

## 作业说明 (35 分)

一、作业描述：给定类似如下的 SQL 查询和对应的查询计划，使用机器学习算法估算其查询规模 (cardinality)

二、所需软件 (仅推荐)

1. Anaconda
2. Pytorch/TensorFlow
3. Sklearn, numpy, pandas

三、数据说明

1. 提供训练样本 60000 条，待预测样本 1070 条
2. 需要大家提供待预测样本的预测结果，并以如下方式提供：

**命名要求：** 预测结果\_姓名\_学号.csv (如 预测结果\_张三\_2021000000.csv) 提交首行需要包括对应的列名如下。

```
Query ID,Predicted Cardinality
0,0
1,0
2,2
3,6
4,12
5,20
```

```
query: "SELECT * FROM movie_companies mc,title t,movie_info_idx mi_idx WHERE t.id=mc.movie_id AND t.id=mi_idx.movie_id AND mi_idx.info_type_id=112 AND mc.company_type_id=2;"
query_id: 0;
explain_result: [{"Plan": [{"Node Type": "NestLoop", "Parallel Aware": false, "Async Capable": false, "Join Type": "Inner", "Startup Cost": 1.29, "Total Cost": 1163.84,
"Plan Rows": 49, "Plan Width": 168, "Actual Startup Time": null, "Actual Total Time": null, "Actual Rows": null, "Actual Loops": null, "Inner Unique": false, "Plans": [{"Node Type": "NestLoop",
"Parent Relationship": "Outer", "Parallel Aware": false, "Async Capable": false, "Join Type": "Inner", "Startup Cost": 0.86, "Total Cost": 1168.84, "Plan Rows": 92, "Plan Width": 128,
"Actual Startup Time": null, "Actual Total Time": null, "Actual Rows": null, "Actual Loops": null, "Inner Unique": true, "Plans": [{"Node Type": "Index Scan", "Parent Relationship": "Outer",
"Parallel Aware": false, "Async Capable": false, "Scan Direction": "Forward", "Index Name": "info_type_id_movie_info_idx", "Relation Name": "movie_info_idx", "Alias": "mi_idx", "Startup Cost":
0.43, "Total Cost": 131.67, "Plan Rows": 92, "Plan Width": 26, "Actual Startup Time": null, "Actual Total Time": null, "Actual Rows": null, "Actual Loops": null, "Index Cond": "(info_type_id = 112)"
}, {"Rows Removed by Index Recheck": 0}, {"Node Type": "Index Scan", "Parent Relationship": "Inner", "Parallel Aware": false, "Async Capable": false, "Scan Direction": "Forward", "Index Name":
"title_play", "Relation Name": "title", "Alias": "t", "Startup Cost": 0.43, "Total Cost": 0.45, "Plan Rows": 1, "Plan Width": 94, "Actual Startup Time": null, "Actual Total Time": null,
"Actual Rows": null, "Actual Loops": null, "Index Cond": "(id = mi_idx.movie_id)", "Rows Removed by Index Recheck": 0}], [{"Node Type": "Index Scan", "Parent Relationship": "Inner", "Parallel
Aware": false, "Async Capable": false, "Scan Direction": "Forward", "Index Name": "movie_id_movie_companies", "Relation Name": "movie_companies", "Alias": "mc", "Startup Cost": 8.43, "Total
Cost": 0.97, "Plan Rows": 3, "Plan Width": 40, "Actual Startup Time": null, "Actual Total Time": null, "Actual Rows": null, "Actual Loops": null, "Index Cond": "(movie_id = t.id)", "Rows Removed
by Index Recheck": 0}, {"Filter": "(company_type_id = 2)", "Rows Removed by Filter": 381111}]]}]
```

3. 数据格式：

提供 JSON 格式的训练和测试数据，query 表示对应的 SQL 查询，query\_id 表示查询的 id，explain\_result 表示查询计划，可以用 Python 中的 json 包处理 (json.loads)

4. column\_min\_max\_vals.csv 表示数据库各个列的最大值、最小值、基数和列所包含不同值的数量。

四、作业提交

1. 提交内容：

- a) 实验报告：实现代码及具体实现报告 (不超过 10 页)。
- b) 源代码。
- c) 预测结果文件。
- d) 验收 PPT

2. 提交地点：报告、代码及最优结果提交到课堂派

3. 提交截止时间点：2025 年 5 月 29 日 24 点 (晚一天扣 1 分，本次作业分值扣完为止)

4. 预计验收日期：2025 年 5 月 30 日上课时间

五、模型评测

大家可以通过 [10.77.110.133:19052](http://10.77.110.133:19052) 评测自己的预测效果 (需内网访问)

在网站评测自己的结果时，**务必**每次提交都使用同样的文件名 (预测结果\_姓名\_学号.csv)，需要大家自己记录每次提交的不同结果，否则后台文件太多会影响大家评测效率。