University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, July 22, 2021

MSc Thesis
Implementing Learned Cardinality Estimation in a Database Systems Context

Query optimization in database management systems heavily relies on the cardinalities of result relations. For instance, to choose between different query plans the database system must accurately estimate the cardinalities of intermediate results. Cardinality estimation in most database systems is based on the assumption that attributes are independent, which rarely holds in reality and easily leads to estimations that are orders of magnitude off.

In order to provide more accurate estimations, recent efforts have been trying to apply learned models to replace the traditional cardinality estimation component in database systems. For example, Kipf et al. [4] proposed to apply multi-set convolutional neural networks to estimate the cardinalities whereas Sun et al. [6] proposed to use segmentation to get a better estimation. Although these research works have shown their advantages, they have not been integrated into existing database systems and have not been evaluated in production environments.

The goal of this master thesis is to implement a learned cardinality estimation method and integrate it into, e.g., the PostgreSQL database system. After the integration, an evaluation between the learned cardinality estimation and the built-in estimation inside PostgreSQL should be performed.

Tasks

1. **Task 1: Literature review and prerequisite study**
    - Study the relevant research work on learned cardinality estimation, including but

not limited to [2], [4], [6], and [7].

2. **Task 2: Implement cardinality estimation with base tables only**
   - Based on the content of base tables, implement the learned cardinality estimation approach with linear regression and possibly other machine learning techniques, e.g., neural networks.
   - The implemented approach should only depend on the content of the table. Thus, the estimation is based on the data and not on any queries.
   - Test the results with selected queries, for example:

     ```
     SELECT x, y FROM test WHERE x<11 AND y<21;
     ```

     where table `test` is populated by

     ```
     INSERT INTO test
     SELECT generate_series(1,100), generate_series(11,110);
     ```

3. **Task 3: Implement cardinality estimation with workload**
   - Based on workload information, implement the learned cardinality estimation described by Hilprecht and Binnig [3] for selection and join queries, such as:

     ```
     SELECT x, y, m FROM test INNER JOIN test2 on x>m and y<m;
     ```

   - In contrast to Task 2, learned cardinality estimation with workload information shall also learn from workload information, i.e., the current query, previous queries and possibly system resources. Thus, the implemented approach shall not only utilise the content of the table but also the queries and other useful information.

4. **Task 4: System Implementation**
   - Investigate the feasibility of integrating learned cardinality estimation into PostgreSQL or another suitable system.
   - Integrate the learned cardinality estimation with base tables that was implemented in the Task 2 into a system context.
   - Integrate the learned cardinality estimation with workload information that was implemented in the Task 3 into a system context.

5. **Task 5: Evaluate different approaches on synthetic and real-world datasets**
   - Evaluate different cardinality estimation approaches with synthetic datasets, e.g., the dataset generated with `generate_series(1, 100)`.
   - Use four different real-world datasets (Census, Forest, Power [1] and DMV[5]) to evaluate the learned cardinality estimation with the built-in estimation technique in PostgreSQL.
   - Analyse the training time of learned cardinality estimation and the running time

of different cardinality estimation techniques. Identify their advantages, disadvantages and limitations.

6. **Task 6: Write and defend the thesis**
    - Describe the implementations, results and evaluations in your Master's thesis.
    - Present and defend your Master's thesis in the DBTG group meeting.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[2] Rojeh Hayek and Oded Shmueli. Improved cardinality estimation by learning queries containment rates. In Angela Bonifati, Yongluan Zhou, Marcos Antonio Vaz Salles, Alexander Böhm, Dan Olteanu, George H. L. Fletcher, Arijit Khan, and Bin Yang, editors, *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02, 2020*, pages 157–168. OpenProceedings.org, 2020. doi:10.5441/002/edbt.2020.15.

[3] Benjamin Hilprecht and Carsten Binnig. One model to rule them all: Towards zero-shot learning for databases. *CoRR*, abs/2105.00642, 2021. URL: https://arxiv.org/abs/2105.00642.

[4] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter A. Boncz, and Alfons Kemper. Learned cardinalities: Estimating correlated joins with deep learning. In *9th Biennial Conference on Innovative Data Systems Research, CIDR 2019, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings*. www.cidrdb.org, 2019. URL: http://cidrdb.org/cidr2019/papers/p101-kipf-cidr19.pdf.

[5] Open NY. Vehicle, snowmobile, and boat registration, 2017. URL: https://data.ny.gov/Transportation/Vehicle-Snowmobile-and-Boat-Registrations/w4pv-hbkt.

[6] Ji Sun, Guoliang Li, and Nan Tang. Learned cardinality estimation for similarity queries. In Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 1745–1757. ACM, 2021. doi:10.1145/3448016.3452790.

[7] Xiaoying Wang, Changbo Qu, Weiyuan Wu, Jiannan Wang, and Qingqing Zhou. Are we ready for learned cardinality estimation? *Proc. VLDB Endow.*, 14(9):1640–1654, 2021. URL: http://www.vldb.org/pvldb/vol14/p1640-wang.pdf.

Supervisors:

- Prof. Dr. Michael Böhlen (boehlen@ifi.uzh.ch)
- Prof. Dr. Anton Dignös (adignoes@ifi.uzh.ch)

- Qing Chen (qing@ifi.uzh.ch)

Start Date: July 15th, 2021

End Date: January 14th, 2022

University of Zurich
Department of Informatics

Prof. Dr. Michael Böhlen
Professor