NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

TED-Learn: Towards Technology-Enabled Database Education

Sourav S Bhowmick

School of Computer Science & Engg

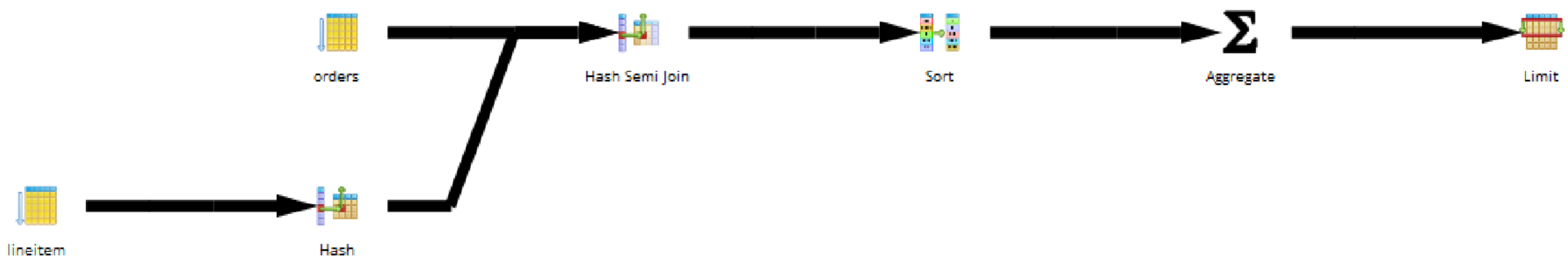
Nanyang Technological Univ

Singapore

Query Optimization in RDBMS

```
select o_orderpriority, count(*) as order_count from orders where
o_orderdate >= date '1996-03-01' and o_orderdate < date '1996-03-01'
+ interval '3' month and exists (select * from lineitem where
l_orderkey = o_orderkey and l_commitdate < l_receiptdate) group by
o_orderpriority order by o_orderpriority limit 1;
```

	QUERY PLAN text
1	Limit (cost=283240.33..283261.81 rows=1 width=24) (actual time=2517.131..2517.131 rows=1 loops=1)
2	-> GroupAggregate (cost=283240.33..283347.75 rows=5 width=24) (actual time=2517.129..2517.129 rows=1 loops=1)
11	Rows Removed by Filter: 1443054
12	-> Hash (cost=187509.83..187509.83 rows=2000182 width=4) (actual time=1733.473..1733.473 rows=3793295 loops=1)
13	Buckets: 131072 (originally 131072) Batches: 64 (originally 32) Memory Usage: 3098kB
14	-> Seq Scan on lineitem (cost=0.00..187509.83 rows=2000182 width=4) (actual time=0.037..1246.522 rows=3793295 loops=1)
15	Filter: (l_commitdate < l_receiptdate)
16	Rows Removed by Filter: 2207919
17	Planning time: 4.656 ms
18	Execution time: 2517.741 ms



Circa 2017

Student: I've been trying to understand the QEPs in X but it's really hard to follow….the descriptions are very different from what we learn from textbooks and lecture slides ☹

Me: Yeah! DBMS vendors use vendor-specific implementation and language. You may refer to their manuals for details...

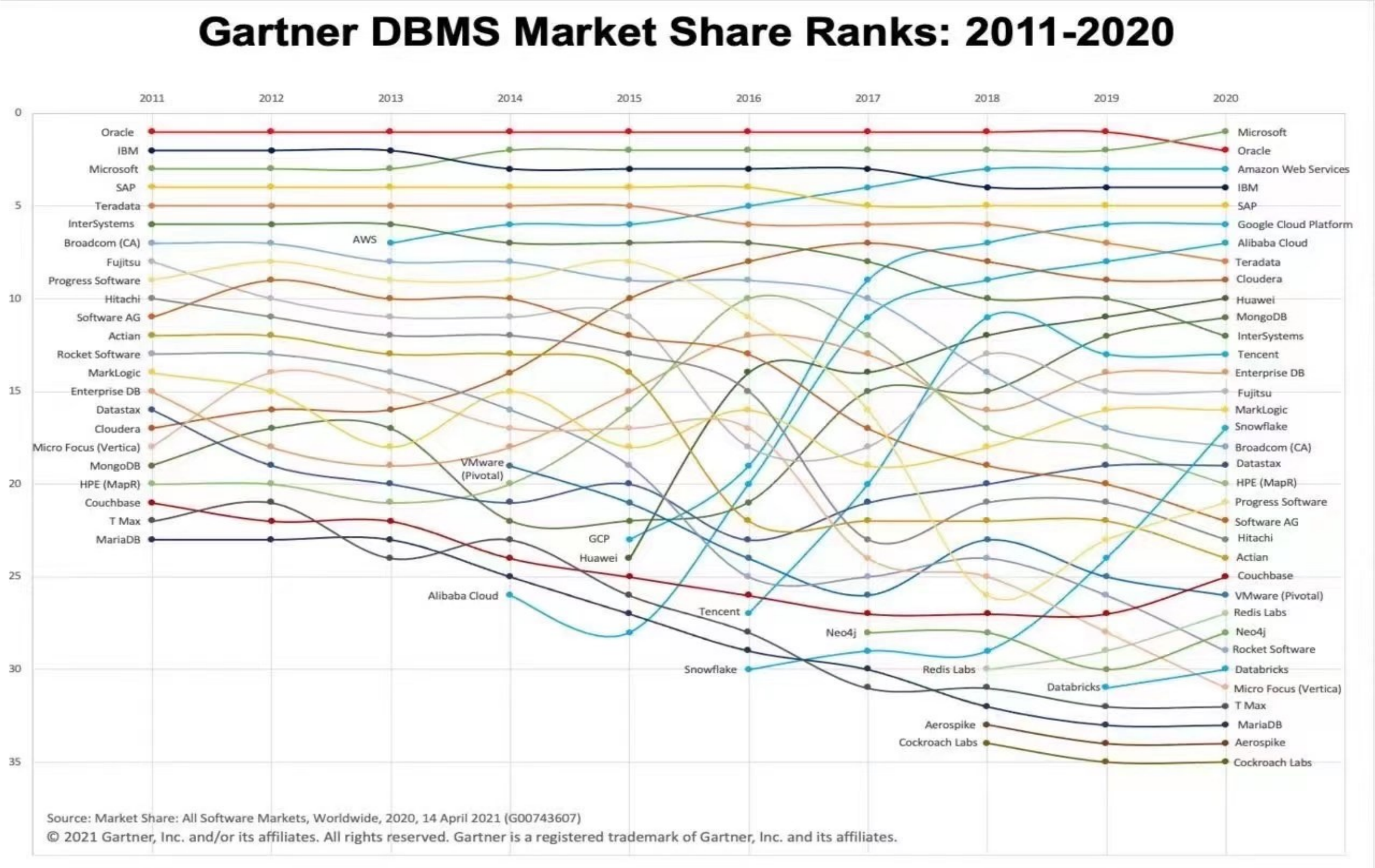
Student: It's boring to read manuals! In any case, for different DBMS I have to peruse different manuals…it's such a waste of time!

Me: Well, existing RDBMS are designed primarily for enterprises and not for students…. ☺

Student: Can you please make things easy for us to learn? I am really struggling to understand query plans….

Hit Like A Brick Through the Window

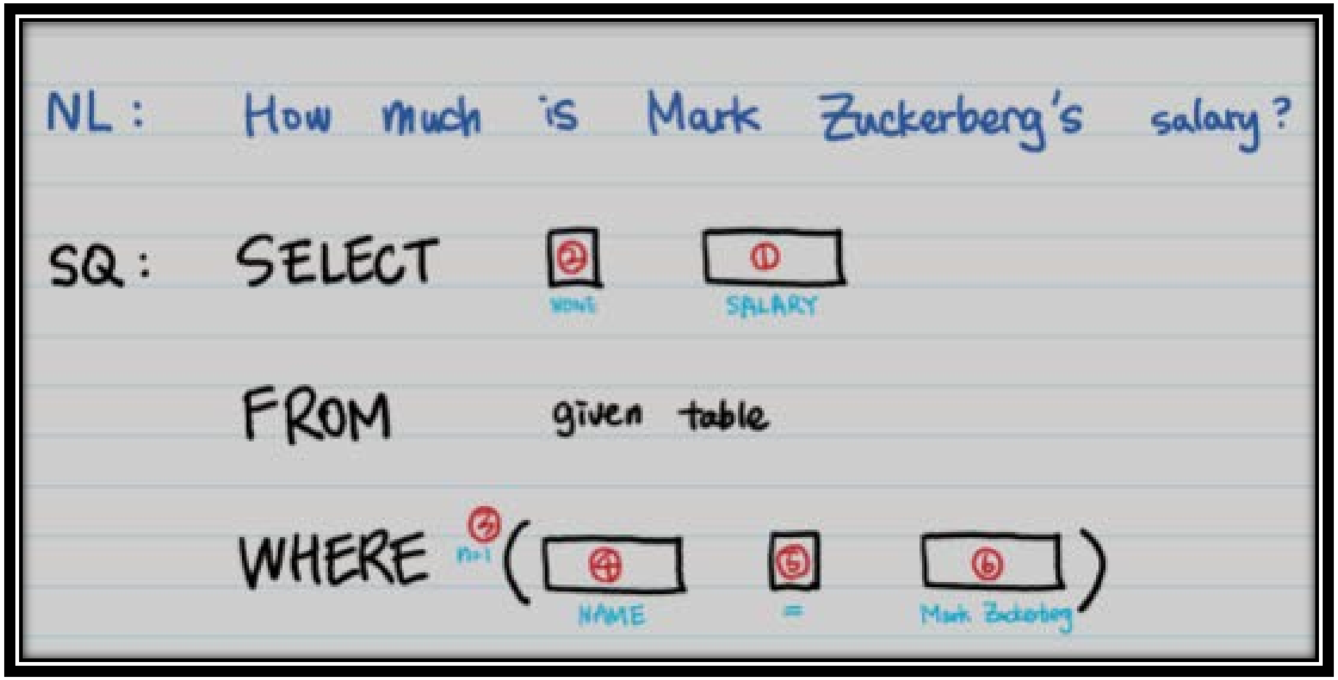
Database Research for Enterprises



Key Question

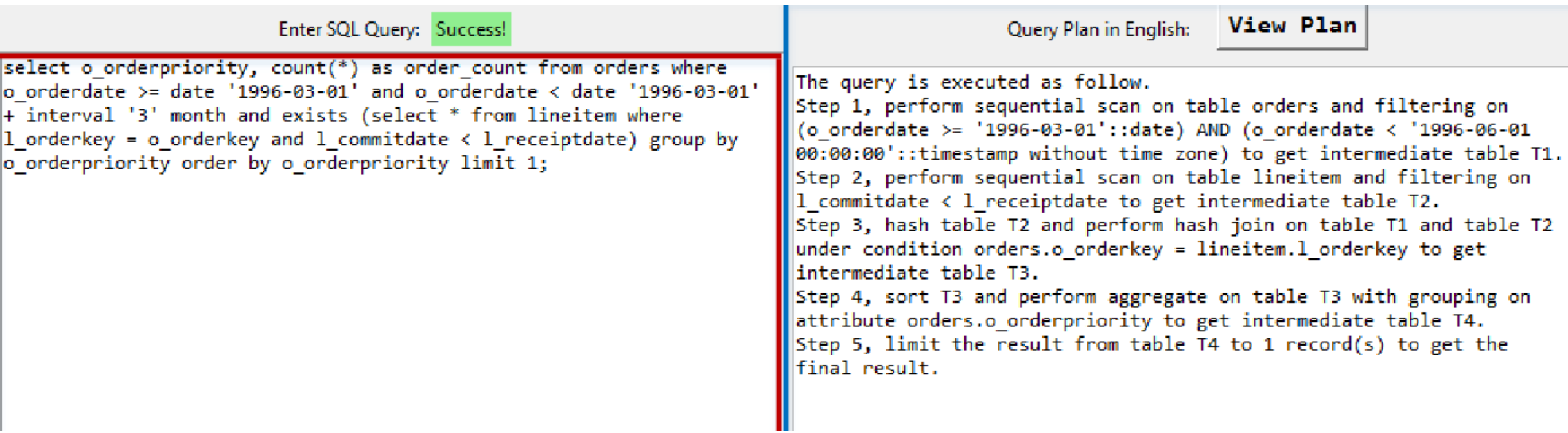
Can we describe a QEP using natural language to enhance DB education?

Juxtapose Ideas That Often Don’t Go Together



https://towardsdatascience.com/text-to-sql-learning-to-query-tables-with-natural-language-7d714e60a70d

Juxtapose Ideas That Often Don’t Go Together

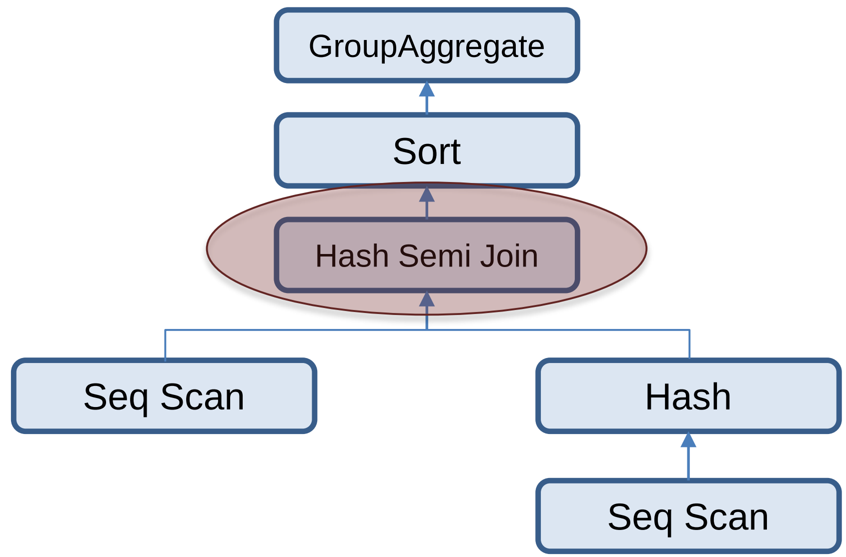


Overview of NEURON [SIGMOD 2019]

NEURON

- Rule-based natural language description
- Build on top of PostgreSQL

```
select o_orderpriority, count(*) as
order_count
from orders
where
        o_totalprice > 100
        and exists (
                select *
                from lineitem
                where
                        l_orderkey = o_orderkey
                        and l_extendedprice > 100
        )
group by o_orderpriority
order by o_orderpriority;
```

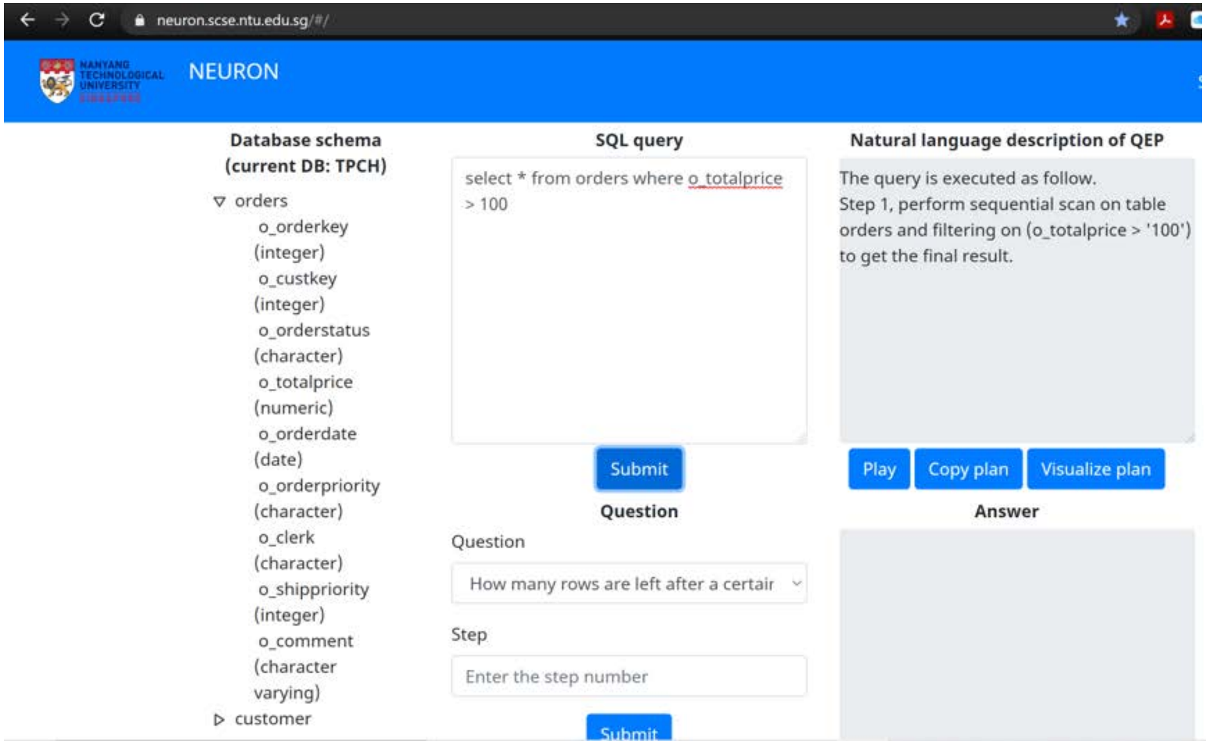


RULE TEMPLATE

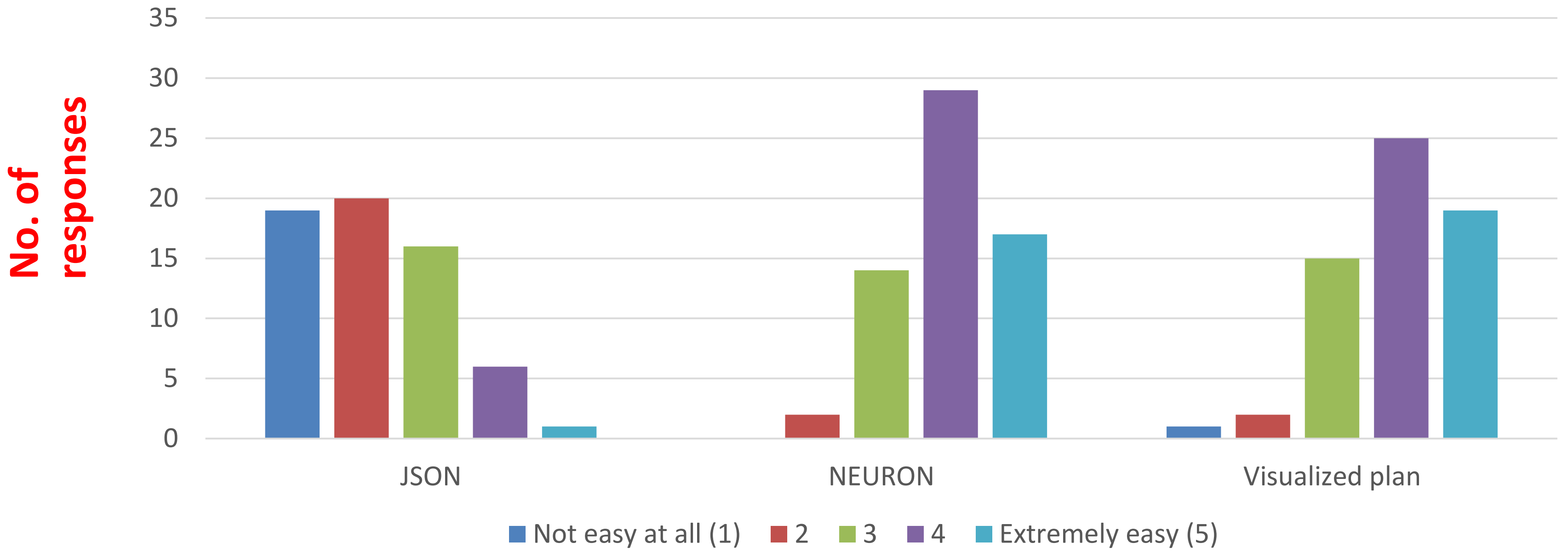
hash table <T> and perform hash semi join on table tablename and table <T> under condition (<C>) to get intermediate table <TN> .

NEURON v1.0

https://neuron.scse.ntu.edu.sg/#/

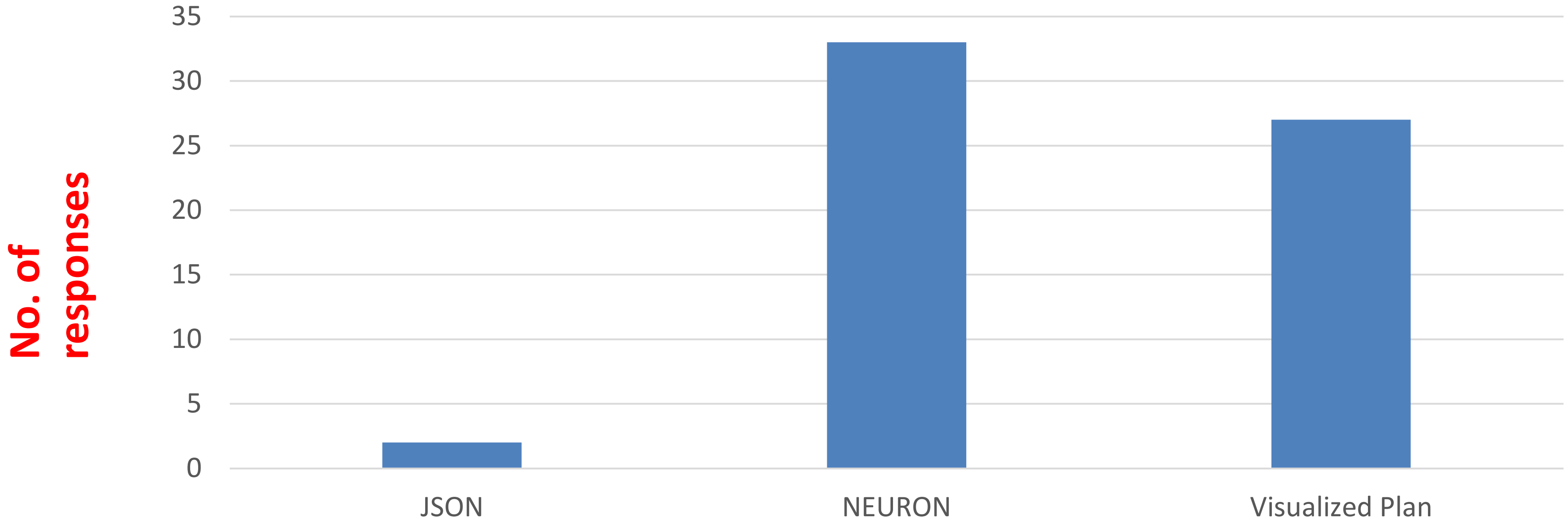


User Feedback: How easy is it to understand the query plan presented in various formats?



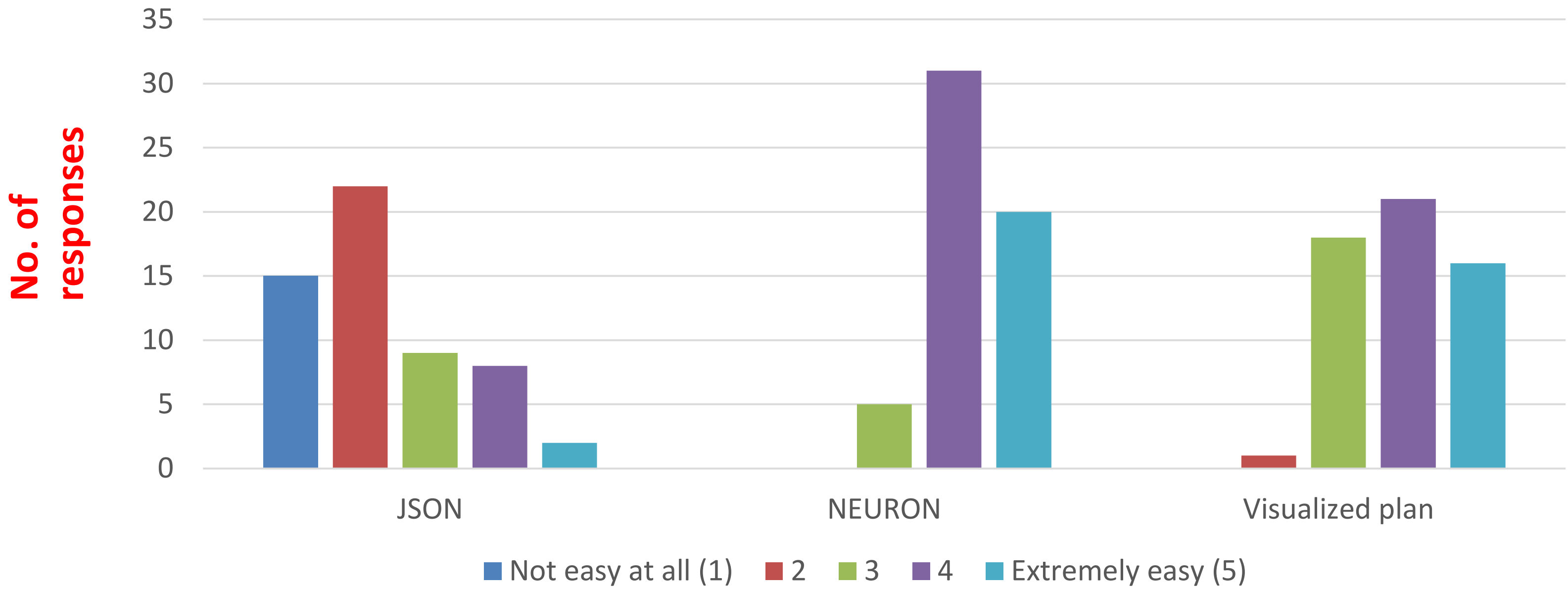
- 62 students volunteered for the survey (Oct 2019)
- NEURON is the easiest format (74.2%) to understand.
- Visualized plan (71%) is comparable with NEURON (74.2%).
- Majority (62.9%) of respondents found JSON format difficult to understand.

User Feedback: Which query plan format is most preferred?



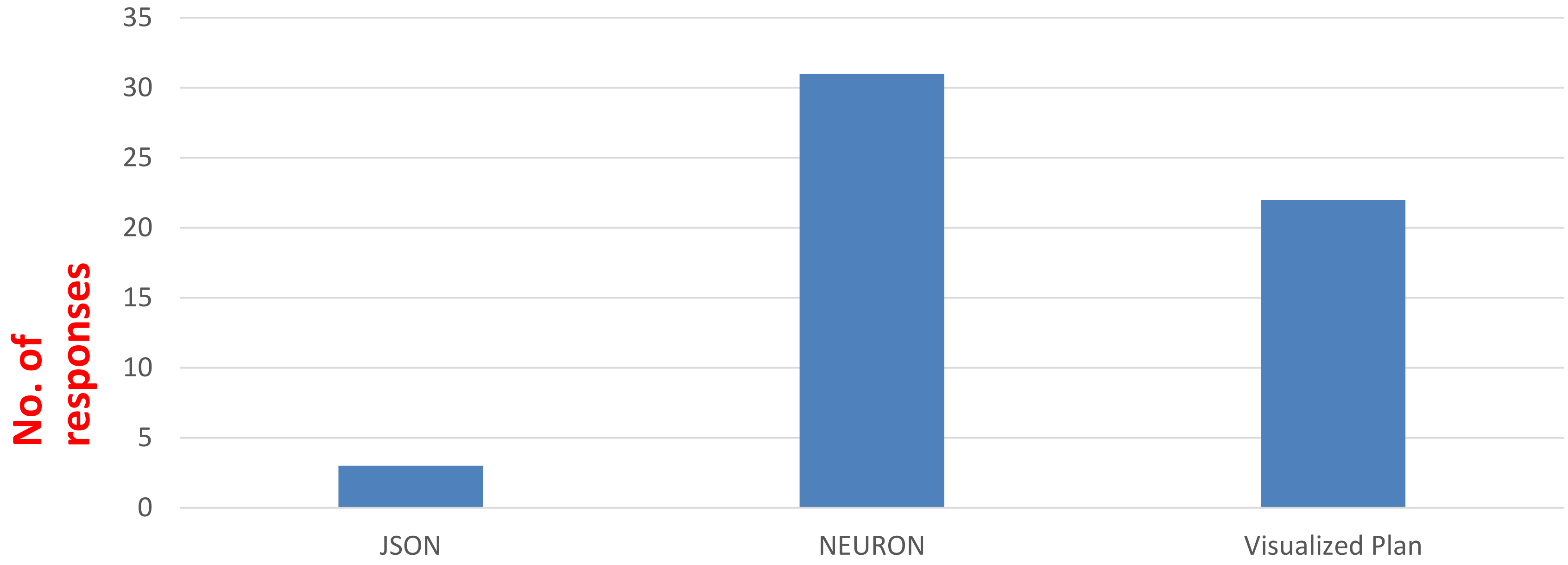
- The survey participants preferred NEURON the most (53.2%)
- Very few participants (3%) chose JSON as the most preferred choice.

User Feedback: How easy is it to understand the query plan presented in various formats?



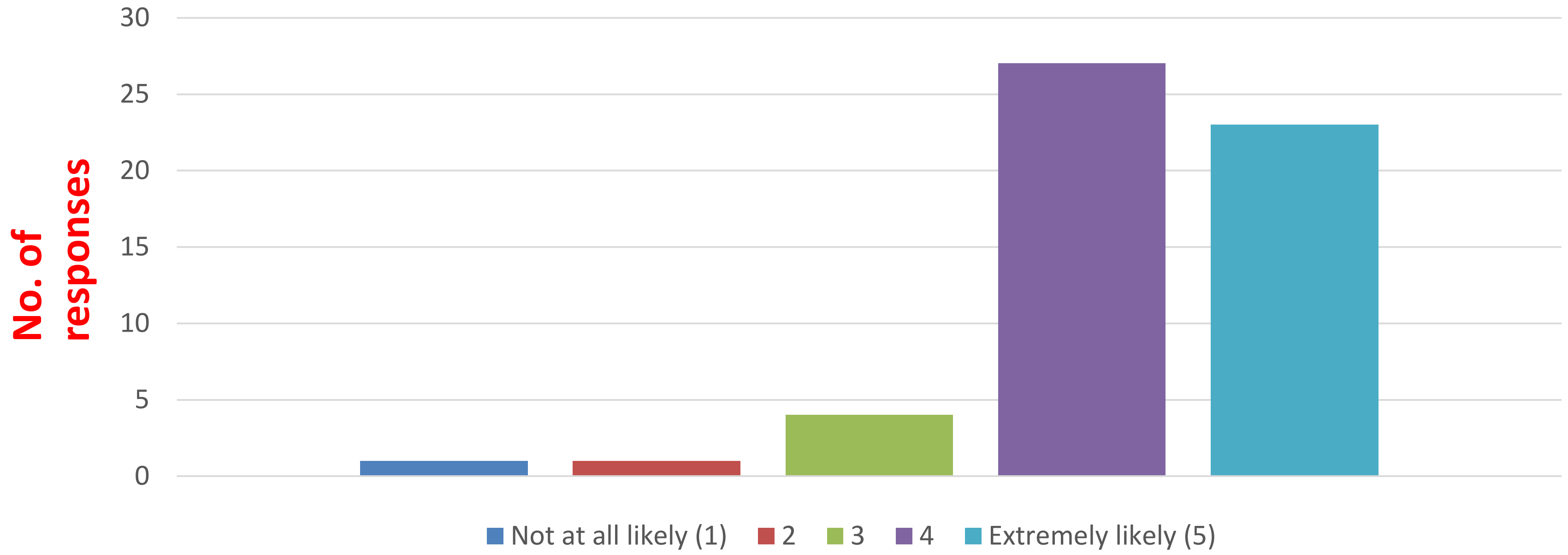
- 56 students volunteered for the survey (Oct 2020)
- NEURON is the easiest format (91.1%) to understand.
- Majority (66.1%) of respondents found JSON format difficult to understand.
- In term of ease of understanding (ranking of 4 or 5):
  NEURON (91.1%) > Visualized plan (66.1%) > JSON (17.9%)

User Feedback: Which query plan format is most preferred?



- The survey participants preferred NEURON the most (55.4%)
- Very few participants (5.4%) chose JSON as the most preferred choice.

User Feedback: How likely are you to recommend NEURON to a course mate?



- 89.3% of respondents are quite likely to recommend NEURON to a course mate.

User Feedback

"I use SQL Server. I can't use your tool! It only works on PostgreSQL!

"After running few queries and reading the descriptions, I feel bored and skip sentences as the language is repetitive!

"The natural language translation can be improved to summarize complex conditions of the QEP

Issues

Issues

- How do we generalize NEURON across different RDBMS?
- How can we alleviate boredom?

Challenges

- Different RDBMS have physical operators with different names
- Rule-based algorithms naturally generates similar text descriptions.

LANTERN [SIGMOD 2021]

Key Idea

- Instead of mapping an entire QEP to its NL description, map the set of physical operators in a RDBMS to corresponding NL descriptions.
- Stitch them together to generate the description of a specific QEP.
- More manageable to label physical operators.
- Enables generalization to handle any application-specific database.
- Orthogonal to the complexities of SQL queries.

Two Variants

- RULE-LANTERN
- NEURAL-LANTERN

Towards Enhancing Database Education: Natural Language Generation Meets Query Execution Plans. Weiguo Wang, Sourav S Bhowmick, Hui Li, Siyuan Li, Shaq Joty, Peng Chen. In SIGMOD, 2021.

RULE-LANTERN

Labeling Physical Operators

A declarative framework where a subject matter expert (SME) can create and manipulate the labels using a declarative language called POOL (Physical Operator Object Language).

POEM Data Model

- A simple and flexible graph model where all entities are objects.
- Each object -> a physical operator of a relational query engine.
- Objects are either atomic or complex having attribute-value pairs.
- source, name, alias, defn, desc, type, cond, and target.
- Values of all attributes are from the atomic type string.

RULE-LANTERN: POOL

```
CREATE POPERATOR hashjoin FOR pg
(ALIAS = null,
TYPE = 'binary',
DEFN = null,
DESC = 'perform hash join',
COND = 'true',
TARGET = null)
```

Compose Operator

- Specify generation of an NL description template of an operator.
- Uses the desc, type, and cond attributes of operators to generate.

POperators(<u>oid</u>, source, name, alias, type, defn,cond, targetid)
PDesc(<u>oid, desc</u>)

POEM Store

```
COMPOSE     hash
FROM        pg
```

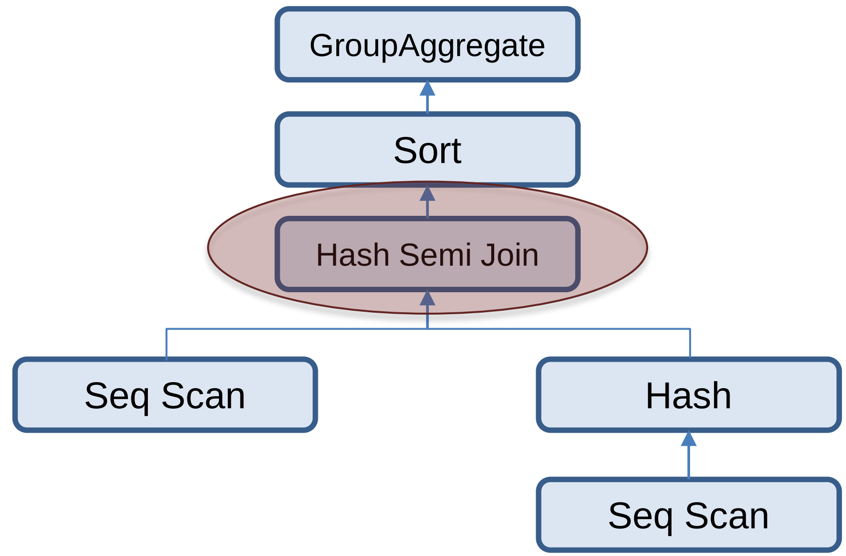
"hash $R1$"

```
COMPOSE     hash, hashjoin
FROM        pg
USING       hashjoin.desc = 'perform hash join'
```

"hash $R1$ and perform hash join on $R2$ and $R1$ on condition $cond$".

RULE-LANTERN Algorithm

```
select o_orderpriority, count(*) as
order_count
from orders
where
        o_totalprice > 100
        and exists (
                select *
                from lineitem
                where
                        l_orderkey = o_orderkey
                        and l_extendedprice > 100
        )
group by o_orderpriority
order by o_orderpriority;
```



RULE TEMPLATE

hash table <T> and perform hash semi join on table tablename and table <T> under condition (<C>) to get intermediate table <TN> .

User Feedback (Revisited)

"After running few queries and reading the descriptions, I feel bored and skip sentences as the language is repetitive!

NEURAL-LANTERN

Leverage DL

- Regard a QEP as an input language while the NL description as the output.
- Interpreting QEP into NL can be viewed as a machine translation task.

Challenges

- DL-based techniques need massive training sets of labeled examples to learn from.
- Prohibitively expensive as they demand database experts to translate thousands of QEPs.
- The platform needs to be generalizable and application domain-independent for ease of deployment and usage.

NEURAL-LANTERN: Training Data

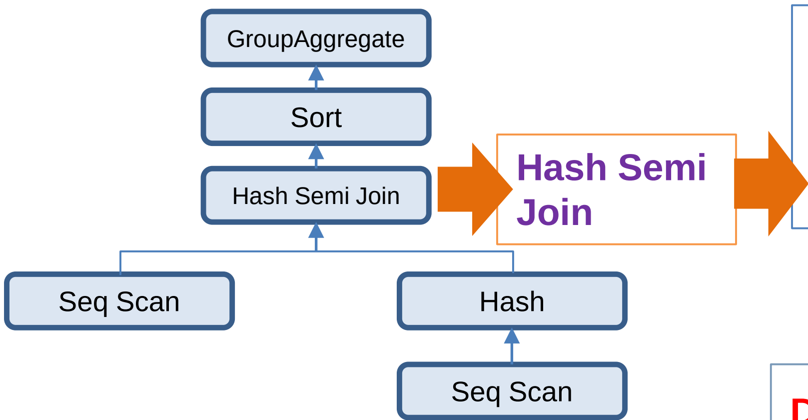
Training Data Generation

- We adopt Kipf et al. [CIDR 2019] to generate a set of SQL queries given a particular schema and database instance.
- A collection of QEPs corresponding to these queries.
- Each QEP is decomposed into a set of acts, each of which corresponds to a set of operators in an operator tree (subtree).
- For each act-> RULE-LANTERN to generate NL description.

Diversifying Text

- For each RULE-LANTERN result, we apply three state-of-the-art synonymous sentence generation tools and acquire a set of synonymous sentences.
- Remove duplicates and manually eliminate invalid sentences.

Example



List of special tags used in the output

Tag	Description	Example
<I>	indexed column name	
<F>	filtering condition	*c_mktsegment* = 'BUILDING'
<C>	join condition	*c_custkey* = *o_custkey*
<T>	an existing temporary table name	
<TN>	new temporary table name	
<A>	column name for sort	order by *revenue* desc ...
<G>	column name for groupby	group by *l_orderkey* ...

Training Data Generation

RULE-LANTERN:

hash table <T> and perform hash semi join on table tablename and table <T> under condition (<C>) to get intermediate table <TN> .

Diversifying Translation:

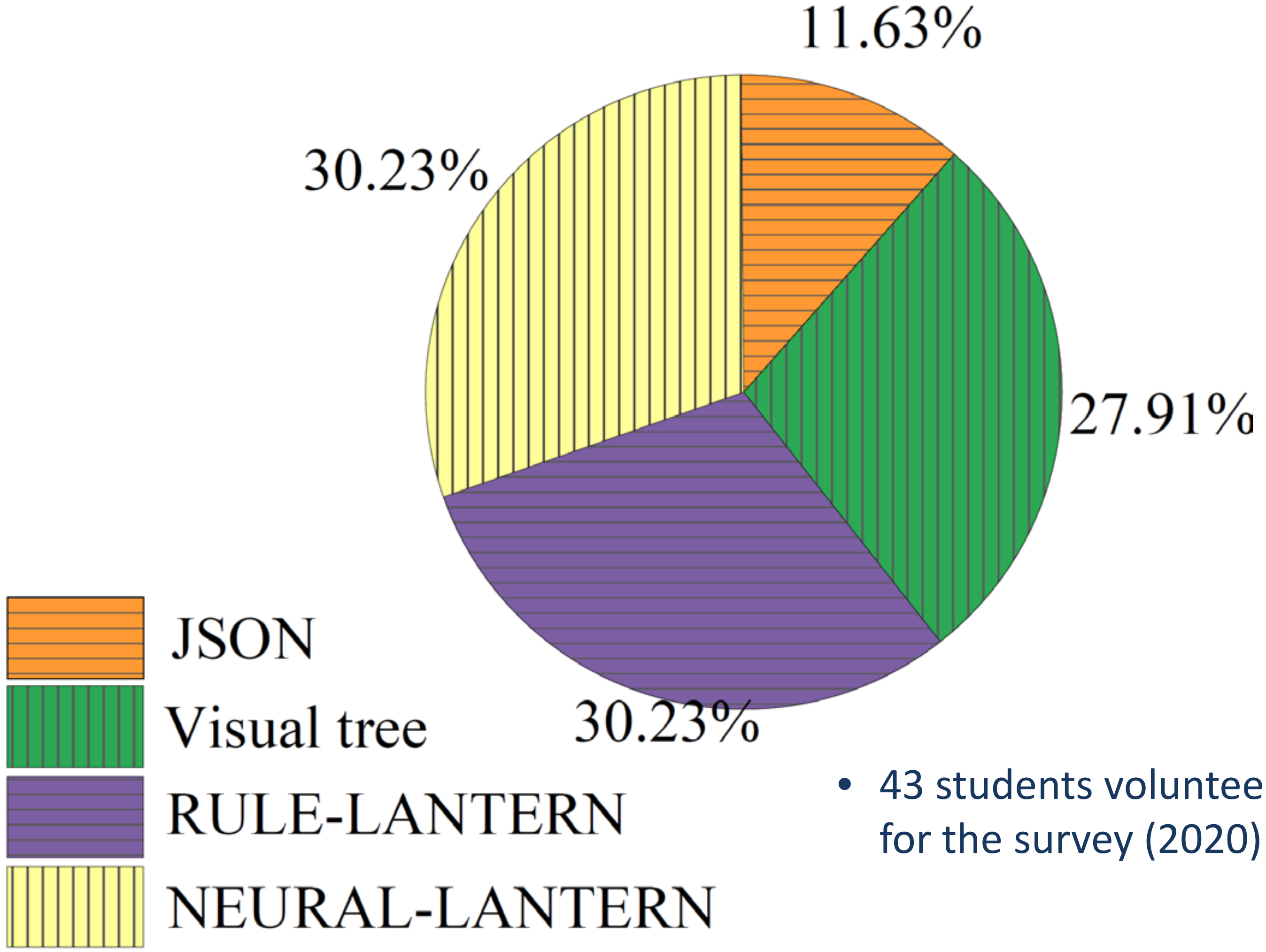
1. hash table <T> and hash semi enter under condition (<C>) on table tablename and table <T> to obtain intermediate table <TN> .
2. hash table <T> and do a half hash join on table tablename and table <T> under condition (<C>) to get intermediate table <TN> .
3. hash table <T> and perform hash semi join on table tablename and table <T> under condition (<C>) to get transitional table <TN> .

NEURAL-LANTERN: Translation Model

QEP2Seq Model

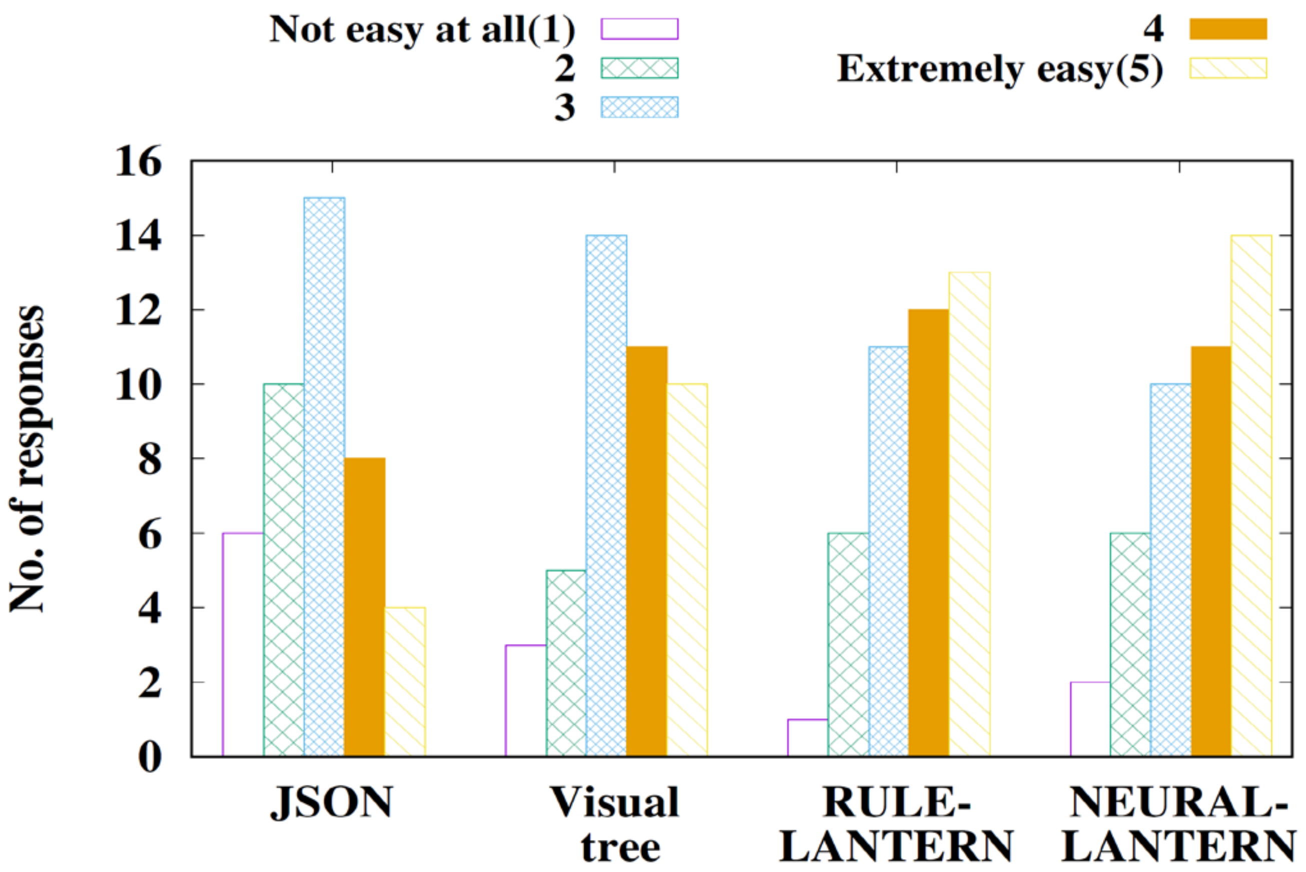
- Follows the Seq2Seq structure.
- The acts collection $actCol$ is composed of a series of acts $L_1, L_2, \ldots, L_n$, each of which is derived from the QEP.
- The Encoder RNN encodes each word in Li into the corresponding hidden state ht using an LSTM layer
- We use an LSTM decoder with an attention mechanism to let the decoder focus on the relevant portion of the encoder while generating a token.
- We adopt both static (Word2Vec and GloVe) and contextual word embeddings (ELMo and BERT) in decoder.
- Training data: workloads in TPC-H (22 queries) and SDSS (71 queries)
- Apply trained model on IMDB (1000 SQL queries)
- Pre-trained word embeddings can reduce the validation set loss while improving validation set accuracy and alleviate overfitting problem.

User Feedback: Which query plan format is most preferred?



- 43 students volunteered for the survey (2020)

User Feedback: How easy is it to understand the query plan presented in various formats?



Impact of Boredom

Do learners feel bored?

- We presented a set of output generated by each approach in random order
- Asked the subjects to rate the degree of boredom (boredom index) they felt perusing these plans to understand QEPs using the Likert scale of 1-5

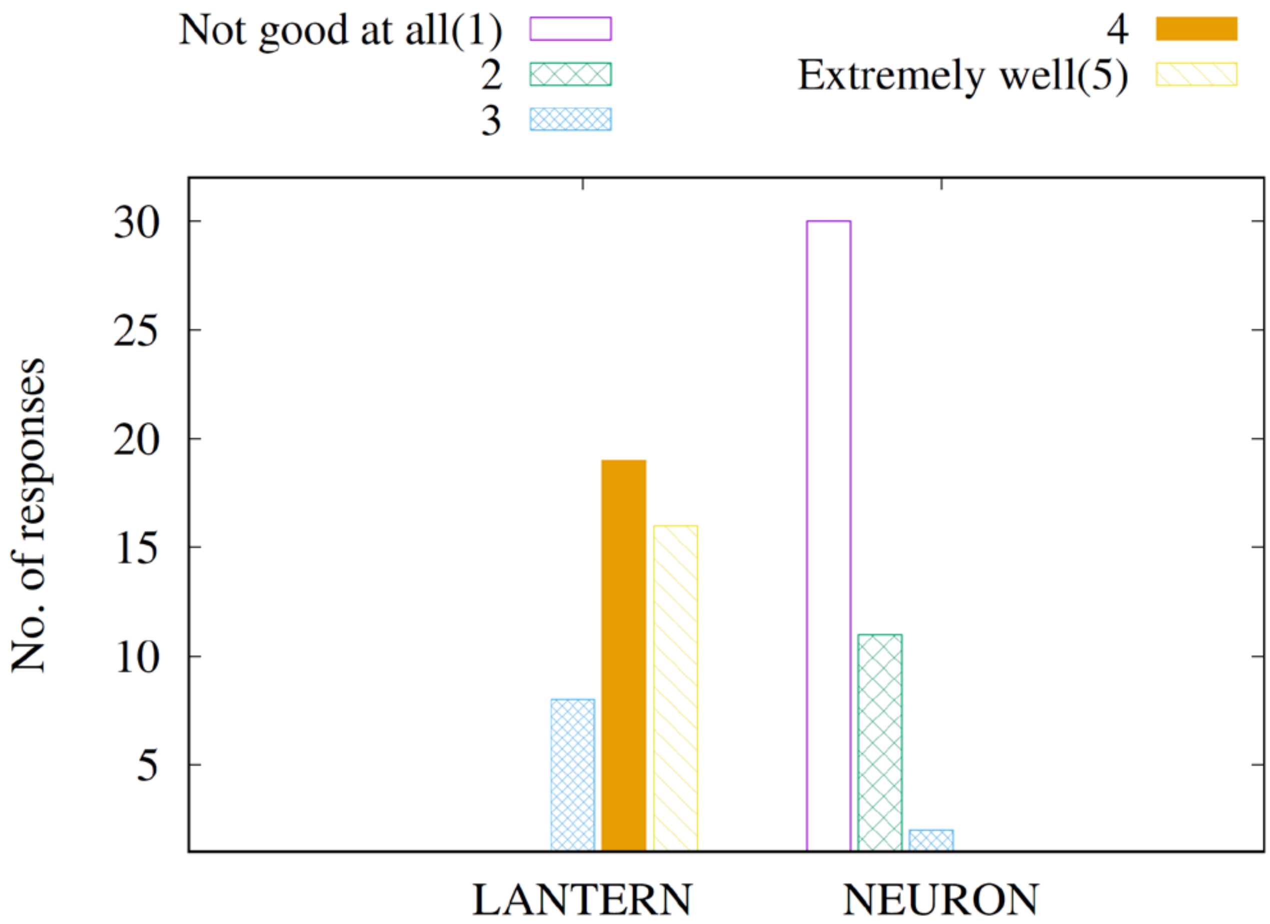
Method	Boredom index (not boring → extremely boring)				
	1	**2**	**3**	**4**	**5**
RULE-LANTERN	2	7	19	10	5
NEURAL-LANTERN	6	11	22	3	1
NEURON	2	8	16	11	6
LANTERN	6	12	21	2	2

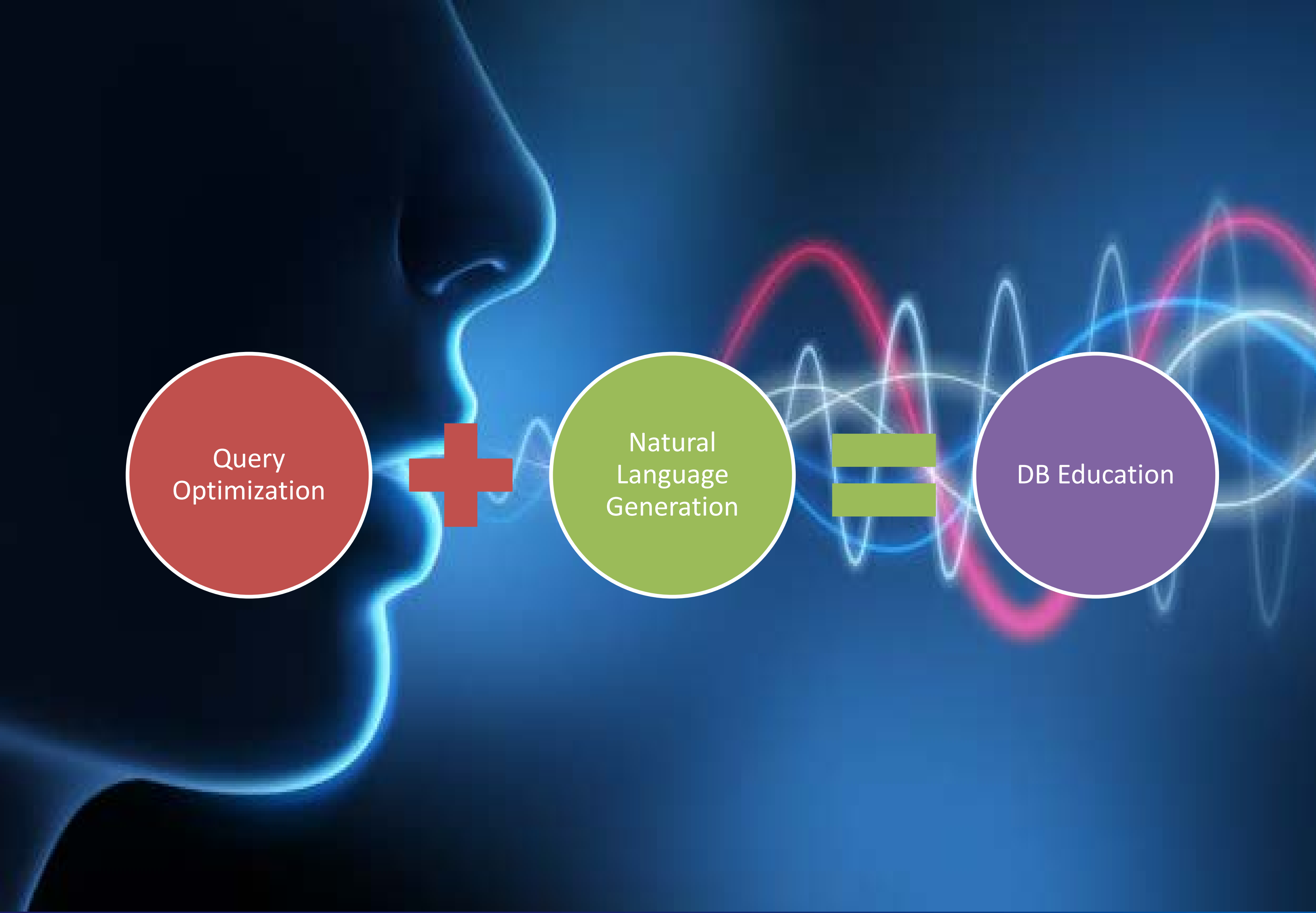
NEURON vs LANTERN

Integrate RULE-LANTERN and NEURAL-LANTERN

- Track (QEP, NL description) pairs viewed by each participants.
- By default, the NL description of each physical operator is generated using RULE-LANTERN.
- Whenever an operator appeared more than a pre-defined frequency threshold (i.e., 5) in total in different QEPs associated with a participant, NEURAL-LANTERN is invoked to generate the description for the operator.

User Feedback: LANTERN vs NEURON




Query Optimization
+
Natural Language Generation
=
DB Education

Summary

Simple and Communicable

- Generating natural language descriptions of QEPs

Timely and Fundamental

- Lifelong learning is increasingly becoming a reality
- By fundamental we mean that the idea touches on something basic to humans in collective settings

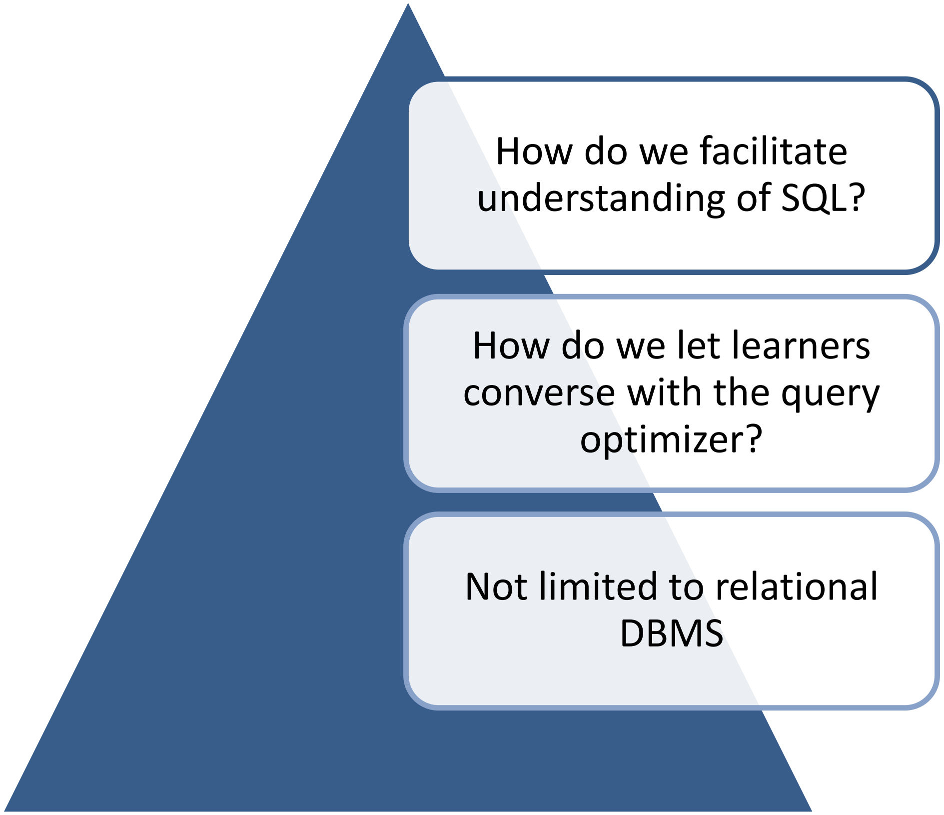
Personally Relevant

- Education is personal!

Interesting Issues based on Feedbacks

- How do we summarize the descriptions?
- How can learners explore the impact of physical operators on a QEP?
- How can learners converse with a query optimizer to aide learning?

DBMS Needs to Break Out from the Enterprise Jar



Acknowledgements

Hui Li, Xidian Univ, China

Shafiq Joty, NTU

Patricia Chen, Dept of Psychology, NUS

- Siyuan Liu, NTU
- Weiguo Wang, Xidian
- Peng Chen, Xidian
- Zheng Li, Xidian

- Student volunteers in NTU and Xidian


Thank You!