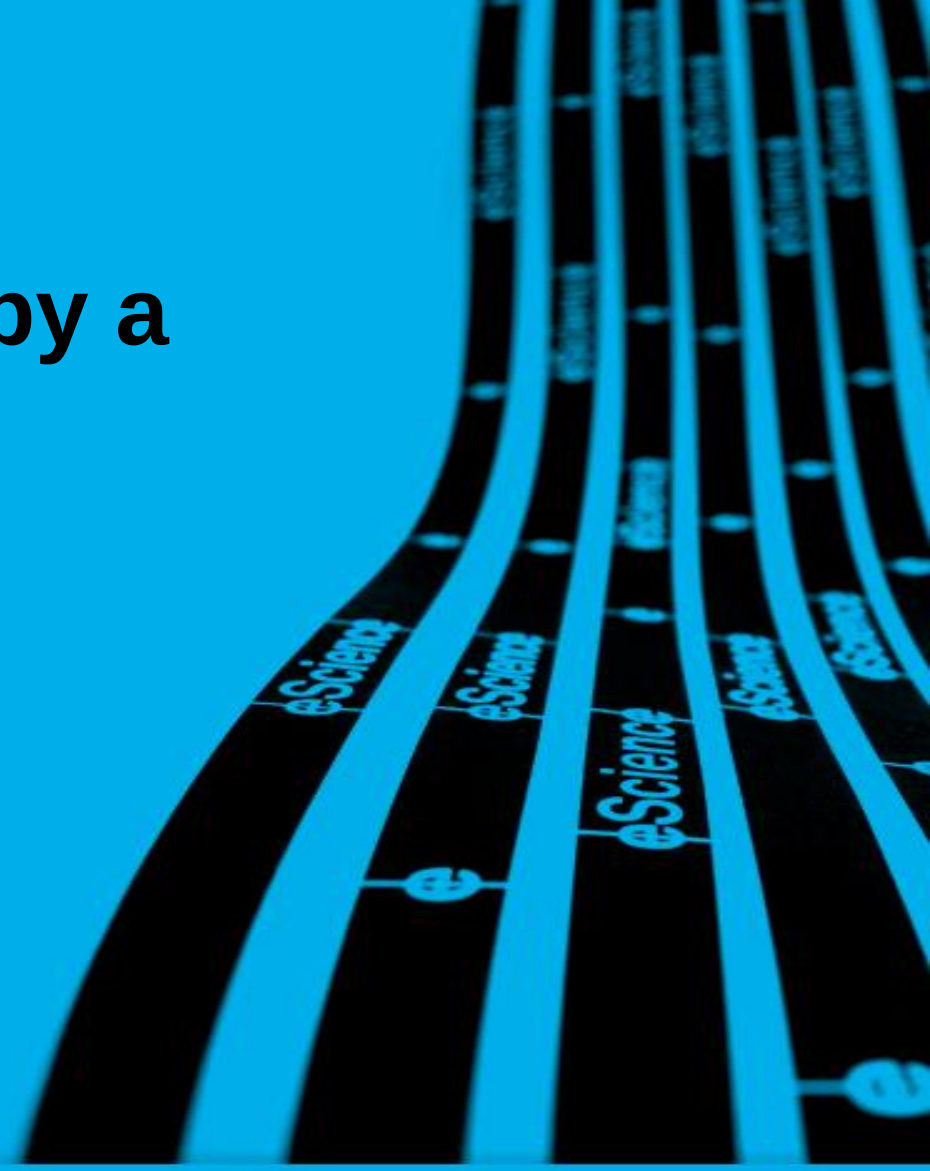
LiDAR data exploration boosted by a column-store.

NLeSC together with CWI DA group

Romulo Goncalves, Kostis Kyzirakos and Dimitar Nedev



netherlands eScience center

by SURF & NWO

CWI Centrum Wiskunde & Informatica

monetdb

Point Cloud usage…

- **Urban planning to develop smart cities**
  - Improve human comfort such as avoiding heat islands
- **Monitoring coastal erosion**
- **Risk management**
  - High resolution flood simulations
  - Align and compare data sets at different resolutions to study the spatial evolution over time of an area or structure.

Aerial LiDAR scanner

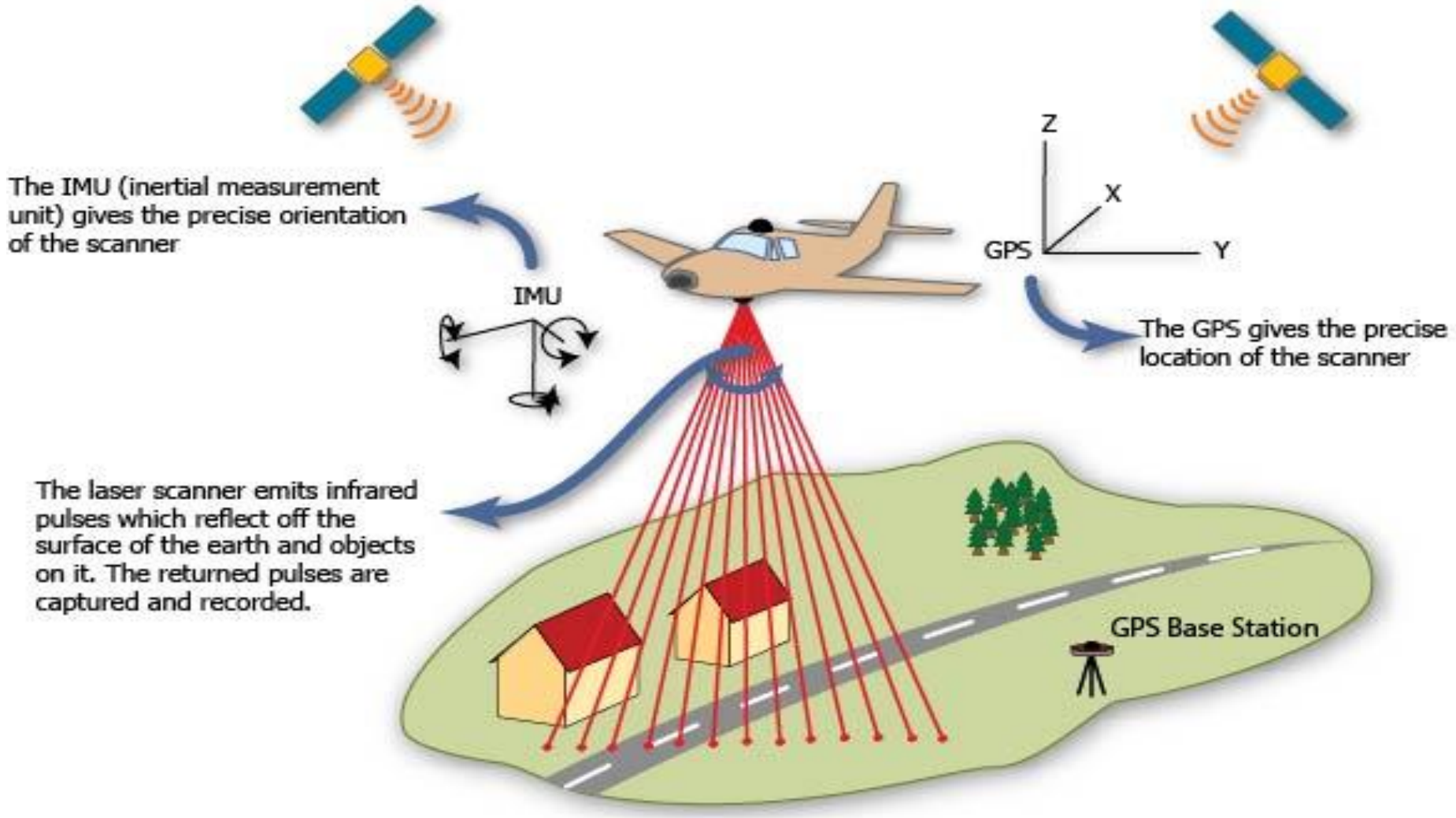


Image from: http://gmv.cast.uark.edu/wp-content/uploads/2013/01/ALS_scematic.jpg

Massive Point Clouds for eScience

- **AHN (actual height model of the Netherlands)**
  - AHN2 640B (60 000 files) -> Size 1 TB (everything merged 2.5 TB)

- **[1]: P. van Oosterom, O. Martinez-Rubi, and et al. Massive point cloud data management: design, implementation and execution of a point cloud benchmark. Computer Graphics, 2015.**

- **Databases for Massive Point Clouds Presentation in SPAR Europe Conference 2014**

- **The website is** http://pointclouds.nl

Technology spectrum

- Specialized libraries
  - File-based systems rich in functionality
  - Domain specific formats
  - Efficient access to the data in its original format

  - Data isolation
  - Applications become dependent on data formats and humans
  - Poor scalability
  - The data interchange is tedious

Technology spectrum

- Specialized libraries
  - File-based systems rich in functionality
  - Domain specific formats
  - Efficient access to the data in its original format
  - Data isolation
  - Applications become dependent on data formats and humans
  - Poor scalability
  - The data interchange is tedious

- Data Management Systems
  - Split between physical and logical structure
  - SQL – declarative language
  - Scalability
  - Data conversion and data loading
  - Basic data filtering reads not as efficient as file-based systems
  - Black boxes
  - Front-ends not user friendly

Solutions

- **Rapidlasso LAStools**
  - **Elegant tool and known for its efficiency**
  - **Advocates LAS/LAZ file format as a standard**
  - **Point of reference for LiDAR data extraction**

- **Spatial DBMS**
  - **Combine heterogeneous data sets**
  - **For analytical workload**
  - **Complex queries**

Benchmark setup

- Single Machine 32 cores and 256GB main memory and 2 RAIDs (24 disks each)

Benchmark setup

- Single Machine 32 cores and 256GB main memory and 2 RAIDs (24 disks each)

- LAStools
  - Open source, unlicensed LAStools
    - It adds noise to the data
  - Data is sorted and indexed in parallel
  - Query execution
    - Single thread
    - List of files obtained using PostGIS
    - LAS for performance reasons

Benchmark setup

- MonetDB
  - Jul2015 release

Benchmark setup

- MonetDB
  - Jul2015 release
  - Partition table
    - A view over set of tables
      - Create table AHN2 (…);
      - Alter table AHN2 add ahn_t1;

Benchmark setup

- MonetDB
  - Jul2015 release
  - Partition table
    - A view over set of tables
      - Create table AHN2 (…);
      - Alter table AHN2 add ahn_t1;
  - Queries
    - Create table result as select… x between… y between… and contains() with data;
      - Filtering step
      - Refinement step

Data Conversion

Data Conversion

- Data conversion
  - Las2col
  - double or decimal(9,2)
  - Binary format for append
    - COPY BINARY INTO

Data Conversion

- Data conversion
  - Las2col
  - double or decimal(9,2)
  - Binary format for append
    - COPY BINARY INTO

- Convert 64 000 laz files
  - 20 000: 21458.994 **secs (5:57 hours)**
    2 000: 29523.665 **secs (8:12 hours)**

Data Conversion

- Data conversion
  - Las2col
  - double or decimal(9,2)
  - Binary format for append
    - COPY BINARY INTO

- Convert 64 000 laz files
  - 20 000: 21458.994 **secs (5:57 hours)**
    2 000: 29523.665 **secs (8:12 hours)**

- On going:
  - Improve parallelism for data conversion

Data Loading

Data Loading

- Binary format for append
  - COPY BINARY INTO
  - 20 000 : 3738 secs (1:02 hours)
    2 000 : 279 secs

Data Loading

- Binary format for append
  - COPY BINARY INTO
  - 20 000 : 3738 secs (1:02 hours)
    2 000 : 279 secs

- Indexing
  - Triggered by a range selection, one table at the time
  - 20 000: 14564.885 (4 hours)
    2 000: 12285.073 (3:24 hours)

Data Loading

- Binary format for append
  - COPY BINARY INTO
  - 20 000 : 3738 secs (1:02 hours)
    2 000 : 279 secs

- Indexing
  - Triggered by a range selection, one table at the time
  - 20 000: 14564.885 (4 hours)
    2 000: 12285.073 (3:24 hours)

- Ongoing:
  - Parallelism for data append and index creation

Data Preparation

- **LAStools (with LAS)**
  - **12h:30min**
  - **11TB storage footprint**

Data Preparation

- **LAStools (with LAS)**
  - **12h:30min**
  - **11TB storage footprint**

- **MonetDB**
  - **11h:18min**
  - **7.5 TB storage footprint**

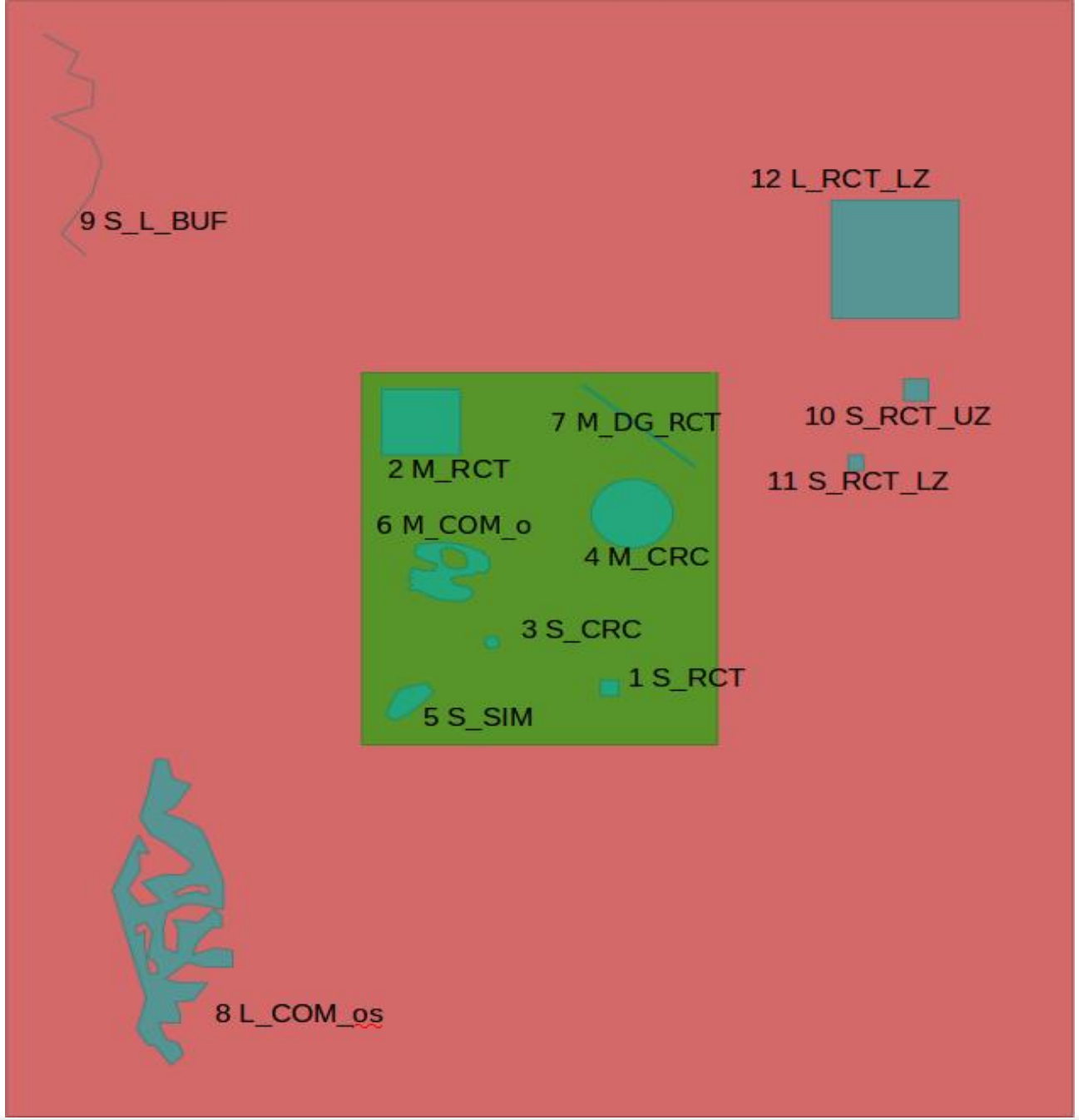
Data Preparation

- **LAStools (with LAS)**
  - **12h:30min**
  - **11TB storage footprint**

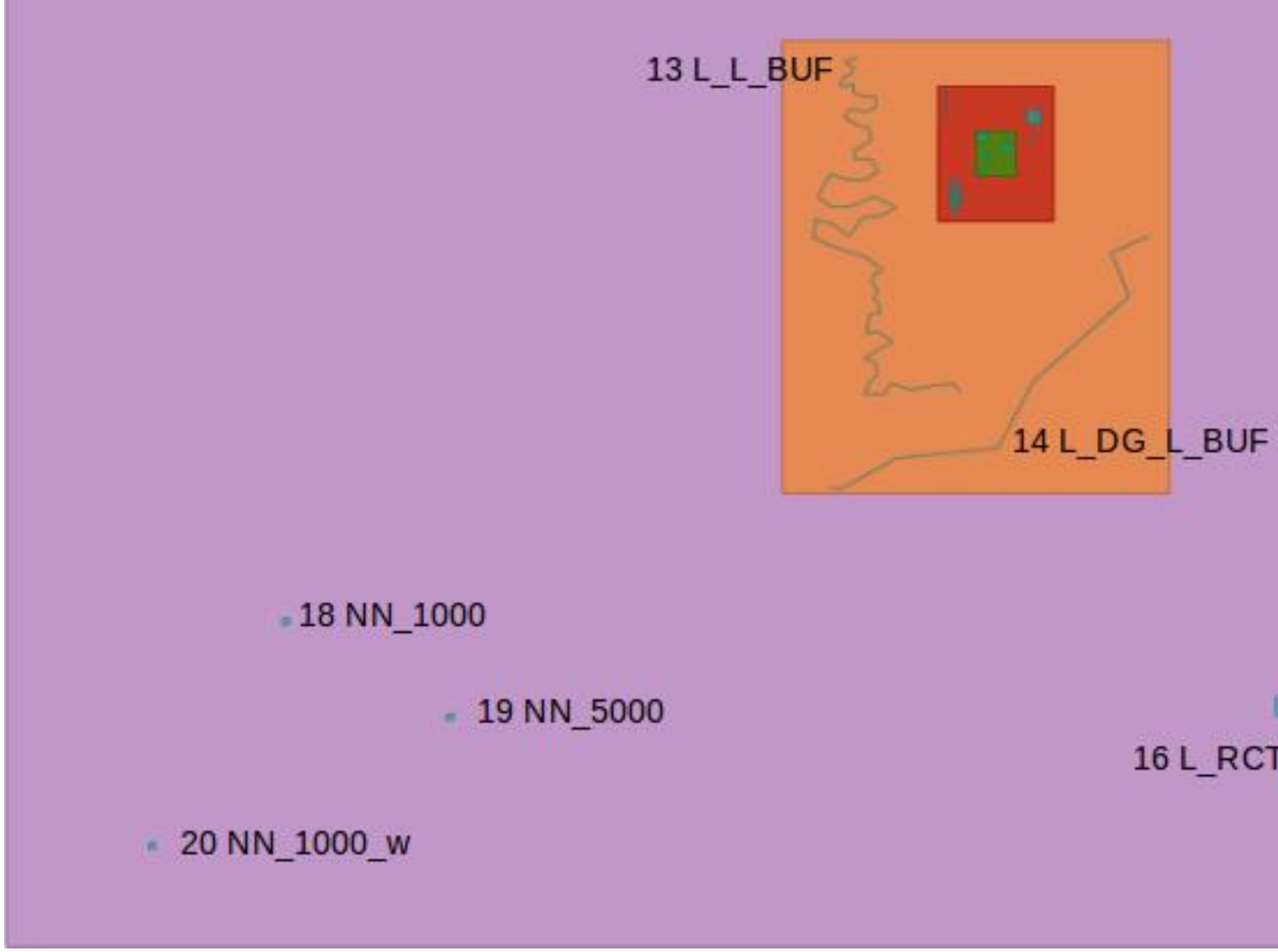
- **MonetDB**
  - **11h:18min**
  - **7.5 TB storage footprint**

  - **Ongoing:**
    - **Compression 2,5TB**
    - **In-situ data access**

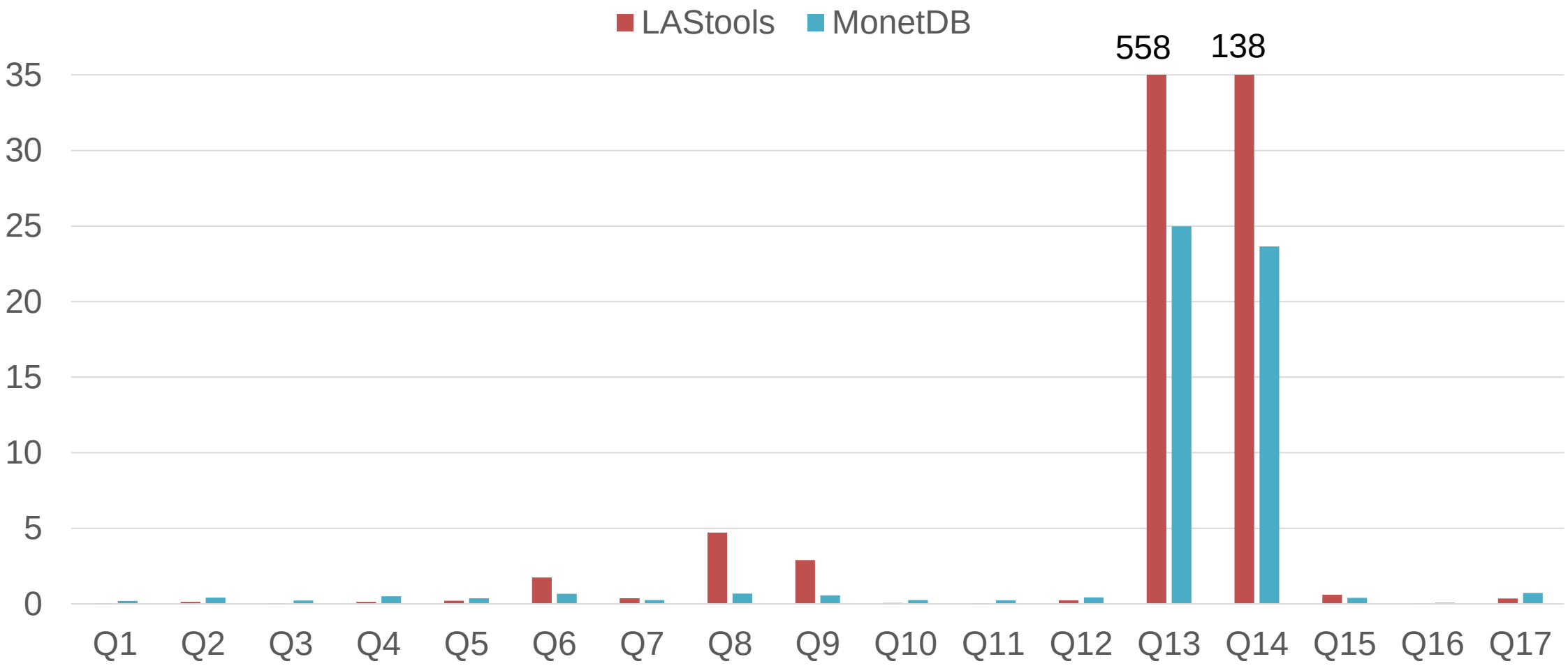
Queries…


9 S_L_BUF
12 L_RCT_LZ
7 M_DG_RCT
10 S_RCT_UZ
2 M_RCT
11 S_RCT_LZ
6 M_COM_o
4 M_CRC
3 S_CRC
1 S_RCT
5 S_SIM
8 L_COM_os

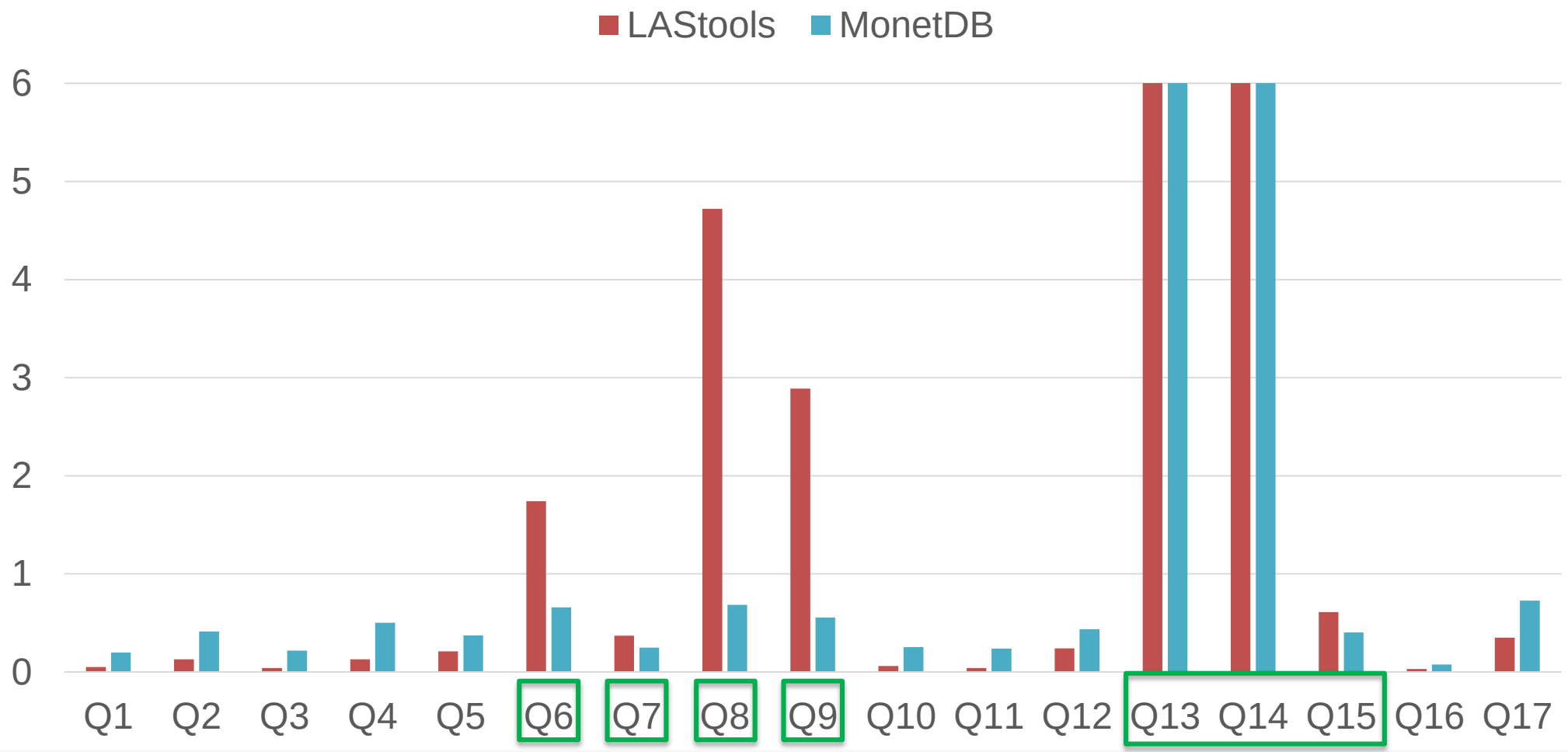
Queries…



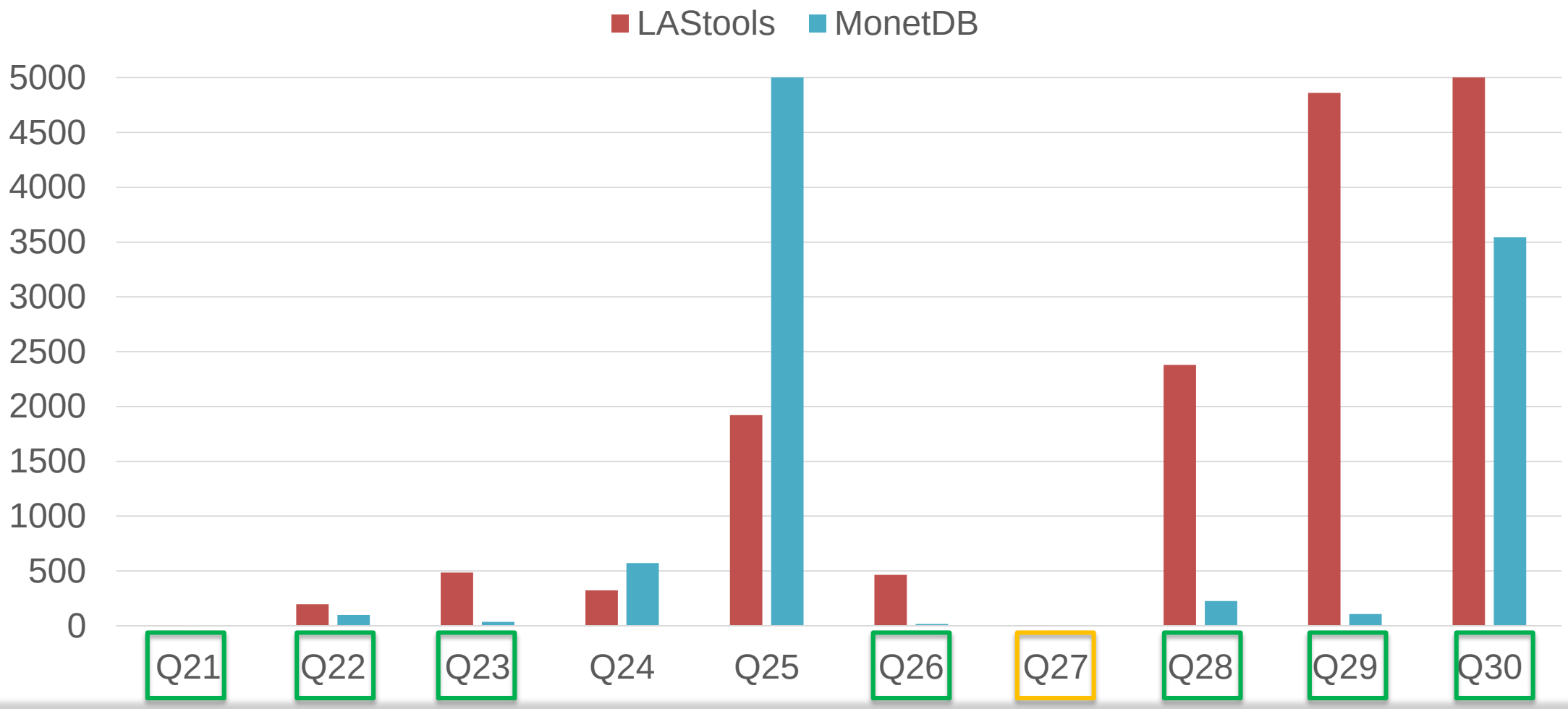
Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)



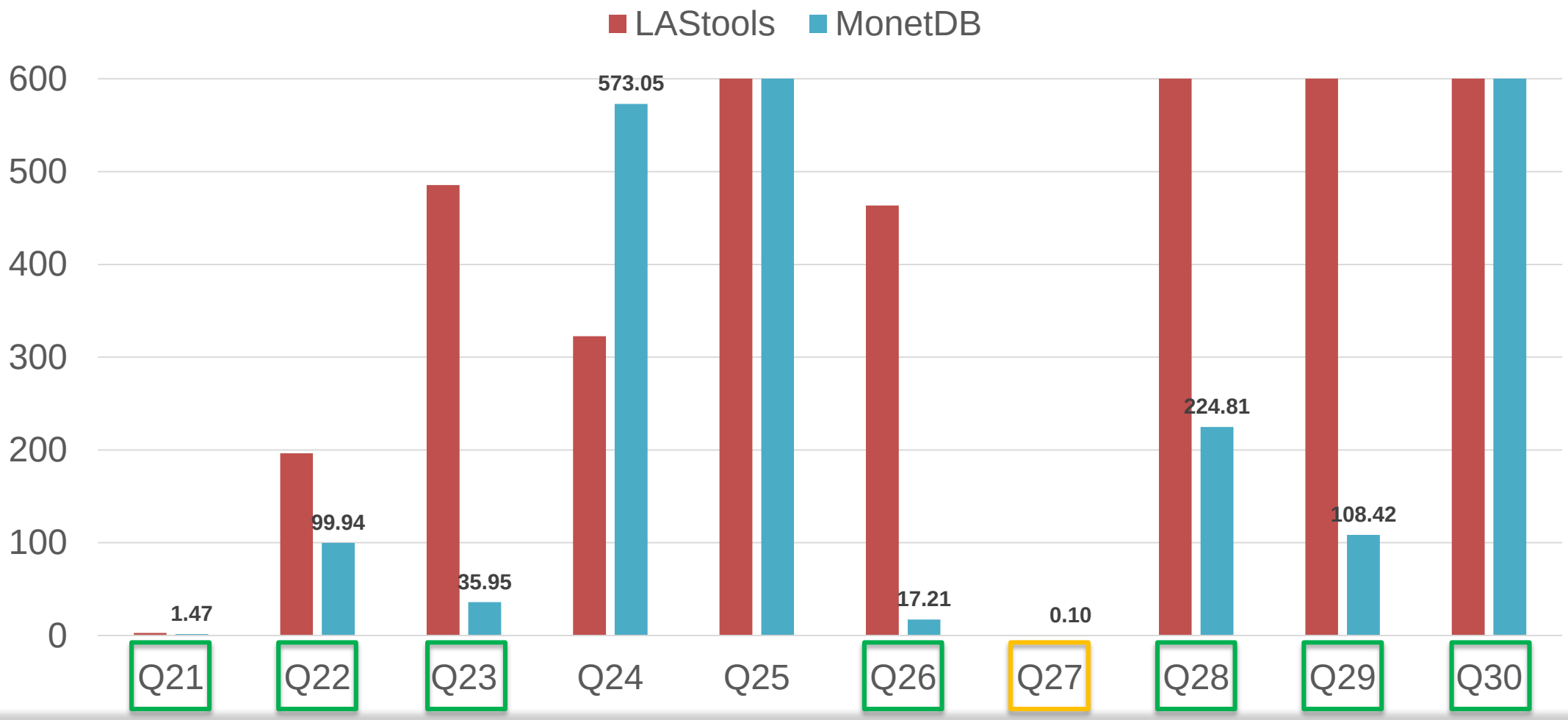
Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)



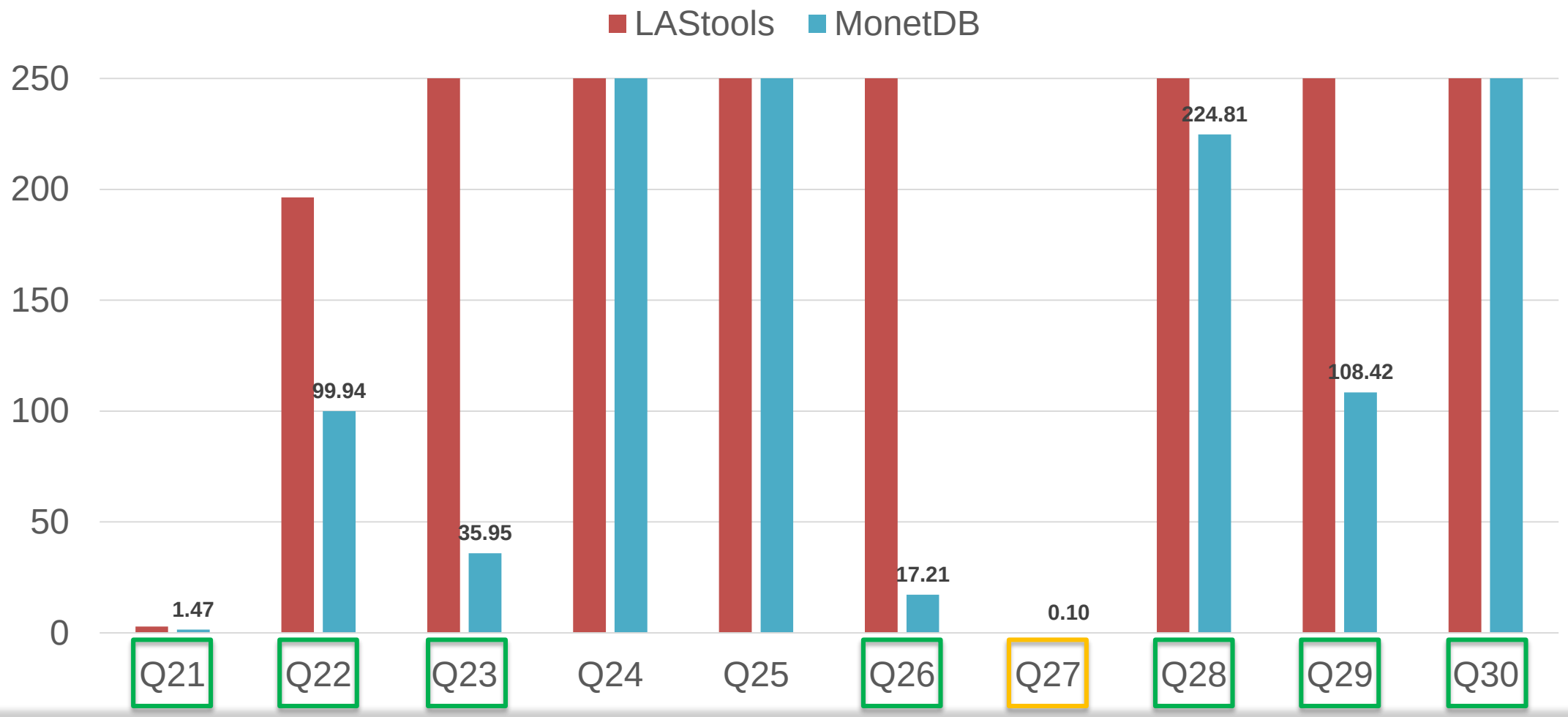
Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)



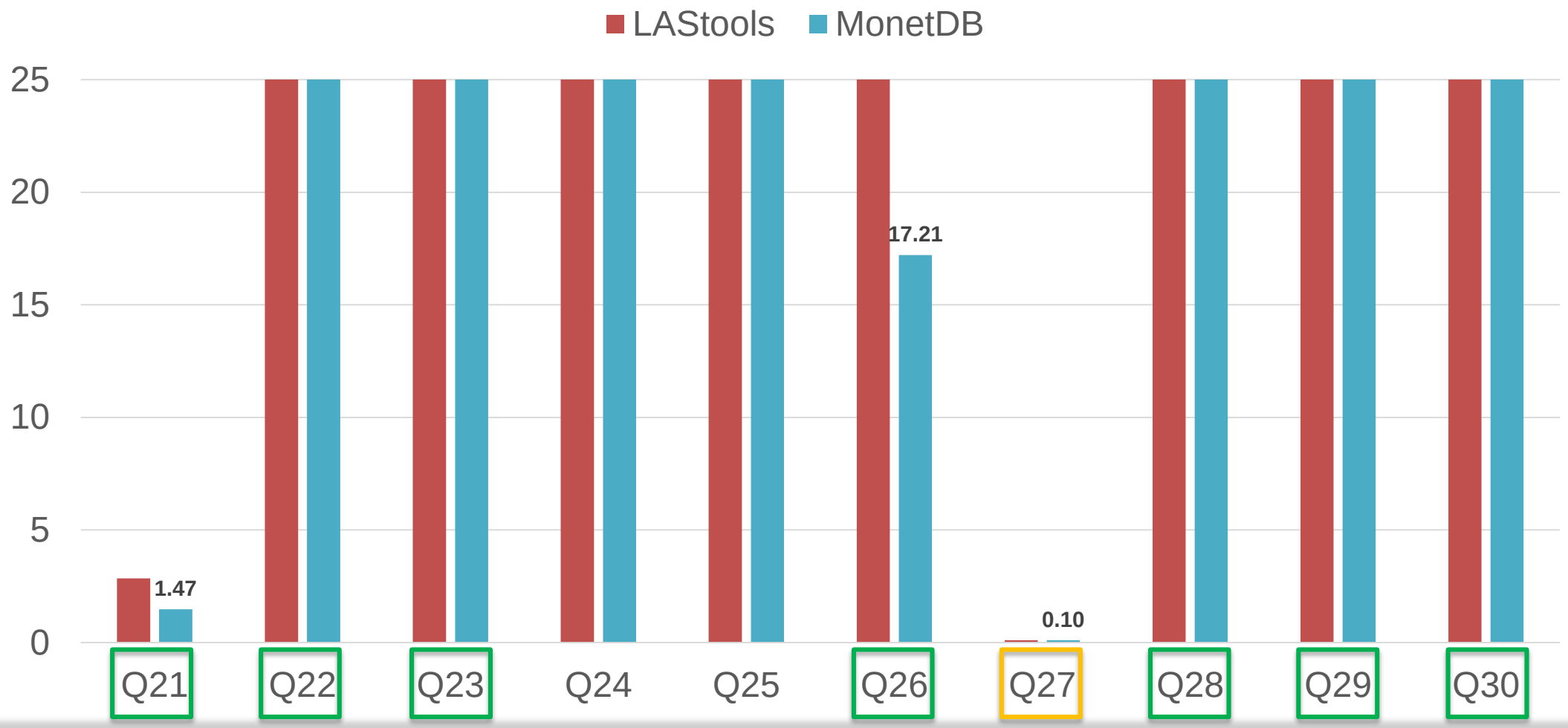
Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)



Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)

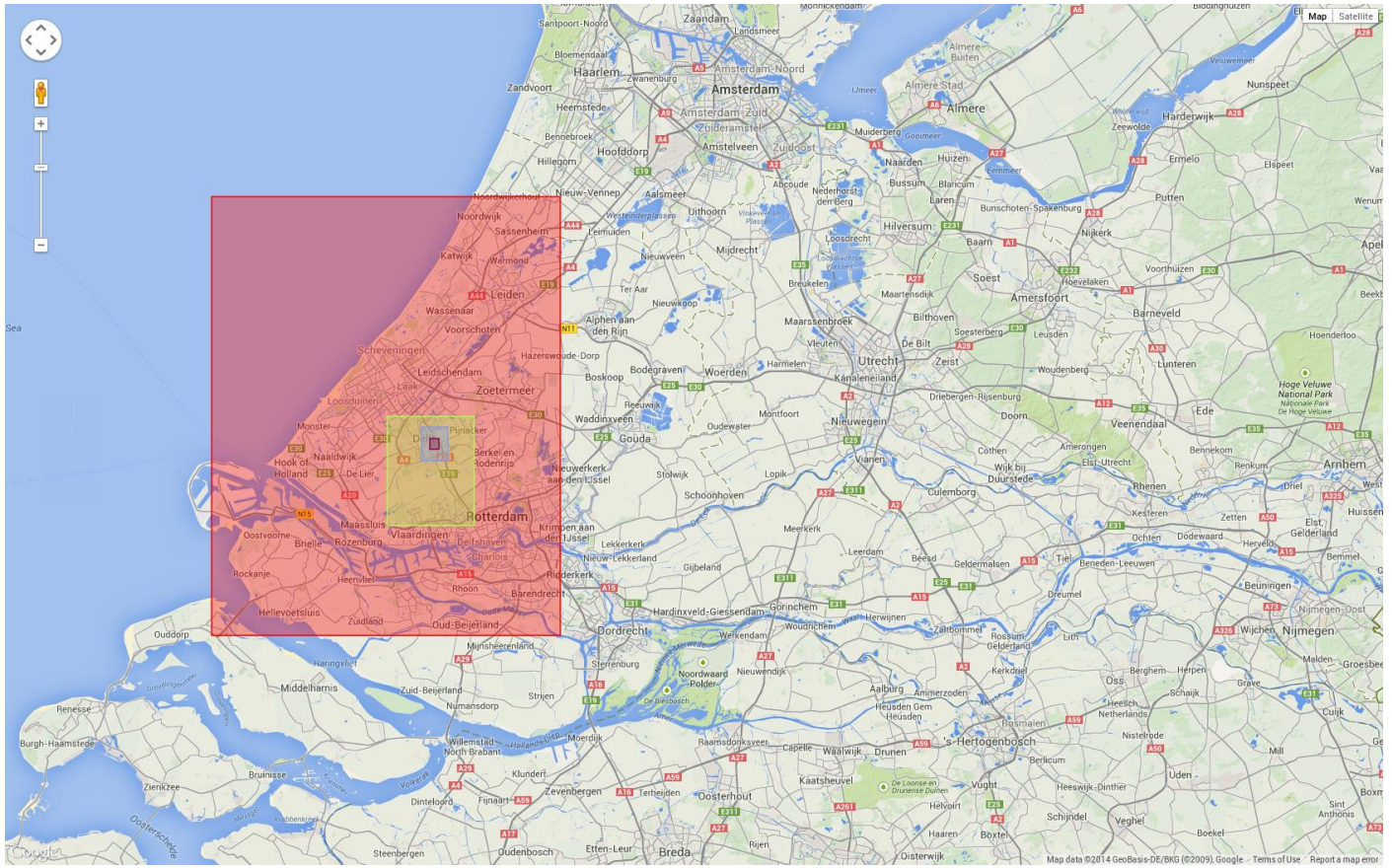


Query Processing (full-ahn2)
LAStools and MonetDB (20 000 part)

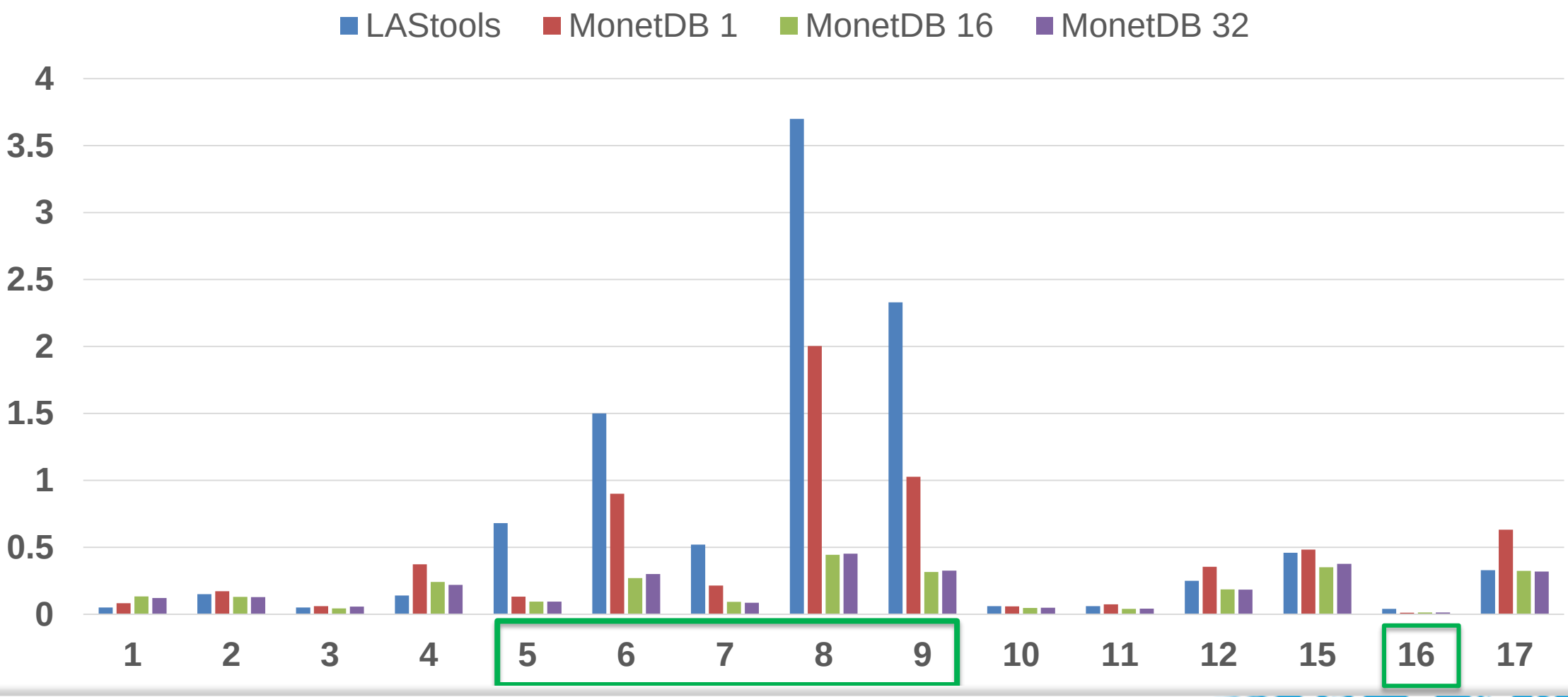


To understand vertical scalability

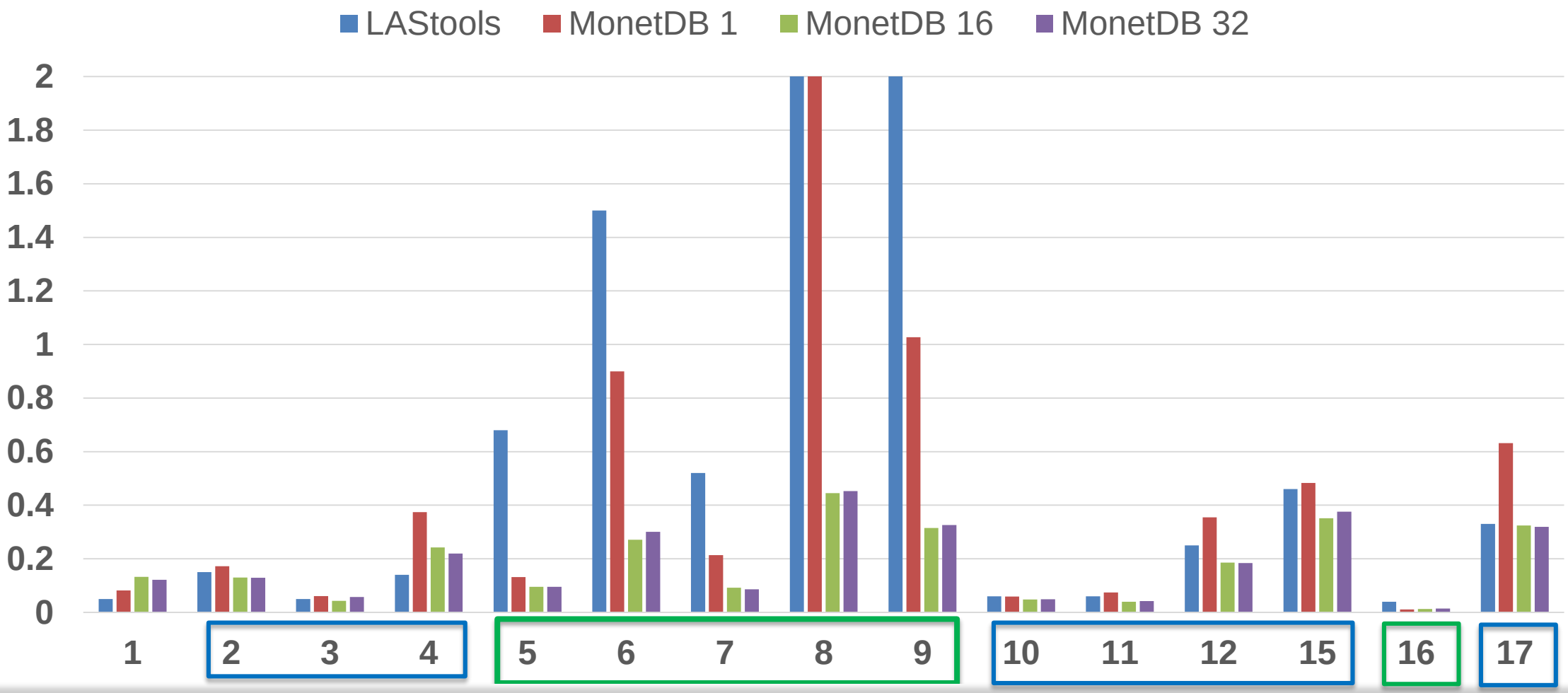
- **Sub-set composed by 23B**
  - **1492 files**
  - **1492 partitions**
  - **For processing**
    - **1, 16 and 32 threads**



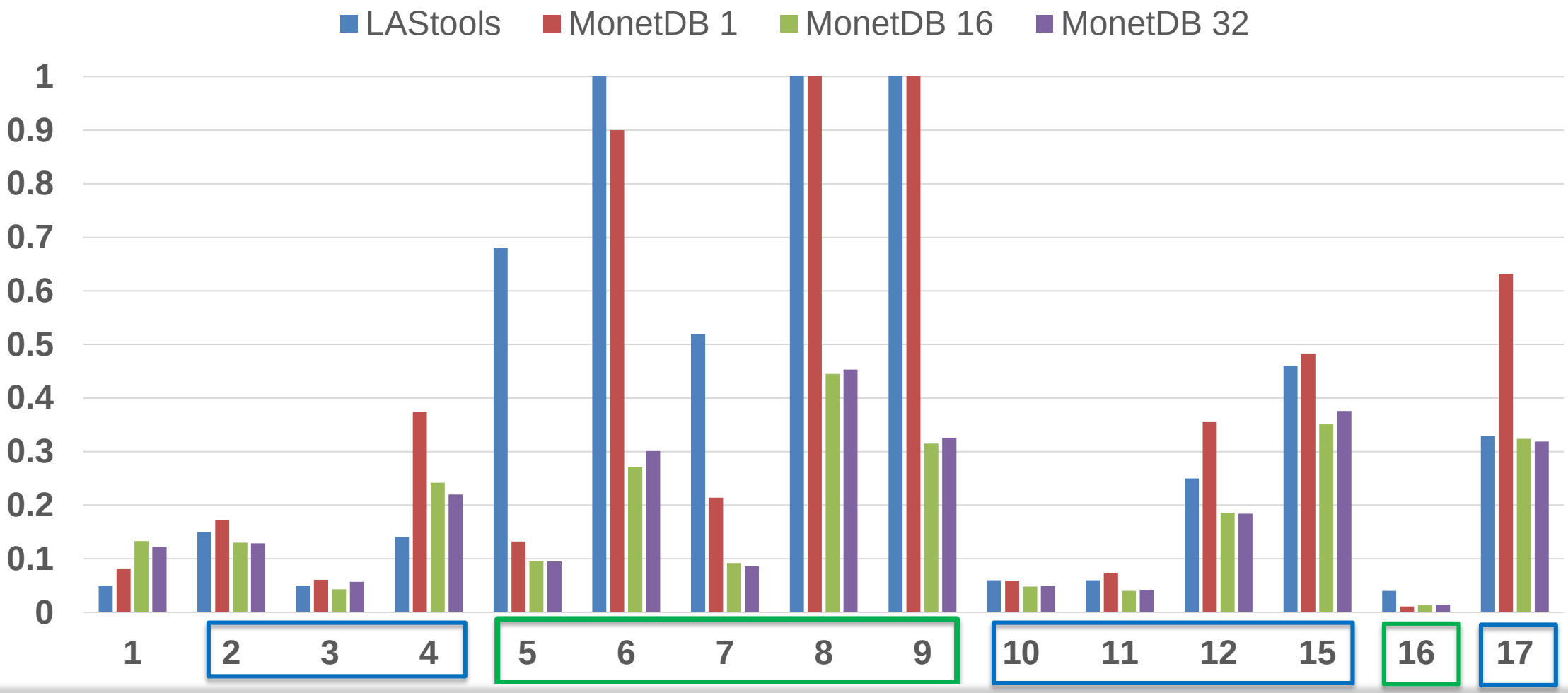
Query Processing (AHN2 - 23B)
LAStools and MonetDB



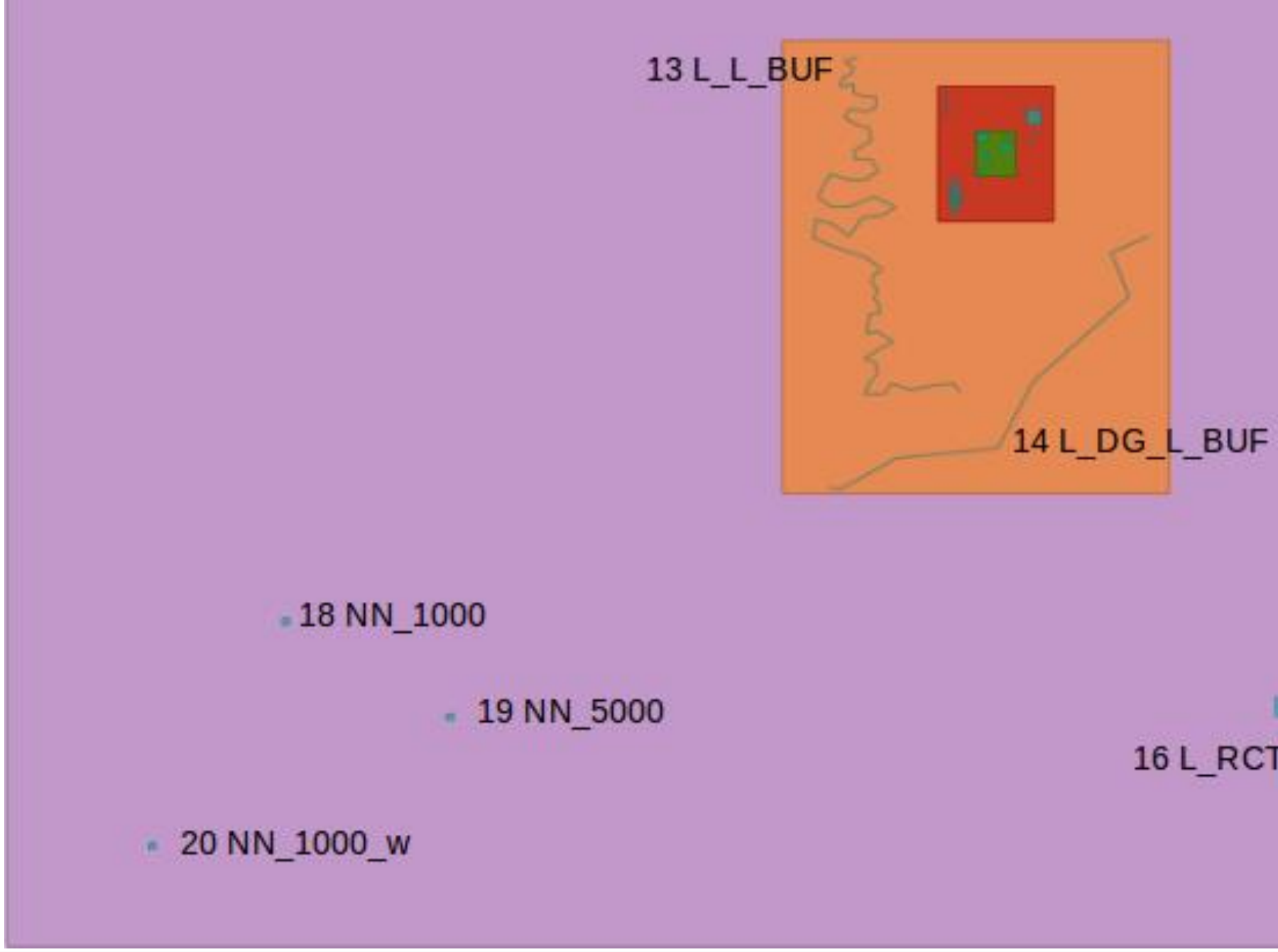
Query Processing (AHN2 - 23B)
LAStools and MonetDB



Query Processing (AHN2 - 23B)
LAStools and MonetDB



Queries…



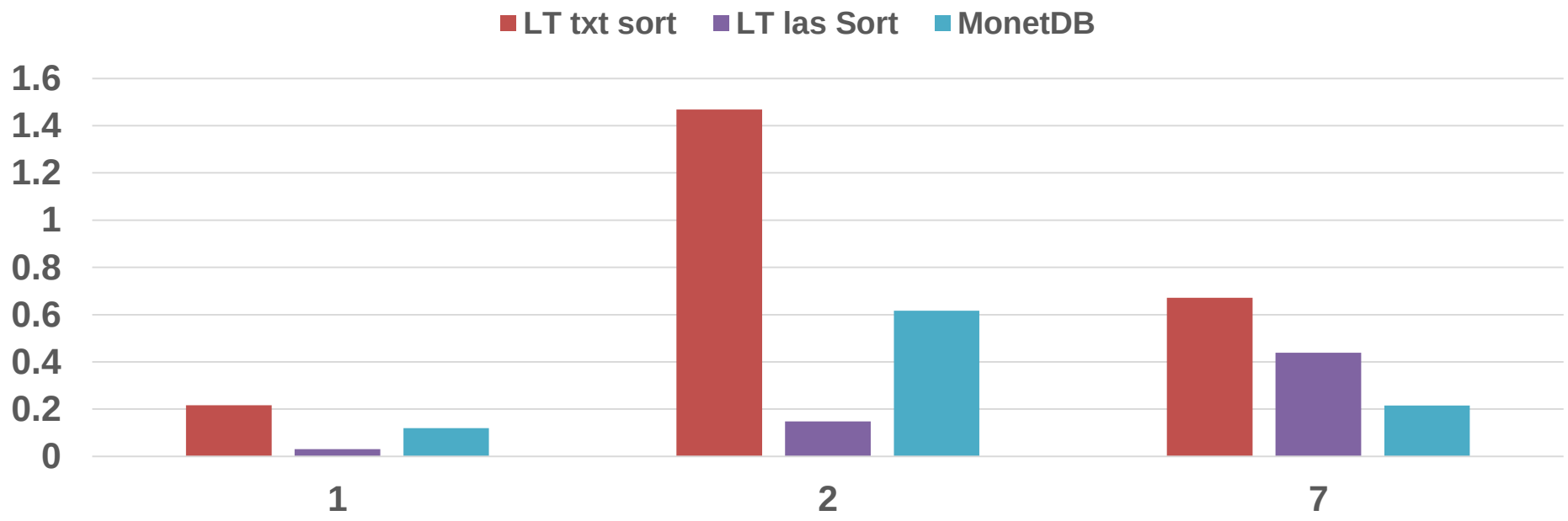
Query Processing (AHN2 - 23B)
LAStools and MonetDB

Query	LAStools	MonetDB 1	MonetDB 16	MonetDB 32
13	394.03	259.09	28.156	23.681
14	84.86	128.518	23.11	20.97

Query Processing (AHN2 - 23B)

Data output

Query	LT txt sort	LT las Sort	MonetDB	Points
1	0.216	0.03	0.119	74885
2	1.468	0.148	0.617	718132
7	0.671	0.439	0.215	45815



Ongoing work to be released…

- **GPU operators**

- **Operate over compressed data**
  - **Internal compressed format**
  - **LAS/LAZ**

Ongoing work to be released…

- **GPU operators**
- **Operate over compressed data**
  - **Internal compressed format**
  - **LAS/LAZ**
- **Output compressed data**
- **Scale out to improve throughput on spatial selections**
- **Compare with PDAL approach**

In a nutshell…

- **Column-stores are up to the task**
  - **Vertical partitioning for free**
    - **Late materialization**
    - **Indexing and filtering points when mixed with different attributes/dimensions [1]**
    - **Integration with other data sets**
  - **Horizontal partitioning**
  - **In-memory indexes**
  - **Vector processing**
  - **Column compression**
  - **Lazy data loading**

[1] László Dobos, István Csabai, János M. Szalai-Gindl, Tamás Budavári, and Alexander S. Szalay. 2014. **Point cloud databases**, SSDBM '14.

A DataScope for Geo-Spatial data

- Combine the best of both (open-source) technologies

A DataScope for Geo-Spatial data

- Combine the best of both (open-source) technologies
  - Data access
    - Efficiently combine heterogeneous data sets
    - In-situ data access

A DataScope for Geo-Spatial data

- Combine the best of both (open-source) technologies
  - Data access
    - Efficiently combine heterogeneous data sets
    - In-situ data access
  - Data processing
    - New functionality
    - Specialized libraries
    - External accelerators (HPC and HTC)

A DataScope for Geo-Spatial data

- Combine the best of both (open-source) technologies
  - Data access
    - Efficiently combine heterogeneous data sets
    - In-situ data access
  - Data processing
    - New functionality
    - Specialized libraries
    - External accelerators (HPC and HTC)
  - Data presentation
    - A set of frond-ends to explore the data sets in various ways
    - Service
      - News feed
      - Geoserver
    - Interactive 4D visualization

A DataScope for Geo-Spatial data

External data banks

A DataScope for Geo-Spatial data

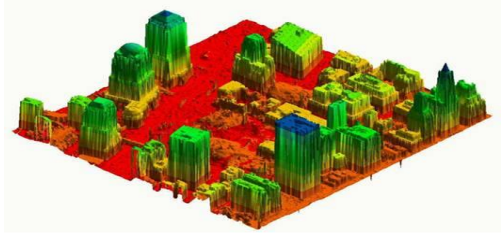
WEATHER UNDERGROUND

TOP10RASTER

kadaster

External data banks

AHN2 = Open Data!



Point Cloud Data

A DataScope for Geo-Spatial data

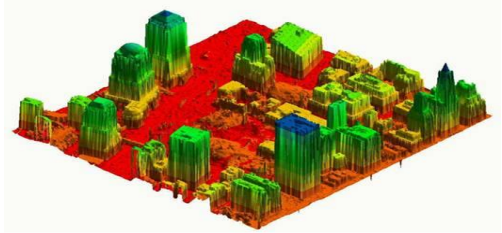
WEATHER UNDERGROUND

TOP10RASTER

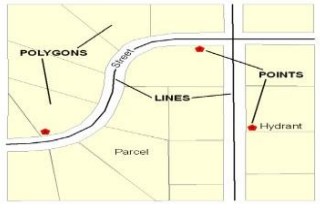
kadaster

External data banks

AHN2 = Open Data!



Point Cloud Data



Vector Data

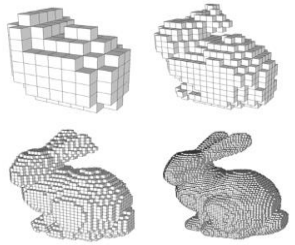
A DataScope for Geo-Spatial data

WEATHER UNDERGROUND

TOP10RASTER

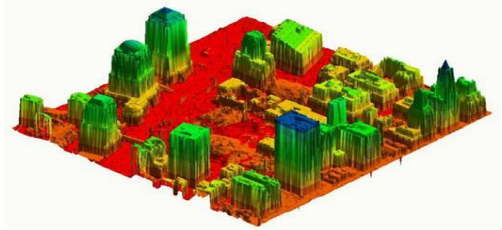
kadaster

External data banks

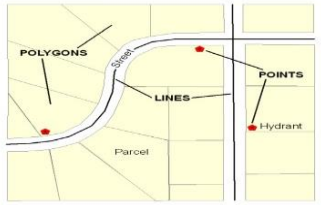


Voxel Data

AHN2 = Open Data!

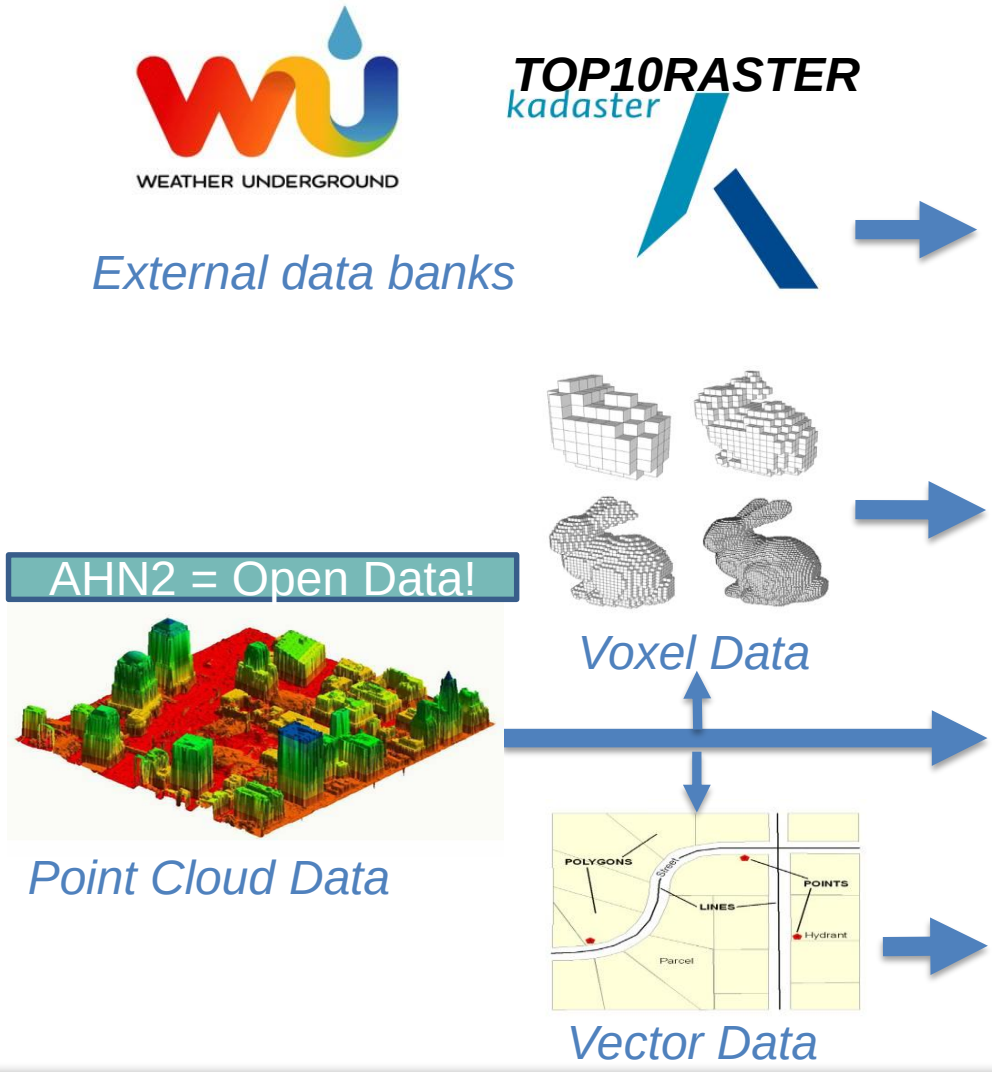


Point Cloud Data

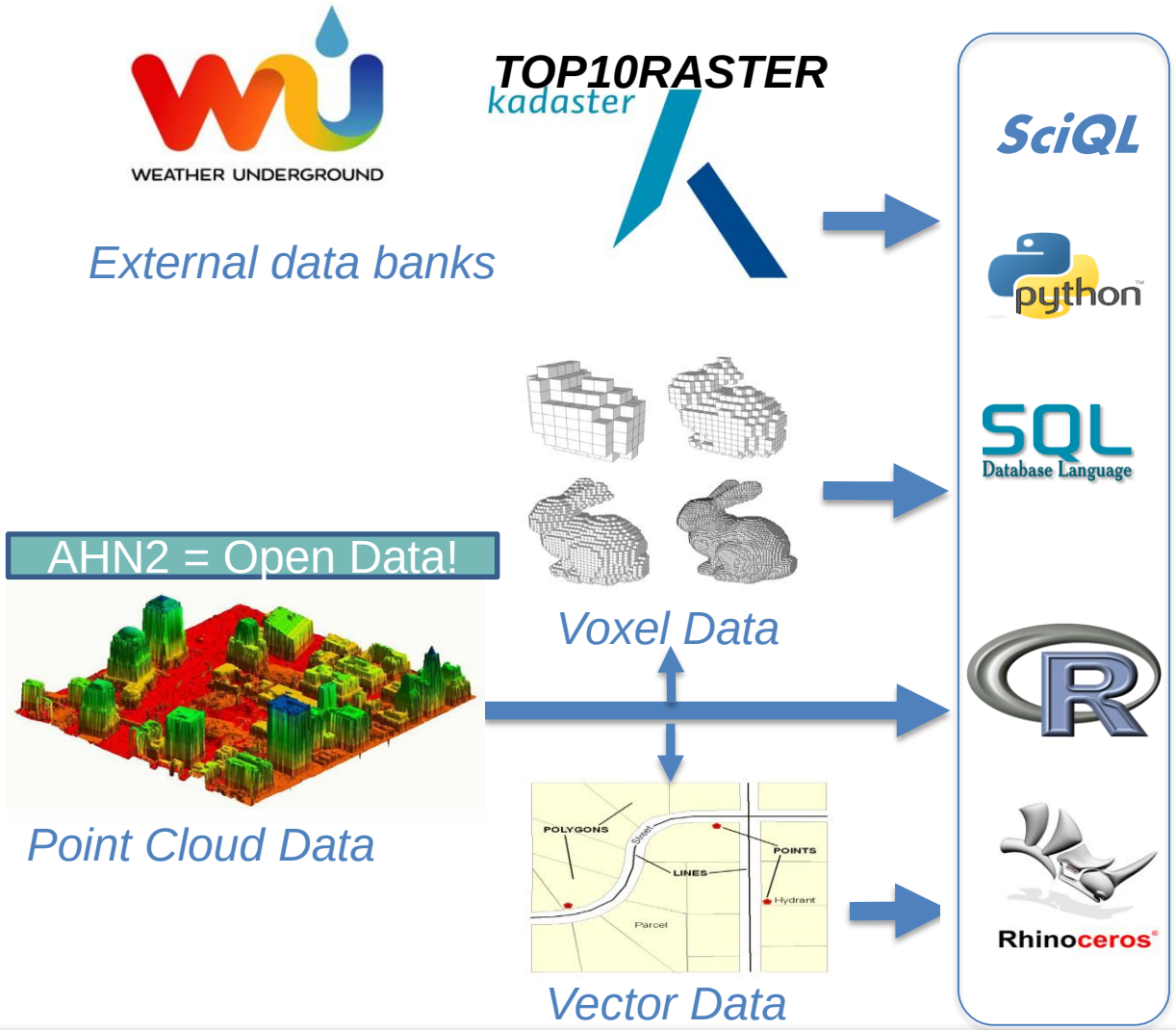


Vector Data

A DataScope for Geo-Spatial data



A DataScope for Geo-Spatial data



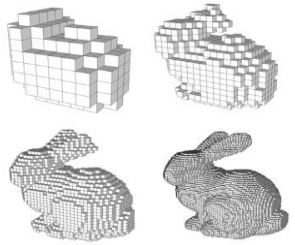
A DataScope for Geo-Spatial data

WEATHER UNDERGROUND

External data banks

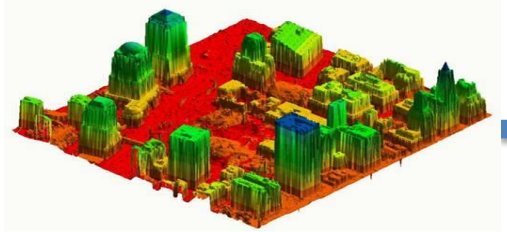
TOP10RASTER

kadaster



Voxel Data

AHN2 = Open Data!



Point Cloud Data

Vector Data

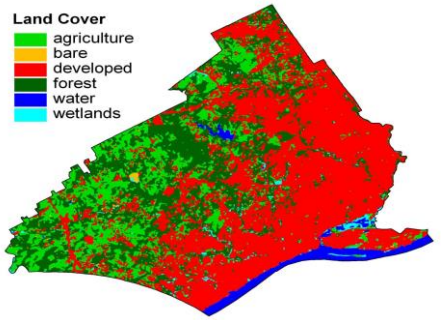
SciQL

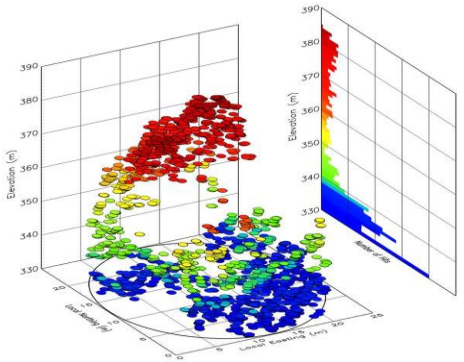
python

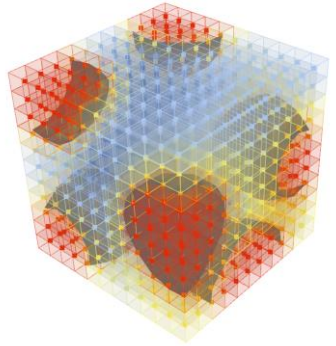
SQL
Database Language

R

Rhinoceros





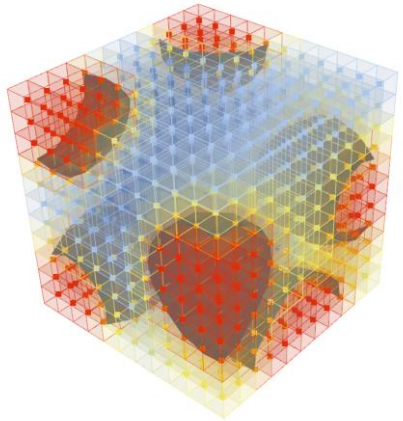


Exploration

- **DataScope allows:**

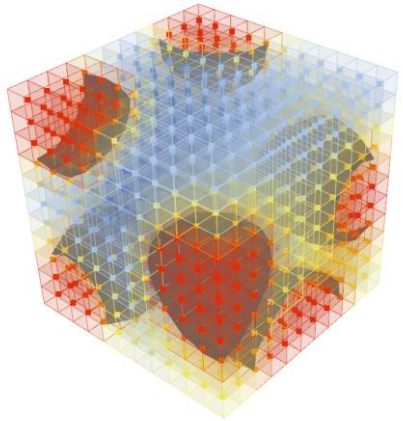
Exploration

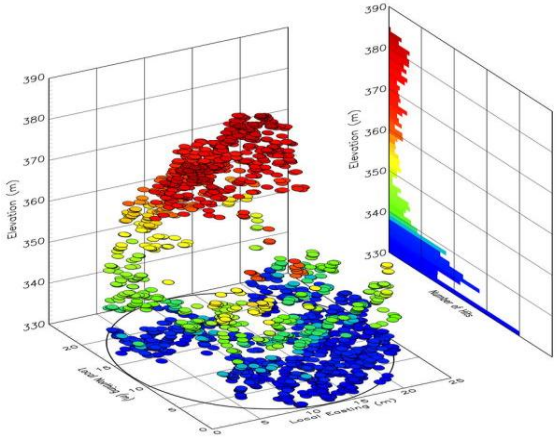
- **DataScope allows:**
  - **3D analyses such as 3D intersections**



Exploration

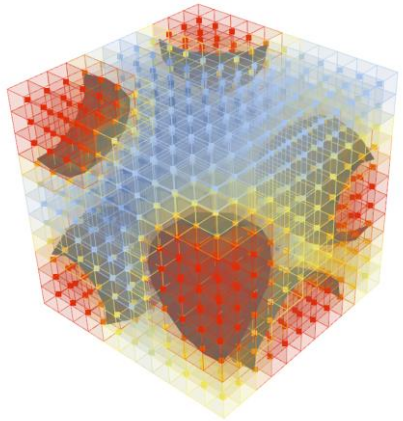
- **DataScope allows:**
  - **3D analyses such as 3D intersections**
  - **Collection of statistics**

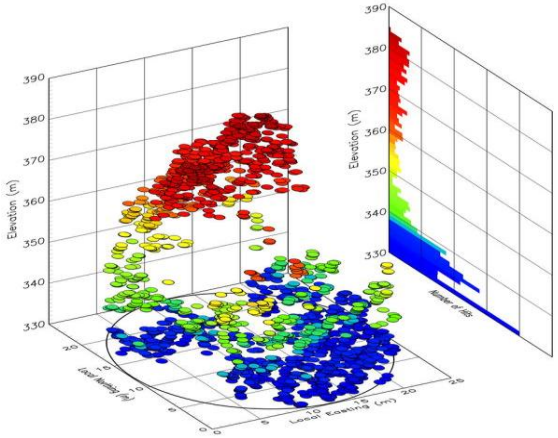


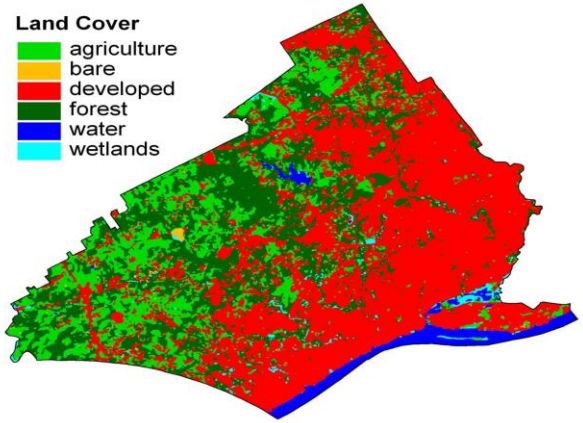


Exploration

- **DataScope allows:**
  - **3D analyses such as 3D intersections**
  - **Collection of statistics**
  - **On-demand retrieval of city areas for interactive exploration**







New knowledge

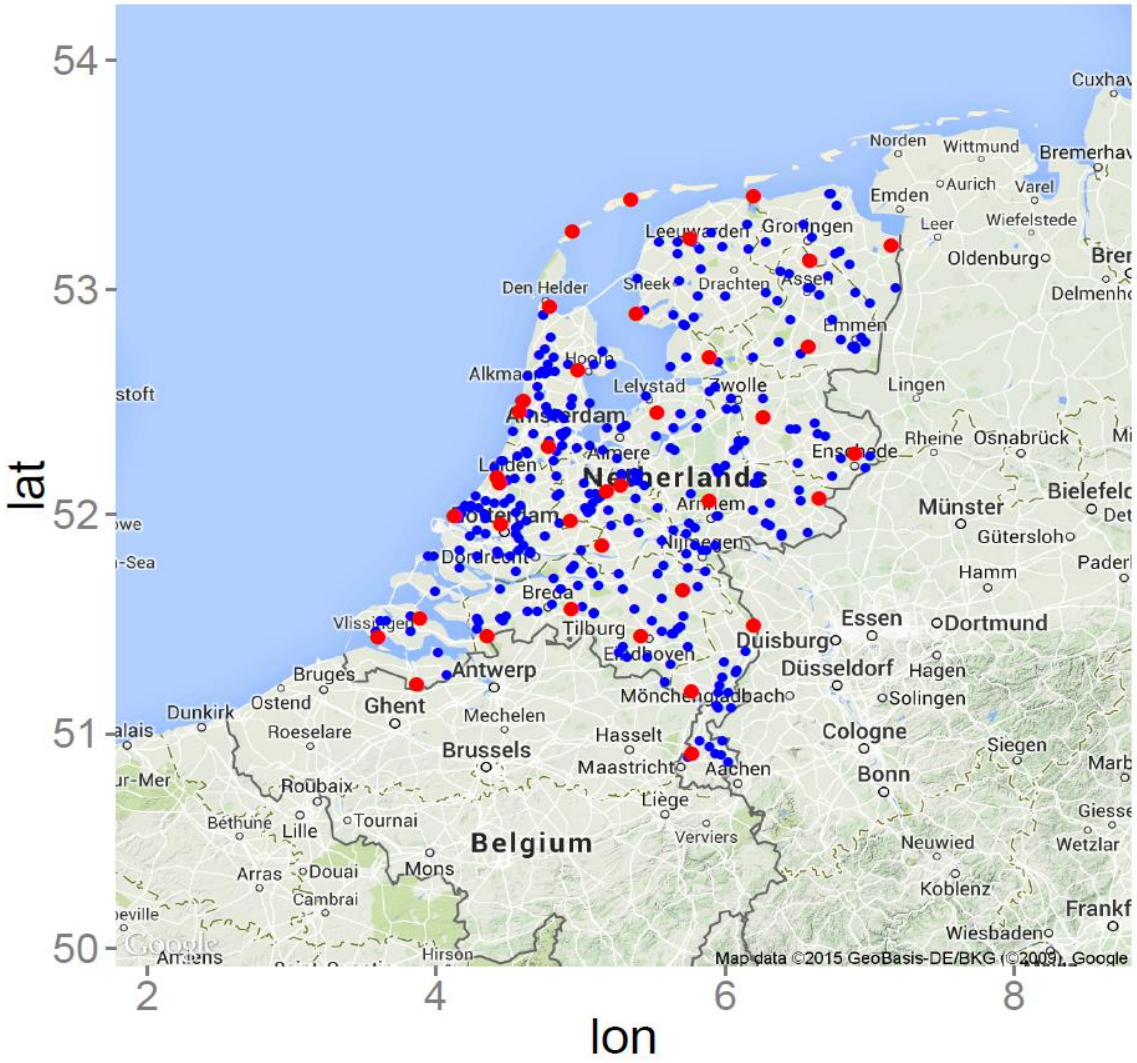
- **Urban planning to develop smart cities**
  - **To increase wealth**
  - **Improve human comfort**
    - **Such as avoiding heat islands**

  - **A Round Table for Multi-disciplinary Research on Geospatial and Climate Data**

- [30] L. van Hove and et al. Temporal and spatial variability of urban heat island and thermal comfort within the Rotterdam agglomeration. *Building and Environment*, 2015.

New knowledge

- **Urban planning to develop smart cities**
  - **To increase wealth**
  - **Improve human comfort**
    - **Such as avoiding heat islands**

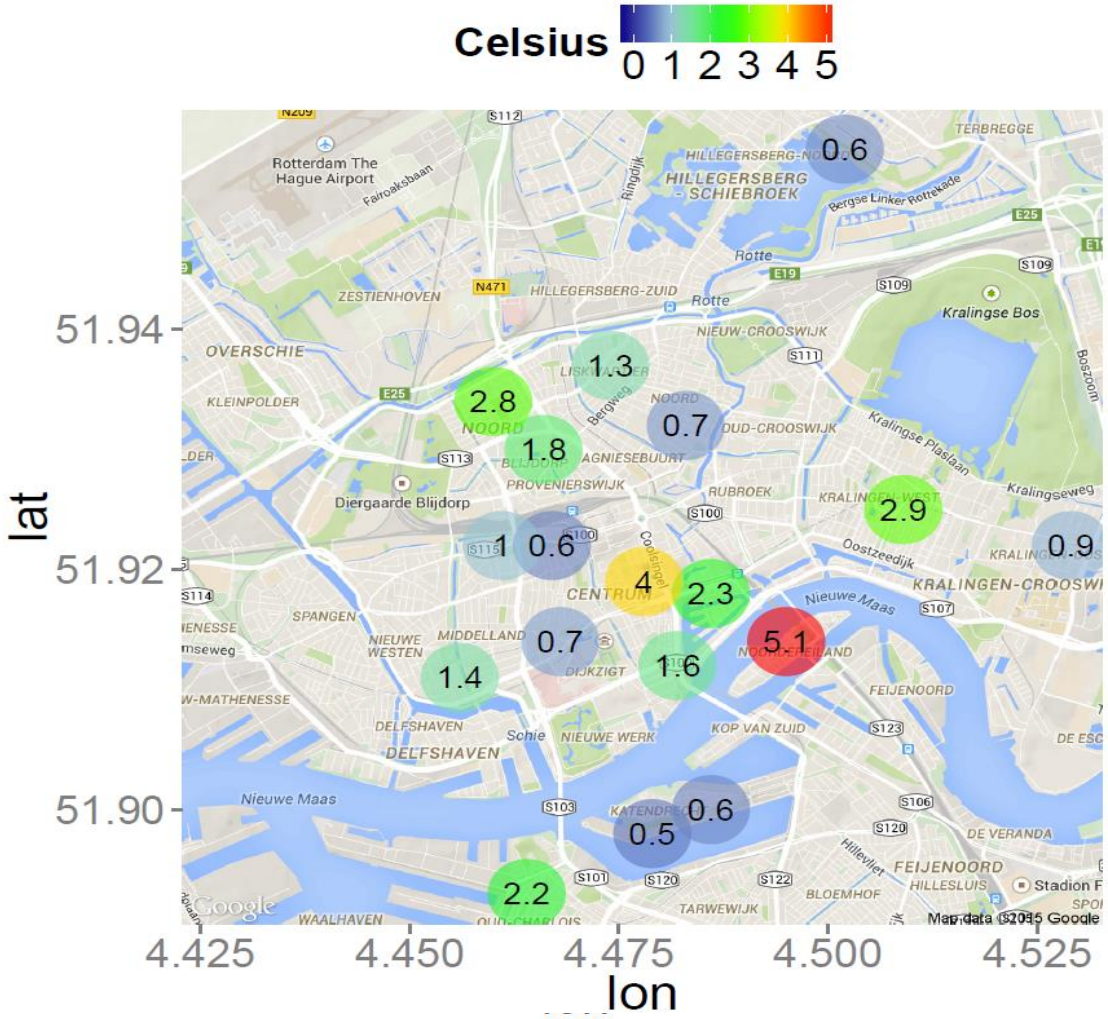


  - **A Round Table for Multi-disciplinary Research on Geospatial and Climate Data**

- [30] L. van Hove and et al. Temporal and spatial variability of urban heat island and thermal comfort within the Rotterdam agglomeration. *Building and Environment*, 2015.

New knowledge

- **Urban planning to develop smart cities**
  - **To increase wealth**
  - **Improve human comfort**
    - **Such as avoiding heat islands**

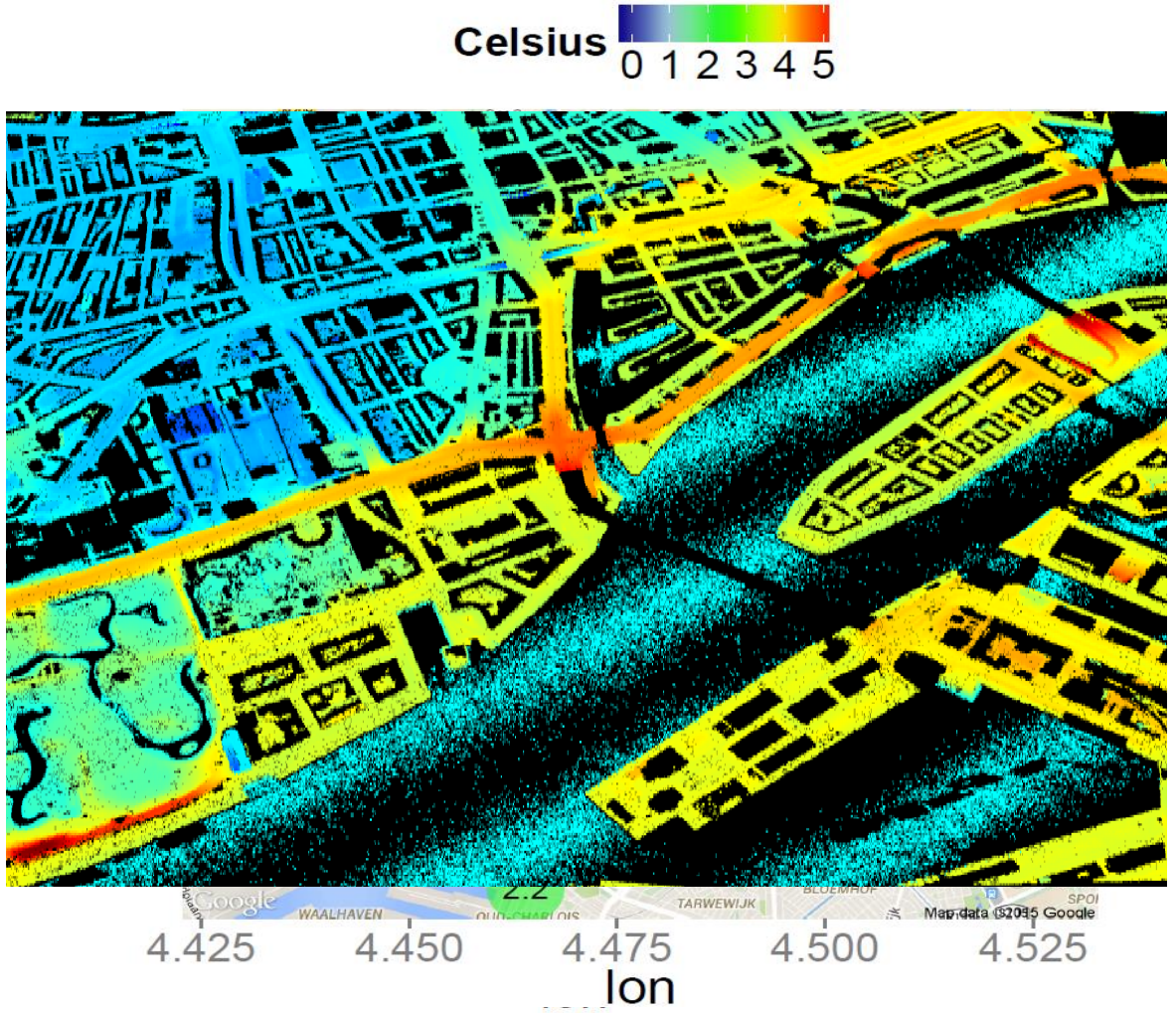


  - **A Round Table for Multi-disciplinary Research on Geospatial and Climate Data**

- [30] L. van Hove and et al. Temporal and spatial variability of urban heat island and thermal comfort within the Rotterdam agglomeration. *Building and Environment*, 2015.

New knowledge

- **Urban planning to develop smart cities**
  - **To increase wealth**
  - **Improve human comfort**
    - **Such as avoiding heat islands**



  - **A Round Table for Multi-disciplinary Research on Geospatial and Climate Data**

- [30] L. van Hove and et al. Temporal and spatial variability of urban heat island and thermal comfort within the Rotterdam agglomeration. *Building and Environment*, 2015.

Questions?

Image from: http://aboutinterviews.com/5-questions-to-not-ask-in-a-job-interview/