CSP

innovazione nelle ICT

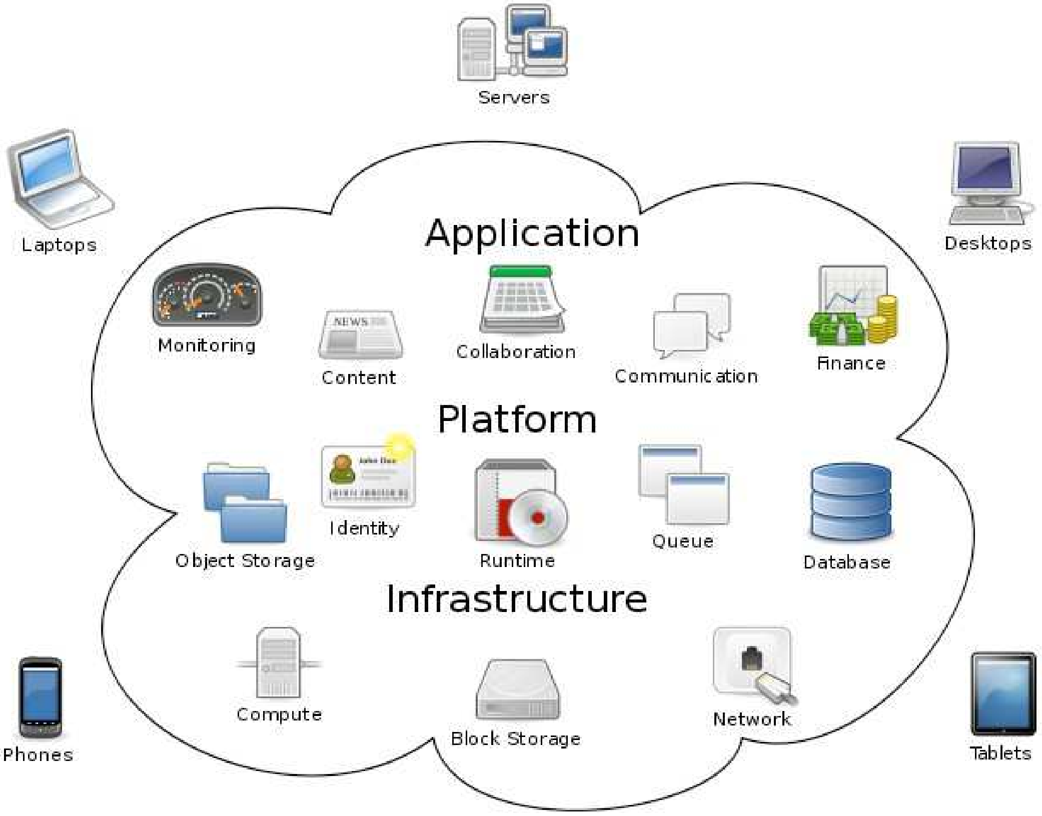
Cloud computing and M2M
Storing large amounts of interlinked data

Assago (MI), 14/05/13 - M2M Forum 2013

Rodolfo BORASO
Diego GUENZI

Cloud computing

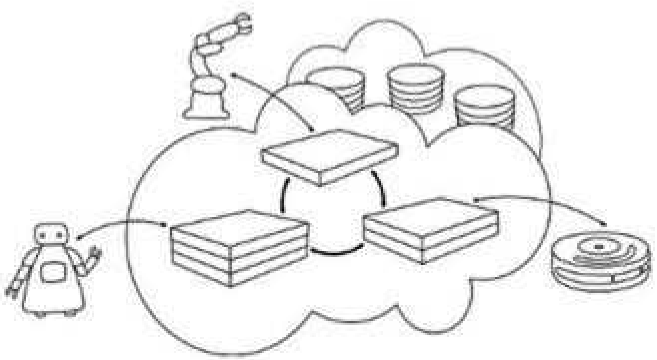
- Possible solution to obtain **computing resources** where they are not directly available (smartphones, thin clients...)
- Also usable to manage and analyze **large amounts of data**

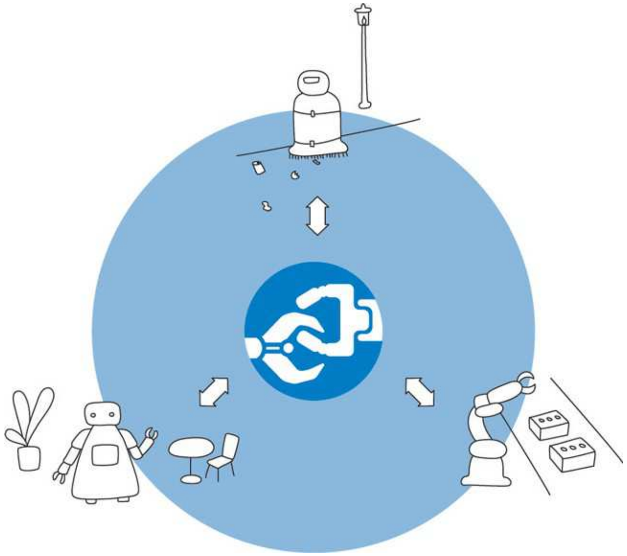


Cloud Computing

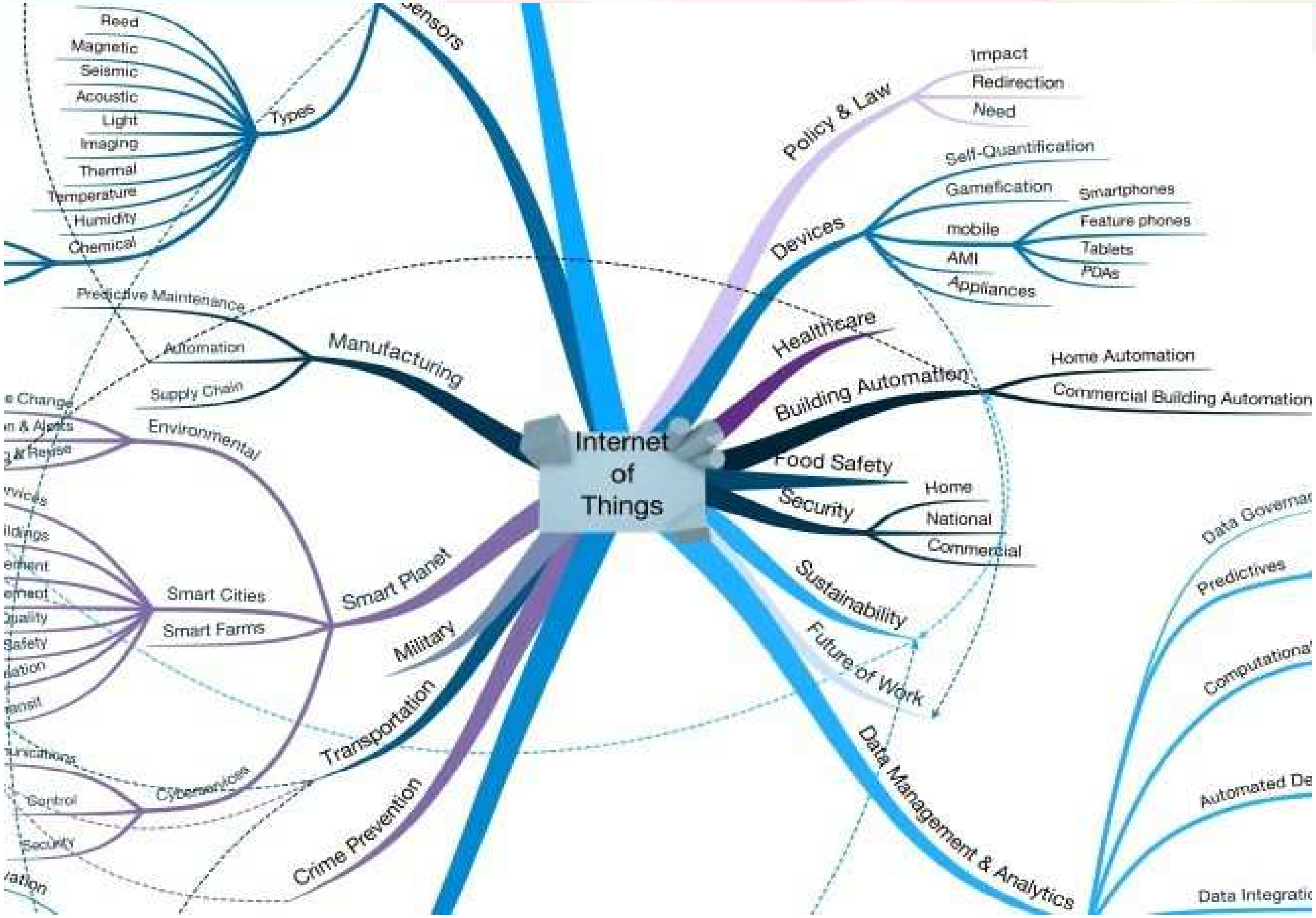
An example – RoboEarth and Rapyuta

- RoboEarth project (FP7 on robotics and cognitive systems)
- Based on open source Rapyuta platform
  - PaaS cloud realized by 5 European universities
- Objective: give robots a simple access to remote resources
  - Powerful computing resources dedicated to **heavy, CPU bound tasks** (that run on the cloud, not directly on the robot's board)
    - Lower hardware cost and better performance
  - A **large, shared knowledge database** where every robot can connect to learn new information and to share their own experiences
    - Accurate and re-used knowledge bases
- Usage examples: drones and autonomous vehicles





An example – Internet of things



- For the near future, there is a forecast of **70 billion interconnected devices** that generate, compute and transmit data over Internet
- They require a method of storing and managing this large data set

Data memorization – Object store

- Useful to build data repository on the cloud, via HTTP
- A lot of applications:
  - Document management software
  - Personal backup & sync (like Dropbox)
  - Media file archive system
  - Repository for ISO images in private cloud system (IaaS)
  - Repository for objects to be used by 3D printers (like Thingiverse)
- Two major standards adopted: Amazon **S3** and Openstack **Swift**
- Ceph and Swift are some examples of mature, stable and open source projects of object store

amazon webservices™
S3 Simple Storage Service

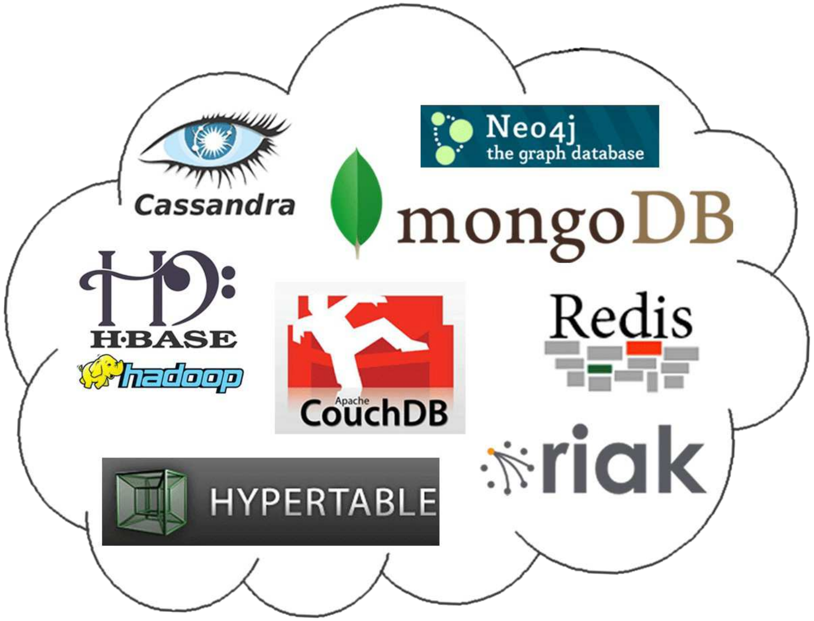
openstack™

ceph

Data memorization – Databases

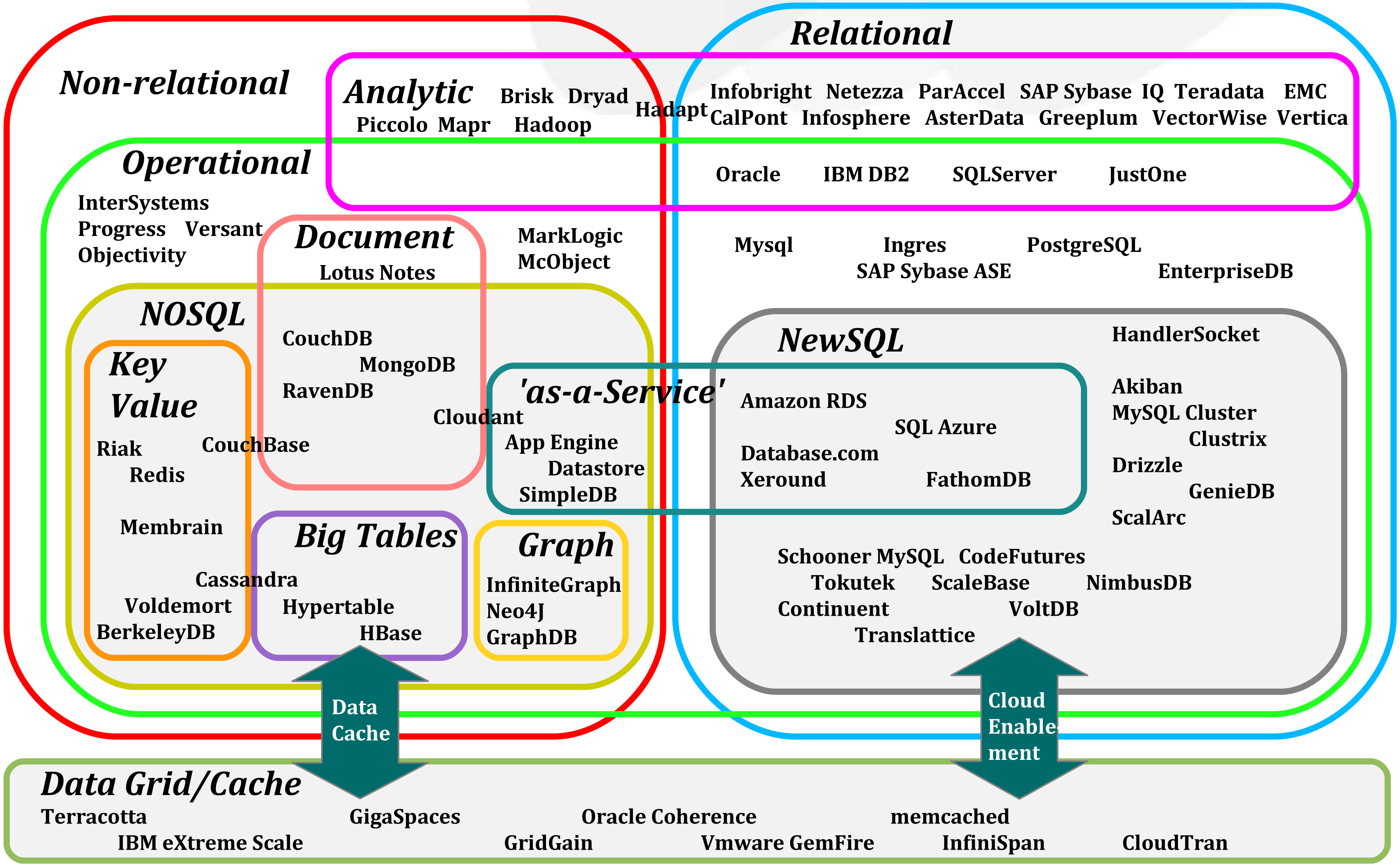
- **RDBMS** (SQL)
  - Standard, popular and well known
- **NOSQL**
  - Distributed, redundant, fault tolerant and well suited for Big Data
- **NewSQL** (or scalable SQL)
  - Standard, distributed, redundant, fault tolerant and well suited for Big Data

IBM
DB2
PostgreSQL
MySQL
ORACLE



VOLTDB
MySQL
Cluster
Drizzle

A wide choice of databases



Non-relational

Relational

Analytic

Brisk Dryad Hadapt Piccolo Mapr Hadoop

Infobright Netezza ParAccel SAP Sybase IQ Teradata EMC CalPont Infosphere AsterData Greeplum VectorWise Vertica

Operational

InterSystems Progress Versant Objectivity

Oracle IBM DB2 SQLServer JustOne

Document

Lotus Notes

MarkLogic McObject

Mysql Ingres PostgreSQL SAP Sybase ASE EnterpriseDB

NOSQL

CouchDB MongoDB RavenDB Cloudant

Key Value

Riak CouchBase Redis Membrain Cassandra Voldemort BerkeleyDB

NewSQL

HandlerSocket

'as-a-Service'

App Engine Datastore SimpleDB

Amazon RDS SQL Azure Database.com Xeround FathomDB

Akiban MySQL Cluster Clustrix Drizzle GenieDB ScalArc

Big Tables

Hypertable HBase

Graph

InfiniteGraph Neo4J GraphDB

Schooner MySQL CodeFutures Tokutek ScaleBase NimbusDB Continuent VoltDB Translattice

Data Cache

Cloud Enablement

Data Grid/Cache

Terracotta IBM eXtreme Scale GigaSpaces GridGain Oracle Coherence Vmware GemFire memcached InfiniSpan CloudTran

NOSQL

- NOSQL = Not Only SQL
  - Not a movement against SQL
  - An **alternative** to traditional RDBMS
  - A new way to see persistence
  - Applications that works in **distributed systems**, well suited for cloud computing

- Different from RDBMS
  - Do not adopt SQL language
  - Do not use fixed table schema (often, they manage semi-structured data)
  - Avoid join operations
  - Scale easily on low cost **commodity hardware**
- Complementary to RDBMS
  - **The right tool for the job**
  - Cover areas where traditional RDBMS are weak

- For some problems, other storage solutions are better suited!

NewSQL

- The **tradeoff** between NOSQL and traditional RDBMS
  - Use of relational tables and SQL
  - Same scalability as NOSQL DBMS
  - A lot of products are coming on the market: VoltDB, MySQL Cluster (NDB), ScaleDB, Xeround, Clustrix…
  - Most of them are storage engine for MySQL

Clustrix

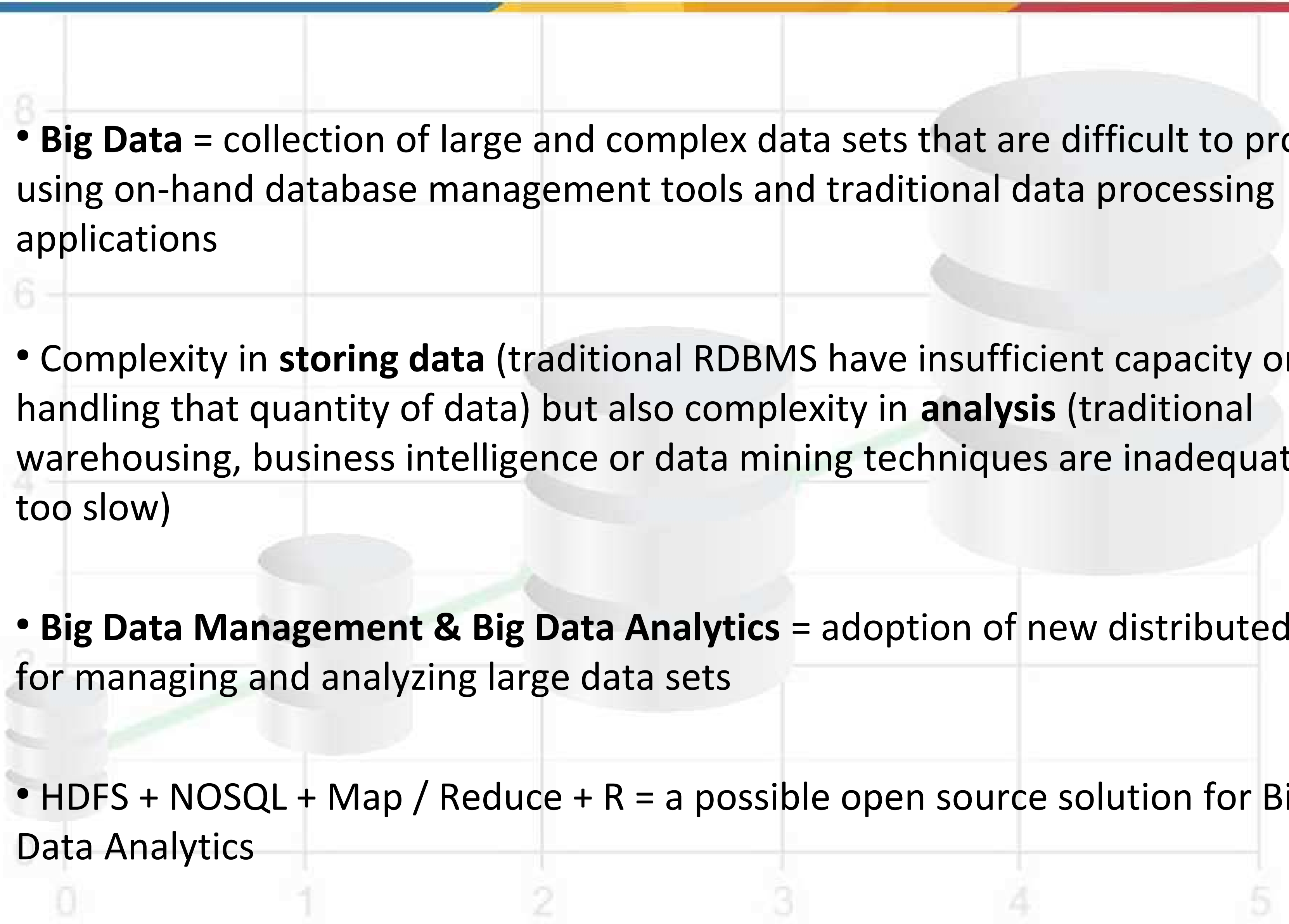
Microsoft® SQL Azure™

MySQL® Cluster

Drizzle

VOLTDB

Big Data



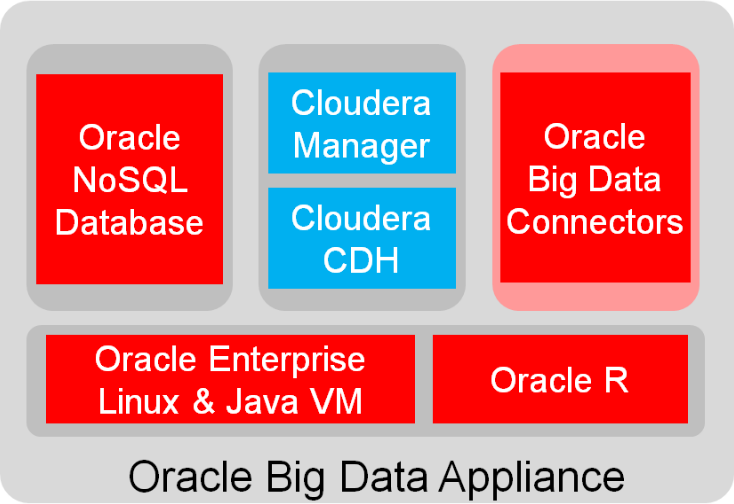
- **Big Data** = collection of large and complex data sets that are difficult to process using on-hand database management tools and traditional data processing applications

- Complexity in **storing data** (traditional RDBMS have insufficient capacity on handling that quantity of data) but also complexity in **analysis** (traditional warehousing, business intelligence or data mining techniques are inadequate or too slow)

- **Big Data Management & Big Data Analytics** = adoption of new distributed tools for managing and analyzing large data sets

- HDFS + NOSQL + Map / Reduce + R = a possible open source solution for Big Data Analytics

An example – Oracle Big Data Appliance



Oracle NoSQL Database

Cloudera Manager

Cloudera CDH

Oracle Big Data Connectors

Oracle Enterprise Linux & Java VM

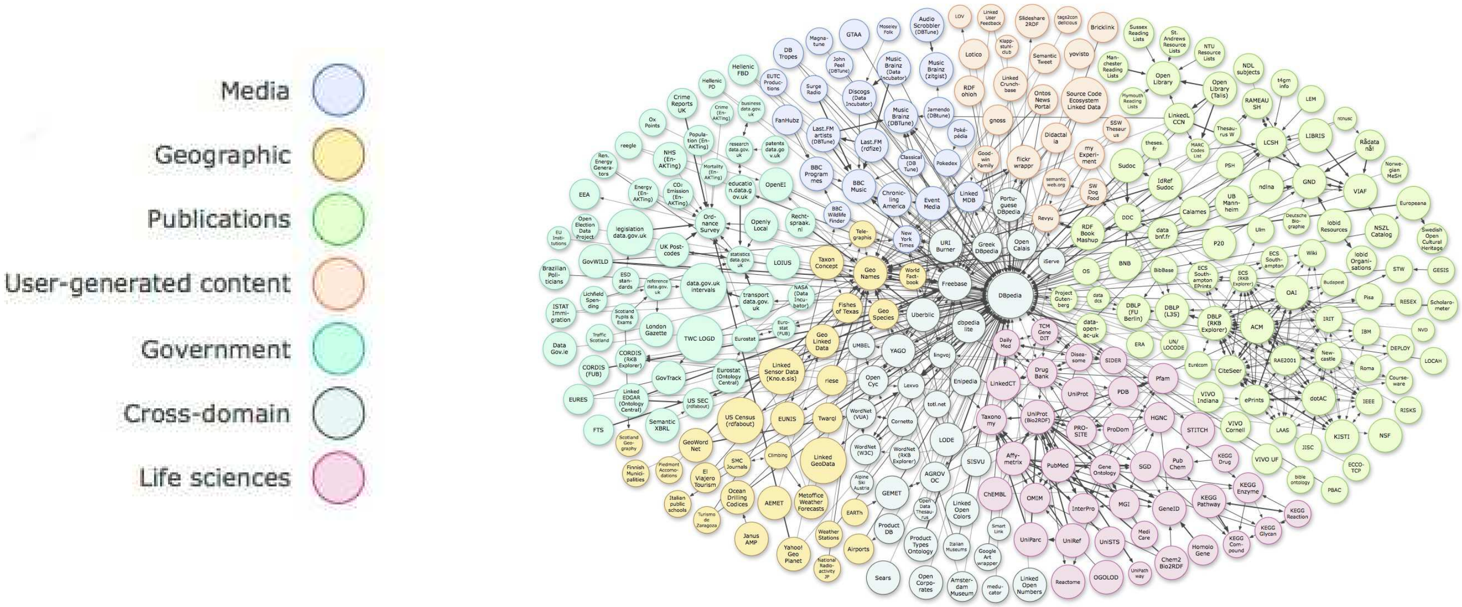
Oracle R

Oracle Big Data Appliance

- An Oracle appliance for Big Data Analytics
  - Oracle Enterprise Linux 5.6
  - CDH - Cloudera's Distribution including Hadoop
  - Oracle NOSQL Database (BerkleyDB)
  - Open source R

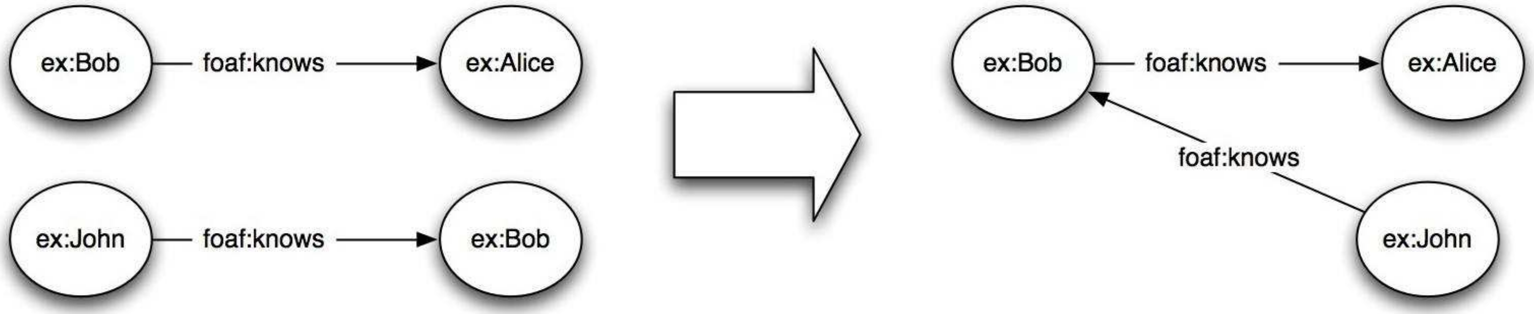
Linked Open Data

- **Linked Data** = a method of publishing structured data in a interlinked way, following the semantic web idea (Tim Berners-Lee)
- **Open Data** = freely accessible data, without copyright or restrictions of any sort
- **Linked Open Data** (LOD) = Linked Data + Open Data
  - Shared among a lot of entities, without a single owner
  - Objective: see the web as a single, big database
  - Requires a standard query language (SPARQL) that permits easy cooperation among remote data set and that uses meta-data catalogs (CKAN) to index and address real data



Data memorization – RDF store

- **Standard proposed by W3C** for application interoperability
  - Represents pieces of information about web resources
  - Based on a graph model (vertex = resource, edge = attribute)
- RDF is not the only mechanism to store LOD
  - It is the most used and flexible
- A lot of serializations
  - RDF/XML (XML file, one of the most well known and adopted)
  - RDF/JSON (JSON instead of XML file)
  - N-Triples (set of triples in the format subject – predicate – object)
  - Notation3 / Turtle (languages that describe resources with their properties, always triple based)



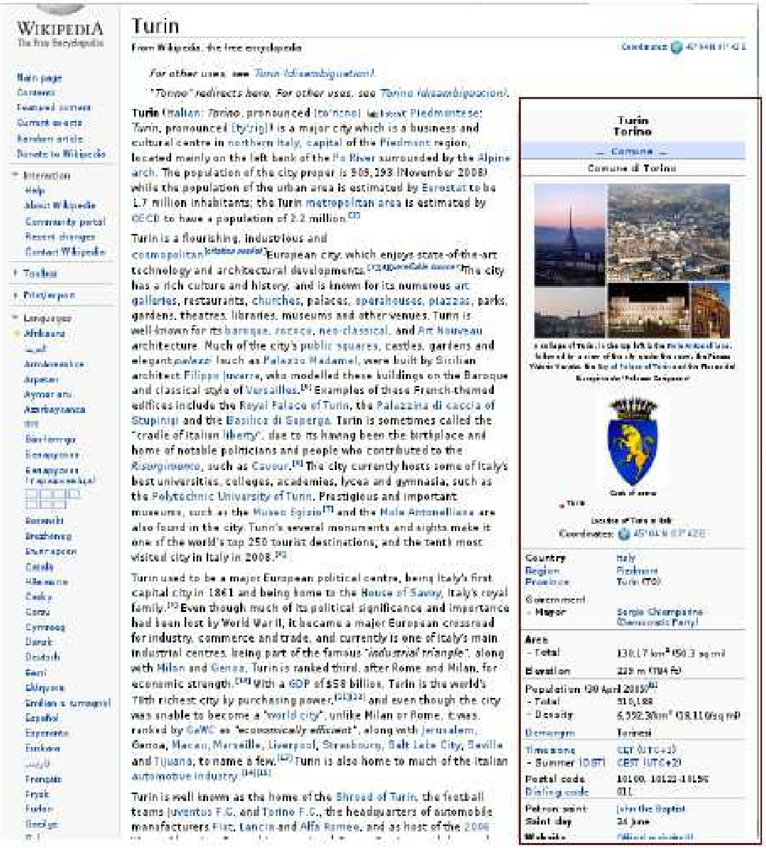
RDF repository example – DBPedia

WIKIPEDIA
L'enciclopedia libera

- Web of documents VS **web of data**
- Human centric VS **machine centric**
- Queryable SPARQL endpoint (http://dbpedia.org/sparql)

DBpedia

Wikipedia



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
[illegible]
Armãneashce
[illegible]
Aymar aru
Azərbaycanca
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
Català
[illegible]
Česky
Corsu
Cymraeg
Dansk
Deutsch
Eesti
Ελληνικά
Emiliàn e rumagnòl
Español
Esperanto
Euskara
فارسی
Français
Frysk
Furlan
Gaeilge

Turin

From Wikipedia, the free encyclopedia

Coordinates: 45°04′N 07°42′E

For other uses, see Turin (disambiguation).

"Torino" redirects here. For other uses, see Torino (disambiguation).

Turin (Italian: *Torino*, pronounced [toˈriːno]; Piedmontese: *Turin*, pronounced [tyˈriŋ]) is a major city which is a business and cultural centre in northern Italy, capital of the Piedmont region, located mainly on the left bank of the Po River surrounded by the Alpine arch. The population of the city proper is 909,193 (November 2008) while the population of the urban area is estimated by Eurostat to be 1.7 million inhabitants; the Turin metropolitan area is estimated by OECD to have a population of 2.2 million.[2]

Turin is a flourishing, industrious and cosmopolitan[citation needed] European city which enjoys state-of-the-art technology and architectural developments.[citation needed] The city has a rich culture and history, and is known for its numerous art galleries, restaurants, churches, palaces, operahouses, piazzas, parks, gardens, theatres, libraries, museums and other venues. Turin is well-known for its baroque, rococo, neo-classical, and Art Nouveau architecture. Much of the city's public squares, castles, gardens and elegant *palazzi* (such as Palazzo Madama), were built by Sicilian architect Filippo Juvarra, who modelled these buildings on the Baroque and classical style of Versailles.[3] Examples of these French-themed edifices include the Royal Palace of Turin, the Palazzina di caccia of Stupinigi and the Basilica di Superga. Turin is sometimes called the "cradle of Italian liberty", due to its having been the birthplace and home of notable politicians and people who contributed to the *Risorgimento*, such as Cavour.[6] The city currently hosts some of Italy's best universities, colleges, academies, lycea and gymnasia, such as the Polytechnic University of Turin. Prestigious and important museums, such as the Museo Egizio[7] and the Mole Antonelliana are also found in the city. Turin's several monuments and sights make it one of the world's top 250 tourist destinations, and the tenth most visited city in Italy in 2008.[8]

Turin used to be a major European political centre, being Italy's first capital city in 1861 and being home to the House of Savoy, Italy's royal family.[9] Even though much of its political significance and importance had been lost by World War II, it became a major European crossroad for industry, commerce and trade, and currently is one of Italy's main industrial centres, being part of the famous "*industrial triangle*", along with Milan and Genoa. Turin is ranked third, after Rome and Milan, for economic strength.[10] With a GDP of $58 billion, Turin is the world's 78th richest city by purchasing power,[11][12] and even though the city was unable to become a "world city", unlike Milan or Rome, it was ranked by GaWC as "*economically efficient*", along with Jerusalem, Genoa, Macau, Marseille, Liverpool, Strasbourg, Salt Lake City, Seville and Tijuana, to name a few.[13] Turin is also home to much of the Italian automotive industry.[14][15]

Turin is well known as the home of the Shroud of Turin, the football teams Juventus F.C. and Torino F.C., the headquarters of automobile manufacturers Fiat, Lancia and Alfa Romeo, and as host of the 2006

Turin
Torino

— Comune —

Comune di Torino

[illegible]

Coat of arms

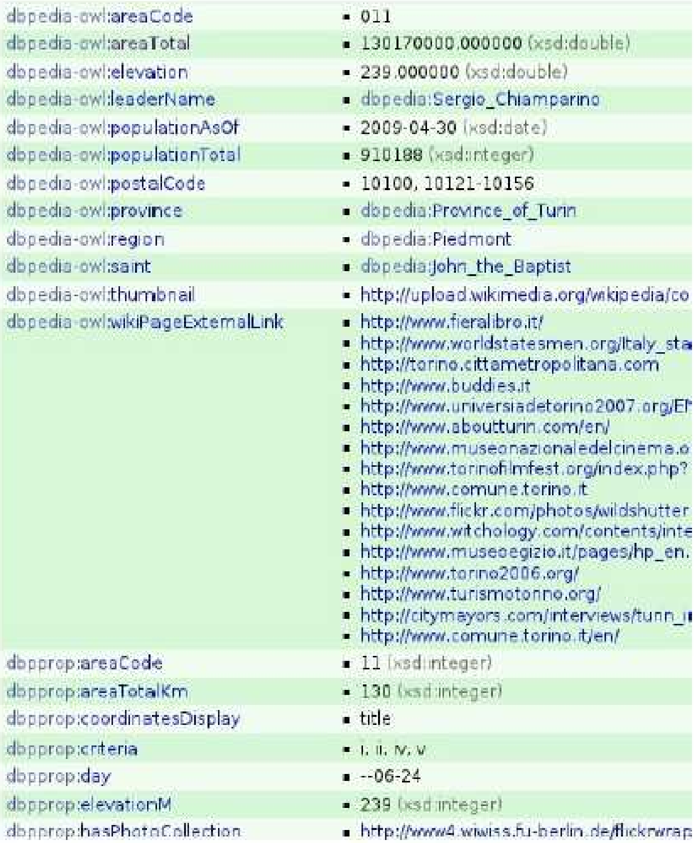
Turin

Location of Turin in Italy

Coordinates: 45°04′N 07°42′E

Country	Italy
Region	Piedmont
Province	Turin (TO)
Government	
- Mayor	Sergio Chiamparino (Democratic Party)
Area	
- Total	130.17 km² (50.3 sq mi)
Elevation	239 m (784 ft)
Population (30 April 2009)[4]	
- Total	910,188
- Density	6,992.3/km² (18,110/sq mi)
Demonym	Torinesi
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postal code	10100, 10121-10156
Dialing code	011
Patron saint	John the Baptist
Saint day	24 June
Website	Official website

DBpedia



dbpedia-owl:areaCode	▪ 011
dbpedia-owl:areaTotal	▪ 130170000.000000 (xsd:double)
dbpedia-owl:elevation	▪ 239.000000 (xsd:double)
dbpedia-owl:leaderName	▪ dbpedia:Sergio_Chiamparino
dbpedia-owl:populationAsOf	▪ 2009-04-30 (xsd:date)
dbpedia-owl:populationTotal	▪ 910188 (xsd:integer)
dbpedia-owl:postalCode	▪ 10100, 10121-10156
dbpedia-owl:province	▪ dbpedia:Province_of_Turin
dbpedia-owl:region	▪ dbpedia:Piedmont
dbpedia-owl:saint	▪ dbpedia:John_the_Baptist
dbpedia-owl:thumbnail	▪ http://upload.wikimedia.org/wikipedia/co
dbpedia-owl:wikiPageExternalLink	▪ http://www.fieralibro.it/ ▪ http://www.worldstatesmen.org/Italy_sta ▪ http://torino.cittametropolitana.com ▪ http://www.buddies.it ▪ http://www.universiadetorino2007.org/EN ▪ http://www.aboutturin.com/en/ ▪ http://www.museonazionaledelcinema.o ▪ http://www.torinofilmfest.org/index.php? ▪ http://www.comune.torino.it ▪ http://www.flickr.com/photos/wildshutter ▪ http://www.witchology.com/contents/inte ▪ http://www.museoegizio.it/pages/hp_en. ▪ http://www.torino2006.org/ ▪ http://www.turismotorino.org/ ▪ http://citymayors.com/interviews/turin_ ▪ http://www.comune.torino.it/en/
dbpprop:areaCode	▪ 11 (xsd:integer)
dbpprop:areaTotalKm	▪ 130 (xsd:integer)
dbpprop:coordinatesDisplay	▪ title
dbpprop:criteria	▪ i, ii, iv, v
dbpprop:day	▪ --06-24
dbpprop:elevationM	▪ 239 (xsd:integer)
dbpprop:hasPhotoCollection	▪ http://www4.wiwiss.fu-berlin.de/flickrwrap

SPARQL

- SPARQL (Sparql Protocol And Rdf Query Language)
  - **Query language** for RDF
  - W3C standard
  - **SOL-like** syntax, based on Turtle notation
- RDF describes concepts and relations as graphs
- SPARQL searches sub graphs matching user's query
  - SPARQL : RDF = XQuery : XML
  - SPARQL : RDF = SQL : relational model
- SPARQL query example: *list all episodes of Star Trek – The Original Series* (http://dbpedia.org/sparql)

```
SELECT ?numEpisodio, ?titolo, ?episodio WHERE {
  ?episodio dbpedia-owl:series <http://dbpedia.org/resource/Star_Trek:_The_Original_Series> .
  ?episodio dbpprop:episode ?numEpisodio .
  ?episodio dbpprop:title ?titolo
}
ORDER BY ?numEpisodio
```

Conclusions

- RDBMS are widespread (often used as a simple and well known **back-end for RDF** – like D2R)
  - Have problems in horizontal scaling
  - Have problems in managing and storing large amounts of data, in particular in distributed systems
- Object store, NOSQL and RDF share the **same goals**:
  - Simple horizontal scalability
  - Capability of storing large amounts of data
  - No fixed schema
- NOSQL / NewSQL can learn a lot from RDF
  - Decentralization
  - Inferences
- RDF can learn a lot from NOSQL / NewSQL
  - Scalability techniques
  - Sharding and data localization techniques
- Chose carefully your data storage tool...
- ...but remember: large amounts of data does not means only storage but also **data accessibility**
  - We need high performance and **scalable web server** to manage a lot of connections to large data sets (Nginx, Tornado, Cherokee...)

Diego GUENZI
Rodolfo BORASO

Distributed computing group
Services design and planning area

E-mail:
diego.guenzi@csp.it
rodolfo.boraso@csp.it
Tel:
+39 011 4815159
+39 011 4815160

CSP - Innovazione nelle ICT s. c. a r. l.

via Nizza, 150 – 10126 Torino
(entrance from via Alassio, 11/c)

Tel: +39 011 4815111
Fax: +39 011 4815001
E-mail: innovazione@csp.it

www.csp.it