Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise Quantitativa sobre o Desempenho na Indexação de *Big Geospatial Data* em Ambiente de Nuvem Computacional

João Bachiega Jr.

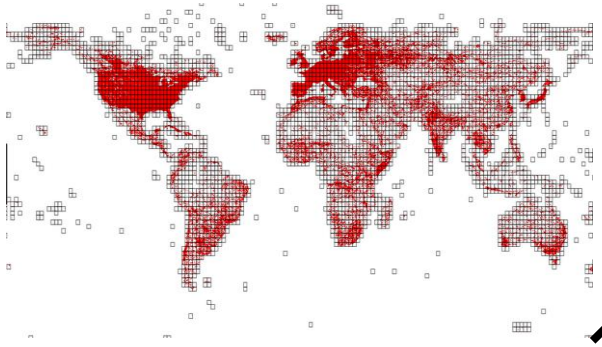
Seminário da Disciplina
Métodos Quantitativos em Computação

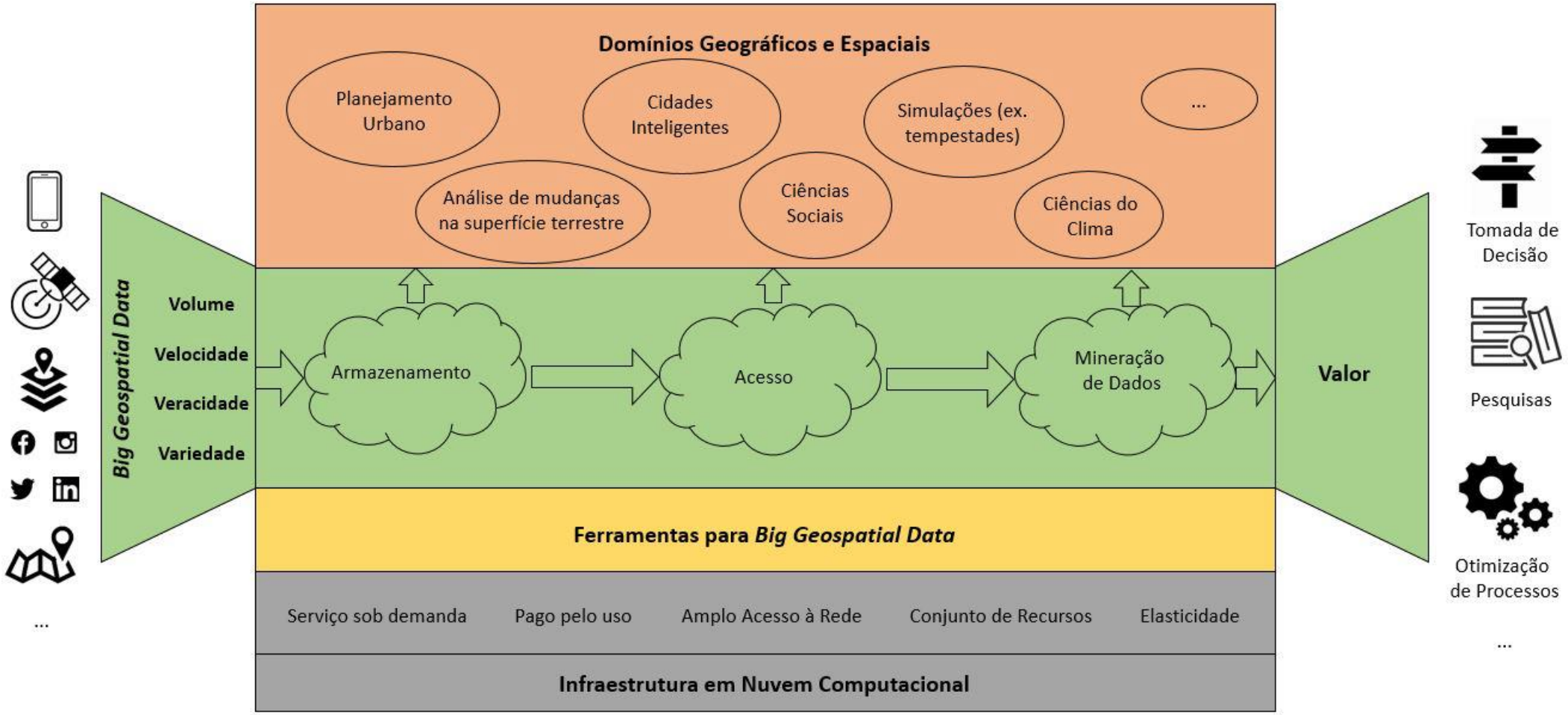
SUMÁRIO

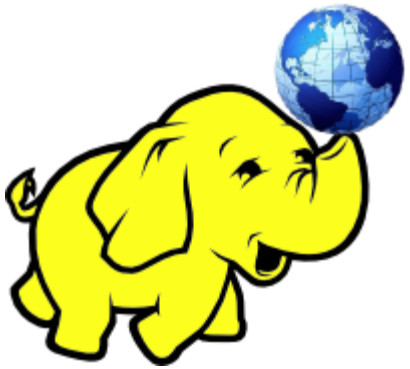
OBJETIVO

Fazer uma análise quantitativa sobre o desempenho da tarefa de indexação para grandes volumes de dados geográficos, executados em ambiente de nuvem computacional, realizando comparações entre tipos diferentes de indexações e entre configurações distintas de cluster.

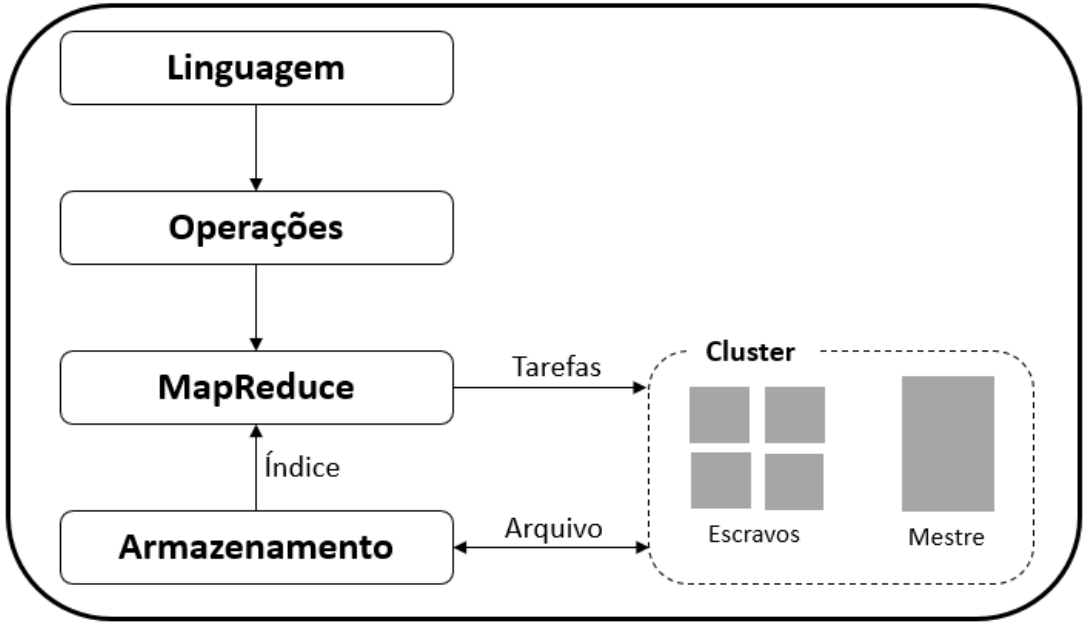
Big Geospatial Data

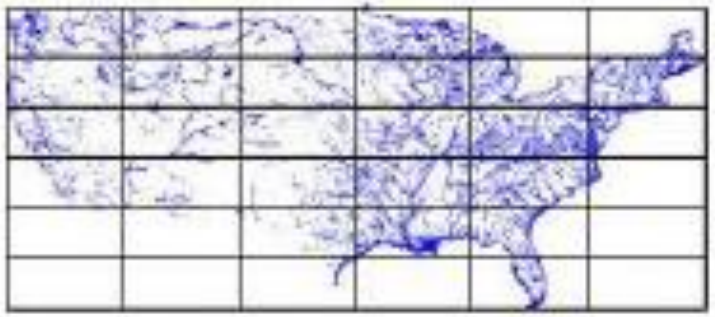



Domínios Geográficos e Espaciais
Planejamento Urbano
Cidades Inteligentes
Simulações (ex. tempestades)
...
Análise de mudanças na superfície terrestre
Ciências Sociais
Ciências do Clima
Big Geospatial Data
Volume
Velocidade
Veracidade
Variedade
Armazenamento
Acesso
Mineração de Dados
Valor
Tomada de Decisão
Pesquisas
Otimização de Processos
...
Ferramentas para Big Geospatial Data
Serviço sob demanda
Pago pelo uso
Amplo Acesso à Rede
Conjunto de Recursos
Elasticidade
Infraestrutura em Nuvem Computacional

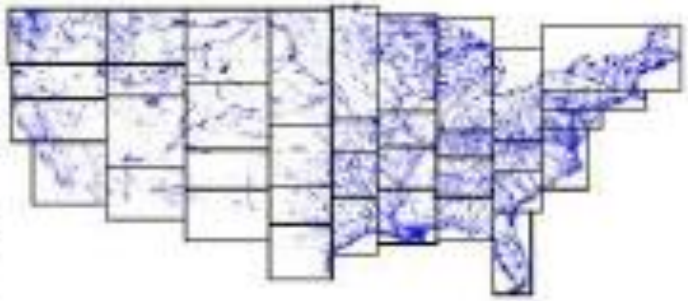


É um *framework* que está incorporado no *Hadoop*, ou seja, implementa as funcionalidades espaciais no interior do núcleo do *Hadoop*, tornando-se mais eficiente no processamento de consultas espaciais.





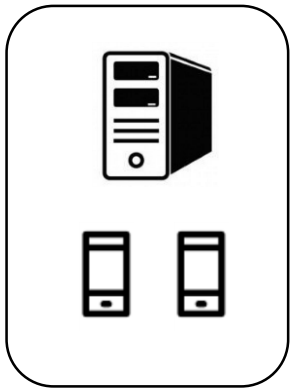
Grid



R-Tree

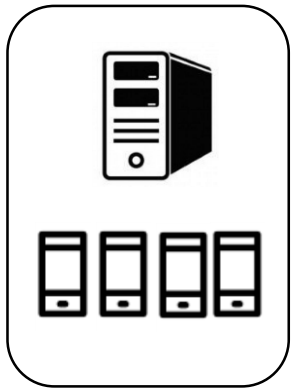
INFRAESTRUTURA

Elastic Map Reduce



1 masternode

2 datanodes



1 masternode

4 datanodes

BASES DE DADOS

Dataset	Conteúdo	Qtde. Registros
BUILDINGS	Contornos de construções no mundo	115 milhões
WAYS	Estradas mapeadas no mundo	164 milhões
RAILS	Ferrovias mapeadas no mundo	181 milhões
OBJECTS	Objetos geográficos mapeados no mundo	263 milhões

Teste 1: Intervalo de Confiança

Objetivo: Avaliar se o desempenho das indexações Grid e e R-Tree são significativamente diferentes.

Dados Coletados

Experimento	Grid (minutos)	R-Tree (minutos)
1	56	56
2	58	57
3	56	56
4	59	57
5	55	56

Resultados Obtidos

Métrica	*Grid*	*R-Tree*
Média	56,80	56,40
Desvio Padrão	1,64	0,55
Diferença das médias	0,40	
Desvio padrão das diferenças (s)	0,77	
IC	$0,40 \pm 1,65$	

Base: Rails – 181 milhões de registros
Cluster com 4 datanodes

TESTE 2: REGRESSÃO LINEAR

OBJETIVO: Estabelecer um modelo de regressão linear entre o tempo de execução e o tamanho da base de dados.

Dados Coletados

Base de Dados	Registros em milhões	Experimentos (em seg)				
		1	2	3	4	5
Buildings	115	51	50	51	50	50
Ways	164	76	76	76	77	75
Rails	181	79	79	80	80	79
Objects	263	86	87	87	87	87

Indexação Grid
Cluster com 2 datanodes

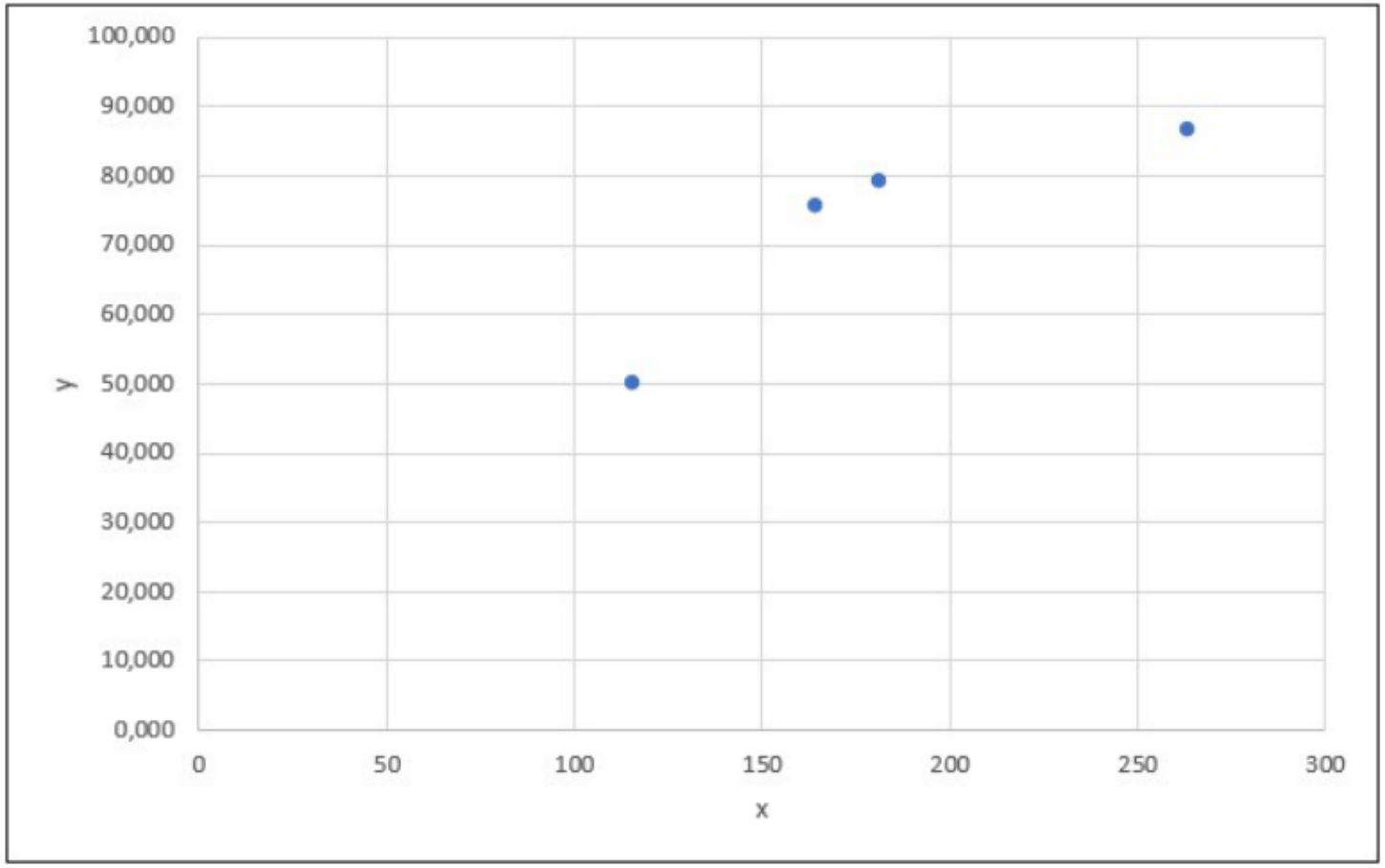
Resultados Obtidos

$n = 4$
$\overline{x} = 180,75$
$\overline{y} = 73,15$
$\Sigma xy = 55460,00$
$\Sigma x^2 = 142051,00$
$\Sigma x = 723,00$
$\Sigma y = 292,60$
$\Sigma y^2 = 22155,00$
$b_1 = 0,23$

$b_0 = 32,25$
$SST = 751,07$
$SSE = 169,04$
$R2 = 0,77$
$MSE = 84,52$
$se = 9,19$
$t = 1,886$
$s_{b_0} = 16,25$
$s_{b_1} = 0,09$

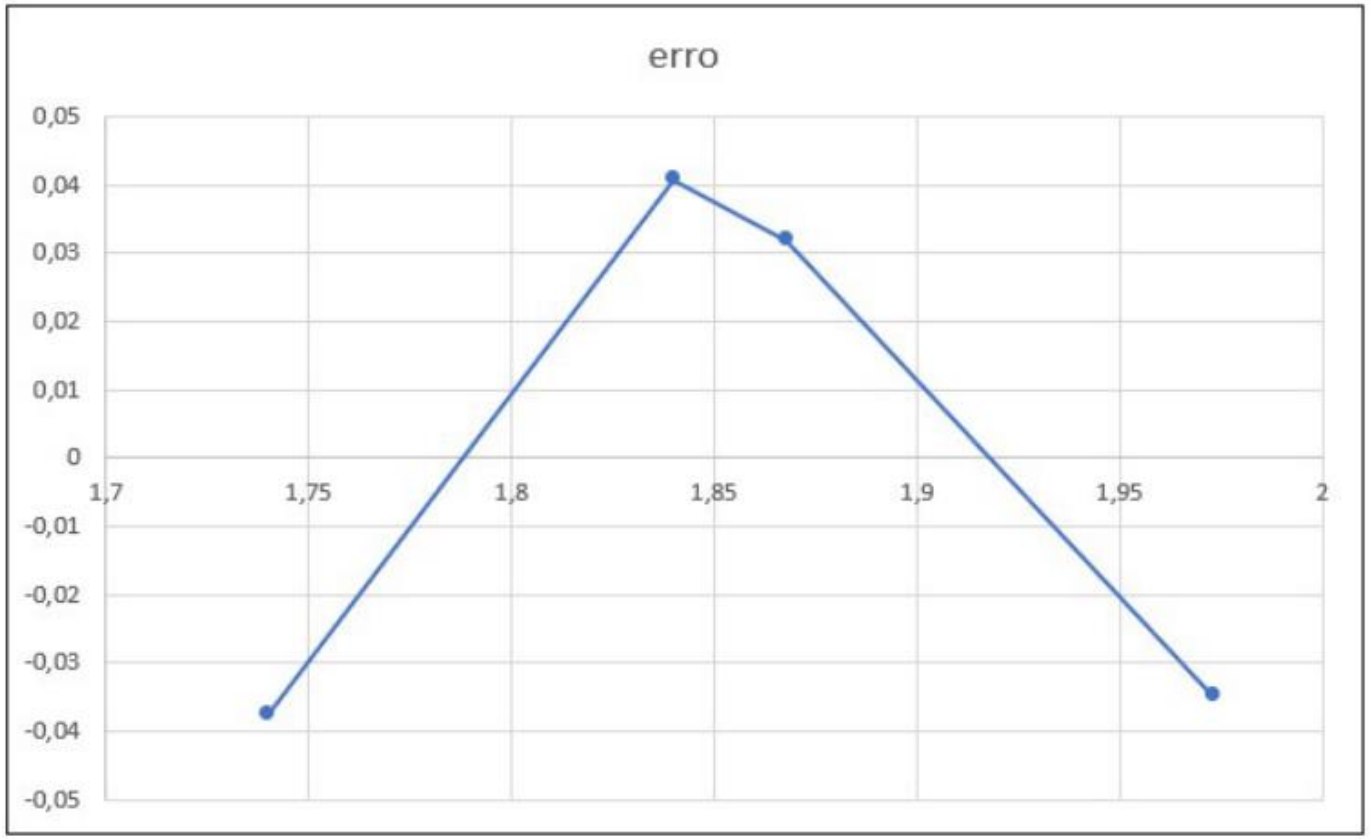
Teste 2: Regressão Linear

Mas...



Teste 2: Regressão Linear

Mas...



TESTE 3: FATORIAL 2^2

OBJETIVO: Determinar qual fator é mais significativo para o tempo de indexação.

Dados Coletados

	2 Datanodes (-1)	4 Datanodes (+1)
Baixo (-1)	59	40
Alto (+1)	90	56

Resultados Obtidos

$$q_0 = 61,25$$
$$q_A = -13,25$$
$$q_B = 11,75$$
$$q_{AB} = -3,75$$

O fator (X_A), que representa o número de *datanodes*, é o que tem o maior impacto (+/-13,25).

- A indexação tem papel fundamental no desempenho de aplicações que processam *Big Geospatial Data*;
- Não existem diferenças significativas entre as indexações *Grid* e *R-Tree*;
- Não foi possível estabelecer um modelo linear que seja capaz de estimar o tempo da indexação de uma base de dados geográfica considerando a quantidade de registros a ser processado;
- Na análise do Projeto Fatorial 2^2, a variação da quantidade de *datanodes* em um cluster *SpatialHadoop* é significativo para o tempo da indexação.

https://github.com/joaobachiegajr/MetQuant