Context and Background

Chaitan Baru
Associate Director Data Initiatives, San Diego Supercomputer Center
Director, Center for Large-Scale Data Systems Research (CLDS)
University of California San Diego

Workshops on Big Data Benchmarking

- First WBDB workshop, May 2012, San Jose. Hosted by Brocade.
  - ~60 attendees from ~45 different organizations
- Second WBDB workshop, December 2012, Pune, India. Hosted by Persistent Systems / Infosys.
  - ~40 attendees from ~25 different organizations
- Third WBDB workshop, July 2013, Xi'an, China. Hosted by Xi'an University.
- Fourth WBDB workshop, October 9-10, 2013, San Jose. Hosted by Brocade
  - Collocated with IEEE Big Data Conference

Publication of WBDB Papers

- Paper from First Workshop
  - Setting the Direction for Big Data Benchmark Standards by C. Baru, M. Bhandarkar, R. Nambiar, M. Poess, and T. Rabl, published in *Selected Topics in Performance Evaluation and Benchmarking*, Springer-Verlag
- Article in inaugural issue of Big Data Journal
  - *Big Data Benchmarking and the Big Data Top100 List* by Baru, Bhandarkar, Nambiar, Poess, Rabl, Big Data Journal, Vol.1, No.1, 60-64, Anne Liebert Publications.
- Selected WBDB papers to be published in two volumes of Springer Verlag LNCS
  - 2012 Volume: 1st and 2nd WBDB
  - 2013 Volume: 3rd and 4th WBDB

Benchmarking Issues

- Audience: Who is the audience for such a benchmark?
- Application: What is the application that should be modeled?
- Single benchmark spec: Is it possible to develop a single benchmark to capture characteristics of multiple applications?
- Component vs. end-to-end benchmark. Is it possible to factor out a set of benchmark "components", which can be isolated and plugged into an end-to-end benchmark(s)?
- Paper and Pencil vs Implementation-based. Should the implementation be specification-driven or implementation-driven?
- Reuse. Can we reuse existing benchmarks?
- Benchmark Data. Where do we get the data from?
- For Innovation or competition? Should the benchmark be for innovation or competition?
- Verifiability of results. Should the benchmark results be audited/ verified in some way? Yes!

Different types of Benchmarks

- System-level versus function-level benchmarks
  - System-level: performance of hardware and software, for a given dataset and workload (ie a given application scenario)
  - Function-level: Performance of a specific function on given hardware, software, dataset.
- Micro-benchmarks
  - E.g. A Micro-benchmark Suite for Evaluating HDFS Operations on Modern Clusters, Panda et al, OSU
- Functional benchmarks
  - Terasort
- Genre-specific benchmarks
  - Graph500
- Application-level benchmarks
  - TPC benchmarks: E.g., TPC-C, TPC-D, TPC-DS

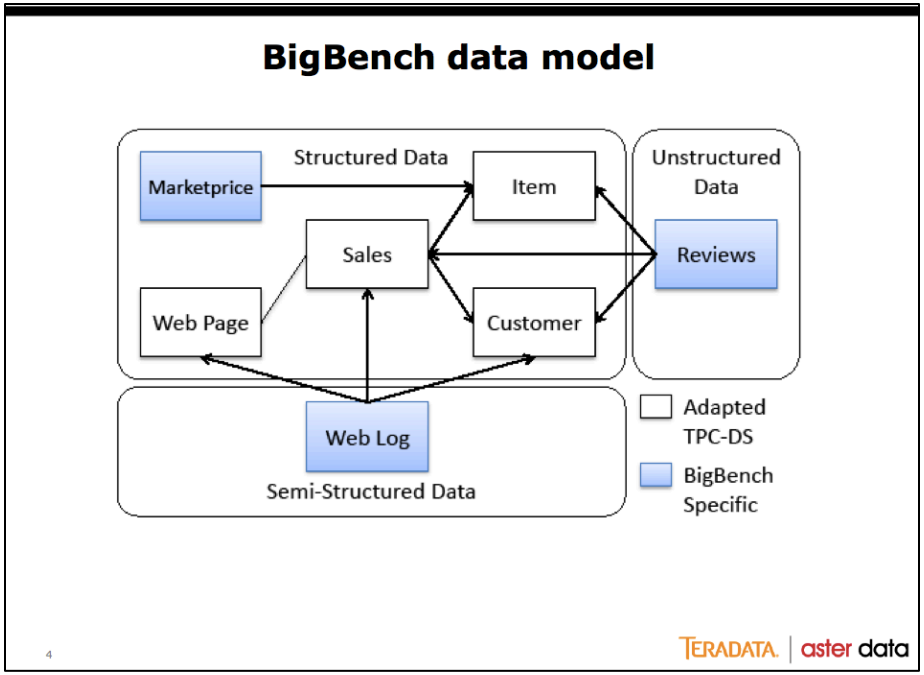
Big Data Characteristics

- Includes data at the "edges" of the enterprise
  - Outside of enterprise transaction systems
  - "Interactions" data
    - E.g. log data, social network data, etc.
  - Data from multiple sources
- Application characteristics
  - Data structuring determined by agile applications
  - Need for loose (flexible) schemas, and "late binding" of schemas
  - ELT rather than ETL
  - Data runs through processing pipelines
  - Execute models using data
- Used for "event detection"
  - User clicks
  - Device failures
  - Hospital re-admissions
  - …

Proposal 1: BigBench

- By Ghazal et al: Teradata, Oracle, U.of Toronto, InfoSizing
- Derived from TPC-DS
- TPC-DS:
  - Multiple snowflake schemas with shared dimensions
  - 24 tables with an average of 18 columns
  - 99 distinct SQL 99 queries with random substitutions
  - More representative skewed database content
  - Sub-linear scaling of non-fact tables
  - Ad-hoc, reporting, iterative and extraction queries
  - ETL-like data maintenance

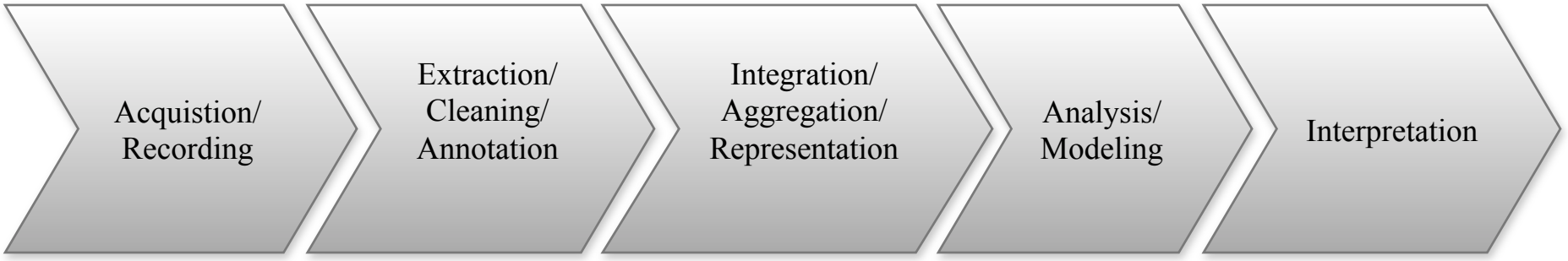
BigBench Data Model



- Workload = Set of queries
  - On structured, semistructured, unstructured data
  - Data mining, ML

Proposal 2: Deep Analytics Pipeline

- Sequence of processing steps:
  - From data ingestion to data cleaning and transformation (ELT, sorting, SQL queries)
  - To Machine Learning and Predictive Analytics
  - Feed data from one step to the next
- Data generation: Need to create a suitable input dataset
  - Synthetic event-based data…with perturbations?



A Deep Analytics Pipeline Application: User Modeling

- Objective: Determine user interests by mining user activities
- Large dimensionality of possible user activities
- Typical user has sparse activity vector
- Event attributes change over time

User Modeling Pipeline

- Data Acquisition
- Sessionization
- Feature and Target Generation
- Model Training
- Offline Scoring & Evaluation
- Batch Scoring & Upload to serving