

MLBDA - MU4IN801

Modèles et Langages Bases de Données Avancées

M1 Informatique

BERND AMANN

BERND.AMANN@LIP6.FR

2022

Organisation

Site web:

- Moodle:
<https://moodle-sciences-22.sorbonne-universite.fr/course/view.php?id=535>

Note :

- 2 examens répartis : ER1 50% - ER2 50%

Equipe pédagogique:

- Cours Mer 16:00-18:00 Bernd Amann
- Gr 1 Lun 8:30-10:30 et 10:45-12:45
- Gr 2 Mer 8:30-10:30 et 10:45-12:45
- Gr 3 Jeu 8:30-10:30 et 10:45-12:45

Planning

Sem	Cours	TD	TME
1	Introduction / Modélisation des données	-	-
2	SQL3: Modèle objet-relationnel / Schéma	Requêtes SQL avancées	TME SQL base Mondial
3	SQL3: Requêtes	SQL3: schémas	TME SQL3: les types
4	XML: Modèles arbres semi-structurés	SQL3 : instanciation, requêtes	TME SQL3: instanciation, requêtes
5	XML: XSchema	SQL3 : méthodes, récursion	TME SQL3: méthodes
6	XML: XPath	XML : Modèle et DTD	TME DTD
7	XML: XQuery	XML : XSchema	TME XSchema
	Congés Toussaint		
	ER 1		
	Interruption enseignement		
8	RDF : Modèle	XML: XPath	TME XPATH
9	RDF: SPARQL	XML : XQuery	TME XQUERY
10	Bases de données NoSQL et JSON	RDF / SPARQL	TME RDF et SPARQL
11	--	JSON / SQL++	TME N1QL
	ER 2		

Bibliographie

S. Abiteboul, I. Manolescu, P. Rigaux, MC.Rousset, P. Senellart :
Web Data Management, 2011, Cambridge University Press
<http://webdam.inria.fr/Jorge/index9213.html?action=chapters>

G. Gardarin : Bases de Données – objet et relationnel, Eyrolles, 2003.

G. Gardarin : XML : des bases de données aux services Web, Dunod, 2002.

F. Gandon, C. Faron-Zucker, O. Corby : Le Web sémantique, Dunod, 2012

Recommendations W3C : <https://www.w3.org/2013/data/>

MLBDA

Master Informatique - DAC

COURS 1

GESTION DE DONNÉES : ÉVOLUTIONS ET BESOINS

Des données accumulées depuis toujours

Ce qui change avec le temps

- Les technologies d'acquisition
- Les supports de stockage
- Les méthodes de traitement

Constat aujourd'hui :

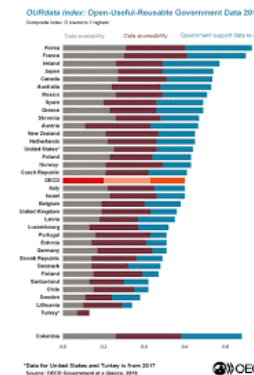
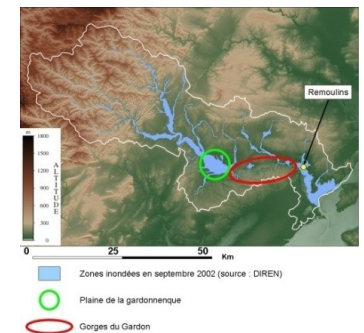
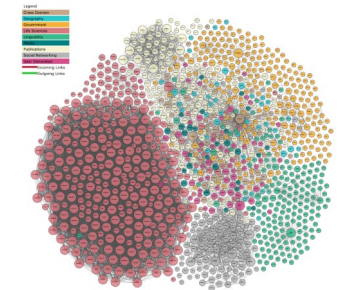
- La numérisation intensive de la société
- La diversité des données et la multiplicité des usages



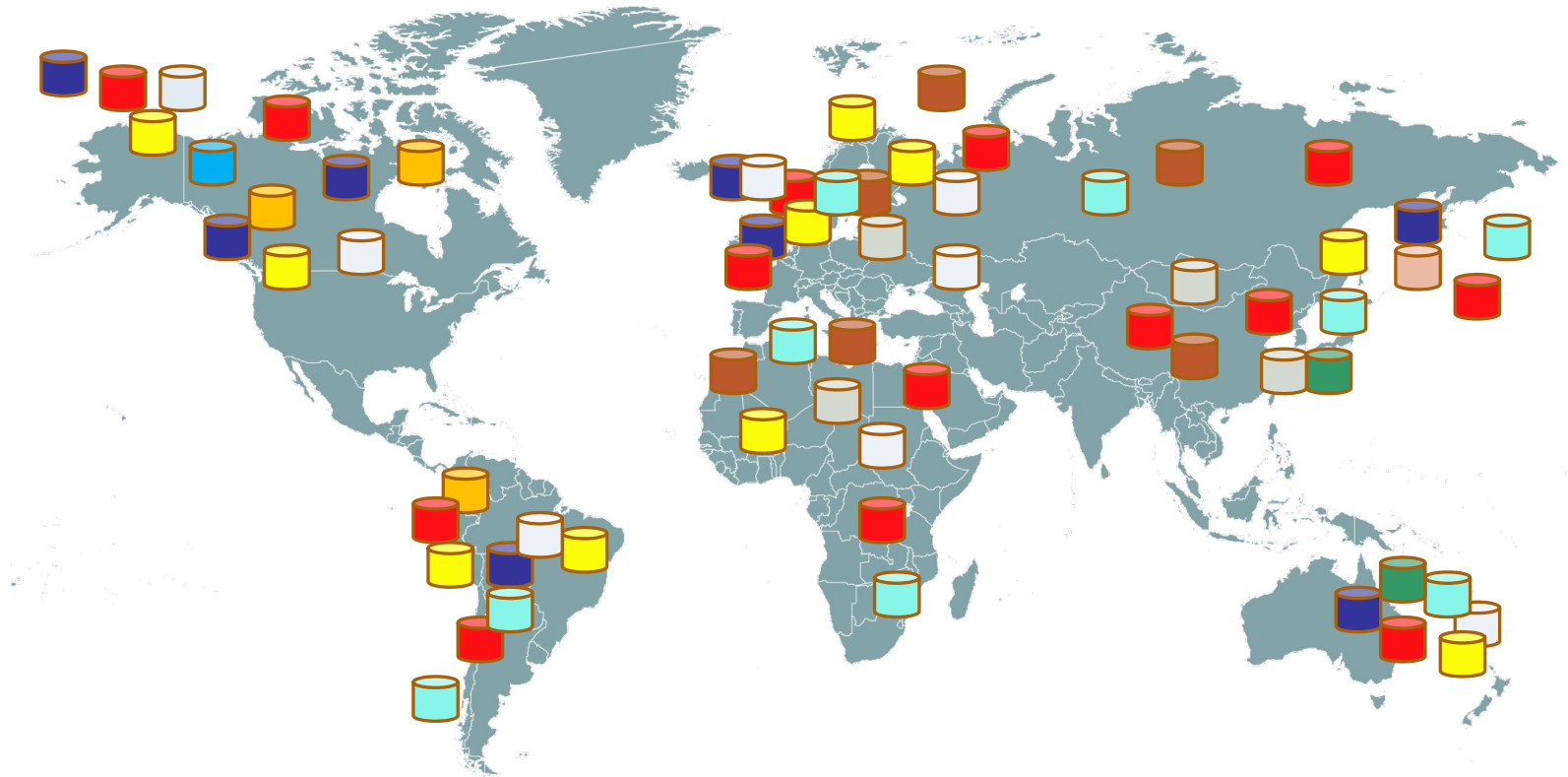
Diversification des données

- Données de références
 - Cadastre, SIG, catalogues de produits, ...
- Données d'observations
 - Satellites, capteurs, expériences scientifiques, ...
- Données transactionnelles
 - Transactions commerciales, requêtes BD/Web, ...
- Données sociales
 - Web, Facebook, Twitter, Crowdsourcing, ...
- Données du patrimoine
 - Culture, architecture, publications, ...
- ...

ProductID	Name	ProductNumber	Manufacturer	ReorderPoint	Color
1	Adjustable Ball	BA-3301	FALSE	FALSE	
2	Ball Bearing	BA-3327	FALSE	FALSE	
3	Ball Bearing	BA-3349	TRUE	FALSE	
4	Headset Ball Bearings	BE-2908	FALSE	FALSE	
5	Ball Bearing	BE-3305	TRUE	FALSE	
6	11-11 Chainring	CA-3905	FALSE	FALSE	Black
7	11-11 Chainring	CA-4738	FALSE	FALSE	Black
8	11-11 Chainring	CA-7957	FALSE	FALSE	Black
9	11-11 Chainring	CA-7957	FALSE	FALSE	Black
10	Chainring Bolts	CR-2901	FALSE	FALSE	Silver
11	Chainring Bolt	CR-6137	FALSE	FALSE	Silver
12	Chainring	CR-7813	FALSE	FALSE	Black
13	Chain Ring	CR-8983	FALSE	FALSE	
14	Chain Ring	CR-2812	TRUE	FALSE	
15	Chain Ring	CR-8752	FALSE	FALSE	
16	Chain Ring	CR-8983	FALSE	FALSE	
17	Chain Ring	CR-2177	TRUE	FALSE	
18	Chain Ring	CR-8983	FALSE	FALSE	
19	Chain Ring	CR-8983	FALSE	FALSE	
20	Chain Ring	CR-2177	TRUE	FALSE	
21	Chain Ring	CR-2177	TRUE	FALSE	
22	Chain Ring	CR-2177	TRUE	FALSE	



Distribution à l'échelle mondiale



Croissance exponentielle

En 2 jours nous avons produit plus de données que l'Humanité n'en a produit en 2 millions d'années !

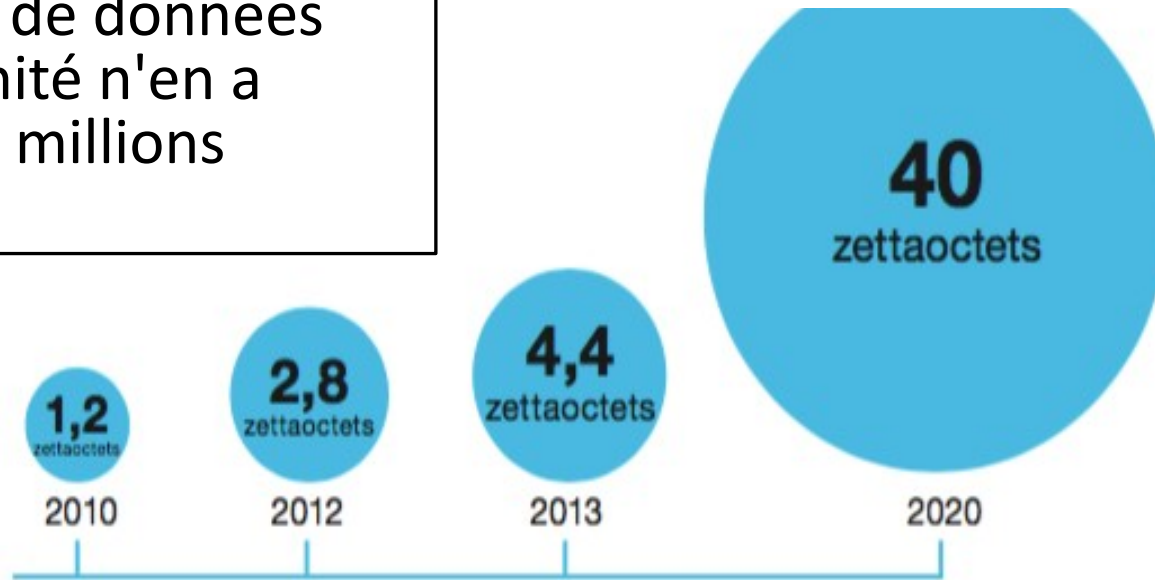
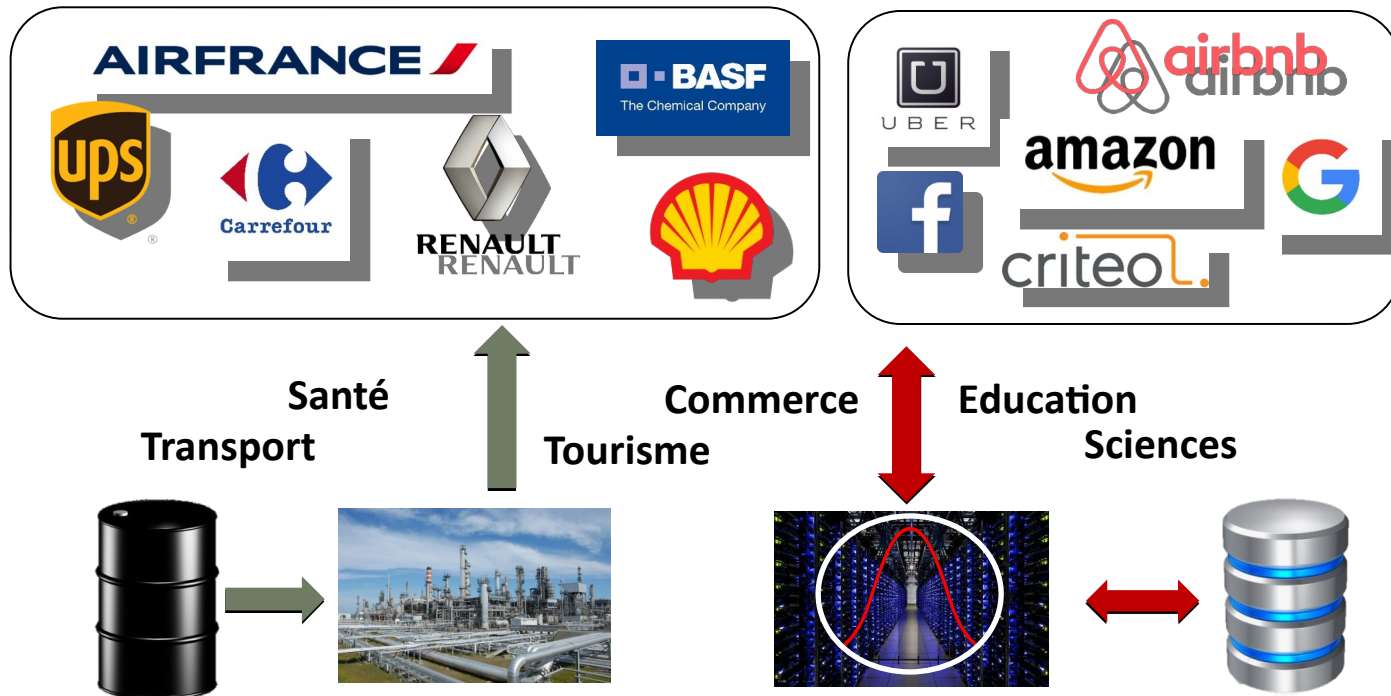
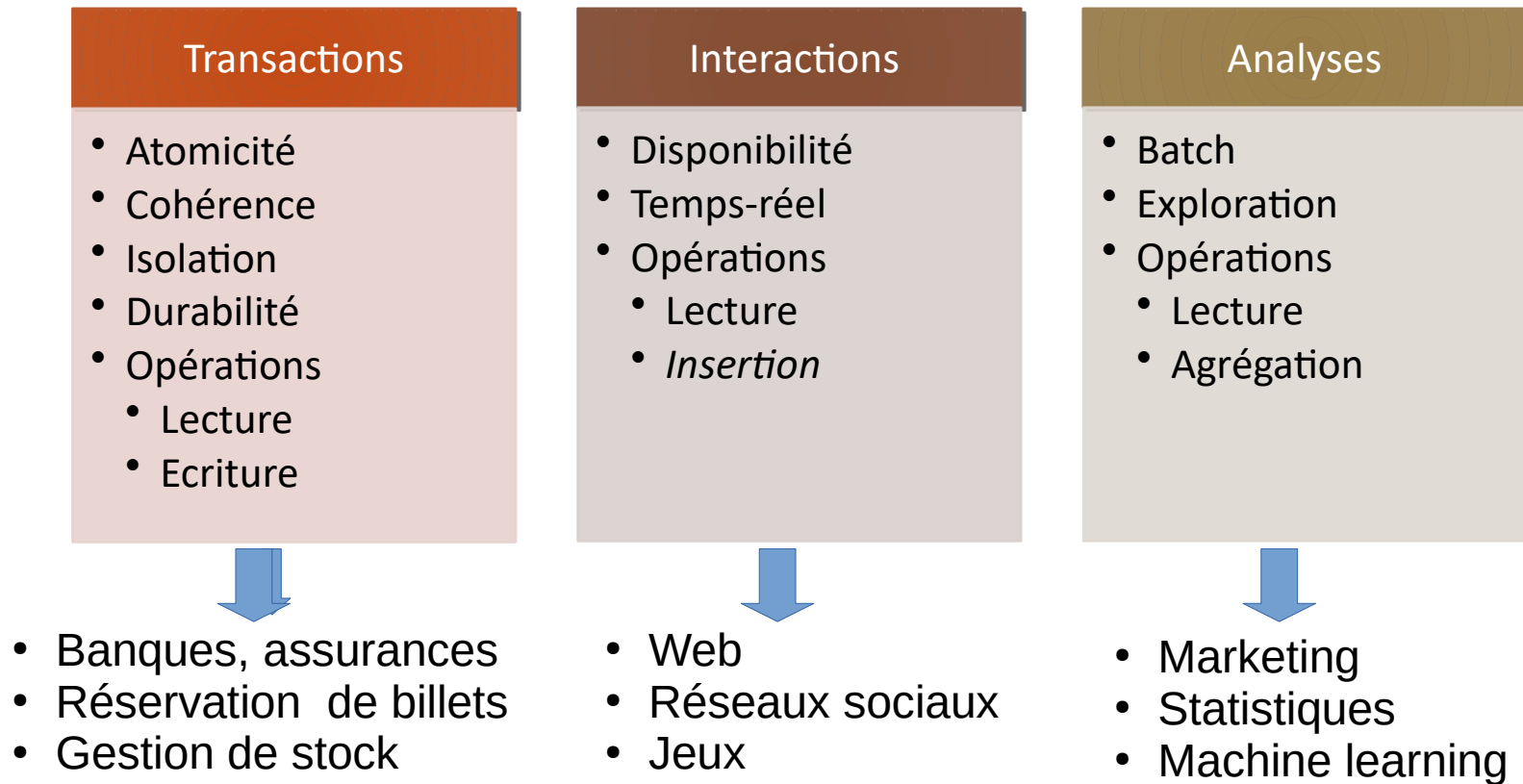


Fig. 1 – Croissance passée et prévisible de la quantité de données créées

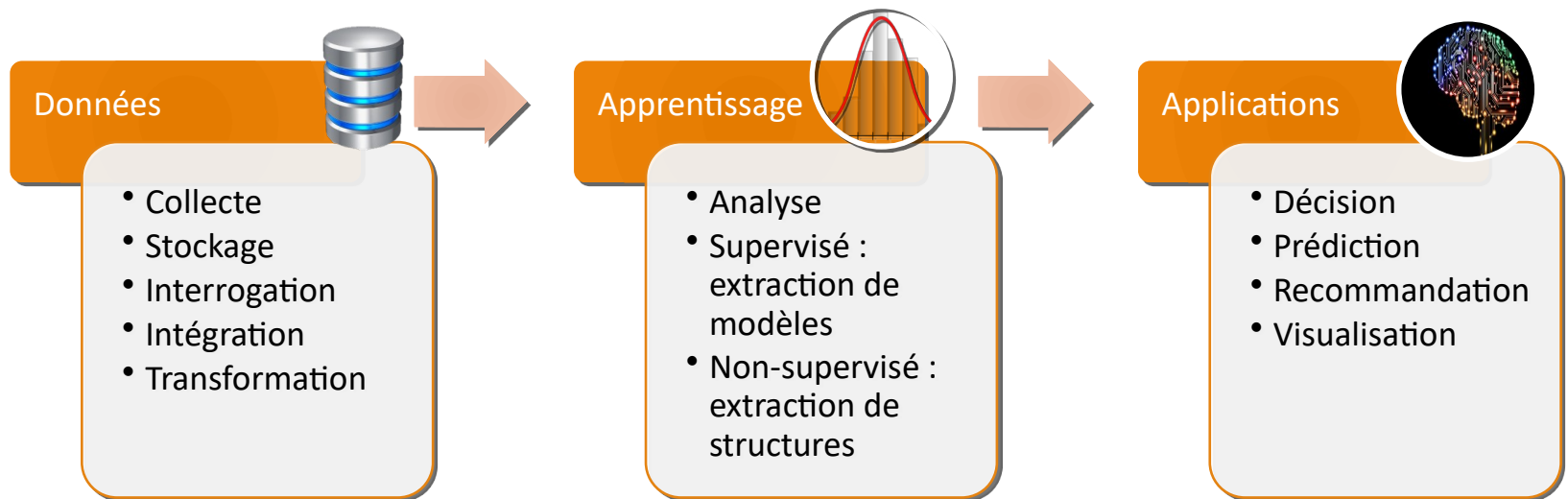
Les Données : le pétrole du 21e siècle



Types de traitements de données



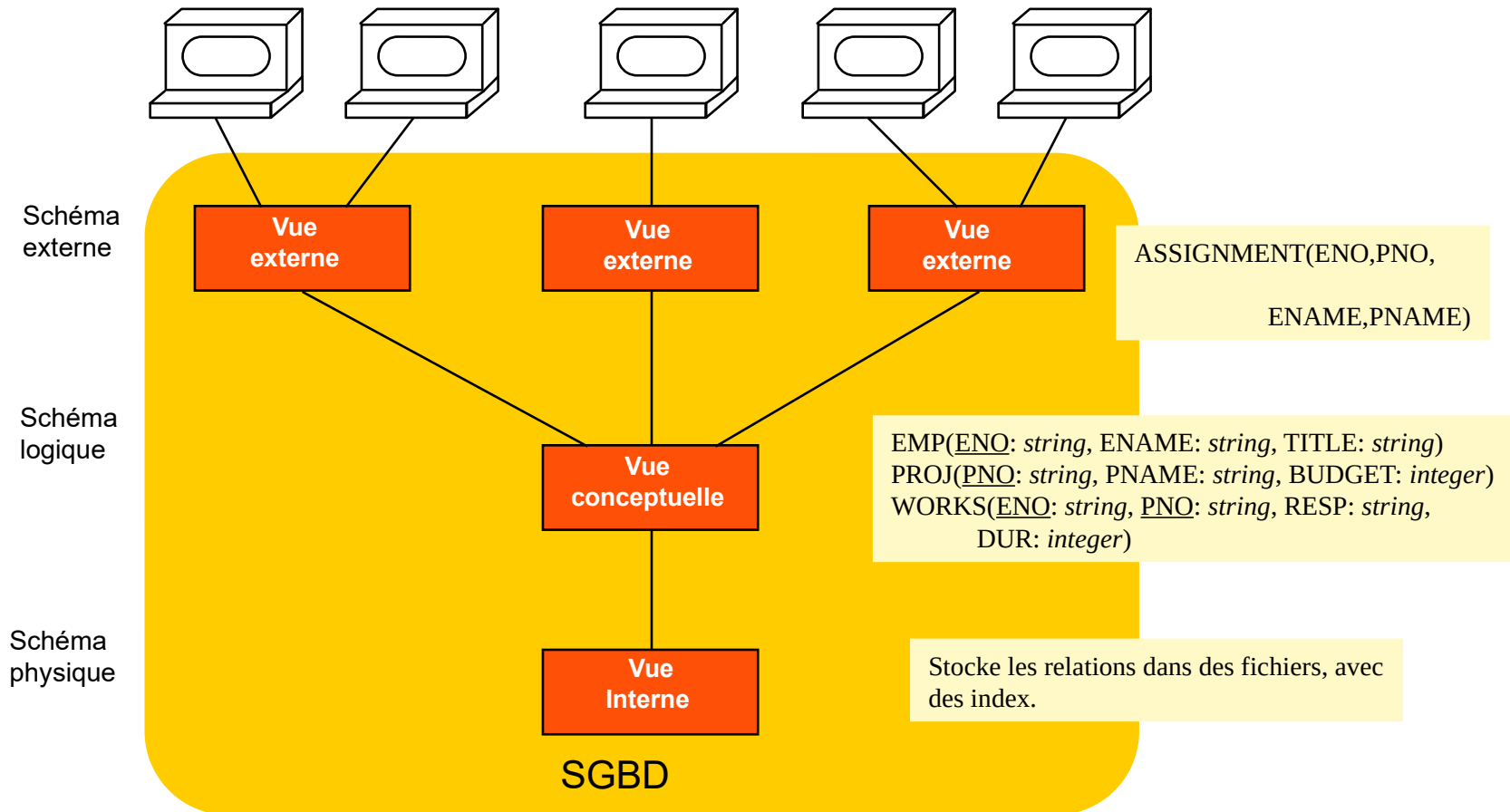
Applications « Sciences de Données »



Evolution des modèles et systèmes

Architecture classique ANSI/SPARC (1975)

Licence (L2 et L3)



Couche physique : services BD

Licence (L2 et L3)

Traitement et optimisation de requêtes

- performances garanties par le système indépendamment de la formulation de la requête
- Index, plans d'exécutions

Transactions

- exécution des requêtes / mises-à-jour par des unités atomiques (tout ou rien)
- concurrence d'accès et gestion de pannes gérée par le système

Administration système

- outils d'audit et de réglage (tuning)
- visualisation des plans d'accès

→ Séparation entre les services BD et l'architecture matériel

Couche logique : Programmation

Licence (L2 et L3)

Schéma logique

- vue uniforme des données, par ex. sous formes de relations (ou tables)

Cohérence

- $24000 \leq \text{Salaire} \leq 250000$
- l'utilisateur spécifie et le SGBD valide

Vues

- réorganisation de relations pour certaines classes d'utilisateurs

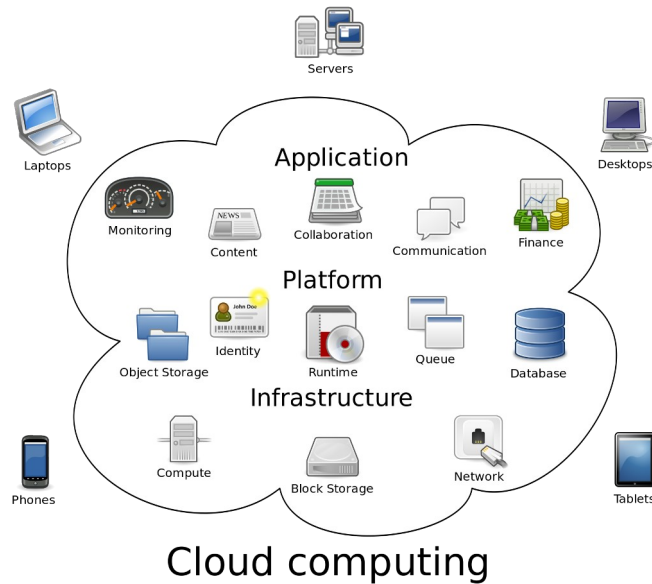
Accès déclaratif

- avec un langage de requête (SQL), l'utilisateur spécifie ce qu'il veut obtenir et non ce qu'il faut faire pour l'obtenir (le quoi et non le comment)

➔ Séparation entre les interfaces de programmation et les services BD

Évolutions récentes :

1. Virtualisation des Services



© Sam Johnston

Infrastructure (IaaS)

Plateforme (PaaS)

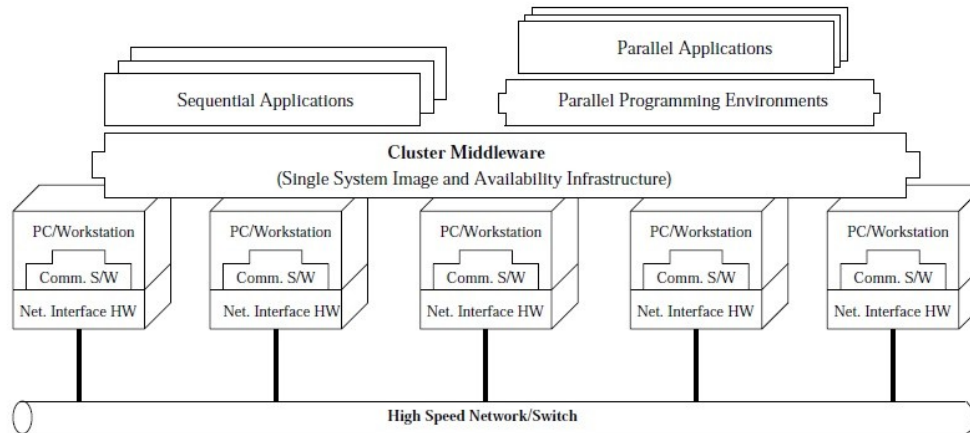
Services (SaaS)

Elasticité

➔ Séparation entre les programmes informatiques et leur environnement d'exécution

Evolution récentes:

2. Parallélisation des traitements



Nœuds de calcul et de stockage *indépendants*

Réseau *très haut débit*

Coordination par un *middleware*

Scalabilité

➔ **Augmentation des performances**

Systemes Big Data

SGBD « classiques »

Architecture verticale

Serveurs centraux avec une grande capacité de stockage et de calcul

Séparation entre la couche de calcul et la couche de stockage

Données structurées (tables)



Difficile à adapter à l'évolution des données et des besoins

Faible rapport coût / bénéfice

Systèmes Big Data

Architecture parallèle

Réseau extensible de nœuds avec « faibles » capacités de calcul et stockage

Virtualisation des ressources et services

Données hétérogènes (tables, documents, graphes, ...)



Adaptation à l'évolution des données et besoins

Coût proportionnel au bénéfice

Elasticité

Scalabilité

Modéliser et interroger des données

Données, informations, connaissances

Vue « théorique / philosophique » :

- Données : **symboles** représentant des **messages** (générées par des humains, processus,)
- Informations : **sens** attribué aux messages par le **récepteur** (humain, ordinateur, ...)
- Connaissances : **informations** « **utile** » pour le récepteur (pour prendre des décisions, raisonner, ...)

Vue « informatique » :

- Données : **séquences binaires** qui représentent des informations (ou des connaissances)
- Informations : **entités et associations** + filtrage / transformation
- Connaissances: **concepts et propriétés** + raisonnement logique et statistique

Modèles et langages bases de données :

- Transformer et manipuler des données, des informations et des connaissances

Modèle relationnel (SQL)

EMP	ENO	ENAME	TITLE
	01	Max	Ingénieur
	02	Paul	Développeur
	03	Marie	Développeur
	04	Léa	Ingénieur
	05	Luc	Développeur

PROJ	PNO	PNAME	BUDGET
	P1	Paye	500M
	P2	Stocks	800M
	P3	Livraisons	200M

Select ENAME, TITLE from EMP;

WORKS	ENO	PNO	RESP	DUR
	01	P1	01	24
	02	P2	04	20
	03	P1	01	20
	04	P2	04	18
	05	P1	01	15

Select ENAME from EMP where Title='Ingénieur';

Select ENAME
from EMP E, PROJ P, WORKS W
where E.ENO=W.ENO
and W.PNO = P. PNO
and P.PNAME = 'Paye';

Apports du modèle relationnel

Modèle fondé sur la **logique du premier ordre**

- Fondement théorique solide (expressivité et complexité bien maîtrisée)
- Sémantique SQL
- Contraintes d'intégrité

Langage déclaratif (SQL)

- Interrogation et traitements données séparées des programmes
- Indépendance d'une architecture ou d'un paradigme de programmation

Systèmes efficaces (SGBD)

- Très bonnes performances sur des grands volumes (téraoctets)
- Cohérence transactionnelle (sérialisabilité, pannes)
- Séparation des couches physique, logiques et externe (architecture ANSI/SPARC)
- Optimisation physique (index) et logique

⇒ **Un standard et une technologie efficace, sûre, éprouvée pour gérer des données structurées**

Limites du modèle relationnel

1ère Forme Normale (Codd)

- Attributs de types simples (entier, réel, chaîne, date, ...)
- Extension avec types plus complexes possible mais difficile (évaluation / optimisation)

Modélisation d'informations complexes difficile

- Éclatement des entités, informations dispersées dans plusieurs relations
- Multiplication des relations, nombreuses jointures
- Informations redondantes, non factorisées

Limites du langage SQL

Au niveau théorique :

- **expressivité limitée** : pas de récursion / itération

Au niveau pratique :

- **interactivité limitée** : utilisation difficile pour les non experts

Solutions :

- SQL + langage de programmation (Java, Python, ...)
- Intégration :
 - Programmation : "API" SQL
 - Transformation de données : table ↔ listes, objets, ...
 - Échange de données : disque ↔ mémoire (curseurs)

Limites des SGBD relationnels

Architecture verticale

- Programmes en mémoire centrale (capacité de calcul)
- Serveurs SGBD centraux (couche de calcul)

Optimisation « locale »:

- Optimiseur SQL + algorithmes externes
- Utilisation d'index limitée à l'évaluation de requêtes SQL
- Indexation limitée de données complexes

Evolution limitée:

- Difficile à adapter à l'évolution des données (formats, besoins)
- Faible rapport coût / bénéfice (scalabilité limitée)

Approche noSQL

NoSQL : **Not only SQL**

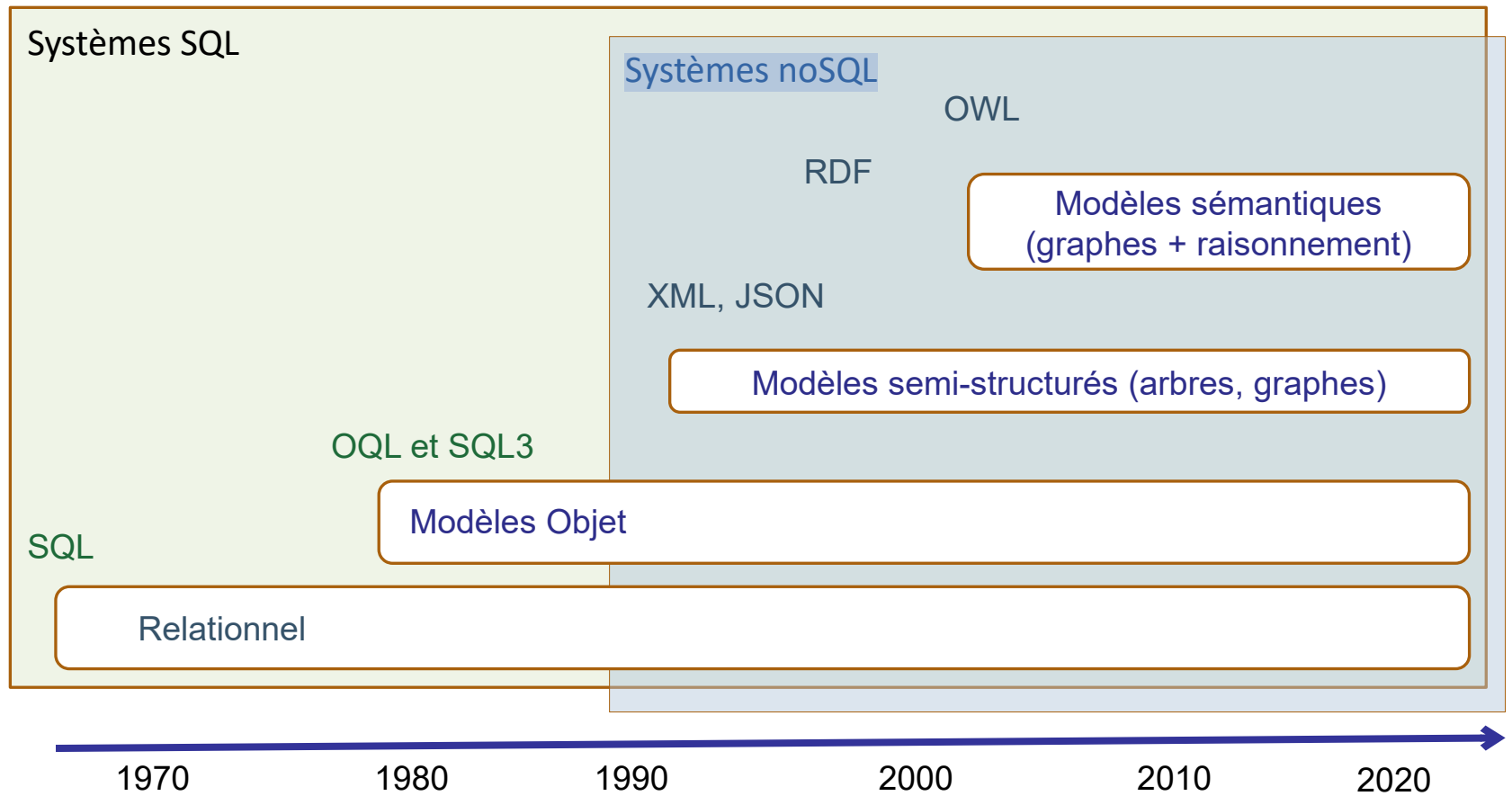
Représentation « proche » des information et des besoins applicatifs :

- Données hétérogènes : tables, graphes, documents, images
- Nouveaux traitements : échange, transformation, intégration

→ **Nouveaux modèles et langages**

- Modèles de données « plus riches » : documents, arbres, graphes
- Schéma de données : validation, intégration, transformation, inférence, ...
- Langages de requêtes : données et structure (schéma), motifs de graphes, navigation, exploration, ..

Evolution des modèles et noSQL



Systèmes noSQL

Modèle de données plus flexible, plus intuitif

Exploitation de ce modèle pour distribuer plus facilement et plus efficacement (cloud, MapReduce)

Langages spécialisés, API simples

Abandon des contraintes fortes des SGBDR (propriétés ACID)

4 catégories :

- Bases orientées colonnes,
- Stockage de couples clé-valeur,
- Bases de documents,
- Bases de graphes

De nombreux produits : MongoDB, Neo4J, Cassandra, CouchDB, BigTable, Hadoop/Hbase, MonetDB, Redis, Titan, OrientDB, BigData, graphBase, etc.

Modèle relationnel

Rois

<u>Id</u>	nom	épouse
R1	François I	NULL
R2	Charles IX	R8
R3	Henri II	R5
R4	François II	R6

Reine

<u>Id</u>	nom	époux
R5	Catherine de Medicis	NULL
R6	Mary Stuart	R6
R8	Elisabeth d'Autriche	R2
R9	Henry III	NULL

Fils

<u>Id</u>	<u>fils</u>
R1	R3
R5	R2
R5	R4
R5	R9

Schéma

- Tables nommés avec plusieurs attributs nommés
- Valeurs NULL : données absentes, manquantes, inconnues
- Contraintes d'intégrité
 - Clés et clés étrangères (intégrité référentielle)
 - Dépendances fonctionnelles
- **Obligatoire** : donnée = instantiation d'un schéma

Requêtes

- Expressive et déclarative
 - Requête SQL = formule logique
- Optimisation :
 - Index, algèbre relationnel, réécriture
- Maîtrise du schéma obligatoire

XML

Document XML

- Représentation textuelle → échange de données

Arbre DOM:

- Représentation arborescente → accès structuré
- Nœud : éléments / attributs, texte
- Arcs
 - Parent / enfant
 - Identifiants et références (attributs d'éléments)

Schéma

- Grammaire d'arbres
- Contraintes :
 - Types
 - Clés et références, ...
- Optionnel : permet la *validation* de données existantes

Requêtes :

- Navigation dans arborescence (Xpath)
- Langage fonctionnel ensembliste (Xquery)

```
<?xml version="1.0" encoding="utf-8"?>
<Dynastie>
  <Roi id="r1">
    <fils>
      <Roi ref= "#r3">
        <nom>Henri II</nom>
        <epouse id="r5"/>
      </Roi>
    </fils>
    <nom>Francois I</nom>
  </Roi>
  <Roi id="r9">
    <nom>Henry III</nom>
  </Roi>
  <Reine id="r8">
    <nom>Elisabeth d Autriche</nom>
    <epoux>
      <Roi id="r2">
        <epouse ref="#r8"/>
        <nom>Charles IX</nom>
      </Roi>
    </epoux>
  </Reine>
  ....
```



JSON

XML « simplifié »

- Représentation textuelle (document) et structurée

Modèle souple:

- valeurs simples
- ensemble de **clés-valeurs** (objets)
- liste d'objets et de valeurs

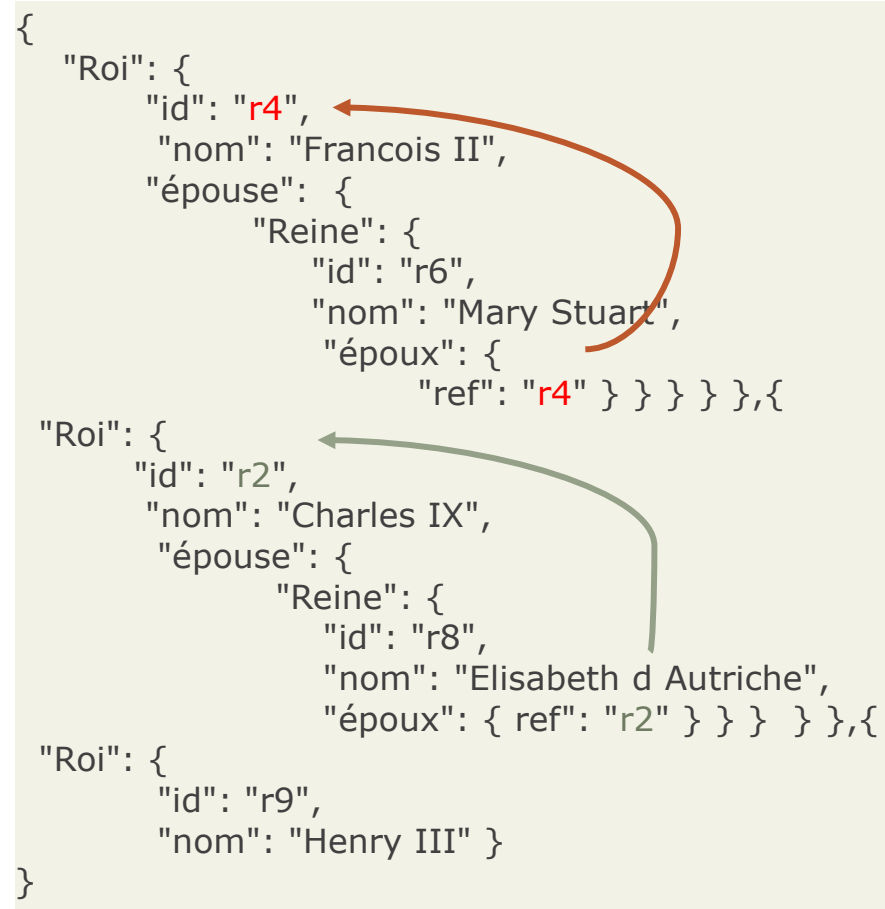
Schéma

- Optionnel pour la validation et l'optimisation

Requêtes

- **SQL « semi-structuré »** (SQL+ +)

```
{
  "Roi": {
    "id": "r4",
    "nom": "Francois II",
    "épouse": {
      "Reine": {
        "id": "r6",
        "nom": "Mary Stuart",
        "époux": {
          "ref": "r4" } } } } },{
  "Roi": {
    "id": "r2",
    "nom": "Charles IX",
    "épouse": {
      "Reine": {
        "id": "r8",
        "nom": "Elisabeth d Autriche",
        "époux": { ref": "r2" } } } } },{
  "Roi": {
    "id": "r9",
    "nom": "Henry III" }
}
```



RDF

Modèle

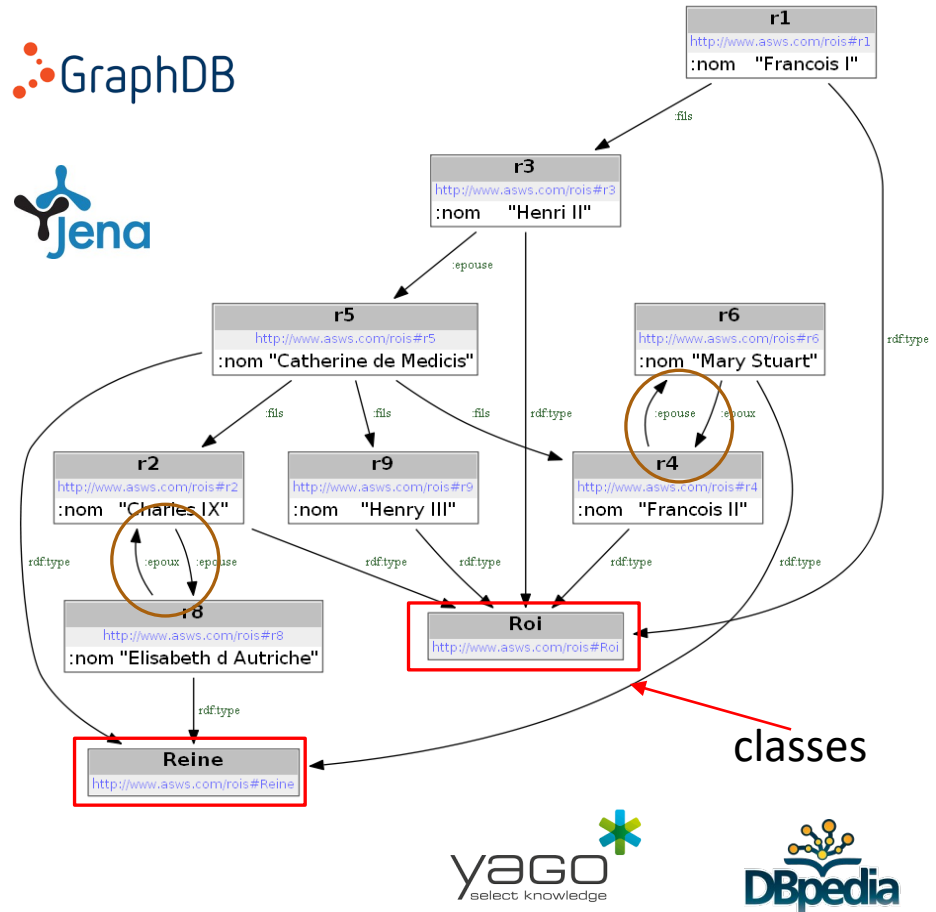
- Représentation textuelle, ensembliste et graphe

Schéma

- Classes et types de propriétés
- Très peu de contraintes → flexibilité
- Optionnel, utilisé pour faire de l'inférence sémantique

Requêtes

- **Motifs de graphes**
- **Inférence sémantique**



RDF

@prefix : <> .

:r1 a :Roi ;
:fils :r3 ;
:nom "Francois I" .

:r3 a :Roi ;
:epouse :r5 ;
:nom "Henri II" .

:r5 a :Reine ;
:fils :r2,
:r4,
:r9 ;
:nom "Catherine de Medicis"
.

:r6 a :Reine ;
:epoux :r4 ;
:nom "Mary Stuart" .

:r8 a :Reine ;
:epoux :r2 ;
:nom "Elisabeth d Autriche" .

:r9 a :Roi ;
:nom "Henry III" .

:r2 a :Roi ;
:epouse :r8 ;
:nom "Charles IX" .

:r4 a :Roi ;
:epouse :r6 ;
:nom "Francois II" .

On résume...

Observation :

- Evolution importante des modèles en fonction du type des données et des applications

Besoins:

- Modèles et langages proches des données et des besoins
- Adaptation des techniques des bases de données existantes (relationnelles) et de définition de nouvelles techniques

Objectifs du cours

Comprendre les problèmes et les solutions de modélisation et de manipulation d'informations complexes :

- Comparer différents modèles de données pour la représentation d'informations (documents, objets, flux, graphes, ...)
- Apprendre les modèles / langages associés pour définir et manipuler ces données (SQL3, XML, JSON, RDF)
- Utiliser les outils et technologies de gestion de données