Discovering Unique Column Combinations on Dynamic Data

Motivation

Production of big datasets at very fast rates:

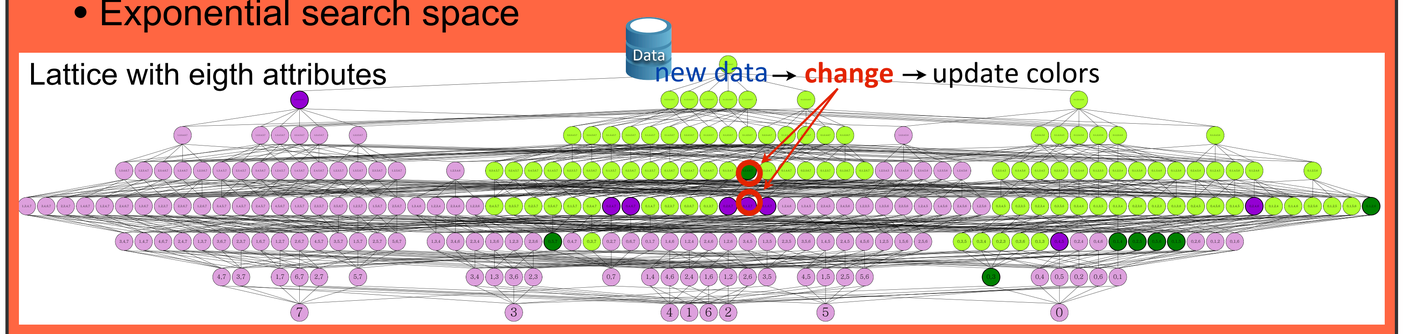
- Social networks
- Scientific applications
- Transactional applications
- ...

Knowing uniques is crucial for:

- Query optimization
- Anomaly detection
- Data modeling
- Indexing
- ...

Problem

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



- Finding unique column combinations is an NP-Hard problem
- Big datasets change quite often

Swan

Dealing with Inserts

Repository
Minimal Uniques: {Name, Age}, {Phone}
Maximal Non-Uniques: {Name}, {Age}

Tuple ID	Name	Phone	Age
($tuple_1$)	Lee	345	20
($tuple_2$)	Payne	245	30
($tuple_3$)	Lee	234	30

($insert_1$)	**Payne**	**245**	**31**

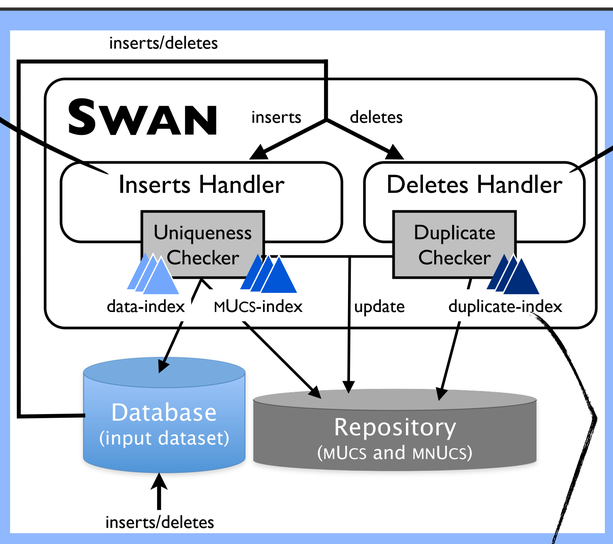
Inverted Indexes: I_{age} I_{Phone}

Retrieved tuples IDs with I_{Age}: {}

Retrieved tuples IDs with I_{Phone}: {($tuple_2$)}

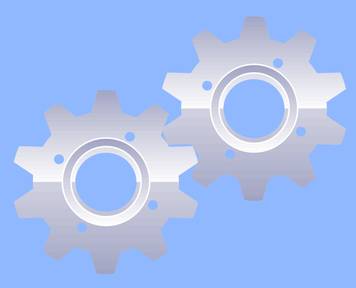
($tuple_2$)	Payne	245	30
($insert_1$)	Payne	245	31

Repository
Minimal Uniques: {Name, Age}, **{Age, Phone}**
Maximal Non-Uniques: **{Name, Phone}**, {Age}



Swan Architecture

Index Creation Mechanism



- Full scans for each change?
- Indexing all attributes?

- Minimal set of indexes:
  - cover all minimal uniques
- Speed-up indexes:
  - reducing number of tuples to retrieve

Dealing with Deletes

Repository
Minimal Uniques: {Name, Age}, {Phone}
Maximal Non-Uniques: {Name}, {Age}

Tuple ID	Name	Phone	Age
($tuple_1$)	Lee	345	20
($tuple_2$)	Payne	245	30
($tuple_3$)	Lee	234	30

Delete:

~~($tuple_1$)~~	~~Lee~~	~~345~~	~~20~~

Duplicate Indexes

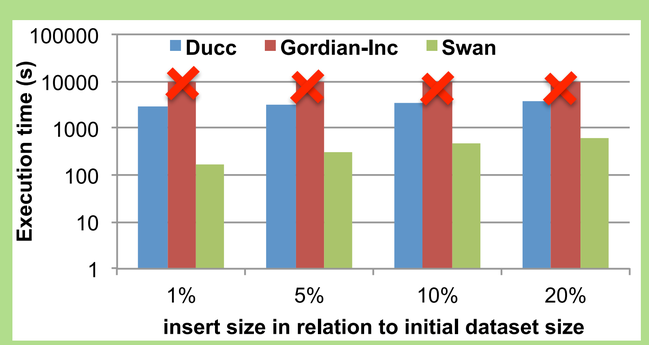
Name: {(~~$tuple_1$~~, ~~$tuple_3$~~)} → {}

Phone: {} → {}

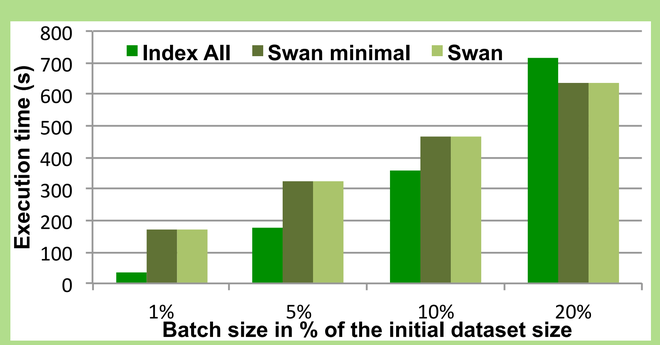
Age: {($tuple_2$, $tuple_3$)} → {($tuple_2$, $tuple_3$)}

Repository
Minimal Uniques: **{Name}**, {Phone}
Maximal Non-Uniques: **{Age}**

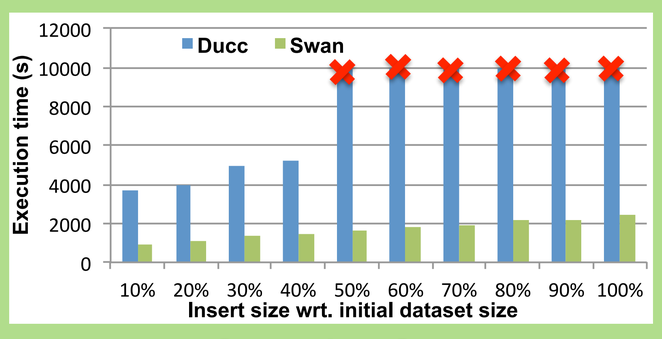
Results



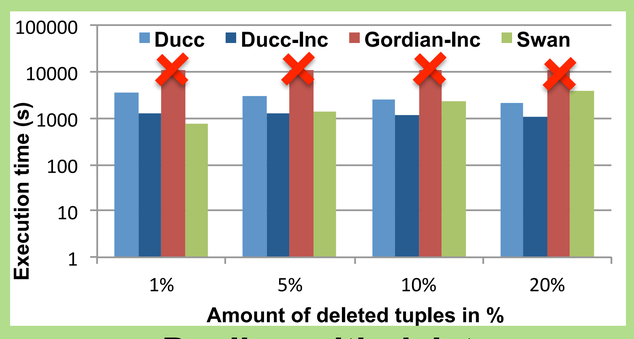
Dealing with inserts



Index evaluation



Stressing Swan



Dealing with deletes

5 millions rows, Lineitem TPC-H

Related Work

Gordian:
- Row-based approach
- Prefix-tree data organization

[Gordian: efficient and scalable discovery of composite keys. VLDB'06]

HCA:
- Column-based approach
- Histograms- and value-counting-based

[Advancing the discovery of unique column combinations. CIKM'11]

Ducc:
- Hybrid (row- and column-based) approach
- Depth-first + Random walk lattice traversal

[Scalable discovery of unique column combinations. VLDB'14]

Metanome

- It is a joint project between **HPI** and the **QCRI**
- It provides a fresh view on data profiling
- It aims at providing scalability for Big Data

Website: http://www.hpi.uni-potsdam.de/naumann/projekte/metanome_data_profiling.html

Ziawasch Abedjan
Hasso Plattner Institute

Jorge-Arnulfo Quiané-Ruiz
Qatar Computing Research Institute

Felix Naumann
Hasso Plattner Institute

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر