# Trabalho Final
## Processamento de Dados Massivos em Nuvem

**Danilo Pimentel** [1] **– 2016058077**
**Gabriel Bastos** [1] **– 2016058204**

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

## 1. Introdução

Dados gerados a partir da atividade de usuários em plataformas de música são um dos aspectos mais valiosos para este tipo de negócio. Tais dados constituem informações sobre as tendências e preferências de seus usuários, e portanto são o objeto principal de estudo para tais plataformas.

Além disso, tais dados são insumo para funcionalidades de recomendação, um dos diferenciais mais relevantes das plataformas modernas. Sistemas de recomendação não são uma novidade, tal que há uma considerável diversidade na literatura de pesquisa. Neste trabalho, buscamos aplicar metodologias já bem estabelecidas em uma base de dados dos hábitos musicais de usuários reais.

## 2. Base de Dados

A base de dados adotada para estudo provém da plataforma Last.fm[1], e contém os hábitos musicais de seus usuários no período de 2002 até 2009. Tais dados estão dispostos em duas tabelas:

1. Reproduções: id de usuário; timestamp; id e nome do artista; id e nome da faixa.
2. Usuários: id de usuário; gênero; idade; país; data de inscrição.

Em particular, a base apresenta 19.150.868 registros de reprodução de 992 usuários distintos. Tal volume de dados se mostra suficiente para a aplicação de metodologias de aprendizado de máquina em sistemas de recomendação, como por exemplo filtragem colaborativa[2].

## 3. Objetivos

Definimos objetivos em dois principais aspectos. O primeiro deles visa uma análise geral, com o propósito de identificar a disposição dos dados. Tal objetivo é importante para identificar características desbalanceadas, valores inesperados, e obter uma intimidade geral com a base de dados. O segundo aspecto visa a elaboração de funcionalidades de recomendação para os usuários, baseada em tendências presentes nos dados.

Tais objetivos foram implementados sobre a plataforma Spark, de forma a explorar os conhecimentos adquiridos na disciplina. Baseado nas demais experiências obtidas no curso, acreditamos que o Spark é uma plataforma que nos empodera a realizar tais análises sem dificuldades adicionais.

### 3.1. Análise geral

Propomos os seguintes tópicos de análise geral:

1. Em qual faixa etária se encontram a maioria dos usuários? Elaboramos um histograma de idade dos usuários para entender melhor o público alvo do Last.fm.
2. Quais foram as faixas mais populares entre os usuários no mundo todo? Produzimos um ranking de popularidade das faixas.
3. No mundo todo, quais são os artistas mais populares entre os usuários? Construímos um ranking de popularidade dos artistas na plataforma.
4. Quem são os *heavy users* do Last.fm? Elaboramos um ranking de atividade dos usuários do sistema.
5. Onde estão os usuários do Last.fm? Analisamos a distribuição de localização dos usuários da plataforma no mundo.
6. Onde ocorrem a maioria das reproduções de faixas? Construimos um mapa de calor de reproduções por país.
7. Quais foram os *hits* do momento? Identificamos as faixas que tiveram os maiores surtos de reproduções no mundo.

### 3.2. Sistema de recomendação

Propomos a construção de um sistema de recomendação baseado em filtragem colaborativa. O sistema deve considerar as interações dos usuários como implícitas, visto que não há informação explícita sobre preferência na base de dados. A interação implícita se dá pela reprodução de uma faixa.

A filtragem colaborativa é uma técnica oportuna para este conjunto de dados, uma vez que não são disponibilizadas características das faixas em si. Sem informações inerentes às faixas, como o gênero musical, não acreditamos que a utilização de algoritmos baseados em proximidade da vizinhança seja eficaz.

### 3.3. Sugestões adicionais

Devido à limitação do tempo disponível para o trabalho, desconsideramos a sugestão de implementação do sistema de recomendação por regras de associação. Além disso, consideramos que tal metodologia seria de certa forma redundante com a nossa proposta baseada em filtragem colaborativa, e portanto não traria experiências adicionais no tocante à conceitos estudados na disciplina. Acreditamos que o sistema de recomendação proposto neste trabalho é suficiente para o propósito estipulado.

## 4. Resultados

### 4.1. Faixa Etária

Em qual faixa etária se encontram a maioria dos usuários?

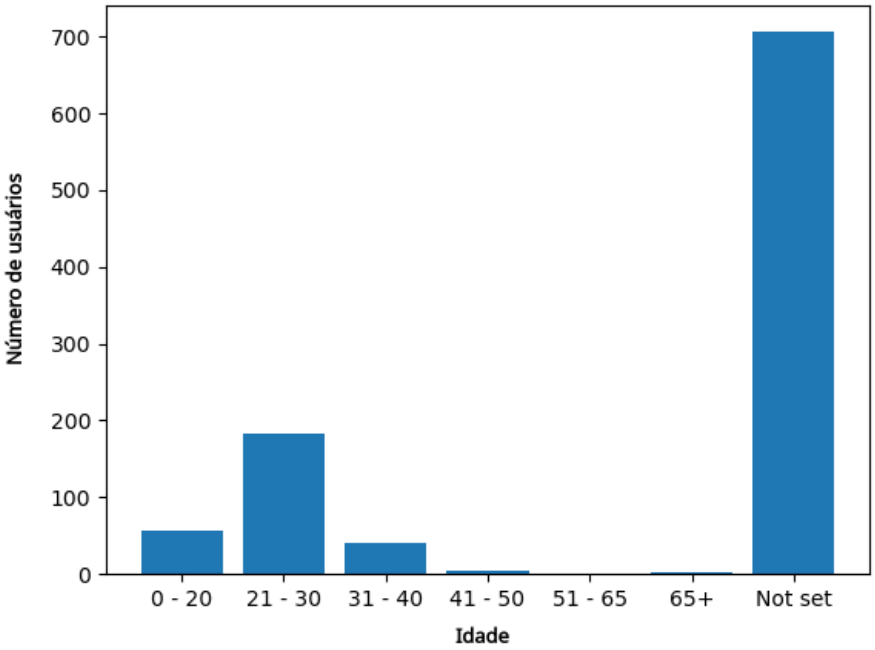


**Figura 1. Histograma de idade dos usuários**

Na figura 1, observamos que a idade da maioria dos usuários não foi informada. Adiante, obervamos que dos usuários com idade informada, mais da metade estão na faixa etária entre 21 e 30 anos. Além disso, temos apenas uma minoria com 41 ou mais anos. Tal concentração etária certamente influenciará os hábitos musicais refletidos na base, provavelmente em favor de tendências do público jovem.

### 4.2. Faixas Populares

Quais foram as faixas mais populares entre os usuários no mundo todo?

| Faixa | Reproduções |
|---|---|
| Such Great Heights | 3992 |
| Love Will Tear us Apart | 3663 |
| Karma Police | 3534 |
| Supermassive Black Hole | 3483 |
| Soul Meets Body | 3479 |
| Heartbeats | 3156 |
| Starlight | 3060 |
| Rebellion (Lies) | 3048 |
| Gimme More | 3004 |
| When You Were Young | 2998 |

**Tabela 1. Faixas com mais reproduções**

Na tabela 1, apresentamos as dez faixas mais reproduzidas na base de dados. Observamos a presença de alguns clássicos como *Love Will Tear us Apart*, por *Joy Division*, e *Karma Police*, por *Radiohead*.

### 4.3. Artistas Populares

No mundo todo, quais são os artistas mais populares entre os usuários?

| Artista | Reproduções |
|---|---|
| Radiohead | 115209 |
| The Beatles | 100338 |
| Nine Inch Nails | 84421 |
| Muse | 63346 |
| Coldplay | 62251 |
| Depeche Mode | 59910 |
| Pink Floyd | 58561 |
| Death Cab For Cutie | 58083 |
| Placebo | 53518 |
| Elliott Smith | 50278 |

**Tabela 2. Artistas com mais reproduções**

Na tabela 2, apresentamos os artistas com mais reproduções na plataforma. Notamos que todos constituem artistas ou bandas de *rock music*, apesar da presença de diversas estrelas *pop* na base dados. Concluímos que talvez o *Last.fm* constitua uma plataforma mais procurada pelos ouvintes deste gênero musical.

### 4.4. *Heavy Users*

Quem são os *heavy users* do Last.fm?

| ID de usuário | Gênero | Ano de registro | País | Reproduções |
|---|---|---|---|---|
| 000949 | F | 2005 | Estados Unidos | 36742 |
| 000791 | M | 2004 | Reino Unido | 32012 |
| 000544 | F | 2006 | Estados Unidos | 31645 |
| 000861 | M | 2005 | Estados Unidos | 30899 |
| 000800 | - | 2005 | Rússia | 28363 |
| 000691 | M | 2006 | Estados Unidos | 26468 |
| 000274 | M | 2006 | México | 24978 |
| 000155 | M | 2006 | Venezuela | 24243 |
| 000233 | M | 2005 | Estados Unidos | 23551 |
| 000349 | M | 2005 | Brasil | 22760 |

**Tabela 3. Usuários mais ativos**

Na tabela 3, apresentamos os usuários com mais reproduções na plataforma. Observamos uma variedade de países de origem, o que indica que o *Last.fm*, apesar de bem focado nos Estados Unidos da América, possui aficcionados ao redor do globo.

### 4.5. Demografia

Onde estão os usuários do Last.fm?

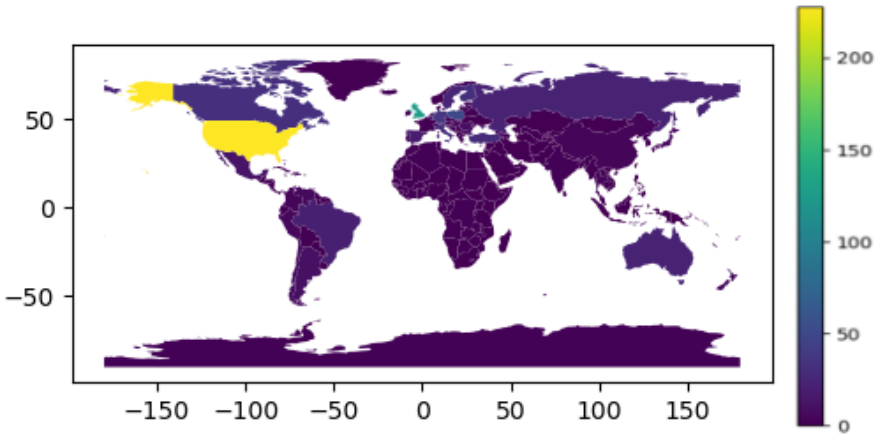


**Figura 2. Mapa de calor da localização dos usuários no mundo**

Na figura 2 observamos uma grande concentração dos usuários nos Estados Unidos da América, seguido de alguns países da Europa. Ainda temos uma quantidade perceptível de usuários em países como Brasil, Argentina, Rússia, Canadá e Austrália. As demais regiões não apresentaram quantidades significativas de usuários.

### 4.6. Geografia

Onde ocorrem a maioria das reproduções de faixas?

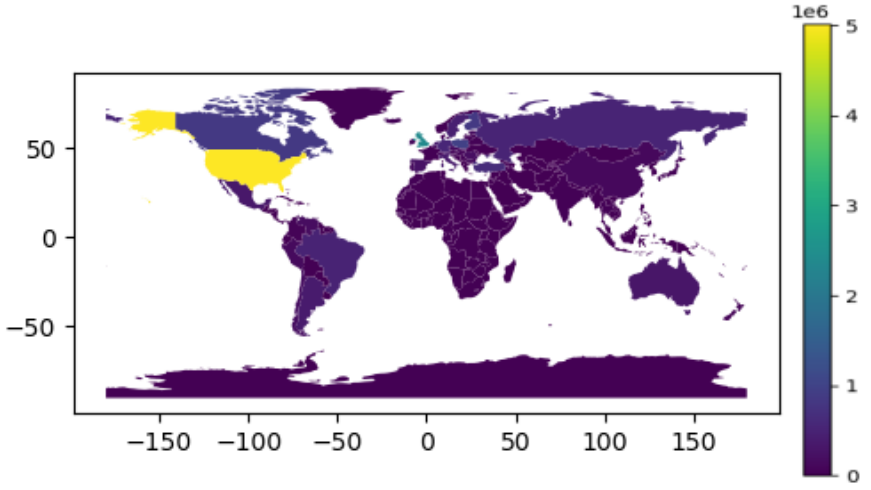


**Figura 3. Mapa de calor de reproduções no mundo**

De forma correspondente à disposição dos usuários, as reproduções se concentram nos Estados Unidos da América, seguido de países europeus. Na retaguarda, temos novamente Brasil, Argentina, Rússia, Canadá e Austrália. Observamos portanto uma correspondência direta com a disposição dos usuários, sem anomalias nesta relação.

### 4.7. *Hits*

Quais foram os hits do momento?

| Semana | Reproduções | Faixa |
|---|---|---|
| 2005-08-25 | 901 | Tv On The Radio - Staring At The Sun |
| 2006-01-05 | 569 | Cake - Jolene |
| 2008-09-25 | 538 | Britney Spears - Womanizer |
| 2007-05-03 | 529 | The Knife - Heartbeats |
| 2009-02-26 | 469 | Zeigeist - Fight With Shattered Mirrors |
| 2005-09-15 | 465 | Broken Social Scene - Anthems For A Seventeen Year Old Girl |
| 2006-11-09 | 446 | Radiohead - Everything In Its Right Place |
| 2007-02-08 | 446 | Soilwork - Distortion Sleep |
| 2007-04-12 | 388 | Mellowdrone - Beautiful Day |
| 2007-08-30 | 370 | Britney Spears - Gimme More |

**Tabela 4. Faixas que obtiveram maiores surtos de reproduções**

Apresentamos na tabela 4 as faixas que obtiveram um surto repentino de reproduções, constituindo efetivamente *hits* na indústria da música. Aqui já observamos algumas faixas do gênero pop, o que se justifica pela sua contemporaneidade. As faixas mais reproduizadas no total, em sua grande maioria, são do século passado, e portanto é menos provável que ocorram surtos em sua frequência de reproduções.

### 4.8. Sistema de Recomendação

Foi construído o sistema de recomendação baseado em filtragem colaborativa. O algoritmo utilizado para aplicar esta técnica foi o ALS - Alternating Least Squares[3]. De maneira geral, o algoritmo tenta definir características para os usuários e músicas, baseado nas avaliações feitas. Após definidas as características, estas podem ser utilizadas para recomendar músicas para usuários, e vice-versa.

O algoritmo ALS recebe alguns hiper-parâmetros. Nossa implementação explora algumas diferentes combinações de valores para iterações e regularização. Após a escolha do melhor modelo, dez recomendações foram feitas para todos os usuários da base. Cada faixa recomendada acompanha a estimativa de quantidade de reproduções pelo usuário.

| Faixa | Est. de Reproduções |
|---|---|
| Christine | 1.99 |
| Stupid Girl | 2.05 |
| Ziggy Stardust | 2.14 |
| Sweetest Perfection | 1.97 |
| A Little Respect | 1.96 |
| Charlotte Sometimes | 2.10 |
| Life On Mars? | 2.12 |
| Vogue | 2.10 |
| Material Girl | 2.03 |
| Queer | 2.19 |

**Tabela 5. Recomendação para o usuário com ID 000861**

A tabela 5 mostra as recomendações feitas para um dos *heavy users* da plataforma, o usuário com ID 000861. Podemos ver que as músicas recomendadas são predominantemente das décadas de 70 a 90, com gêneros variando entre Pop, Rock e Dance. É interessante notar que a quantidade de reproduções não passa de 3, sendo o limite observado nas recomendações feitas.

| Faixa | Est. de Reproduções |
|---|---|
| Ace Of Spades | 0.01 |
| The Trooper | 0.01 |
| Highway To Hell | 0.01 |
| The Number Of The Beast | 0.01 |
| Hallowed Be Thy Name | 0.01 |
| Still Loving You | 0.01 |
| Thunderstruck | 0.01 |
| Poison | 0.01 |
| Run To The Hills | 0.01 |
| Enter Sandman | 0.01 |

**Tabela 6. Recomendação para o usuário com ID 000533**

A tabela 6 mostra recomendações com valor baixo para quantidade estimada de reproduções. Acreditamos que tal resultado se dá pela infrequência do usuário na plataforma. Porém, mesmo com baixa atividade, as faixas recomendadas se mantém coerentes com as tendências do usário, variando entre os gêneros Heavy metal e Hard Rock.

| Faixa | Est. de Reproduções |
|---|---|
| Amazing (Feat. Young Jeezy) | 0.99 |
| Street Lights | 1.00 |
| Pinocchio Story (Freestyle Live From Singapore) | 1.00 |
| Heartless | 1.05 |
| Welcome To Heartbreak (Feat. Kid Cudi) | 1.01 |
| Love Lockdown | 1.09 |
| Coldest Winter | 1.04 |
| Paranoid (Feat. Mr. Hudson) | 1.01 |
| Robocop | 1.01 |
| Bad News | 0.99 |

**Tabela 7. Recomendação para o usuário com ID 000488**

A tabela 7 mostra recomendações para um usuário regular. Neste resultado, é interessante notar a predominância de músicas do gênero Hip-Hop, do artista Kanye West. Este e outros resultados abaixo mostram a capacidade do modelo de reproduzir as tendências dos usuários em relação aos diferentes gêneros musicais.

Como curiosidade, são apresentados os resultados abaixo nas tabelas 8 e 9. Estes mostram gostos musicais díspares da maioria dos usuários. A tabela 9 mostra as recomendações de faixas no gênero Pop da década de 2000. A tabela 8 mostra recomendações com gêneros alternativos.

| Faixa | Est. de Reproduções |
|---|---|
| Rehab | 1.8035021 |
| Love Is A Losing Game | 1.5938014 |
| Back To Black | 1.8022593 |
| Gimme More | 1.6374944 |
| Wake Up Alone | 1.6190151 |
| Tears Dry On Their Own | 1.8812811 |
| Guilty Pleasure | 1.7952653 |
| Womanizer | 1.6756923 |
| Radar | 1.5699006 |
| Poker Face | 1.7283481 |

**Tabela 8. Recomendação para o usuário com ID 000458**

| Faixa | Est. de Reproduções |
|---|---|
| Unravel | 0.18480842 |
| Gorecki | 0.18318476 |
| Risingson | 0.18048202 |
| Group Four | 0.18025853 |
| Life In Mono | 0.19038083 |
| La Noyée | 0.19454736 |
| Jóga | 0.18207338 |
| Gabriel | 0.18400241 |
| La Valse Des Monstres | 0.18377545 |
| Sur Le Fil | 0.20105311 |

**Tabela 9. Recomendação para o usuário com ID 000540**

## 5. Conclusão

Com este trabalho, consolidamos os conhecimentos adquiridos na disciplina sobre processamento distribuído de dados massivos. Em particular, exploramos a plataforma Spark para diversas tarefas de análise de dados, incluindo limpeza, transformações, agregações e aprendizado de máquina. Reconhecemos que o Spark foi fundamental para permitir a aplicação de tais técnicas em uma base de dados massiva, coletada em uma plataforma com milhões de usuários. Tal experiência nos empodera a executar demais tarefas de processamento de dados massivos em trabalhos futuros.

## Referências


[1] LAST.FM. *The last.fm website*. 2010. `https://www.last.fm/`.

[2] SPARK. *Collaborative Filtering in Spark*. `https://spark.apache.org/docs/2.2.0/ml-collaborative-filtering.html`.

[3] KOREN, Y. e. a. *Matrix Factorization Techniques for Recommender Systems*. `https://dl.acm.org/doi/10.1109/MC.2009.263`.