Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

SimDB

Numerical Simulations using a Multidimensional Array Model

Hermano Lustosa, Fabio Porto
{hermano, fporto}@lncc.br

Laboratório Nacional de Computação Científica

DEXL LAB EXTREME DATA LAB

CNPq

Agenda

- Context
- SciDB (DBMS)
- Chunks Partitioning
- Simulation Data
- Sparsity and Irregularity Problems
- SimDB
- Tests and Results

Big-Data (in science) Data Challenges

- Data Representation
  - Different Data Models:
    - Data structure and query languages
    - Graphs, Matrixes, Key-Value,…
- Data Uncertainty
  - Data is uncertain
  - uncertainty quantification on data
- Data Partitioning
  - in sync with data processing
- Data Heterogeneity
  - Data Granularity

Big-Data (in science) Data Challenges

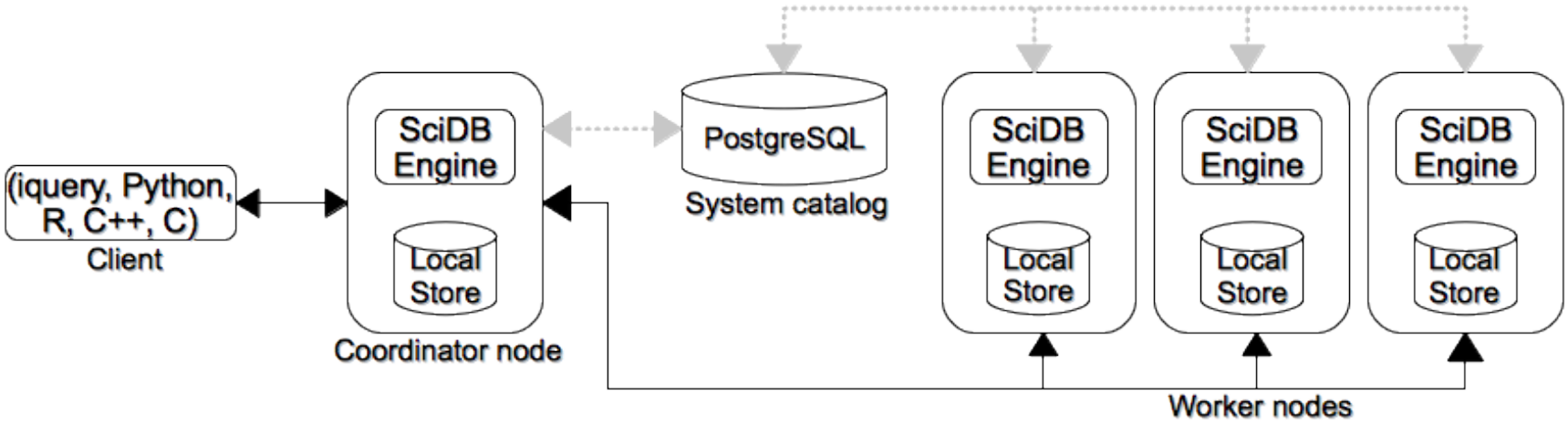
- **Data Representation**
  - **Different Data Models:**
    - **Data structure and query languages**
    - **Graphs, Matrixes, Key-Value,…**
- Data Uncertainty
  - Data is uncertain
  - uncertainty quantification on data
- **Data Partitioning**
  - **in sync with data processing**
- Data Heterogeneity
  - Data Granularity

Context

- Large amount of data generated by numerical simulations
- Stored in simple text files
- RDBMS are inadequate for storing scientific data
- Use of DBMS implemeting a different data model more adequate for analysing scientific data

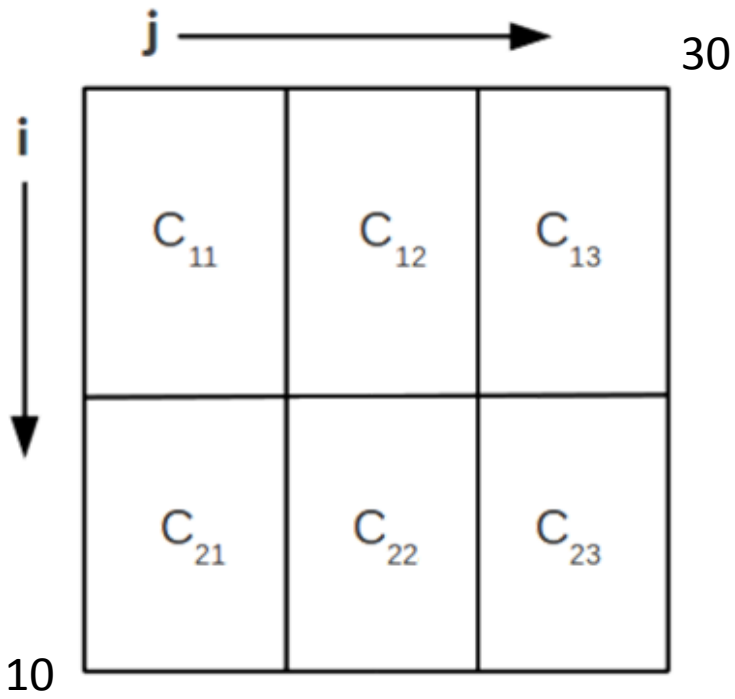
SciDB

- Multidimensional arrays as the basic storage unit
- Arrays defined as a set of named dimensions
- Combinations of indexes for each dimension identify a cell
- Cells can contain associated data (atributes)



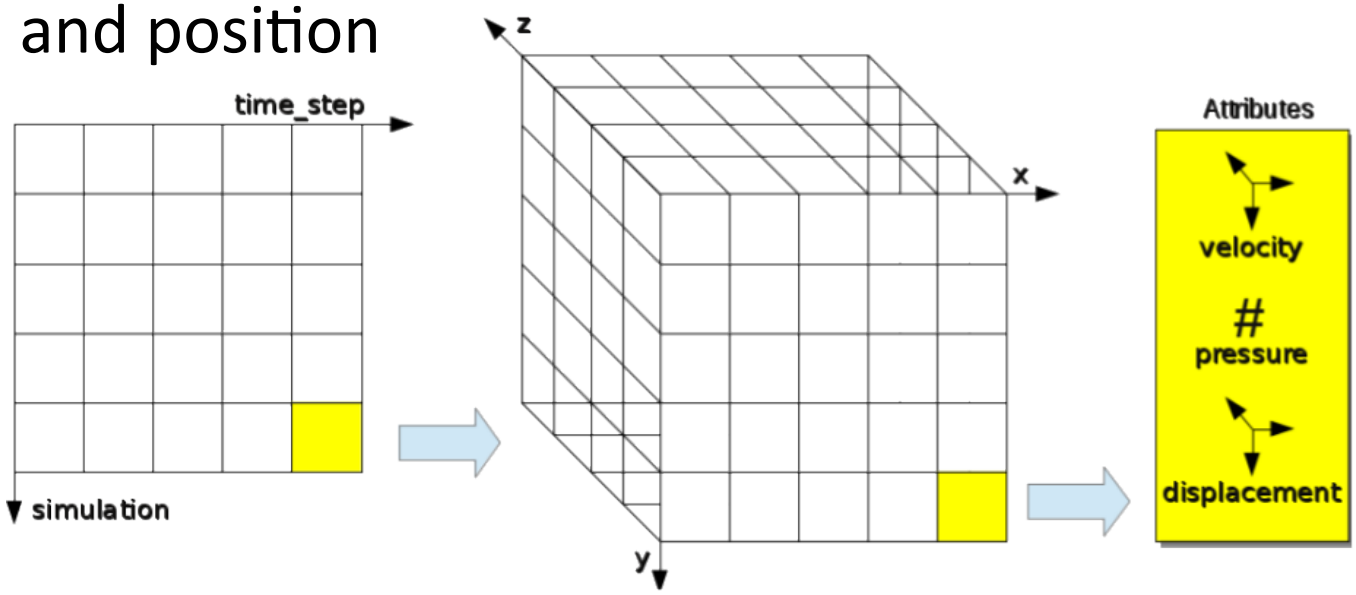
Chunks Partitioning

- Partitioning based on chunks

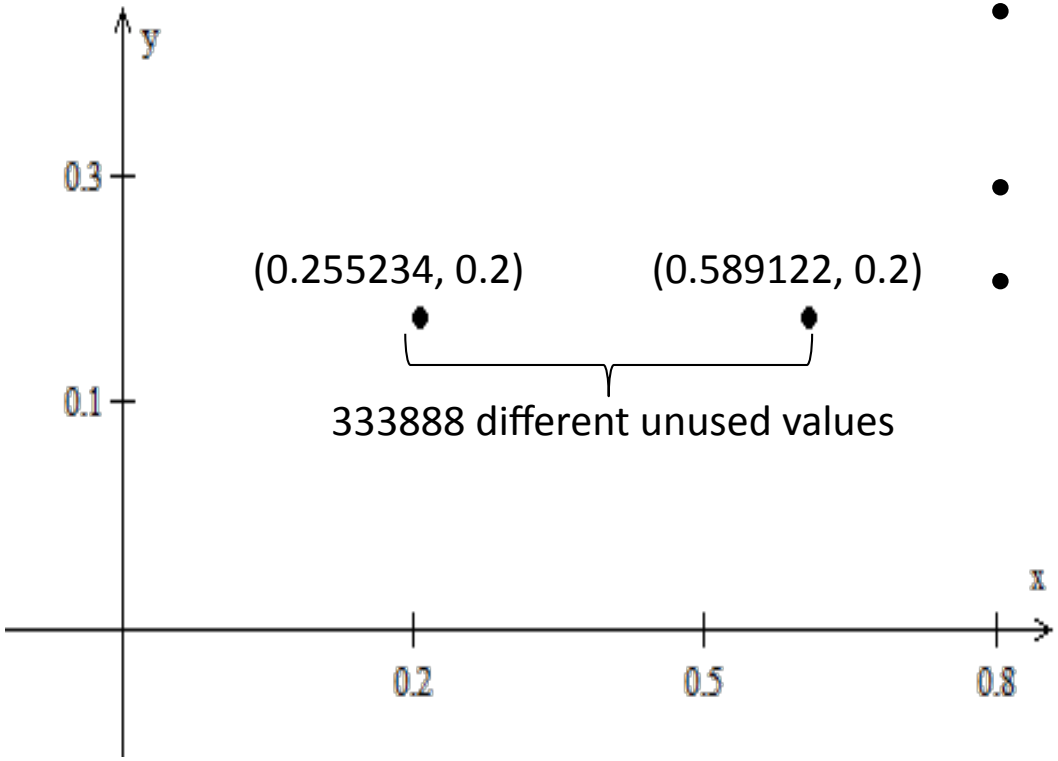


Simulation Data

- Simulation Data provided by HeMoLab
- Geometrical representation of arteries on Multidimensional Array Data Model
- Physical quantities calculated for each simulation, time step and position



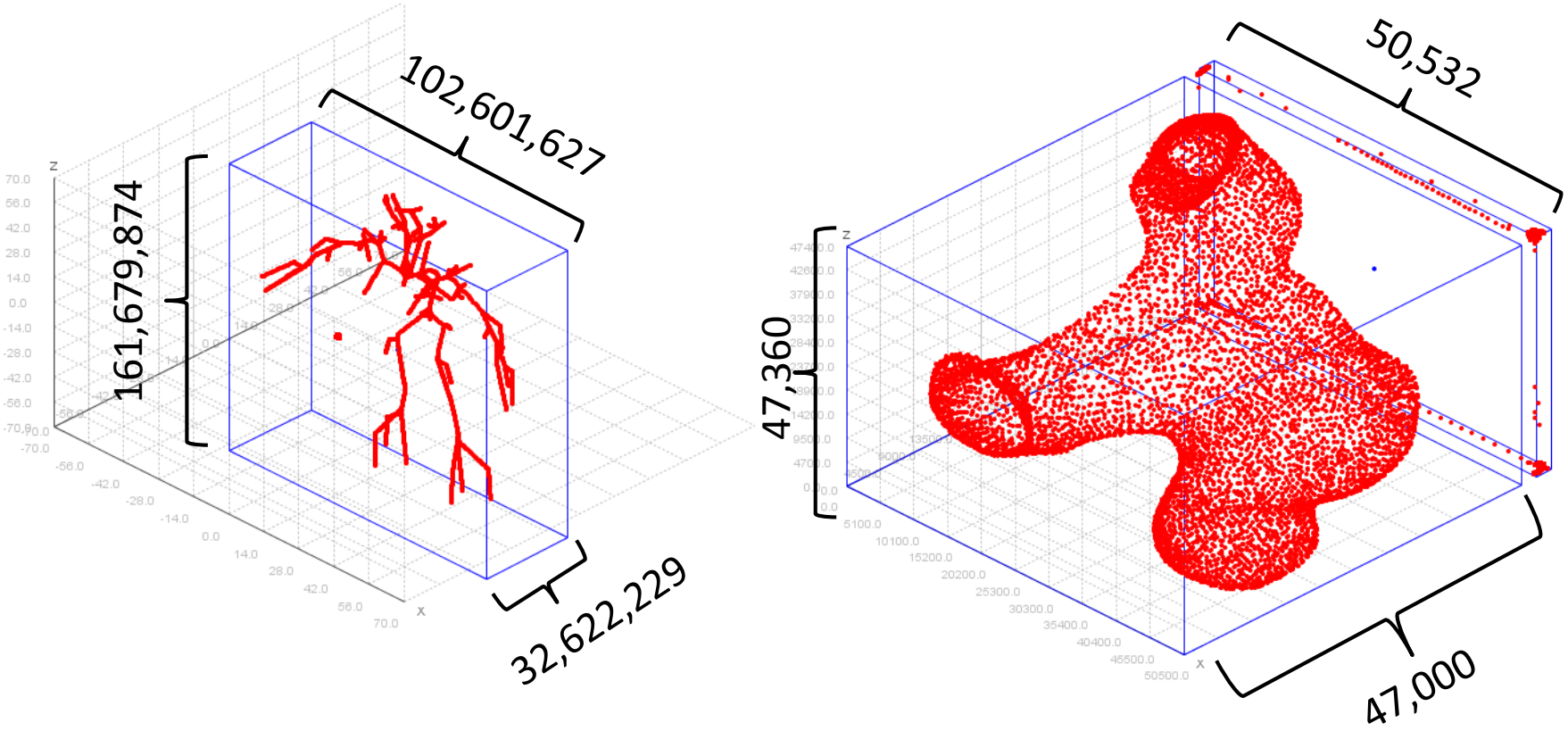
Sparsity and Irregularity Problems



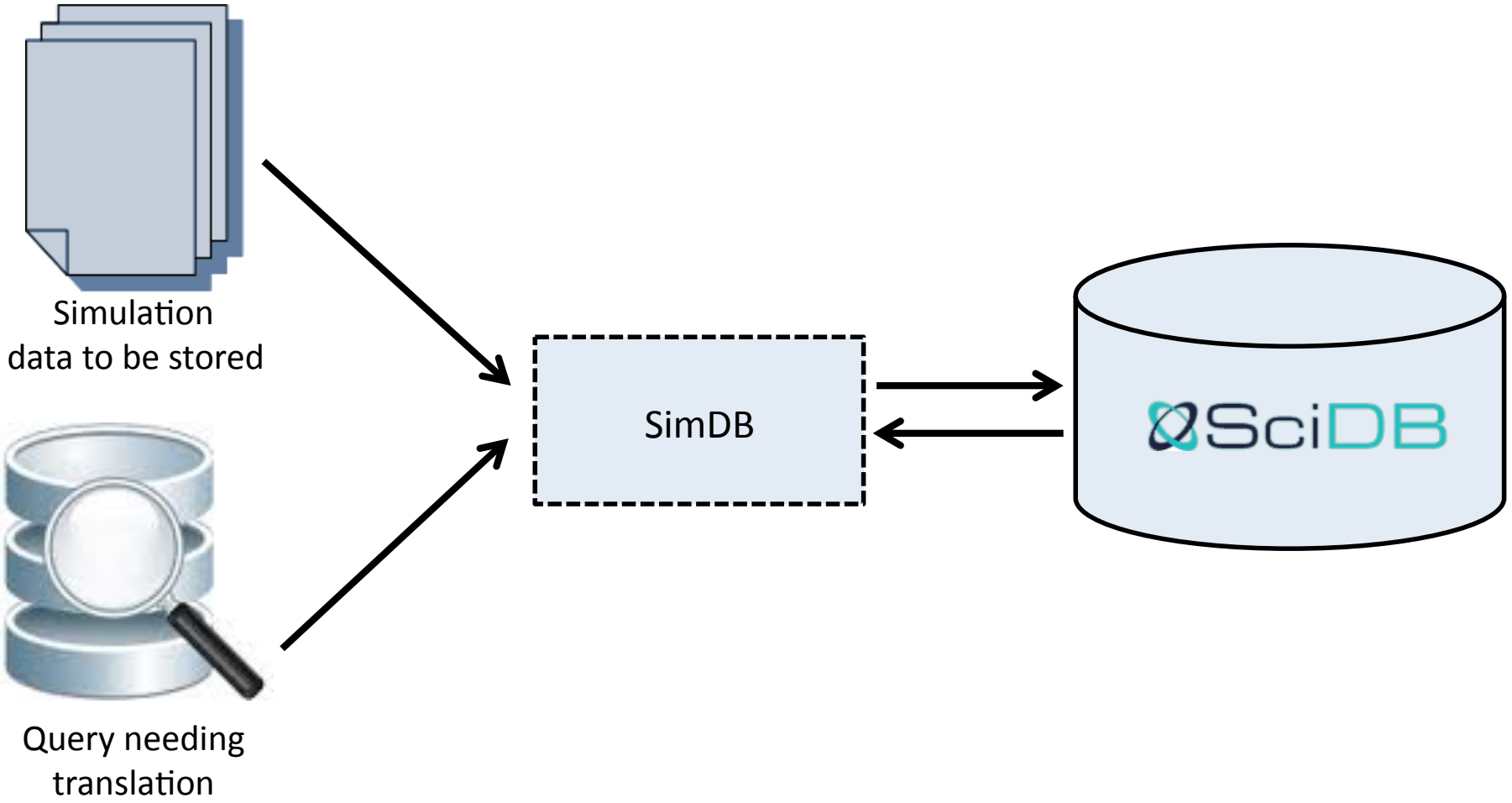
- Coordinates specified with 6 decimal places of precision
- Points distributed irregularly
- Solutions (transform real values into integer indexes for the array):
  - Normalizing (multiplying each value for 10^6)
  - Mapping the coordinates

Sparsity and Irregularity Problems

- Normalizing vs. Mapping



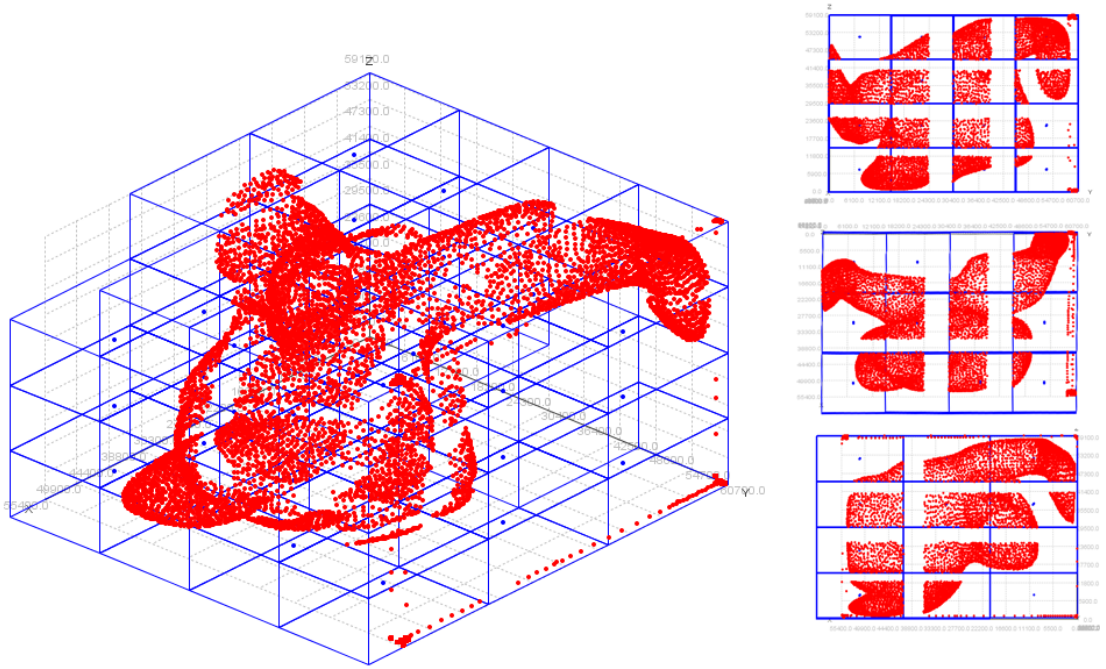
SimDB


Simulation
data to be stored
SimDB
SciDB
Query needing
translation

SimDB

- Step 1 – Using Equi-depth histograms to determine partitioning
- Step 2 -Redistribution of the indexes in such manner that all ranges are stretched to be the same size
- Step 3 - Creation of an artificial dimension to distribute data more evenly among the chunks

SimDB



Tests and Results

- Tests executed using the original data (normalized) and 2 arrays generated by SimDB (mapping + index redistribution).
- Array 3D generated by executing steps 1 and 2 only. And Array 4D generated after the execution of all steps through 1 to 3.
- Query types:
  - Full Scan
  - Range Query
  - SciDB native array operators

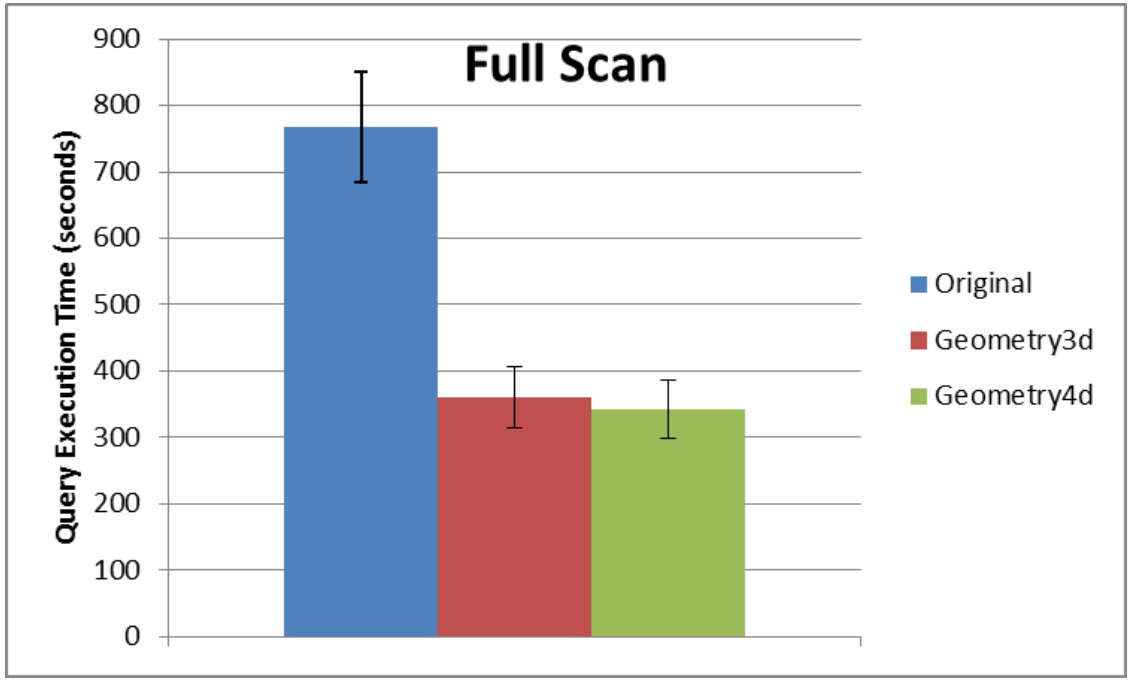
Tests and Results

- Information about the arrays (all of them have aprox. 200kk cells).

Array	Number of Chunks	Average Nº of cells per chunk
Original	1,171,260	170.93
3D	86	2,402,413.33
4D	936	214,348.58

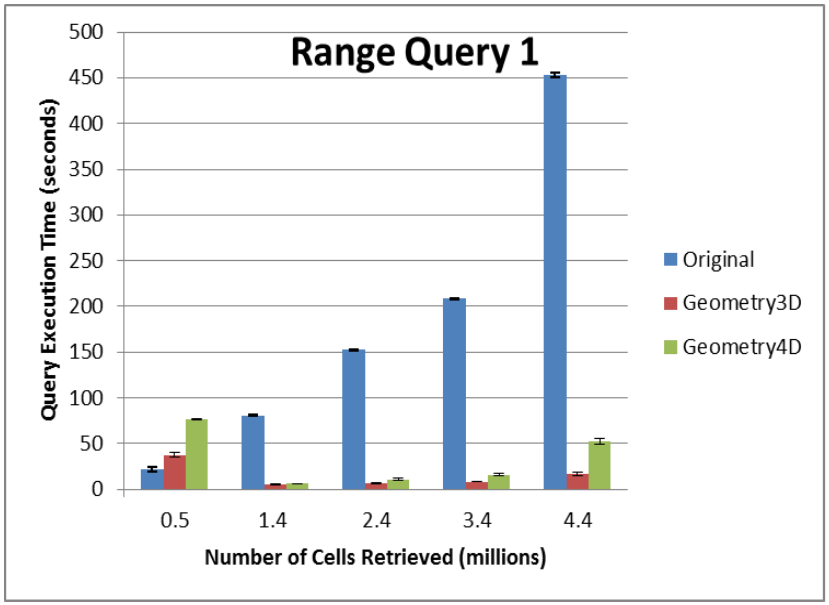
Tests and Results

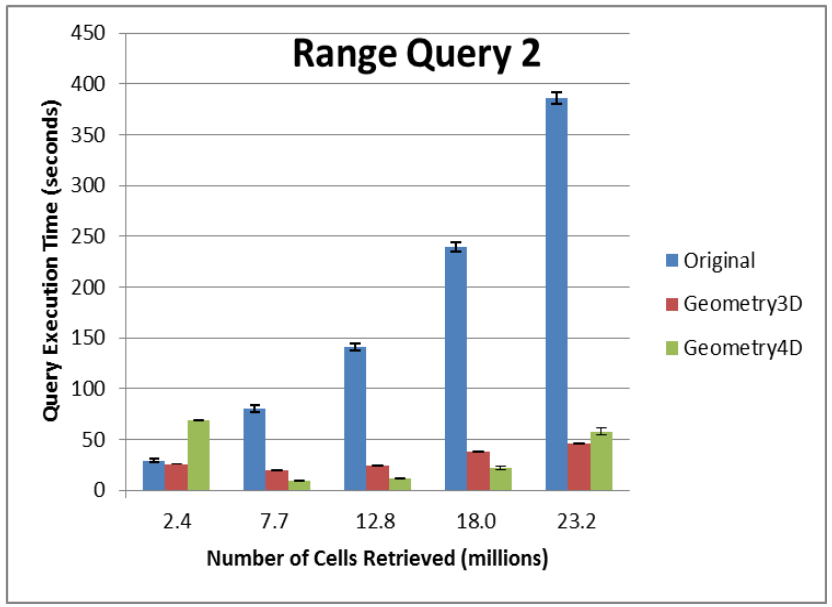
- Full scan



Tests and Results

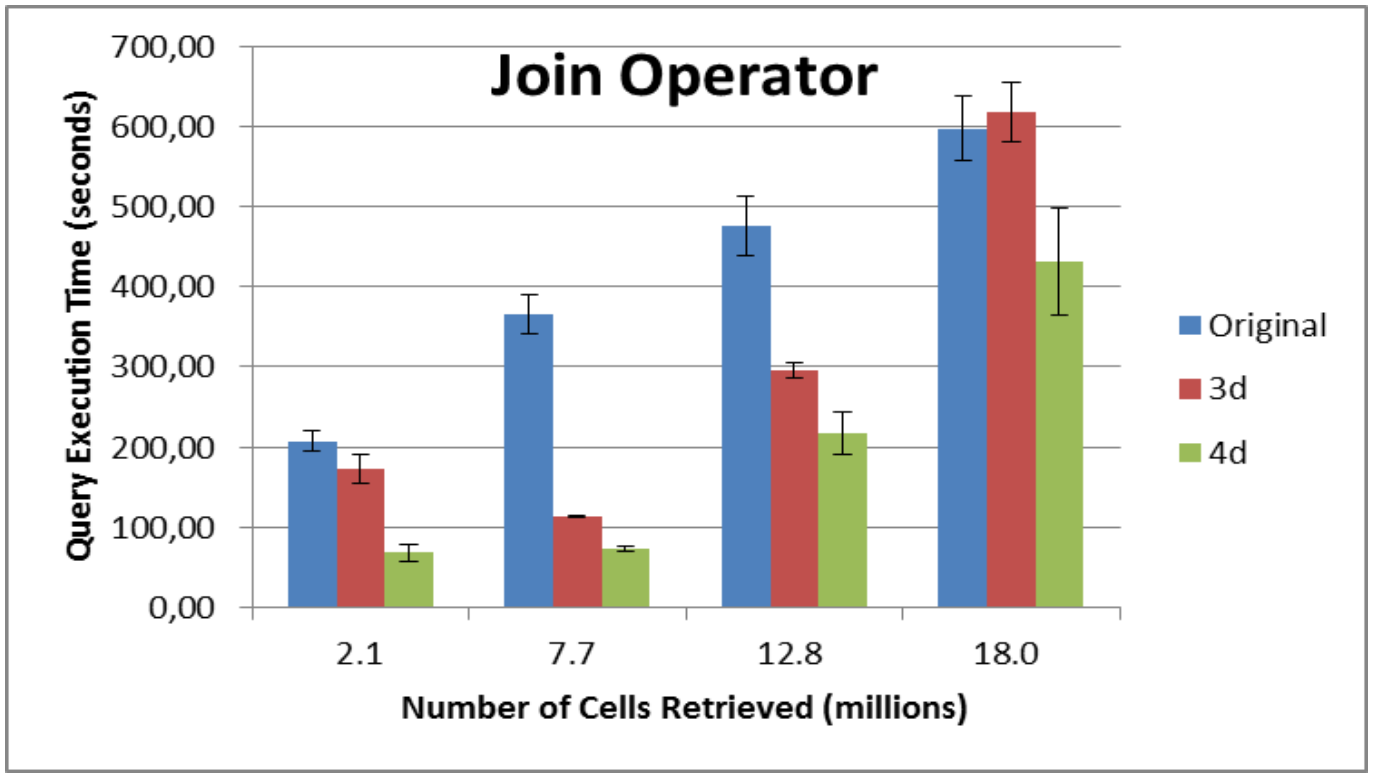
- Range Queries





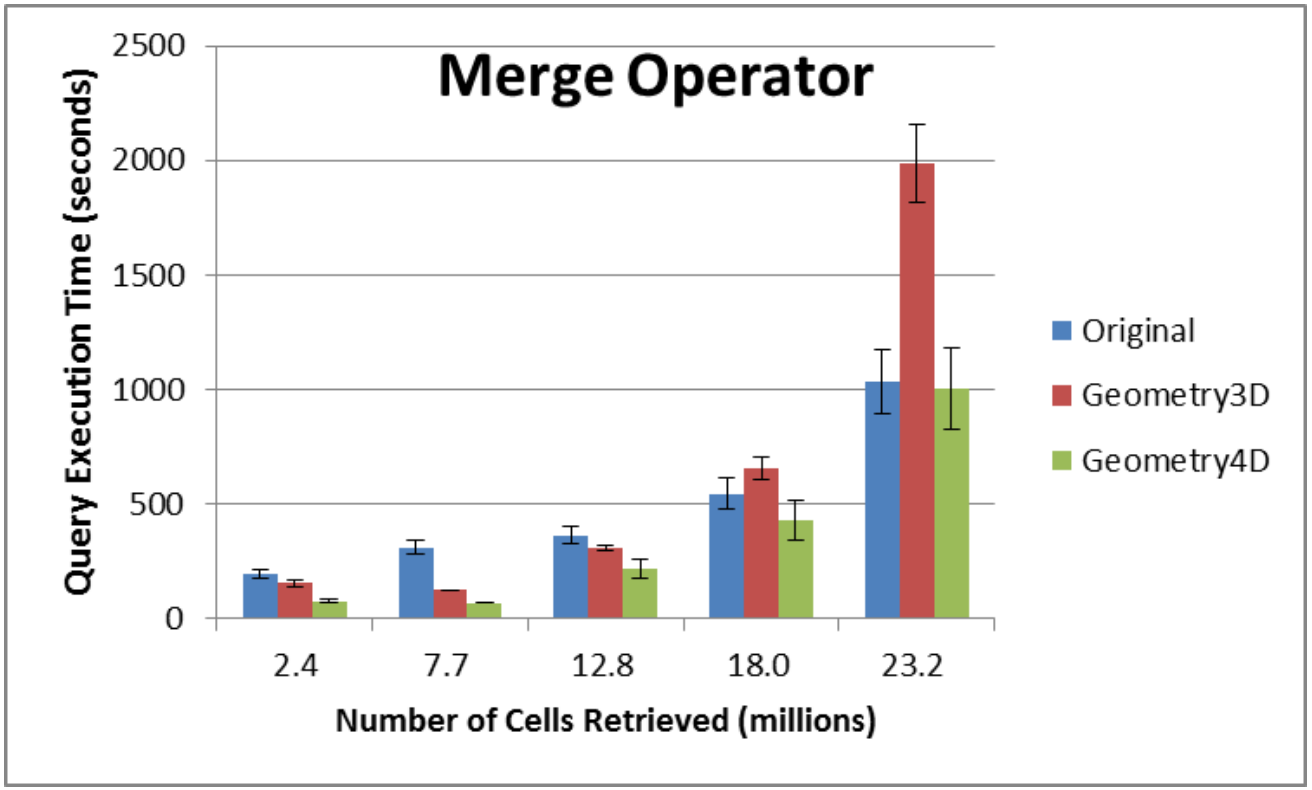
Tests and Results

- Join Operator



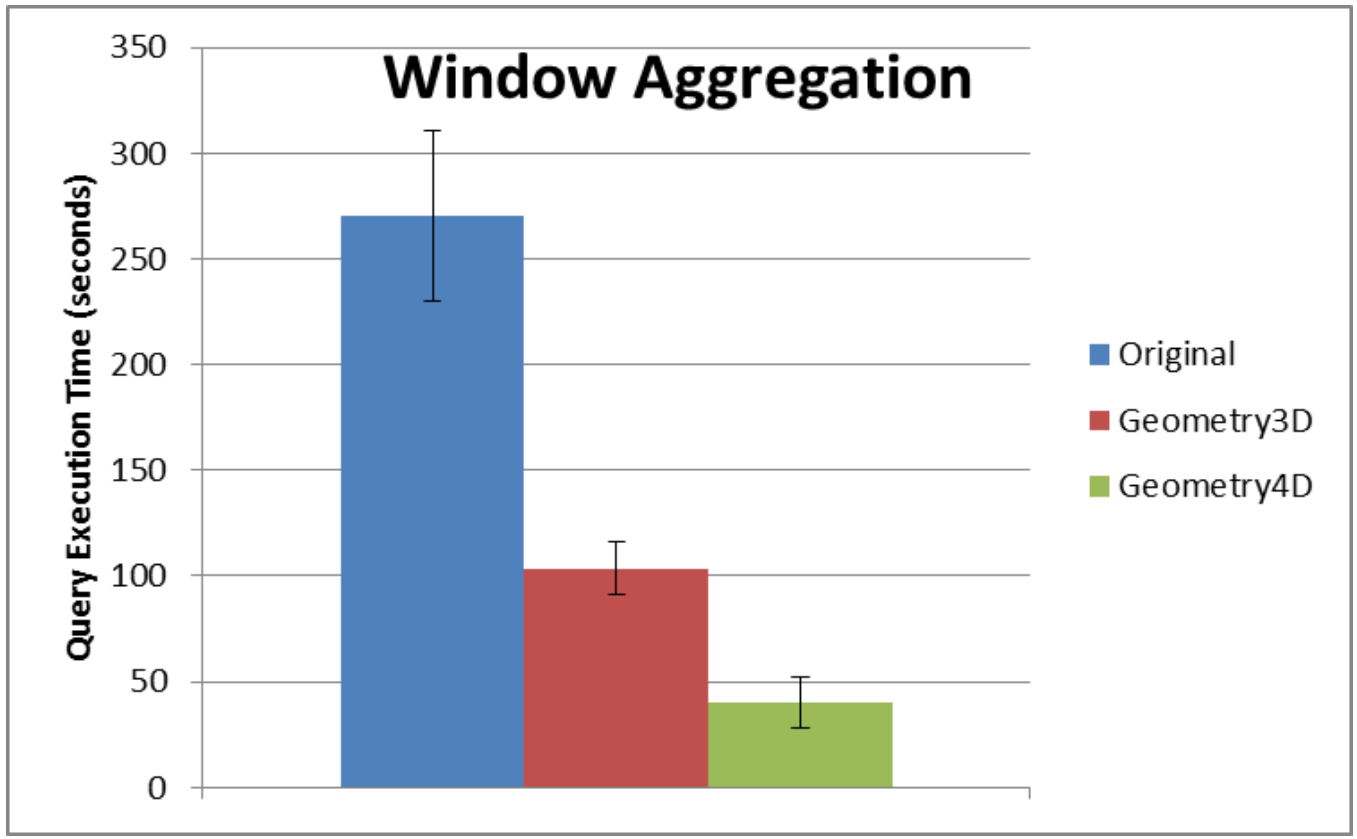
Tests and Results

- Merge Operator



Tests and Results

- Window Aggregation Operator



End of Presentaton

Thank you!