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

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.

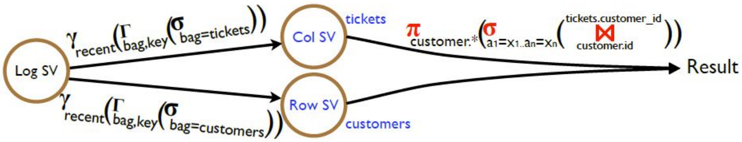
Our Goal

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

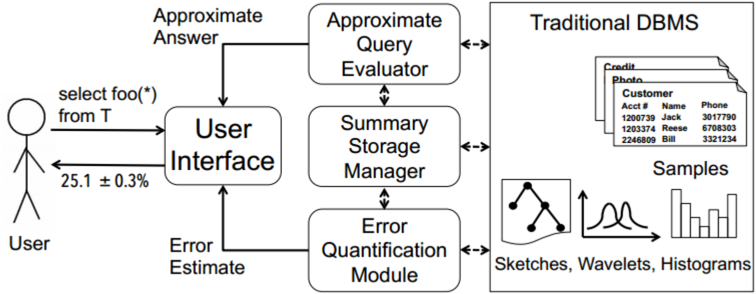
Octopus in Blinktopus



(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).

- Easy to implement.
- Through adhering the same data structure as original tables support the widest range of queries.
- Unbiased estimators for SUM/AVG queries are straightforwardly built.
- Due to the immediate construction after issuing user query do not incur a delay.
- Collecting more samples incrementally enhance imprecise estimates of a query results.
- Without using the attribute value independence assumption allow high quality selectivity estimations.

Types of Synopses: Samples

- Poor estimations for less results.
- Larger relations need more advanced techniques to make the sampling more scalable.
- Selectivity estimations over larger datasets are less efficient.
- Sensitive to skew and outliers.
- Hard to use with (NOT-)/IN, DISTINCT, EXISTS queries.
- Might be difficult to interpret by statistically unsophisticated users.

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.
- Conceptual simplicity allows an effective use of 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 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 significant features in a frequency domain and can capture combinations of high and low frequency information.

- Useful for range-sum queries.
- Appropriately defined AQP algebra allows applying general SPJ (select, project, join) queries on relation summaries.
- The linearity of the basic Haar transform manages a better maintenance under dynamic data than histograms.
- Non-trivial to maintain wavelet coefficients under arbitrary update patterns.
- Large number of coefficients must be retained to guarantee an accurate reconstruction of the data distribution in the multi-dimensional wavelet.

Types of Synopses: Sketches

- Especially appropriate for streaming data.
- Process fasters and easily paralleliazed if each new piece of data is independent of the current state of the summary.
- Can be used as primitives within more complex mining operations, and to extract wavelet and histogram representations of streaming data.

Types of Synopses: Sketches

- Mostly focused on answering a single type of query.
- Number of parameters affects 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).

Types of Synopses: Sketches

- Transformation
- Data structure
- Set of estimator algorithms

The HyperLogLog is a famous sketch supporting count-distinct.

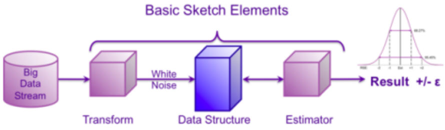


Figure: Distinct Count Sketch. High-level View.

5

[5] Source: https://yahooeng.tumblr.com/post/135390948446/data-sketches

Building a Blinktopus. Recall

First, the Octopus:

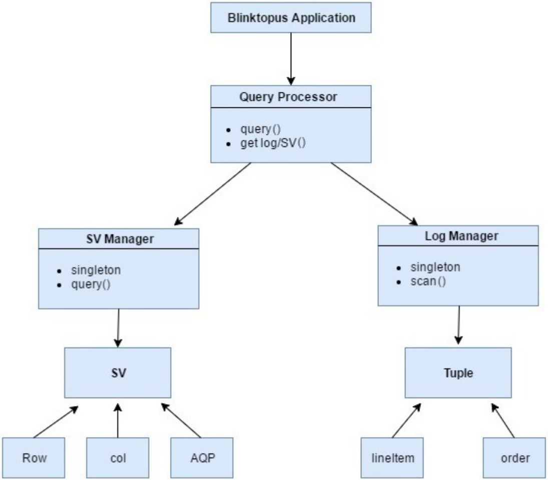
- Store incoming data in logs.
- Query the logs (just a filter query).
- Allow users to create views (row, column) over certain logs.
- List all views and logs.
- Launch the query over views or over logs, see the changes in performance.

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. Workflow



Building a Blinktopus. IDE

- Back end
- Front end [6]

[6] 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.

Meetings:

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

Project Organisation.Schedule

Task	System Part	Task Type	Begin Date	Finish Date	Responsible
Team meeting	All	Meeting	18.04.2017		Guzel Mussilova
Meeting with Supervisor	All	Meeting	19.04.2017		Gabriel Campero
MS-I. Kick-Off	All	Presentation	23.04.2017	01.05.2017	Pavlo Shevchenko, Ali Hashaam
Team meeting	All	Meeting	24.04.2017		Guzel Mussilova
Meeting with Supervisor	All	Meeting	26.04.2017		Gabriel Campero
Team meeting	All	Meeting	01.05.2017		Guzel Mussilova
MS-I. Kick-Off	All	Presentation	03.05.2017		Pavlo Shevchenko, Ali Hashaam
Meeting with Supervisor	All	Meeting	03.05.2017		Gabriel Campero
DropWizard	Back-end	Research	03.05.2017	10.05.2017	Pavlo Shevchenko
Jupyter Notebook	Front-end	Research	03.05.2017	10.05.2017	Ali Hashaam
AQP Synopses	All	Research	03.05.2017	11.05.2017	Guzel Mussilova, Ali Raza
Load data from CSV	Front-end	Implementation	04.05.2017	10.05.2017	Ali Hashaam
Team meeting	All	Meeting	08.05.2017		Guzel Mussilova
Meeting with Supervisor	All	Meeting	10.05.2017		Gabriel Campero
Classes (Lineitem, Order)	Back-end	Implementation	10.05.2017	15.05.2017	Pavlo Shevchenko
Resource class(Primary Log)	Back-end	Implementation	10.05.2017	15.05.2017	Pavlo Shevchenko
Data Import	Back-end	Implementation	10.05.2017	15.05.2017	Pavlo Shevchenko
StorageView POJO	Back-end	Implementation	10.05.2017	15.05.2017	Pavlo Shevchenko
StorageView Manager	Back-end	Implementation	10.05.2017	15.05.2017	Pavlo Shevchenko
Handling user inputs	Front-end	Implementation	10.05.2017	15.05.2017	Ali Hashaam
AQP(architecture/components).	All	Research	11.05.2017	14.05.2017	Guzel Mussilova, Ali Raza
Implementation of Sketches in JAVA	All	Research	14.05.2017	14.05.2017	Ali Raza
Team meeting	All	Meeting	15.05.2017		Guzel Mussilova
MS-II. Concepts	All	Presentation	15.05.2017	22.05.2017	Guzel Mussilova, Ali Raza
Meeting with Supervisor	All	Meeting	17.05.2017		Gabriel Campero
HyperlogLog(working principle)	All	Research	18.05.2017	22.05.2017	Guzel Mussilova, Ali Raza
Integration DP&JN	All	Implementation	18.05.2017	22.05.2017	Ali Hashaam, Pavlo Shevchenko
Processing JSONHandler Function	Front-end	Implementation	18.05.2017	25.05.2017	Ali Hashaam
Query Manager	Back-end	Implementation	20.05.2017	22.05.2017	Pavlo Shevchenko
Team meeting	All	Meeting	22.05.2017		Guzel Mussilova
MS-II. Concepts	All	Presentation	23.05.2017		Guzel Mussilova, Ali Raza

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. Mozafari, Barzan. "Approximate query engines: Commercial challenges and research opportunities." SIGMOD, 2017.
5. 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.
6. https://yahooeng.tumblr.com/post/135390948446/data-sketches