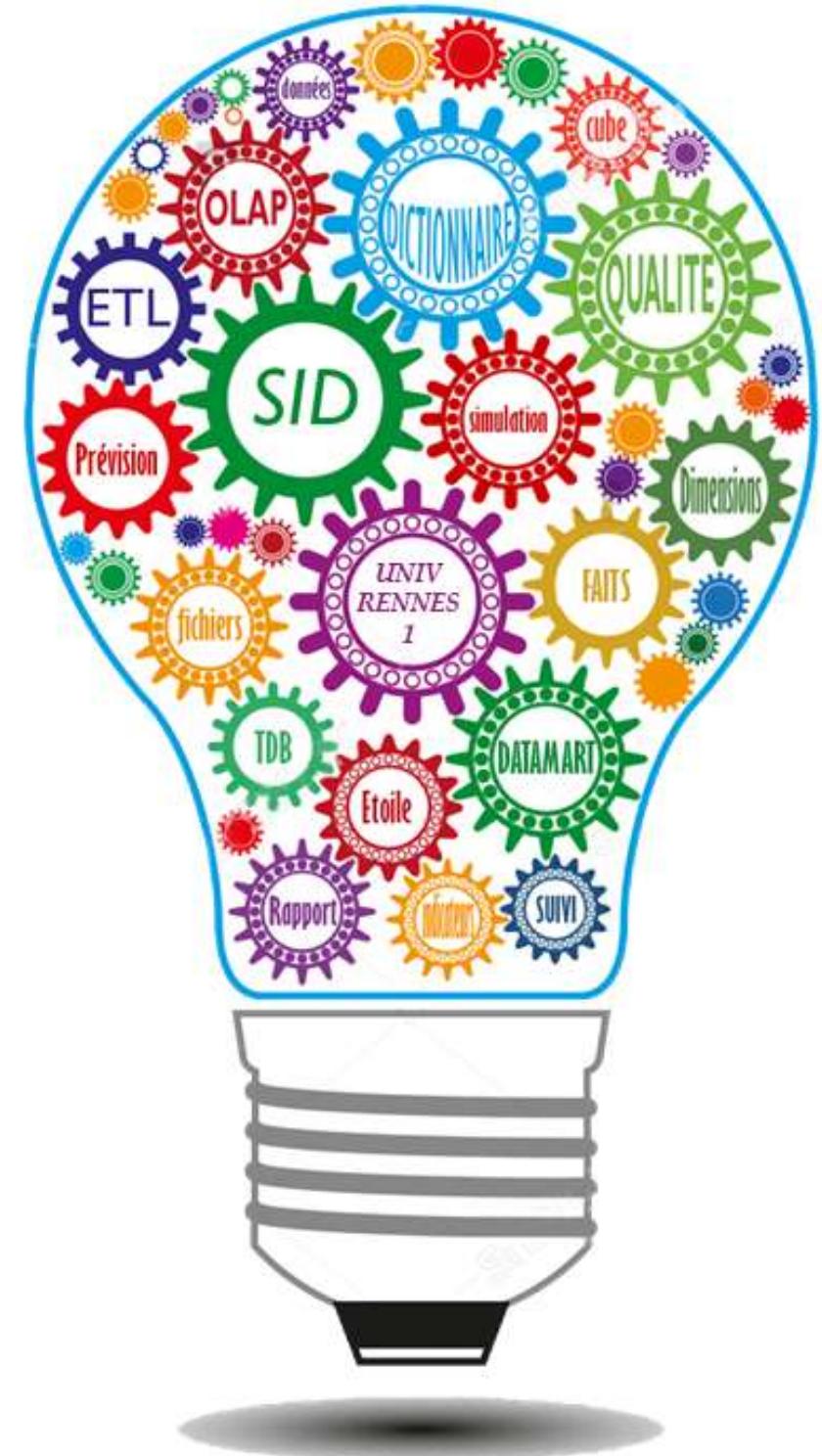


Système D'Information Décisionnel ou Business Intelligence



PLAN DE COURS

Généralités
sur la B.I.

Les modules
décisionnels

Le
dictionnaire
des données

La
modélisation
des données

L'ETL

La base de
données

Les outils
d'analyse et
restitution

Gestion de la
plate-forme
décisionnelle

Le Big Data
dans tout ça

C'est quoi la BI / SID



C'est quoi l'objectif de la BI / SID

L'objectif du décisionnel est

**L'amélioration des
performances de l'entreprise**

**par la mise en œuvre d'un processus automatisé de
collecte et analyse de données issues des systèmes
de l'entreprise, ou de données externes**

In fine : aider à la prise de décision



C'est quoi l'objectif de la BI / SID

D'après le Forrester

« ***La Business intelligence est un ensemble de méthodologies, processus, architectures et technologies qui transforment les données brutes en informations utiles, afin d'améliorer les connaissances stratégiques, tactiques et opérationnelles, ainsi que la prise de décision.*** »



C'est quoi l'objectif de la BI / SID

Attention au terme réducteur BI

Dans beaucoup d'entreprises, le terme « BI » est réducteur, et se cantonne à la partie « restitution » uniquement.



Un peu d'histoire

60 ans d'histoire....

Paternité : Hans Peter Luhn en 1958, chercheur chez IBM

« A business Intelligence System » paper

“Abstract: An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize data-processing machines for auto-abtracting and auto-encoding of documents and for creating interest profiles for each of the “action points” in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points. This paper shows ...”



Un peu d'histoire

De 1958 à 1980 : c'est un concept puisque « tout est papier »

De 1980 à 1990 : Les entreprises s'informatisent, mais uniquement pour la partie « production ». Il est encore délicat d'aller chercher de l'information de manière globale.

Début des « infocentres » : On copie les données de production sur un autre serveur (tout ou partie). C'est le début réel de « l'informatique décisionnelle »

1990 à 2000 : Professionnalisation des solutions : organisation des structures de données dans un objectif d'analyse, ETL, OLAP, tableaux de bord, ...



Un peu d'histoire

De 2000 à 2010 : Le décisionnel devient une composante très importante dans l'entreprise, des équipes et budgets importants sont dédiés pour son bon fonctionnement.

Les performances des ordinateurs n'ont plus beaucoup de limite, et il est possible de réaliser des « reports » mais surtout de prédire avec de plus en plus de finesse ce qui va se passer.

Une des difficultés pour les entreprises : comment gérer le volume de données qui devient gigantesque puisque tous les domaines de l'entreprise sont couverts

Depuis 2010 :

Architectures Big Data, démocratisation des systèmes grâce au solutions SaaS (Software As A Service). Des coûts « abordables », des boîtes à outils accessibles à tous, dans tous les langages...



Système décisionnel versus Système Opérationnel

Quelque constats

- **Vision des systèmes opérationnels (SIO ou SIOP)**
 - S.I. historiquement conçus pour le traitement de l'information
 - -> Origine du terme français « informatique »
 - Automatisation de tâches
 - Mise en œuvre et déroulement de processus de travail
 - Abordés indépendamment par secteur fonctionnel
 - Besoins et budgets distincts propres à chaque Direction
 - Expertise fonctionnelle interne
 - Fédération difficile d'un point de vue technique



Système décisionnel versus Système Opérationnel

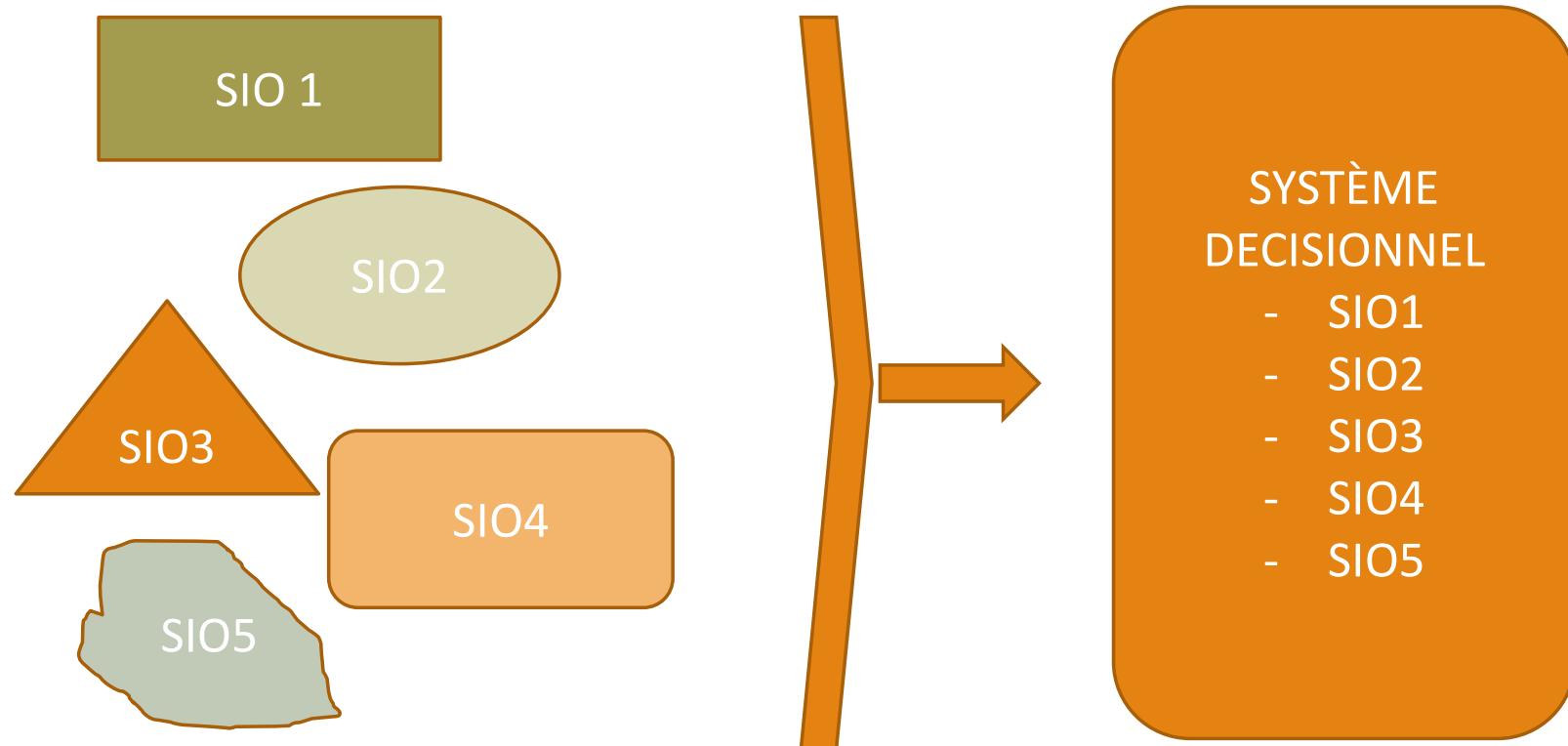
Données du SIO

- Difficiles d'accès
- Présentes sous une forme éloignée de la vision d'un utilisateur (profil fonctionnel par essence)
- Redondantes, non uniformisées
- Volatiles
- Qualité souvent non pas mauvaise mais incertaine (difficulté à mesurer le niveau de qualité fourni)



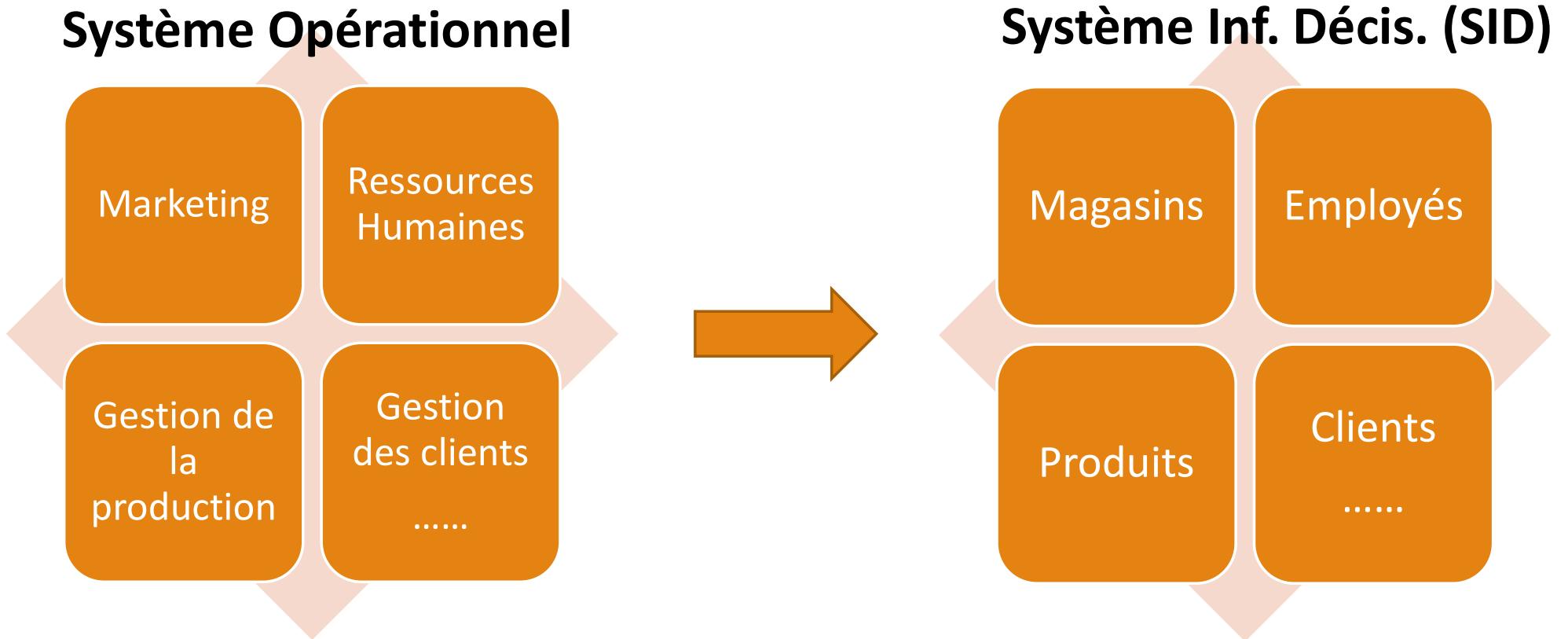
Système décisionnel versus Système Opérationnel

L'OBJECTIF : CONSOLIDER **ET HISTORISER** EN UN LIEU UNIQUE LES DONNEES DES SYSTEMES OPERATIONNELS QUI EVOLUENT TOUS DIFFEREMMENT !

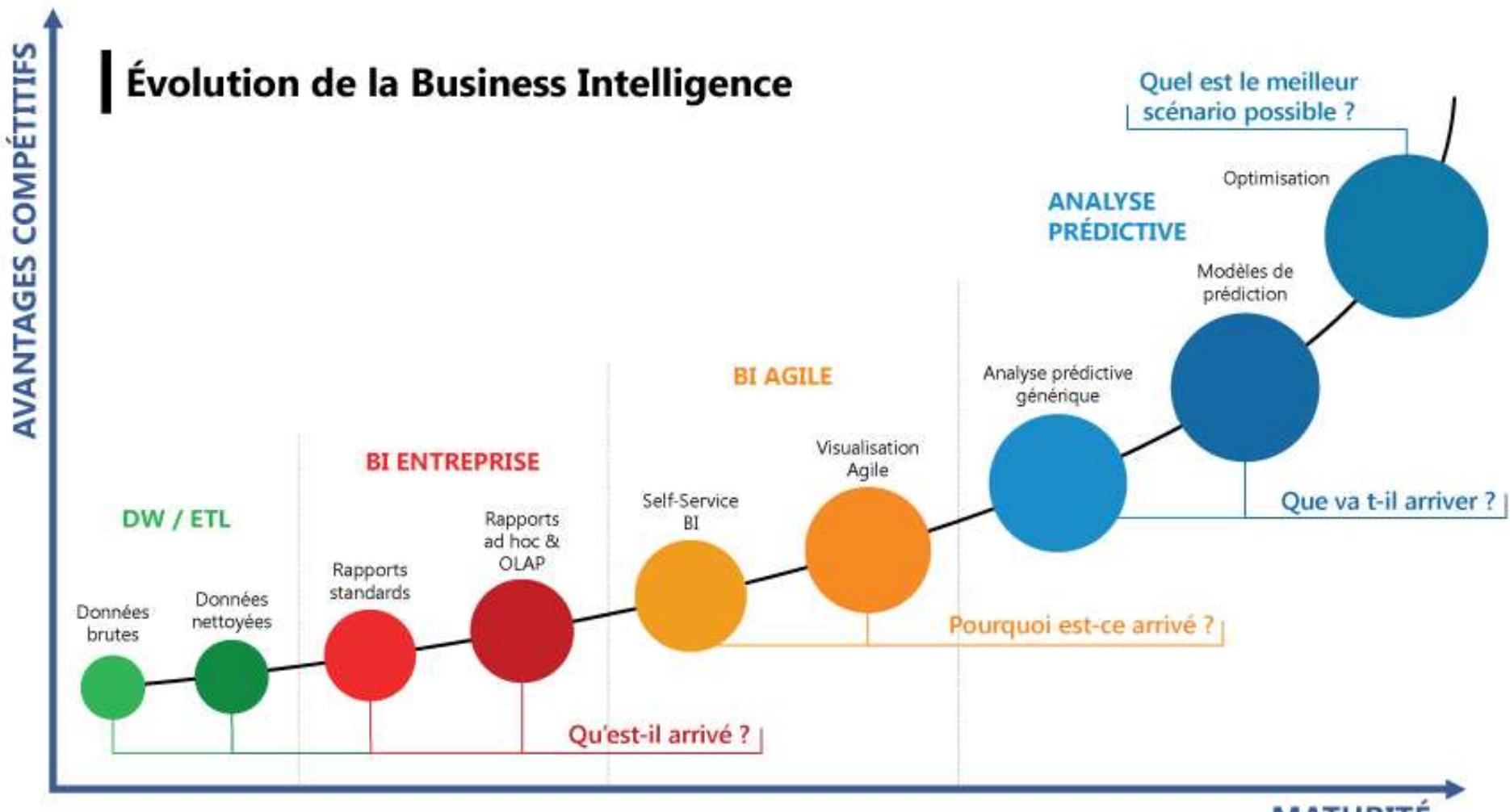


Système décisionnel versus Système Opérationnel

Travailler sur la transformation du modèle des données. Au départ, on a des applications, on va devoir « penser » plutôt en « entités fonctionnelles », ou « entités métiers »



Niveau de maturité d'une entreprise



<http://www.decivision.com/evolution-business-intelligence>



Une panoplie de solutions

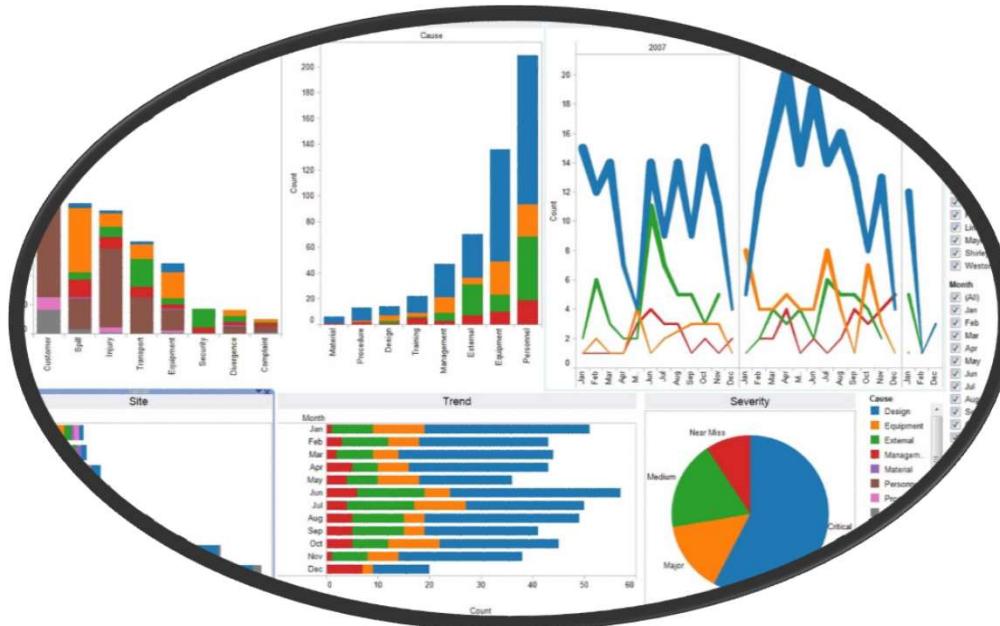
Le plus dur est de choisir !

“Analytics and business intelligence (ABI) platforms enable less technical users, including businesspeople, to model, analyze, explore, share and manage data, and collaborate and share findings, enabled by IT and augmented by artificial intelligence (AI). ABI platforms may optionally include the ability to create, modify or enrich a semantic model including business rules.

Security, Governance, Cloud-Enabled, Data Source Connectivity, Catalog, Automated Insights, DataViz, Natural Language Query, Data Storytelling, Natural Language Generation, reporting “



Le SID vu par les métiers

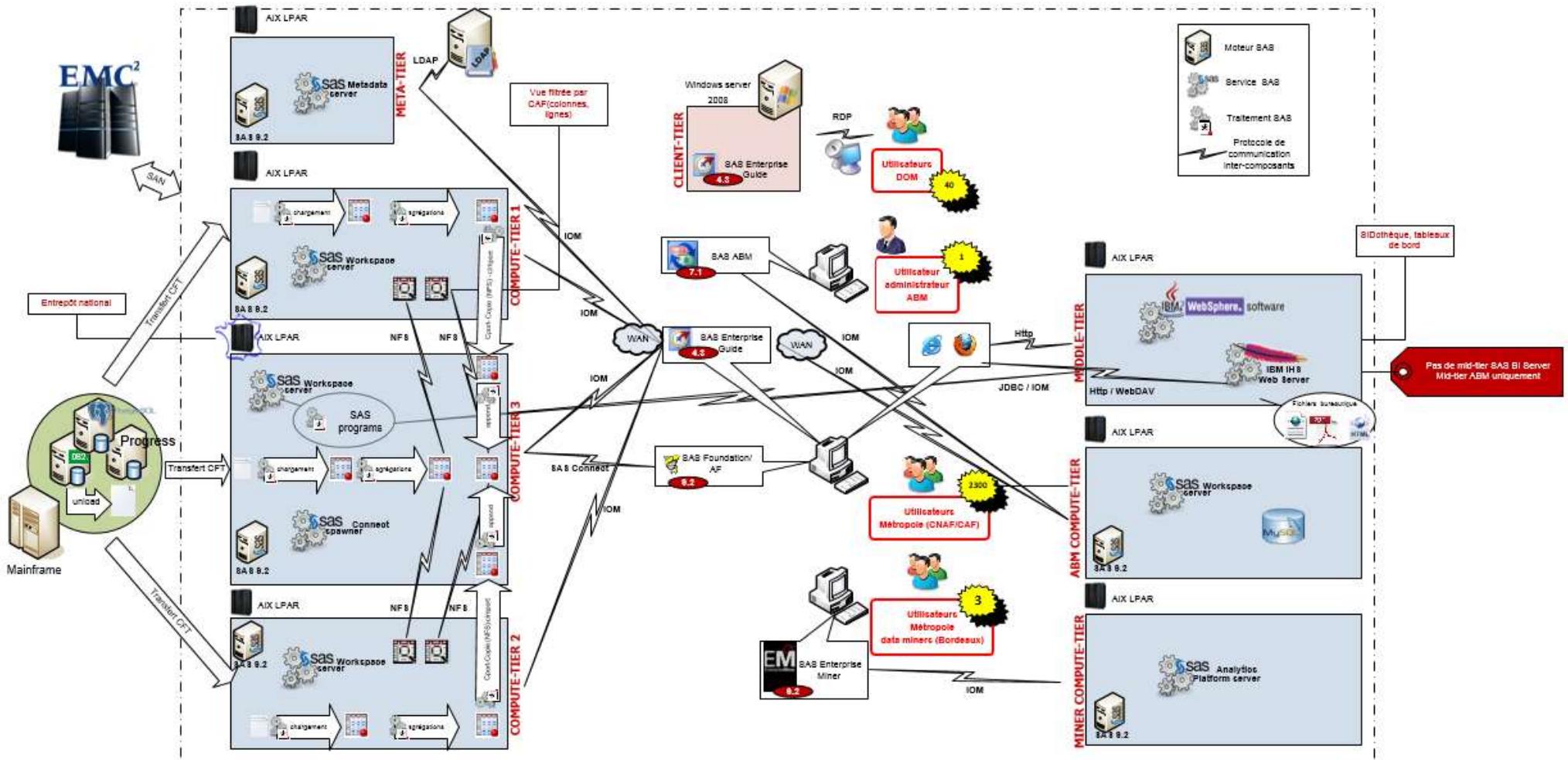


This figure shows a detailed data grid with numerous columns, likely representing a database or a large dataset. The columns include various identifiers and numerical values, such as:

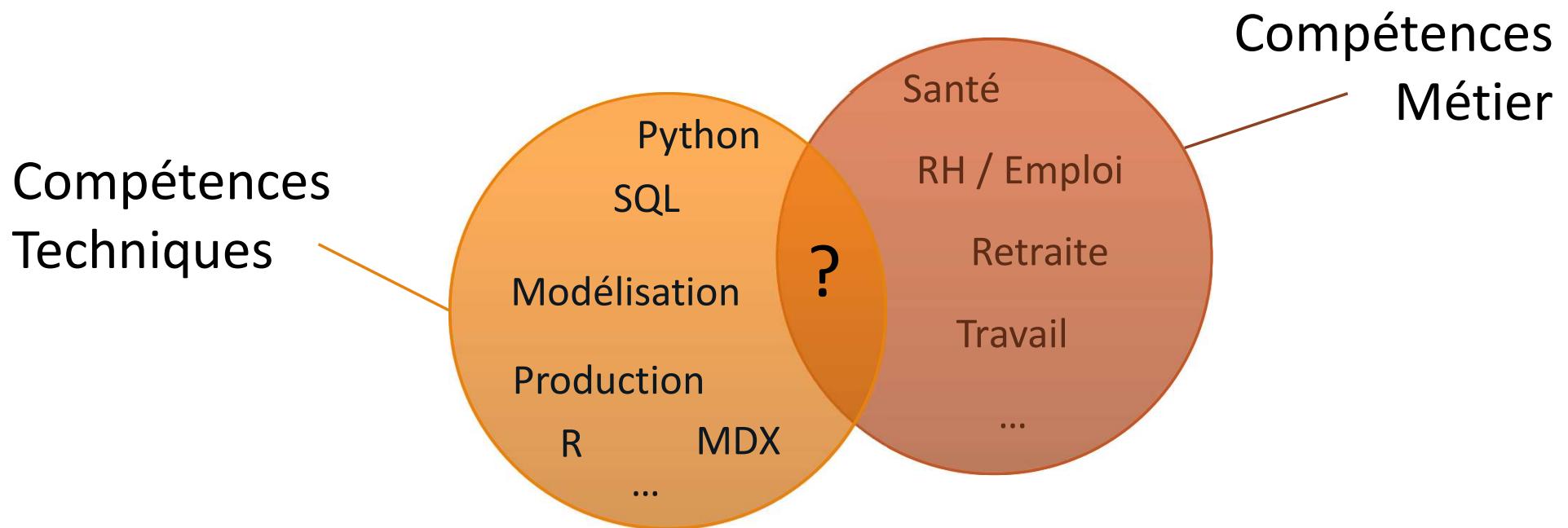
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500
--	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----



Le SID vu par la DSI



Le SID... mélange de compétences



Vers une même finalité

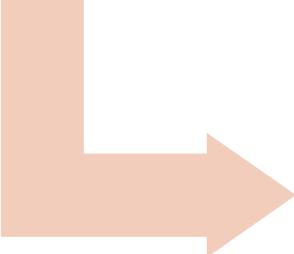
-> la valorisation du travail des équipes pour plus de performance



Comment mettre en place un système décisionnel ?

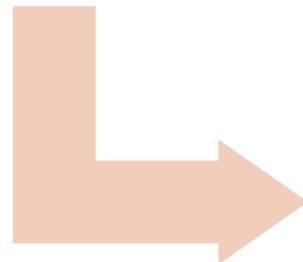
Créer une bonne équipe pluri-disciplinaire !

- Des personnes qui connaissent le métier, capables de définir des indicateurs
- Des personnes qui connaissent bien les systèmes sources
- Des personnes du service informatique pour la mise en œuvre



Travailler sur un cas simple

- Ne créer que quelques indicateurs
- Documenter la définition des indicateurs dans un référentiel consultable par tous
- Mettre l'ensemble de la chaîne décisionnelle en œuvre (du chargement à la restitution)



Généraliser le processus

- À plusieurs processus du même service
- A plusieurs services en gardant une homogénéité dans les outils utilisés



Relations Métiers/informatique

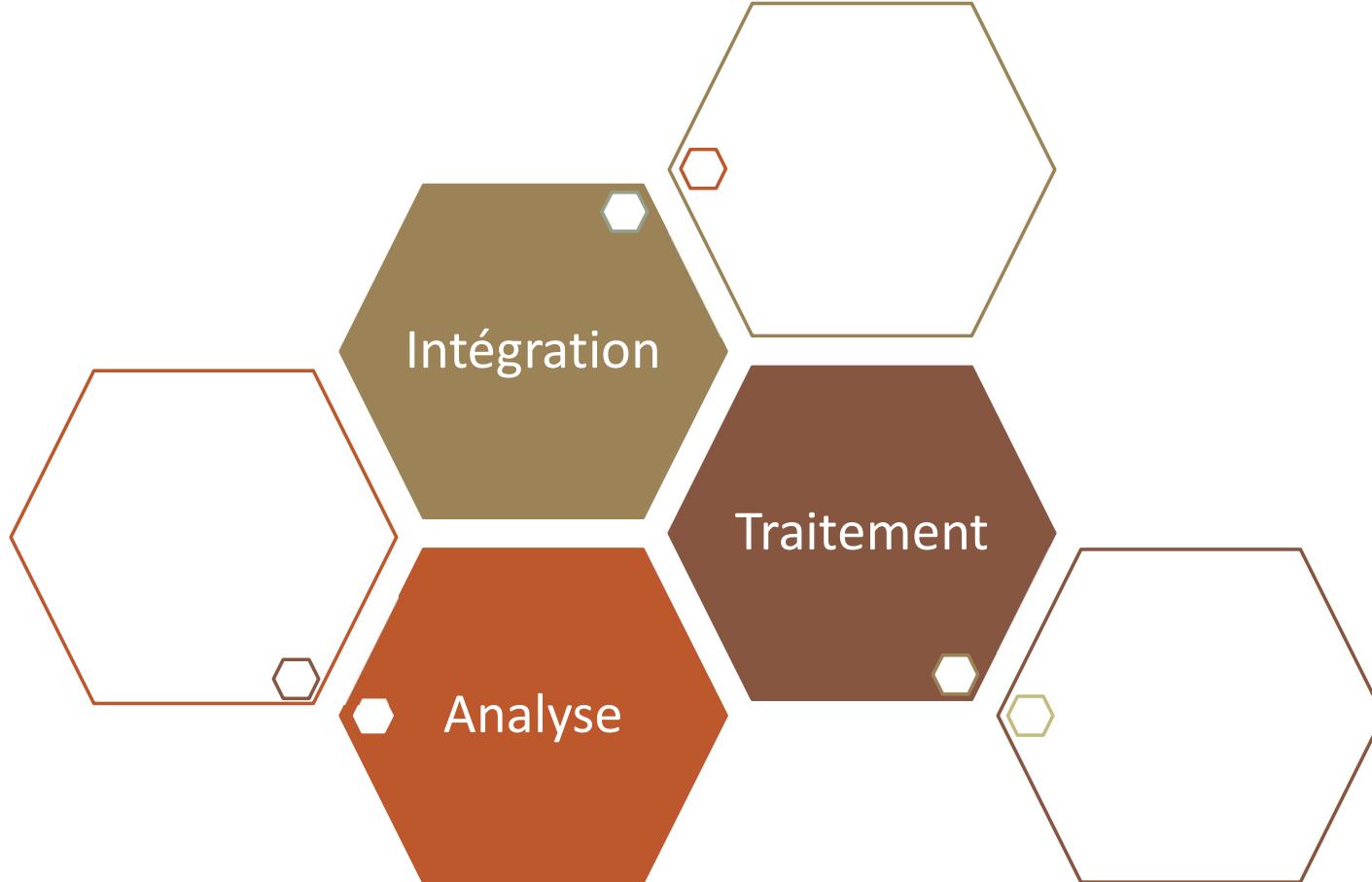
Toujours partir du besoin exprimé par les utilisateurs

**Ne pas se laisser embarquer par la DSI
(service informatique)**

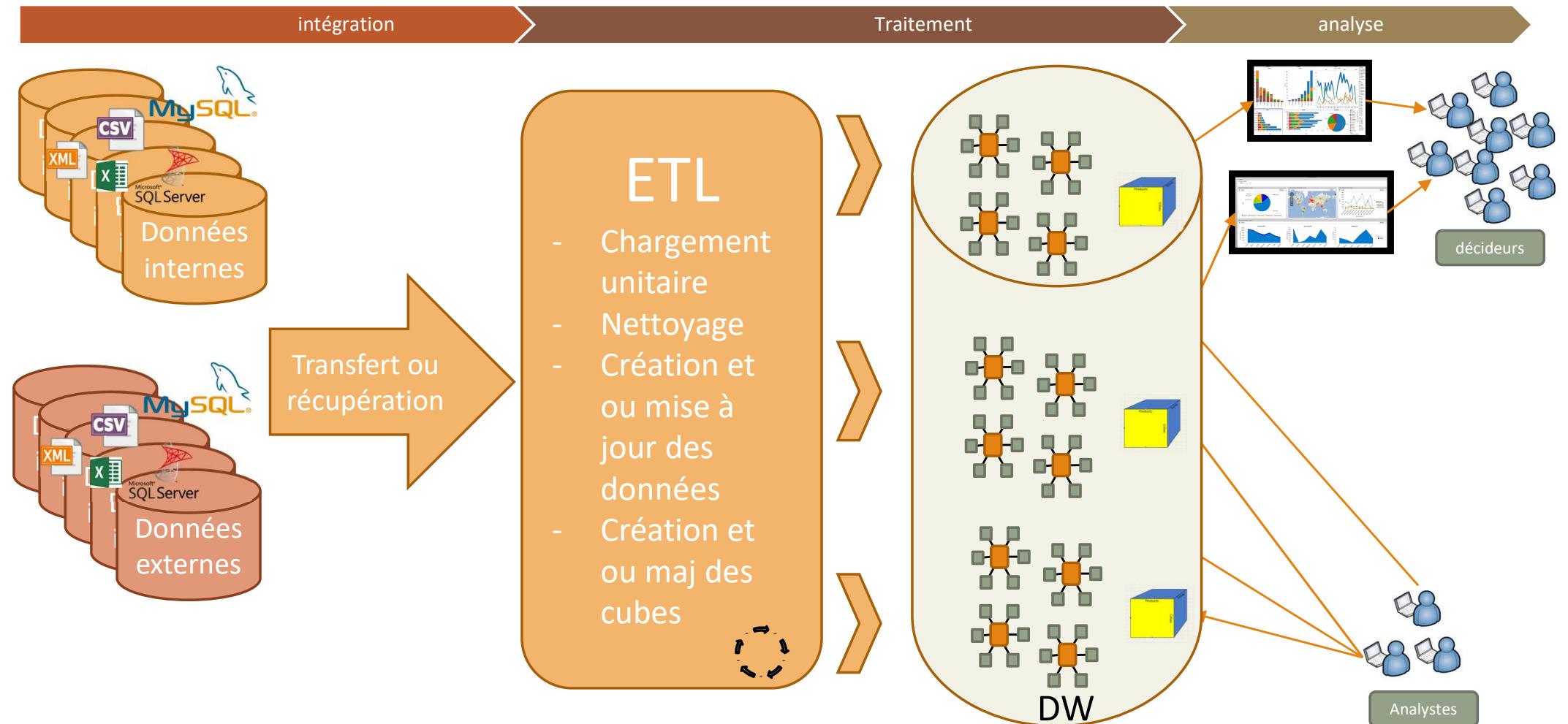
Ce type de projet peut très vite partir sur des aspects techniques, il faut toujours se raccrocher au besoin formulé par le métier.



Les modules d'une plate-forme décisionnelle



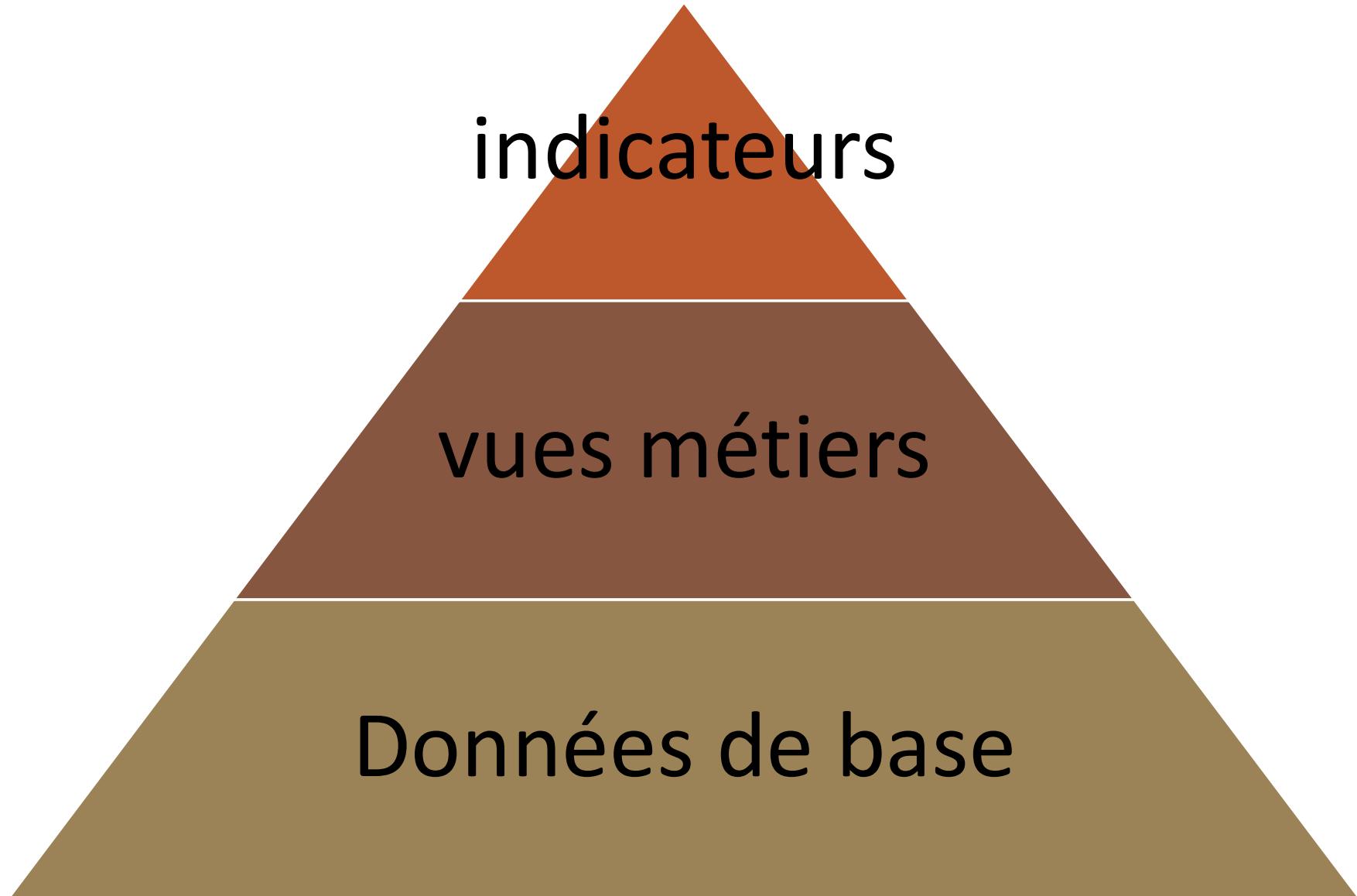
Architecture générale d'une plate-forme décisionnelle



Gouvernance des données (dictionnaire, indicateurs)



Le dictionnaire des données



Le dictionnaire des données

Les données doivent être référencées dans un « DICTIONNAIRE ». Ce dictionnaire doit :

1. Être accessible à tous les utilisateurs de l'entreprise
2. Être compréhensible par tous, et donc contenir des informations fonctionnelles détaillées
3. Contenir l'ensemble des données unitaires, agrégées, indicateurs, avec leurs méthodes de calcul
4. Idéalement : contenir un outil de « datalineage », qui permettra d'avoir graphiquement l'ensemble des opérations réalisées sur la donnée. Du « Système opérationnel » jusqu'à l'indicateur présent dans les tableaux de bord.



Le dictionnaire des données

Exemple de fiche décrivant un « indicateur ».

On peut voir une importance particulière concernant :

- la temporalité
- les axes d'analyse
- la thématique



exemple-fiche-indicateur



fiche indicateur 2

Intitulé court (arborescence)	Réf Prisme		
Antériorité du solde <= 15 jours ouvrés	Axe Prisme 03 - processus		
	N° d'ordre 02		
Intitulé long (titre tableau & information bulle)	Réf autre tableau bord		
Antériorité du solde <= 15 jours ouvrés	TBCAF 11.01.08		
	RNDC 11.01.08		
Type	Périodicité	Format	Unité de l'indicateur
Taux	J Journalier	xx,x	%
Appartenance aux listes figées d'indicateurs			
Intérressement			
Part variable des directeurs			

OBJECTIFS

Fournir un indicateur complémentaire aux indicateurs solde de pièces à traiter en jour et délai de traitement, pour mieux apprécier le niveau de réalisation de l'engagement de service relatif à la rapidité des traitements.
Cet indicateur entre dans le calcul de la part locale de l'intéressement des Caf. Cf. Accord d'intéressement – Indicateurs associés à l'amélioration de la relation de service : « la part de l'antériorité annuelle moyenne du solde mensuel de pièces à traiter supérieure à 15 jours ne dépassant pas 15 % ». Ainsi, au moins 85% des pièces en solde doivent avoir une antériorité inférieure ou égale à 15 jours.

DEFINITION

Antériorité de pièces en solde depuis 15 jours ouvrés à partir de l'année 2011. L'indicateur est mesuré en pourcentage.



Le dictionnaire des données

Pour un indicateur, la fiche devra contenir :

1. Les données de synthèse vues précédemment
2. **L'objectif de l'indicateur**
3. Sa définition
4. Les interprétations et limites
5. La liste des données constitutives de l'indicateur
6. La méthode de calcul
7. Facultatif : ses modes de représentation

Exercice



Le dictionnaire des données

DOS : TABLE DOSSIER CRISTAL

Traçabilité

Description de la table

Vue parente

[INFOCENTRE CRISTAL](#)

Structure

■ INFORMATIONS DOSSIER ALLOCATAIRE

Nom	Libellé
MATRICULE	MATRICULE
LETCLE	LETTRE CLE
TERMINIMATRICULE	TERMINAISON MATRICULE
CODEGES	CODE GESTION
IDCODEGES	IDENTIFIANT CODE GESTION
ORIIMMADOS	ORIGINE IMMATRICULATION
MATRICULEETR	MATRICULE A ETRANGE
SITFAMANT	SITUATION FAMILIALE ACTUELLE
DOSITFAM	DATE DEBUT SITUATION
SITFAM	SITUATION FAMILIALE
DTEMM	DATE EMMENAGEMENT
DTMAJNUMTEL	DATE MISE A JOUR NUMERO TEL
NUMTELDOS	NUMERO TELEPHONE Dossier
AUTORUTINUMTELDOS	AUTORISATION UTILISATION TELEPHONE DOSSIER

NBPIETRAI : NOMBRE PIECES TRAITEES

Traçabilité

Caractéristiques de la donnée

Donnée sémantique	NBPIETRAI : NOMBRE PIECES TRAITEES
Nom physique	NPIET
Nature	QT - QUANTITE
Type de format	Numérique
Donnée signée	NON
Longueur	8
Fils d'ariane	GESCAF - DONNEES DE GESTION CAF > SUIVI DES DOSSIERS / COURRIERS / COURRIELS / PIECES > PIECES TRAITEES > Données pièces traitées > GSSDPTRA . GESCAF - SDP PIECES TRAITEES GESCAF - DONNEES DE GESTION CAF > SUIVI DES DOSSIERS / COURRIERS / COURRIELS / PIECES > DETAIL PRODUCTION : ARRIVEES/STOCK /TRAITEMENT > Données détail prod , arrivées/stock/traitement > GSSDPDET . GESCAF - DETAIL SUIVI DE PRODUCTION GESCAF - DONNEES DE GESTION CAF > SUIVI DES DOSSIERS / COURRIERS / COURRIELS / PIECES > SUIVI DES PIECES PAR NATURE DE PIÈCE > Données suivi des pièces par nature > GSSDPNAT . GESCAF - SDP PIECES PAR NATURE DE PIÈCE

Description

Nombre de pièces (masses et flux) traitées sur la période de référence Les traitements qui font suite à une transmission (TRA) sont exclus, ainsi que les états SDP: VEI, VEV ou NVV dont l'état antérieur a déjà été compté "traité" sur un jour précédent

Données source

Règle d'élaboration

Depuis infocentre SDP, si ($\text{DTETATSDP}=\text{DTJOUR}=DFREP$) et ($\text{DTJOUR}=\text{DTFINTRAIERPIE}$ ou ($\text{DTJOUR}>\text{DTFINTRAIERPIE}$ et $\text{ETATFINTRAIERPIE} <> \text{TRA}$)) et ($\text{ETATSDP}=\text{ADE}, \text{ADI}, \text{APJ}, \text{INS}, \text{REP}, \text{RET}, \text{DET}, \text{ERR}, \text{SFG}, \text{LAV}$ ou LAT ou ($\text{ETATSDP}=\text{ACE}, \text{ACI}, \text{CLO}, \text{CPL}, \text{DEC}, \text{ECI}, \text{INC}, \text{NRC} \dots$)
(...) ou ORD et $\text{DTETATSDP}=\text{DTFINTRAIERPIE}$) ou ($\text{DTETASDPT}=\text{DTJOUR}$ et $\text{ETATSDPANT}=\text{LAV}$ et $\text{ETATSDP}=\text{VEI}, \text{VEV}$ ou NVV)) ou ($\text{DTETATSDP}<\text{DTFINTRAIERPIE}$ et $\text{DTJOUR}=\text{DTFINTRAIERPIE}$ et $\text{ETATFINTRAIERPIE}=\text{NVV}$) alors +=1 ; fin si.

Exemple d'outil permettant d'accéder à la description des données de détail



Une modélisation différente des modèles de production, pourquoi ?

Comme expliqué précédemment, la modélisation des modèles de production n'est pas adaptée, il est nécessaire de :

1. Transformer les données normalisées en modèles « simples »
2. Permettre d'avoir des requêtes performantes
3. Permettre de « naviguer » dans les données en allant de l'information générale, jusque dans le détail
4. Utiliser les techniques de modélisations utilisées par l'ensemble des outils BI
5. Avoir le moins de maintenance possible lors de l'évolution des données source



Une modélisation « multidimensionnelle »

Le « mot » est compliqué, mais c'est en réalité très simple

L'objectif de cette modélisation est de répondre à n'importe quelle question liée à un processus. Pour une question donnée, les éléments suivants apparaissent :

- le contexte : les **dimensions** (quelle année, quel gestionnaire, quelle prestation)
- l'indicateur : la **mesure** ou le **fait** (combien de bénéficiaires, quel montant versé)

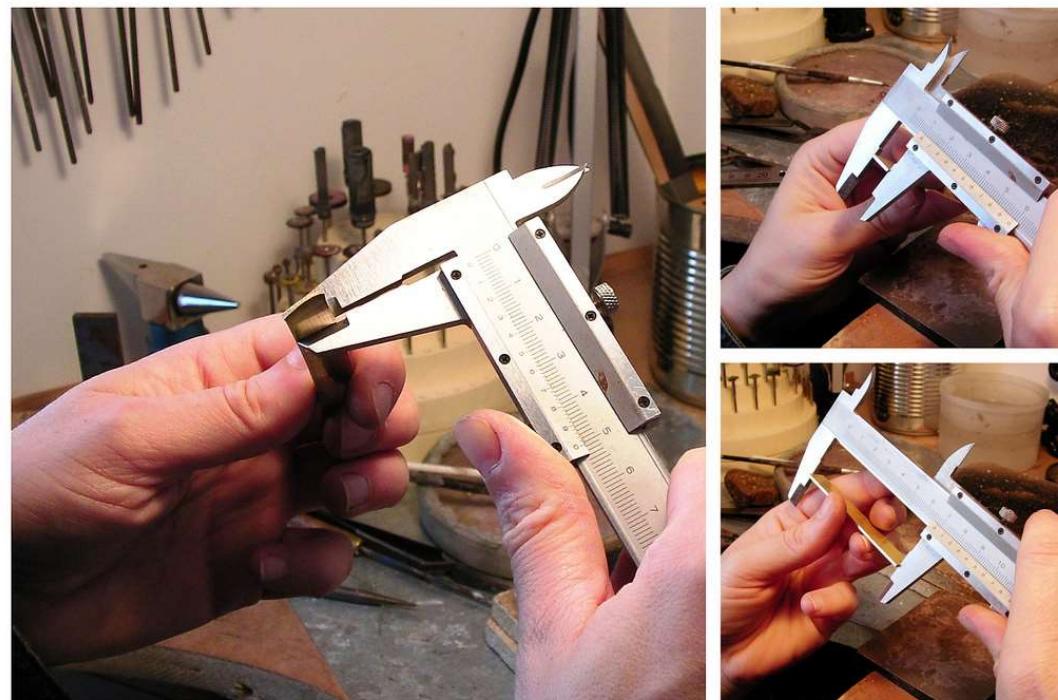
L'objectif de la modélisation est de ne pas recréer des bases dès que l'utilisateur se pose une nouvelle question. Ex :

- Hier je me posais des questions de manière globale
- Aujourd'hui, je souhaite connaître le détail par entité géographique



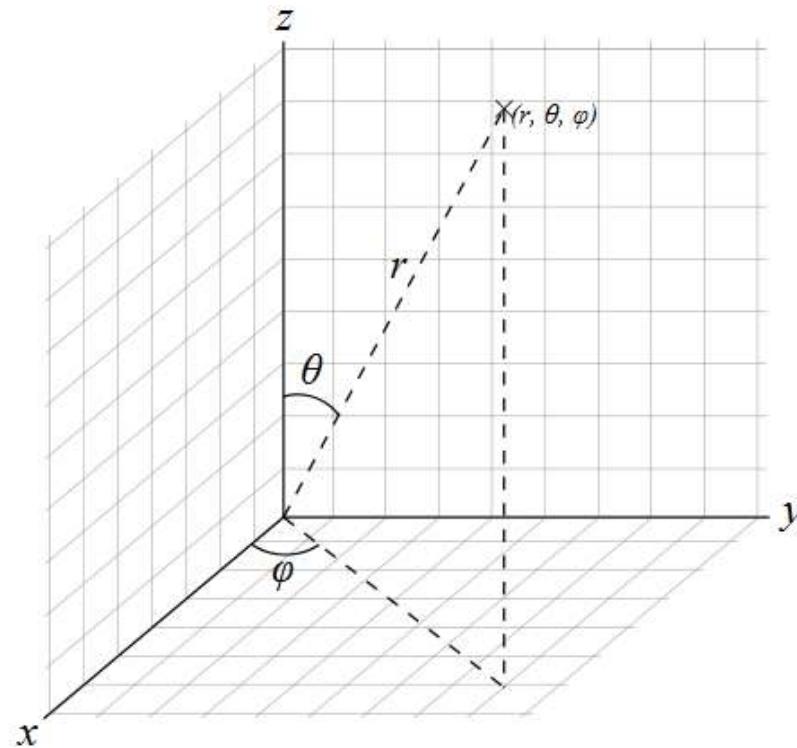
Une modélisation simple, en 3 mots

1 : La mesure



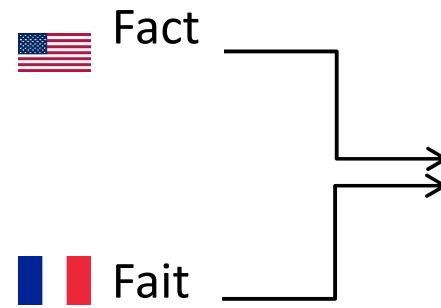
Une modélisation, simple, en 3 mots

2 : Les dimensions



Une modélisation, simple, en 3 mots

3 : Les faits



Factum : Acte, événement

Il s'est passé quelque chose, et on l'a mesuré selon notre référentiel, nos dimensions



On range

Un fait, c'est une ligne, dans une table de faits

Table de Faits

Il s'est passé quelque chose

Il s'est passé autre chose

Il s'est passé quelque chose d'autre



On ordonne

Les dimensions donnent le contexte du fait

Table de Faits		
Quand	Où	
Hier	Ici	Il s'est passé quelque chose
Hier	Là bas	Il s'est passé autre chose
Aujourd'hui	Ici	Il s'est passé quelque chose d'autre



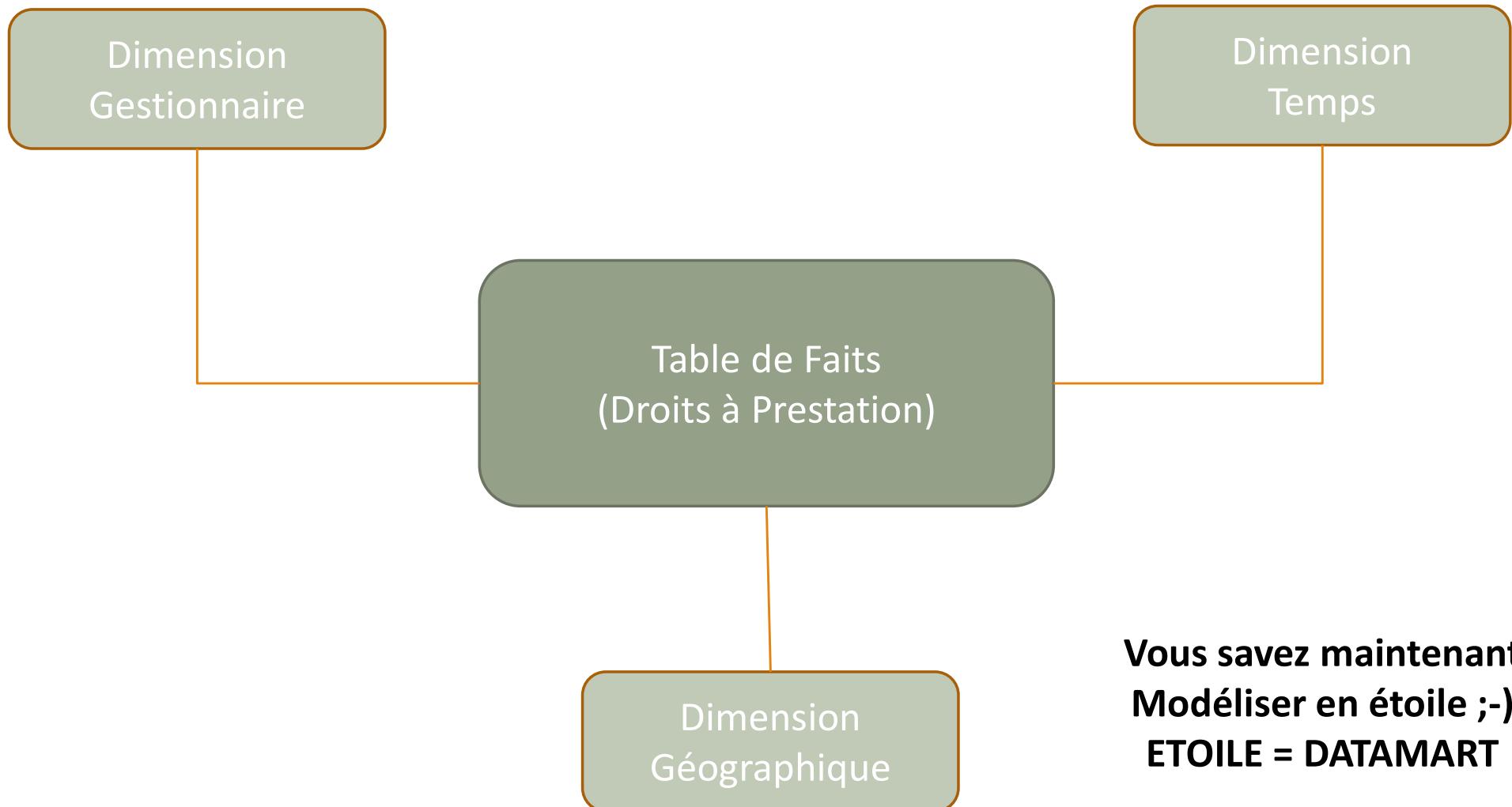
On compte

Les mesures donnent les valeurs numériques du fait

Table de Faits : Crédits alloués aux entreprises par la Banque					
Date du prêt	Client	Type de crédit	Pays	MT alloué (M€)	MT utilisé (M€)
2018/12/01	TOTAL	Exploitation	France	50	20
2017/11/10	TOTAL	Investissement	U.S.A.	100	60
2017/10/10	CARREFOUR	Exportation	Suisse	40	40



Et puis... on assemble !



En SQL

```
SELECT      G.NOM, G.CATEGORIE,B.SEXE,sum(D.MONTANT),sum(D.NB)  
FROM        DROIT_PRESTATION D, GESTIONNAIRES G, TEMPS T  
WHERE       D.TID = T.TID (Jointure)  
AND         D.GID = G.GID (Jointure)  
AND         D.BID = B.BID (Jointure)  
AND         T.ANNEE = 2016 (Sélection)  
AND         B.SEXE = 'F' (Sélection)  
GROUP BY    G.NOM, G.CATEGORIE,B.SEXE
```

Requêtes de jointure en étoile

- Plusieurs jointures
- Suivies par des sélections

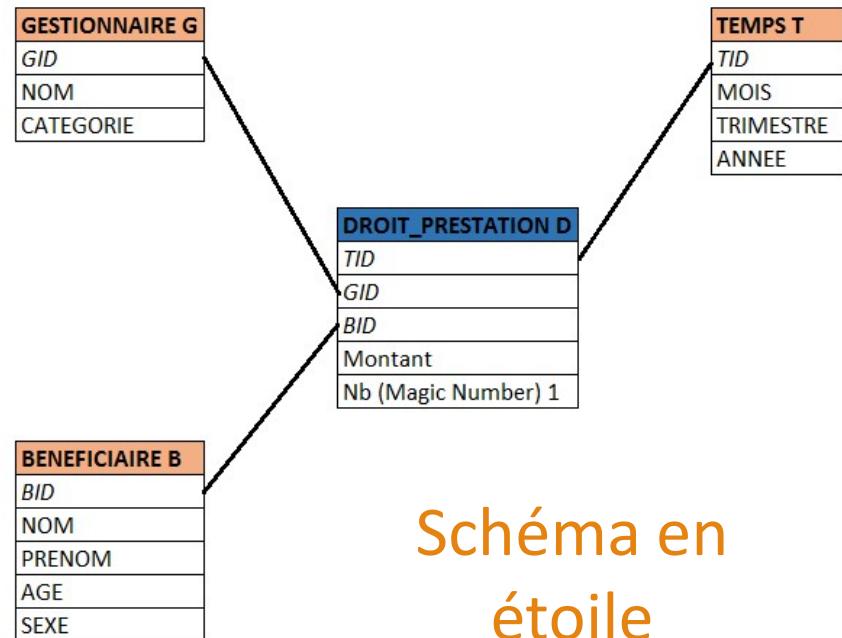


Schéma en
étoile



Le process de modélisation



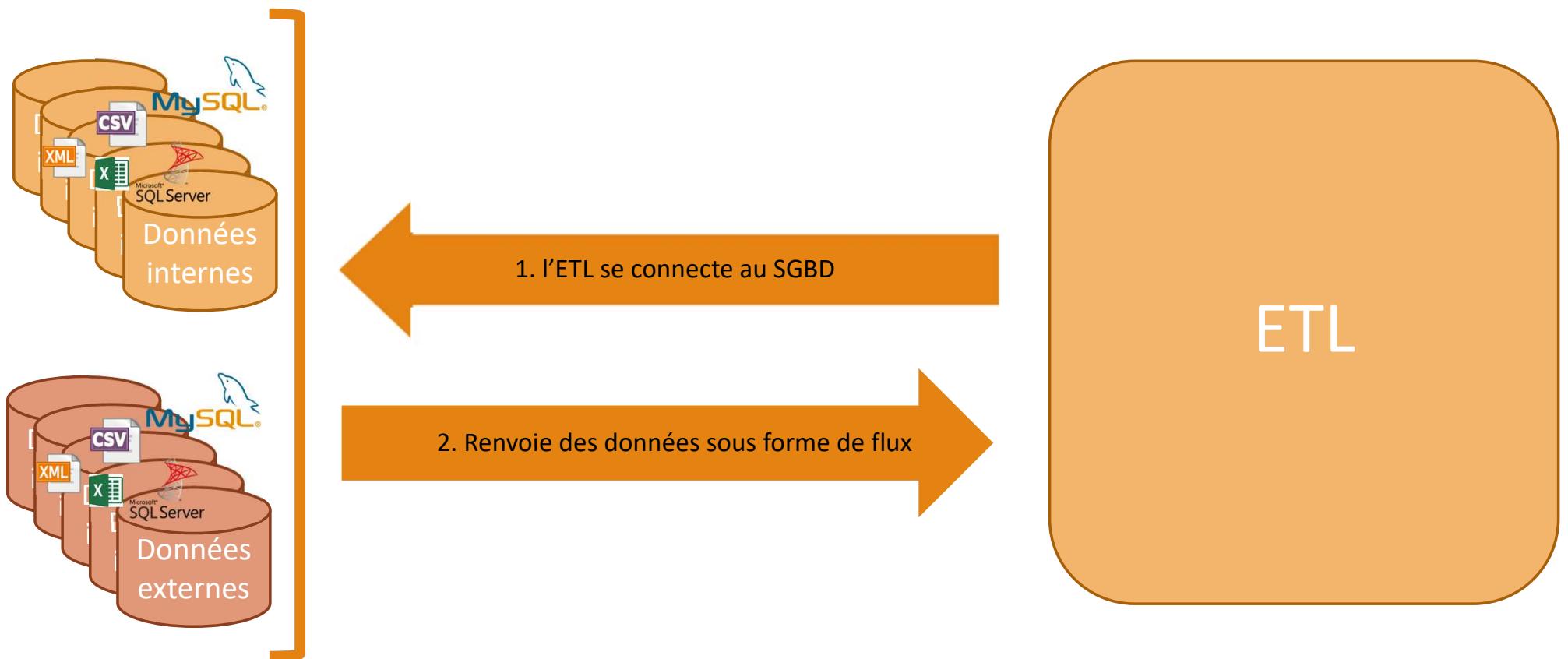
l'acquisition des données (internes / partenaires)

intégration

Traitement

analyse

- **Choix 1** : Vous allez chercher les données dans le système de gestion externe



l'acquisition des données (internes / partenaires)

intégration

Traitement

analyse

- **Choix 1** : Vous allez chercher les données dans le système de gestion externe

Avantages

Les données sont à jour en permanence

Aucune latence dans le processus

L'accès aux nouvelles données est aisé

Inconvénients

Complexité du système de gestion

Habilitations, difficulté d'accès

Diversité des formats de bases à lire

Peut perturber le système de gestion

Pas de séparation fonctionnelle (qui a la RESPONSABILITE de l'extraction ?)



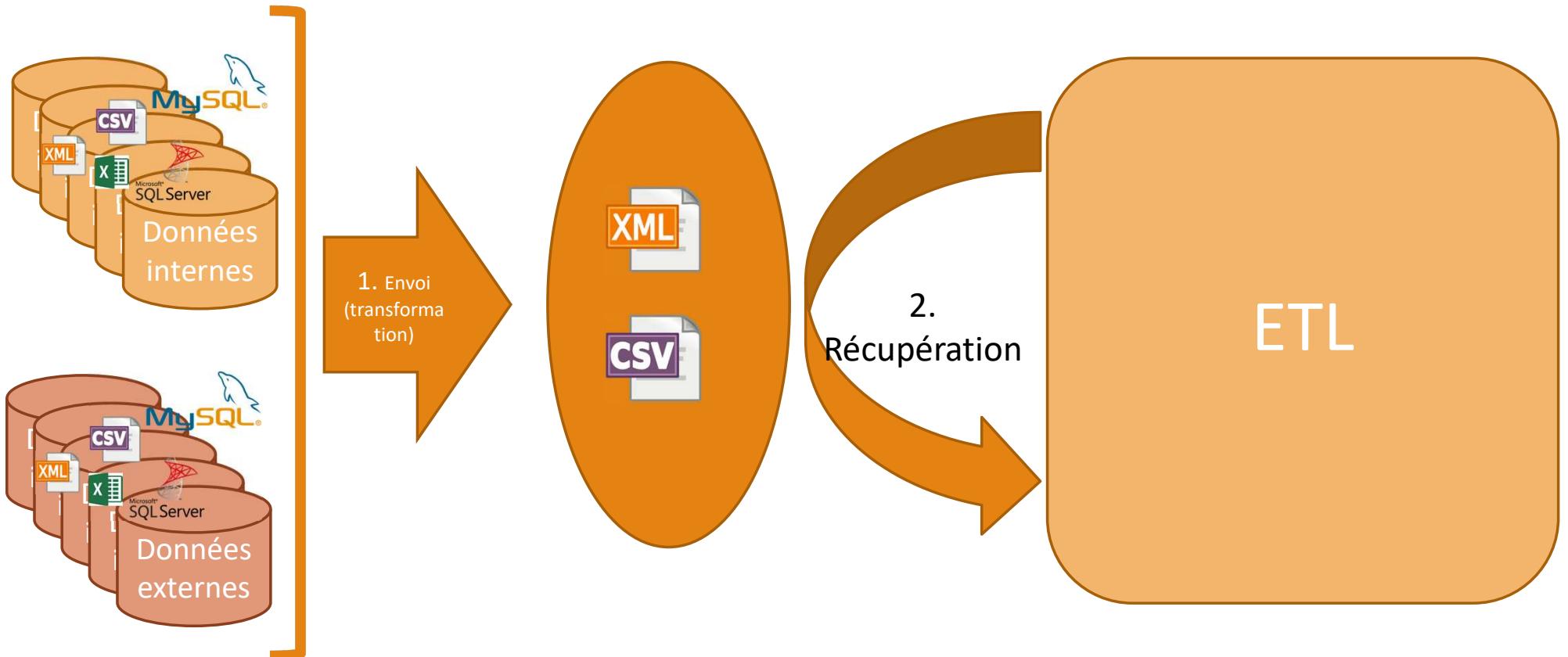
l'acquisition des données (internes / partenaires)

intégration

Traitement

analyse

- **Choix 2** : Le système de gestion externe vous envoie les données (sur une plate-forme d'échange)



l'acquisition des données (internes / partenaires)

intégration

Traitement

analyse

- **Choix 2** : Le système de gestion externe vous envoie les données (sur une plate-forme d'échange)

Avantages

La RESPONSABILITE des acteurs est claire

Un dialogue se construit entre les 2 équipes

Le format des données à transférer peut être varié

Le partenaire gère l'envoi en fonction de son planning de production

La SID ne vient pas perturber le système de gestion

Inconvénients

Nécessite du travail côté « envoyeur »

En cas d'évolution des données, les 2 équipes sont impactées

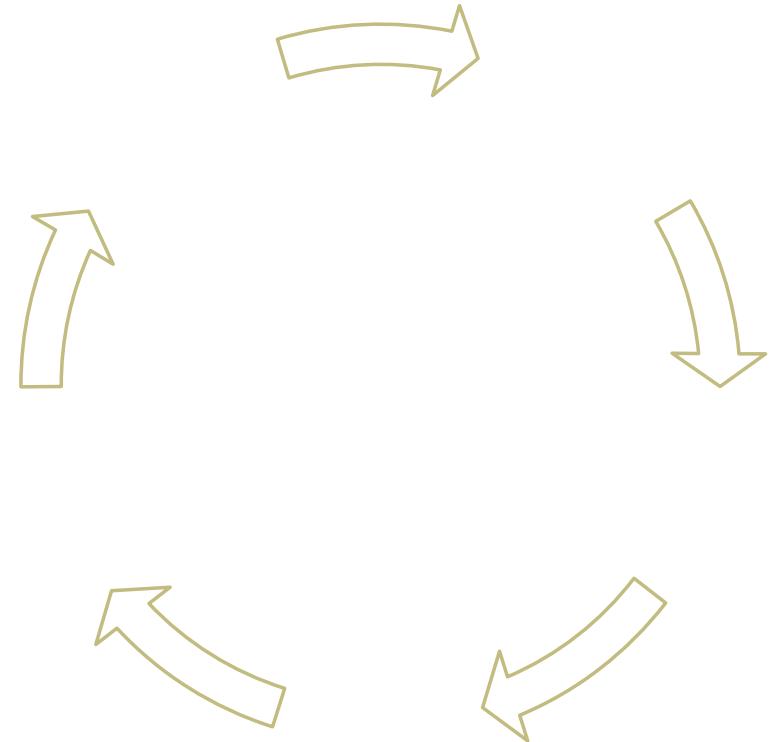
Le SID n'a pas la main sur les extractions, il est en attente de réception



l'outil de transformation de la données (ETL)



- **E**xtract : Extraire la donnée du modèle source
- **T**ransform : Transformation de la donnée
 - **Filtrer**
 - **Trier**
 - **Homogénéiser**
 - **Nettoyer**
 - ...
- **L**oad : Chargement pour l'utilisateur



l'outil de transformation de la données (ETL)



- **De nombreux outils ETL existent**



Outils Open Source relativement complets (version payante souvent nécessaire en production)



l'outil de transformation de la données (ETL)

intégration

Traitement

analyse

Avantages/Inconvénients de l'ETL par rapport à du développement pur (Java / Python)

+

- Homogénéité des traitements
- Force à externaliser les ressources de type bd
- Pas de lien avec la base de données (possible d'en changer « assez » facilement). On en utilise pas ses spécificités
- Intègre Git ou SVN
- Facile d'accès
- Traitements plus pérennes

-

- Peu devenir compliqué à lire sur de gros traitements
- Souvent moins performant
- Certaines transformations peuvent être difficiles à traiter

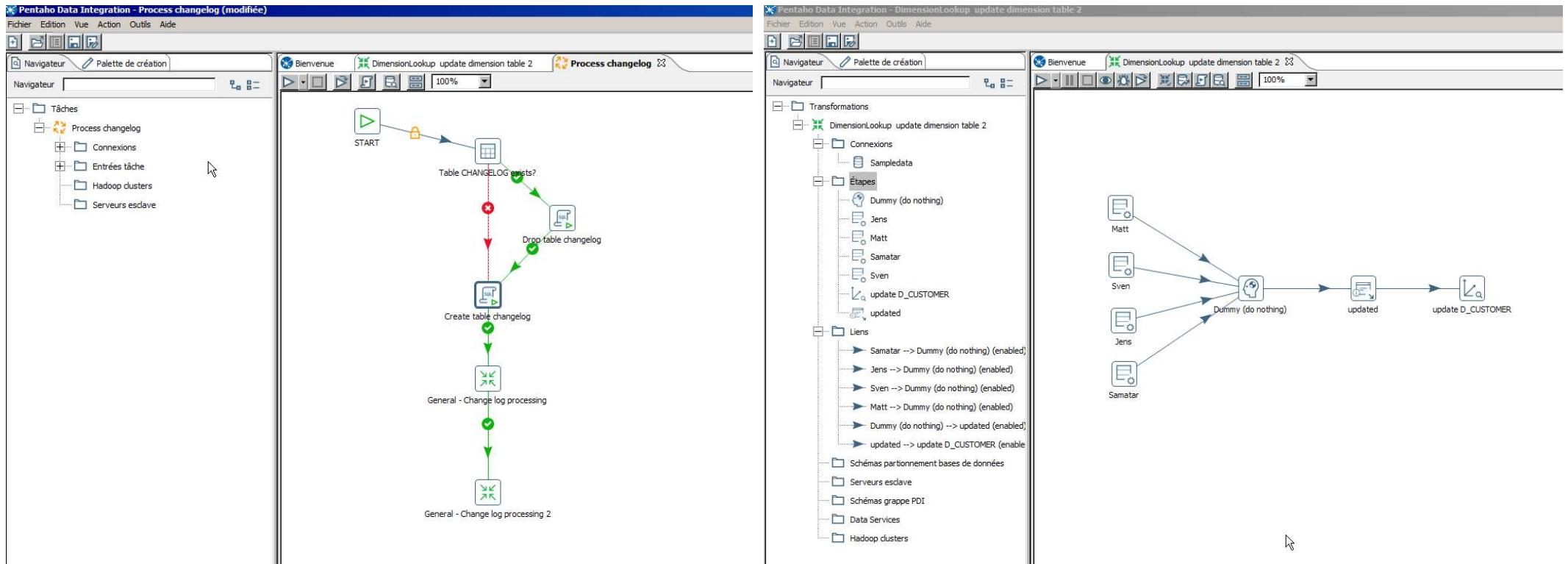


l'outil de transformation de la données (ETL)

intégration

Traitement

analyse



l'outil de transformation de la données (ETL) – via SAS EG

intégration

Traitement

analyse

Exercice :

Utilisation de SAS Enterprise Guide

1. Import de 3 fichiers « plats »

- Nb de bénéficiaires de l'allocation logement (par département)
- Référentiel INSEE des communes associées aux départements
- Nb d'étudiants et population par commune (source INSEE)

2. Créer via un processus ETL la table suivante :

CD_DEPARTEMENT	LB_DEPARTEMENT	Taux d'étudiants dans la population	Taux de bénéficiaires d'allocations logement dans la population française
35	ILLE-ET-VILAINE	10.7%	5.8%
31	HAUTE-GARONNE	10.6%	8.4%
86	VIENNE	10.5%	6.1%
49	MAINE-ET-LOIRE	10.4%	5.4%
54	MEURTHE-ET-MOSELLE	10.2%	5.3%
21	COTE-D'OR	10.0%	5.0%
69	RHONE	9.9%	5.8%
14	CALVADOS	9.9%	4.8%
59	NORD	9.8%	3.8%



l'outil de transformation de la données (ETL) – via SAS EG

intégration

Traitement

analyse

Exercice :

Utilisation de Power BI Desktop

1. Import de 3 fichiers « plats »

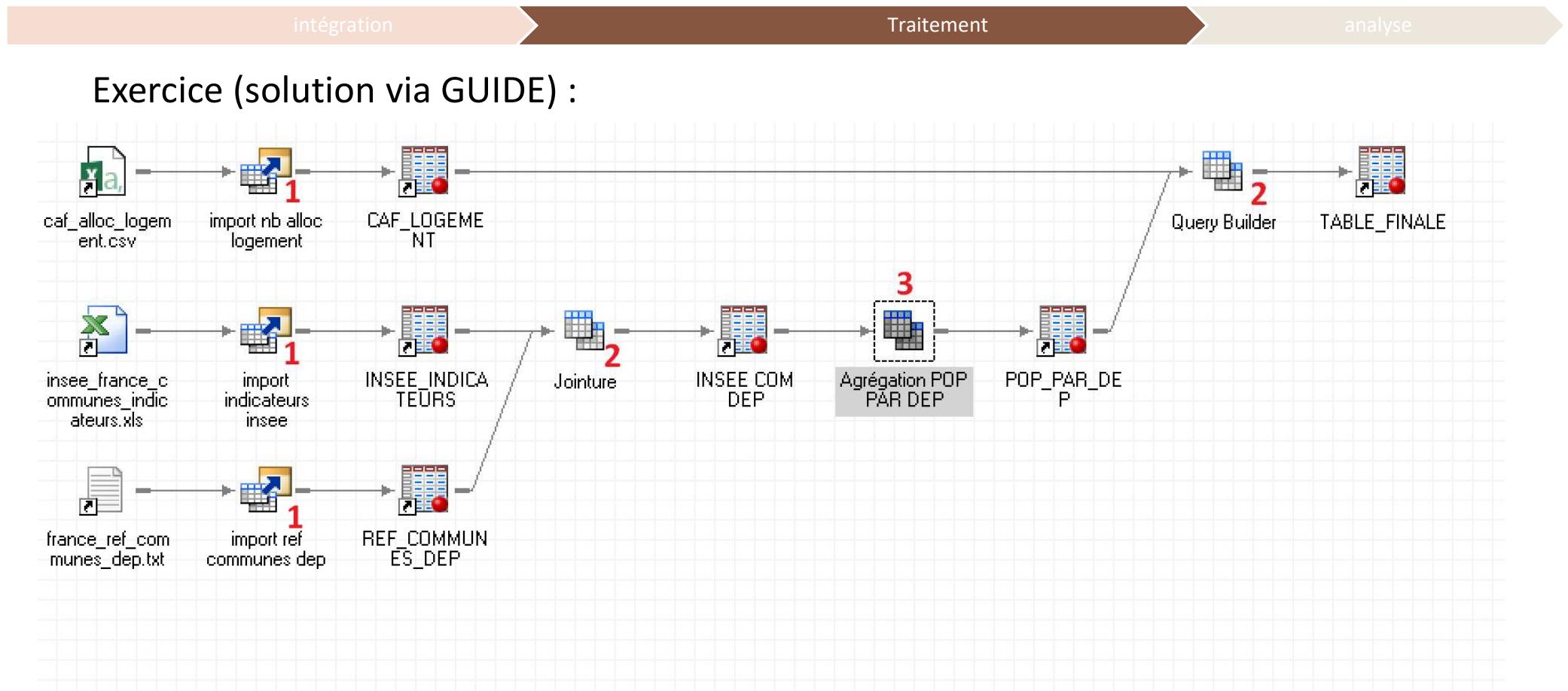
- Nb de bénéficiaires de l'allocation logement (par « code département »)
- Référentiel INSEE des communes associées aux départements
- Nb d'étudiants et population par commune (source INSEE)

2. Crée via un processus ETL, la table suivante :

	CD_DEPARTEMENT	france_ref_commun..._LB_DEPARTE...	% Taux d'étudiants dans la population	% Taux de bénéficiaires d'alloc. logement dans la pop française
1	35	ILLE-ET-VILAINE	10,68 %	4,88 %
2	31	HAUTE-GARONNE	10,59 %	7,59 %
3	86	VIENNE	10,51 %	5,32 %
4	49	MAINE-ET-LOIRE	10,39 %	4,76 %
5	54	MEURTHE-ET-MOSSELLE	10,17 %	4,57 %
6	21	COTE-D'OR	10,05 %	4,29 %
7	69	RHONE	9,88 %	5,00 %
8	14	CALVADOS	9,87 %	4,27 %
9	59	NORD	9,82 %	3,43 %
10	51	MARNE	9,67 %	3,28 %
11	34	HERAULT	9,66 %	7,60 %
12	25	DOUBS	9,65 %	3,43 %
13	44	LOIRE-ATLANTIQUE	9,55 %	4,31 %
14	38	ISERE	9,47 %	3,19 %
15	95	VAL-D'OISE	9,36 %	1,37 %
16	37	INDRE-ET-LOIRE	9,32 %	4,18 %
17	63	PUY-DE-DOME	9,31 %	4,81 %
18	33	GIRONDRE	9,21 %	5,65 %
19	75	PARIS	9,11 %	5,49 %
20	78	YVELINES	9,07 %	1,22 %
21	80	SOMME	9,07 %	3,99 %
22	76	SEINE-MARITIME	8,92 %	3,47 %



l'outil de transformation de la données (ETL) – via SAS EG



1. Utiliser la tâche : File / Import Data
2. Utiliser la tâche : Task / Data / Query builder
3. Utiliser la tâche : Task / Data / Query Builder avec un attribut "SUM" pour les var à sommer



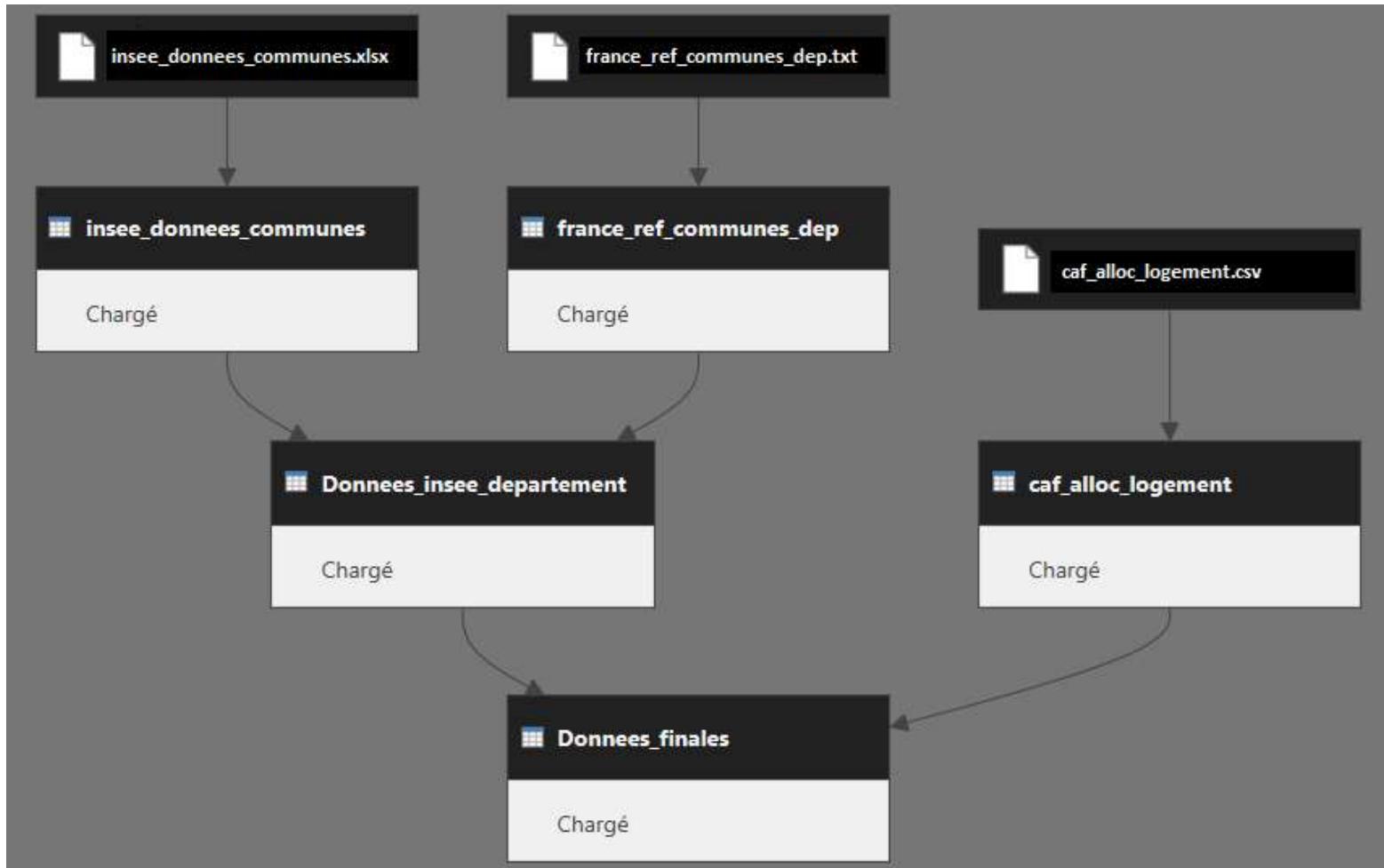
l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice :



l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 1 : import des données communales, transformation en caractère du code_commune (qui est sur 5 positions)

Requêtes [2]

	A ^B _C CD_COMMUNE	A ^B _C CD_DEPARTEMENT	1 ² 3 Population	1 ² 3 Nb_Etudiants
1	01001	01	725	51
2	01002	01	167	5
3	01004	01	11432	904
4	01005	01	1407	97
5	01006	01	86	2
6	01007	01	2144	173
7	01008	01	586	47
8	01009	01	275	10

Requêtes [2]

	A ^B _C CD_COMMUNE	A ^B _C LB_DEPARTEMENT
1	01001	AIN
2	01002	AIN
3	01004	AIN
4	01005	AIN
5	01006	AIN
6	01007	AIN



l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 2 : jointure des 2 tables pour récupérer le libellé du département

Sans titre - Éditeur Power Query

Fichier Accueil Transformer Ajouter une colonne Affichage Outils Aide

Fermer & appliquer Nouvelle source récentes Entrer des données Paramètres de la source de données Gérer les paramètres Actualiser l'aperçu Gérer Requête Choisir les colonnes Supprimer les colonnes Conserver les lignes Supprimer les lignes Réduire les lignes Trier Fractionner la colonne Regrouper par Utiliser la première ligne Remplacer les valeurs Fusionner des requêtes

Requêtes [2]

	CD_COMMUNE	LB_DEPARTEMENT
1	01001	AIN
2	01002	AIN

Fusionnez cette requête avec une autre requête dans ce fichier pour créer une nouvelle requête.

Fusionner

Sélectionnez des tables et les colonnes correspondantes pour créer une table fusionnée.

insee_donnees_communes

CD_COMMUNE	CD_DEPARTEMENT	Population	Nb_Etudiants
01001	01	725	51
01002	01	167	5
01004	01	11432	904
01005	01	1407	97
01006	01	86	2

france_ref_comunes_dep

CD_COMMUNE	LB_DEPARTEMENT
01001	AIN
01002	AIN
01004	AIN
01005	AIN
01006	AIN

Type de jointure

Interne (seules les lignes correspondantes)

Utiliser la correspondance approximative pour effectuer la fusion

Options de correspondance approximative

La sélection correspond à 35380 des 36677 lignes de la première table et...

Requêtes [3]

CD_COMMUNE	CD_DEPARTEMENT	Population	Nb_Etudiants	france_ref_comunes_dep
1	01	725	51	
2	01	167	5	
3	01004	11432	904	
4	01005	1407	97	
5	01006	86	2	
6				
7				
8				
9				
10				
11				
12				
13				

Rechercher les colonnes à développer

Développer Agréger

(Sélectionner toutes les colonnes) CD_COMMUNE LB_DEPARTEMENT

Utiliser le nom de la colonne d'origine comme préfixe

OK Annuler

V. F. uni

ETL : CHARGEMENT DES DONNEES

52

l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 3 : calculer la population et le nb d'étudiants par département

The screenshot shows the 'Regrouper par' (Group By) dialog box in PowerBI. On the left, there's a sidebar with icons for 'Fusionner' (Merge), 'Typ', 'Regrouper', and 'Ajouter Regroupement'. The main area has a title 'Regrouper par' and a subtitle 'Spécifiez les colonnes de regroupement et une ou plusieurs sorties.' Below this, there are two radio buttons: 'De base' (Basic) and 'Avancé' (Advanced), with 'Avancé' selected. A dropdown menu 'CD_DEPARTEMENT' is open. To the right of the dropdown are two rows of settings: 'france_ref_commun... dep.LB_DEPARTE...' and 'Ajouter un regroupement'. Below these are sections for 'Nouveau nom de colonne' (New column name) and 'Opération' (Operation). For 'Population', the operation is 'Somme' (Sum) and the new column name is 'Population'. For 'Nb_Etudiants', the operation is 'Somme' and the new column name is 'Nb_Etudiants'. At the bottom right are 'OK' and 'Annuler' buttons.

	CD_DEPARTEMENT	france_ref_commun... dep.LB_DEPARTE...	1.2 Population	1.2 Nb_Etudiants
1	01	AIN	513326	37176
2	02	AISNE	534024	38861
3	03	ALLIER	343966	21383
4	04	ALPES-DE-HAUTE-PROVENCE	139396	8118
5	05	HAUTES-ALPES	119571	7621
6	06	ALPES-MARITIMES	1011866	71060
7	07	ARDÈCHE	286160	17922
8	08	ARDENNES	288082	20390
9	09	ARIEGE	137321	7785
10	10	AUDE	292166	21355
11	11	AUDE	309722	18609
12	12	AVEYRON	257653	15947
13	13	BOUCHES-DU-RHÔNE	1835407	160284
14	14	CORSE	505001	58817

Renommer la table « Fusionner1 » en
« Donnees_insee_departement »



l'outil de transformation de la données (ETL) – via PowerBI



Exercice (solution) :

Etape 4 : récupérer le nb de bénéficiaires de l'allocation logement en important le fichier CSV (récupérer sept_18 uniquement, et renommer en « nb_benef_logement »)

The screenshot shows the PowerBI desktop application. On the left, the 'Requêtes [4]' pane lists four queries: 'insee_donnees_communnes', 'france_ref_communes_dep', 'Donnees_insee_departement', and 'caf_alloc_logement'. The fourth query is currently selected and highlighted with a yellow bar. On the right, a preview pane displays a table with two columns: 'cd_departement' and 'nb_benef_logement'. The data is as follows:

	cd_departement	nb_benef_logement
1	01	7588
2	02	12249
3	03	12880
4	04	6041
5	05	4092
6	06	40469



l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 5 : Faire une jointure entre « Donnees_insee_departement » et « caf_alloc_logement », en réutilisant « Fusionner des requêtes comme nouvelles »

Fusionner

Sélectionnez des tables et les colonnes correspondantes pour créer une table fusionnée.

Donnees_insee_departement

CD_DEPARTEMENT	france_ref_comunes_dep.LB_DEPARTEMENT	Population	Nb_Etudiants
01	AIN	513326	37176
02	AISNE	534024	38861
03	ALLIER	343966	21383
04	ALPES-DE-HAUTE-PROVENCE	139396	8118
05	HAUTES-ALPES	119571	7621

caf_alloc_logement

cd_departement	nb_benef_logement
01	7588
02	12249
03	12880
04	6041
05	4092

Type de jointure

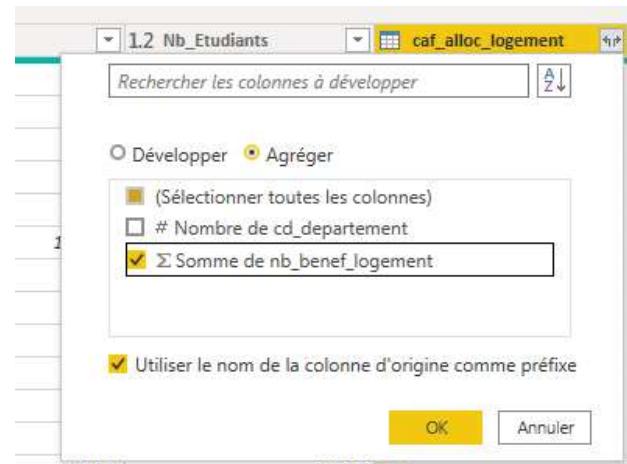
Interne (seules les lignes correspondantes)

Utiliser la correspondance approximative pour effectuer la fusion

↳ Options de correspondance approximative

La sélection correspond à 96 des 100 lignes de la première table et à 96 d...

OK Annuler



Renommer la table « Fusionner1 » en « Donnees_finales»



l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 6 : Calculer les taux demandés dans l'exercice

Requêtes [5]

	CD_DEPARTEMENT	A ^B _C france_ref_communes_dep.LB_DEPARTE...	1.2 Population	1.2 Nb_Etudiants	A ^B _C Somme de caf_alloc_logement.nb_benef_logem...
1	01	AIN	513326	37176	7588
2	02	AISNE	534024	38861	12249
3	03	ALLIER	343966	21383	12880
4	04	ALPES-DE-HAUTE-PROVENCE	139396	8118	6041
5	05	HAUTES-ALPES	119571	7621	4092
6	06	ALPES-MARITIMES	1011866	71060	40469
7	07	ARDECHE	286160	17922	8962
8	08	ARDENNES	288082	20390	6505
9	09	ARIEGE	137321	7785	6415
10	10	AUBE	292166	21355	8932
11	11	AUDE	309722	18609	15505
12	12	AVEYRON	257653	15947	9029

Sans titre - Éditeur Power Query

Fichier Accueil Transformer Ajouter une colonne Af

Colonne à partir d'exemples Colonne personnalisée Appeler une fonction personnalisée Colonne cor
Colonne d'ir Duplication

Nouveau nom de colonne Taux d'étudiants dans la population

Formule de colonne personnalisée ⓘ = [Nb_Etudiants]/[Population]

Colonne personnalisée

Ajoutez une colonne calculée à partir des autres colonnes.

Nouveau nom de colonne Taux de bénéficiaires d'alloc. logement dans la pop française

Formule de colonne personnalisée ⓘ = [Somme de caf_alloc_logement.nb_benef_logement]/[Population]

	CD_DEPARTEMENT
1	01
2	02
3	03
4	04
5	05
6	06
7	07

Colonne personnalisée

Ajoutez une colonne calculée à partir des autres colonnes.

Nouveau nom de colonne Taux de bénéficiaires d'alloc. logement dans la pop française

Formule de colonne personnalisée ⓘ = [Somme de caf_alloc_logement.nb_benef_logement]/[Population]

l'outil de transformation de la données (ETL) – via PowerBI

intégration

Traitement

analyse

Exercice (solution) :

Etape 7 : Formater les pourcentages, sélectionner les bonnes colonnes, trier le taux d'étudiants pour faire apparaître l'Ille et Vilaine !

	A ^B C CD_DEPARTEMENT	A ^B C france_ref_communes_dep.LB_DEPARTE...	% Taux d'étudiants dans la population	% Taux de bénéficiaires d'alloc. logement dans la pop française
1	35	ILLE-ET-VILAINE	10,68 %	4,88 %
2	31	HAUTE-GARONNE	10,59 %	7,59 %
3	86	VIENNE	10,51 %	5,32 %
4	49	MAINE-ET-LOIRE	10,39 %	4,76 %
5	54	MEURTHE-ET-MOSELLE	10,17 %	4,57 %
6	21	COTE-D'OR	10,05 %	4,29 %
7	69	RHONE	9,88 %	5,00 %
8	14	CALVADOS	9,87 %	4,27 %
9	59	NORD	9,82 %	3,43 %
10	51	MARNE	9,67 %	3,28 %
11	34	HERAULT	9,66 %	7,60 %
12	25	DOUBS	9,65 %	3,43 %
13	44	LOIRE-ATLANTIQUE	9,55 %	4,31 %
14	38	ISERE	9,47 %	3,19 %
15	95	VAL-D'OISE	9,36 %	1,37 %
16	37	INDRE-ET-LOIRE	9,32 %	4,18 %
17	63	PUY-DE-DOME	9,31 %	4,81 %



la base de données, les fonctionnalités attendues

intégration

Traitement

analyse

- **M**odélisation classique : modèle relationnel 3NF
 - Tables, attributs, tuples, vues
 - Normalisation (aucune redondance)
 - Requêtes « simples »
- **L**e Temps
 - Représentation du passé et du futur. Une difficulté pour le modèle 3NF

Comment analyser, prédire, piloter !!



la base de données, les fonctionnalités attendues

intégration

Traitement

analyse

- **Quelle est l'évolution du nb d'étudiants du M2 « Prévision, Prédiction Économiques » qui travaillent dans la finance sur les 10 dernières années ?**
- **Comment va évoluer le CA de carrefour par catégorie de produits dans les 6 prochains mois**

Le temps ! La base de l'analyse décisionnelle



la base de données, les fonctionnalités attendues

intégration

Traitement

analyse

- **P**our répondre à ces problématiques une « modélisation » spécifique est nécessaire dans la B.D., mais sauf si le volume est important (> 10 Go), une base de données « classique » sera suffisante
- **A**grégation des données performante
- **E**ncaisser de fortes volumétries



Cubes... Quesako

Schéma en étoile

Ensemble de tables, avec des relations, permettant de récupérer l'information le plus rapidement possible



Peu volumineux

Une requête est exécutée pour obtenir tous les résultats, pendant la consultation

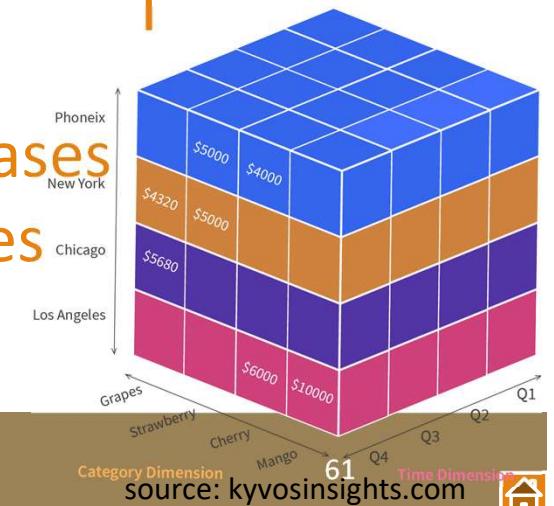
Cube OLAP

Objet physique créé par un éditeur de logiciel, permettant de pré-générer les résultats de toutes combinaisons un schéma en étoile (il se base donc sur un schéma en étoile au départ)

L'accès à un résultat est instantané

Temps de traitement pour constituer le cube si bcp d'indicateurs ou dimensions

Les Cube OLAP sont en « perte de vitesse » face aux bases de données « in « memory » qui permettent d'avoir des résultats de manière quasi instantanées



la base de données, les fonctionnalités attendues

intégration

Traitement

analyse

Figure 1: Magic Quadrant for Cloud Database Management Systems



20/09/2023 11:01

Gartner Report

Gartner: Licensed for Distribution

Magic Quadrant for Cloud Database Management Systems

Published 13 December 2022 - ID G00763557 - 71 min read

By Henry Cook, Merv Adrien, and 2 more

The market is converging on a set of advanced capabilities resulting in a complex landscape ready to launch into the next wave of disruption. This Magic Quadrant will help data and analytics leaders make the right cloud DBMS choices in a complex and fast-evolving market.

Strategic Planning Assumptions

By 2025, 90% of new data and analytics deployments will be through an established data ecosystem, causing consolidation across the data and analytics market.

Through 2024, organizations that adopt aggressive metadata analysis across their complete data management environment will decrease time to delivery of new data assets to users by as much as 70%.

Market Definition/Description

Gartner defines the Cloud DBMS market as follows:

Core capabilities are that vendors fully supply provider-managed public or private cloud software systems that manage data on cloud storage. Data is stored in a cloud storage tier. Optionally, they may cater to multiple data models and data types – relational, nonrelational (document, key value, wide column, graph), geospatial, time series and others.

These DBMSs reflect optimization strategies designed to support transactions and/or analytical processing for more than one of the following use cases:

- OLTP transactions
- Lightweight transactions
- Augmented transactions
- Stream event processing
- Traditional data warehouse

<https://www.gartner.com/docspages?id=12C000AS&lnk=22129&sl=6>



les outils permettant de restituer les données, quel outil pour quel public ?

intégration

Traitement

analyse

Tableau de bord

- ✓ Cible visée : Directeur, Chef de service
- ✓ Objectif : Obtenir un suivi d'activité, piloter une activité, être alerté en cas de problème



les outils permettant de restituer les données, quel outil pour quel public ?

intégration

Traitement

analyse

Reporting / Listing

- ✓ Cible visée : Chef de service, Analyste, Technicien
- ✓ Objectif : Diffuser de manière automatique des rapports sous forme « de listes » (peu visuels en général) permettant d'obtenir de l'information assez détaillée, qui servira à réaliser des actions via les services spécialisés.
- ✓ Ces listings sont récurrents (mensuels, annuels,...) et « statiques », aucune possibilité de navigation.
- ✓ Conçus pour « l'impression »



les outils permettant de restituer les données, quel outil pour quel public ?

intégration

Traitement

analyse

Requêtage

- ✓ Cible visée : Analyste
- ✓ Objectif : Le tableau de bord ou le reporting automatique n'a pas permis de comprendre un problème, il faut « descendre » dans le détail des données pour comprendre... le requêtage sur les données devient indispensable



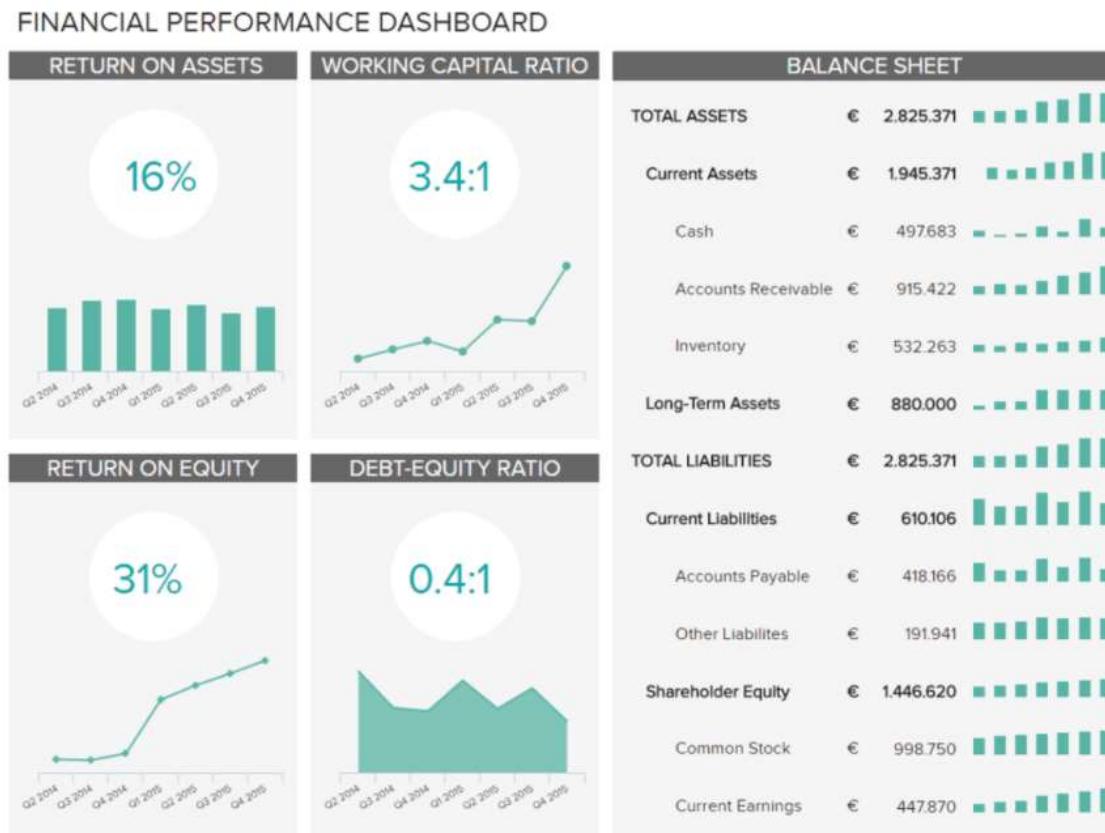
A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage



A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage

Referring MD Report										
Charges by Specialty										
Specialty	Group	Physician	January	February	March	April	May	June	Total	
GI	Practice A	Physician A	18,000	18,000	16,000	19,000	10,250	5,250	86,500	
		Physician B	18,500	15,555	18,000	18,000	22,000	24,000	116,055	
		Physician C	16,500	17,555	18,600	18,000	19,500	20,000	110,155	
		Physician D	15,500	16,555	17,000	17,500	19,220	18,500	104,275	
			68,500	67,665	69,600	72,500	70,970	67,750	416,985	
% of Total Charges by Specialty			53%	52%	53%	54%	53%	52%		
OBGYN	Practice B	Physician A	1,000	1,650	1,850	1,900	2,200	2,450	11,050	
		Physician B	1,500	1,750	2,000	2,000	2,300	2,500	12,050	
		Physician C	2,200	2,400	2,600	2,600	2,700	2,800	15,300	
		Physician D	3,600	3,200	2,900	2,550	3,200	3,400	18,850	
		Physician E	2,600	2,600	3,200	2,200	2,200	3,000	15,800	
		Physician F	1,850	2,200	1,650	1,000	1,250	1,550	9,500	
		Physician G	4,200	3,600	3,800	4,400	3,600	3,200	22,800	
			16,950	17,400	18,000	16,650	17,450	18,900	105,350	
% of Total Charges by Specialty			13%	13%	14%	12%	13%	14%		
GU	Practice C	Physician A	16,500	18,000	22,000	23,550	19,500	22,500	122,050	
		Physician B	19,000	18,000	14,000	12,000	11,000	10,500	84,500	
		Physician C	6,500	6,000	4,500	3,800	5,500	4,800	31,100	
		Physician D	2,600	3,200	4,400	2,500	3,000	2,200	17,900	
			44,600	44,600	44,600	44,600	44,600	44,600	255,550	
% of Total Charges by Specialty			34%	34%	34%	33%	34%	34%		
Total			\$ 130,050	\$ 129,665	\$ 132,200	\$ 133,750	\$ 133,020	\$ 131,250	\$ 777,885	



A vous de jouer

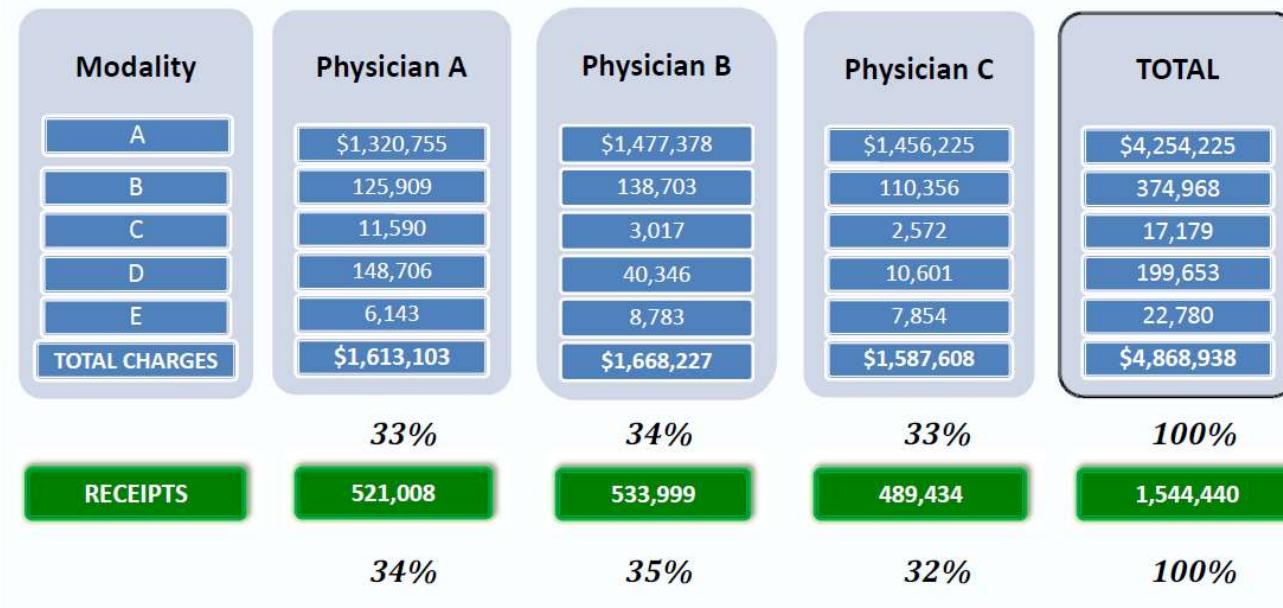
intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage

Billing Summary by Physician {Processing: Jan-June}



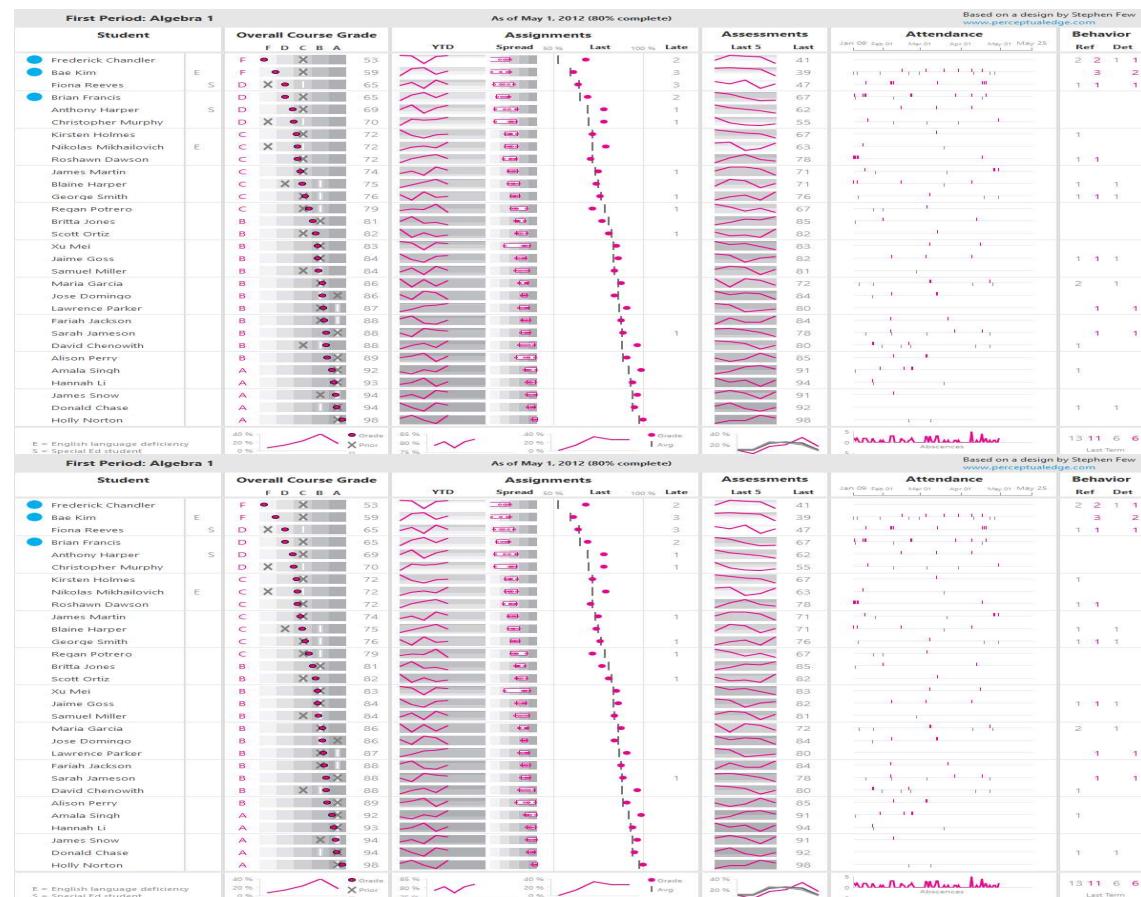
A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requête



A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requête

Tivoli **IBM.**

Service Level Exception

SLA: 1001
Applies To: INCIDENT
Status: ACTIVE
Service Group: IT
Service:
Date From:

Description: IT Generic P1 - Respond in 30 mins., Resolve in 2 hrs.
Type: CUSTOMER
Vendor:
Organization: EAGLENA
Site: BEDFORD
Date To:

# of Times SLA Applied within Time Frame	# of Times Commitments Met	# of Times Commitments were Violated	% of Compliance	% of Violations
3	3	6	33.33%	66.67%

Tickets

ID	Description	Type	Status	Classification	Contact Violation	Response Violation	Resolution Violation
1049	Failure when connecting to Moon Server	INCIDENT	RESOLVED	End User Issue \ Network \ Connection	N	Y	Y
1050	Error when trying to login to the network	INCIDENT	RESOLVED	End User Issue \ Network \ Connection	N	Y	Y
1051	Error message: Can't login to the network	INCIDENT	RESOLVED	End User Issue \ Network \ Connection	N	Y	Y

Response Commitment

ID	Description	Status	Classification	Target Response	Actual Response	Commt Delta
1049	Failure when connecting to Moon Server	RESOLVED	End User Issue \ Network \ Connection	10/6/04 12:55:58 PM	10/6/04 2:45:00 PM	1:49
1050	Error when trying to login to the network	RESOLVED	End User Issue \ Network \ Connection	10/6/04 12:57:10 PM	10/6/04 4:30:00 PM	3:33
1051	Error message: Can't login to the network	RESOLVED	End User Issue \ Network \ Connection	10/6/04 12:59:12 PM	10/6/04 2:20:00 PM	1:22

Resolution Commitment

ID	Description	Status	Classification	Target Resolution	Actual Resolution	Commt Delta
1049	Failure when connecting to Moon Server	RESOLVED	End User Issue \ Network \ Connection	10/6/04 2:25:58 PM	10/6/04 6:55:00 PM	4:28
1050	Error when trying to login to the network	RESOLVED	End User Issue \ Network \ Connection	10/6/04 2:27:10 PM	10/6/04 8:45:00 PM	6:18
1051	Error message: Can't login to the network	RESOLVED	End User Issue \ Network \ Connection	10/6/04 2:28:12 PM	10/6/04 3:15:00 PM	0:47



A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage



A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requête

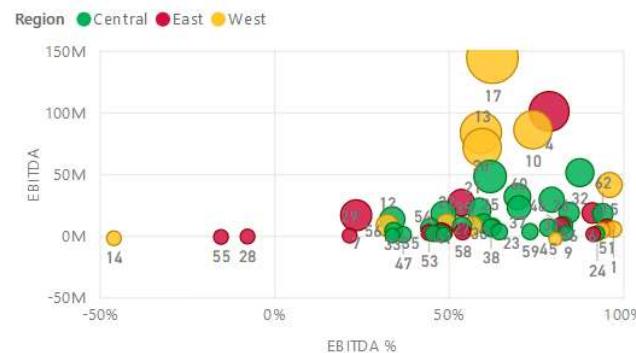
December

Revenue (YTD): **1 664M**
EBITDA (YTD): **1 064M**
BOE: **992M**

EBITDA par Division



Revenue and Profitability by District



EBITDA and BOE Trend

EBITDA (Blue bar) BOE (Red line)



Revenue and Operating Expense Trend

Revenue (Green bar) OPEX % of Revenue (Yellow line)



A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage

```
import mysql.connector

mydb = mysql.connector.connect(
  host="localhost",
  user="yourusername",
  passwd="yourpassword",
  database="mydatabase"
)

mycursor = mydb.cursor()

mycursor.execute("SELECT * FROM customers")

myresult = mycursor.fetchall()

for x in myresult:
  print(x)
```



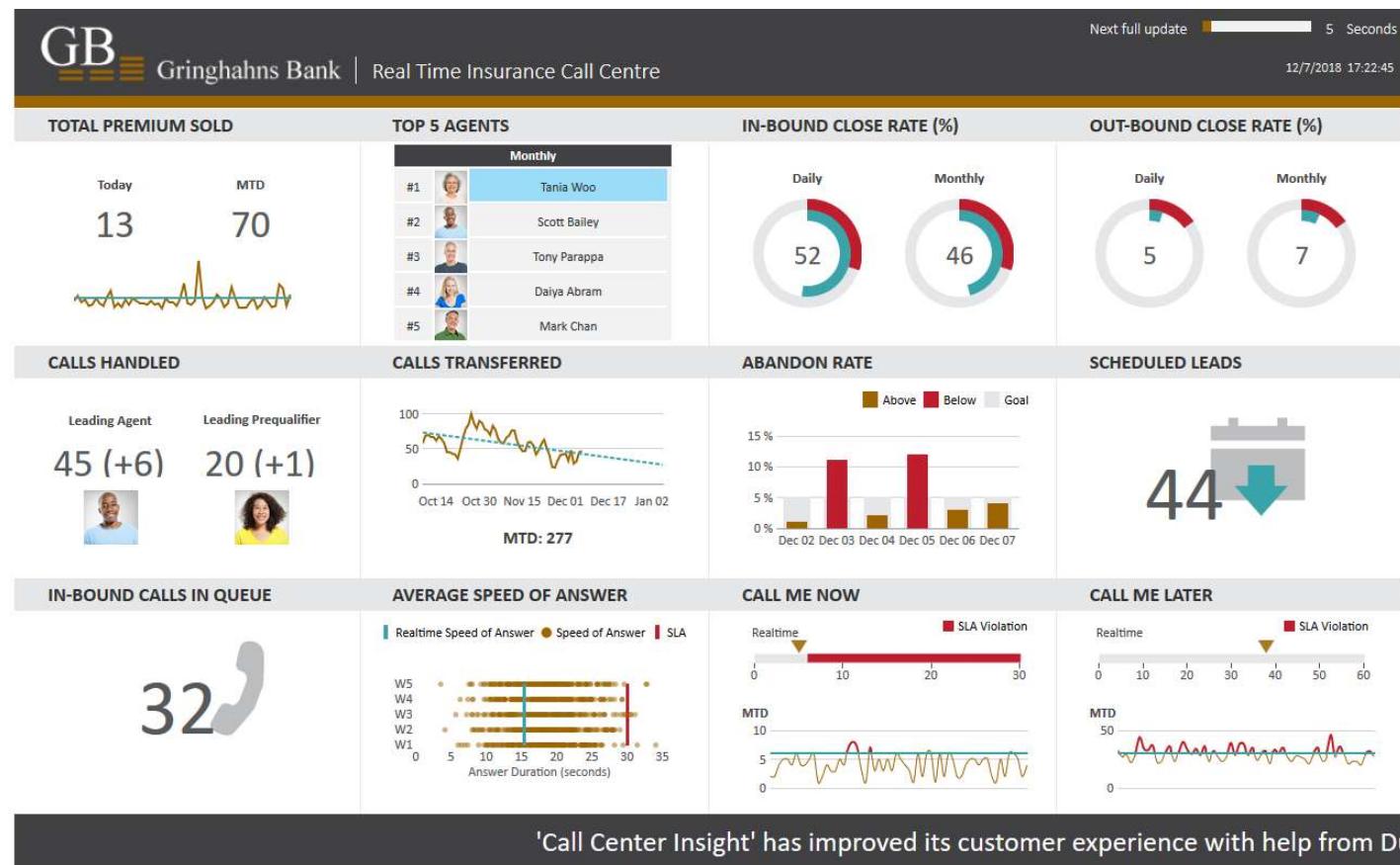
A vous de jouer

intégration

Traitement

analyse

Tableau de bord Stratégique/Opérationnel ou Reporting ou Requête



A vous de jouer

intégration

Traitement

analyse

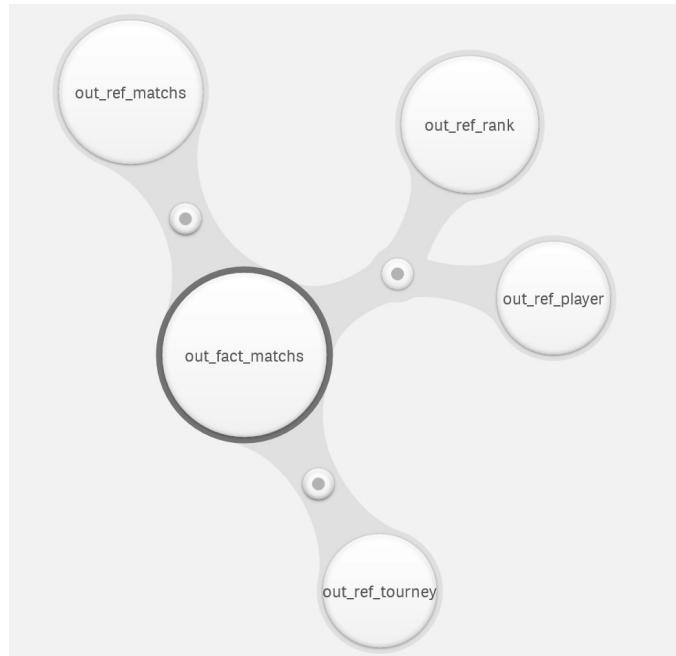
Tableau de bord Stratégique/Opérationnel ou Reporting ou Requêteage

```
library(RMySQL)  
  
mydb = dbConnect(MySQL(), user='username', password='password', host='rennes1.db.com', dbname="dbUnivRennes1")  
rs <- dbSendQuery(mydb, "select NAME_STUDENT,mean(NOTE_EXAM) as MEAN from FACT_NOTE_STUDENTS where CD_COURSE = 'M2_MAS_PPE' group by NAME_STUDENT")  
dbFetch(rs)  
dbClearResult(rs)  
...|
```



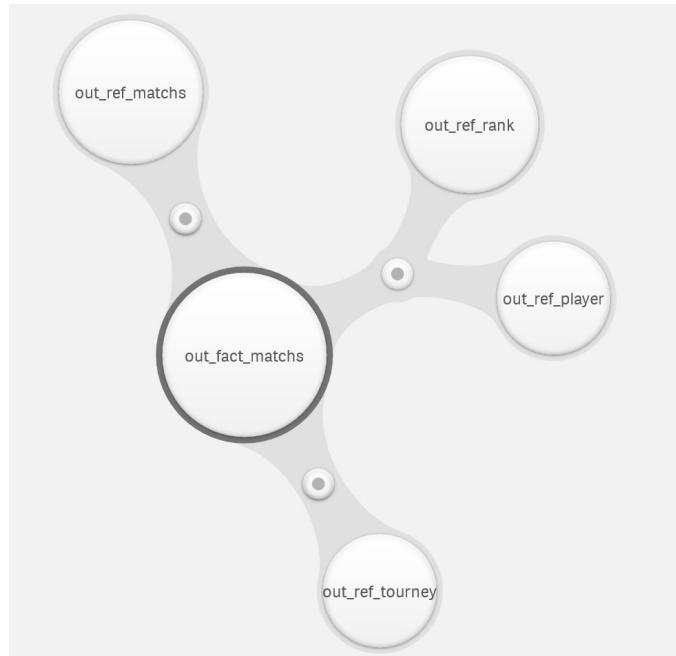
DataViz - exercice

1. Créer un compte sur wwwqlikcloud.com
2. Importer les données présentes dans le zip exercices/dataviz/atp_data.zip
3. Créer les associations entre les tables



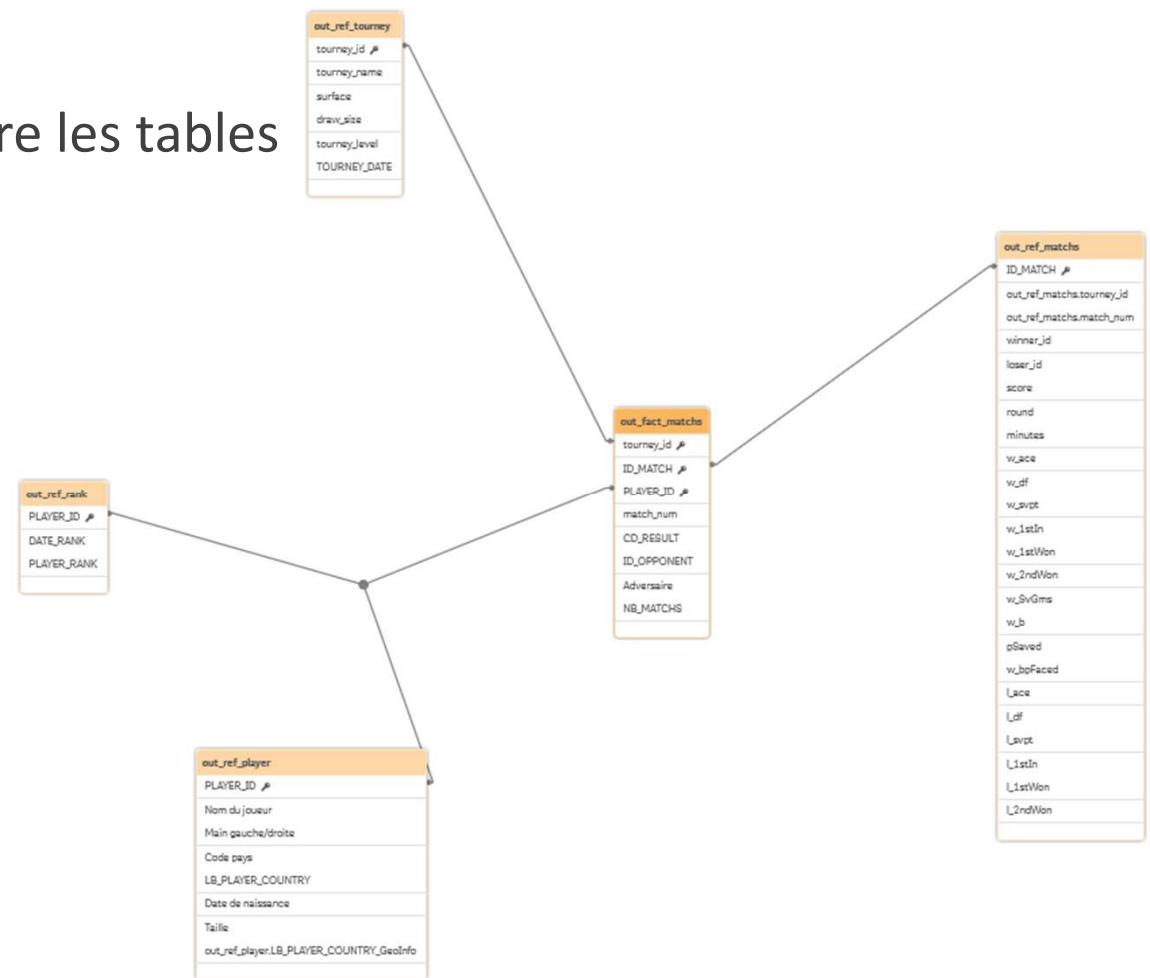
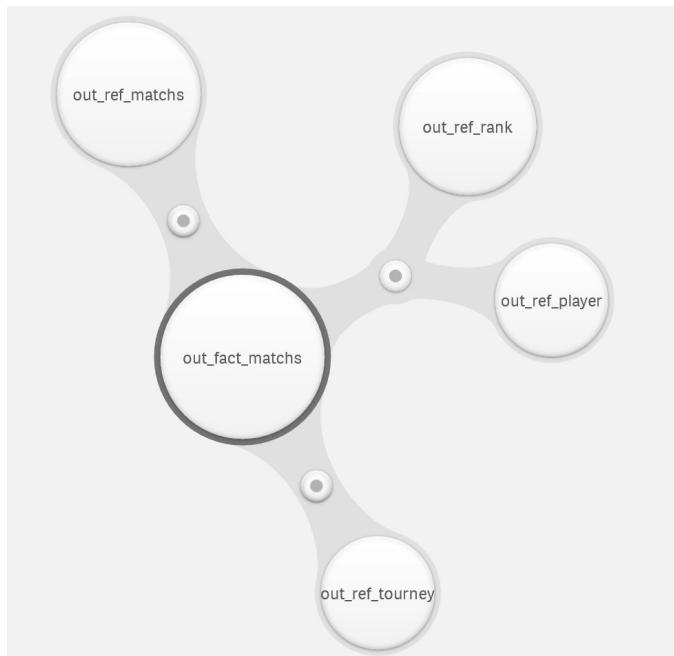
DataViz - exercice

1. Créer un compte sur wwwqlikcloud.com
2. Importer les données présentes dans le zip exercices/dataviz/atp_data.zip
3. Créer les associations entre les tables



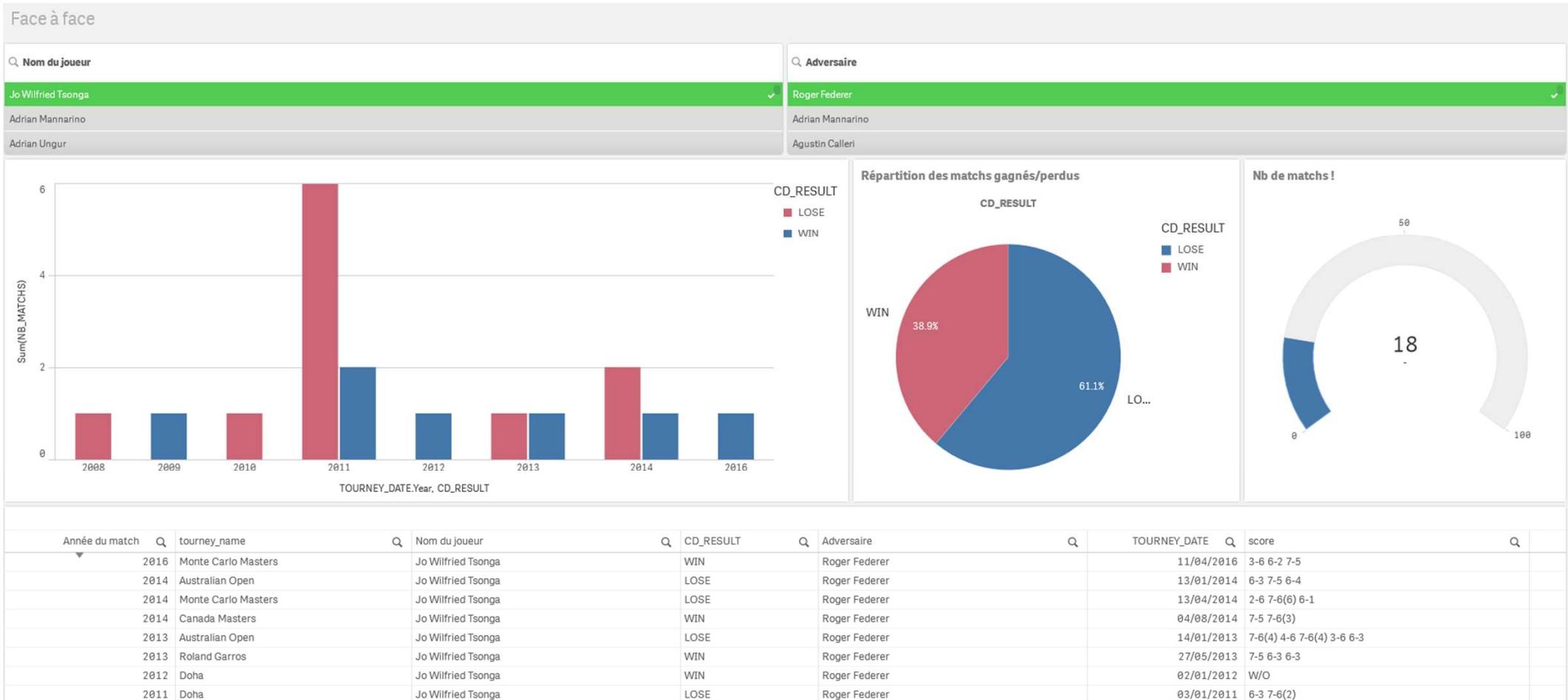
DataViz - exercice

1. Aller dans l'onglet analyse
2. Importer les données présentes dans le zip exercices/dataviz/atp_data.zip
3. Créer les associations entre les tables



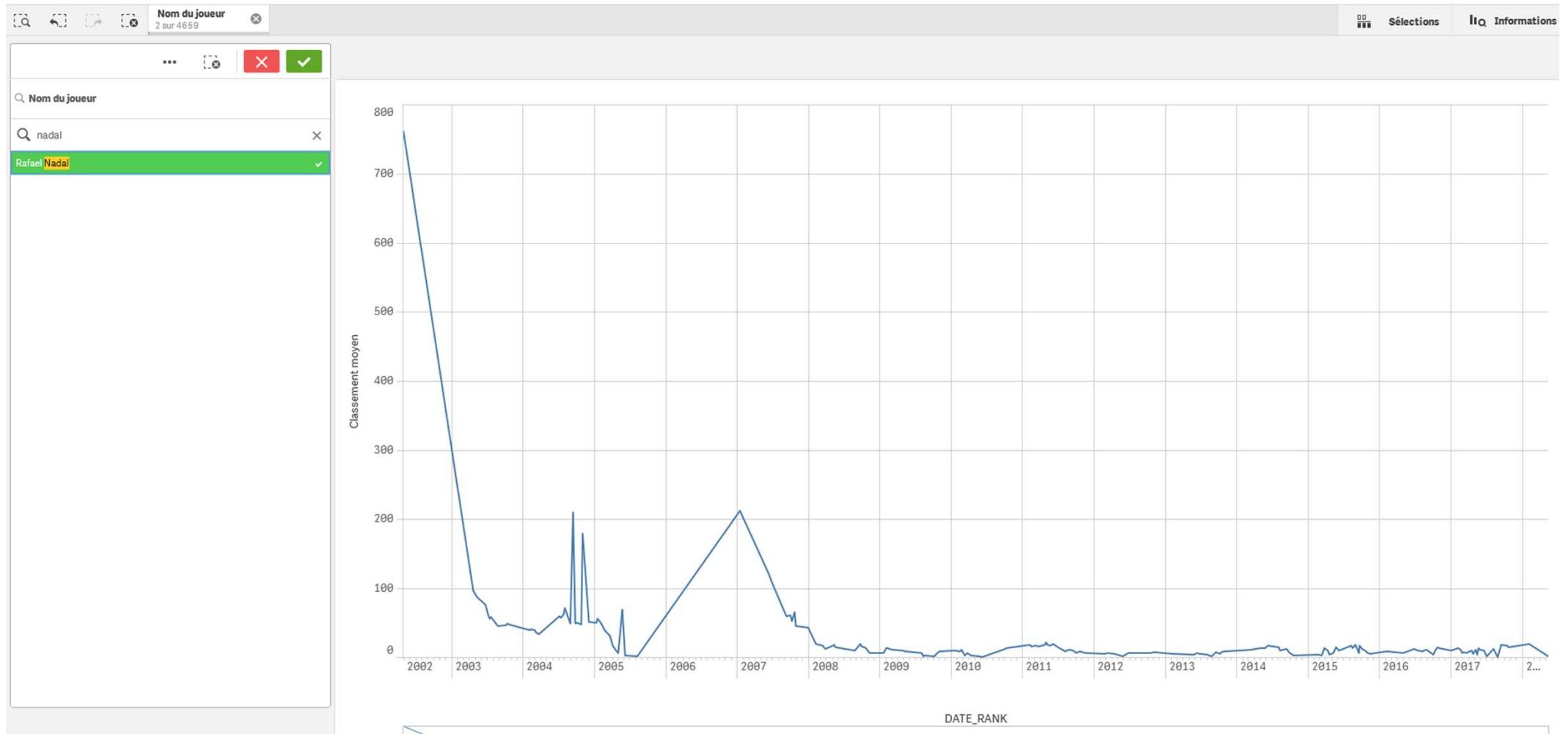
DataViz - exercice

Créer une première visualisation permettant d'avoir des statistiques sur des « Face à Face »



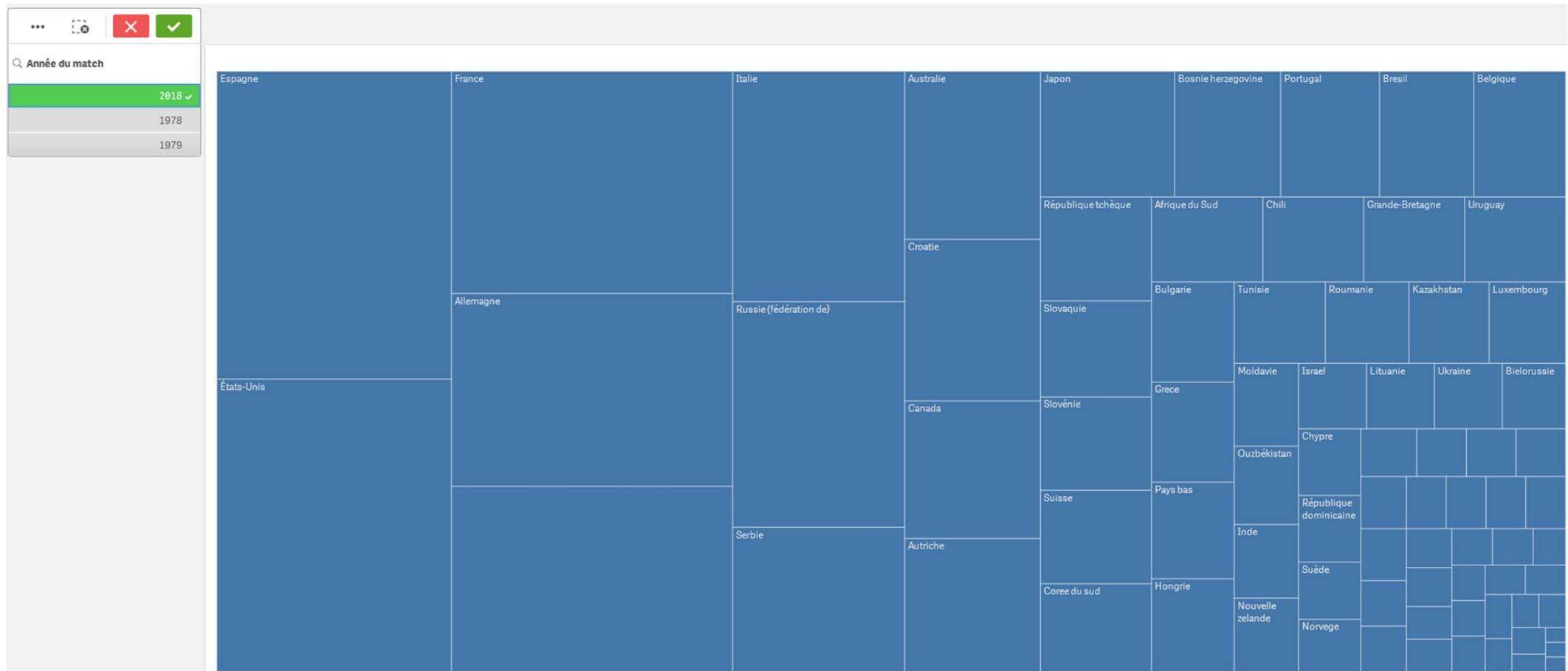
DataViz - exercice

Créer une deuxième visualisation permettant d'avoir l'évolution du classement



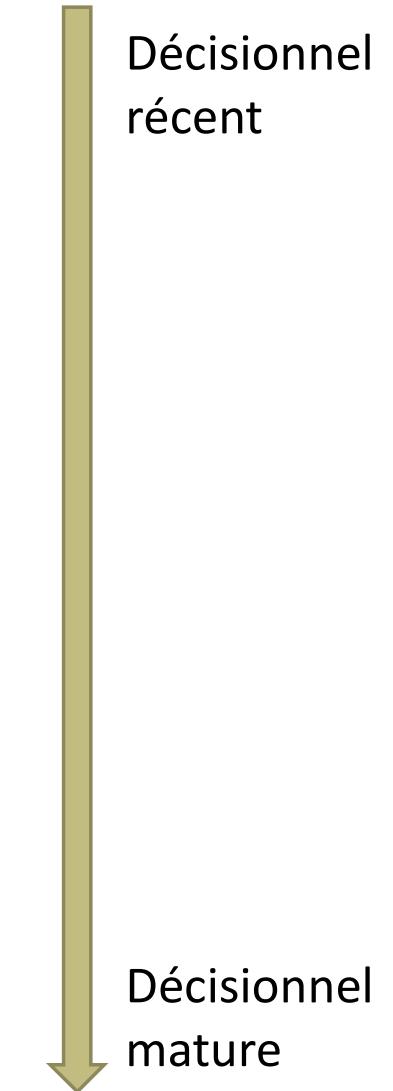
DataViz - exercice

Créer une troisième visualisation permettant d'avoir la répartition du nombre de matchs avec une hiérarchie : PAYS -> JOUEUR -> ADVERSAIRE

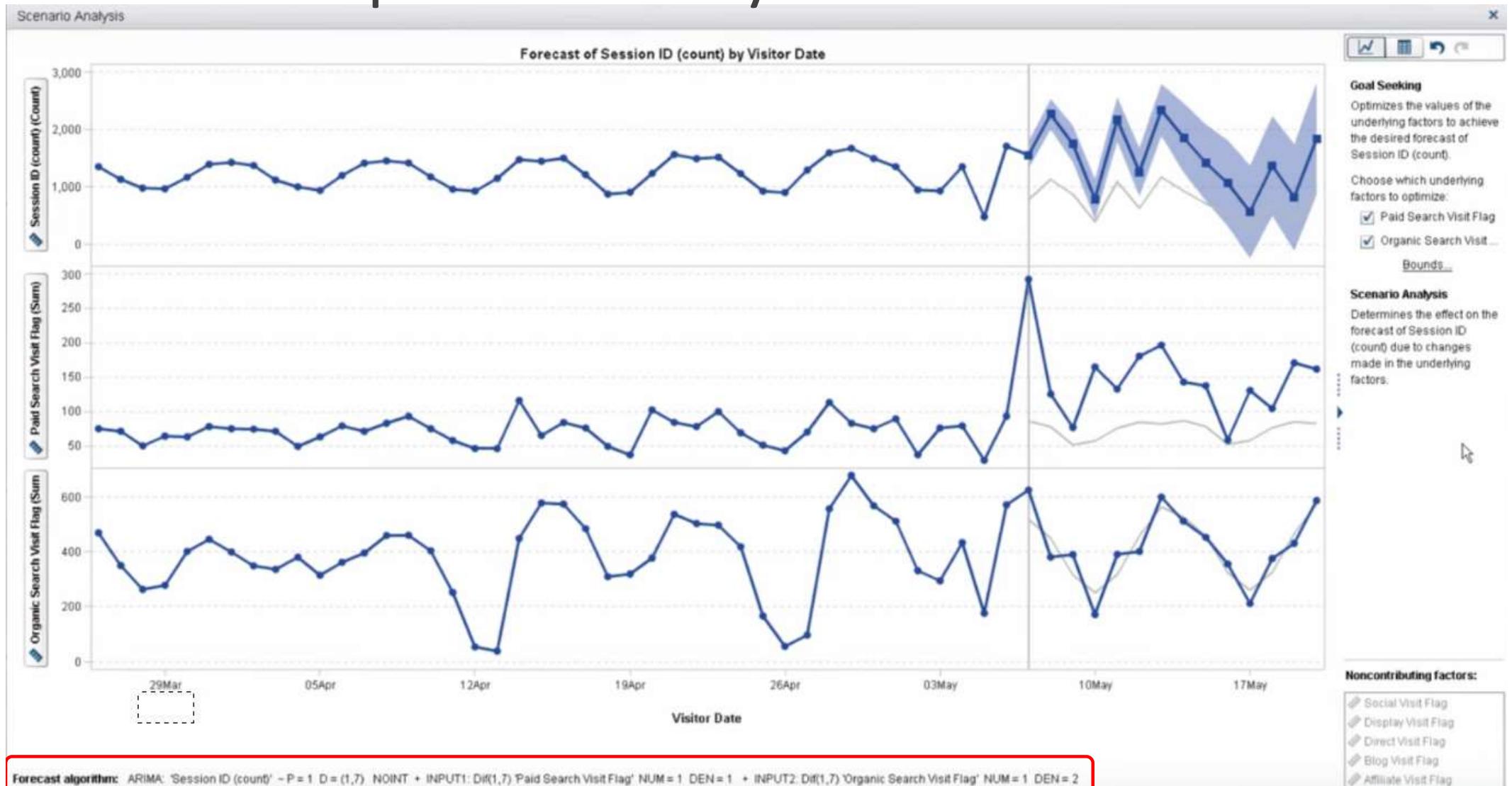


La prévision, la simulation

- **Etape 1 « Stocker »** : créer un entrepôt décisionnel avec un certain nombre de données, historiser cette information, réaliser des rapports, Tableau de bord
Factuel, Etat en cours
- **Etape 2 « Prévoir »** : utilise les données disponibles dans l'entrepôt pour prévoir (avec des modèles statistiques plus ou moins complexe) ce qui va se passer dans les prochains jours, semaines,...
Combien aurais-je de bénéficiaires de la nouvelle prestation de service que nous allons mettre en place ?
- **Etape 3 « Simuler »** : Réaliser des prévisions en fonction de scénarios prédéfinis.
Si je baisse le plafond de ressource de 15%, quel impact cela aura sur l'évolution des bénéficiaires d'ici la fin de l'année ?



Exemple d'analyse de scénario



Rôle des analystes en amont : trouver les bons modèles !



La gestion opérationnelle de la plate-forme décisionnelle

3 règles importantes pour la gestion opérationnelle

1. L'automatisation
2. L'automatisation
3. L'automatisation



La gestion opérationnelle de la plate-forme décisionnelle

Actions à réaliser

1. Vérifier l'arrivée des fichiers des différents partenaires aux dates convenues
2. Vérifier le bon déroulement des chargements (réalisées par l'outil ETL)
3. Vérifier que les ressources système (disque, mémoire, CPU) de la plate-forme
4. Gérer la mise en place des évolutions lors de la mise à jour de nouveaux indicateurs (gestion d'environnements de développement, validation,...)
5. Etre en contact avec les utilisateurs pour les remontées de problèmes, ou alerter les utilisateurs d'un problème dans les données



La gestion opérationnelle de la plate-forme décisionnelle

Ordonnanceur

Pour permettre une bonne supervision, un ordonnanceur doit être utilisé. Il permettra d'envoyer des alertes en cas de problème.

Fonctionnement Ordonnanceur

L'ordonnanceur appelle les traitements réalisés par l'ETL et envoie des alertes (visuelles, mail,...) en cas de problème dans les traitements



Les habilitations / le traçage de l'accès à la donnée

Mot magique : **Les métadonnées : « les données sur les données »**

Toutes les solutions décisionnelles contiennent une surcouche à la base de données, en général appelée « **métadonnées** ». Cette couche permet d'ajouter des informations à la base de données :

- ❖ ajouter une « couche métier » dans les outils de reporting permettant d'avoir des noms de données non techniques et compréhensibles par l'utilisateur
 - ❖ « Nature de la prestation » au lieu de NATPF
- ❖ gestion des accès (utilisateurs, rapports, données...)
- ❖ Tracer les accès



Les habilitations / le traçage de l'accès à la donnée

Mot magique : **Les métadonnées**

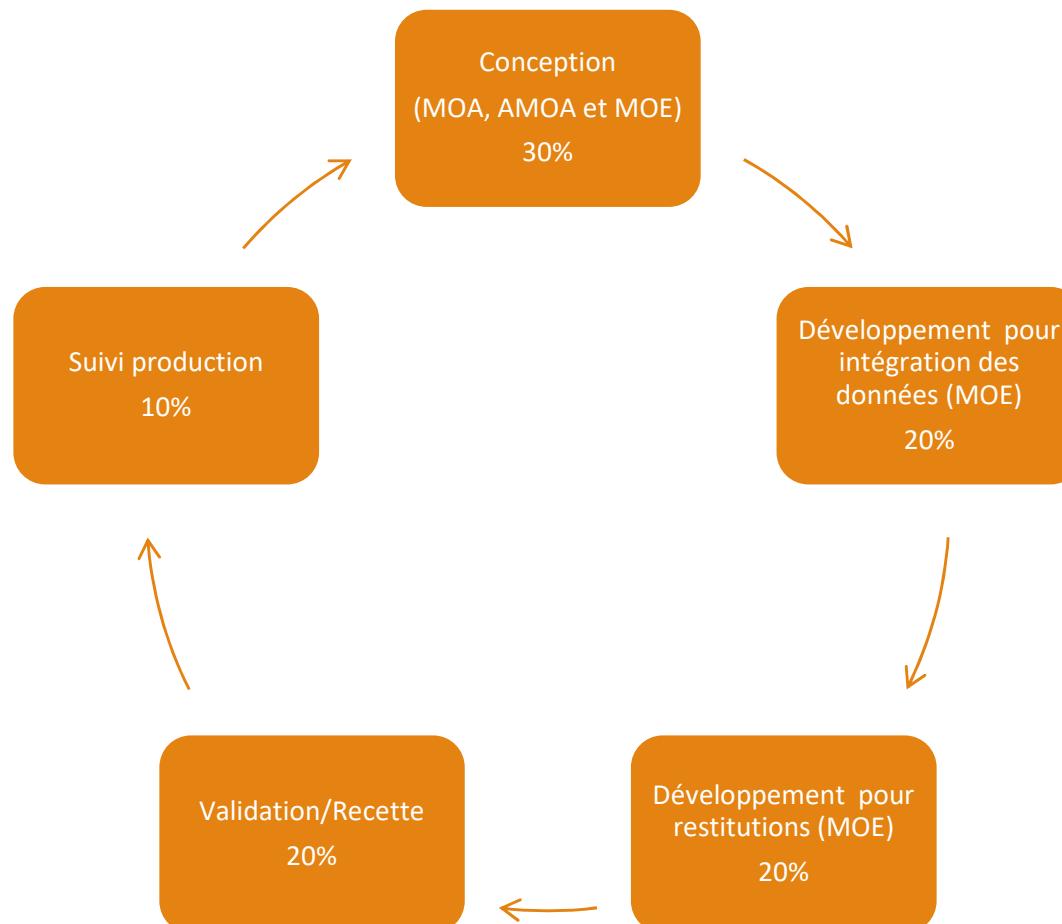
Ce module est lié à chaque l'outil. Là où le langage SQL est connu comme standard pour les accès aux bases de données, il n'y a pas de langage spécifique pour la partie métadonnées.

De manière générale, cette partie est gérée à la main, à travers des interfaces fournies par les outils de la suite décisionnelle. Les suites complètes peuvent intégrer des API (accessibles via JAVA par exemple), permettant d'automatiser la gestion de cette sécurité



L'équipe BI

Voici la répartition en termes de temps du process BI, de la création des indicateurs, jusque la mise en œuvre (une fois que l'ensemble de la solution est en place, bien sûr !)



Les métiers de la BI

Data Analyst

Un analyste de données, est un quelqu'un qui est **capable d'interroger des sources de données** pour en faire des rapports et des visualisations graphiques (graphes camemberts, histogrammes etc...). Un Data Analyst a une **compréhension forte du domaine métier** dans lequel il opère. Ce qui lui permet de mieux communiquer avec les gens du métier.

Pour mieux explorer les données, un Data Analyst est généralement à l'aise avec les outils statistiques. Toutefois, il n'est pas forcément aussi "calé" techniquement qu'un *software engineer* pour traiter les grands volumes de données (Big Data).

Compétences et outils : Excel, Access, SQL, SAS, SPSS, Tableau, Statistiques...

Source : <https://mrmint.fr>



Les métiers de la BI

Business Intelligence Developer

Les développeurs B.I. (Business Intelligence / informatique décisionnelle) vont mettre en place des outils de B.I. pour les besoins de l'entreprise. Ces outils se présentent généralement sous forme de *Data warehouses*, *Datamart*, ainsi que des *bases de données multidimensionnelles* construits à partir **d'agrégation de données** en provenance de plusieurs bases de données. La construction des *Data warehouse* et *les bases OLAP* est généralement effectuée à travers des Job ETL (Extract, Transform, Load) en utilisant l'outil **Talend** par exemple.

Ces Bases de données multidimensionnelles et *Data warehouses* sont par la suite utilisées par les développeurs B.I pour construire des tableaux de bords (*Dashboards*) et des rapports utiles pour les manageurs et les décideurs.

Les développeurs de B.I. ont généralement une connaissance métier moindre que celle d'un Data Analyst. Cependant, ils sont plus "calés" techniquement pour s'interfacer avec les différentes sources de données.

Compétences et outils : SQL, OLAP, Data warehouses, Cubes, SSAS, SSIS, ETL (Talend...)

Source : <https://mrmint.fr>



Les métiers de la BI

Data Engineer

Un Data Engineer est quelqu'un ayant un **background technique** en développement logiciel. Il peut être un *Software Engineer* qui s'est reconvertis dans le Big Data.

Les Data Engineers vont mettre en place des systèmes de Big Data pour traiter ces dernières. Ils opteront pour des outils de stockage performants comme les **bases de données NoSQL** et se baseront sur **Hadoop, Spark, Map/Reduce** pour traiter convenablement ces grands volumes de données.

Les Data Engineer vont collecter, transformer les données de différentes sources. Ce travail préparatoire permettra d'avoir **des données "propres"**, prêtes pour qu'on leur applique dessus des techniques de Machine Learning.

En d'autres termes, le travail d'un Data Engineer est de **préparer le terrain** pour qu'un Data Scientist puisse se servir des données propres pour en tirer des tendances (Insights).

Compétences et outils : SQL, NoSQL, Hadoop, Data Lake, Big Data, Spark, Software Engineering, Map/Reduce...

Source : <https://mrmint.fr>



Les métiers de la BI

Data Scientist

Un Data Scientist est un **profil pluridisciplinaire** qui aura pour mission première de tirer de l'**information utile** (*insights*) depuis des données brutes. Le métier du Data Scientist est à l'**intersection** entre *Data Analyst* et de *Data Engineer*. Tout en ayant des connaissances métiers dans le domaine dans lequel il évolue.

En effet, un *Data scientist* va **explorer et exploiter les gisements de données de l'entreprise** pour leur appliquer des techniques de machine learning. Il s'agit donc d'une forme de *Data Analysis* poussée sur de grands volumes de données. L'exposition au contexte Big Data exige qu'un Data Scientist soit familier avec des concepts comme **Map/Reduce, Hadoop, Data lake** etc...

L'information utile recherchée par un Data Scientist est **spécifique à une entreprise** et plus généralement à un domaine métier. Pour cela, un Data Scientist doit être à l'aise avec le domaine métier dans lequel il opère. Pour cela, il côtoiera les gens du métier pour creuser avec eux les différentes pistes de réflexion.

Finalement, un data scientist doit être **un bon communicant** pour mieux communiquer ses retrouvailles. Il usera pour cela des différents supports de présentation comme les présentations *PowerPoint*, ainsi que des visualisations graphiques (histogrammes, camemberts...) plus parlantes aux décideurs.

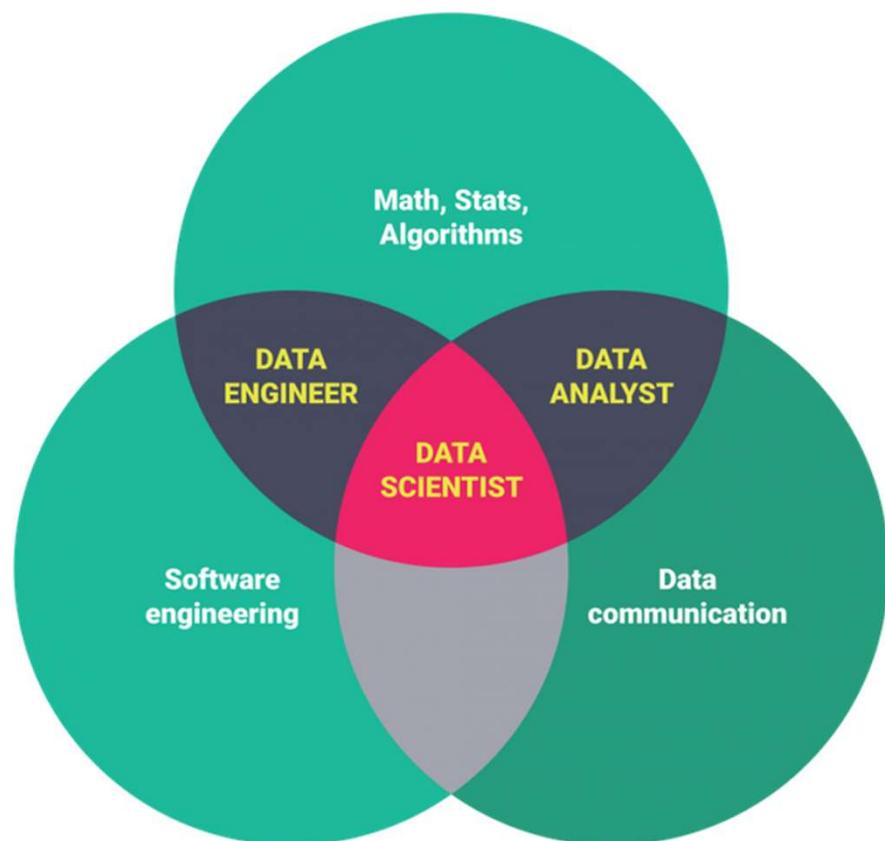
Compétences et outils : SQL, NoSQL, Python, R, Machine Learning, Deep Learning, Statistiques, Software Engineering...

Source : <https://mrmint.fr>

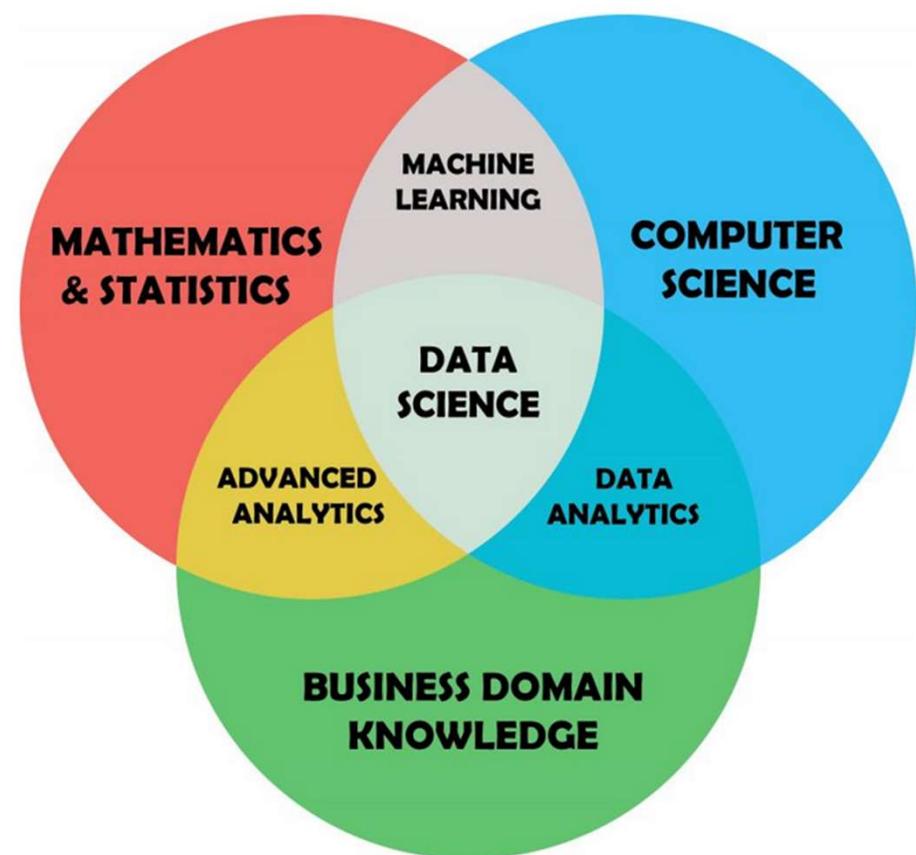


Les métiers de la BI

Métier



Domaine



Source : <https://mrmint.fr>



Quelques solutions commerciales



<https://www.appsruntheworld.com/top-10-analytics-and-bi-software-vendors-and-market-forecast/>



Quelques solutions Open Source

ETL	Entrepôt de données	OLAP	Reporting	Data Mining
<ul style="list-style-type: none">■ Octopus■ Kettle■ CloverETL■ Talend	<ul style="list-style-type: none">■ MySql■ Postgresql■ Greenplum/Bizgr es	<ul style="list-style-type: none">■ Mondrian■ Palo	<ul style="list-style-type: none">■ Birt■ Open Report■ Jasper Report■ JFreeReport■ Apache Superset	<ul style="list-style-type: none">■ Weka■ R-Project■ Orange■ Xelopes

Intégré

- Pentaho (Kettle, Mondrian, JFreeReport, Weka)
- Knowage (100% Open Source – payants pour l'ensemble des modules)



LE « BIG DATA » DANS TOUT CA



The W. EDWARDS
Deming
Institute

W. Edwards Deming

In God we trust, all others must bring data.

*attribution disputed,
see source link

source: quotes.deming.org/3734



« Big Data », c'est quoi ?

Littéralement, ces termes signifient **mégadonnées**, grosses données ou encore **données massives**.

Cette appellation est apparue en **octobre 1997** dans des articles scientifiques concernant des défis technologiques à relever pour visualiser de grands ensemble de données.

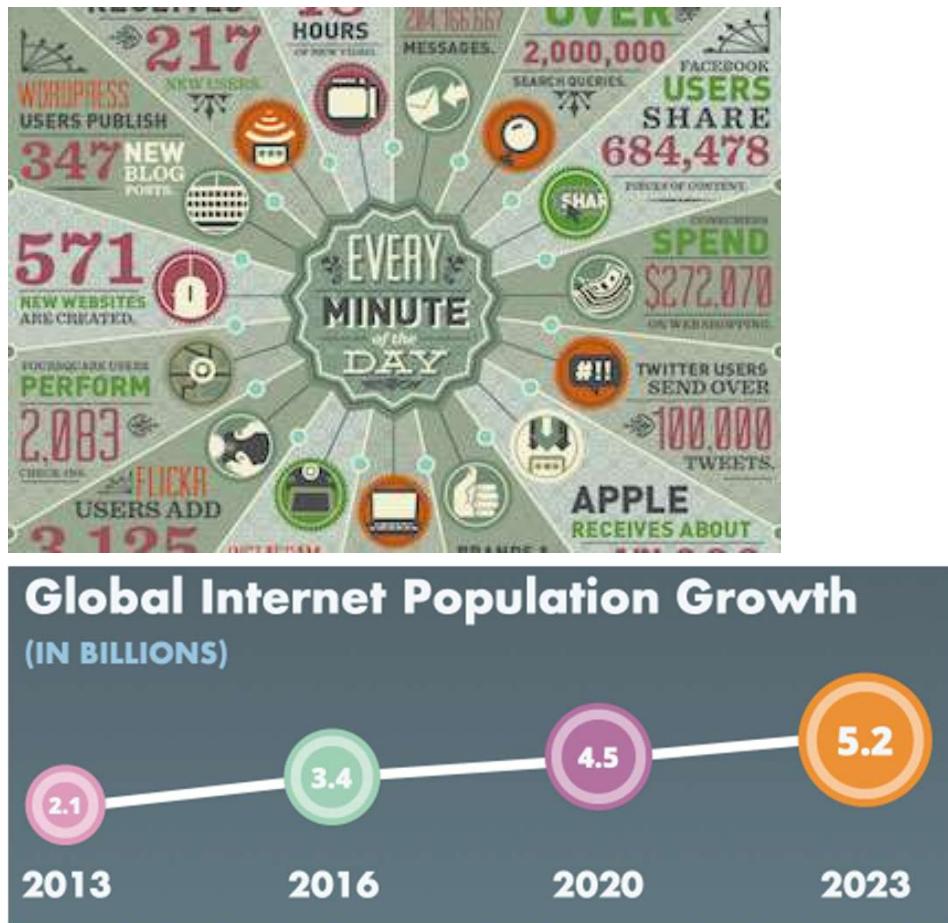
Aucune définition universelle ne peut être donnée, la définition change en fonction des usagers, communautés ou fournisseurs de services.



LE « BIG DATA » DANS TOUT CA

(domo.com)

2013

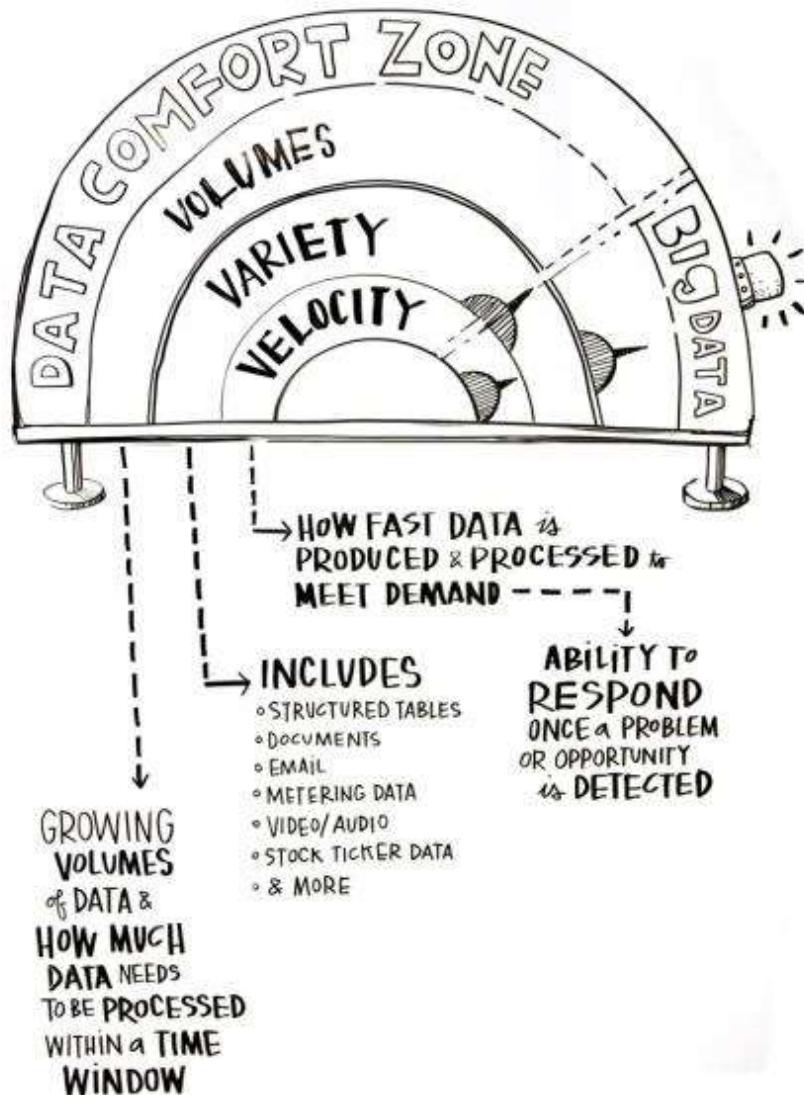


Source : domo.com

2023



Small ou Big Data ?



Volume

- De quel volume j'ai besoin pour réaliser mes différents traitements ou analyse

Variété

- Transactions, données structurées en tables, pdf, images, vidéo, audio, logs WEB,...

Vitesse

- Pourquoi faire du batch, quand nous pouvons Faire du TEMPS REEL !



Côté "Machines" comment répondre au besoin ?



NOUVEAUX SYSTÈMES DE FICHIERS

Les données sont réparties sur plusieurs machines

Les calculs sont distribués en simultanés sur plusieurs machines

« JE VEUX DOUBLER MA PERFORMANCE, JE DOUBLE LE NOMBRE DE MACHINES », le système est scalable à l'infini.



Côté "système de fichiers" comment répondre au besoin ?



« NOUVEAUX » SYSTÈMES DE FICHIERS

HDFS (Hadoop File System)

CFS (Cassandra File System)

GridFS (MongoDB)

The **Teradata** Database File System

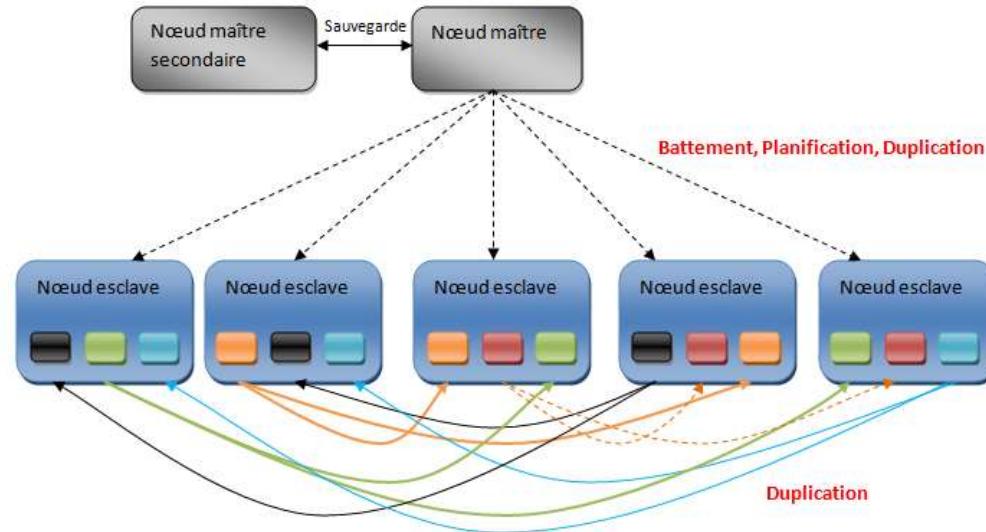
S3 (Amazon ou minIO)

MapRFS

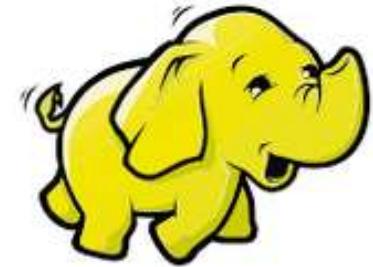
...



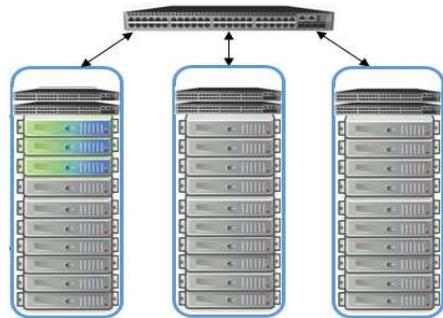
Côté "système de fichiers" comment répondre au besoin ?



Hadoop : le gagnant



hadoop



Hadoop est un framework (contient un ensemble de librairies/outils)

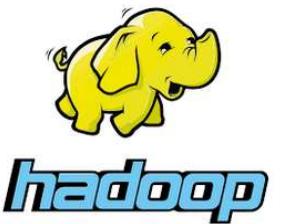
Hadoop est distribué par la fondation Apache, en Open Source

Hadoop permet le stockage des fichiers (HDFS)

Hadoop permet de traiter ces gros volumes de données à partir de différents moteurs

- Hive : moteur simili SQL
- Hbase : NoSQL orienté colonne
- Spark : moteur In Memory pour l'analyse de données
- Impala (incubating) : base de données orientée « décisionnel »





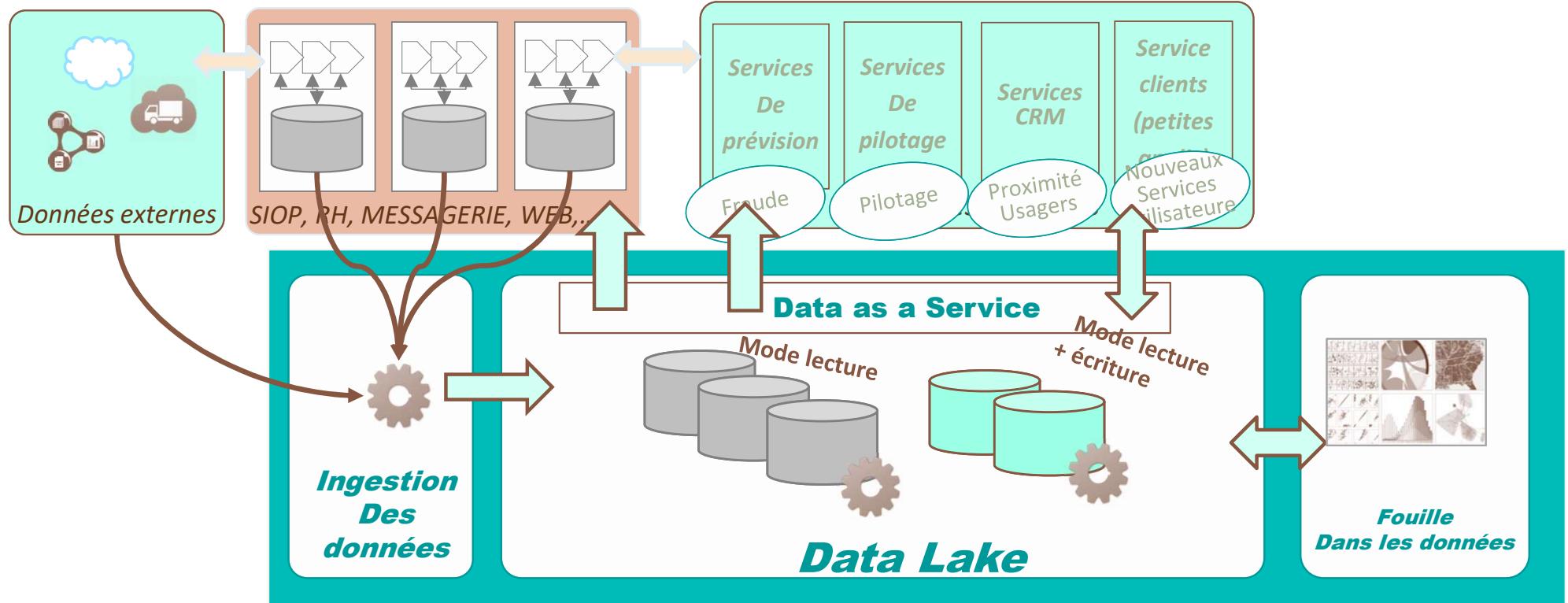
Distributions Hadoop

Pour éviter de passer du temps en R&D et intégration des différents modules listés précédemment, il est utile pour une grande société de passer par des distributions commerciales.

- Framework Hadoop personnalisé, stable
- Fonctionnalités supplémentaires pour la sécurité et l'administration
- Documentation, formation
- Modules SQL spécifiques



VISION SI ORIENTEE Big Data



BIG DATA et analyse

Concrètement : Comment exploiter toutes ces données, et comment rendre interprétable une forte masse de données !

- Outils de DataVisualisation vont se plonger sur les architectures Big Data pour en utiliser la performance (disque / mémoire) et offrir des temps de requête optimums

sasvisualanalytics
qliksense
microstrategy
sisense shiny
domo
powerbi
tableau thoughtspot



BIG DATA et Machine Learning

Des librairies Open Source complètes qui utilisent la performance des clusters Big Data

(classification, régression, clustering, ...)

Des solutions payantes (SAS Viya, Dataiku,...) intégrant ces librairies open source

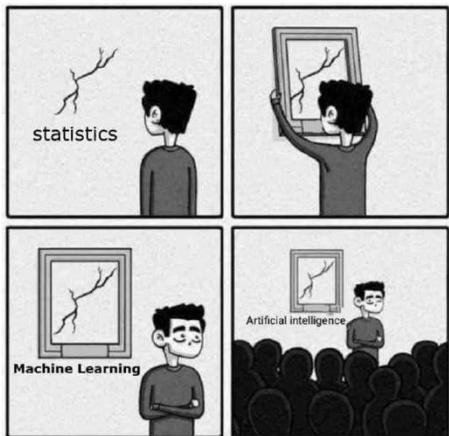
- Spark ML ou MLLib (accessible en Java, R, Python ou Scala)
- H2O (accessible en Java, R, Python ou Scala)

Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms



Un mot sur l'Intelligence Artificielle

Définissons l'IA



*When you're fundraising,
it's Artificial intelligence*

*When you're hiring to do it,
it's Machine Learning*

*When you put it into production,
It's Linear Regression*

Aleksander Callebat – Data scientist @ microsoft



Mat Velloso
@matvelloso

[Follow](#)

Difference between machine learning
and AI:

If it is written in Python, it's probably
machine learning

If it is written in PowerPoint, it's
probably AI

5:25 PM - 22 Nov 2018

1,186 Retweets 3,333 Likes



41 1.2K 3.3K



Sources

- Modélisation décisionnelle, Thibault Bourcy, Eni Editions – 2017
- Gartners Magic quadrant
- domo.com



Sujet : la planète et l'énergie

Vous êtes membre de « Réseau action climat », et vous devez fournir un outil sur le site WEB permettant de fournir quelques indicateurs, par pays, sur la consommation énergétique, la production énergétique, le bilan carbone. Vous devrez trouver des indicateurs qui « comparent ce qui est comparable ».

La temporalité sera un facteur important de votre travail, puisque l'objectif in fine, est de s'assurer que les pays sont dans des dynamiques de réduction de leur consommation, et réduction d'émission de CO₂.

- Livrable 1 : expliquer votre démarche/méthodologie pour mettre en place ce tableau de bord (livrable : *nom_equipe_livrable1.doc*)
- Livrable 2 : Lister 4 indicateurs qui permettent de répondre au besoin. Créer les « fiches indicateur » associées (livrable : *nom_equipe_livrable2.doc*)
- Livrable 3 : Fournir une maquette visuelle de votre tableau de bord (livrable : *nom_equipe_livrable3.doc* – images réalisées en Excel / PPT ou scans de papiers)
- Livrable 4 : Tout ou partie du tableau de bord, réalisé avec PowerBI (livrable : *nom_equipe_livrable4.pbix*)

