# Apresentação de projeto final

GCC273 – BIG DATA

Arthur Nunes
Caio Bastos
Daniel Assis
Paulo Pompeu
Pedro Rabelo

# Motivação e Proposta

- **Problema:** Alto volume de microdados do ENEM dificulta análises locais.
- **Objetivo:** Processar, analisar e extrair insights dos dados de forma distribuída.
- **Solução:** Pipeline de dados usando Apache Spark + HDFS em Docker Swarm.
- **Volume de dados analisado:** ~13 milhões de registros por ano.

# Fonte de dados

- **Dataset escolhido:** Microdados do ENEM 2020, 2021 e 2023 (Exame Nacional do Ensino Médio)
- **Disponibilidade:** Disponível publicamente via Inep (dados abertos) e Kaggle
- **Contexto e informações:** Este conjunto fornece dados detalhados de ~2,7 milhões de participantes do ENEM por ano, incluindo informações socioeconômicas, desempenho nas provas e muito mais
- **Campos utilizados:** inscrição, ano, UF, tipo de escola, faixa de renda (Q006), nota geral, nota de matemática, entre outros.

# Processamento a ser Realizado

Usaremos Apache Spark para processar os dados do ENEM, explorando paralelismo e memória distribuída. As principais tarefas planejadas incluem:

- Cada componente roda em um container.
- Comunicação por rede Docker interna (hadoop).
- Escalável com --scale spark-worker=X --scale datanode=Y.

# Pipeline de Processamento

- Verifica/baixa/extrai os dados localmente.
- Envia os arquivos .csv para o HDFS (/user/enem/csv_raw/<ano>/).
- Lê os dados via Spark com schema tipado.
- Processa e salva em .parquet no HDFS e local.
- Executa análises e salva resultados.
- Durante o processo faz o log de métricas de análise.

# Análises Realizadas

- [AN1] Média por UF: Média de notas da prova fechada por estado/ano.
- [AN2] Média por Tipo de Escola: Comparação entre pública, privada, etc.
- [AN3] Correlação Renda x Nota: Relação entre Q006 (faixa de renda) e nota.
- [AN4] Desigualdade Regional: Média, desvio padrão e nº de alunos por região.
- [AN5] Faixas de Renda: Agrupamento por 4 faixas (até 1k, 1k–3k, etc.)

# Análises Realizadas

## (AN1) Média Geral por UF

Exemplo de destaques (2020):

- São Paulo (SP): **541,20**
- Minas Gerais (MG): **534,08**
- Acre (AC): **480,82**
- Amapá (AP): **476,80**

## (AN2) Correlação Renda x Nota

Exemplo focado em 2020:

| Tipo de Escola | Descrição | Média ENEM |
|---|---|---|
| 1 | Não respondeu | 520.03 |
| 2 | Pública | 499.52 |
| 3 | Privada | 610.63 |

# Análises Realizadas

## [AN3] Correlação de Renda Familiar

| Ano | Correlação (Pearson) |
|---|---|
| 2020 | 0.3945 |
| 2021 | 0.3745 |
| 2023 | 0.3824 |

## (AN4) Desigualdade Regional

| Região | Média | Desvio Padrão | Número Participantes |
|---|---|---|---|
| Sudeste | 559.36 | 123.15 | 2.531.820 |
| Sul | 550.79 | 119.85 | 813.382 |
| Centro-Oeste | 528.26 | 121.08 | 627.912 |
| Nordeste | 508.94 | 115.88 | 2.744.535 |
| Norte | 487.11 | 103.82 | 818.062 |

# Análises Realizadas

## [AN5] Média por Faixa de Renda

| Faixa de Renda | Média ENEM |
|---|---|
| Até 1k | 490.49 |
| 1k–3k | 544.49 |
| 3k–6k | 598.89 |
| Acima de 6k | 650.86 |

# Resultados Obtidos

- Total de registros: 13+ milhões
- Registros processados: Aproximadamente 7535711 registros.
- Throughput médio:  20679.14 linhas/s
- Correlação renda x nota: positiva, mas moderada(~0.38 nos 3 anos)
- Diferença clara entre regiões e tipos de escola.
- Região Sudeste e escolas privadas com maiores médias.

# Experimentos de performance

| Configuração | Tempo (s) | Registros | Throughput | CPU Total (%) | RAM (MB) por Worker | Treads por Worker |
|---|---|---|---|---|---|---|
| 1W / 1D | 415.66 | 7.535.711 | 18.129,37 | 50.0% | 4096 MB | 3.0 |
| 2W / 1D | 362.80 | 7.535.711 | 20.770,75 | 50.0% | 4096 MB | 3.0 |
| 2W / 2D | 365.82 | 7.535.711 | 20.599,35 | 50.0% | 4096 MB | 3.0 |

Podemos verificar que ao aumentar a quantidade de Workers diminuiu o tempo de execução, e aumento o throughput

Testando com 3 Workers tivemos problemas de alocamento de memória de RAM. Com isso mantivemos os testes até 2 Workers

# Conclusões

- Pipeline completo, escalável e com análises relevantes.
- Custo de distribuição precisa ser balanceado (memória vs performance).
- Possível expansão para outros anos, análises e dashboards.
- Desafios com permissões HDFS e alocação de memória
- Overhead ao aumentar workers (troca entre nós)