KNOW MORE IN A DASH: dashDB

Easy Webscale Data Warehousing in the Cloud

Sam Lightstone, IBM Canada | email: light@ca.ibm Visit: http://dashDB.com

RESEARCH TOPICS:
- Speed is king! Fastest possible query times
- Structured data *and* NoSQL as data sources
- Make data warehousing simple
- Support many SQL and stored proc. dialects (Oracle, PostgreSQL, DB2, Netezza etc)

ENGINEERING TOPICS:
- Webscale – megabytes to petabytes
- Bring spatial and statistical analytics to the data

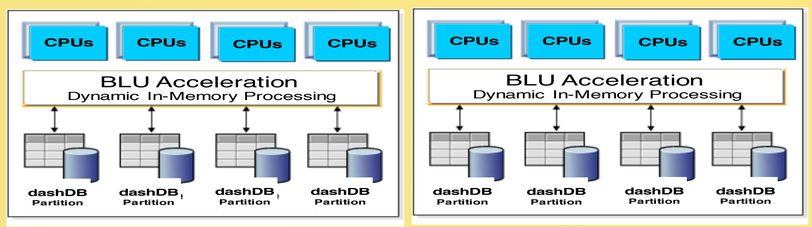
STATE of the ART (*i.e. elsewhere*)
- No dual SQL and JSON
- No true load-and-go simplicity
- No in-database analytics on cloud
- No polyglot cloud warehouses

SELECTED PAPERS:

1. Lanjun Wang, et al, **Schema Management for Document Stores**. PVLDB 8(9): 922-933 (2015)
2. Ronald Barber, et al, **Memory-Efficient Hash Joins**. PVLDB 8(4): 353-364 (2014)
3. Ronald Barber, et al, **In-memory BLU acceleration in IBM's DB2 and dashDB: Optimized for modern workloads and hardware architectures.** ICDE 2015: 1246-1252
4. Vijayshankar Raman, et al, **DB2 with BLU Acceleration: So Much More than Just a Column Store**. PVLDB 6(11): 1080-1091 (2013)
5. Eva Kwan, et al, **Automatic Database Configuration for DB2 Universal Database: Compressing Years of Performance Expertise into Seconds of Execution**. BTW 2003: 620-629

ARCHITECTURUAL FOUNDATIONS

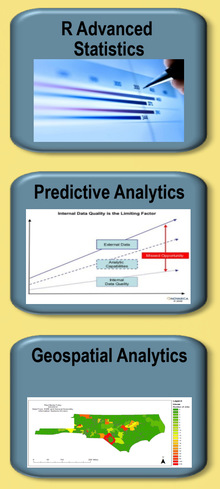
Elastic & Webscale



- Megabytes to Petabytes
- Elastic growth with hash partitioned shared nothing architecture
- MPP cluster can expand to 8X without rehashing data via initial over partitioning (Combined SMP and MPP parallelism exploited to ensure all cores leveraged)
- In-progress:
  - Scale to 10 PB
  - Load from on-premises at Terabyte-scale (e.g. 1TB/hr)

In-database analytics

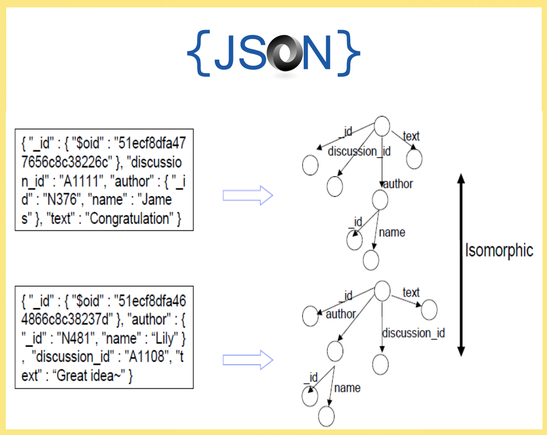
- Built in **R**, **predictive** and **spatial** analytics services
- Run within the memory space of the database
- Convert some algorithms to SQL and harness the scale of the query engine
- Provides 4-40X speedup over outside database analytics
- Memory limited by database memory, not the analytic engine



RECENT & ONGOING RESEARCH AREAS

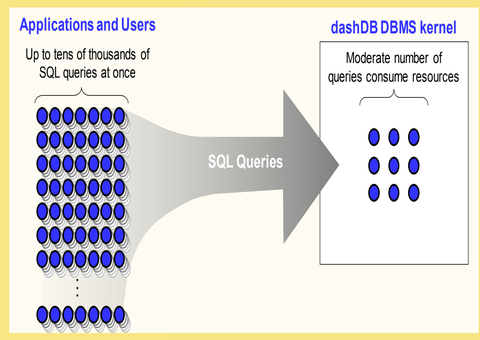
SQL and NoSQL Integrated

- Automatic schema discovery from JSON (see VLDB paper)
- Seamlessly integrated with Cloudant.com NoSQL
  - Pushbutton creation of data warehouse for JSON with continuous synchronization
- In-progress:
  - Schema discovery is slow. Improve by creating an index for real-time schema metadata discovery



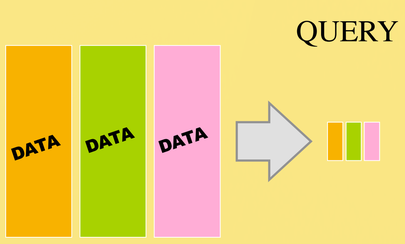
Load-and-Go Simplicity

- No tuning
  - Automatic detect RAM, cores
  - Apply heuristics to memory configuration, process model, database statistics etc.
- Admission control ensures manageable number of active queries do not over-consume resources
- Cloud provider manages everything other than table creation, data load, and user access control



Fast In-memory Columnar

- Leverages IBM's BLU Acceleration columnar query engine
  - Operations on compressed data
  - Parallel vector processing
  - In-memory optimized, with L3 & L2 cache optimized architecture, operates on strides of data.
- In progress:
  - 3X better performance through: a. Partitioned grouping. b. Reduced join overhead (see VLDB paper). c. New sort and OLAP implementation



Polyglot SQL

- Provides SQL and stored proc. language compatibility with many variants
- Reduction of SQL to common representation through query graph modeling
  - Oracle
  - PostgreSQL
  - DB2 for LUW
  - DB2 for z/OS
  - Netezza
- In-progress
  - Netezza PL/SQL