OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

MS-II. Concepts

Ali Hashaam, Ali Memon, Guzel Mussilova, Pavlo Shevchenko

Scientific Project: Databases for Multi-Dimensional Data, Genomics and Modern Hardware

May 23, 2017

Table of Contents

Blinktopus

- Octopus + BlinkDB = Blinktopus
- Our Goal

Concepts

- AQP.Architecture
- AQP.Synopses Manager
- Types of Synopses
- Types of Synopses: Samples
- Types of Synopses: Histograms
- Types of Synopses: Wavelets
- Types of Synopses: Sketches

Building a Blinktopus. Recall

Building a Blinktopus. Progress

Project Organisation

- Roles

Octopus + BlinkDB = Blinktopus

Create a new type of database system without fixed store that will mimic several existing systems

The goal is to provide approximate answers with acceptable accuracy in orders of magnitude less time than that for the exact query processing.[1]

[1] Liu, Qing. Approximate Query Processing (Reference work entry) in: Liu, Ling, and M. Tamer zsu. Encyclopedia of database systems. Vol. 6. Berlin, Heidelberg, Germany: Springer, 2009.

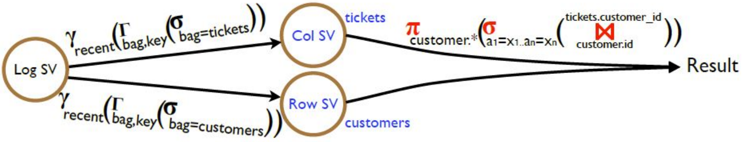
Our Goal

To provide a **framework** that gives user a chance to act as *Holistic SV Optimizer* like in OctopusDB

Add **Approximate Query Processing (AQP)** techniques

Evaluate performance depending on choice of SV

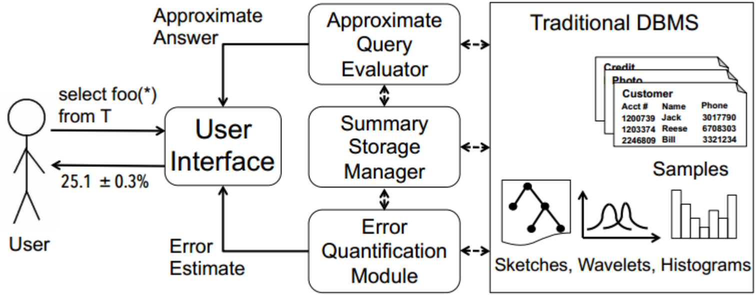
Octopus in Blinktopus


Log SV
$\gamma_{recent}(\Gamma_{bag.key}(\sigma_{bag=tickets}))$
$\gamma_{recent}(\Gamma_{bag.key}(\sigma_{bag=customers}))$
Col SV
Row SV
tickets
customers
$\pi_{customer.*}(\sigma_{a1=x1..an=xn}(\bowtie_{customer.id}^{tickets.customer\_id}))$
Result

(d) SV Transformation: use more efficient Row SV and Col SV [2]

[2] Jindal, Alekh. "OctopusDB: flexible and scalable storage management for arbitrary database engines." (2012).

AQP.Architecture



[3]

[3] The general anatomy of approximate query processing system

AQP.Synopses Manager

A synopsis captures essential properties of the real data while taking less space. The synopses manager is responsible for:

- Type of summary to use(Samples, histograms, sketches, wavelets etc.)
- When to build it (offline vs. online)
- How to store it (to use overlapping samples, how to structure/index/cache the synopses)
- When to update it (batch or online)

Types of Synopses

4 main families of synopses[4]:

- Samples
- Histograms
- Wavelets
- Sketches

[4] Cormode, Graham, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. "Synopses for massive data: Samples, histograms, wavelets, sketches." Foundations and Trends in Databases 4, no. 13 (2012): 1-294.

Types of Synopses: Samples

Representative subset, chosen by stochastic sampling methods (e.g. Bernoulli, stratified,simple random with and without replacement).

Types of Synopses: Samples

Representative subset, chosen by stochastic sampling methods (e.g. Bernoulli, stratified,simple random with and without replacement).

- The same schema as the data is used to sample, thus they the specialized operators are not needed.
- Unbiased estimators for SUM/AVG queries are straightforwardly built.
- Can be constructed immediately after user query has been issued, without incurring a delay.
- Imprecise estimates of a query result can be incrementally enhanced by collecting more samples.
- Due to general-purpose data structure can be used to answer a wide variety of arbitrary queries.

Types of Synopses: Samples

- Poor estimations for less results.
- For larger relations more advanced techniques to make the sampling scalable are to be implemented.
- Selectivity estimations over larger datasets might be less efficient.
- Sensitive to skew and outliers.
- Hard to use with (NOT-)/IN, DISTINCT, EXISTS queries.

Types of Synopses: Histograms

A binned representation of the data distribution. The summary and bucket information is used to (approximately) reconstruct the data in the bucket in order to approximately answer the query.

Types of Synopses: Histograms

A binned representation of the data distribution. The summary and bucket information is used to (approximately) reconstruct the data in the bucket in order to approximately answer the query.

- A natural solution for range-sum queries.
- Due to conceptual simplicity can be effectively used for a broad variety of estimation tasks(E.g. set-valued queries, real-valued data, and aggregate queries over predicates that more complex than simple ranges.
- Relatively simple in interpretation.
- Practically acceptable accuracies, provided that the sufficient storage space are allocated.

Types of Synopses: Histograms

- Sensitive to dimensionality.
- Performance strongly depends on bucketing schemes(how the buckets are chosen, what statistics are stored, how estimates are extracted, and what classes of query are supported).
- Incremental maintenance.
- Might provide too loose error estimates over the class of queries.

Types of Synopses: Wavelets

Transform the data to represent the most significant features in a frequency domain and can capture combinations of high and low frequency information.

Types of Synopses: Wavelets

Transform the data to represent the most significant features in a frequency domain and can capture combinations of high and low frequency information.

- Useful for range-sum queries.
- Through an appropriately defined AQP algebra general SPJ (select, project, join) queries can be applied on relation summaries.
- More amenable to maintenance under dynamic data due to the linearity of the basic Haar transform.
- Large number of coefficients must be retained in order to guarantee accurate reconstruction of the data distribution in the multi-dimensional wavelet.

Types of Synopses: Sketches

- Especially appropriate for streaming data.
- Each new piece of data might be independent of the current state of the summary, which makes them faster to process, and easy parallelizable.
- Sketch summaries can be used as primitives within more complex mining operations, and to extract wavelet and histogram representations of streaming data.

Types of Synopses: Sketches

- Tends to be focused on answering a single type of query.
- Number of parameters affect the accuracy and probability of failure.
- Techniques do not extend well to more complex queries which combine multiple sub-queries.
- The only complexity is mathematical (for complete accuracy estimation).

Building a Blinktopus. Recall

First, the Octopus:

- Store incoming data in logs.
- Query the logs (just a filter query)
- Allow users to create views (row, column) over certain logs.
- List all views and logs
- Launch the query over views or over logs, see the changes in performance.

Building a Blinktopus. Recall

Enter AQP:

- What synopsis can we easily support as a view for a specific query? Which will we choose to test? (Samples, histograms?)
- Do Octopuses and AQP match well together?
- How will we allow users to build this view?
- How will we support queries using this view?

Building a Blinktopus. IDE

- Back-end

- Front-end

[5]

[5] Sources: http://jupyter.org/
http://honstain.com/new-dropwizard-1-0-5-java-service/

Project Organisation.Roles

Team:

Guzel - Team Leader-Researcher
Pavlo - Developer
Ali H. - Developer
Ali M. - Researcher

Supervisor:

Gabriel Campero Durand

Changing roles after each milestone.

Project Organisation.Schedule

Milestones

Date	Phase	Status
02.05.2017	MS-I (Kick-Off)	done
23.05.2017	MS-II (Concepts)	done
13.06.2017	MS-III (Implementation)	
04.07.2017	MS-IV (Final)	

Meetings

Team Meetings: Mo 14-15
Meetings with supervisor: We 10-11

Thank you! Any questions?

Literature

1. Jindal, Alekh. "The mimicking octopus: Towards a one-size-fits-all database architecture." VLDB PhD Workshop. 2010.
2. Dittrich, Jens, and Alekh Jindal. "Towards a One Size Fits All Database Architecture." CIDR. 2011.
3. Jindal, Alekh. "OctopusDB: flexible and scalable storage management for arbitrary database engines." (2012).
4. Idreos, Stratos, Martin L. Kersten, and Stefan Manegold. "Database Cracking." In CIDR, vol. 7, pp. 68-78. 2007.
5. Mozafari, Barzan. "Approximate query engines: Commercial challenges and research opportunities." SIGMOD, 2017.
6. Cormode, Graham, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. "Synopses for massive data: Samples, histograms, wavelets, sketches." Foundations and Trends in Databases 4, no. 13 (2012): 1-294.