Efficient Selection of Various k-objects for a keyword Query based on MapReduce Skyline Algorithm

Md. Anisuzzaman Siddique
and
Yasuhiko Morimoto

Hiroshima University

Overview

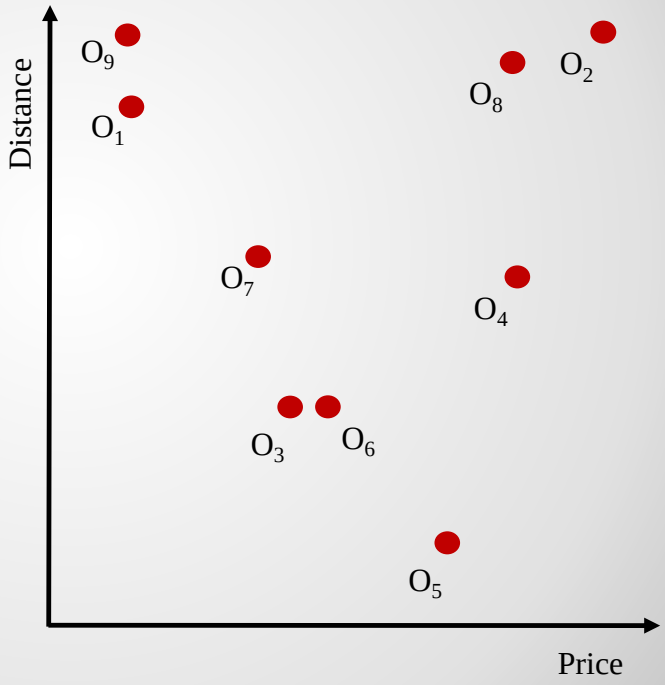
1. Top-k Query
2. Skyline Query
3. Related work
4. Top-k and Skyline Query
5. Contributions
6. k-Objects Selection using MapReduce
7. Performance Evaluation
8. Conclusions

Top-k Query processing

Find *k* objects that minimize a scoring function.

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20
:	:	:

Hotels Dataset



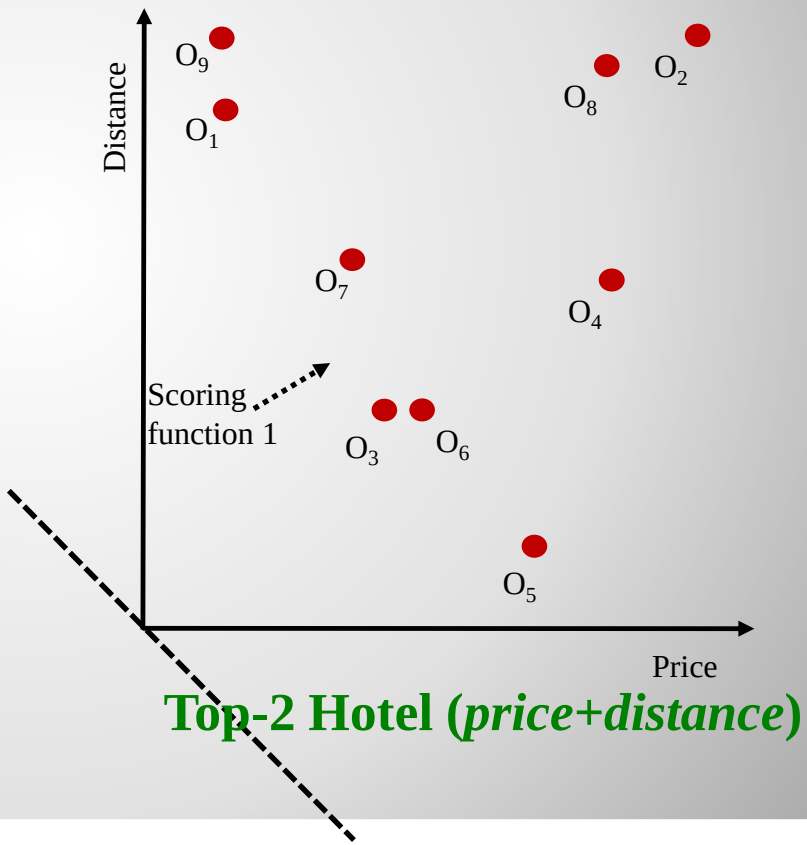
Geometric Interpretation

Top-k Query processing

Find the 2 hotels that minimizes (***price+distance***).

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20
:	:	:

Hotels Dataset



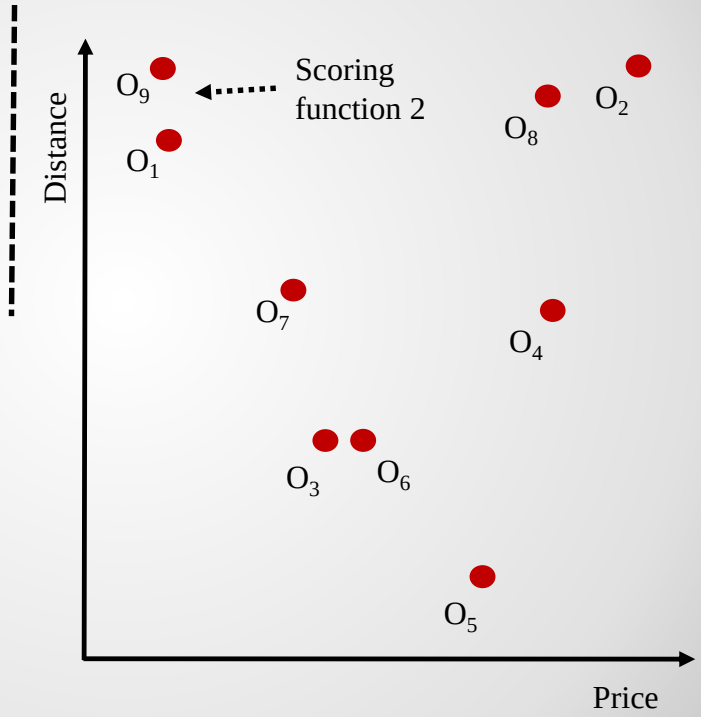
Top-2 Hotel (*price+distance*)

Top-k Query processing

Find the 2 hotels that minimizes (***price***) only.

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20
:	:	:

Hotels Dataset

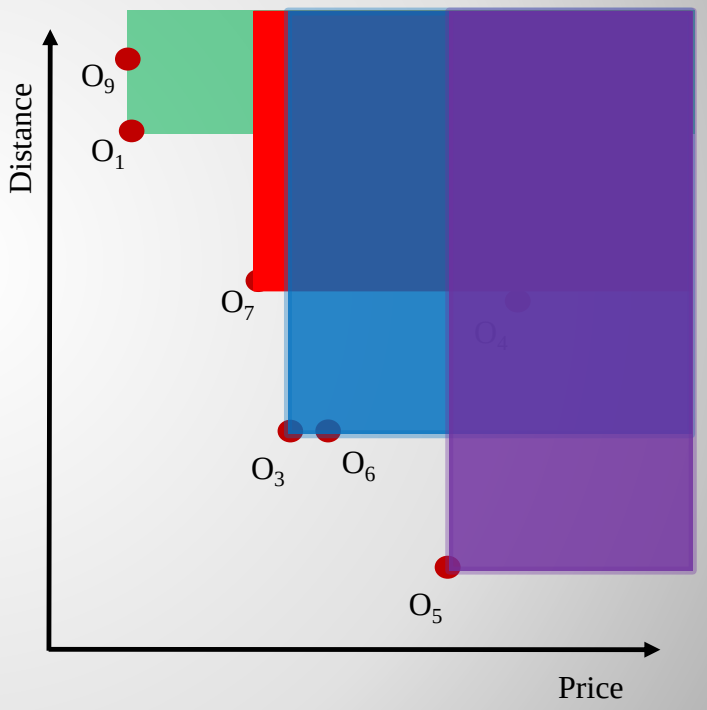


Top-2 Hotel (*price*)

Skyline Query processing

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20
:	:	:

Hotels Dataset

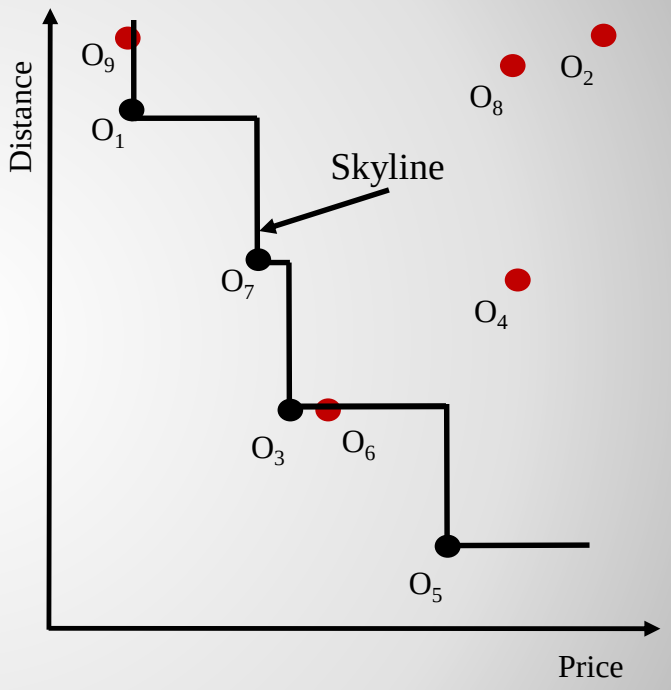


Dominance regions

Skyline Query processing

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20
:	:	:

Hotels Dataset



Skyline of Hotels Dataset

Related work

Top-k

- No Random Access (NRA), *Fagin et al., PODS'01*
- Threshold Algorithm (TA), *Fagin et al., PODS'01*
- PREFER, *Hristidis et al., SIGMOD'01*
- etc.

Skyline

- Sort Filter Skyline (SFS), *Chomicki et al., ICDE'03*
- Linear Elimination Sort for Skyline (LESS), *Parke et al., VLDB'05*
- Nearest Neighbor (NN), *Kossmann et al., VLDB'02*
- Branch and Bound Skyline (BBS), *Papadias et al., SIGMOD'03*
- etc.

Top-k and Skyline Query

Top-k

Advantage: Always select k-objects according to scoring function.

Disadvantage: It require a concrete scoring function.

Skyline

Advantage: It does not require any scoring function.

Disadvantage: No control on retrieved objects. It is difficult to identify interesting objects from large result.

Our Contributions

- We combine the strength of both skyline and top-*k* query and propose an algorithm to select various *k* objects without scoring function.
- We consider an efficient algorithm to select the *k* objects with MapReduce framework.

Problem Definition

A user looking for ***k*** hotels that are **cheap** and **close to beach**. (keyword = ***Aizu hotel*** and set ***k=2)***.

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20

Aizu Hotels Dataset

How to choose 2 hotels without scoring function?

k-Objects Selection Procedure

Skyline computation

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20

Aizu Hotels Dataset

ID	Price	Distance
O_1	3	17
O_3	9	6
O_5	13	3
O_7	8	12

Skyline result

k-Objects Selection Procedure

Case 1: No. of skyline objects >= *k* (say, *k* = 2)

ID	Price	Distance
O_1	3	17
O_3	9	6
O_5	13	3
O_7	8	12

Skyline Dataset

ID	Price	Distance	Pri.+Dist.
O_1	3 (1)	17 (4)	20 (3)
O_3	9 (3)	6 (2)	15 (1)
O_5	13 (4)	3 (1)	16 (2)
O_7	8 (2)	12 (3)	20 (3)

Skyline Dataset

ID	Comb. Rank
O_1	8
O_3	6
O_5	7
O_7	8

Skyline Dataset

ID	Comb. Rank
O_3	6
O_5	7

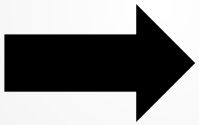
Top-2 Result

k-Objects Selection Procedure

Case 2: $k >$ No. of skyline objects (say, $k = 5$)

ID	Price	Distance
O_1	3	17
O_3	9	6
O_5	13	3
O_7	8	12

Skyline result

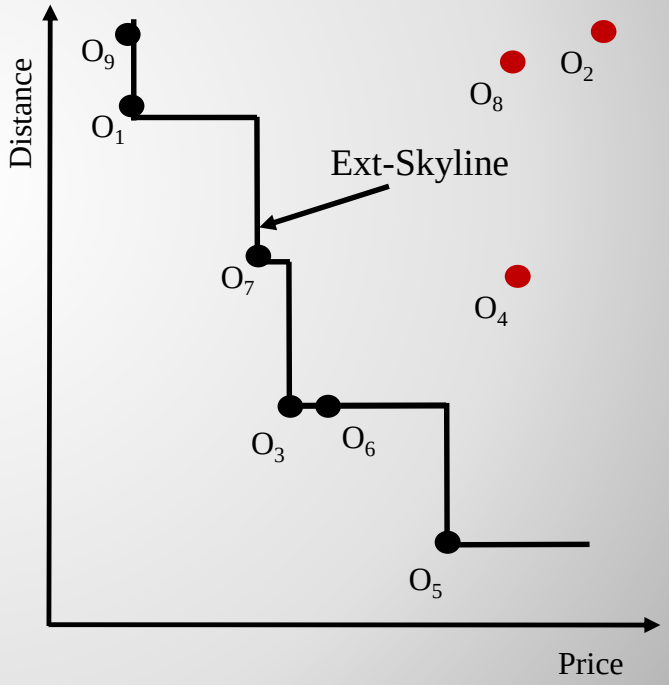


No solution!

We compute extended skyline objects and select *k* objects.

ID	Price	Distance
O_1	3	17
O_2	17	19
O_3	9	6
O_4	16	11
O_5	13	3
O_6	10	6
O_7	8	12
O_8	13	18
O_9	3	20

Aizu Hotels Dataset



Ext-Skyline of Hotels Dataset

k-Objects Selection Procedure

Case 2: k > No. of skyline objects (say, $k = 5$)

ID	Price	Distance
O_1	3	17
O_3	9	6
O_5	13	3
O_6	10	6
O_7	8	12
O_9	3	20

Ext. Skyline result

ID	Price	Distance	Pri.+Dist.
O_1	3 (1)	17 (5)	20 (4)
O_3	9 (4)	6 (2)	15 (1)
O_5	13 (6)	3 (1)	16 (2)
O_6	10 (5)	6 (2)	16 (2)
O_7	8 (3)	12 (4)	20 (4)
O_9	3 (1)	20 (6)	23 (6)

Ext. Skyline result

ID	Comb. Rank
O_1	10
O_3	7
O_5	9
O_6	9
O_7	11
O_9	13

Skyline Dataset

ID	Comb. Rank
O_1	10
O_3	7
O_5	9
O_6	9
O_7	11

Top-5 result

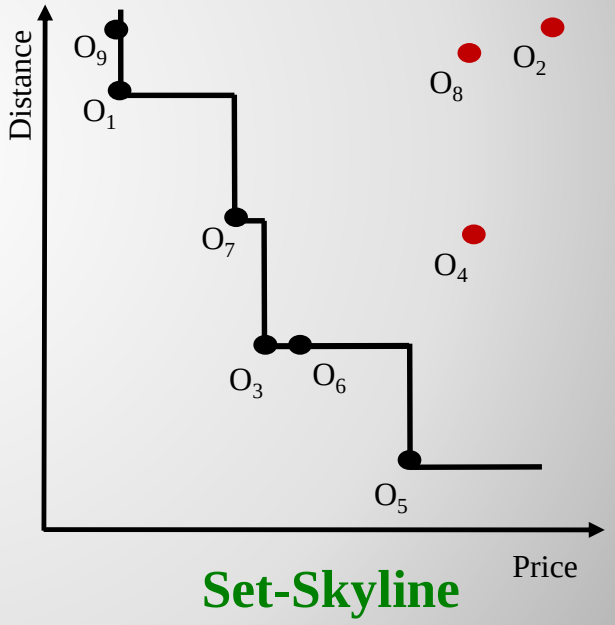
k-Objects Selection Procedure

Case 3: does not follow case 1 & 2 (*k > ext. skyline objects*)

We need to consider other procedure.

Set skyline can be a solution

We discussed on skyline sets query in DNIS 2010



Set-Skyline

How to select *k* objects from set skyline. (Future work)

k-Objects Selection using MapReduce

It has three phases:

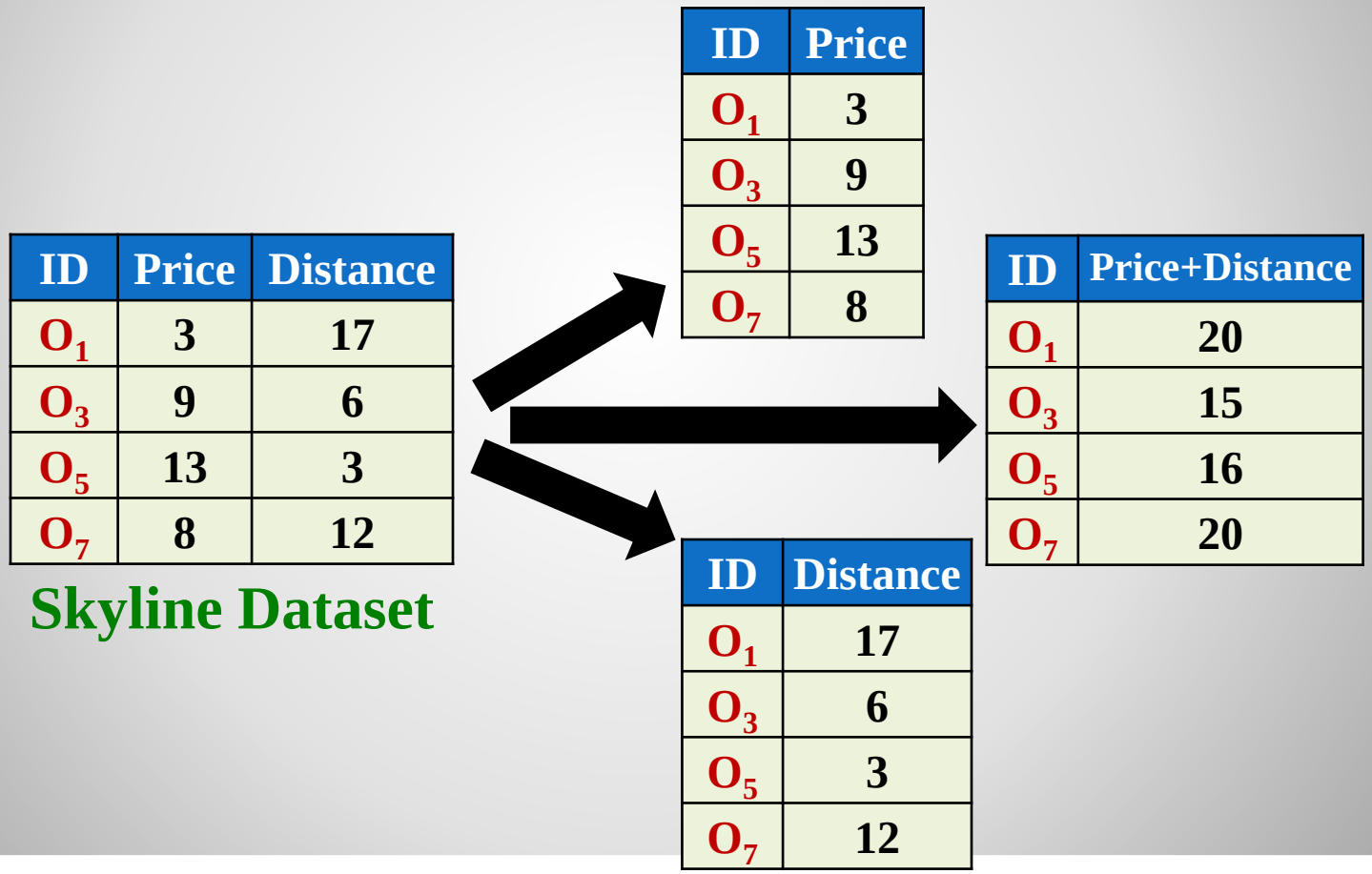
- Data splitting phase
- Map and ranking phase
- Reduce and selection phase

ID	Price	Distance
O_1	3	17
O_3	9	6
O_5	13	3
O_7	8	12

Skyline result

Skyline and Splitting phase

Phase 1: Split skyline dataset vertically

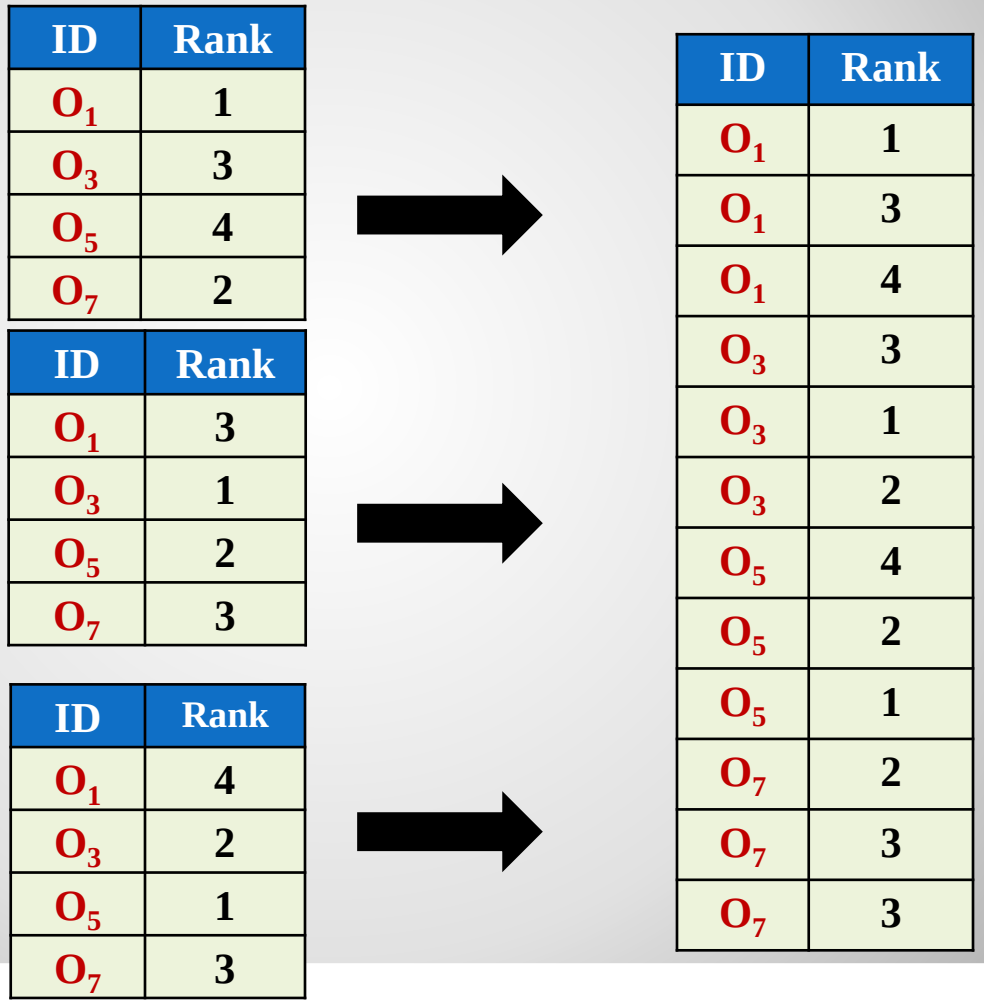


Map and Ranking Phase

Phase 2(a): Map all data objects

Map and Ranking Phase

Phase 2(b): Send (*ID*, *Rank*) pairs to the combiner



Reduce and Selection Phase

Phase 3(a): Reducer produce (*ID*, *list(Rank)*) pairs

ID	Rank
O_1	1
O_1	3
O_1	4
O_3	3
O_3	1
O_3	2
O_5	4
O_5	2
O_5	1
O_7	2
O_7	3
O_7	3

Reducer Input

ID	Comb. Rank
O_1	8
O_3	6
O_5	7
O_7	8

Reducer Output

Reduce and Selection Phase

Phase 3(b): Select *k* objects

ID	Comb. Rank
O_1	8
O_3	6
O_5	7
O_7	8

Reducer Output

ID	Comb. Rank
O_3	6
O_5	7

Result for k = 2

Experimental Datasets

Synthetic Datasets

- **Independent**
- **Anti-Correlated**

Experimental Setup

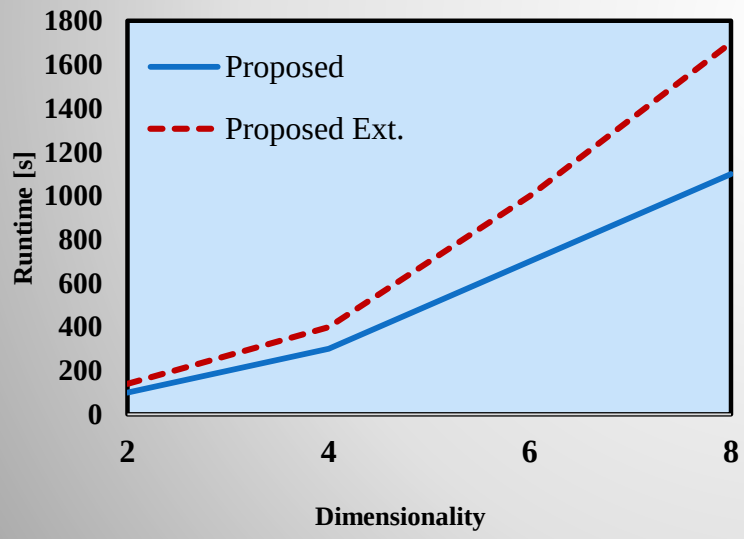
- 4 PCs with Intel Core i7 3.4GHz CPU
- 4GB RAM
- Windows 8.0 OS.
- 100 Mbits/S LAN connection

Varying *Dimensionality*

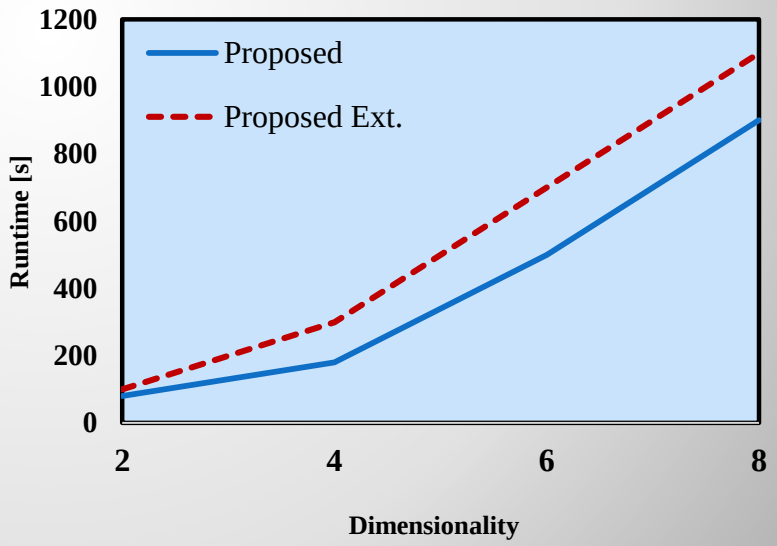
No of records = 100k

Dimension = 2 - *8*

Query Number = *500*



a) Anti-Correlated



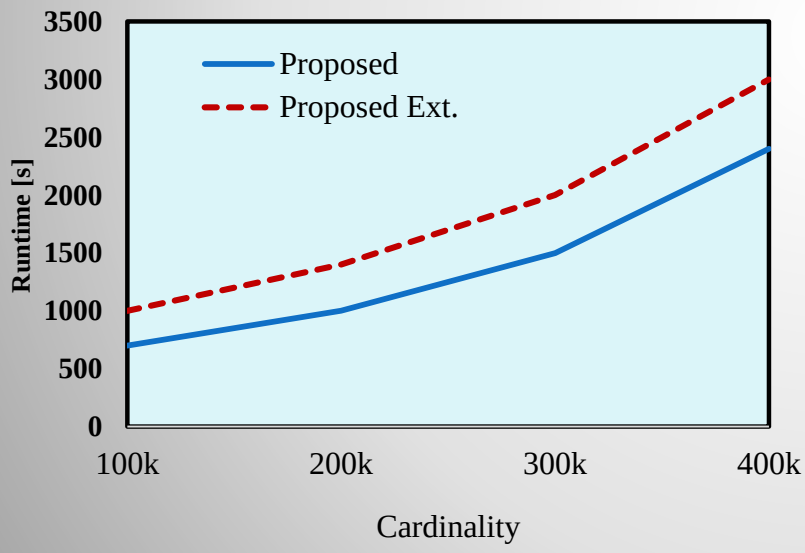
b) Independent

Varying *Cardinality*

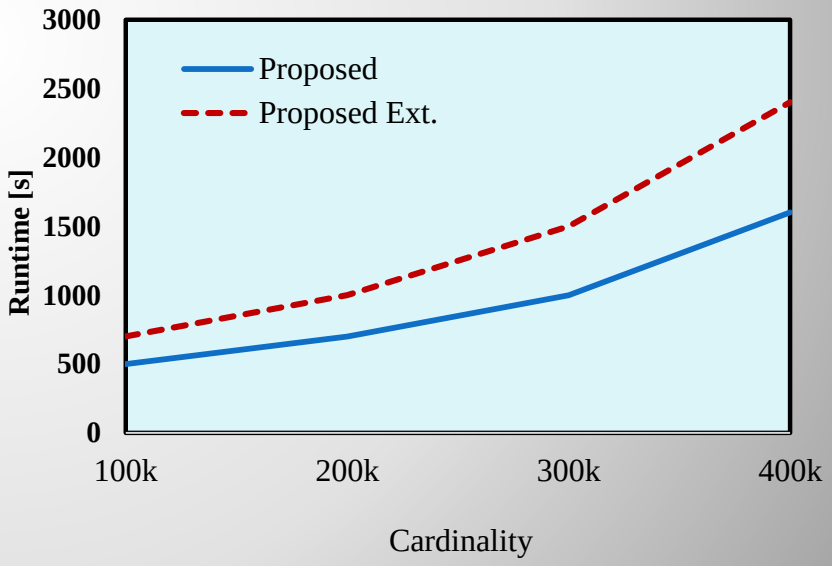
No of records = 100- 500k

Dimension = 6

Query Number = *500*



a) Anti-Correlated



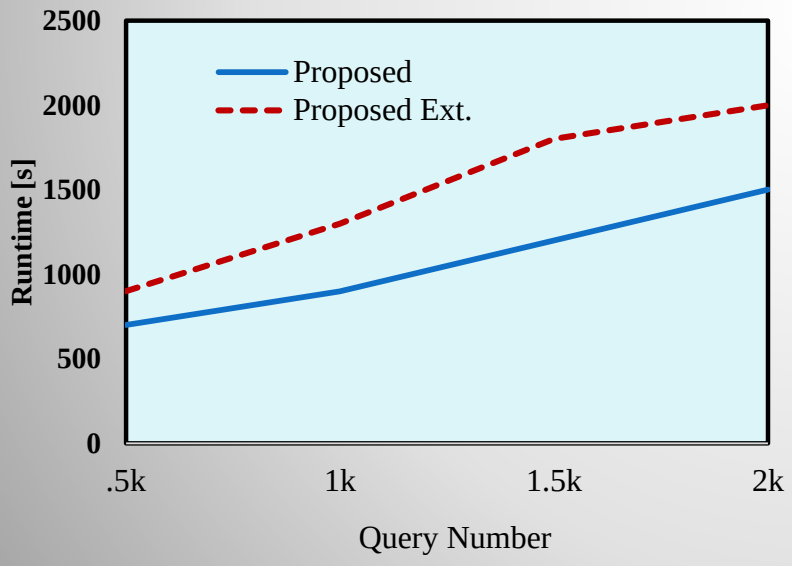
b) Independent

Varying *Query Number*

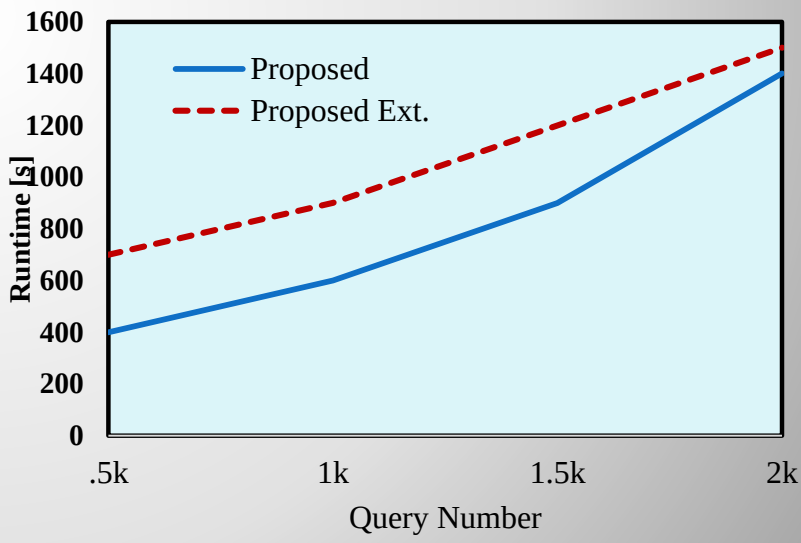
No of records = 100k

Dimension = 7

Query Number = *500-2000*



a) Anti-Correlated



b) Independent

Conclusions

- *This work addresses k-objects selection problem and give guideline how to selects k objects.*
- *We applied the idea of skyline queries to select the k objects and proposed an efficient algorithm.*
- *Experimental results show the effectiveness of the proposed algorithm.*

THANK YOU

Questions ?