CS 8803-MDS

Human-in-the-loop Data Analytics

Lecture 8

09/19/22

Logistics

Milestone 1: project proposal

- Assignment: 5PM Wednesday
- Presentation: https://bit.ly/3Bg0Kst
  - 5 min per group
  - a brief version of your introduction
  - order according to canvas signup
  - *not graded

Questions?

Today's class

Hillview: A trillion-cell spreadsheet for big data

Author: Akshay

Archaeologist: Eric

Researcher: Vishnu

Practioner: Ashmita

VLDB 2019 | General Information | Conference Program | Call for Contributions | Dates & Guidelines | Sponsorship | Participant Information

VLDB 2019: Overview

VLDB is a premier annual international forum for data management and database researchers, vendors, practitioners, application developers, and users. The VLDB 2019 conference will feature research talks, tutorials, demonstrations, and workshops. It will cover issues in data management, database and information systems research, since they are the technological cornerstones of the emerging applications of the 21st century.

VLDB 2019 will take place in Los Angeles, California, from August 26th to August 30th, 2019.

Tweets from @VLDB2019

VLDB 2019 @VLDB... · Sep 2, 2019

Sending a special shout-out to VLDB 2019's organizing committee and sponsors for making VLDB 2019 a grand success. And thanks to everyone for being a part of it. #VLDB2019 #VLDBSponsors

Hillview Demo

Contribution/Strengths

- No existing spreadsheet system scale to a trillion rows
- Compute only what you can display
- Provided data layer that allows connection to other tools – can operate directly on data stored in SQL, NOSQL, JSON, CSV, Hadoop, Spark, etc
- Makes no assumption about the distribution of data across its servers
- Hillview provides the user with a partial summary that gradually progresses to the final result
- Authors recognize and leverage the “short-lived” characteristic of spreadsheet results to optimize the system
  - Stateless worker nodes, independent GC etc.

Limitation/Weaknesses

- Data modifications are paused while Hillview runs.
- No discussion of the time required for rendering sudden changes made by a user (e.g., a user magnifies a histogram by 20x, the computation required to reach this resolution may take some time.)
- Vizketches require implementation for every type of visualization. This is not a generic scalable solution to various number of visualizations in practice.
- No experiments with different display setups
- Lack of comparison with other spreadsheet system

Limitation/Weaknesses

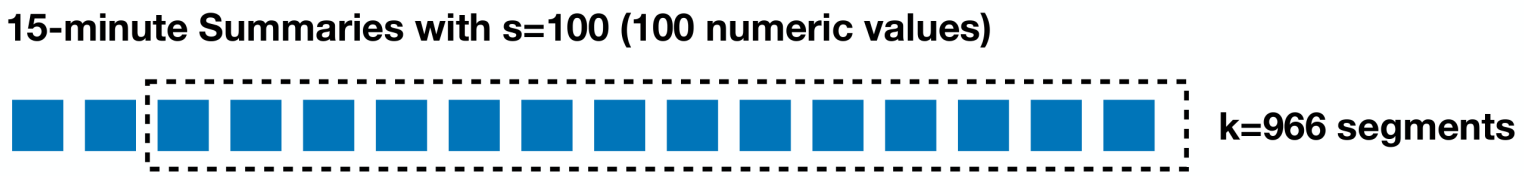
Barriers to adoption

- Hillview is not a readily usable system to augment existing spreadsheets. It serves as an independent web app and may provide limited commercial value with real users.
- It requires data to be horizontally partitioned - any current system that hasn't done this would be reluctant to adopt Hillview
- A specialized backend like Hillview can considerably improve the performance of spreadsheets. However, they do not justify this over the potential expense of bringing a new backend system to the servers. Shall the servers run a generic backend for other queries and another specialized backend for spreadsheets?
- The authors noticed that most of the time is spent thinking about how to best translate a question into UI operations rather than processing the operations themselves.

Clarifications

- No support for privacy
- The evaluation section did not include any results for scrolling through the spreadsheet
- The final vizketch is calculated by merging multiple summaries but won't that add up all the errors in the individual summary?

Limited Space: SOTA summaries using space s can provide error O(1/s)



Mergeable Summaries*: error preserved when combining summaries

Query error is 1/100 = 1%

Clarifications

More on mergeable summaries:

merging summaries is much less accurate than using larger summaries

15-minute Mergeable Summaries with s=100

k=966 segments

Query error 1/100 = 1%

Flexible but inaccurate

Ideal: one large summary with s=100*k

Query error $1/(sk) \approx .001\,\%$

Inflexible but accurate

Did Hillview have users?

- They have a pretty extensive user manual. Wonder if this was just a research project or if it was actually put into production and used in VMWare.
- I saw the demo video of the HillView tool and while it seemed performant in the demo, the interface itself did not seem very polished - primarily because the authors have implemented the frontend from scratch.

From authors of Hillview

Reasons why Hillview had no users:

- It's a Swiss army knife, but does not do any job particularly well
- Too generic for customers with specific needs
  - e.g., log browsing
- Lack of support for coding/scripting

Future work ideas

- I wonder whether we can precompute a set of mergeable sketches before the queries and use them for fast visualization of the data. Currently, it seems that caching some of the sketches helps, but should we actively compute and store some sketches?
- An abstraction on top of the existing datasets to mitigate the two limitations of partitioned data and static data.

Next class

Project proposal presentation!

- 12 groups (order according to canvas)
- 5min each

Learn about your classmate's ideas