DJ Byg Data

Diego
Trajche
Tribhuvanesh

Continuing...

- Goal: Is it possible to determine the genre of a song based solely on its lyrics?

- K-Means Clustering

- Scaling up
  - > 237,000 tracks
  - > 20000 artists
  - 5000 unique (stemmed) words
  - 5-20 clusters

The Big Data Challenge

- Problem 1: Preprocessing
  Process and clean 280GB of data

- Problem 2: Clustering
  280k songs to cluster, vectors of dimension 5000

- Problem 3: Visualization and Analysis
  9M pairs, 280k songs, 20.5k artists

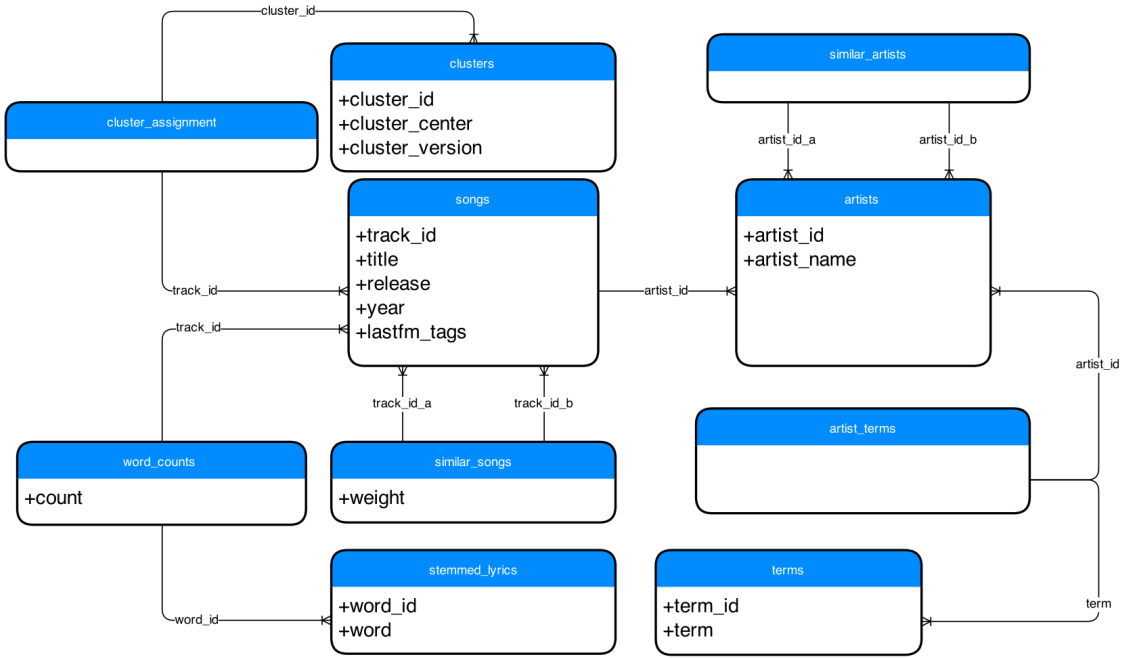
The Big Data Challenge

- Problem 1: Preprocessing
  Solution: S3 and EMR Hadoop MapReduce

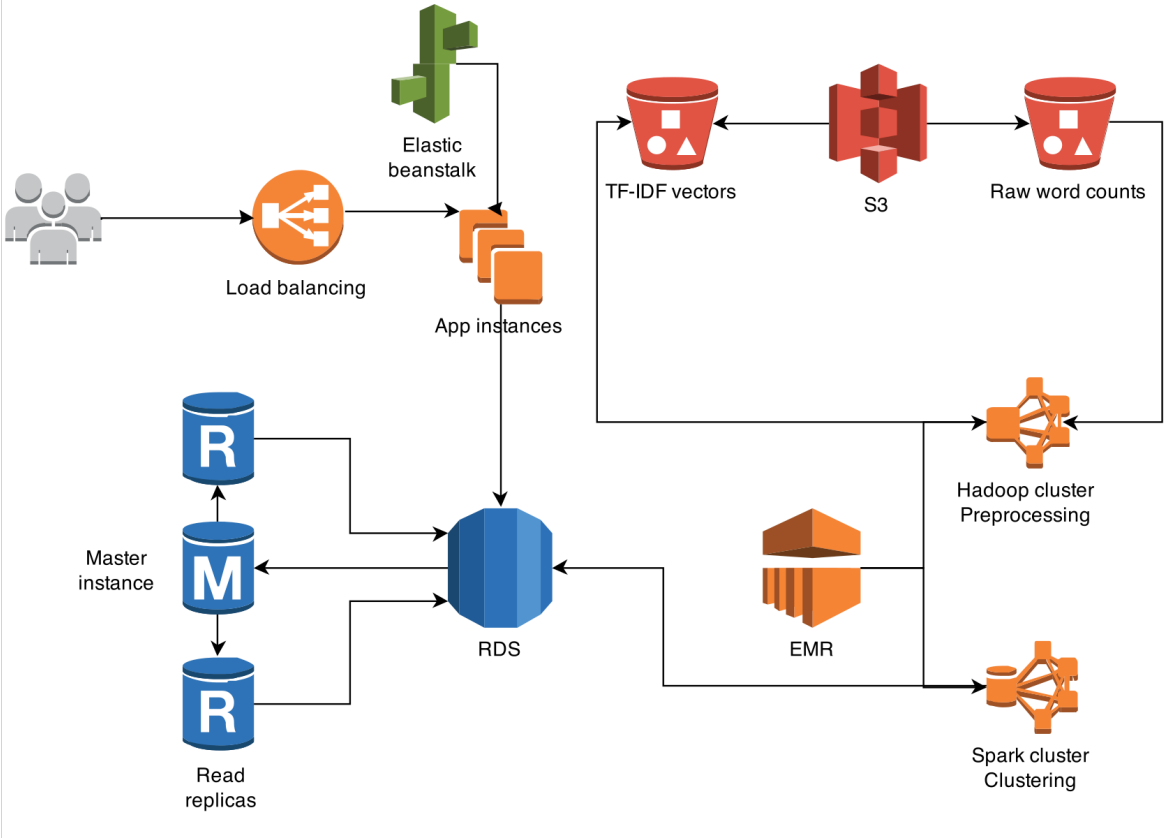
- Problem 2: Clustering
  Solution: Spark on EMR

- Problem 3: Visualization and Analysis
  Solution: Relational database, Replicated, Scalable web cluster

Data Model - Relational



Architecture



Quality Metrics

- Clustering parameters
  - Number of clusters, i.e. k
  - Number of top words removed when calculating the clusters
- Using the similarity information from the MSD
  - Similar artists - Clusters should group similar artists together
  - Similar songs - Clusters should preserve similar songs graph
- Using the fact that artists are expected to be contained in a single cluster
- Using the distribution of songs among the clusters

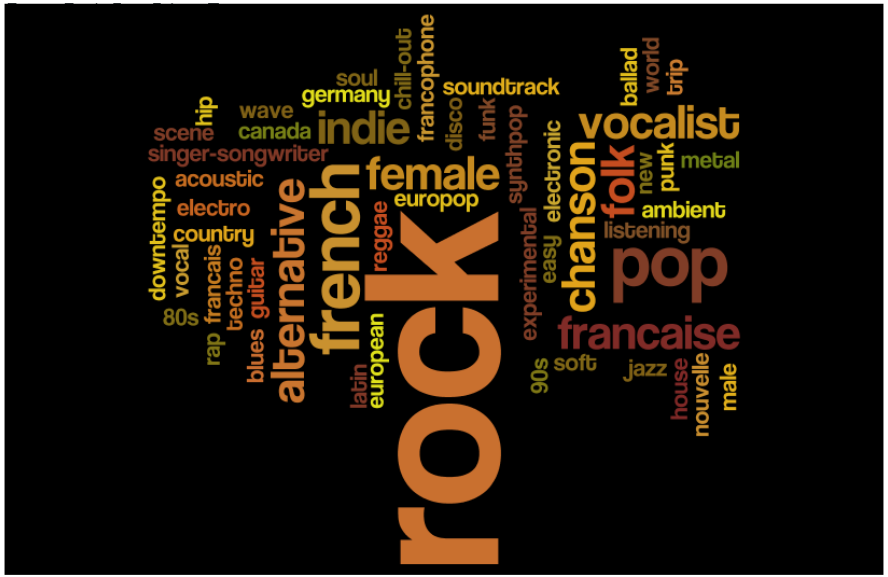
Results


rock
german
pop
metal
alternative
gothic
vocalist
wave
folk
rap
indie
electronic
ambient
female
80s
ebm
experimental
house
germany
deutschsprachig
chanson
jazz
reggae
emo
techno
acoustic
soul
trance
disco
hardcore
european
hip electro
europop
male
heavy
punk
vocal
downtempo
singer-songwriter
industrial
black
hard
chill-out
classic
ballad
new
dark
artist
soft
synthpop
lyrics

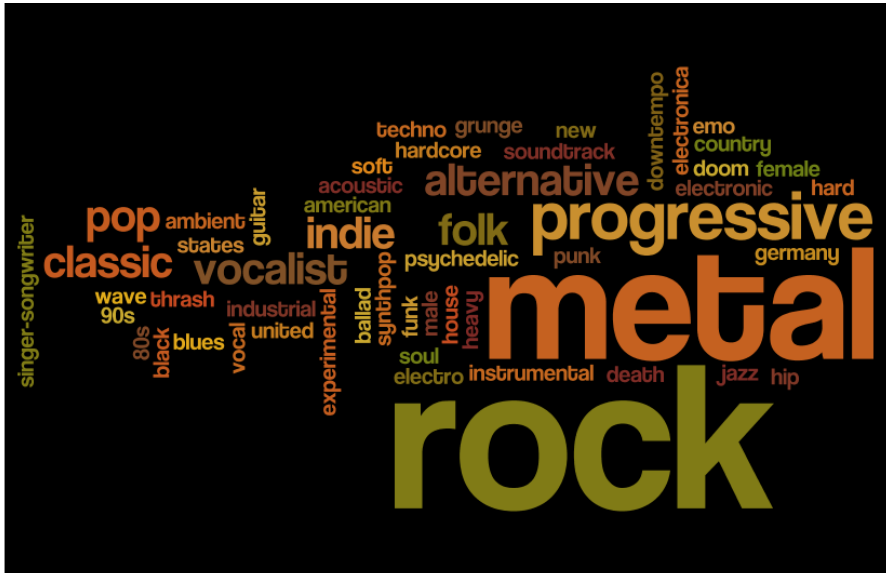

Udo
Bosse
Naidoo
Annett
Cicero
Wahl
Juliane
Stund
Söhne
Mantus
Louisan
Puhdys
Jürgens
Freiheit
Al-Deen
Christina
Pur
Obducta
Mutabor
Münchener
Silbermond
Grönemeyer
Reinhard
Fantastischen
Toe
Murphy
Roger
Lindenberg
2raumwohnung
Mey
Kampfhörspiele
Adversus
Japanische
Westernhagen
Tocotronic
Onkelz
Revolverheld
Laith
Ohrbooten
Cephalgy
PeterLicht
Tac
Stürmer
Rio
Fehlfarben
Spider
Dannen
Falco
Gang
Rainhard
Mannheims
LaFee
Instanz
Fendrich
Herbert
Reiser
Tic
Erdball
J.B.O
Klee
Dritte
Vier
Massive
Noctе
Stillste
Van
Werding
Xavier
Staubkind
Funny
Selig
Setlur
Planlos
Sabrina
Letzte
Nena
Tomte
Töne
Melotron
Prinzen
Schweisser
Rammstein
Böhse
Welle

Results


JulienDaho
Piaf
Biolay
Louise
Solaar
Miossec
Natasha
Paradis
Kétanou
Brel
Benjamin
Fabian
Vanessa
Attaque
Clerc
Obispo
Goldman
Jean-Jacques
Jacques
Pascal
La
Souchon
Étienne
Higelin
Bruel
Hysteria
Lara
MC
Les
Barback
Bénabar
Mass
De
St-Pier
Alain
Patrick
Édith
Ogres
Rue


soul
germany
chill-out
francophone
soundtrack
ballad
world
trip
hip
wave
indie
disco
funk
vocalist
scene
canada
singer-songwriter
female
synthpop
folk
metal
acoustic
europop
electronic
new
punk
downtempo
electro
chanson
ambient
country
reggae
experimental
easy
listening
vocal
francais
techno
guitar
alternative
french
pop
80s
rap
blues
rock
latin
european
francaise
90s
soft
jazz
house
nouvelle
male

Results


techno
grunge
new
emo
hardcore
soundtrack
country
downtempo
electronica
soft
alternative
doom
female
acoustic
electronic
hard
american
pop
ambient
guitar
indie
folk
progressive
singer-songwriter
states
classic
vocalist
psychedelic
punk
germany
wave
thrash
90s
industrial
ballad
synthpop
funk
male
house
heavy
80s
black
blues
vocal
united
experimental
metal
soul
electro
instrumental
death
jazz
hip
rock


Old
Pig
Furs
Petra
Baez
Tankard
Truth
Man's
Cradle
Mercyful
Oil
Mercury
Oyster
Voivod
Destroyer
Psychedelic
Carnage
Cephalic
Factory
Blue
Joan
Converge
Slayer
Moonspell
Borknagar
Carcass
Angel
Maniacs
GWAR
Midnight
Dying
Funeral
Morbid
Cattle
Venom
Harum
Fate
Krisiun
Procol
Filth
Child
CMX

Demo