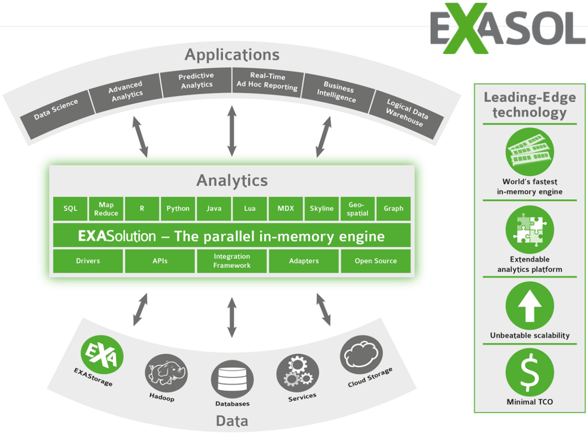
EXASOL

Query Processing in EXASolution

stefan.mandl@exasol.com

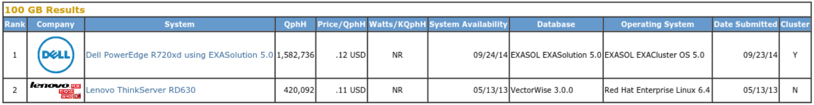
June 09 2015



- Distributed and parallel standard compliant SQL database system (Transactions, ACID, Backup and Recovery, . . .)
- Optimized for analytical workloads
- . . . and it is fast

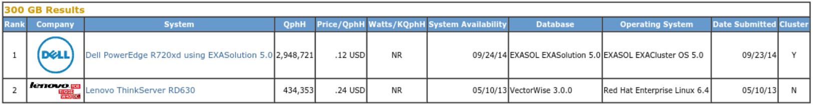
How fast? ⇒ http://www.tpc.org/tpch/[1]

(TPC-H is an ad-hoc, decision support benchmark.)



100 GB Results

Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	1,582,736	.12 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	lenovo FOR THOSE WHO DO	Lenovo ThinkServer RD630	420,092	.11 USD	NR	05/13/13	VectorWise 3.0.0	Red Hat Enterprise Linux 6.4	05/13/13	N



300 GB Results

Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	2,948,721	.12 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	lenovo FOR THOSE WHO DO	Lenovo ThinkServer RD630	434,353	.24 USD	NR	05/10/13	VectorWise 3.0.0	Red Hat Enterprise Linux 6.4	05/10/13	N

[1] Last checked on June 7th 2015

1,000 GB Results

Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	5,246,338	.14 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	CISCO	Cisco UCS C460 M4 Server	588,831	.97 USD	NR	12/16/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard	12/15/14	N
3	inspur	INSPUR K1	585,319	3.42 CNY	NR	09/04/14	Actian Analytics Database - Vector 3.5.1	K-UX2.2	09/03/14	N
4	IBM	IBM System x3850 X6	519,976	1.36 USD	NR	04/16/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard	04/15/14	N
5	inspur	INSPUR K1	485,242	4.03 CNY	NR	06/04/14	Actian Vector 3.0.0	K-UX2.2	06/03/14	N
6	hp invent	DL380 Gen9	390,590	.97 USD	NR	09/08/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard	09/07/14	N
7	FUJITSU	SPARC M10-4S	326,454	1,524.25 JPY	NR	02/07/14	Oracle Database 11g R2 Enterprise Edition w/Partitioning	Oracle Solaris 11.1	02/06/14	N
8	CISCO	Cisco UCS C240 M3 Server	304,361	.73 USD	NR	08/20/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	08/19/14	N
9	HUAWEI	Huawei FusionCube v2.01	258,474	7.08 USD	NR	12/01/13	Sybase IQ 16.0 SP02	Red Hat Enterprise Linux 6.2	11/16/13	Y

3,000 GB Results

Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	7,808,386	.15 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	lenovo	Lenovo System x3850 X6	700,392	.99 USD	NR	05/26/15	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows 2012 R2 Standard Edition	05/01/15	N
3	hp invent	DL580 G8	461,837	2.04 USD	NR	04/16/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard Edition	04/15/14	N
4	ORACLE	SPARC T5-4 Server	409,721	3.94 USD	NR	09/24/13	Oracle Database 11g R2 Enterprise Edition w/Partitioning	Oracle Solaris 11.1	06/07/13	N
5	CISCO	Cisco UCS C420 M3 Server	230,119	1.29 USD	NR	12/30/13	Sybase IQ 16.0 SP02	Red Hat Enterprise Linux 6.4	10/31/13	N

10,000 GB Results										
Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	10,133,244	.17 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	lenovo FOR THOSE WHO DO	System x3950 X6	652,239	2.43 USD	NR	04/07/15	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard	04/06/15	N
3	hp invent	HP Proliant DL580 Gen9	606,821	1.82 USD	NR	05/05/15	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard	05/04/15	N
4	hp invent	DL580 G8	404,005	2.34 USD	NR	04/16/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 R2 Standard Edition	04/15/14	N
5	ORACLE	SPARC T5-4 Server	377,594	4.65 USD	NR	11/26/13	Oracle Database 11g R2 Enterprise Edition w/Partitioning	Oracle Solaris 11.1	11/25/13	N
6	hp invent	HP ProLiant DL980 G7	158,108	6.49 USD	NR	04/15/13	Microsoft SQL Server 2012 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	04/15/13	N

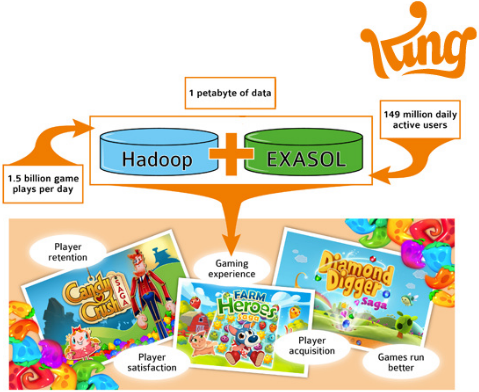
30,000 GB Results										
Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	11,223,614	.23 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y

100,000 GB Results										
Rank	Company	System	QphH	Price/QphH	Watts/KQphH	System Availability	Database	Operating System	Date Submitted	Cluster
1	DELL	Dell PowerEdge R720xd using EXASolution 5.0	11,612,395	.37 USD	NR	09/24/14	EXASOL EXASolution 5.0	EXASOL EXACluster OS 5.0	09/23/14	Y
2	HITACHI	Hitachi BladeSymphony BS2000	82,678	18,912.00 JPY	NR	10/20/13	Hitachi Advanced Data Binder 01-02	Red Hat Enterprise Linux 6.2	10/19/13	Y

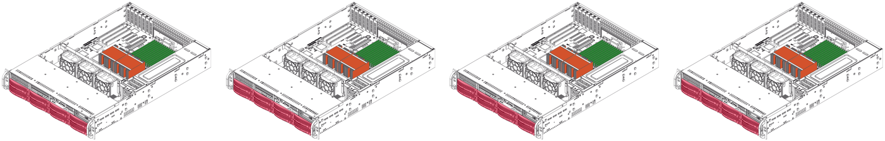
EXASOL History

- 90ies: Research Project at University of Erlangen-Nürnberg
- 2000: Company is founded
- 2006: Pilot customer Karstadt-Quelle uses EXASolution in production
- 2008: Many new records in TPC-H benchmark
- 2010: Most successful vendor on analytical database systems in Germany (BARC)
- 2012: Inclusion in Gartner's "Magic Quadrant for Data Management Systems"
- 2014: 80 customers in 12 countries, 100 TB TPC-H benchmark

> “EXASOL has solved problems for us that were previously deemed intractable. The investment has rapidly paid off and delivered value to our business. Our users get answers to their business queries in a fraction of the time it took previously.”
>
> — Andy Done, Data Platform Lead at King



EXASolution Environments



- A cluster of standard servers
  - Standard server hardware: 2 Quad/Hexa/Ten Core CPUs, 16–786 GB RAM, 2–24 SAS/SATA HDD, GBit Ethernet (1GiB, 10GiB), Free vendor choice: Dell, HP, IBM, FSC, Oracle (Sun), . . .
  - Our own cluster operating system: EXAClusterOS
- Cloud Platforms
  - EXACloud
  - BigStep (`bigstep.com`)
  - Soon: Azure (`http://azure.microsoft.com/`)
- EXAOne (commercial one node version)
- EXASolo (free-to-use one node VM image, DBRAM limitation)

Query Processing

- Basic Ideas
- Compiler
- Engine
- Data

Basic ideas

- massive parallel processing: utilize all the available cores in every single machine
- distributed processing: execute even single queries on all available machines at the same time (SPMD paradigm)
- avoid synchronization (no master node, no main execution thread)
- low-level hardware characteristics matter
- EXASOL's in-memory design principle:
  - Algorithms are written, *as if* all data is stored in RAM
  - Clever machinery takes care of guaranteeing this assumption (most of the time)

The Compiler

EXASOL SQL

- SQL
  - In practice, many versions of SQL
  - EXASOL's version is designed for compatibility
  - Query preprocessor enables users to adopt the input lanuage
- Standard Compiler Architecture
  1. Tokenizer: String → Tokens
  2. (If preprocessor script is defined: Tokens → SQL Tokens)
  3. Parser: SQL Tokens → SQL Syntax Tree
  4. Analyzer: SQL Syntax Tree → Query Graph
  5. (check if the result is cached)
  6. Optimizer: (Query Graph$_i$ → Query Graph$_{i+1}$)*, finally creates Execution Graph

Optimization and Statistics

- Rule based optimization
  - Evaluate constants
  - Remove empty tables (**WHERE** 0=1)
  - Move conditions from **HAVING** to **WHERE**
  - Push filters into sub-selects
  - . . .
- Cost based optimization of Join order
  - Table/column statistics
  - Estimation of filter selectivity
  - ⇒ Try to minimize the size of intermediate results
  - Statistics are always up-to-date

EXASolution by-and-large is tuning free
EXASOL considers sub-optimal execution graphs like bugs!

Table Replication and Index Creation

- Tables are distributed across all machines in the cluster
- Small tables may additionally be replicated completely on every machine → some operations are much faster
- Trade-off: replicating arbitrary large tables defies the purpose of a distributed database
- BUT: typical BI schemas (star, snowflake) involve very large and very small tables ⇒ replication allows to perform Joins without global communication
- Indexes are created, updated, and deleted completely automatically as seen fit by the optimizer as many algorithms in databases can be implemented efficiently with index-based approach.

The Engine

Parallel Processing via Pipelining

- Reminder: The compiler generates an Execution Graph
- Execution Graphs typically consist of several Excution Pipelines which are executed one after the other
- Each Execution Pipeline consists of a number of Pipeline Stages

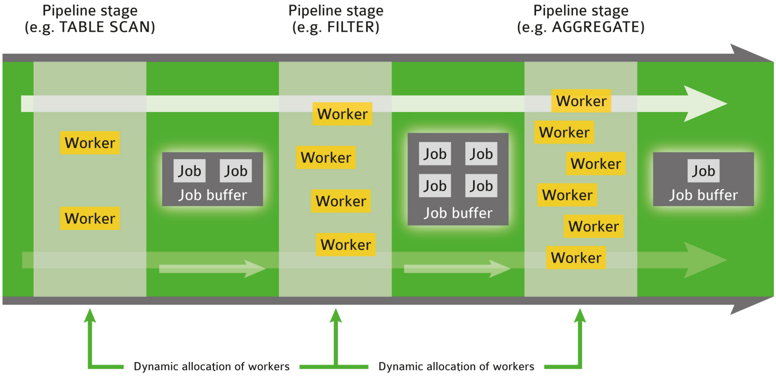
Main concepts for Execution Pipelines:

- Pipeline Stage: represents an operation to be performed
- Pipeline Job: encapsulates the state of computations
- Job buffer: Enables asynchronous control flow between Pipeline Stages
- Worker: basically a thread of execution
- Scheduler: Allows to adapt execution behaviour at runtime!

Provides a powerful abstraction which allows to describe local (per-node) parallel execution.

Pipeline for `SELECT sum(y) FROM T WHERE x > 5;`

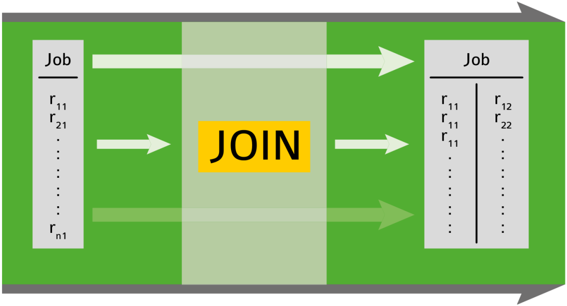
Pipeline



Pipeline Jobs

- Most jobs reference data via row number
- Data access is handled by lower levels
- Only load data that is actually needed!
- → index based filtering can eliminate many rows

Jobs before and after a Join operation



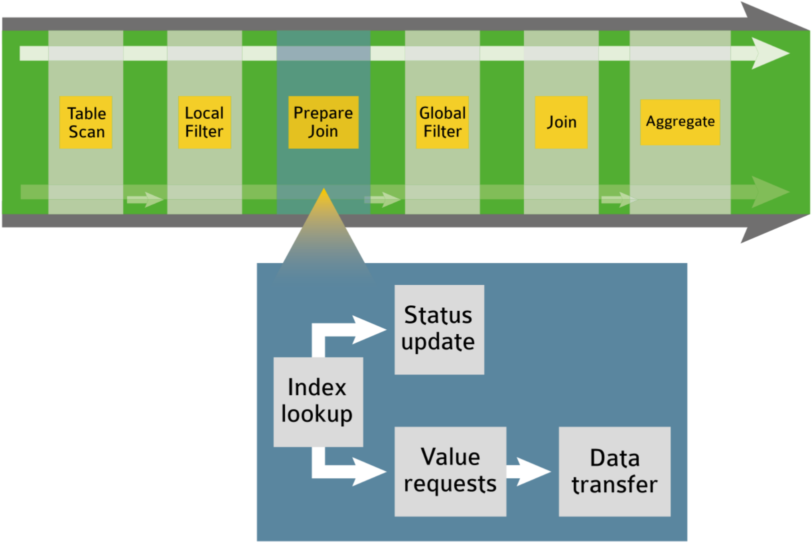
Distributed Processing

Distributed Pipelines

- Create jobs on remote machines
- Jobs may reference data on different machines
- Data can be transfered on demand
- Various communication primitives (sync, global AND, global OR, 1:1, 1:n message passing)
- Distributed exception handling
- Most distributed algorithms in EXASolution do not need to track jobs they create on remote machines!

Distributed Processing – 2

Example: Global Join (both tables are distributed)



The Data

Columns and Tables: Columnar layout

- Tables are stored *column-oriented*

+ Keeps only columns in RAM which are actually needed

+ Column values are ‘close to each other’ ⇒ hardware caching!

+ Column values have the same struture which enables sophisticated column-local compression techniques (e.g. `ALTER TABLE ADD COLUMN ...` does not influence compression)

– INSERT, UPDATE, and DELETE are harder to implement efficiently, but: EXASolution uses

  - INSERT buffers
  - DELETE markers

  for tables and *indexes*

Columns and Tables: Horizontal partitioning

Tables are partitioned horizontally across all machines

→ For each table, every machine holds a subset of the rows

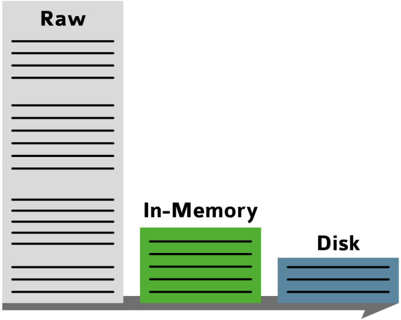
+ A given query is executed on all machines

+ Many algorithms can be written in a cluster-agnostic style

! Much better than vertical partitioning, where machines wich do not contain relevant columns are idle

Compression

- Compression in EXASolution is applied per-column
- In-memory compression is dictionary based (single elements can still be accessed)
- On-Disk compression is block-based
- Raw values are needed for complex computations
- Equi-Joins and some filters can be performed with in-memory compressed values
- On-Disk compression is used for creating backups



Block Management

Remember the in-memory design principle!

- Data is in memory and if not, has to be read before access
- Our block management is basically the same as:

```
void* column = mmap(0,1024,PROT_READ|PROT_WRITE,
                                  MAP_SHARED,fd,0);
```

  but highly tuned for EXASolution

- Hence, the acutally available RAM does not not limit functionality, but performance
- In addition of putting data in RAM, make sure that processor caches are utilized
- Typical RAM size: 10% of the raw data size.
- Transactions are mainly handled at this level
- We usually do not use files but our own distributed storage system: EXAStorage

Wrap-up

Why is EXASolution fast for analytical workloads?

- We cover the complete stack:
  - EXAClusterOS
  - EXAStorage
  - In-memory technology
  - Parallel Processing: avoid locks
  - Distributed Processing: avoid synchronization
- We pay attention to low level details (with $\mathcal{O}(n)$ algorithms, the linear factors matter!)
- A number of tricks we learned over the last 15 years
  ... want to know more? ⇒ yes, we are hiring!

Free Download (EXASolo)
`https://www.exasol.com/portal/display/DOWNLOAD/Free+Trial`

Let's take a look ...