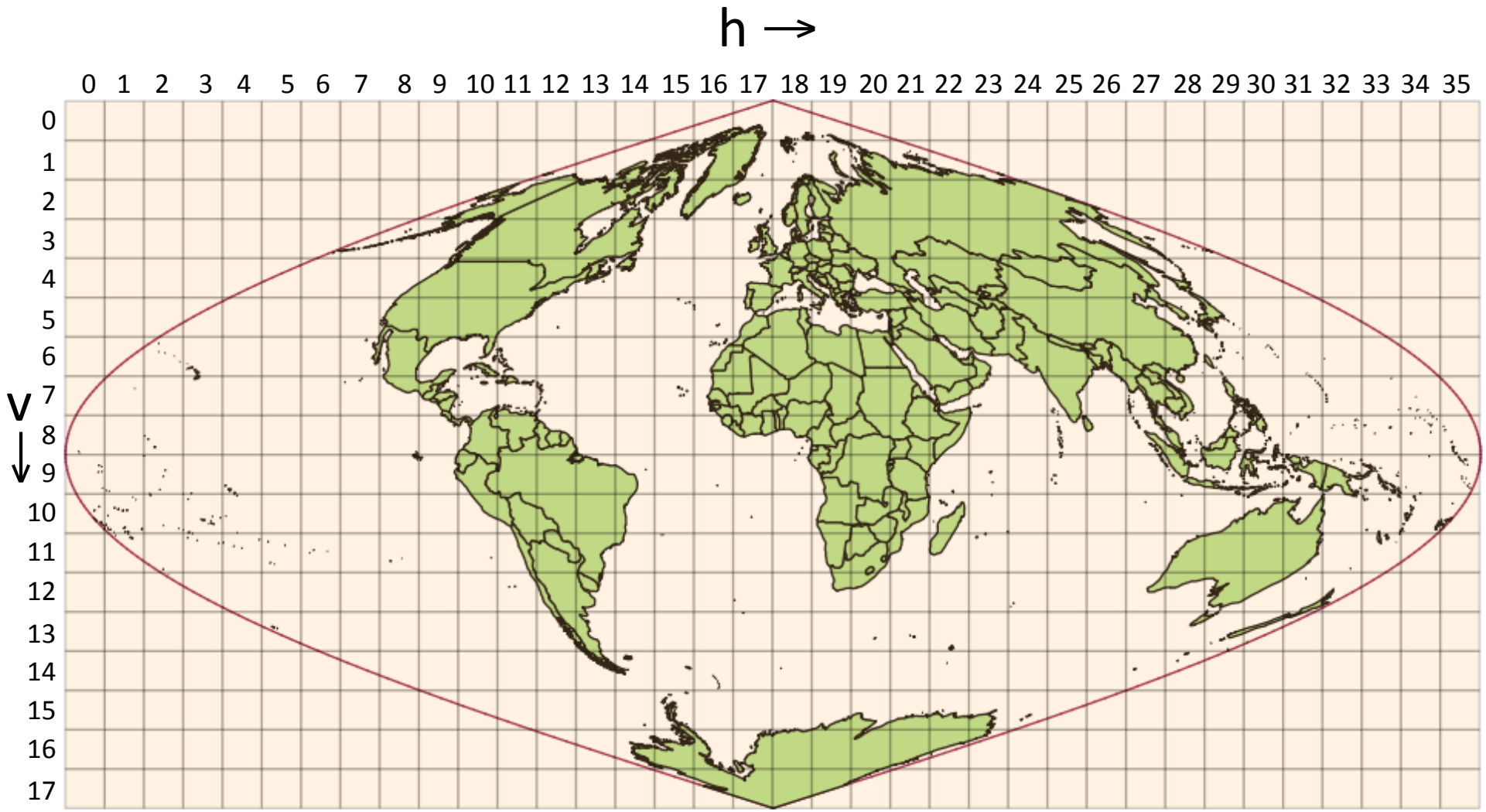
Bancos de Dados Geográficos

Dr. Gilberto Ribeiro de Queiroz
<gribeiro@dpi.inpe.br>

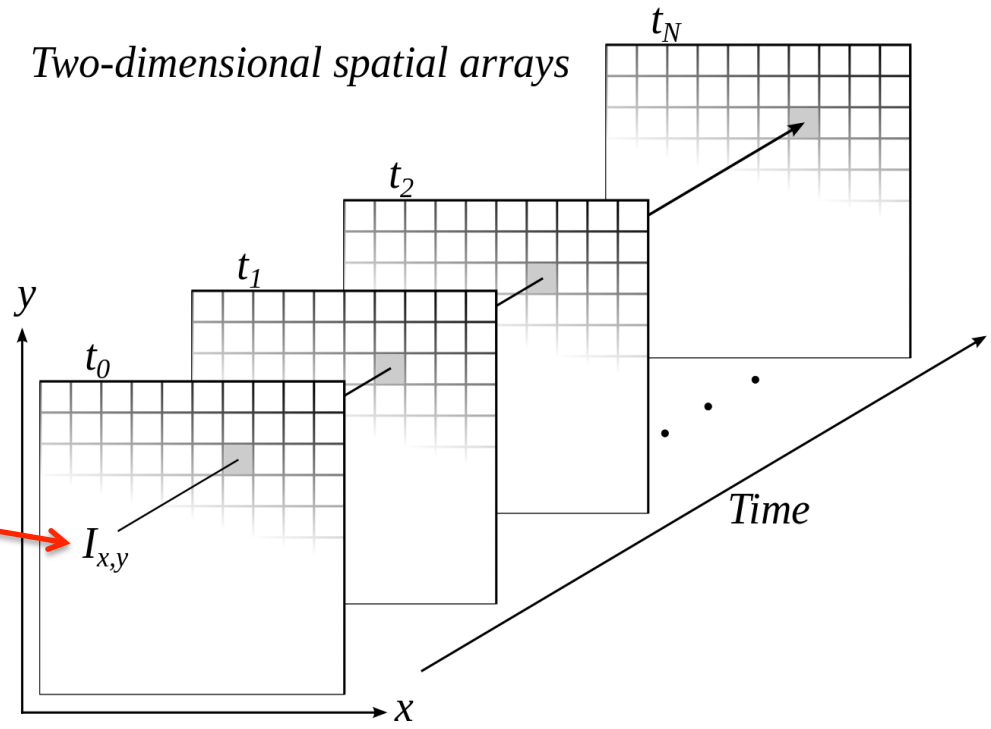
Discussão: Existe algum problema com o projeto do suporte Raster dos atuais SGBD-R?

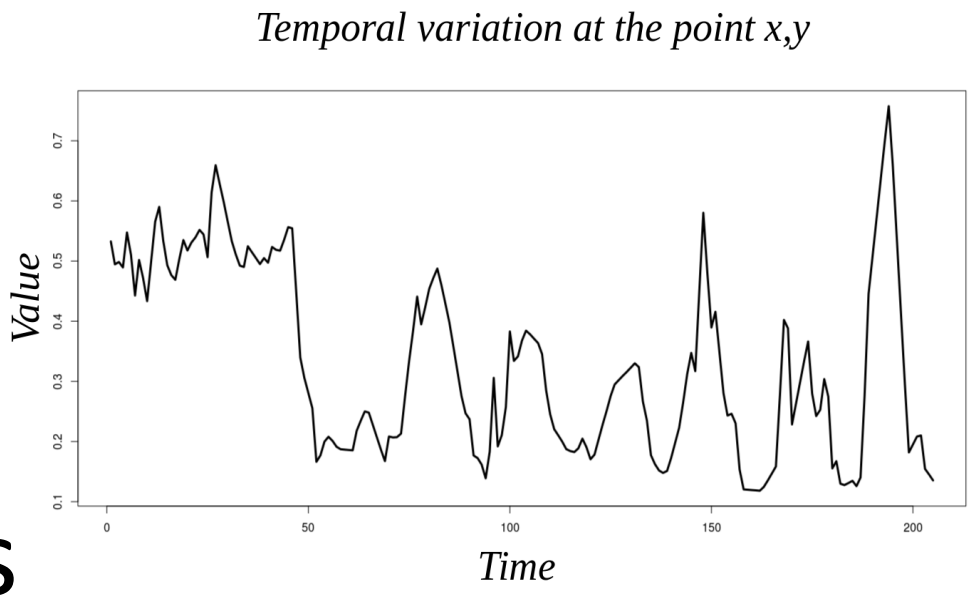
Como lidar com os requisitos de aplicações de EO que podem necessitar como entrada dados massivos?

MODIS: Grade Sinusoidal

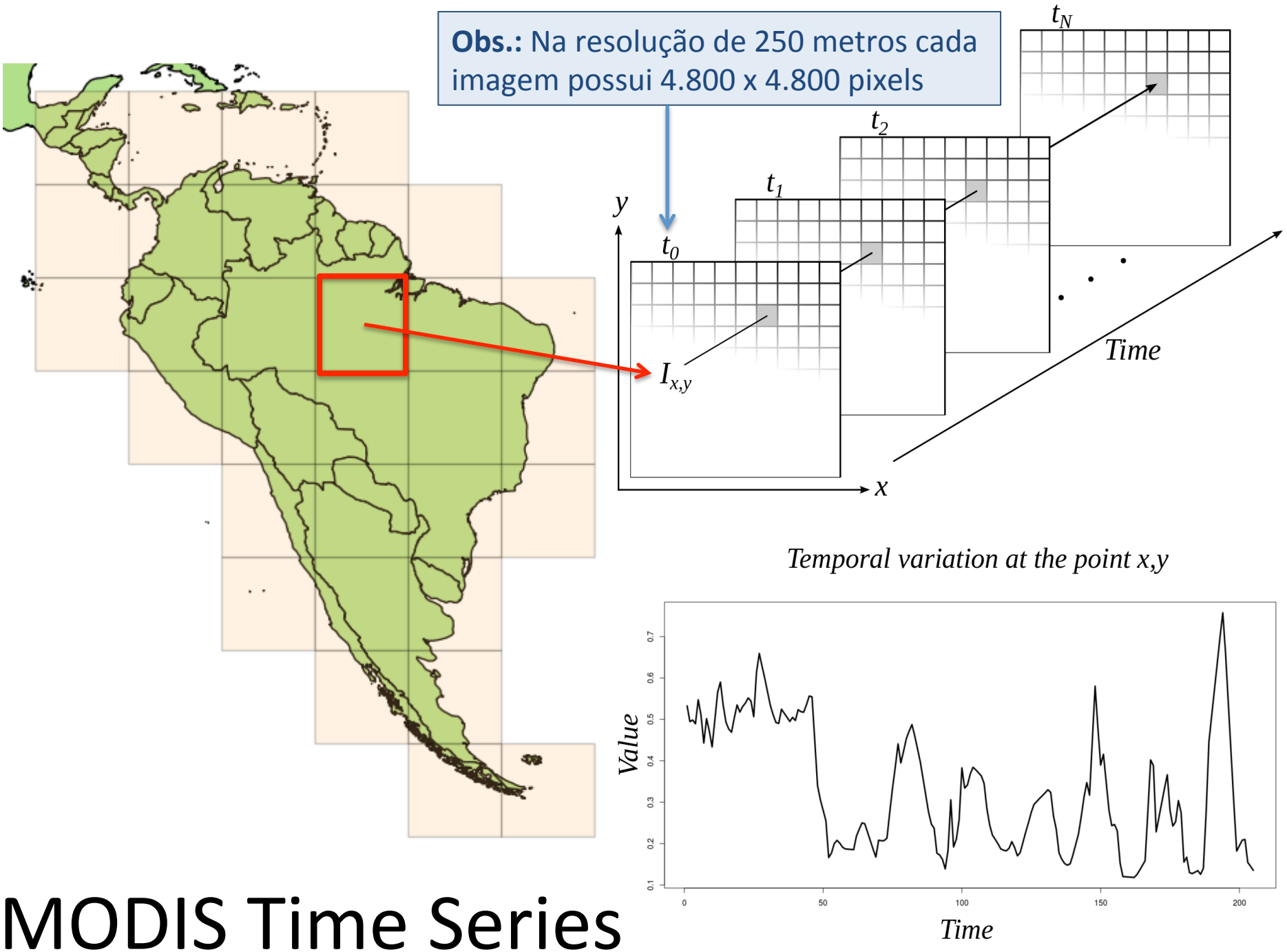


Fonte: Adpatado de http://nsidc.org/data/modis/data_summaries/landgrid.html





MODIS Time Series

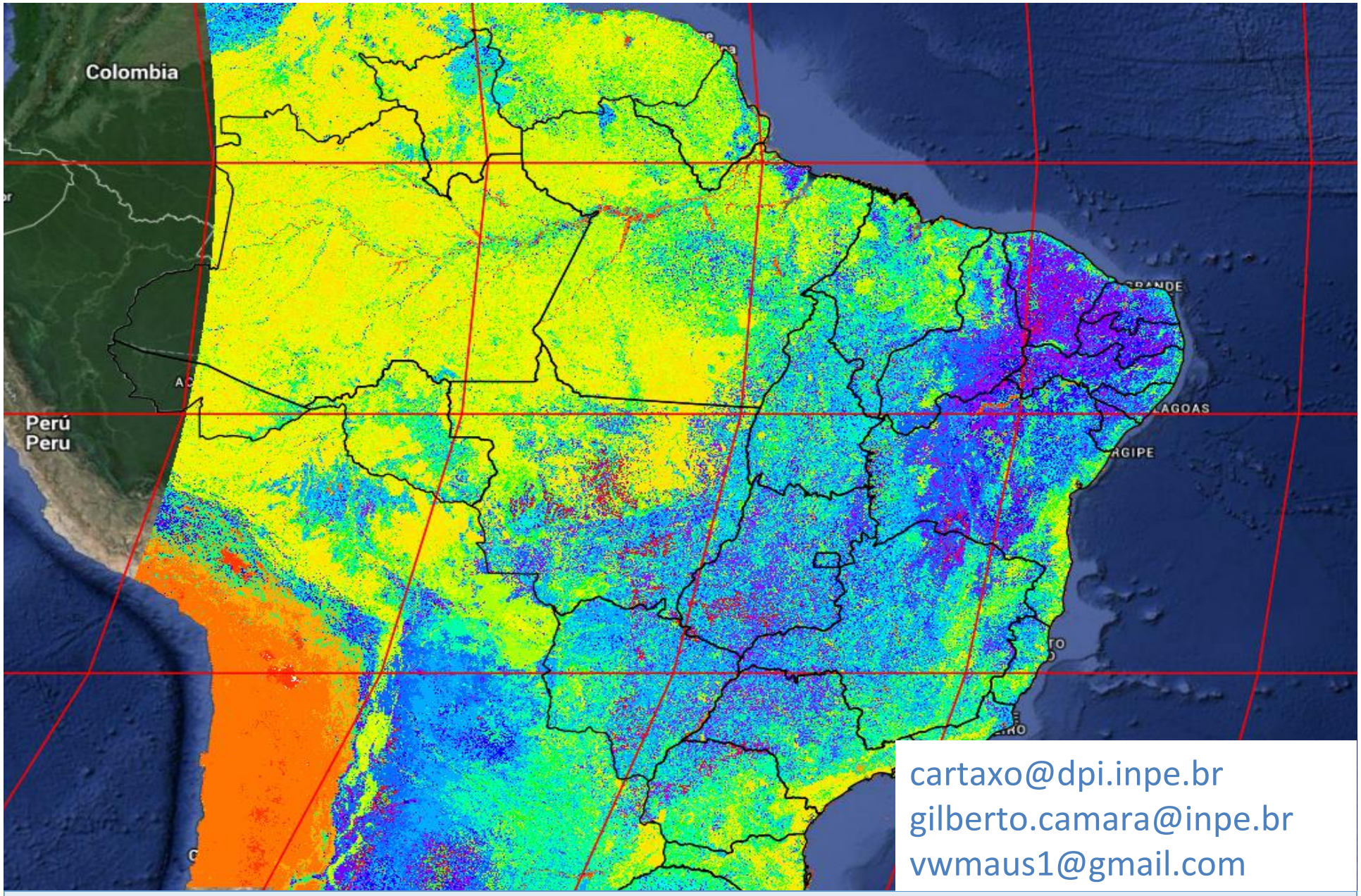


MODIS Time Series

Quais são os desafios?

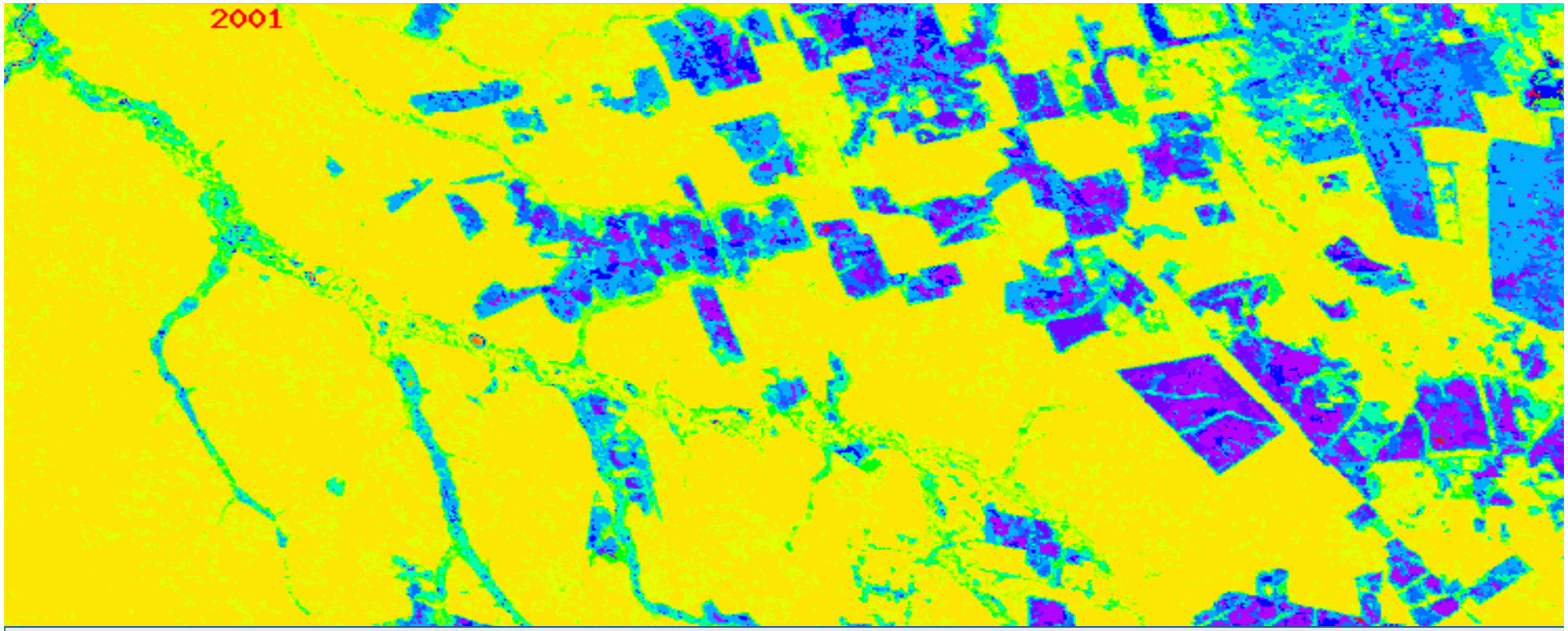
Exemplo: Mapa de Vegetação Global de Fevereiro de 2000 a Fevereiro de 2014

Fonte: NASA Earth Observatory (23 de Abril , 2014)

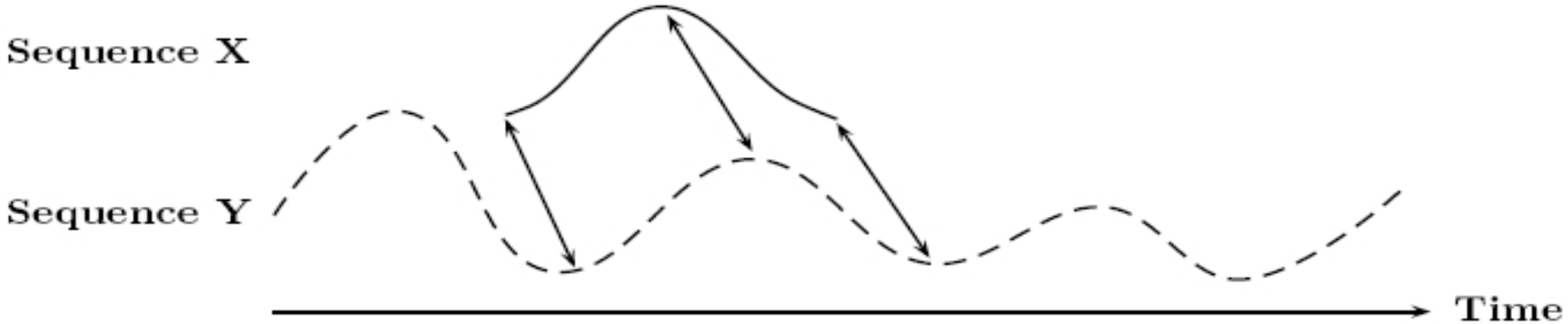


cartaxo@dpi.inpe.br
gilberto.camara@inpe.br
vwmaus1@gmail.com

Temporal Land Cover Classification

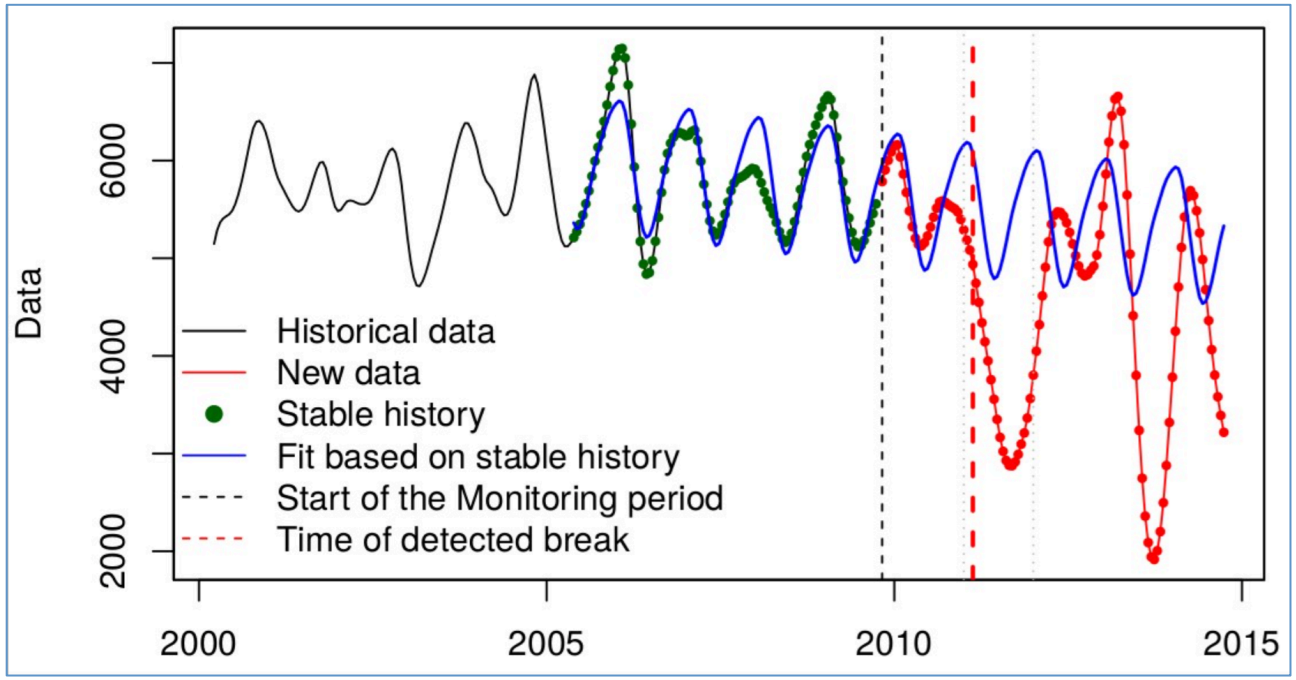


Basic Problem: identify a signature on a time series



Approach: use Dynamic Time Warping (DTW)

Automated Real-Time Deforestation Monitoring With Satellite Image Time Series



Breaks For Additive Season and Trend (BFAST)

Christopher Stephan
(c_step03@uni-muenster.de)

Como fornecer melhor infraestrutura computacional para nossos pesquisadores?

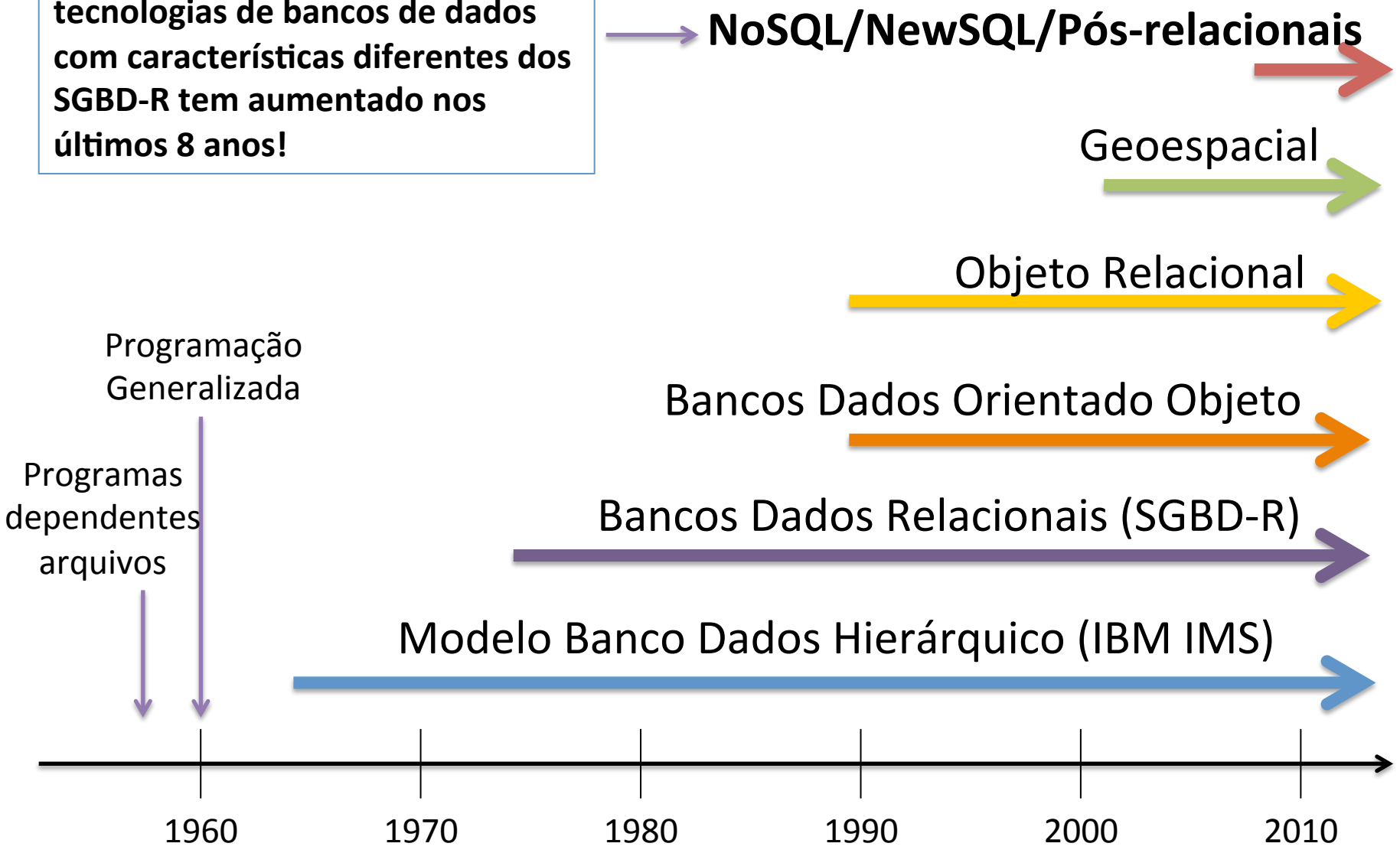
Como lidar com big geospatial data?

Satellites: Landsat, Terra, Aqua, Sentinel

(Time Series of Remote Sensing Imagery)

Evolução das Tecnologias de Bancos Dados

Interessante: o número de tecnologias de bancos de dados com características diferentes dos SGBD-R tem aumentado nos últimos 8 anos!



O “cardápio” de opções aumentou?

- *Sistemas Não-Relacionais* ou *Not Only SQL* ou *Pós-relacionais*:
  - http://nosql-database.org/
  - https://en.wikipedia.org/wiki/NoSQL

- Diferentes modelos de dados:
  - Document Oriented: MongoDB, CouchDB;
  - Column Stores: Cassandra;
  - Graph Databases: OrientDB, Neo4J;
  - Array Databases: SciDB, Rasdaman.

- Nem todos são baseados no paradigma de transações ACID.

- Escalabilidade: Horizontal x Vertical

Dados Científicos

Sensoriamento Remoto

Os dados utilizados em diversas áreas da Ciência encontram-se na forma de **Arrays**

Arrays = Matrizes

Solução 2: Usando um Array Database

rasdaman

SciDB

Solução 2: Usando um Array Database

rasdaman

SciDB 

Novas Tecnologias de Bancos de Dados para Dados Matriciais

Array Databases

"Arrays as first class citizens"

Source: Wikipedia

ACM Turing Award (2014)

SciDB

"SciDB is an open-source analytical database oriented toward the data management needs of scientists."

(Stonebraker et al., 2011)

O que é o SciDB?

- Mix de plataforma para gerenciamento e análise de dados

- Array Database:
  - O modelo de dados trata de “Arrays” e não de “Tabelas”:
    - Array = Nome + Definição de Atributos das Células + Dimensões
  - Linguagem de consulta é baseada em um álgebra de arrays:
    - AQL and AFL: filter, aggregation, join

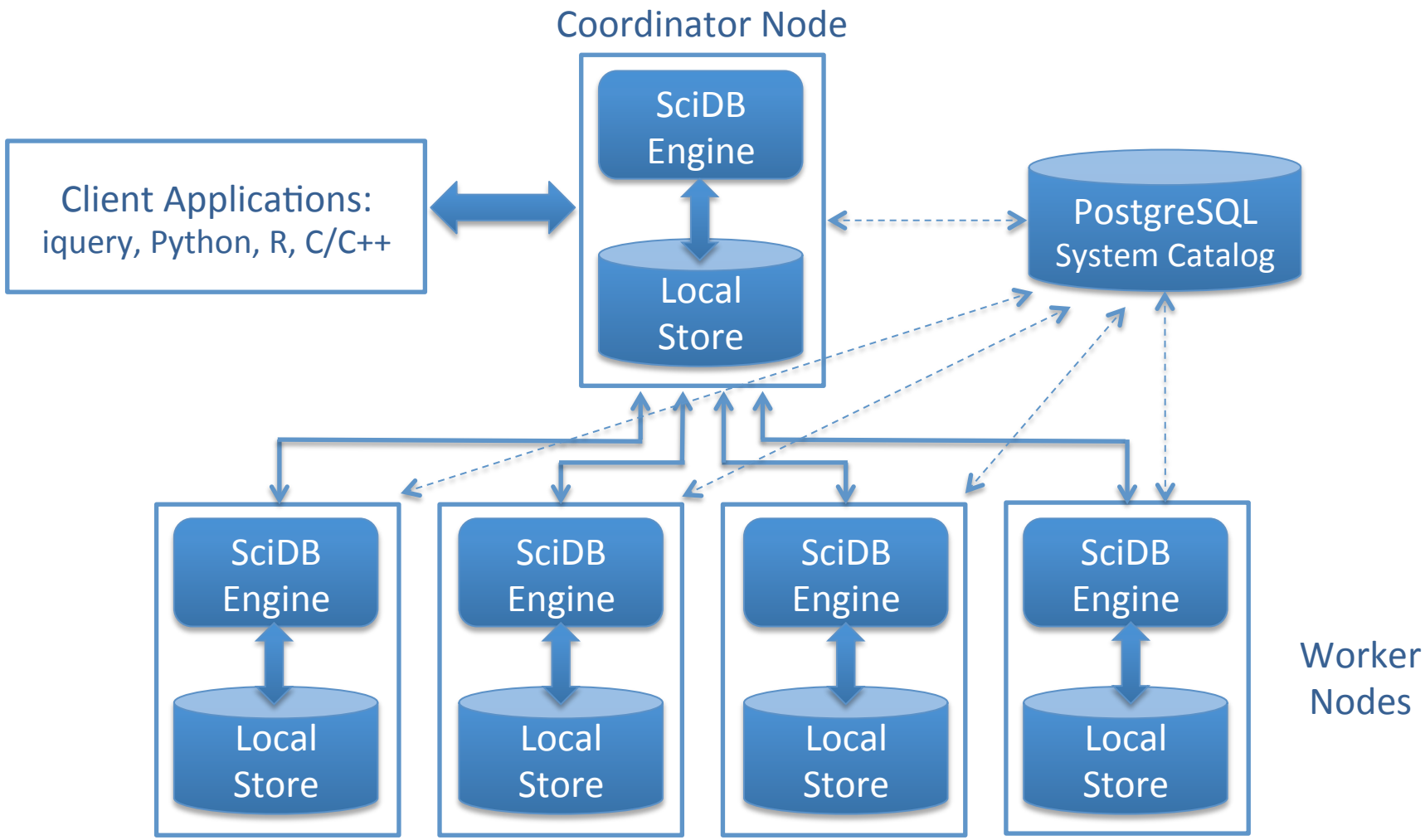
- Projetado para trabalho em *clusters*:
  - Opção: arquitetura ***shared nothing***

SciDB

- Site: http://www.paradigm4.com
- License: AGPL v3.
- Version: 15.7.
- Currently supported platforms:
  - Linux: Ubuntu 12.04 e 14.04, RHEL 6, CentOS 6.
- Principal desenvolvedor: Paradigm4.
- Código fonte encontra-se disponível como pacotes tar.gz:
  - Community Edition: não há acesso aberto ao repositório de código fonte .
  - Enterprise Edition: acesso SVN/GIT.

Instância

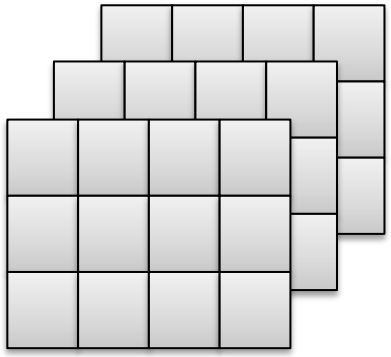
Arquitetura



Source: Adapted from PARADIGM4

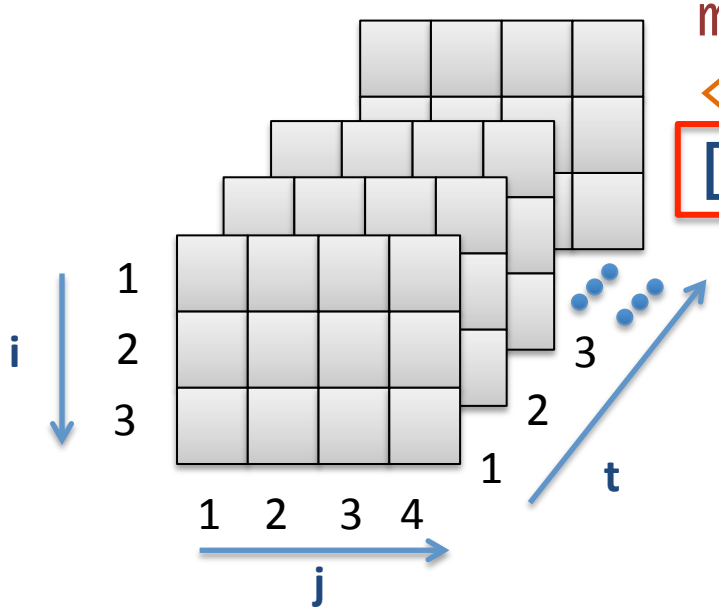
array

name <attributes> [dimensions]



Arrays: Dimensões

- Em geral utiliza-se valores inteiros de 64-bit.
- Dimensões com limites bem definidos (*bounded dimension*):
  - Quando sabemos a priori o número total de células em uma dada dimensão (ou a cardinalidade de uma dimensão).
- Dimensões com limites indefinidos (*unbounded dimension*):
  - Quando não sabemos a cardinalidade do array em tempo de criação.



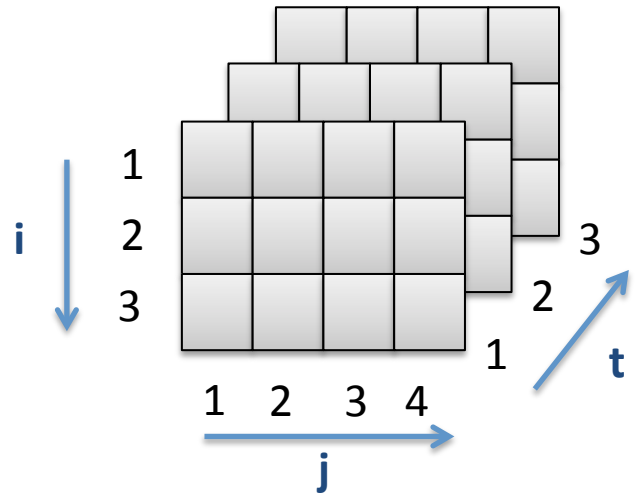

```
mod09q1
<attributes>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```

O Array mod09q1 terá 3 dimensões: 4 x 3 x *

Arrays: Atributos

- Cada Célula pode estar associada a múltiplos valores (ou atributos), cada um pertencente a um tipo de dados específico:
  - int8, uint8, int16, uint16, int32, uint32, int64, uint64, float, double, string, datetime, datetimetz...
- Cell = (d_1, d_2, ..., d_n).

```
mod09q1
<red:int16, nir:int16, quality:uint16>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```

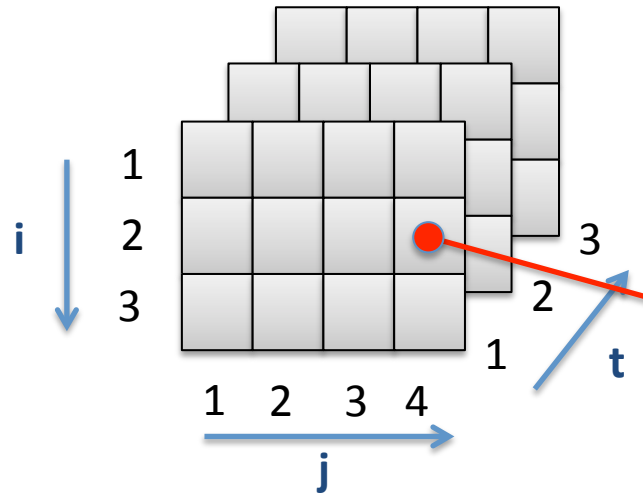


Cada célula do array mod09q1 terá 3 atributos

Arrays: Atributos

- Cada Célula pode estar associada a múltiplos valores (ou atributos), cada um pertencente a um tipo de dados específico:
  - int8, uint8, int16, uint16, int32, uint32, int64, uint64, float, double, string, datetime, datetimetz...
- Cell = (d_1, d_2, ..., d_n).

```
mod09q1
<red:int16, nir:int16, quality:uint16>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```



Cada célula do array mod09q1 terá 3 atributos

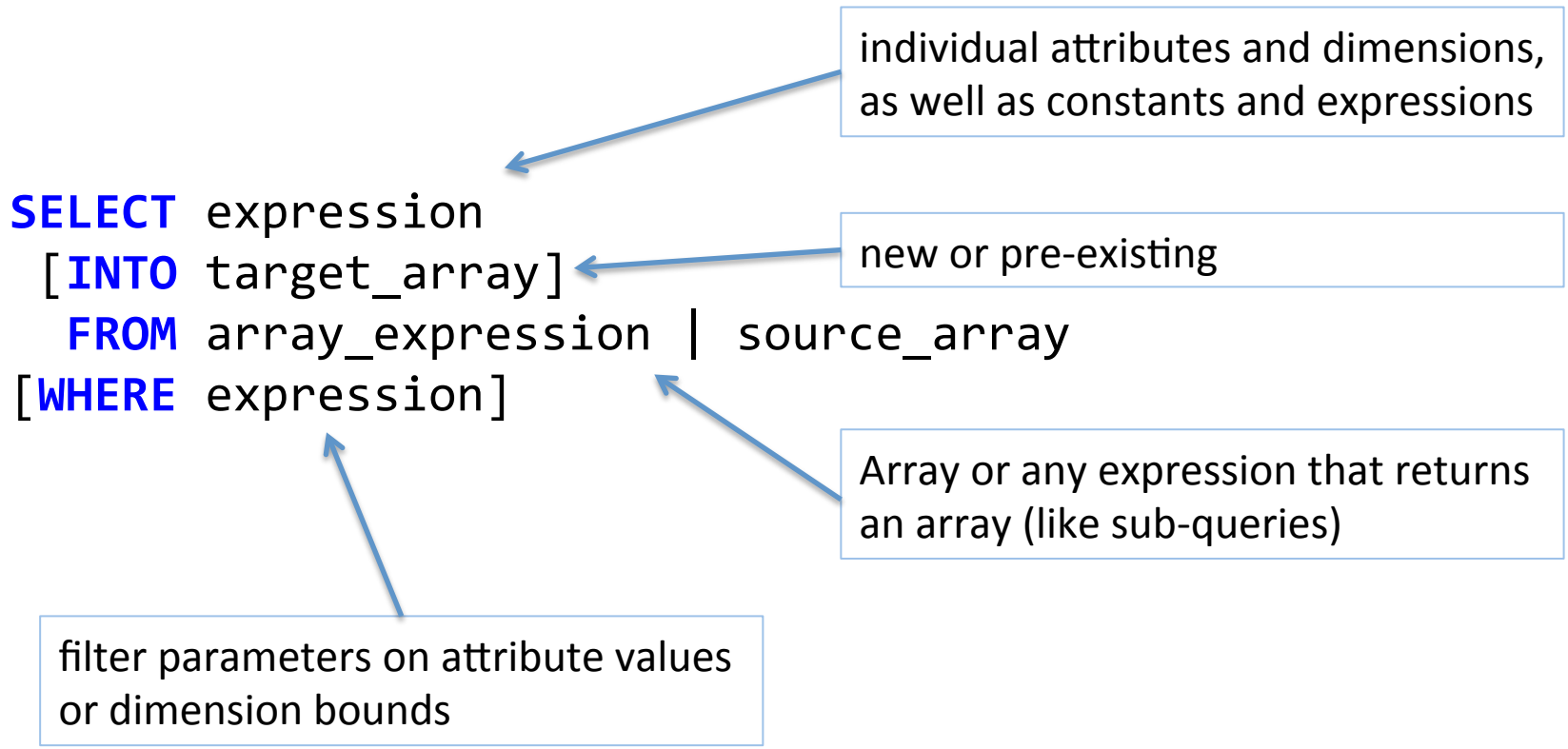
Ex: a célula(4, 2, 1) pode estar associada aos valores:
red = 474; nir = 3109; quality = 4096

Linguagens de Consulta

Array **Q**uery **L**anguage (AQL)

Array **F**unctional **L**anguage (AFL)

Array **Q**uery **L**anguage: AQL



There are DML and DDL clauses

Array **F**unctional **L**anguage (AFL)

```
store(build(<num:double>[x=0:8,1,0, y=0:9,1,0],
            random()),
      random_numbers);
```

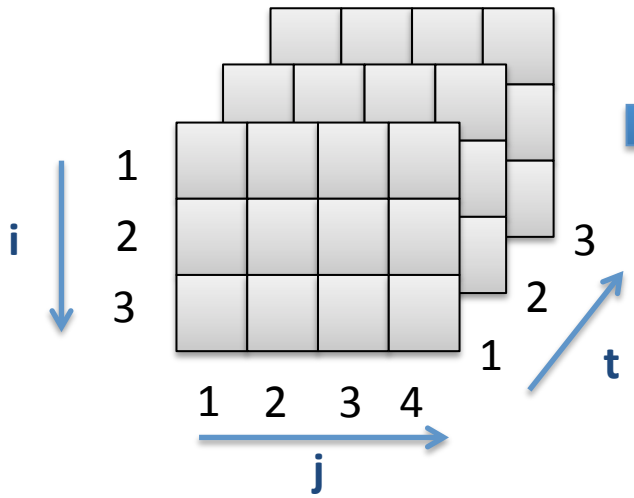
Particionamento de Dados

Chunks

Vertical Partitioning

Arrays: Particionamento

```
mod09q1
<red:int16, nir:int16, quality:uint16>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```



Array é dividido em chunks
chunk-size: 2 x 1 x 3

particionamento
vertical p/ atributo

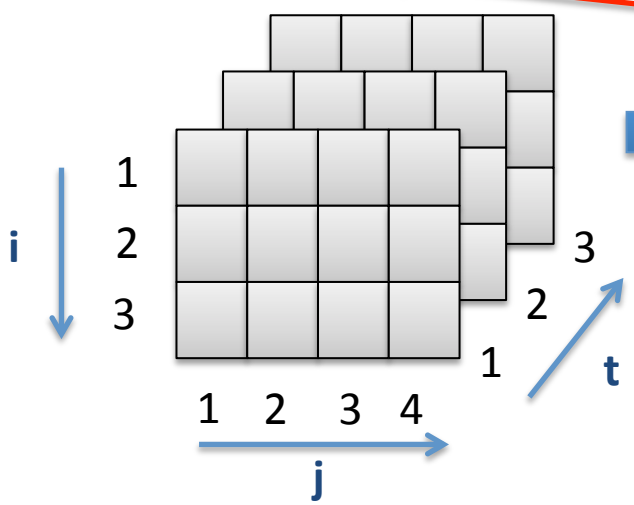
chunk with
red data

chunk with
quality data

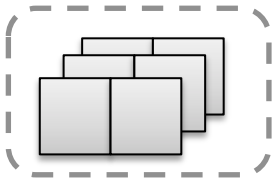
chunk with
quality data

Arrays: Particionamento

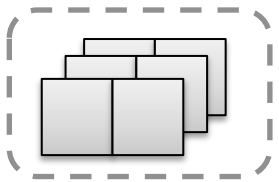
```
mod09q1
<red:int16, nir:int16, quality:uint16>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```



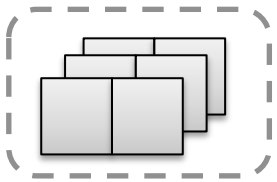
Array é dividido em chunks
chunk-size: 2 x 1 x 3



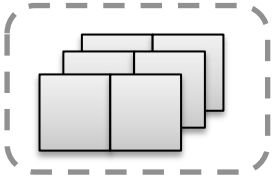
particionamento
vertical p/ atributo



chunk with
red data



chunk with
quality data



chunk with
quality data

Arrays: Particionamento

- Não existe uma B^+-tree ou R-tree:
  - As dimensões formam a base da indexação dos dados

- O dado é "quebrado" em *chunks* e mapeados através de uma função *hash* para cada instância do cluster.

- Uma questão importante trata-se de como especificar o número de células ao longo de cada dimensão que será usado para estabelecer o tamanho do *chunk*.

Replicação

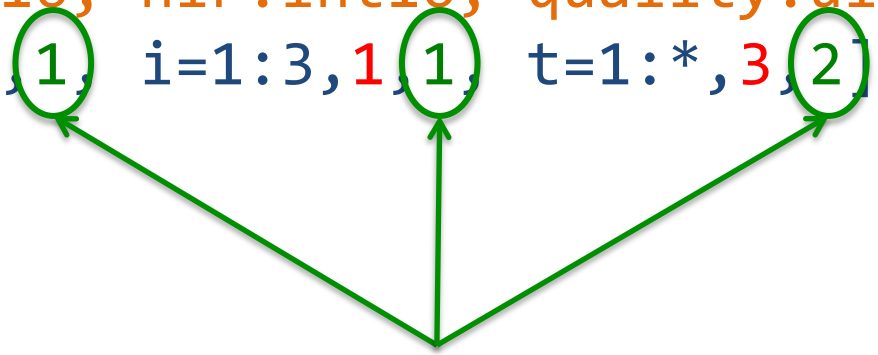
Overlap de Células

Replicação de Chunks entre as Instâncias do *Cluster*

Arrays: Replicação Células de Borda

- É possível definir um fator chamado de *overlap* para as células de borda de um *chunk*:
  - Trata-se de uma boa estratégia para acelerar consultas envolvendo vizinhança (*neighborhood*).

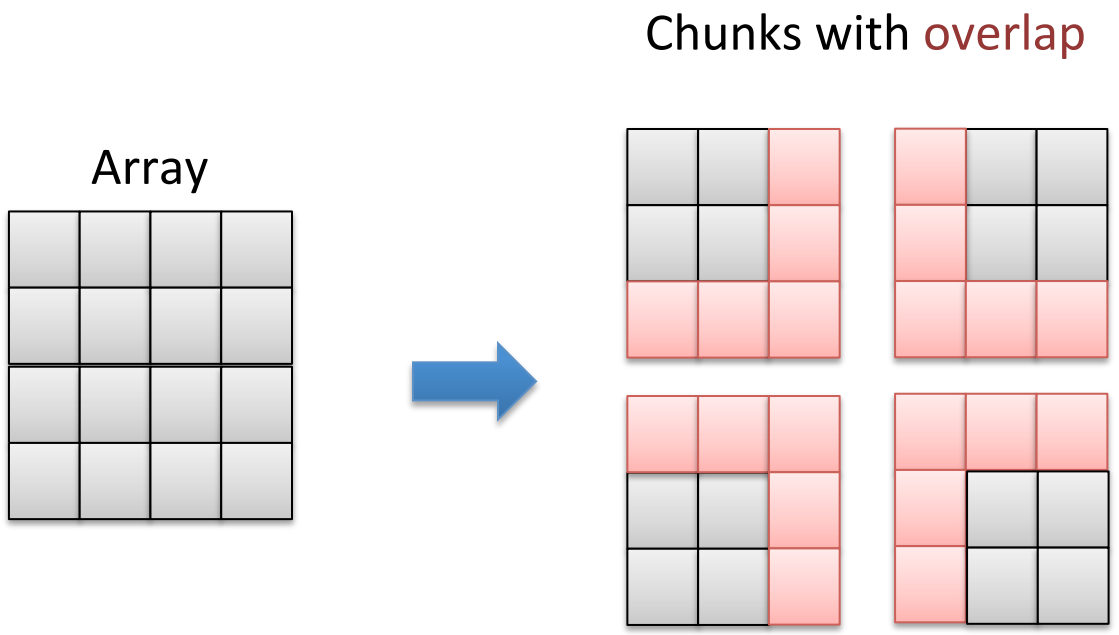
```
mod09q1
<red:int16, nir:int16, quality:uint16>
[j=1:4,2,1, i=1:3,1,1, t=1:*,3,2]
```



Cada *chunk* conterá uma célula de borda nas dimensões *j* e *i*
Na dimensão *t* serão duas células de borda
Essas células são utilizadas apenas pelo SciDB durante as consultas

Arrays: Replicação Células de Borda

- Ilustração da replicação das células de borda:



Arrays: Replicação de Chunks entre as Instâncias

Tolerância a falhas

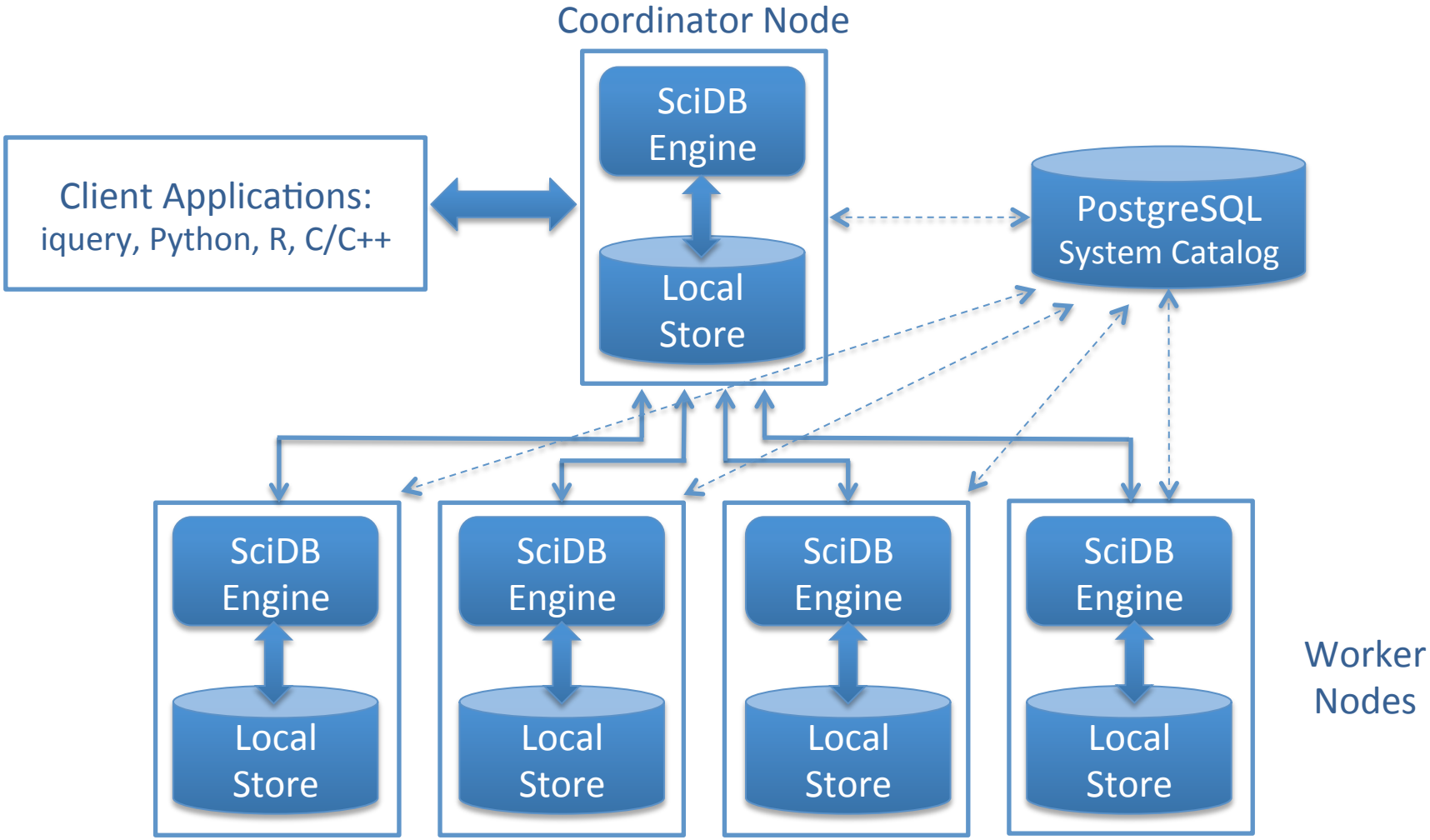
Outras Características/Funcionalidades

- Can use RLE for data compression.
- Cache for frequently used chunks.
- Versioning:
  - The storage model is based on the idea of a “no overwrite”.
  - Update queries write a new full chunk or a delta chunk.
- Storage segments:
  - Reserved contiguous space on disk for a collection of chunks belonging to the same array.
- Provision for temporary storage during query execution.
- Single statement (ACID) transactions:
  - Array-level locking.
- Robust handling of empty cells, including NULL and code for missing data

Integração com Diversos Ambiente

- REST API:
  - Shim
- Linguagens de Programação:
  - C++ e Python
  - Novidade: Julia
- Ambientes estatísticos:
  - R
- Álgebra linear:
  - ScaLAPACK — Scalable Linear Algebra PACKage

Arquitetura



Source: Adapted from PARADIGM4

Referências

Livros

- John R. Jensen. ***Remote Sensing of the Environment***.

- ELMASRI, R.; NAVATHE, S. B. ***Fundamentals of database systems***. Addison Wesley, 2006. 1139p.

- DATE, C. J. ***An introduction to database systems***. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1991.

Artigos

- STONEBRAKER, M.; BROWN, P.; POLIAKOV, A.; RAMAN, S. ***The architecture of SciDB***. In Proceedings of the 23rd international conference on Scientific and statistical database management (SSDBM'11), Judith Bayard Cushing, James French, and Shawn Bowers (Eds.). Springer-Verlag, Berlin, Heidelberg, 2011, 1-16.

- TAFT, R.; VARTAK, M.; SATISH, N. R.; SUNDARAM, N.; MADDEN, S.; STONEBRAKER, M. ***Genbase: a complex analytics genomics benchmark***. Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2013-028, November 19, 2013.

Artigos

- E. F. Codd. 1970. ***A relational model of data for large shared data banks***. *Communications of the ACM, v.* 13, n. 6, June 1970, pp. 377-387.

- Chen, P. ***The Entity-Relationship Model-Toward a Unified View of Data***. ACM Transactions on Database Systems, vl. 1, n. 1. March 1976, pp. 9-36.

- GRAY, J. ***Evolution of Data Management***. IEEE Computer 29(10): 38-46, 1996.

- Vijlbrief, T., and P. van Oosterom. ***The GEO++ System: An Extensible GIS***. Proc. 5th Intl. Symposium on Spatial Data Handling, Charleston, South Carolina, 1992, 40-50.

Especificações e Padrões

- OGC. ***OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 1: Common architecture***. Available at: http://www.opengeospatial.org. Access: October, 2012.

- OGC. ***OpenGIS Implementation Specification for Geographic information - Simple feature access - Part 2: SQL option***. Available at: http://www.opengeospatial.org. Access: October, 2012.

- ISO. ***SQL Multimedia and Application Packages – Part 3: Spatial***.

Slides

- NAUGHTON, J. F. ***DBMS Research: First 50 Years, Next 50 Years***. Kynote speaker' slides at ICDE 2010. Disponível em: http://pages.cs.wisc.edu/~naughton/naughtonicde.pptx. Acesso: Abril de 2013.

Videos

- THAKAR, A (Johns Hopkins University). **Billions of Stars on Off-the-shelf Software / The SDSS SkyServer and Beyond: Why "The Not-So-Little-Engine that Could" is Still Chugging Along**. 7th Extremely Large Databases Conference, September 9-12, 2013, Stanford University, California, USA. Video, Slides, Access: 04th April, 2014.

Cursos

- FONSECA, L. M. G.; KÖRTING, T. S. **PDI: Fundamentos**. Notas de Aula SER 437 - Processamento Digital de Imagens de Sensores Remotos. Disponível em: http://www.dpi.inpe.br/cursos/ser437. Acesso: 02 de Maio de 2014.

Manuais

- SciDB Reference Manual: Community and Enterprise Editions Document Version 14.12.2013. Copyright © 2010–2015 Paradigm4, Inc.

Entrevistas

- ***On the SciDB array database***. Interview with Mike Stonebraker and Paul Brown. Disponível em: http://www.odbms.org/blog/2014/04/interview-mike-stonebraker-paul-brown/. Acesso: Novembro, 2015. Blog Roberto V. Zicari.