# Scalable Discovery of Unique Column Combinations

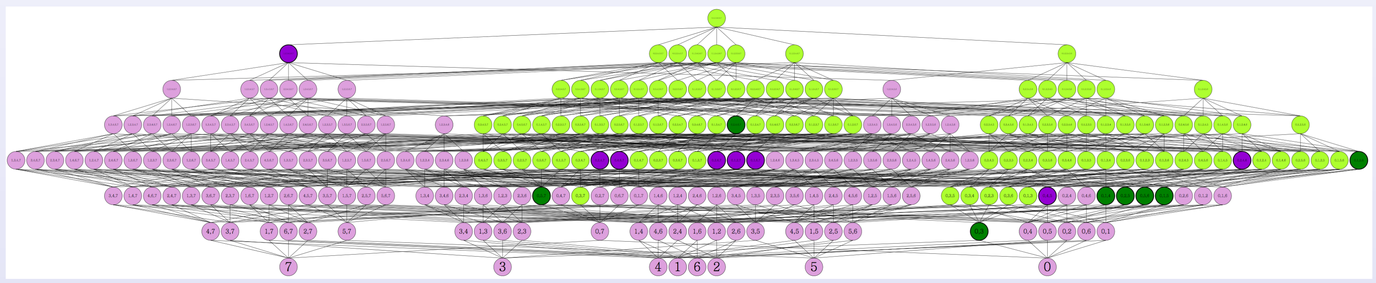
## Motivation

**Large datasets** at very **fast rates**:
- Social networks
- Scientific applications
- Transactional applications
- ...

Finding **uniques** is crucial for:
- Query optimization
- Anomaly detection
- Data modeling
- Indexing
- ...

## Problem

- Unique column combinations often **unknown** in big datasets
- **Exponential** search space

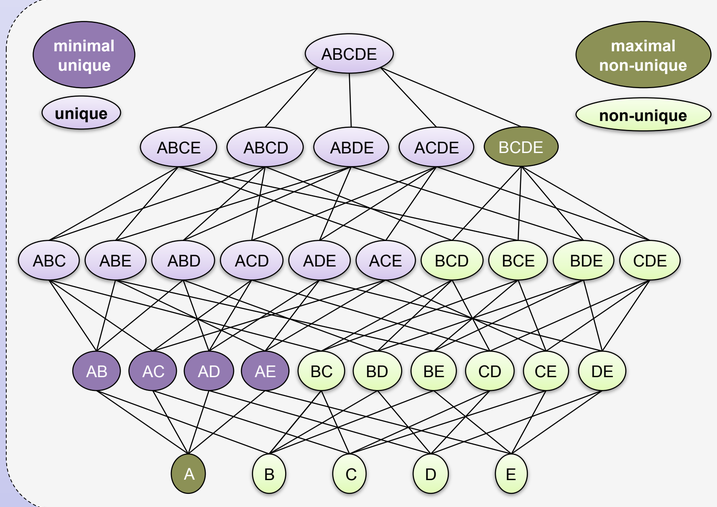


Lattice with eight attributes

Finding unique column combinations is an **NP-Hard** problem

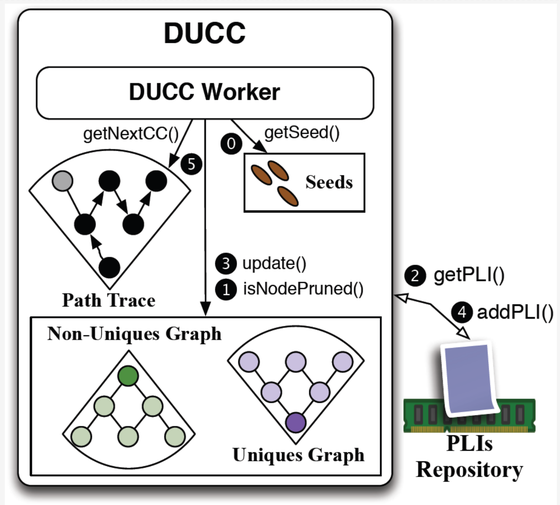
## DUCC

### Modeled as graph coloring problem



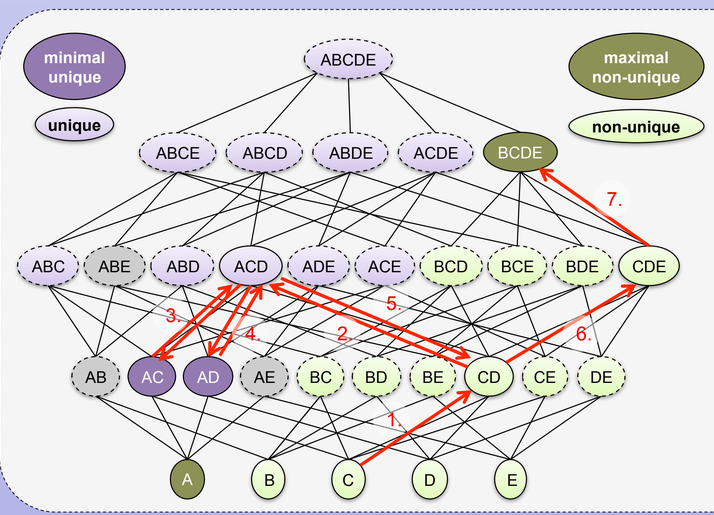

- Graph **divided** into **uniques** and **non-uniques**
- **Minimal uniques** summarize uniques
- **Maximal non-uniques** summarize non-uniques
- DuCC **simultaneously** detects (non-)uniques
- **Completeness** verified (proof in paper)

### Modular architecture



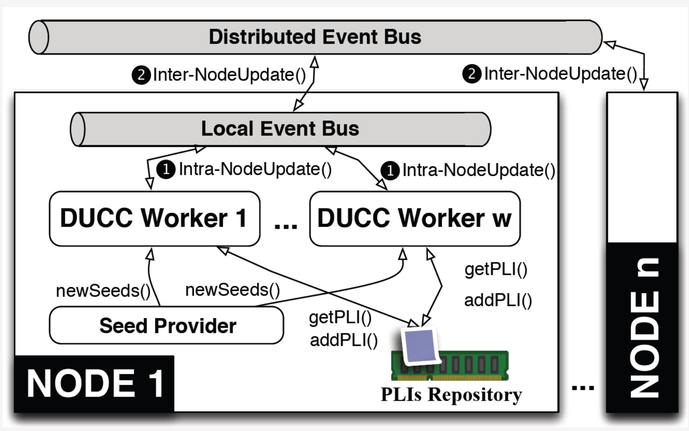

- Worker receives next CC from traversal algorithm
- Performs **fast** check with **PLI intersection**
- Maintains **pruning** data structures
- **Seed provider** detects holes and calculates restart CC

### Random walk graph traversal




- **Randomly** pick next CC from current CC
- Go **upwards** if CC is **non-unique**
- Go **downwards** if CC is **unique**
- **Trace back** if no pruned CC is left
- Check for **holes** by comparing min uniques and max non-uniques

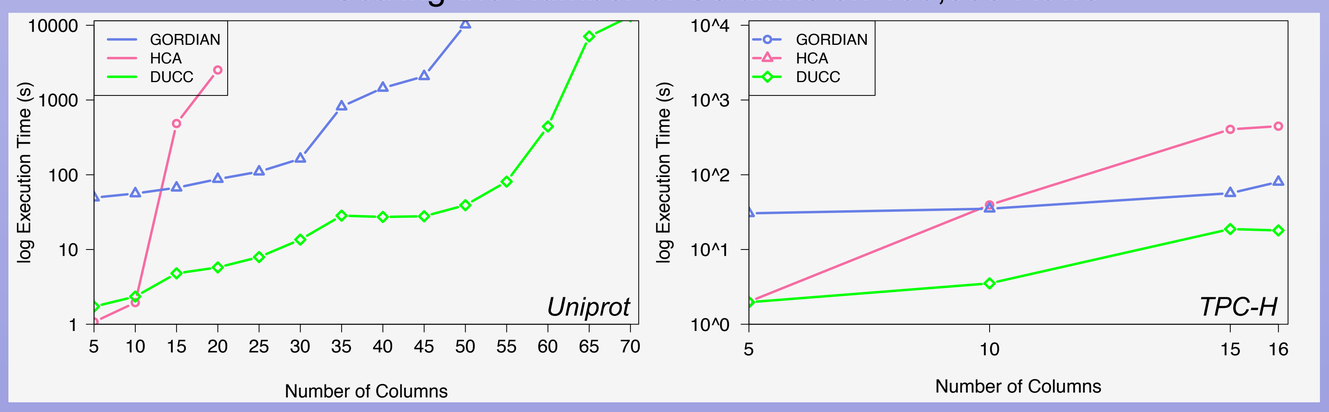
### Scalable architecture



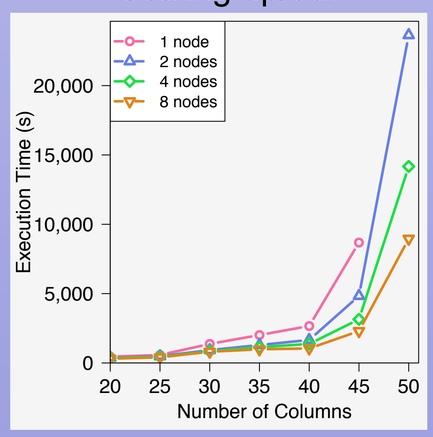

- Workers **exchange** minimal uniques and maximal non-uniques
- **Scale up** with local event bus
- **Scale out** with **distributed event bus** (ZooKeeper)
- **Fault-tolerance** with Map-only Hadoop Job

## Results

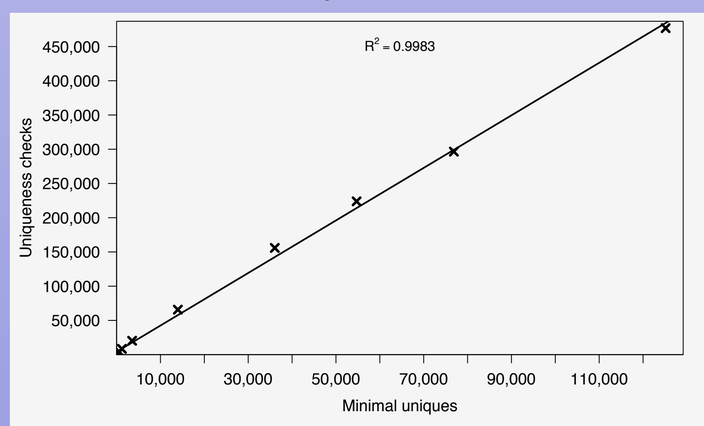
### Scaling the Number of Columns on 100,000 Rows



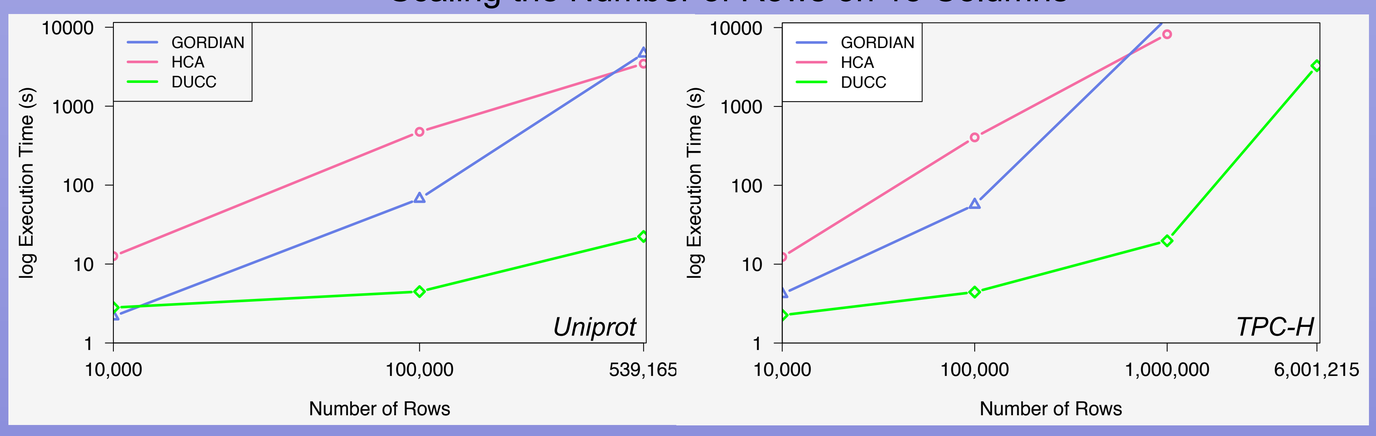

### Scaling up/out




### Solution space and checks




### Scaling the Number of Rows on 15 Columns




## Related Work

Gordian: [*Gordian: efficient and scalable discovery of composite keys. VLDB'06*]
- Row-based approach
- Prefix-tree data organization

HCA: [*Advancing the discovery of unique column combinations. CIKM'11*]
- Column-based approach
- Histograms- and value-counting-based

Swan: [*Detecting Unique Column Combinations on Dynamic Data. ICDE'14*]
- Builds on top of DUCC
- Focus on dealing with incremental data

* Work done in the context of ***Metanome***: joint project between HPI and QCRI that provides a fresh view on *data profiling* and aims at providing *scalability* for *Big Data*.
**Website**: http://www.hpi.uni-potsdam.de/naumann/projekte/metanome_data_profiling.html


Arvid Heise[1] Jorge-Arnulfo Quiané-Ruiz[2] Ziawasch Abedjan[1] Anja Jentzsch[1] Felix Naumann[1]

[1]Hasso Plattner Institute [2]Qatar Computing Research Institute