Computação Paralela e Análise de Big Data: Ferramentas e Estratégias

GRUPO 02

01

SSC0903 - COMPUTAÇÃO DE ALTO DESEMPENHO

INTRODUÇÃO

- **Computação Paralela**
  - Computador e programação paralela para aumentar o desempenho.

- **Big Data**
  - Análises automáticas de padrões, a partir de grandes quantidades de dados.

Estratégias de Computação Paralela e Big Data

MPP - Massively Parallel Processing

Massively Parallel Processing, ou MPP, é uma arquitetura de processamento projetada para lidar com grandes volumes de dados e executar cálculos complexos em ambientes distribuídos.

Data Parallelism

Data Parallelism é uma estratégia que distribui grandes conjuntos de dados em partes menores, processando-as simultaneamente em diferentes unidades de processamento, como CPUs, GPUs ou clusters distribuídos.

MPP - Massively Parallel Processing

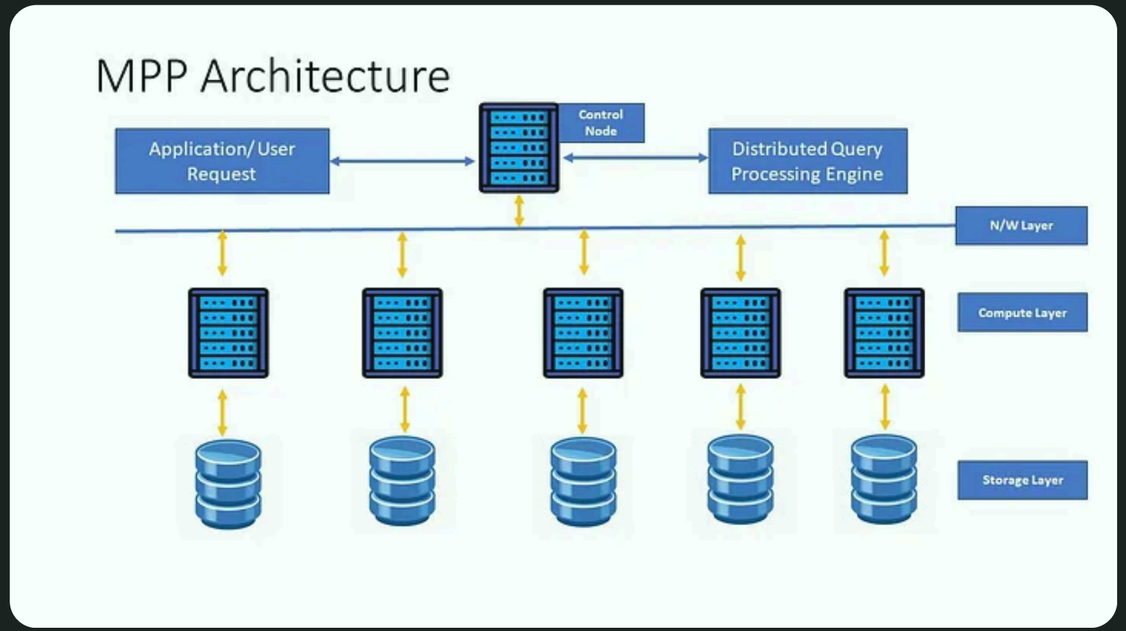
O QUE É?

- Sistema composto por múltiplos nós independentes;
- Em geral, cada nó tem sua própria memória e um sistema operacional;

MPP - Massively Parallel Processing

COMO FUNCIONA

- Opera com a divisão do trabalho entre os nós;
- Um **nó lider** recebe a tarefa principal e a divide em sub tarefas menores;
- Essas **sub tarefas** são divididas para os nós, que trabalham de forma independente;
- Os resultados individuais são combinados pelo nós líder;



FONTE:HTTPS://MEDIUM.COM/@JAGADESHJAMJALANARAYANAN/5-KEY-CONCEPTS-OF-MASSIVELY-PARALLEL-PROCESSING-86D993552F8C

MPP - Massively Parallel Processing

ARQUITETURAS

- **Shared-Nothing:** cada nó trabalha com recursos isolados;
- **Shared-Disk:** os nós compartilham um armazenamento comum;

MPP - Massively Parallel Processing

CONCLUSÃO

Massively Parallel Processing é uma tecnologia essencial em uma era orientada por dados. Sua arquitetura distribuída e escalável permite que empresas e cientistas de dados lidem com o crescimento exponencial de informações, transformando dados brutos em insights práticos.

Data Parallelism

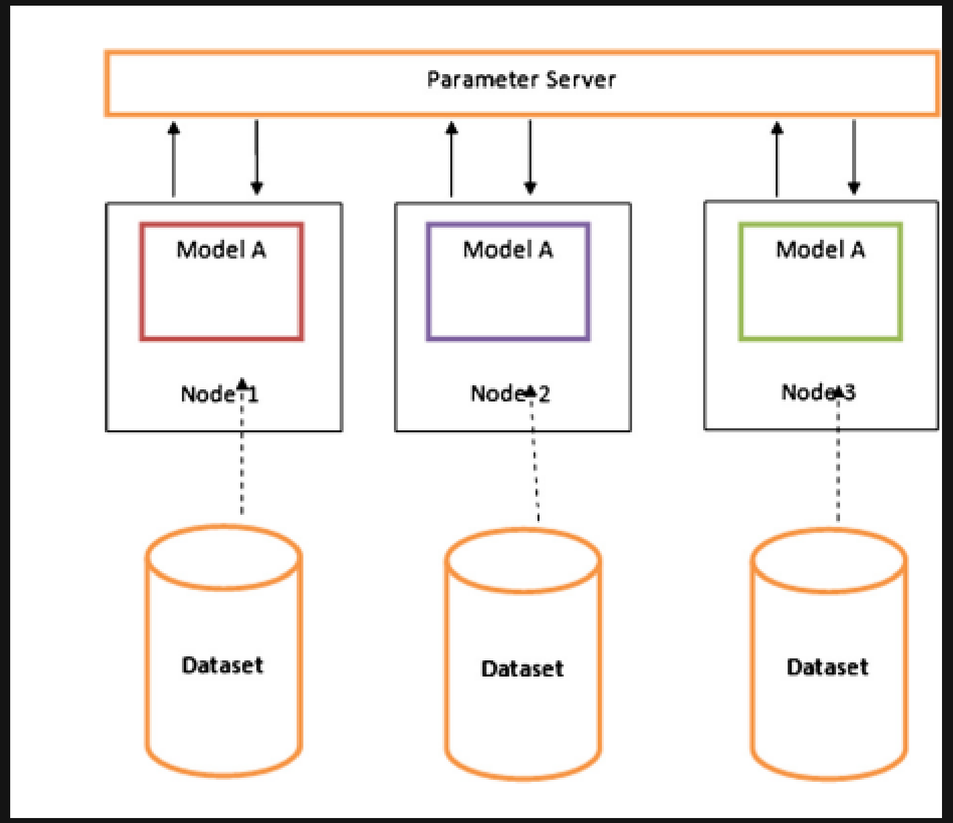
O QUE É?

- Computação paralela e análise de Big Data são práticas que dividem problemas grandes em partes menores, resolvendo-os simultaneamente.
- Essenciais para lidar com volumes massivos de dados e tarefas intensivas, aumentando eficiência e velocidade de execução.

Data Parallelism

COMO FUNCIONA

- Processos são distribuídos entre múltiplos núcleos ou máquinas, que trabalham em paralelo.
- Dados são particionados ou organizados em blocos menores, garantindo processamento simultâneo.
- O resultado final é consolidado a partir das saídas individuais de cada unidade de processamento.



FONTE: MADIAJAGAN, 2019

Data Parallelism

CONCLUSÃO

Data Parallelism é uma estratégia essencial para processar grandes volumes de dados, dividindo-os em partes menores para execução simultânea. Essa abordagem reduz tempos de processamento, otimiza recursos e suporta escalabilidade, sendo amplamente utilizada em análises de Big Data.

Ferramentas
de Computação Paralela e Big Data

Open Computing Language (OpenCL)

Padrão de programação em ambiente computacional heterogêneo

Apache Spark

Plataforma de processamento de dados em larga escala

BigQuery

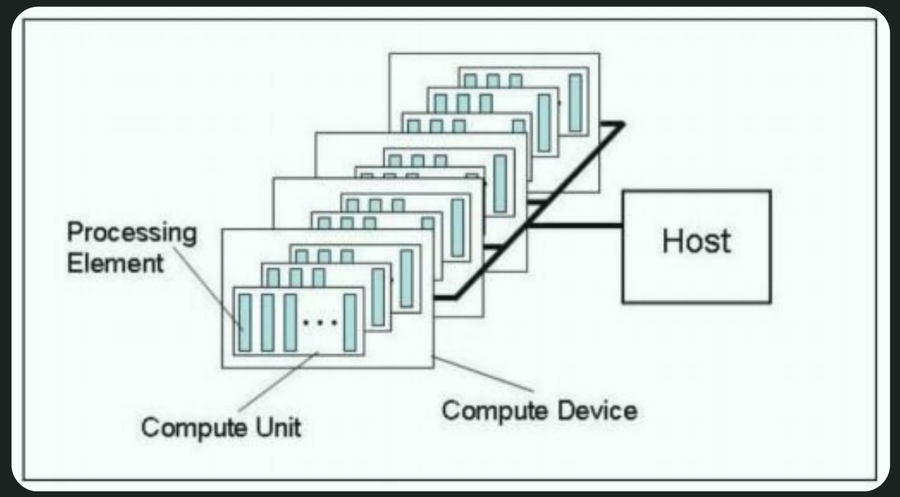
Data warehouse oferecido pelo Google Cloud

OpenCL

CURIOSIDADES

- Criado pela Apple.
- Padronizada pelo Khronos Group

MODELO DE PLATAFORMA:



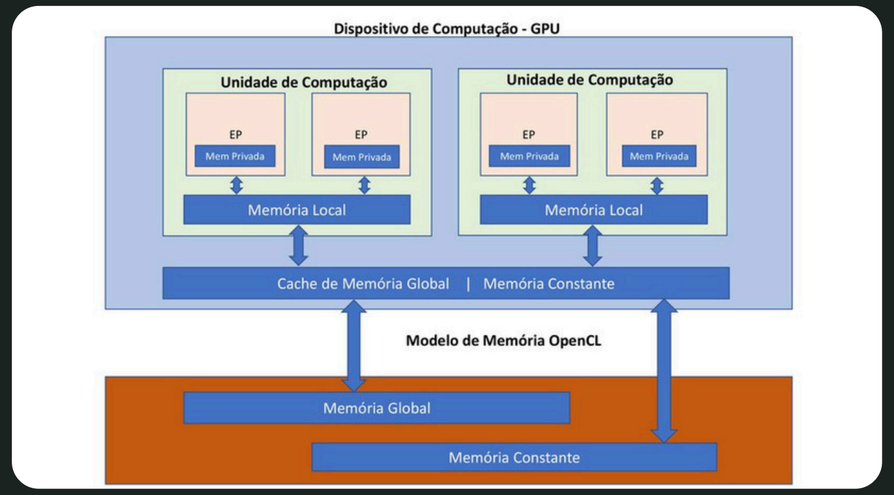
Fonte: https://www.researchgate.net/figure/Figura-1-Modelo-da-Plataforma-OpenCL-8_fig1_311740814

- Plataforma de computação para sistemas heterogêneios;
  - Modelo: *host* responsável por inicialização e transferência de dados/tarefas dos dispositivos.
- Baseado em C99.
- Abordagens:
  - Execução de algoritmos de IA;
  - Paralelismo de dados;
  - Paralelismo de tarefas.

OpenCL

- Níveis de acesso ao modelo de memória;
  - Global: compartilhada por todos os itens para leitura e escrita;
  - Local: compartilhada por itens de um mesmo grupo para leitura e escrita;
  - Privada: restrita a cada item de trabalho para escrita e leitura
  - Constante: compartilhada por todos os itens para leitura.

MODELO DE MEMÓRIA:



Fonte: https://slideplayer.com.br/slide/16901281/

- Consistência de leitura e escrita.
- Memórias global e local são consistentes entre itens de trabalho de um mesmo grupo de trabalho em uma barreira

Apache Spark

VISÃO GERAL

O QUE É O APACHE SPARK?

Plataforma de computação em cluster que fornece uma API para programação distribuída para processamento de dados em larga escala.

USOS

- Processamento de Dados em Tempo Real
- Modelos de Machine Learning
- Processamento de Dados Estruturados

CARACTERÍSTICAS

- Permite a divisão de dados e tarefas em clusters com vários nós
- Cada nó funciona processa apenas uma parte parte do volume total de dados
- Compatível com várias linguagens: Python, Scala, Java e R

Apache Spark

MÓDULOS DO APACHE SPARK

SPARK SQL

Spark SQL é usado para processamento de dados estruturados

SPARK MLLIB

MLlib é uma biblioteca de aprendizado de máquina escalonável

SPARK STREAMING

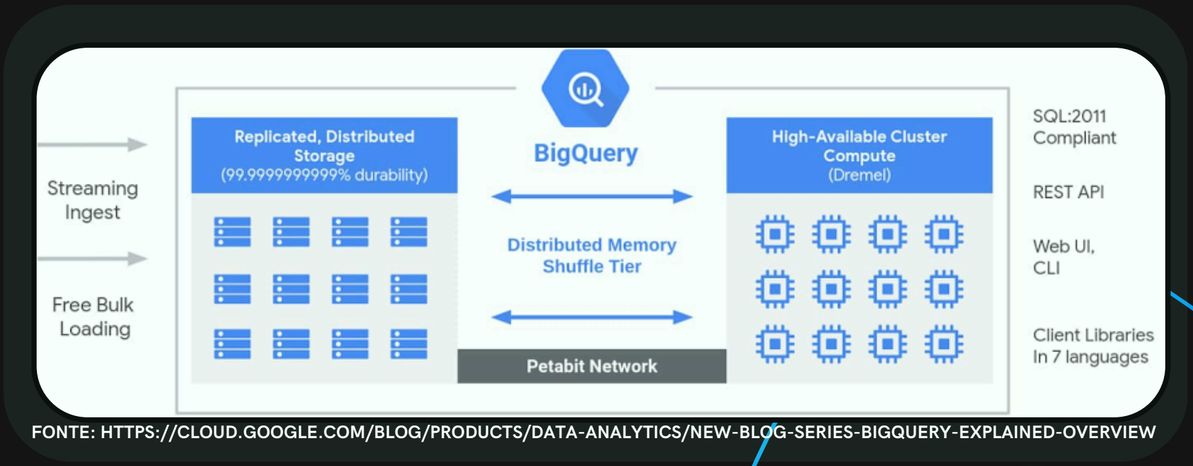
Spark Streaming possibilita o uso de poderosas aplicações interativas e analíticas em streaming e dados históricos

BigQuery

VISÃO GERAL E ARQUITETURA

O QUE É?

- Um *data warehouse* gerenciado pela Google Cloud.



FONTE: HTTPS://CLOUD.GOOGLE.COM/BLOG/PRODUCTS/DATA-ANALYTICS/NEW-BLOG-SERIES-BIGQUERY-EXPLAINED-OVERVIEW

PRINCIPAIS CARACTERÍSTICAS

- Armazenamento de dados *serverless*
- Armazenamento colunar
- Integrações
- Processamento Massivamente Paralelo (MPP)

BigQuery

COMPUTAÇÃO PARALELA NO BIGQUERY

DREMEL

- Divide consultas em uma árvore de execução paralela

COLOSSUS

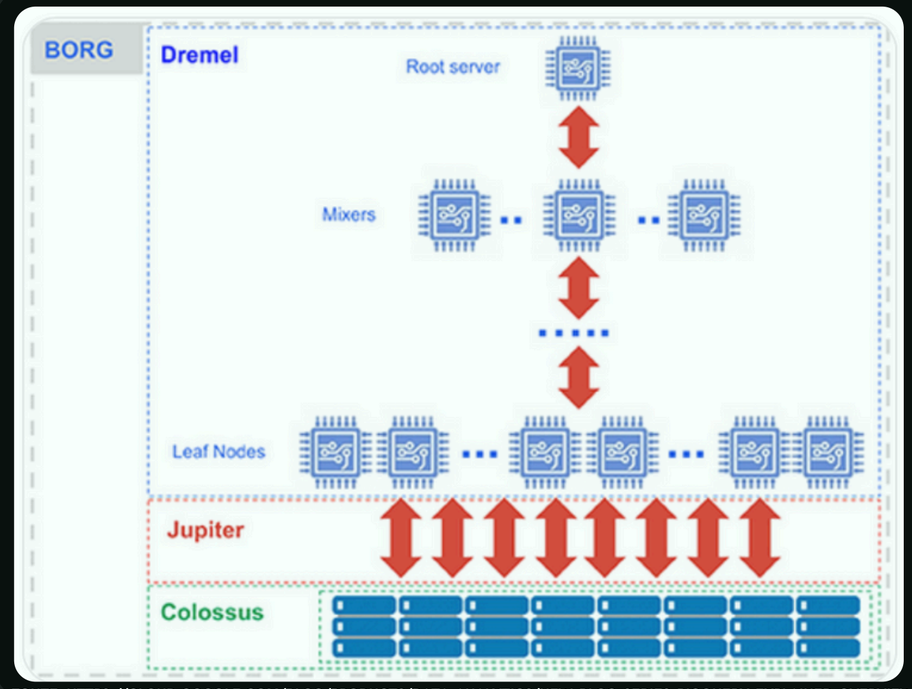
- Sistema de arquivos distribuído

JUPITER

- Rede de alta capacidade (ordem de 1 Petabit/s)

BORG

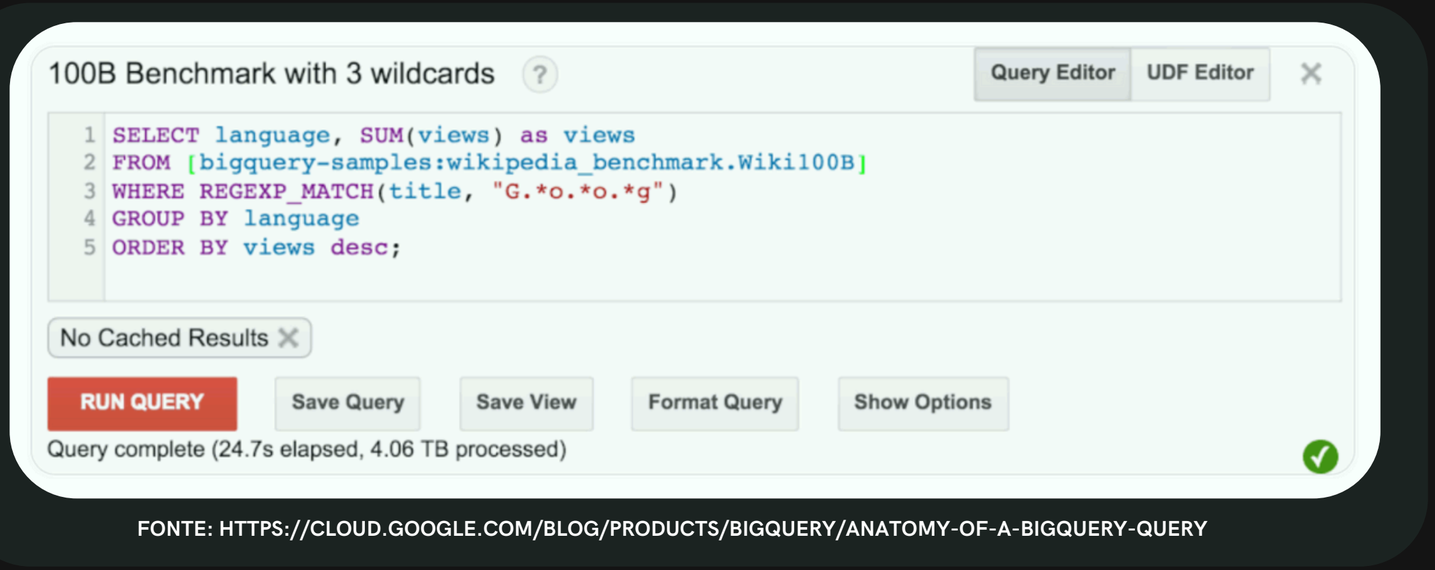
- Gerenciador de clusters que distribui tarefas em milhares de núcleos



FONTE: HTTPS://CLOUD.GOOGLE.COM/BLOG/PRODUCTS/DATA-ANALYTICS/NEW-BLOG-SERIES-BIGQUERY-EXPLAINED-OVERVIEW

BigQuery

DESEMPENHO



FONTE: HTTPS://CLOUD.GOOGLE.COM/BLOG/PRODUCTS/BIGQUERY/ANATOMY-OF-A-BIGQUERY-QUERY

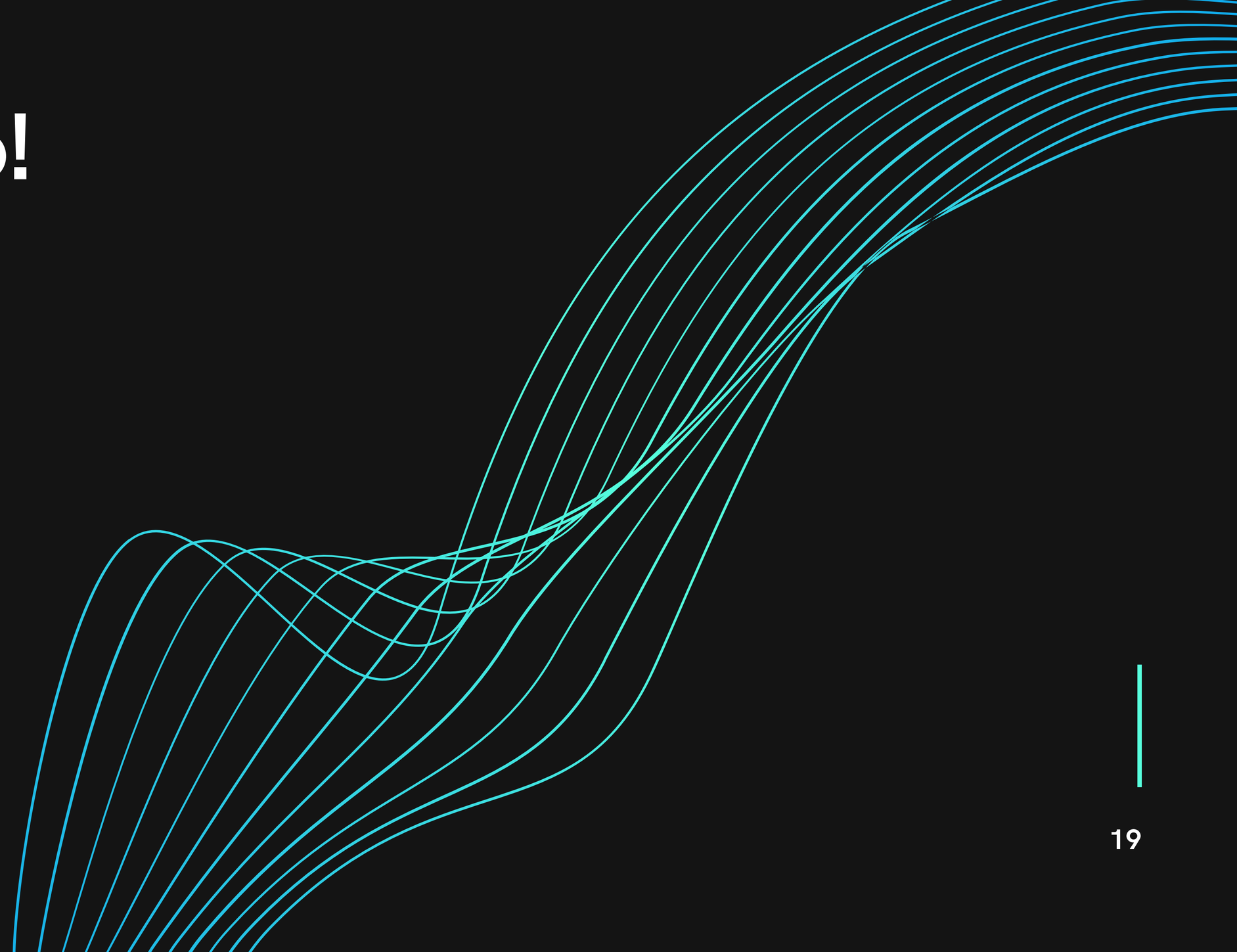
O QUE FOI FEITO?

- Ler cerca de 1TB de dados e desempacotá-los para 4 TB.
- Executar 100 bilhões de expressões regulares.
- Distribuir 1,25 TB de dados pela rede.

CLUSTER EQUIVALENTE

- Cerca de 330 discos rígidos dedicados de 100 MB/s.
- Uma rede de 330 Gigabits para transferir os 1,25 TB de dados.
- 3.300 núcleos.

Obrigado!



INTEGRANTES

• CARLOS HENRIQUE HANNAS DE CARVALHO	NUSP: 11965988
• CARLOS NERY RIBEIRO	NUSP: 12547698
• GABRIEL RIBEIRO RODRIGUES DESSOTTI	NUSP: 12547228
• LUCAS CARVALHO FREIBERGER STAPF	NUSP: 11800559
• PEDRO MANICARDI SOARES	NUSP: 12547621

REFERÊNCIAS

[1] Massively Parallel Processing - an overview | ScienceDirect Topics. Disponível em: https://www.sciencedirect.com/topics/computer-science/massively-parallel-processing. Acesso em: 15 nov. 2024.

[2] What is Massively Parallel Processing? | TIBCO. Disponível em: https://www.tibco.com/glossary/what-is-massively-parallel-processing. Acesso em: 15 nov. 2024.

[3] M. Madiajagan, S. Sridhar Raj, Chapter 1 - Parallel Computing, Graphics Processing Unit (GPU) and New Hardware for Deep Learning in Computational Intelligence Research, Editor(s): Arun Kumar Sangaiah, 2019, ISBN 9780128167182, https://doi.org/10.1016/B978-0-12-816718-2.00008-7.
(https://www.sciencedirect.com/science/article/pii/B9780128167182000087)

[4] X. Li, G. Zhang, K. Li, W. Zheng, Chapter 4 - Deep Learning and Its Parallelization, Editor(s): Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi, 2016, ISBN 9780128053942, https://doi.org/10.1016/B978-0-12-805394-2.00004-0.
(https://www.sciencedirect.com/science/article/pii/B9780128053942000040)

REFERÊNCIAS

[5] SILVEIRA, César L. B.; SILVEIRA JUNIOR, Luiz G. da; CAVALHEIRO, Gerson Geraldo H.. Programação em OpenCL: Uma introdução prática. Pelotas - Rs: Universidade Federal de Pelotas, 2009. 33 slides, P&B. Disponível em: http://www.inf.ufsc.br/~bosco.sobral/ensino/ine5645/Programacao_OpenCL_Introd_Pratica.pdf. Acesso em: 15 nov. 2024.

[6] O que é Apache Spark? Disponível em: https://cloud.google.com/learn/what-is-apache-spark?hl=pt-BR. Acesso em: 15 nov. 2024

[7] GOOGLE. New blog series: BigQuery explained - Overview. Disponível em: https://cloud.google.com/blog/products/data-analytics/new-blog-series-bigquery-explained-overview. Acesso em: 14 nov. 2024.

[8] GOOGLE. BigQuery under the hood. Disponível em: https://cloud.google.com/blog/products/bigquery/bigquery-under-the-hood. Acesso em: 14 nov. 2024.

[9] GOOGLE. Anatomy of a BigQuery query. Disponível em: https://cloud.google.com/blog/products/bigquery/anatomy-of-a-bigquery-query. Acesso em: 14 nov. 2024.