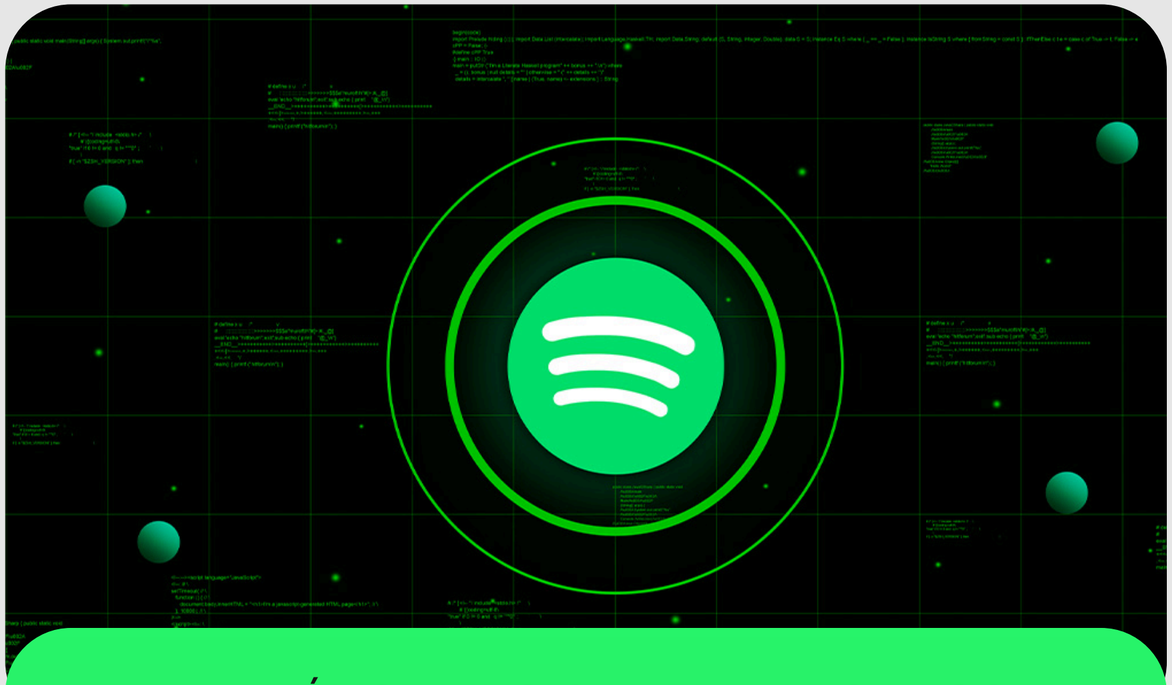
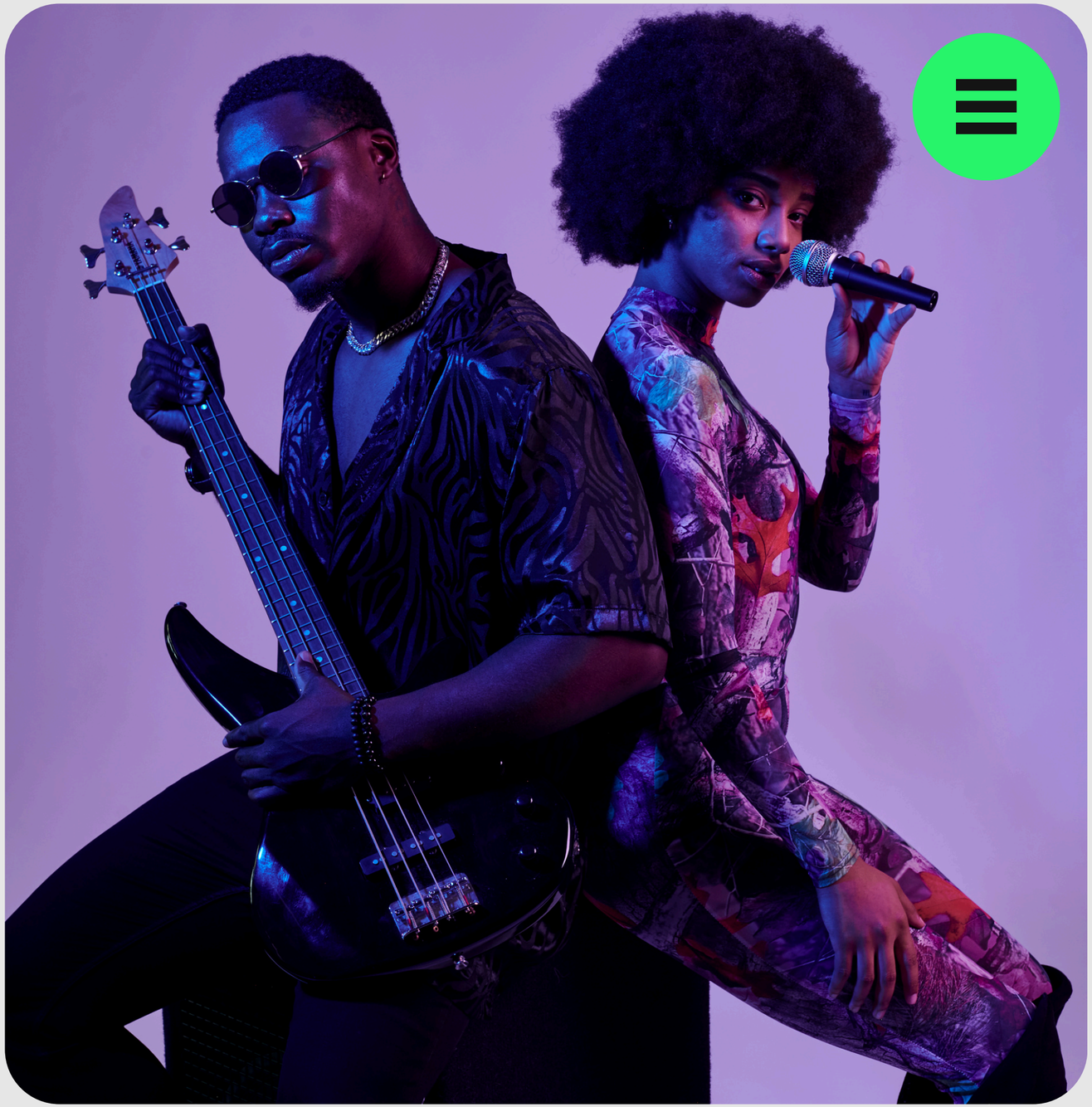


ANÁLISE MUSICAL EM LARGA ESCALA

SPOTIFY

JOÃO PAULO SOUZA
KATARINA FRIEDRICH
LEANDRO CASTRO
PEDRO COBUCCI
SAMUEL SANTOS

NEXT



DATASET

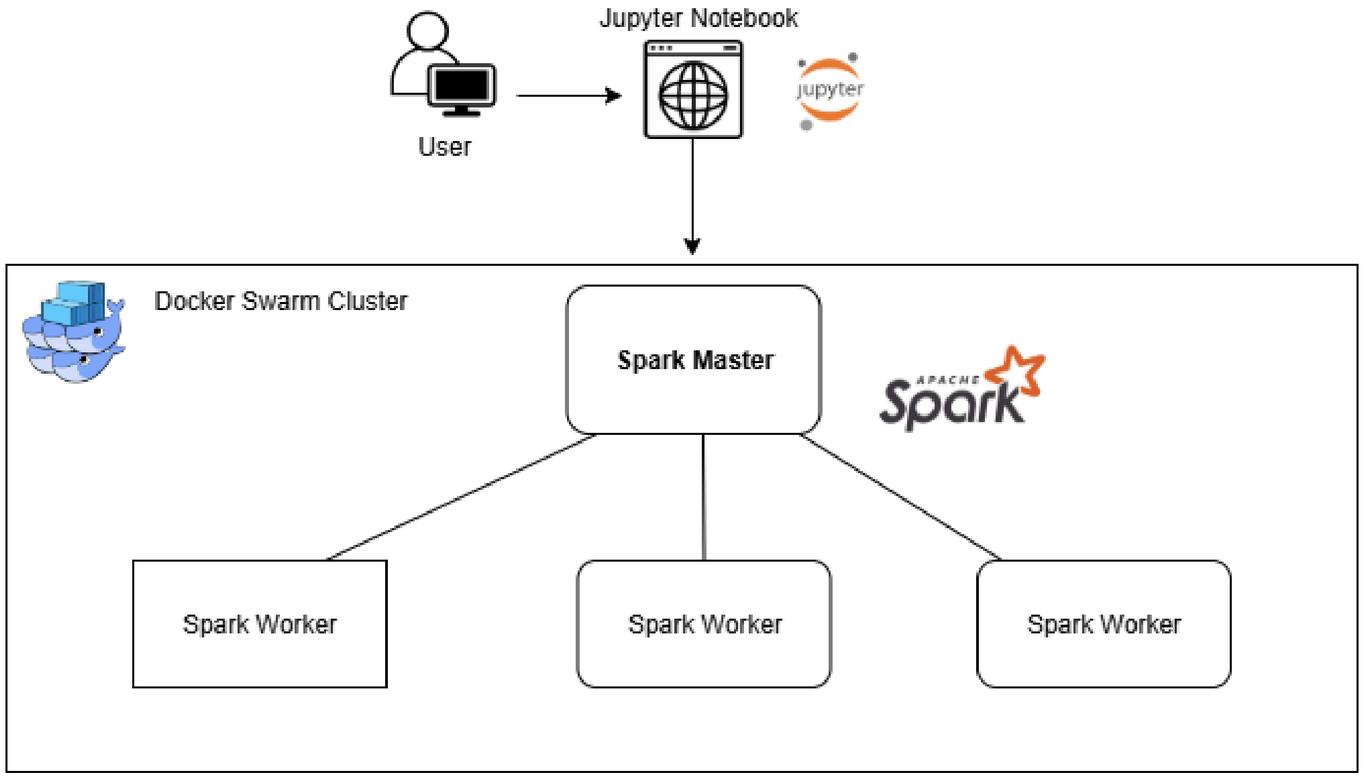
Cada linha representa uma música e seus dados

500.000 linhas
10 colunas

CSV para Parquet

424 MB (parquet)

ARQUITETURA



[CSV] → [TRANSFORMAÇÃO PARQUET] → [LIMPEZA] → [PROCESSAMENTO] → [RESULTADOS]

- Usuário (navegador web): Acessa a interface do Jupyter Notebook em http://localhost:8888 para escrever e executar scripts PySpark. O navegador não integra o cluster, mas é o ponto de entrada para o usuário.
- Jupyter Notebook (Spark Driver): Executado em um contêiner, fornece a interface web e atua como Spark Driver. Inicia a SparkSession e envia os jobs para o cluster via spark://spark-master:7077. Compartilha o volume /spark-data com os demais serviços.
- Spark Master: Coordena o cluster, recebendo jobs do Driver e distribuindo-os entre os Workers. Roda em contêiner próprio e expõe sua interface de monitoramento em http://localhost:8080
- Spark Workers: Executam as tarefas de forma distribuída. Conectam-se automaticamente ao Master e acessam o volume /spark-data para leitura e escrita. São contêineres escaláveis no Docker Swarm.
- Volume /spark-data: Pasta compartilhada entre todos os contêineres, usada para armazenar dados de entrada (.csv) e saída (.parquet). Elimina a necessidade de transferência via rede.
- Rede spark-net: Interliga todos os contêineres, permitindo comunicação entre os serviços.

[WORKLOAD-1]

AGRUPAMENTO DE LETRAS DE MÚSICAS POR EMOÇÃO

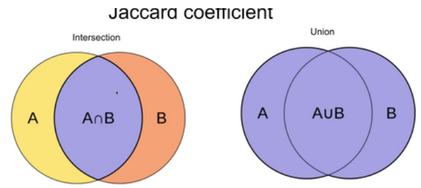
Objetivo: Mostrar quais palavras são mais frequentes nas letras das músicas em cada emoção.

Etapas:

- Leitura do dataset no formato Parquet.
- Remoção de pontuação e símbolos usando regex_replace.
- Remoção de StopWords (palavras sem sentido semântico).
- Tokenização.
- Explosão das palavras em linhas individuais usando explode.
- Identificação da principal emoção associada a cada palavra
- Ranqueamento de palavras mais recorrentes por emoção.

[WORKLOAD-2]

CÁLCULO DA SIMILARIDADE DE JACCARD PARA ENCONTRAR OS PARES DE ARTISTAS MAIS SIMILARES DENTRO DOS PRINCIPAIS GÊNEROS

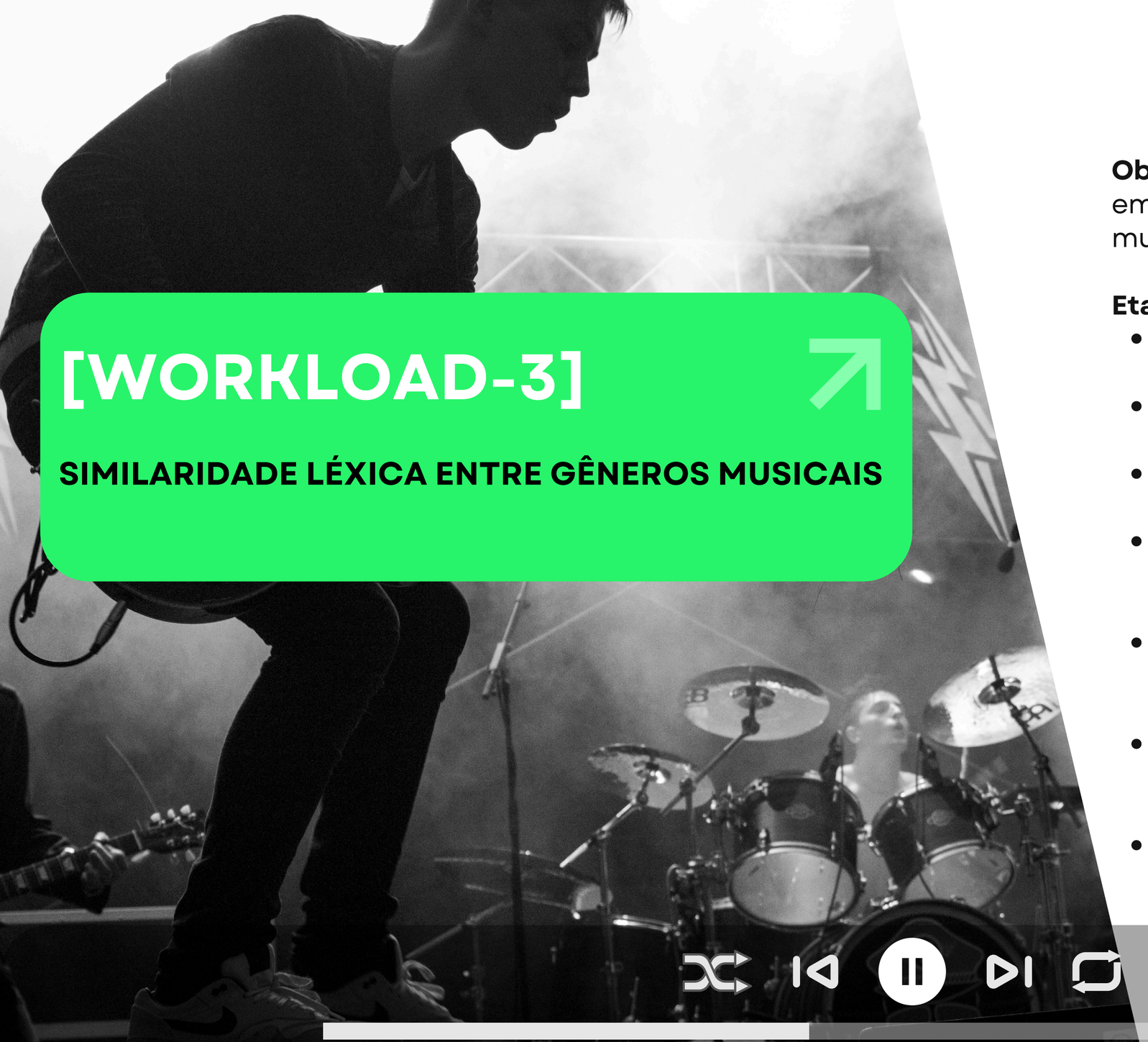


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Objetivo: Comparar todos os pares possíveis de artistas dentro do mesmo gênero utilizando a métrica de Jaccard com base no vocabulário textual.

Etapas:

- Leitura do dataset no formato Parquet.
- Remoção de pontuação e símbolos usando regex_replace.
- Conversão para minúsculas.
- Explosão das palavras em linhas individuais usando explode.
- Realização de um self-join (cruzamento) entre os artistas por gênero.
- Cálculo de interseção (array_intersect) e união (array_union) de vocabulários.
- Cálculo da similaridade de Jaccard: interseção / união.
- Ranqueamento dos pares mais similares por gênero.



[WORKLOAD-3]

SIMILARIDADE LÉXICA ENTRE GÊNEROS MUSICAIS

Objetivo: Encontrar o número de palavras em comum entre os principais gêneros musicais, medindo similaridade léxica.

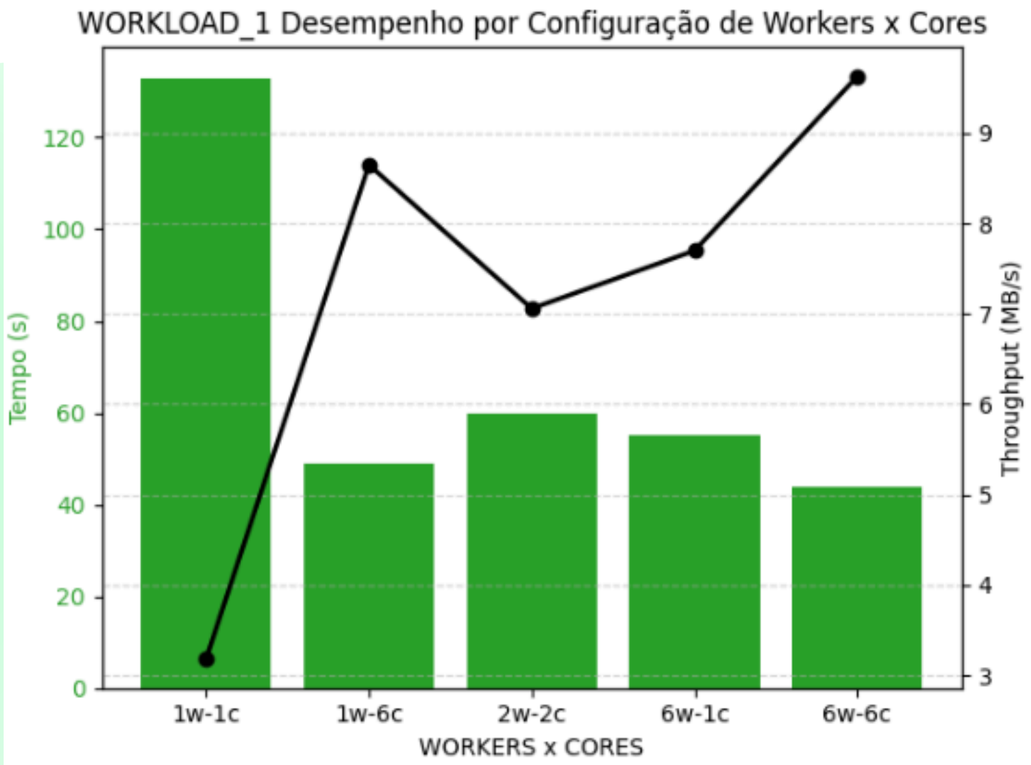
Etapas:

- Leitura do dataset musicas_limpas.parquet.
- Seleção dos 10 gêneros musicais com mais músicas no dataset.
- Explosão das palavras com explode e split.
- Agrupamento por main_genre usando collect_set("palavra") para obter vocabulário único por gênero.
- Self-join do vocabulário de gêneros usando crossJoin, comparando pares distintos (g1 < g2).
- Cálculo do número de palavras em comum por par de gêneros usando array_intersect + size.
- Ordenação decrescente pelo número de palavras em comum.

CÓDIGO

EXPERIMENTS AND RESULTS

WORKLOAD 1



WORKLOAD 1

1W-1C

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input
Active(2)	0	276.3 KiB / 827.8 MiB	0.0 B	1	1	0	3	4	2.1 min (2 s)	171.7 MiB
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B
Total(2)	0	276.3 KiB / 827.8 MiB	0.0 B	1	1	0	3	4	2.1 min (2 s)	171.7 MiB

1W-6C

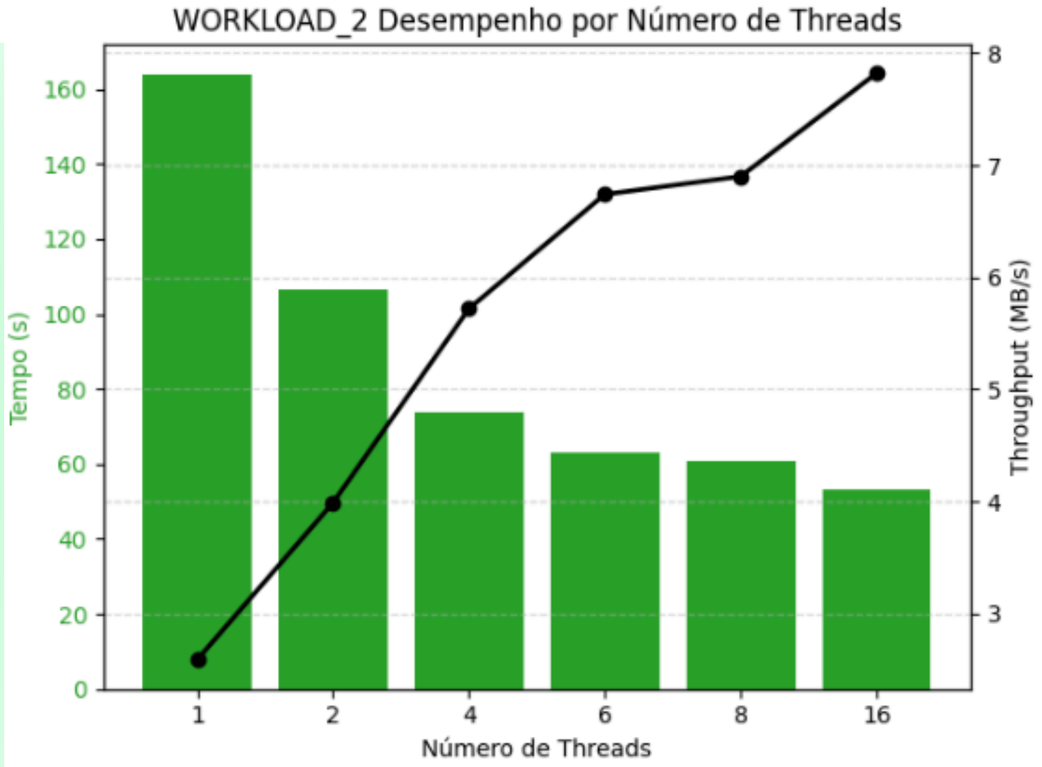
	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input
Active(2)	0	276.2 KiB / 848.3 MiB	0.0 B	6	6	0	1	7	18 s (0.5 s)	0.0 B
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B
Total(2)	0	276.2 KiB / 848.3 MiB	0.0 B	6	6	0	1	7	18 s (0.5 s)	0.0 B

6W-6C

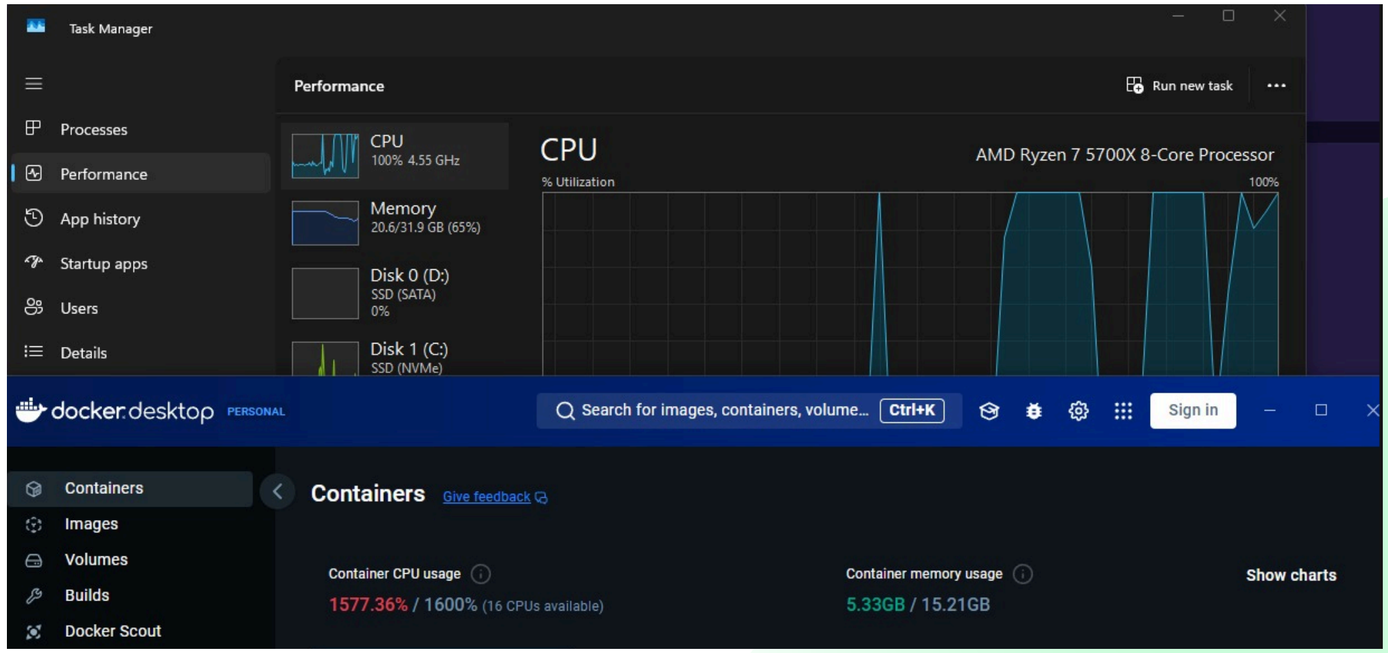
	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input
Active(7)	0	594 KiB / 2.9 GiB	0.0 B	36	10	0	29	39	2.8 min (6 s)	430.9 KiB
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B
Total(7)	0	594 KiB / 2.9 GiB	0.0 B	36	10	0	29	39	2.8 min (6 s)	430.9 KiB

EXPERIMENTS AND RESULTS

WORKLOAD 2

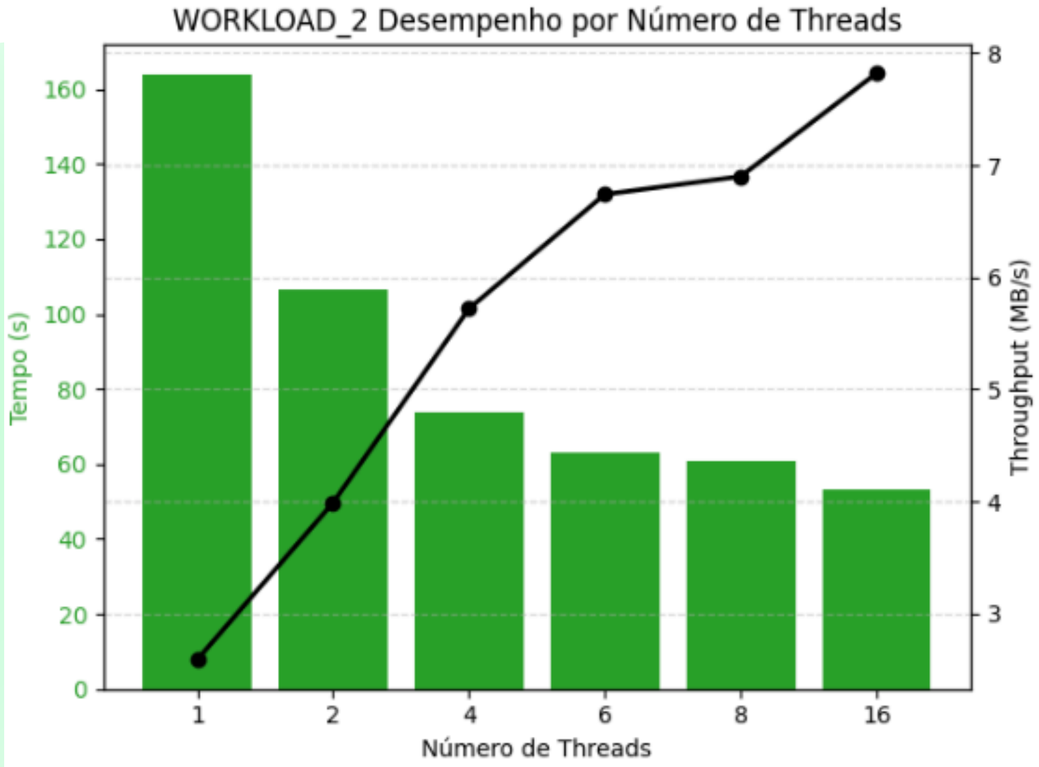


WORKLOAD 2



EXPERIMENTS AND RESULTS

WORKLOAD 2



WORKLOAD 2

Active(17)	0	2.3 MiB / 6.9 GiB	0.0 B	16	10	0	30	40	2.0 min (5 s)	1 MiB	55.7 KiB	55.7 KiB	0			
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0			
Total(17)	0	2.3 MiB / 6.9 GiB	0.0 B	16	10	0	30	40	2.0 min (5 s)	1 MiB	55.7 KiB	55.7 KiB	0			

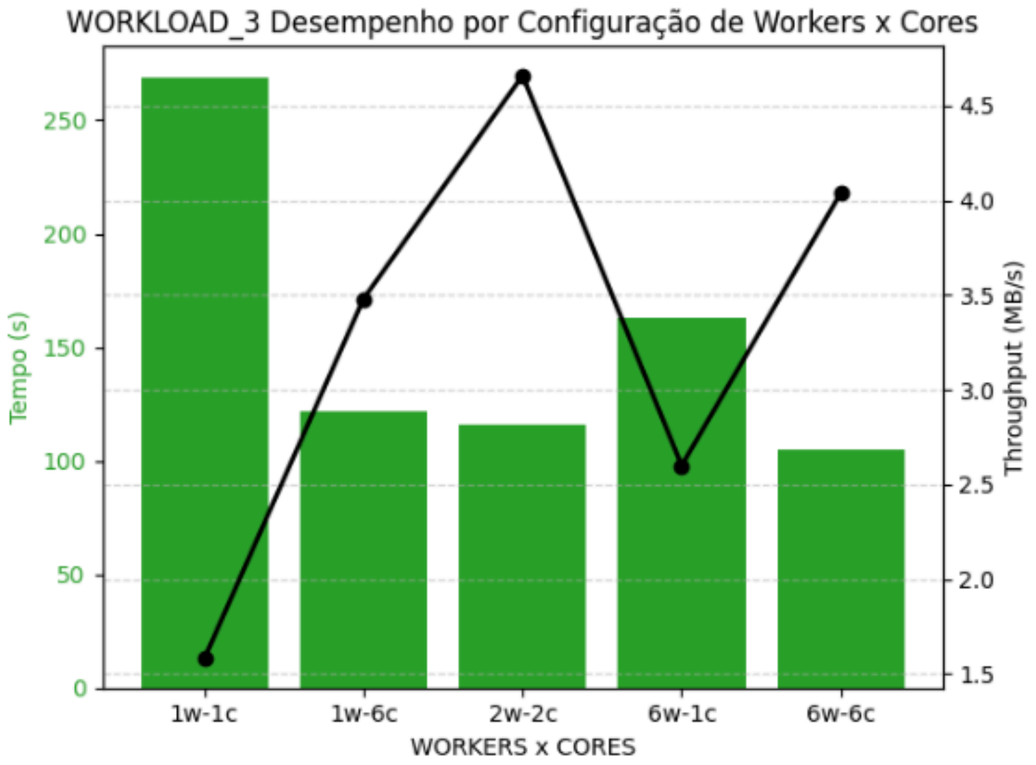
Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump	Heap Histogram	Add Time	Remove Time
driver	4a576f9fe9f6:42211	Active	0	234.3 KiB / 413.9 MiB	0.0 B	0	0	0	0	0	33 s (0.5 s)	0.0 B	0.0 B	0.0 B		Thread Dump	Heap Histogram	2025-07-09 17:55:44	-
0	10.0.2.32:41061	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	5 s (0.3 s)	81.1 KiB	0.0 B	5.5 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
1	10.0.2.38:44525	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	5 s (0.3 s)	71.2 KiB	0.0 B	5.6 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
2	10.0.2.29:41211	Active	0	159.6 KiB / 413.9 MiB	0.0 B	1	0	0	4	4	7 s (0.3 s)	110.7 KiB	55.7 KiB	5.6 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
3	10.0.2.26:43273	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	6 s (0.5 s)	57.2 KiB	0.0 B	5.5 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
4	10.0.2.33:40117	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	0	0	2	2	6 s (0.3 s)	93.6 KiB	0.0 B	5.5 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
5	10.0.2.37:39433	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	0	0	3	3	6 s (0.3 s)	102 KiB	0.0 B	5.6 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
6	10.0.2.35:41655	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	5 s (0.1 s)	17.2 KiB	0.0 B	0.0 B	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:52	-
7	10.0.2.25:35805	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	5 s (0.3 s)	79.2 KiB	0.0 B	5.5 KiB	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-
8	10.0.2.27:32921	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	0	0	3	3	6 s (0.3 s)	47.4 KiB	0.0 B	0.0 B	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:52	-
9	10.0.2.39:40209	Active	0	122.4 KiB / 413.9 MiB	0.0 B	1	1	0	1	2	5 s (0.3 s)	14.4 KiB	0.0 B	0.0 B	stdout stderr	Thread Dump	Heap Histogram	2025-07-09 17:55:51	-

EXPERIMENTS AND RESULTS

WORKLOAD 3



WORKLOAD 3

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Sh
Active(7)	200	92 MiB / 2.8 GiB	0.0 B	6	7	0	7991	7998	6.2 min (18 s)	5.3 GiB	1(
Dead(1)	0	181.6 KiB / 413.9 MiB	0.0 B	1	0	2	2	3	37 s (2 s)	43.6 MiB	0.
Total(8)	200	92.1 MiB / 3.2 GiB	0.0 B	7	7	2	7993	8001	6.8 min (20 s)	5.3 GiB	1(

Executors

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
driver	0f86dc356fad:37521	Active	0	470.9 KiB / 413.9 MiB	0.0 B	0	0	0	0	0	1.4 min (2 s)	0.0 B	0.0 B	0.0 B		Thread Dump
0	10.0.1.15:45011	Active	50	22.3 MiB / 413.9 MiB	0.0 B	1	1	0	1978	1979	1.1 min (3 s)	1.3 GiB	54.1 MiB	45.3 MiB	stdout stderr	Thread Dump
1	10.0.1.21:37049	Active	0	224.7 KiB / 413.9 MiB	0.0 B	1	0	0	6	6	19 s (2 s)	45.6 MiB	0.0 B	13.7 MiB	stdout stderr	Thread Dump
2	10.0.1.14:41543	Active	47	21.5 MiB / 413.9 MiB	0.0 B	1	2	0	1963	1965	1.0 min (3 s)	1.3 GiB	33.4 MiB	36.5 MiB	stdout stderr	Thread Dump
3	10.0.1.17:42327	Active	55	24.7 MiB / 413.9 MiB	0.0 B	1	2	0	2036	2038	1.1 min (3 s)	1.3 GiB	61.7 MiB	47.9 MiB	stdout stderr	Thread Dump
4	10.0.1.19:37169	Dead	0	181.6 KiB / 413.9 MiB	0.0 B	1	0	2	2	3	37 s (2 s)	43.6 MiB	0.0 B	13.7 MiB	stdout stderr	Thread Dump
5	10.0.1.18:38503	Active	48	22.8 MiB / 413.9 MiB	0.0 B	1	2	0	2004	2006	1.2 min (5 s)	1.4 GiB	35.7 MiB	41.3 MiB	stdout stderr	Thread Dump
6	10.0.1.19:38647	Active	0	18.5 KiB / 413.9 MiB	0.0 B	1	0	0	4	4	2 s (99.0 ms)	1.7 MiB	0.0 B	236 B	stdout stderr	Thread Dump

OBRIGADO PELA ATENÇÃO