Streaming analytics better than batch - when and why ?

Adam Kawa - Dawid Wysakowicz

getindata

About Us

- **At GetInData, we build custom Big Data solutions**
  - Hadoop, **Flink**, Spark, Kafka and more
- **Our team is today represented by**

Adam Kawa

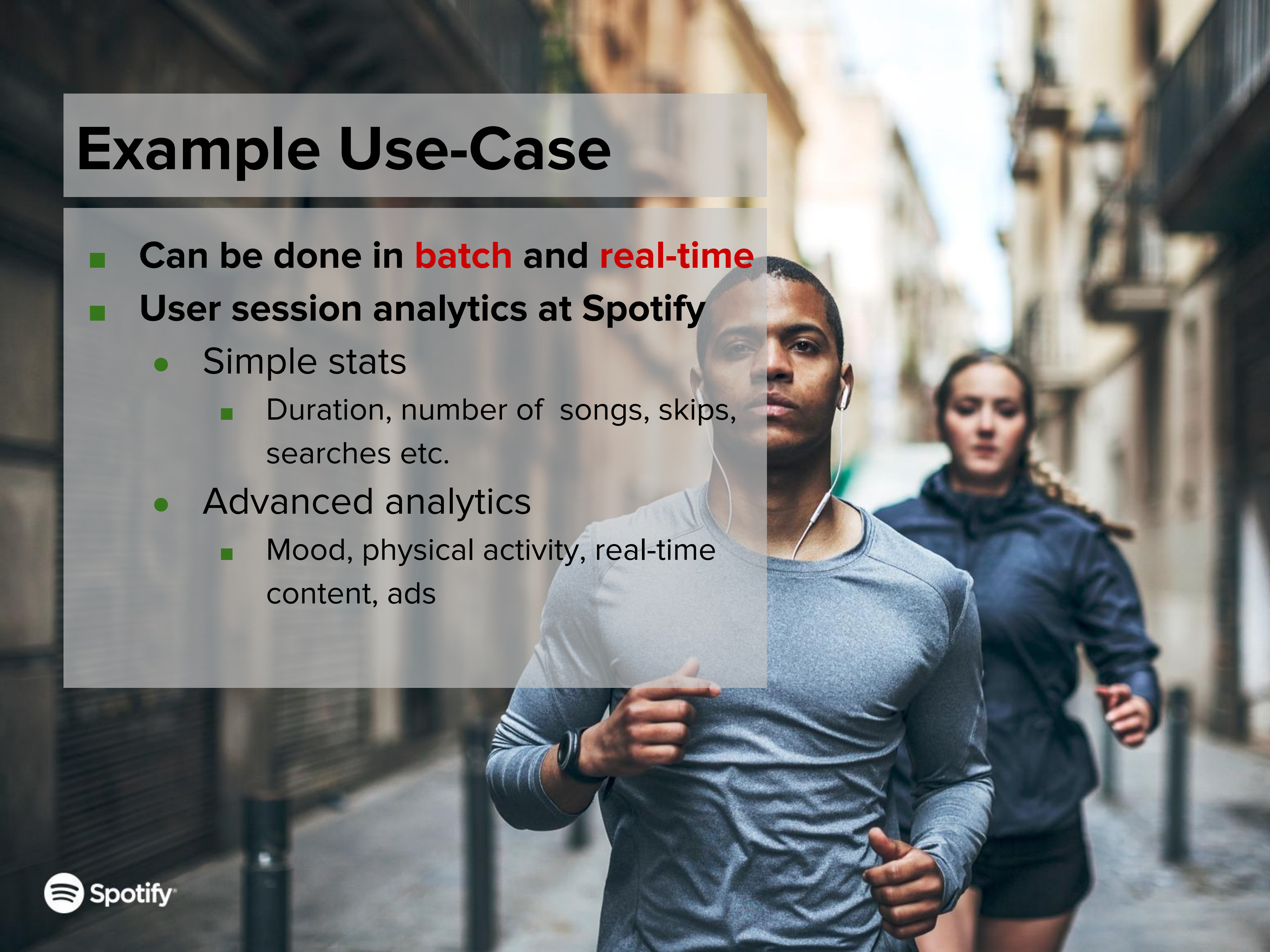
Dawid Wysakowicz

Have you ever built cool Big Data pipelines?

NOT COOL BRO

Example Use-Case

- **Can be done in batch and real-time**
- **User session analytics at Spotify**
  - Simple stats
    - Duration, number of songs, skips, searches etc.
  - Advanced analytics
    - Mood, physical activity, real-time content, ads



Example Output

1. Dashboards

How long do users listen to a **new** edition of Discover Weekly?

Example Output

1. Dashboards

How long do users listen to a **new** edition of Discover Weekly?

2. Alerts

Australian users are listening to Discover Weekly **too short** !!!

Example Output

1. Dashboards

2. Alerts

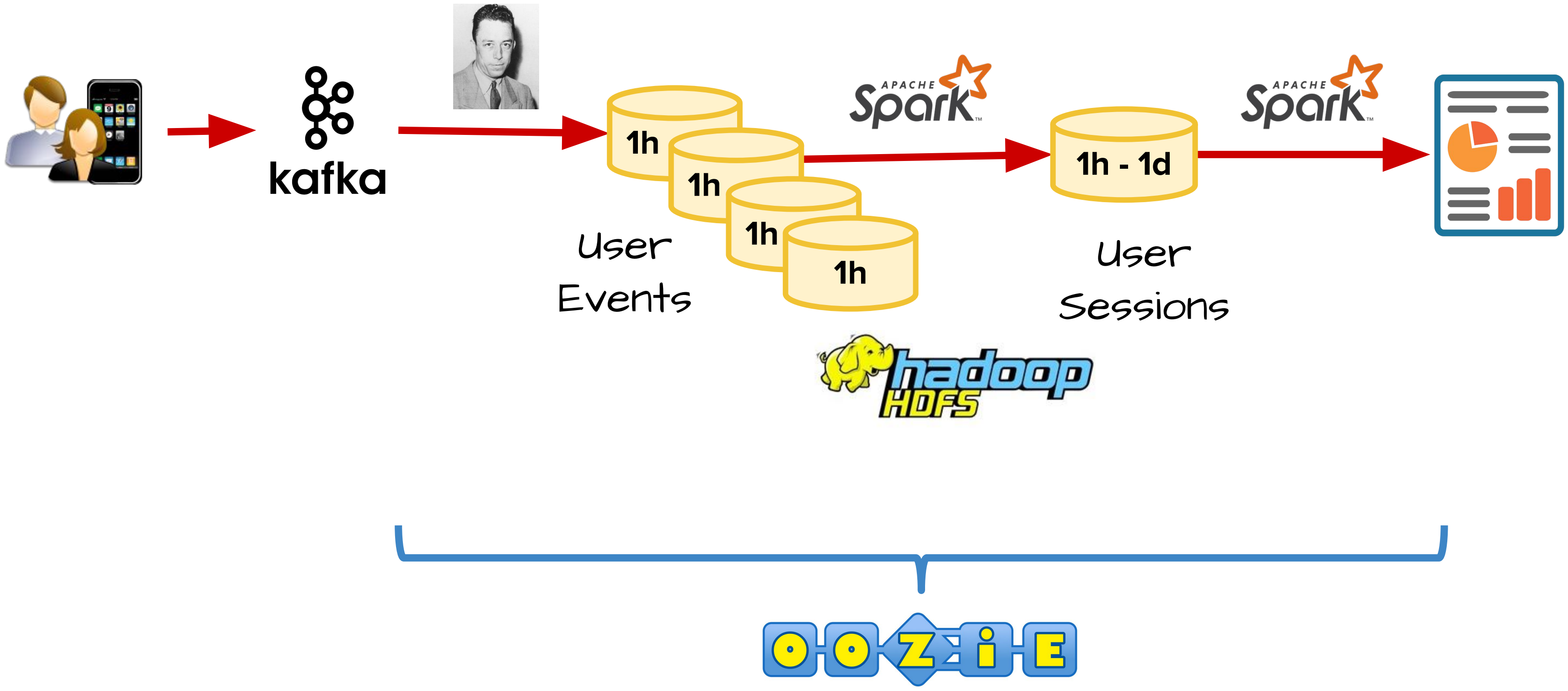
3. Content

How long do users listen to a **new** edition of Discover Weekly?

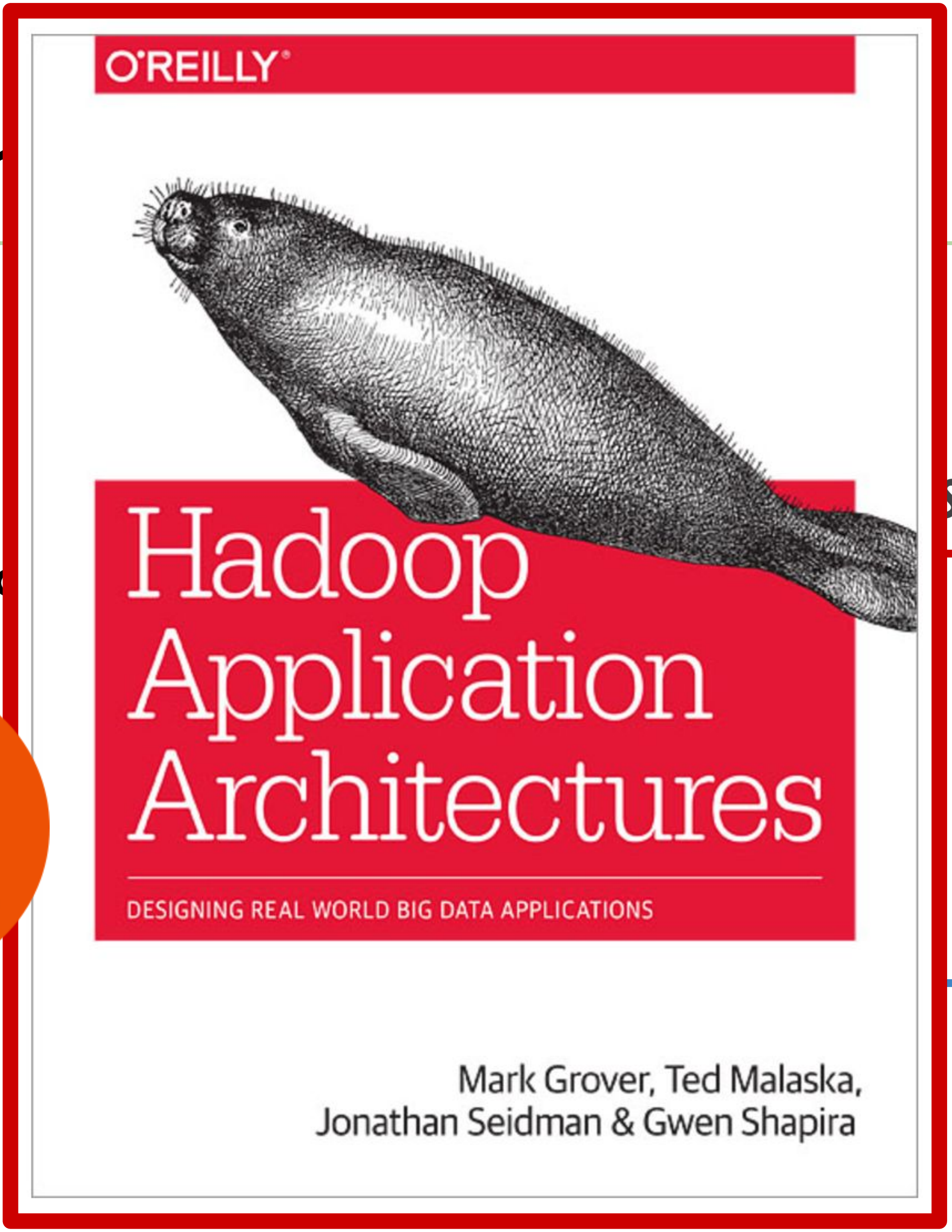
Australian users are listening to Discover Weekly **too short** !!!

Recommend songs and ads based on **current** activity.

1st - Batch Architecture



1st - Batch

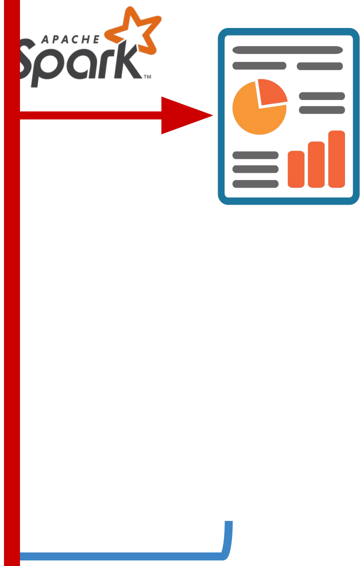


O'REILLY

Hadoop Application Architectures

DESIGNING REAL WORLD BIG DATA APPLICATIONS

Mark Grover, Ted Malaska,
Jonathan Seidman & Gwen Shapira



FIND OUT MORE!

Many Moving Parts

- ⬇ The higher learning curve
- ⬇ The more gluing code
- ⬇ The larger administrative effort
- ⬇ The more error-prone solution

Long Waiting Time

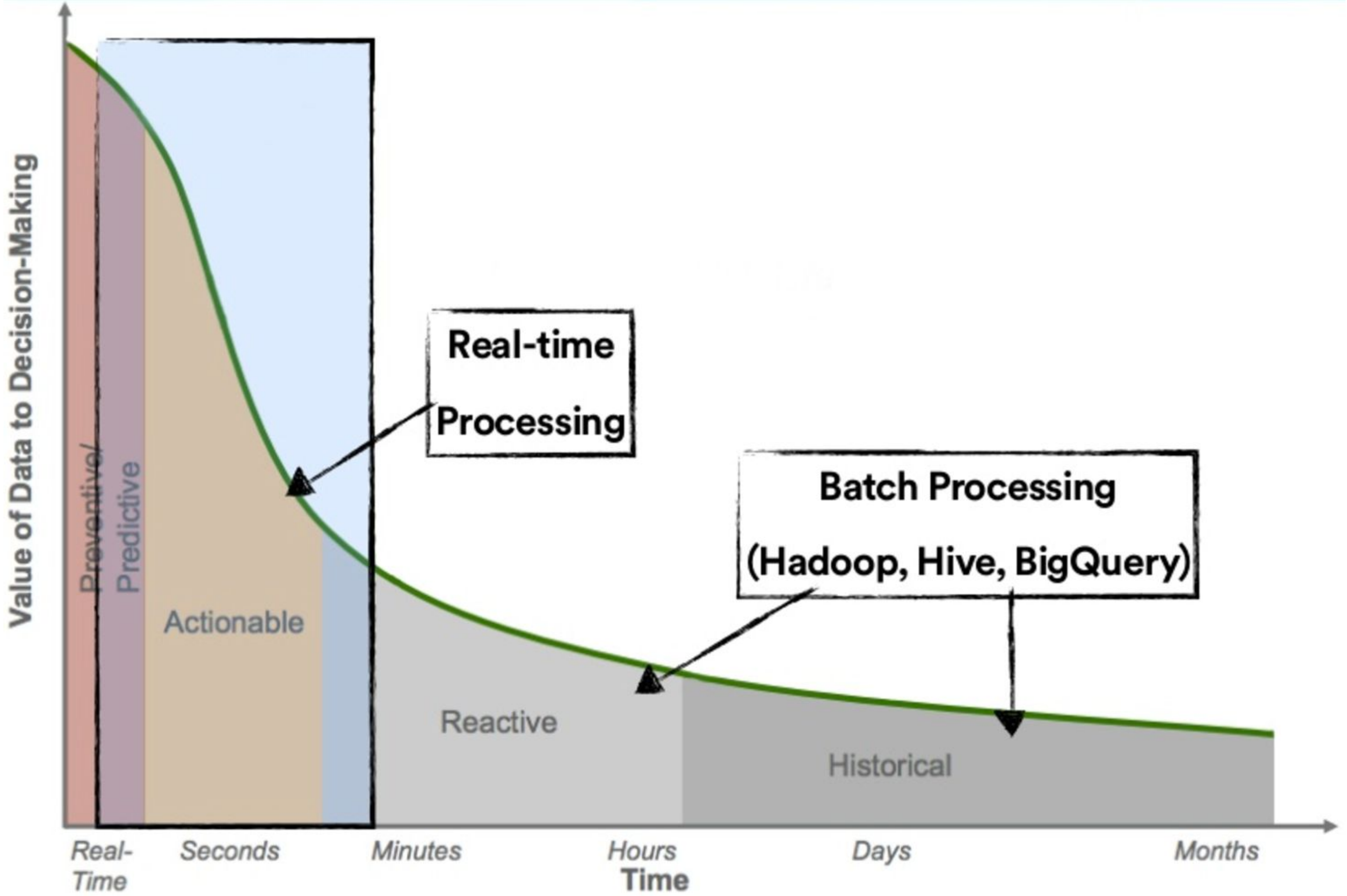
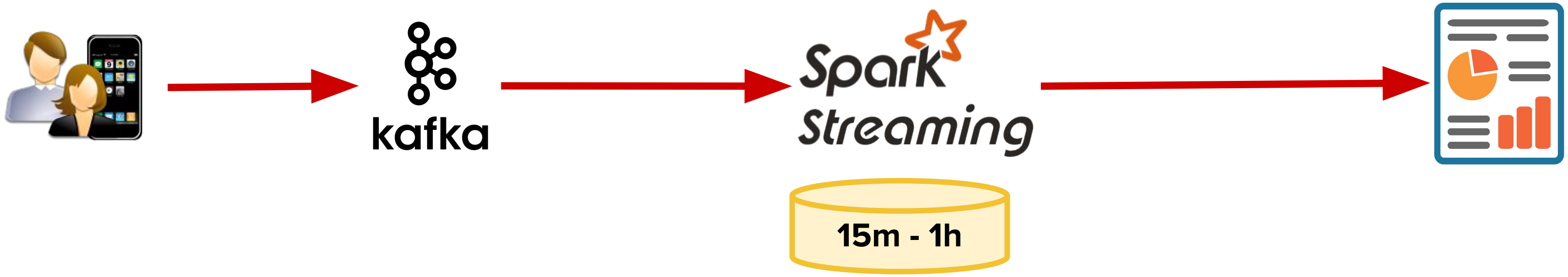
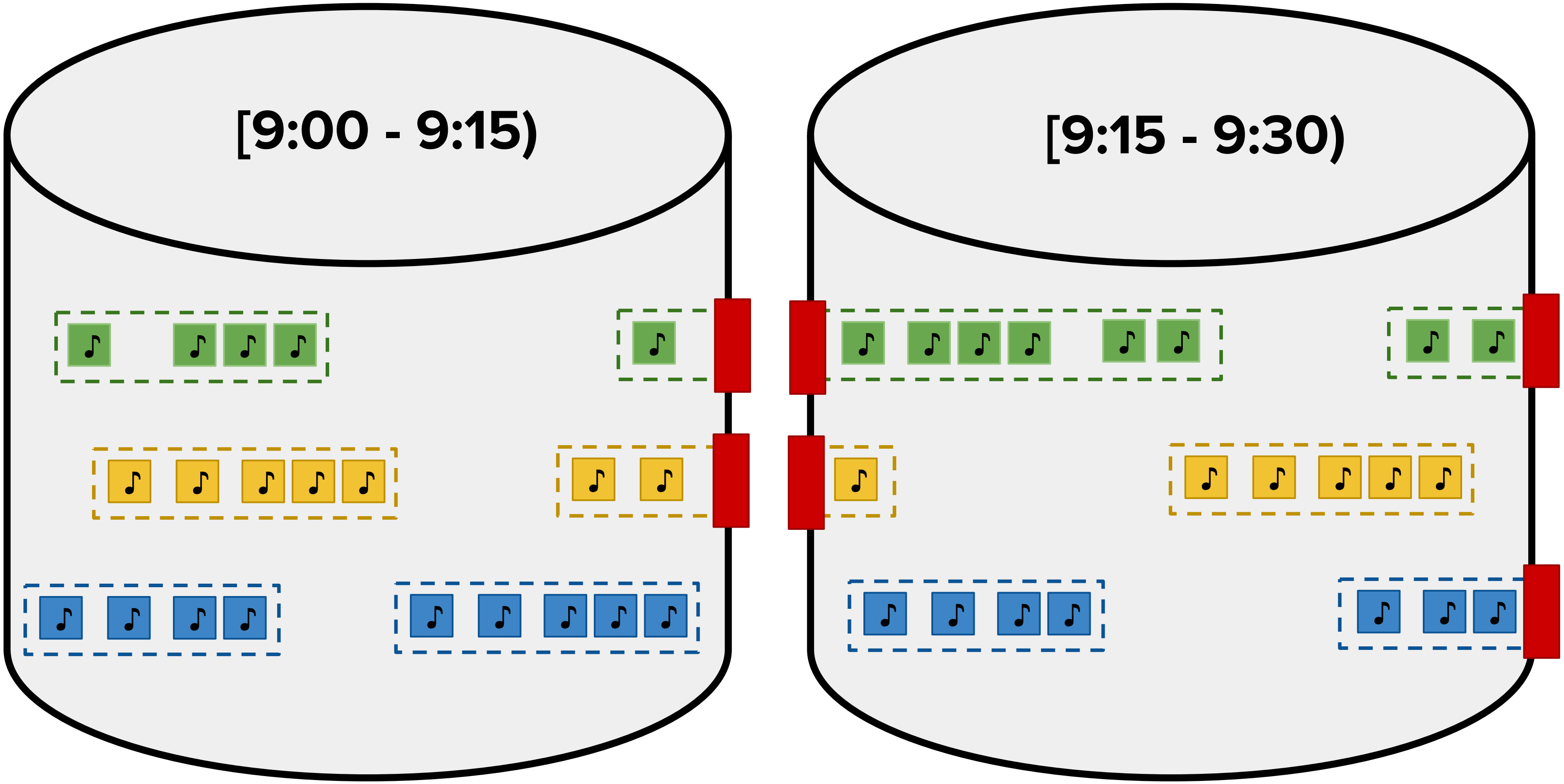


Image source: "Continuous Analytics: Stream Query Processing in Practice", Michael J Franklin, Professor, UC Berkley, Dec 2009 and http://www.slideshare.net/JoshBaer/shortening-the-feedback-loop-big-data-spain-external

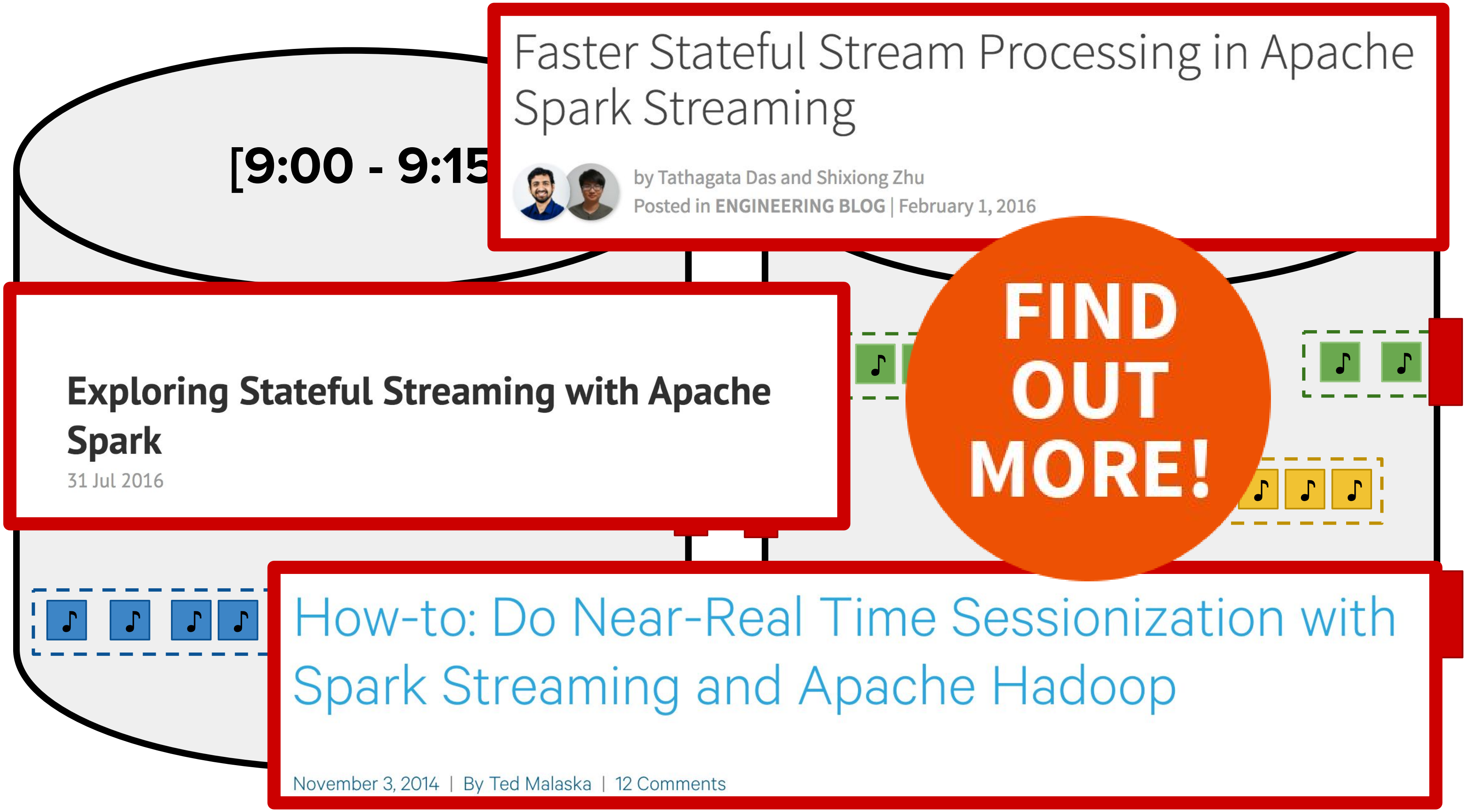
2nd - Micro-Batch Architecture



No Built-In Session Windows



No Built-In Session Windows



Faster Stateful Stream Processing in Apache Spark Streaming

by Tathagata Das and Shixiong Zhu
Posted in ENGINEERING BLOG | February 1, 2016

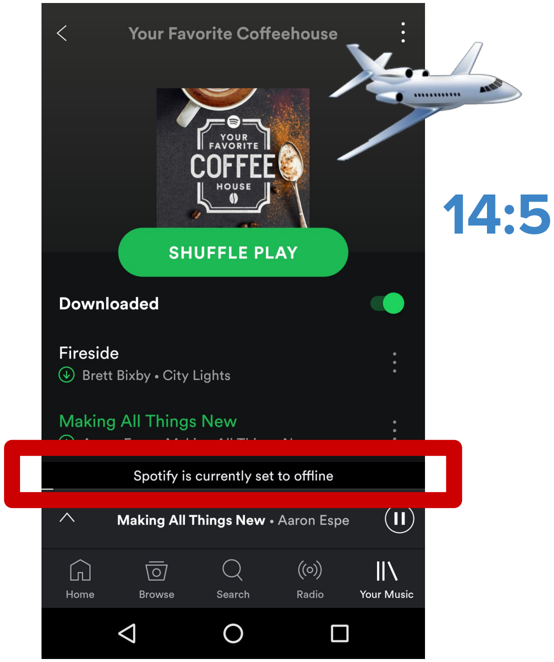
Exploring Stateful Streaming with Apache Spark

31 Jul 2016

How-to: Do Near-Real Time Sessionization with Spark Streaming and Apache Hadoop

November 3, 2014 | By Ted Malaska | 12 Comments

Late Data ...



Your Favorite Coffeehouse

Downloaded

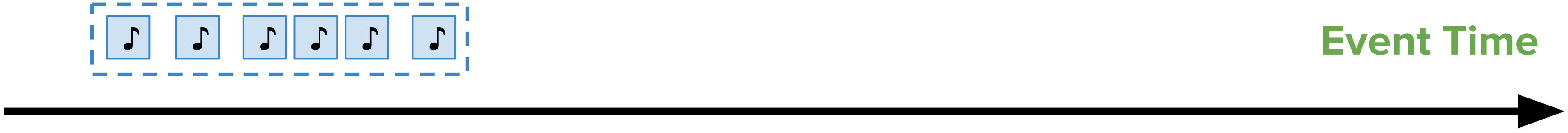
Fireside
Brett Bixby • City Lights

Making All Things New

Spotify is currently set to offline

Making All Things New • Aaron Espe

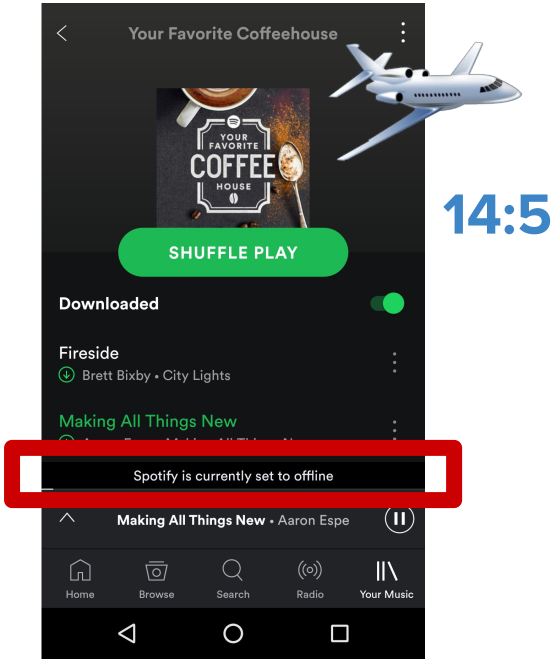
14:55 - 16:45



Event Time

Processing Time

… Included in Current Batch



Your Favorite Coffeehouse

SHUFFLE PLAY

Downloaded

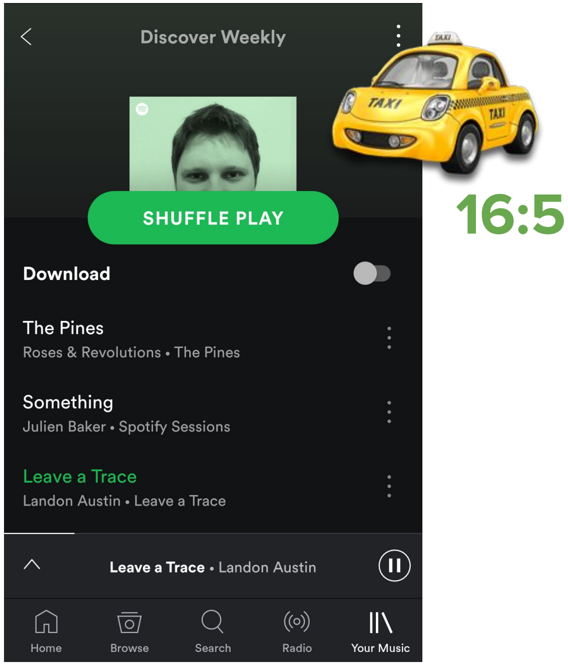
Fireside
Brett Bixby • City Lights

Making All Things New

Spotify is currently set to offline

Making All Things New • Aaron Espe

14:55 - 16:45



Discover Weekly

SHUFFLE PLAY

Download

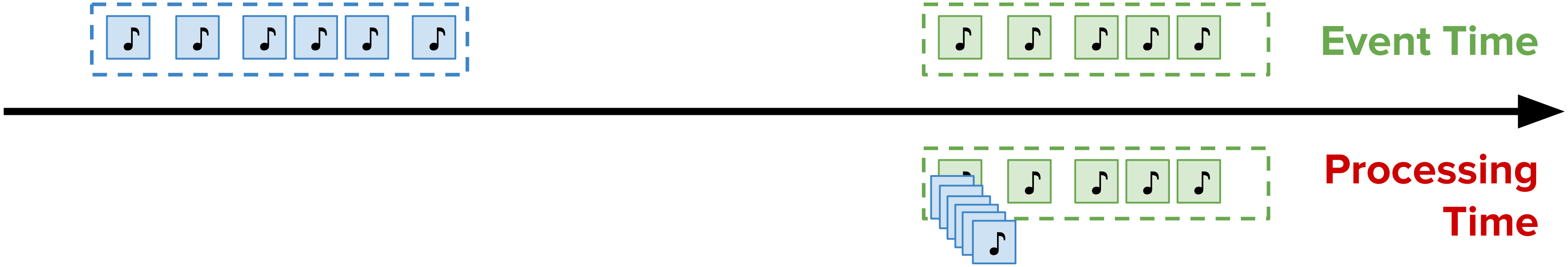
The Pines
Roses & Revolutions • The Pines

Something
Julien Baker • Spotify Sessions

Leave a Trace
Landon Austin • Leave a Trace

Leave a Trace • Landon Austin

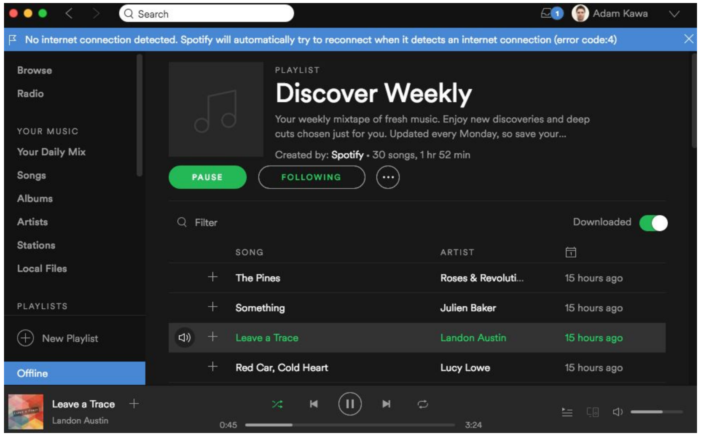
16:55 - …



Event Time

Processing Time

Out-Of-Order Data …



No internet connection detected. Spotify will automatically try to reconnect when it detects an internet connection (error code:4)

PLAYLIST

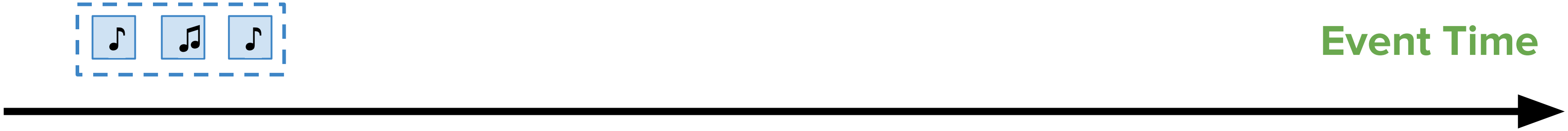
Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your...

Created by: Spotify • 30 songs, 1 hr 52 min

SONG	ARTIST	
The Pines	Roses & Revoluti...	15 hours ago
Something	Julien Baker	15 hours ago
Leave a Trace	Landon Austin	15 hours ago
Red Car, Cold Heart	Lucy Lowe	15 hours ago

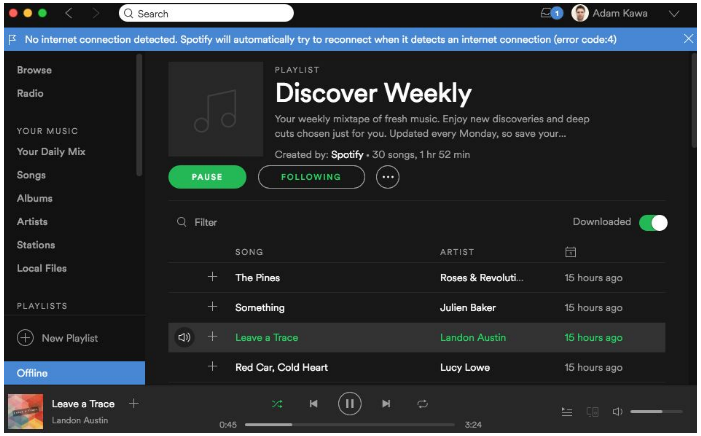
Leave a Trace
Landon Austin
0:45 3:24



Event Time

Processing Time

Out-Of-Order Data ...



No internet connection detected. Spotify will automatically try to reconnect when it detects an internet connection (error code:4)

PLAYLIST

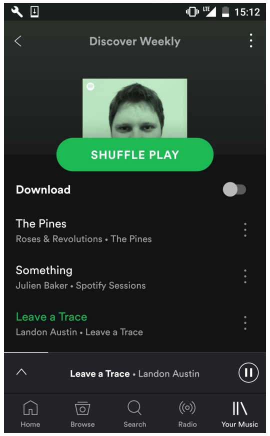
Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your...

Created by: Spotify • 30 songs, 1 hr 52 min

SONG	ARTIST	
The Pines	Roses & Revoluti...	15 hours ago
Something	Julien Baker	15 hours ago
Leave a Trace	Landon Austin	15 hours ago
Red Car, Cold Heart	Lucy Lowe	15 hours ago

Leave a Trace
Landon Austin

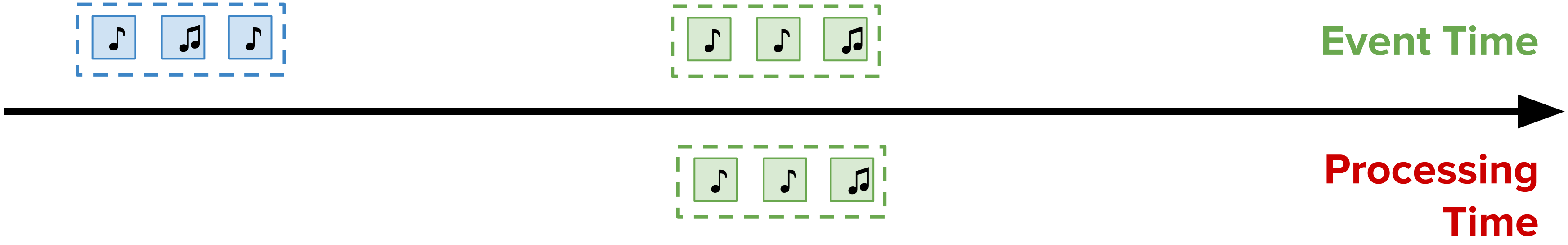


The Pines
Roses & Revolutions • The Pines

Something
Julien Baker • Spotify Sessions

Leave a Trace
Landon Austin • Leave a Trace

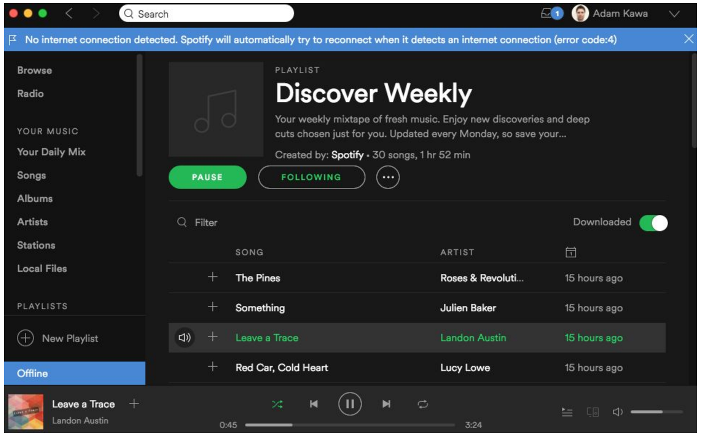
Leave a Trace • Landon Austin



Event Time

Processing Time

Out-Of-Order Data ...



Search

Adam Kawa

No internet connection detected. Spotify will automatically try to reconnect when it detects an internet connection (error code:4)

Browse
Radio

YOUR MUSIC
Your Daily Mix
Songs
Albums
Artists
Stations
Local Files

PLAYLISTS
New Playlist

Offline

PLAYLIST

Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your...

Created by: Spotify • 30 songs, 1 hr 52 min

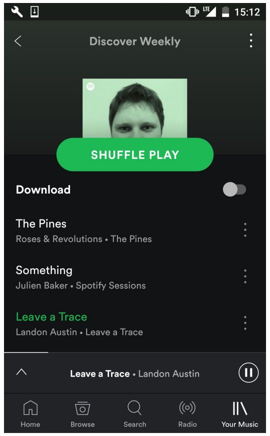
PAUSE FOLLOWING

Filter Downloaded

SONG	ARTIST	
The Pines	Roses & Revoluti...	15 hours ago
Something	Julien Baker	15 hours ago
Leave a Trace	Landon Austin	15 hours ago
Red Car, Cold Heart	Lucy Lowe	15 hours ago

Leave a Trace
Landon Austin

0:45 3:24



15:12

Discover Weekly

SHUFFLE PLAY

Download

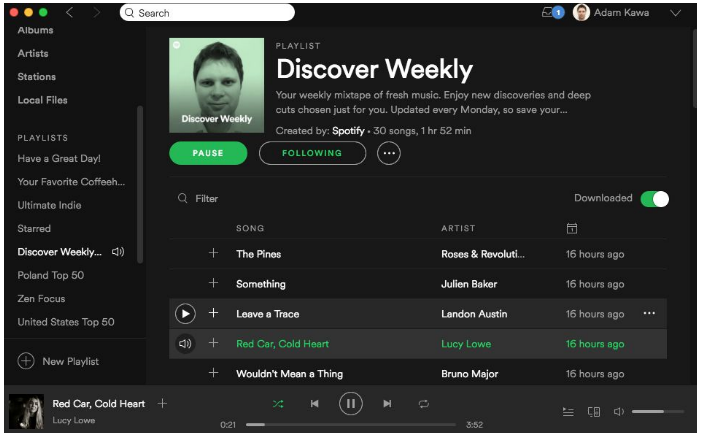
The Pines
Roses & Revolutions • The Pines

Something
Julien Baker • Spotify Sessions

Leave a Trace
Landon Austin • Leave a Trace

Leave a Trace • Landon Austin

Home Browse Search Radio Your Music



Search

Adam Kawa

Albums
Artists
Stations
Local Files

PLAYLISTS
Have a Great Day!
Your Favorite Coffeeh...
Ultimate Indie
Starred
Discover Weekly...
Poland Top 50
Zen Focus
United States Top 50

New Playlist

PLAYLIST

Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your...

Created by: Spotify • 30 songs, 1 hr 52 min

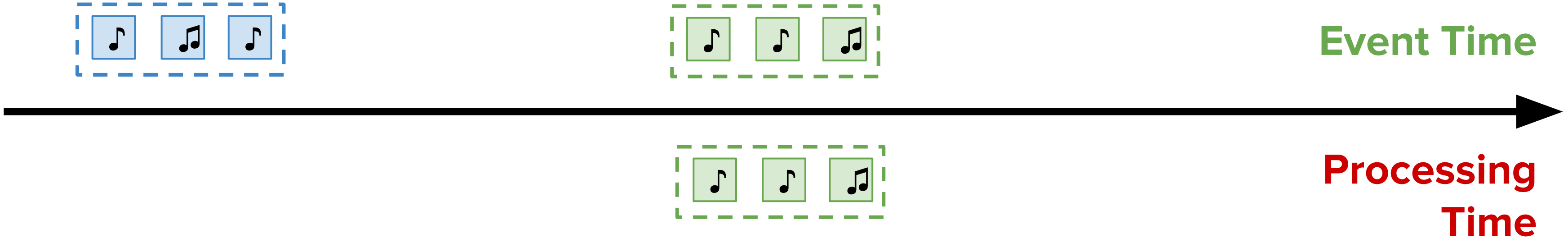
PAUSE FOLLOWING

Filter Downloaded

SONG	ARTIST	
The Pines	Roses & Revoluti...	16 hours ago
Something	Julien Baker	16 hours ago
Leave a Trace	Landon Austin	16 hours ago
Red Car, Cold Heart	Lucy Lowe	16 hours ago
Wouldn't Mean a Thing	Bruno Major	16 hours ago

Red Car, Cold Heart
Lucy Lowe

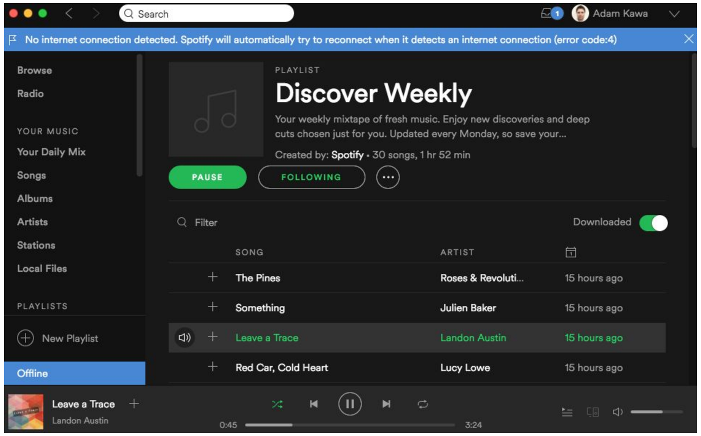
0:21 3:52



Event Time

Processing Time

… Breaks Correctness



No internet connection detected. Spotify will automatically try to reconnect when it detects an internet connection (error code:4)

PLAYLIST

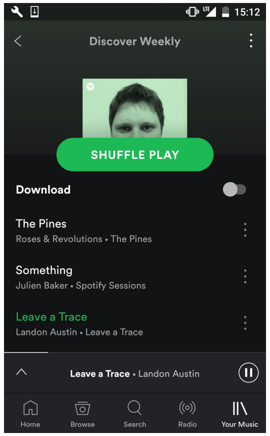
Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your…

Created by: Spotify • 30 songs, 1 hr 52 min

SONG	ARTIST	
The Pines	Roses & Revoluti…	15 hours ago
Something	Julien Baker	15 hours ago
Leave a Trace	Landon Austin	15 hours ago
Red Car, Cold Heart	Lucy Lowe	15 hours ago

Leave a Trace
Landon Austin
0:45 3:24



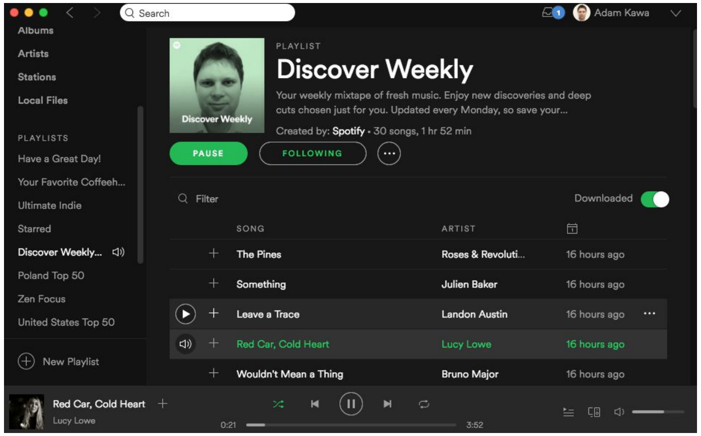
Discover Weekly

The Pines
Roses & Revolutions • The Pines

Something
Julien Baker • Spotify Sessions

Leave a Trace
Landon Austin • Leave a Trace

Leave a Trace • Landon Austin



PLAYLIST

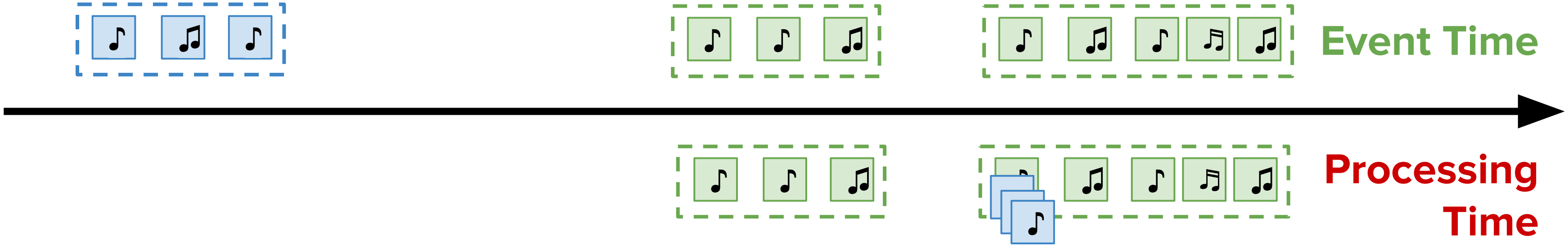
Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your…

Created by: Spotify • 30 songs, 1 hr 52 min

SONG	ARTIST	
The Pines	Roses & Revoluti…	16 hours ago
Something	Julien Baker	16 hours ago
Leave a Trace	Landon Austin	16 hours ago
Red Car, Cold Heart	Lucy Lowe	16 hours ago
Wouldn't Mean a Thing	Bruno Major	16 hours ago

Red Car, Cold Heart
Lucy Lowe
0:21 3:52



Event Time

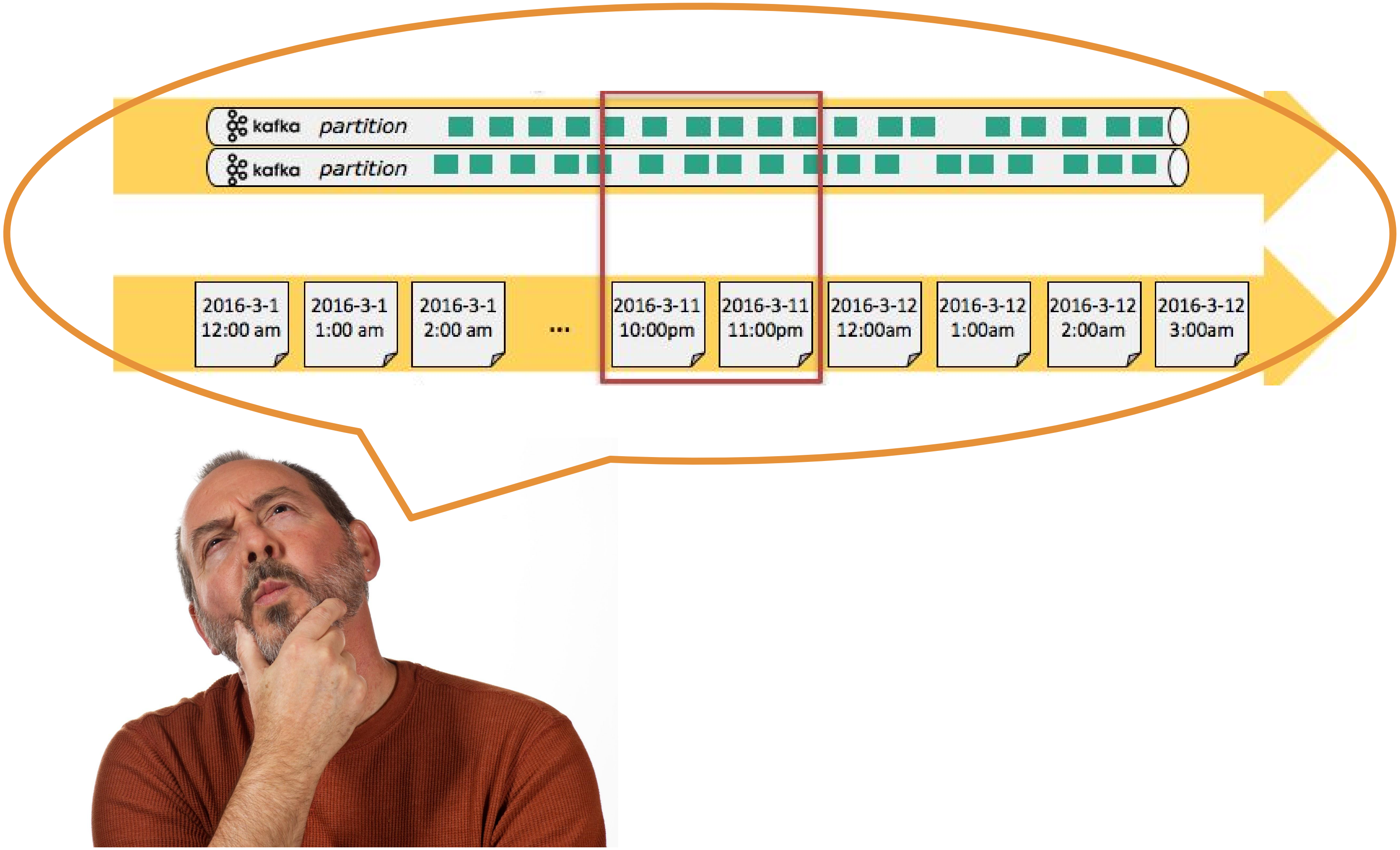
Processing Time

Problems

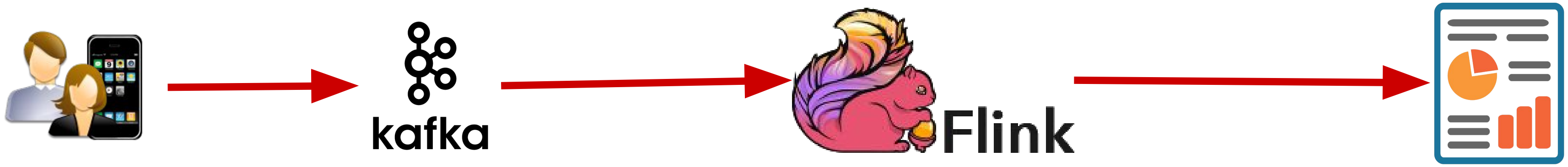
FILES,
BATCHES,
DATA LAKES
Hive
Spark

HISTORICAL INACCURACIES
HISTORICAL INACCURACIES EVERYWHERE
quickmeme.com

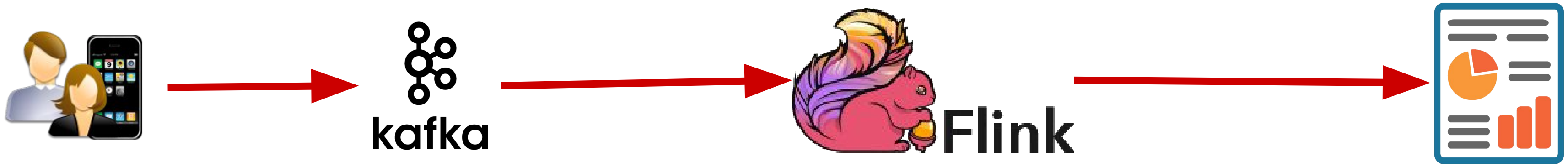
Solving Streaming Problem With Batch?

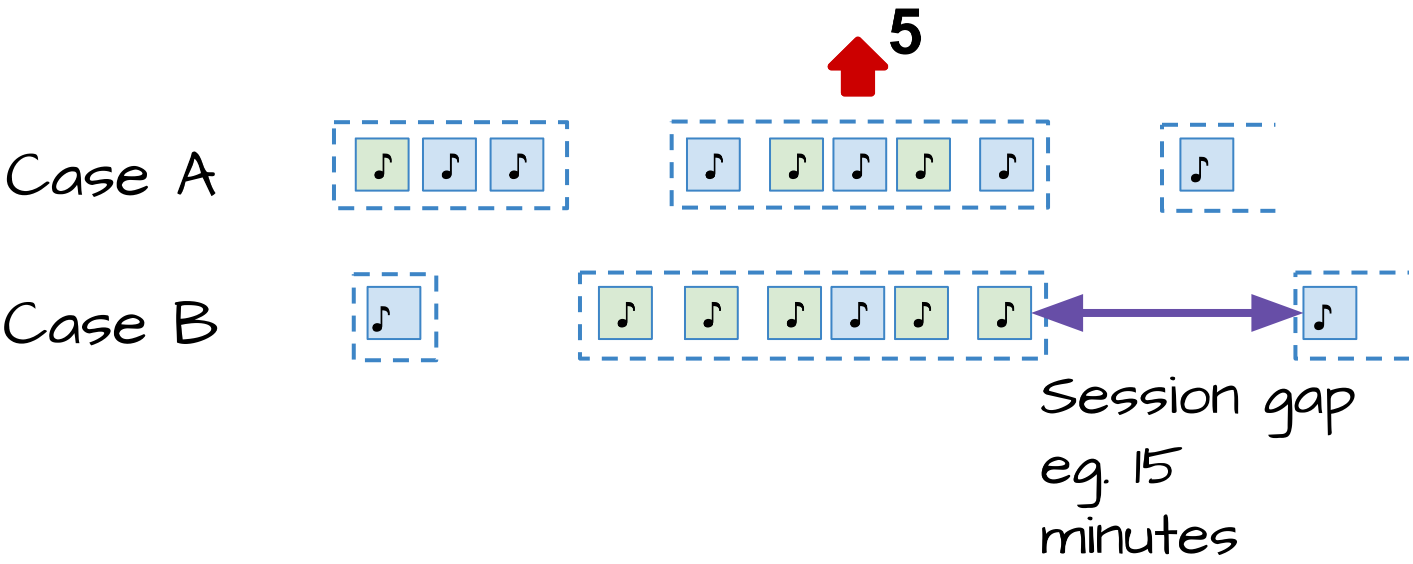


3rd - Streaming-First Architecture



User Session Windows




5

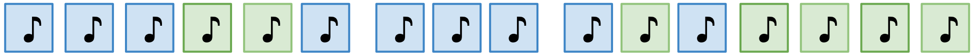
Case A

Case B

Session gap
eg. 15
minutes

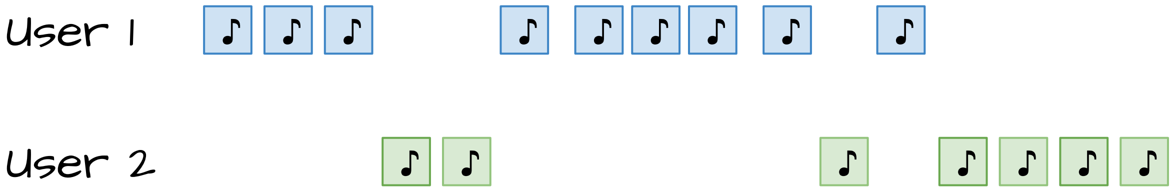
Reading From Kafka

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
```



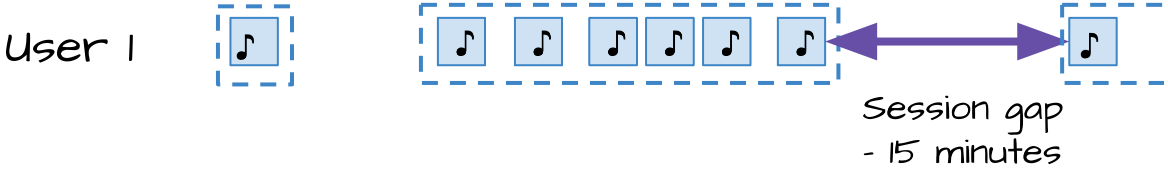
Session Windows With Gap

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
```



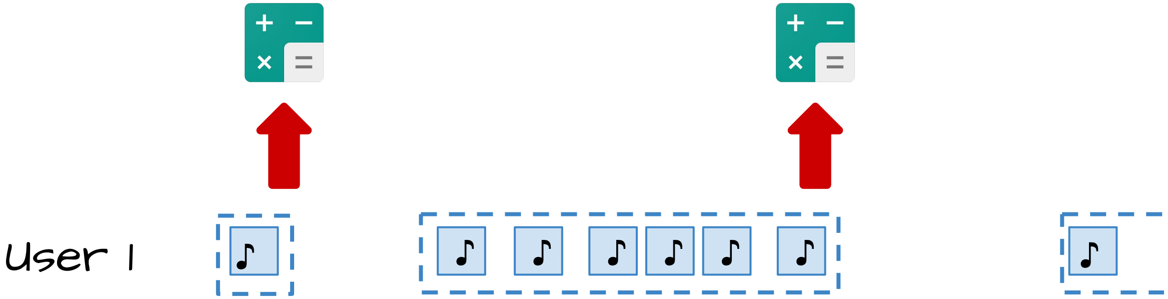
Session Windows With Gap

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
    .window(EventTimeSessionWindows.withGap(Time.minutes(15)))
```



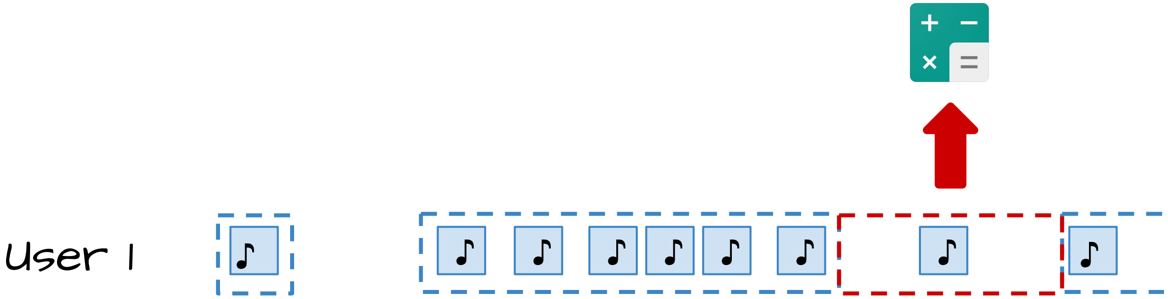
Analyzing User Session

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
    .window(EventTimeSessionWindows.withGap(Time.minutes(15)))
    .apply(new CountSessionStats())
```



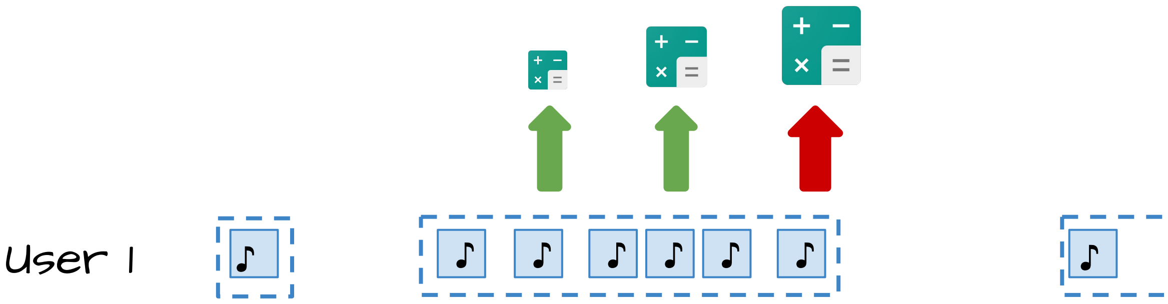
Handling Late Events

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
    .window(EventTimeSessionWindows.withGap(Time.minutes(15)))
    .allowedLateness(Time.minutes(60))
    .apply(new CountSessionStats())
```



Triggering Early Results

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
    .window(EventTimeSessionWindows.withGap(Time.minutes(15)))
    .trigger(EarlyTriggeringTrigger.every(Time.minutes(10)))
    .allowedLateness(Time.minutes(60))
    .apply(new CountSessionStats())
```



Sessionization Example

```
val sessionStream : DataStream[SessionStats] = sEnv
    .addSource(new KafkaConsumer(...))
    .keyBy(_.userId)
    .window(EventTimeSessionWindows.withGap(Time.minutes(15)))
    .trigger(EarlyTriggeringTrigger.every(Time.minutes(10)))
    .allowedLateness(Time.minutes(60))
    .apply(new CountSessionStats())
```

FIND OUT MORE!

Working example:
https://github.com/getindata/flink-use-case

Modern Stream Processing Engines

- **Rich stream processing semantic**
  - Built-in support for event-time windows
  - Accurate results for late / out-of-order events and replays
  - Early triggers
- **Low latency and high-throughput**
- **Exactly-once stateful processing**

Flink

Modern Stream Processing Engines

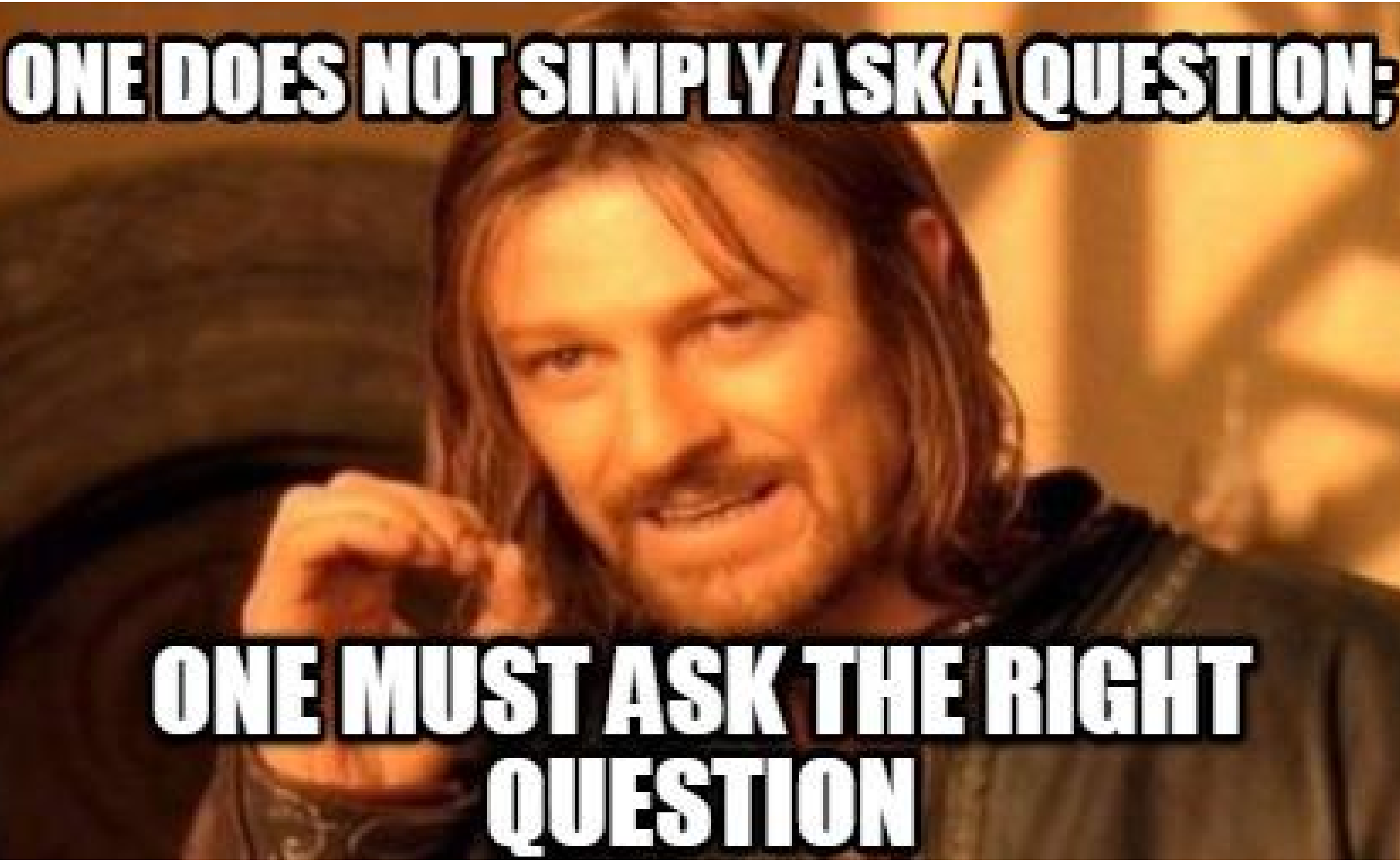
- **Rich stream processing semantic**
  - Built-in support for event-time windows
  - Accurate results for late / out-of-order events and replays
  - Early triggers
- **Low latency and high-throughput**
- **Exactly-once stateful processing**

FIND OUT MORE!

User survey:

http://data-artisans.com/flink-user-survey-2016-part-1

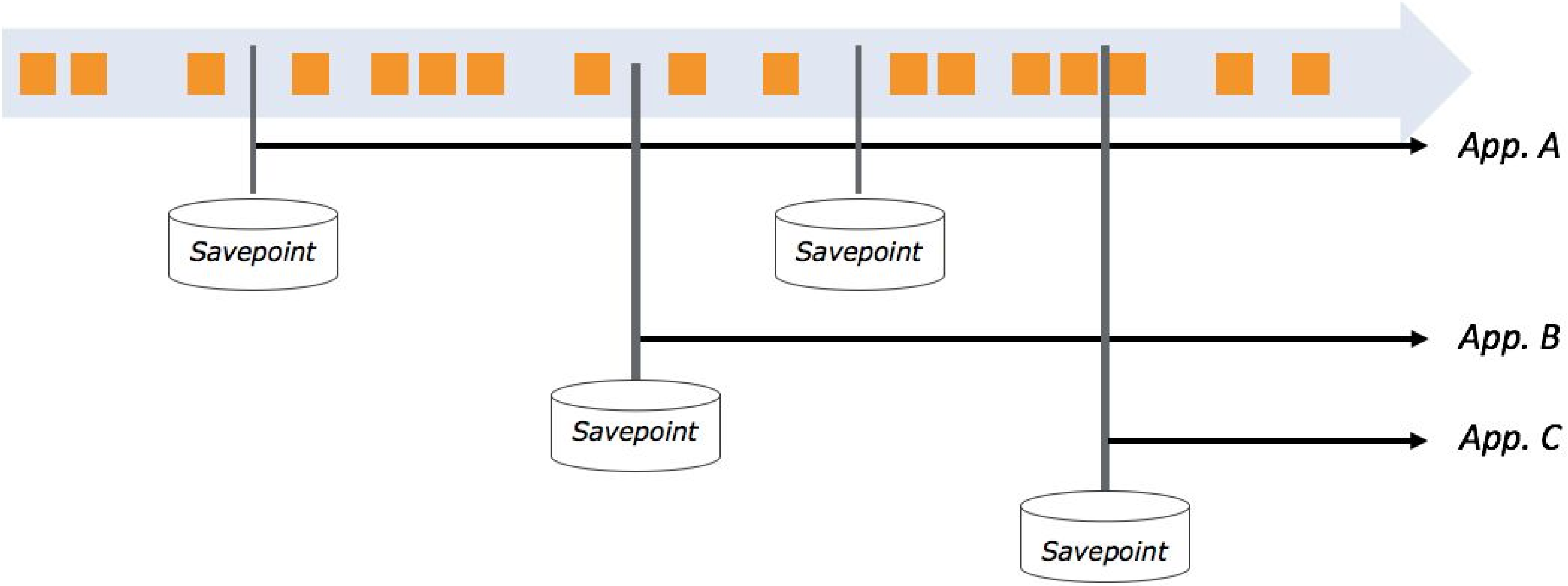
http://data-artisans.com/flink-user-survey-2016-part-2


ONE DOES NOT SIMPLY ASK A QUESTION;
ONE MUST ASK THE RIGHT QUESTION

How can I reprocess data?

Reprocessing Events In Flink

1. **Take periodic snapshots of a job**
   - It stores Kafka offsets, on-flight sessions, application state
2. **Restart a job from a savepoint rather than from a beginning**



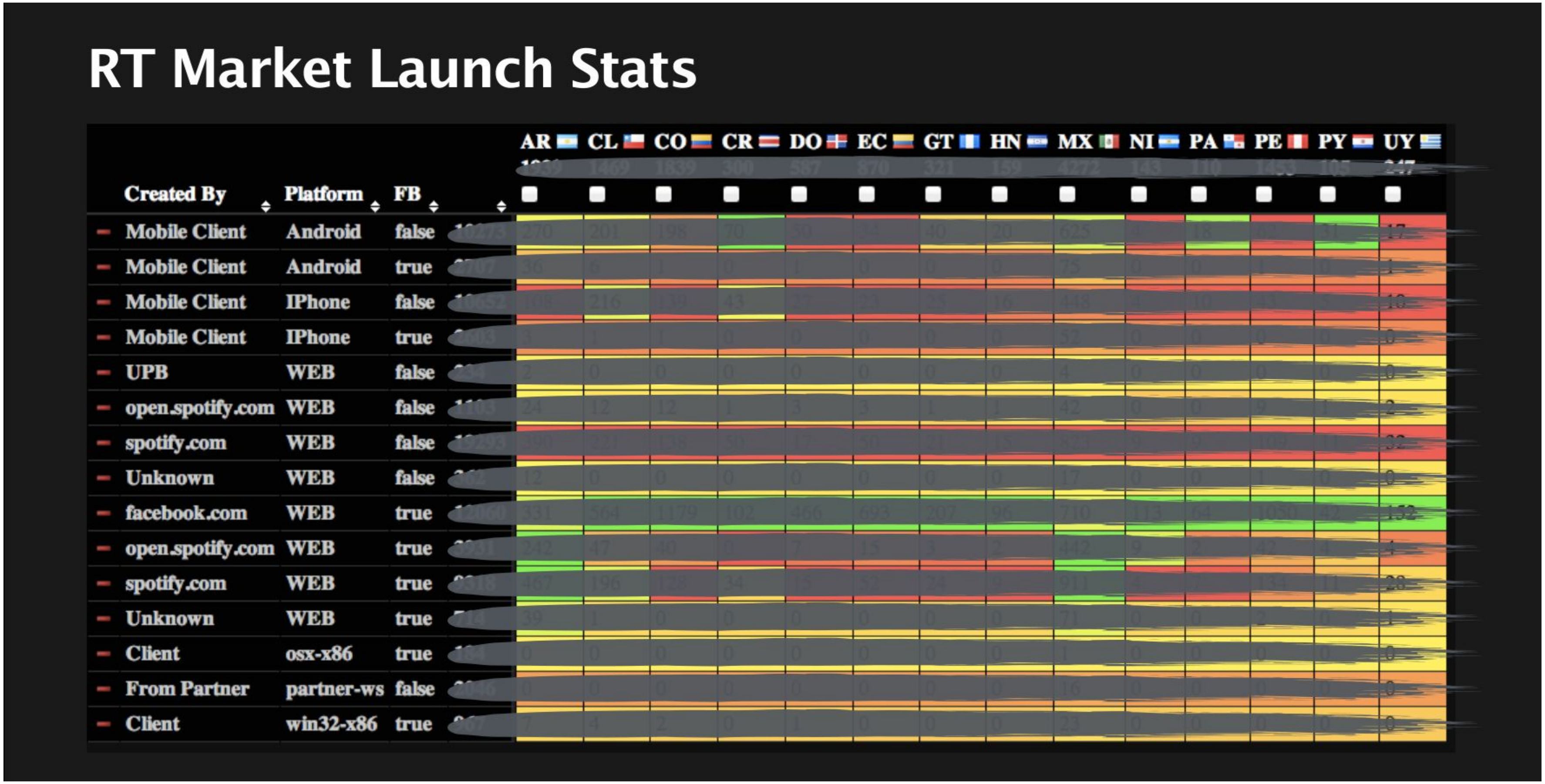
What if data is no longer in Kafka?

Consuming Data From HDFS

- **Run your streaming code on HDFS (bounded data)**
  - You need to read data in event-time based order
  - Implement mechanism of proper watermark generation

What are usual stream processing applications?

Stream Analytics

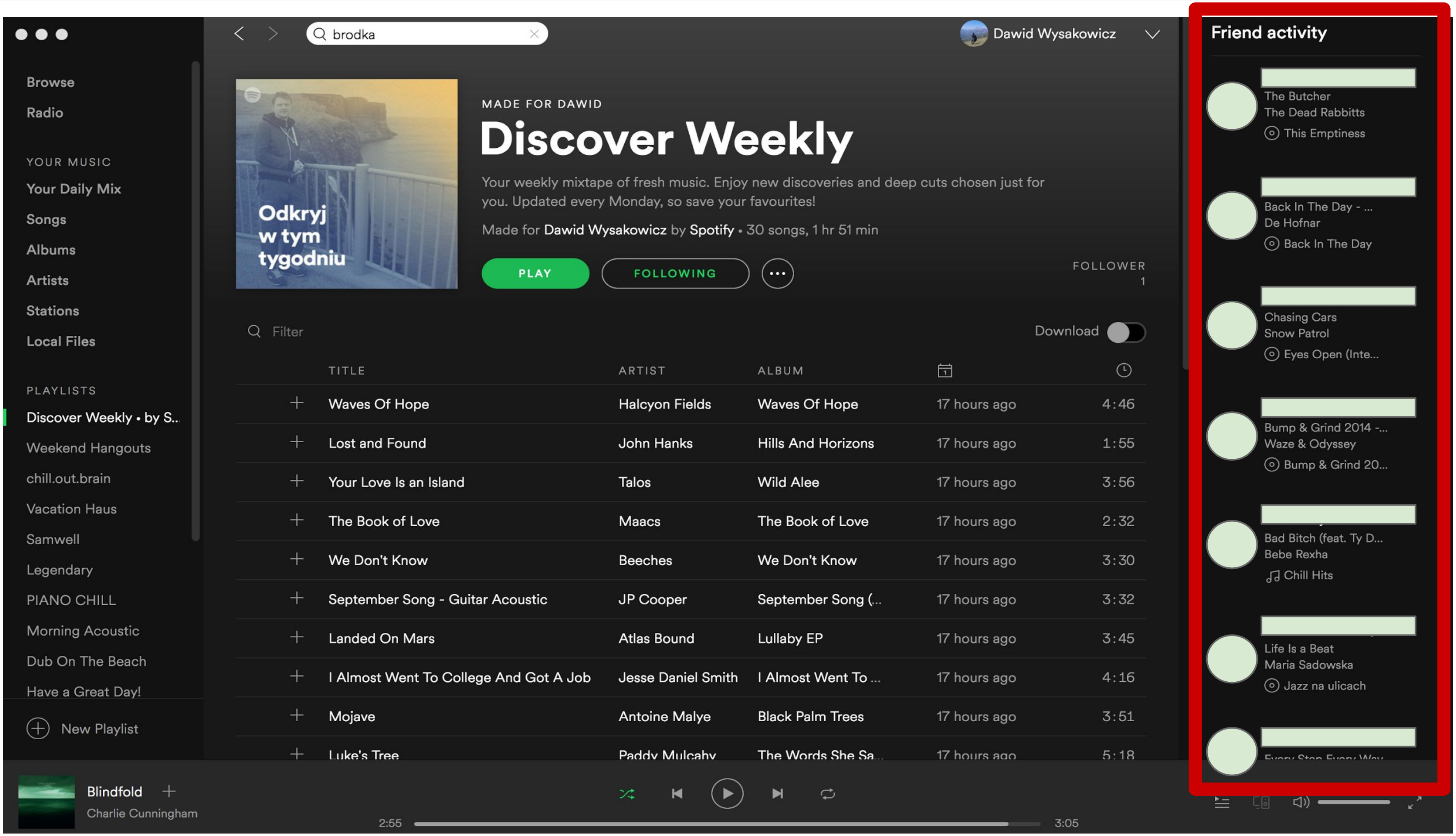


RT Market Launch Stats

	Created By	Platform	FB		AR	CL	CO	CR	DO	EC	GT	HN	MX	NI	PA	PE	PY	UY
					[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
					□	□	□	□	□	□	□	□	□	□	□	□	□	□
−	Mobile Client	Android	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Mobile Client	Android	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Mobile Client	IPhone	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Mobile Client	IPhone	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	UPB	WEB	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	open.spotify.com	WEB	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	spotify.com	WEB	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Unknown	WEB	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	facebook.com	WEB	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	open.spotify.com	WEB	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	spotify.com	WEB	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Unknown	WEB	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Client	osx-x86	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	From Partner	partner-ws	false	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
−	Client	win32-x86	true	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Image source: https://www.slideshare.net/sinisalyh/storm-at-spotify

Stream 24/7 Applications



brodka

Dawid Wysakowicz

Browse
Radio

YOUR MUSIC

Your Daily Mix
Songs
Albums
Artists
Stations
Local Files

PLAYLISTS

Discover Weekly • by S...
Weekend Hangouts
chill.out.brain
Vacation Haus
Samwell
Legendary
PIANO CHILL
Morning Acoustic
Dub On The Beach
Have a Great Day!

New Playlist

Odkryj w tym tygodniu

MADE FOR DAWID

Discover Weekly

Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Updated every Monday, so save your favourites!

Made for Dawid Wysakowicz by Spotify • 30 songs, 1 hr 51 min

PLAY FOLLOWING ···

FOLLOWER 1

Filter

Download

TITLE	ARTIST	ALBUM		
Waves Of Hope	Halcyon Fields	Waves Of Hope	17 hours ago	4:46
Lost and Found	John Hanks	Hills And Horizons	17 hours ago	1:55
Your Love Is an Island	Talos	Wild Alee	17 hours ago	3:56
The Book of Love	Maacs	The Book of Love	17 hours ago	2:32
We Don't Know	Beeches	We Don't Know	17 hours ago	3:30
September Song - Guitar Acoustic	JP Cooper	September Song (...	17 hours ago	3:32
Landed On Mars	Atlas Bound	Lullaby EP	17 hours ago	3:45
I Almost Went To College And Got A Job	Jesse Daniel Smith	I Almost Went To ...	17 hours ago	4:16
Mojave	Antoine Malye	Black Palm Trees	17 hours ago	3:51
Luke's Tree	Paddy Mulcahy	The Words She Sa...	17 hours ago	5:18

Blindfold
Charlie Cunningham

2:55 3:05

Friend activity

[redacted]
The Butcher
The Dead Rabbitts
This Emptiness

[redacted]
Back In The Day - ...
De Hofnar
Back In The Day

[redacted]
Chasing Cars
Snow Patrol
Eyes Open (Inte...

[redacted]
Bump & Grind 2014 -...
Waze & Odyssey
Bump & Grind 20...

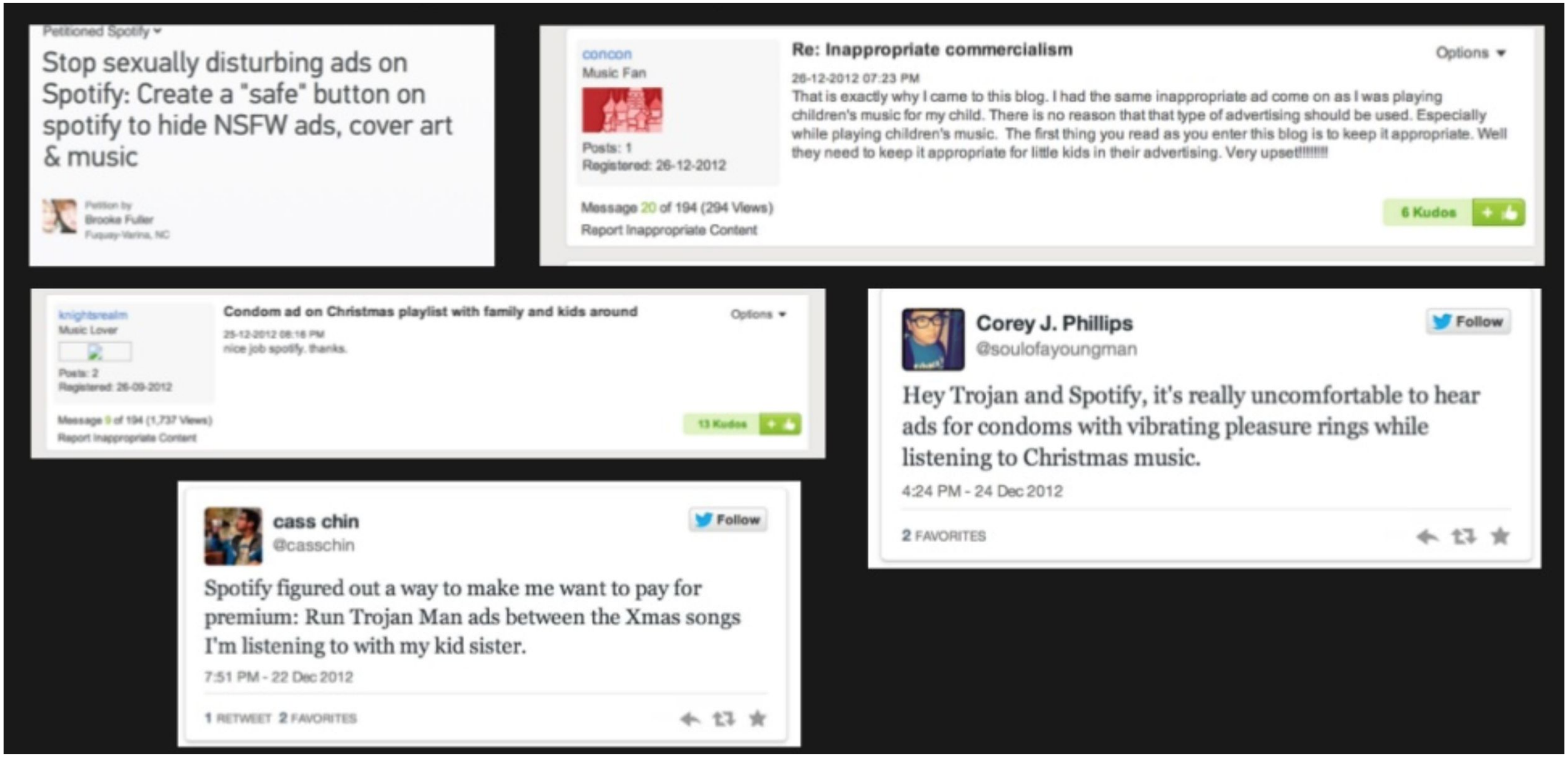
[redacted]
Bad Bitch (feat. Ty D...
Bebe Rexha
Chill Hits

[redacted]
Life Is a Beat
Maria Sadowska
Jazz na ulicach

[redacted]
Every Step Every Way

What can happen if you **stick to batch** ?

Real-Time Personalization



Stop sexually disturbing ads on Spotify: Create a "safe" button on spotify to hide NSFW ads, cover art & music

Petition by
Brooke Fuller
Fuquay-Varina, NC

concon
Music Fan
Posts: 1
Registered: 26-12-2012

Re: Inappropriate commercialism

26-12-2012 07:23 PM
That is exactly why I came to this blog. I had the same inappropriate ad come on as I was playing children's music for my child. There is no reason that that type of advertising should be used. Especially while playing children's music. The first thing you read as you enter this blog is to keep it appropriate. Well they need to keep it appropriate for little kids in their advertising. Very upset!!!!!!!

Message 20 of 194 (294 Views)
Report Inappropriate Content

knightsrealm
Music Lover
Posts: 2
Registered: 26-09-2012

Condom ad on Christmas playlist with family and kids around

25-12-2012 08:16 PM
nice job spotify. thanks.

Message 9 of 194 (1,737 Views)
Report Inappropriate Content

Corey J. Phillips
@soulofayoungman

Hey Trojan and Spotify, it's really uncomfortable to hear ads for condoms with vibrating pleasure rings while listening to Christmas music.

4:24 PM - 24 Dec 2012

2 FAVORITES

cass chin
@casschin

Spotify figured out a way to make me want to pay for premium: Run Trojan Man ads between the Xmas songs I'm listening to with my kid sister.

7:51 PM - 22 Dec 2012

1 RETWEET 2 FAVORITES

Image source:
https://www.slideshare.net/g33ktalk/spotifys-ad-targeting-infrastructure-achieving-realtime-personalization-for-24-million-users

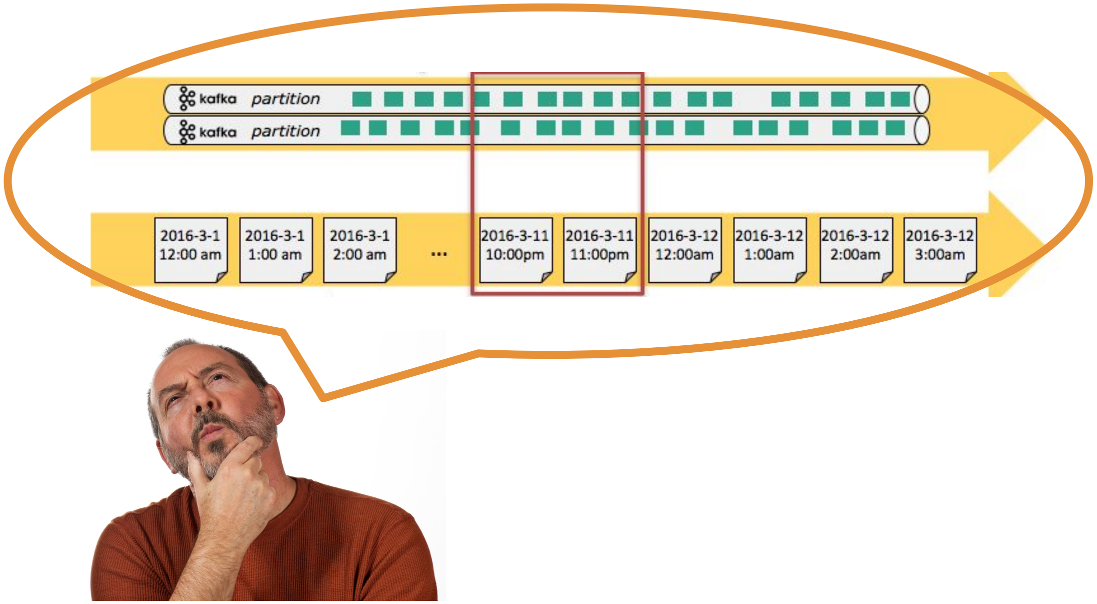
When is batch processing good?

Batch Processing Use-Cases

- **Ad-hoc analytics and data exploration**
  - Notebooks, Spark/**Flink**/Hive, Parquet
  - Complete data sets
- **Technical advantages**
  - A large swaths of historical data in HDFS
  - High-level libraries in mature batch technologies

Batch or Stream?

- **Stream is often a natural representation of data**
- **Stream processing is not only about low latency**
  - Correctness, expressive API, simplicity
- **Stream processing is a great fit for ETL, KPIs, reports**



Take-Away Message

- Stream is often a natural representation of data
- Stream processing is not only about low latency
  - Correctness, expressive API, simplicity
- Stream processing is a great fit for ETL, KPIs, reports

(when data arrives continuously)

don’t solve streaming problem with batch jobs

Thanks Vilnius !

Q&A

getindata

big data. experience. passion.