# flight

# flight

## Scaling Answers Back-End

**Ed Solovey**

Staff Software Engineer, Answers

@edsolovey

# What is Answers?

# Answers Is...

REAL-TIME

OPINIONATED

flight

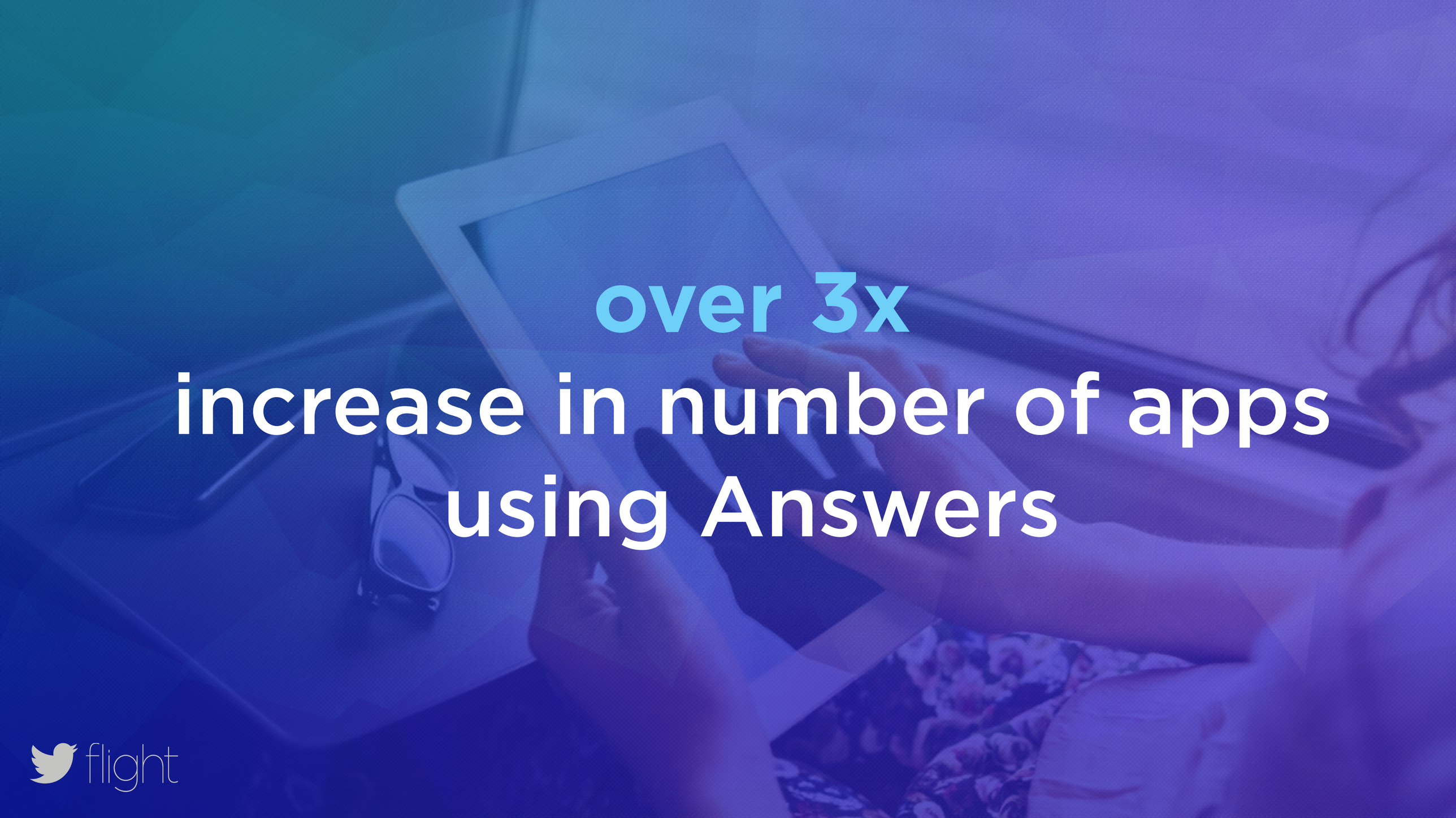

fabric.io
Cannonball 2.3.1
io.fabric.example.cannonball
Jeff Seibert
Twitter, Inc.
ACTIVE USERS
2,019
RIGHT NOW
DAILY ACTIVE USERS
77.9 k
▲8.7%
DAILY NEW USERS
7.1 k
▲11.4%
MONTHLY ACTIVE USERS
2.1 M
▲3.8%
CRASH-FREE USERS
96.5 %
▼-2.6%
SESSIONS
248.2 k
▲10.2%
SESSION LENGTH
0:35 SEC
0.0%
DAILY ACTIVE USERS
77.9k
TODAY
300K
200K
100K
0
SEP 16
SEP 21
SEP 28
OCT 05
OCT 12
OCT 16

flight

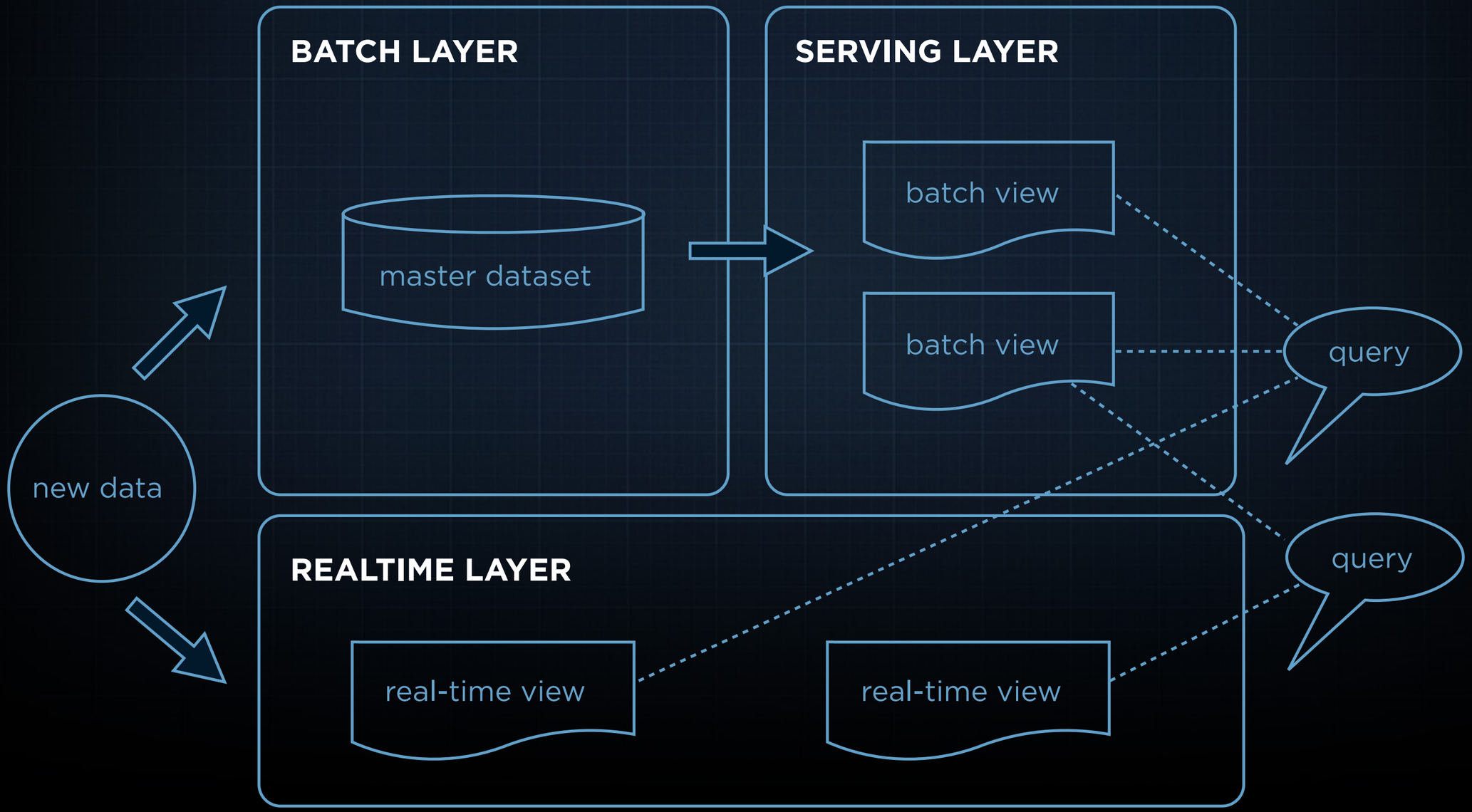
# SDK Built On Events

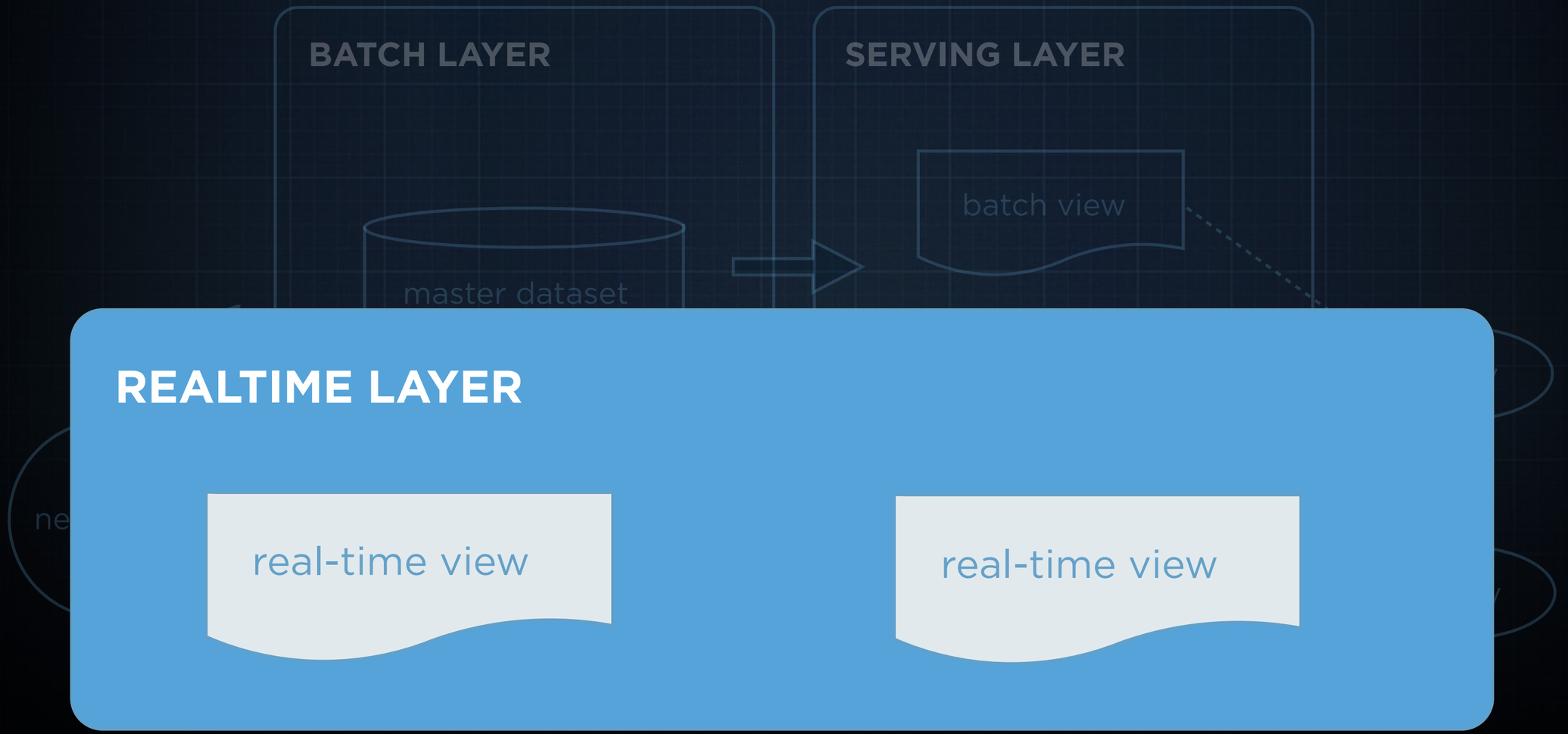
- Install
- App Foreground
- App Background
- Crash
- Predefined Events / Custom Events

flight


over 3x
increase in number of apps
using Answers
flight


1,500,000
events every second
flight


BATCH LAYER
master dataset
SERVING LAYER
batch view
batch view
query
query
new data
REALTIME LAYER
real-time view
real-time view

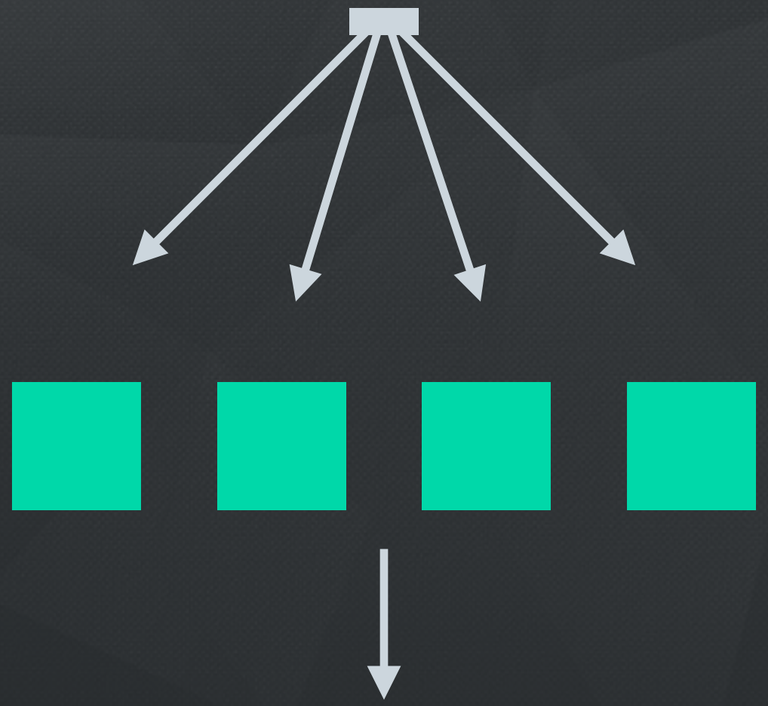

BATCH LAYER
SERVING LAYER
batch view
master dataset
REALTIME LAYER
real-time view
real-time view

# Scaling the Speed Layer

IN REAL-TIME

# Ideal: Linear Scaling In # Of Events

**1,000,000 Events**

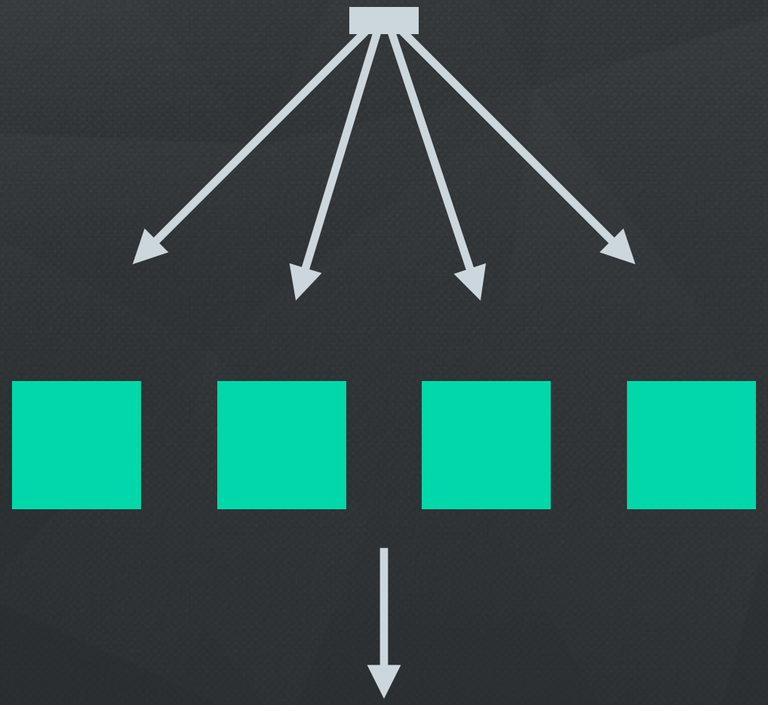


flight

# Ideal: Linear Scaling In # Of Events

**1,000,000 Events**

**ios.cannonball, foreground, device-one**

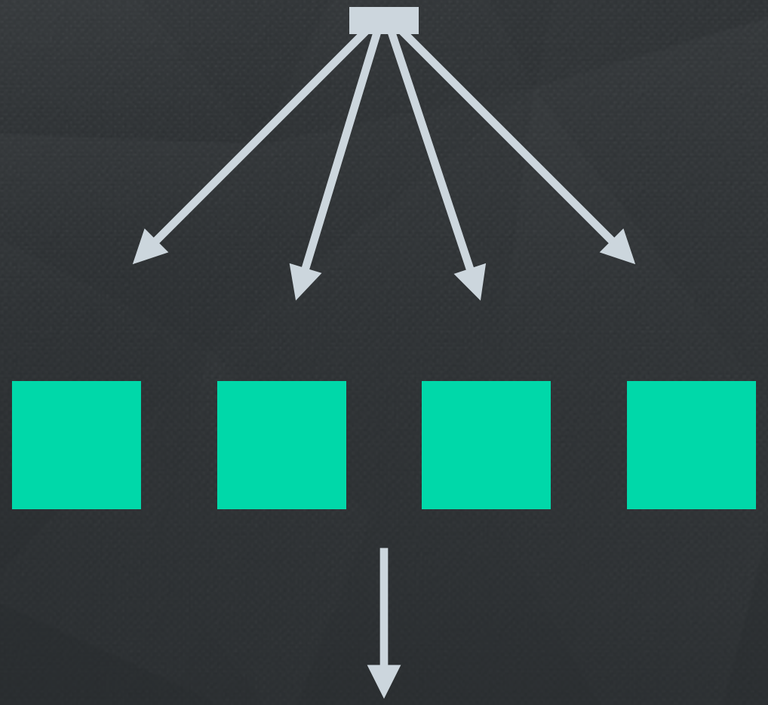


flight

# Ideal: Linear Scaling In # Of Events

**1,000,000 Events**

**com.twitter, install, device-two**

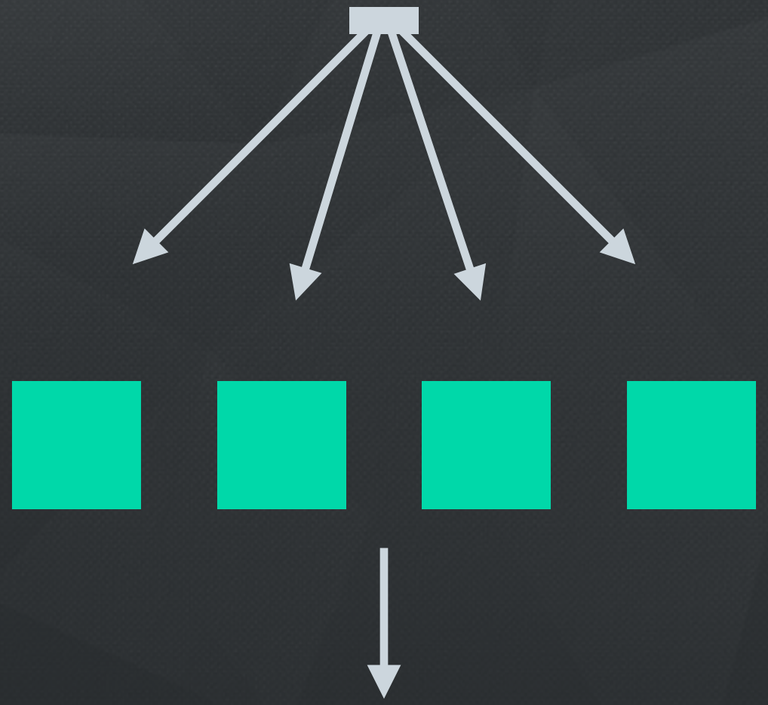


flight

# Ideal: Linear Scaling In # Of Events

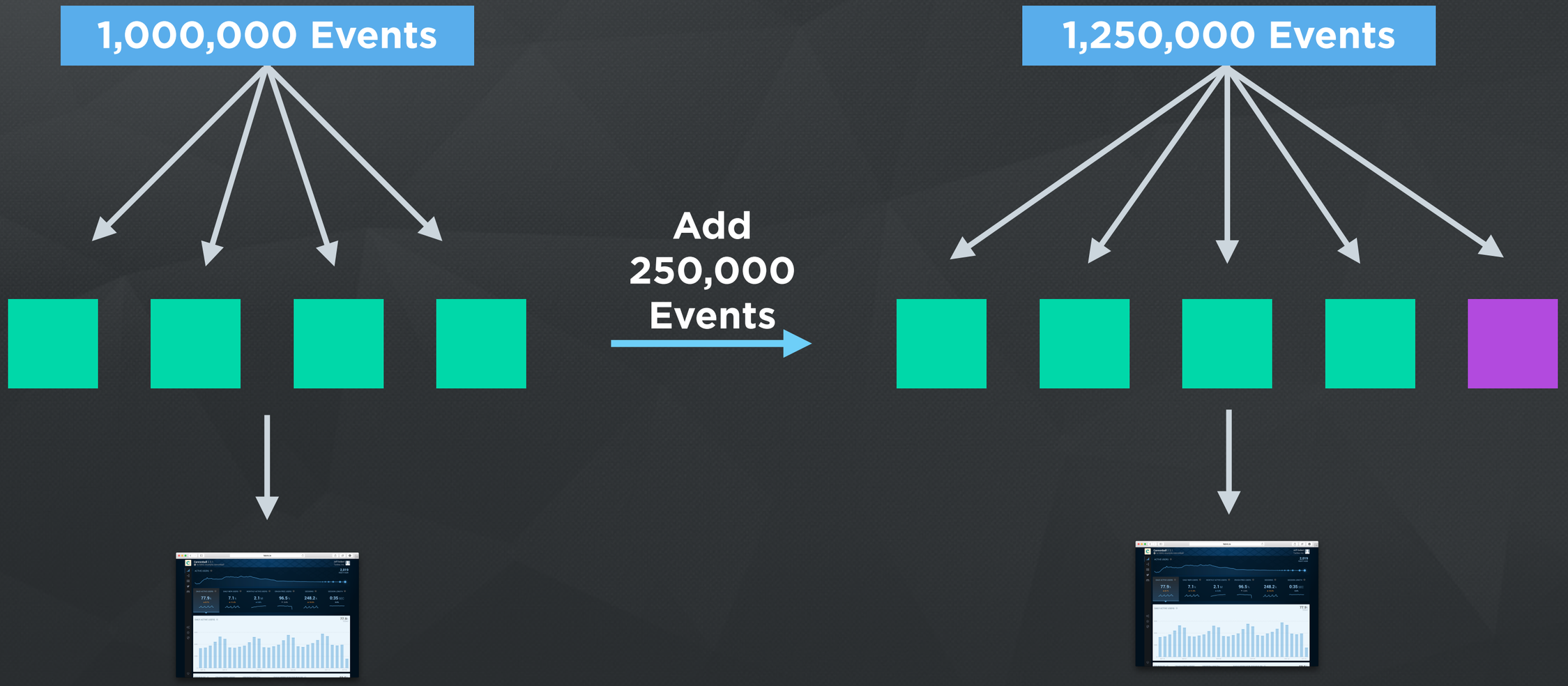
**1,000,000 Events**

**ios.cannonball, background, device-three**



flight

# Ideal: Linear Scaling In # Of Events



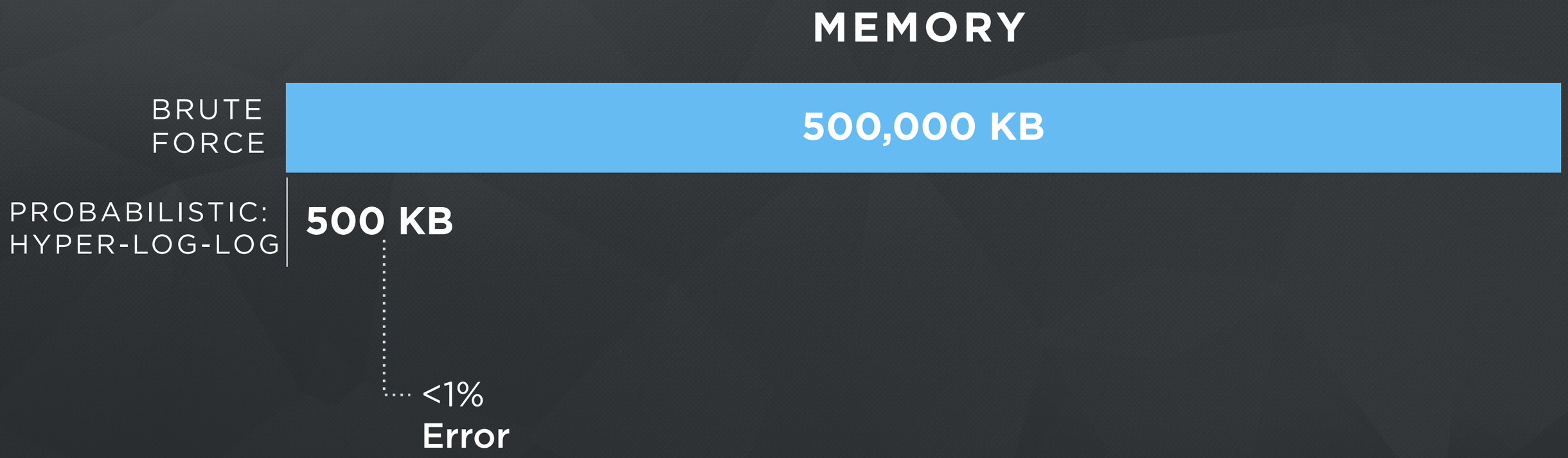

flight

# Complexity In Scaling

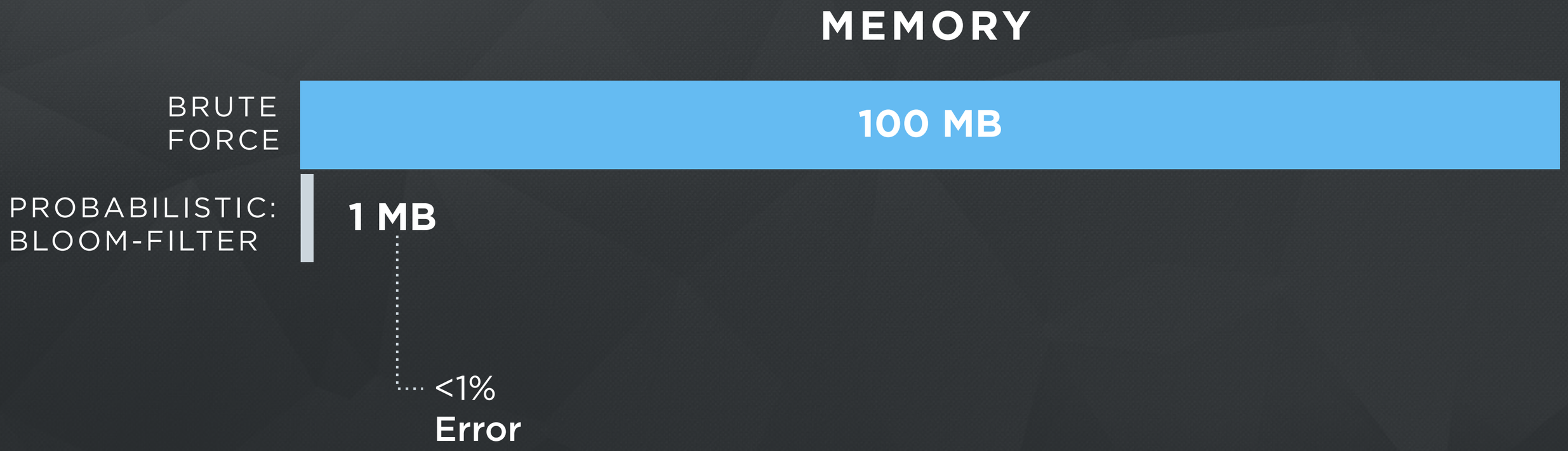
- **Daily Active Users - Cardinality** - Need a set that holds millions of members
- **Retention - Set Membership** - Need multiple sets that hold thousands of members
- **Session Duration - Median of Millions of Data Points** - Need sorted list of millions of data points

# Probabilistic Data Structures

# Daily Active Users - Cardinality

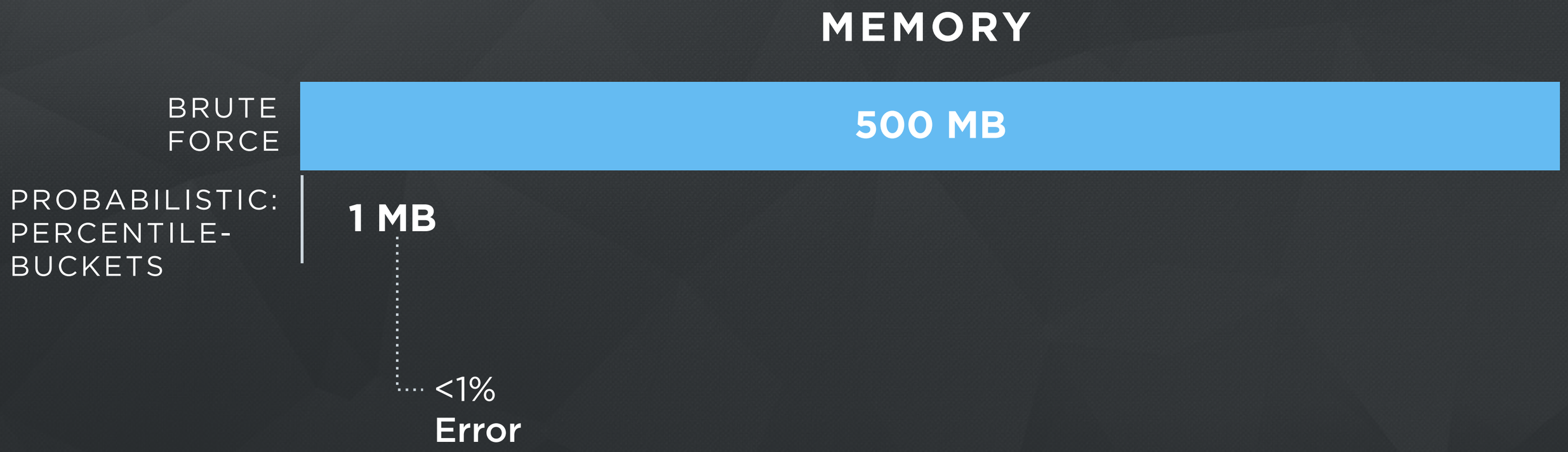


# Retention - Set Membership




flight

# Session Duration - Median Of Millions Of Data Points




flight

# Session Duration - Median Of Millions Of Data Points

1.321 s, 2.134 s,  2.231 s, 3.345 s, 5.011 s, 5.451 s, 6.019 s, 7.778 s, 8.345 s, 9.123 s, 12.125 s, 30.124 s ...

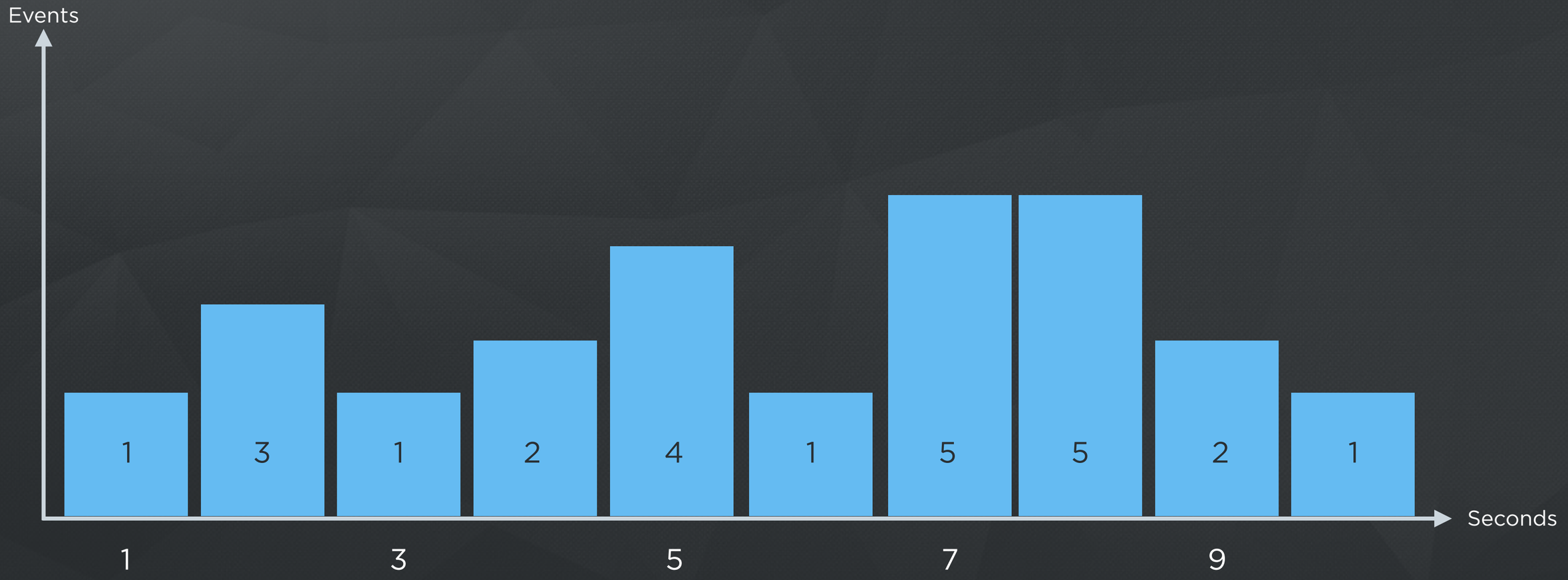
# Session Duration - Median Of Millions Of Data Points

1 s, 2 s, 2 s, 3 s, 5 s, 5 s, 6 s, 7 s, 8 s, 9 s, 12 s, 30 s

# Session Duration - Median Of Millions Of Data Points

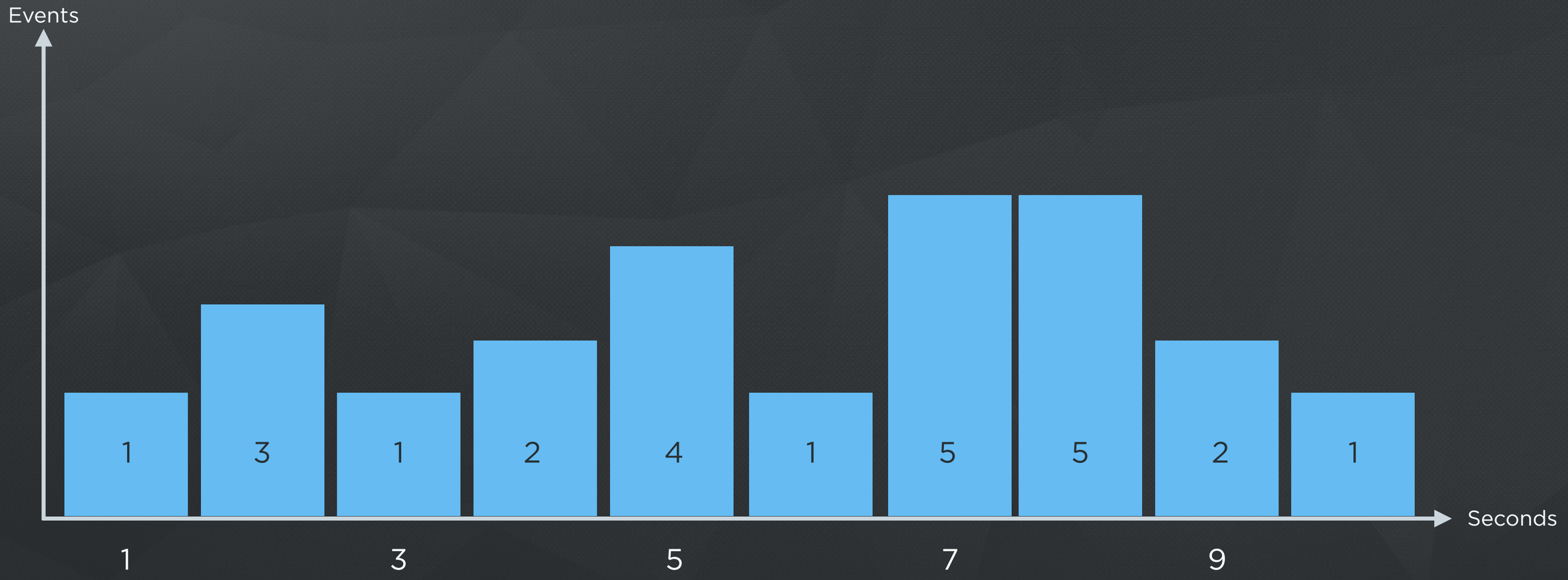

flight

# For Example : Have Seen 25 Sessions



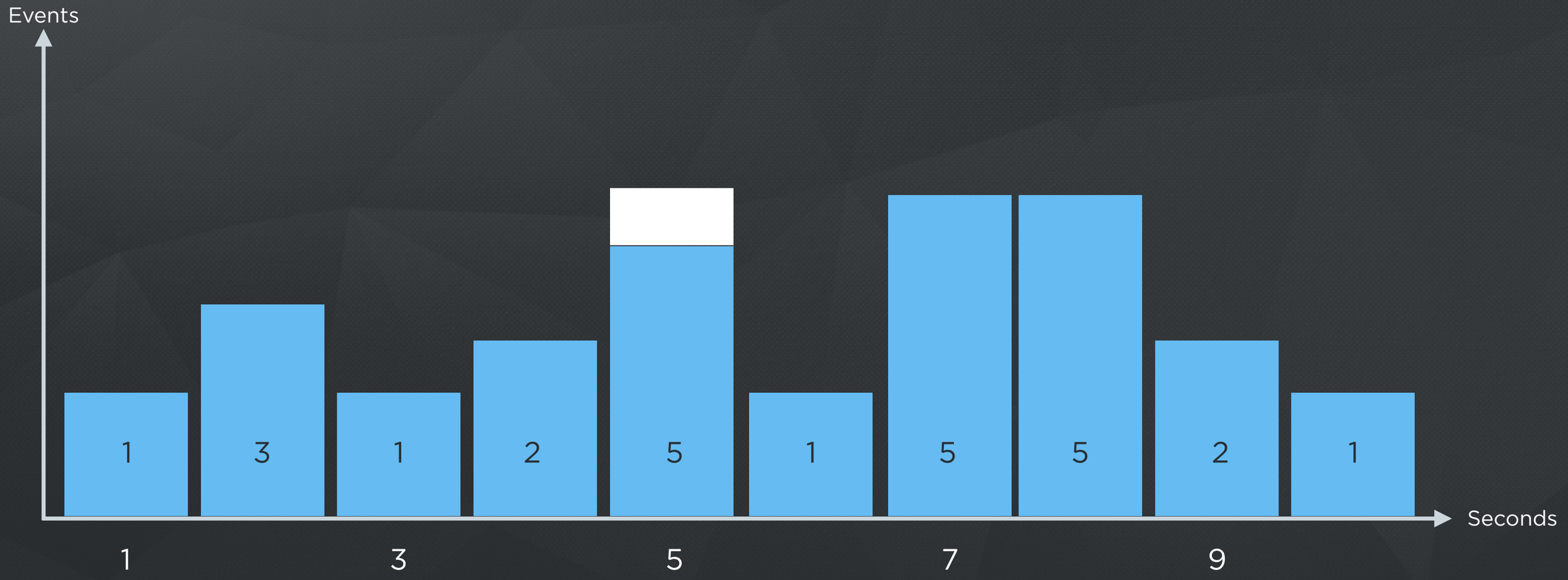

flight

# NEW SESSION : 5 seconds



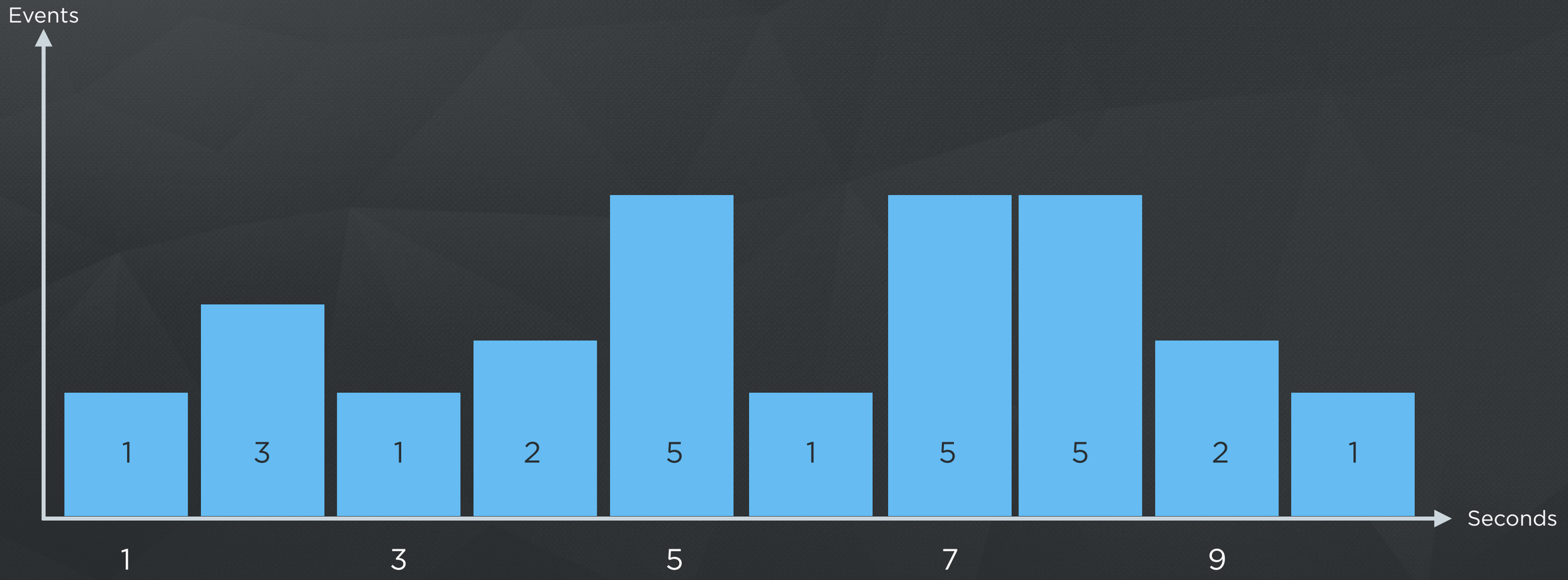

flight

# NEW SESSION : 5 seconds



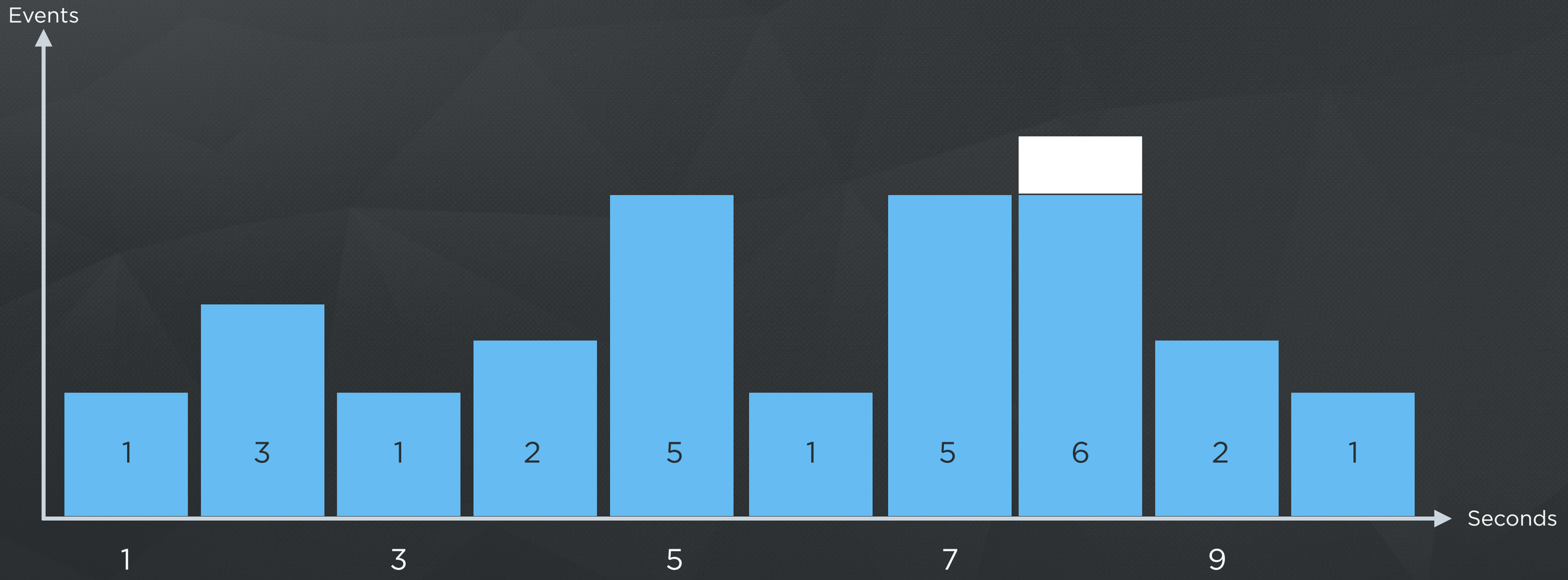

flight

# NEW SESSION : 8 seconds



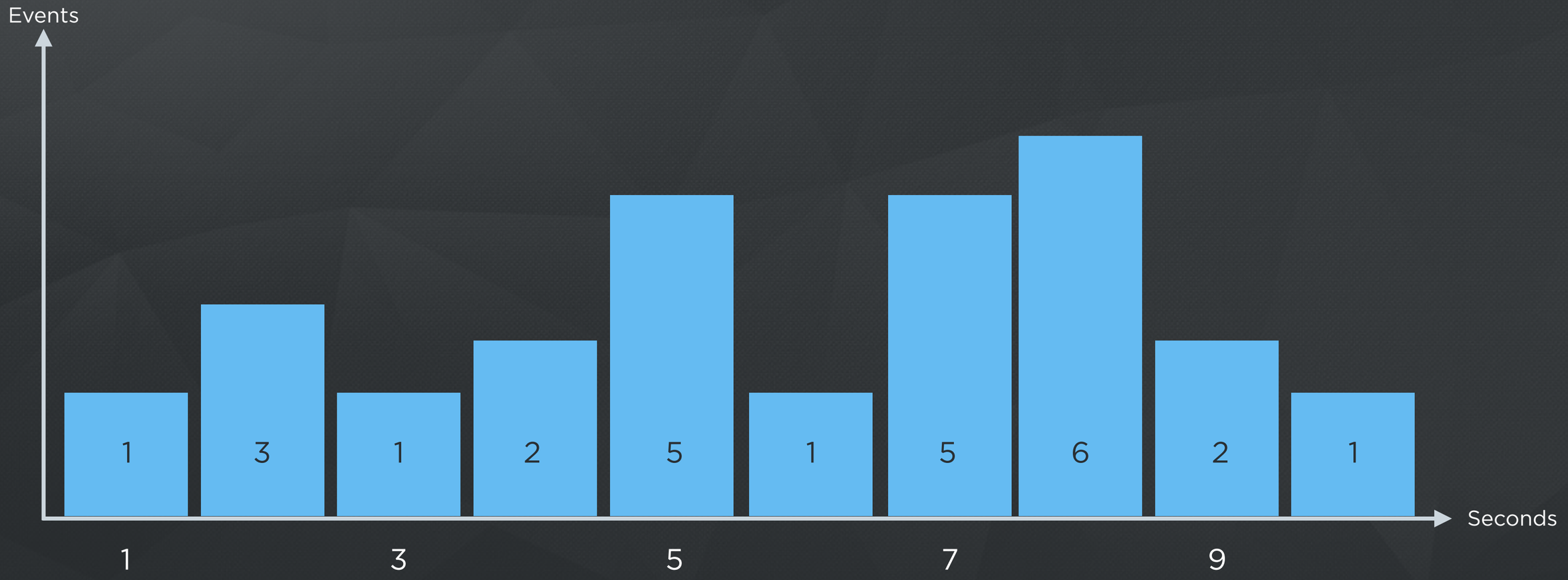

flight

# NEW SESSION : 8 seconds



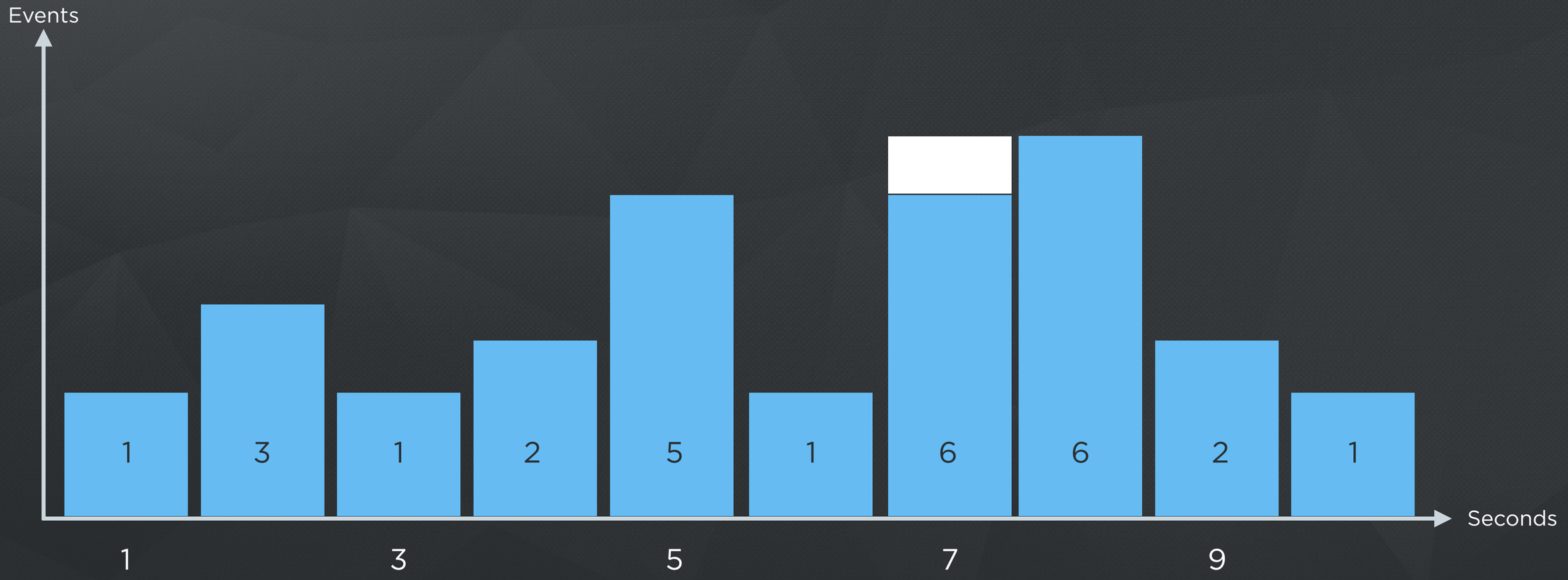

flight

# NEW SESSION : 7 seconds



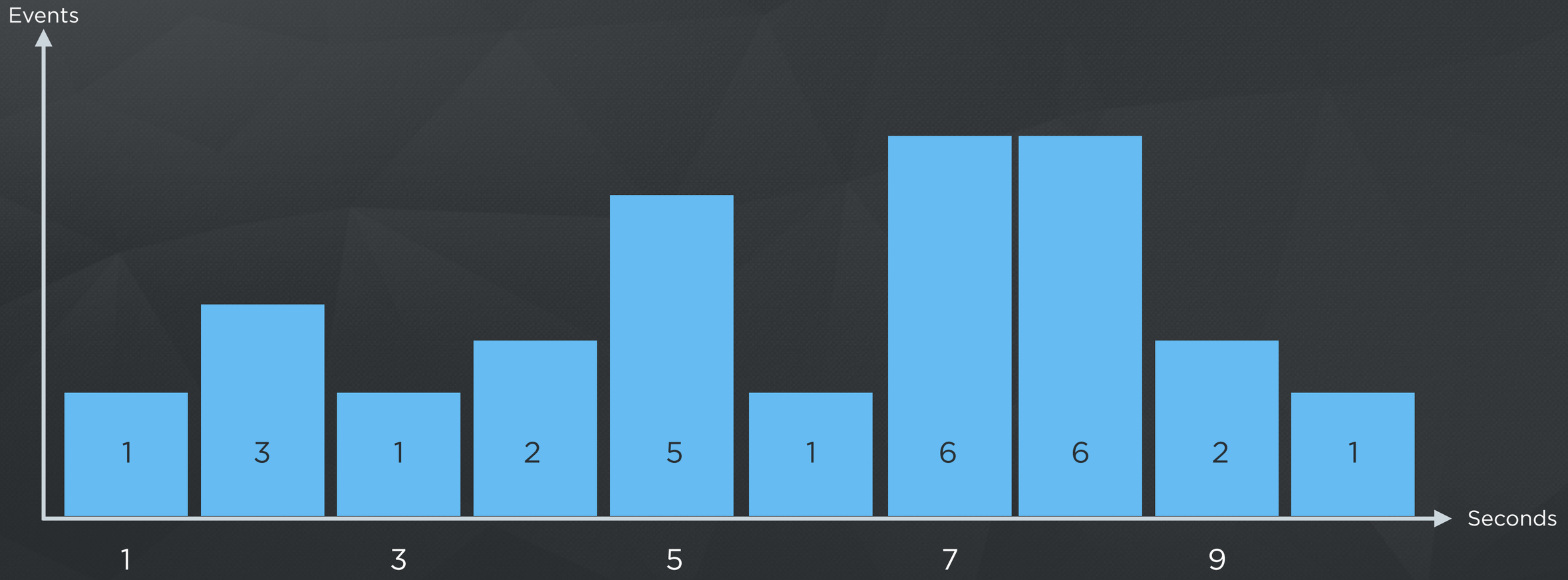

flight

# NEW SESSION : 7 seconds



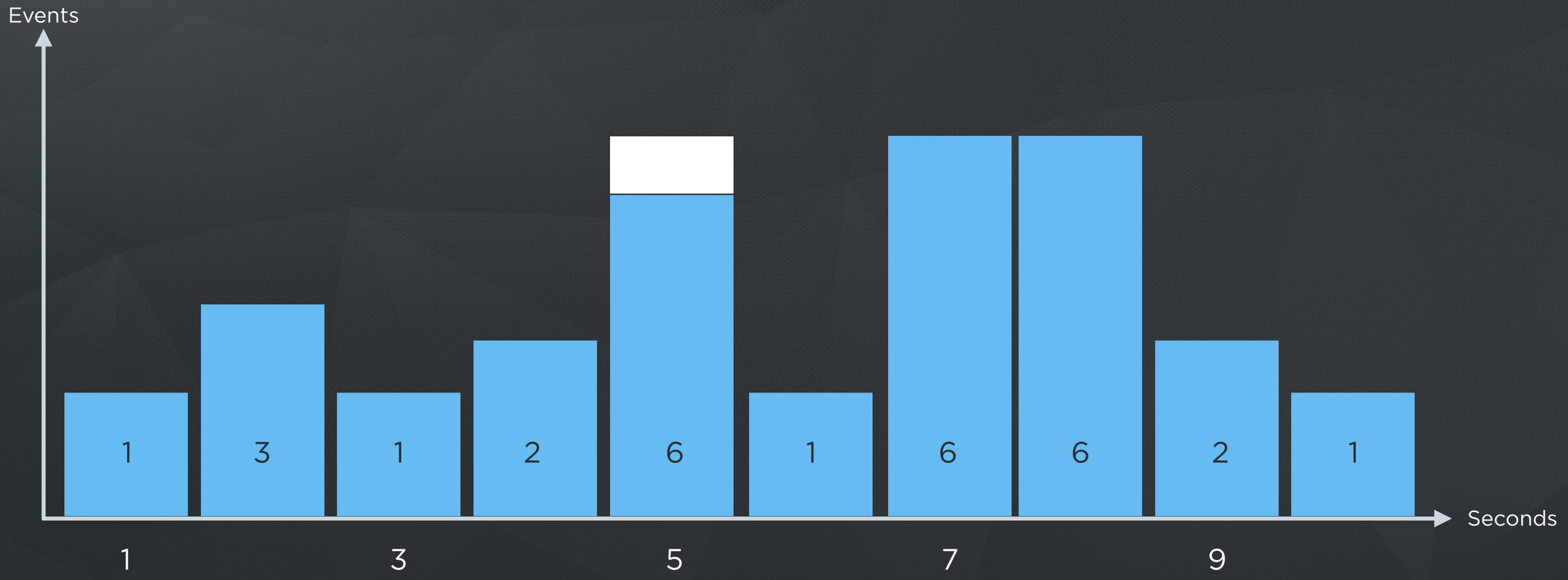

flight

# NEW SESSION : 5 seconds



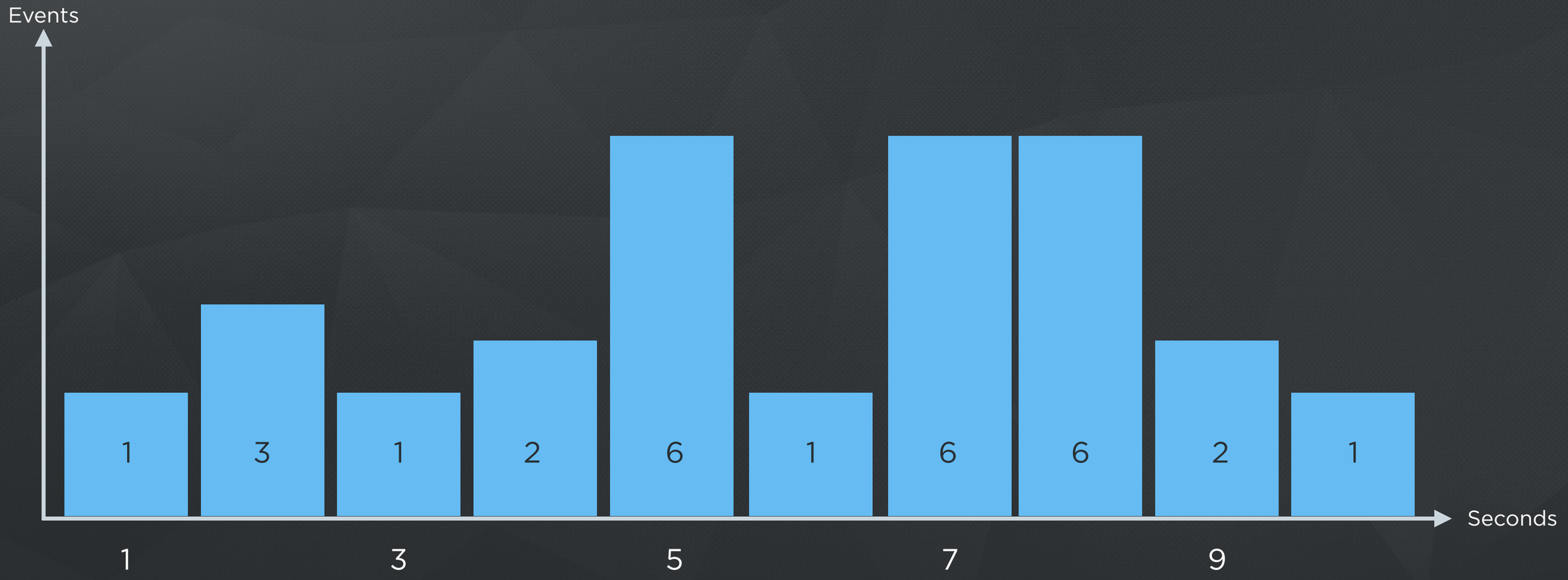

flight

# NEW SESSION : 5 seconds



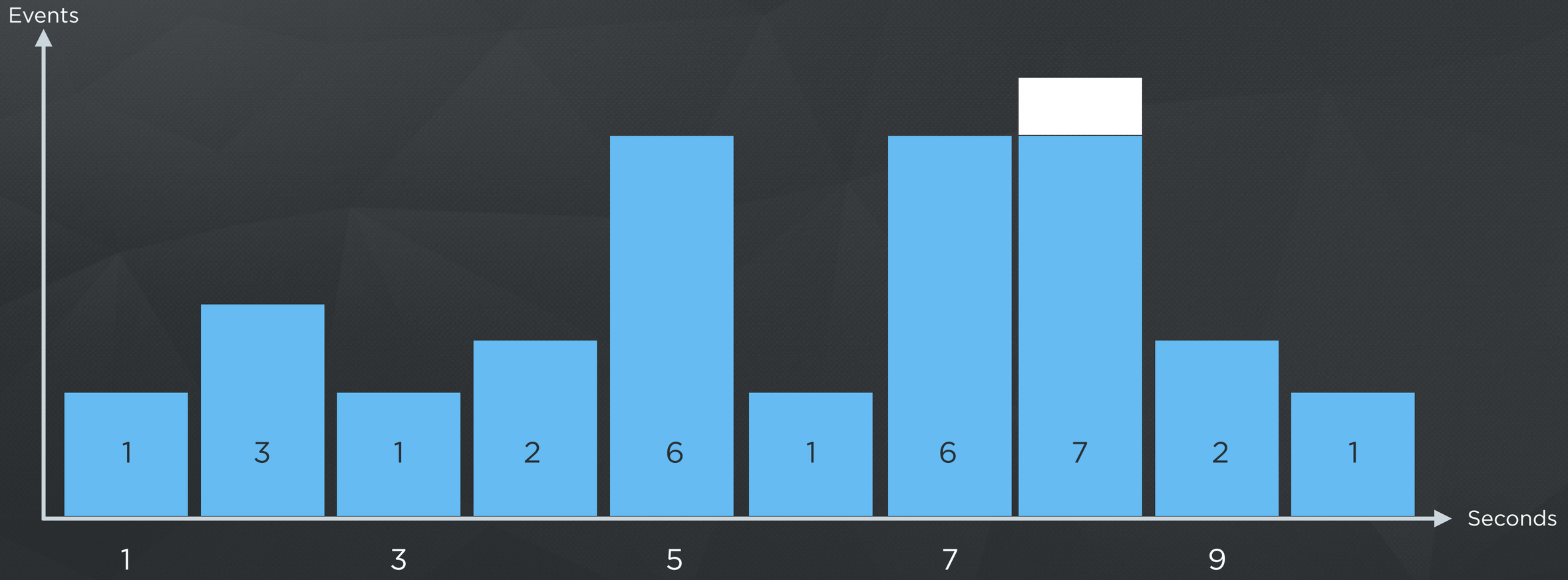

flight

# NEW SESSION : 8 seconds



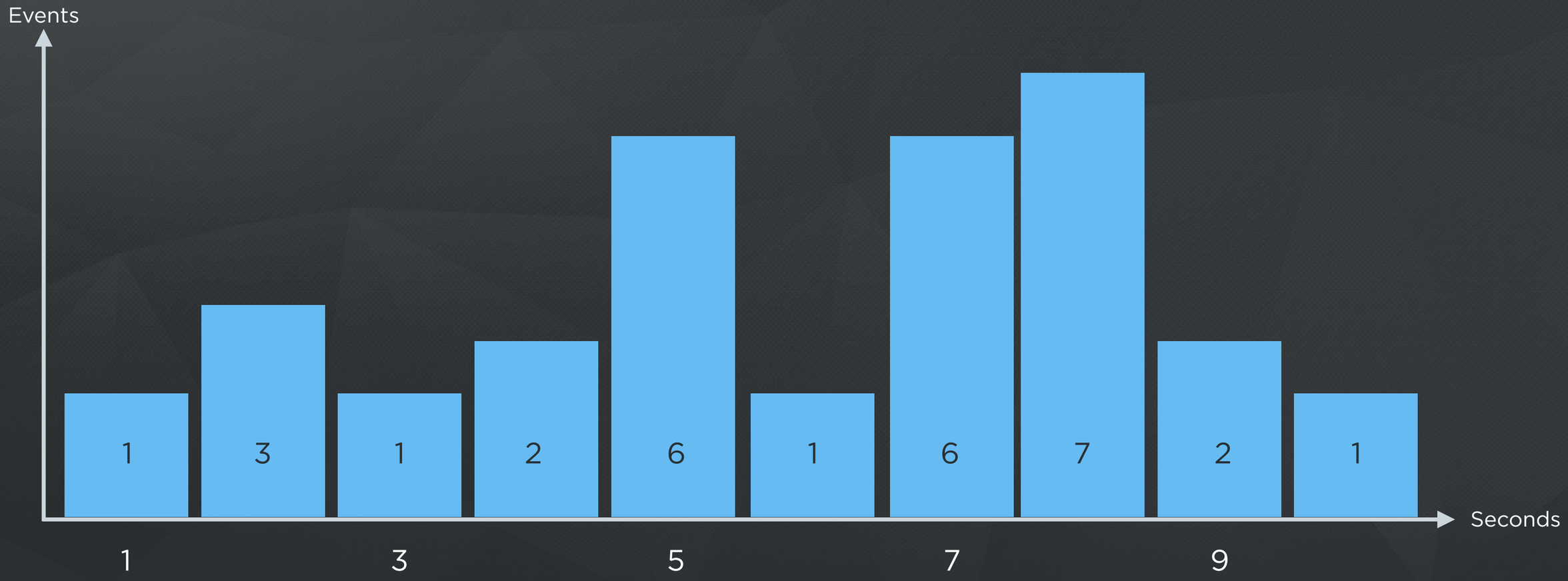

flight

# NEW SESSION : 8 seconds



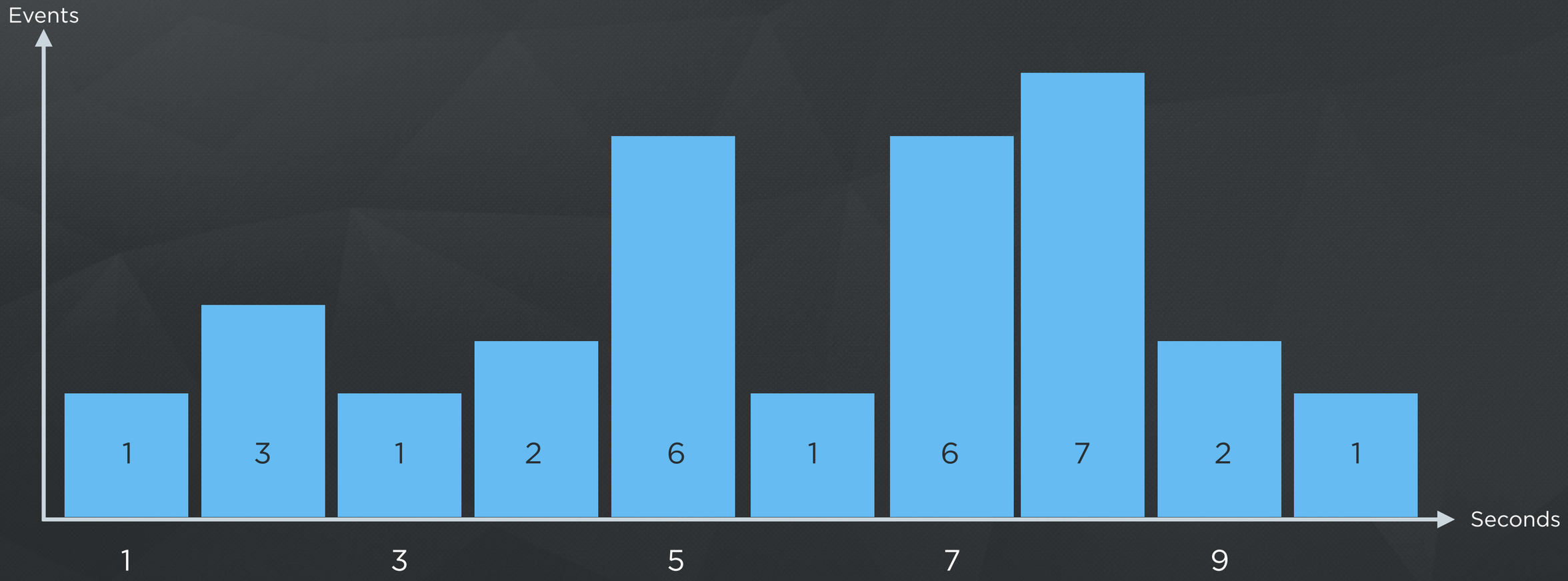

flight

# For Example : Have Seen 30 Sessions



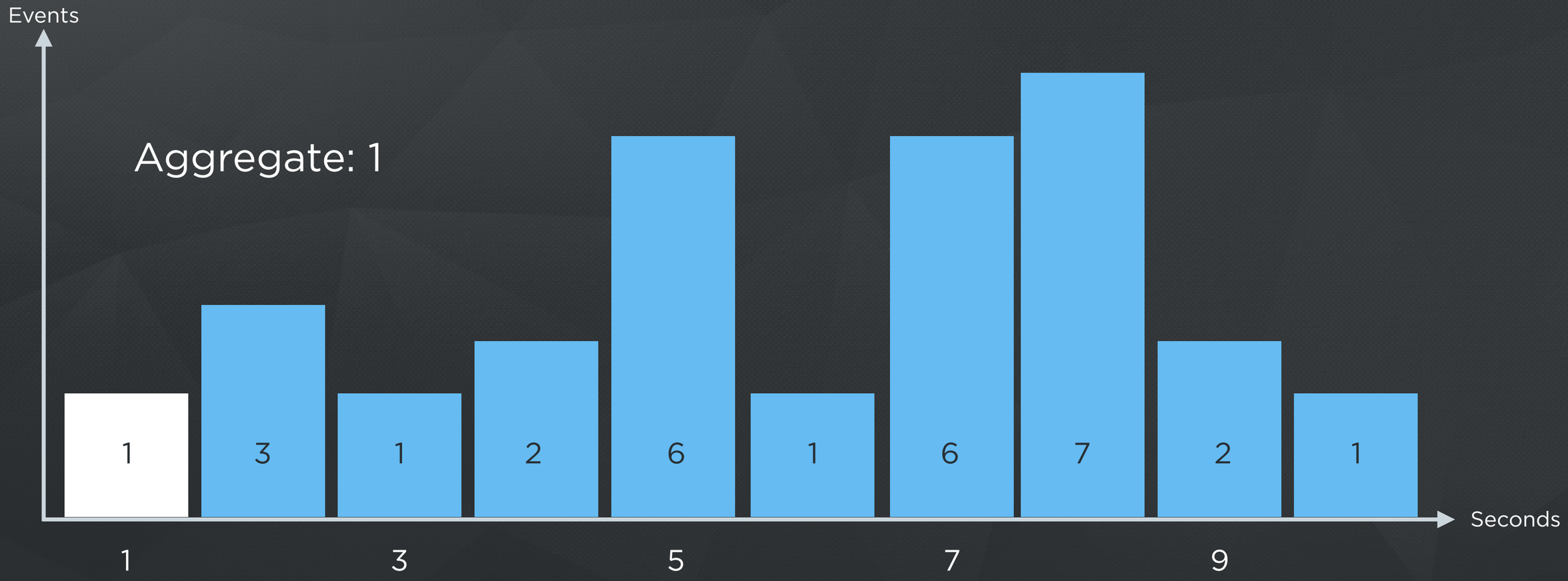

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



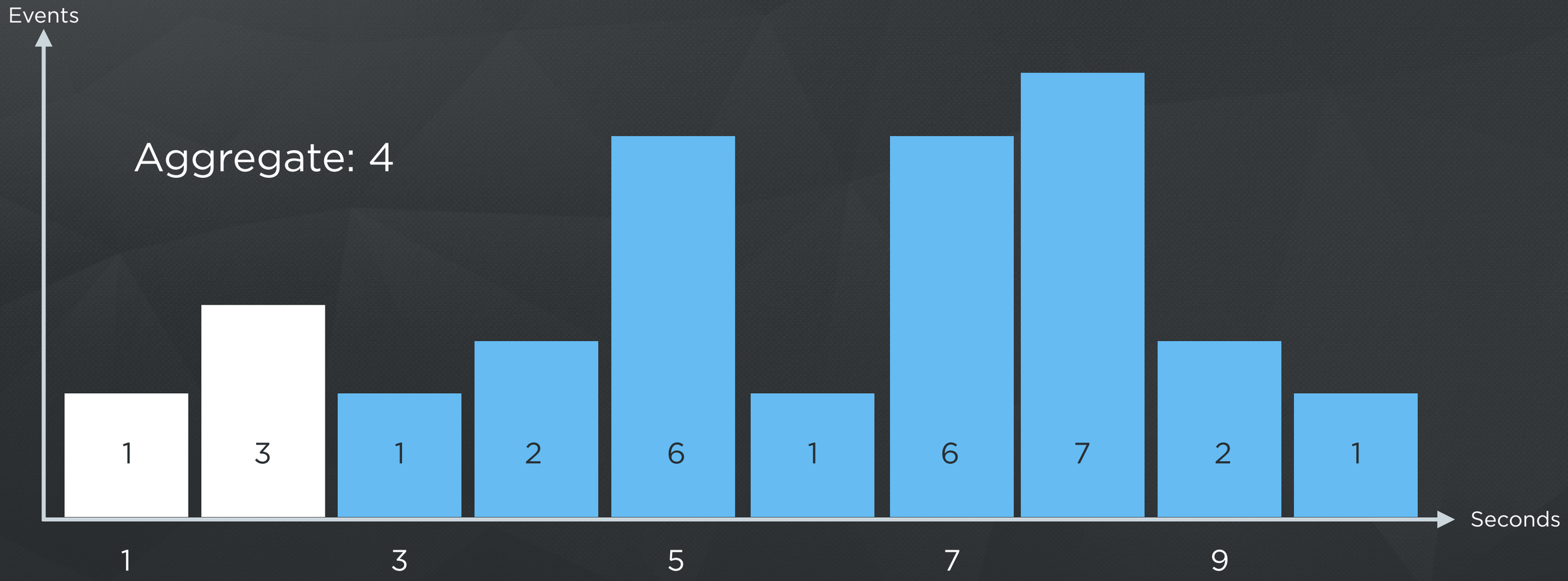

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



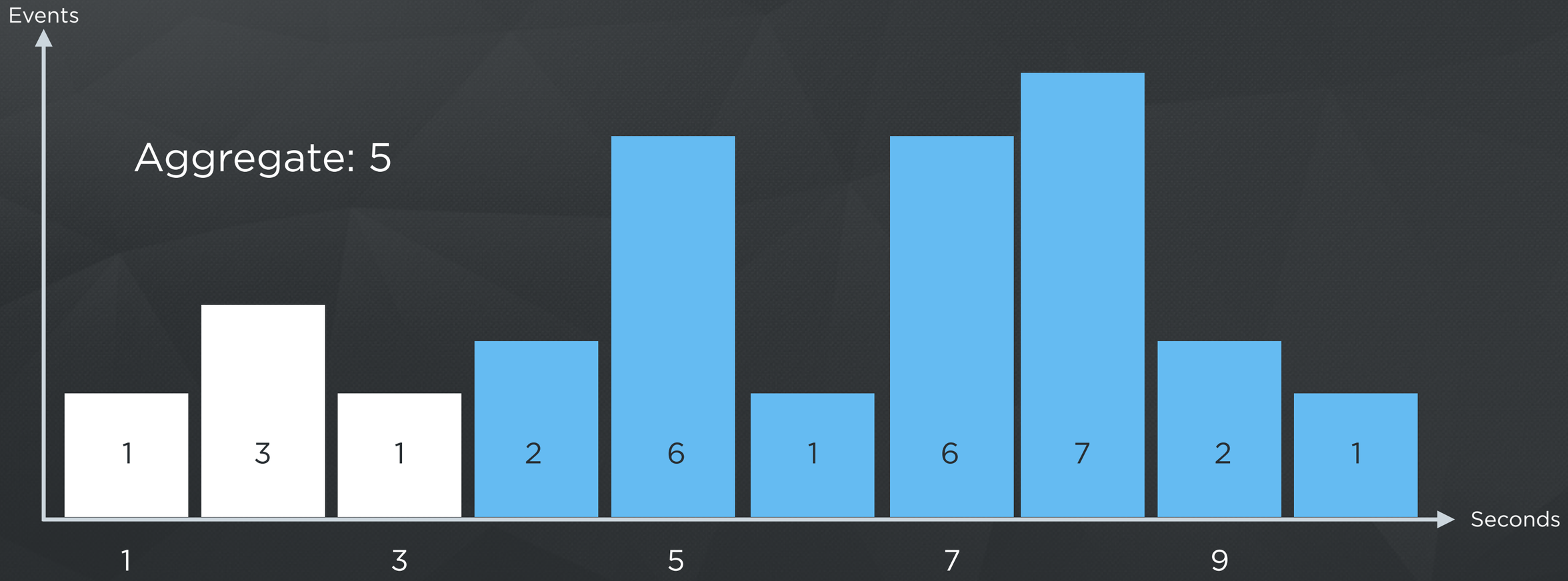

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



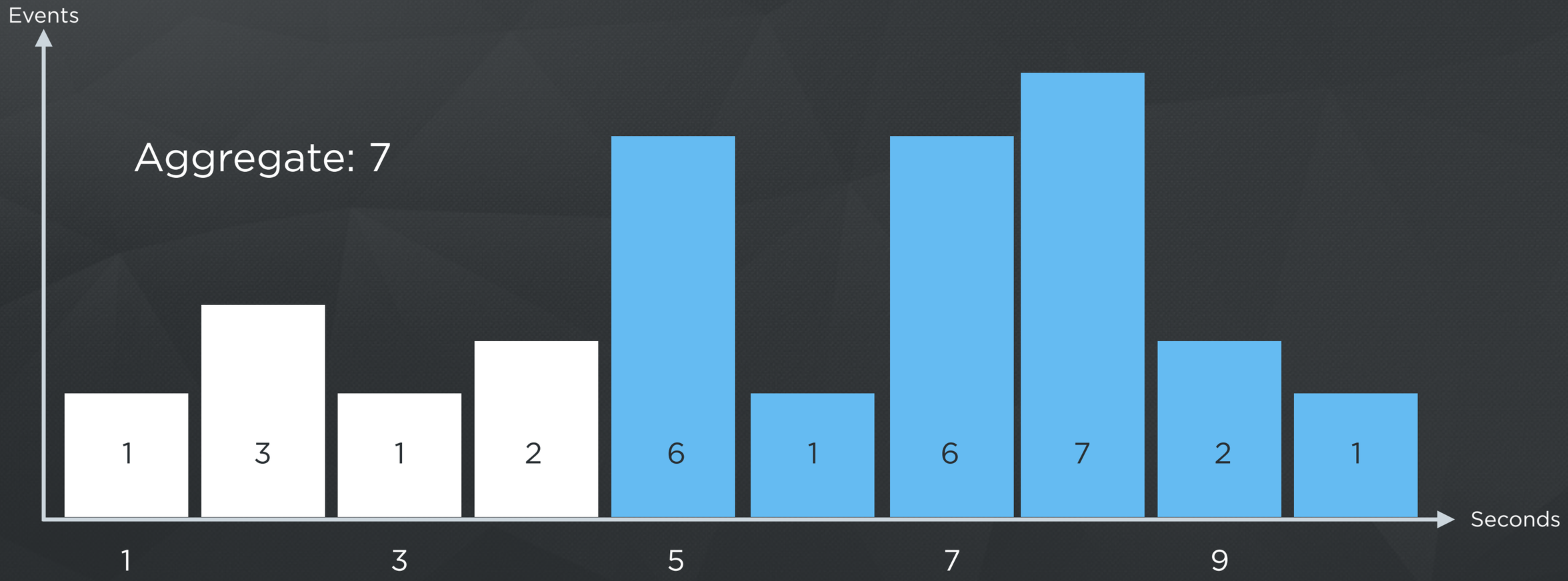

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



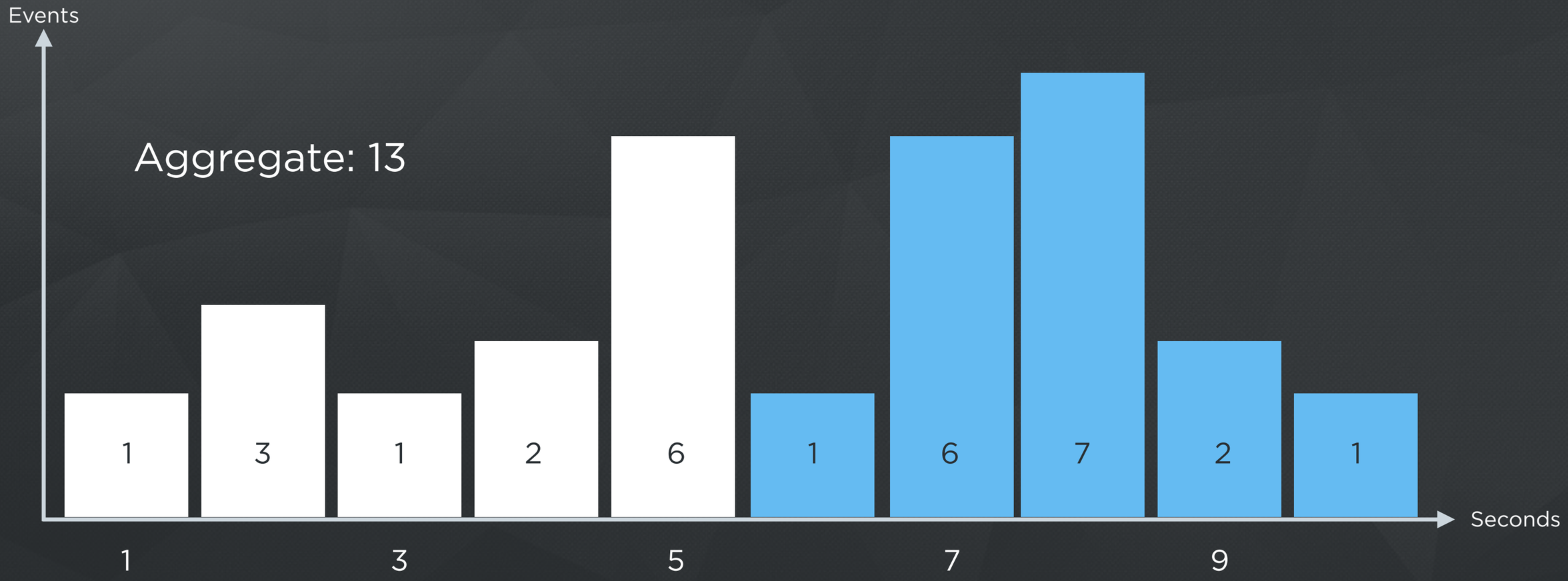

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



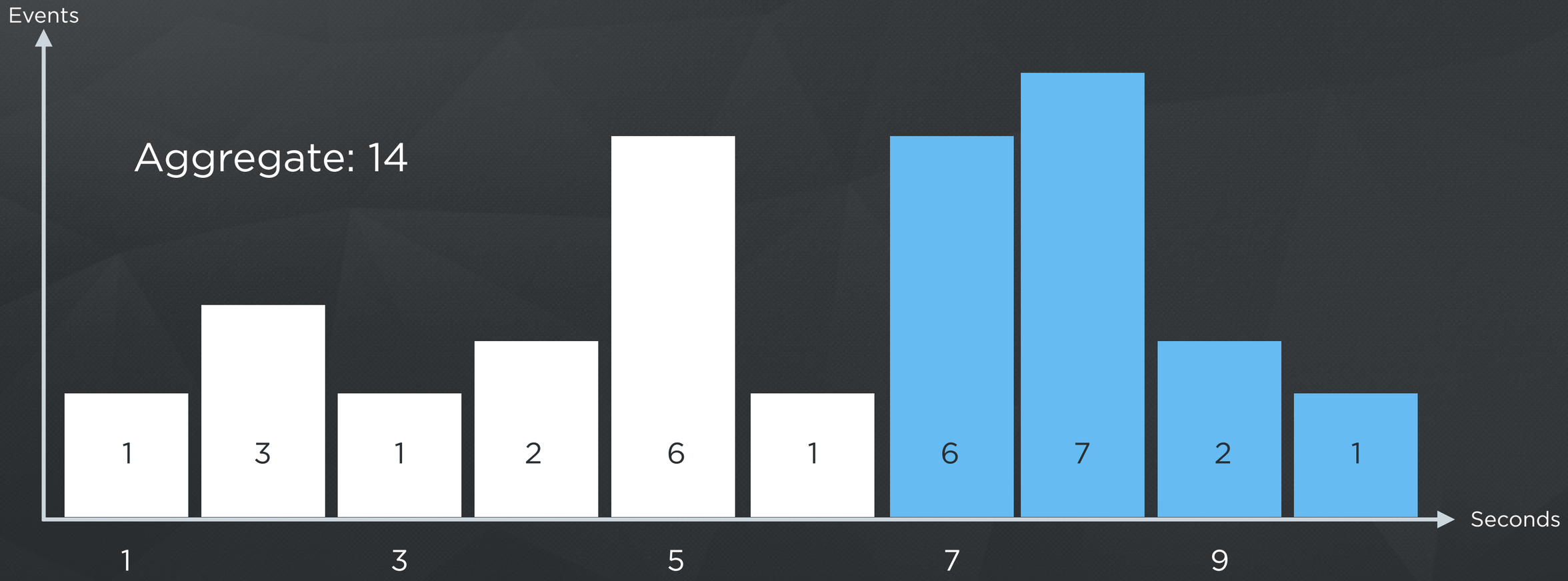

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?



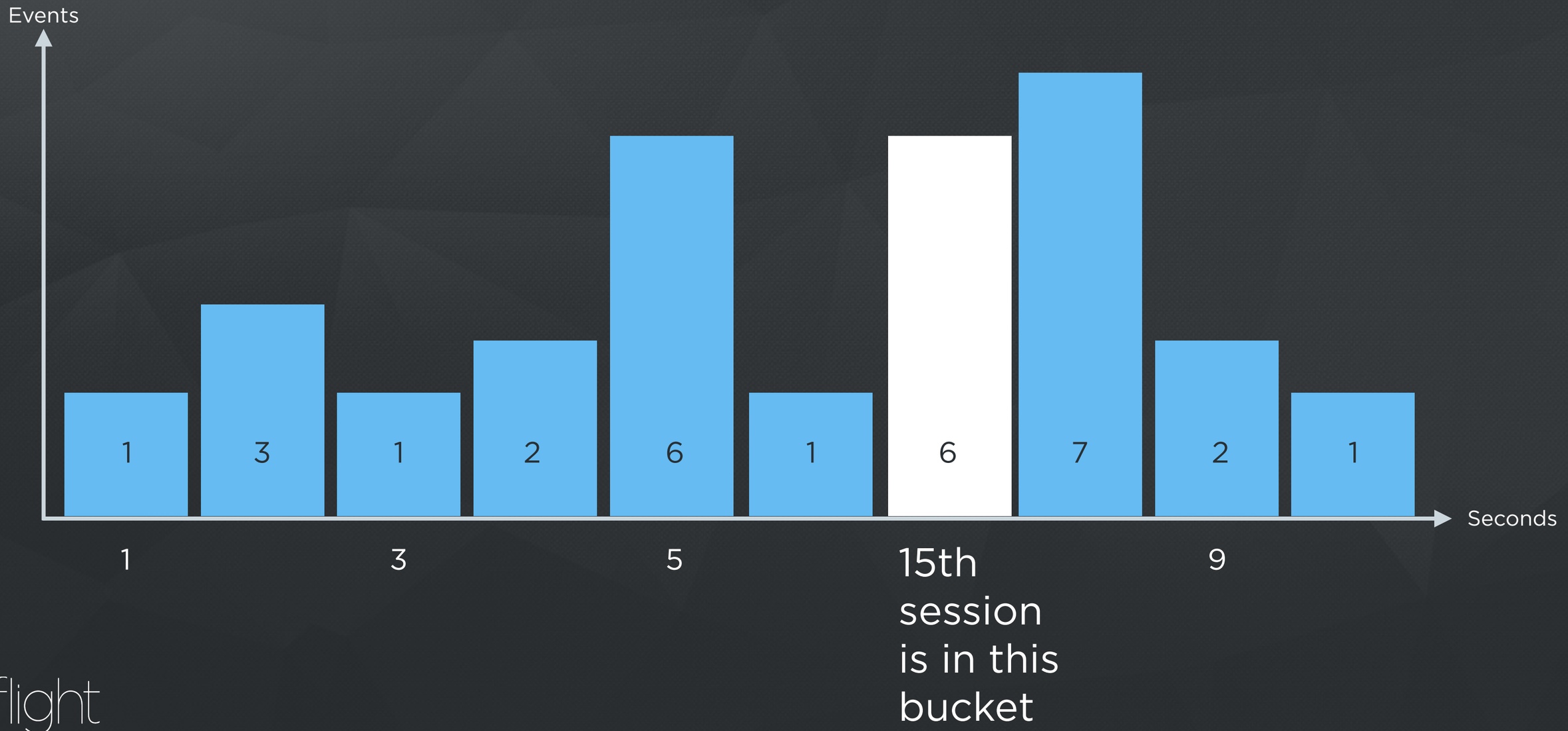

flight

# 30 Sessions : Median Is Fifteenth, How Do We Find It?




flight

# For Example : Have Seen 30 Sessions




flight


Most straightforward if all events for a particular app to go to a single machine so we can maintain one instance of each data structure per app...
flight

# Trade Off

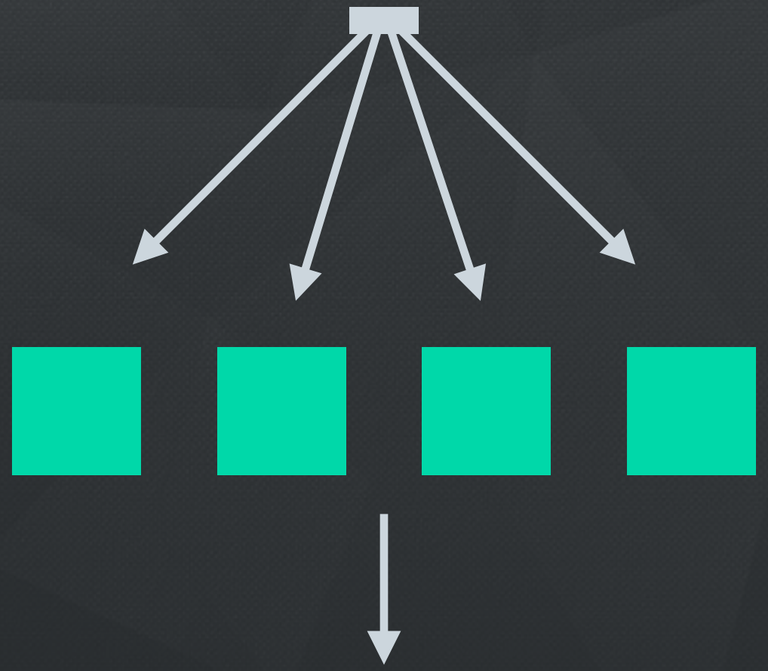
- Simplicity of one data structure per app with complexity of Hot Spotting, OR
- Simplicity of events going anywhere with complexity of combining multiple data structures for a single app, and consuming memory on every box for every app

# New Challenge

## HOT SPOTTING

# Hot Spotting

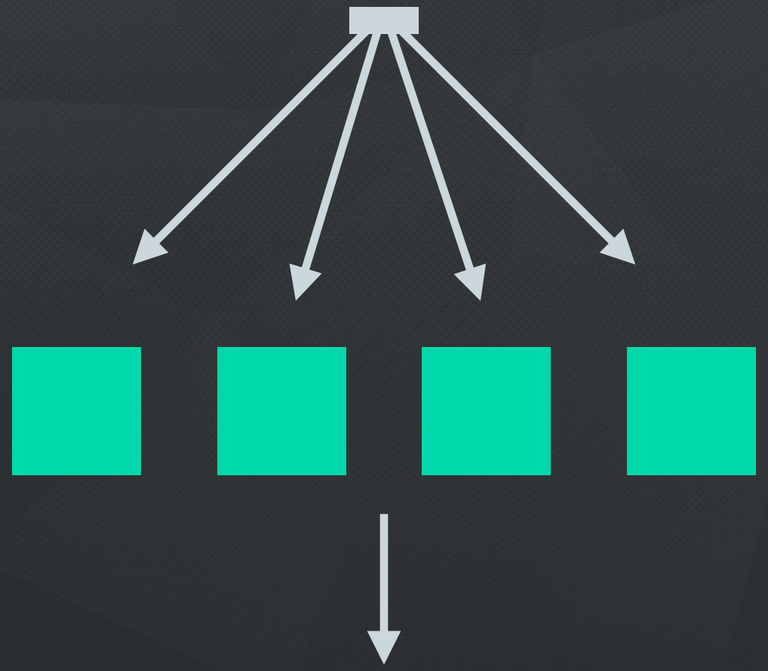
**ios.cannonball, foreground, device-one**



flight

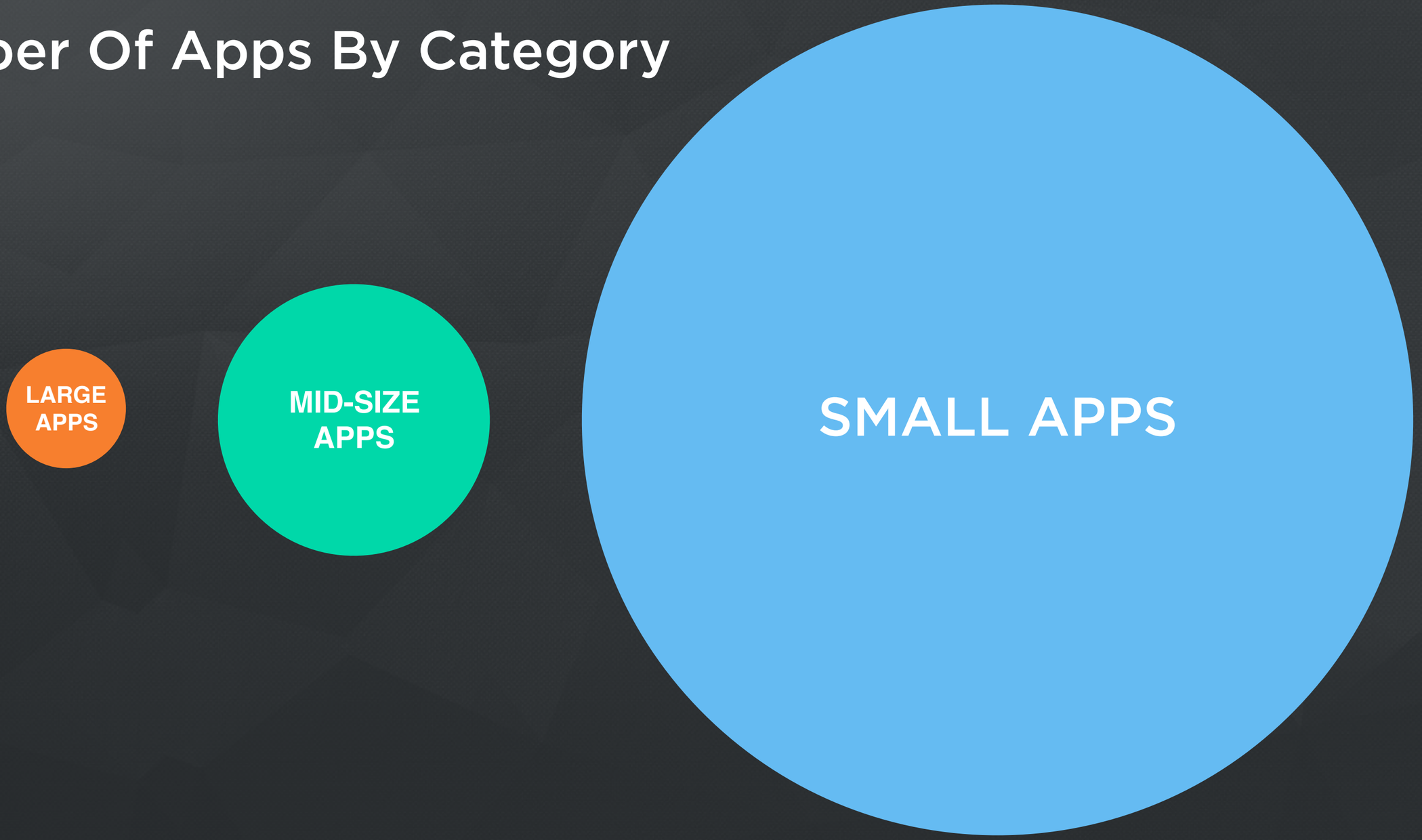
# Hot Spotting

**ios.cannonball, background, device-three**



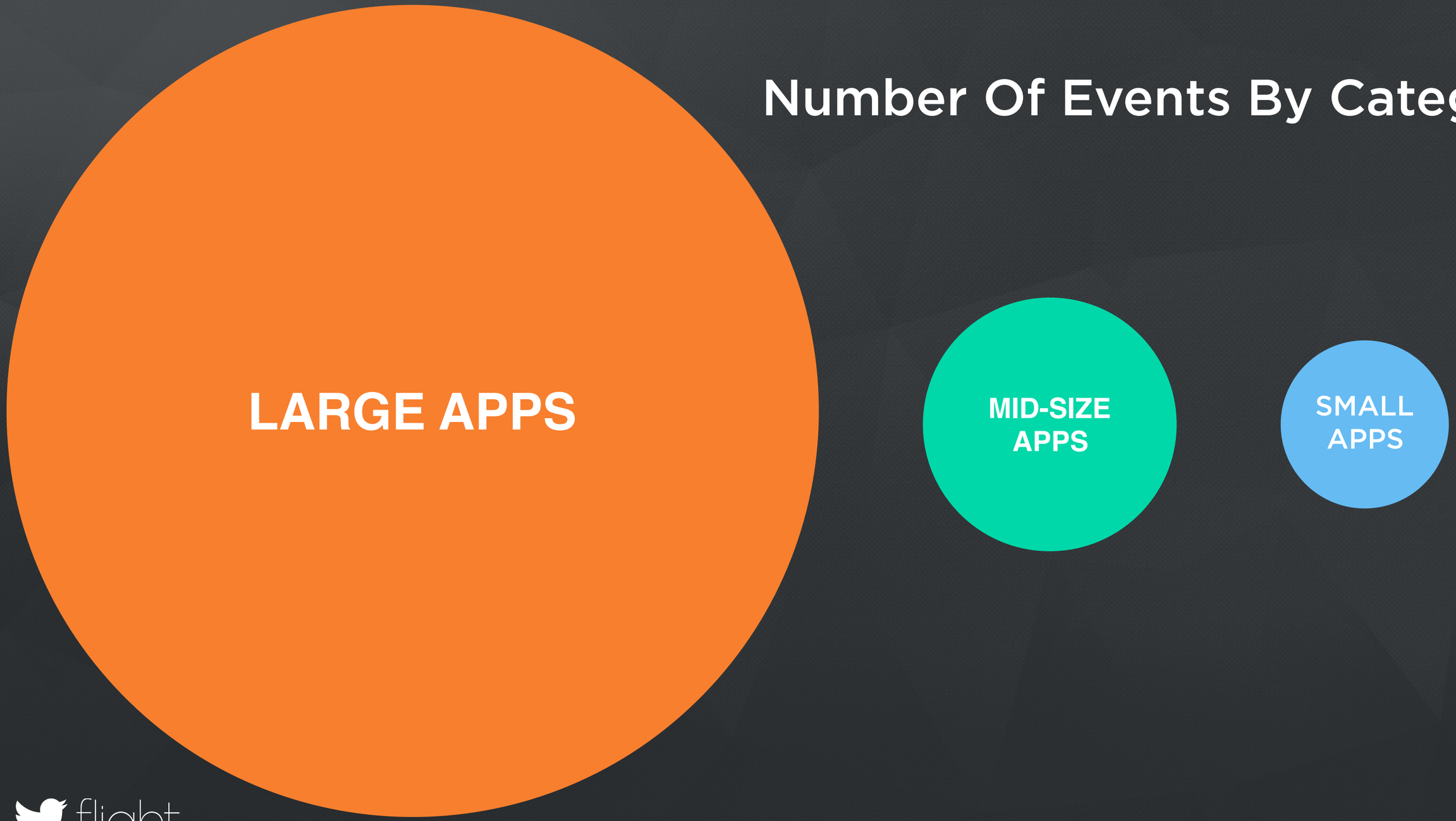
flight

# Number Of Apps By Category



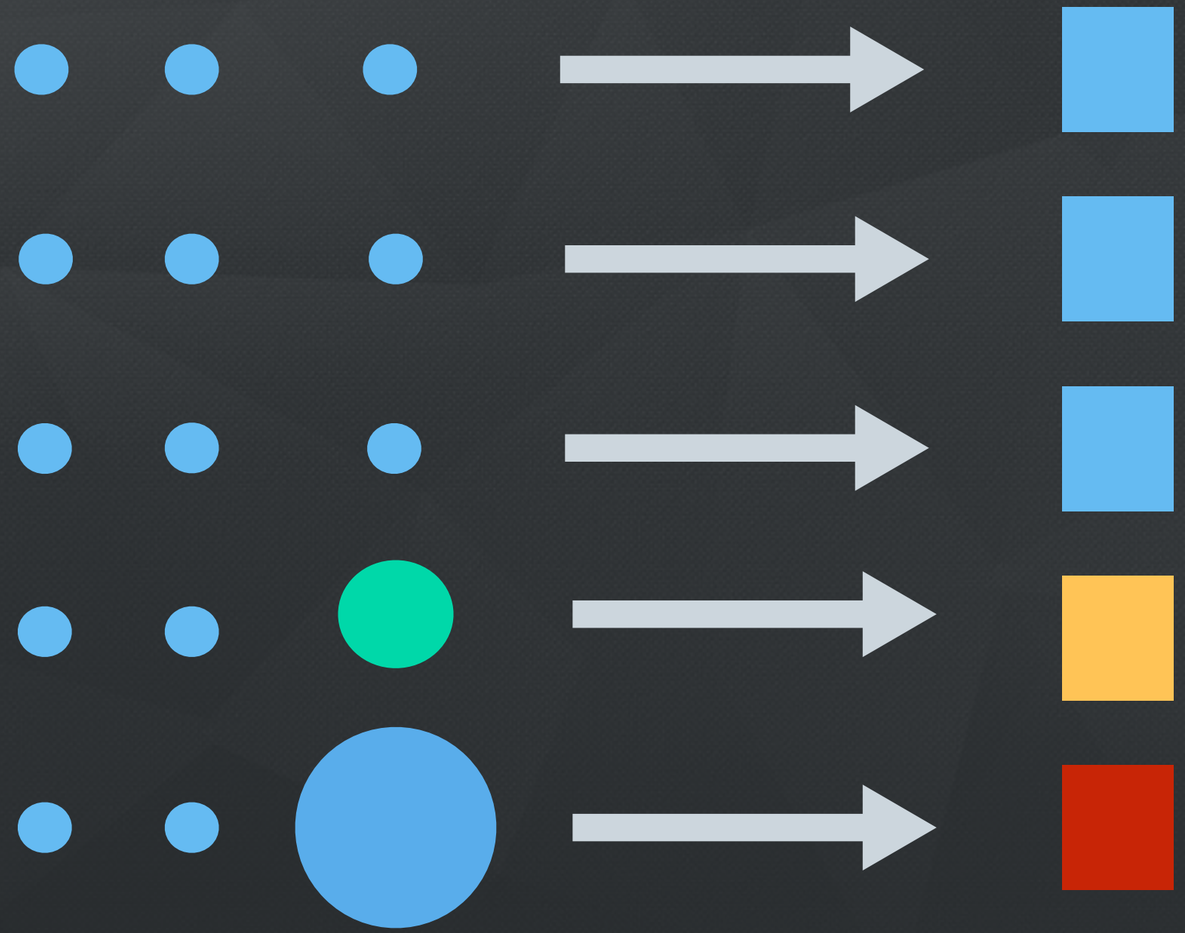

flight

# Number Of Events By Category




flight

# Assignment With Even Distribution Of Apps



flight

# What do we do about Hotspotting?

# Two Approaches

1. Stream-Sized Grouping

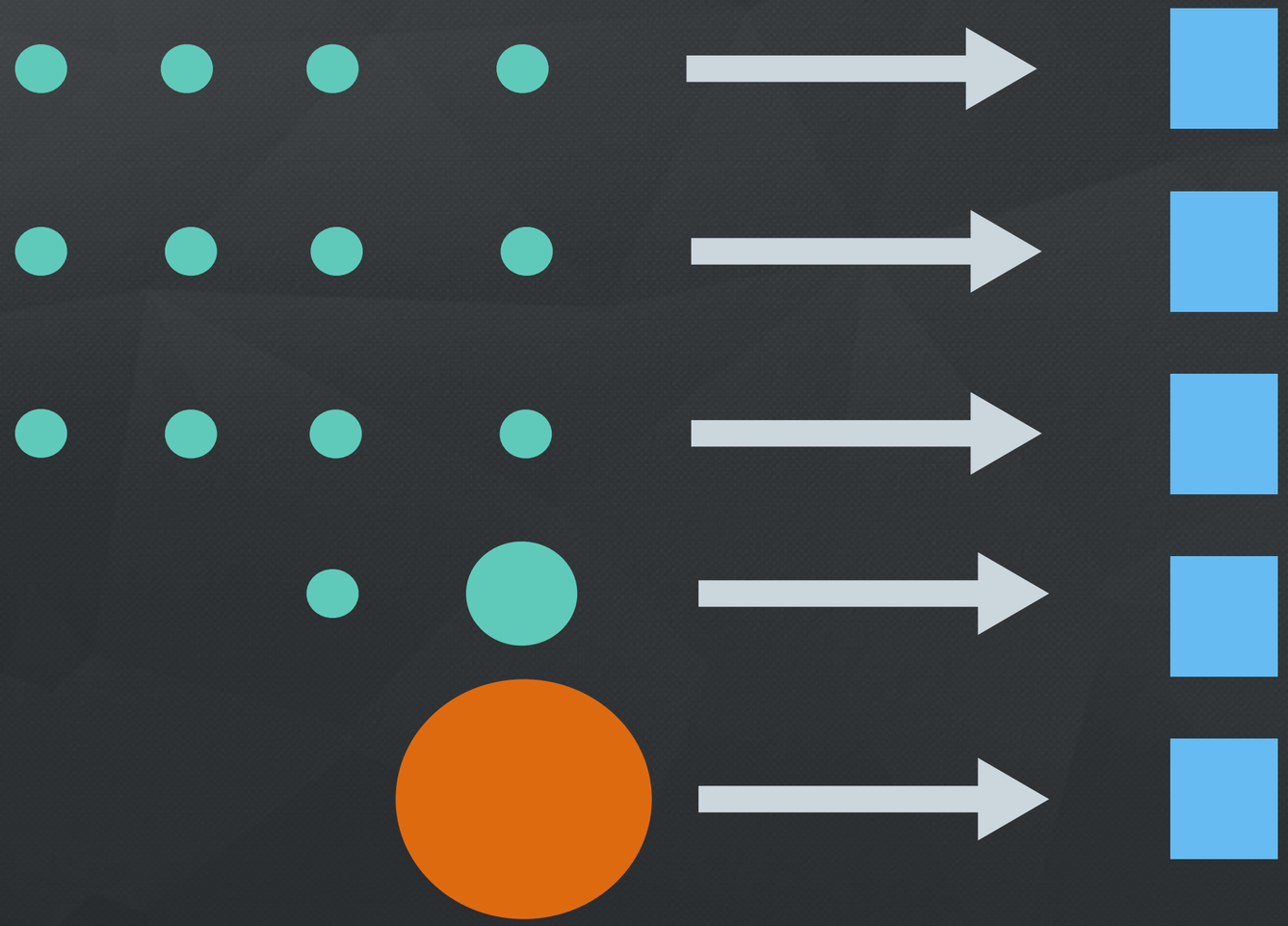
2. Sampling

# Stream-Sized Grouping

# Move Towards Assignment With Even Distribution Of Events

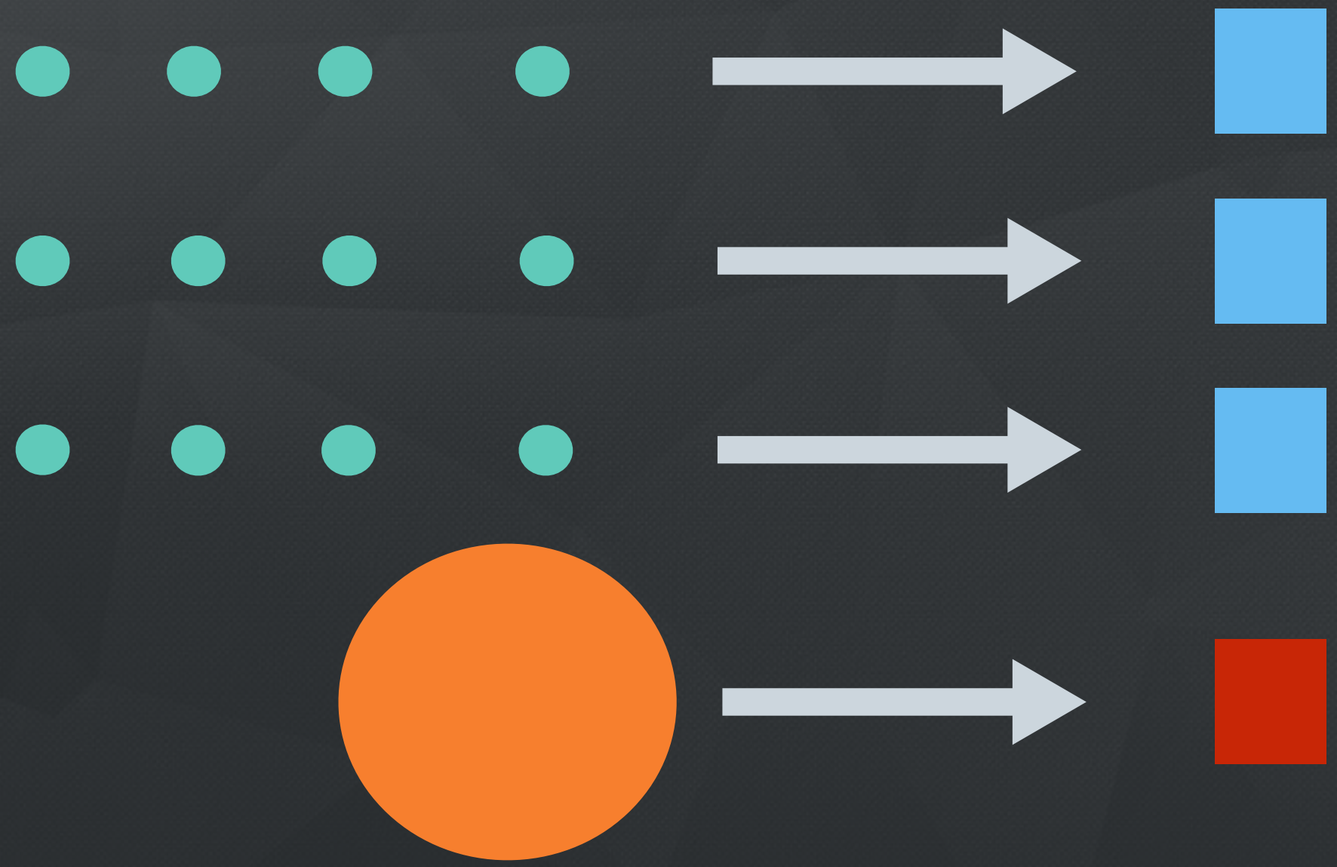
flight

# Assignment With Even Distribution Of Events



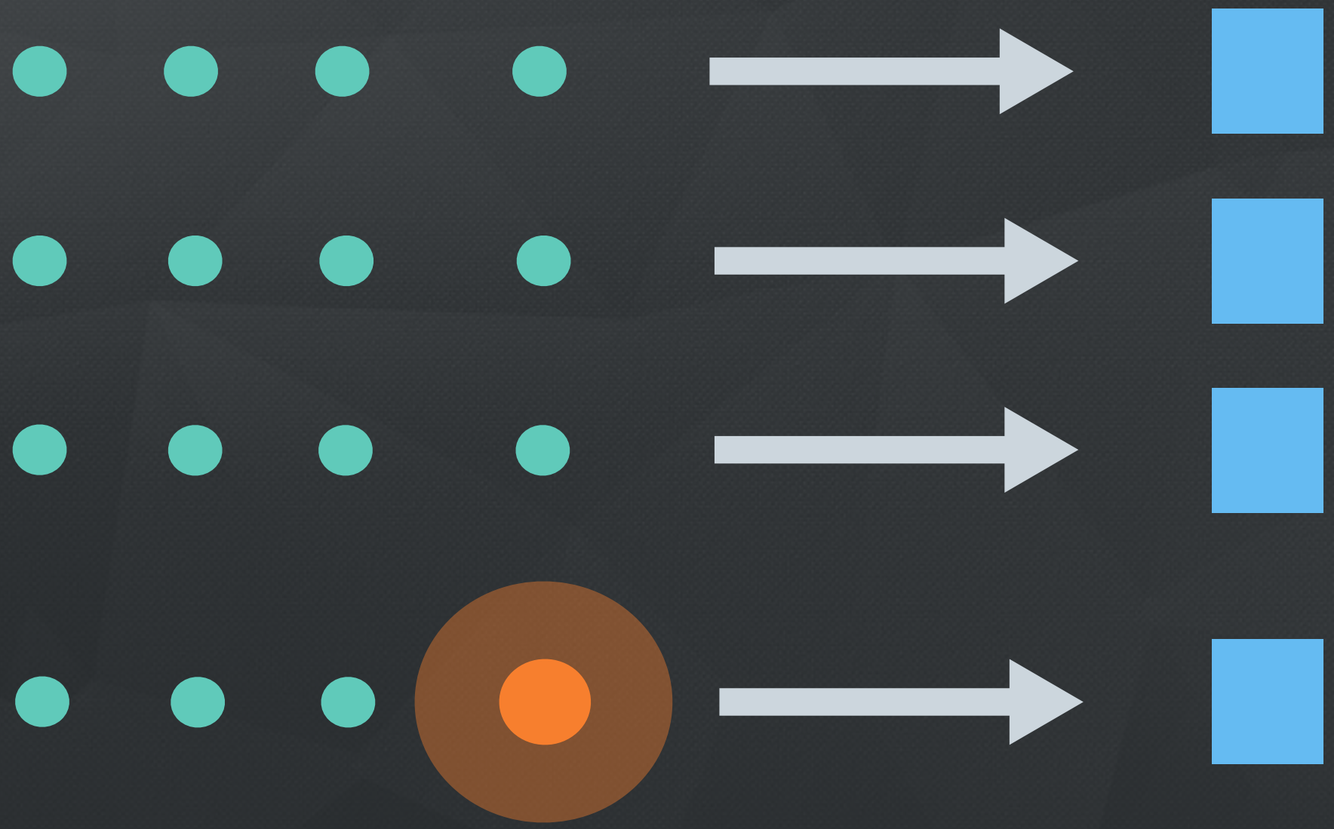
flight

# What If A Single App Is Still Too Big?



flight

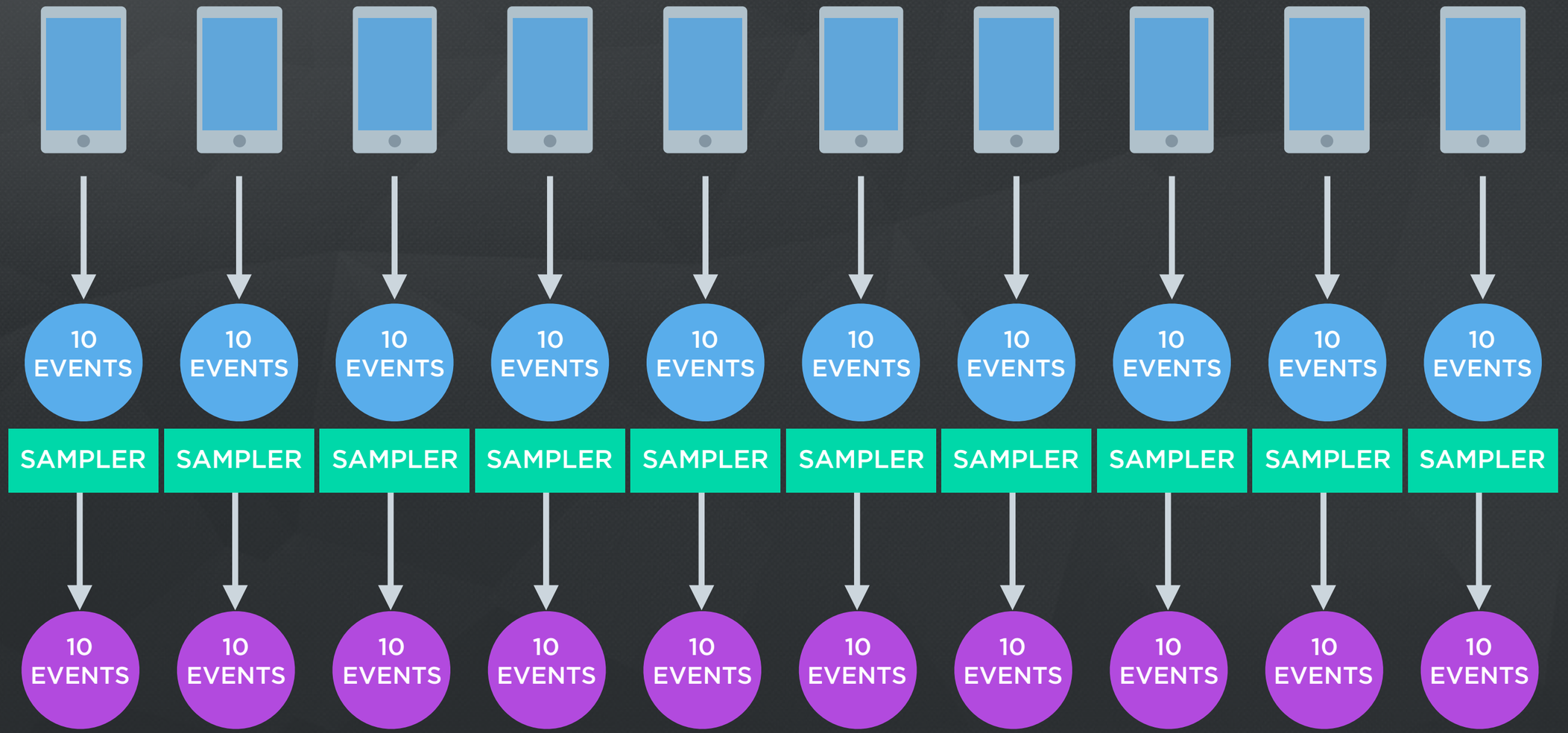
# Sampling



flight

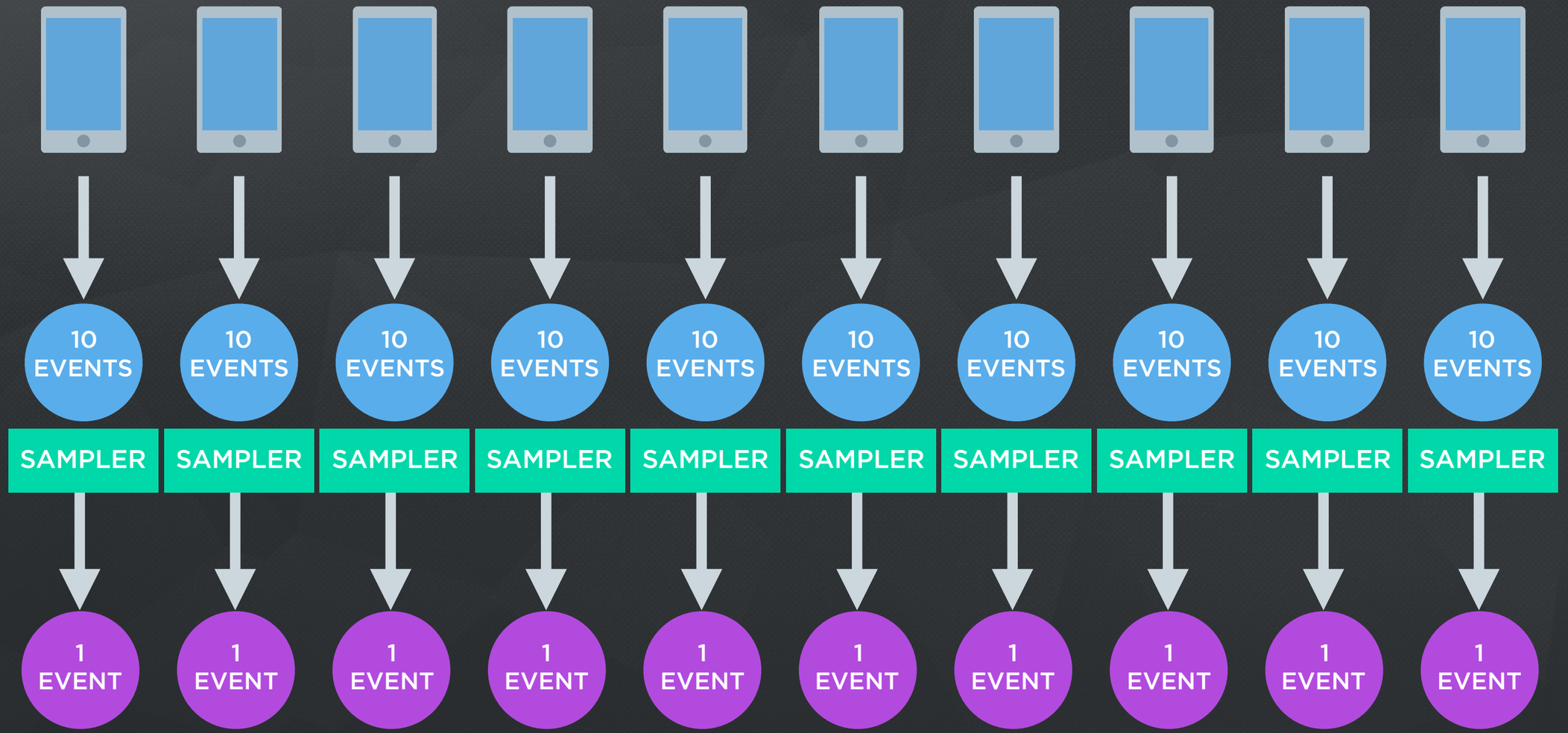
# How Do We Sample?

# Sampling: An Intro



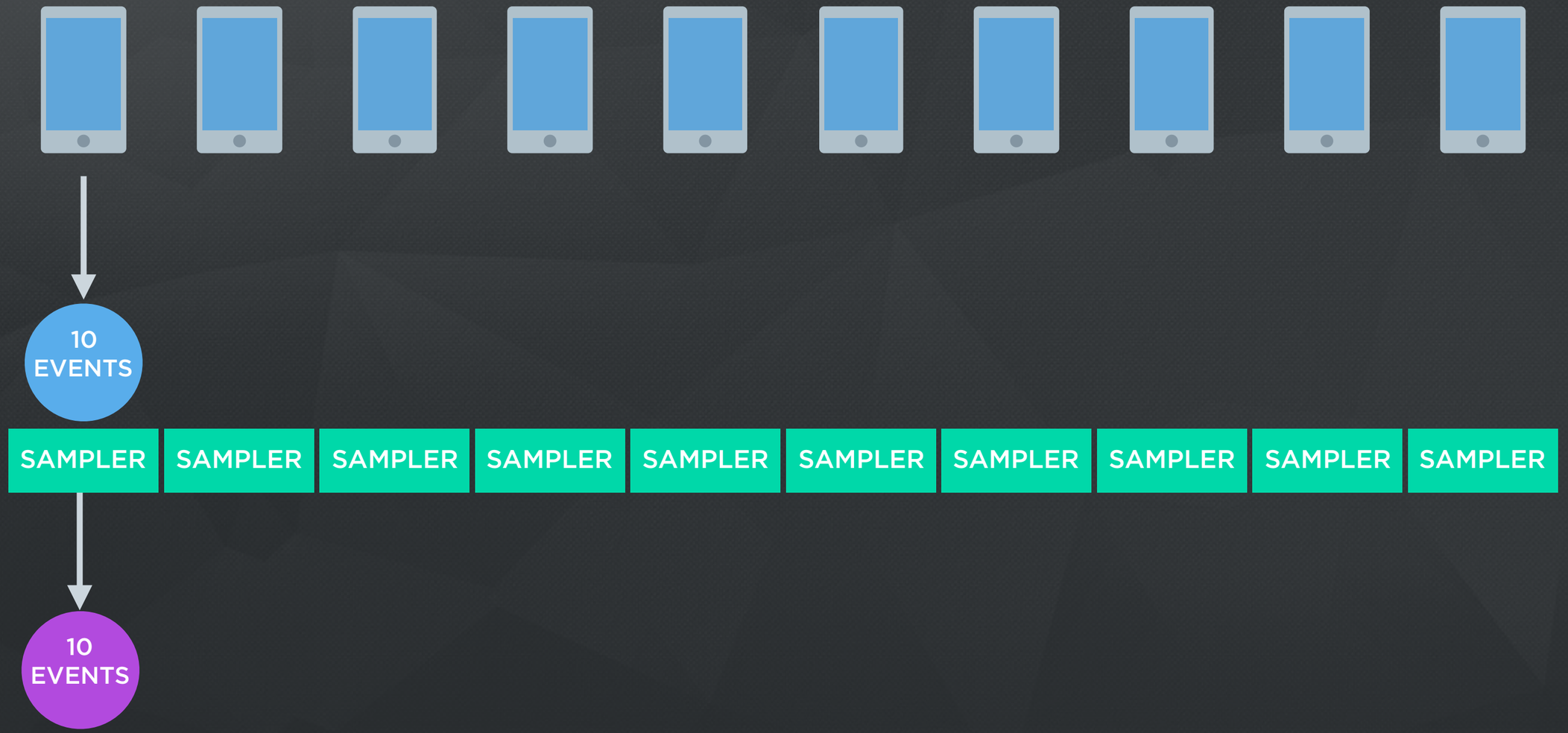

flight

# Sampling: Take 10 % Of Events?




flight

# Sampling: Take 100 % Of Events For 10% Of Devices




flight

# Sampling

- Enabled for 0.2% of our apps; reduces aggregate traffic by 80%
- Rough rule of thumb: 200,000,000 daily events before sampling is enabled
- Empirically tune sampling threshold and rate to achieve < 1% error for apps with sampling enabled
- Don’t sample lower volume events : INSTALL, CRASH, predefined-events, custom-events, Beta tester events, etc...

# SUMMARY

Expect Tradeoffs

Know the Tools To Deal with Trade offs

flight

# Thank You

@edsolovey

flight