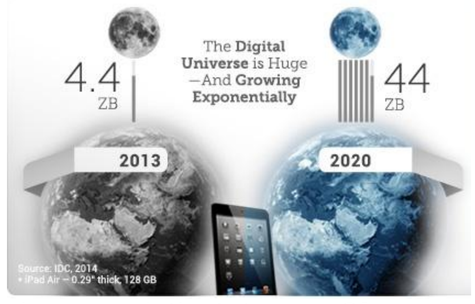
An Experimental Study of Aging Influence on Query Cost estimation

Chihiro Kato[1], Yuto Hayamizu[1], Kazuo Goda[1], Masaru Kitsuregawa[1,2]
[1]The University of Tokyo, [2]National Institute of Informatics

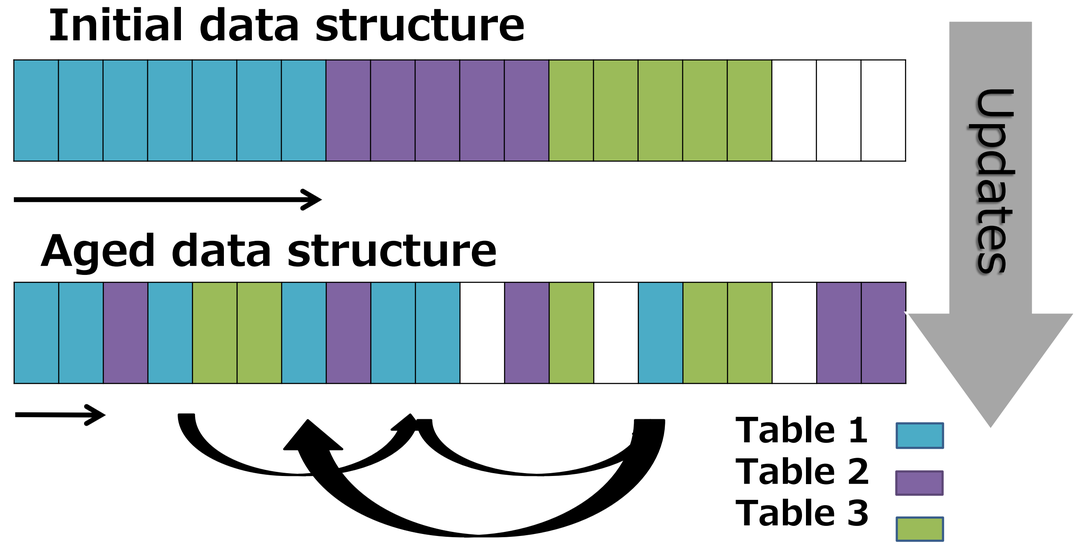
Database Aging

- Big Data draws many attentions from the society
  - Volume: large scale data store
  - Velocity: Intensive updates



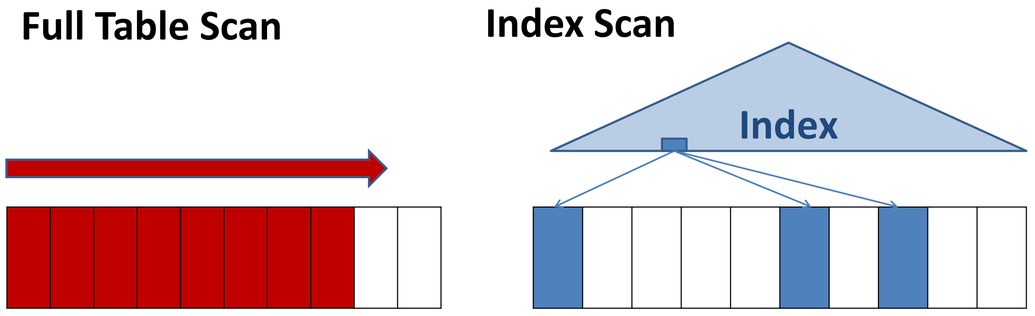
From 『The Digital Universe and Big Data – EMC』
(http://www.emc.com/leadership/digital-universe/index.htm)

- Aging
  - Deterioration of database physical structure caused by intensive database updates
  - It is impossible to invoke aging all the time

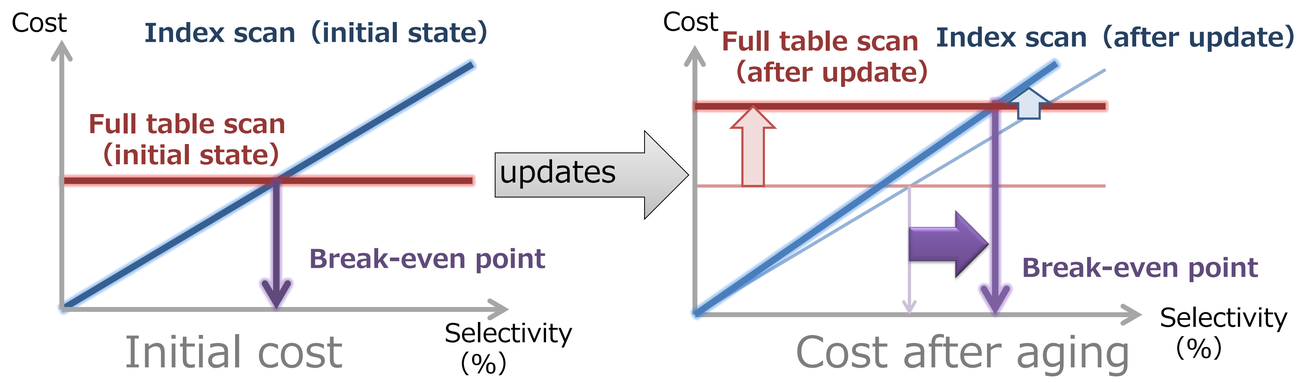


Aging influence on query optimizer

- Two major access methods
  - Full table scan: Scan entire table
  - Index scan: Access partial table via index
  - Choice of access method is a crucial key for query optimization



- Aging affects access cost
  - And, sometimes even change optimal plan



- Consideration of database aging may be beneficial for query optimization

Experiment on aging influence

Setup

Dell Power Edge R720xd
WD9001BKHG（900GB）×10

PostgreSQL version 9.4.0: Shared buffer 128MB
TPC-H benchmark: Scale factor 100 (100GB)

Measurement query

Query (A): on clustered index

```
SELECT SUM(l_extendedprice)
FROM lineitem
WHERE l_orderkey < x
```

Query (B): on secondary index

```
SELECT SUM(l_extendedprice)
FROM lineitem
WHERE l_partkey < x
```

- Investigate aging influence using two access methods
  - Compare actual execution cost and estimated cost
  - There were non-negligible errors observed
    - up to 66.3% (Query(A)) and up to 50.5% (Query(B))

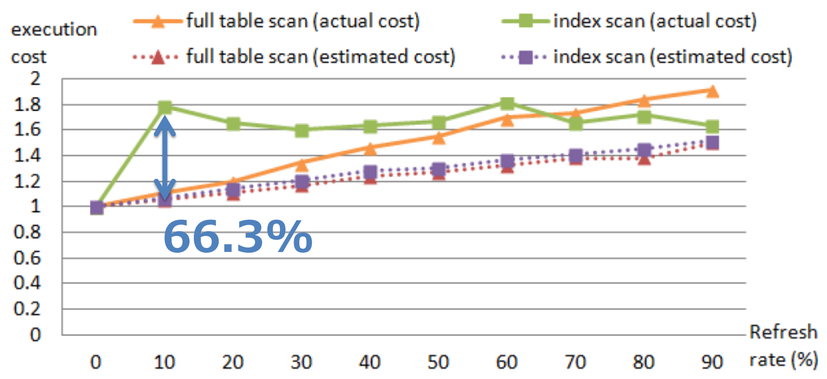


Figure1: actual and estimated cost in Query (A)

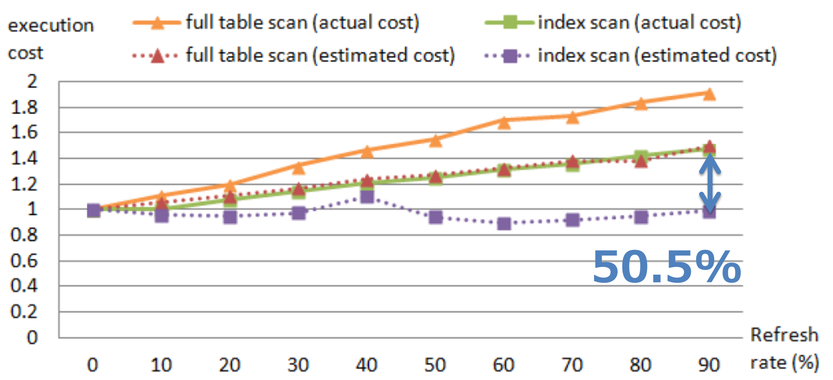


Figure2: actual and estimated cost in Query (B)

Conclusion

- Examine aging influence on query cost estimation
  - Execution time eventually increased
- Analyze cost estimation error
  - a conventional cost-based optimizer could choose a non-optimal query plan