2023

SENTIMENT ANALYSIS ON TWEETS

A Big Data Project

MOTIVATION

To build a real time sentiment analysis application on tweets, that enables users to get an overview of the mood on Twitter among the most popular tweets.

01

Use Case

02

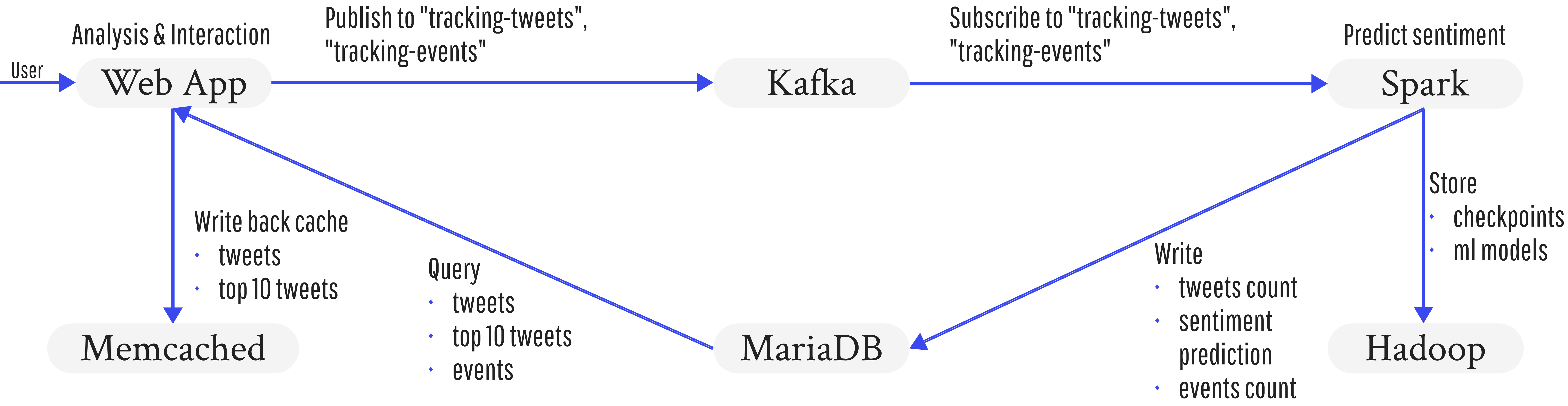
Implementation

03

Demo

Section 01

USE CASE



User

Analysis & Interaction

Web App

Publish to "tracking-tweets", "tracking-events"

Kafka

Subscribe to "tracking-tweets", "tracking-events"

Predict sentiment

Spark

Write back cache

- tweets
- top 10 tweets

Memcached

Query

- tweets
- top 10 tweets
- events

MariaDB

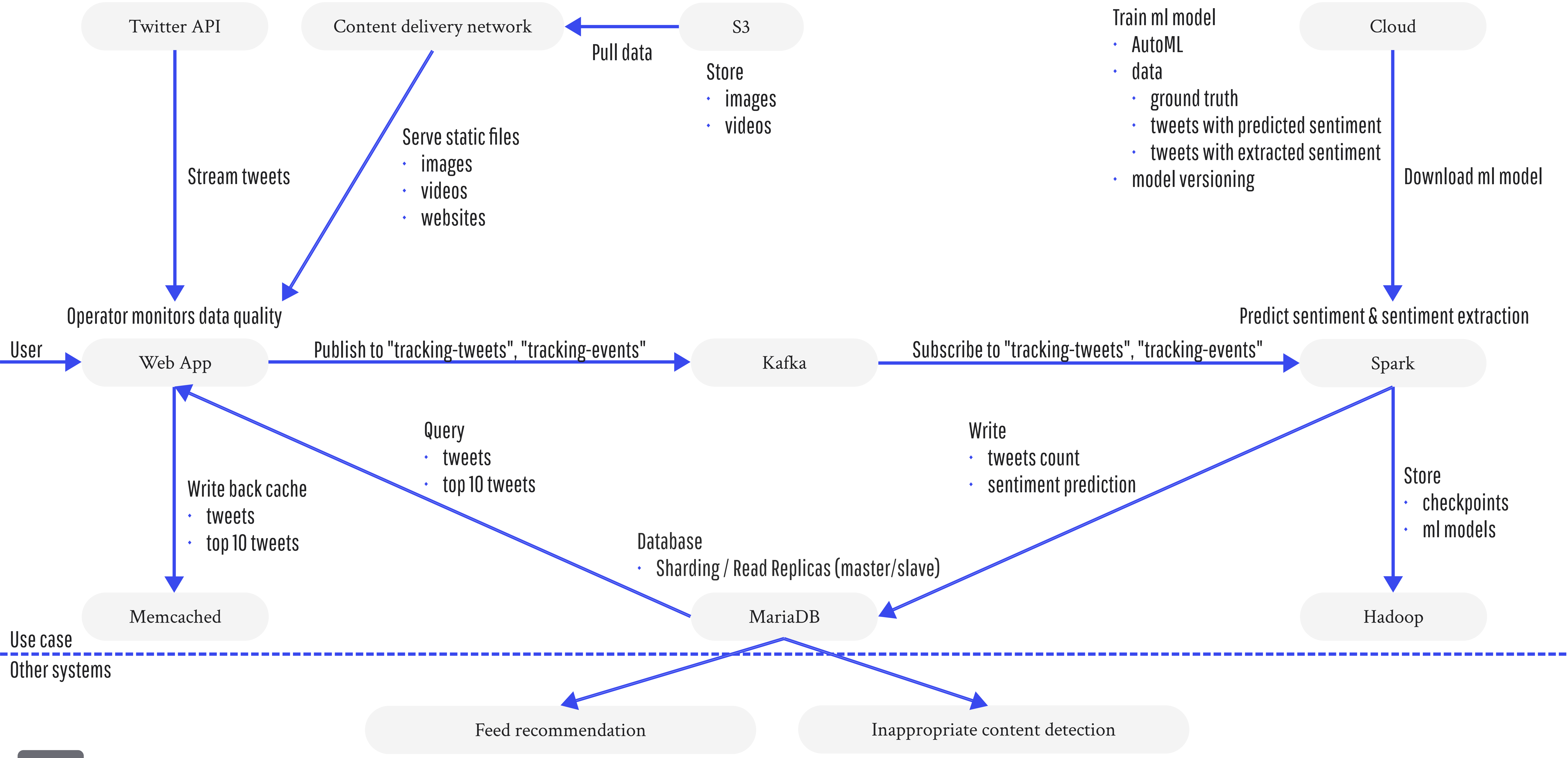
Write

- tweets count
- sentiment prediction
- events count

Store

- checkpoints
- ml models

Hadoop



Twitter API

Content delivery network

S3

Pull data

Store
- images
- videos

Train ml model
- AutoML
- data
  - ground truth
  - tweets with predicted sentiment
  - tweets with extracted sentiment
- model versioning

Cloud

Stream tweets

Serve static files
- images
- videos
- websites

Download ml model

Operator monitors data quality

Predict sentiment & sentiment extraction

User

Web App

Publish to "tracking-tweets", "tracking-events"

Kafka

Subscribe to "tracking-tweets", "tracking-events"

Spark

Query
- tweets
- top 10 tweets

Write
- tweets count
- sentiment prediction

Write back cache
- tweets
- top 10 tweets

Store
- checkpoints
- ml models

Database
- Sharding / Read Replicas (master/slave)

Memcached

MariaDB

Hadoop

Use case

Other systems

Feed recommendation

Inappropriate content detection

Section 02

IMPLEMENTATION

KEY FACTS

Code is fully functional

Implemented new MySQL connector, adjusted Docker files, database and Kafka

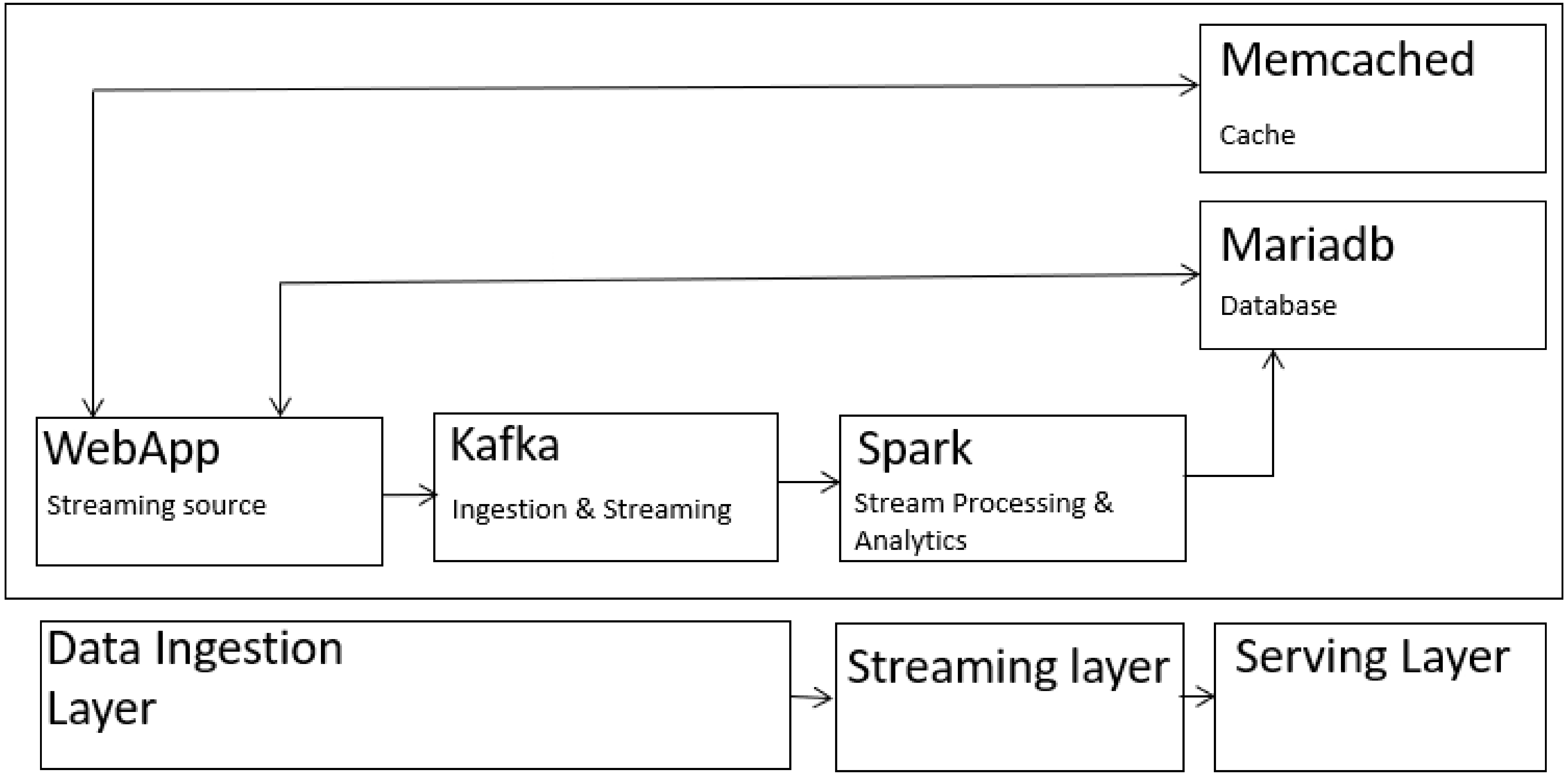
Consistent code style following Google Style Guides

Code splitting in index, config, database, kafka, memcached, utils

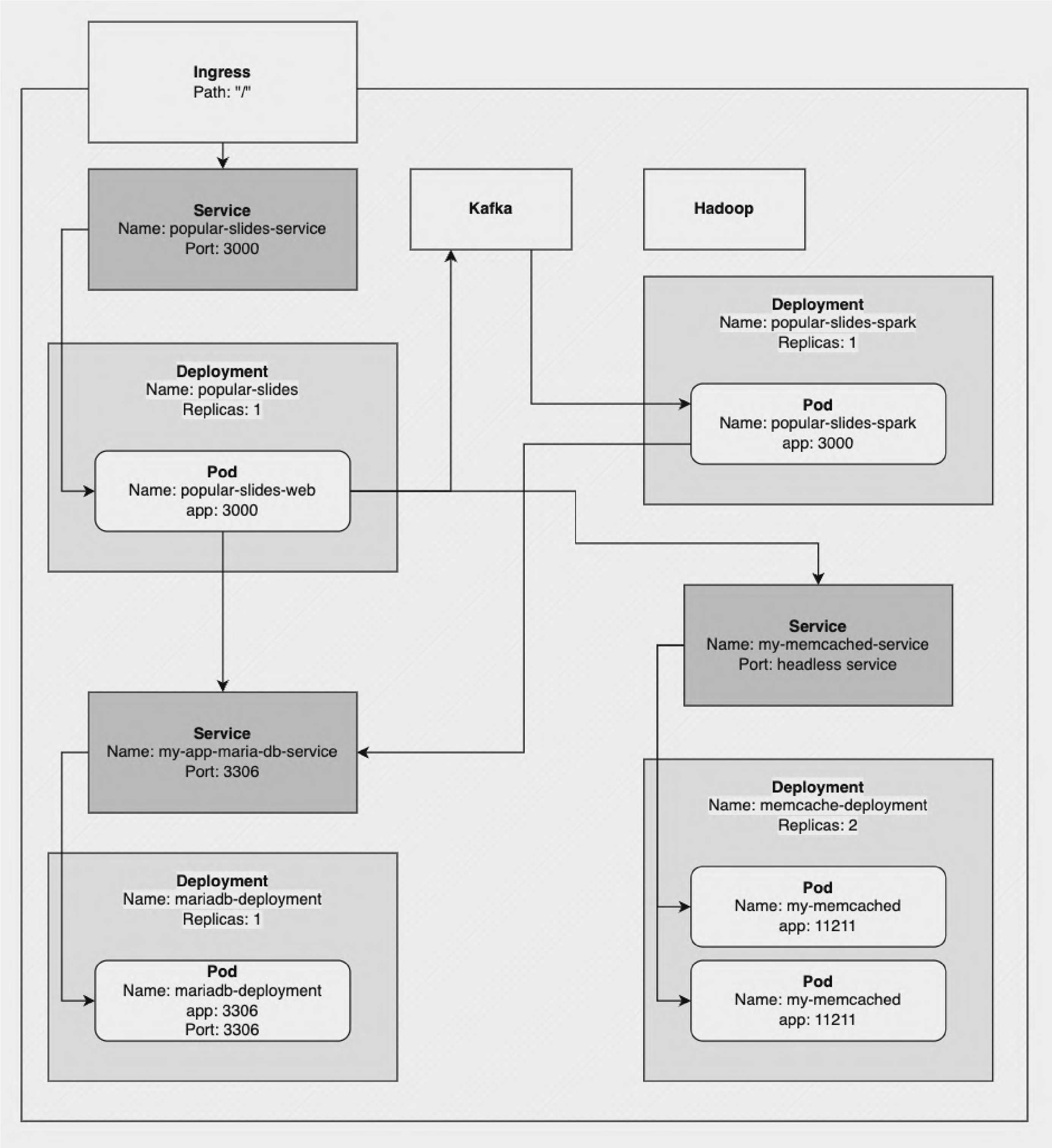
Development of web server in TypeScript for better tooling at any scale

Tweets are fetched via a button, clicked individually or streamed in background (1 tweet/10 sec)

ARCHITECTURE



KUBERNETES



Ingress
Path: "/"

Service
Name: popular-slides-service
Port: 3000

Kafka

Hadoop

Deployment
Name: popular-slides
Replicas: 1

Pod
Name: popular-slides-web
app: 3000

Deployment
Name: popular-slides-spark
Replicas: 1

Pod
Name: popular-slides-spark
app: 3000

Service
Name: my-memcached-service
Port: headless service

Service
Name: my-app-maria-db-service
Port: 3306

Deployment
Name: mariadb-deployment
Replicas: 1

Pod
Name: mariadb-deployment
app: 3306
Port: 3306

Deployment
Name: memcache-deployment
Replicas: 2

Pod
Name: my-memcached
app: 11211

Pod
Name: my-memcached
app: 11211

KAFKA MESSAGES

- Regex used to standardize tweet content
- **Performance optimization**
  - Kafka topics are published as batches
  - Spark uses only one driver to subscribe

Kafka topic used for ingesting application events

```
{
    "event_type": "streamed" // "clicked" or "fetched",
    "timestamp": 1604326237
}
```

Kafka topic used for ingesting tweets

```
{
    "tweet_id": 0,
    "tweet": "content",
    "timestamp": 1604326237
}
```

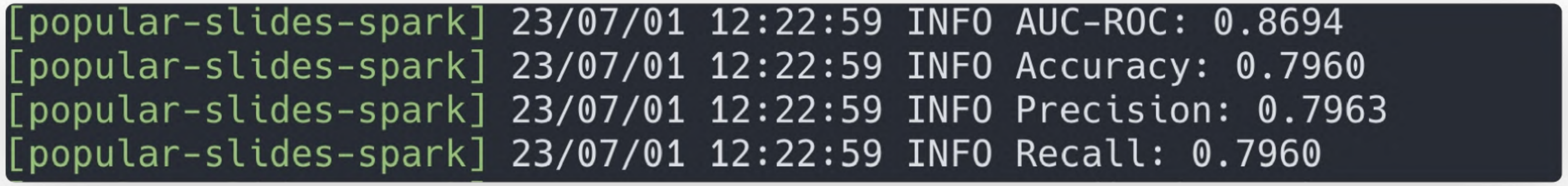
DATASET

- Sentiment140
- 1.6 million tweets
- Columns: sentiment, tweet_id, timestamp, username, tweet
- sentiment = negative, neutral, positive
- transformed → negative (0), positive (1)
- **Balanced** dataset → Accuracy most important metric

```
"0","1467811184","Mon Apr 06 22:19:57 PDT 2009","ElleCTF","I broke my leg "
```

ML MODELLING

- Binary classification problem
- NLP-Preprocessing (Tokenization, Count Vectorizer, Inverse Document Frequency)
- Logistic regression trained on Spark and saved in HDFS
- **Real time sentiment classification of a tweet as positive (class = 1) or negative (class = 0)**



```
[popular-slides-spark] 23/07/01 12:22:59 INFO AUC-ROC: 0.8694
[popular-slides-spark] 23/07/01 12:22:59 INFO Accuracy: 0.7960
[popular-slides-spark] 23/07/01 12:22:59 INFO Precision: 0.7963
[popular-slides-spark] 23/07/01 12:22:59 INFO Recall: 0.7960
```

TRY IT YOURSELF

http://xxx.xxx.xxx.xxx:8080/

Section 03

DEMO

Sentiment Analysis on tweets - a big data application

THANK YOU FOR YOUR ATTENTION