ROVIO

HOW TO KEEP OUR FLOCK HAPPY WITH APACHE FLINK ON AWS

ROVIO

AGENDA

ANALYTICS @ ROVIO

WHAT IS ROVIO AND HOW DOES ROVIO UTILIZE DATA?

DO WE NEED STREAMING?

WHY WOULD A COMPANY WITH A WORKING BATCH PROCESSING PIPELINE INVEST IN STREAMING? WHAT USE CASES ROVIO CURRENTLY RUNS IN STREAMING?

RUNNING FLINK IN PRODUCTION ON AWS

HOW DOES ROVIO OPERATE ITS PRODUCTION PIPELINE IN AWS?

ROVIO

STREAMLINE.

FAST
REACTIVE
ANALYTICS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 688191.

ROVIO

WE ARE HIRING!

ROVIO

POSITIONING

GAMES FIRST ENTERTAINMENT COMPANY

ROVIO

OUR MISSION

SET THE WORLD ON FIRE BY CRAFTING **THE BEST GAMES** AND **ENTERTAINMENT** IN THE WORLD

ROVIO

ROVIO IN AN EGGSHELL

FOUNDED IN

2003

CREATOR OF

ANGRY BIRDS™

BUSINESS UNITS

GAMES
3 STUDIOS

BRAND
LICENSING

PEOPLE

< 400
ROVIANS
ESPOO, STOCKHOLM, LONDON, SHANGHAI, LA

DOWNLOADS

4B
SINCE 2009

BOX OFFICE

$350M
THE ANGRY BIRDS™ MOVIE

SEQUEL

THE ANGRY BIRDS 2 MOVIE

SEPTEMBER 2019

ROVIO

HOW DO WE USE DATA?

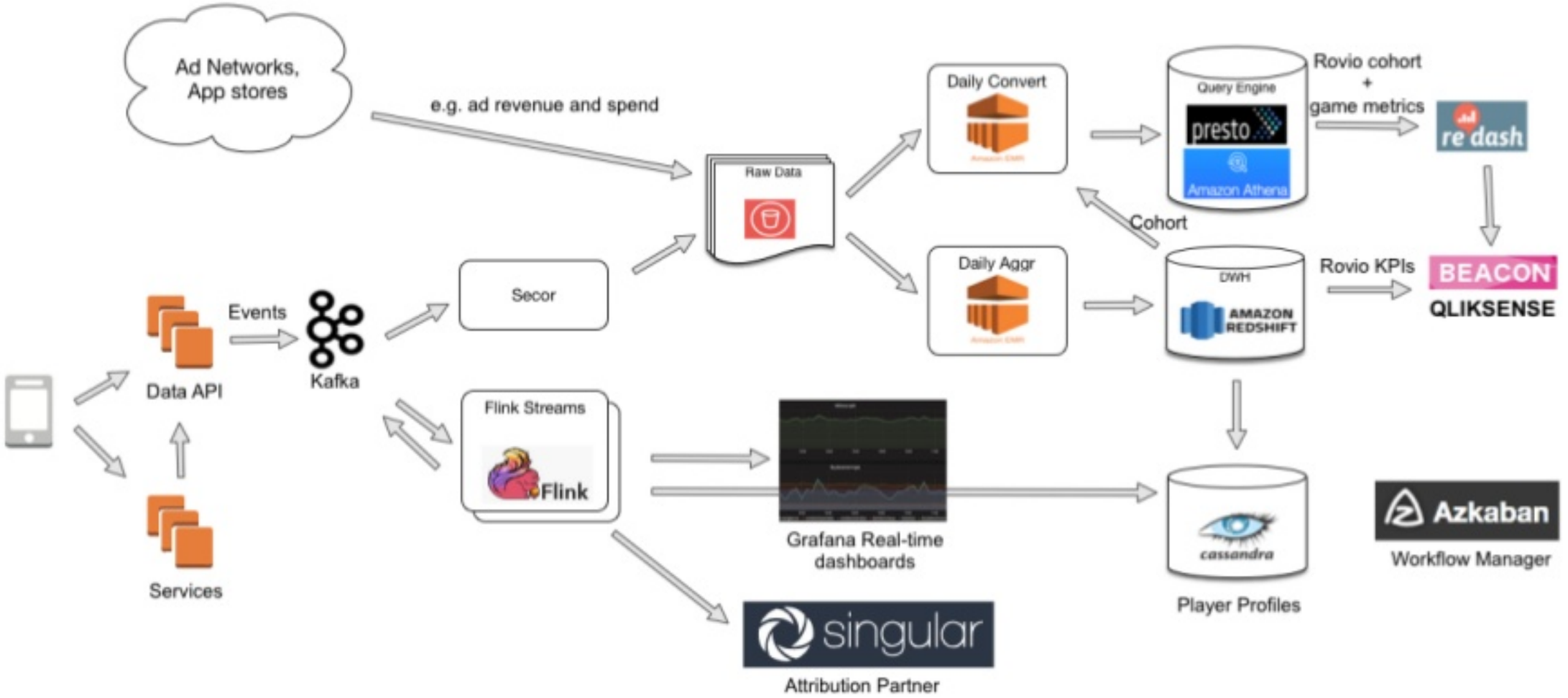
MARKET RESEARCH

PERFORMANCE MARKETING

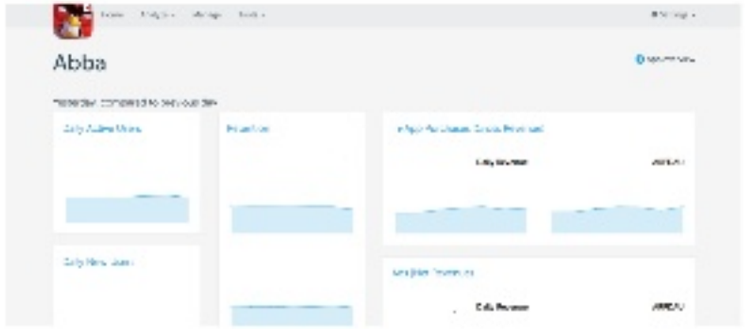
GAME OPTIMIZATION

ROVIO

PIPELINE



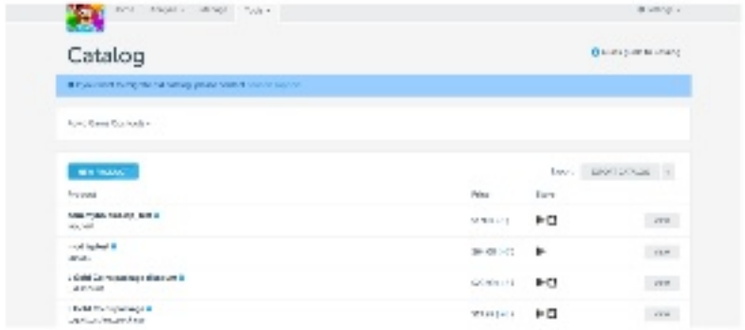
BEACON



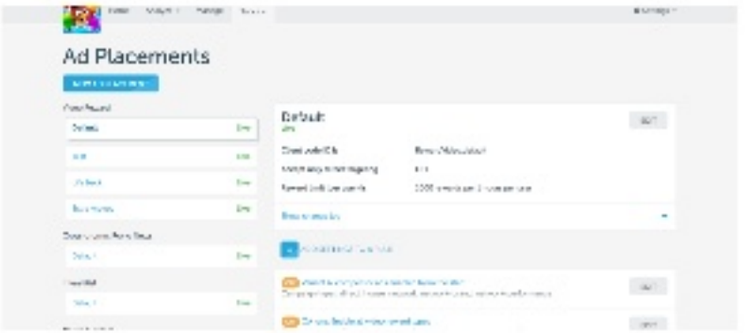
ANALYTICS

A/B TESTING

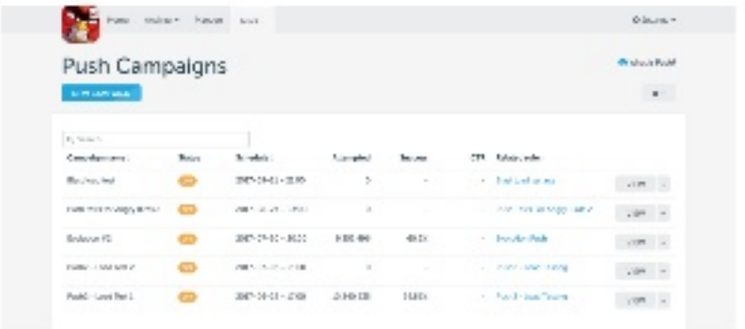
PERSONALIZATION



PAYMENT

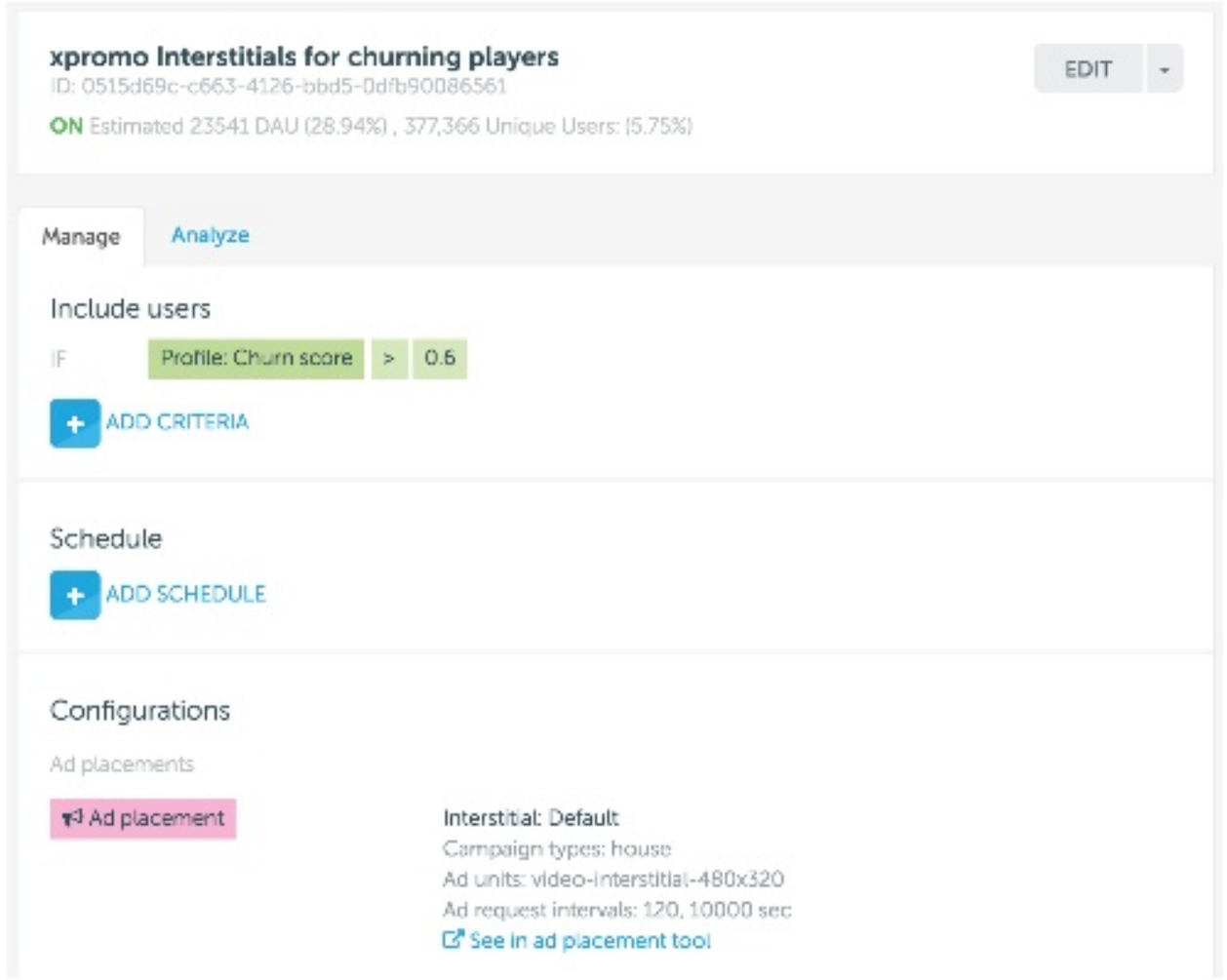


ADS



PUSH

BEACON - MANAGE



Show one forced interstitial and only show xpromo reward videos for churning players

27% more xpromo users

0% retention impact

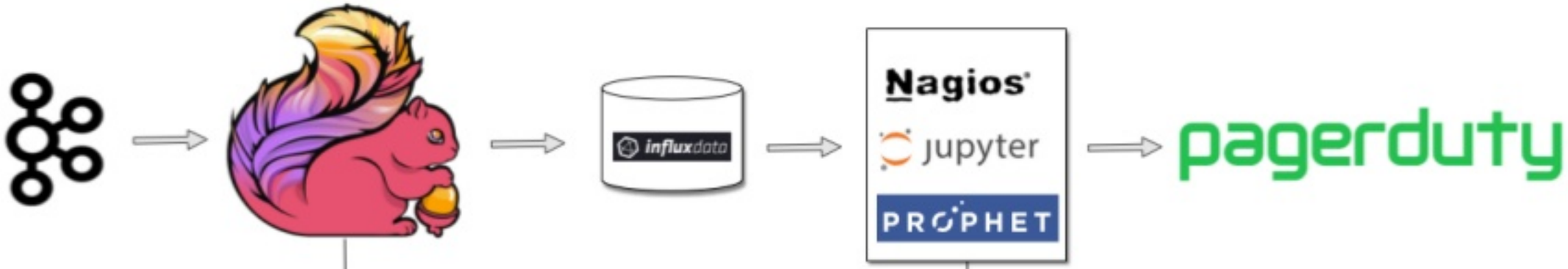
DO WE NEED STREAMING?

ROVIO

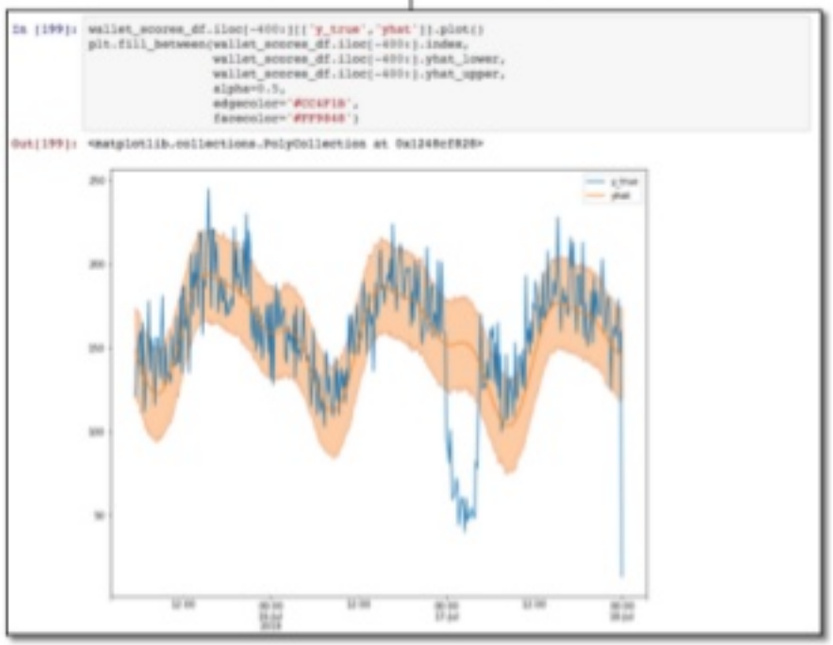
TIME TO DATA

ROVIO

ANOMALY DETECTION




```
-- jsonFields{ s.cid, s.dcid, s.cver, s.aid1 }
-- measurementName{ active_session }
SELECT
  s_cid AS app_id,
  s_dcid AS distribution_channel,
  s_cver AS client_version,
  COUNT(*) AS event_count,
  COUNT(DISTINCT s_aid1) AS unique_users,
  TUMBLE_START(rowtime, INTERVAL '60' SECOND) AS ts
FROM topics{ 'audit.session, audit.identity' }
GROUP BY
  s_cid,
  s_dcid,
  s_cver,
  TUMBLE(rowtime, INTERVAL '60' SECOND)
```

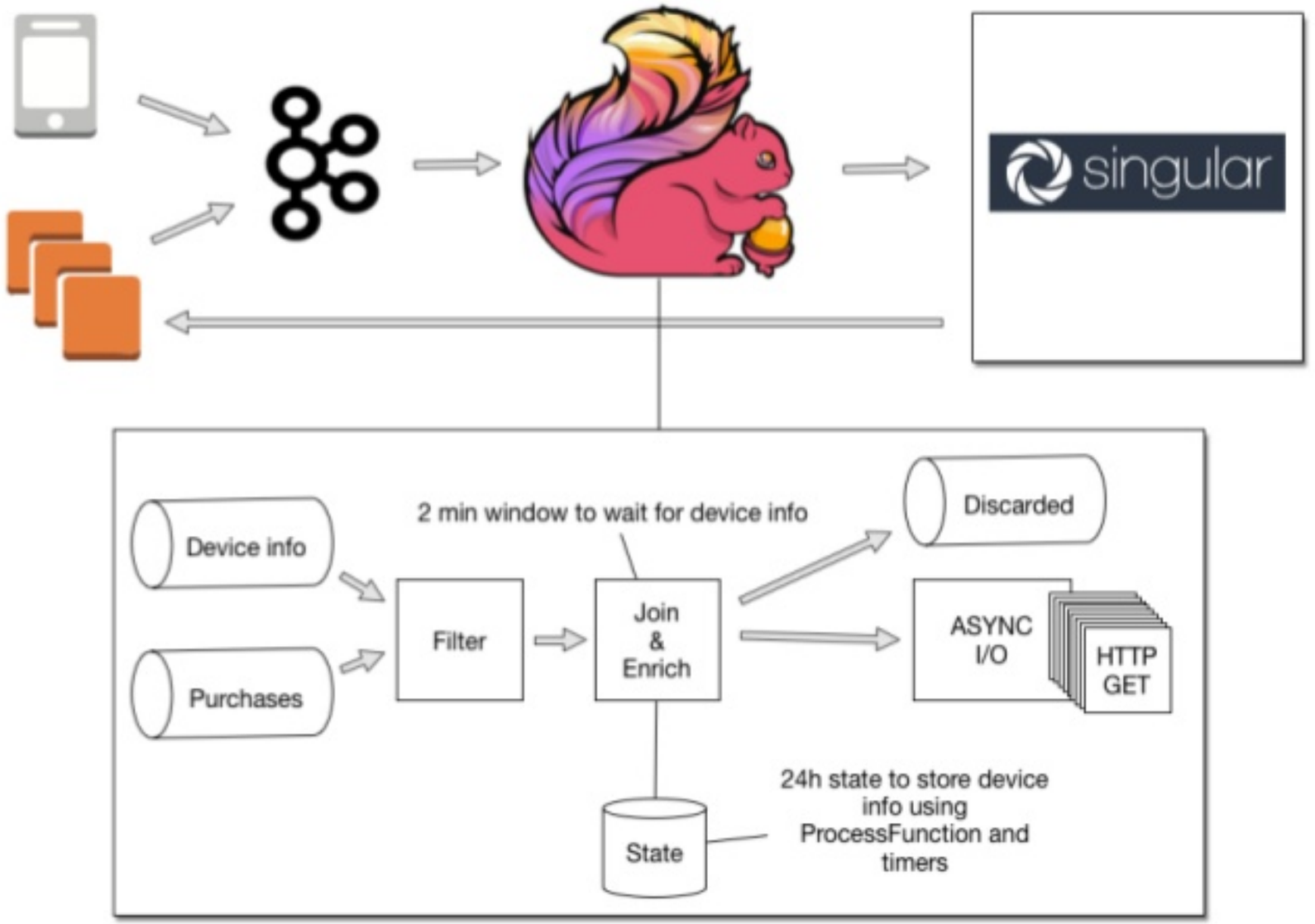


ROVIO

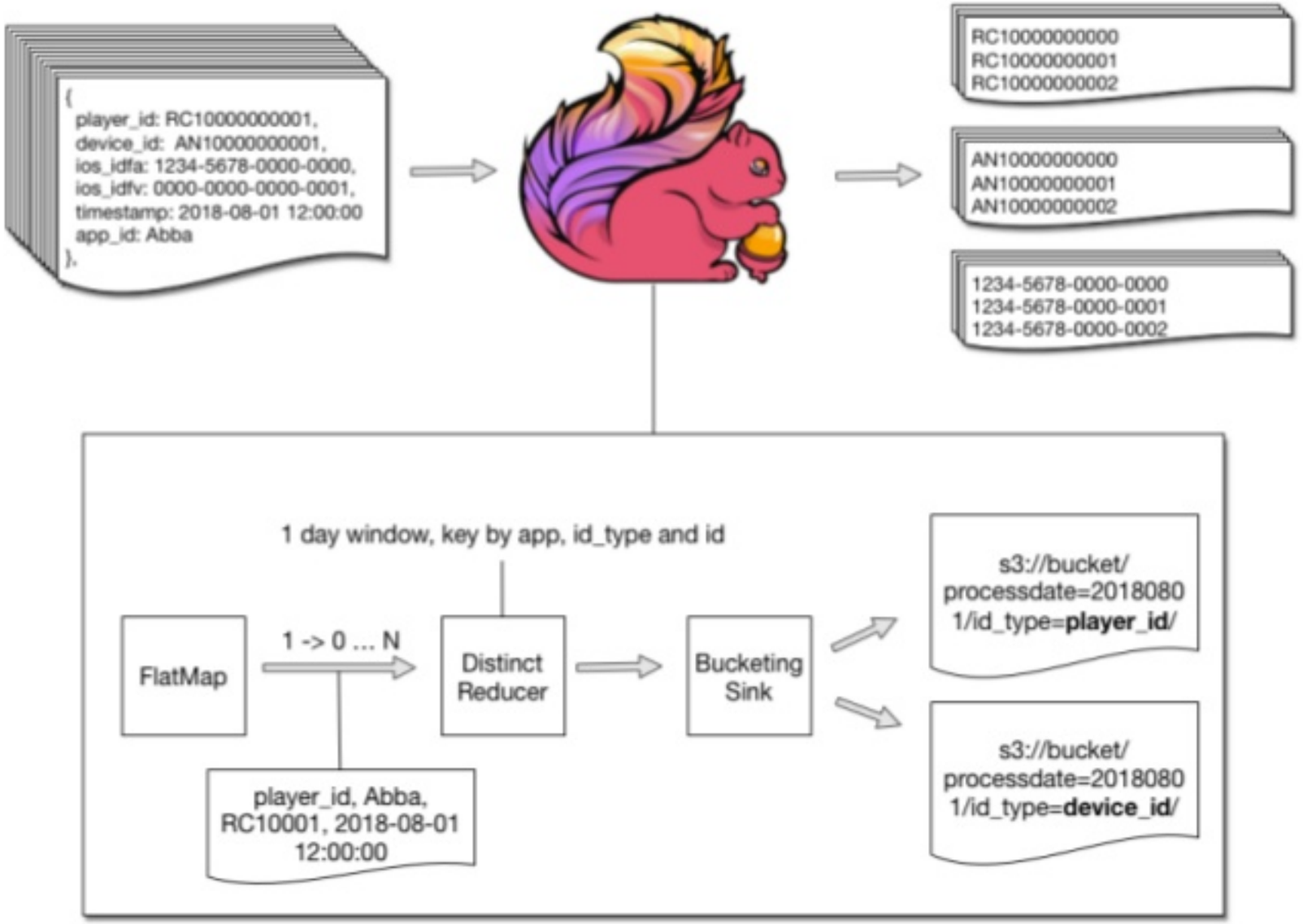
PERSONALIZATION


```
{
  "dynamicAppFieldMappings": [
    {
      "id": "progress_level",
      "name": "Progress Level",
      "type": "COUNT",
      "group": "USER_ACTIVITY",
      "appProfileColumn": "abba_gdpr.dynamic_attributes_full_profile.progress_level",
      "pushProfileColumn": "abba_gdpr.push_grooming.progress_level",
      "cassandraColumn": null,
      "cassandraType": null,
      "options": null,
      "defaultValue": 0,
      "topic": "collector.Abba",
      "hiveType": "INTEGER",
      "eventType": "level_finished",
      "eventField": "m.progress_level",
      "redisKey": "",
      "filterExpression": ""
    }
  ]
}
```

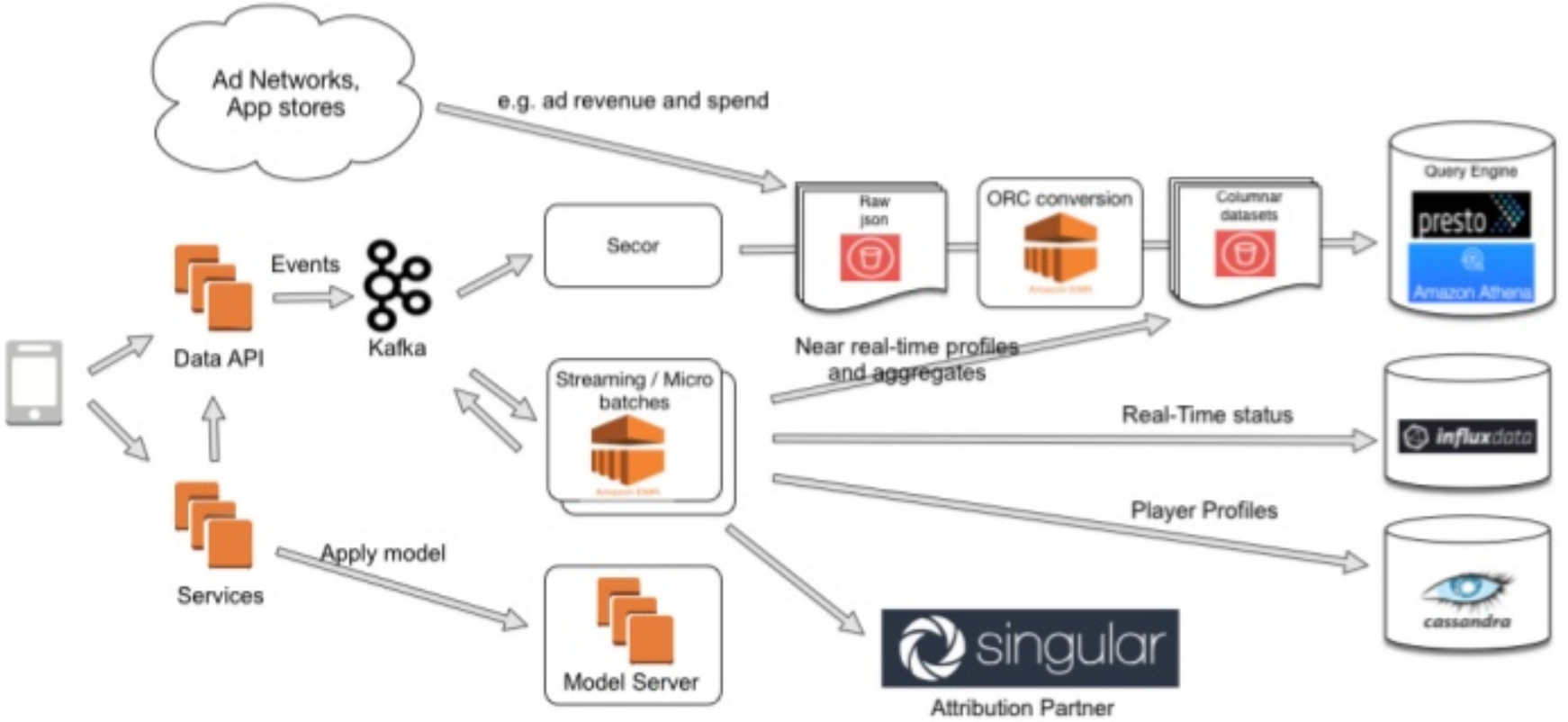
INTEGRATION WITH 3RD PARTY SYSTEMS



STREAMING DEDUPLICATION



INPROGRESS: STREAMING 4 ALL?



WHAT WE NEED TO RUN STREAMING PRODUCTION PIPELINE?

ROVIO

WE NEED...

DATA

Backfilling the historical data

SLA

System must operate 24/7

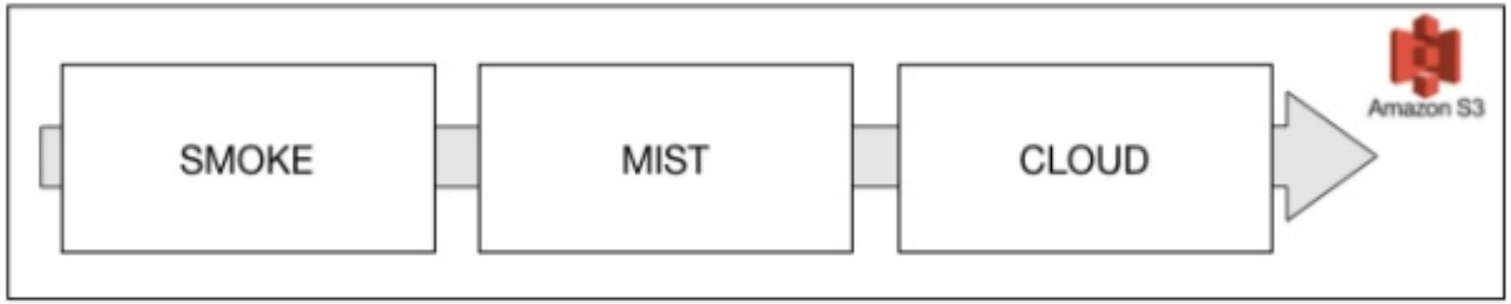
CI

Implement and deploy features care free

ROVIO

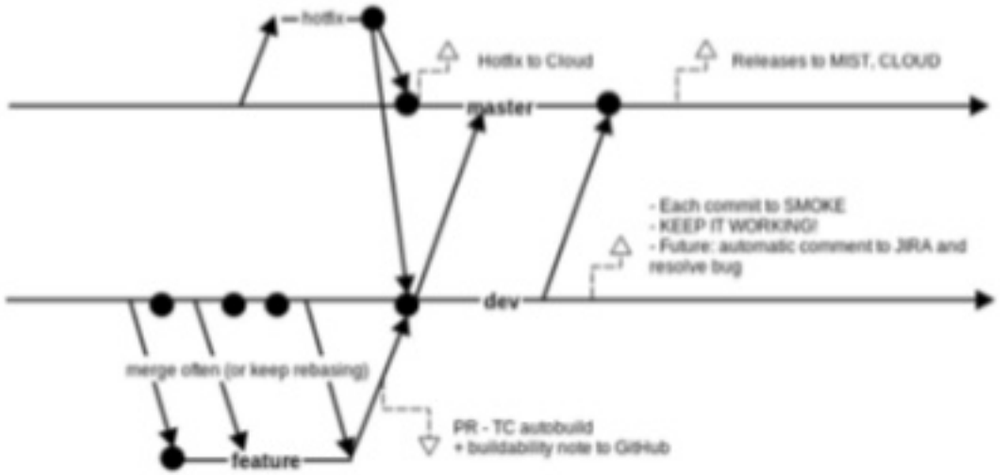
HOW DO WE OPERATE AND MONITOR FLINK STREAMS?

ROVIO



TEAMCITY PROJECTS TO BUILD AND DEPLOY FLINK STREAM UBER JARS TO DEVELOPMENT, STAGING AND PRODUCTION S3 BUCKETS.

GITFLOW INSPIRED BRANCHING STRATEGY IN GITHUB.

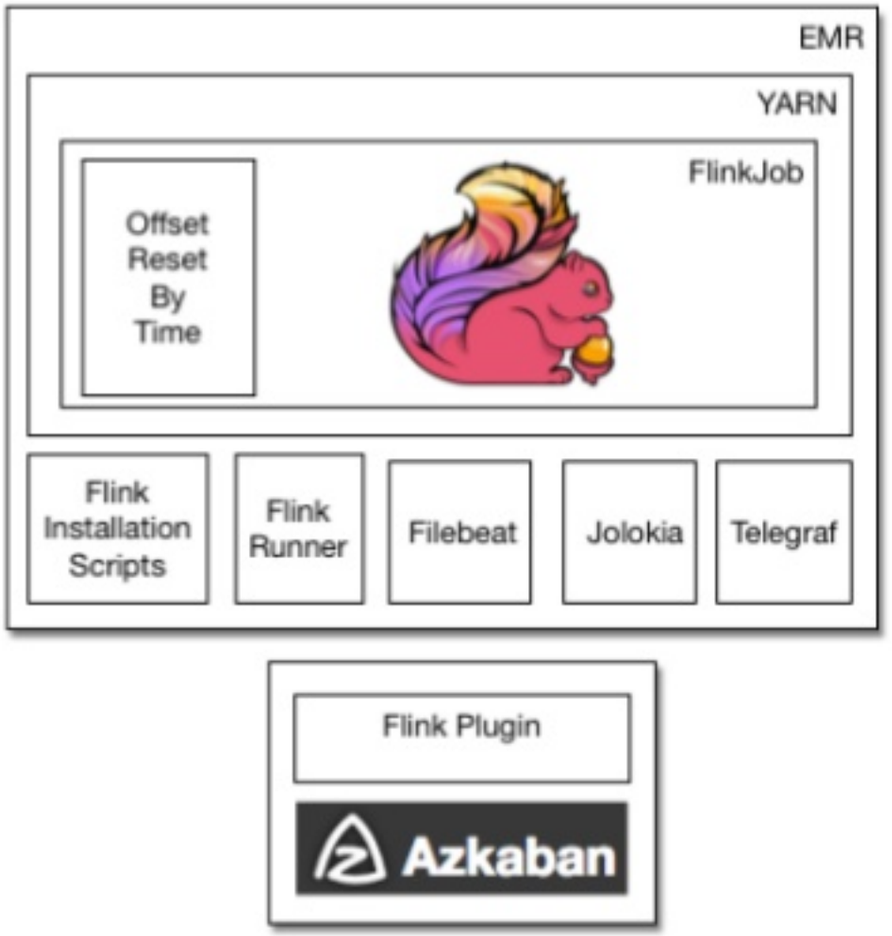


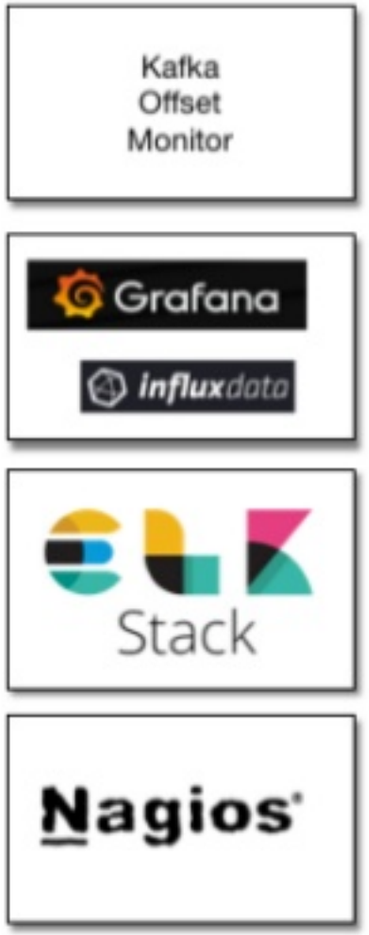
ROVIO

ELASTIC MAPREDUCE WITH YARN

BOOTSTRAP ACTIONS TO SETUP FLINK, LOGGING AND MONITORING

CUSTOM TOOLS AND LIBRARIES MOSTLY FOR STATE HANDLING





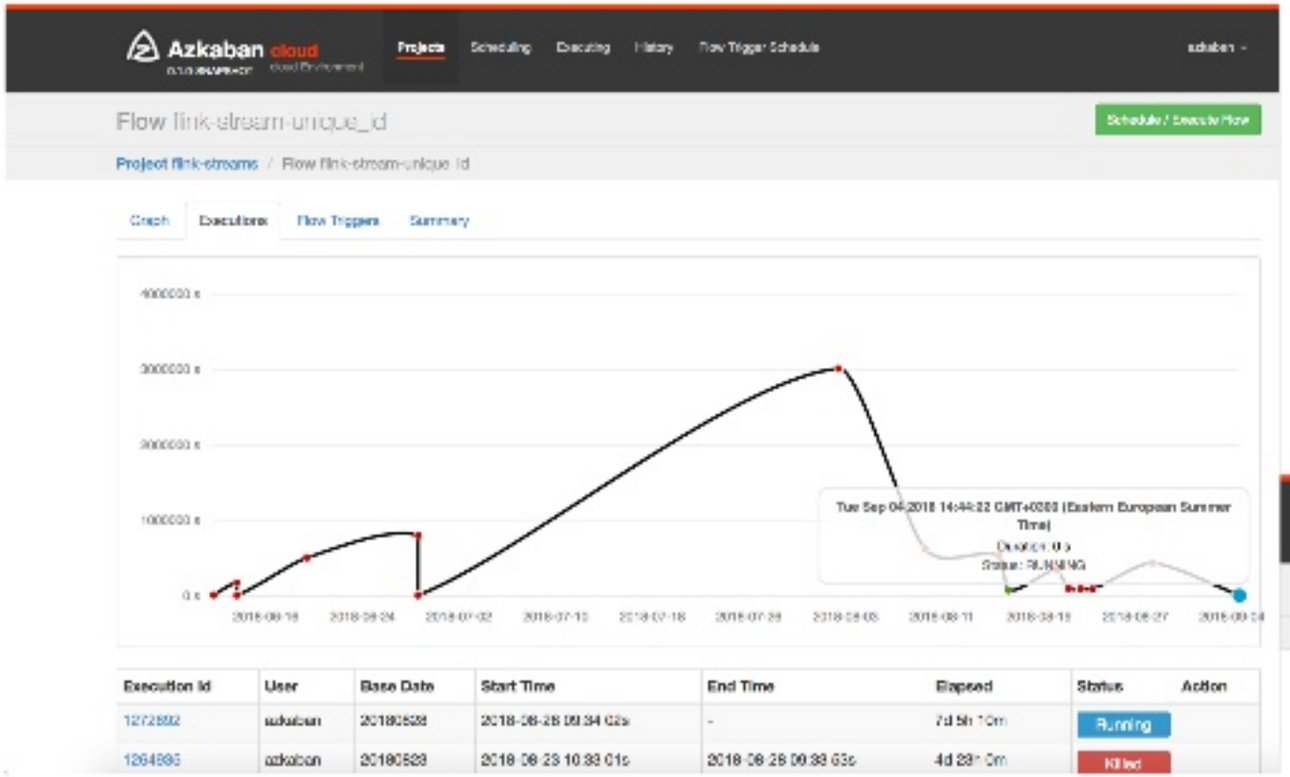
KAFKA OFFSET MONITOR TO MONITOR CONSUMER LAG

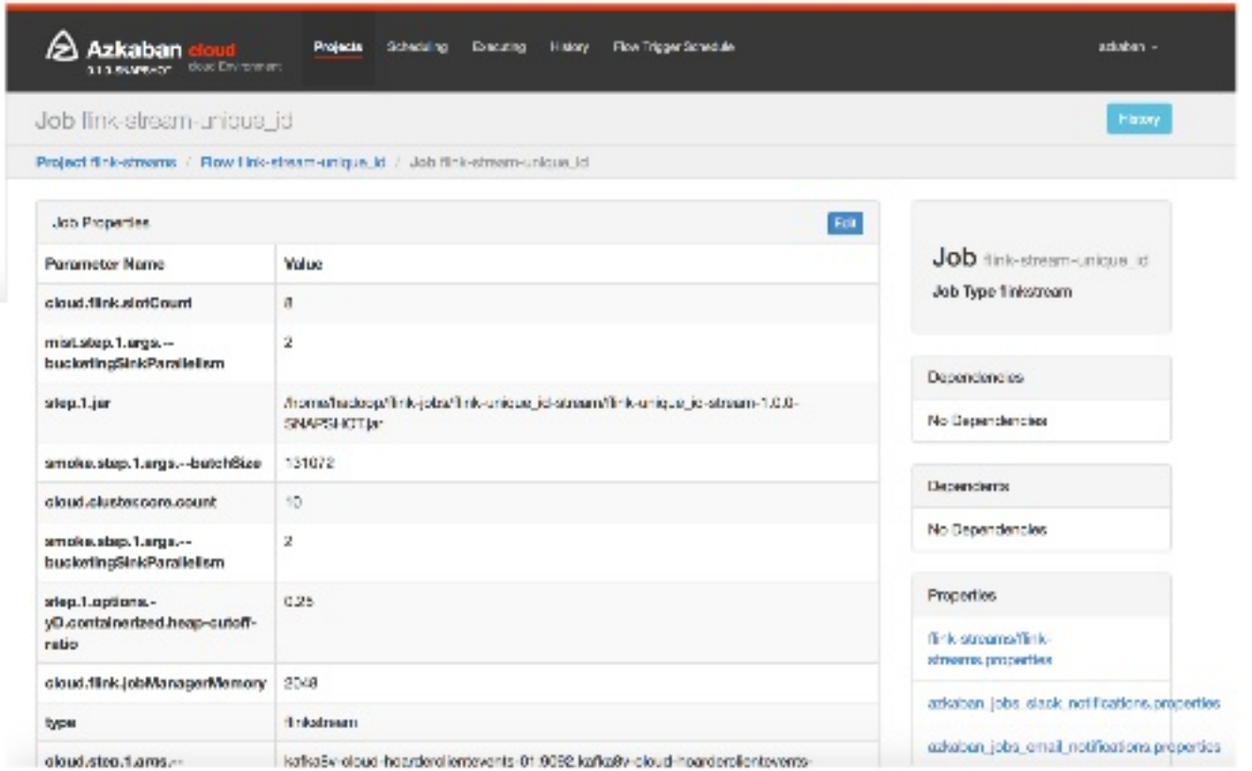
REAL-TIME TIME SERIES DATA IN GRAFANA

LOGS IN KIBANA

MONITORS OFFSET LAG, ACCUMULATORS, BACKPRESSURE AND ANOMALIES

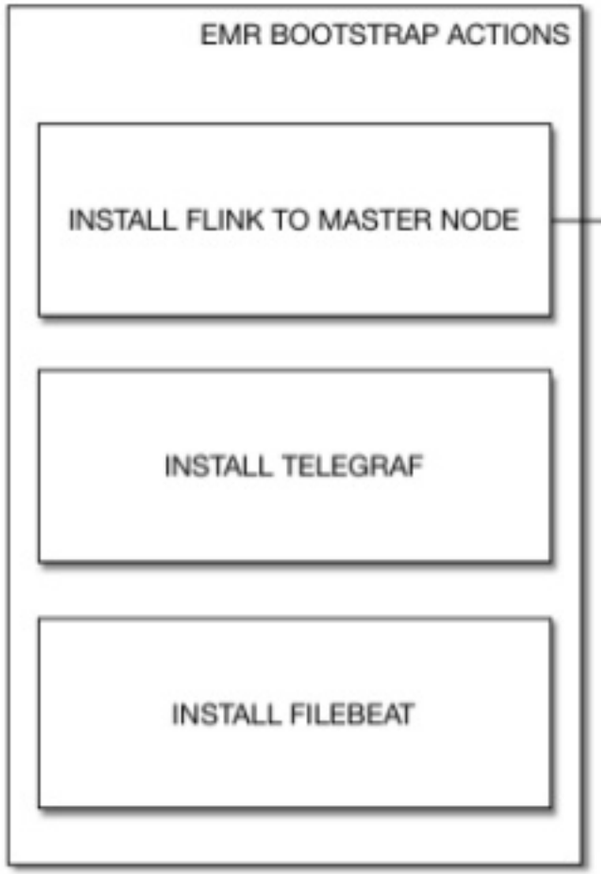
AZKABAN WORKFLOW MANAGER WITH FLINK PLUGIN USED TO ORCHESTRATE ENABLED EMR CLUSTERS


Azkaban cloud
Projects
Scheduling
Executing
History
Flow Trigger Schedule
azkaban
Flow flink-stream-unique_id
Schedule / Execute Flow
Project flink-streams / Flow flink-stream-unique_id
Graph
Executions
Flow Triggers
Summary
Execution Id
User
Base Date
Start Time
End Time
Elapsed
Status
Action
Running
Killed


Azkaban cloud
Projects
Scheduling
Executing
History
Flow Trigger Schedule
azkaban
Job flink-stream-unique_id
History
Project flink-streams / Flow flink-stream-unique_id / Job flink-stream-unique_id
Job Properties
Edit
Parameter Name
Value
cloud.flink.slotCount
8
step.1.jar
cloud.flink.jobManagerMemory
2048
type
flinkstream
Job flink-stream-unique_id
Job Type flinkstream
Dependencies
No Dependencies
Dependents
No Dependencies
Properties

HOW TO BOOTSTRAP FLINK TO EMR CLUSTER?

ROVIO




```
#!/bin/bash
set -e
#
# This script installs Apache Flink on master node
#
JOLOKIA_PCK=jolokia-jvm-1.3.7-agent.jar
aws s3 cp s3://ds-analytics-emrjobs-${var:env}/lib/${JOLOKIA_PCK} /home/hadoop/

if [ -z $4 -o $4 = " " ]; then
    echo "not enough args"
    exit 1
fi
JOB_NAME=$2
CUSTOM_CONF=$3
START_TIME=$4

IS_MASTER=$(grep "\"isMaster\": true" /mnt/var/lib/info/instance.json | wc -l)
if [ ${IS_MASTER} = "1" ]; then
    echo "We are on master node... Install Flink."
    pushd /home/hadoop
    FLINK_PCK="$1"
    aws s3 cp s3://ds-analytics-emrjobs-${var:env}/lib/${FLINK_PCK} /home/hadoop/
    tar -xvf ${FLINK_PCK}
    popd
    FLINK_DIR=$(find /home/hadoop/flink-[0-9]* -maxdepth 0 -type d)
    sed "s/\${JOB_NAME}/${JOB_NAME}/g;\
     s/\${JOLOKIA_PCK}/${JOLOKIA_PCK}/g;\
     s/\${START_TIME}/${START_TIME}/g\
     " ${CUSTOM_CONF}\
     >> ${FLINK_DIR}/conf/flink-conf.yaml
else
    echo "Not a master... Skipping Flink installation."
fi
```

ROVIO

HOW TO START FROM THE LATEST RESTORE POINT?

ROVIO

```
public String getLatestRestorePoint() {

    log.info("Searching for latest savepoint..");
    S3SourceWithDate latestSavePoint = getLatestSavepoint();
    log.info("latestSavePoint: " + latestSavePoint);

    log.info("Searching for latest checkpoint..");
    S3SourceWithDate latestCheckPoint = getLatestCheckpoint();
    log.info("latestCheckPoint: " + latestCheckPoint);

    S3SourceWithDate restorePoint = Stream.of(latestSavePoint, latestCheckPoint)
            .filter(Objects::nonNull).max(Comparator.comparing(S3SourceWithDate::getLastModified)).orElse(null);
    if (restorePoint == null) {
        return null;
    }
    return restorePoint.source.resolveNativeLink(flatBirdHelper);
}
```

FIND THE LATEST CHECKPOINT OR SAVEPOINT…

```
// flink.fromSavepoint allows manual/static override of where to restore from
String restorePointPath = props.getString("flink.fromSavepoint", null);
if (StringUtils.isBlank(restorePointPath) && savepointGateway.isEnabled()) {
    restorePointPath = savepointGateway.getLatestRestorePoint();
    if (restorePointPath == null) {
        log.warn("Checkpointing is enabled but couldn't find savepoints or checkpoints to resume from");
    }
}
```

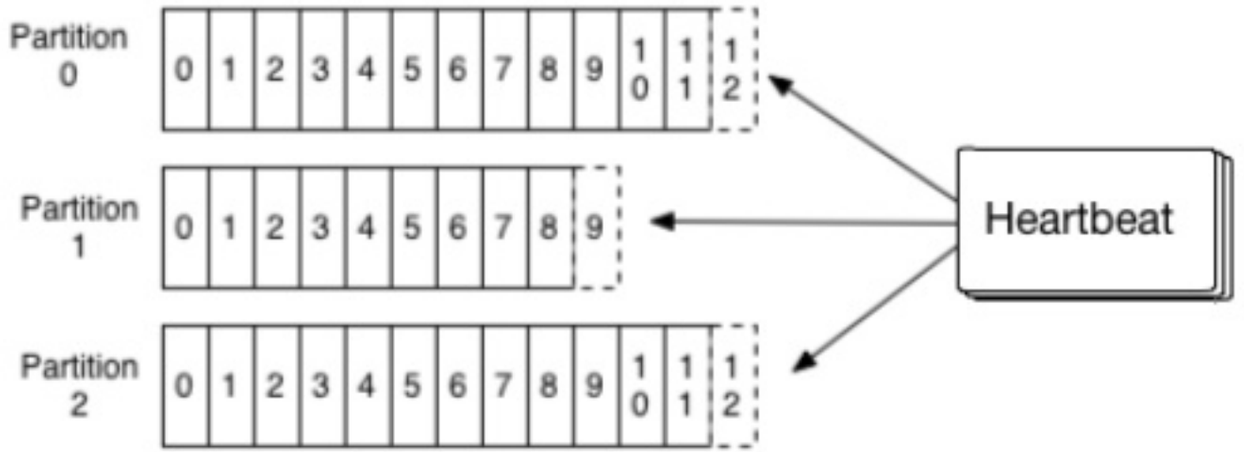
… AND PASS IT AS ARGUMENT!

ROVIO

HOW TO HANDLE IDLE PARTITIONS?

ROVIO

PRODUCER SENDS EMPTY MESSAGE WITH EVENT TIMESTAMP TO EVERY PARTITION



Source: https://kafka.apache.org/intro

ROVIO

HOW TO INITIALIZE THE STATE?

ROVIO

1. Use same uid and state name in the init and main job steps
2. Backfill history from files using streaming and PROCESS_CONTINUOSLY mode
3. Monitor Flink job and save the state when no data is being processed anymore
4. Start your streaming job from the save point and start processing Kafka events from specific point of time

ROVIO

WITH FLINK 1.5, KAFKA 0.10 AND ABOVE

1. **Add timestamp to Kafka message**
2. **Use FlinkKafkaConsumer setStartFromTimestamp(…)**

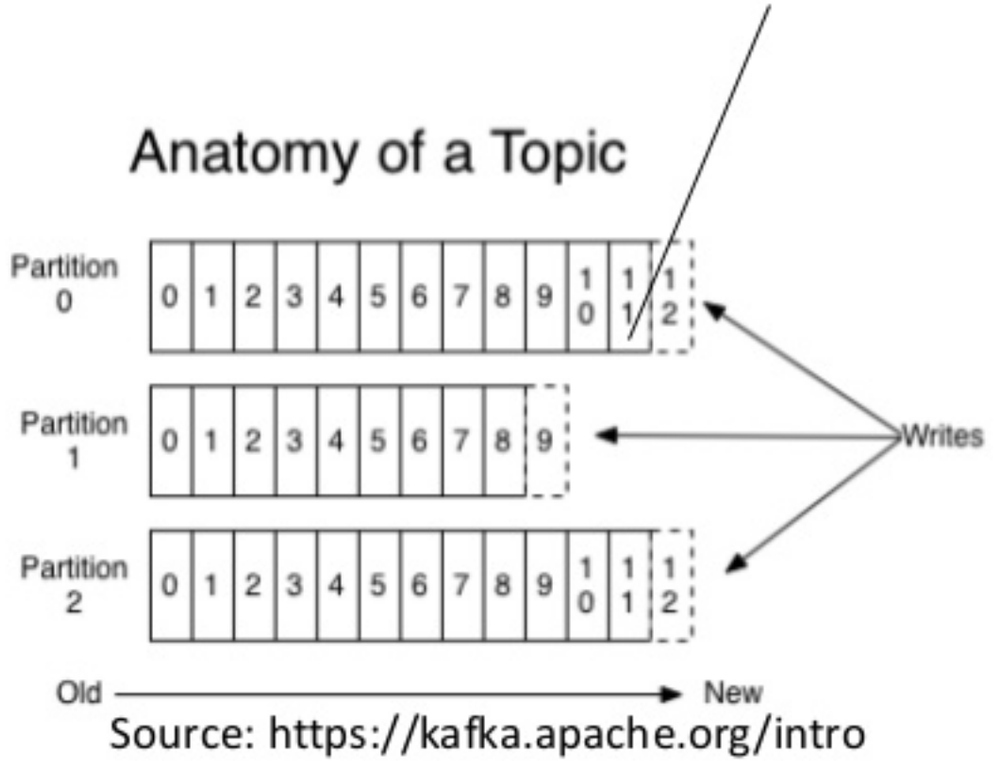
How did we do it for older versions?

ROVIO

ANATOMY OF KAFKA TOPIC

```
{
  timestamp: 2018-08-01 12:00:00
  .
  .
  .
}
```

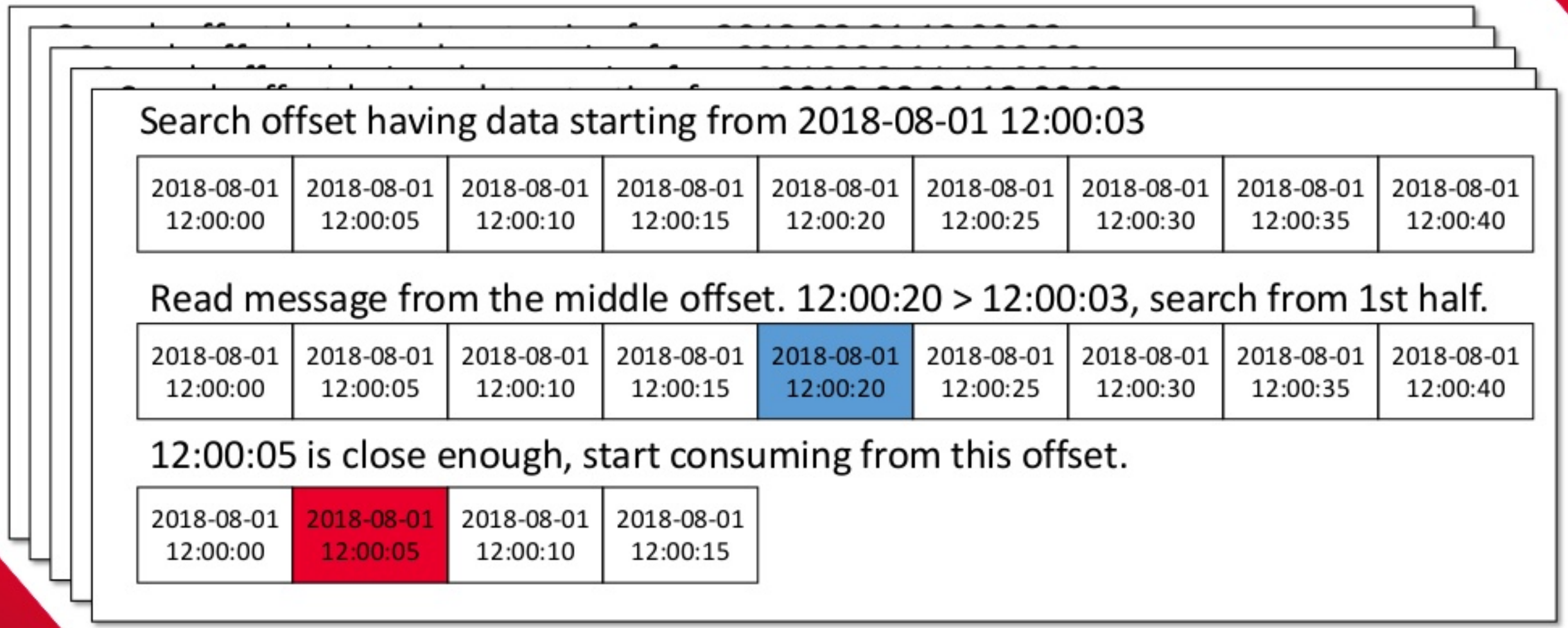
ARRAYS!!!



Source: https://kafka.apache.org/intro

How does one search data from arrays?

BINARY SEARCH



ROVIO

SUMMARY

ROVIO MAKES GREAT GAMES AND DATA IS HEAVILY USED IN EVERY STEP OF THE GAMES LIFE-CYCLE

STREAMING IS USED TO IMPROVE TIME TO DATA ESPECIALLLY IN CASES WHERE AUTOMATED DECISION MAKING IS BEING DONE

FOR PRODUCTION USE ONE NEEDS ADDITIONAL FEATURES ON TOP OF THE VANILLA FLINK PLATFORM ESPECIALLY AROUND OPERATIONS AND STATE MANAGEMENT

ROVIO

THANKS!

ROVIO