4th Workshop on Big Data Benchmarking

MPP SQL Engines: architectural choices and their implications on benchmarking

09 Oct 2013

Agenda:

Big Data Landscape

Market Requirements

Benchmark Parameters

Benchmark Wish List

Some Results

Who I am

Ravi Chandran, CTO & Co-Founder

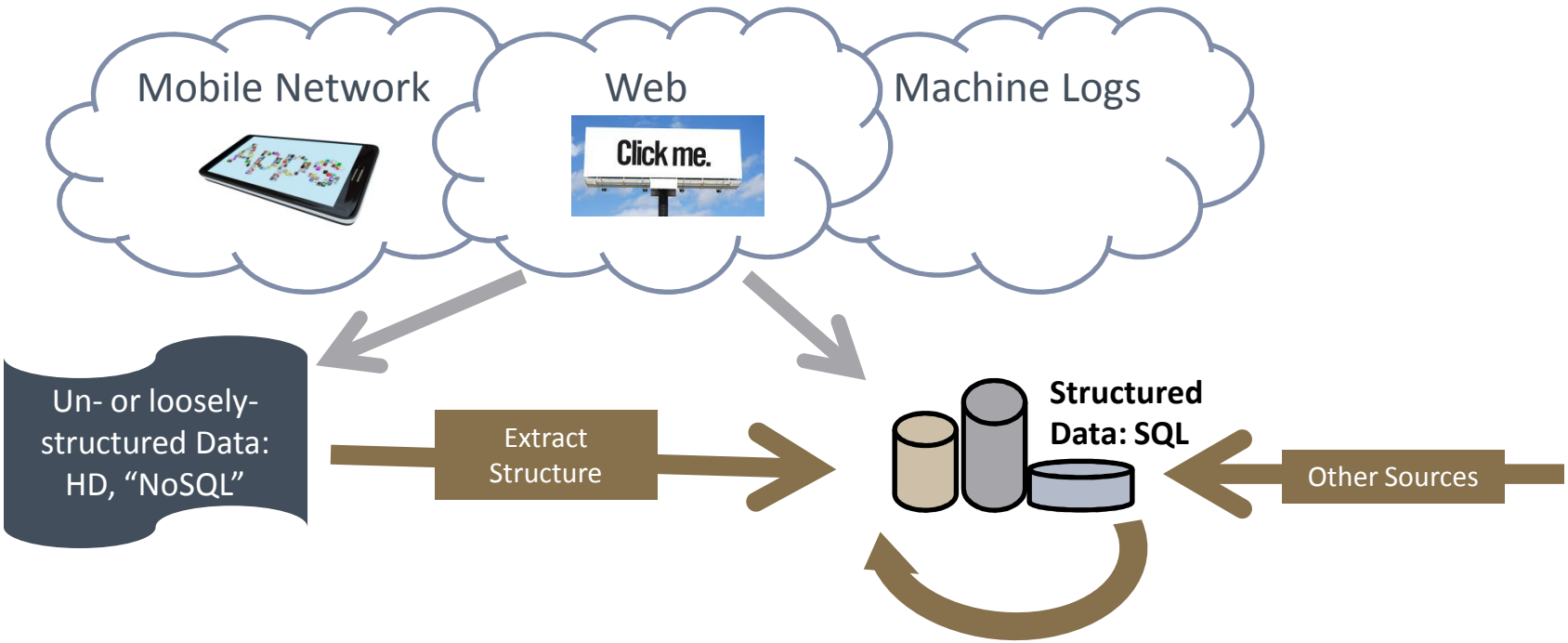
ravi.chandran@xtremedata.com

847-224-8907

Decades of experience designing high-performance computing systems:
ASICs, DSPs, x86 clusters, embedded RTOSes, Linux …

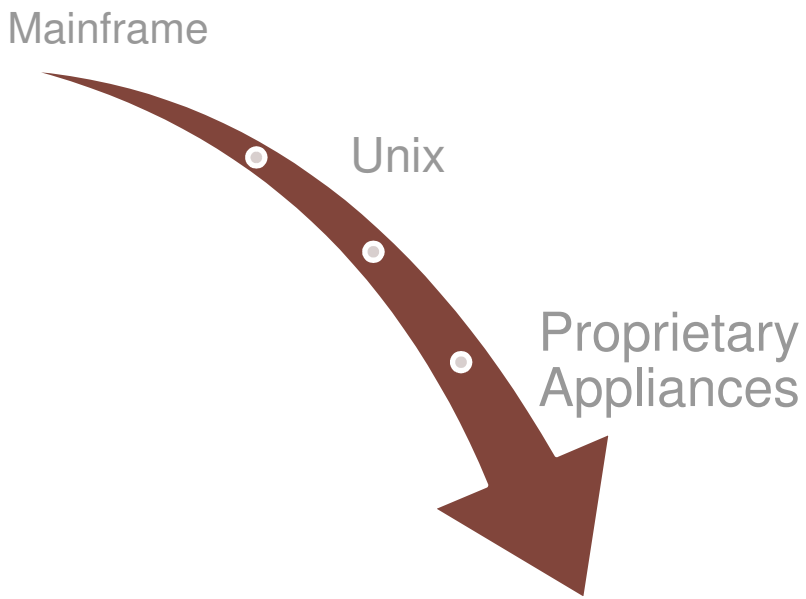
… definitely not a database benchmark expert!

Big Data Domain



- HD, NoSQL and SQL all have a place in big data domain
- SQL engines are good at integrating structured data from multiple sources and complex, iterative analytics
- Example: Twitter & Blog data captured in HD; key info extracted and fed into SQL analytic DataWarehouse

Big Data: Hardware Platform



- **Era of horizontally scalable commodity hardware, virtualization and cloud**
- **All big data solutions need to deploy on this infrastructure**

Hardware Infrastructure

Hardware infrastructure in today's converged data center: server-storage-network

- All servers today are (more or less) equal…
- All storage configurations (for big data) are equal - local attached, distributed
- All storage options are equal: multi-tiered: DRAM, Flash, Disk

- Network is much more interesting …
  - Step function, not continuous : 1 or 10 gigE, DDR or QDR IB …
  - Switches have a range of capabilities:
    - Bi-section bandwidth
    - Lane throttling / bandwidth guarantees
  - NICs also have a range of capabilities:
    - Lane throttling / bandwidth guarantees

Converged Data Center

Why SQL?

- Well-understood by 100's of 1000's worldwide
- Large, mature ecosystem of tools
- Portable code

Most important (from our POV):

- Declarative language: can parallelize/optimize at run-time

SQL: Requirements

SQL: requirements for big data analytics

- Run efficiently in parallel (MPP) on converged hardware
- Scale-out at large scale: distributed, shared-nothing architecture
- Tackle the "hard" problems at scale: Large table multi-way Joins, Group-Aggregates, Window Functions

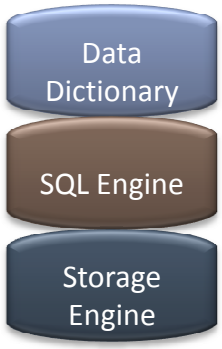
Example:

- In both Wall Street and Digital Advertising, there is a daily deluge of multiple data streams: Bids, Impressions, Clicks; Quotes, Orders, Trades, ….
- Not unusual for Billions of rows/day & TBs/day
- Need to correlate data between multiple streams: multi-way Joins

Approaches to MPP SQL

Components of a DB Engine:

- **postgreSQL:** common starting point for many of us
- Code is monolithic, single-threaded, single-node, …



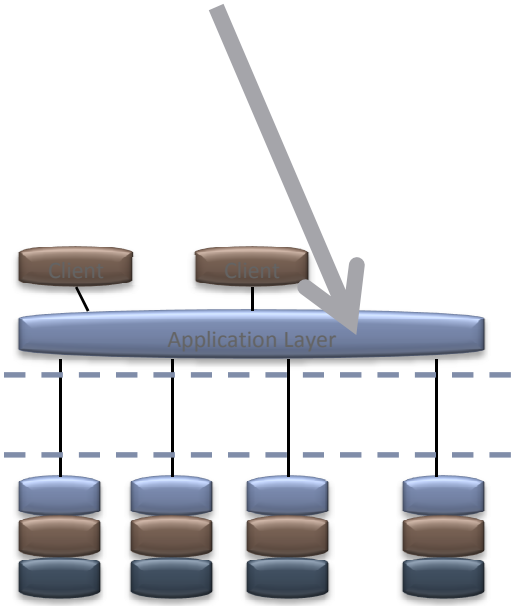
Challenges: How to increase scalability &performance of postgreSQL?

- Scalability via parallelism is possible at many different levels: Sharding, Federated, True-MPP, ..
- Performance improvement requires significant code re-write…

Parallelism

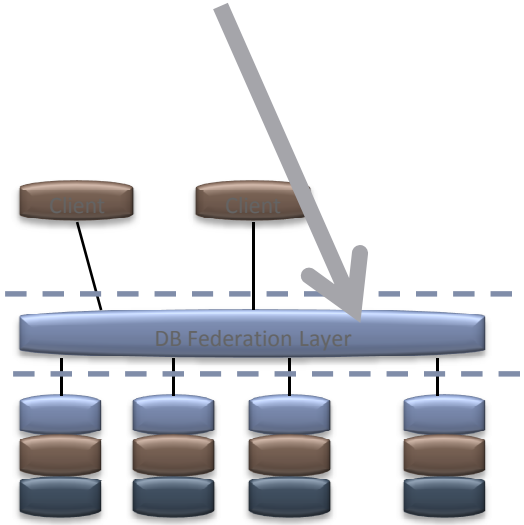
Sharding:

Multiple DBs unified at Application layer



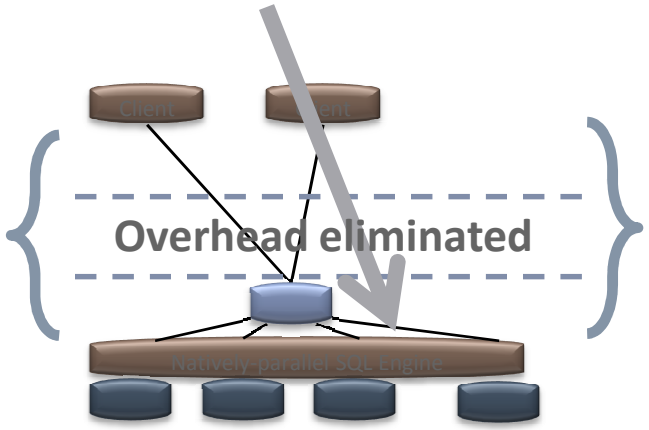
Federation:

Multiple DBs unified at DB Federation layer



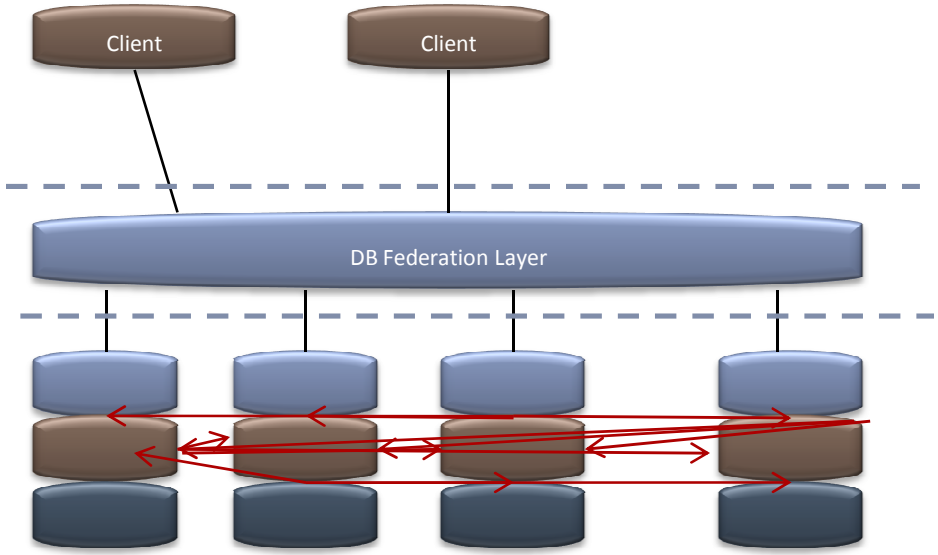
True MPP:

Single DB with distributed storage and SQL execution - **XtremeData**



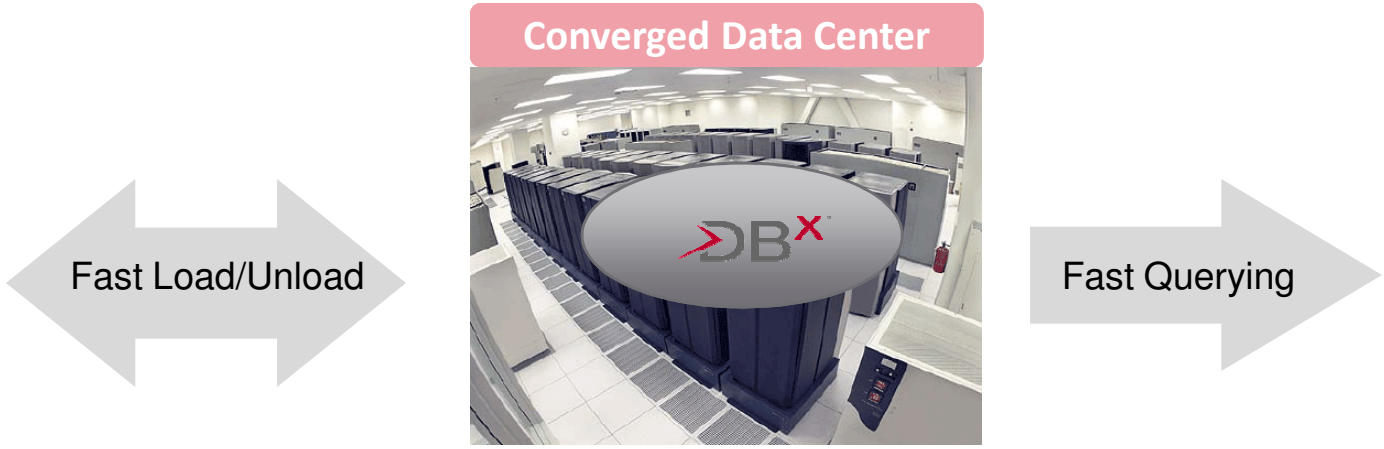
The higher the level at which parallelism is implemented, the higher the overhead: more hardware required to do the same job.

Parallelism: the N^2 problem



- Most general scenario: Join two large tables:
- Tables distributed across nodes at “random” w.r.t. Join key
- N-to-N re-distribution of data is unavoidable … N^2
- Okay for small N…
- If N is dictated by # of Cores, rapidly becomes unmanageable: 256 Cores in a single rack today …

dbX: Product Highlights



- Full-featured SQL, scale-out, deployable in the Cloud
- High-speed parallel ingest
- High-performance querying across multiple large tables
- Scale out to 100's of Nodes and 100's of TBs
- "Logical Node" concept – can be mapped to many physical configurations

Implications for benchmarks: H/W

Assessing underlying hardware:

- Servers:
  - CPU capability (# of Cores)
  - Memory size and bandwidth
  - Network bandwidth
- Storage:
  - # of tiers
  - size and bandwidth for each tier
- Network:
  - Topology: point-to-point latency
  - Switch bisectional bandwidth

Implications for benchmarks: S/W

Assessing SQL engine (big data analytics):

- Functionality:
  - SQL language support
  - Partitions, Indexes, Cursors, Window Functions

- Performance:
  - Parallel load from external source
  - Single table tests:
    - Scan-Filter-Complex compute
    - Group-Aggregate
    - Window Functions
  - Multi-table tests:
    - Joins
    - Joins + all of the above
  - Table creation within DB (CTAS) – for data-intensive, iterative processing

Implications Summary

Would be great if benchmark for big data analytics could:

- Combine assessment of hardware and SQL engine
- Scale DB size while holding system size constant
- Scale system size holding DB size constant

We have made an attempt at this … merely a starting point …

Our Benchmark

- Written completely in portable SQL
- Data generation and tests
- ~50 queries in ~6 groups
- Multiple DB sizes: typically use 6 scale factors
- L0:5 ranging from 0.33 to 10.56 TB for DB size

Largest Table:			
# Rows (000's)	Size, GB	DB Size, GB	ScaleFactor:
524,288	165.000	330.00	L0
1,048,576	330.000	660.00	L1
2,097,152	660.000	1,320.00	L2
4,194,304	1,320.000	2,640.00	L3
8,388,608	2,640.000	5,280.00	L4
16,777,216	5,280.000	10,560.00	L5

Deficiencies:

- Synthetic data with known, fixed distribution
- All tables have same column schema
- Each new table is 2x previous table
- SQL data generation (INSERT) – can be slow!

Proven useful for us, provides a lot of data on both hardware and software … but still, merely a starting point …

Benchmark developed by K.T.Sridhar & Sakkeer Ali of XtremeData

Sample Results: Hardware

- Benchmark has been run on many, many hardware platforms, a sample:
  - dbX-x: Commodity rack-mount servers with InfiniBand network
  - SeaMicro: dense mesh of Atom CPUs (http://www.seamicro.com/products/sm15K_overview)
  - HP-980: High-end HP DL980 8xCPU server plus SSD or SAN

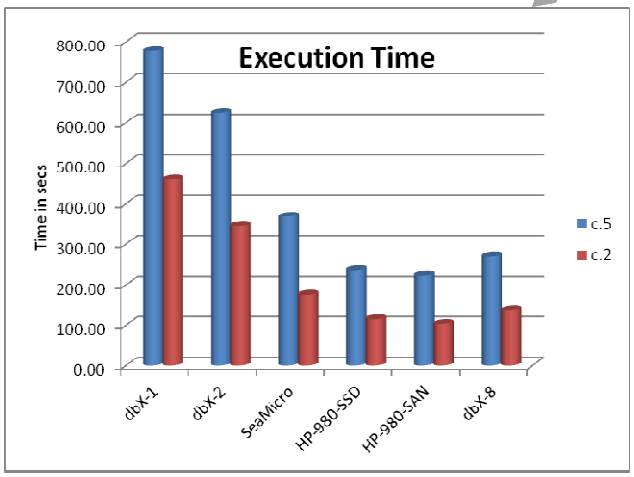
	dbX-1	dbX-2	SeaMicro	HP-980-SSD	HP-980-SAN	dbX-8
CPU type	Nehalem	Xeon	Atom	Xeon	Xeon	Opteron
Storage type	Direct-attached	Direct-attached	NAS	SSD	SAN	Direct-attached
Network	InfiniBand	InfiniBand	10GigE	Fiber	FC	InfiniBand

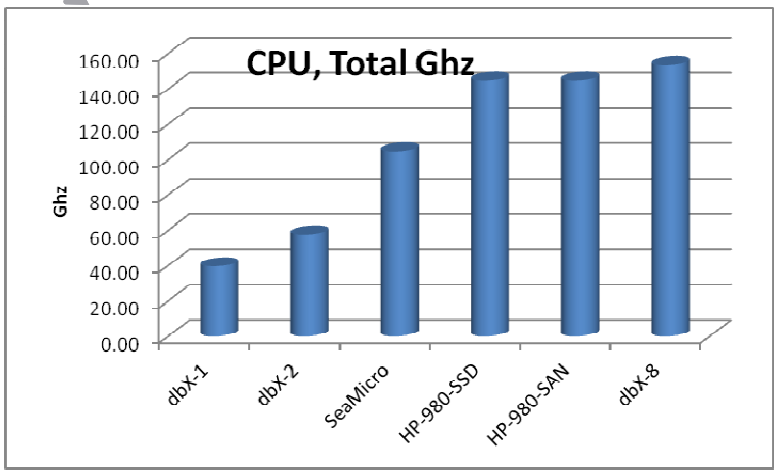
Sample Results: Hardware-CPU

- For CPU-limited queries (c2 and c5: Group, Join, Distinct in Benchmark), performance correlates well with CPU power.

	dbX-1	dbX-2	SeaMicro	HP-980-SSD	HP-980-SAN	dbX-8
CPU Core type	Nehalem	Xeon	Atom	Xeon	Xeon	Opteron
Clock Speed, Ghz	3.3	2.4	1.9	2.26	2.26	2.4
# Cores	6	6	1	8	8	4
# Sockets	2	4	55	8	8	16
Total, Ghz	**39.60**	**57.60**	**104.50**	**144.64**	**144.64**	**153.60**

Reasonable correlation



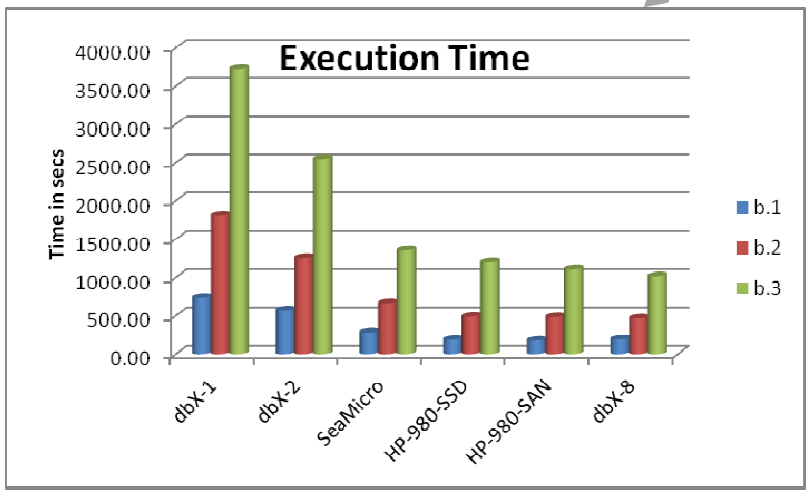


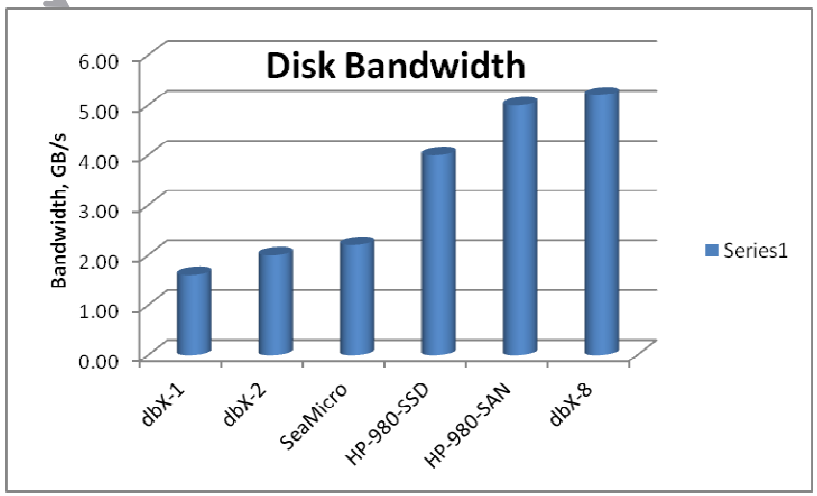
Sample Results: Hardware-Disk

- For Disk-limited queries, (b1, b2 and b3: 2, 3 and 4 Tables Joins in Benchmark), performance correlates well with Disk bandwidth.

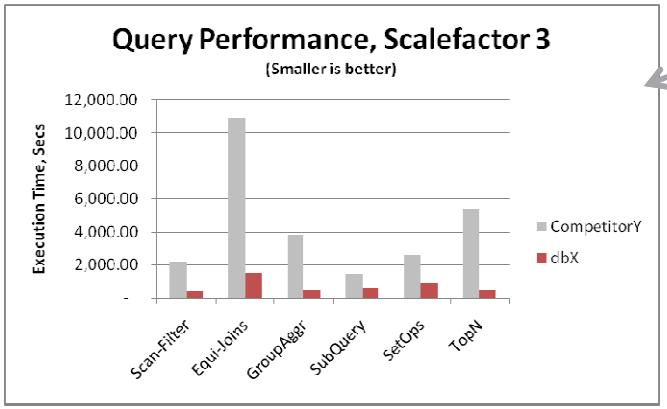
	dbX-1	dbX-2	SeaMicro	HP-980-SSD	HP-980-SAN	dbX-8
Disk Bandwidth, GB/s	1.60	2.0	2.20	4.00	5.00	5.20

Reasonable correlation





Sample Results: Software

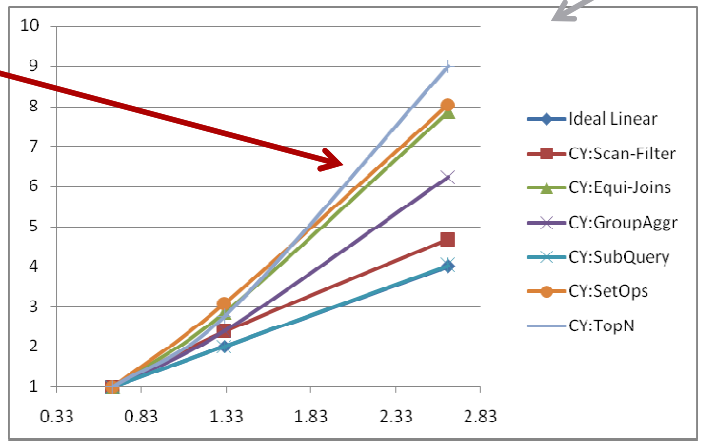


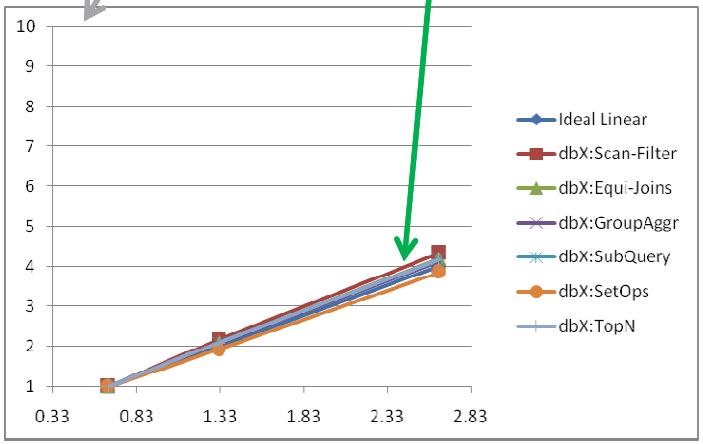
Raw performance at one DB scale factor

Relative performance across DB scale factors

Clustering close to Ideal is good

Divergence from Ideal is bad!





Query Time, Normalized to Scale Factor 0

DB Size, TB

Thank You

Questions?

Ravi Chandran, CTO & Co-Founder

ravi.chandran@xtremedata.com

847-224-8907