aster data

more data. big insights.

BigBench: Big Data Benchmark Proposal

Ahmad Ghazal, Minqing Hu, Tilmann Rabl, Alain Crolotte, Francois Raab, Meikel Poess, Hans-Arno Jacobsen

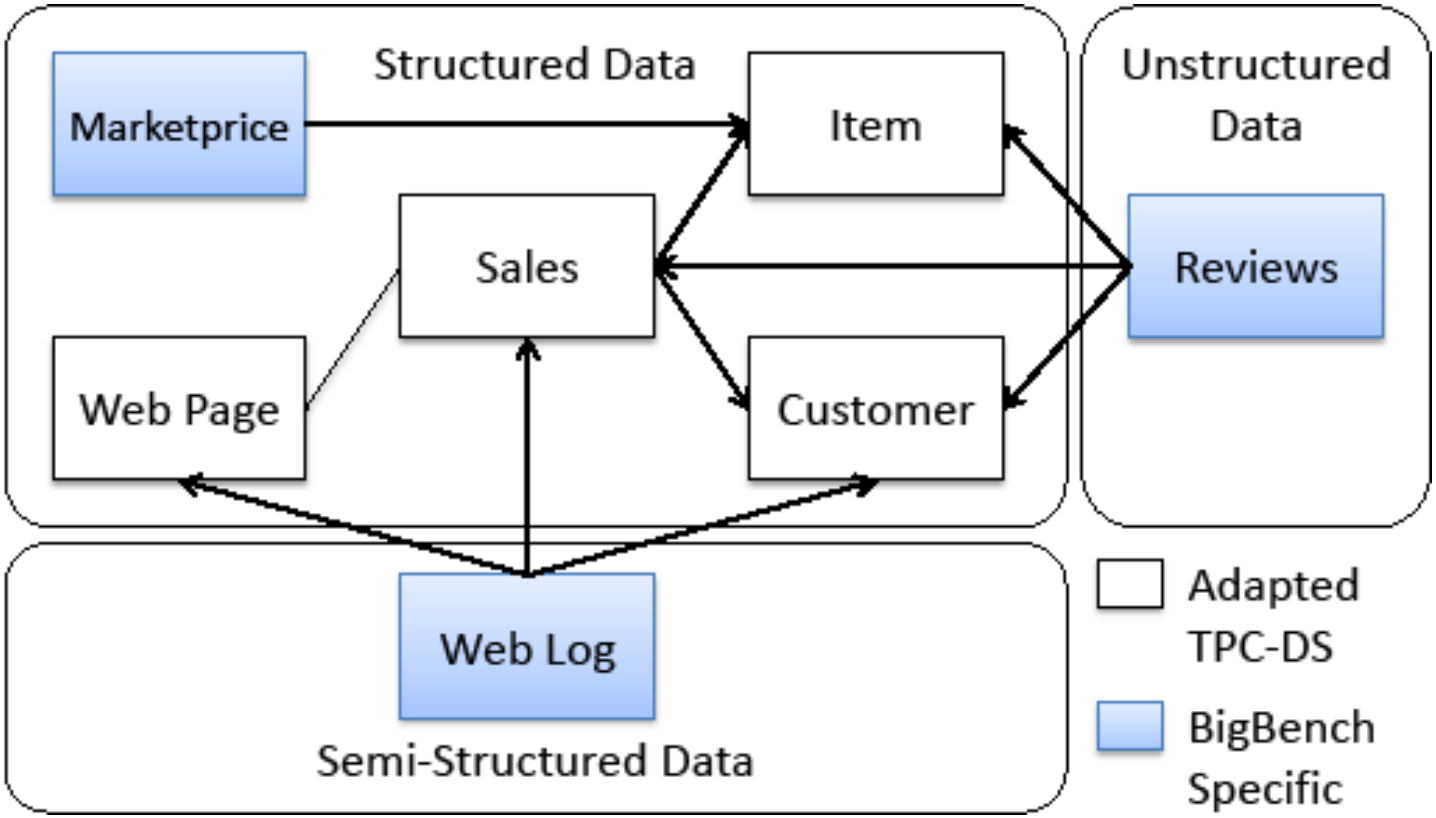
BigBench

- Initial work presented at 1st WBDB, San Jose
- Based on a product retailer
- End to end benchmark
- Focus on
  - Parallel DBMS
  - MR engines
- Collaboration with Industry & Academia
  - Teradata
  - University of Toronto
  - InfoSizing
  - Oracle
- Full paper submitted to SIGMOD 2013

BigBench (outline)

- **Data Model**
  - Variety, Volume, Velocity
  - Variety :
    - structured from TPC-DS
    - Semi-structured: web logs
    - Un-structured: review text
- **Data Generator**
  - PDGF for structured data
  - Enhancement : Semi-structured & Text generation
- **Workload specification**
  - Main driver: retail big data analytics
  - Covers : data source, declarative & procedural and machine learning algorithms.
- **Evaluation**
  - Done on Teradata Aster
  - Queries written using SQL-MR

BigBench data model



BigBench data model

- **Volume**
  - Based on scale factor
  - Similar to TPC-DS scaling
  - Weblogs & product reviews also scaled

- **Velocity**
  - Periodic refreshes for all data
  - Different velocity for different areas
    - $V_{structured}$
    - $V_{unstructured}$
    - $V_{semistructured}$
  - Queries run with refresh

BigBench data generator

- "Parallel Data Generation Framework" PDGF
  - For the structured part of model
  - Scale factor similar to TPC-DS
- Extensions to PDGF for web logs & product reviews
- Web logs: retail customers/guests visiting site
  - Web logs similar to apache web server logs
  - Coupled with structured part
  - Sizing based on scale factor
- Product reviews: Customers and guest users
  - Algorithm based on Markov chain
  - Real data set sample input
  - Coupled with structured and based on scale factor as well

BigBench Workload

- 30 queries
- Specified in English
- No required syntax
- Driven by big data retail analytics
  - Adapted from McKinsey

BigBench Workload (continued)

Retail analytics 5 areas

- **Marketing**
  - Cross-selling
  - Customer micro-segmentation
  - Sentiment analysis
  - Enhancing multichannel consumer experiences
- **Merchandising**
  - Assortment optimization
  - Pricing optimization
- **Operations**
  - Performance transparency
  - Product return analysis
- **Supply chain**
  - Inventory management
- **Reporting (customers and products)**

BigBench Workload (continued)

Technical Functions

- **Data source dimension**
  - Structured
  - Semi-structured
  - Un-structured
- **Processing type dimension**
  - Declarative (SQL, HQL)
  - Procedural
  - Mix of both
- **Analytic technique dimension**
  - Statistical analysis: correlation analysis, time-series, regression
  - Data mining: classification, clustering, association mining, pattern analysis and text analysis
  - Simple reporting: ad hoc queries not covered above

BigBench Evaluation

- BigBench proof of concept
- Can be done On DBMS
  - Typically data loaded into tables
  - Possibly parsing weblogs to get schema
  - Reviews captured as VARCHAR or BLOB fields
  - Queries run using SQL + UDF

- Can be done on MR engine
  - Data can be loaded on DFS like HDSF
  - MR, HQL, PigLatin can be used

- DBMS and MR engine
  - DBMS with Hadoop connectors
  - Data can be placed and split among both
  - Processing can also be split among two

BigBench Evaluation (continued)

- Done on Teradata Aster
  - Has functionality to run BigBench
- Data generation
  - DSDGen produced structured part
  - PDGF+ produced semi-structured and un-structured
- Data loaded into tables
  - Weblogs table
  - Product reviews table
- Queries
  - SQL-MR syntax

BigBench Evaluation (continued)

- Example query
- Perform category affinity analysis for products purchased online together.
  - Computes the probability of browsing products from a category after customers viewed items from another category.
  - Referred as market basket as well

- Business case: Marketing
  - cross-selling
- Type of source: structured
- Processing type : mix of declarative and procedural
- Analytic type: data mining
  - Affinity analysis

BigBench Evaluation (continued)

```
SELECT
category_cd1 AS category1_cd ,
category_cd2 AS category2_cd , COUNT (*) AS cnt
FROM
   basket_generator (
                ON
                   ( SELECT i. i_category_id AS category_cd ,
                   s. ws_bill_customer_sk AS customer_id
                   FROM web_sales s INNER JOIN item i
                   ON s. ws_item_sk = i_item_sk
                   )
                   PARTITION BY customer_id
                   BASKET_ITEM (' category_cd ')
                   ITEM_SET_MAX (500)
    )
GROUP BY 1,2
order by 1 ,3 ,2;
```

Next steps

- BigBench: industry standard benchmark.
  - Data, workload and metric speciation details.
- Provide a downloadable kit
  - Finalize implementation of data and query generators.
- Proof of concept
  - Include velocity and multi-user test.
  - Run the benchmark on one the Hadoop ecosystem