Efficient Structure-aware OLAP Query Processing over Large Property Graphs

by

Yan Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Property graph model is a popular semantic rich model for real-world applications concerning graph structure data, like social networks, financial transaction networks and ect. On-Line Analytical Processing(OLAP) provides an important tool for data analyses by allowing users to perform data aggregation through different combinations of dimentions. For instance, over a Q&A forum dataset, in order to study if there is a correlation between age and post quality,one may ask what is the average user's age group by post score. In the field of music industry, we may process a query like what is total sales of records with respect to music company and year to study market activities.

State-of-art graph databases like neo4j do not have efficient support for OLAP aggregation queries. Neo4j processes each OLAP query in two steps. First expands nodes and edges to the query structure, and then perform aggregation. Even if a query is repeatedly executed for multiple times, in each round Neo4j still processes the query from scratch, without caching any structure-wise "knowledge" from previous workload. When it comes to large property graphs, current graph databases' efficiency is far from satisfaction. It is unacceptable for users to wait for hours for results of a single query.

We propose a system that greatly improves efficiency of OLAP over large property graphs. The idea is to smartly materialize some views(in main memory or hard disk) that a client was interested in based on a client's previous workload. Hopefully such materialization can be used to accelerate future query processing.

We implemented our system on top of Neo4j and compared our system with orginal Neo4j system. We wrote some practical OLAP queries and randomly partition them into previous workload and future workload, and them executed future workload using both our system and Neo4j. Result shows that with an acceptable cost of memory or disk usage, we are able to improve OLAP processing efficiency by 10-30 times.

Acknowledgements

I would like to thank Professor Tamer Ozsu and Dr. Xiaofei Zhang who made this thesis possible.

Dedication

This is dedicated to my mother Limei Leng whom I love.

Table of Contents

List of Tables

List of Figures

Glossary

computer A programmable machine that receives input data, stores and manipulates the data, and provides formatted output 1

Abbreviations

AAAAZ American Association of Amature Astronomers and Zoologists 1

Nomenclature

dingledorf A person of supposed average intelligence who makes incredibly brainless misjudgments 1

List of Symbols

$\mathbf{v}$ Random vector: a location in n-dimensional Cartesian space, where each dimensional component is determined by a random process 1

Chapter 1

Introduction

In the beginning, there was π:

$$e^{\pi i} + 1 = 0 \tag{1.1}$$

A computer could compute π all day long. In fact, subsets of digits of π's decimal approximation would make a good source for psuedo-random vectors, $\mathbf{v}$.

1.1 State of the Art

See equation 1.1 on page 1.[1]

1.2 Some Meaningless Stuff

The credo of the American Association of Amature Astronomers and Zoologists (AAAAZ) was, for several years, several paragraphs of gibberish, until the dingledorf responsible for the AAAAZ Web site realized his mistake:

”Velit dolor illum facilisis zzril ipsum, augue odio, accumsan ea augue molestie lobortis zzril laoreet ex ad, adipiscing nulla. Veniam dolore, vel te in dolor te, feugait dolore ex vel erat duis nostrud diam commodo ad eu in consequat esse in ut wisi. Consectetuer

[1] A famous equation.

dolore feugiat wisi eum dignissim tincidunt vel, nostrud, at vulputate eum euismod, diam minim eros consequat lorem aliquam et ad. Feugait illum sit suscipit ut, tation in dolore euismod et iusto nulla amet wisi odio quis nisl feugiat adipiscing luptatum minim nisl, quis, erat, dolore. Elit quis sit dolor veniam blandit ullamcorper ex, vero nonummy, duis exerci delenit ullamcorper at feugiat ullamcorper, ullamcorper elit vulputate iusto esse luptatum duis autem. Nulla nulla qui, te praesent et at nisl ut in consequat blandit vel augue ut.

Illum suscipit delenit commodo augue exerci magna veniam hendrerit dignissim duis ut feugait amet dolor dolor suscipit iriure veniam. Vel quis enim vulputate nulla facilisis volutpat vel in, suscipit facilisis dolore ut veniam, duis facilisi wisi nulla aliquip vero praesent nibh molestie consectetuer nulla. Wisi nibh exerci hendrerit consequat, nostrud lobortis ut praesent dignissim tincidunt enim eum accumsan. Lorem, nonummy duis iriure autem feugait praesent, duis, accumsan tation enim facilisi qui te dolore magna velit, iusto esse eu, zzril. Feugiat enim zzril, te vel illum, lobortis ut tation, elit luptatum ipsum, aliquam dolor sed. Ex consectetuer aliquip in, tation delenit dignissim accumsan consequat, vero, et ad eu velit ut duis ea ea odio.

Vero qui, te praesent et at nisl ut in consequat blandit vel augue ut dolor illum facilisis zzril ipsum. Exerci odio, accumsan ea augue molestie lobortis zzril laoreet ex ad, adipiscing nulla, et dolore, vel te in dolor te, feugait dolore ex vel erat duis. Ut diam commodo ad eu in consequat esse in ut wisi aliquip dolore feugiat wisi eum dignissim tincidunt vel, nostrud. Ut vulputate eum euismod, diam minim eros consequat lorem aliquam et ad luptatum illum sit suscipit ut, tation in dolore euismod et iusto nulla. Iusto wisi odio quis nisl feugiat adipiscing luptatum minim. Illum, quis, erat, dolore qui quis sit dolor veniam blandit ullamcorper ex, vero nonummy, duis exerci delenit ullamcorper at feugiat. Et, ullamcorper elit vulputate iusto esse luptatum duis autem esse nulla qui.

Praesent dolore et, delenit, laoreet dolore sed eros hendrerit consequat lobortis. Dolor nulla suscipit delenit commodo augue exerci magna veniam hendrerit dignissim duis ut feugait amet. Ad dolor suscipit iriure veniam blandit quis enim vulputate nulla facilisis volutpat vel in. Erat facilisis dolore ut veniam, duis facilisi wisi nulla aliquip vero praesent nibh molestie consectetuer nulla, iriure nibh exerci hendrerit. Vel, nostrud lobortis ut praesent dignissim tincidunt enim eum accumsan ea, nonummy duis. Ad autem feugait praesent, duis, accumsan tation enim facilisi qui te dolore magna velit, iusto esse eu, zzril vel enim zzril, te. Nisl illum, lobortis ut tation, elit luptatum ipsum, aliquam dolor sed minim consectetuer aliquip.

Tation exerci delenit ullamcorper at feugiat ullamcorper, ullamcorper elit vulputate iusto esse luptatum duis autem esse nulla qui. Volutpat praesent et at nisl ut in consequat blandit vel augue ut dolor illum facilisis zzril ipsum, augue odio, accumsan ea augue molestie lobortis zzril laoreet. Ex duis, te velit illum odio, nisl qui consequat aliquip qui blandit hendrerit. Ea dolor nonummy ullamcorper nulla lorem tation laoreet in ea, ullamcorper vel consequat zzril delenit quis dignissim, vulputate tincidunt ut."

Chapter 2

Observations

This would be a good place for some figures and tables.

Some notes on figures and photographs...

- A well-prepared PDF should be
  1. Of reasonable size, *i.e.* photos cropped and compressed.
  2. Scalable, to allow enlargment of text and drawings.
- Photos must be bit maps, and so are not scaleable by definition. TIFF and BMP are uncompressed formats, while JPEG is compressed. Most photos can be compressed without losing their illustrative value.
- Drawings that you make should be scalable vector graphics, *not* bit maps. Some scalable vector file formats are: EPS, SVG, PNG, WMF. These can all be converted into PNG or PDF, that pdflatex recognizes. Your drawing package probably can export to one of these formats directly. Otherwise, a common procedure is to print-to-file through a Postscript printer driver to create a PS file, then convert that to EPS (encapsulated PS, which has a bounding box to describe its exact size rather than a whole page). Programs such as GSView (a Ghostscript GUI) can create both EPS and PDF from PS files. Appendix A shows how to generate properly sized Matlab plots and save them as PDF.
- It's important to crop your photos and draw your figures to the size that you want to appear in your thesis. Scaling photos with the includegraphics command will cause

loss of resolution. And scaling down drawings may cause any text annotations to become too small.

For more information on LaTeX see the uWaterloo Skills for the Academic Workplace course notes. [1]

The classic book by Leslie Lamport [3], author of LaTeX, is worth a look too, and the many available add-on packages are described by Goossens *et al* [1].

[1] Note that while it is possible to include hyperlinks to external documents, it is not wise to do so, since anything you can't control may change over time. It *would* be appropriate and necessary to provide external links to additional resources for a multimedia "enhanced" thesis. But also note that if the **hyperref** package is not included, as for the print-optimized option in this thesis template, any `\href` commands in your logical document are no longer defined. A work-around employed by this thesis template is to define a dummy `\href` command (which does nothing) in the preamble of the document, before the **hyperref** package is included. The dummy definition is then redifined by the **hyperref** package when it is included.

References

[1] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1994.

[2] Donald Knuth. *The TeXbook.* Addison-Wesley, Reading, Massachusetts, 1986.

[3] Leslie Lamport. *LaTeX — A Document Preparation System.* Addison-Wesley, Reading, Massachusetts, second edition, 1994.

APPENDICES

Appendix A

Matlab Code for Making a PDF Plot

A.1 Using the GUI

Properties of Matab plots can be adjusted from the plot window via a graphical interface. Under the Desktop menu in the Figure window, select the Property Editor. You may also want to check the Plot Browser and Figure Palette for more tools. To adjust properties of the axes, look under the Edit menu and select Axes Properties.

To set the figure size and to save as PDF or other file formats, click the Export Setup button in the figure Property Editor.

A.2 From the Command Line

All figure properties can also be manipulated from the command line. Here's an example:

```
x=[0:0.1:pi];
hold on % Plot multiple traces on one figure
plot(x,sin(x))
plot(x,cos(x),'--r')
plot(x,tan(x),'.-g')
title('Some Trig Functions Over 0 to \pi') % Note LaTeX markup!
legend('{\it sin}(x)','{\it cos}(x)','{\it tan}(x)')
hold off
```

```
set(gca,'Ylim',[-3 3]) % Adjust Y limits of "current axes"
set(gcf,'Units','inches') % Set figure size units of "current figure"
set(gcf,'Position',[0,0,6,4]) % Set figure width (6 in.) and height (4 in.)
cd n:\thesis\plots % Select where to save
print -dpdf plot.pdf % Save as PDF
```