C.R.E.A.M.

Codd Rules Everything Around Me

Toby DiPasquale
toby@cbcg.net

NoSQL and Big Data

Image credit: http://www.b2bsaleslounge.com/wp-content/uploads/2010/10/DoubleBarrelShotgun.jpg

what is NoSQL?

DEFINITION

NoSQL describes any data store that is not one of the following:

- Oracle
- SQL Server
- MySQL
- PostgreSQL*

Oh hai Bruce

Image credit: http://www.flickr.com/photos/obartunov/5735115601/sizes/l/

what is Big Data?

“big data”

bullshit marketing term

1. More data than you can analyze with Excel/R

MOTIVATION

RDBMS can do everything

not optimal for anything

PROPERTIES

no SQL
(duh)

usually some kind
of distributed story

CAP-awareness

Basically-

Available

Soft state

Eventual consistency

hard to taxonimize

CATEGORIES

Key/Value

get/put/delete API

(semi-)opaque records, no JOINs

data modeling is...
uh... a thing

scaling/HA stories range from none to excellent

relaxed consistency
is common

lots and lots of these, all kinds of types

EXAMPLES

- Riak
- Redis
- Citrusleaf
- Memcached
- Cassandra
- GT.M
- BerkeleyDB
- Voldemort
- Dynamo (Amazon-internal)

USE CASES

fast access to huge number of records

persistent “cache”

large transaction volume on individual records

Document Stores

hierarchical,
schema-less
records

(a.k.a. objects)

collection vs. table
document vs. row

content-
addressable

still no JOINs

same distribution stories as RDBMS

EXAMPLES

- CouchDB
- MongoDB
- MarkLogic
- Lotus Notes
- Amazon SimpleDB

USE CASES

web apps, mostly

SQL is hard,
lets go shopping

cargo cult

Image credit: http://purgingtalon.com/diabologue/wp-content/uploads/2009/01/cargocult1.jpg

Graph Databases

graphs, edges,
vertices, etc.

rich graph traversal

rely mostly on
vertical scalability

EXAMPLES

- neo4j
- AllegroCache
- FlockDB
- Pregel

USE CASES

social network analysis

spread of diseases

sociological research

anywhere you have
a big graph, really

Object Databases

native object stores

basically the same
as document stores

most are much
older than “NoSQL”

trivial application integration

usually highly integrated into particular runtime

EXAMPLES

- Gemstone/S
- db4o
- MarkLogic
- InterSystems Cache
- ZODB

USE CASES

need fast transactions and are willing to lock-in to platform to get it

some handle XML natively, if you're into that sort of thing (not that there's anything wrong with that...)

not very popular out of certain verticals

Big Tables

huge table
distributed across
many machines

single-row transactions/ consistency

no JOINs, but range scans available

column-oriented with no fixed record schema*

(*) not quite the same storage as column stores

scalability/HA stories are good to amazing

some offer ISAM
(never call it that,
though)

EXAMPLES

- Cassandra
- HBase
- BigTable
  - GAE DataStore
- Yahoo! PNUTS

USE CASES

TF–IDF index

when you search Google, this is where your results come from

fast access to very large data volumes

simple analytics on big data

Column Stores

SQL front-ends

based on RDBMS technology

JOINs!

(finally!)

(wtf?)

optimized for data warehousing/ analytics

column-oriented for fast reads on star schemas

can use efficient compression for better I/O

rely heavily on vertical scaling and/or custom hardware

mostly commercial
big $$$

EXAMPLES

- Vertica (HP)
- Aster Data (Teradata)
- Greenplum (IBM)
- kdb (Kx Systems)
- MonetDB (open source)

USE CASES

data warehousing/ data marts

business analytics
dashboards/
reporting

statistical modeling

File Systems

file systems?

Hadoop rules the world

store giant quantities of data and actually use it later

lingua franca: MapReduce

Hive, BigQuery, etc for non-dev querying

most “big data” is sitting in a file system, not a DB

most data mining/ machine learning is happening here, too

ETL still a PITA

honorable mentions

- multivalue DBs (Pick et al)
- IMDGs (e.g. GigaSpaces, Terracotta)
- IMDBs (VoltDB, MemSQL, HANA)
- scientific (array) DBs (e.g. SciDB)

OUTRO

datastore choice is exploding

most systems use
more than one type

requires more careful thought about app requirements

significant perf/ops boost if you get it right

hair and job loss if you get it wrong

data science =
growth field

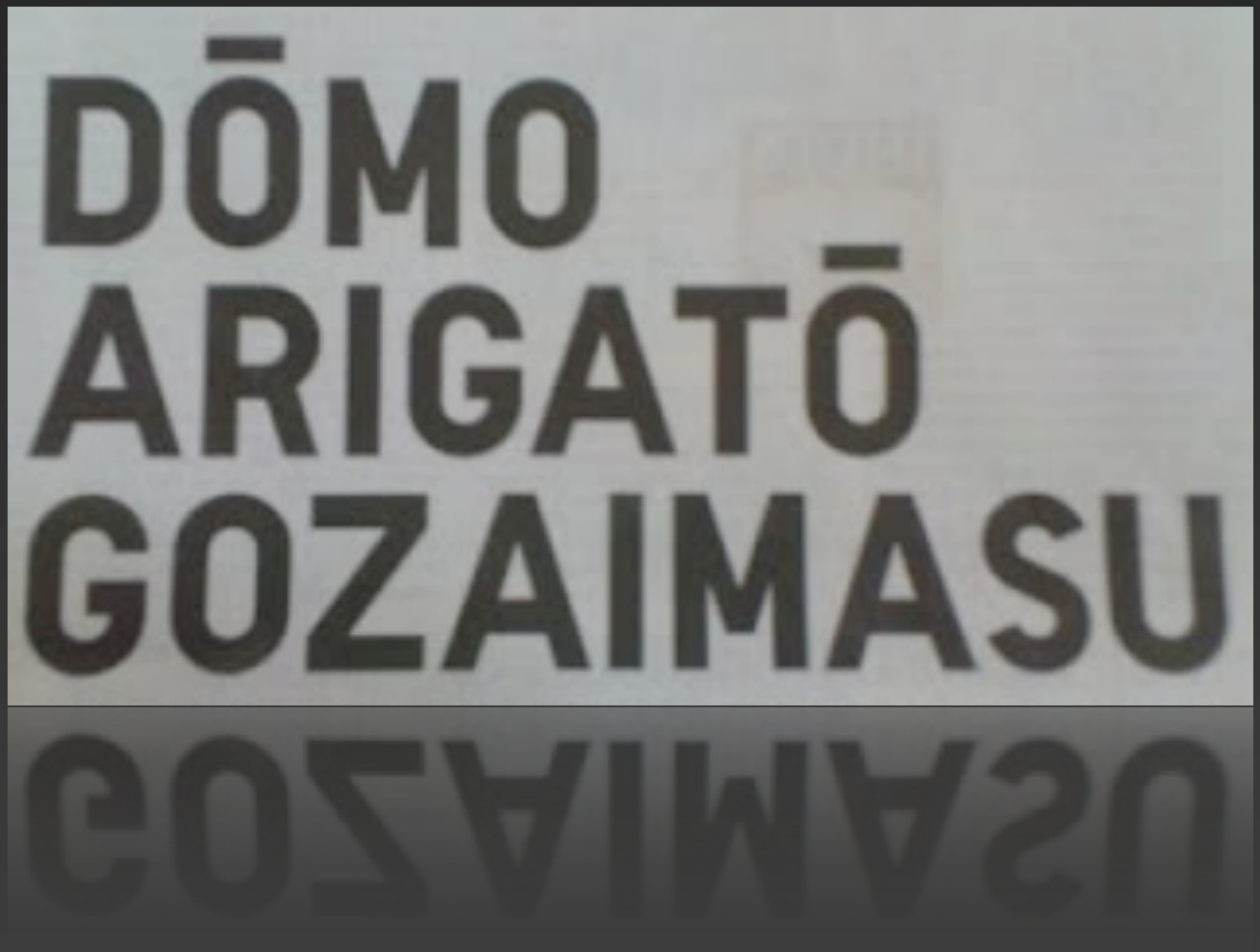

Image credit: http://1.bp.blogspot.com/_sWF5m34IYBQ/SgF4kk6TM1I/AAAAAAAAAjE/4r1PbXJwk64/s320/DSC01636.JPG

LEARN STATISTICS NOW WHILE YOU STILL CAN!!

SERIOUSLY, TAKE AS MUCH STATS AS YOU CAN