Dot-Product Join: An Array-Relation Join Operator for Big Model Analytics

Chengjie Qin, Martin Torres, and Florin Rusu

University of California Merced

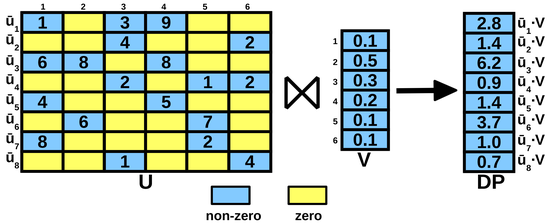
Background & Motivation

Example: Gradient in logistic regression where $\vec{x}_i$ is the training example; $\vec{w}$ is the model.

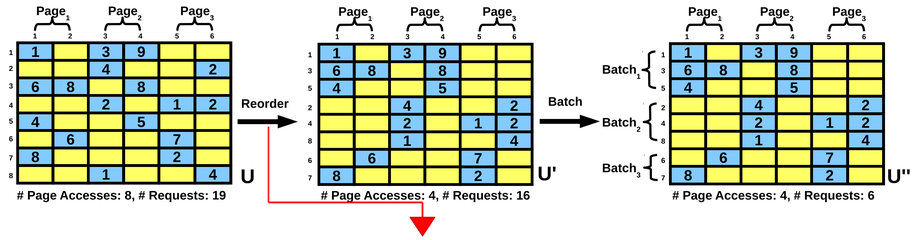
$$\sum_i \frac{e^{-y_i \vec{w} \cdot \vec{x}_i}}{1 + e^{-y_i \vec{w} \cdot \vec{x}_i}} \left(-y_i \cdot \vec{x}_i\right)$$

Model size grows beyond the memory of a single machine.

Dot-Product Join Operator



Optimization Strategies



Reordering strategies: LSH-based Nearest Neighbor, Radix Sort, K-Center Clustering

Dot-Product Join Algorithm

Require:
U (index INTEGER[],value NUMERIC[],tid INTEGER)
V (index INTEGER,value NUMERIC)
memory budget M

Ensure: DP (tid INTEGER,product NUMERIC)

1: **for each** page $u_p \in U$ **do**

OPTIMIZATION

2: Reorder vectors $\vec{u}_i$ to cluster similar vectors together

3: Group vectors $\vec{u}_i$ into batches B_j that access at most M pages from V

BATCH EXECUTION

4: **for each** batch B_j **do**

5: Collect pages $v_p \in V$ accessed by vectors $\vec{u}_i \in B_j$ into a set $v_{B_j} = \{v_p \in V | \exists \vec{u}_i \in B_j$ that accesses $v_p\}$

6: Request access to pages in v_{B_j}

DOT-PRODUCT COMPUTATION

7: **for each** vector $\vec{u}_i \in B_j$ **do**

8: dp $\leftarrow$ *Dot-Product*($\vec{u}_i$,V)

9: Append ($\vec{u}_i.tid$, dp) to DP

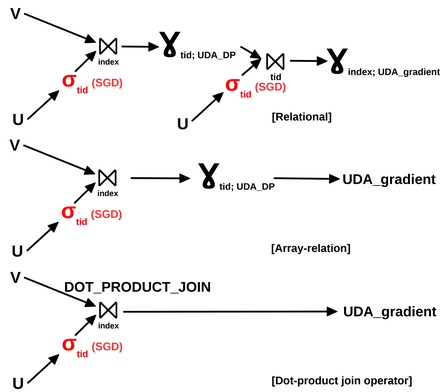
10: **end for**

11: **end for**

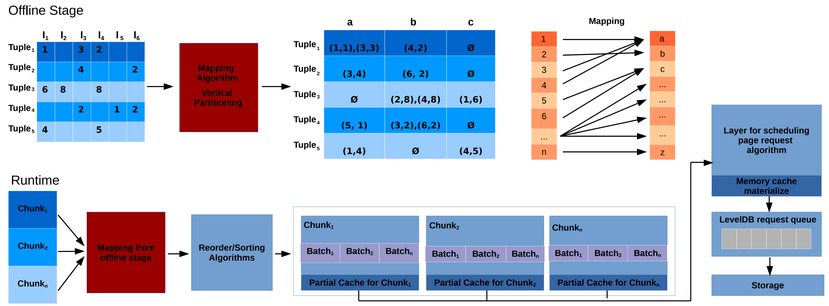
12: **end for**

13: **return** DP

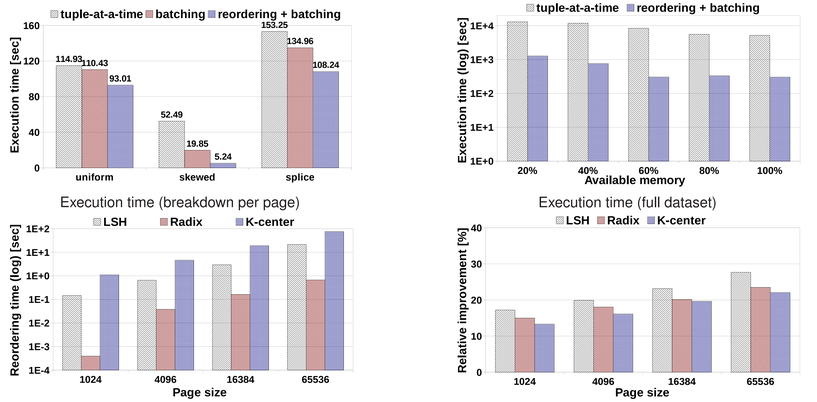
Gradient Descent Integration



Parallel Dot-Product



Experimental Evaluation



Reordering heuristics execution time

Relative improvement over basic LRU

	GLADE R	PG R	PG ARRAY-RELATION	MonetDB R	DOT-PRODUCT JOIN
	[28 GB]	[28 GB]	[28 GB]	[28 GB]	
uniform	37851 [3.2 GB]	N/A	N/A	4280	**3914** [3.2 GB]
skewed	13000 [3.2 GB]	N/A	N/A	4952	**786** [3.2 GB]
matrix	N/A [16 GB]	N/A	N/A	N/A	**9045** [16 GB]
splice	9554 [40 MB]	81054	N/A	**2947**	6769 [40 MB]
MovieLens	905 [32 MB]	5648	72480	N/A	**477** [32 MB]

Dot-product join execution time (in seconds) with memory budget in []. N/A stands for not finishing the execution in 24 hours.

University of California Merced

Email: {cqin3, mtorres58, frusu}@ucmerced.edu