pensaient qu'il n'était pas possible de le relever. L'algorithme d'apprentissage pourrait en effet se retrouver coincé dans un minimum local, d'où il ne pourrait sortir en ajustant les paramètres par petites touches.

Or avec mes collègues, nous avons montré que si le réseau de neurones est assez grand, le problème des minima locaux est fortement réduit. Dans un tel réseau, la plupart des minima locaux correspondent à un niveau d'apprentissage quasi équivalent à celui du minimum global.

Bien que les problèmes théoriques de l'optimisation soient, en principe, résolubles, la construction de grands réseaux comportant plus de deux ou trois couches avait souvent échoué. À partir de 2005, les recherches financées par l'Icra ont commencé à porter leurs fruits et les obstacles ont été surmontés progressivement. En 2006, nous sommes parvenus à entraîner des réseaux neuronaux plus profonds en utilisant une technique qui opérait couche par couche.

Par la suite, en 2011, nous avons trouvé un moyen d'entraîner des réseaux encore plus profonds (avec davantage de couches de neurones virtuels) en modifiant les opérations effectuées par chacun des neurones, ce qui les rapprochait davantage des neurones biologiques dans leur façon de traiter l'information. Nous avons aussi découvert qu'en injectant du bruit aléatoire dans les signaux transmis entre les neurones au cours de l'apprentissage – comme ce qui a cours dans le cerveau –, les réseaux apprenaient mieux à identifier correctement une image ou un son.

Par ailleurs, deux facteurs essentiels ont contribué au succès des techniques d'apprentissage profond. Le premier est l'augmentation d'un facteur 10 de la puissance de calcul des ordinateurs grâce aux processeurs dédiés au traitement d'image, ce qui a permis d'entraîner des réseaux plus grands en un temps raisonnable. Le second facteur est que l'on avait désormais accès à d'énormes bases de données étiquetées, avec lesquelles les algorithmes d'apprentissage ont pu s'exercer à reconnaître un « chat », par exemple.

Une autre raison aux récents succès de l'apprentissage profond est que le réseau est capable d'apprendre à effectuer un ensemble de calculs qui construisent ou analysent pas à pas une image, un son ou un autre type de donnée. Avec une profondeur suffisante, ces dispositifs excellent dans beaucoup de tâches de reconnaissance visuelle ou auditive. Des travaux théoriques et expérimentaux récents ont d'ailleurs montré qu'il est impossible d'effectuer efficacement certaines de ces opérations mathématiques sans un réseau assez profond.

Mais que se passe-t-il au cœur du réseau de neurones profond ? Chaque couche transforme son entrée et produit une sortie qui est envoyée à la couche suivante (*voir l'encadré page 45*). Les premières couches se concentrent sur les détails aux échelles les plus petites, puis les couches suivantes agrandissent les échelles considérées. Plus les couches sont profondes, plus elles vont représenter des concepts abstraits.

## RECONNAÎTRE DES VIDÉOS

Jusqu'à récemment, les réseaux de neurones artificiels se distinguaient en grande partie par leur capacité à effectuer des tâches telles que la reconnaissance de motifs dans des images statiques. Mais d'autres types de réseaux neuronaux affichent des résultats intéressants sur des événements dynamiques. Les « réseaux neuronaux récurrents », par exemple, ont démontré leur capacité à effectuer une séquence de calculs pour traiter la parole ou une vidéo. Les données séquentielles sont constituées d'unités (phonèmes ou mots dans le cas de la parole) qui se suivent. La façon dont les réseaux neuronaux récurrents traitent leurs entrées ressemble un peu au fonctionnement du cerveau. Les signaux qui se propagent parmi les neurones changent constamment à mesure que les entrées sensorielles sont traitées.

Les réseaux récurrents parviennent à prédire quel sera le mot suivant dans une phrase et peuvent ainsi produire une nouvelle séquence de mots, l'un après l'autre. Ils peuvent aussi s'atteler à des tâches plus complexes. Après avoir « lu » tous les mots d'une phrase, le réseau est à même de deviner le sens de la phrase entière. Un réseau récurrent distinct peut alors utiliser le traitement sémantique du premier réseau pour traduire la phrase dans une autre langue.

La recherche sur les réseaux neuronaux récurrents a connu sa propre période de stagnation à la fin des années 1990 et au début des années 2000. Mes propres travaux théoriques suggéraient qu'ils rencontreraient des difficultés à apprendre à récupérer de l'information du passé lointain, à savoir les premiers éléments de la séquence en cours de traitement. Mais certains de ces problèmes ont été résolus grâce à des techniques consistant à stocker l'information pour qu'elle persiste durablement. Ces réseaux neuronaux utilisent la mémoire temporaire (mémoire cache) d'un ordinateur pour traiter des informations multiples et dispersées, comme des idées contenues dans différentes phrases réparties au fil d'un document.

Le retour en force des réseaux neuronaux à apprentissage profond après le long sommeil de l'intelligence artificielle n'est pas uniquement un triomphe technique. C'est aussi une leçon de sociologie des sciences. En particulier, il souligne la nécessité de soutenir des idées qui font fi du *statu quo* technologique et d'encourager la tenue d'un portefeuille de recherches diversifié, laissant de la place à des champs provisoirement passés de mode. ■

## BIBLIOGRAPHIE

**D. SILVER *ET AL.*,** ***Mastering the game of Go without human knowledge,*** **Nature, vol. 550, pp. 354–359, 2017.**

**Y. LECUN *ET AL.*,** ***Deep learning*****, Nature, vol. 521, pp. 436-444, 2015.**

**Y. BENGIO *ET AL.*,** ***Representation learning: A review and new perspectives*****, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35(8), pp. 1798-1828, 2013.**

**ImageNet classification with deep convolutional neural networks, 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), 2012.**

## SUR LE WEB

**Cours de Yann LeCun au Collège de France, 2015-2016: https://www.college-de-france.fr/site/yann-lecun/index.htm**


© Air France


© Pixabay


© SNCF

# « Quand les “ data scientists ” nous simplifient la vie »

**En ce début de XXIe siècle,** la science des données s’immisce tous les jours un peu plus dans notre quotidien. Reconnaissance vocale, moteurs de recherche, publicités ciblées, recommandations d’achats… Elle est bien sûr déjà omniprésente dans nos activités numériques. Mais cette *data science* gagne aussi progressivement tous les secteurs d’acti-vités économiques. Décoller en A380, emprunter sereinement son train quotidien ou bien encore se remettre d’une catastrophe naturelle: voici trois cas très concrets où des algorithmes aussi discrets qu’efficaces sont déjà à la manœuvre. Autant d’applications qui nécessitent les compétences de *data scientists* chevronnés… Un secteur d’activité florissant qui pourrait créer trois millions d’emplois en Europe! Et grâce au nombre et à la qualité de ses formations, la France est bien placée pour ne pas laisser filer sa part du gâteau… Rien que sur l’Hexagone, le recrutement de 130 000 *data scientists* sera nécessaire d’ici à 2020. En ferez-vous partie?

## SOMMAIRE



**Cahier spécial réalisé en partenariat avec**

AIRFRANCE MAIF assureur militant SNCF

# Décoller à l'heure

Les Airbus A380 d'Air France qui atterrissent à Roissy sont aujourd'hui analysés par des algorithmes de maintenance prédictive. Objectif : empêcher toute panne inopinée source d'importants retards… Reportage.

**Lundi 11 décembre 2017**, 10 heures du matin. Dans un froid glacial, un Airbus A380 d'Air France en provenance de Washington D.C. atterrit sur le tarmac de l'aéroport parisien Roissy-Charles-de-Gaulle. Les passagers commencent à s'ébrouer puis quittent leurs sièges, tandis que tout un ballet se met en place autour du géant des airs pour sortir les bagages des soutes, récupérer les déchets, faire des contrôles sur l'appareil, etc. Mais ce que les passagers ne savent pas, c'est qu'un autre ballet démarre en même temps, numérique celui-là…

En effet, à peine a-t-il posé les roues sur la piste que notre A380 commence à transmettre par wifi toutes les données enregistrées en continu sur son dernier aller-retour *via* les centaines de milliers de capteurs dont ses équipements sont truffés. Consommation électrique, vitesse, angle, température… L'ensemble de ces *data* (1 gigaoctet en moyenne) est alors traité par un logiciel de maintenance prédictive baptisé Prognos. Ce dernier transmet ensuite ses résultats en temps réel aux ingénieurs du département maintenance avion d'Air France-KLM.

## 1 Go de données par vol

C'est là que nous rencontrons Paul-Louis Vincenti, *data analyst* au département recherche opérationnelle chez Air France Industries-KLM Engineering & Maintenance. «*Grâce à une infrastructure big data distribuée dédiée* [1]*, Prognos détecte en moins de trente minutes les éventuelles signatures de prémices de pannes cachées dans cette masse de données*», explique-t-il. Le cas échéant, le service maintenance agit de manière préventive… De quoi éviter toute panne inopinée qui empêcherait l'avion de décoller, générant un retard important, voire une annulation de dernière minute.

Dans la pratique, les algorithmes [2] de Prognos mettent en œuvre des méthodes de *machine learning* par apprentissage supervisé * qui s'entraînent sur l'historique de précédents vols. «*Nous sommes en effet très attachés à l'explicabilité des résultats,* précise P. L. Vicenti. *Voilà pourquoi nous évitons volontairement le* deep learning*, et les réseaux de neurones* **: *nous voulons toujours pouvoir justifier d'un point de vue physique les décisions prises*».

En 2017, grâce à Prognos, les équipes d'Air France ont remplacé dans les A380 huit pompes d'alimentation moteur, et quatre capteurs de rotation logés dans le nez des trains d'atterrissage. «*Jusqu'à présent,* confirme P. L. Vincenti*, toutes ces déposes concernaient bien des équipements déclarés défectueux après inspection*». Fort de ces résultats, l'outil de maintenance prédictive a aussi été mis en place sur les dix-huit Boeing 747 de KLM.

* Apprentissage automatique à partir d'exemples annotés
** Méthodes mimant la profondeur des couches d'un cerveau

1. Hadoop Distributed File System
2. Développés en langages Python et Spark, ils s'appuient sur la librairie Scikit learn


© Air France

*« Les algorithmes de maintenance prédictive des A380 doivent toujours rester explicables »*

**PAUL-LOUIS VICENTI**
***Data analyst* chez Air France-KLM**

### AMÉLIORER LA RELATION AVEC LES PASSAGERS

**Chez Air France-KLM, une seule et même plateforme *big data* centralise aussi toutes les données que la compagnie collecte sur ses clients *via* divers canaux: *call centers*, site web, réseaux sociaux, dans l'aéroport, lors de salons… ou bien encore à bord des avions. Toutes ces données sont passées à la moulinette d'algorithmes qui en tirent des recommandations adaptées au profil de chaque client: options pour le menu à bord du prochain vol, propositions de destinations et de tarifs personnalisés, temps de transport jusqu'à l'aéroport de départ, guidage du passager en correspondance dans l'aéroport… Sur ordinateur ou tablette, tous les personnels au sol ou en cabine auront bientôt accès au dossier de chaque client, et à ses éventuels échanges en cours avec d'autres services de la compagnie. Le tout se fait bien sûr dans le respect de la vie privée définie par la Cnil.**

# Des sinistrés pris en charge plus vite

Grâce à la *data science*, on commence à prédire l'étendue des dégâts… avant même qu'une catastrophe naturelle s'abatte sur une région donnée. De quoi mieux anticiper la prise en charge des futurs sinistrés. Explications…

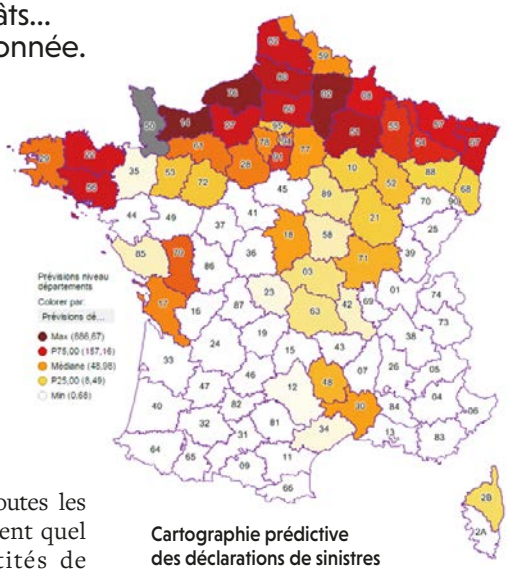
**Inondations, tempêtes, cyclones, ouragans…** Entre 1988 et 2013, les catastrophes naturelles ont coûté 1,9 milliard d'euros en indemnisations, avec 431 000 sinistrés en moyenne chaque année. Quand des catastrophes de ce type s'abattent sur un territoire, les assureurs activent immédiatement leurs plans de gestion de crise. Principal objectif: traiter au plus vite l'arrivée en masse de dossiers de demande d'indemnisation d'assurés impactés par la catastrophe et qui risquent d'affluer en agences. Mais aujourd'hui les assureurs veulent aller plus loin…

## «Data visualisation» des futurs sinistrés

**ORLANE MONNET**

«*À la MAIF, nous développons par exemple un outil capable de prédire le nombre de déclarations de sinistres sur telle ou telle zone, avant même qu'elle ne soit touchée par une tempête hivernale ou une inondation cévenole*», lance Orlane Monnet du département Études statistiques de cet assureur. Calibré sur la sinistralité des événe-ments climatiques passés les plus significatifs (tempêtes Klaus, Xynthia, Joachim…), ce modèle de prévision utilise deux grands types de données. Tout d'abord, des données sur les communes et les assurés qui risquent d'être touchés: maisons ou appartements, domiciles ou résidences secondaires, propriétaires ou locataires, densité de population, existence ou non d'un plan de prévention des risques, etc. Ensuite, le modèle utilise aussi bien sûr des données météorologiques: force du vent prévue, durée estimée, quantité de précipi-tations à venir, etc.

«*Écrit en langage R*, notre outil se lance et s'utilise* via *une seule et unique interface de* data visualisation** *nommée Tibco Spotfire,* indique Orlane Monnet. *Elle permet de visualiser sur une carte le nombre de déclarations de sinistres prévisibles – plus les zones sont touchées plus elles apparaissent foncées.*» Par simple clic sur des curseurs, elle permet aussi de mesurer la sensibilité du modèle de façon interactive; exemple, si on majore par sécurité toutes les rafales de 10 km/h, on voit directement quel pourcentage augmente les quantités de déclarations de sinistres sur la carte. La MAIF a commencé à utiliser son outil sur des inondations fin 2016 et sur de petites tempêtes début 2017.




Cartographie prédictive des déclarations de sinistres par département

## Anticiper les réparations

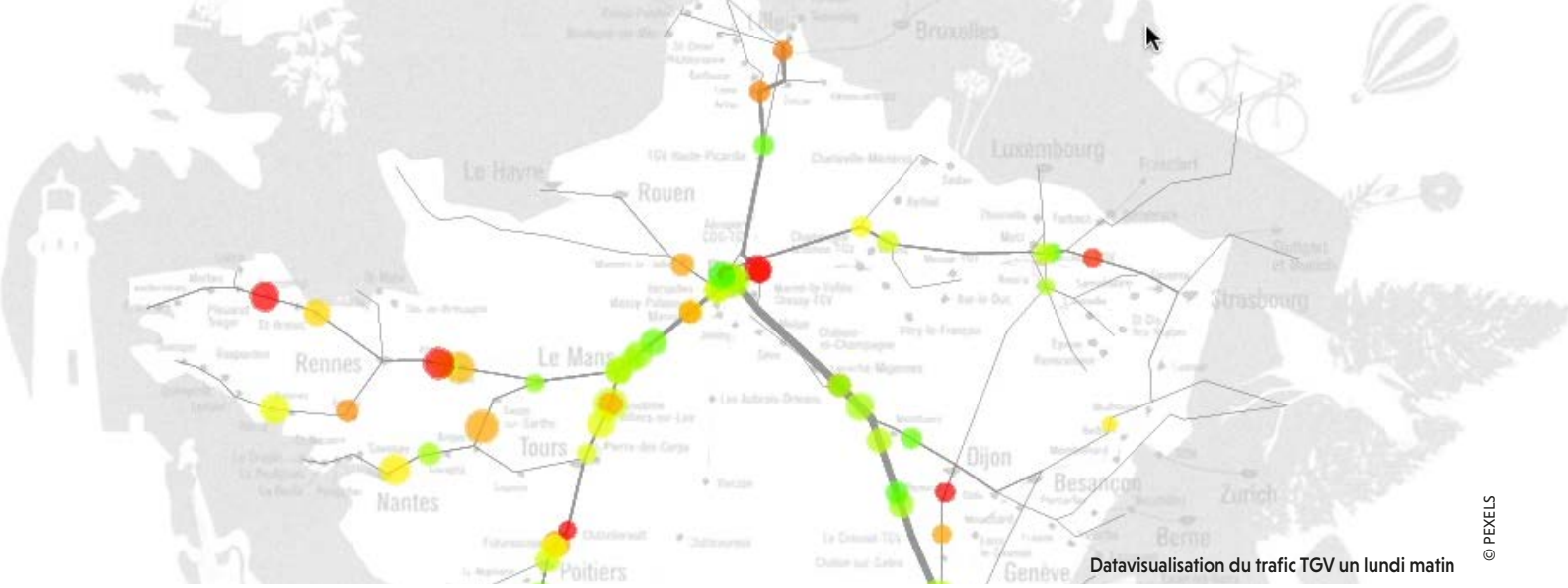
À l'avenir, ce type d'outil pourrait encore être amélioré pour évaluer plus précisément les coûts attendus, et même commencer à mobiliser les entreprises réparatrices (loueurs de motopompes et de groupes électrogènes, réparateurs de carrosseries de voitures…). Mais pour cela, il faudrait collecter des données encore plus précises sur les assurés: équipements, habite à tel étage, arbres sur son terrain, type de toiture, parking aérien ou souterrain, terrain en forte ou faible pente… Enfin, ce type d'outil pourrait aussi servir à alerter par mail ou SMS les assurés concernés et leur prodiguer des conseils de prévention personnalisés.

*« L'usage d'algorithmes de machine learning nous aidera à encore mieux assurer les biens de nos sociétaires »*

© Darri MAIF

**STÉPHANE RENOUX**
**responsable du programme Data & intelligence artificielle à la MAIF**

* Langage informatique dédié aux statistiques et à la science des données
** En mettant en images les données, la « dataviz » est un moyen d'identifier des tendances et corrélations




Datavisualisation du trafic TGV un lundi matin

© PEXELS

# Améliorer l'information temps réel

SNCF met au point des algorithmes de *big data* capables de prédire en temps réel l'affluence à bord d'un train, et l'heure prévisionnelle de ses futurs passages en gare en cas de situation perturbée. Explications avec Maguelonne Chandesris qui dirige l'équipe Innovation & Recherche « Statistique, Économétrie et Datamining » de SNCF.

**MAGUELONNE CHANDESRIS**

## *Des futurs du passé... au futur du présent*

### Quel est l'objectif principal de ce projet ?

**M. C.** Baptisé RELUANCE (qui contient « retards et affluence en avance »), ce projet vise deux grands objectifs : prédire en temps réel le niveau de confort à bord (par exemple si on trouvera une place assise) et à quelle heure est prévue l'arrivée d'un train dans telle ou telle gare en cas de situation perturbée... Si le second point intéresse l'ensemble des 5 millions de voyageurs qui empruntent nos trains chaque jour, le premier concerne ceux qui prennent des trains régionaux, sans réservation. Pour y parvenir, nous avons développé des algorithmes de *machine learning* déjà testés sur plus de 3000 trains en circulation. Chaque jour, ils s'alimentent de centaines de milliers de données, à partir desquelles ils fournissent des millions de prédictions, au fur et à mesure de la remontée en temps réel des informations !

### Sur quels types de données se basent-ils ?

**M. C.** Dans la pratique, nos algorithmes utilisent deux grands types de données. Tout d'abord celles qui concernent la localisation précise des trains en temps réel, collectée par des capteurs placés au sol et à bord des trains. Mais aussi des données sur le nombre de passagers qui montent et descendent des trains à chaque gare, obtenues par des capteurs installés à bord, au niveau des portes. Toutes ces données sont acheminées en temps réel, *via* des interfaces de programmation (API), vers nos algorithmes de calcul, hébergés sur nos serveurs situés près de la gare de Lyon.

### Comment fonctionnent ces algorithmes ?

**M. C.** Leur principe général est facilement compréhensible. Très concrètement, ils vont chercher dans l'historique de ce type de données les situations les plus semblables à la situation actuelle. Une fois celles-ci identifiées, ils s'appuient sur ce qui est advenu après ces situations semblables passées (les « futurs du passé »)... pour prédire ce qui va se passer maintenant (le « futur du présent »). Point important : ces algorithmes dits « d'apprentissage non-paramétrique » réalisent leur phase d'apprentissage en continu, sans aucun reparamétrage à effectuer, facilitant la maintenance industrielle. Au final, ils améliorent sensiblement les prédictions de l'heure d'arrivée des trains à destination, et une marge d'erreur sur l'affluence inférieure à vingt voyageurs... pour des trains pouvant transporter plus de neuf cents personnes.

### Quelles sont les prochaines étapes ?

**M. C.** Forts de ces résultats, nous avons déjà démarré les travaux d'industrialisation. Il nous faut notamment implémenter ces algorithmes de manière robuste dans « l'immeuble numérique » du groupe SNCF, afin que ces prédictions puissent alimenter tous nos systèmes d'analyse et de diffusion : affichages en gare, applications clients, centres opérationnels de supervision...
Ces informations en temps réel seront en effet aussi très utiles aux agents qui régulent la circulation des trains et les flux de passagers en gare. Au final, tous nos trains devraient bénéficier de ces prédictions algorithmiques à l'horizon 2020.

© SNCF

# « Et si vous deveniez “ data scientist ” ? »

Comme l'illustrent les trois précédents articles de ce cahier, les entreprises ont de plus en plus besoin de *data scientists*. Mais comment se former à cette nouvelle profession ? Le point sur les différentes voies pour y parvenir…

« *Métier le plus sexy du XXIe siècle* ». C'est en ces termes que la prestigieuse Harvard Business Review qualifie le job de *data scientist* ! Un métier en plein boom consistant à extraire de nouvelles connaissances à partir de données (*data*), grâce à des techniques issues des mathématiques appliquées, de la statistique et de l'informatique. En quelques années à peine, les formations se sont multipliées. Alors, comment s'y retrouver ?

Les formations de qualité remplissent généralement trois critères : elles sont souvent d'un niveau master, adossées à des laboratoires ou des chaires… et à un écosystème industriel. L'objectif étant de pouvoir ainsi adapter les enseignements au rythme des évolutions de la recherche et de l'industrie. On trouve de tels masters dans les universités et les grandes écoles, le plus réputé restant le MVA (Mathématiques, Vision, Apprentissage) de l'ENS Paris-Saclay… qui a fêté ses vingt ans en 2017 ! Certains sont plutôt à dominante « maths », d'autres plus à dominante informatique.

## Trois casquettes

Car dans la pratique, le métier de *data scientist* en regroupe plusieurs. Il y a bien sûr le concepteur d'algorithmes, nécessitant une solide formation générale en maths, en informatique scientifique et en modélisation. Mais il y a aussi l'« intégrateur » ou « développeur informatique » qui lui met en œuvre des algorithmes et des méthodes d'apprentissage, et doit gérer problèmes opérationnels et informatiques. Sans oublier le « spécialiste métier » ou « utilisateur » : véritable expert dans la manipulation de ces outils, et capable de comprendre le contexte d'utilisation (médical, marketing, transports…).

CEPE, École Polytechnique, TelecomParisTech, université Paris-Descartes… de nombreux organismes proposent aussi des formations continues qui se sont fortement développées ces dernières années, notamment dans le domaine du *big data*. Enfin, même s'ils ne remplacent pas les formations classiques, les cours en ligne ouverts à tous (Mooc) peuvent aussi se révéler très utiles.

« *Les* data scientists *devront de plus en plus se spécialiser.* »
**NICOLAS VAYATIS**
**Responsable du master MVA de l'ENS Paris-Saclay**

« *L'enjeu est à la fois de faire évoluer nos statisticiens en interne vers ces nouveaux métiers et de recruter en externe.* »
**OLIVIER BAES**
**Responsable du Datalab de la MAIF**

« *Nous avons créé avec l'École des ponts et chaussées la première Chaire en France regroupant optimisation et apprentissage.* »
**ISABEL GOMEZ GARCIA DE SORIA**
**Directrice de la Recherche Opérationnelle chez Air France**

« *Nous initions des concours en sciences des données depuis 2014.* »
**MAGUELONNE CHANDESRIS**
***Data scientist* chez SNCF Innovation & Recherche**

## En chiffres

- La France aura besoin de **130 000** *data scientists* d'ici à 2020
- La *data science* pourrait créer **3 millions d'emplois** en Europe
- Un *data scientist* gagne en moyenne de **45 000 à 55 000 euros** par an

### DATA CHALLENGES : DE VRAIS DÉFIS !

**Aujourd'hui, de nombreuses formations intègrent des « data challenges », compétitions dans lesquelles des équipes d'étudiants data scientists s'affrontent pour résoudre des problèmes très concrets : prévoir la consommation électrique d'une zone, la fréquentation d'une gare, l'intérêt de consommateurs pour tel ou tel produit, améliorer des algorithmes de recommandations d'achat en ligne, etc. Souvent proposés par des entreprises, y compris les grandes (Air France, SNCF…), ces *data challenges* permettent aussi à certains étudiants de se faire repérer pour décrocher un stage… voire un job !**

© Shutterstock

## L'ESSENTIEL

- Les données, de diverses natures et toujours plus abondantes, sont interrogées *via* des requêtes ou analysées pour y dénicher des corrélations, des indicateurs...
- Les applications nécessitent en amont des algorithmes qui optimisent la collecte, la gestion et le prétraitement des données.
- Ces traitements sont facilités par divers outils, tels que l'algèbre relationnelle, le calcul distribué, le modèle MapReduce...
- Au-delà du calcul proprement dit, la gestion des données passe aussi par leur indexation et leur protection.


## LES AUTEURS

MOKRANE BOUZEGHOUB
est professeur à l'université de Versailles et directeur adjoint scientifique de l'INS2I, au CNRS.

ANNE DOUCET
est professeure à l'université Pierre-et-Marie-Curie, déléguée scientifique à l'international de l'INS2I, au CNRS.


# Traiter les BIG DATA avec raffinement

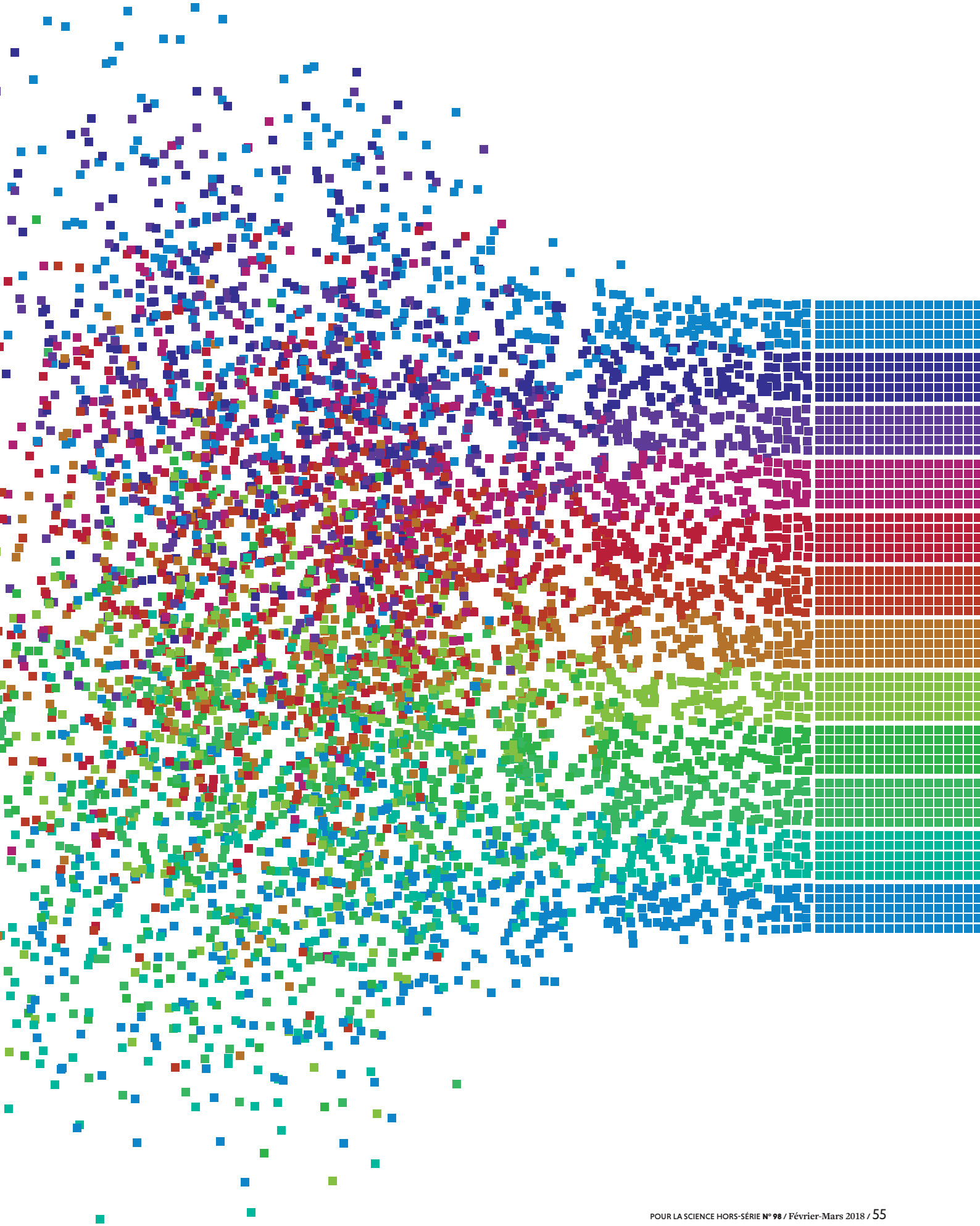
**La gestion des *big data*, de la collecte jusqu'à l'analyse, requiert de nombreux algorithmes. Tous sont confrontés à des problèmes de volume, d'hétérogénéité et de qualité des données... Comment relever le défi ?**

**Au sortir d'un puits, le pétrole est un mélange complexe de nombreux hydrocarbures, d'autres composés et d'impuretés.** Pour être utile, ce cocktail doit être raffiné de façon à séparer les produits exploitables: bitume, fioul, kérosène, gazole, essence... De même, les données réunies sous le terme de *big data* subissent un long processus de raffinage, de sélection et de transport avant de pouvoir livrer une quelconque information, c'est-à-dire d'être interrogées ou analysées. En quoi consistent ces traitements?

Lorsqu'on évoque les *big data*, on pense aussitôt à l'analyse de données plus qu'à leur exploration par des requêtes, comme si le volume croissant des données, en atteignant un certain niveau, rendait caduque leur interrogation directe. En réalité, ce sont deux approches complémentaires, toutes deux contraintes par les limites technologiques dues au traitement massif des données. Qui plus est, l'analyse au sens large nécessite souvent une exploration et un prétraitement individualisé des données. La chaîne de valorisation des *big data* est longue et semée d'embûches, la partie analyse ne représentant qu'un petit maillon de l'ensemble en bout de chaîne.

Les algorithmes sous-jacents à la collecte des données, leur indexation, leur nettoyage, leur stockage, leur annotation... peu visibles de l'extérieur, constituent pourtant l'ossature du pipeline >

Les données disparates doivent rentrer dans le rang pour être exploitables. C'est le rôle des algorithmes.



> qui alimente les applications de décision ou d'apprentissage. Qu'il s'agisse d'algorithmes simples ou complexes, déterministes ou prédictifs, opérant sur des données exactes ou bruitées, leur compréhension est indispensable pour entrevoir en quoi consiste l'exploitation des *big data*.

Occultés par les succès récents de l'apprentissage machine et autres programmes de jeux, ces algorithmes enfouis sont donnés pour acquis, avec un coût nul pour leur mise en œuvre et leur exécution. Nous verrons que ce n'est pas le cas, un travail permanent de recherche et d'ingénierie est nécessaire pour les adapter aux nouveaux contextes applicatifs, les étendre, les optimiser et en inventer de nouveaux.

Il y a autant de processus de gestion et de valorisation des données que de raisons de valoriser ces données. On y retrouve cependant de grandes activités génériques (*voir l'encadré page ci-contre*), sur chacune desquelles existent des verrous importants.

Pour réaliser ces grandes activités, il est important de spécifier les tâches constitutives en fonction des applications visées. La consultation intensive des données n'a pas les mêmes besoins qu'un traitement des flux de données en temps réel. Une application d'aide à la décision n'a pas les mêmes besoins qu'une application de prédiction du panier d'un consommateur ou d'aide au diagnostic d'une tumeur. La nature de l'application et des tâches algorithmiques qui en découlent doivent être mises en perspective avec le modèle de calcul le plus adapté.

## DEUX MODÈLES DE CALCUL

La gestion des données n'est pas constituée d'algorithmes *ad hoc* réalisés au coup par coup, dans un langage donné, en fonction des applications. C'est une science qui a ses fondements théoriques, ses propres objets d'étude, ses propres abstractions et ses propres méthodes. Les technologies produites ont pour vocation de réduire l'effort nécessaire à la gestion des données, de contrôler l'accès à ces données et de faciliter l'évolution des systèmes sans réécriture de leur code. Pour simplifier, il y a deux modèles de calcul sur les grands volumes de données: celui dédié à la recherche d'information et celui adapté à l'analyse de données. Même si les deux approches partagent de nombreux concepts et techniques et peuvent se combiner, elles diffèrent d'un point de vue sémantique.

Dans le premier cas, le sens est porté par la requête de l'utilisateur; si la requête est pertinente, les données délivrées seront bien conformes aux critères de recherche spécifiés. Par exemple, «rechercher les références de produits vendus en quantité supérieure à 200 dans chacun des magasins de la région parisienne durant la période de Noël», est une requête complexe, mais qui définit clairement le calcul à faire et les critères de sélection de données. Ainsi spécifié, le système de gestion de données se chargera de générer le code nécessaire à l'exécution de cette requête. D'autres requêtes sont plus approximatives, leurs critères étant spécifiés de façon lâche. C'est le cas des recherches sur le Web. Ainsi, «rechercher les magasins parisiens pratiquant les prix les moins chers pour le chocolat et le champagne durant la période de Noël» est une requête floue. Elle impose un traitement approché du critère «moins cher» obligeant à présenter les résultats dans un certain ordre.

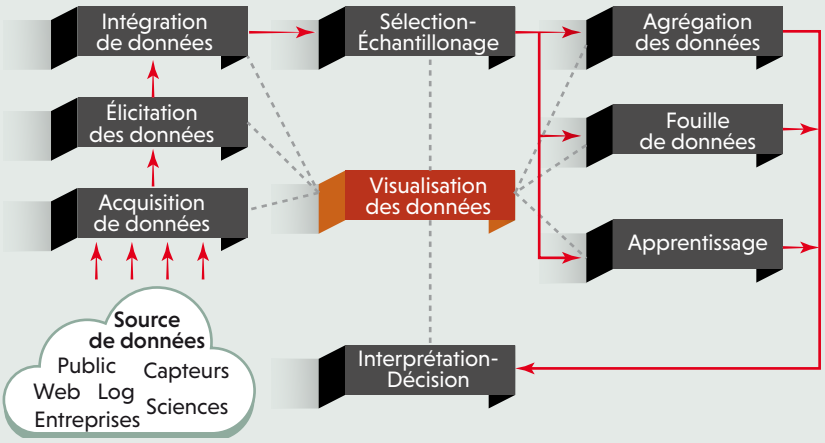
Dans le second cas, la question du sens est plus complexe, car elle interroge à la fois la méthode utilisée pour l'analyse des données et l'interprétation par l'utilisateur des résultats produits. Pour une requête d'analyse de l'évolution des cours de bourse des entreprises du CAC40 ou de prédiction des ventes de smartphones par région et par période de l'année, les résultats ne sont pas interprétables sans connaissance du métier et des modèles de prédiction sous-jacents.

Ce problème peut devenir plus aigu dans des algorithmes encore moins transparents. Par exemple, dans un système de reconnaissance fondé sur l'apprentissage, la confusion entre un chat et un caniche fait sourire, mais celle d'un honnête citoyen avec un criminel peut être dramatique. Les algorithmes sous-jacents à ces systèmes posent des défis à la société.

**Il y a autant de processus de valorisation des données que de raisons de valoriser ces données**

D'un point de vue conceptuel, un modèle de calcul est défini par les structures de données qu'il reconnaît et les opérations qu'il offre sur ces structures de données. Les structures de données peuvent être des tables (ou matrices), des arbres, des graphes ou des séquences finies ou infinies. Les opérations sont associées à chaque structure: les opérations de tri, projection, produit et union sur

## LES MILLE ET UNE VIES DES DONNÉES




Le cycle de vie des données au terme duquel elles peuvent livrer des informations.

**Le traitement des *big data* implique des activités que l'on retrouve quelle que soit l'application. Passons-les en revue. Le stockage pose la question du choix des données brutes à enregistrer, de leur acquisition et des données à résumer, ainsi que les méthodes d'indexation pour accélérer leur recherche ultérieure. La validation des données, c'est-à-dire leur nettoyage et leur annotation (on parle d'élicitation), nécessite souvent une interaction avec un expert, ce qui est hors de portée d'un humain si le volume de données est considérable. L'intégration de données est une phase très complexe car elle concerne des sources de données distribuées à grande échelle où l'hétérogénéité est très élevée. Le modèle d'intégration dépendra des objectifs assignés au système : la recherche d'information, la prise de décision, l'extraction de connaissances… C'est là que se situent les principaux verrous liés aux performances des systèmes et à la pertinence des données produites. La sélection et l'échantillonnage constituent des vues sur les données intégrées, c'est-à-dire une façon de réduire les données aux sous-ensembles nécessaires à un type de requête ou un type d'analyse. Un mauvais choix de ce sous-ensemble hypothèque la qualité de la tâche ultérieure. Quant à l'analyse elle-même, elle peut être réalisée par différentes méthodes : l'agrégation pour générer des indicateurs statistiques (somme, moyenne, tendance…), la fouille de données pour identifier des règles d'association ou des motifs fréquents, l'apprentissage automatique (profond ou par renforcement) pour définir un modèle de prédiction permettant de décider ultérieurement si un nouvel objet peut être classé de façon pertinente. Enfin, la visualisation des données est utile à toutes les étapes. Elle offre souvent une vue synthétique des grandes données et aide à leur compréhension.**

les tables ; la recherche d'éléments ou le parcours de chemin sur un arbre ; l'appariement ou l'extraction de sous-graphes ; une copie instantanée d'une séquence. Ces structures et ces opérations ont été étudiées depuis des décennies pour formaliser leurs sémantiques, expliciter leurs propriétés et trouver les meilleurs programmes qui leur correspondent dans différents environnements.

Plusieurs facteurs interviennent dans leur mise en œuvre : la hiérarchie et la taille des mémoires (mémoire principale, mémoire cache, disque…), la localisation des données (un site centralisé, de multiples serveurs…), les ressources de calcul mobilisables (une ou plusieurs machines…) et les caractéristiques techniques de ces ressources (puissance de calcul, partage de mémoire…).

### DES TABLEAUX MALLÉABLES

L'algèbre relationnelle a représenté un saut technologique et qualitatif dans la gestion des données. En quoi consiste-t-elle ? Rappelons d'abord qu'une table relationnelle est décrite par un ensemble de variables (ou attributs) désignant les colonnes et un ensemble de valeurs à ces variables organisées en lignes, chacune représentant un objet du réel. En d'autres termes, il s'agit de tableaux à deux dimensions, on parle aussi de tableau à double entrée. Or avec cinq opérateurs de base (*voir la figure page suivante*), on sait aujourd'hui transformer ces tableaux et en dériver d'autres.

C'est d'une simplicité confondante, et pourtant, la combinaison de ces opérations permet d'exprimer une large palette de requêtes, sans programmer une seule ligne de code, puisque le code correspondant à chacune des opérations est défini une fois pour toutes et pour toutes les applications. Cette façon de représenter le réel par un ensemble de tableaux indépendants et d'exprimer une requête sur ces tableaux au moyen d'une expression algébrique a notablement modifié notre rapport à la programmation.

Parmi ces opérateurs, certains sont plus sensibles que d'autres au volume des données. Ainsi la jointure de deux tables comportant des milliards de lignes pose de réels problèmes de calcul pour comparer ces milliards de lignes. De nombreux travaux de recherche proposent des algorithmes de jointure tenant compte ou non du tri des données, de la taille de la mémoire centrale, de l'indexation des données, de la distribution des données sur différents disques, du parallélisme des machines et du débit sur les réseaux.

Ce coût dépend aussi du choix initial de la représentation d'une table selon ses lignes ou ses colonnes. Par exemple, pour les requêtes >

PROJECTION

| MAGASIN | PRODUIT | PRIX | DATE | SITE |
|---|---|---|---|---|
| Auchan | Iphone | 700 | 12/9/2017 | Paris Est |
| Auchan | Iphone | 760 | 12/9/2017 | Paris Est |
| Leclerc | IPad | 480 | 27/9/2017 | Essonne |
| Fnac | IPad | 870 | 30/9/2017 | La Défense |
| Carrefour | PC Dell | 600 | 24/9/2017 | Paris Ouest |
| Leclerc | Galaxy | 400 | 28/9/2017 | Essonne |
| Boulanger | Galaxy | 310 | 4/10/2017 | Plaisir |
| Carrefour | Galaxy | 470 | 4/11/2017 | St. Quentin |
| Leclerc | MacBookAir | 970 | 7/10/2017 | Yvelines |
| Auchan | MacBookAir | 1100 | 7/10/2017 | Yvelines |
| Darty | PC Dell | 860 | 19/10/2017 | Velizy |
| Darty | PC Dell | 640 | 23/10/2017 | Parly |
| Darty | iPad | 480 | 27/10/2017 | Parly |

RESTRICTION

| MAGASIN | PRODUIT | PRIX | DATE | SITE |
|---|---|---|---|---|
| Auchan | Iphone | 700 | 12/9/2017 | Paris Est |
| Auchan | Iphone | 760 | 23/9/2017 | Paris Est |
| Carrefour | PC Dell | 600 | 24/9/2017 | Paris Ouest |
| Leclerc | IPad | 400 | 27/9/2017 | Essonne |
| Leclerc | Galaxy | 400 | 28/9/2017 | Essonne |
| Fnac | IPad | 870 | 30/9/2017 | La Défense |
| Fnac | iMac | 1220 | 1/10/2017 | La Défense |
| Fnac | iMac | 2100 | 3/10/2017 | Montparnasse |
| Boulanger | Galaxy | 310 | 4/10/2017 | Plaisir |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |

| PRODUIT | MARQUE | BREVET |
|---|---|---|
| Iphone | Apple | AP2056XY |
| PC Dell | Dell | DL3467GH |
| Galaxy | Samsung | SM734DG |
| IPad | Apple | AP205VY |

JOINTURE

| MAGASIN | PRODUIT | PRIX | DATE | SITE | MARQUE | BREVET |
|---|---|---|---|---|---|---|
| Auchan | Iphone | 700 | 12/9/2017 | Paris Est | Apple | AP2056XY |
| Auchan | Iphone | 760 | 23/9/2017 | Paris Est | Apple | AP2056XY |
| Carrefour | PC Dell | 600 | 24/9/2017 | Paris Ouest | Dell | DL3467GH |
| Leclerc | IPad | 400 | 27/9/2017 | Essonne | Apple | AP205VY |
| Leclerc | Galaxy | 400 | 28/9/2017 | Essonne | Samsung | SM734DG |
| Fnac | IPad | 870 | 30/9/2017 | La Défense | Apple | AP205VY |
| Boulanger | Galaxy | 310 | 4/10/2017 | Plaisir | Samsung | SM734DG |
| ... | ... | ... | ... | ... | | |
| ... | ... | ... | ... | ... | | |

**L'algèbre relationnelle offre cinq opérateurs de base grâce auxquels on peut exprimer un grand nombre de requêtes sur des données. À partir d'une table relationnelle *(le tableau bleu)*, la projection réduit le nombre de colonnes, la restriction celui des lignes, l'union fusionne deux tables de même structure, l'intersection calcule les lignes communes à deux tables de même structure, la jointure apparie deux tables sur un ou plusieurs critères.**

> d'exploration de données, les tables seront rangées par lignes alors que pour les requêtes d'analyse agrégeant les valeurs, la représentation par colonne est plus adaptée.

Dans le même élan de définition de ce modèle élégant, des extensions ont été apportées pour traiter des objets plus complexes que les tables relationnelles, notamment les tables imbriquées et les tables à plus de deux dimensions. L'algèbre relationnelle a également été étendue à d'autres domaines spécialisés grâce à l'introduction, par exemple, d'opérations géométriques ou d'autres sur des données temporelles. Le champ d'investigation du modèle relationnel est loin d'être couvert !

## DES ARBRES ET DES GRAPHES

Avec le Web, les bases de données ont évolué pour intégrer des données XML dont la structure est formalisée par des arbres. Une algèbre est aussi associée à ces arbres pour effectuer des recherches d'éléments et pour transformer leurs structures. La sélection et la projection permettent d'extraire des sous-arbres vérifiant certaines conditions portant soit sur les données (les valeurs d'un élément), soit sur les métadonnées (les éléments descriptifs de ces valeurs). Il est possible également de restructurer un arbre, ou de l'aplatir, pour obtenir une organisation différente des données et les visualiser sous des angles variés.

Enfin, l'appariement permet de comparer deux structures, une opération très utile pour de nombreuses applications, comme la bioinformatique, l'analyse d'images, les bases de documents... On associe à ce type d'opération une mesure de similarité qui peut être le nombre de transformations à effectuer pour obtenir un arbre à partir d'un autre. Plus la distance est petite, plus les arbres sont similaires. Cette notion est cruciale dans le contexte des *big data*. En effet, le calcul de ces distances est un problème difficile, dont la complexité augmente fortement avec le volume des données et avec la diversité structurelle des graphes à comparer. Les données dont nous avons parlé sont durables, mais qu'en est-il des flux de données, qui disparaissent juste après leur observation ?

## LES FLUX DE DONNÉES

La logique de gestion est ici bien différente. Cette fois, ce sont les requêtes qui sont persistantes. Elles sont définies une fois pour toutes et agissent comme des filtres, déterminant les données observées qu'on souhaite conserver. Le premier opérateur indispensable sur un flux définit la taille de la fenêtre d'observation et un pas de glissement pour suivre le flux. Ces paramètres peuvent être définis par des durées, par le nombre d'événements attendus ou la combinaison des deux. Les données observées peuvent être stockées temporairement dans des tables relationnelles et filtrées par des requêtes définies en amont.

Pour faciliter l'interprétation des flux, il est souvent nécessaire de les corréler avec des données de référence persistantes dans des catalogues, ce qui impose des opérations de jointures potentiellement coûteuses si ces catalogues sont gigantesques. Par exemple, l'observation d'une température de −110 °C dans un bâtiment est sans doute liée à la défaillance d'un capteur. En revanche, une forte température trahit sans