Using BigBench to Evaluate an Automated Physical Design of Materialized Views

Jiang Du, Boris Glavic, Wei Tan, Renée J. Miller
University of Toronto, Illinois Institute of Technology,
IBM T. J. Watson Research Center
Motivation

• OLAP queries are not ad-hoc.

• Materialization has the potential to improve query performance.

• Early attempt to extend materialization to scalable data analytics
  • [Elghandour and Aboulnaga PVLDB’12, Perez and Jermaine ICDE’14]

• Current research does not consider the physical design of materialized views.

• Materialized views are only used to rewrite a query on the logical level.

• An appropriate physical design of materialized views can greatly improve query performance.
DeepSea: Workload-aware Physical Design

- Workload-aware combination of materialization with partitioning.
  
  - Materialize intermediate query results as views to speed up future queries.
  
  - Views are incrementally partitioned based on incoming queries
    
    - We start with a coarse partitioning and split partitions that are frequently accessed
  
  - Bookkeeping of (potential) view/partition benefits to make future decisions on materialization and repartitioning
Approach

• View Creation
  • Materialize and partition intermediate query results as views
  • Incrementally refine partitions based on workload
  • Can piggy-back on fault tolerance mechanisms

• Query Answering with Views
  • Automatically match views/partitions to incoming queries using enhanced index structure from [Goldstein and Larson SIGMOD ’01].

• View Selection
  • Keep track of (potential) benefits of views and repartitioning and use that to inform materialization and repartitioning decisions
Experimentation

- We need a benchmark framework that reflects the characteristics of big data analytics.
  - Traditional benchmarks, e.g., TPC-H
    - No user defined functions
    - Inflexible data generator (e.g., control skew)
  - We choose BigBench as our testing framework.
    - Use of user defined functions
    - Flexible data generator
Empirical Evaluation

![Diagram](image)

**Partitioning in 1TB dataset**

- **Original Hive**
- **Materialization w/o partitioning**
- **Equidepth (6 Buckets)**
- **DeepSea Partitioning**

- **Queries in sequence (selectivity: 0.01, heavily skewed)**
- **Total time elapsed (seconds)**

- **Dim.**
- **Fact**
- **Reduce**
- **Map**