Delivering Music Recommendations to Millions

Sriram Malladi

Rohan Singh (@rohansingh)

A little bit about us

Over 24 million active users

55 countries, 4 data centers

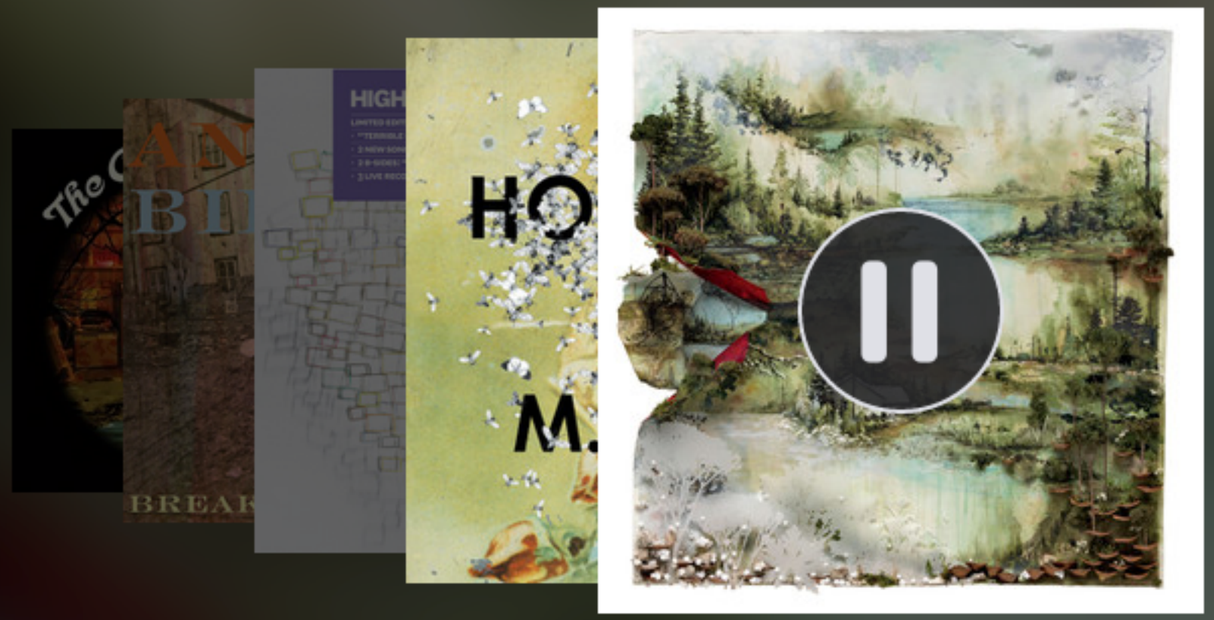
20+ million tracks

Discovering music — then

Artist Radio based on

The Tallest Man On Earth

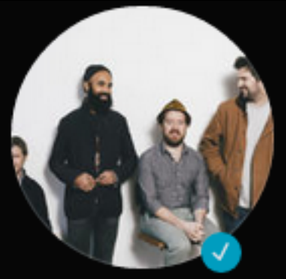
CREATE NEW STATION



Hinnom, TX

Bon Iver

YOUR STATIONS



ARTIST · VIEW PROFILE

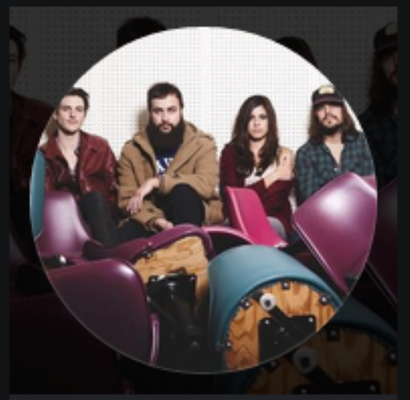
The Cave Singers

PLAY FOLLOWING ···

FOLLOWERS
11,529

OVERVIEW **RELATED ARTISTS** BIOGRAPHY

Jessica Lea Mayfield



Houndmouth

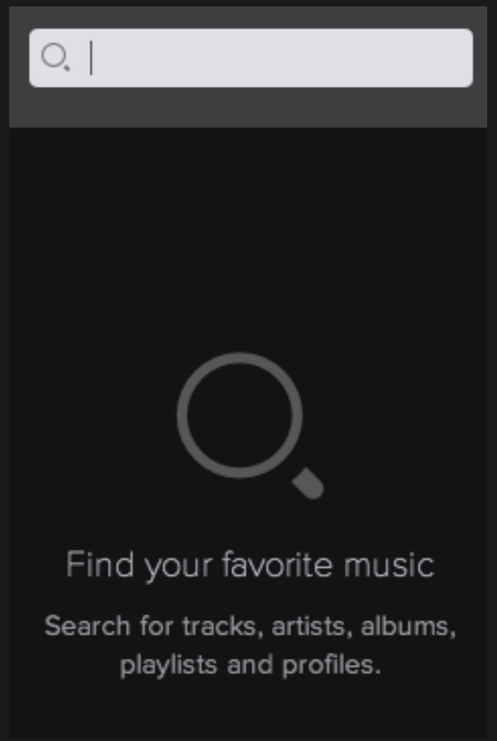
Jenny O.

Frontier Ruckus

Those Darlins

Langhorne Slim

What do you want to listen to?



The right music for every moment

The Plan

1. Collect lots of data
2. Generate personalized recommendations
3. Serve those to millions of listeners each day

Data, data, data

Track plays

Radio feedback

Playlists

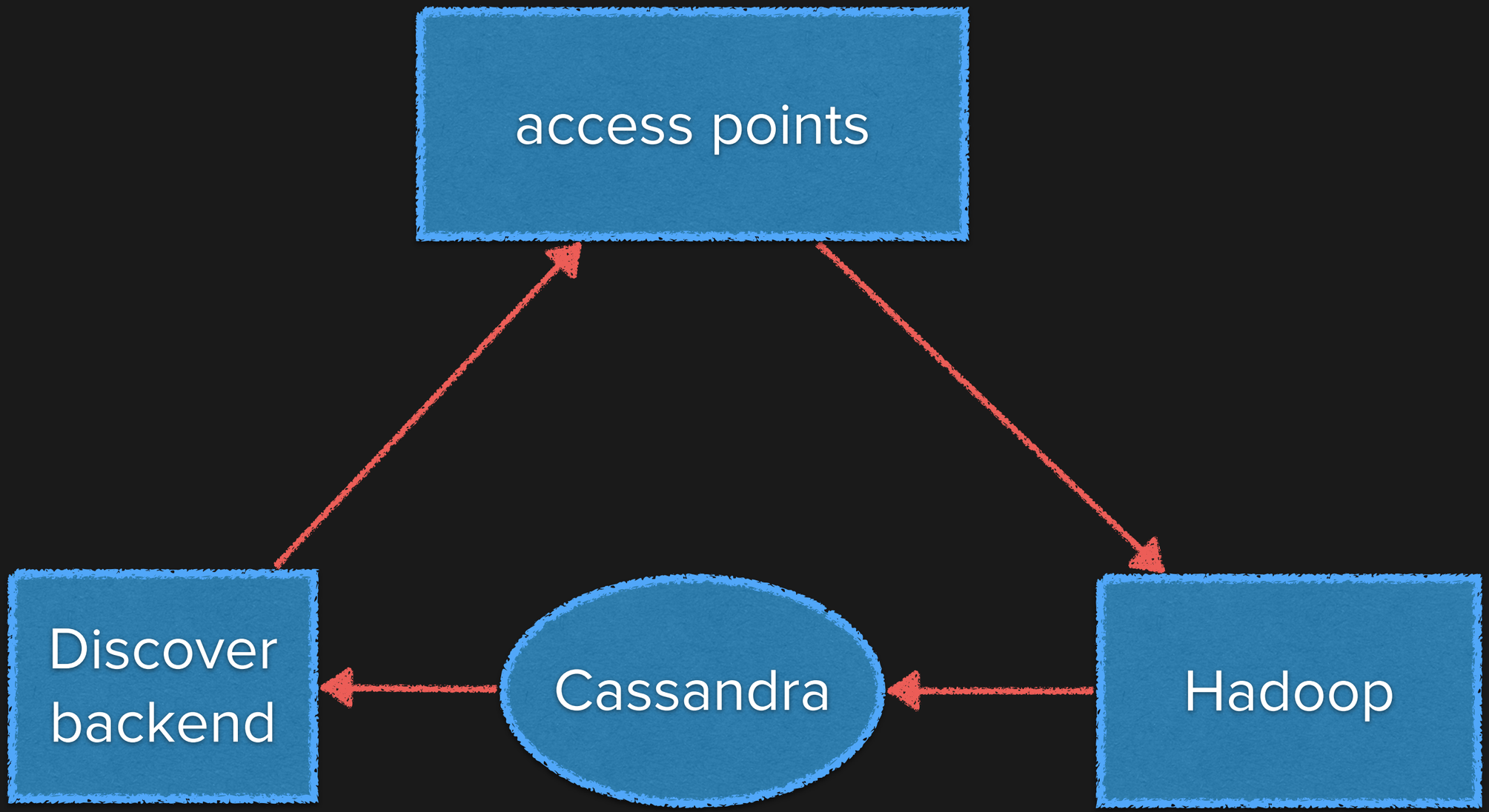
Follows

Collecting it

All data into and out of Spotify goes through **access points**

Access points and services log data

Logs are aggregated and shipped to Hadoop


access points
Discover backend
Cassandra
Hadoop

Hadoop

Framework to store and process big, distributed data

Hadoop Distributed File System (HDFS)

Hadoop MapReduce

hadoop.apache.org

Crunch the data

What you listen to

Artists you follow

What **similar users** listen to

Your buddy listened to Daft Punk all day

An artist you follow released a new album

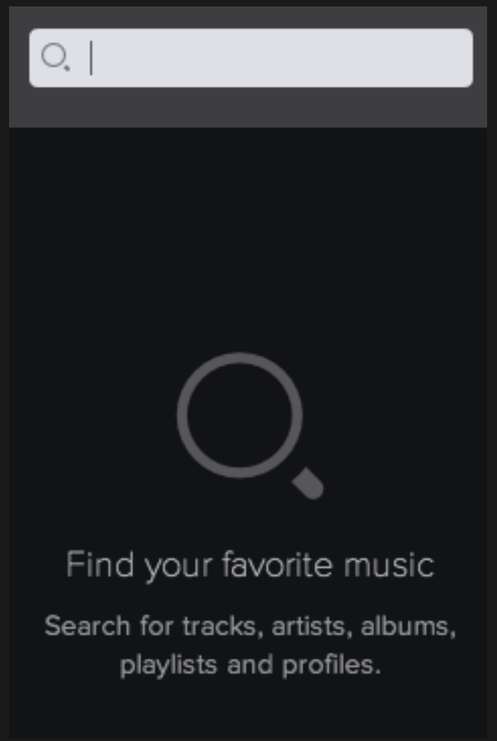
A friend just created a new playlist

Region

Age group

Gender

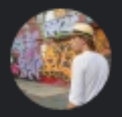
What do you want to listen to?



You listened to **Angus Stone**. You might like this song.

Friends Make Garbage (Good Friends Take It Out)

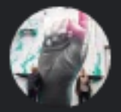
Low Roar



Blixt shared a song by **Purity Ring**.

A DAY AGO

"I'm really enjoying this song right now."



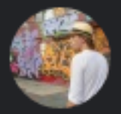
Chris Johnson has been listening to a lot of **Cheatahs** this week.

A DAY AGO

Cheatahs

Cheatahs

0 ▶ 1 Like



Blixt shared a song by **OVERWERK**.

2 DAYS AGO

People who listen to **Sufjan Stevens** are also listening to **Andrew Bird**.

Andrew Bird

99,750 FOLLOWERS

Jonathan McKay has been listening to a lot of **Metric** this week.

15 HOURS AGO

You listened to **Deorro** yesterday. Want to try **Julian Jordan**?

Julian Jordan

2,427 FOLLOWERS

People who listen to **AWOLNATION** are also listening to **Two Door Cinema Club**.

Roll it out

Storing recommendations

About 80 kilobytes of data per user

Regenerated daily for all active users

> 1 TB of recommendations overall

Storing recommendations

Terabyte a day is a fair amount of data to write and index
(+ replicas)

Tradeoff between storing data or
looking things up on the fly

Apache Cassandra

Highly available, distributed, scalable

Fast writes, decent reads

We have one cluster per data center

Transferring worldwide

Recommendations all generated in Hadoop, in London

Need to be shipped to our data centers worldwide

So much Internet weather

hdfs2cass

Internal tool to copy data from Hadoop to Cassandra

Creates table from data in HDFS

Loads tables into Cassandra using a bulk loader

github.com/spotify/hdfs2cass

Serve it up

1. Aggregate data from all our data sources
2. Decorate it with metadata
3. Shuffle!

Serving at scale…

Discover is Spotify's home page
500+ requests per second

Thousands of requests to other services

Database calls for each unique user

...or not?

Naive, first-stage prototype:

10 to 20 seconds per request

(doesn't really scale)

Fail a lot

Make it webscale!

Find & rewrite slow sections

Switch to C++ from Python for critical code

More improvements

Pregenerate more data

Cache aggressively

Throw hardware at it

Switch to SSD's

Scale horizontally

Takeaways

You will need hardware —
big data can still be expensive

Less data can be better

Prototype, iterate, optimize—
fail early but improve

Questions?