CS 5151 / 6051
Database Theory (DBT)

Group Project

Seokki Lee

Group Project

- Studying various systems in database field
- Objectives
  - To study recent techniques on multiple topics
  - To learn and share knowledge of the (research) works and systems

Topics

- (Big) Data management
  - Approximate Query Processing (AQP)
  - Incremental (view) maintenance
  - Analytics and data warehouse
- Database Management System (DBMS) in various domains
  - Machine Learning (ML)
  - Data visualization with DBMS
  - Blockchain database
- (Big) Data Provenance
- Datalog

Research in (Big) Data Management

- Approximate Query Processing (AQP)
  - DeepDB
    - Literature review
      - DeepDB: Learn from Data, not from Queries!
    - Resource
      - https://github.com/DataManagementLab/deepdb-public

Research in (Big) Data Management

- Incremental (view) maintenance
  - DBSP
    - Literature review
      - DBSP: Automatic Incremental View Maintenance for Rich Query Languages
      - LINVIEW: incremental view maintenance for complex analytical queries
    - Resource
      - https://github.com/vmware/database-stream-processor

Research in (Big) Data Management

- Analytics and Data Warehouse
  - Spark SQL
    - Literature review
      - Spark sql: Relational data processing in spark
      - Integration of Skyline Queries into Spark SQL
      - ...
    - Resource
      - https://spark.apache.org/sql/
      - https://github.com/Lukas-Grasmann/skyline-queries-spark

Research in (Big) Data Management

- Analytics and Data Warehouse
  - Snowflake
    - Literature review
      - The snowflake elastic data warehouse
      - An improved join-free schema for ETL and OLAP of data warehouse
      - …
    - Resource
      - https://www.snowflake.com/en/

Database Research in Various Domains

- ML on big data systems
  - SystemDS
    - Literature review
      - SystemDS: A declarative machine learning system for the end-to-end data science lifecycle
      - Loupe: A Visualization Tool for High-Level Execution Plans in SystemDS
    - Resource
      - https://github.com/apache/systemds
      - https://github.com/tugraz-isds/systemds

Database Research in Various Domains

- ML on big data systems
    - VolcanoML
        - Literature review
            - VolcanoML: speeding up end-to-end AutoML via scalable search space decomposition
            - Efficient End-to-End AutoML via Scalable Search Space Decomposition
        - Resource
            - https://github.com/VolcanoML

Database Research in Various Domains

- Data Visualization with DBMS
    - Interactive visualization interfaces
        - Literature review
            - PI2: End-to-end Interactive Visualization Interface Generation from Queries
            - NL2INTERFACE: Interactive Visualization Interface Generation from Natural Language Queries
        - Resource
            - https://github.com/learnedinterfaces/PI2

Database Research in Various Domains

- Blockchain Database
  - MongoDB
    - Literature review
      - Trends in Development of Databases and Blockchain
      - Databases fit for blockchain technology: A complete overview
    - Resource
      - https://www.mongodb.com/databases/blockchain-database

Research in (Big) Data Provenance

- BreadCrumb
  - Literature review
    - To not miss the forest for the trees-A holistic approach for explaining missing answers over nested data
    - Debugging missing answers for spark queries over nested data with breadcrumb
  - Resource
    - https://github.com/UniStuttgart-DataEngineering/breadcrumb

Research in (Big) Data Provenance

- Titian
  - Literature review
    - Titian: Data provenance support in spark
    - Adding data provenance support to Apache Spark
  - Resource
    - https://github.com/maligulzar/bigdebug/blob/titian-2.1/vldb2016-p301-interlandi.pdf
    - https://github.com/maligulzar/bigdebug

Research in Datalog

- RaDlog (Former BigDatalog)
  - Literature review
    - Formal semantics and high performance in declarative machine learning using Datalog
    - Big Data Analytics with Datalog Queries on Spark
    - http://wis.cs.ucla.edu/deals/
  - Resource
    - https://github.com/radlog-web/radlog
    - https://github.com/ashkapsky/BigDatalog

Research in Datalog

- Souffle
  - Literature review
    - https://souffle-lang.github.io/publications
  - Resource
    - https://github.com/souffle-lang/souffle

Workload

- **Select a system** from the list
  - Maximally 2 teams for each system
  - Fill out the excel sheet (CS5151-6051-Fall2023-Formed-Group-Info)

- Study the system
  - Using the resources available on **google scholar** and **search**
  - Understanding the research work
    - Motivations and contributions
    - Technical details / Novelty of their approaches
    - Limitations
    - Experimental evaluation
    - …

Workload

- Setting up and **running the system**
  - On your local
  - Cloud available for free if possible / necessary

- Prepare **presentation**
  - Theoretical / Algorithmic aspect
    - Sharing the system and research work to the class
    - What you have learned
  - Demo
    - Functionality of the system
    - Showcase 2 scenarios to the class
      - 1 simple + 1 complex

Workload

- 1st report (1 page)
  - Due by **Oct 1st**
  - Status of preparation for the project
    - What is the reference you focus on?
    - Understanding of the system and research work
  - Scenarios for the demo
  - Optional: experiment plan

- 2nd report (2 pages)
  - Due by **Oct 31th**
  - Summary of research work including technical details
  - The system must be installed.
  - Running scenarios
  - Optional: running experiment(s)

Format:

- Arial 11pt
- 1 inch margin

Schedule

- Uploading slides for the presentation
  - Due by **Nov 17 at 11:59 pm**
  - Submission to Canvas under assignment

- Presentation (10 mins) + Q&A (2 ~ 3 mins)
  - Scheduling poll will be released by **Nov 2**
  - Selection due by **Nov 7**
  - On **Nov 19, 21, 26,** and **Dec 3**

Grading

- Reports (10%)
  - Report 1 (5%)
  - Report 2 (5%)
- Presentation (20%)
  - Literature review
    - Outline the system
    - Lesson learned
  - Demo
    - Set up properly
    - Functionality the systems
    - Running a simple and complex scenarios

Grading

- Bonus points
  - **Reproducing** an existing experiment
  - Providing a **new** experiment
  - ...

- **Late** policy
  - **-10%** per day
  - No exception (except health issue)