Scalable and Dynamic Generation of Big Data Volumes

Anupam Sanghi[†], **Raghav Sood, Jayant Haritsa**
Indian Institute of Science, Bangalore, India

Srikanta Tirthapura
Iowa State University, Ames, USA

Database Systems Laboratory
Indian Institute of Science

Motivation

Database vendors need access to client data for:

- Testing their engines to resolve client problems.
- Assessing impacts of a planned engine upgrade.

However, client data is often unavailable!

- Privacy concerns
- Transfer cost (especially at Big Data Scale)

Objective

Dynamically generate client's representative data during query execution while ensuring similar performance on original and synthetic data

Mechanism

Volumetric Similarity: When a matching client plan is executed at the vendor, the number of output rows for each operator should be similar.

Toy Example

Workload Dependent Data Generation

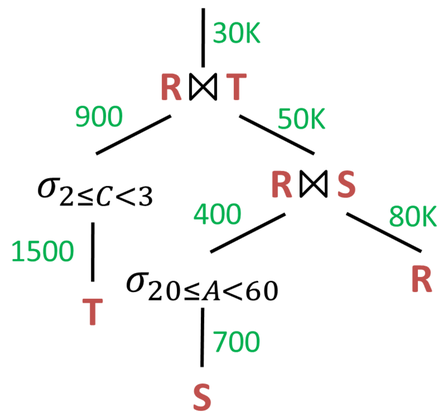
Schema

R (R_pk, S_fk, T_fk)
S (S_pk, A, B)
T (T_pk, C)

SQL Query

SELECT * **FROM** R, S, T
WHERE R.S_fk = S.S_pk
AND R.T_fk = T.T_Pk
AND S.A >= 20 AND S.A < 60
AND T.C >=2 AND T.C < 3

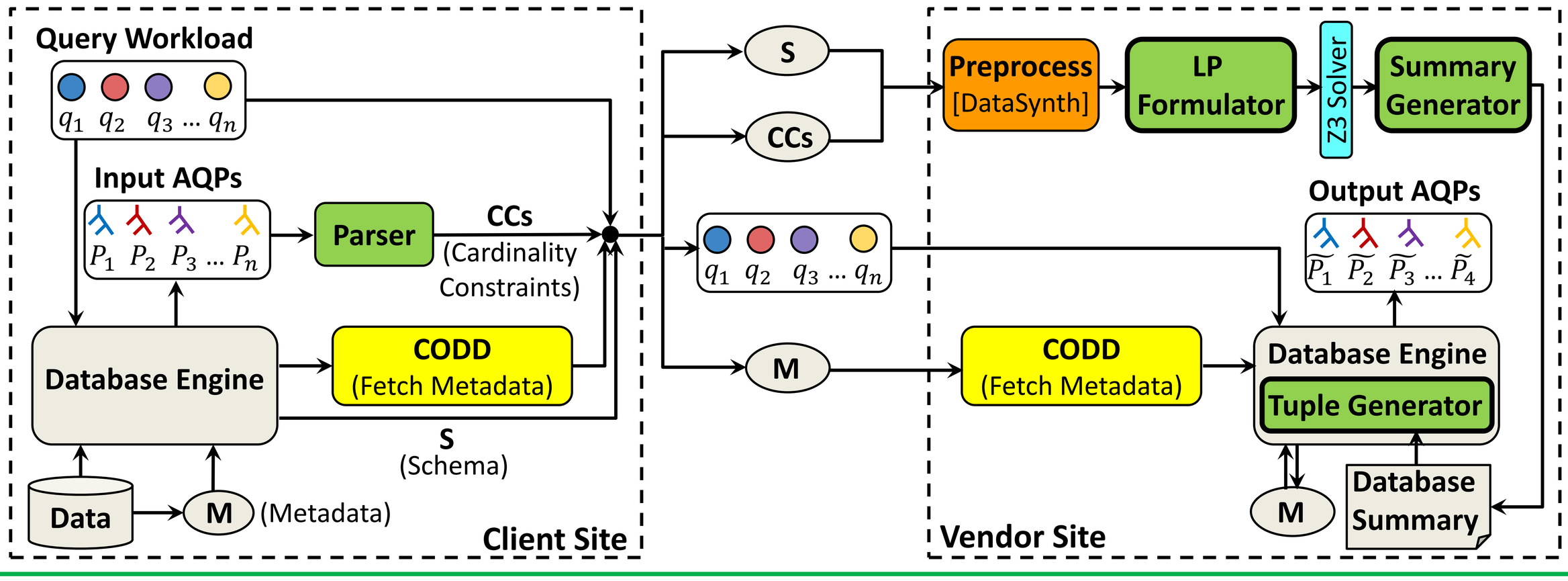
Annotated Query Plan



Cardinality Constraints

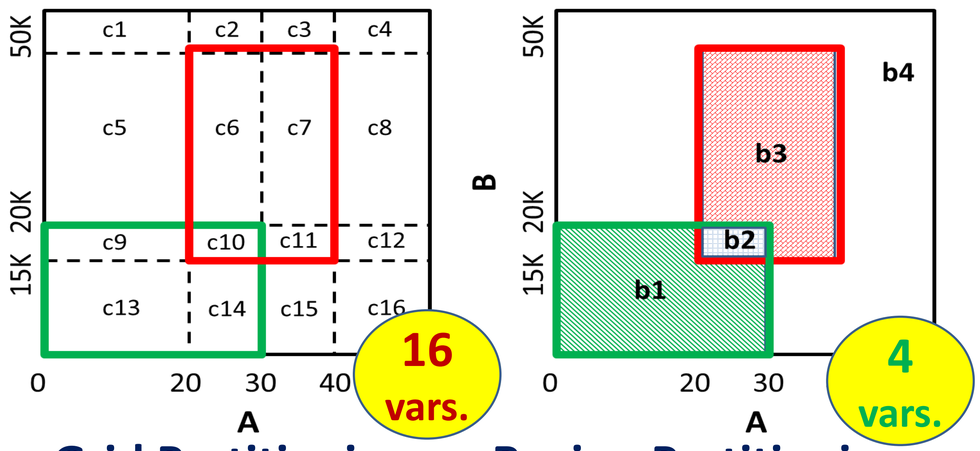
$|R| = 80K$
$|S| = 700$
$|T| = 1500$
$|\sigma_{S.A\in[20,60)}(S)| = 400$
$|\sigma_{T.C\in[2,3)}(T)| = 900$
$|\sigma_{S.A\in[20,60)}(R\bowtie S)| = 50K$
$|\sigma_{S.A\in[20,60)\ \wedge\ T.C\in[2,3)}(R\bowtie S\bowtie T)| = 30K$

Hydra Architecture



Hydra Contributions

Extended Workload Coverage



Grid Partitioning vs. Region Partitioning

Dynamic Data Generation

R		
R_pk	S_fk	T_fk
1 – 30K	321	1
30001 – 50K	621	601
50001 – 60K	71	601
60001 – 70K	121	1
70001 – 80K	1	1

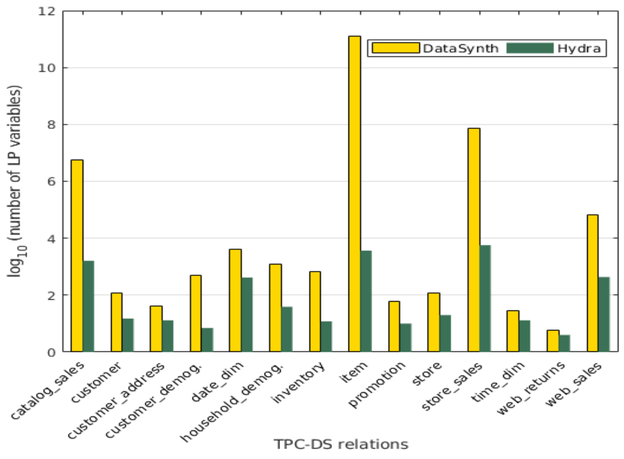
S		
S_pk	A	B
1–100	0	15
101–250	20	15
251–500	20	10
501–700	0	5

T	
T_pk	C
1–600	0
601–1500	2

Example Database Summary

Experimental Evaluation

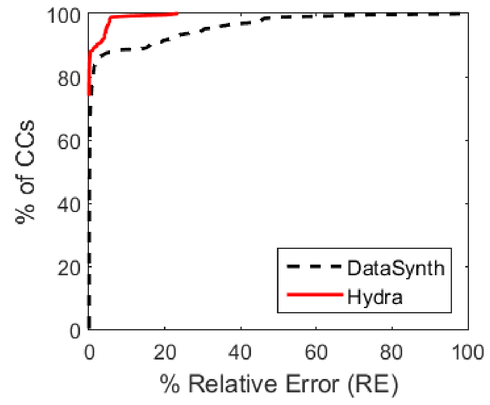
Workload Complexity



LP Processing Time

Workload 1		Workload 2	
DataSynth	Hydra	DataSynth	Hydra
50 min	13 sec	Crash	58 sec

Volumetric Similarity



Materialization Time

Size (in GB)	DataSynth	Hydra
10	4 hours	2 min
100	42 hours	11 min
1000	> 1 week	1.6 hours

Hydra generated the database summary from a set of 131 AQPs in less than 2 minutes for an Exabyte (10^{18}) scale database!

[†] Contact: anupamsanghi@iisc.ac.in

Project Website: dsl.cds.iisc.ac.in/projects/HYDRA