Projects and Groups

Kexin Rong

CS197 10/3/19

Agenda

- Project overview
- Form group
- Assignment 2 and misc

What you can and can’t expect from me

Can:

- Answer your questions to the best of my knowledge and work with you to find answers
- Provide feedback and suggestions

Can’t

- Fix all your segfaults
- A step by step guide to solve your research project

What I expect from you

- Respect each other. Most sections will function like research group meetings. This mean you will share out your progress and get feedback from each other. For that to work you need to be listening
- This is a collaborative space and we are here to support each other. Research is not an independent endeavor. Share ideas and encouragement, but also accept criticism as constructive
- By the end of the quarter, you should know more about the project than I do

Your preferences

- Independence
  - 122333345
- Metadata
  - 111222225
- Design Sketches
  - 112334445
- Visualization
  - 344454455
- Hash
  - 111233555

#1 Independence Assumption in Real Life

CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies

- Scrape datasets
- Brute force correlation computation
- Implementing CORDS and extending to triplets, quadruplets etc.
- How much does improved statistics help with query optimization

Suggested reading:

- How Good Are Query Optimizers, Really?

#2 Answering Queries with Metadata

Implementing Data Cubes Efficiently

- Aggregates: SUM(), AVG(), COUNT(), COUNT(DISTINCT)
- Predicates: ranges, equalities and inequalities, regex
- Group by
- Joins

Suggested reading:

- Mergeable Summaries

#3 Designing Sketches in End-to-end Systems

Ray: A Distributed Framework for Emerging AI Applications

- Build a few applications in Ray
- Profile the object store
- Focus on cost of serialization/deserialization
    - https://arrow.apache.org/blog/2017/10/15/fast-python-serialization-with-ray-and-arrow/

Suggested reading:

- Making Sense of Performance in Data Analytics Frameworks
- Filter Before You Parse: Faster Analytics on Raw Data with Sparser

#5 Hash Table Bake off

[A Seven-Dimensional Analysis of Hashing Methods and its Implications on Query Processing]

- Reproduce part of the benchmark
- Try to make things continuous
  - e.g. with snapshots at 50%, 60% load factor, can I predict the performance at 55%?

Suggested reading:

- Modeling LSH for Performance Tuning
- The Case for Learned Index Structures

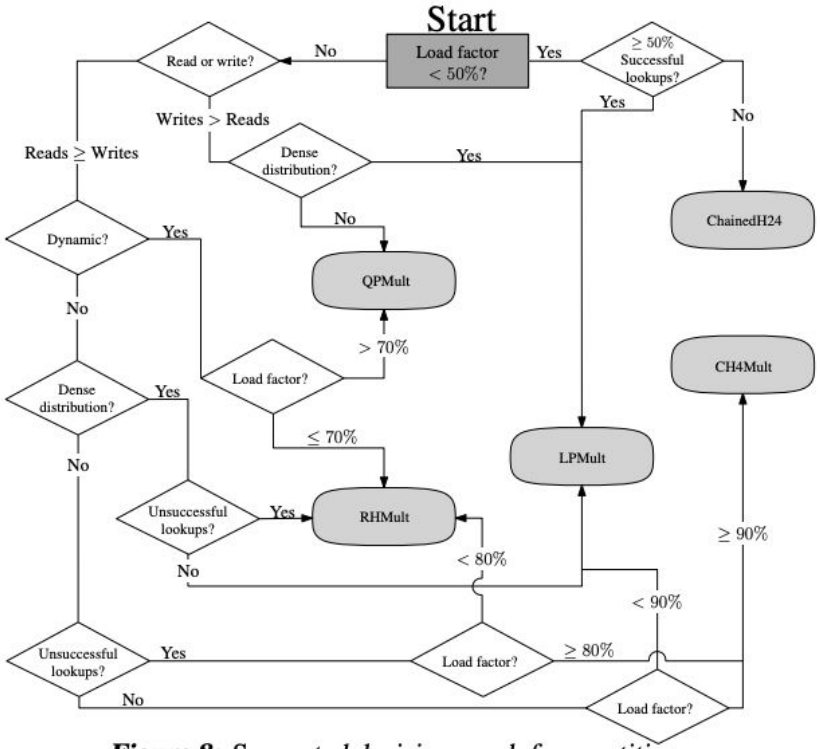

Figure 8: *Suggested decision graph for practitioners.*

Top venues

- Databases:
  - SIGMOD, VLDB
- Systems
  - OSDI, SOSP
  - ATC, Eurosys, SysML
- Networking/Distributed Systems
  - NSDI, SIGCOMM
- Data Mining
  - KDD, WWW
- Just in case: machine learning
  - ICLR, ICML, NeurIPS/NIPS

Some useful tools

- tmux
- htop
- jupyter notebook
- CLion, IntelliJ, PyCharm

Work your group

- Get to know each other
- Exchange contact information
- Find a time to work together
- Set up Github/Overleaf etc.
- Start talking about literature search and how to divide up the work