Kangaroo: Workload-Aware Processing of Range Data and Range Queries in Hadoop

Ahmed M. Aly[1], Hazem Elmeleegy[2], Yan Qi[2], Walid G. Aref[1]

[1]Purdue University, West Lafayette, IN, USA [2] Turn Inc., Redwood City, CA, USA

Motivation

Analytical (Range) Queries for Temporal (Range) Data:

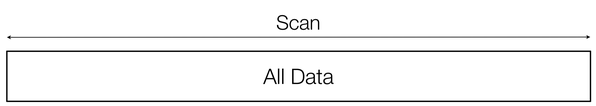
- **Digital Advertising:** Analytics for digital ad campaigns over cookies that are active within certain periods of time
- **Television:** Analytics for viewed TV channels (and for how long)
- **Telecom:** Analytics for usage and billing over certain periods of time

Example: Ad campaign analysis:
At Turn Inc., each analytical query has a filtering time interval to retrieve the cookies that are *active* within that interval

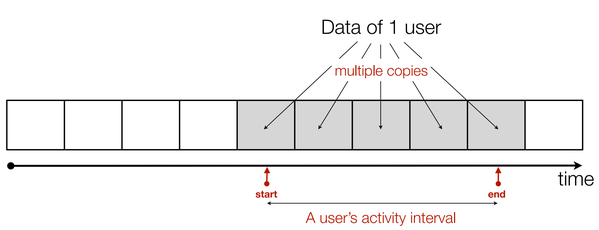
SELECT ...
FROM ..
DATES [2011-12-01, 2012-03-24]

Straightforward Approaches:

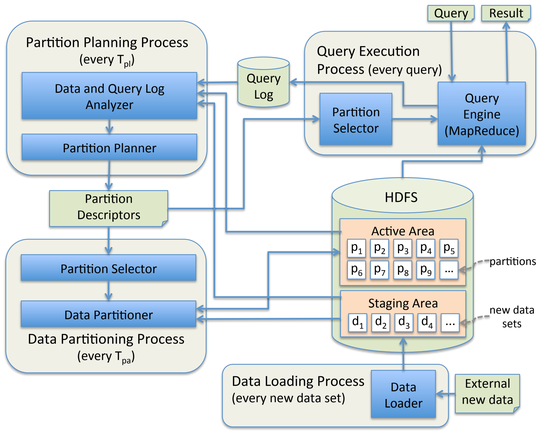
- Scan all data. Inefficient!



- Partition the data across time. Incurs redundancy due to data duplication.



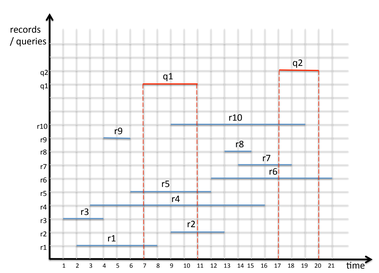
Kangaroo



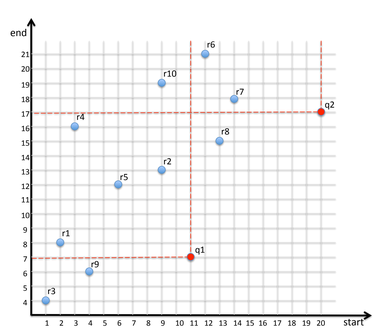
Main Features:

- No data duplication
- Workload awareness
- No partition overlap
- Bounded partition sizes and number of partitions

Space Transformation:



Sample range data and range queries in the raw format (1D)



Sample range data and range queries in the transformed format (2D)

Workload Aware Partitioning

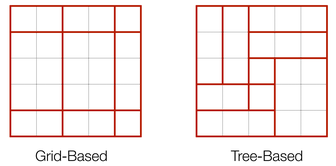
Kangaroo finds the best partitioning layout that minimizes the total processing time of a given query workload:

$$\sum_{\forall p} C(p)$$

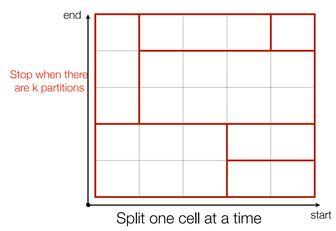
$$C(p) = N(p) \times Q(p)$$

where $N(p)$ is the number of users in p, and $Q(p)$ is the number of queries overlapping p.

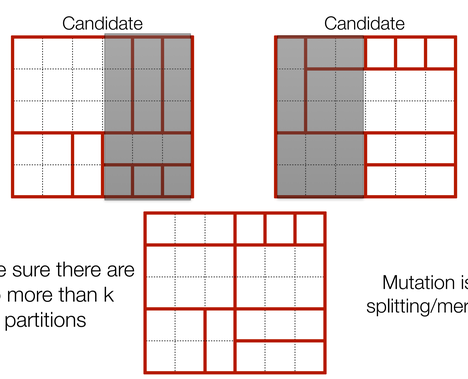
Partitioning Strategies:



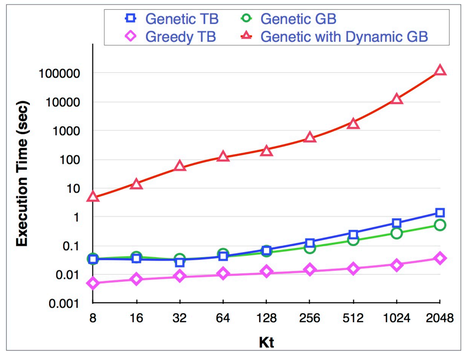
- Greedy



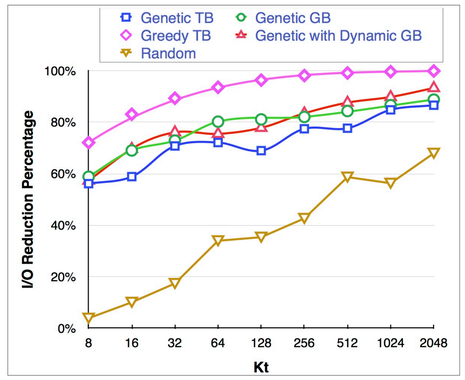
- Genetic



Experiments



I/O reduction percentage achieved at a given K_t



Execution time of the partitioning algorithms at a given K_t

Acknowledgements

This research was supported in part by National Science Foundation under Grant IIS 1117766.