THE NOSQL MOUVEMENT

GENOVEVA VARGAS SOLAR
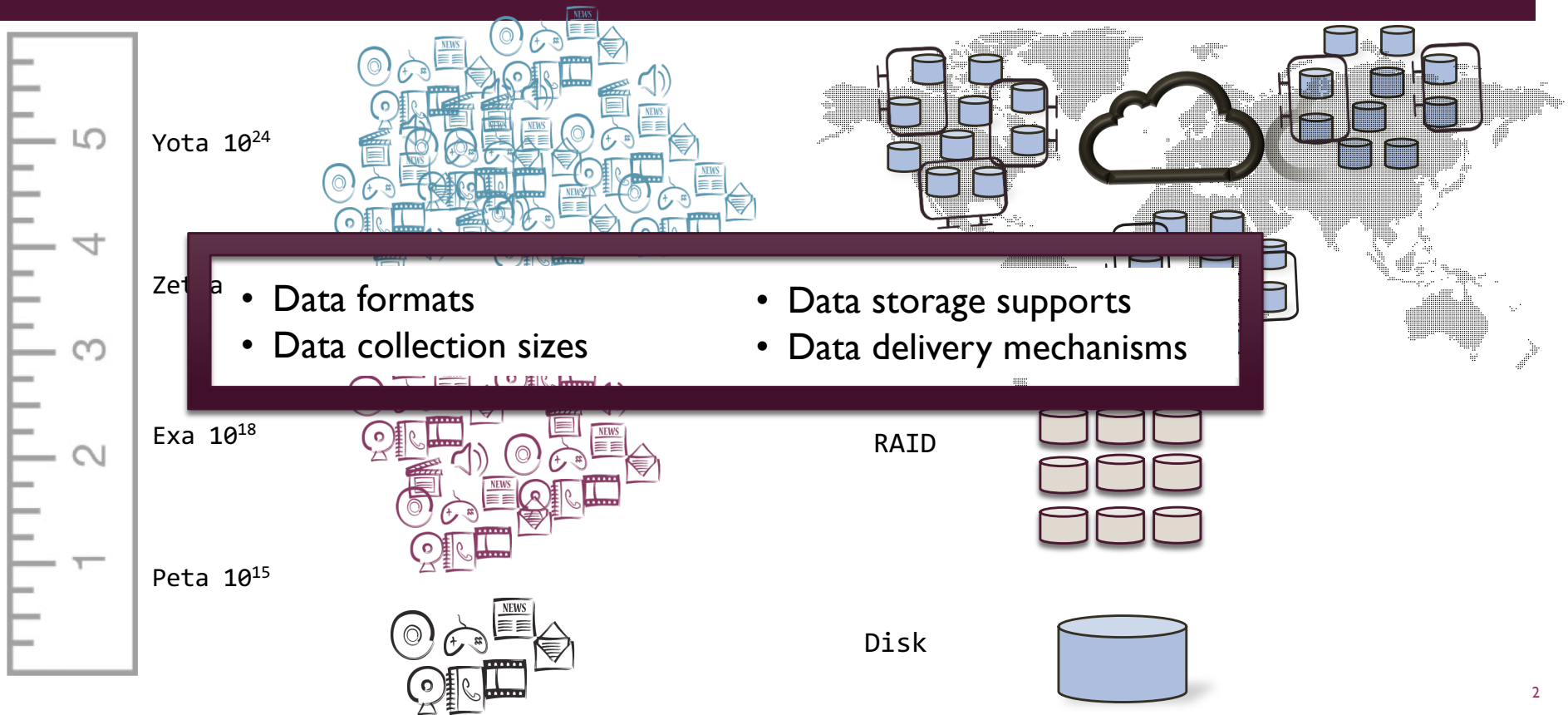
FRENCH COUNCIL OF SCIENTIFIC RESEARCH, LIG-LAFMIA, FRANCE

Genoveva.Vargas@imag.fr

http://www.vargas-solar.com/bigdata-managment

STORING AND ACCESSING HUGE AMOUNTS OF DATA

Yota 10^{24}

Zet a

- Data formats
- Data collection sizes
- Data storage supports
- Data delivery mechanisms

Exa 10^{18}

RAID

Peta 10^{15}

Disk

DEALING WITH HUGE AMOUNTS OF DATA

Yota 10^{24}

Zetta 10^{21}

Exa 10^{18}

Peta 10^{15}

Relational

Graph

Key value

Columns

Concurrency

Consistency

Atomicity

Cloud

RAID

Disk
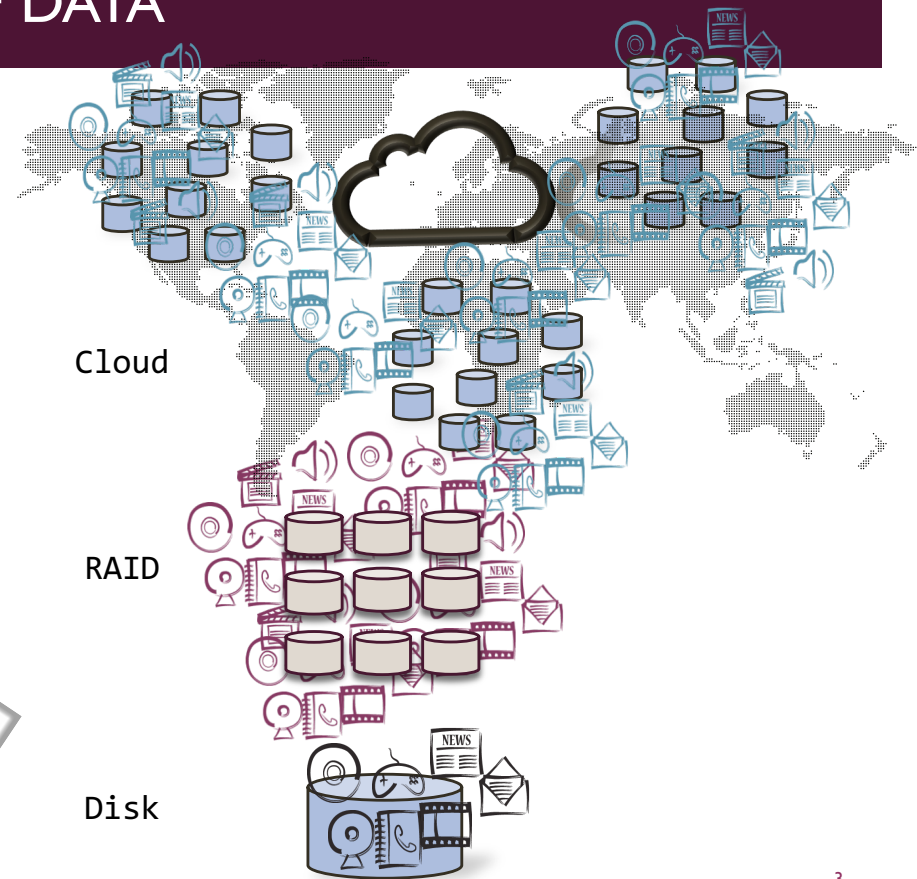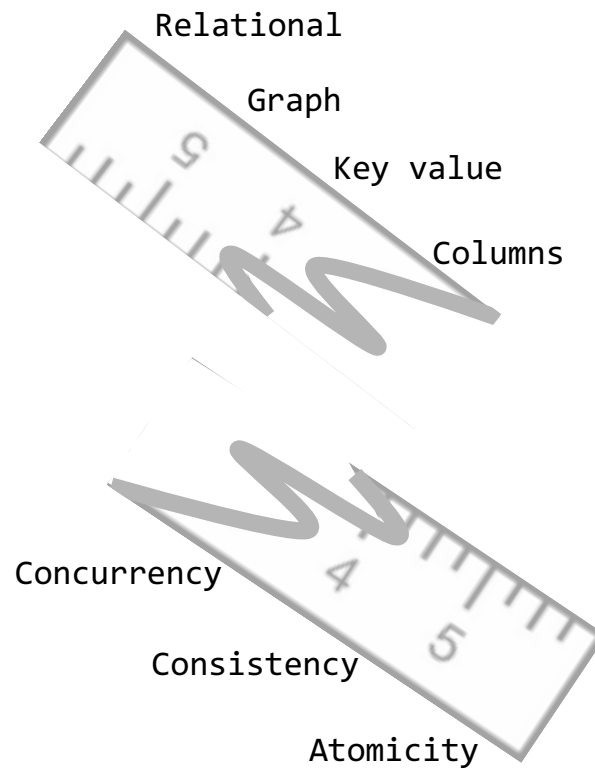
NOSQL STORES CHARACTERISTICS

There is no standard definition of what NoSQL means. The term began with a workshop organized in 2009, but there is much argument about what databases can truly be called NoSQL.

But while there is no formal definition, there are some common characteristics of NoSQL databases

☐ they don't use the relational data model, and thus don't use the SQL language

☐ they tend to be designed to run on a cluster

☐ they tend to be Open Source

☐ they don't have a fixed schema, allowing you to store any data in any record

- Simple operations
  - Key lookups reads and writes of one record or a small number of records
  - No complex queries or joins
  - Ability to dynamically add new attributes to data records
- Horizontal scalability
  - Distribute data and operations over many servers
  - Replicate and distribute data over many servers
  - No shared memory or disk
- High performance
  - Efficient use of distributed indexes and RAM for data storage
  - Weak consistency model
  - Limited transactions

Next generation databases mostly addressing some of the points: being **non-relational, distributed, open-source** and **horizontally scalable** [http://nosql-database.org]

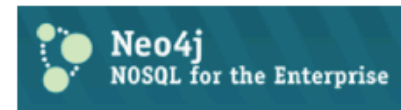# so now we have NoSQL databases

examples include

- Data model
- Consistency
- Storage
- Durability
- Availability
- Query support

Data stores designed to scale simple

OLTP-style application loads

Read/Write operations
by thousands/millions of users

mongoDB
{name: "mongo", type:"DB"}

CouchDB
relax

riak

Cassandra

APACHE
HBASE

Neo4j
NOSQL for the Enterprise

redis

We should also remember Google's
Bigtable and Amazon's SimpleDB. While
these are tied to their host's cloud
service, they certainly fit the general
operating characteristics

DATA MODELS

- Tuple
  - Row in a relational table, where attributes are pre-defined in a schema, and the values are scalar

- Document
  - Allows values to be nested documents or lists, as well as scalar values.
  - Attributes are not defined in a global schema

- Extensible record
  - Hybrid between tuple and document, where families of attributes are defined in a schema, but new attributes can be added on a per-record basis

DATA STORES

- Key-value
  - Systems that store values and an index to find them, based on a key
- Document
  - Systems that store documents, providing index and simple query mechanisms
- Extensible record
  - Systems that store extensible records that can be partitioned vertically and horizontally across nodes
- Graph
  - Systems that store model data as graphs where nodes can represent content modelled as document or key-value structures and arcs represent a relation between the data modelled by the node
- Relational
  - Systems that store, index and query tuples

KEY-VALUE STORES

- "Simplest data stores" use a data model similar to the memcached distributed in-memory cache

- Single key-value index for all data

- Provide a persistence mechanism

- Replication, versioning, locking, transactions, sorting

- API: inserts, deletes, index lookups

- No secondary indices or keys

| SYSTEM	ADDRESS
Redis	code.google.com/p/redis
Scalaris	code.google.com/p/scalaris
Tokyo	tokyocabinet.sourceforge.net
Voldemort	project-voldemort.com
Riak	riak.basho.com
Membrain	schoonerinfotech.com/products
Membase	membase.com

```
SELECT   name, pic, profile_url
FROM     user
WHERE    uid = me()
```
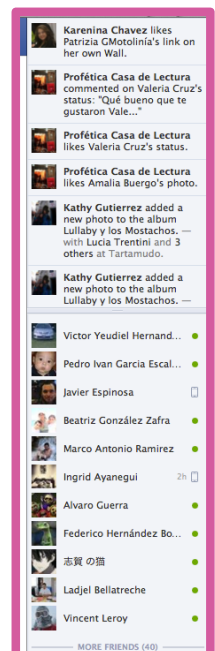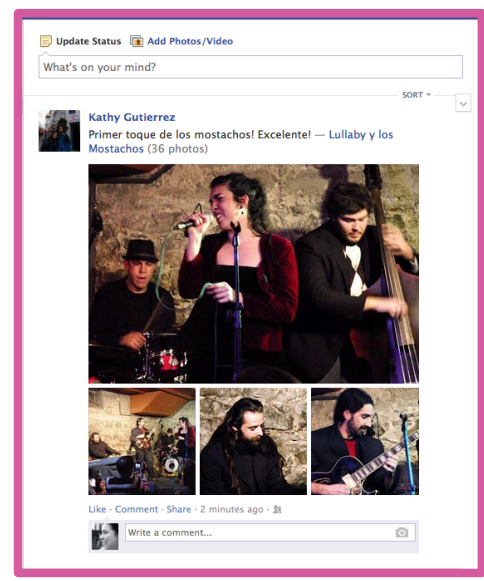
```
SELECT   message, attachment
FROM     stream
WHERE    source_id = me() AND type = 80
```

```
SELECT   name
FROM     friendlist
WHERE    owner = me()
```
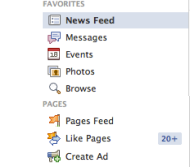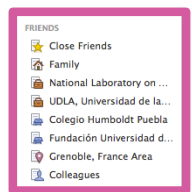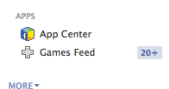
```
SELECT   name
FROM     group
WHERE    gid IN ( SELECT  gid
                  FROM    group_member
                  WHERE   uid = me() )
```

```
SELECT   name, pic
FROM     user
WHERE    online_presence = "active"
         AND
         uid IN ( SELECT  uid2
                  FROM    friend
                  WHERE   uid1 = me() )
```

9

<805114856,

DOCUMENT STORES

- Support more complex data: pointerless objects, i.e., documents

- Secondary indexes, multiple types of documents (objects) per database, nested documents and lists, e.g. B-trees

- Automatic sharding (scale writes), no explicit locks, weaker concurrency (eventual for scaling reads) and atomicity properties

- API: `select, delete, getAttributes, putAttributes` on documents

- Queries can be distributed in parallel over multiple nodes using a map-reduce mechanism

System	Address
SimpleDB	amazon.com/simpledb
Couch DB	couchdb.apache.org
Mongo DB	mongodb.org
Terrastore	code.google.com/terrastore

```json
{
    "name": "Genoveva Vargas-Solar",
    "id": "805114856"
}
```

Search for people, places and things

Home Genoveva

```json
{
  "data": [
    {
      "name": "Genoveva Vargas-Solar",
      "pic": "https://fbcdn-profile-a.akamaihd.net/hprofile-ak-ash4/275915_805114856_16986061_s.jpg",
      "profile_url": "https://www.facebook.com/genoveva.vargas"
    }
  ]
}
```

Genoveva Vargas Solar

Edit Profile

```json
{
  "data": [
    {
      "name": "$$$ Se Vende Jeep Compass 2008 - 60,000kms. $$$"
    },
    {
      "name": "Découvre ce film qui s'engage pour le mariage pour tous"
    },
    {
      "name": "emepink"
    },
    {
      "name": "Such Good People - an indie screwball comedy"
    },
    {
      "name": "Comunidad Mexicana de Tecnologías Semánticas"
    },
    {
      "name": "TI-502 Administración de Datos"
    },
    {
      "name": "exaUDLAP Sistemas Computacionales"
    },
    {
      "name": "\"Hombre Nuevo\" artículos de valores humanos del P. Otaolaurruchi"
    },
    {
      "name": "LACCIR"
    },
    {
      "name": "Monis - groupe de soutien"
    },
    {
      "name": "Red Temática de las TIC"
    },
```

FRIENDS
Close Friends
Family
National Laboratory on ...
UDLA, Universidad de la...
Colegio Humboldt Puebla
Fundación Universidad d...
Grenoble, France Area
Colleagues

FAVORITES
News Feed
Messages
Events
Photos
Browse

PAGES
Pages Feed
Like Pages   20+
Create Ad

Agresados UDLAP
Such Good People – an i...
Monis – groupe de soutien
Découvre ce film qui s'e...
AIDONS LE REFUGE
Create Group...

App Center
Games Feed   20+

MORE▾

```json
{
  "data": [
    {
      "message": "",
      "attachment": {
        "media": [
          {
            "href": "https://www.facebook.com/photo.php?fbid=10151871935952502&set=a.99396912501.109184.98871212501&type=1",
            "alt": "",
            "type": "photo",
            "src": "https://fbcdn-photos-e-a.akamaihd.net/hphotos-ak-ash3/1146527_10151871935952502_258686255_s.jpg",
            "photo": {
              "aid": "98871212501_109184",
              "pid": "98871212501_1073742168",
              "fbid": "10151871935952502",
              "owner": 98871212501,
              "index": 1,
              "width": 611,
              "height": 458,
              "images": [
                {
                  "src": "https://fbcdn-photos-e-a.akamaihd.net/hphotos-ak-ash3/1146527_10151871935952502_258686255_s.jpg",
                  "width": 130,
                  "height": 97
                }
              ]
            }
          }
        ]
      },
      "name": "Timeline Photos",
      "href": "https://www.facebook.com/album.php?fbid=99396912501&id=98871212501&aid=109184",
      "caption": "El sutil arte de cantinflear.\r\n\r\nvía – Lectura Cinematográfica",
      "description": ""
    }
  ]
}
```

Vincent Leroy
MORE FRIENDS (40)
Aby Aragon
Alee Merlo

12

EXTENSIBLE RECORD STORES

- Basic data model is rows and columns

- Basic scalability model is splitting rows and columns over multiple nodes

  - Rows split across nodes through sharding on the primary key

    - Split by range rather than hash function

    - Rows analogous to documents: variable number of attributes, attribute names must be unique

    - Grouped into collections (tables)

    - Queries on ranges of values do not go to every node

- Columns are distributed over multiple nodes using "column groups"

  - Which columns are best stored together

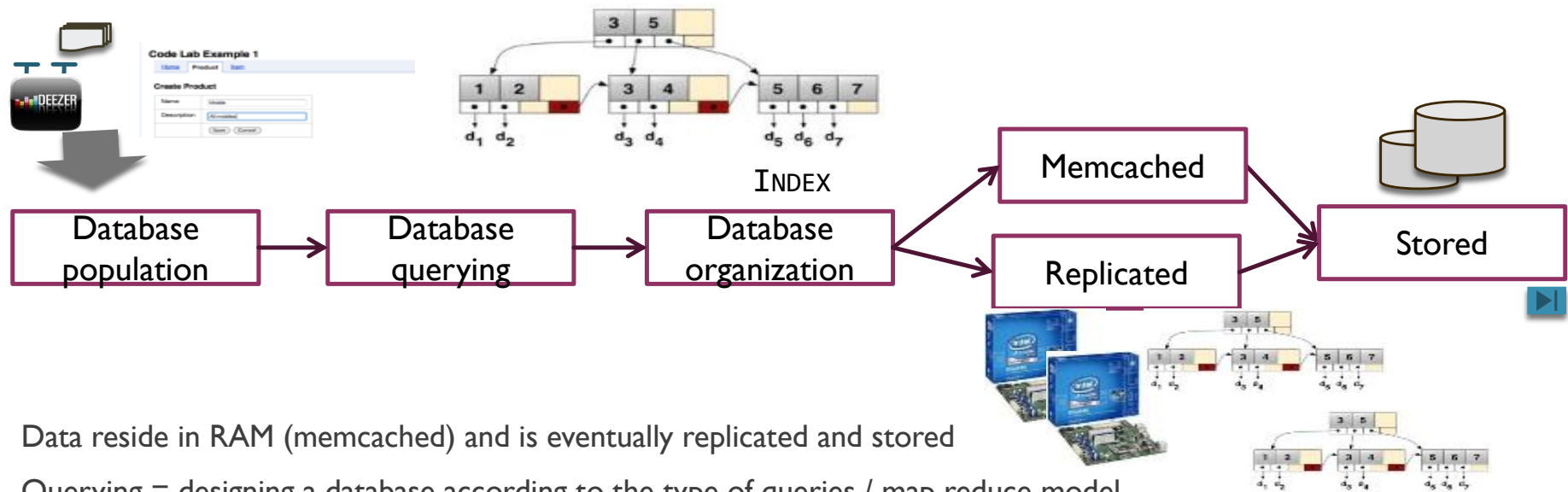 - Column groups must be pre-defined with the extensible record stores

SYSTEM	ADDRESS
HBase	hbase.apache.com
HyperTable	hypertable.org
Cassandra	incubator.apache.org/cassandra

SCALABLE RELATIONAL SYSTEMS

- SQL: rich declarative query language

- Databases reinforce referential integrity

- ACID semantics

- Well understood operations:
  - Configuration, Care and feeding, Backups, Tuning, Failure and recovery, Performance characteristics

- Use small-scope operations
  - Challenge: joins that do not scale with sharding

- Use small-scope transactions
  - ACID transactions inefficient with communication and 2PC overhead

- Shared nothing architecture for scalability

- Avoid cross-node operations

System	Address
MySQL C	mysql.com/cluster
Volt DB	voltdb.com
Clustrix	clustrix.com
ScaleDB	scaledb.com
Scale Base	scalebase.com
Nimbus DB	nimbusdb.com

NOSQL DESIGN AND CONSTRUCTION PROCESS

INDEX

| Database population | → | Database querying | → | Database organization | → | Memcached |
| | | | | | → | Replicated |

Stored

- Data reside in RAM (memcached) and is eventually replicated and stored

- Querying = designing a database according to the type of queries / map reduce model

- "On demand" data management: the database is virtually organized per view (external schema) on cache and some view are made persistent

- An elastic easy to evolve and explicitly configurable architecture

Use the right tool for the right job…

Key-Value | Ordered Key-Value | Big Table | Document, Full-Text Search | Graph | SQL
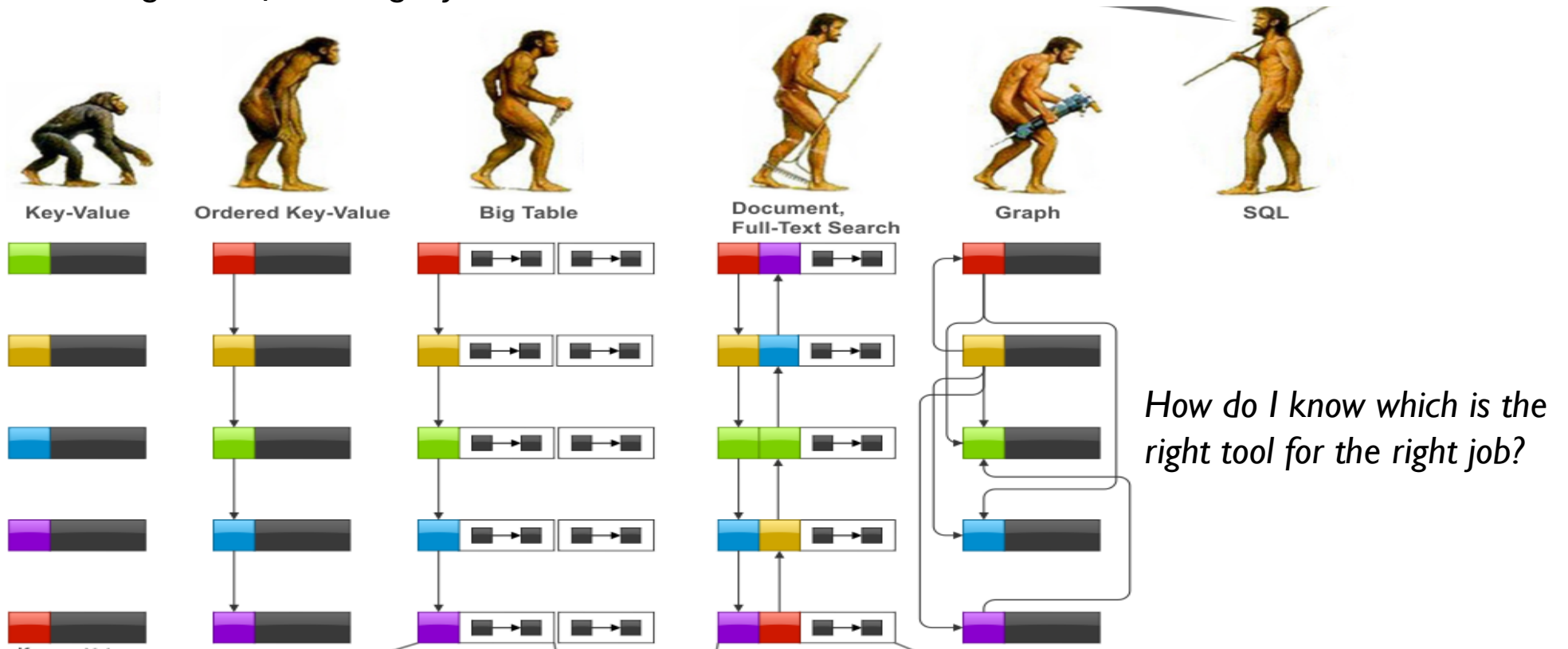
How do I know which is the right tool for the right job?

(Katsov-2012)

Genoveva.Vargas@imag.fr
http://www.vargas-solar.com/bigdata-management

REFERENCES

- Eric A., Brewer "Towards robust distributed systems." PODC. 2000

- Rick, Cattell "Scalable SQL and NoSQL data stores." ACM SIGMOD Record 39.4 (2011): 12-27

- Juan Castrejon, Genoveva Vargas-Solar, Christine Collet, and Rafael Lozano, ExSchema: Discovering and Maintaining Schemas from Polyglot Persistence Applications, In Proceedings of the International Conference on Software Maintenance, Demo Paper, IEEE, 2013

- M. Fowler and P. Sadalage. NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Pearson Education, Limited, 2012

- C. Richardson, Developing polyglot persistence applications, http://fr.slideshare.net/chris.e.richardson/developing-polyglotpersistenceapplications-gluecon2013