Unified Pipeline Architecture:

The Evolution of Data Processing at Spotify.

Erin Palmer
Applied Data Scientist @ Spotify

What’s Spotify, and What Data Do We Process in Creator?

Search

erinapalmer

MAIN

- Browse
- Radio

YOUR MUSIC

- Songs
- Albums
- Artists
- Stations
- Local Files

PLAYLISTS

- Piano Ballads • by Spotify
- Good Times Great Oldies • by Spotify
- Alyssa - Skating Competition Music
- Soft Rock • by Spotify
- The Ultimate Toddler Playlist • by Kidz Bop ...
- A Gift For Christmas • by Idina Menzel
- Ballroom Dance • by Laura Gray
- Afternoon Acoustic • by Spotify
- Discover Weekly • by Spotify
- Rock Ballads • by Spotify
- In The Arms Of A Woman • by Spotify
- 50 Great Female Voices • by Spotify

New Playlist

I'll Be There for You (Friends Theme
Boyce Avenue

OVERVIEW | CHARTS | GENRES & MOODS | NEW RELEASES | DISCOVER | CONCERTS

Fresh Playlists!

#ThrowbackThursday

Good things come in pairs... Come on, get happy! #SpotifyTBT

1,122,690 FOLLOWERS

This Is: Bon Iver

The best of Bon Iver, and a bit more of founder Justin Vernon's other projects.

6,010 FOLLOWERS

Rock This

These are the bands you need to know right now. Period. #RockThis / Cover: Ghost B.C.

2,093,675 FOLLOWERS

Wild Cards

All your favorite recently released emerging songs all in one place! Photo: Bishop Briggs

156,559 FOLLOWERS

See what your friends are playing

FIND FRIENDS

Julia Eger
Heartbeats – José Gonzá...
The Indie Mix

Ben Khan
Night Changes – One Directi...
Guilty Pleasures

Sean Aquilina
Flashing Lights – Kanye West
Graduation

Maddie Stocker
12m
Get Lucky – Daft Punk
French Touch Radio

Ella Child
23m
Hail To The Thief – Radioh...
Radiohead Essentials

Laura Stephens
1h
Legend – Drake
New Music Tuesday

0:41

2:18

Monthly Listeners Daily Listeners Fans

7.2M Monthly Listeners

Down 8.8k yesterday

Compare to 2 more artists

Spotify

The Beatles • 6,596 listeners now

Your Artists Help Sign Out

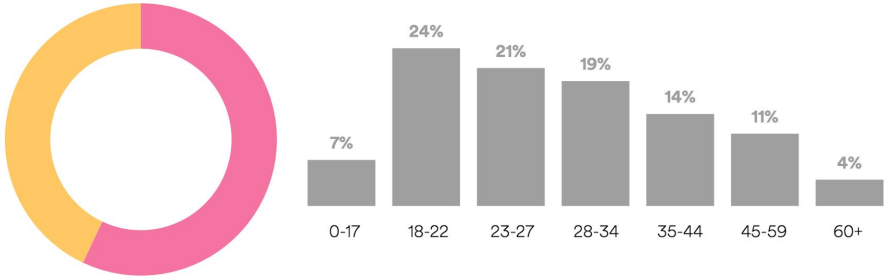
AUDIENCE SONGS PLAYLISTS

How They Listen on Spotify

Monthly Listeners Fans

Who They Are

Monthly Listeners Fans



What's Our Primary Source?

EndSong

- Log record containing:
  - trackGid
  - userId
  - Ms Played
  - Time started
  - Play Context
  - Location
  - Etc
- From it we can derive most of the useful information for listening history

Spotify

Secondary Sources

- User Meta Data
- Playlist Snapshot
- Artist/Track Meta Data
- etc

Spotify

Current Architecture

Characteristics

- Independent code paths compute necessary endpoints
- Each endpoint has 3-6 intermediate computation stages
- Intermediate Data is not reusable

Spotify

Current Architecture

Advantages

- Very Flexible
- New pipelines can be easily built
- Adaptable to constantly changing requirements

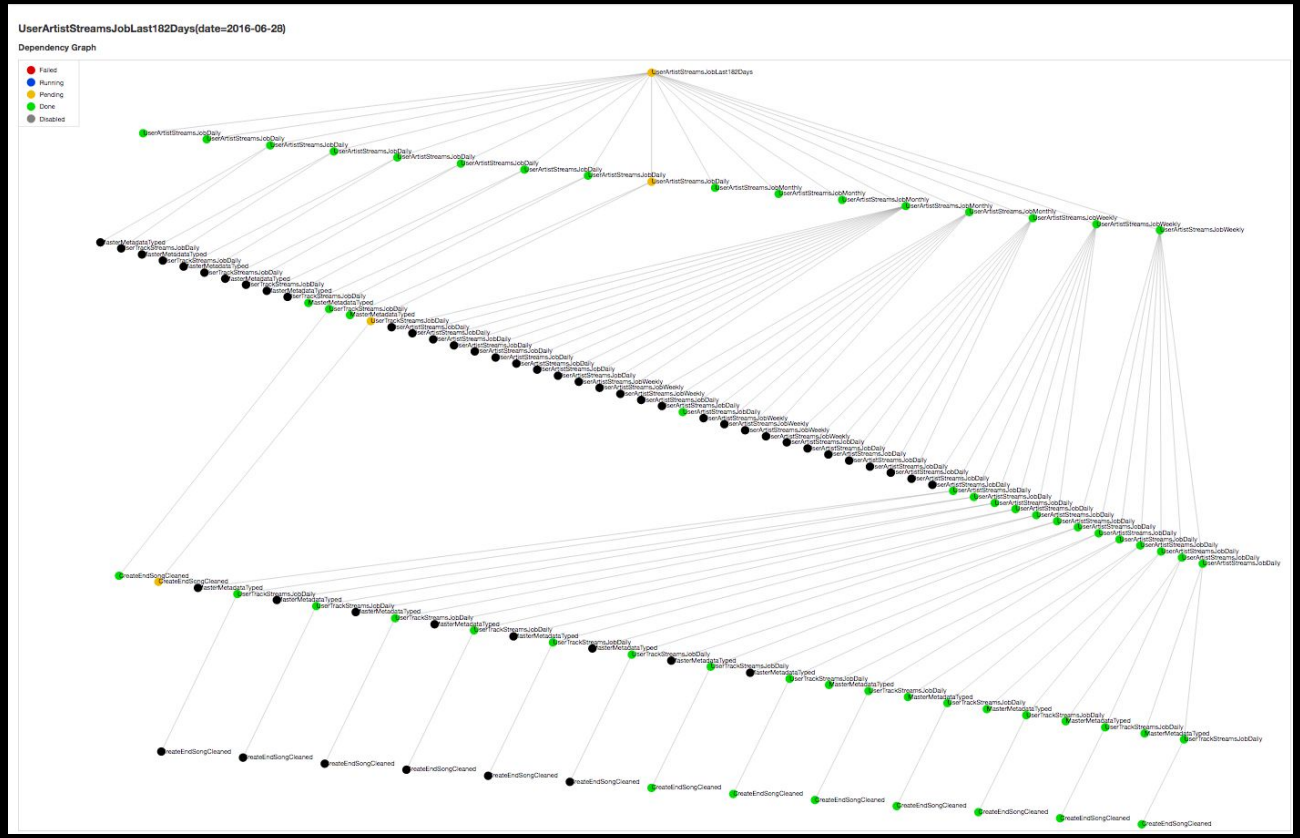
Spotify

Current Architecture

Issues

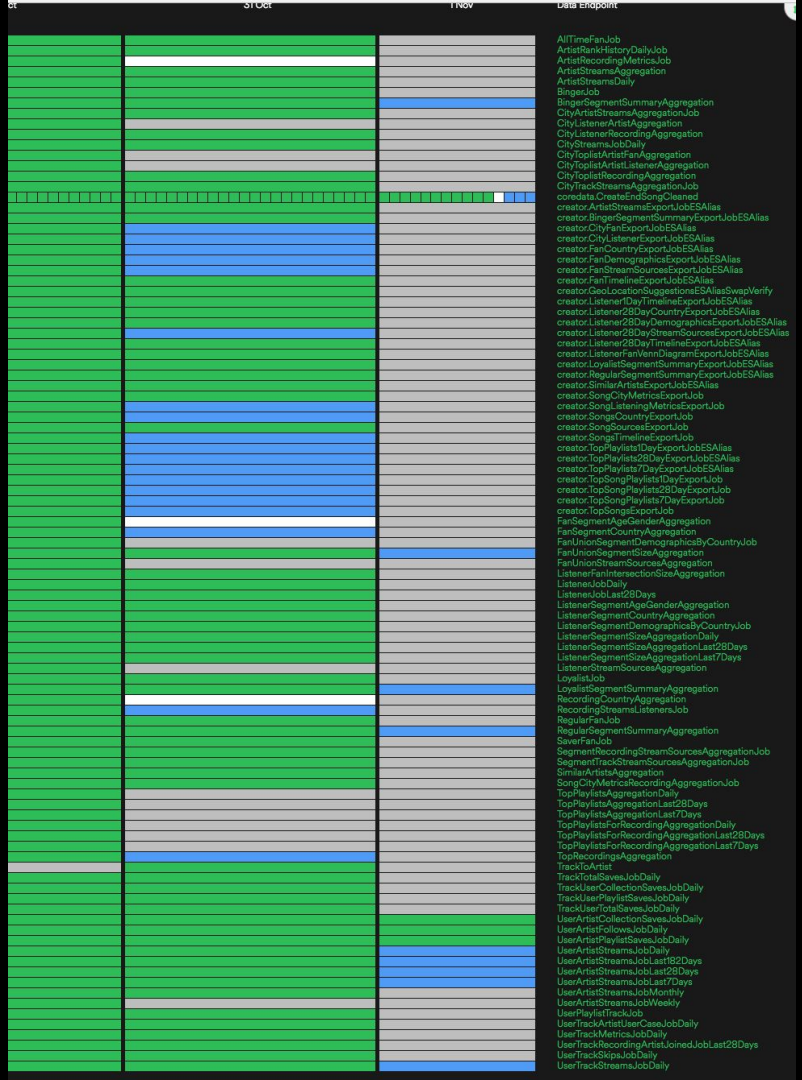
- Redundant Dependencies
- Multi-layer dependencies cause cascading effects if there are any issues/delays
- Inconsistencies

Redundant Dependencies



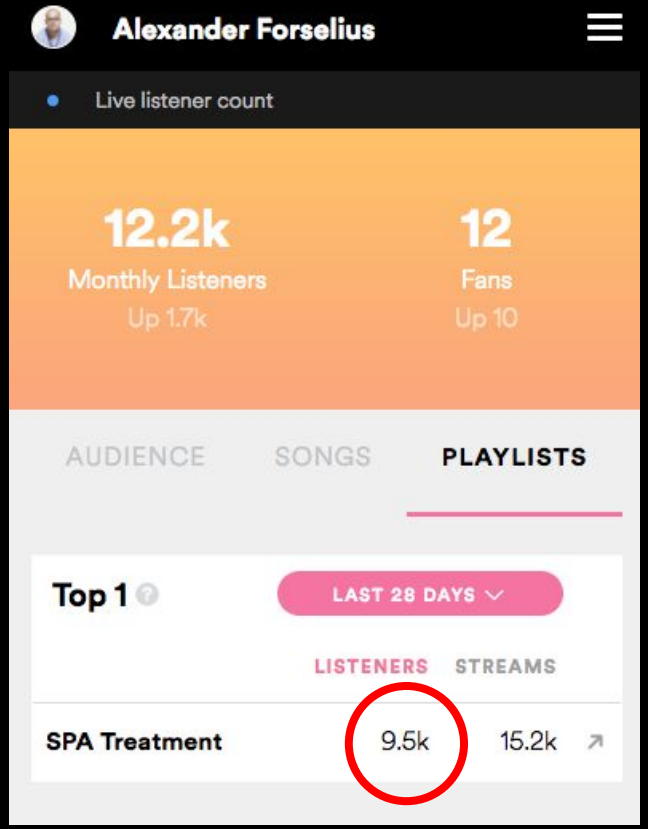
Spotify

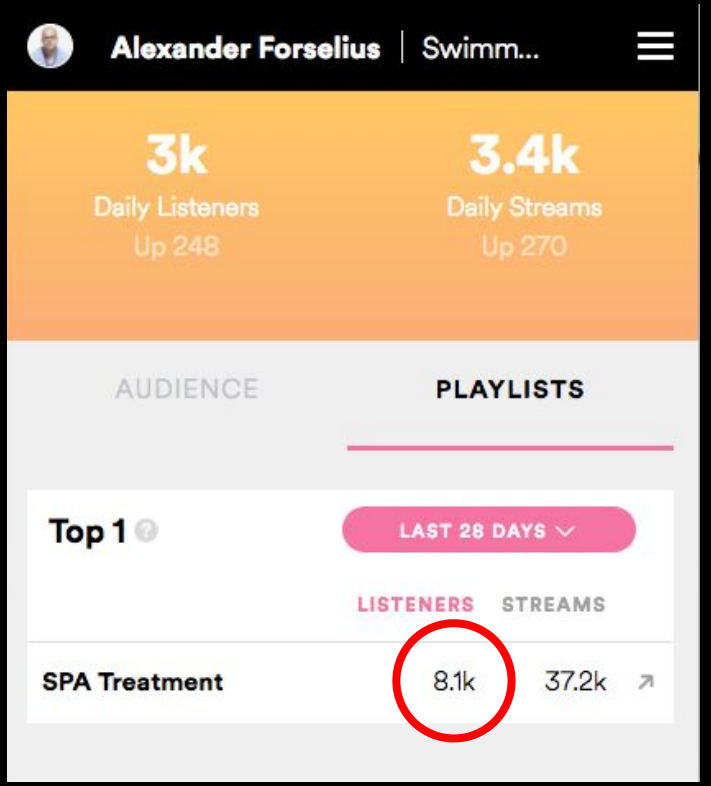
Redundant Dependencies



Spotify

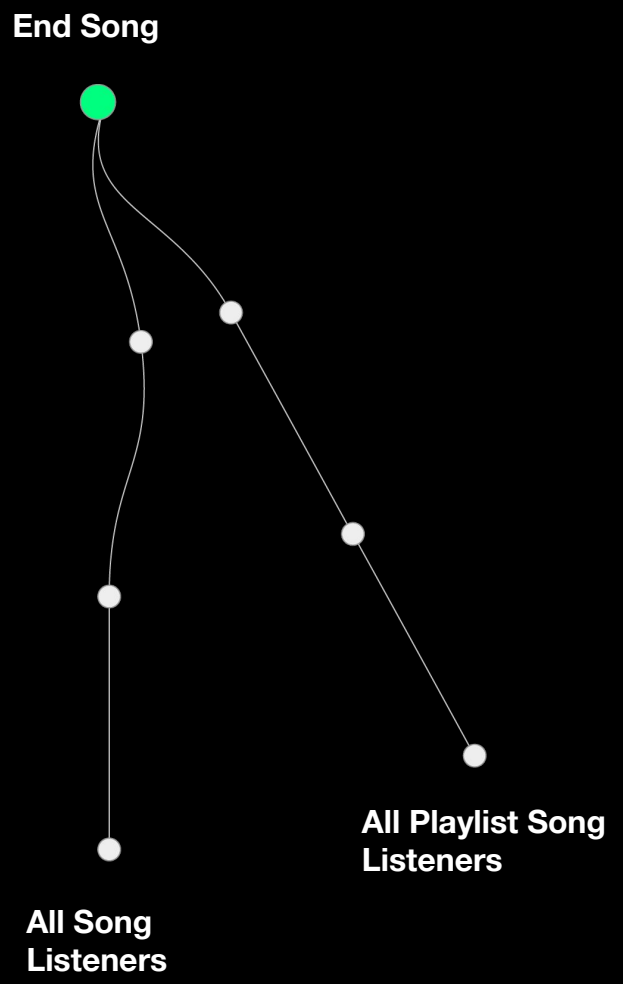
Data Inconsistencies



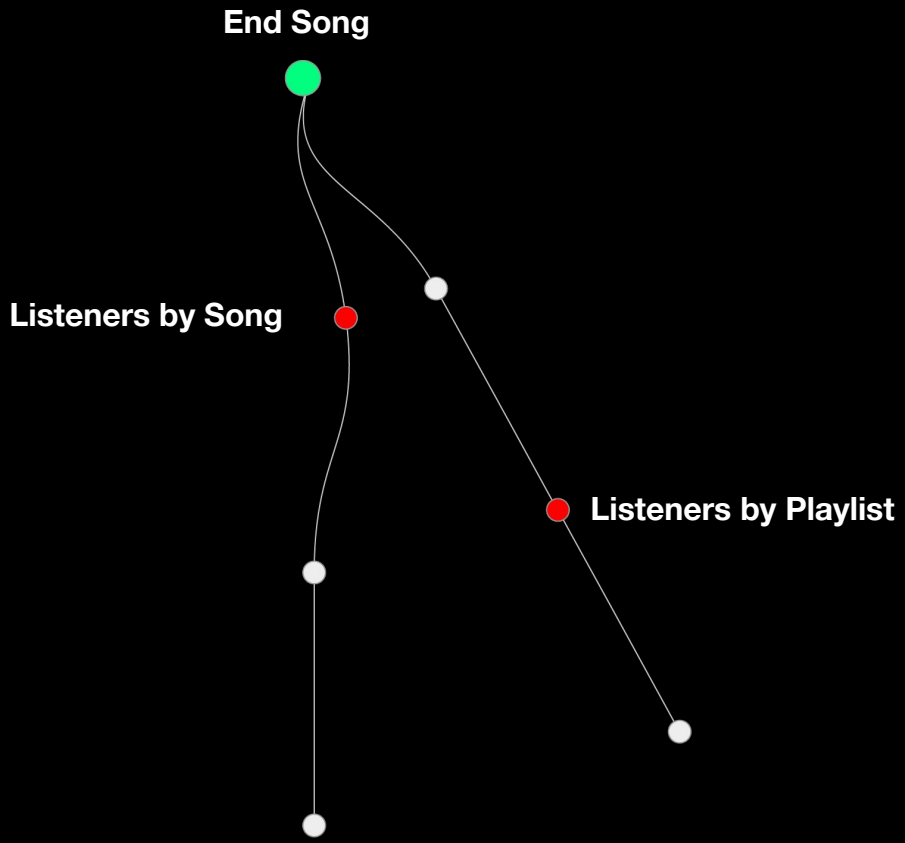


Spotify

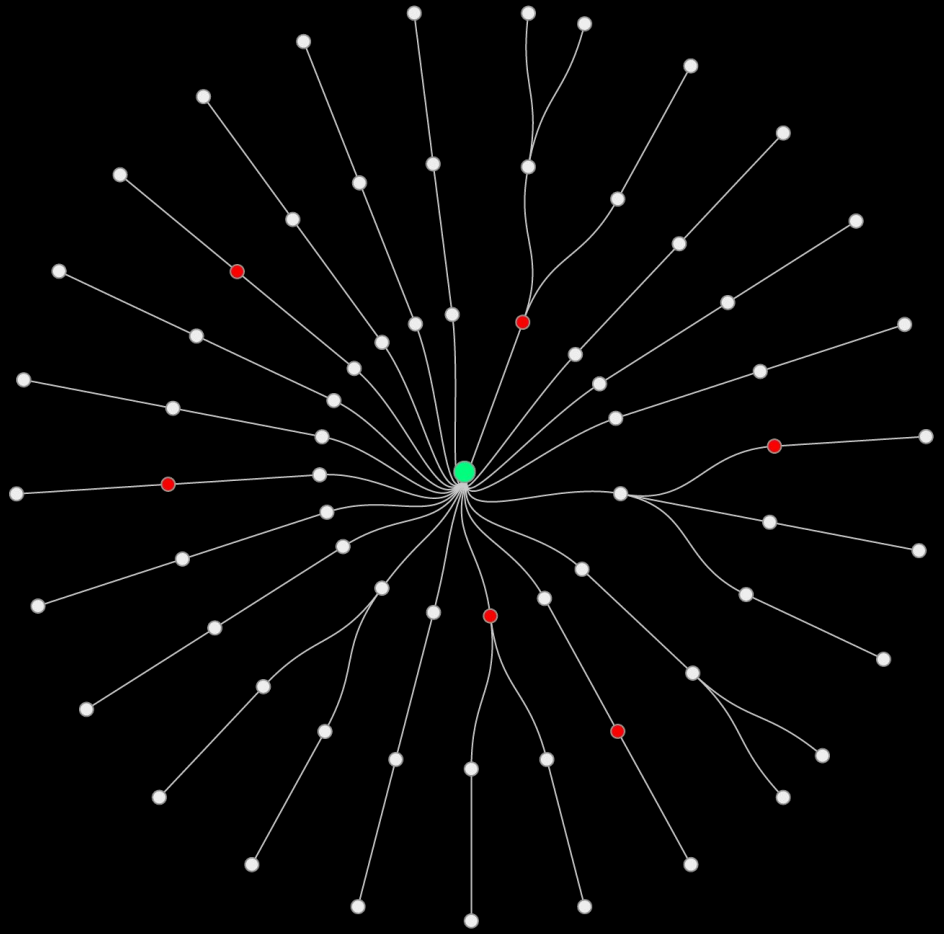
Current Architecture



Duplicative Data Sets



Duplicative Data Sets

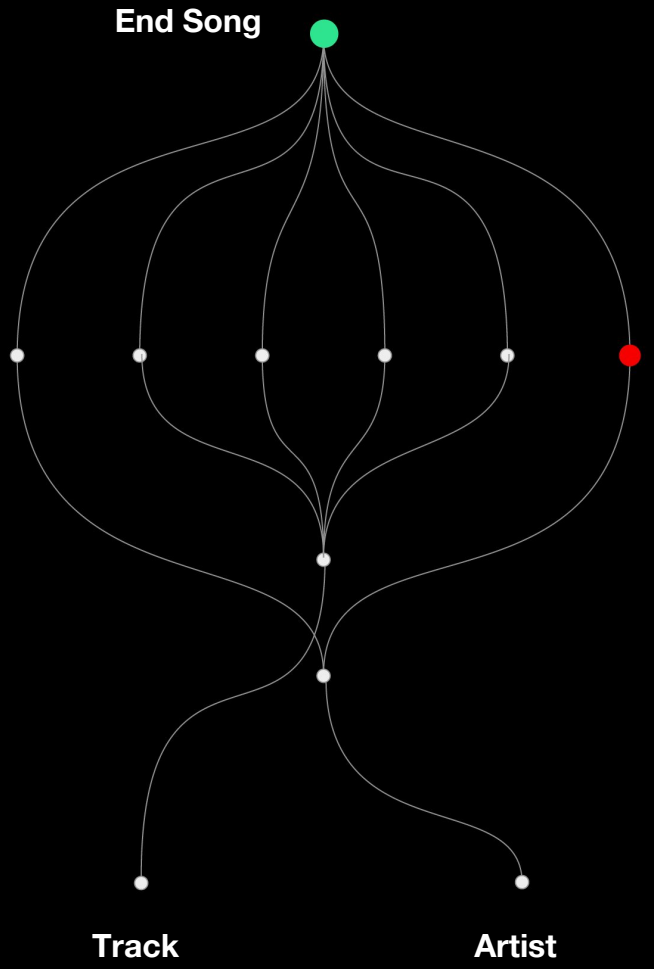


Considerations

1. Can we only read each source once?
2. How can we reduce the computation time?
3. Can we join in all the secondary sources as they comes, rather than waiting until the end of the day?
4. Retain flexibility to add new datasets or new fields

Spotify

New Architecture



Spotify

Architecture Components

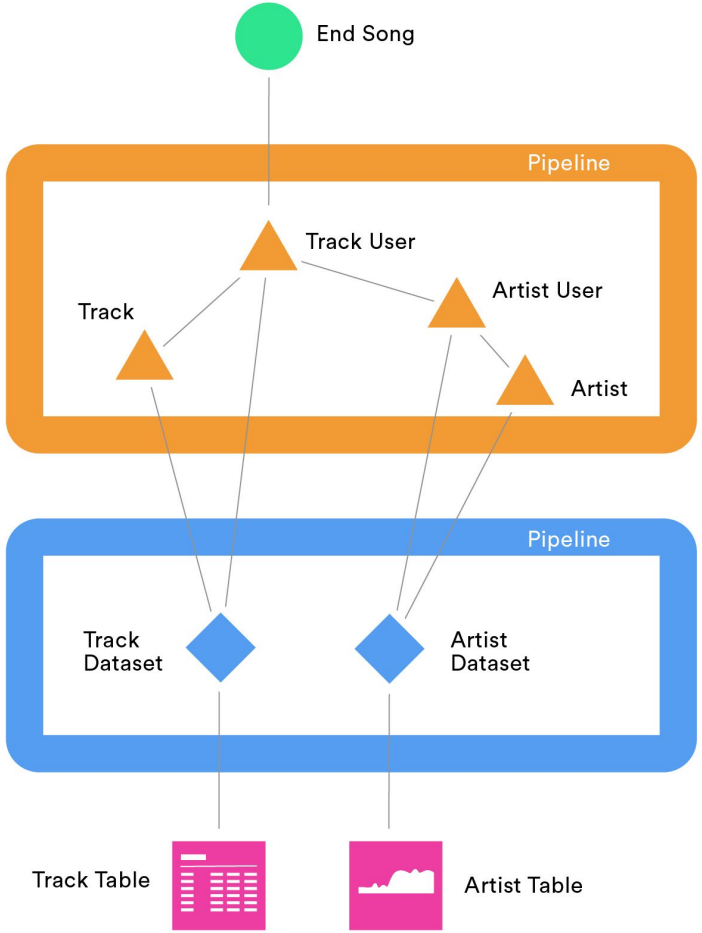
- Sources (EndSong)
- Entities
- Datasets
- Exports (Entries in a DB)

Raw Data

Entities

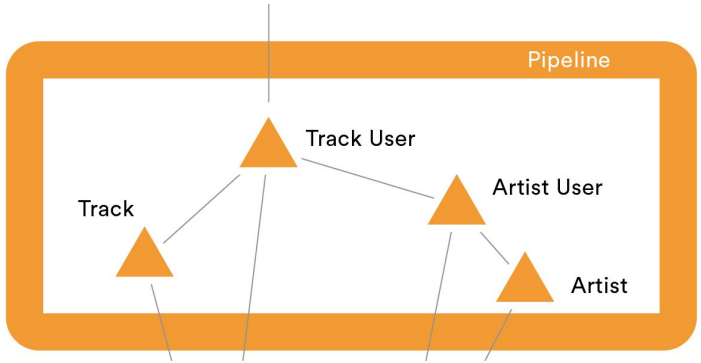
Datasets

Exports

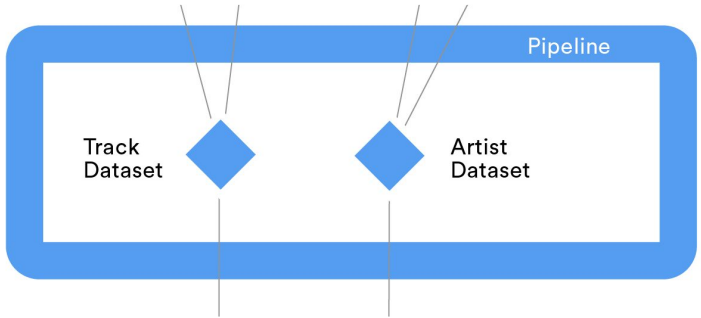


Raw Data

Entities



Datasets



Exports

Track Table

Artist Table

Spotify

Pipeline Architecture

- Google Computing Engine
  - Google Cloud Store
  - Dataflow
  - Datastore
- Framework: Scala / Scio
  - Built on top of Google’s Dataflow
- Schema: Protobufs
  - Easy Iteration
  - Built in versioning

Scio

Ecclesiastical Latin IPA: /ˈʃi.o/, [ˈʃiː.o], [ˈʃi.i̯o]

Verb: I can, know, understand, have knowledge.

Spotify

Google Dataflow with Scala => Scio

Dataflow

- Hosted, fully managed, no ops
- GCP ecosystem - BigQuery, Bigtable, Datastore, Pubsub
- Unified batch and streaming model

Scala

- High Level DSL
- Functional Programing is a natural fit for data
- Numerical Libraries: Breeze, Algebird

Spotify

Example: Word Count

```
val sc = ScioContext()
sc.textFile("shakespeare.txt")
  .flatMap { _
    .split("[^a-zA-Z']+")
    .filter(_.nonEmpty)
  }
  .countByValue
  .saveAsTextFile("wordcount.txt")
sc.close()
```

Spotify

Protocol Buffers

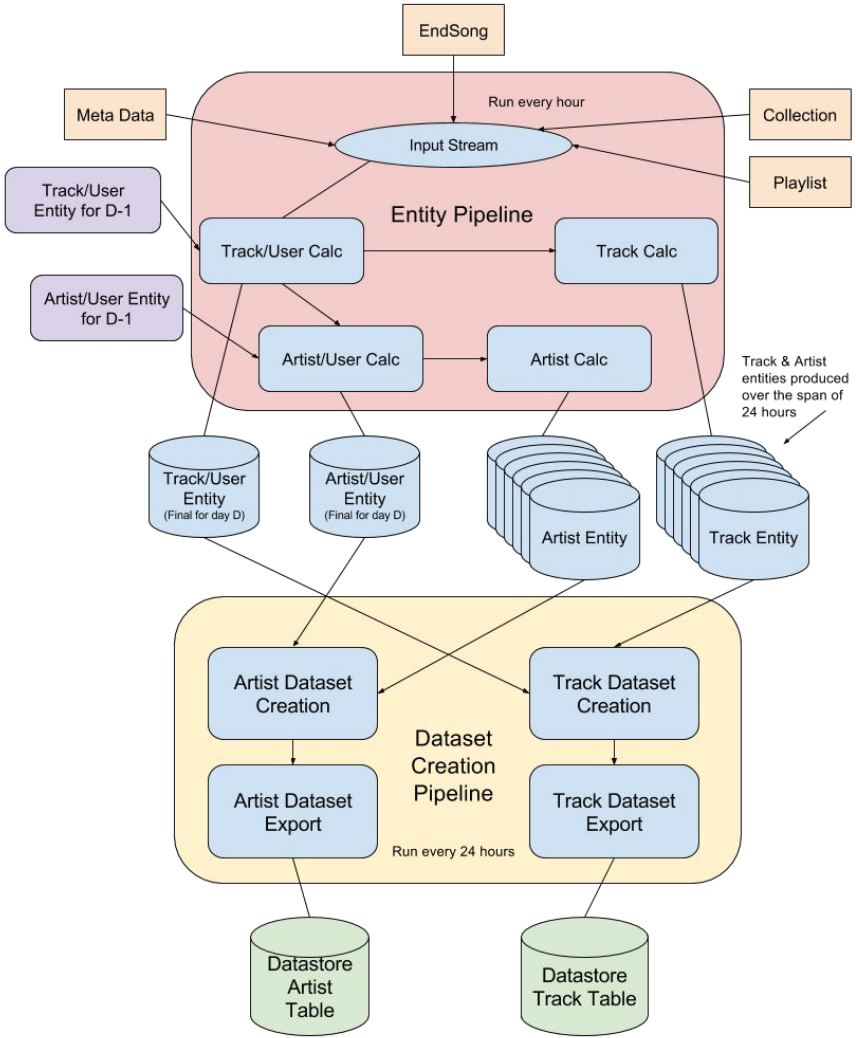
Why Protobufs?

- Land themselves to high level schema organization
- Easy to read and manipulate
- Features
  - Allows repeated fields
  - Field numerical tags allow for schema compatibility
  - Can be built to be arbitrarily large
- Protobufs are compact when stored: they are serialized into binary format

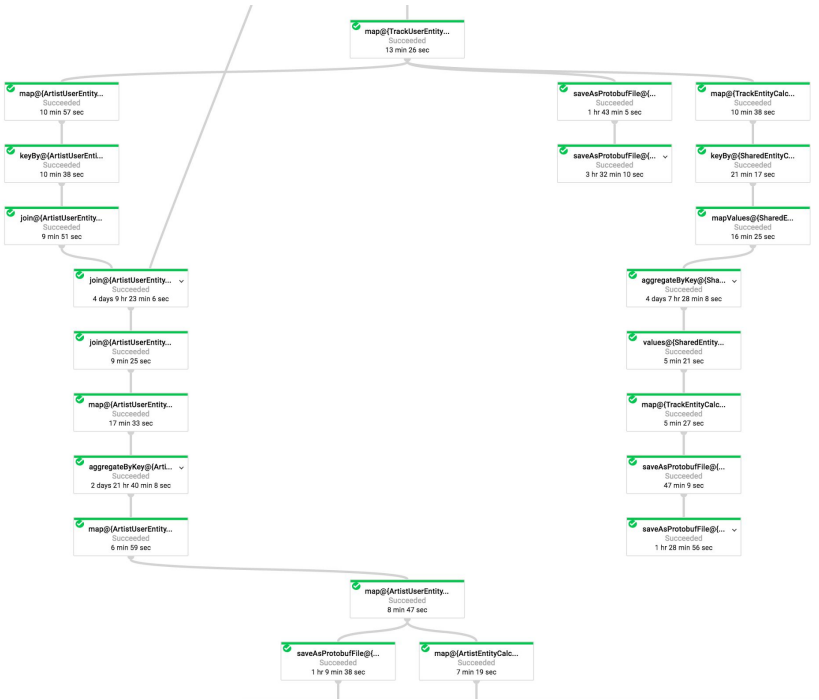
Example

```
// Next ID: 6
message AggregateKeyPB {
  enum Type {
    TRACK = 0;
    ARTIST = 1;
  }
  optional Type type = 1;
  optional string identifier = 2;
  optional string date = 3;

  // Geo and Listener type are part of the key
  optional GeographyInfoPB location = 4;
  optional ListenerTypePB listenerType = 5;
}

// Next ID: 3
message AggregateListenersByGeoPB {
  optional AggregateKeyPB key = 1;
  optional DataByTimeframePB listeners = 2;
}
```

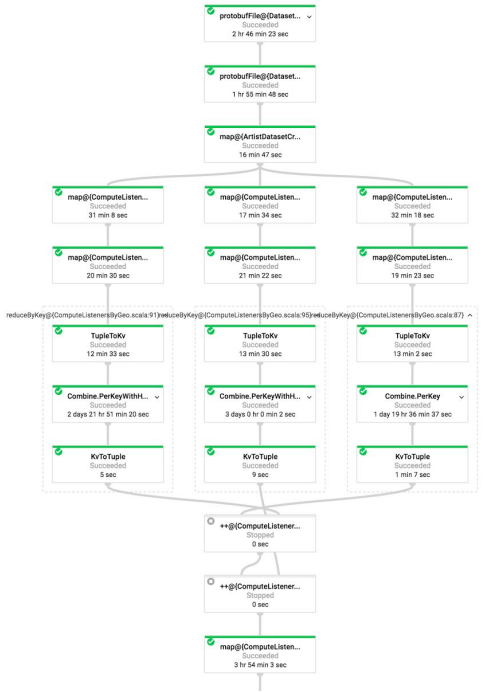
Architecture



Entities

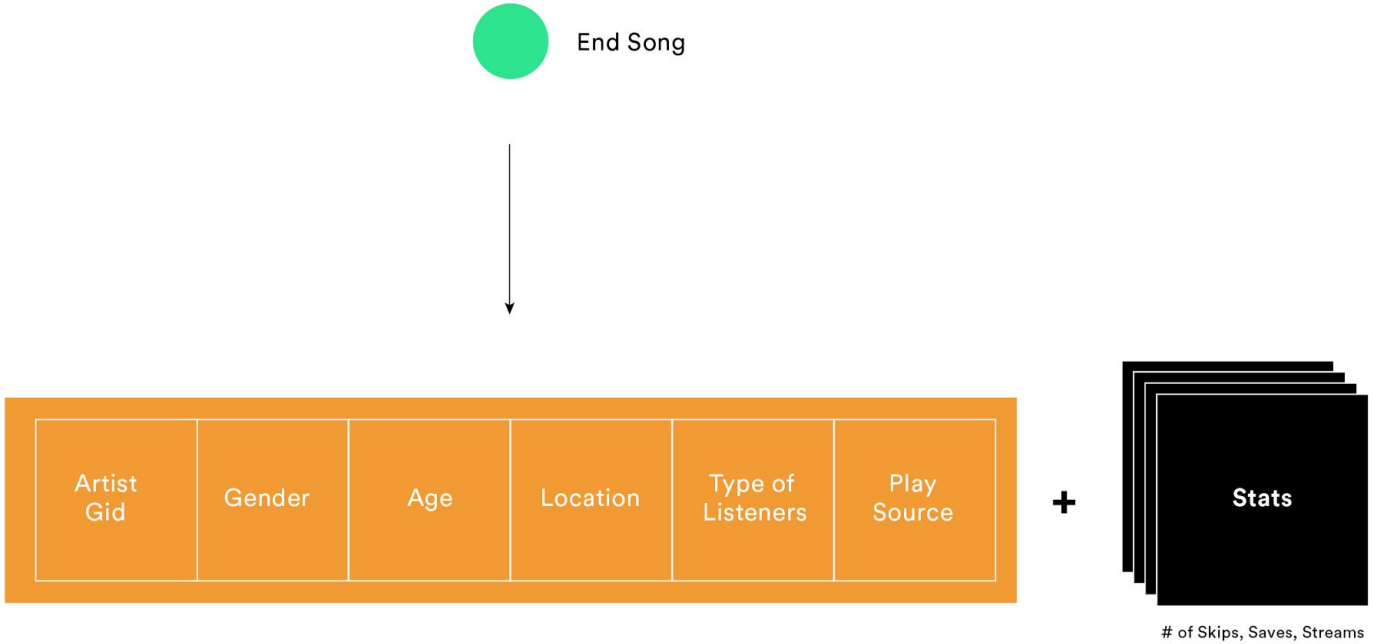


Datasets



Spotify

Entities Layer


End Song
Artist Gid
Gender
Age
Location
Type of Listeners
Play Source
+
Stats
# of Skips, Saves, Streams
Key

Gender

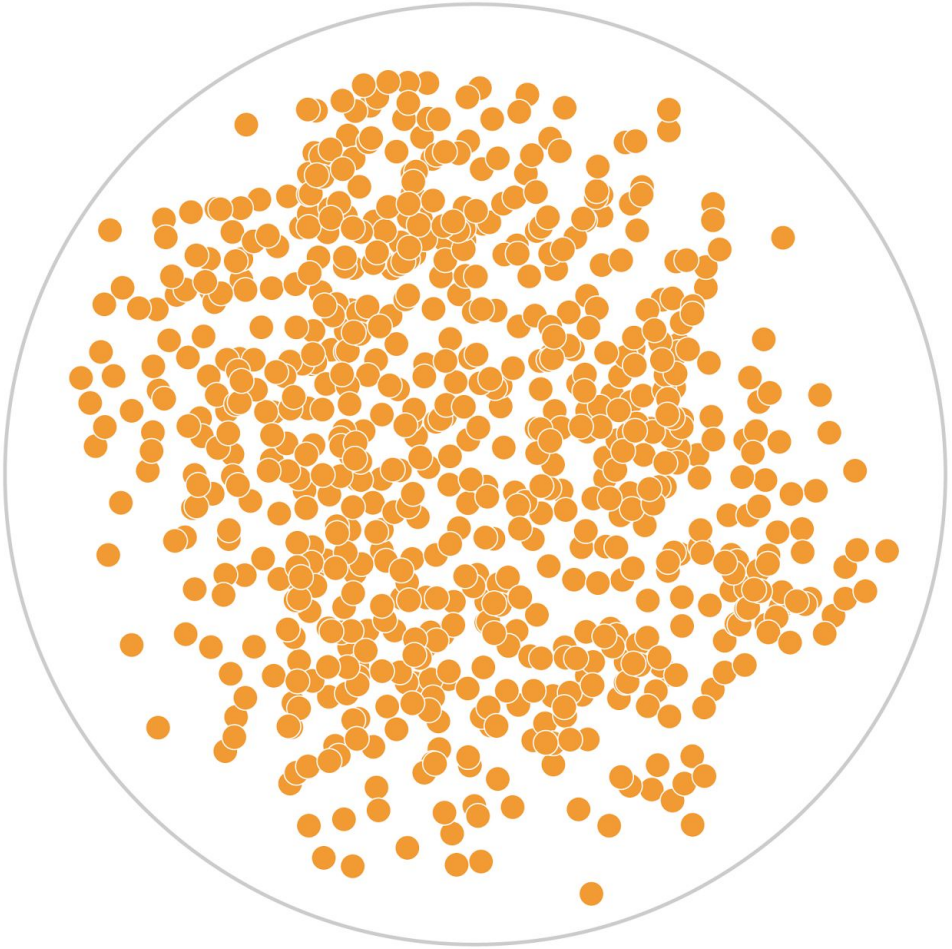
Age

Time

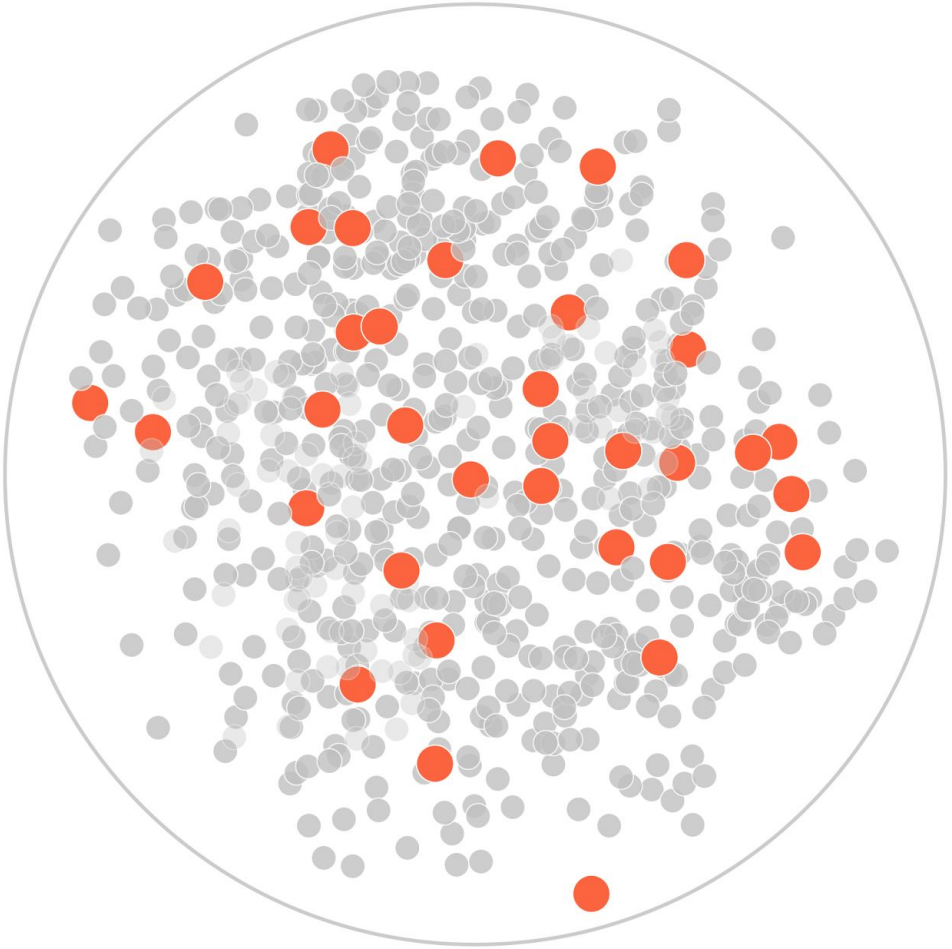
Location

Listener Type

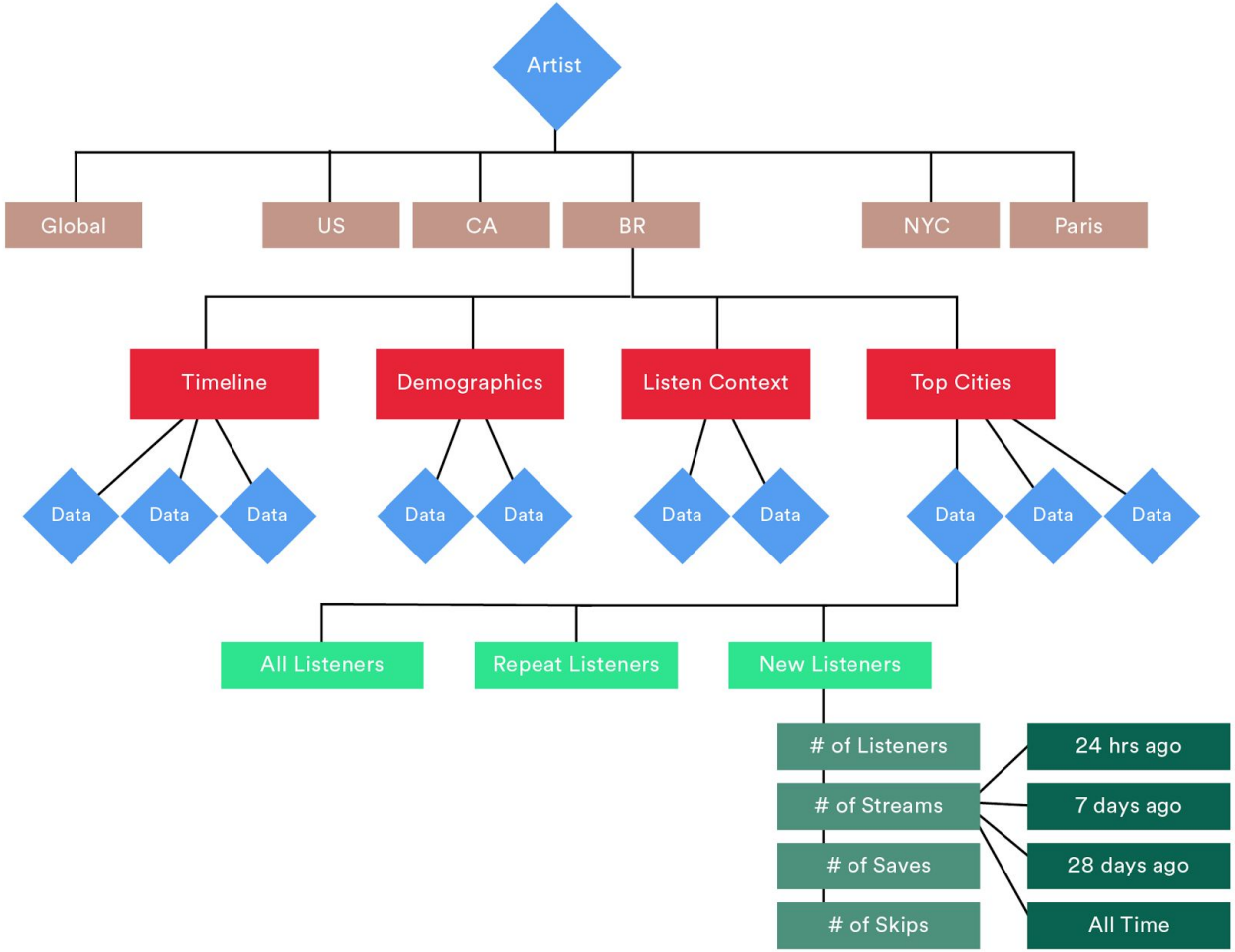
Play Source

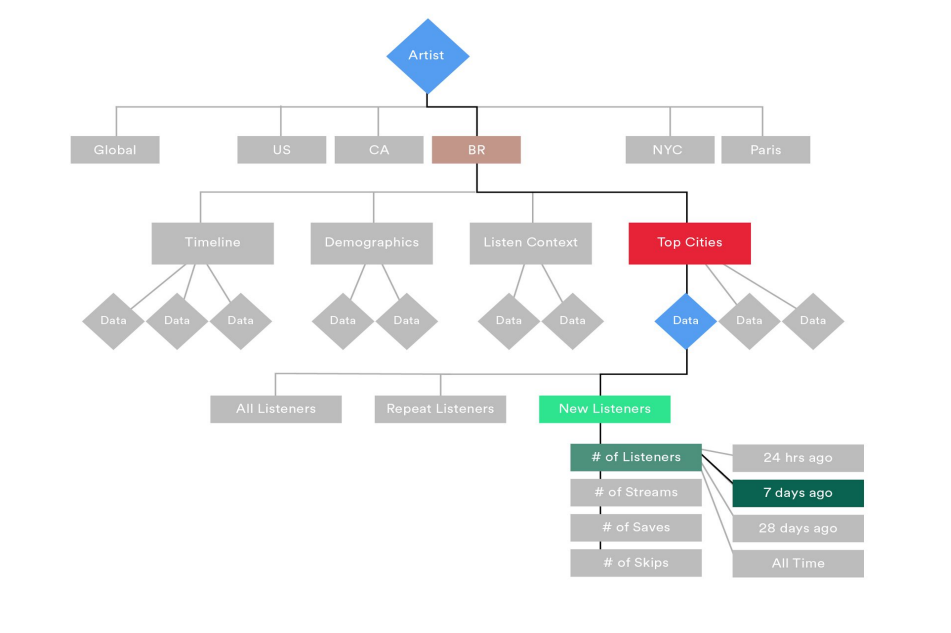


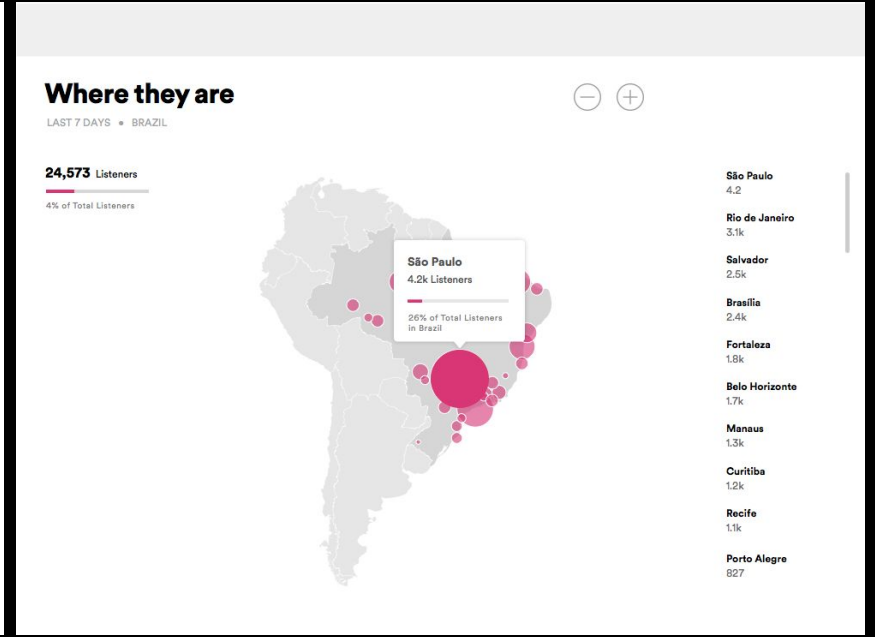
Gender	Male
Age	19-24
Time	
Location	Brazil
Listener Type	Fans
Play Source	

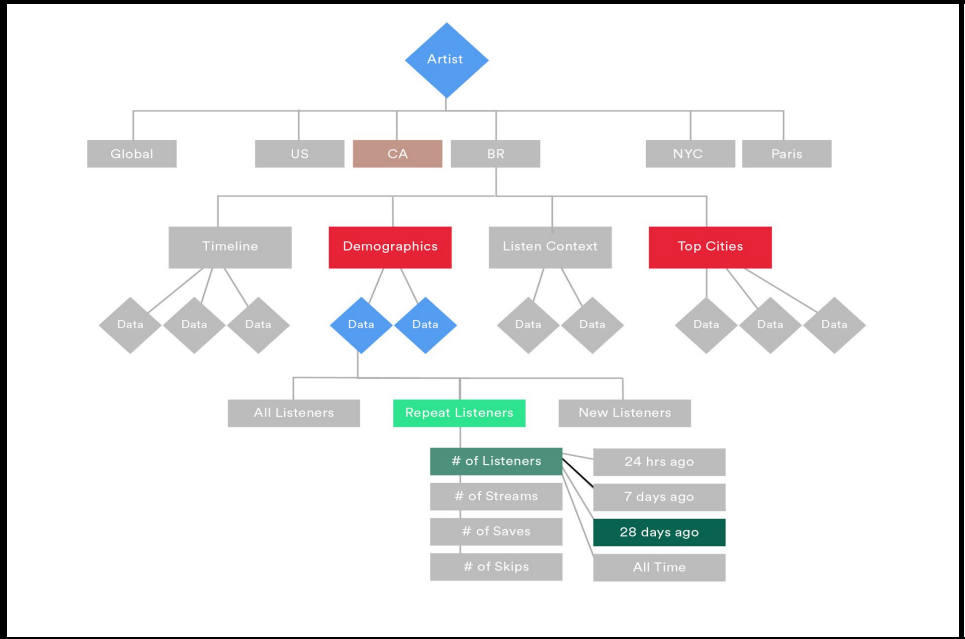


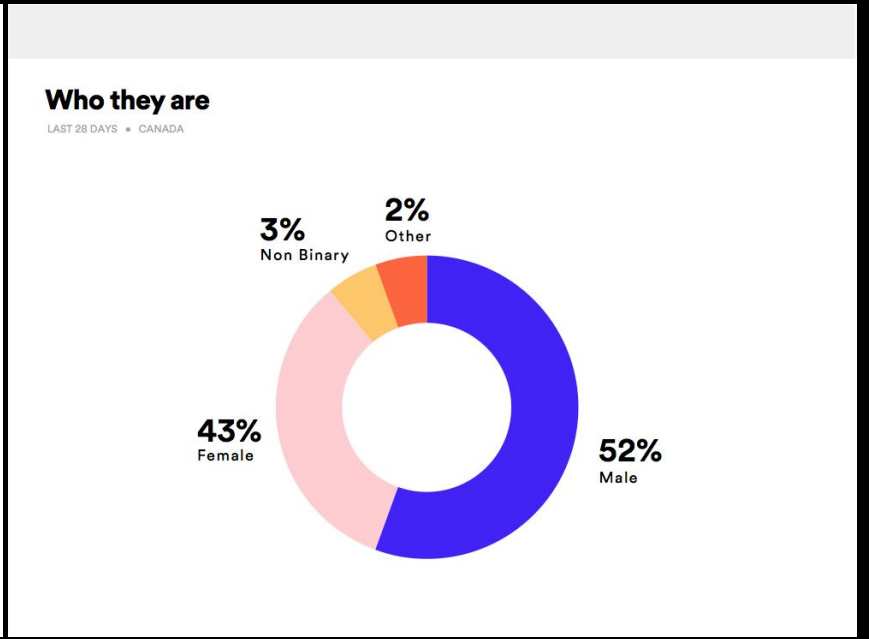
Dataset Layer


Artist
Global
US
CA
BR
NYC
Paris
Timeline
Demographics
Listen Context
Top Cities
Data
Data
Data
Data
Data
Data
Data
Data
Data
All Listeners
Repeat Listeners
New Listeners
# of Listeners
# of Streams
# of Saves
# of Skips
24 hrs ago
7 days ago
28 days ago
All Time


Artist
Global
US
CA
BR
NYC
Paris
Timeline
Demographics
Listen Context
Top Cities
Data
Data
Data
Data
Data
Data
Data
Data
Data
Data
All Listeners
Repeat Listeners
New Listeners
# of Listeners
# of Streams
# of Saves
# of Skips
24 hrs ago
7 days ago
28 days ago
All Time


Where they are
LAST 7 DAYS • BRAZIL
24,573 Listeners
4% of Total Listeners
São Paulo
4.2k Listeners
26% of Total Listeners in Brazil
São Paulo
4.2
Rio de Janeiro
3.1k
Salvador
2.5k
Brasília
2.4k
Fortaleza
1.8k
Belo Horizonte
1.7k
Manaus
1.3k
Curitiba
1.2k
Recife
1.1k
Porto Alegre
827

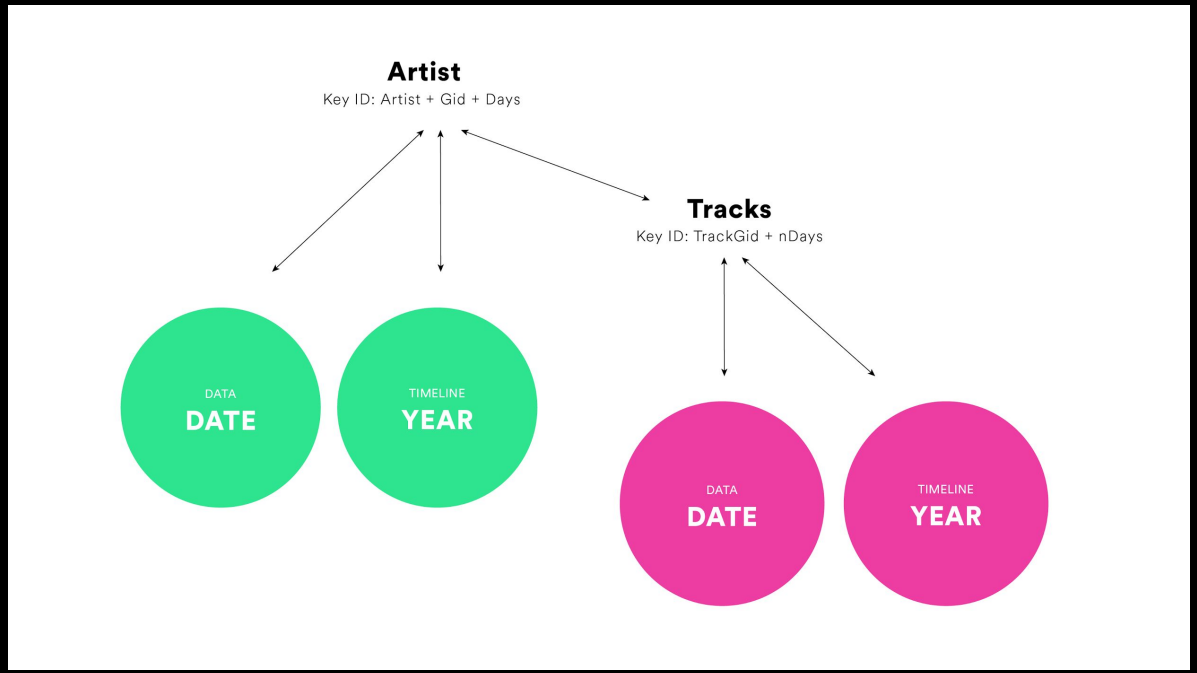

Artist
Global
US
CA
BR
NYC
Paris
Timeline
Demographics
Listen Context
Top Cities
Data
Data
Data
Data
Data
Data
Data
Data
Data
Data
All Listeners
Repeat Listeners
New Listeners
# of Listeners
# of Streams
# of Saves
# of Skips
24 hrs ago
7 days ago
28 days ago
All Time


Who they are
LAST 28 DAYS • CANADA
3%
Non Binary
2%
Other
43%
Female
52%
Male

Export Layer

Spotify

Datastore Storage Model



Spotify

Conclusions

- Unified Pipeline Architecture allows you to parallelize your processing while consolidating the logic
  - No redundant computations
- Minimize Number of joins
- Read in sources only once and decorate them once for further processing
- Think About the Future!

Spotify

Questions?

- Come to my office hours!
- My colleague: Deepti Deshpande will also be there to help answer questions