Achieving 100,000 Transactions Per Second with a NoSQL Database

Eric David Bloch
@eedeebee

19 jun 2012

A bit about me

- I've written software used by millions of people.

  Apps, libraries, compilers, device drivers, operating systems

- This is my second QCon and my first talk
- I'm the Community Director at MarkLogic, last 2 years.
- Born here in NY in 1965; now in CA
- I survived having 3 kids in less than 2 years.

Me, 1967


PHOTOS / PEGGY ZARNEK—PHOTO IDEAS
TOWERTOWN QUALITY
ERIC
BLOCH
BORN: 4-14-65
HEIGHT: 36"
WEIGHT: 29 lbs.
SIZE: 2-3 Toddler
SHOE: 6½
EYES: Dark Brown
HAIR: Dark Brown
PIX DATE: Apr. '67

Me, today

Musical Form for the Talk

A: Why?

B: How?

- Whirl-wind database architecture tour
- Melody from part A again
- Techniques to get to 100K

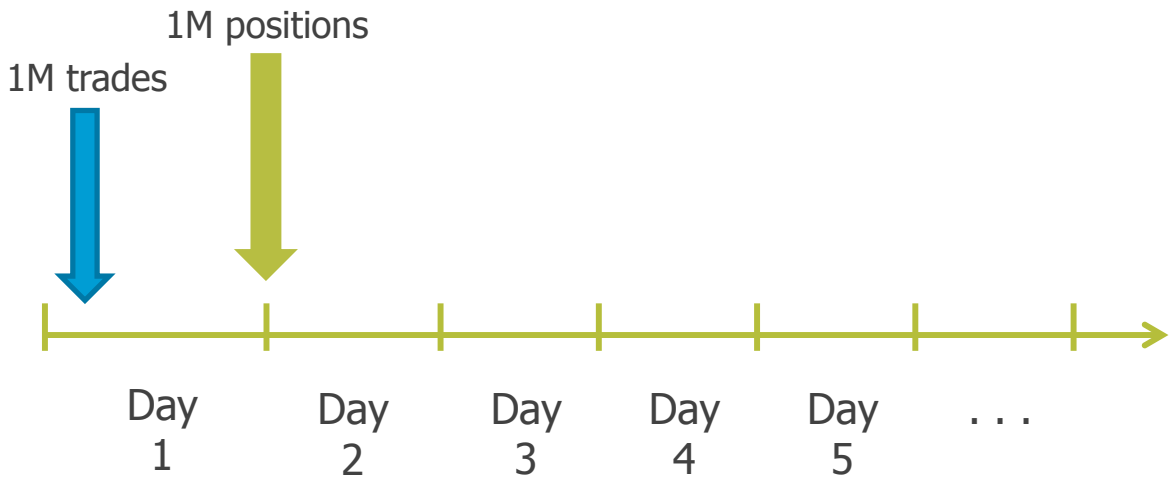

20
JG 26182531
G7

- It's about money.

- Top 5 bank needed to manage trades
- Trades look more like documents than tables
- Schemas for trades change all the time
- Transactions
- Scale and velocity ("Big Data")

Trades and Positions

- 1 million trades per day
- Followed by 1 million position reports at end of day
  - Roll up trades of current date for each "book, instrument" pair
  - Group-by, with key = "date, book, instrument"



Trades and Positions

```
<trade>
    <quantity>8540882</quantity>
    <quantity2>1193.71</quantity2>
    <instrument>WASAX</instrument>
    <book>679</book>
    <trade-date>2011-03-13-07:00</trade-date>
    <settle-date>2011-03-17-07:00</settle-date>
</trade>


<position>
    <instrument>EAAFX</instrument>
    <book>679</book>
    <quantity>3</quantity>
    <business-date>2011-03-25Z</business-date>
    <position-date>2011-03-24Z</position-date>
</position>
```

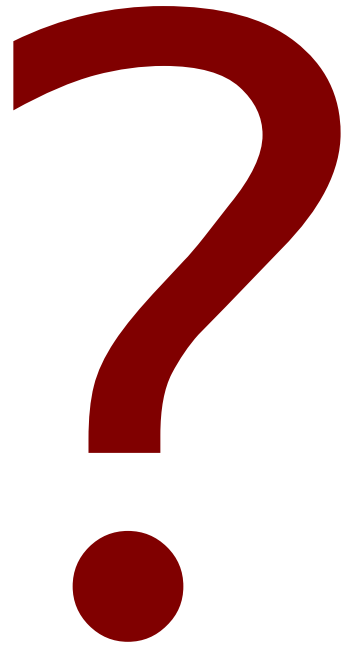
Now show us

- 15K inserts per second
- Linear scalability

Requirements

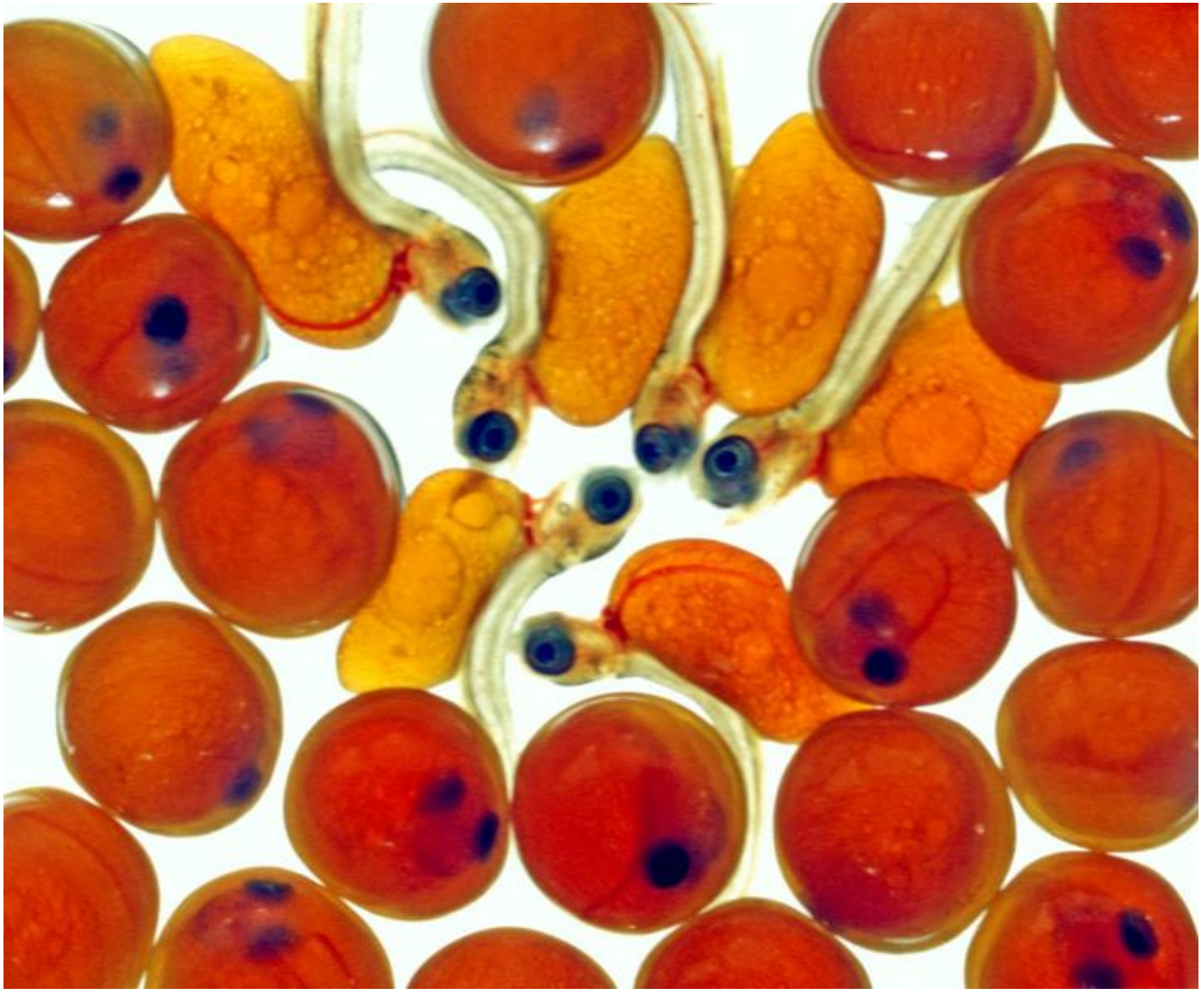
NoSQL flexibility,
performance & scale

Enterprise-grade
transactional guarantees



What NoSQL, OldSQL, or NewSQL database out there can we use?


APACHE
HBASE
VoltDB
riak
Neo4j
the graph database
mongoDB
PostgreSQL
Apache
CouchDB
Cassandra
ORACLE



What is MarkLogic?

- Non-relational, document-oriented, distributed database
  - Shared nothing clustering, linear scale out
  - Multi-Version Concurrency Control (MVCC)
  - Transactions (ACID)
- Search Engine
  - Web scale (Big Data)
  - Inverted indexes (term lists)
  - Real-time updates
  - Compose-able queries

MarkLogic

Architecture

- Data model
- Indexing
- Clustering
- Query execution
- Transactions

Question

What does data model have to do with scalability?

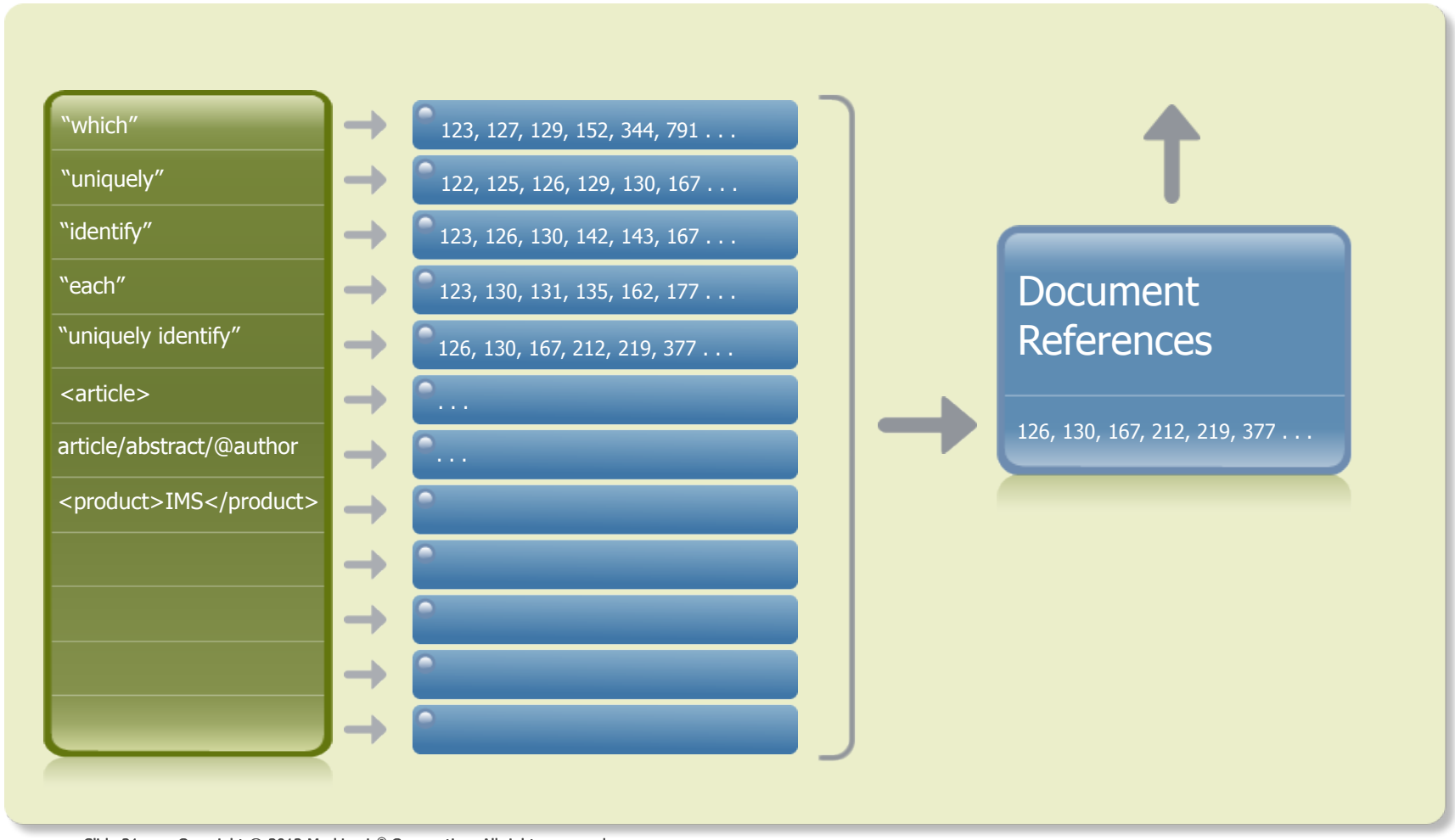
MarkLogic

Key (URI)	Value (Document)
/trade/153748994	<trade> <id>8</id> <time>2012-02-20T14:00:00</time> <instrument>BYME AAA</instrument> <price cur="usd">600.27</price> </trade>

/user/eedeebee	{
	  "name" : "Eric Bloch",
	  "age" : 47,
	  "hair" : "gray",
	  "kids" : ["Grace", "Ryan", "Owen"]
	}

/book5293	It was the best of times, it was the worst of times, it was the age of wisdom, ...
/2012-02-20T14:47:53/01445	.mp3 .avi [your favorite binary format]

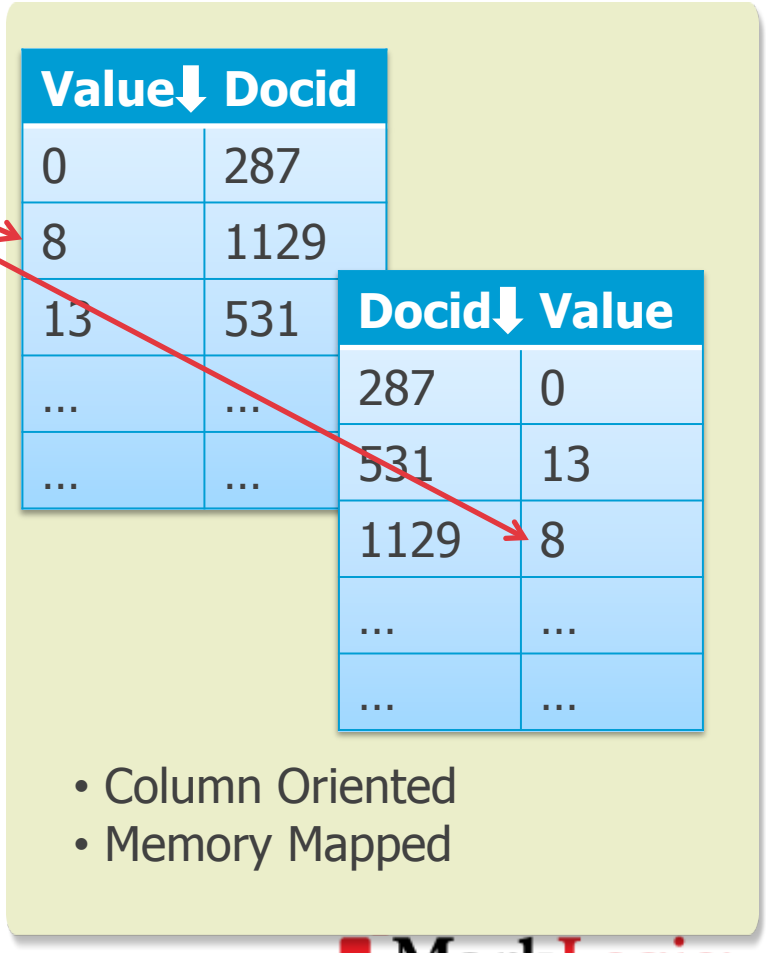
Inverted Index



Range Index

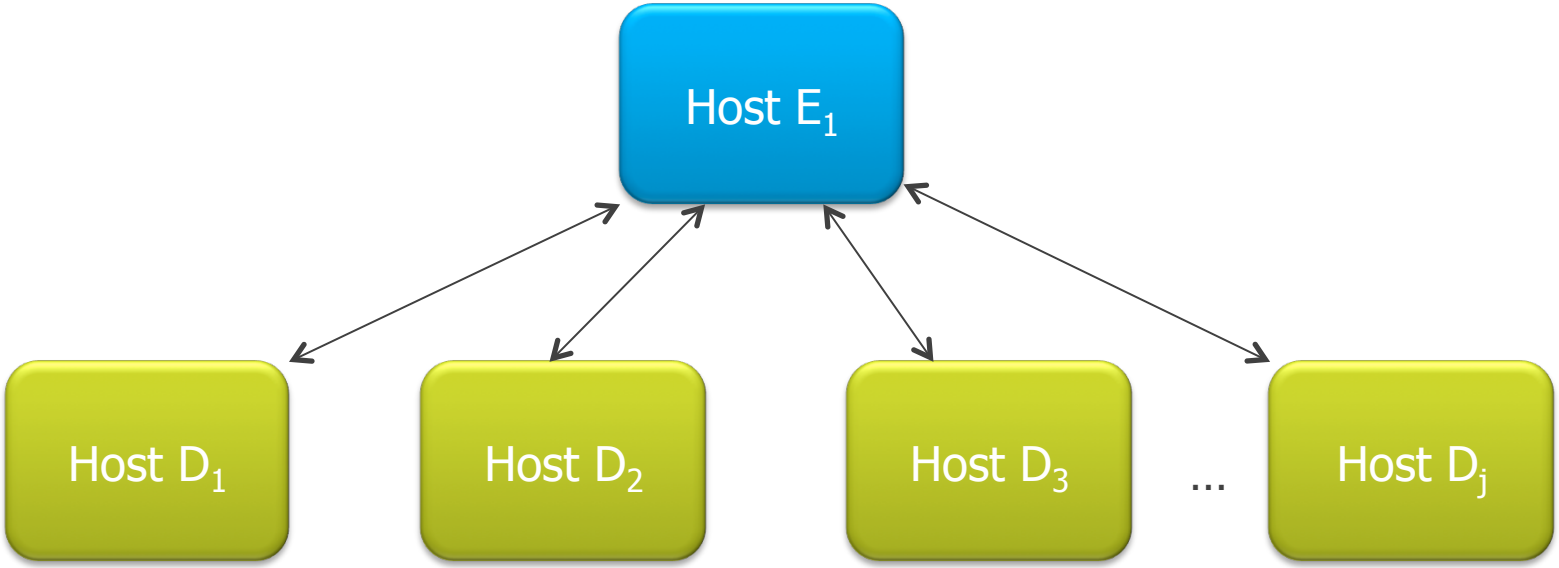
Rows

```
<trade>
  <trader_id>8</trader_id>
  <time>2012-02-20T14:00:00</time>
  <instrument>IBM</instrument>
  …
</trade>

<trade>
  <trader_id>13</trader_id>
  <time>2012-02-20T14:30:00</time>
  <instrument>AAPL</instrument>
  …
</trade>

<trade>
  <trader_id>0</trader_id>
  <time>2012-02-20T15:30:00</time>
  <instrument>GOOG</instrument>
  …
</trade>
```



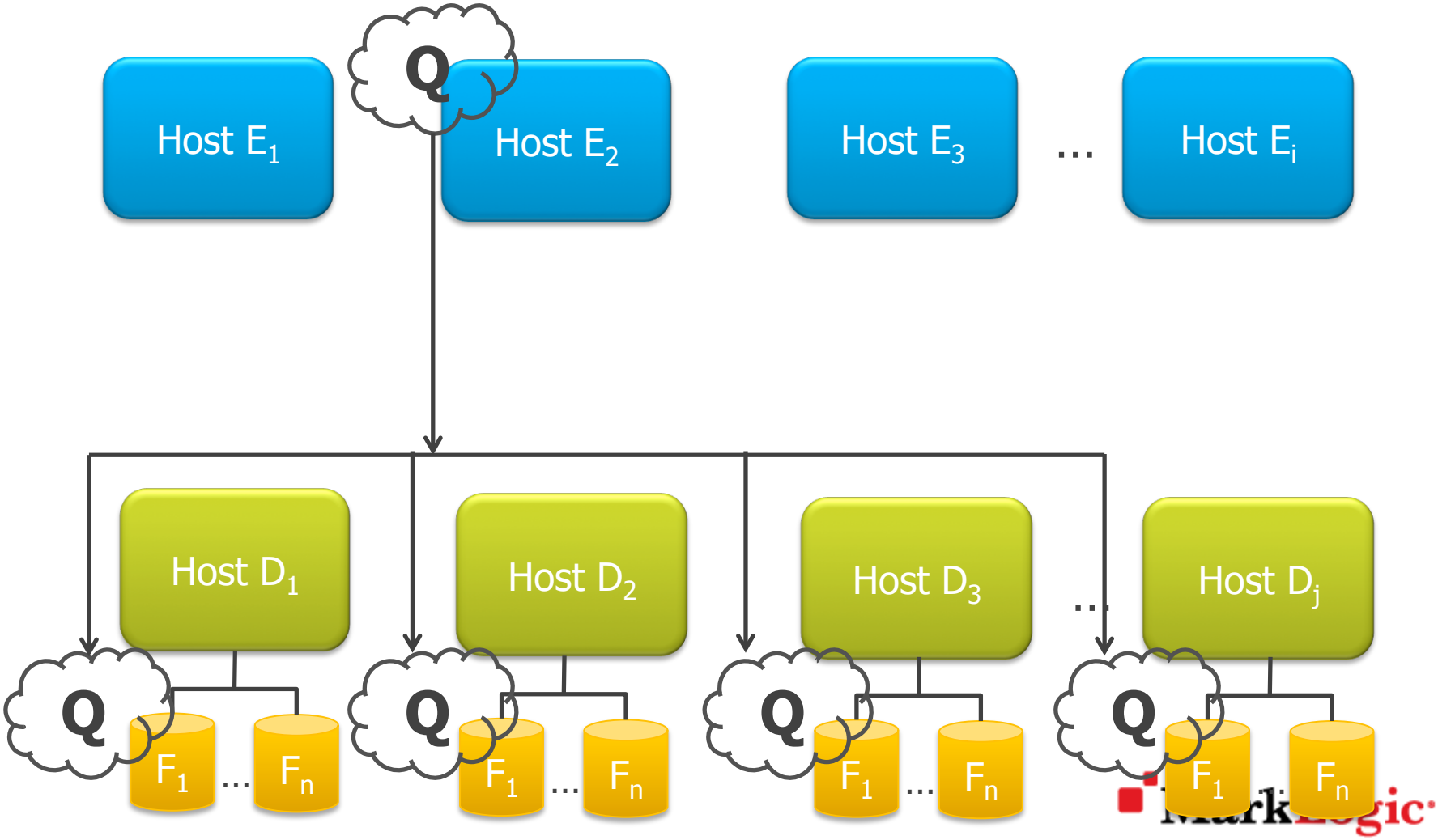
MarkLogic

Shared-Nothing Clustering

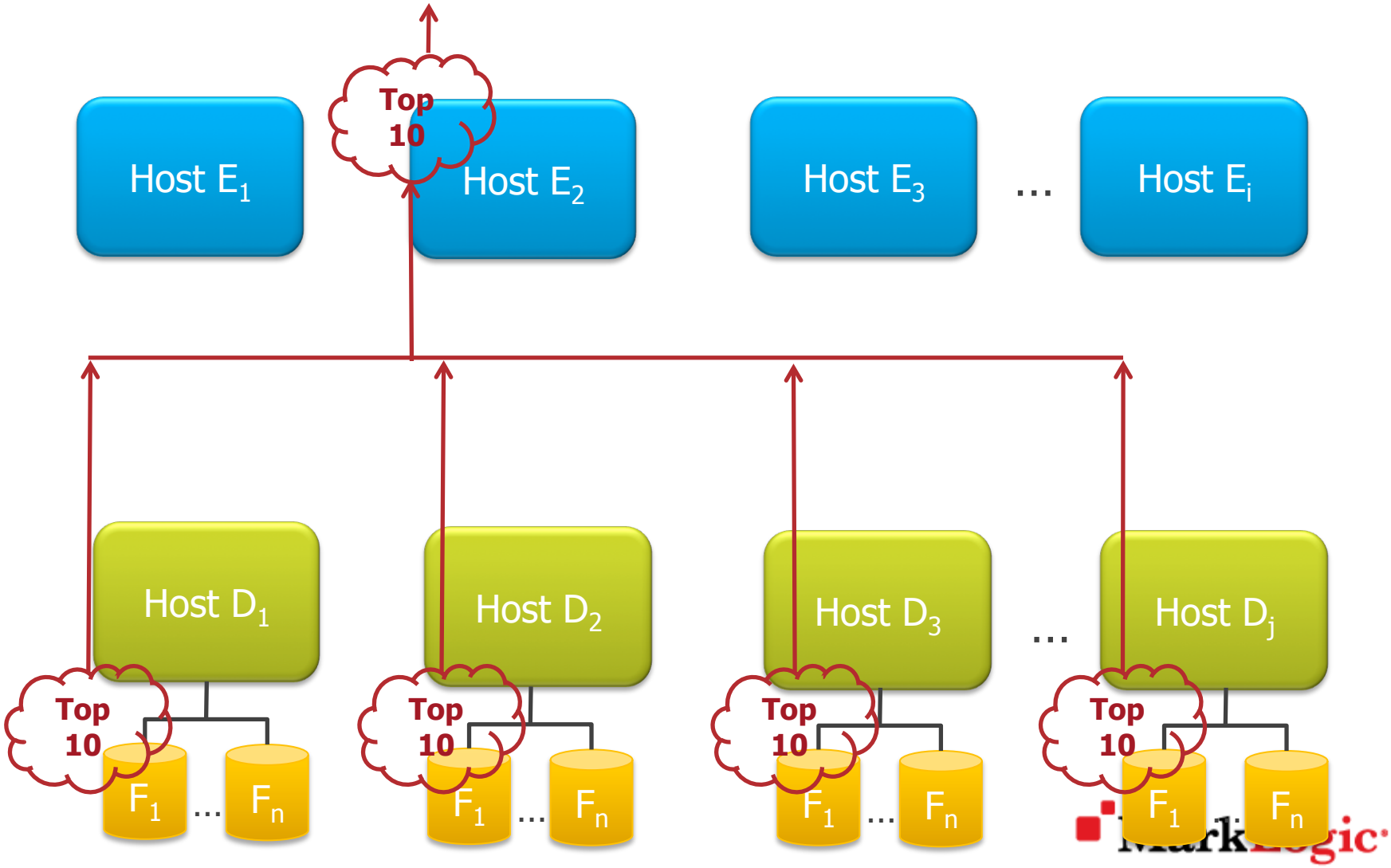


MarkLogic

Query Evaluation – "Map"



Query Evaluation – "Reduce"



Queries/Updates with MVCC

- Every query has a timestamp
- Documents do not change

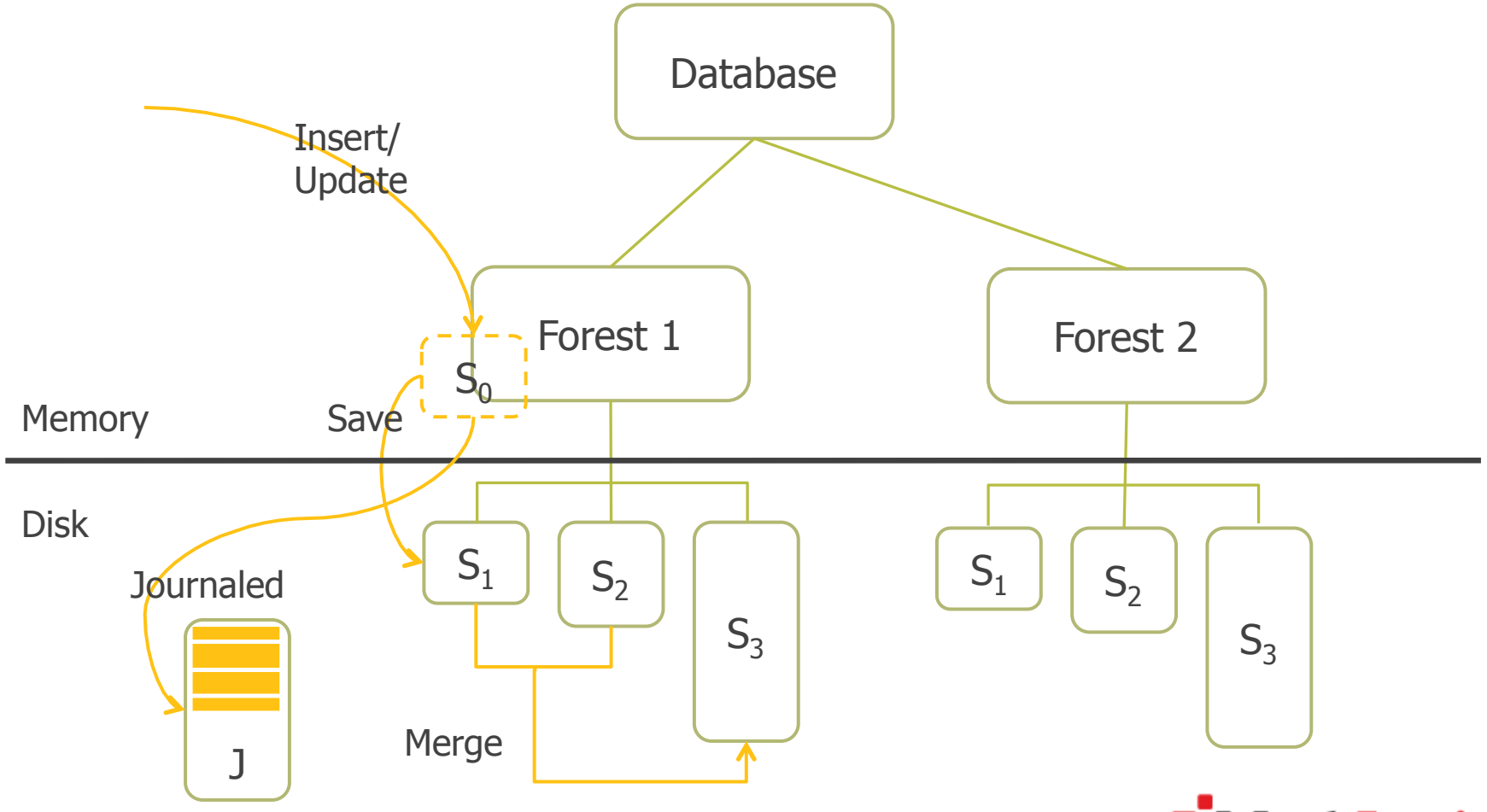
- Reads are lock-free
- Inserts – see next slide
- Deletes – mark as deleted
- Edits –
  - copy
  - edit
  - insert the copy
  - mark the original as deleted

Insert Mechanics

1) New URI+Document arrive at E-node
2) URI Probe – determine whether URI exists in any forest
3) URI Lock – write locks taken on D node(s)
4) Forest Assignment – URI is deterministically placed in Forest
5) Indexing
6) Journaling
7) Commit – transaction complete
8) Release URI Locks – D node(s) are notified to release lock

MarkLogic

Save-and-merge
(Log Structured Tree Merge)

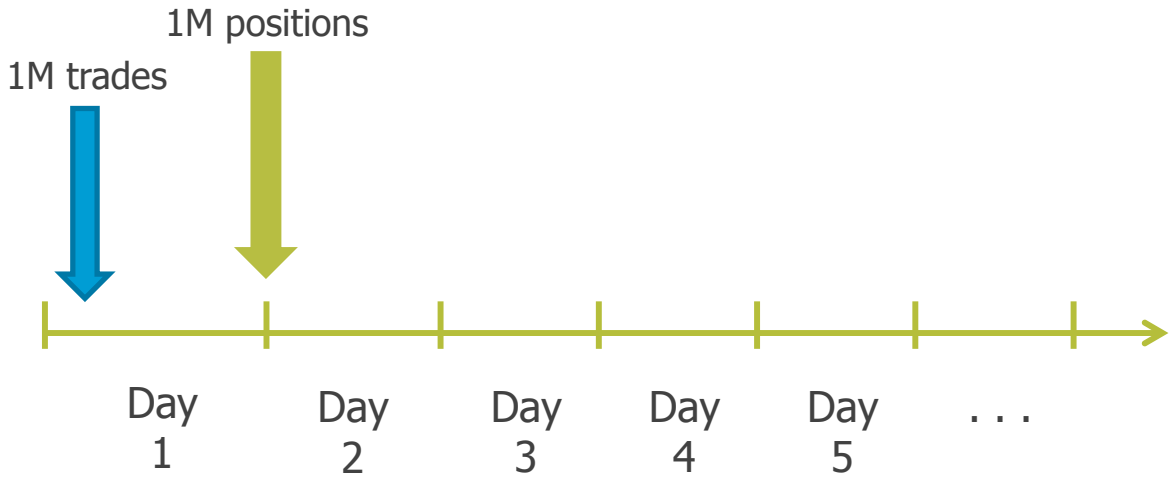


MarkLogic

Back to
the money

Trades and Positions

- 1 million trades per day
- followed by 1 million position reports at end of day
  - Roll up trades of the current "date" for each "book:instrument" pair
  - Group-by, with key = "book:date:instrument"



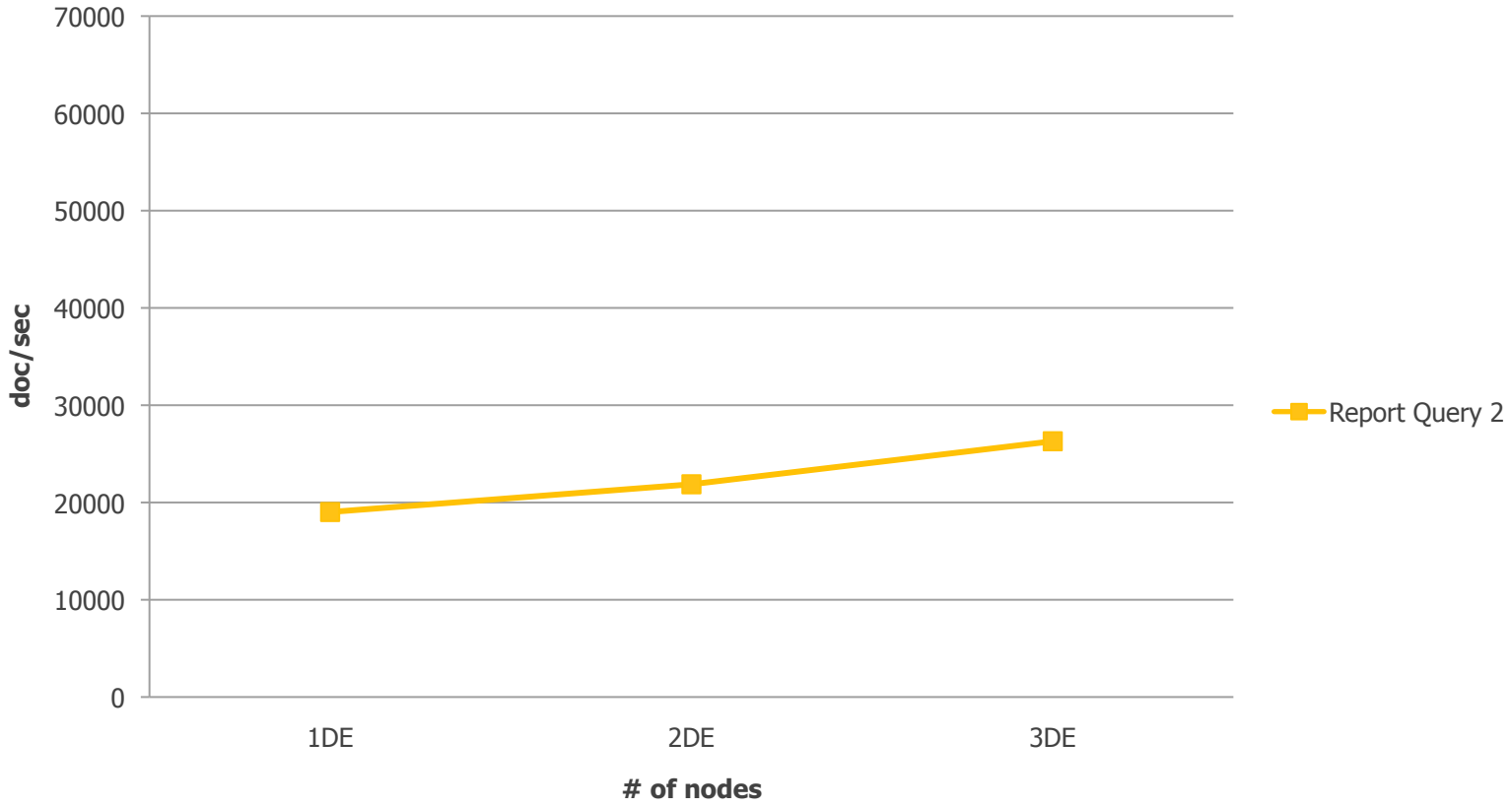
Naive Query Pseudocode

```
for each book
 for each instrument in that book
    position = position(yesterday, book, instrument)
    for each trade of that instrument in this book
      position += trade(today, book, instrument).quantity
    insert(today, book, instrument, position)
```

Initial results

Single node – 19,000 inserts per second

Initial results with cluster



Techniques to get to 100K

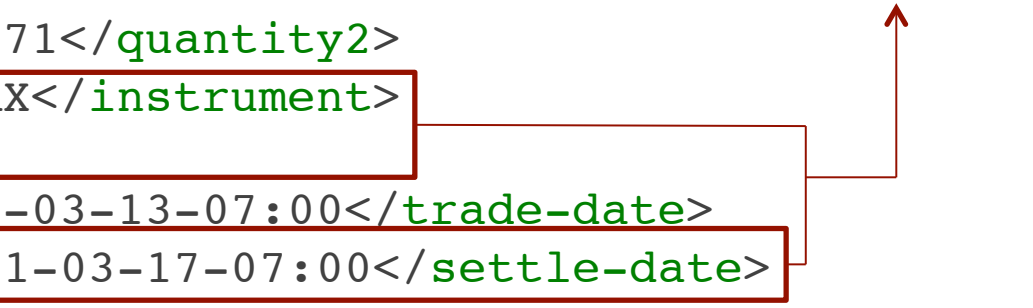
- Insert Query for Computing New Positions
  - Materialized compound key, Co-Occurrence Query and Aggregation
- Insert of New Positions
  - Batching
  - Optimized insert mechanics

Materializing a compound key

```
<trade>

    <quantity2>1193.71</quantity2>
    <instrument>WASAX</instrument>
    <book>679</book>
    <trade-date>2011-03-13-07:00</trade-date>
    <settle-date>2011-03-17-07:00</settle-date>
</trade>
```

Materializing a compound key

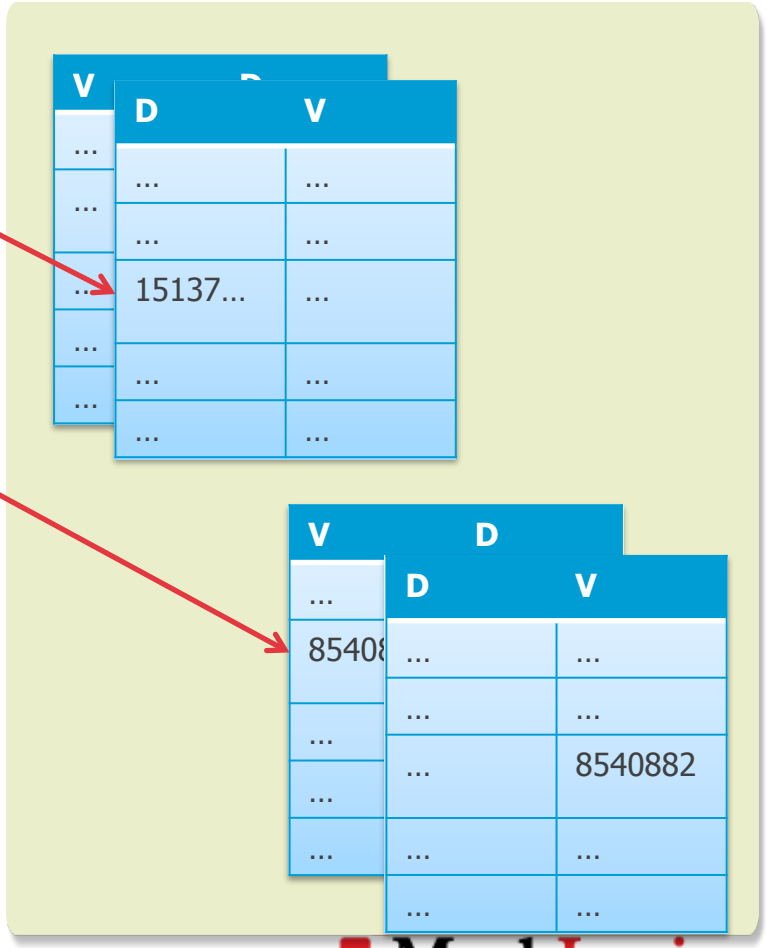
```
<trade>
    <roll-up book-date-instrument="15133344556782303"/>
    <quantity2>1193.71</quantity2>
    <instrument>WASAX</instrument>
    <book>679</book>
    <trade-date>2011-03-13-07:00</trade-date>
    <settle-date>2011-03-17-07:00</settle-date>
</trade>
```



MarkLogic

Co-Occurrence and Distributed Aggregation

```
<trade>
  <roll-up
     book-date-instrument="151373445566703"/>
  <quantity>8540882</quantity>
  <instrument>WASAX</instrument>
  <book>679</book>
  <trade-date>2011-03-13-07:00</trade-date>
  <settle-date>2011-03-17-07:00</settle-date>
</trade>
```



- Co-occurrences:
  Find pairings of range indexed values
- Aggregate on the D nodes (Map/Reduce):
  Sum up the `quantities` above

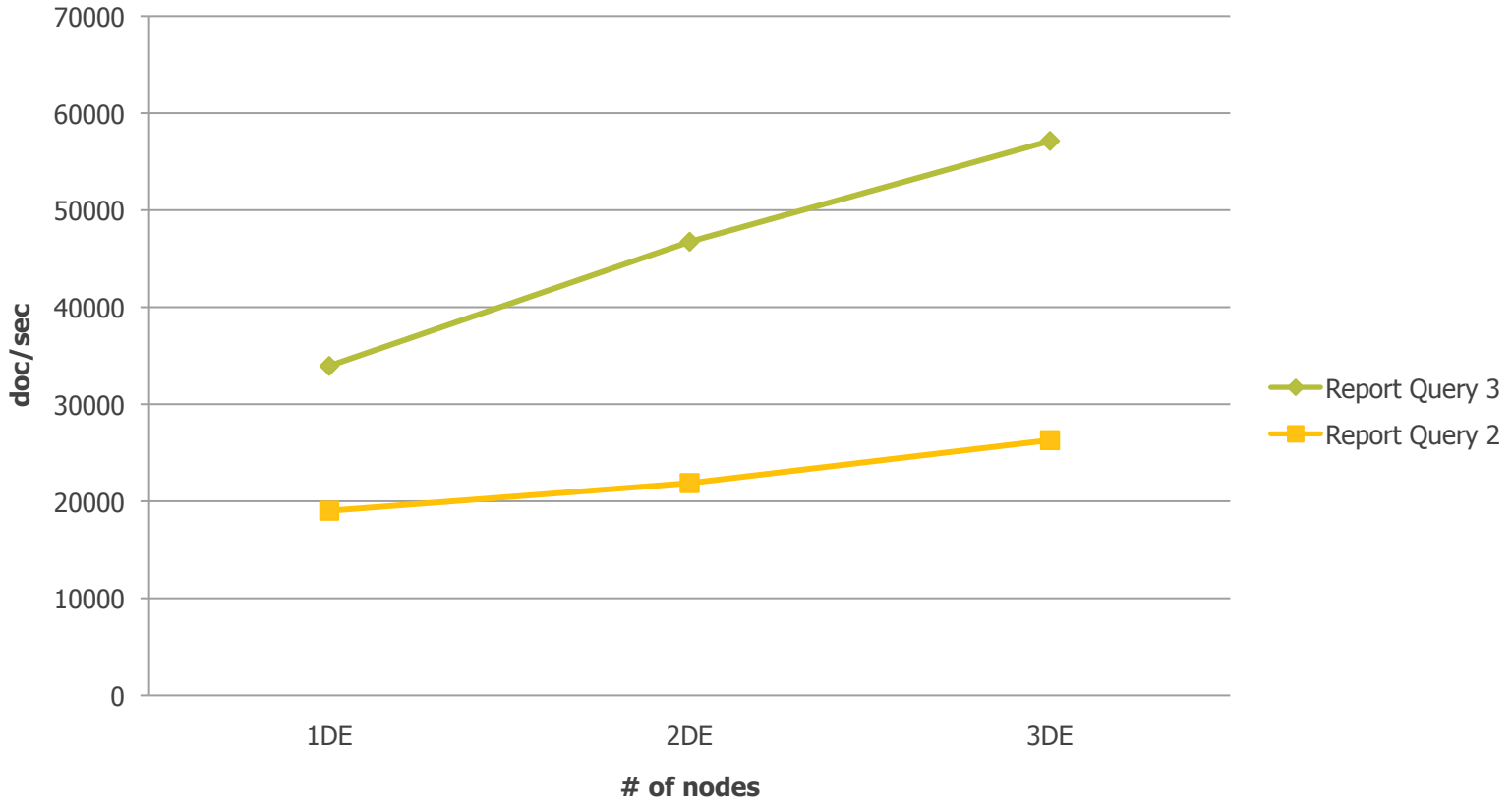
- Similar to a Group-by,
  - in a column-oriented, in-memory database

MarkLogic

Initial results with new query

> 30K inserts/second on a single node

Co-Occurrence + Aggregate versus Naïve approach

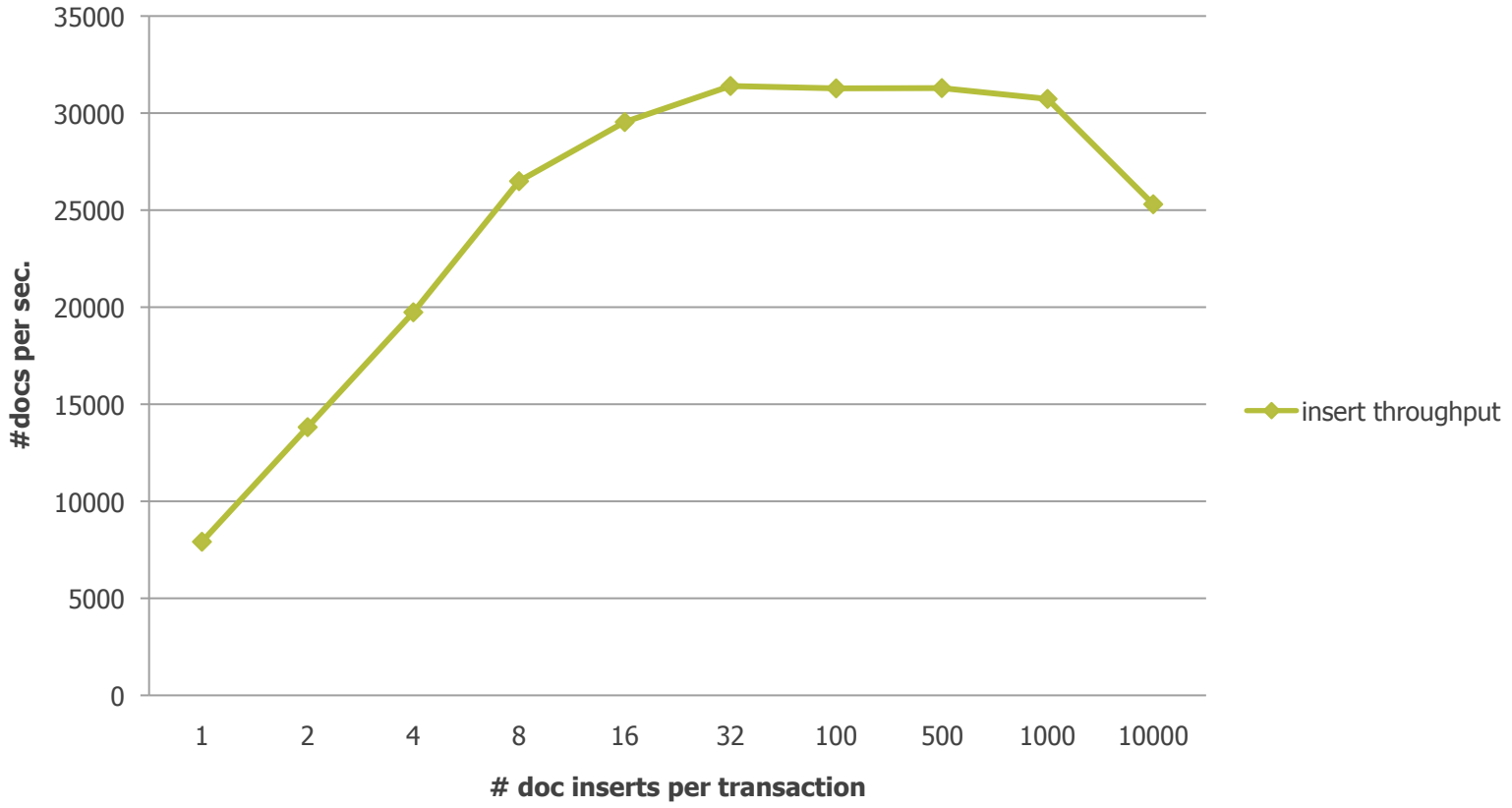


MarkLogic

Techniques

- Computing Positions
  - Materialized compound key, Co-Occurrence Query and Aggregation
- Updates
  - Batching
  - Optimized insert mechanics

Transaction Size and Throughput



Techniques

- Computing Positions
  - Materialized compound key, Co-Occurrence Query and Aggregation
- Updates (Transaction)
  - Batching
  - Optimized insert mechanics

MarkLogic

Insert Mechanics, Again

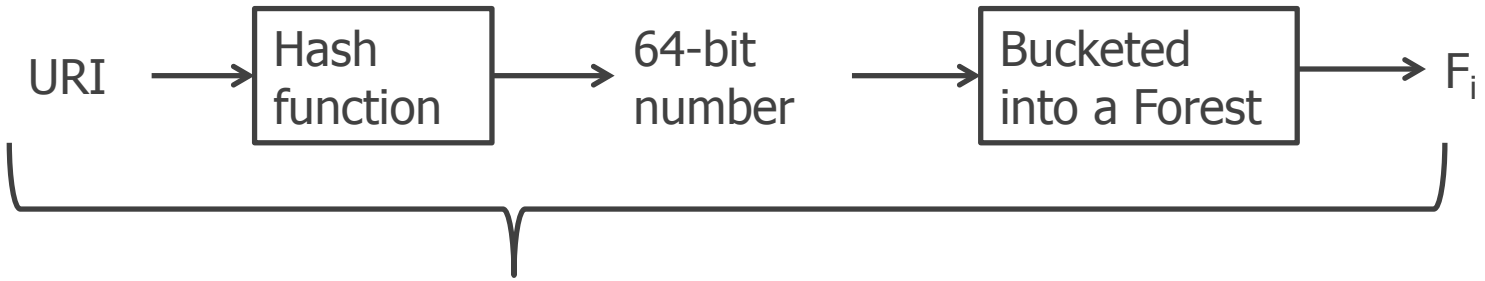
1) New URI+Document arrive at E-node
2) *URI Probe – determine whether URI exists in any forest
3) *URI Lock – write locks are created
4) Forest Assignment – URI is deterministically placed in Forest
5) Indexing
6) Journaling
7) Commit – transaction complete
8) *Release URI Locks – D nodes are notified to release lock

*** Overhead of these operations increases with cluster size**

- X

Deterministic Placement

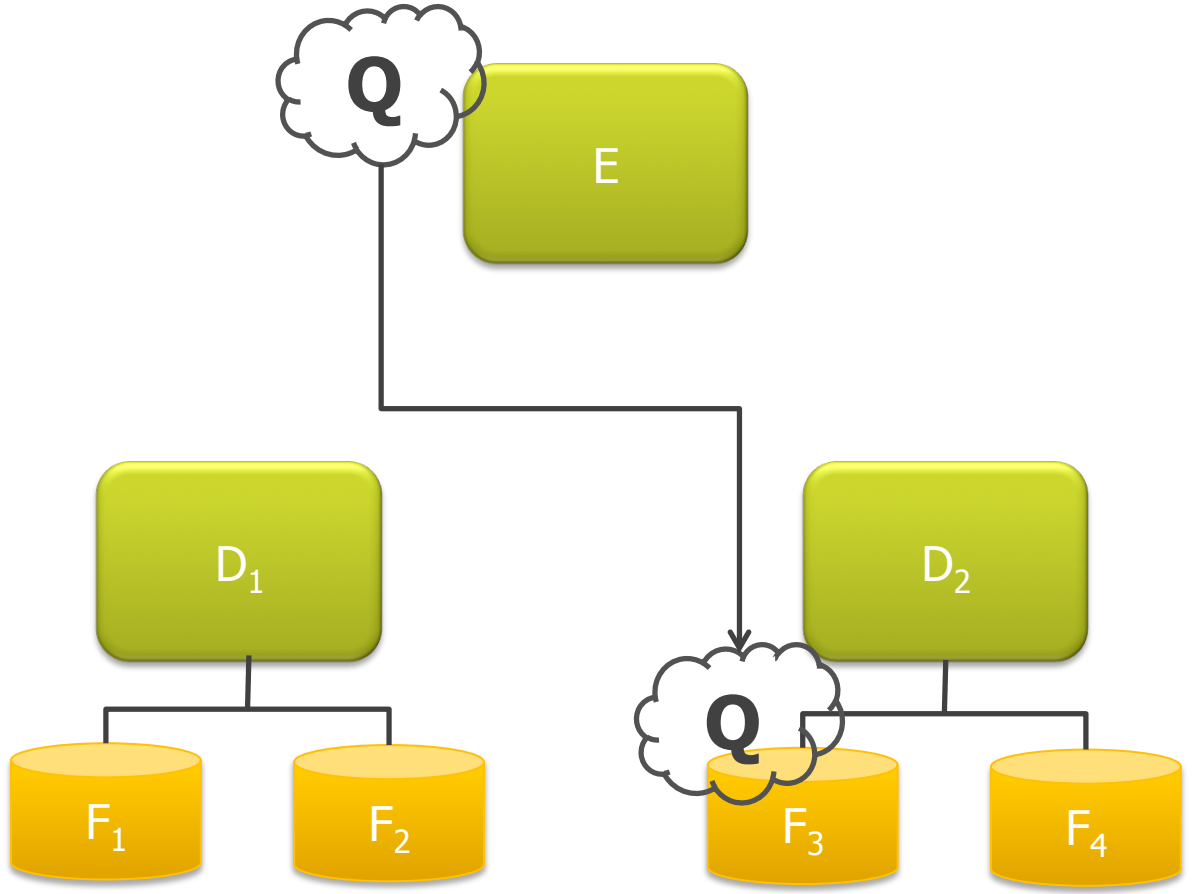
4) Forest Assignment – URI is deterministically placed in Forest



- Done in C++ within server
- But…
  - Can also be done in the client
  - Server allows queries to be evaluated against only one forest…

MarkLogic

Optimized Insert



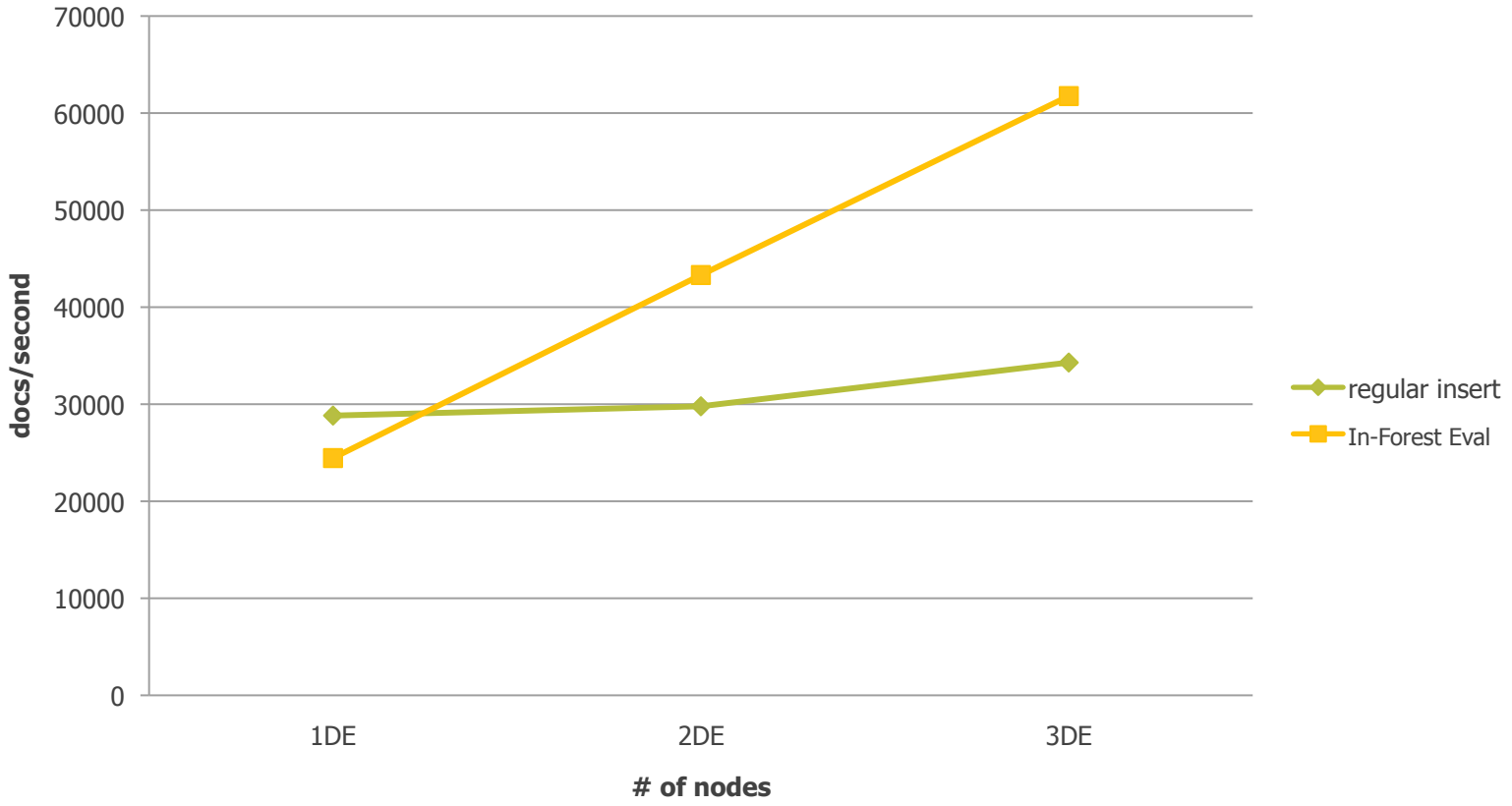
MarkLogic

Optimized Insert Mechanics

1) New URI+Document arrive at E-node
   a) Compute F_i using
   b) Ask server to evaluated the insert query only against F_i
2) URI Probe – F_i Only
3) URI Lock – F_i Only
4) Forest Assignment – F_i Only
5) Indexing
6) Journaling
7) Commit – transaction complete
8) Lock Release - F_i Only

MarkLogic

Regular Insert Vs. Optimized Insert



MarkLogic

We weren't content with 15K

So we showed them…

100,000 INSERTS PER SECOND

A quote from the bank

"We threw everything we had at MarkLogic and it didn't break a sweat"

Enterprise-grade NoSQL document-oriented database

with real-time full-text search and transactions

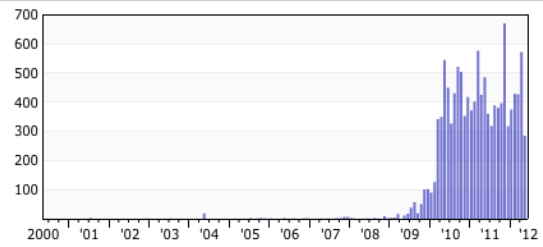
Doing 100K transactions per second

MarkMail

Search 8,534 lists for: nosql Search

Home | Want your own MarkMail? Tell us about it. | Sign In or Sign Up (Why?)

Messages per Month (Swipe to refine by date)



What List? View more

com.googlegroups.jquery-br	2,304
com.googlegroups.mongodb-user	824
com.googlegroups.python-cn	691
org.apache.incubator.isis-commits	464
org.apache.incubator.cassandra-user	402
org.apache.hadoop.hbase-user	312
org.apache.incubator.isis-dev	303
com.basho.lists.riak-users	283
com.googlegroups.nodejs	221

Who Sent It? View more

Suissa	1,007
danh...@apache.org	414
Apache Wiki	223
buil...@apache.org	140
Apache Jenkins Server	113
Ruan Carlos	90
Mark Phillips	88
Sam Millman	83
Apache Hudson Server	75

Any Attachments? View more

gif	152
jpg	83
png	28
pdf	14
patch	10
Other	7
tiff	7
txt	6
py	4

Sort by Relevance 1 to 10 of about 12121

Re: Exporting from Oracle to Mongo
... commercial RDBMS to **NoSQL**, hence my question.
Mar 28, 2012 - NoSQL Guy - com.googlegroups.mongodb-user

Nightly Build
i'm want to run hadoop (hbase) in an IBM JVM. I've seen that there were several patches for that reason. I am not a developer so my knowleges in building java jars fromsources are very limited and the link with the nightly builds do not work. I only need hadoop-core-1.0.3.jar. Where can i find it e
Apr 29, 2012 - nosql - org.apache.hadoop.core-user

Re: ANN: Globals - new **NoSQL** DB with a native Node.js interf...
...in effect, a generic **NoSQL** database engine. See ...ery high-performance **NoSQL** database should really check this out.
Aug 25, 2011 - rtweed - com.googlegroups.nodejs

Re: ANN: Globals - new **NoSQL** DB with a native Node.js interf...
...e) are battle tested **noSQL** solutions. Mumps has been around since the 70's; ... ...s in a "packaged up" **NoSQL** database such as a graph database - it's just tha...
Aug 26, 2011 - Sam Habiel - com.googlegroups.nodejs

Re: ANN: Globals - new **NoSQL** DB with a native Node.js interf...
...o model all kinds of **NoSQL** databases. Of course the schemas you show are exa ... storage: retrieval and atomicity. What specialised **NoSQL** data storage provide are ways to easily retrieve ... ...in effect, a generic **NoSQL** database
engine. Seehttp://gradvs1.mgateway.com/docs/**nosql**_in_globals.pdf
Aug 26, 2011 - Floby - com.googlegroups.nodejs

Re: ANN: Globals - new **NoSQL** DB with a native Node.js interf...
...s in a "packaged up" **NoSQL** database such as a graph database - it's just tha... ...ren't something many **NoSQL** databases do, almost by definition! However in t...

Re: Exporting from Oracle to Mongo 16 messages in com.googlegroups.mongodb-user

Nathan Ehresman	Aug 10, 2011 7:20 am
Steve Francia	Aug 10, 2011 7:21 am
Steve Francia	Aug 10, 2011 8:03 am
Rajat Hubli	Aug 10, 2011 8:30 am
Steve Francia	Aug 10, 2011 8:55 am
Rajat Hubli	Aug 10, 2011 8:58 am
NoSQL Guy	Mar 28, 2012 10:58 am

Subject: Re: Exporting from Oracle to Mongo
From: NoSQL Guy ("nos...@att.net)
Date: Mar 28, 2012 10:58:11 am
List: com.googlegroups.mongodb-user

Rajat Hubli <hublirajat@...> writes:

> Thats perfect. I would work on that now..Thanks a lot
>
> On Aug 10, 5:55 pm, Steve Francia <st...@...> wrote:
> > the script will directly be querying your oracle database and inserting it directly into mongo. No exporting or files at all.

Nathan, Rajat and all,

Why are you moving from Oracle to MongoDB?
Is it for a new application that needs some function that Oracle cannot provide?
If so, what function? or is it primarily to save cost?
Rajat, sounds like you are trying to take an existing application and
rewrite it so that you could use MongoDB, and I am curious what is driving that?

I have not seen many existing apps being rewritten from using a commercial RDBMS to **NoSQL**, hence my question.

Thanks!

MarkLogic COMMUNITY

Search the COMMUNITY with MarkLogic

Sign up | Log in

DOCUMENTATION | DOWNLOADS | CODE | PEOPLE | BLOG | DISCUSS | MEET

MarkLogic 5 is THE operational database for Big Data. Built from the ground up for performance and scale, it includes ACID transactions and real-time search. **Read about MarkLogic 5 »**

Free Download

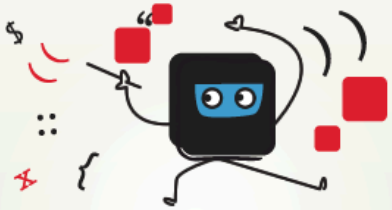
Documentation

Learn about MarkLogic

Get Coding Quickly
Tutorial using a REST API

5 min — Start Now!

Learn the Ninja Way
Interactive coding tutorials using XQuery



45 min — Start Now!

Watch and Learn
Visual Tools

5 min — Start Now!

New & Notable

Community blogging update: Big Data apps, XQuery tricks, code performance, etc.
by Evan Lenz, May 11, 2012
See what MarkLogic Developers have been blogging about the last couple of months.

Highlights from MarkLogic World 2012
by Evan Lenz, May 7, 2012
Check out some visual and Twitter highlights from MarkLogic World 2012.

MarkLogic: RT: @davenielsen: Using #BigData, US healthcare alone could create > $300 billion in value every year (via

People & Meetups

MarkLogic World lightning talk slides
We've published up a number of the slide decks from the awesome series of lightning talks this year...

The Ontologist: Semantic Web and Controlled Term Lists
May 24, 2012 6:30 PM
American Psychological Association
Lists are everywhere. From lists of people in your organization to lists of stores of a given type in your area to lists of documents concerning sales figures...

MarkLogic User Group London: Jeni

Questions & Answers

0 vote | 2 answer
How to get total number of documents in Marklogic database?
I have around 20 lacs documents in Marklogic Database. I want the total number of documents in my search application for
2012-05-18 Puneet Pant 45

0 vote | 1 answer
Issue with cts:and-query in Marklogic
I have some xml document. The structure of the documents are like this :-
2012-05-18 Puneet Pant 45

0
How to change element name while loading data with Record Loader?

Thank You!

@eedeebee
http://community.marklogic.com
eric.bloch@marklogic.com