BIG DATA ON THE ROAD

Aggregating and Managing Big rEaltime Data in the Cloud: application to intelligent transport for Smart Cities

Parisa GHODOUS, U. Claude Bernard, LIRIS, parisa.ghodous@univ-lyon1.fr

Genoveva VARGAS-SOLAR, CNRS, LIG-LAFMIA, genoveva.vargas@imag.fr

Catarina Ferreira da Silva, U. Claude Bernard, LIRIS, catarina.ferreira@univ-lyon1.fr

Did U take public transport today?

TRANSPORT ISSUES

Mio. People per kilometre 2013

30000

25000

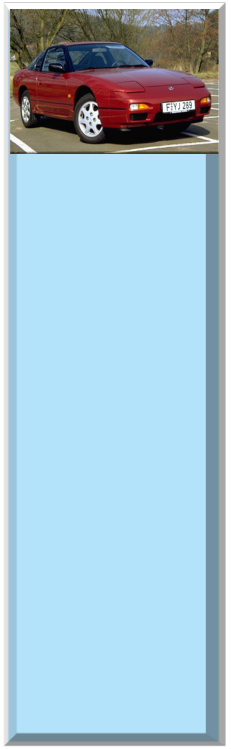
20000

15000

10000

5000

0



Passengers per public transport medium*

INTELLIGENT TRANSPORT

(Tele)Communication

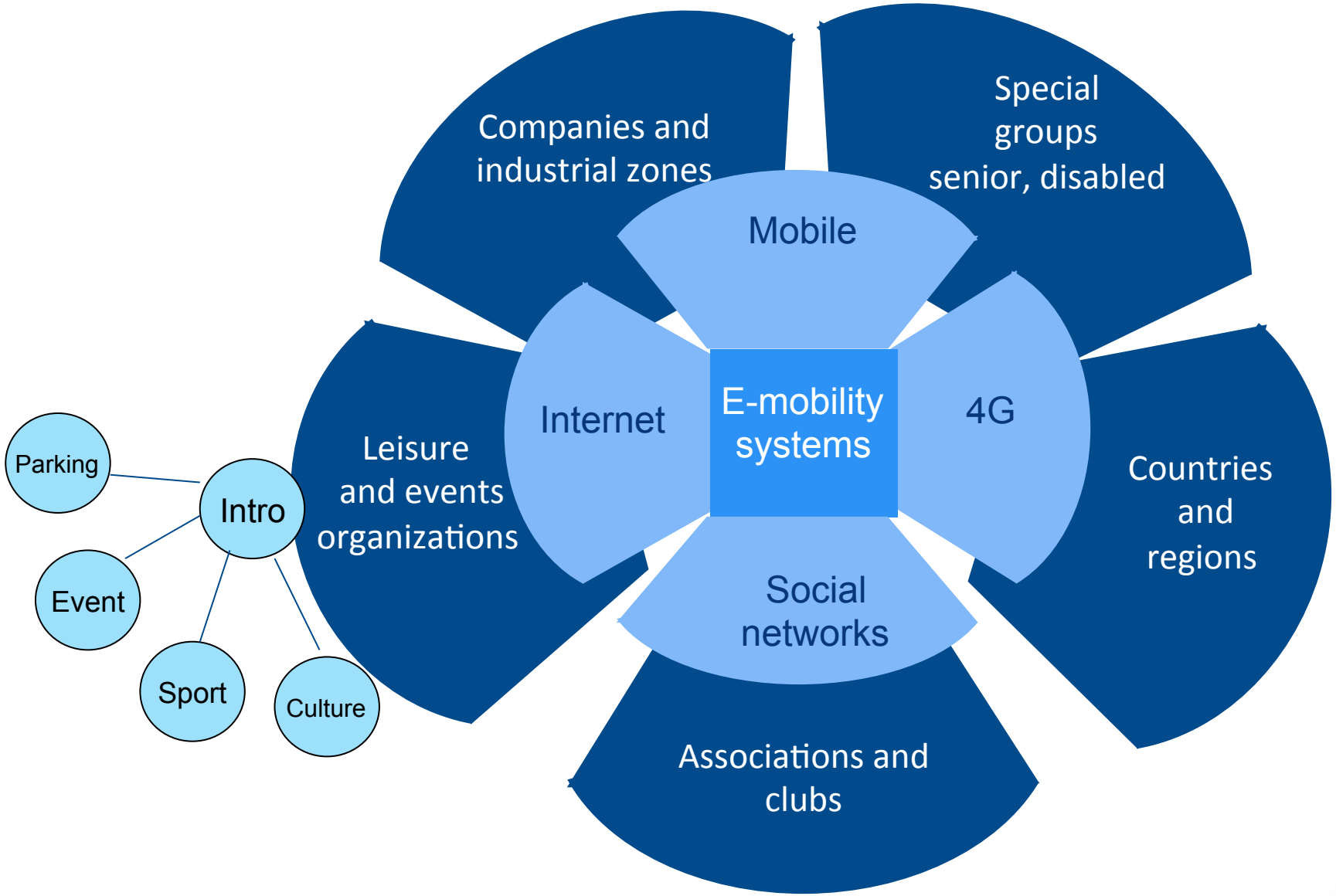
Optimize time, money and space: less stress, high avalability, easy motion

Clean environment

Parking places management

Vehicles and people pooling

E-MOBILITY RESSOURCES



OBJECTIVES

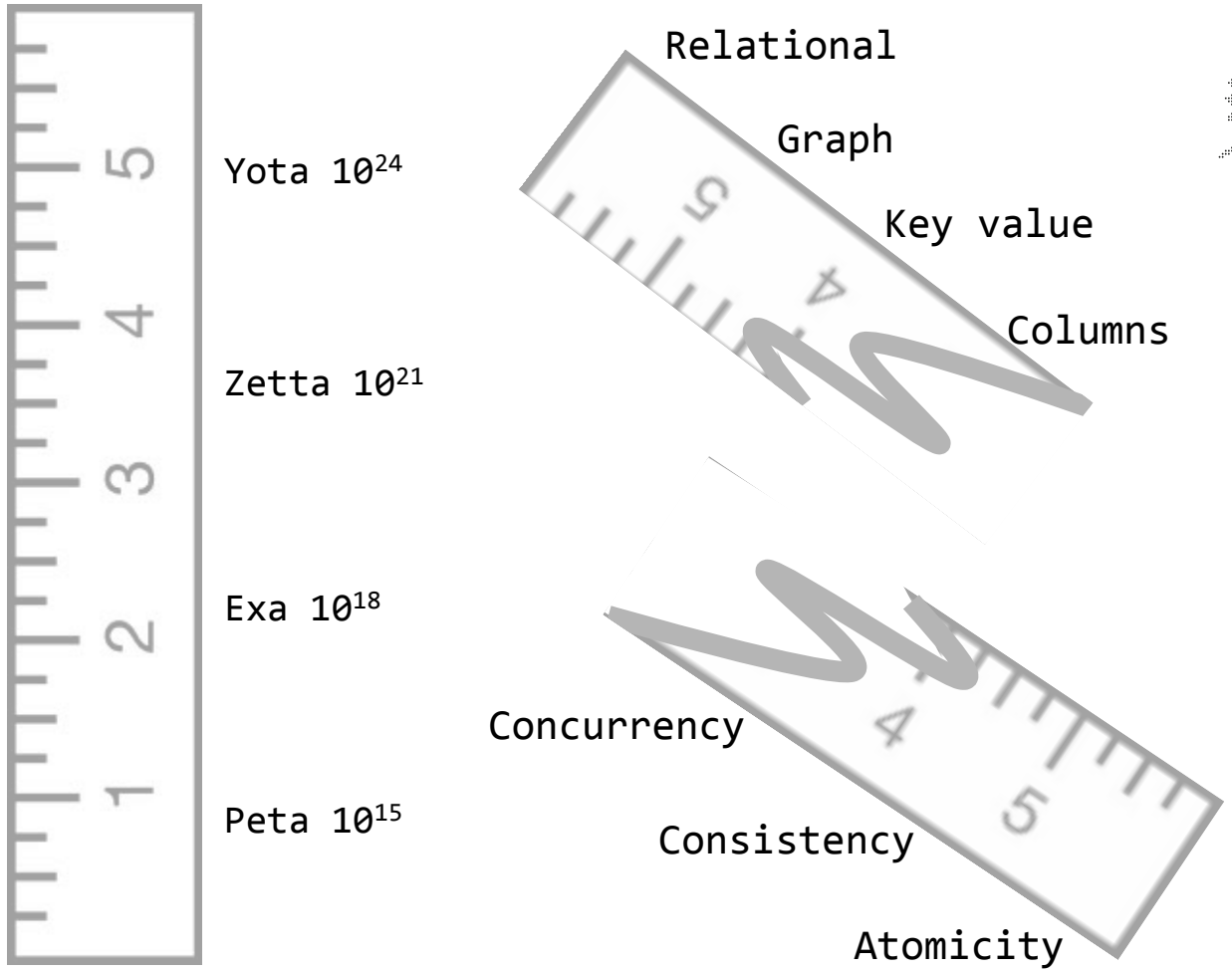
Study and analyse the problem of real-time streaming transportation big data

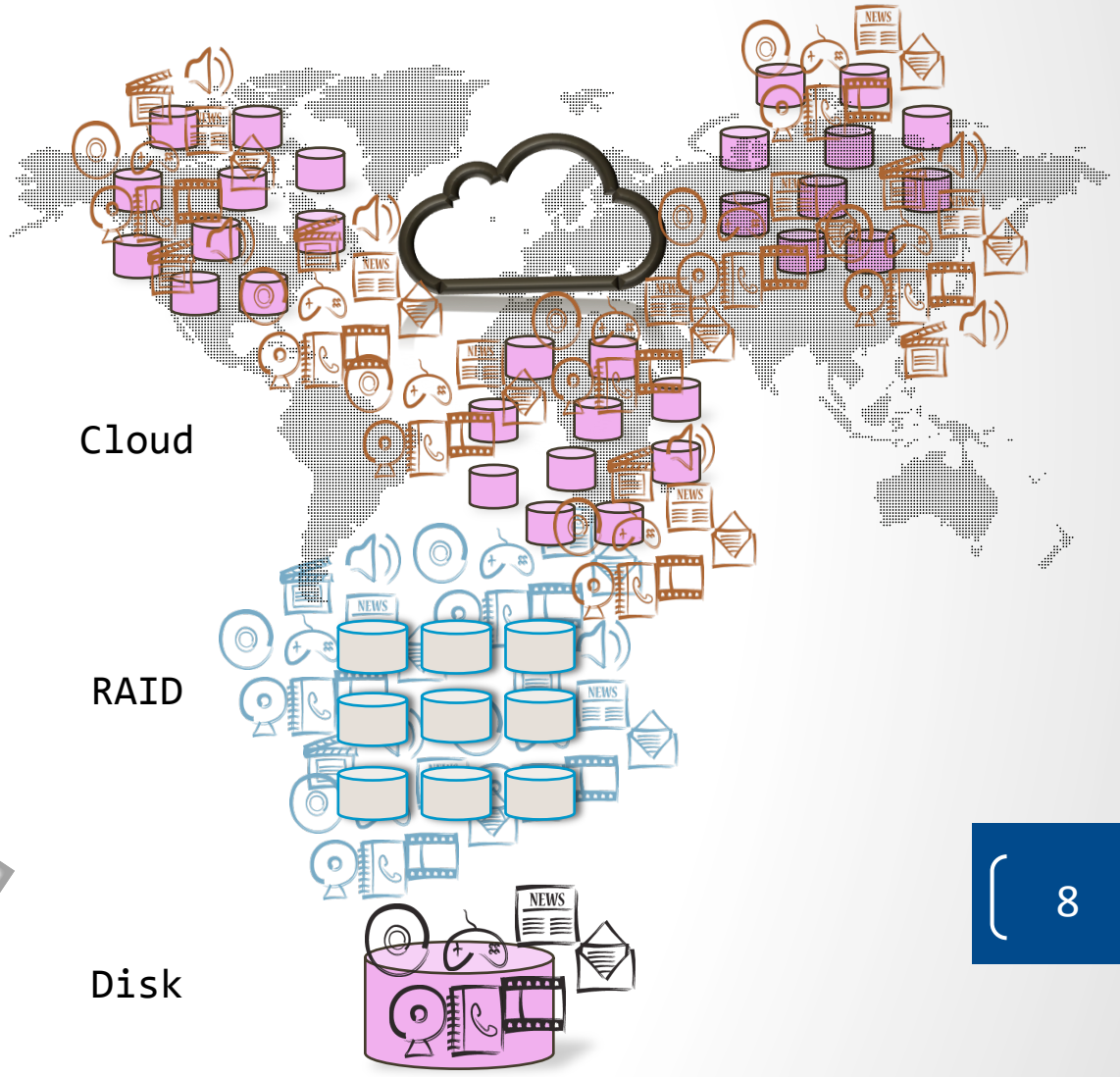
- Propose efficient algorithms and strategies to manage (integrate, store and deliver) big real-time data streams/ collections
- Support decision making processes of intelligent transport actors according to their needs and requirements: *civilians, providers, local and regional planning authorities*

ROADMAP

- **Smart data for smart transport systems**
  - E-Mobility resources
  - Collecting, curating and storing data
  - Getting benefits from the XXL
- **The big data economy**
  - How big is big?
  - Where are the big data collections? Who owns them? Who gets Return of Investment?
- **Conclusions and perspectives**

DEALING WITH HUGE AMOUNTS OF DATA



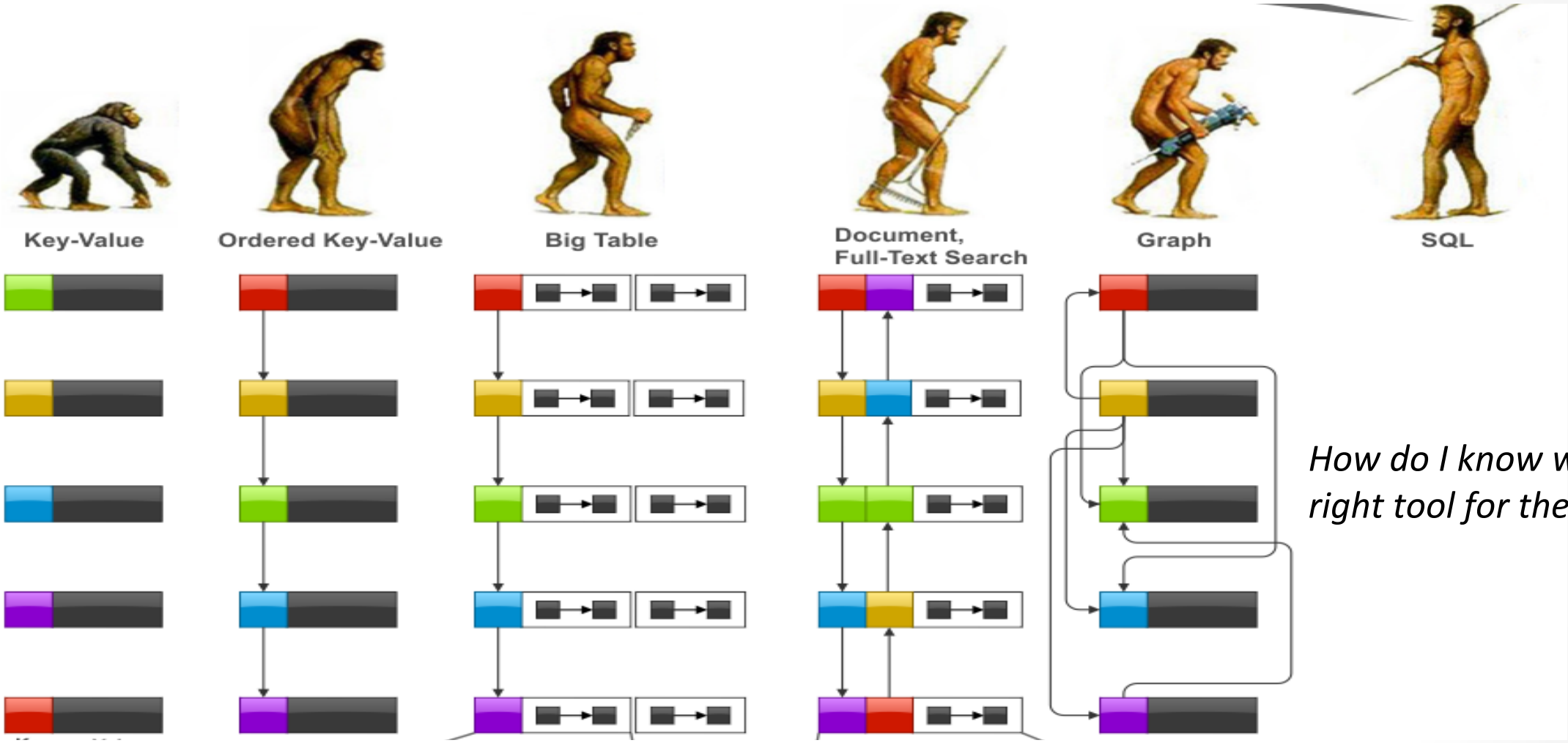


POLYGLOT PERSISTENCE

- ***Polyglot Programming:*** applications should be written in a mix of languages to take advantage of different languages are suitable for tackling different problems
- ***Polyglot persistence***: any decent sized enterprise will have a variety of different data storage technologies for different kinds of data
  - a new strategic enterprise application should no longer be built assuming a relational persistence support
  - the relational option might be the right one - but you should seriously look at other alternatives

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

Use the right tool for the right job…



How do I know which is the right tool for the right job?

(Katsov-2012)

So now we have NoSQL databases

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

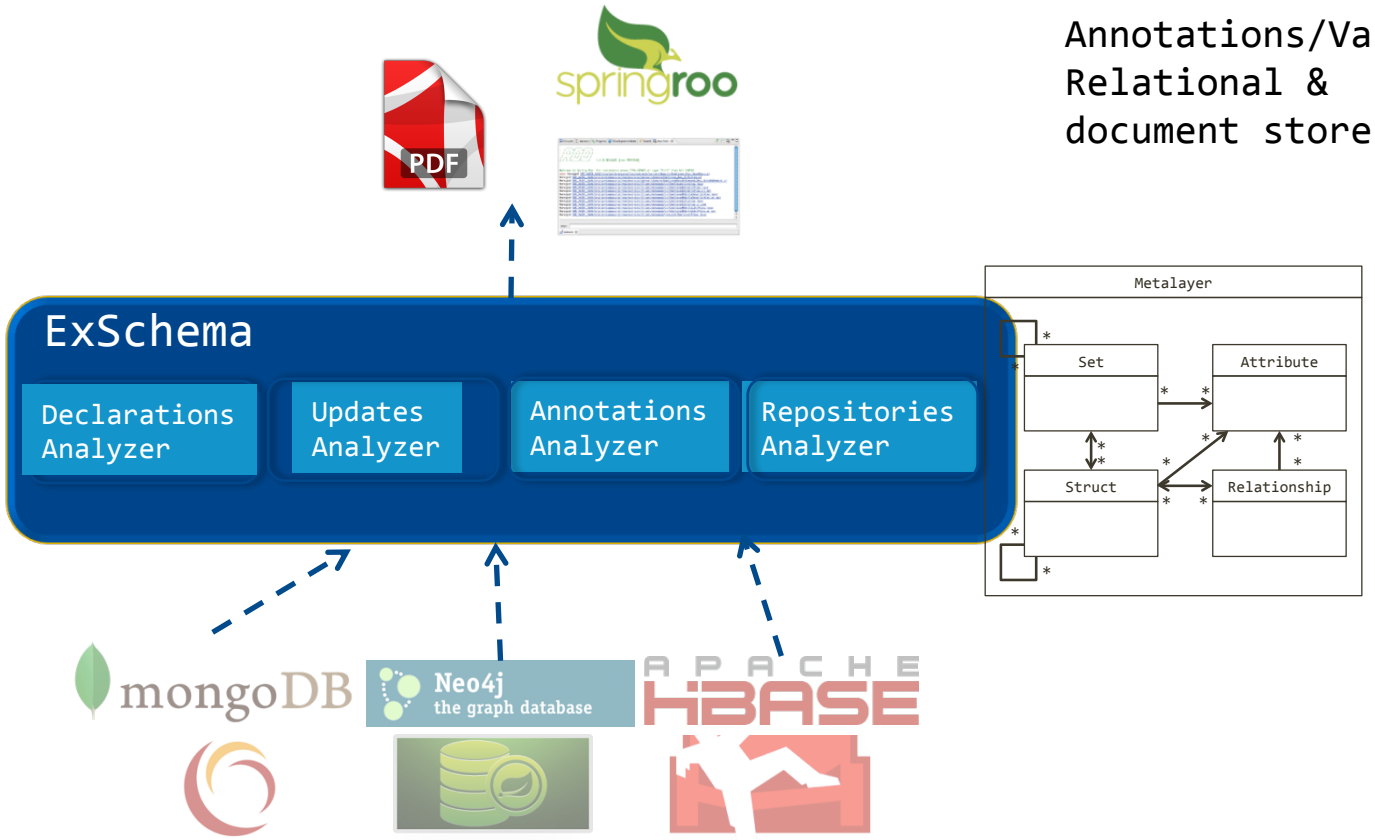
Data stores designed to scale simple OLTP-style application loads

Read/Write operations by **thousands/millions** of users

examples include

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

EXTRACTING SCHEMATA FROM A POLYGLOT DATABASE



Application source code

Annotations/Variables
Relational &
document stores

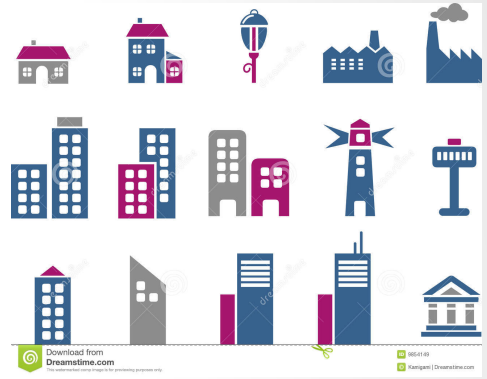
```
import javax.persistence.Entity;
import org.springframework.data.mongodb.crossstore.RelatedDocument;

@Entity                                          <- Annotation
public class Customer {

  @Id   private Long id;                         <- Declaration
  private String firstName;
  private String lastName;

  @RelatedDocument
  private SurveyInfo surveyInfo;
  ...
}
```

```
import org.neo4j.graphdb.Node;
import org.neo4j.graphdb.Relationship;

class ActorImpl implements Actor {
  private static final String NAME_PROPERTY = "name";

  private final Node underlyingNode;             <- Declaration

  public void setName( final String name ) {
    underlyingNode.setProperty( NAME_PROPERTY, name );   <- Update
  }

  public Role createRole( final Actor actor, final Movie movie, final String roleName )
    Node actorNode = ((ActorImpl) actor).getUnderlyingNode();
    Node movieNode = ((MovieImpl) movie).getUnderlyingNode();
    Relationship rel = actorNode.createRelationshipTo( movieNode, RelTypes.ACTS_IN );
  ...
}
```

Variables/Methods Graph Store

MULTI-CLOUD APPROACH FOR MANAGING TRANSPORT BIG DATA SETS

Smart city



Curation

Sanitation

MySQL

xeround

The Cloud Database

Neo4j

the graph database

OPENSHIFT

mongoDB

mongolab

JSON documents

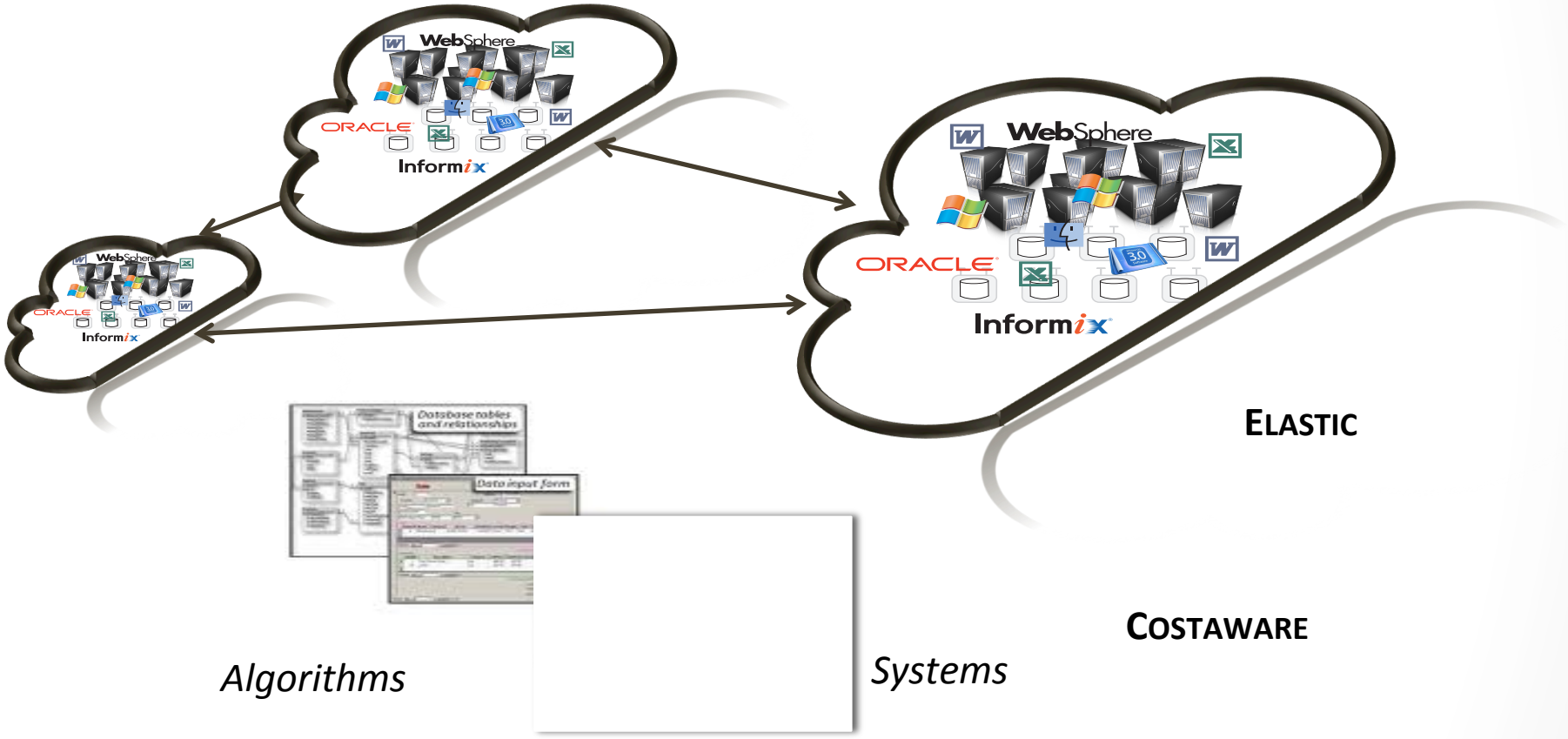
Delivery

Analysis

CLOUD FOUNDRY

Collection

DATA MANAGEMENT WITHOUT RESOURCES CONSTRAINTS



Reasonably costly manage and exploit data sets according to unlimited storage, memory and computation resources

ROADMAP

- **Smart data for smart transport systems**
  - E-Mobility resources
  - Collecting, curating and storing data
  - Getting benefits from the XXL
- **The big data economy**
  - How big is big?
  - Where are the big data collections? How to get a Return of Investment?
- **Conclusions and perspectives**

DIGITAL INFORMATION SCALE

Megabyte(10^6) Gigabyte (10^9) Terabytes (10^{12}), petabytes (10^{15}), exabytes (10^{18}) and zettabytes (10^{21})

Unit	Size	Meaning
Bit (b)	1 or 0	Short for binary digit, after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	2^{10} bytes	From "thousand' in Greek. One page of typed text is 2KB
Megabyte (MB)	2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5 MB. A typical pop song is 4 MB.
Gigabyte (GB)	2^{30} bytes	From "giant" in Greek. A two-hour film ca be compressed into 1-2 GB
Terabyte (TB)	2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB.
Petabyte (PB)	2^{50} bytes	All letters delivered in America's postal service this year will amount ca. 5PB. Google processes 1PB per hour.
Exabyte (EB)	2^{60} bytes	Equivalent to 10 billion copies of The Economist
Zettabyte (ZB)	2^{70} bytes	The total amount of information in existence this year is forecast to be around 1,27ZB
Yottabyte (YB)	2^{80} bytes	Currently too big to imagine

Big Data

- Collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- Challenges include capture, curation, storage, search, sharing, analysis, and visualization **within a tolerable elapsed time**
- Data growth challenges and opportunities are three-dimensional (**3Vs model**)
  - increasing volume (amount of data)
  - velocity (speed of data in and out)
  - variety (range of data types and sources)

"Big data are high volume, high velocity, and/or high variety information assets that require **new forms of processing** to enable enhanced decision making, insight discovery and process optimization**.”**

MASSIVE DATA

MASSIVE DATA COLLECTIONS

- **Information-sensing** mobile devices, aerial sensory technologies (remote sensing)
- **Software logs**, posts to social media sites
- Telescopes, cameras, microphones, digital pictures and videos posted online
- **Transaction records** of online purchases
- **RFID readers**, wireless sensor networks (number of sensors increasing by 30% a year)
- **Cell phone GPS signals** (increasing 20% a year)

DATA PROVIDERS

- **Sloan Digital Sky Survey** (2000-) its archive contains 140 terabytes. Its successor: Large Synoptic Survey Telescope (2016-), will acquire that quantity of data every five days !
- **Facebook** hosts 140 billion photos, and will add 70 billion this year (ca. 1 petabyte). Every 2 minutes today we snap as many photos as the whole of humanity took in the 1800s !
- **Wal-Mart**, handles more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes (10^{15})
- The **Large Hadron Collider** (LHC): nearly 15 million billion bytes per year -15 petabytes (10^{15}). These data require 70,000 processors to be processed!

http://blog.websourcing.fr/infographie-la-vrai-taille-dinternet/

BIG DATA TECH ECONOMY

Big data is not a "thing" but instead a dynamic/activity that crosses many IT borders

Data-driven world guided by a rapid ROD (**Return on Data**)

→Generate value by extracting the right information from the digital universe

- Analysing patterns and spotting relationships/trends that enable decisions to be made faster with more precision and confidence
- Identifying actions and bits of information that are out of compliance with company policies can avoid millions in fines
- Proactively reducing the amount of data you pay ($18,750/gigabyte to review in eDiscovery) by identifying only the relevant pieces of information
- Optimizing storage by deleting or offloading non-critical assets to cheaper cloud storage thus saving millions in archive solutions

http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

GIVING VALUE TO DATA: DATA MARKETS

Data market: economically capitalize the effort of going out for hunting data sources and services of different qualities and stemming from different processing processes, curating and delivering them

- being able to associate a cost model for the data market, i.e., associate a cost to raw data and to processed data according on the amount of data and processing resources used for treating it, for instance
- then being able to combine these cost model and the consumer expectations (service level agreement) with processing resources cost required by data processing
- providing data management and brokering processing mechanisms under ad hoc business models

GIVING VALUE TO DATA

- Economy oriented data brokering will be tuned by the cost of accessing data markets, and the cost of using resources for brokering data

- ***Completeness***: a sample of resulting data that can be completed according to an economic model: guided by predefined fees (1M 10euros, 10M 15 euros), the user specifies whether to buy data to complete results.
- ***Data delivery frequency***: new data can be delivered, while a data market proposes new data. It is up to the user to subscribe according to different fees.
- ***Duration***: volatile/persistent results produced out of series of queries can be used for building a new data market. The owner can require accessing and buying storage services for dealing with her data markets and exporting them as paying services. The associated cost of such service will depend on the QoS and the kind of Service level agreements that the service can honour.
- ***Content quality***: data provenance, data freshness and degree of aggregation.

ROADMAP

- **Smart data for smart transport systems**
  - E-Mobility resources
  - Collecting, curating and storing data
  - Getting benefits from the XXL
- **The big data economy**
  - How big is big?
  - Where are the big data collections? Who owns them? Who gets Return of Investment?
- **Conclusions and perspectives**

“With hardware, networks and software having been commoditized to the point that all are essentially free, it was inevitable that the trajectory toward maximum entropy would bring us the current age of Big Data.”

-Shomit Ghose, Big Data, The Only Business Model Tech has Left; CIO Network
http://www.bigdatabytes.com/big-data-is-the-new-tech-economy/

OUTLOOK

Big data open issues: we are struggling to capture, storage, search, share, analyze, and visualize it

- Current technology is not adequate to cope with such large amounts of data (requiring massively parallel software running on tens, hundreds, or even thousands of servers)
- Expertise in a wide range of topics including
  - Machine architectures (HPC), Service-oriented-architecture
  - Distributed/cloud computing, operating and database systems
  - Software Engineering, algorithmic techniques and networking

Merci

Thanks

Parisa GHODOUS, U. Claude Bernard, LIRIS, parisa.ghodous@univ-lyon1.fr
Genoveva VARGAS-SOLAR, CNRS, LIG-LAFMIA, genoveva.vargas@imag.fr
Catarina Ferreira da Silva, U. Claude Bernard, LIRIS, catarina.ferreira@univ-lyon1.fr