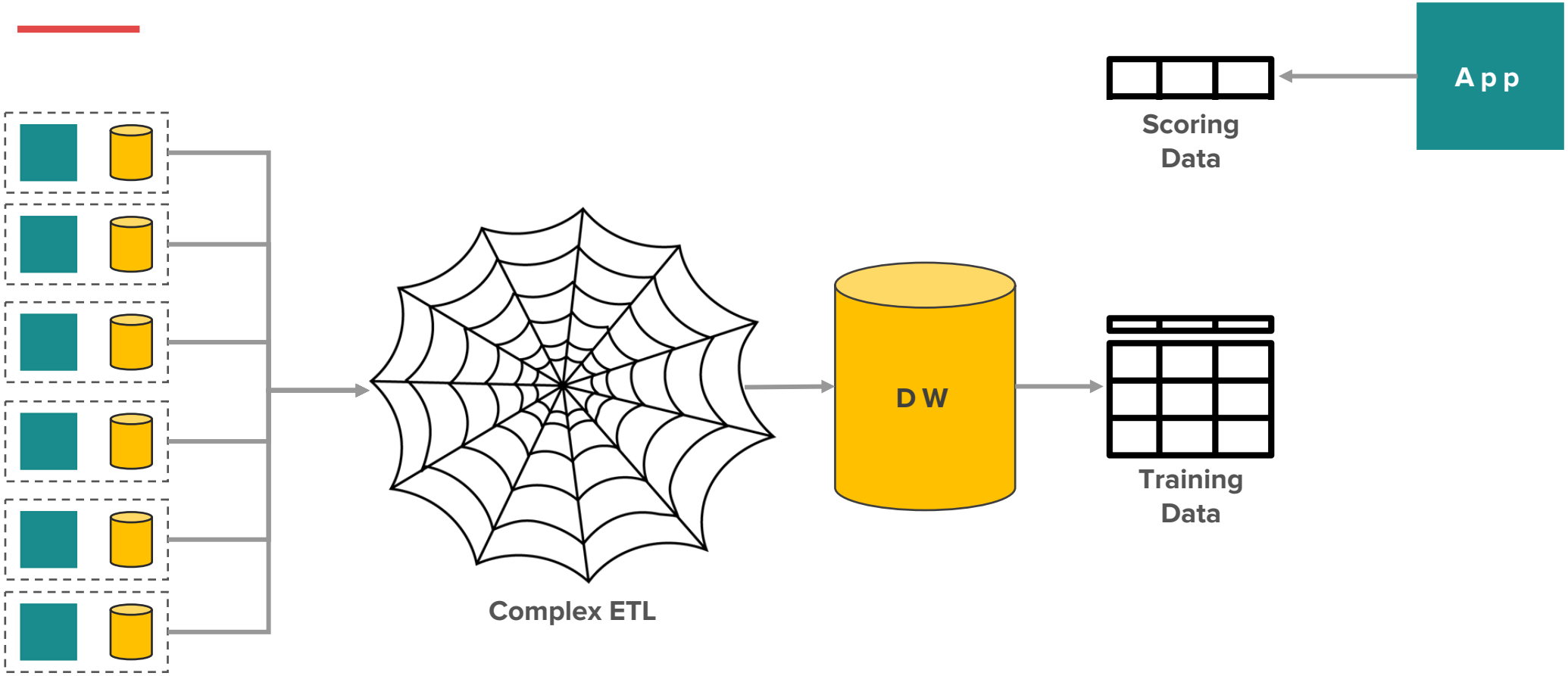
Old Data Architecture



Old Data Architecture



Old Data Architecture



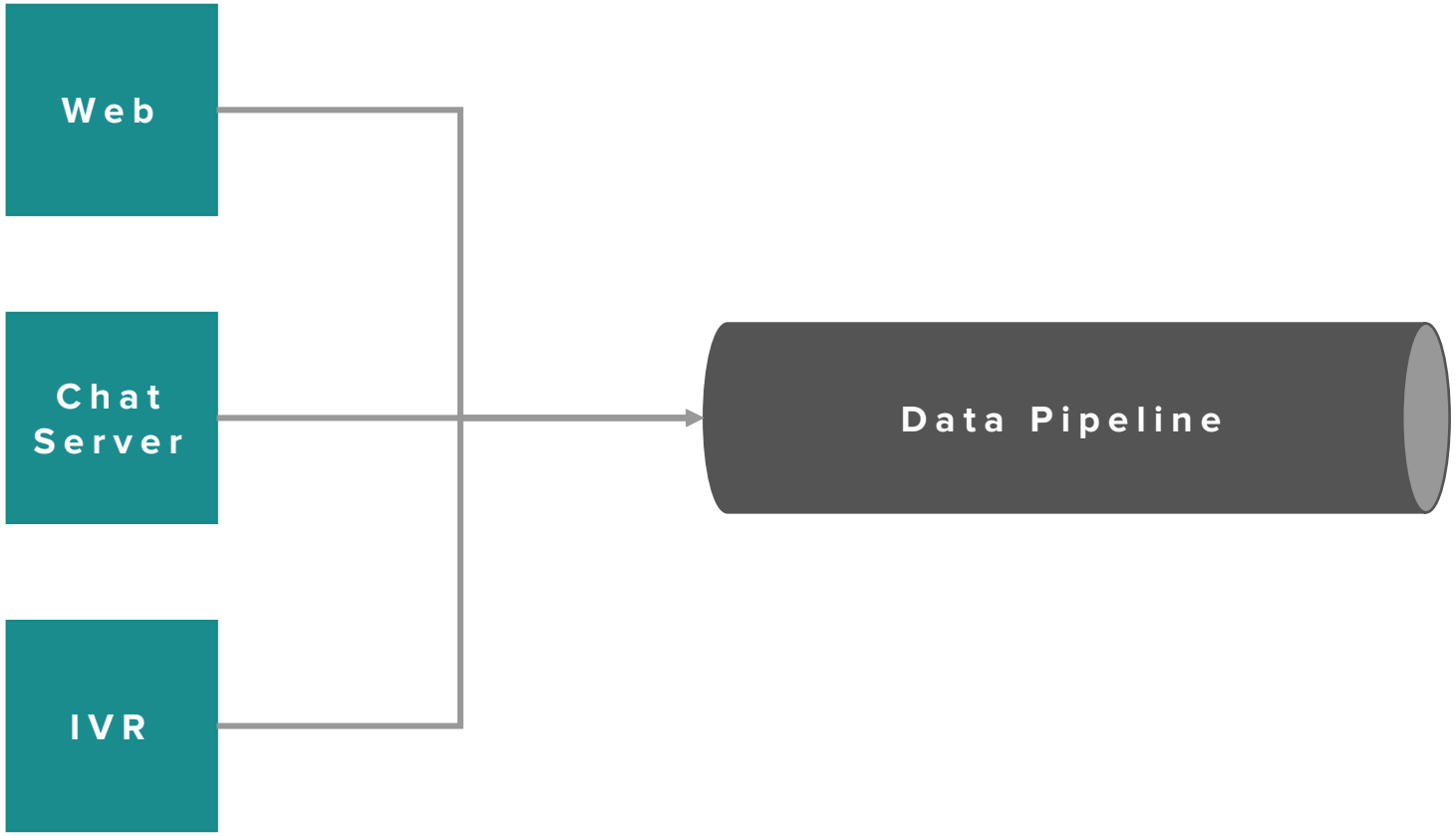
Pain Points

- Duplication of business logic
- Data drift

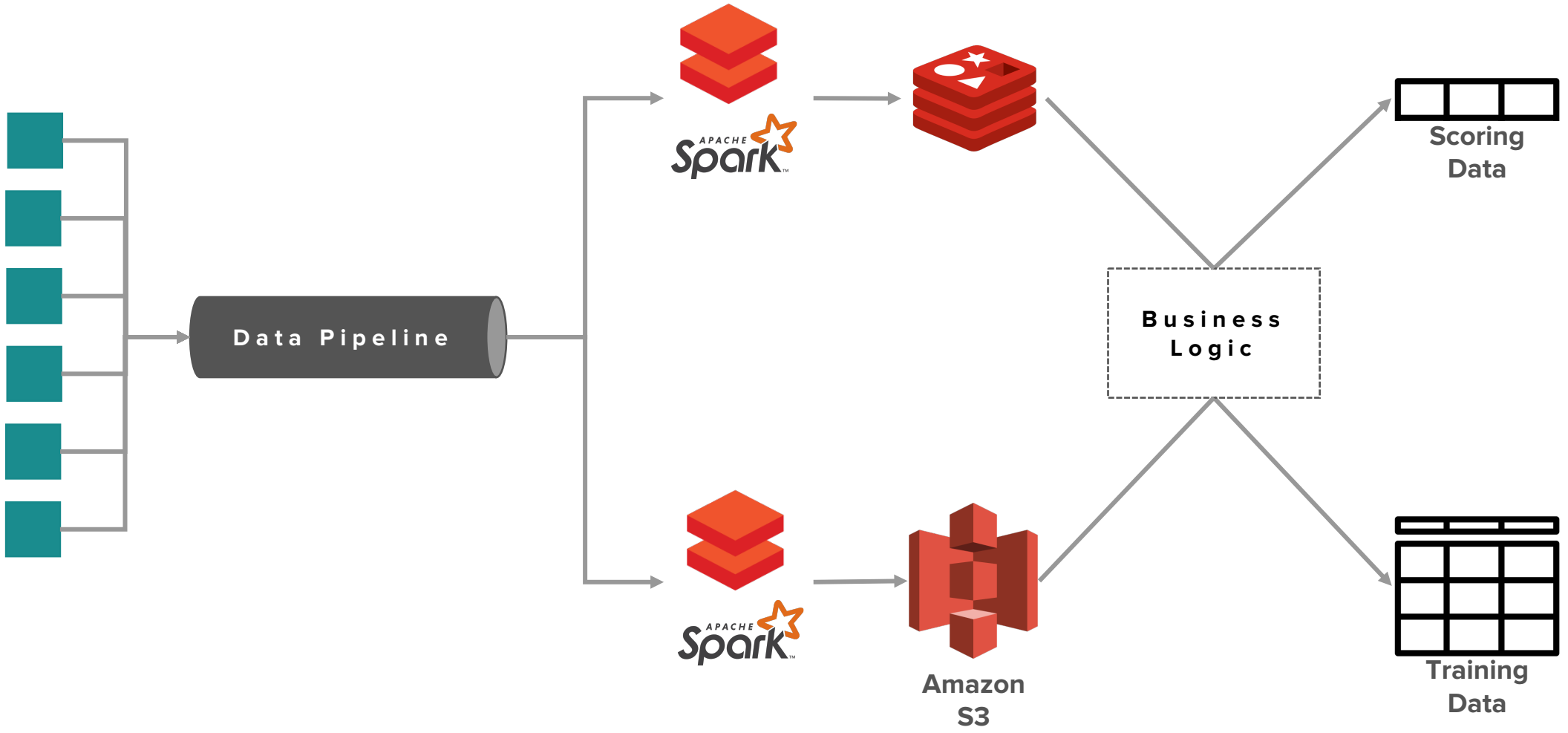
Goals

1. Immutable data
2. Write business logic once
3. Make data available in real-time

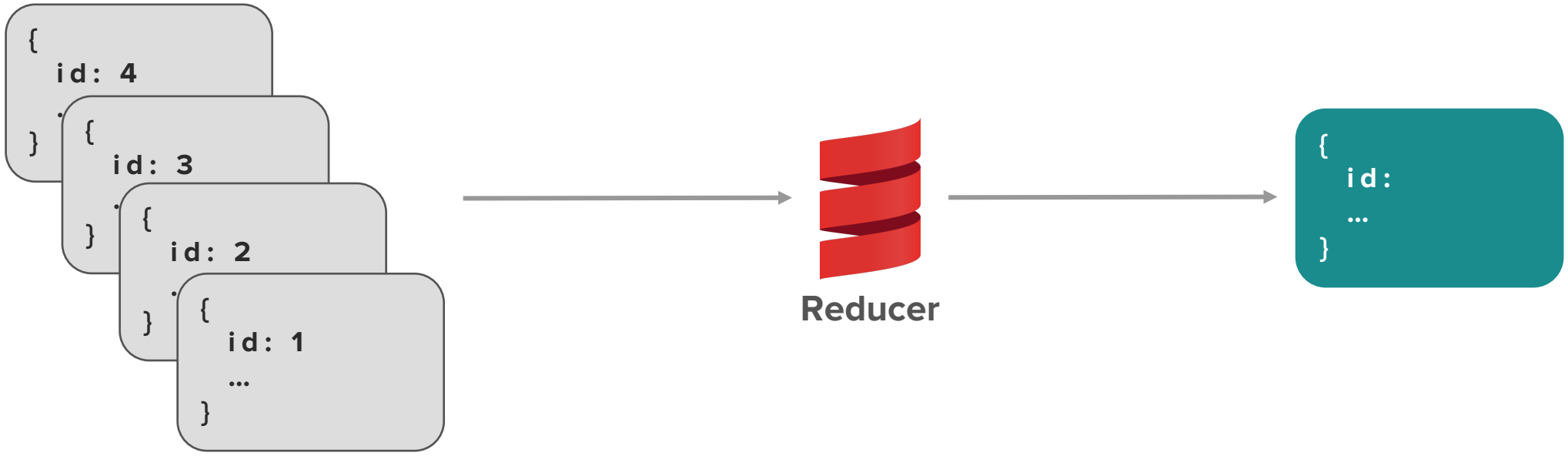
Event-Driven Architecture



New Data Architecture



Projections



Credit Card Recommendation

User Id	Keyword	Page View Count	Card Shown	Clicked
a	best travel cards	2	Travel	1
b	credit cards	3	Cash Back	0
c	top credit cards	1	Cash Back	1
d	credit cards	1	Travel	0

reducer

```
{
    "userId": "",
    "latestSearchKeyword": "",
    "pageViewCount": 0,
    "cardShown": "",
    "clicked": 0
}
```

```
{
    "id": 1,
    "userId": "z",
    "eventType": "Page Viewed",
    "timestamp": "2018-03-01T10:00:00.000Z",
    "data": {
        "url": "www.url.com/?keyword=airline...",
        "path": "/",
        "query": "keyword=airline%20miles%20card",
        "referrer": "www.google.com"
    }
}
```


reducer


```
{
    "userId": "z",
    "latestSearchKeyword": "airline miles card",
    "pageViewCount": 1,
    "cardShown": "",
    "clicked": 0
}
```

```
{
    "id": 2,
    "userId": "z",
    "eventType": "Product Viewed",
    "timestamp": "2018-03-01T10:00:00.300Z",
    "data": {
        "type": "Credit Card",
        "name": "Travel"
    }
}
```

reducer

```
{
    "userId": "z",
    "latestSearchKeyword": "airline miles card",
    "pageViewCount": 1,
    "cardShown": "Travel",
    "clicked": 0
}
```

```
{
    "id": 3,
    "userId": "z",
    "eventType": "Product Clicked",
    "timestamp": "2018-03-01T10:02:00.000Z",
    "data": {
        "type": "Credit Card",
        "name": "Travel"
    }
}
```

reducer

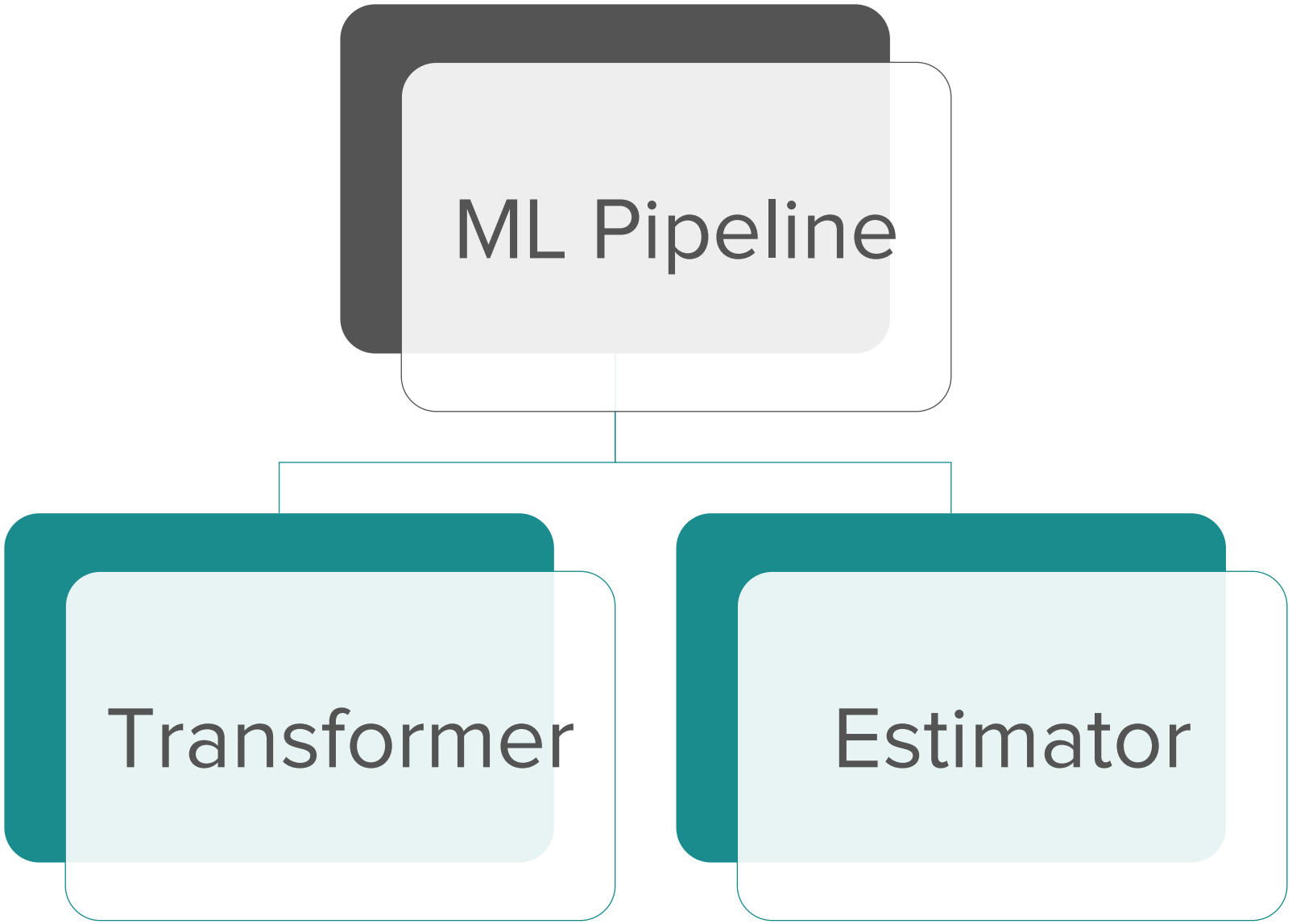
```
{
    "userId": "z",
    "latestSearchKeyword": "airline miles card",
    "pageViewCount": 1,
    "cardShown": "Travel",
    "clicked": 1
}
```

Credit Card Recommendation

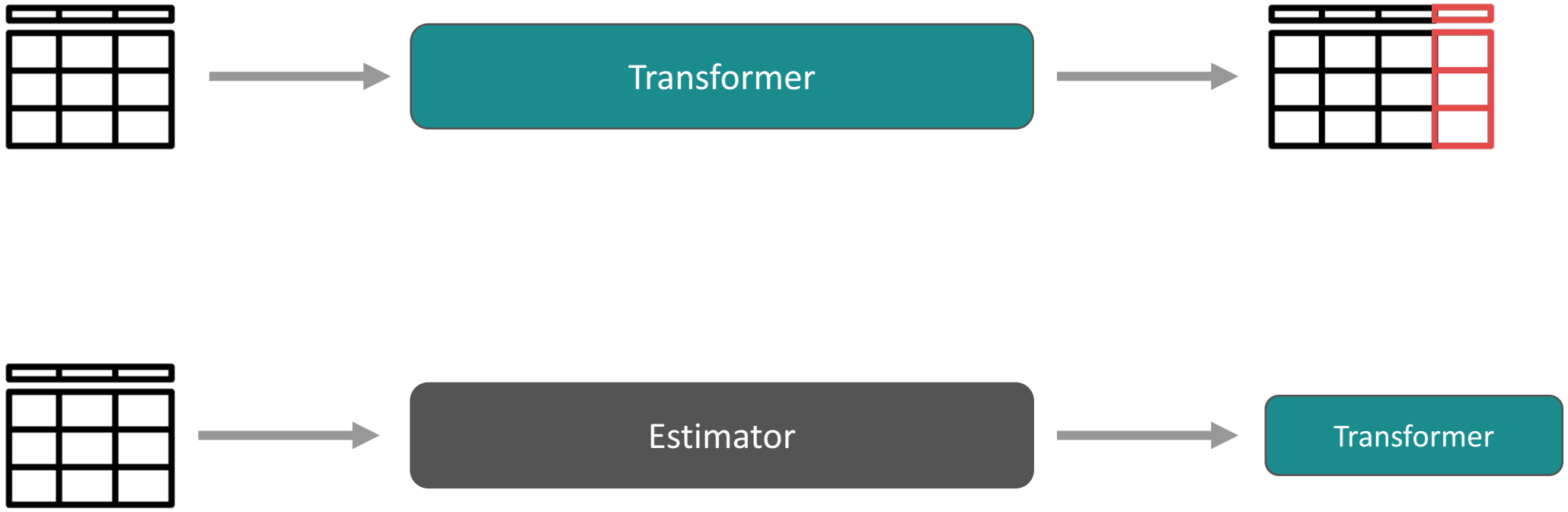
User Id	Keyword	Page View Count	Card Shown	Clicked
a	best travel cards	2	Travel	1
b	credit cards	3	Cash Back	0
c	top credit cards	1	Cash Back	1
d	credit cards	1	Travel	0
z	**airline miles card**	**1**	**Travel**	**1**

Data Science Process

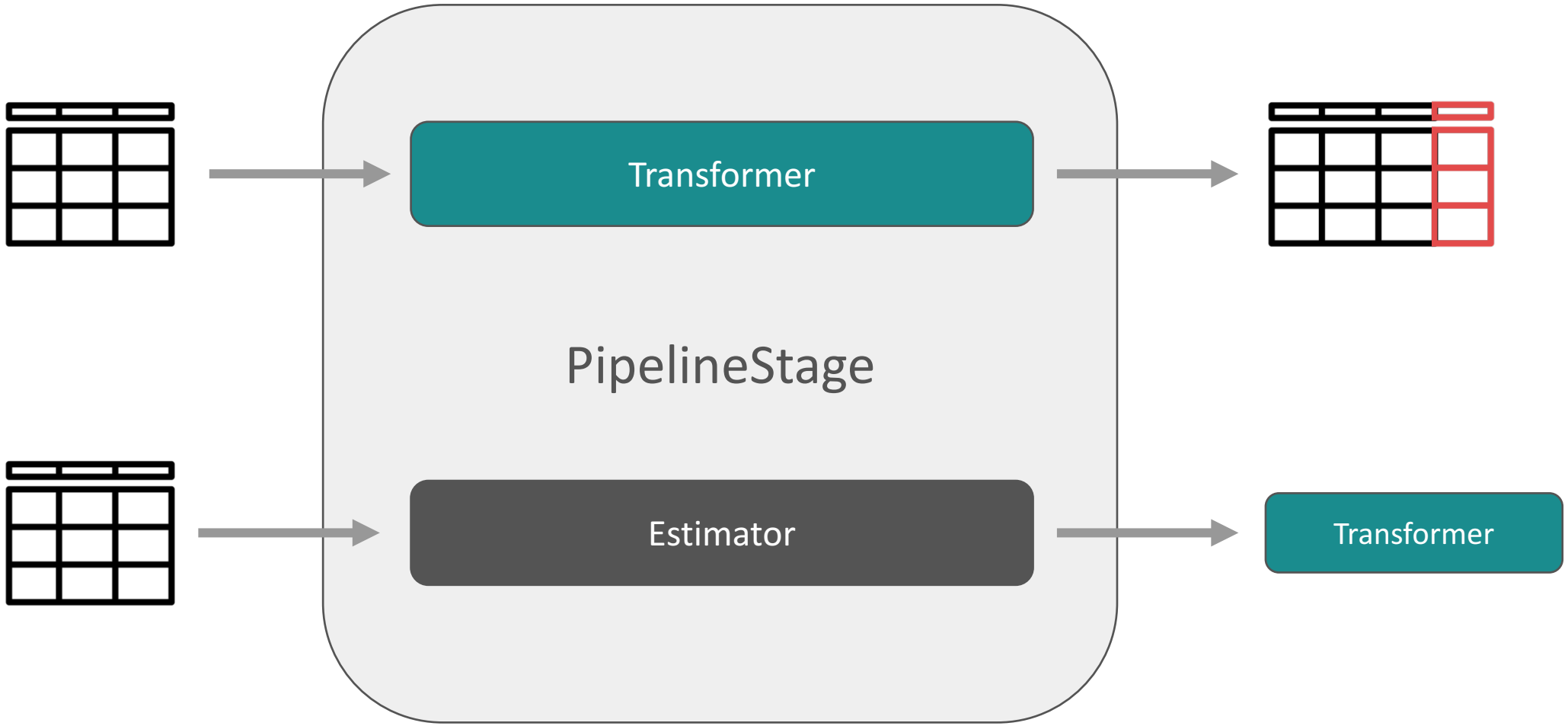
1. Data Collection
2. Machine Learning Pipelines
3. Model Deployment


ML Pipeline
Transformer
Estimator

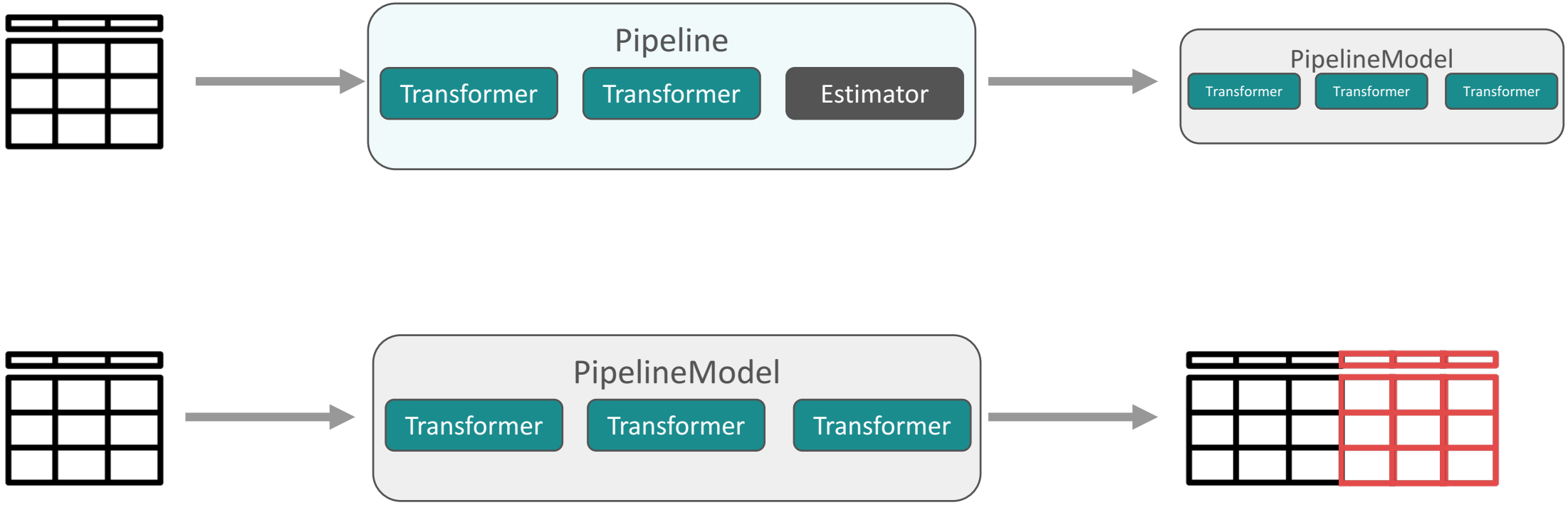
Spark: Estimators and Transformers



Spark: Estimators and Transformers

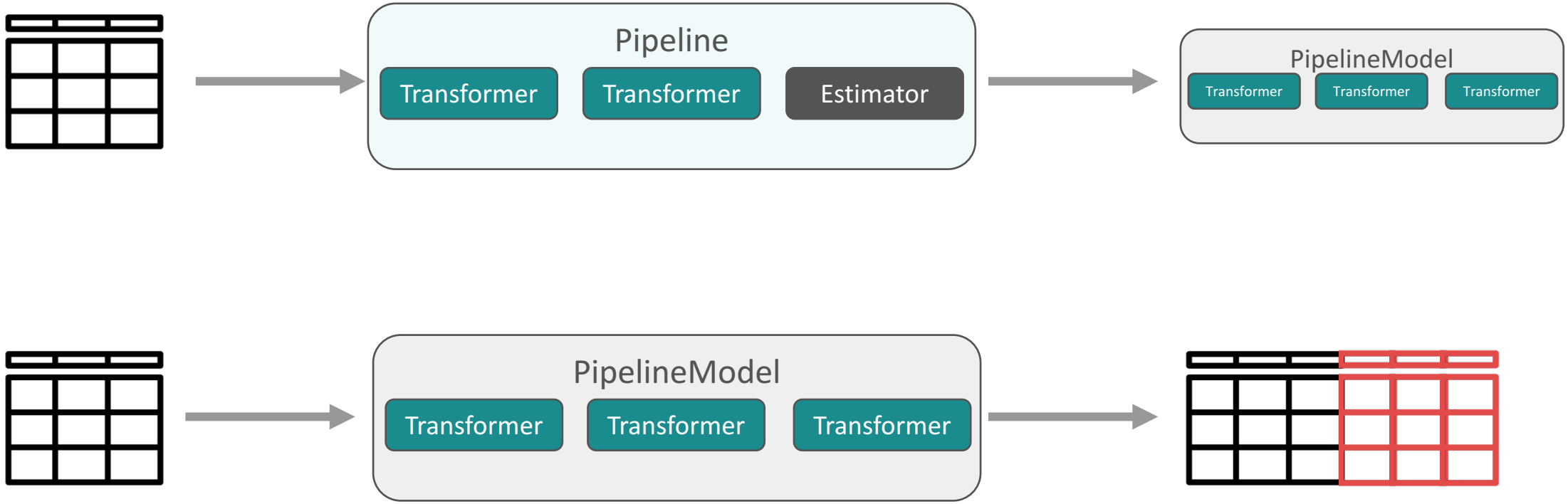


Spark: Estimators and Transformers

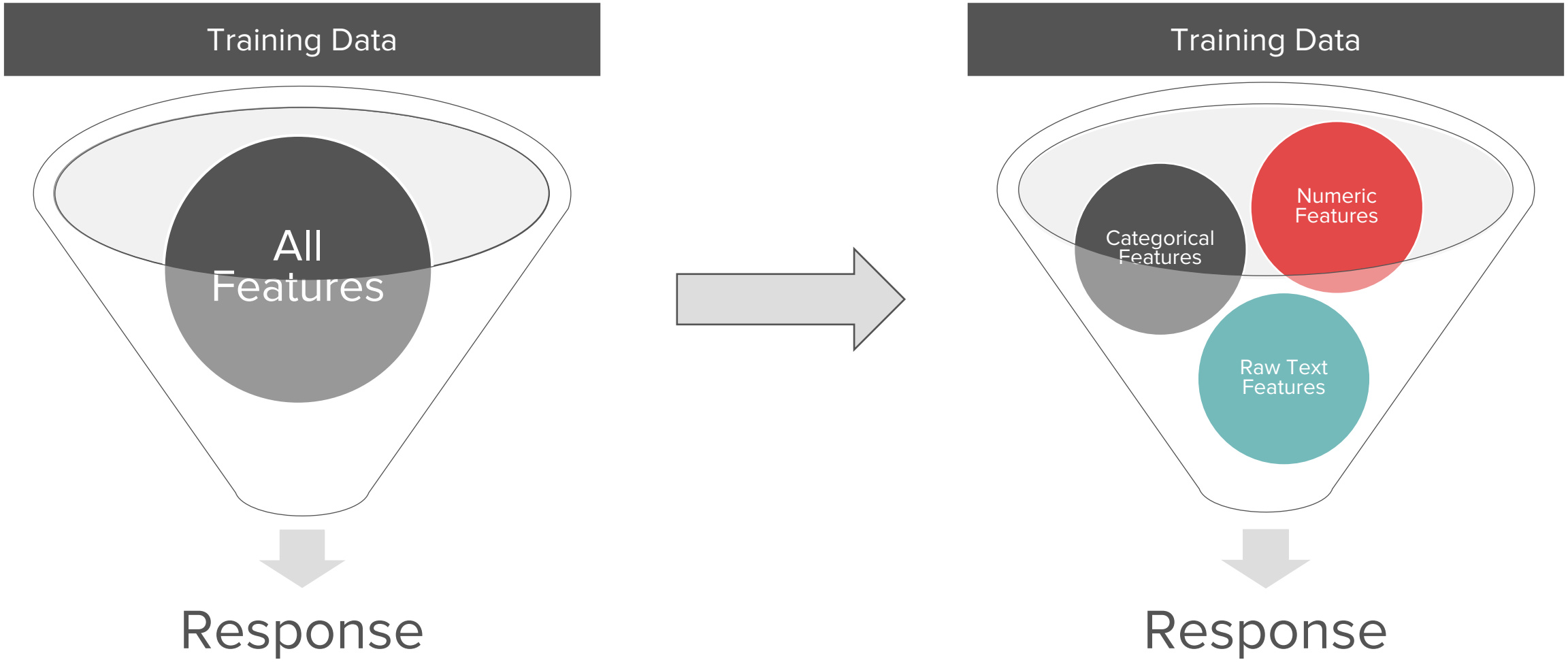


How do ML algorithms fit in?

Spark: Estimators and Transformers



Generalizing Data Science

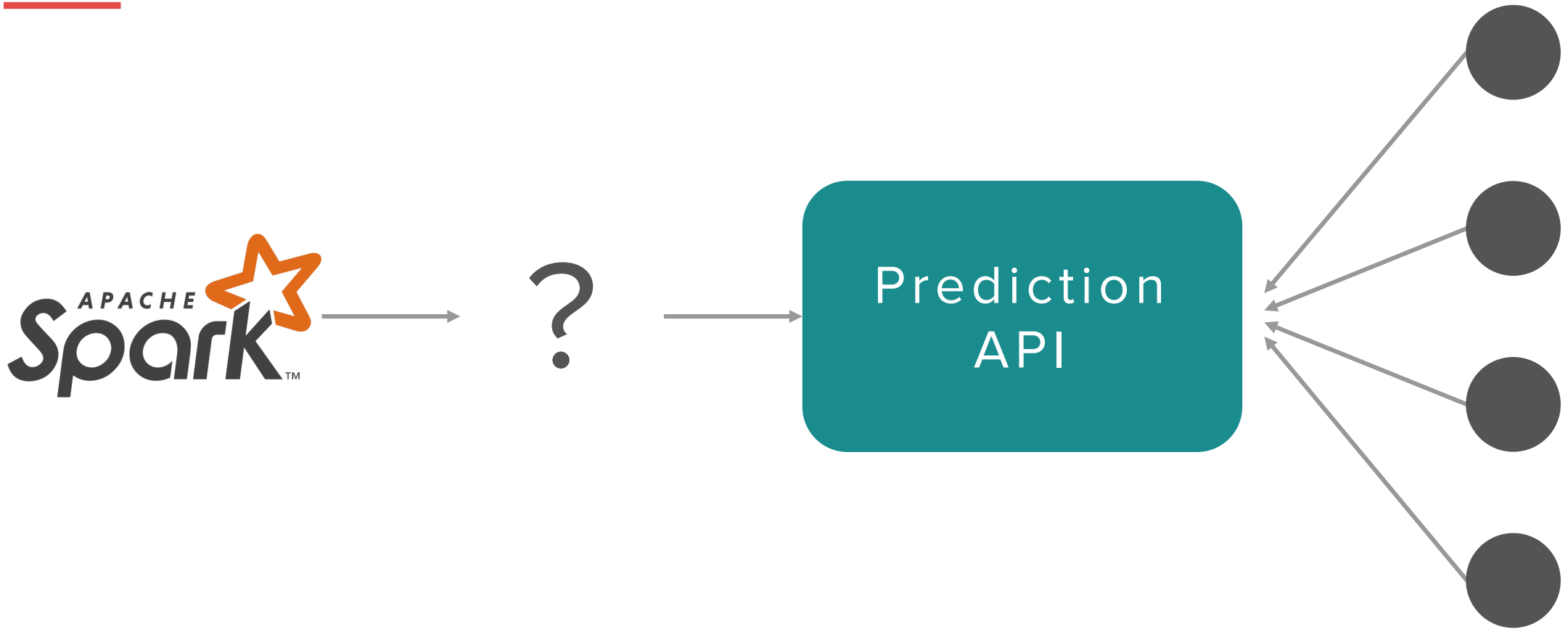


We fit our pipeline… now what?

Data Science Process

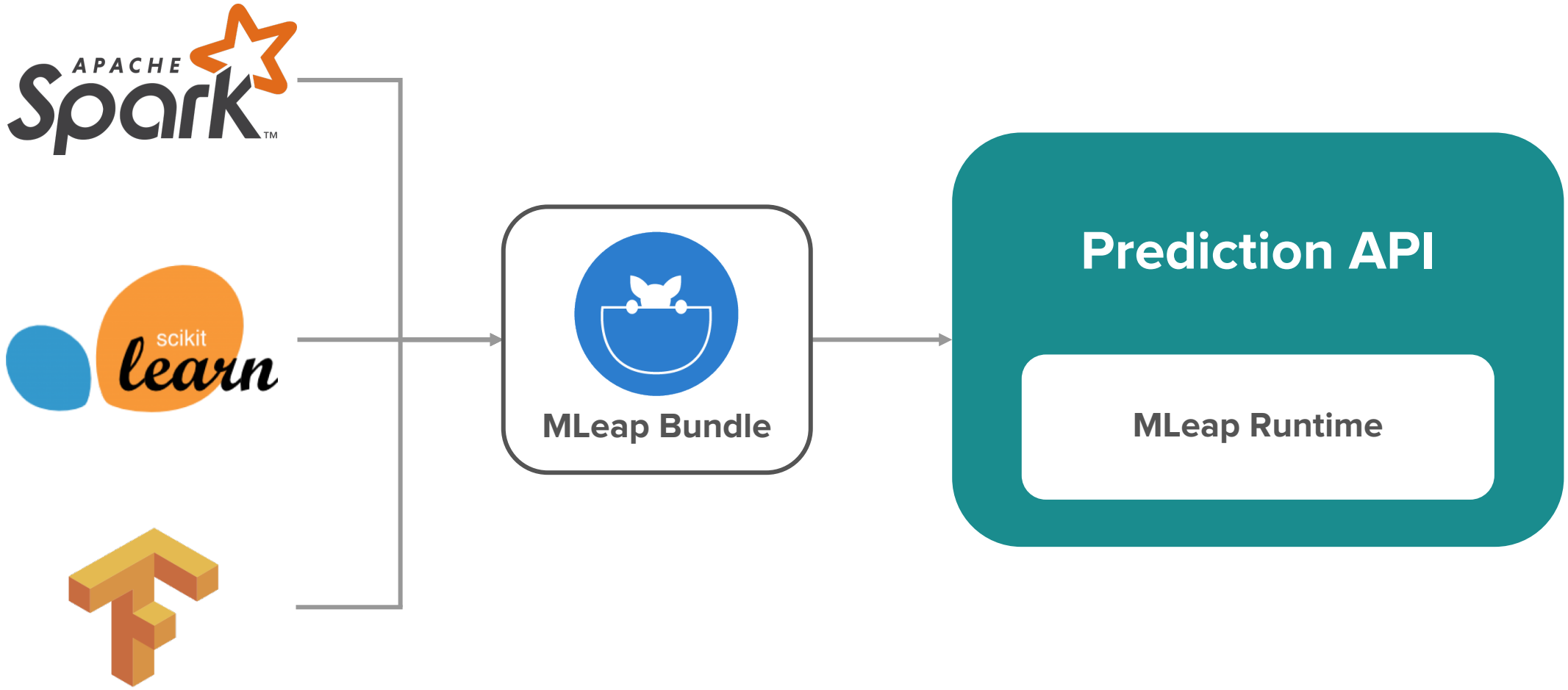
1. Data Collection
2. Machine Learning Pipelines
3. Model Deployment

Real-time scoring paradigm

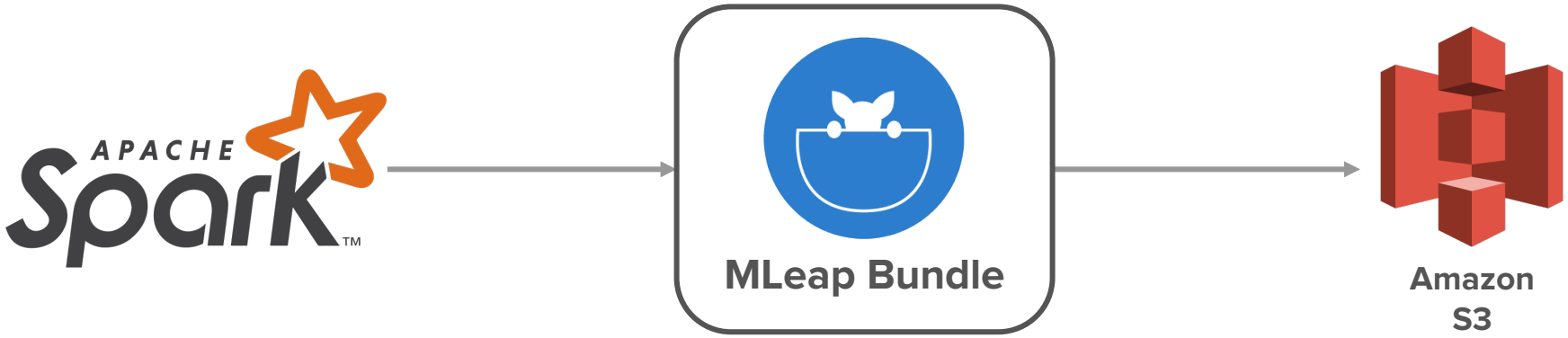


Model evaluation in real-time – with Spark

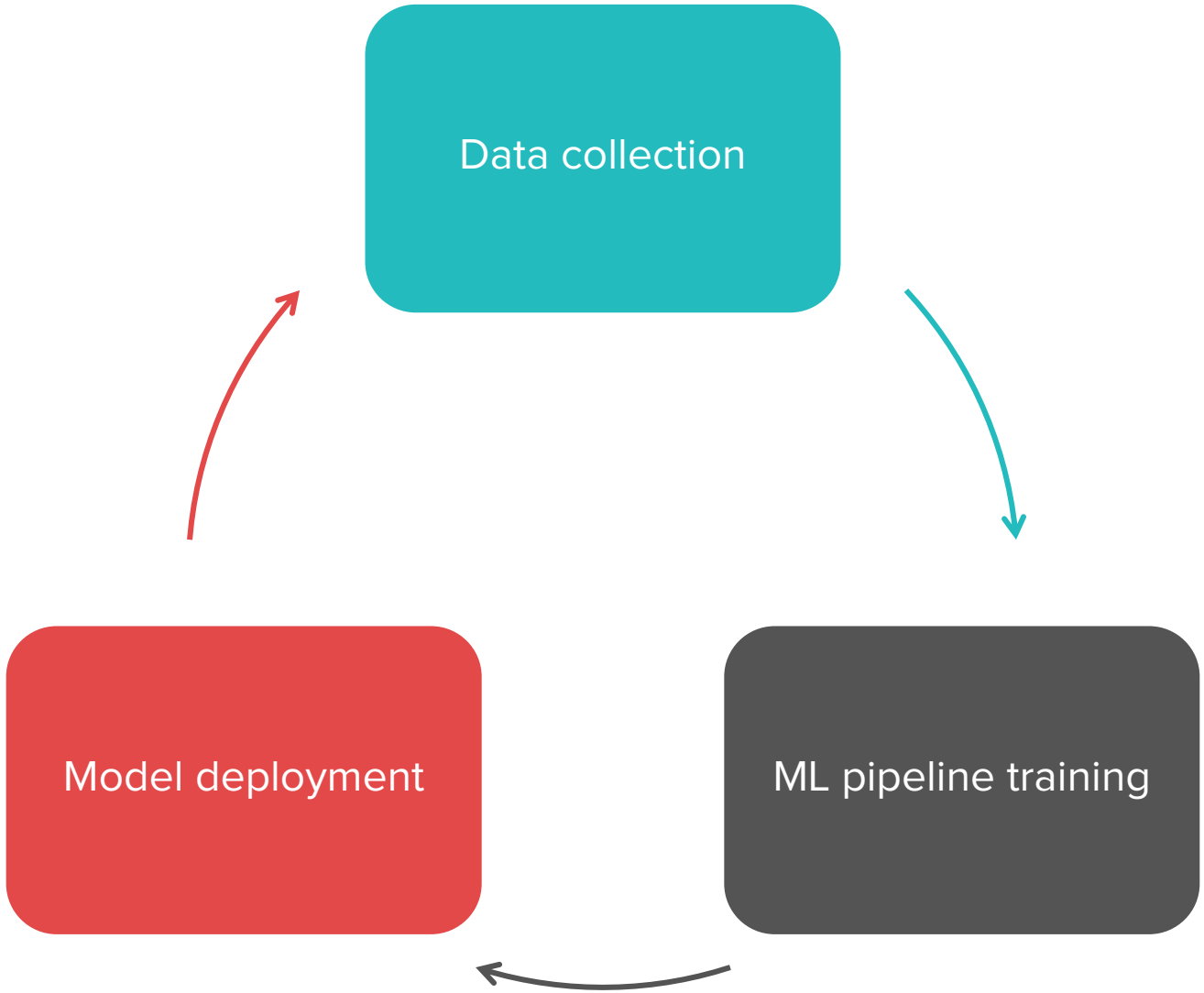
Model evaluation in real-time – with MLeap



Model deployment



Prediction API


Data collection
ML pipeline training
Model deployment

Recap

1. Data Collection
2. Machine Learning Pipelines
3. Model Deployment

New capabilities

>

New algorithms

300+

Real-Time Scoring Models

1

# of people required to productionize model

We're hiring!

- redventures.com/careers