Universidade Federal do Ceará

Mestrado e Doutorado em Ciência da Computação

Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem

Ticiana L. C. da Silva

Flávio R. C. Sousa

José Antônio F. de Macêdo

Javam C. Machado

Antônio Araújo Cavalcante

If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis.

– Prof. Hal Varian, UC Berkeley, Chief Economist at Google

Quem somos nós ?

Advanced Research in Database Group

Laboratorio de Sistemas e Banco de Dados

Fenômeno “Big Data”

Volume de Dados no Mundo

- 2000, 800 Terabytes
- 2006, 160 Exabytes
- 2009, 500 Exabytes(Internet)
- 2012, 2.7 Zettabytes
- 2020, 35 Zettabytes, 2020

Múltiplos do byte					
Prefixo binário (IEC)			Prefixo do SI		
Nome	Símbolo	Múltiplo	Nome	Símbolo	Múltiplo
byte	B	2^{0}	byte	B	10^{0}
kibibyte	KiB	2^{10}	Kilobyte	kB	10^{3}
mebibyte	MiB	2^{20}	megabyte	MB	10^{6}
gibibyte	GiB	2^{30}	gigabyte	GB	10^{9}
tebibyte	TiB	2^{40}	terabyte	TB	10^{12}
pebibyte	PiB	2^{50}	petabyte	PB	10^{15}
exbibyte	EiB	2^{60}	**exabyte**	EB	10^{18}
zebibyte	ZiB	2^{70}	zettabyte	ZB	10^{21}
yobibyte	YiB	2^{80}	yottabyte	YB	10^{24}

Quem está gerando esses dados?

12+ TBs
of tweet data

25+ TBs **of log data every day**

RFID tags today (1.3B in 2005)

camera phones world wide

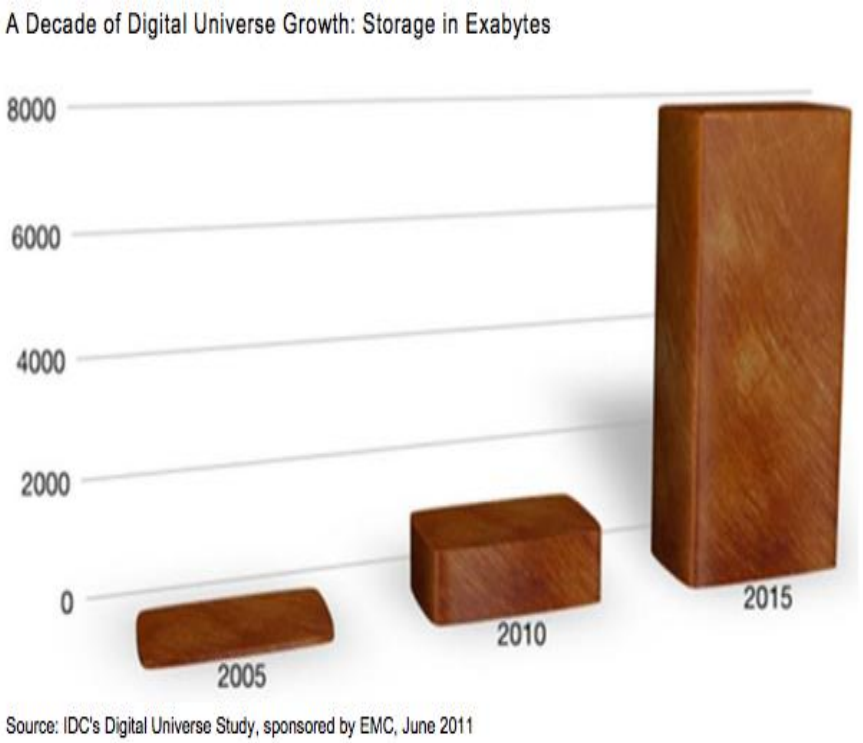
devices sold annually

smart meters in 2009… 200M by 2014

people on the Web by end 2011

IDC Digital Universe (1/2)

- "Extracting Value from Chaos" - a informação mundial está dobrando a cada 2 anos - 1.8 zettabytes foram criados em 2011, crescendo mais que a lei de Moore.



IDC Digital Universe (2/2)

”*As the volume and complexity of data barraging businesses from all angles increases, IT organizations have a choice: they can either succumb to information-overload paralysis, or they can take steps to harness the tremendous potential teeming within all of those data streams*”

Jeremy Burton, Executive Vice President, Product Operations and Marketing for EMC.

Big Data = Volume de Dados ?

Os 3 V's do Big Data

Mais V’s = Variability e Value

Além disso …

- Grande volume de dados está disponível para qualquer pessoa;

- Recursos computacionais estão tambem disponiveis – cloud, internet, dispositivos móveis, sensores, etc.

E ainda mais …

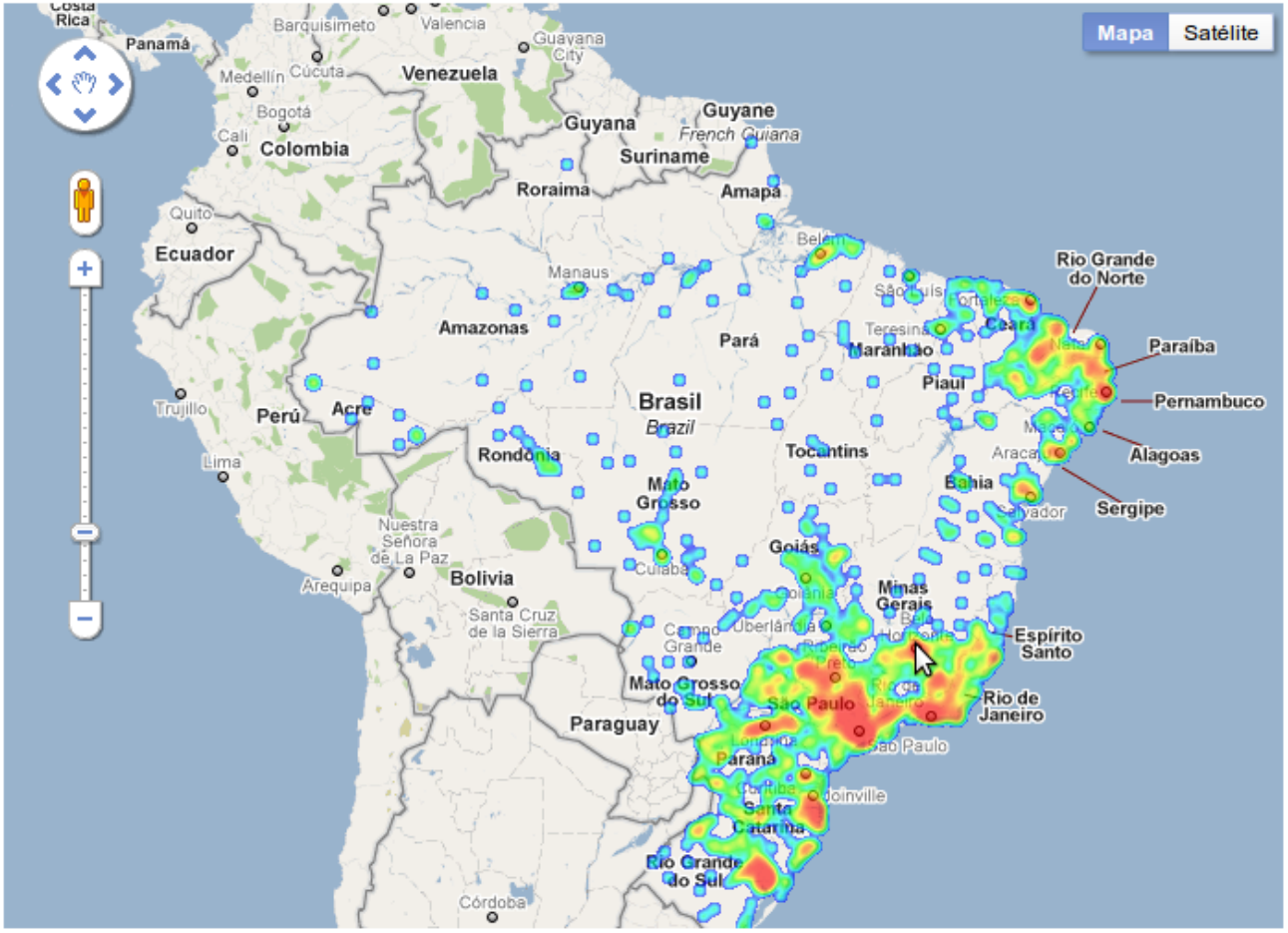
- Diferente das tentativas anteriores:
  - Computação Distribuída
  - Banco de Dados Distribuído
  - Computação em Grade
- Computação em nuvem é REAL:
  - Crescimento Orgânico: Google, Yahoo, Microsoft, e Amazon
  - Governos pretendem criar suas próprias infraestruturas

E se pudessemos analisar o “Big Data?”

Dengue Watch
Heatmap

observatóriodadengue

Menções à dengue no Twitter no mês de fev/2011

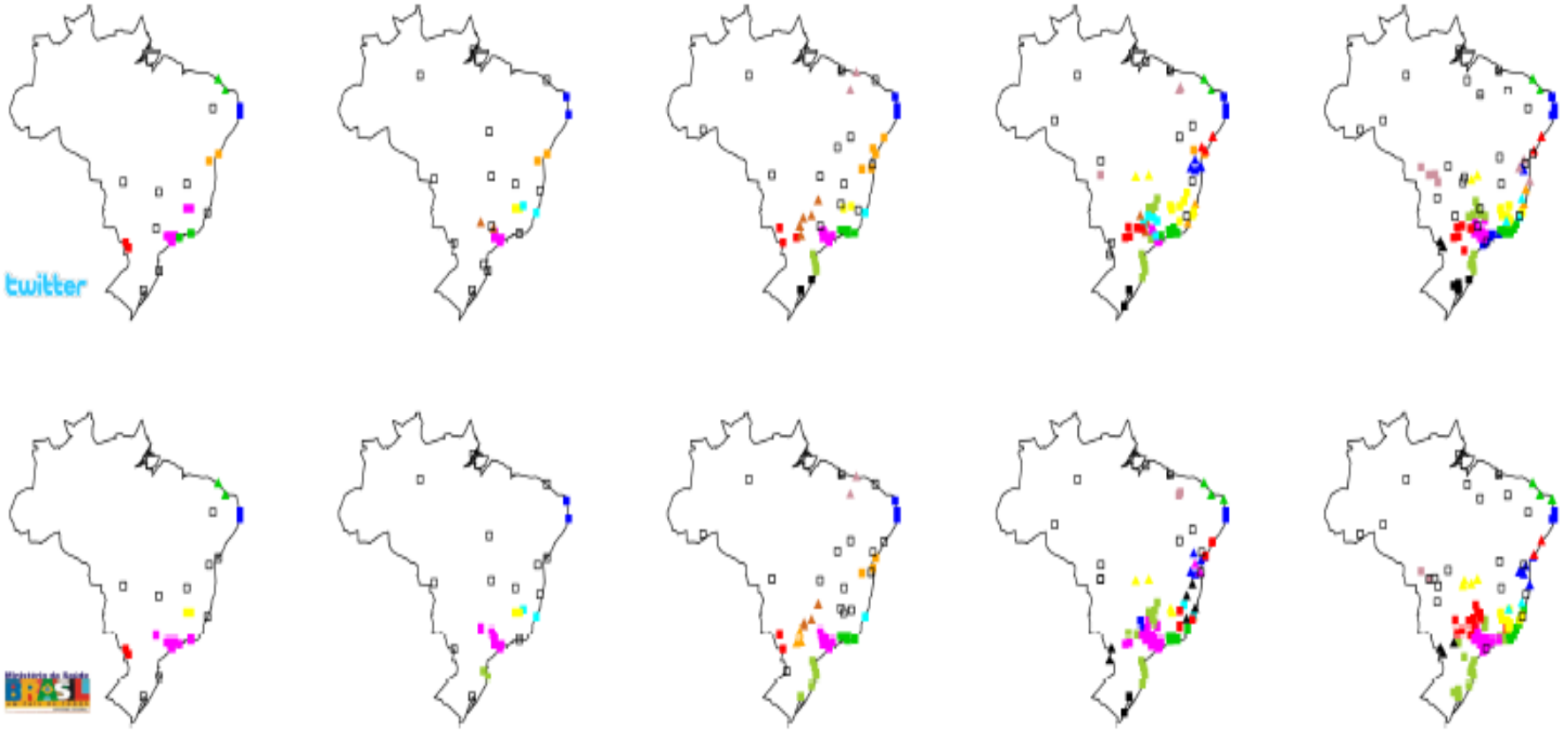


Clique nos pontos do mapa para informações

Cidades: 11 Tweets: 59 População: 1925450 Tx. Inc. Méd.: 1.5334e-04

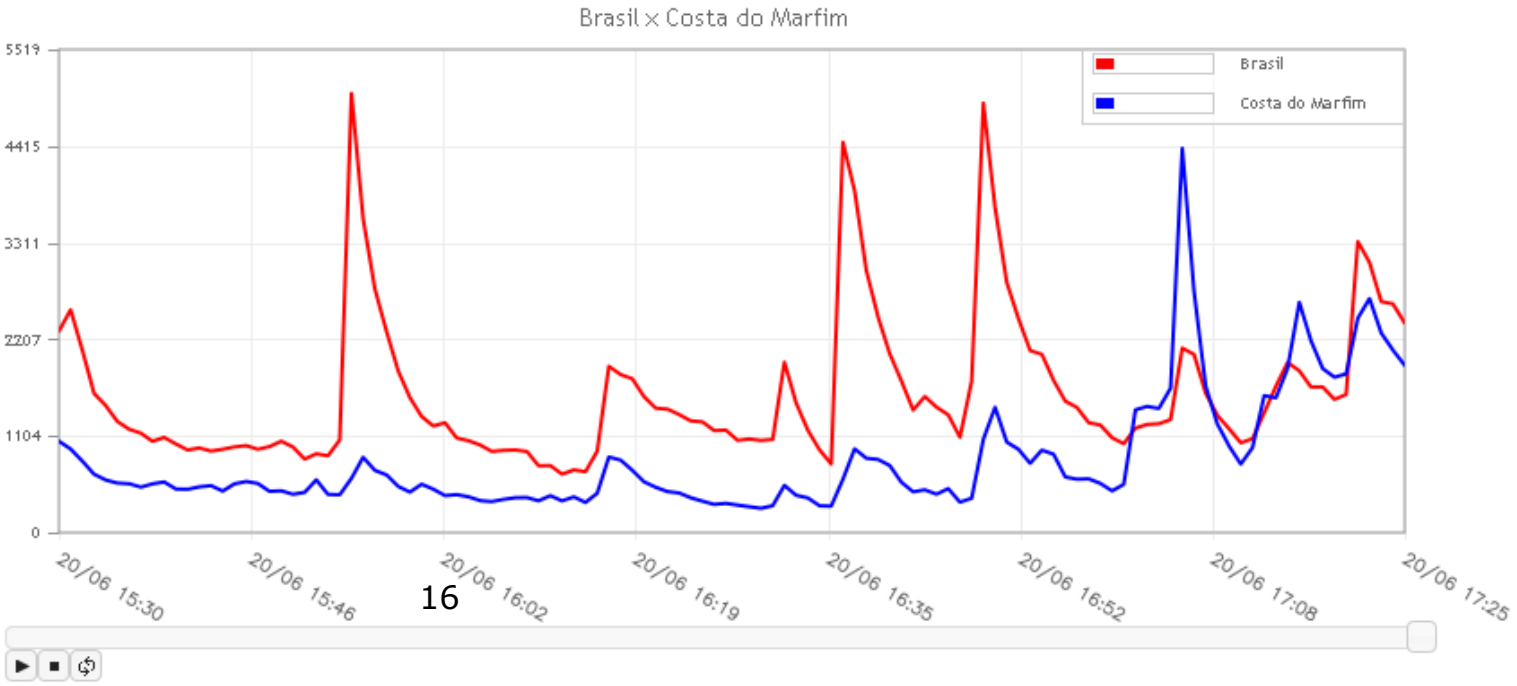
Cidade	Pop.	Tweets	Tx.Inc
Betim	377547	14	1.8230e-04
Brumadinho	34013	1	1.4289e-04
Contagem	603048	22	1.7922e-04
Ibirite	159026	4	1.2110e-04
Itabira	109551	6	2.7305e-04
Itauna	85396	1	5.2124e-05
Joao Monlevade	73451	4	2.7146e-04
Matozinhos	32973	1	1.4765e-04
Ribeirao das Neves	296376	2	2.6665e-05
Sabara	126219	3	1.1399e-04
Santa Barbara	27850	1	1.7627e-04

Dengue Surveillance
Twitter X Official Data



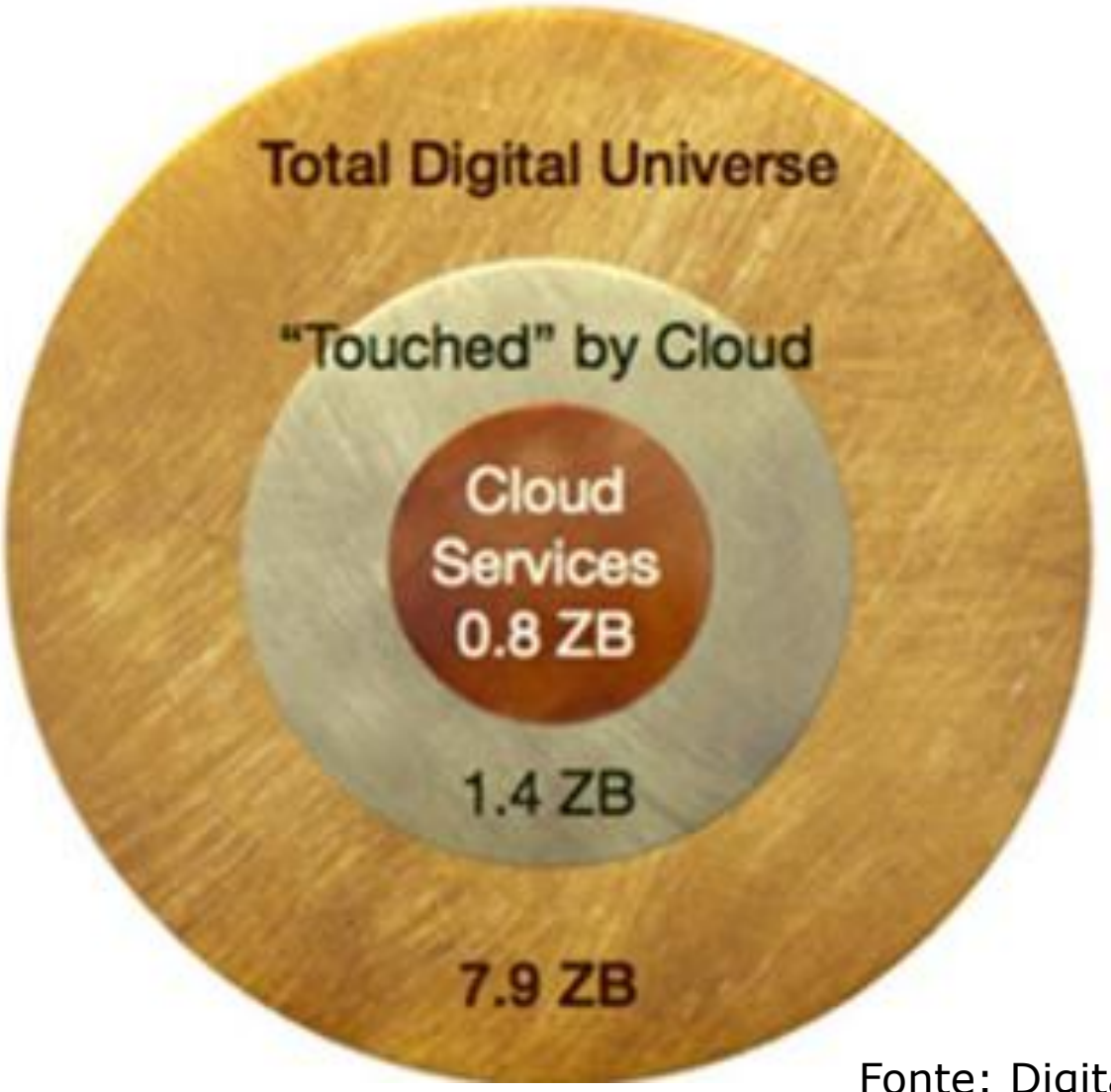
Fonte: Janaína Gomide, Adriano Veloso, Wagner Meira Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, Mauro Teixeira: Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. WebSci 2011: 1-8.

World Cup Watch
Twitter View of a Match



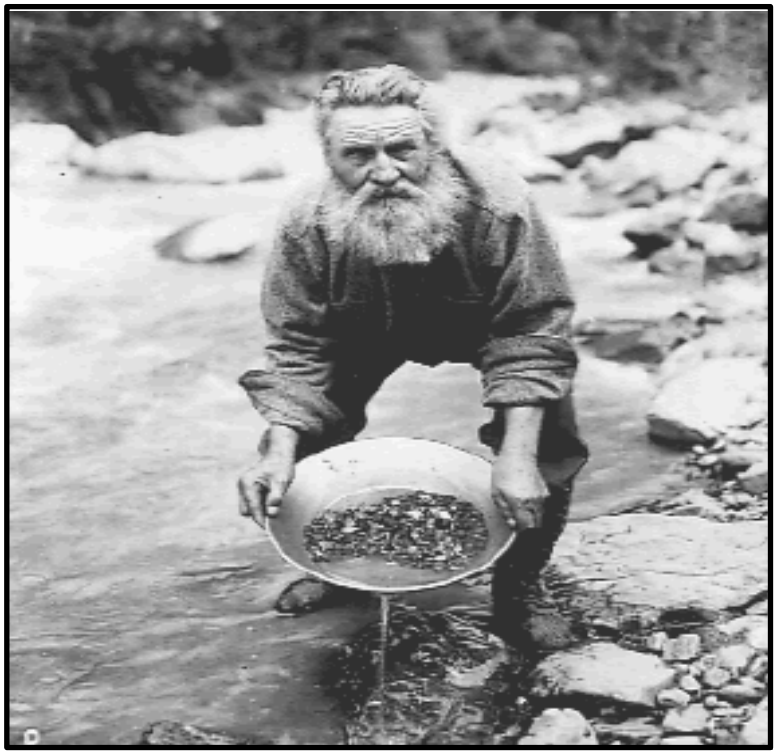
DESAFIOS

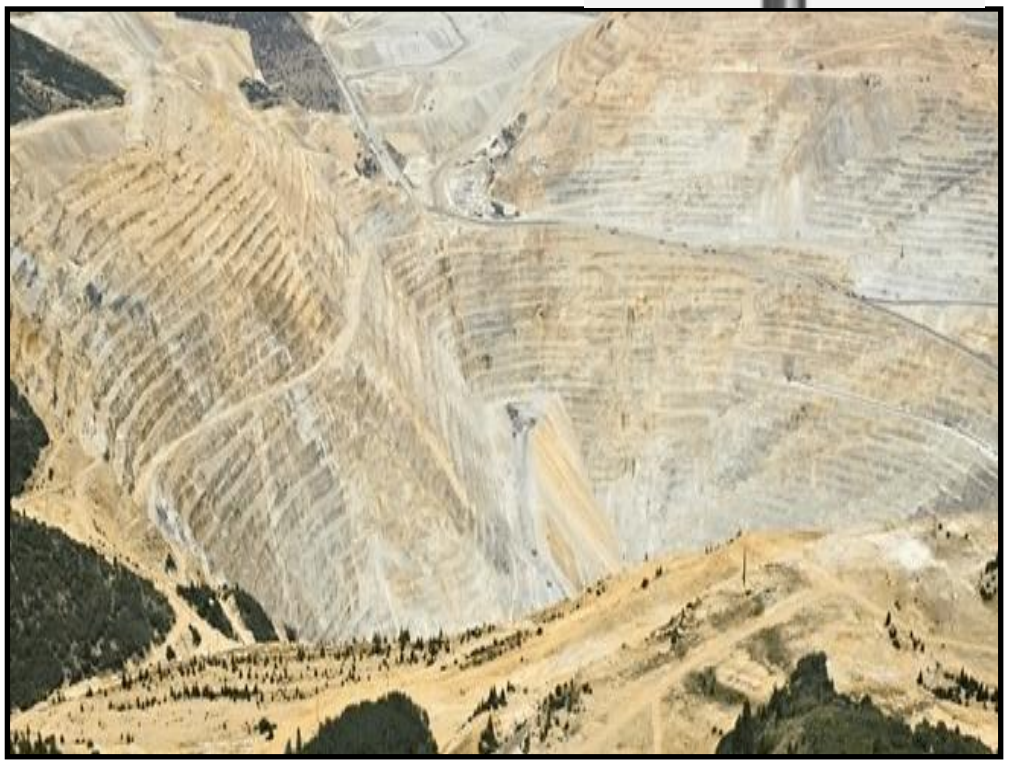
Existe muito dado intocado …



Fonte: Digital Universe Study

As ferramentas de análise são pobres …





Precisamos de ferramentas de análise sofisticadas …

- Lidem com 5 V’s do Big Data;
- Não mais limitada à sumarização e agregação de dados;
- Precisamos realizar análises complexas baseadas em mineração de dados e técnicas estatísticas;
- Precisamos urgentemente de técnicas de gerenciamento de Big Data;

Usar tecnologias maduras …

- 40 anos de pesquisa em BD;
- Tecnologias existentes lidam bem com:
  - Volume de Dados,
  - Variedade de Modelos;
  - Stream de Dados;
  - Fonte de dados heterogeneas e distribuídas;
- Porém, são necessárias novas técnicas:
  - Análise de Padrões Temporais;
  - Processamento em tempo real;
  - Incerteza, subjetividade e ambiguidade;

e adaptá-las ao novo contexto …

- Otimização baseada em I/O não deve ser uma premissa;
- Muitas análises precisam iterar sobre o dado diversas vezes;
- Devemos mover os programas e não os dados;
- Falha é uma regra e não uma exceção !

E ainda tem mais …

- Precisamos de algoritmos eficientes: O(n), O(log n)

O que podemos usar agora?

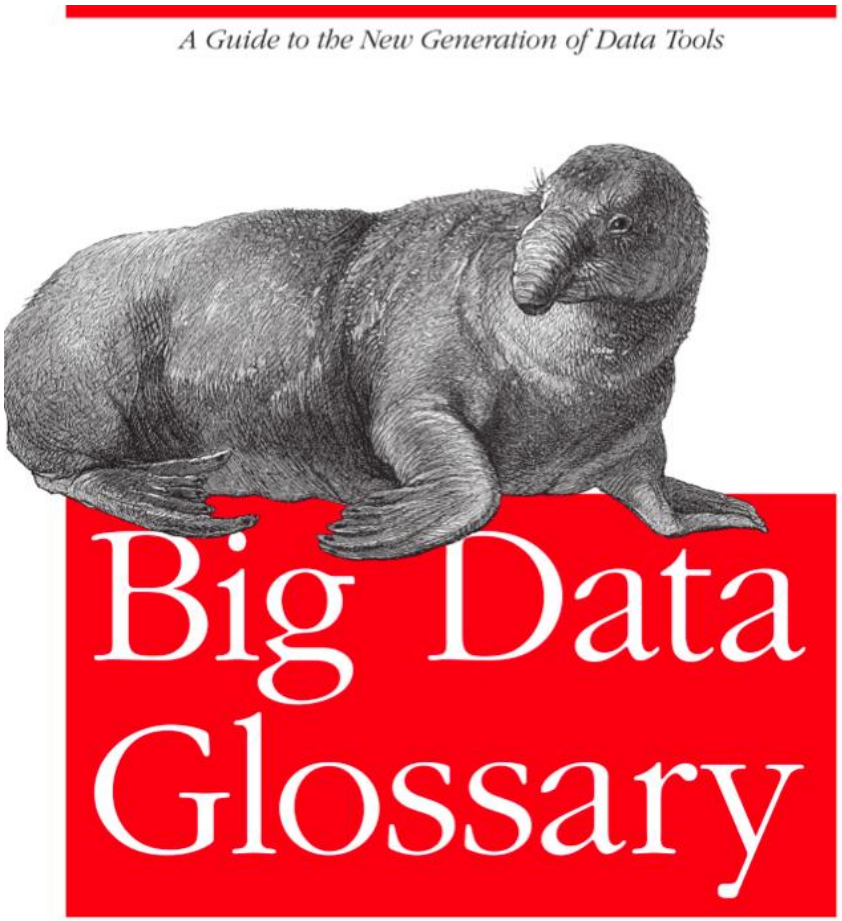
Análise de Dados na Cloud

- Escalabilidade para grandes volumes de dados:
  - Scan 100 TB em1 node @ 50 MB/sec = 23 dias;
  - Scan em 1000-node cluster = 33 minutos;
- Possibilidade de Dividir e Conquistar
  - Estratégias de particionamento dos dados;
- Commodities:
  - Nós (baratos, mas não confiáveis);
  - Rede;
  - Tolerância à falhas automática (poucos administradores);
  - Fácil de usar (menos programadores);

Plataforma para Analise de Dados em Larga Escala

- Tecnologias de Banco de Dados Paralelos:
  - Madura há mais de duas décadas;
  - Multi-billion dollar industry: Proprietary DBMS Engines intended as Data Warehousing solutions for very large enterprises;
- Map Reduce:
  - Criado pelo Google;
  - Popularizado pela Yahoo! (Hadoop);

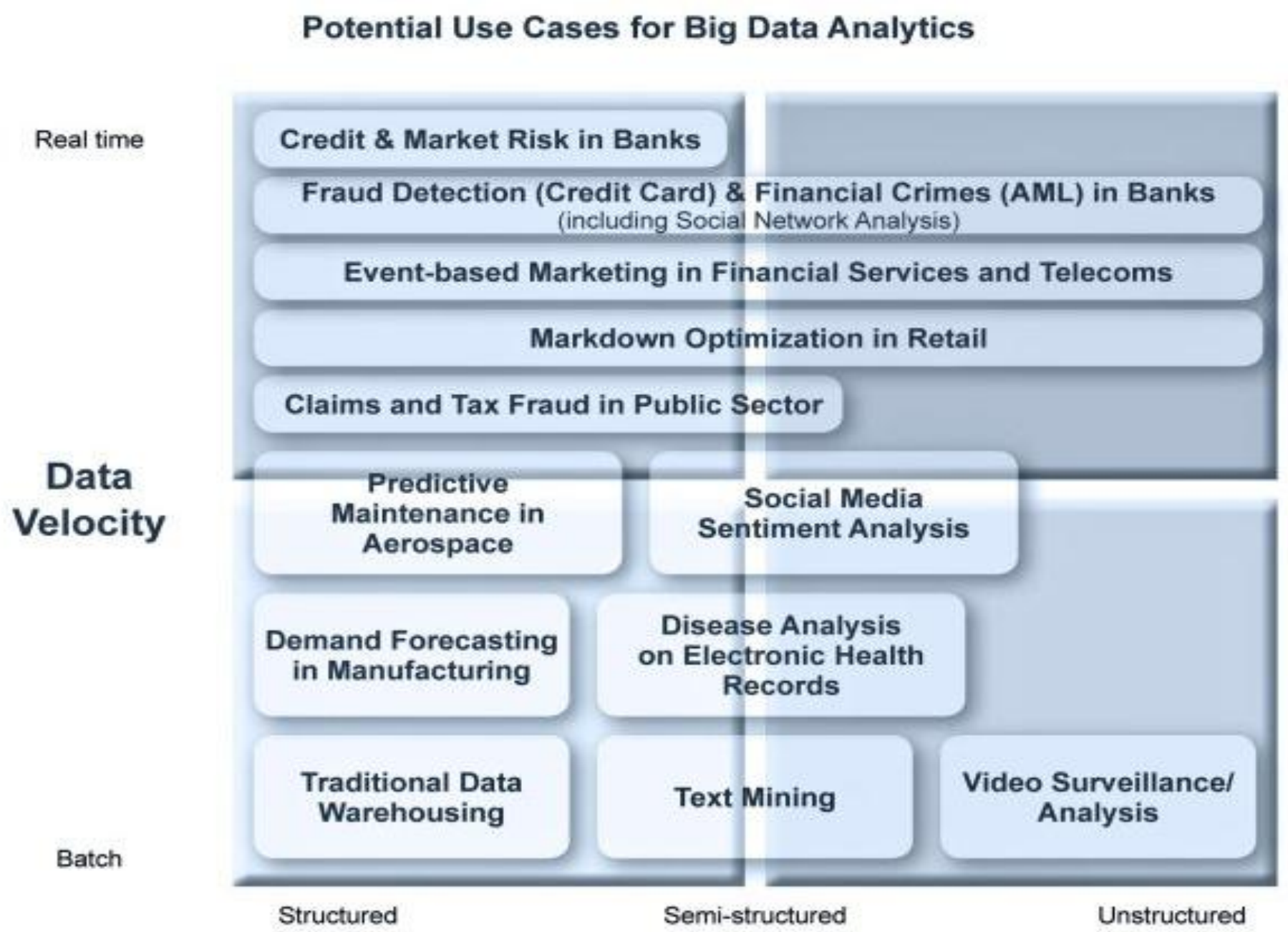
Tecnologias para Big Data



Tecnologias para Big Data

- NoSQL Databases
  - MongoDB, CouchDB, Cassandra
- Map Reduce
  - Hadoop, Hive, Pig
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine
- Processing
  - R, Yahoo! Pipes
- NLP
  - NL Toolkit, OpenNLP
- Machine Learning
  - WEKA, Mahout
- Visualization
  - Gephi, GraphViz
- Serialization
  - JSON, BSON

Oportunidades



Fonte: Alberto Laender Web Science, 2011 Presentation.

Novos Empregos !

“Em 2018, os EUA enfrentará uma escassez de 140.000 a 190.000 profissionais com capacidade de realizar análise detalhada, assim como existirá uma carência de 1.5 M de gerentes e analistas com conhecimento em análise de Big Data...”

McKinsey Report on Big Data

Muitas Oportunidades de Pesquisa

- Analysis and modeling of social behavior
- Social media: propagation and difusion
- Theory and models of social networks
- Link analysis and data mining
- Ranking algorithms
- Recommender systems
- Information extraction
- Replica identification
- Entity recognition
- Query log and clickstream analysis
- Collaborative filtering
- Sentiment analysis
- Opinion mining
- Human computation
- Crowdsourcing
- Security, privacy, risks and confidence
- Community computing
- Outros ...

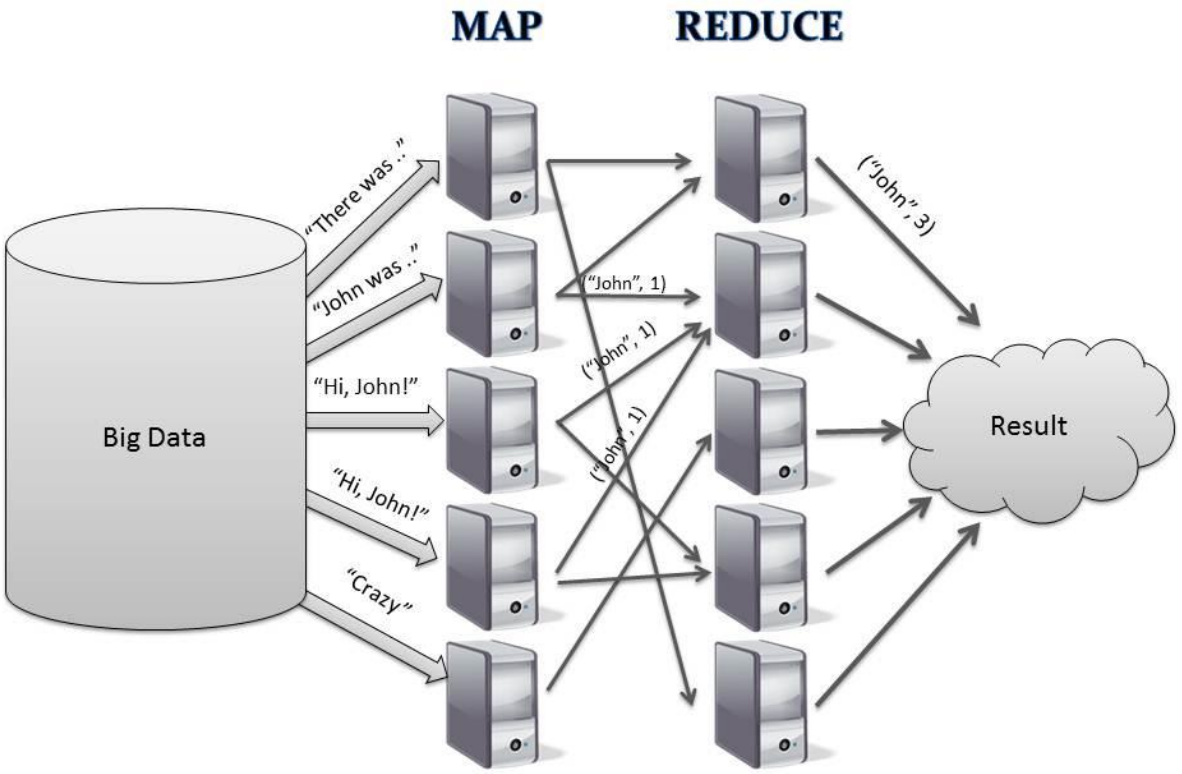
Algoritmos de Mineração de Dados em Big Data

Boas Práticas para Análise em Big Data

- Processamento centralizado não é adequados para Big Data;
- Objetivo: elencar um conjunto de boas práticas para que esses algoritmos sejam modificados e satisfatórios para o contexto de Big Data.

Boas Práticas para Análise em Big Data

- 1º príncipio: Uso do paralelismo e do paradigma de divisão e conquista [Ji et al. 2012], isso pode ser alcançado utilizando MapReduce.



Boas Práticas para Análise em Big Data

- 2º princípio: Distribuição/particionamento dos dados deve ser balanceado entre os nós do cluster, a fim de evitar sobrecarga em um nó em particular. É necessário nesse caso observar o desvio da distribuição do dado.

Boas Práticas para Análise em Big Data

- 3º princípio: É comum a utilização de replicação para paralelizar os algoritmos. No contexto de Big Data, é importante evitar replicação de dados, já que possivelmente pode resultar em crescimento exponencial da quantidade de dados a se processar.

Boas Práticas para Análise em Big Data

- 4º princípio: Os sistemas sobre os quais esse algoritmos executam também devem ser satisfatórios para a análise de dados Big Data [Cohen et al. 2009, Herodotou et al. 2011].
  - *Magnetic*
  - *Agile*
  - *Deep*
  - *Data-lifecycle- awareness*
  - *Elasticity*
  - *Robustness*

Boas Práticas para Análise em Big Data

- *Magnetic:*
  - Capacidade de manipular qualquer fonte de dados;
- *Agile:*
  - Adaptação a evolução dos dados;
- *Deep:*
  - Suporte a análises complexas;

Boas Práticas para Análise em Big Data

- *Data-lifecycle-awareness:*
  - Otimização da movimentação, armazenamento e processamento dos dados BigData;
- *Elasticity:*
  - Ajustar o uso dos recursos aos requisitos dos usuários e do processamento da carga de trabalho;
- *Robustness:*
  - Prover serviços mesmo com adversidades (falhas);

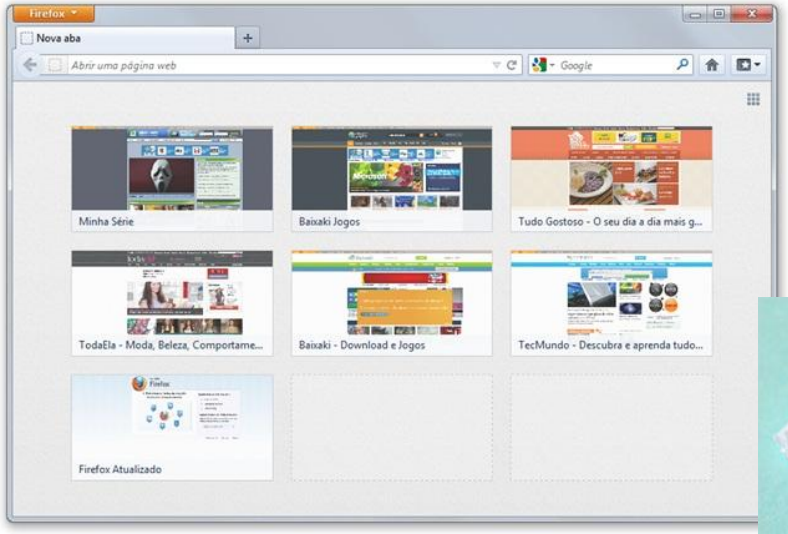
Boas Práticas para Análise em Big Data

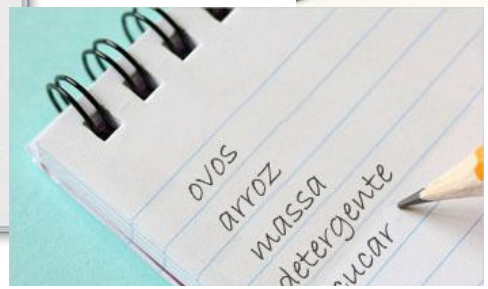
- Os trabalhos apresentados a seguir seguem alguns desses princípios, o que os torna adequados para grandes volumes de dados.

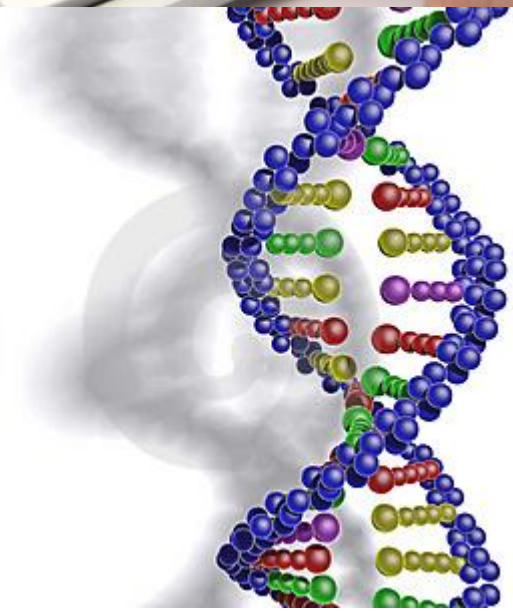
- *“Let’s start with the obvious observation: data intensive processing is beyond the capability of any individual machine and requires clusters—which means that large-data problems are fundamentally about organizing computations on dozens, hundreds, or even thousands of machines. This is exactly what MapReduce does…”*

Itens Frequentes/Regras de Associação

- Identificar padrões e regras de associação entre os itens de dados, ou seja, descobrir quais itens aparecem juntos frequentemente nesse conjunto de dados.







Itens Frequentes/Regras de Associação

- Dado um conjunto de transações, encontre regras que predizem a ocorrência de um item baseado na ocorrência de outros itens na transação.

Lista de compras no supermercado

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Exemplos de Regras de Associação:

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Itens frequentes e Regras de Associação

- Apriori [Agrawal et al. 1994]
  - O tempo gasto pode ser exponencial [Yang et al. 2010];
- [Yang et al. 2010]
  - Apriori paralelizado via MapReduce;
  - Múltiplas iterações de MapReduce;
  - Uma iteração produz um item frequente, e deve-se continuar iterando até que não existam mais itens novos a serem adicionados;
  - Pode ser muito custoso para um grande volume de dados;

Itens frequentes e Regras de Associação

- [Riondato et al. 2012]
  - Não usa replicação de dados;
  - Amostras randômicas e pequenas do conjunto de dados;
  - Processamento nas amostras independentes e em paralelo;
  - Os resultados gerados são coletados, agregados e filtrados para prover uma única saída;
  - Minera subconjuntos randômicos do conjunto de dados, o resultado final é uma aproximação da solução real;

Itens frequentes e Regras de Associação

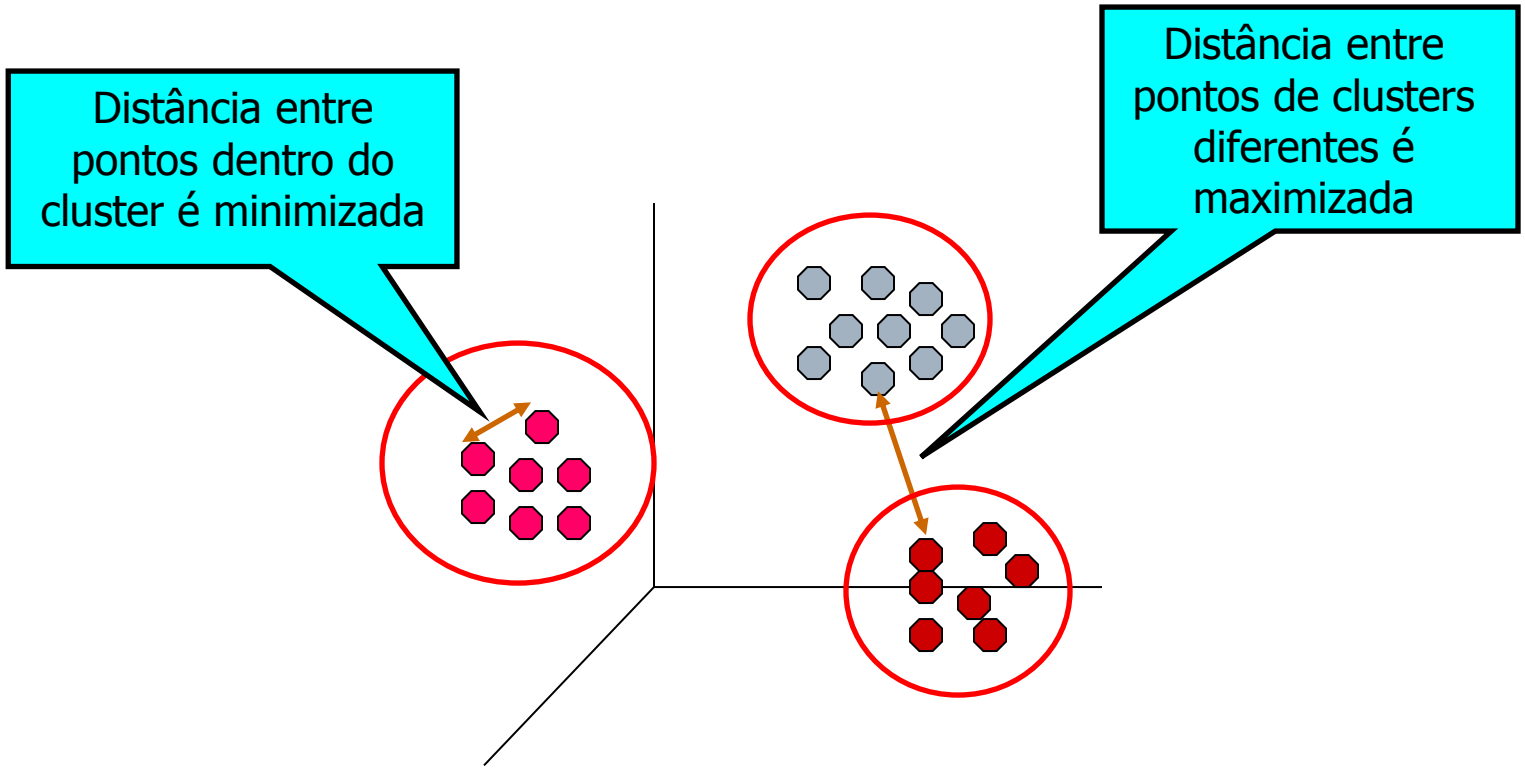
- [Riondato et al. 2012] difere dos outros [Cryans et al. 2010, Ghoting et al. 2011, Hammoud 2011, Li et al. 2008, Li and Zhang 2011, Yang et al. 2010]:
  - Não usa replicação de dados para paralelizar a computação;
  - A extração dos itens frequentes quanto a criação das regras são feitas via MapReduce.

Classificação

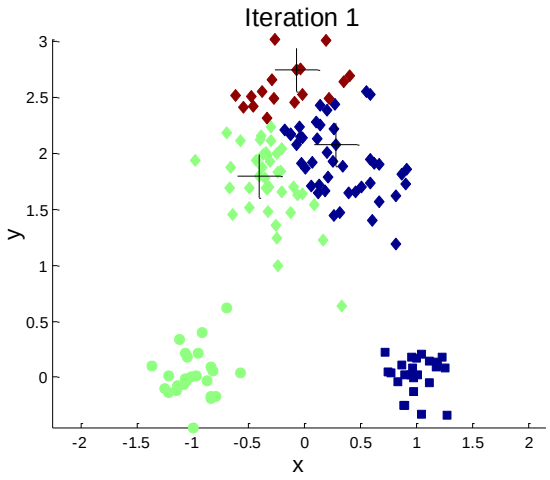
- Capazes de realizar predição de categorias;
- Mais conhecidos:
  - Vizinhos mais próximo [Cover and Hart 1967];
  - Rede bayesianas [Michie et al. 1994];
  - Árvore de decisão [Quinlan 1986];
- Pesquisas focaram em paralelizá-los para serem adequados a grandes volumes de dados;
- [He et al. 2010] propõe uma versão paralelizada para os 3 algoritmos;
- [Kim and Shim 2012], [Lu et al. 2012], [Ghoting et al. 2011], [Zhou et al. 2012], [Panda et al. 2009] aplicam/implementam classificação com MapReduce;

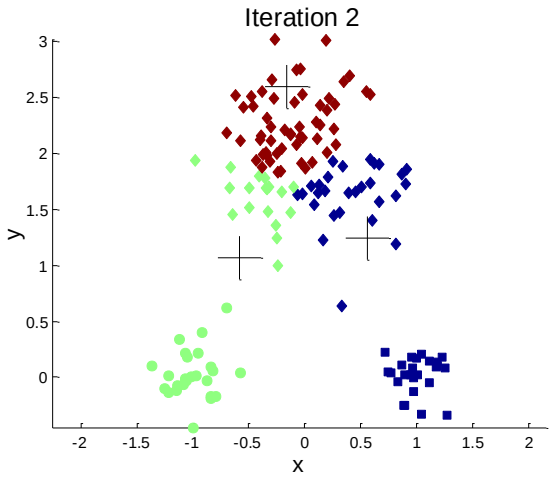
Clusterização

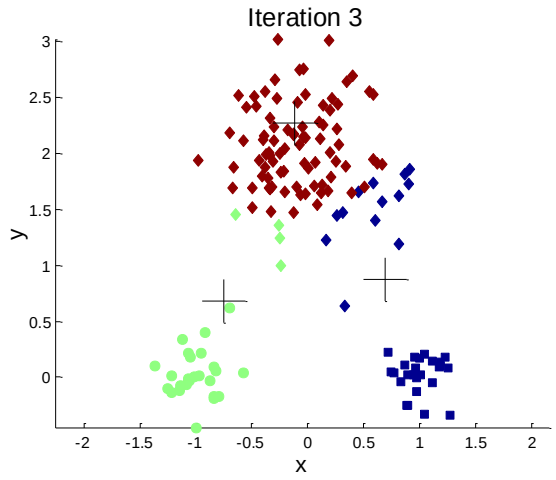
- Encontrar grupos de objetos;
- Objetos que pertencem a um mesmo grupo são similares(ou relacionados);
- Objetos que pertencem a grupos distintos são diferentes (ou não relacionados).

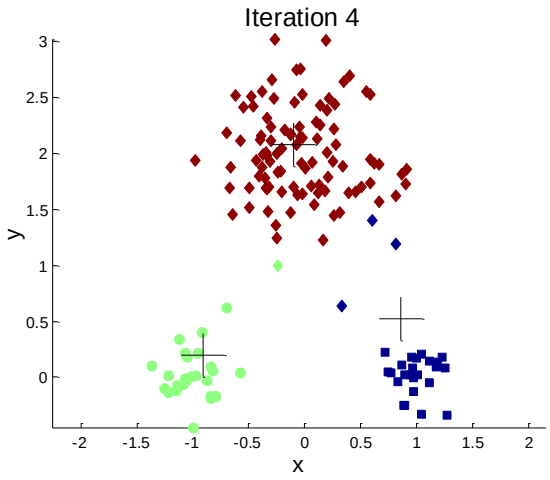


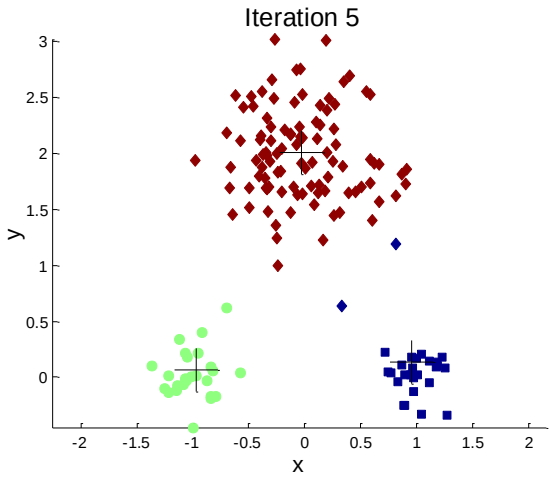
K-means

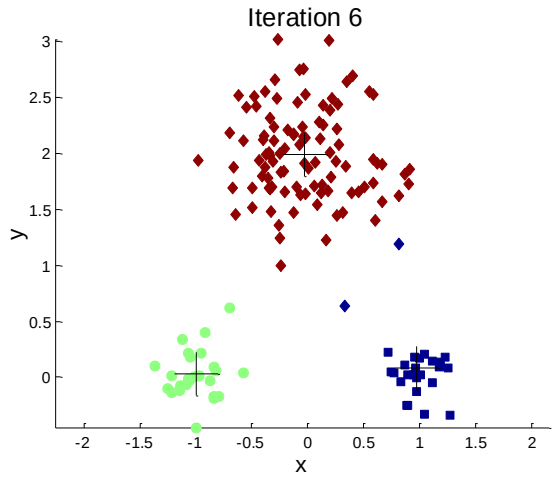

Iteration 1
x
y


Iteration 2
x
y

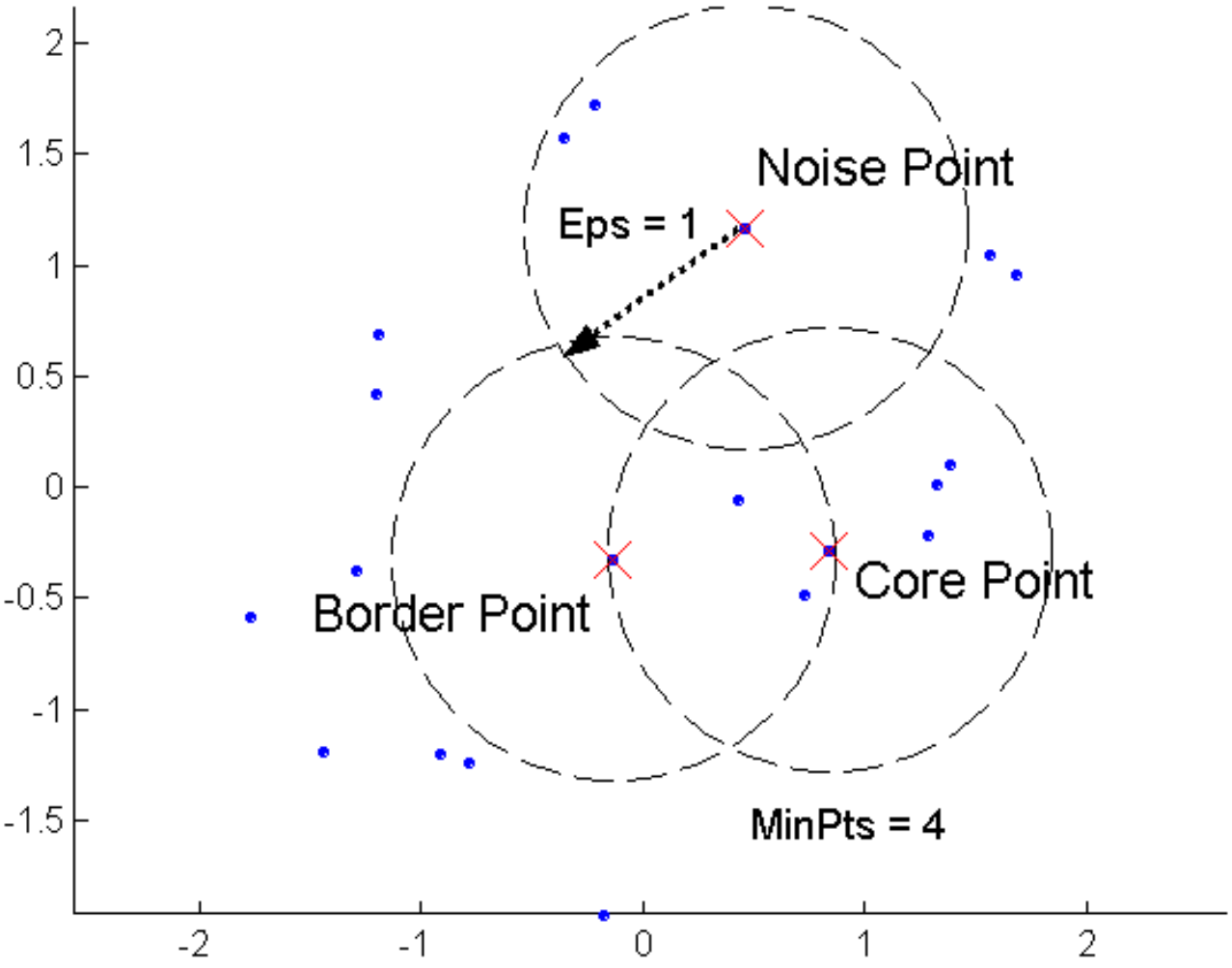

Iteration 3
x
y


Iteration 4
x
y


Iteration 5
x
y


Iteration 6
x
y

DBSCAN: Density Based Spatial Clustering Application with Noise



Clusterização: k-means

- Pode ser exponencial no pior caso;
- Solução pode ser de um ótimo local, e possivelmente longe de se obter um ótimo global;
- k-means++ [Arthur and Vassilvitskii 2007]
  - Evolução do k-means;
  - Oferece melhor estratégia de inicialização dos centróides;
- Aplicabilidade é limitada para grandes volumes de dados:
  - Deve-se iterar k vezes sobre os dados para encontrar um bom conjunto de centróides inicialmente;

Clusterização: k-means

- Ideal:
  - menor número de iterações;
  - Selecionar mais que um ponto em cada iteração da mesma maneira que o K-means++;
- Em [Bahmani et al. 2012]:
  - Estratégia paralela para reduzir o número de iterações necessárias, a fim de obter uma boa inicialização utilizando MapReduce;
  - Aplicabilidade possível para grandes volumes de dados;

Clusterização: DBScan

- [He et al. 2011] e [Dai and Lin 2012] seguem a mesma proposta implementação do DBScan usando MapReduce e particionamento baseado em grid;
- [Dai and Lin 2012]: particionamento centralizado para distribuir bem os dados;
- Estratégia para junção de clusters que estão em partições diferentes e que contém os mesmos pontos em suas fronteiras;
- Esses pontos são replicados;
- Replicação pode afetar a eficiencia da clusterização para grandes volumes de dados;
- **Nosso estudo de caso mostra uma implementação do DBScan usando MapReduce sem replicação para grandes volumes de dados;**

Infraestrutura para Análise em Big Data

Big Data

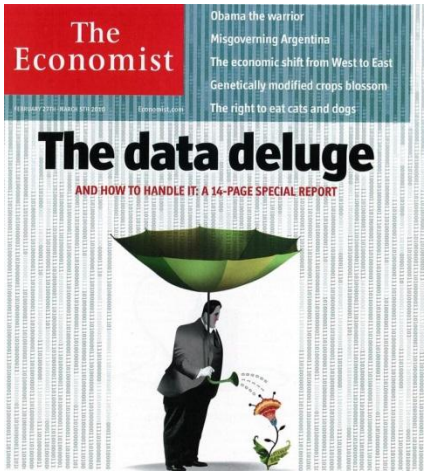
computing
p. 76 "Your Homework" Returns
BIG DATA

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

Science
11 February 2011 | $10
data

Harvard Business Review
GETTING CONTROL OF BIG DATA

CIÊNCIA HOJE
LUIZ WERNECK VIANNA
Câncer de pele
PEDRO NAVA
DESAFIOS DO
BIG DATA
Como escolher e guardar tanta informação?


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Big Data

- **Wikipedia** possui + de **3,5 milhões de páginas**
- **Yahoo!** + de **650M de usuários, 11B visitas a páginas/mês**
- **Flick** + de **5B de fotos**
- **Facebook**: **1B** de usuários, **1,13 Trilhões** de "likes", **219B** de fotos e **140.3B** de relacionamentos
- **Youtube**: **100 horas** de vídeos adicionado a **cada minuto**
- **Twitter**: **80 TB** e **1B** de tweets por dia
- **Boeing**: **640 TB** gerados em um voo transatlântico
- **Wal-Mart**: **2,5 PB** e **1 milhão** de transações/hora

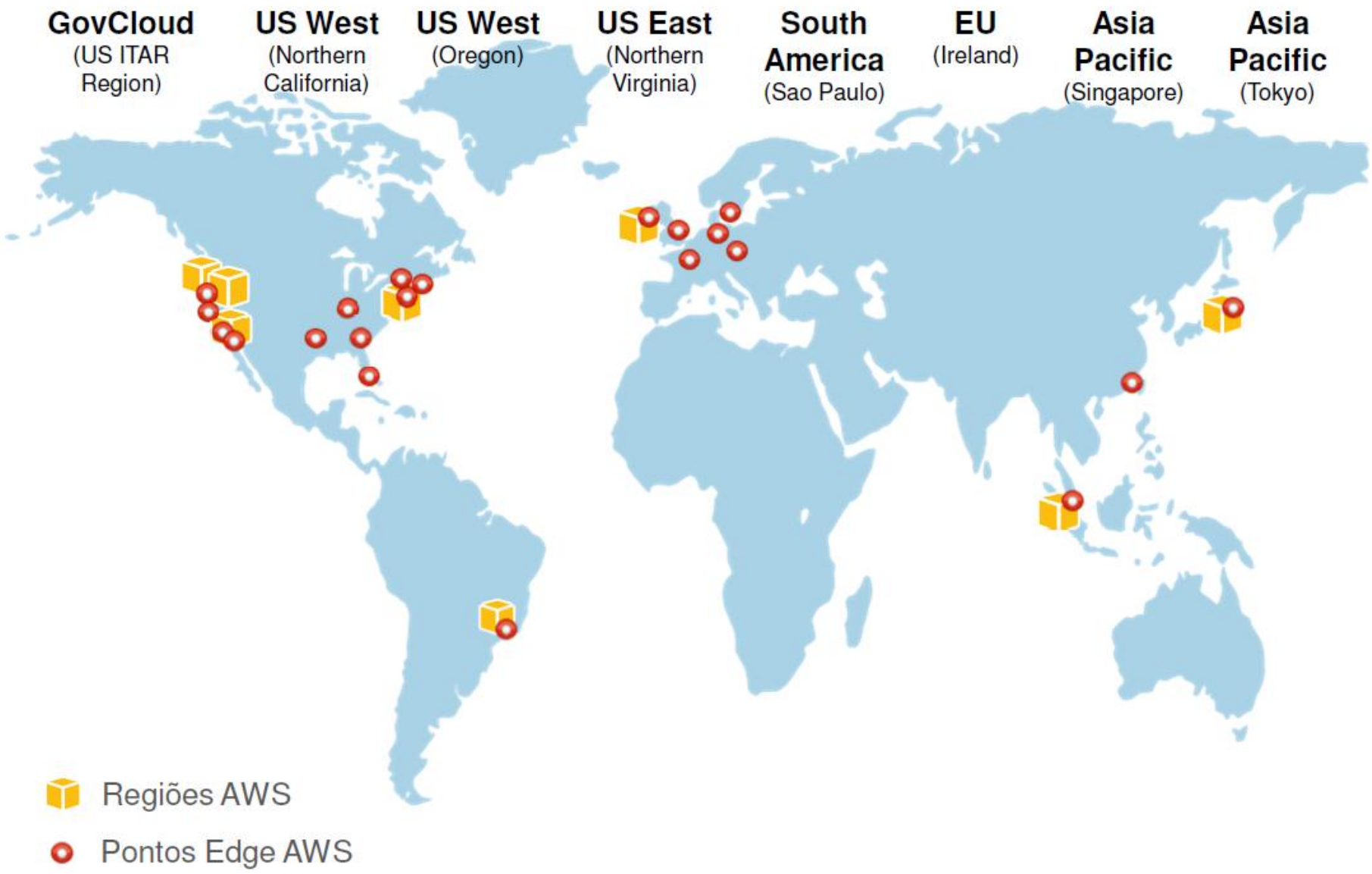
Onde e como armazenar, processar e analisar este volume de dados?

Computação em Nuvem

- Serviços básicos e essenciais são todos entregues de uma forma transparente
- A mesma ideia tem sido aplicada no contexto da informática
  - *Cloud Computing* ou Computação em Nuvem
- Computação em Nuvem
  - Ideia antiga: Software como um Serviço (SaaS)
    - Entrega de aplicações através da Internet
  - Recentemente: "[Hardware, Infraestrutura, Plataforma] como um serviço"
    - "X como um serviço"

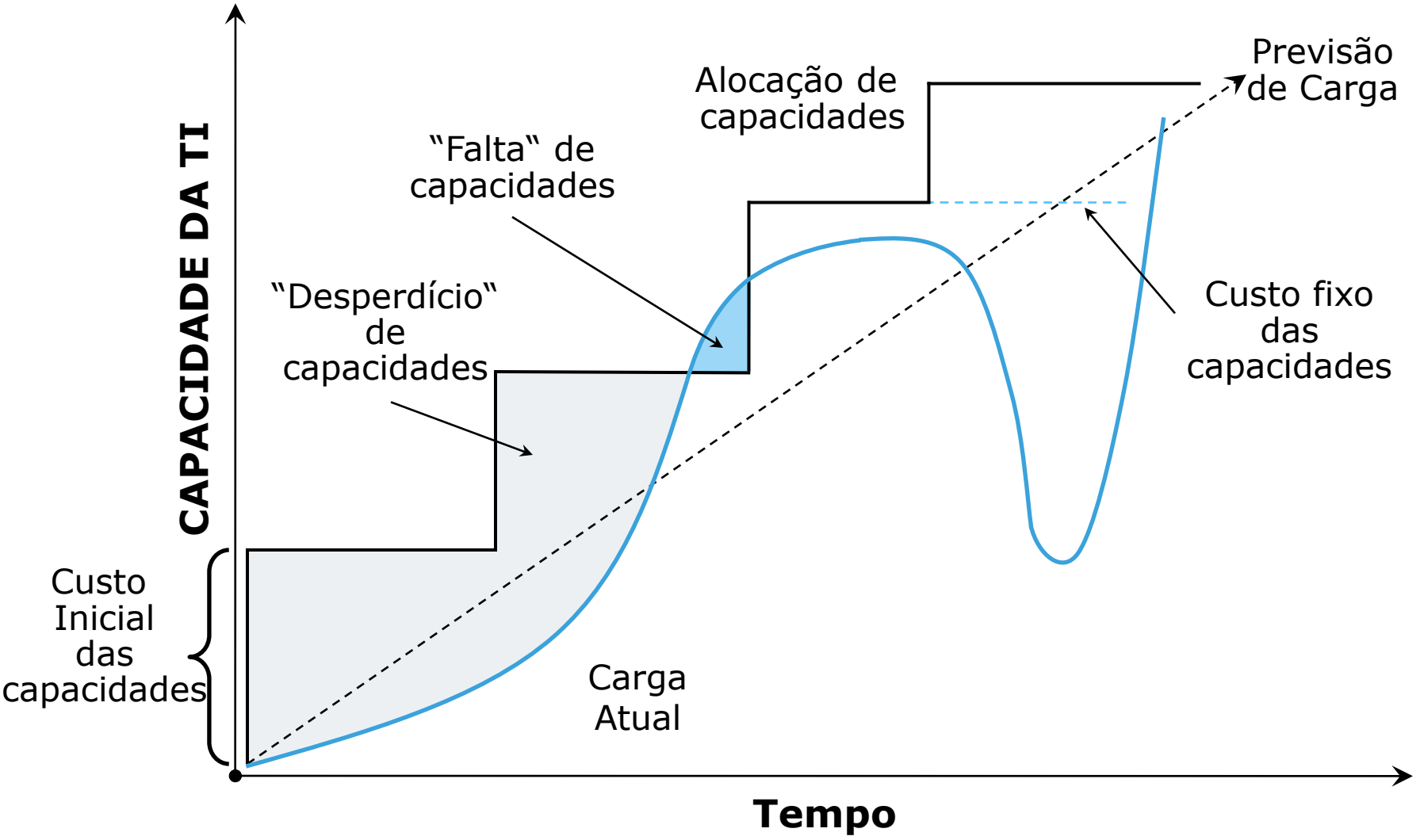
Computação em Nuvem

Computação em Nuvem: Amazon AWS



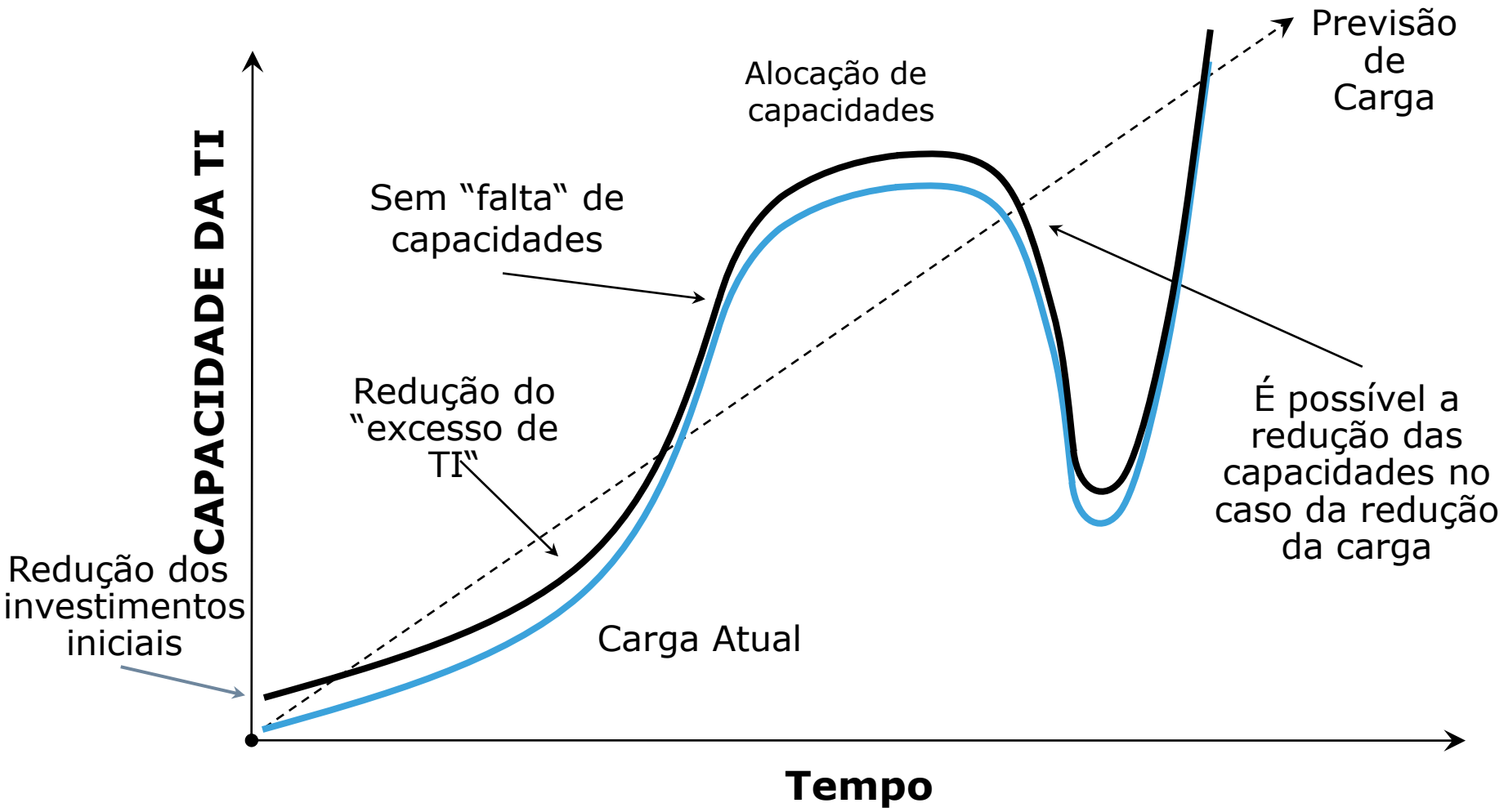
Computação em Nuvem: Promessas

Computação em Nuvem



Fonte: Microsoft

Computação em Nuvem: Elasticidade



Fonte: Microsoft

Computação em Nuvem: Ameaças

- **Privacidade**
  - Seus dados e de seus cliente são armazenados no provedor
- **Dependência** de provedor de nuvem
  - Sem controle físico sobre o hardware
  - E se governo decide encerrar a “nuvem”? ou obrigar o provedor a fornece informações?
  - E se o datacenter “quebra”?
- Limitado ao provedor (lock-in)
  - **Adaptando soluções** para serviços específicos (não-padrão)

Computação em Nuvem

Computação em Nuvem

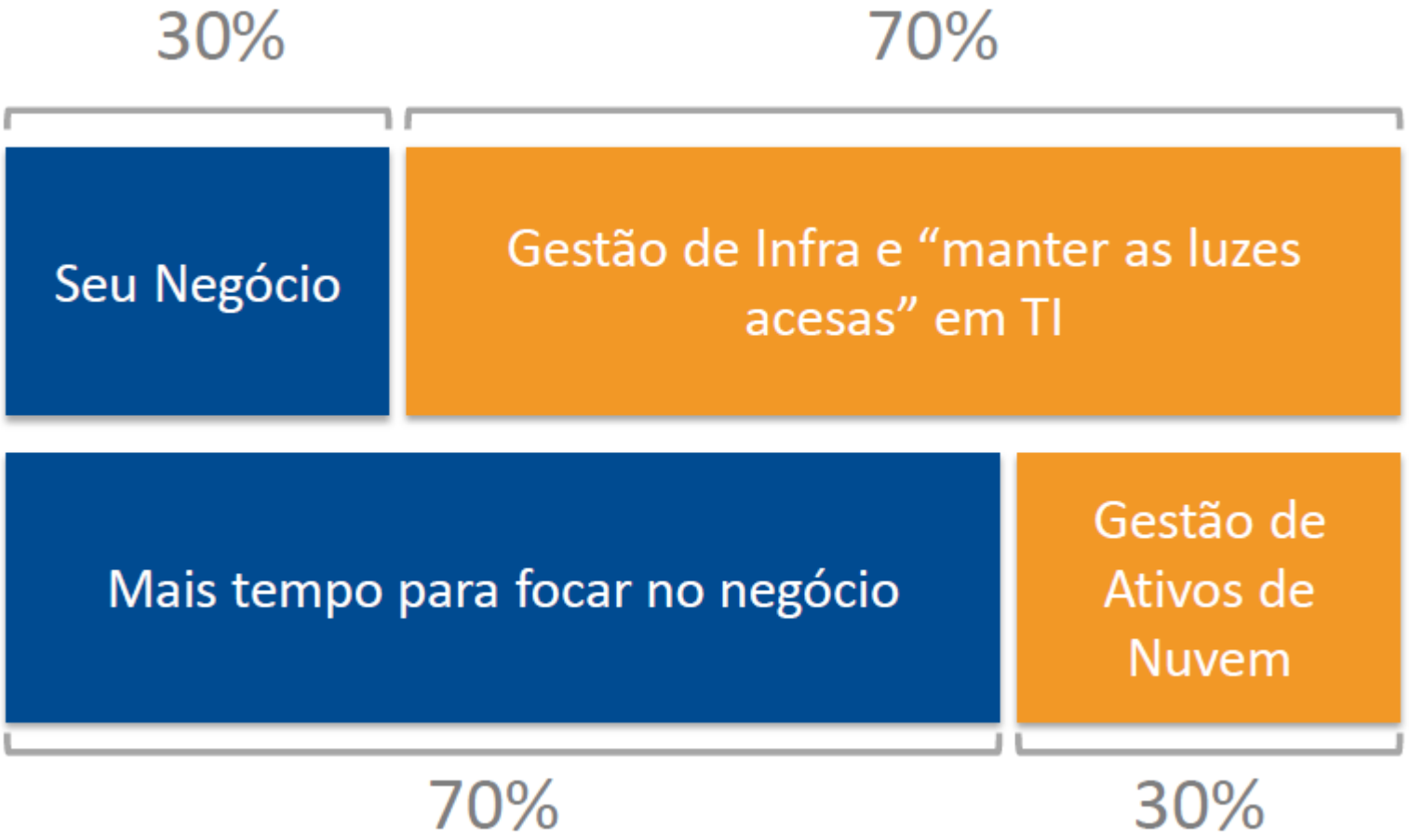
Computação em Nuvem

Computação em Nuvem

- Alto investimento inicial
- Alto custo de manutenção
- Quant. fixa de recursos

- Pagamento pelo uso
- Menor custo de manutenção
- Sob demanda

Computação em Nuvem



Sem preocupação com infra

99.95 de disponibilidade

Fonte: Amazon

Computação em Nuvem: Custo AWS

Tipo	CPU	RAM	Disco	Custo hora
Standard				
Micro	1	0.6 GB	8 GB	**$ 0.02**
Small	1	1.7 GB	160 GB	**$ 0.06**
Large	2	7.5 GB	850 GB	**$ 0.24**
Extra Large	8	15 GB	1690 GB	**$ 0.48**
High CPU				
Extra Large	20	7 GB	1690 GB	**$ 0.58**
High Memory				
4 XL	26	68.4 GB	1690 GB	**$ 1.62**
Cluster GPU				
4 XL	33.5	22 GB	1690 GB	**$ 2.10**

Armazenamento	Custo mês
Standard Storage (GB)	**$ 0.095**
Glacier Storage (GB)	**$ 0.01**

Computação em Nuvem

- Lista **Top 500**
  - Elenca os 500 supercomputadores mais rápidos do mundo
- **1.064 instâncias** do EC2 foram usadas para criar um supercomputador com **17.024 cores**
- **240 teraflops** de velocidade
  - 240 trilhões de operações por segundo
- Esse supercomputador é o **72º computador** mais rápido do mundo
  - Lista do Top 500 (jun/2012)

Você pode alugá-lo por **menos de US$ 1.000/h**

Fonte: Daniel Cordeiro

Computação em Nuvem: Mitos

- Nuvem é apenas uma “moda” passageira
- A nuvem não é confiável
- A nuvem não é segura

- Eu irei perder o controle dos dados
- Custo é a única vantagem da nuvem

Computação em Nuvem

- Aplicações existentes
  - Migração para a nuvem
- Novas aplicações
  - Grande volume de dados
    - Internet of Things (IoT)
    - Big Data
- Diferentes abordagens
  - Sistemas relacionais
  - NoSQL
  - NewSQL

Hadoop

Hadoop

- Framework *open-source*;
- Desenvolvimento e execução de aplicações distribuídas que processam grandes quantidades de dados;
- Como lidar com falhas no cluster?
- Arquitetura permite escalonamento próximo ao linear;
- Se a demanda por capacidade de processamento aumenta, a única restrição é sobre o orçamento previsto para adicionar novas máquinas;
- Análise vai ao encontro dos dados;

Hadoop File System(HDFS)

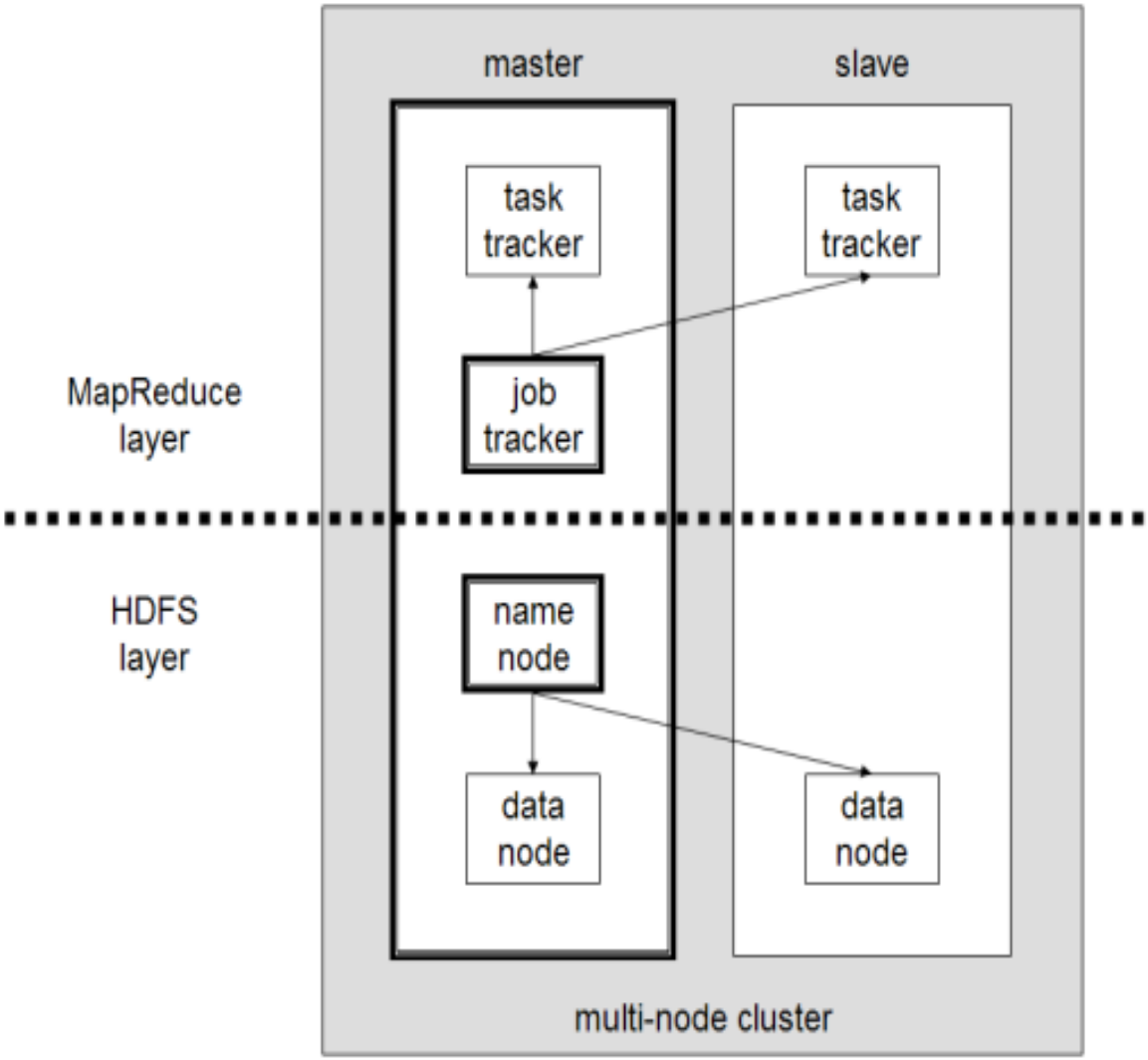
- Particiona os dados no cluster;
- Várias cópias dos dados são replicadas, como controle de falha do hardware;
- O código é enviado para a máquina que contém o dado sobre o qual se pretende operar;
- Como muitos bancos de dados NoSQL, organiza os dados por chave e valor, ao invés de por relações;

MapReduce

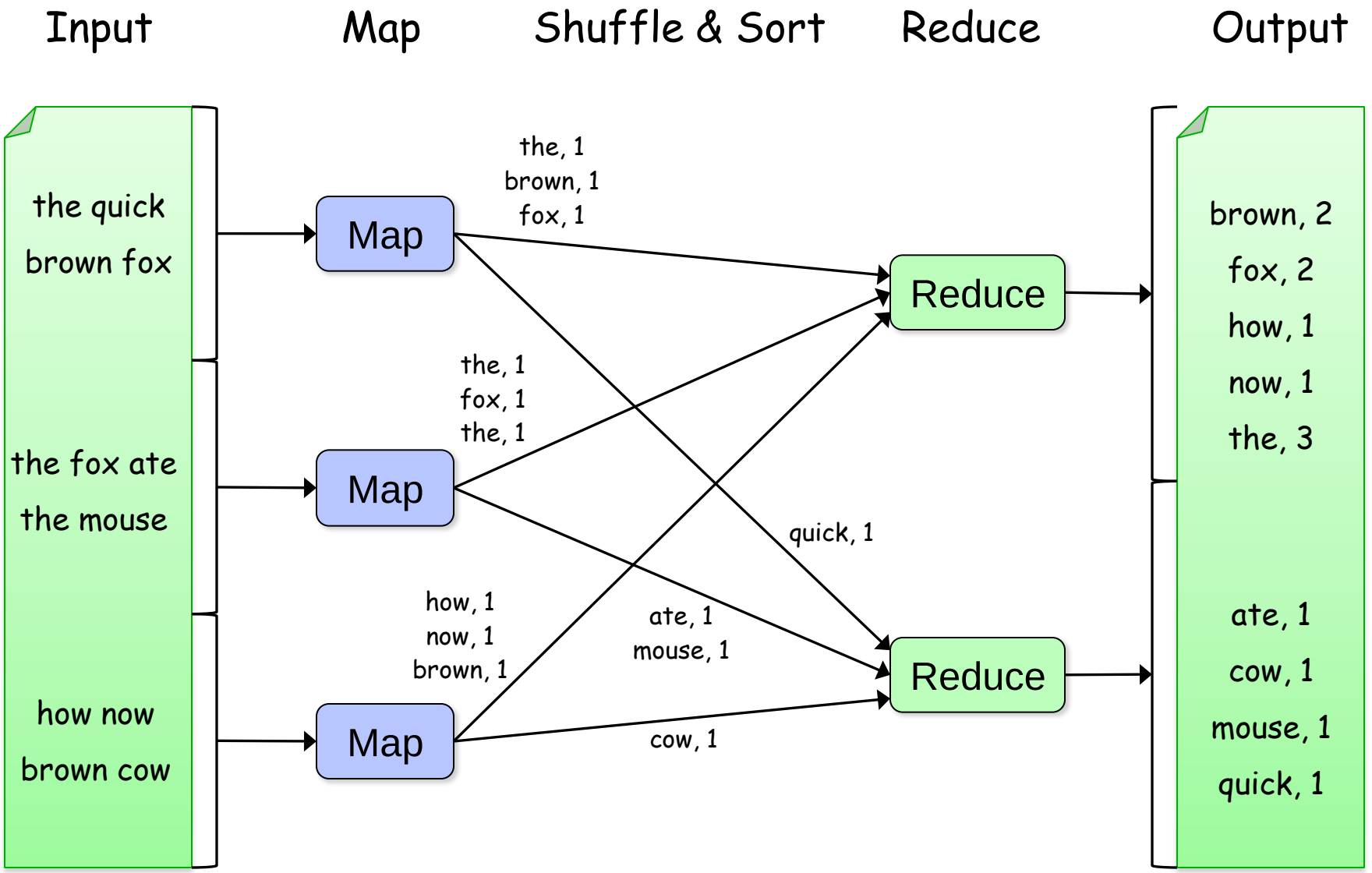
- Paradigma de programação funcional;
- Gerencia o processamento paralelo de um grande volume de dados distribuídos;
- Trabalha em dois passos fundamentais:
  - Map: módulo particionador do Hadoop divide o problema em pequenos subconjuntos e os atribuí a processos de map que o solucionam;
  - Reduce: combina os resultados dos processos de map e prepara a saída da operação MapReduce;

Processos do Hadoop

- NameNode
- DataNode
- TaskTracker
- JobTracker

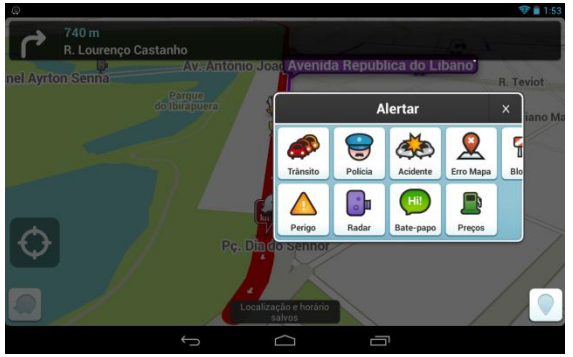


MapReduce

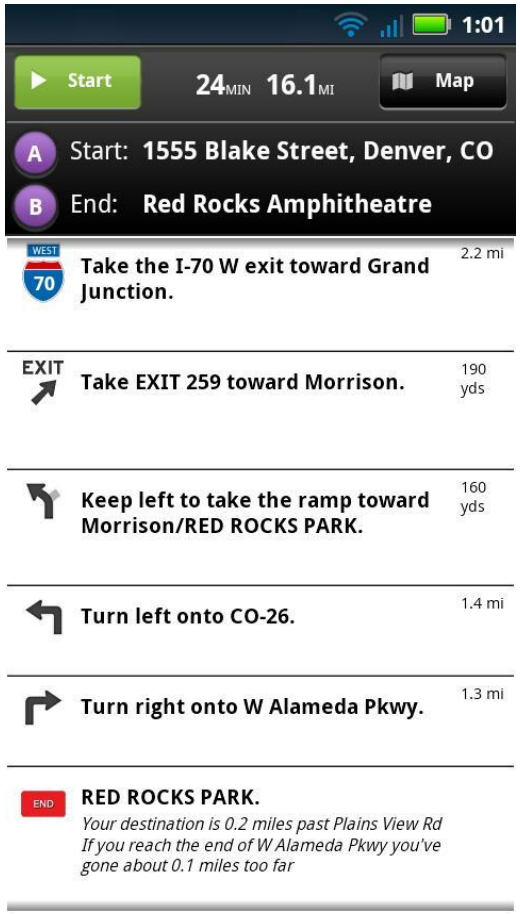

Input
Map
Shuffle & Sort
Reduce
Output
the quick
brown fox
the fox ate
the mouse
how now
brown cow
Map
Map
Map
the, 1
brown, 1
fox, 1
the, 1
fox, 1
the, 1
how, 1
now, 1
brown, 1
quick, 1
ate, 1
mouse, 1
cow, 1
Reduce
Reduce
brown, 2
fox, 2
how, 1
now, 1
the, 3
ate, 1
cow, 1
mouse, 1
quick, 1

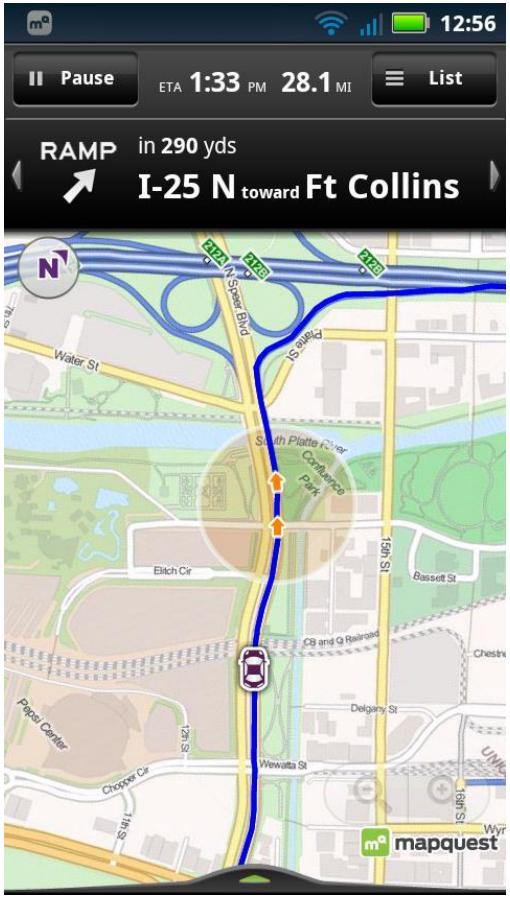
Estudo de Caso

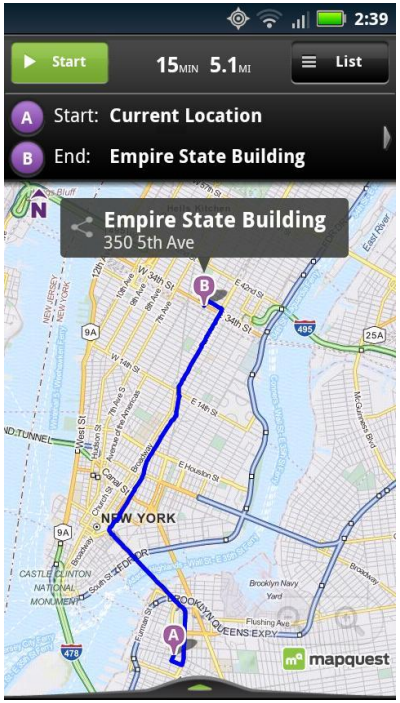
Waze Social GPS Maps & Traffic

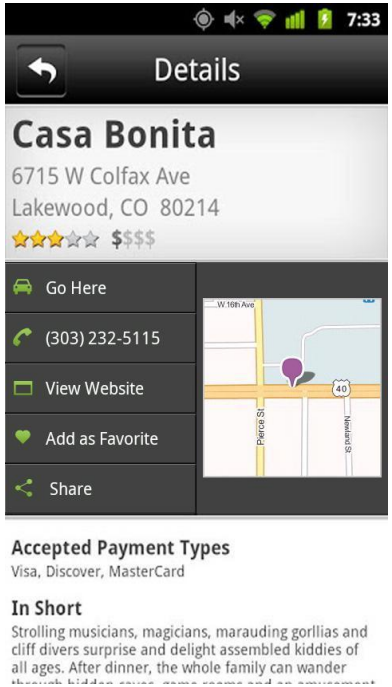


MapQuest: Maps, GPS & Traffic

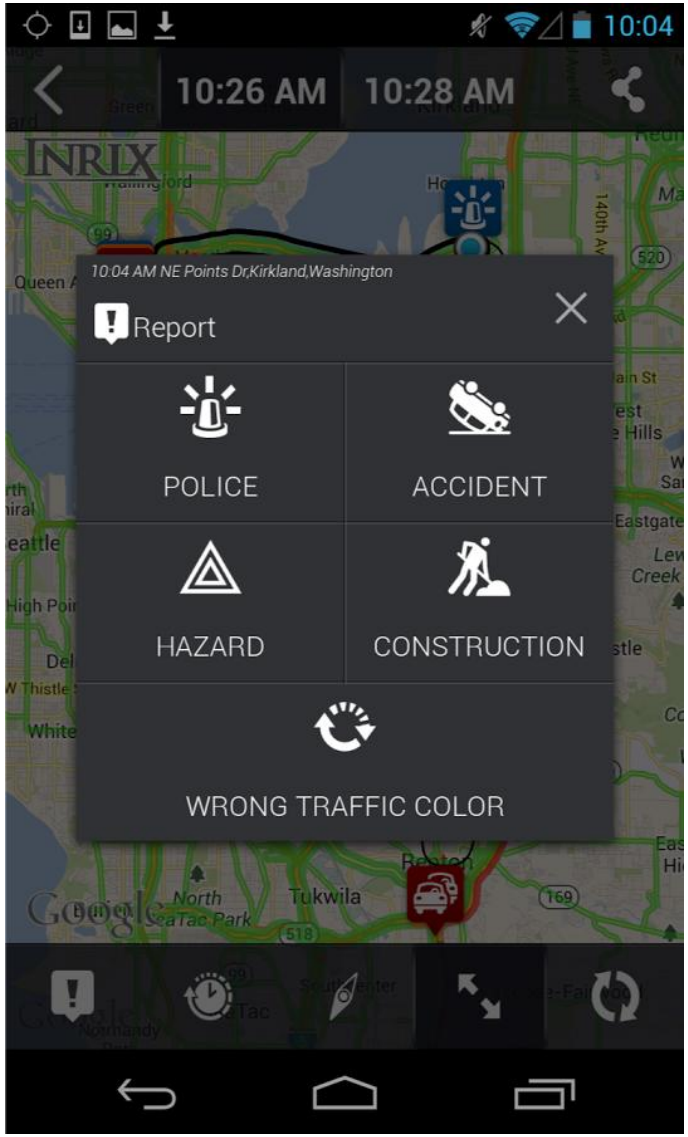


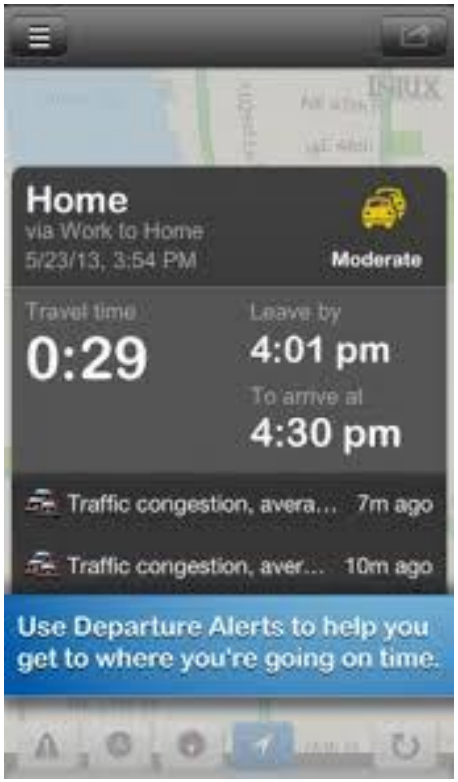






INRIX Traffic, Maps & Alerts





Coleta de dados

- Colaboração Humana
  - *Crowd-sourcing*;
  - Recebe/Reporta acidentes, regiões perigosas e outros eventos vistos na via;
  - Recomendação: posto de gasolina com preço mais barato mais perto de você.
  - Orientar a ação dos agentes públicos;
- Compartilhamento via email, Facebook, Twitter e Foursquare.

O que fazer a mais com esses dados?

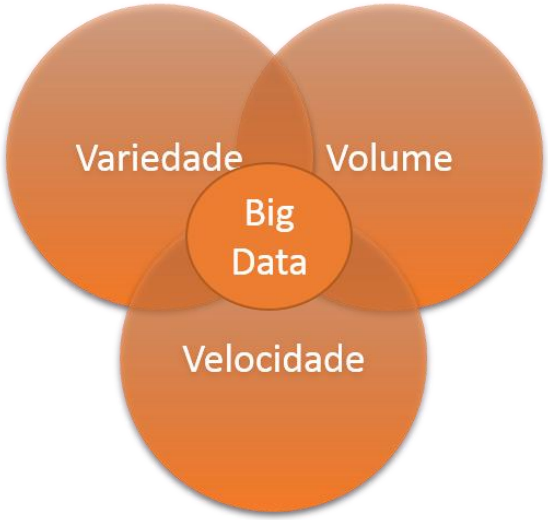
- Analisar o comportamento do trânsito nas cidades e descobrir padrões:
  - Em quais trechos da cidade ocorrem mais congestionamento?
- Busca por soluções eficazes de reengenharia de trânsito no contexto de cidades inteligentes;

Descrição – Estudo de Caso

- Dados coletada de trânsito por meio de sites e GPS;
  - Velocidade média em que os carros trafegam;
  - Latitude, longitude e o horário da coleta;
- Aplicação de algoritmo de mineração para grandes volumes de dados;
- Como paralelizar os algoritmo no contexto de dados de trânsito?
  - Plataforma de nuvem;
  - MapReduce;

Por que é Big Data?

- Grande volume de dados a ser gerenciado;
- Variedade: Dados obtidos a partir de diferentes fontes, não apresentam a mesma estrutura;
- Velocidade: informações chegam aos conjuntos de dados a todo instante de tempo. Os sensores coletam a posição espaço temporal do veículo, bem como a sua velocidade;



Infraestutura de Computação em Nuvem

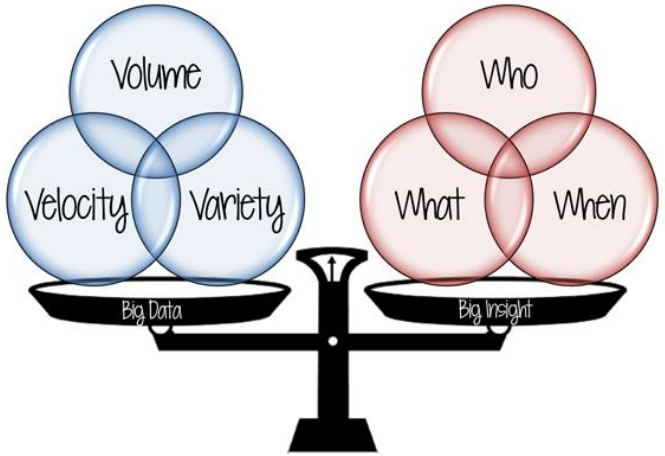
- Ambiente de nuvem privada da Universidade Federal do Ceará (UFC);
- Plataforma OpenNebula;
- #MVs:11 com sistema operacional Ubuntu, 8GB de memória RAM e 4 unidades de CPU;
- Hadoop 1.1.2;

- PostgreSQL 9.1 e biblioteca Spatial and Geographic Objects for PostgreSQL (PostGIS) para manipular objetos espaciais;

OpenNebula.org
The Open Source Toolkit for Cloud Computing

Análise para Big Data

- Objetivo: mostrar como a técnica de mineração foi modificada para ser aplicada em Big Data;



Congestionamento:velocidade<20 km/h

Somente os pontos de baixa velocidade

Grupos de alta densidade por proximidade geográfica

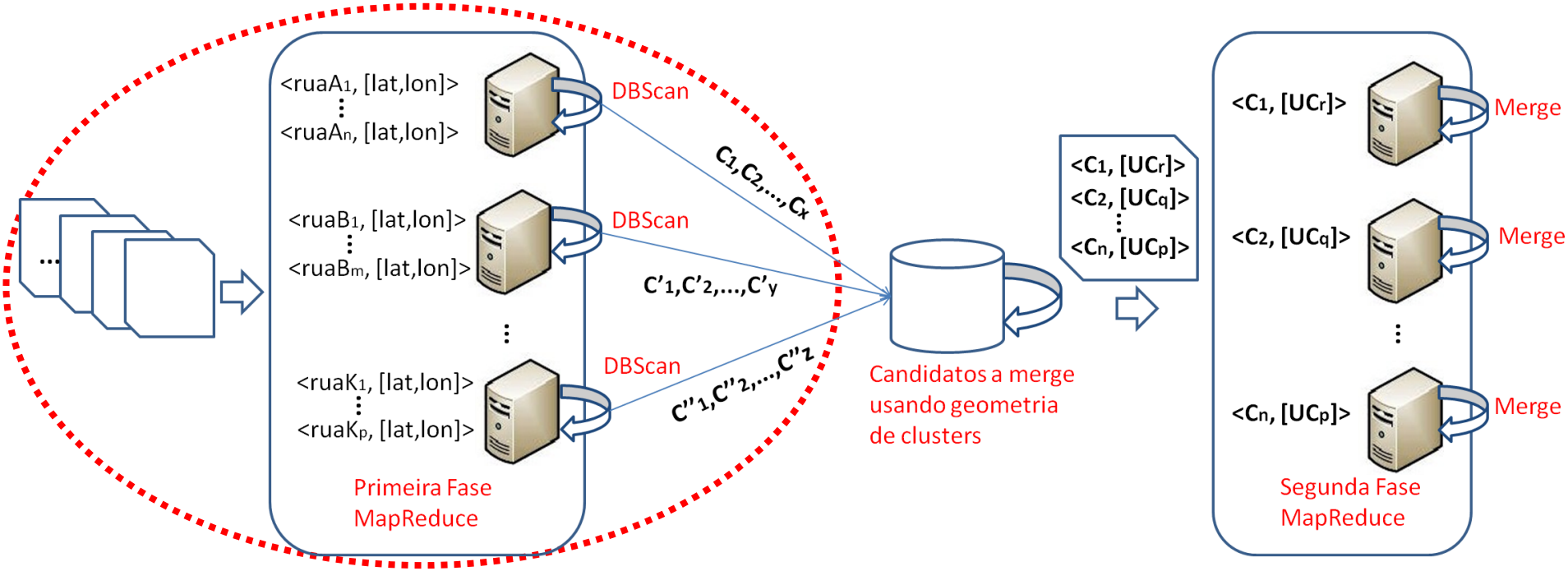
Agrupamento por tempo

Resultado da Análise

Visualização dos potenciais congestionamentos

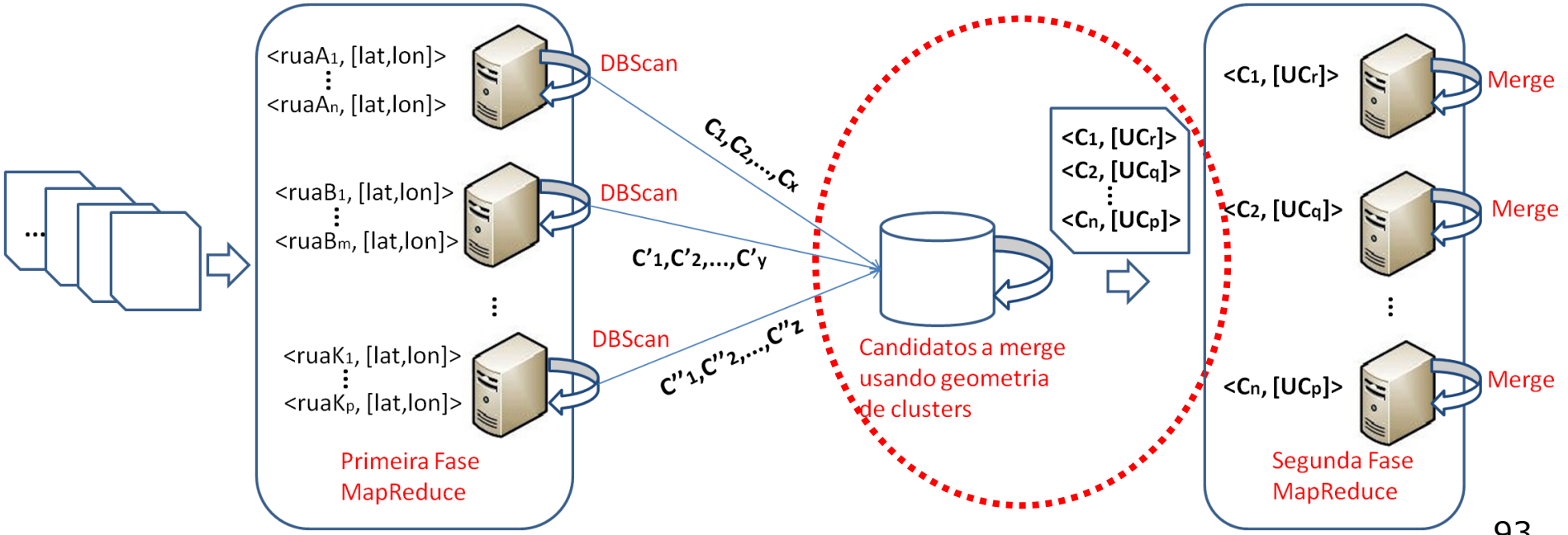
Algoritmo DBScan para Big Data

- Primeira fase:
  - Map: <k:rua,v:posicao_geografica(lat,lon)>;
  - Reduce:
    - <k:rua,v:[lat,lon]> ;
    - DBScan é aplicado;
  - Saída: Clusters armazenados em base de dados;



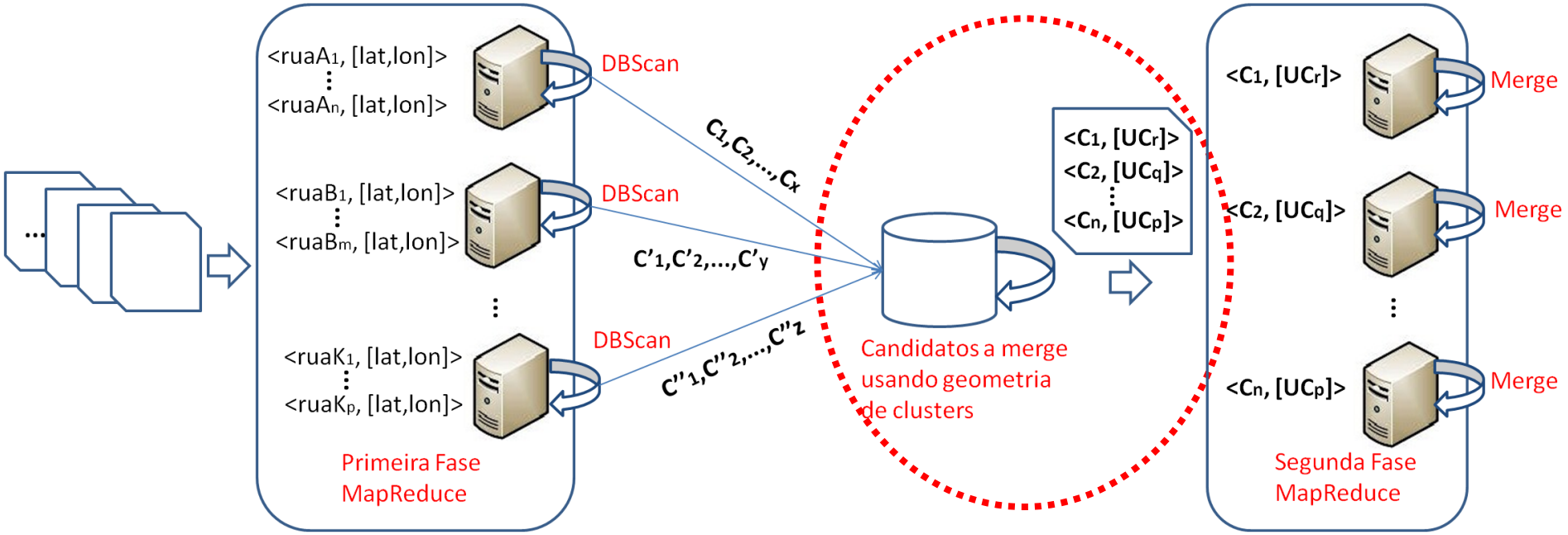
Algoritmo DBScan para Big Data

- Particionamento feito por ruas e as ruas se cruzam;
- Quais clusters de ruas diferentes se interceptam ou podem ser reunidos em apenas um cluster?
- Dois clusters podem ter pontos a uma distância menor que eps, de tal sorte que se os dados fossem processados em um mesmo reduce ou fossem de uma mesma partição, teria-se apenas um único cluster?



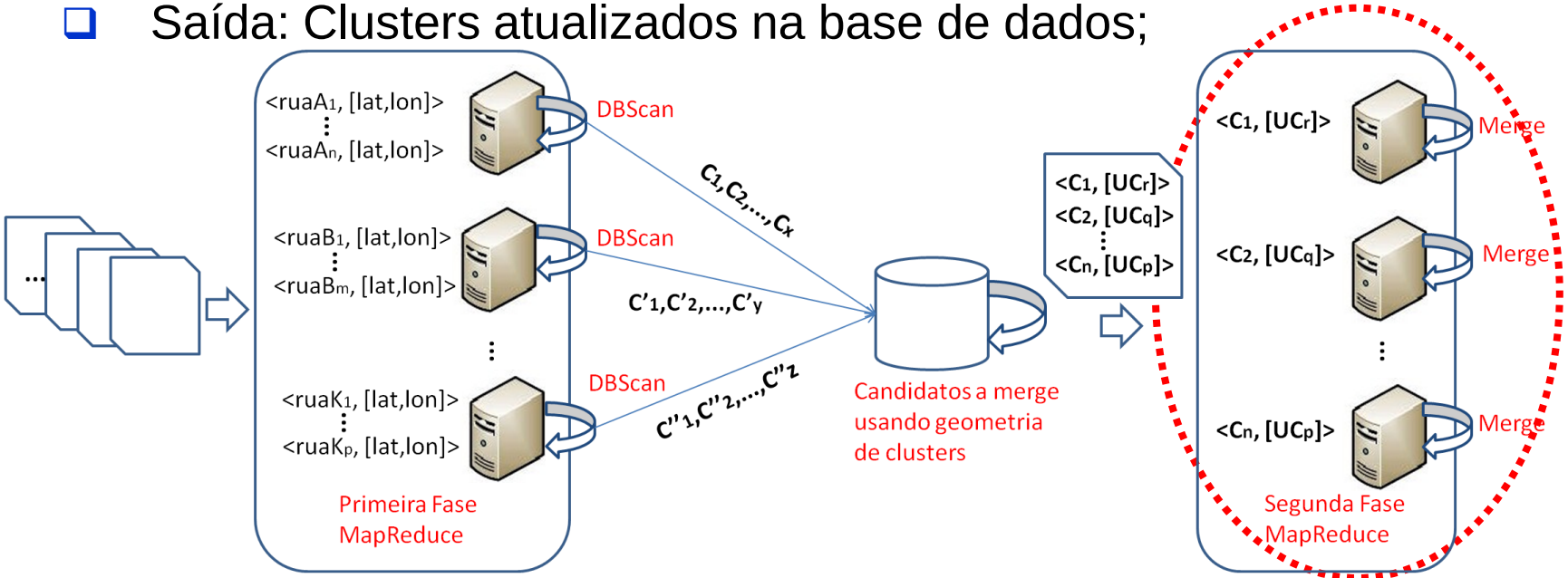
Algoritmo DBScan para Big Data

- Cada cluster é armazenado como um objeto geométrico no banco de dados;
- Objetos que estão a uma distância de no máximo eps para a fase de merge;
- Tuplas com pares de clusters candidatos a merge são passados com mesma chave para o próximo MapReduce;



Algoritmo DBScan para Big Data

- Segunda fase:
  - Map: função identidade;
  - Reduce:
    - key: Menor id de clusters Ci que podem ser reunidos em um só;
    - value: [UCi] - candidatos a merge;
    - Merge quando existem dois core points pi e pj, sendo pi do cluster Ci e pj do cluster Cj;
    - Tratamento de pontos noise;
  - Saída: Clusters atualizados na base de dados;



Tutorial - Instalação Hadoop

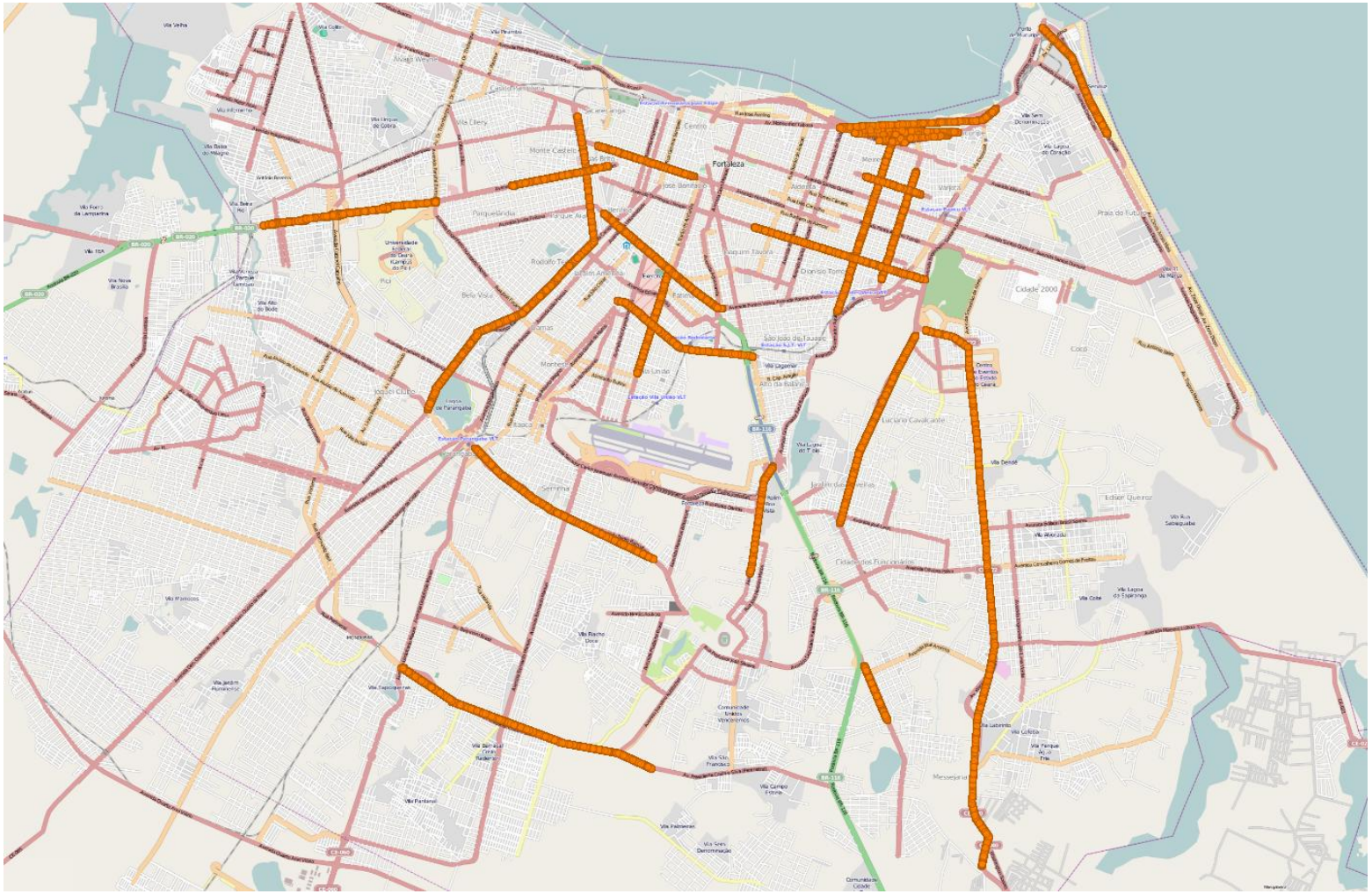
- https://sites.google.com/site/ufcckp8444/wiki/map-reduce
- Instalação Single Node;
- Instalação Multi Node;
- Exemplo de execução de MapReduce;

Utilização da VM – *Hands On*

- https://sites.google.com/site/ufcckp8444/wiki/map-reduce/dbscan-distribuido-utilizando-mapreduce

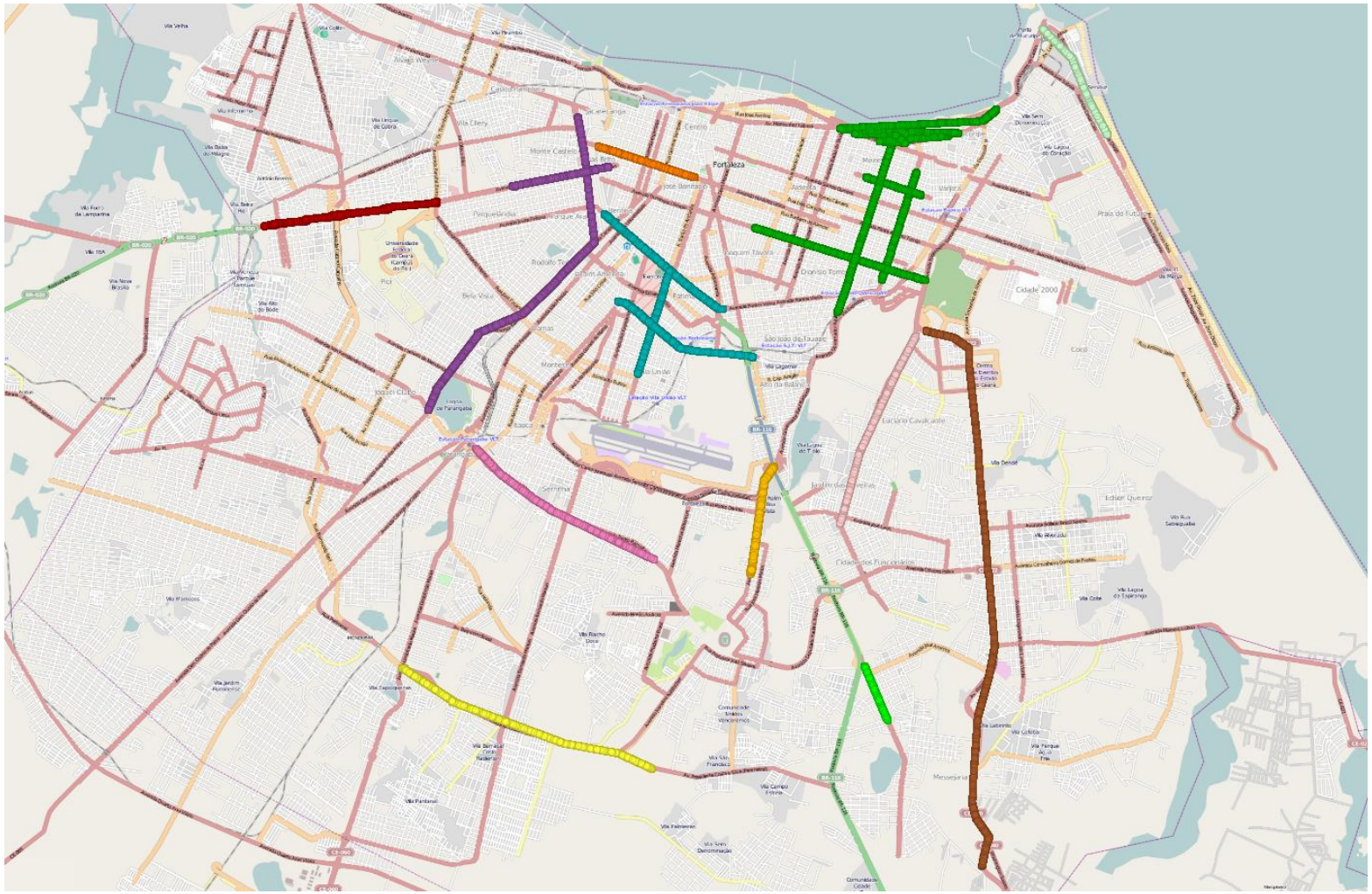
Resultados da Análise

- Regiões com concentração de veículos em Fortaleza - Ceará



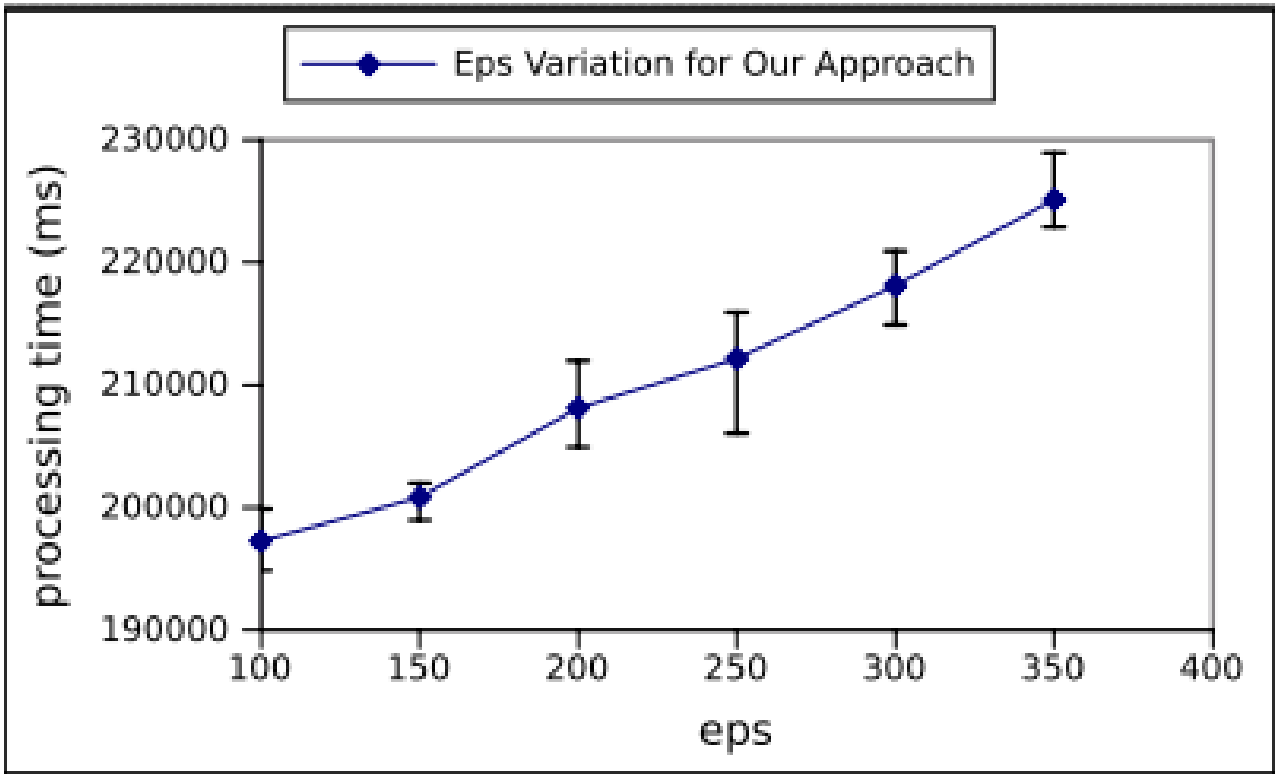
Resultados da Análise

- Regiões de alta densidade de veículos de baixa velocidade em Fortaleza - Ceará



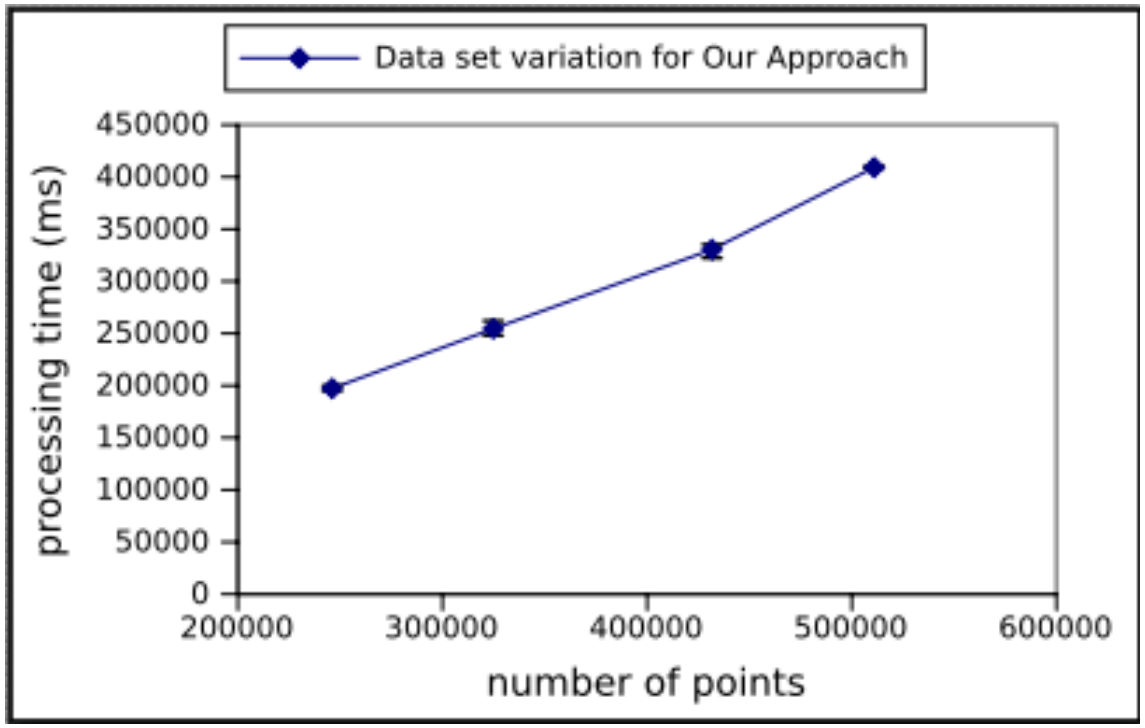
Resultados da Análise – Variação do eps

- Variação eps: 100,150,200,250,300 e 350;
- Estático:
  - Quantidade de pontos: 246142;
  - MinPoints: 50;



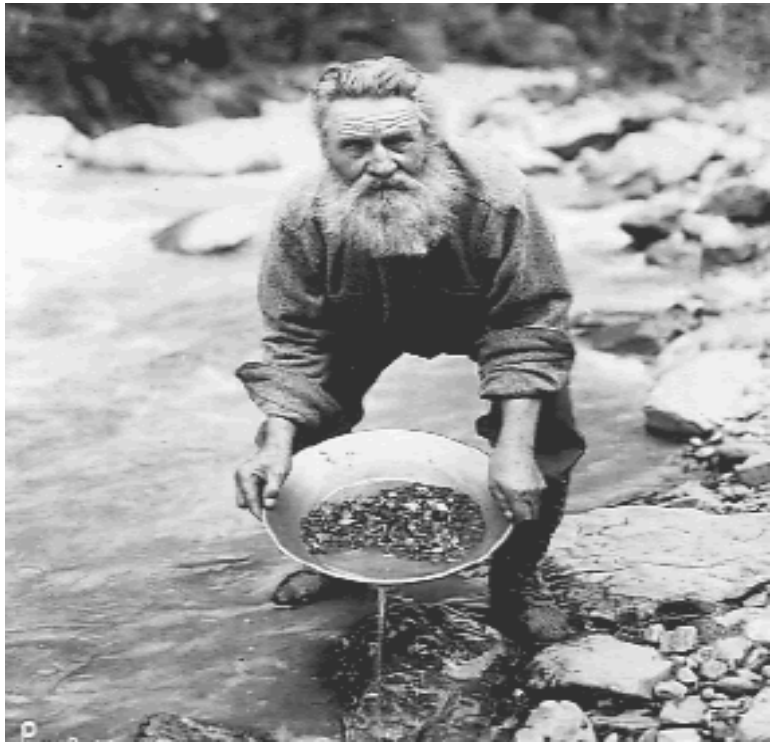
Resultados da Análise – Variação da quantidade de pontos

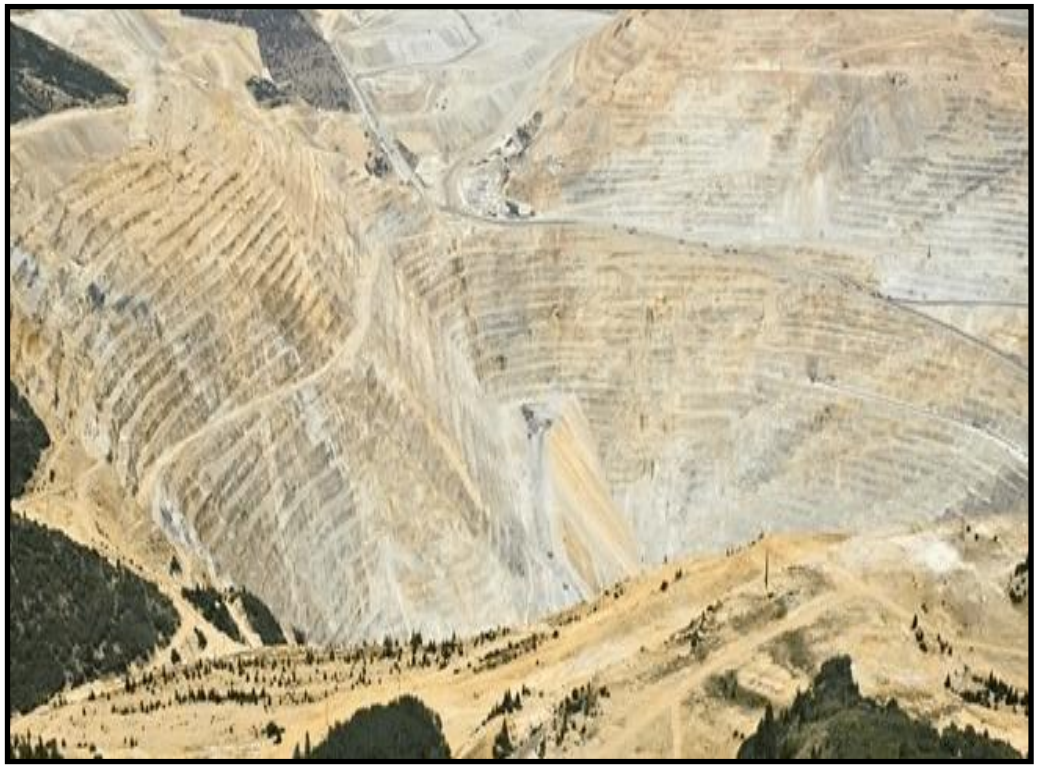
- Solução escalável;
- Variação quantidade de pontos: 246142, 324778, 431698 e 510952 pontos;
- Estático:
  - eps: 100;
  - MinPoints: 50;



Desafios em Big Data e Conclusão

O processo de mineração de antes não serve agora?





Desafios em Big Data

- Gerenciamento e Armazenamento
  - Tecnologias atuais não são adequadas para Big Data;
  - Capacidade de armazenamento cresce mais lentamente do que a quantidade de dados;
  - Algoritmos não são eficientes para trabalhar com heterogeneidade de dados;
  - Incompletude dos dados;

Desafios em Big Data

- Computação e Análise dos Dados
  - Velocidade;
  - Utilização de índice não é adequada para dados complexos;
  - Construir e adaptar algoritmos para trabalhar de forma paralela;
  - Aproveitar a infraestrutura em nuvem;
- Desempenho
- Congestionamento da rede

Desafios em Big Data

- Segurança
  - Permitir análise de dados sem informações confidenciais;
  - Tecnologias atuais para proteção da privacidade consideram dados estáticos;
- Escalabilidade
  - Construir sistemas para processar um volume crescente de dados;
  - Volume está aumentando em uma velocidade maior do que os recursos computacionais;
  - *Crowd-sourcing*

Desafios em Big Data

- Análise para Big Data
  - Alta dimensionalidade dos dados;
  - Acúmulo de *outliers*;
  - Correlação incorreta dos dados;
  - Alto custo computacional;
  - Necessidade de algoritmos complexos;
- Heterogeneidade
  - Múltiplas fontes, diferentes pontos de tempo e diferentes tecnologias;
  - Necessidade de procedimentos robusto e adaptáveis;

Desafios em Big Data

- *Big Data e Cloud*
  - Necessidade de gerenciar grandes infraestruturas;
  - Hadoop é complexo e sem suporte gerencial;
    - Mais de 150 variáveis a serem configuradas;

Conclusão

- Gerenciar e analisar um grande volume de dados é complexo;
- Soluções atuais não tratam de forma adequada este volume
  - Necessidade de paralelizar as soluções;
- Computação em nuvem contribui para a construções destas soluções;
- Muitos desafios:
  - Armazenamento, processamento, análise, segurança, escalabilidade;

REFERÊNCIAS

Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., Li, K., and Qu, W. (2012). Big data processing: Big challenges and opportunities. Journal of Interconnection Networks, 13(03n04).

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., and Welton, C. (2009). Mad skills: new analysis practices for big data. Proceedings of the VLDB Endowment, 2(2):1481–1492.

Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F. B., and Babu, S. (2011). Starfish: A self-tuning system for big data analytics. In CIDR, volume 11, pages 261–272.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, volume 1215, pages 487–499.

Yang, X. Y., Liu, Z., and Fu, Y. (2010). Mapreduce as a programming model for association rules algorithm on hadoop. In Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on, pages 99–102. IEEE.

Riondato, M., DeBrabant, J. A., Fonseca, R., and Upfal, E. (2012). Parma: a parallel randomized algorithm for approximate association rules mining in mapreduce. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 85–94. ACM.

REFERÊNCIAS

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). Machine learning, neural and statistical classification.

Panda, B., Herbach, J. S., Basu, S., and Bayardo, R. J. (2009). Planet: massively parallel learning of tree ensembles with mapreduce. Proceedings of the VLDB Endowment,2(2):1426–1437.

Cryans, J.D., Ratte, S., and Champagne, R. (2010). Adaptation of apriori to mapreduce to build a warehouse of relations between named entities across the web. In Advances in Databases Knowledge and Data Applications (DBKDA), 2010 Second International Conference on, pages 185–189. IEEE

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1):21–27.

Ghoting, A., Kambadur, P., Pednault, E., and Kannan, R. (2011).Nimble: a toolkit for the implementation of parallel data mining and machine learning algorithms on mapreduce. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 334–342.

Hammoud, S. (2011). Mapreduce network enabled algorithms for classification based on association rules.

Li, L. and Zhang, M. (2011). The strategy of mining association rule based on cloud computing. In Business Computing and GlobalInformatization (BCGIN), 2011 International Conference on, pages 475–478.

REFERÊNCIAS

He, Y., Tan, H., Luo, W., Mao, H., Ma, D., Feng, S., and Fan, J. (2011). Mr-dbscan: An efficient parallel density-based clustering algorithm using mapreduce. In Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on, pages 473–480. IEEE.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1):81–106.

Kim, Y. and Shim, K. (2012).Parallel top-k similarity join algorithms using mapreduce. In Data Engineering (ICDE), 2012 IEEE 28th International Conference on, pages 510–521. IEEE.

Lu, W., Shen, Y., Chen, S., and Ooi, B. C. (2012). Efficient processing of k nearest neighbor joins using mapreduce. Proceedings of the VLDB Endowment, 5(10):1016–1027.

Zhou, L., Wang, H., and Wang, W. (2012). Parallel implementation of classification algorithms based on cloud computing environment. TELKOMNIKA Indonesian Journal of Electrical Engineering, 10(5):1087–1092.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pages 1027–1035.

REFERÊNCIAS

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. Proceedings of the VLDB Endowment, 5(7):622–633.

Dai, B.R. and Lin, I.C. (2012). Efficient map/reduce-based dbscan algorithm with optimized data partition. In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, pages 59–66. IEEE

Obrigada!!

ticianalc@ufc.br
ticianalc@gmail.com