WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

CS 839: Topics in Database Management Systems

Lecture 4: Analytical Processing

Xiangyao Yu
9/18/2023

Update on the Schedule

Two lectures (instead of three) on “analytical processing with disaggregation”

“Transaction processing with disaggregation” starts next Monday (9/25) instead of next Wed

Presentation

Please pick your presentation slot early

If you choose a topic with many presenters, please try to add your own presentation material to the signup sheet

Discussion Question

For a multi-cloud analytical database where the data is stored across AWS, Azure, and GCP, is Snowflake architecture a good fit? What architecture would you choose in this scenario?

- Snowflake still works. VWs can talk to storage services in different clouds
- Each cloud is an AZ of storage -> very expensive!!
- Expensive to fetch data from a different cloud
- Computation pushdown to individual clouds, results combined in the virtual warehouse. Computer clusters coordinated like map-reduce.
- VW in the corresponding cloud to process local data -> cannot do cross-cloud join. Shared-nothing across clouds

Discussion Question

Snowflake is promoting the idea of data marketplace, where Snowflake users can share/trade their data and queries (think of App Store). What new applications can this enable in your opinion?

- Data does not need to be copied or collected again
- Reduce duplicate queries
- Selling ML data and model building
- Query statistics on sensitive databases
- Data science and data analytics in healthcare

PushdownDB and Spectrum – Q/A

What was Amazon's motivation for building S3 Select?

How does Amazon Redshift handle concurrent queries, especially when they involve both Redshift tables and S3 data?

Existing work to support arbitrary data sources for a query engine?

Discussion

In the extreme, all computation can be pushed down to the storage, making the system shared-nothing again. Then, we lose the benefits of disaggregation. How do we strike the right balance between disaggregation and computation pushdown? Are there principles that computation pushdown must follow such that the system retain all the benefits of disaggregation?

Please submit your discussion to hotcrp **as a new submission** by the end of Tuesday (9/19)

- Title starts with “[Discussion L4]”
- Set authors properly