BI4people[1] – Le décisionnel pour tous

Jérôme Darmont
Université Lumière Lyon 2, ERIC

Enjeux et état de l'art

Les technologies de l'informatique décisionnelle (*Business Intelligence*, BI), telles que les entrepôts de données, qui permettent de collecter, stocker et structurer des données opérationnelles en données décisionnelles, ainsi que l'analyse en ligne (*On-Line Analytical Processing*, OLAP) qui aide à croiser plusieurs axes de données, sont des outils primordiaux dans l'aide à la décision dans de nombreux domaines comme la finance, mais aussi la santé ou l'enseignement supérieur [6, 8].

Les outils de l'informatique décisionnelle ont longtemps nécessité un investissement financier et humain très lourd. Toutefois, il existait au début du projet BI4people de nombreuses solutions de BI gratuites, qu'elles fussent propriétaires, libres ou infonuagiques [1]. Les logiciels propriétaires se focalisaient cependant sur les tableaux de bord et la visualisation, et avaient tous des fonctionnalités limitées, comme l'absence d'une intégration efficace de données depuis des sources disparates.

Bien que quelques logiciels libres proposassent des explorations OLAP, ils demeuraient techniquement hors de portée des petites entreprises, des associations, des chercheurs, des indépendants comme des journalistes ou des *makers*, et des citoyens actifs [2], que nous ciblions particulièrement dans le projet BI4people.

De plus, la tendance à déporter la BI dans le nuage avait rejoint la demande grandissante d'outils collaboratifs permettant aux usagers et usagères de croiser des données privées, publiques, ainsi que des *self data*,

1. https://eric.univ-lyon2.fr/bi4people/.

DOI:10.48556/SIF.1024.26.59

d'effectuer des analyses conjointes, d'annoter des figures ou des rapports et de communiquer via les réseaux sociaux [10]. Les réponses à l'époque à cette demande globale restaient en deçà des attentes en se limitant au partage en ligne de résultats d'analyse.

Objectifs

L'objectif de BI4people était de rendre accessible la puissance de l'analyse interactive OLAP à la plus large audience possible, en mettant en œuvre le processus d'entreposage de données en mode *software-as-a-service*, de l'intégration de données multisource, hétérogènes (typiquement sous la forme de tableaux issus de tableurs, de documents textuels ou semi-structurés, ou encore du Web) à une analyse OLAP et une visualisation très simples (figure 1).

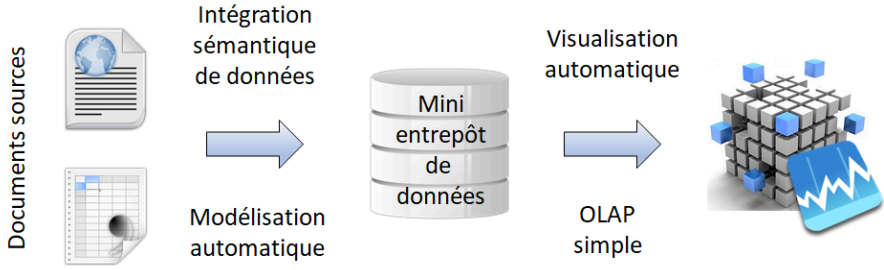

Fig. 1. Processus global.

Pour atteindre ce but, le service de BI devait inclure la *privacy by design*, être autonome, extrêmement simple, ergonomique et intelligible (jargon informatique ou BI interdit !). Dans ce contexte, les étapes classiques de l'entreposage de données s'appliquaient toujours, mais devaient être complètement automatisées [9, 12]. Au début du projet, BI4people était la première plateforme à atteindre complètement ce but, les projets similaires relevant plutôt de l'apprentissage automatique.

De plus, le prototype logiciel que nous proposions comme le délivrable principal du projet prenait en compte la confidentialité des données dans toutes les étapes, permettait des analyses collaboratives et était intelligible par ses usagers. Nous avions en effet insisté sur l'importance de l'appropriation des visualisations fournies par l'outil par les usagers, ce qui a impliqué une collaboration interdisciplinaire entre l'informatique (CNU 27) et les sciences de l'information et de la communication (CNU 71).

Enfin, il faut souligner que l'évaluation de nos prototypes par les usagers a été mise en œuvre tout au long du projet et pas uniquement à la fin.

Afin de mener tous ces travaux, nous avons constitué le consortium suivant :

– équipe de recherche de Lyon en sciences de l'information et de la communication (ELICO, appropriation des outils auprès des usagers) ;
– ERIC (informatique / BI collaborative, sécurité, Lyon) ;
– institut de recherche en informatique de Toulouse (IRIT, informatique / entrepôts de données) ;
– laboratoire d'informatique (LIFAT / visualisation, Tours) ;
– société TRIMANE (Saint-Germain-en-Laye, informatique / intégration, tests) ;

ainsi que des partenaires associés :

– *think-tank* FING, Marseille-Paris (jusqu'au 27/04/2022) ;
– TUBÀ Lyon (*living lab*) ;
– métropole de Lyon ;
– UrbaLyon.

Méthodes et approches

La méthodologie de projet que nous avons adoptée s'inspirait des méthodes agiles ou des *living labs*, mais sur des périodes plus longues, adaptées au rythme de la recherche. Concrètement, nous avions prévu de construire rapidement un premier prototype imparfait, puis de le perfectionner dans deux ou trois versions successives. L'idée était de confronter ce logiciel à différentes catégories d'usagers (des petites entreprises clientes du partenaire TRIMANE et des citoyens actifs recrutés par le TUBÀ), afin d'exploiter leurs retours et d'améliorer le prototype de manière incrémentale. Malheureusement, la pandémie de Covid-19 a bouleversé substantiellement le rythme du projet et nous n'avons pu avancer qu'une seule version, et avec moins de testeurs que prévu.

Néanmoins, nous avons conduit simultanément une étude d'utilité et d'appropriation du prototype par les usagers ainsi que leur évolution dans le temps, afin de mesurer si nos efforts allaient dans la bonne direction et amélioraient « l'expérience utilisateur ».

D'un point de vue opérationnel, le projet est subdivisé entre huit lots de travaux (*work packages*, WP), eux-mêmes scindés en sous-tâches, avec une prise en compte des risques potentiels et des solutions de repli (figure 2).

– WP1 — Coordination du projet ;
– WP2 — Automatisation de l'entreposage des données ;
– WP3 — Analyse collaborative des données ;
– WP4 — Visualisation et exploration des données ;

– WP5 — Protection des données ;
– WP6 — Validation et expériences ;
– WP7 — Évaluation de l'appropriation des usagers / usagères ;
– WP8 — Dissémination et exploitation.

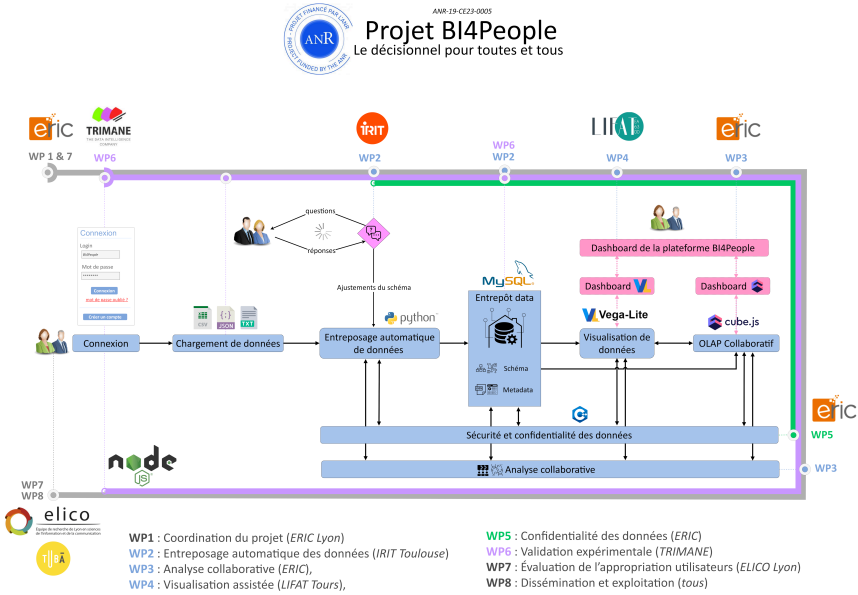

Fig. 2. Schéma global du consortium et des tâches du projet.

Nous avions prévu au sein du WP1 une approche intégrée avec tous les partenaires du projet, afin de prendre en compte pleinement les enjeux transversaux du projet, ainsi que de travailler sur la base d'un cas d'usage structurant. Toutefois, la temporalité des recrutements, les retards dus à la pandémie de Covid-19 et la dissolution de la FING (partenaire associé pourvoyeur de données) en 2022, nous ont forcé à nous adapter et à trouver des données de test plus diverses.

Résultats et faits marquants

Dans cette section, nous résumons les résultats principaux du projet BI4people, WP par WP.

WP1 — Coordination

Ce projet ambitieux traitait de différents volets informatiques, correspondant à des thématiques variées (entreposage de données, visualisation de données, sécurité, analyse collaborative) avec une collaboration avec les sciences de l'information et de la communication pour aborder la question de l'appropriation par les usagers. Il s'avère que chaque partie a pu aboutir à des résultats qui sont décrits dans ce qui suit en fonction des WP. Globalement, les objectifs scientifiques du projet ont été atteints, malgré quelques difficultés (cf. infra).

WP2 — Automatisation de l'entreposage

Nous avons conçu et implémenté une chaîne logicielle largement automatisée permettant de transformer des données tabulaires en entrepôt de données qualifiées [15, 16].

WP3 — Analyse collaborative

Nous avons conçu et implémenté deux prototypes permettant de :

1. enregistrer une démarche d'analyse pour un cas, diffusée ensuite aux autres personnes et qui enrichit la démarche de manière itérative [11] ;
2. utiliser un agent conversationnel dédié à l'aide à l'analyse [5].

WP4 — Visualisation

Nous avons créé un système de recommandation réalisant l'optimisation globale de tableaux de bords par un algorithme génétique [13, 14].

WP5 — Sécurité

Nous avons mené une étude des systèmes de chiffrement homomorphique, qui permettent de faire des calculs directement sur des données cryptées, protocoles de calcul sécurisés multipartites[2] pour sécuriser des cas d'usage [7].

WP6 — Validation

Nous avons mené une évaluation des parties développées. Chaque partie a été expérimentée et testée indépendamment. Les résultats obtenus sont encourageants pour une validation ultérieure plus complète.

2. Cela permet à plusieurs parties de faire un calcul sur leurs données respectives, sans que celles-ci ne puissent être divulguées aux autres parties.

WP7 – appropriation des usagers / usagères

Nous avons mené une étude des modalités d'appropriation de la dataviz à partir de ses prétentions communicationnelles, ses prédicats sémiotiques et les registres appréciatifs qui leurs sont associés par des usagers non-experts [3, 4].

WP8 – Dissémination

Le projet BI4people a produit 27 publications scientifiques, 2 thèses de doctorat, 5 mémoires de master, 3 vidéos, 1 poster ; 3 manifestations scientifiques ont été organisées, 4 présentations du projet ont été effectuées, 6 prototypes logiciels ont été réalisés et 1 jeu de données a été établi. Tous ces résultats sont libres de droits et déposés sur la plateforme HAL[3].

Perspectives générales

Globalement, aider l'analyse en permettant la production de résultats pour des usagers novices qui collaborent constitue une étape importante. Pour parfaire l'objectif d'accessibilité, il s'agit ensuite de soutenir l'étape d'interprétation des résultats. Ceci pourrait passer par une génération de l'analyse automatique ou semi-automatique. Une vigilance dans la recherche d'une telle solution est de présenter l'interprétation de telle ou telle partie d'une visualisation, qui n'est pas neutre, et qu'une interprétation plus globale est nécessaire. Une démarche collaborative trouverait également sa place pour parfaire l'interprétation des résultats et construire une connaissance commune sur les données traitées.

Finalement, l'interprétation de l'analyse rendrait possible d'aller plus loin en termes d'inclusion, notamment par rapport au handicap visuel. Les approches génératives de descriptions d'images sont un support important, mais cette perspective nécessiterait un travail pluridisciplinaire important pour prendre en compte les besoins des usagers.

Aider l'analyse en permettant la production de résultats pour des usagers novices qui collaborent constitue une étape importante. Pour parfaire l'objectif d'accessibilité, il s'agit ensuite de soutenir l'étape d'interprétation des résultats. Ceci pourrait passer par une génération de l'analyse (semi-)automatique. Une vigilance dans la recherche d'une telle solution est de présenter l'interprétation de telle ou telle partie d'une visualisation, qui n'est pas neutre, et qu'une interprétation plus globale est nécessaire. Une démarche collaborative trouverait également sa place pour parfaire l'interprétation des résultats et construire une connaissance commune sur les données traitées. Finalement, l'interprétation de l'analyse rendrait possible

3. https://hal.science/.

d'aller plus loin en termes d'inclusion, notamment par rapport au handicap visuel. Les approches génératives de descriptions d'images sont un support important, mais cette perspective nécessiterait un travail pluridisciplinaire important pour prendre en compte les besoins des usagers.

Enfin, nous souhaitons terminer sur le thème de l'interdisciplinarité, que nous avons voulue dès le début du projet. Sur la base de ce cheminement et de réflexions partagées, nous esquissons un protocole méthodologique qui se décline en trois phases.

1. Des enjeux et des temporalités différentes au sein de chaque WP.
2. Un sens consacré aux données différent au sein des sciences informatiques et des sciences de l'information et de la communication (SIC).
3. Quel sens attribuer aux données par les usagers ?
   – importance du « contexte d'usage » dans le processus d'appropriation ;
   – quelle place donner à l'usager dans la coconception des data-visualisation pour favoriser la « capacitation » ?

Références

[1] Top 30 open source and free business intelligence software. 2019. PAT RESEARCH, `https://www.predictiveanalyticstoday.com/open-source-freebusiness-intelligence-solutions/`.

[2] A. Abello, J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazon, F. Naumann, T.-B. Pedersen, S. Rizzi, J. Trujillo, P. Vassiliadis, and G. Vossen. 2013. Fusion Cubes: Towards Self-Service Business Intelligence. *International Journal of Data Warehousing and Mining* 9, 2: 66–88.

[3] T. Andry, J. Bonaccorsi, and F. Labarthe. 2022. Le décisionnel pour toutes et tous? In *Colloque Intersections du design - 3e édition - Le design dans la démocratie, Montréal (Québec), Canada*, 196–211.

[4] T. Andry, J. Bonaccorsi, and F. Labarthe. 2023. Le décisionnel pour toutes et tous? Retour sur les ambitions transformatrices d'un projet de recherche visant la démocratisation d'un outil d'analyse des données numériques. In *Intersections du design 2022 : Le design dans la démocratie, Montréal, Canada*, 212–227.

[5] O. Cherednichenko, F. Muhammad, J. Darmont, and C. Favre. 2023. A Reference Model for Collaborative Business Intelligence Virtual Assistants. In *6th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2023), Kharkiv, Ukraine* (CEUR), 114–125 vol. III.

[6] J. Darmont et P. Marcel. 2017. Entrepôts de données et OLAP, analyse et décision dans l'entreprise. CNRS Editions, Paris, 132–133.

[7] T.-V.-T. Doan, M.-L. Messai, G. Gavin, and J. Darmont. 2023. A Survey on Implementations of Homomorphic Encryption Schemes. *The Journal of Supercomputing* 79: 15098–15139.

[8] C. Favre, F. Bentayeb, O. Boussaïd, J. Darmont, G. Gavin, N. Harbi, N. Kabachi, and S. Loudcher. 2013. Les entrepôts de données pour les nuls… ou pas ! In *2e Atelier aIde à la Décision à tous les Etages (EGC/AIDE 13), Toulouse*.

[9] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter. 2018. Practical Automated Machine Learning for the AutoML Challenge 2018. In *ICML 2018 AutoML Workshop*.

[10] J. Kaufmann and P. Chamoni. 2014. Structuring Collaborative Business Intelligence: A Literature Review. In *In Proc. HICSS 2014*, 3738–3747.

[11] F. Muhammad and J. Darmont. 2023. An Ontology-based Collaborative Business Intelligence Framework. In *12th International Conference on Data Science, Technology and Applications (DATA 2023), Rome, Italy*, 480–487.

[12] L. De Raedt. 2016–2021. Synth: Synthesising Inductive Data Models.

[13] P. Soni, C. de Runz, F. Bouali, and G. Venturini. 2023. A genetic algorithm for automatic dashboard generation: first results. In *27th International Conference Information Visualisation (IV 2023), Tempere*, 77–82.

[14] P. Soni, C. de Runz, F. Bouali, and G. Venturini. 2024. A survey on Automatic Dashboard Recommendation Systems. *Visual Informatics*: Journal pre-proof.

[15] Y. Yang, F. Abdelhédi, J. Darmont, F. Ravat, and O. Teste. 2022. Automatic Machine Learning-based OLAP Measure Detection for Tabular Data. In *24th International Conference on Big Data Analytics and Knowledge Discovery (DaWaK 2022), Vienna, Austria* (Lecture Notes in Computer Science), 173–188.

[16] Y. Yang, J. Darmont, F. Ravat, and O. Teste. 2021. An Automatic Schema-Instance Approach for Merging Multidimensional Data Warehouses. In *25th International Database Engineering and Applications Symposium (IDEAS 2021), Montreal, Canada* (ICPS), 232–241.