INF FAKULTÄT FÜR INFORMATIK

MS-III. Implementation

Ali Hashaam, Ali Memon, Guzel Mussilova, Pavlo Shevchenko

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

June 13, 2017

Table of Contents

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

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.

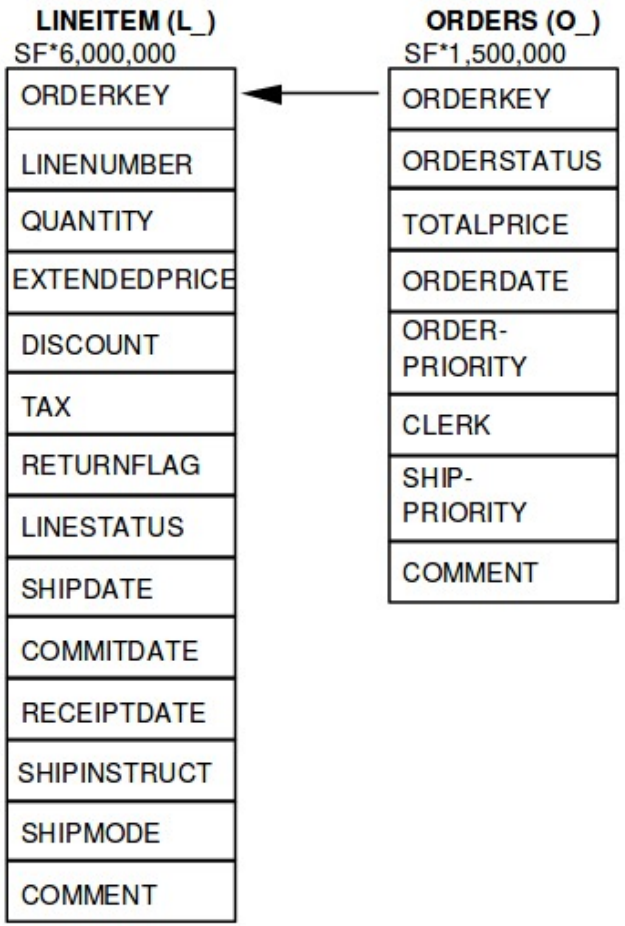
Building a Blinktopus. Recall

Enters Approximate Query Processing (AQP):

- Which synopsis will we choose to test? (Samples, histograms, sketches?)
- Do Octopuses and AQP match well together?
- Build the selected synopsis on the whole data, after data insertions.
- Using the synopsis, answer the user queries by reconstructing the approximate data.

Building a Blinktopus. Implementation

Schema


LINEITEM (L_)
SF*6,000,000
ORDERKEY
LINENUMBER
QUANTITY
EXTENDEDPRICE
DISCOUNT
TAX
RETURNFLAG
LINESTATUS
SHIPDATE
COMMITDATE
RECEIPTDATE
SHIPINSTRUCT
SHIPMODE
COMMENT
ORDERS (O_)
SF*1,500,000
ORDERKEY
ORDERSTATUS
TOTALPRICE
ORDERDATE
ORDER-PRIORITY
CLERK
SHIP-PRIORITY
COMMENT

OctopusDB. Customization/Alteration/Power to User/Variation

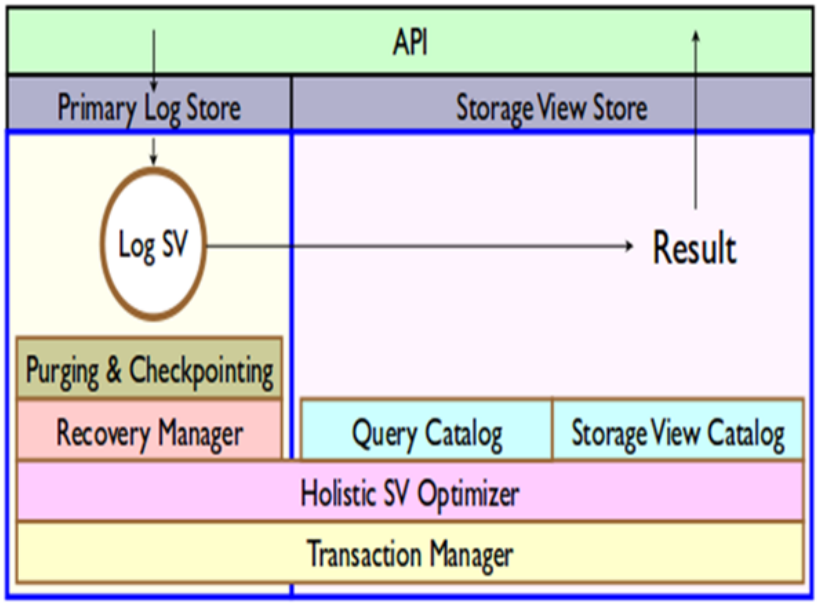


Figure 2: OctopusDB Architecture.

OctopusDB. Customization/Alteration/Power to User/Variation

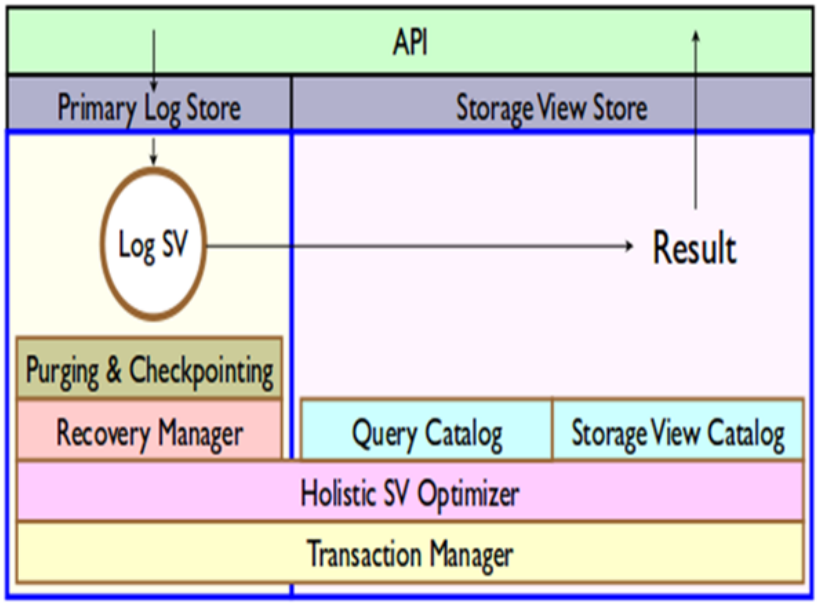


Figure 2: OctopusDB Architecture.

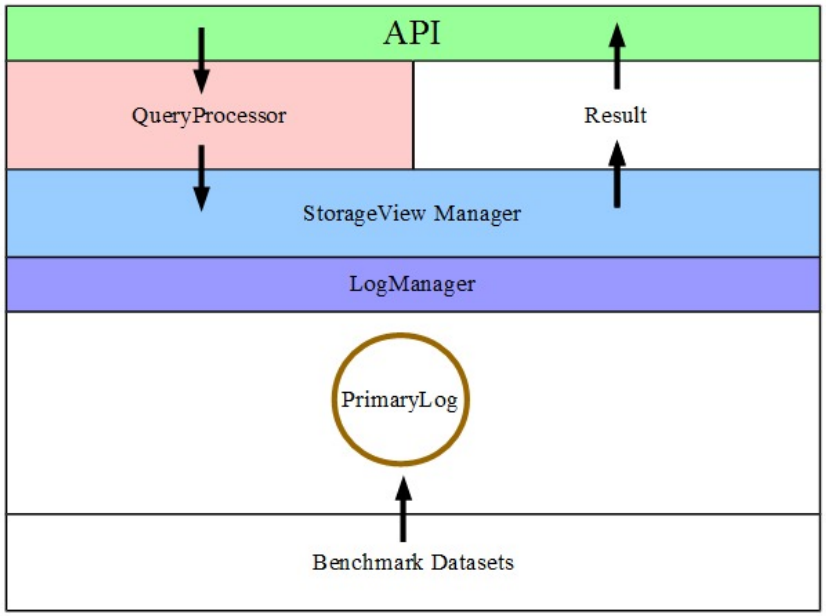


Figure 3: Blinktopus.

OctopusDB. Evaluation

Log	37650411,93
RowSV	10773667,29
ColSV	6965524,85

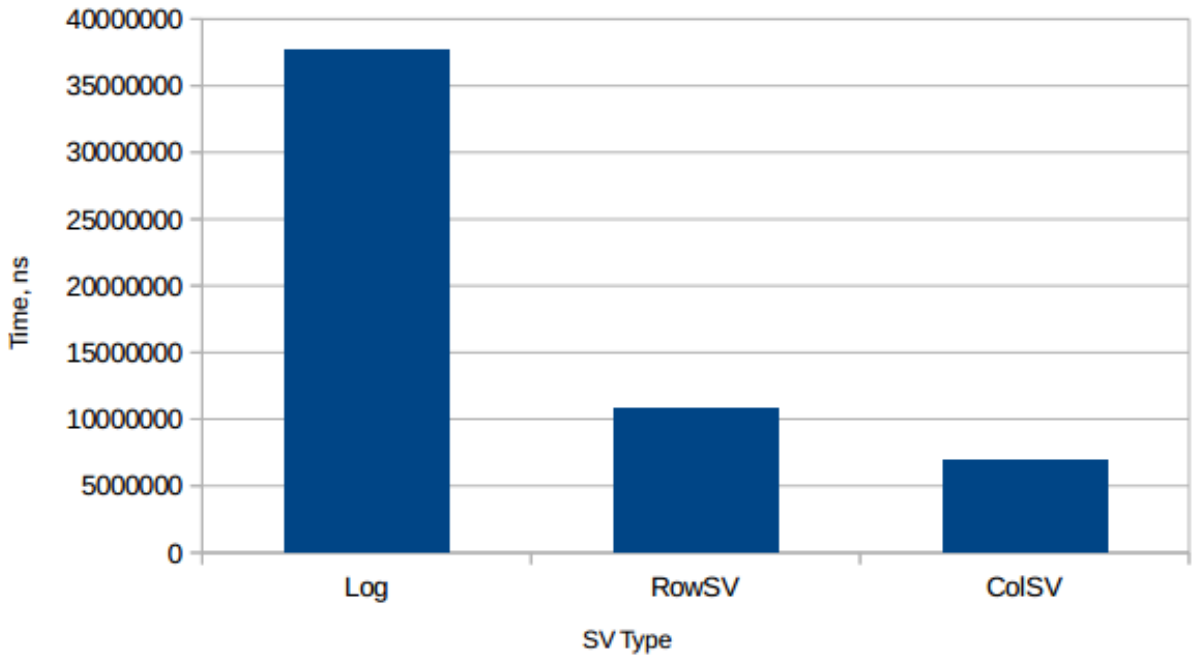


Figure 4: Evaluation result for 100 runs over Totalprice Column in Orders with Range from 50,000 to 200,000.

AQP. Synopses

4 main families of synopses[1]:

- Samples
- Histograms ✓
- Wavelets
- Sketches ✓

[1] 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. 1-3 (2012): 1-294.

AQP. Histograms

In histogram's development, main cornerstones are:

- Partition the dataset into buckets.
- Store summary statistics for each bucket about the data values in the it.
- Store information about the buckets themselves, like bucket boundaries.

At query time, the summary and bucket information is used to approximately answer the query.

AQP. Histograms

Vital Points to consider:

- Bucketing Scheme
- Statistics Stored per Bucket
- Approximation Scheme
- Class of queries answered
- Efficiency
- Accuracy & Error Estimates
- Incremental Maintenance

AQP. Histograms

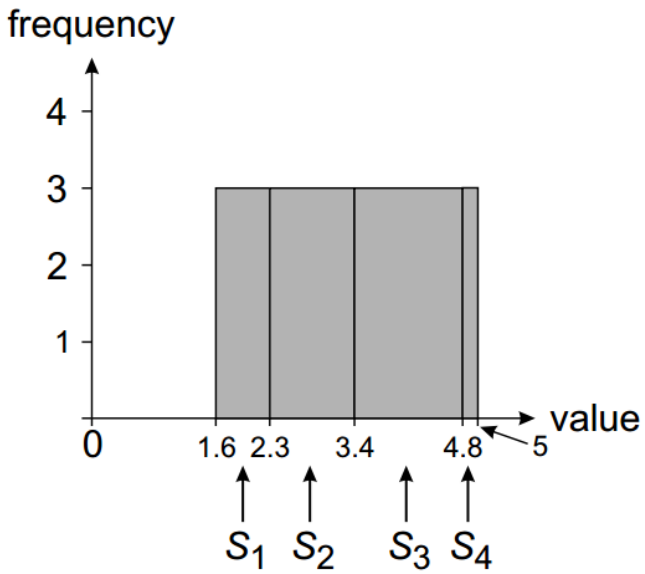


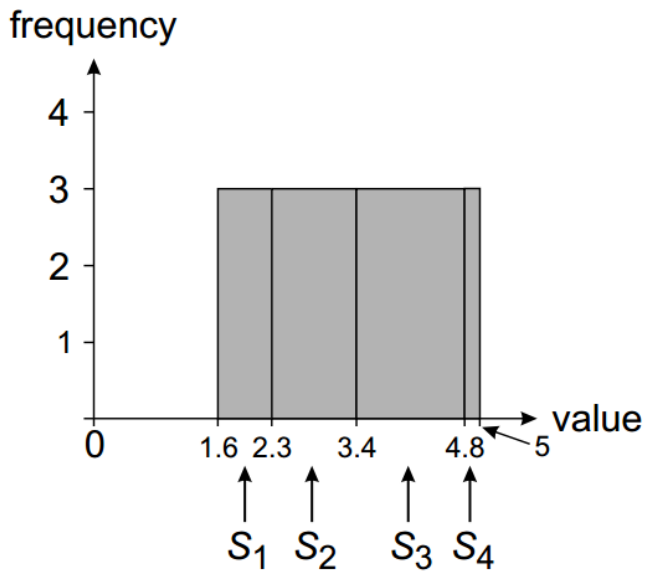
Figure 5: Equi-Depth Histogram

To calculate number of bins 'k':

$$k = 2n^{1/3} \qquad \text{(RICE RULE)}$$

AQP. Histograms

What if the count of the values between 1.1 and 4.5 is required?



Continuous value Assumption allows the estimation of values inside a bucket via interpolation.

$$N = 3 + 3 + ((4.5 - 3.4)/(4.8 - 3.4))3 = 8.4$$

AQP. Histograms

- Histograms are a natural solution for range-sum and range-count queries.
- Conceptual simple and relatively simple in interpretation.
- Practically acceptable accuracies, provided that the sufficient storage space are allocated.

AQP. 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.

AQP. Sketches

AQP. HLL

Building a Blinktopus. IDE

- Back end 

- Front end 

[2]

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

Project Organisation.Roles

Team:

Guzel - Team Leader-Researcher
Pavlo - Developer (Backend - OctopusDB)
Ali H. - Developer (Backend - AQP)
Ali M. - Developer (Frontend - User Views)

Supervisor:

Gabriel Campero Durand

Changing roles after each milestone.

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. Mozafari, Barzan, and Ning Niu. "A Handbook for Building an Approximate Query Engine." IEEE Data Eng. Bull. 38, no. 3 (2015): 3-29.
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.