Adaptive Indexing for In-situ Visual Exploration & Analytics

Stavros Maroulis [1,2] Nikos Bikakis [2] George Papastefanatos [2]
Panos Vassiliadis [3] Yannis Vassiliou[1]

rawVis

[1] Nat. Tech. Univ. of Athens, Greece
[2] IMSI Institute, ATHENA R.C., Greece
[3] University of Ioannina, Greece

Intro

Common challenges in data exploration

- Large datasets that do not fit in main memory
- Users with limited skills in data management & processing
- Limited hardware resources (e.g., no access to a distributed environment)
- Traditional DBMS require full loading & indexing → long data-to-query time

In-situ Data Exploration

On-the-fly exploration & analysis of big raw data files e.g., csv, json

- Avoid full loading & indexing
- Progressive loading & indexing

- Recent works have focused on generic in-situ querying (mainly range queries)

- In this work
  - We study categorical-based operations in in-situ scenarios
    - Group-by Operations: essential for most-known visualization types
    - Categorical Filters: effective exploration (e.g., faceted exploration)

Contributions

- Formulation of exploratory & analytical operations over categorical attributes as data-access operations
- **CET:** a main-memory lightweight tree structure
  - organizes objects & computes statistics based on categorical attributes
- **VETI:** a hybrid index that combines tile & tree structures
  - supports in situ 2D exploration & analytics over categorical, numeric, spatial attributes
- Experimental evaluation using real & synthetic datasets

Conclusions

our technique outperforms competitors both in execution time (~40× faster) & I/O's (~3 orders of magnitude)

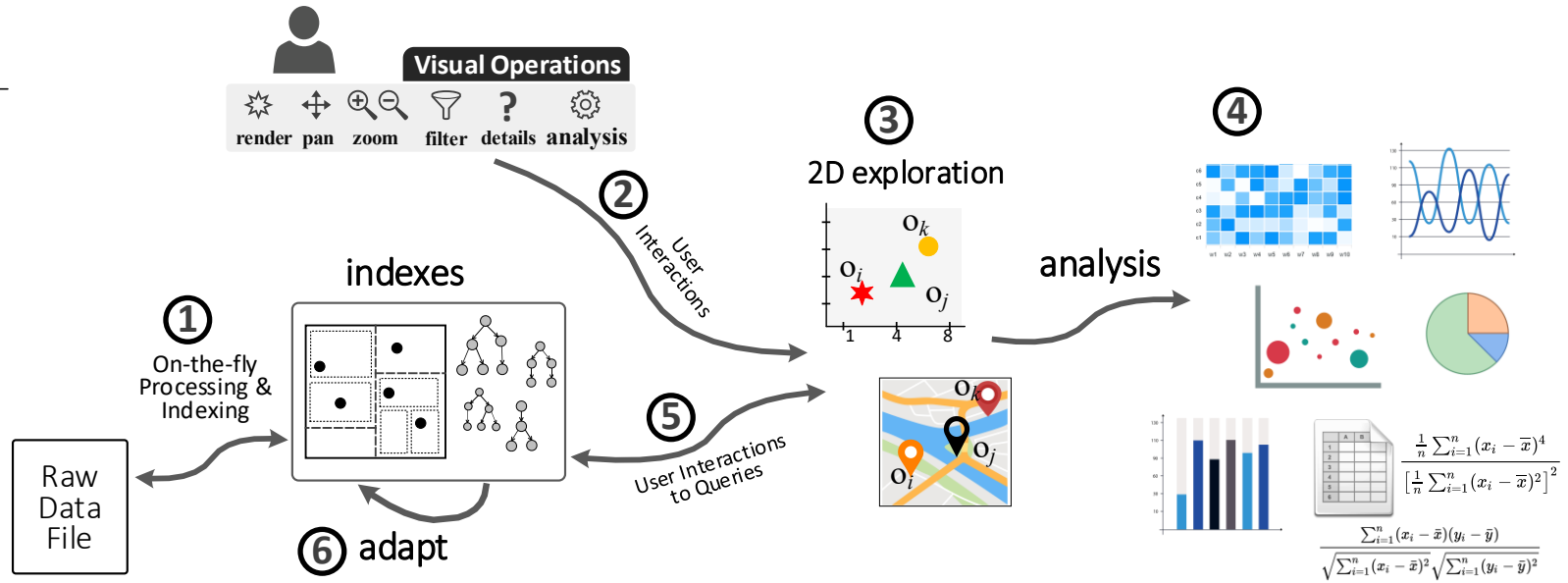
Working Scenario

Objects	Lat	Long	Signal	Width	Brand	Provider	Net
	Attributes						
o_1	21	11	3	7	Samsg	Veriz	3G
o_2	29	18	1	4	Samsg	Veriz	4G
o_3	11	1	7	6	Xiaomi	AT&T	4G
o_4	19	7	2	3	Huawei	AT&T	5G
o_5	23	12	4	8	Huawei	Veriz	5G

(a) **Raw Data File Sample**

A_{brand} = {Apple, Huawei, Samsg, Xiaomi}
$A_{provider}$ = {AT&T, Veriz}
A_{net} = {3G, 4G, 5G}

(b) **Categorical Attributes Domains**



(c) **Working Scenario**

Exploratory Query

User operations → **exploratory queries** (data-access operations over index)

Exploratory query

- Selection clause
  - 2D range query over X and Y attributes
- Filter clause
  - conditions over the non-axis attributes
- Details clause
  - non-axis attributes to retrieve
- Group-by clause
  - attributes based on which to group results
- Analysis clause
  - aggregate functions

Categorical Exploration Tree (CET) Overview

- Lightweight, memory-oriented, trie-like tree structure
- Level-based organization
  - Each tree level corresponds to a different categorical attribute
- Based on the tree hierarchy, each node is associated with a set of objects based on the node path

Categorical Exploration Tree (CET)
Leaf nodes

Object Entries: $\langle a_{i,x'}\ a_{i,y'}\ f_i \rangle$

- $a_{i,x'}\ a_{i,y}$ being the values of the axis attributes
- f_i the offset (a hex value) of o_i in the raw file

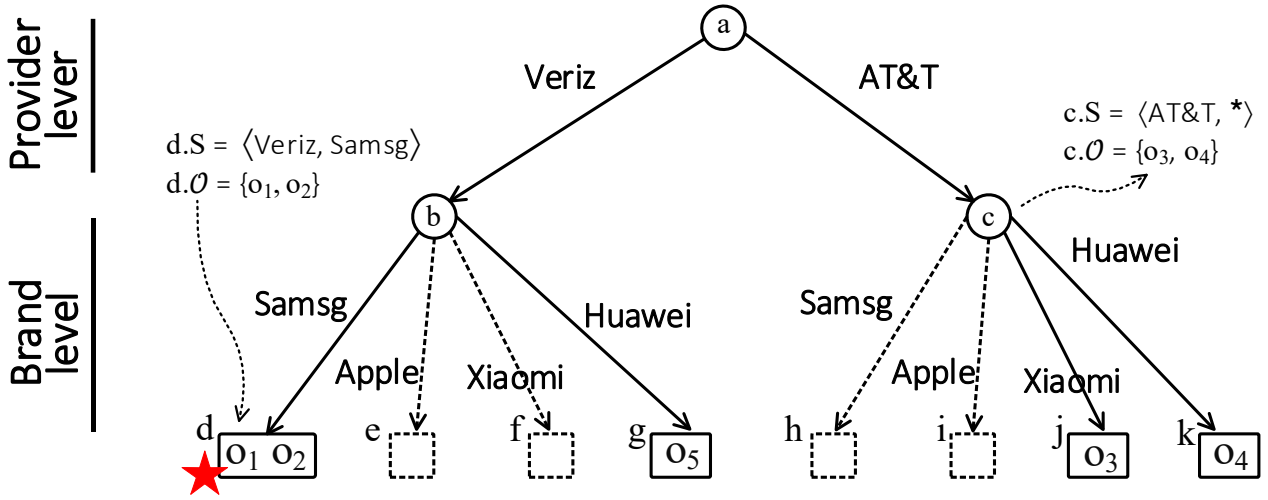
Synopsis Metadata

- algebraic aggregate functions over one or more non-axis numeric attributes

  e.g., sum, mean, sum of squares of deltas, etc.

Categorical Exploration Tree Example

Objects	Lat	Long	Signal	Width	Brand	Provider	Net
	Attributes						
o_1	21	11	3	7	Samsg	Veriz	3G
o_2	29	18	1	4	Samsg	Veriz	4G
o_3	11	1	7	6	Xiaomi	AT&T	4G
o_4	19	7	2	3	Huawei	AT&T	5G
o_5	23	12	4	8	Huawei	Veriz	5G



(a) **CET Tree**

★ **Leaf** d

object entries d.$\mathcal{E}$

	Lat	Long	File off.
o_1 :	⟨21	11	f_1⟩
o_2 :	⟨29	18	f_2⟩

metadata d.$\mathcal{M}$

Signal	Width
min(Signal)=1	max(Width)=7
∑Signal=4	∑Width=11
$\sum Signal^2$ =10	$\sum Width^2$=65

#Obj	Signal & Width
n = 2	∑Signal ∗ Width=25

(b) **Contents of Leaf** d

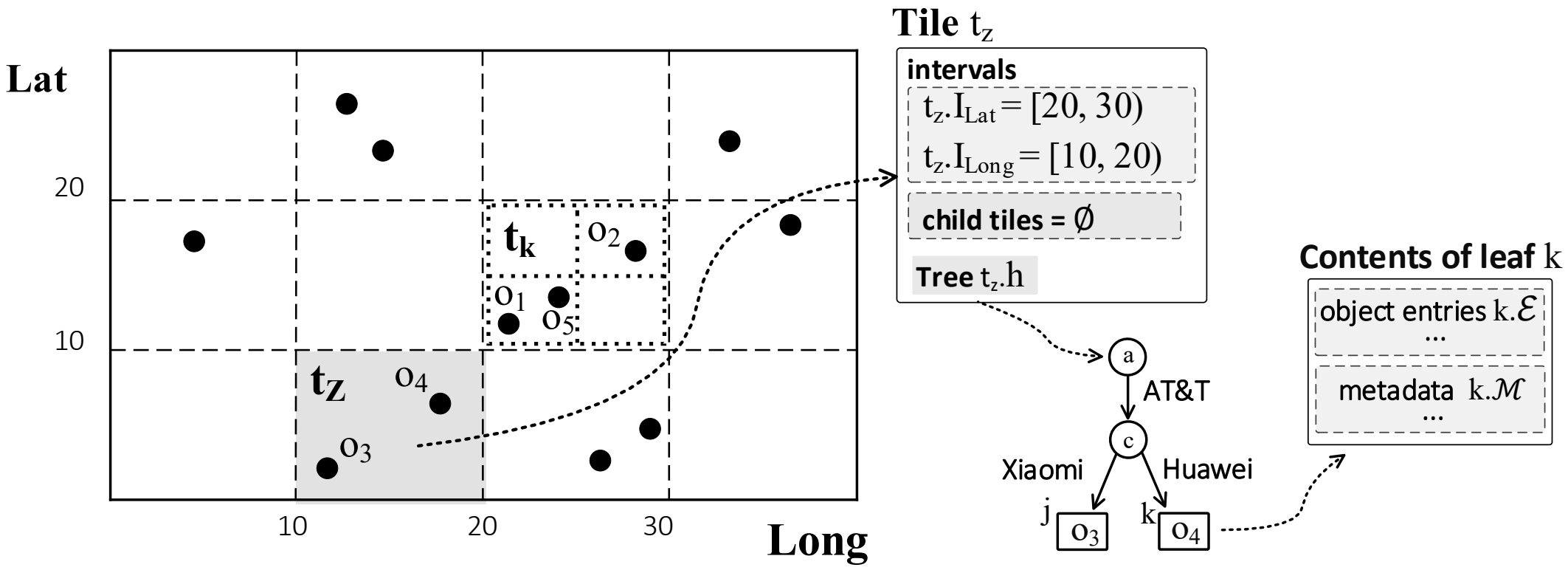
VALINOR Index

- In-memory tile-based multilevel index
- Raw file data objects are organized into hierarchy of **tiles**
- In each level of the hierarchy, all tiles are disjoint & can belong to only one parent tile
- Constructed on-the-fly
- Incrementally adjusted based on user interactions
- User operations may split a tile into more fine-grained ones

More details: Bikakis N. et al. In situ Visual Exploration over Big Raw Data Information Systems, 95, 2021

VETI Index

- Combines the **VALINOR** tile-based index with the **CET** tree
  - Supports categorical based operations & analytics
- Each leaf tile is associated with a CET tree
- Tile objects are stored in the leaf nodes of its CET tree

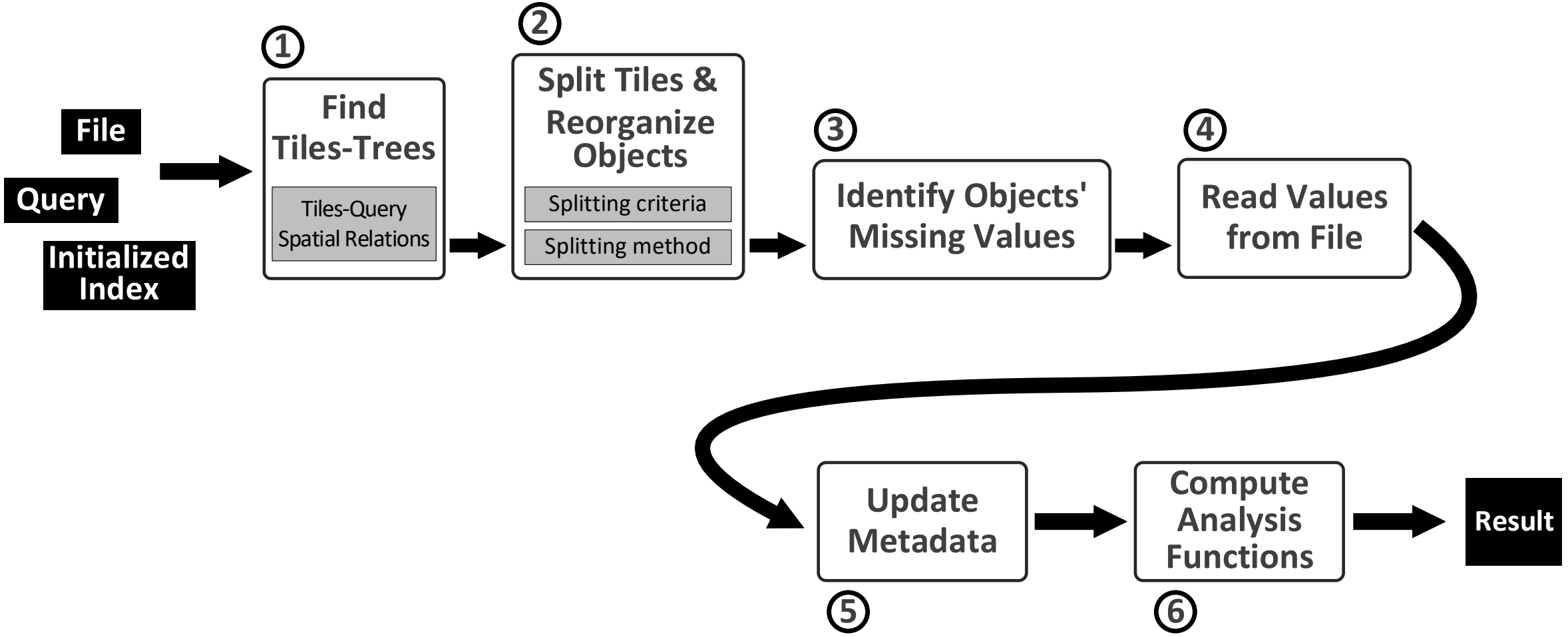


VETI Initialization

- Constructed on-the-fly based on the first user interaction
- Tile structure Initialization
  - Locality-based probabilistic initialization
    - More fine-grained near the initial query
    - Smaller tiles more likely to be fully overlapped by window query and avoid file accesses by utilizing metadata
- Insert objects (single file scan)

  For each object:
  - We find the tile it belong to based on it's X, Y values
  - We insert it into the tile's CET tree based on the categorical attributes

Query Processing and Incremental Adaptation



Experimental Analysis Setting

Both real and synthetic datasets:

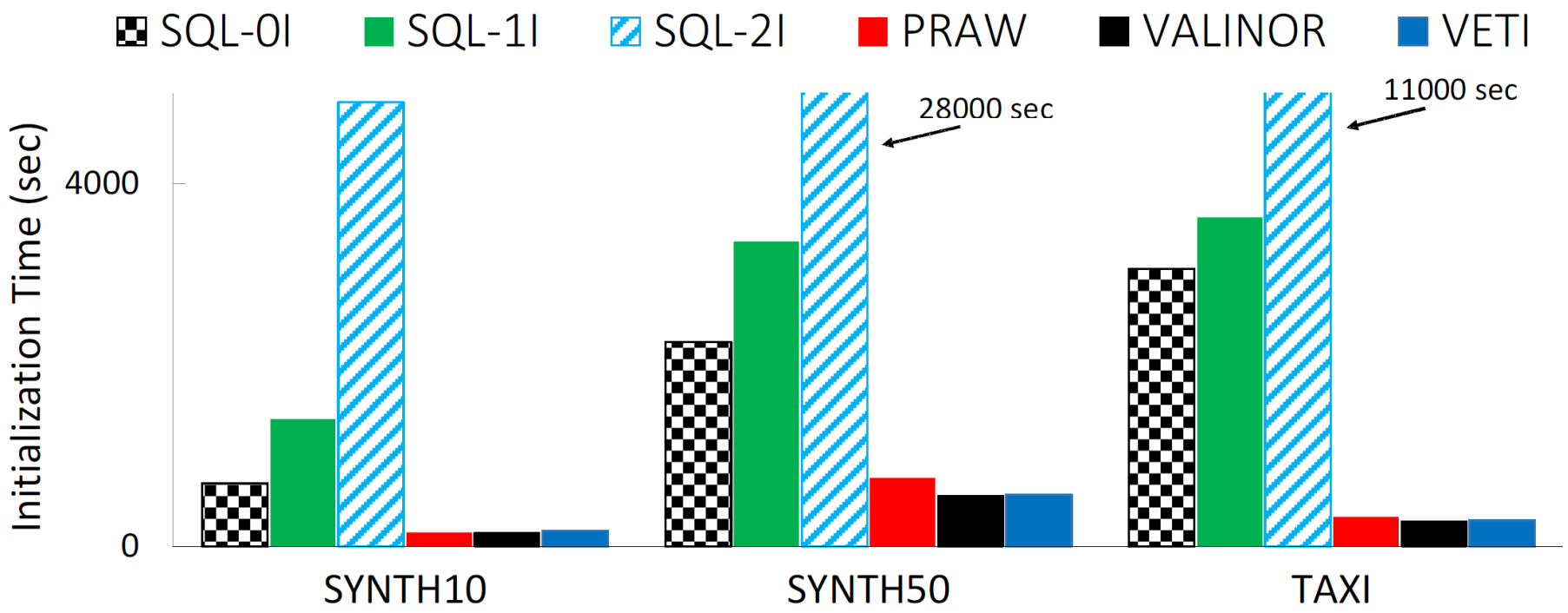
- NYC Yellow Taxi Trip Records (TAXI)
  165M objects, 18 attributes, 26 GB
- Synthetic CSV files: 100M objects - uniform distribution
  100M objects, 10 & 50 attributes (11 & 51 GB, respectively)

Competitors

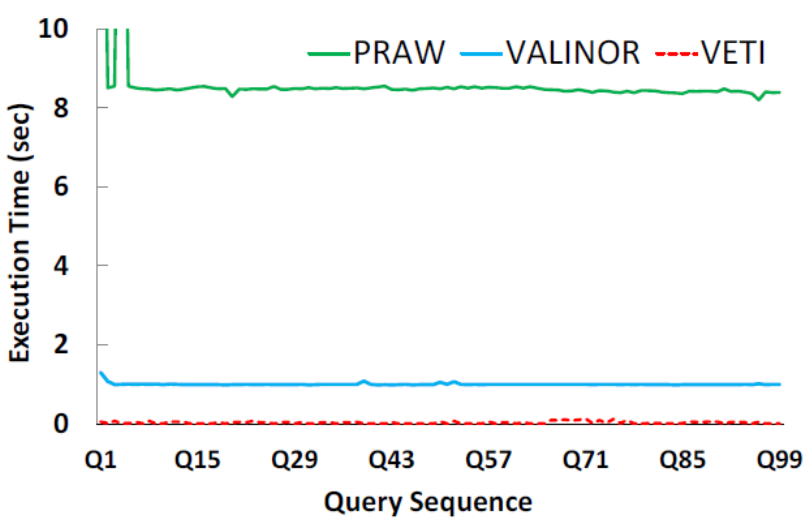
- VALINOR (tile-based index without the CET tree)
- MySQL
- PostgresRaw (platform for in situ querying over raw data)

Initialization Time

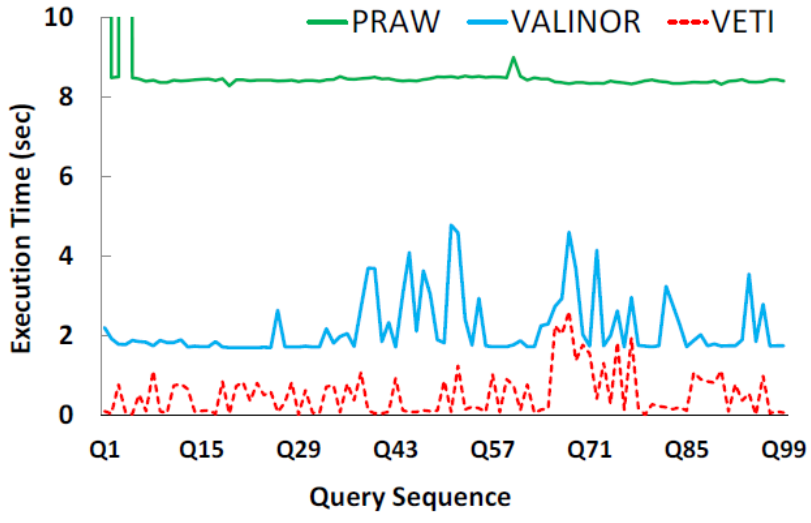
File Parsing, Index Construction & Q_0 Evaluation



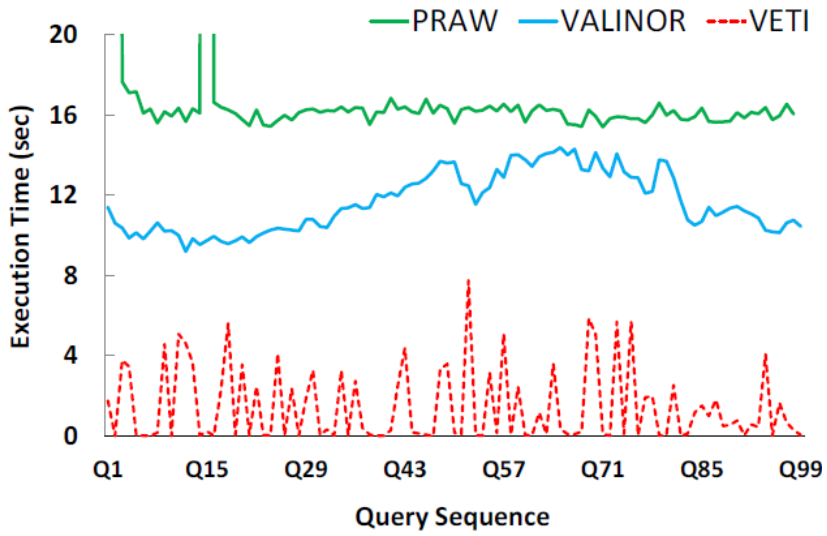
Exploration Scenario: Execution Time



(a) SYNTH10



(b) SYNTH50



(c) TAXI

Conclusions

VETI Index

- lightweight main memory index for in-situ 2D visual exploration & analysis of large raw data files
- Tile-based indexing + CET trees for supporting operations on categorical attributes
- Constructed on-the-fly & adapted based on user interaction

Experimental evaluation using real & synthetic datasets

our technique outperforms competitors both in execution time & I/O's

- **RawVis** System: Open source tool @ SIGMOD 2021

rawVis

Thank you!

https://visualfacts.imsi.athenarc.gr

H.F.R.I. Hellenic Foundation for Research & Innovation

This research is funded by the project VisualFacts (#1614) - 1st Call of the Hellenic Foundation for Research and Innovation Research Projects for the support of post-doctoral researchers