Année universitaire 2021 - 2022

Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J
2 Rue du Doyen-Gabriel-Marty
31042 TOULOUSE

UNIVERSITÉ TOULOUSE 1 CAPITOLE
FACULTÉ D'INFORMATIQUE

MASTER MIAGe IPM
L'INFORMATION MAITRISÉE

MEMOIRE DE STAGE

Développement de l'application pour l'entreposage automatique et l'imputation de données

Stage effectué au sein de la société IRIT

(du 23/05/2022au 16/12/2022)

Haoyang YU

Mémoire soutenu à l'université le 15/11/2022

Encadrant : M. Yuzhao YANG

Tuteur : M. Alain BERRO

Titre de stage : Développement de l'application pour l'entreposage automatique et l'imputation de données

Nom de l'entreprise : IRIT (Institut de recherche en Informatique de Toulouse)

Equipe : SIG (Systèmes d'Informations Généralisées)

Période du stage : 23/05/2022 – 22/07/2022 et 22/08/2022 - 16/12/2022

Nom stagiaire : Haoyang YU

Date de soutenance : 22/11/2022

Encadrant : M. Yuzhao YANG

Tuteur : M. Alain BERRO

Résumé

Pendant mon stage, j'ai intégré l'équipe du projet BI4people, qui visait à développer un système de business intelligence (BI) pour les non-spécialistes des petites et moyennes entreprises.

L'objectif de ce stage est de concevoir et de développer une application web pour l'enregistrement automatisée d'un entrepôt de données à partir des données tabulaires et imputation des données. Au cours de ce stage, j'ai d'abord mis en œuvre quatre méthodes d'apprentissage automatique pour imputer les données manquantes. J'ai ensuite fait des expérimentations d'imputation sur quatre jeux de données différents pour la partie de l'expérimentation de M. Yuzhao YANG. De plus, j'ai ensuite développé une application web permettant d'automatiser entreposage et d'imputer les données manquantes dans les entrepôts de données. L'application démontre les résultats des recherches de M. Yuzhao YANG et facilite également le processus d'entreposage.

Grâce à ce stage, j'ai mis en pratique ce que j'ai appris pendant mon master, telles que la gestion de projet, la modélisation de bases de données et le développement d'applications. Cela m'a aidé à m'adapter rapidement à un nouvel environnement technique et à apprendre de nouvelles connaissances. Ce stage m'a aussi permis de mieux comprendre l'industrie du doctorat et le travail des développeurs, et m'a permis de faire de grands progrès dans le travail avec les autres et la programmation.

Remerciements

Je tiens à remercier toutes les personnes qui m'ont aidé et accompagné au cours de ces six mois.

Tout d'abord, je voudrais remercier M. Franck RAVAT et M. Yuzhao YANG de l'équipe SIG de l'IRIT pour m'avoir offert ce stage et pour leurs conseils professionnels et leur patience tout au long du stage. Cela m'a aidé à m'adapter à l'environnement de travail et à réaliser mes missions de stage le plus sereinement possible.

Ensuite, je voudrais remercier Mme. Yuni CHEN, les membres du projet BI4PEOPLE pour leurs conseils techniques, leur soutien et leurs encouragements, qui m'ont donné la confiance nécessaire pour mener à bien mon stage.

Enfin, je tiens à remercier mon tuteur universitaire, M. Alain BERRO. Il m'a posé des questions qui m'ont permis de réfléchir plus clairement à ce que le stage m'a apporté et de mieux présenter les résultats de mon stage.

Table des matières

Introduction

Dans le cadre de la validation de ma deuxième année de master MIAGE parcours Ingénierie des Processus Métier (IPM) de l'Université Toulouse I Capitole, j'ai effectué un stage de six mois en tant que data ingénieur au sein de l'équipe SIG du laboratoire IRIT du 23 mai au 16 décembre 2022.

Ce stage a été effectué dans le cadre du projet de recherche BI4PEOPLE, qui vise à développer un système de business intelligence utilisable par des non-experts, à promouvoir le développement de business intelligence dans les petites et moyennes entreprises (PME) et les aider à réduire le coût humain et financier de la technologie.

Le stage était divisé en deux parties principales : une partie consiste à aider M. Yuzhao YANG sur ses expérimentations d'imputation de données. Un soutien technique a été fourni pour l'étude du complément des données manquantes dans l'entrepôt de données. L'autre partie consiste à concevoir et à développer une application web pour l'entreposage automatique de données tabulaires. Elle dispose de fonctions telles que l'entreposage automatique de données, la fusion de l'entrepôt de données et l'imputation de données afin de visualiser les résultats de la recherche de M. Yuzhao YANG.

Au cours du stage, j'ai également rencontré quelques enjeux pour réaliser mes missions. Par exemple, l'enjeu technique de l'apprentissage et de mise en œuvre de nouveaux langages de programmation que je n'avais pas utilisés auparavant, tels que Python et Node.js. En outre, il y avait l'enjeu de la connaissance de l'apprentissage automatique, qui nécessite d'apprendre et de comprendre les algorithmes pertinents et de lire des articles scientifiques pour mettre en place des expérimentations.

À la fin de mon stage, j'ai mis en œuvre quatre algorithmes d'apprentissage automatique différents pour des expérimentations d'imputation de données et j'ai acquis une meilleure compréhension de l'apprentissage automatique. De plus, j'ai conçu et développé une application web d'entreposage automatique pour faciliter d'utiliser les algorithmes proposés par M. Yuzhao YANG.

En vue de rendre compte de manière analytique des six mois de stage, ce rapport présentera en cinq parties. La première et deuxième partie du rapport reposent respectivement sur la présentation générale et le contexte de travaux. La troisième partie et la quatrième présentent les démarches que j'ai mises en œuvre pour réaliser les expérimentations et l'application web. Enfin, la dernière partie fait un résumé personnel des apports et des difficultés rencontrées au cours du stage.

1. Présentation générale

Le stage a été effectué dans le cadre du projet BI4PEOPLE, l'un des projets de recherche de l'équipe SIG de l'IRIT. Voici une présentation du laboratoire, de l'équipe et du projet.

1.1. Institut IRIT

L'Institut de Recherche en Informatique de Toulouse (IRIT), une des plus imposantes Unité Mixte de Recherche (UMR 5505) au niveau national, est l'un des piliers de la recherche en Occitanie avec 600 membres permanents et non permanents et plus d'une centaine de collaborateurs extérieurs. Grâce à ses multiples canaux (CNRS, Universités toulousaines), son impact scientifique et ses interactions avec d'autres domaines, le laboratoire est l'une des forces structurantes de l'informatique et de ses applications dans le monde numérique, tant au niveau régional que national.[1]

Les recherches de IRIT se structure autour de cinq grands sujets scientifiques :

- Conception et construction de systèmes
- Modélisation numérique du monde réel
- Concepts pour la cognition et l'interaction
- Etude des systèmes autonomes adaptatifs à leur environnement
- Passage de la donnée brute à l'information intelligible

D'un point de vue organisationnel, les 24 groupes de recherche du laboratoire sont répartis dans sept départements :

Département ASR	Architecture, Systèmes, Réseaux (5 équipes)
Département CISO	Calcul Intensif, Simulation, Optimisation (2 équipes)
Département FSL	Fiabilité des Systèmes et des Logiciels (4 équipes)
Département GD	Gestion de Données (3 équipes)
Département IA	Intelligence Artificielle (3 équipes)
Département ICI	Intelligence Collective, Interaction (3 équipes)
Département SI	Signaux et Images (4 équipes)

[1] https://www.irit.fr/SIG/site/

1.2. Equipe SIG

L'équipe SIG (Systèmes d'Informations Généralisés), l'une des équipes du département GD, constitue l'une des plus importantes équipes du laboratoire avec 20 enseignants-chercheurs, en poste dans quatre universités de la région Occitanie.

Leur domaine de recherche est la Data, et plus particulièrement la gestion des données et le traitement des données de masse (Big Data). Les travaux de recherche de l'équipe SIG visent à concevoir et à développer des méthodes, modèles, langages, algorithmes et outils logiciels qui permettent un accès simple et efficace à l'information pertinente pour en améliorer l'usage, faciliter l'analyse et aider la prise de décision.

Leurs recherches s'orientent sur l'ensemble de la chaîne de traitement des données, des données brutes aux données élaborées accessibles aux utilisateurs qui recherchent des informations, qui visualisent pour des vues synthétiques et qui réalisent des analyses décisionnelles et prédictives.

1.3. Projet BI4PEOPLE

Les technologies de Business Intelligence (BI) telles que l'entreposage de données et le traitement analytique en ligne (OLAP) sont des outils importants pour l'aide à la décision et ont longtemps nécessité de très lourds investissements financiers et humains. De ce fait, il est difficile pour de nombreuses petites et moyennes entreprises d'utiliser la technologie BI pour contribuer à la croissance de leur activité.

L'objectif de BI4people est 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 simple.[2]

En fin de compte, ce projet de recherche fournira une série de solutions automatisées qui permettront aux petites entreprises et aux organisations ne disposant pas de spécialistes de BI d'utiliser les techniques BI pour analyser les données.

[2] https://www.univ-lyon2.fr/recherche/les-projets-de-recherche/bi4people

2. Contexte de travaux

Les systèmes de Business Intelligence (BI) sont largement utilisés dans l'industrie, en particulier dans les grandes entreprises, combinant des données opérationnelles avec des outils analytiques pour présenter les informations de manière structurée et efficace pour soutenir la prise de décision des planificateurs et des décideurs. Avec la tendance actuelle à la numérisation, les petites entreprises, les organisations ou même les particuliers peuvent exploiter chaque jour un grand nombre de données et l'essor des données ouvertes rend les données diverses encore plus accessibles. Pour être compétitives et obtenir des informations précieuses à partir de ces données, ces petites entités s'intéressent également aux systèmes de BI.

Dans le développement d'un système de BI, l'entreposage de données est la tâche la plus difficile, exigeant environ 80 % du temps et des efforts et représentant plus de 50 % des coûts non planifiés du projet [8]. Néanmoins, la conception et l'implémentation d'un entrepôt de données doivent être réalisées par des experts qui possèdent des connaissances professionnelles et des compétences approfondies dans l'entreposage de données. Cependant, il y a un manque général d'une telle expertise technique dans les petites entités. Il est donc nécessaire d'automatiser le processus d'entreposage de données.

Les données se présentent sous de nombreuses formes, mais la plupart des données des petites entreprises et des organisations, et la plupart des données ouvertes, sont sous forme de tableaux [9,10]. Les données tabulaires n'ont normalement pas de schémas. Pourtant, la plupart des méthodes actuelles de conception automatique d'entrepôts de données se concentrent sur les types de données avec schémas [6,7].

En conséquence, M. Yuzhao YANG a proposé de créer un processus automatique de conception et d'implémentation d'entrepôts de données à partir de données tabulaires pour permettre aux PME, aux organisations et même aux particuliers sans expertise technique approfondie d'analyser facilement les données avec les systèmes de BI.

L'ensemble du processus d'entreposage automatique est divisé en trois parties :

- Conception et implémentation automatique d'entrepôts de données multidimensionnelles [2,5]

  La première étape consiste à intégrer des données tabulaires dans un entrepôt de données. En analysant les caractéristiques de la structure des données tabulaires, les attributs des données et les valeurs de données, les relations entre les attributs sont détectées pour automatiquement concevoir un entrepôt de données en modèles multidimensionnels et l'implémenter.

- Fusion des entrepôts de données [3]

  Lorsque plusieurs entrepôts de données partagent des informations en commun, la fusion des entrepôts de données peut permettre aux utilisateurs d'effectuer une analyse globale et intégrée. En analysant les attributs communs des

différents entrepôts de données et les caractéristiques des attributs, les entrepôts de données sont fusionnés aux niveaux de schéma et d'instance pour faciliter l'analyse complète des données par l'utilisateur.

- Imputation de données [1,4]

  En raison des différents attributs de données de l'entrepôt de données original, l'entrepôt de données fusionné peut comporter des valeurs manquantes. Afin d'analyser des données de manière plus complète, l'imputation des données a été utilisée pour remplacer des données manquantes.

2.1. Problématique

Dans mon stage, il y a deux problématiques principales.

Comme décrit dans la section 2, M. Yuzhao YANG a proposé un processus complet d'entreposage automatique à partir des données tabulaires. Dans la partie imputation de données, M. Yuzhao YANG a proposé et a implémenté un algorithme d'imputation pour les données manquantes de dimensions dans les entrepôts de données. Cet algorithme prend en compte la structure et les contraintes dans les entrepôts de données. Il devrait donc théoriquement avoir une meilleure performance en termes d'efficacité et de d'efficience que les autres algorithmes d'imputation existants lorsqu'il s'agit de remplacer des données de dimensions. Pourtant, cette supériorité doit être validée par des expérimentations en comparant l'algorithme de M. Yuzhao YANG avec d'autres algorithmes existants et en utilisant des jeux de données variés. Pour répondre à cette problématique, je dois analyser les algorithmes d'imputation existants, implémenter ceux qui sont adaptés aux données dimensionnelles, chercher les jeux de données d'entrepôt de données et enfin effectuer des expérimentations.

Les algorithmes de différentes parties dans le processus d'entreposage automatique ont déjà été implémentés. Néanmoins, ces différentes parties ont été développées indépendamment. Il est alors difficile d'effectuer le processus complet du début à la fin. De plus, nos utilisateurs cibles sont non-experts, mais il n'y pas d'interface interactive et conviviale permettant aux utilisateurs de consulter les informations et d'exécuter les algorithmes facilement sans connaître les connaissances de BI. Pour répondre à cette problématique, je dois développer une application en reliant les différentes fonctionnalités du processus et en mettant en œuvre une interface pour faciliter les opérateurs d'utilisateurs.

2.2. Objectifs

Avant le début du stage, M. Yuzhao YANG avait déjà effectué des recherches sur la conception d'entreposage automatique pour le système de business intelligence. Il y a deux objectifs dans mon stage. Le premier objectif du stage est de l'aider à réaliser la partie expérimentation de l'imputation de données, qui sera utilisée pour valider les

algorithmes qu'il a proposés. Le deuxième est de concevoir et de développer une application web pour démontrer l'ensemble du processus d'entreposage automatique.

Pour atteindre ces objectifs, mon stage a consisté en deux grandes missions :

- Expérimentations

  Comprendre le contexte de la recherche sur l'imputation de données dans le domaine de la base de données multidimensionnelle. De plus, lire les articles scientifiques pertinents et mettre en œuvre les algorithmes qui y sont proposés. Enfin, appliquer aux jeux de données pour faire les expérimentations.

- Application web

  Collaboration avec une autre stagiaire, Mme. Yuni CHEN, pour concevoir et développer une application web d'entreposage automatique avec génération automatique de bases de données, fusion de bases de données et imputation de données. Cette application sera utilisée pour mettre en œuvre les résultats de recherches de M. Yuzhao YANG.

2.3. Planification

Après notre première rencontre avec M. Yuzhao YANG, nous avons d'abord décidé de diviser le programme de stage en trois parties : l'expérimentation, l'application et le rapport. Afin de rendre le plan adaptable aux changements, un plan détaillé est établi avant le début de chaque partie. La planification finale est présentée à la Figure 1.

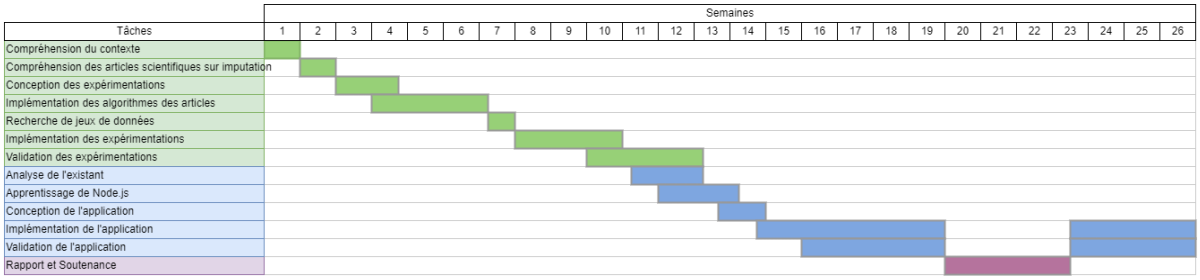


Figure 1 Planning du projet

2.4. Gestion de projet

Le développement agile est une approche itérative et incrémentale de la gestion de projet et du développement de logiciels. Elle permet aux équipes de développement de livrer des produits de manière cohérente, de réduire les risques, d'améliorer continuellement la qualité du développement de l'équipe. Elle peut aussi les aider à répondre avec souplesse à l'évolution des besoins des utilisateurs et d'améliorer la satisfaction des clients grâce au développement itératif.

Comme je travaille sur un projet de recherche, mes tâches peuvent changer de temps en temps en fonction de l'objet de la recherche. Nous avons donc utilisé un certain nombre d'activités agiles pour gérer cet environnement changeant, en exécutant mes missions de manière itérative et incrémentale.

De plus, la partie expérimentations se concentre davantage sur les données et les implémentations de l'algorithme, alors que la mission de l'application web se concentre sur le développement de l'application, les activités agiles utilisées dans les deux parties sont différentes.

2.4.1. Expérimentations

Il n'était pas possible de diviser une expérimentation en plusieurs parties pour le développement et il était également nécessaire de valider chaque étape pour garantir la précision des implémentations d'algorithme avec M. Yuzhao YANG. En outre, en raison des difficultés différentes de mettre en œuvre les algorithmes, nous ne pouvons pas spécifier la durée du sprint. De ce fait, nous n'avons pas fixé chaque sprint à 10 jours ou 15 jours comme la méthode normale de Scrum.

Cependant, nous avons défini des tâches fixes pour chaque sprint pendant l'implémentation d'expérimentations. En utilisant les idées de développement incrémental et itératif de méthode agile, nous avons divisé chaque expérimentation en trois parties, comme le montre la Figure 2. Tout d'abord, nous mettons en œuvre l'algorithme de l'article scientifique, puis nous modifions l'algorithme pour qu'il puisse être appliqué aux données de dimension d'entrepôts de données multidimensionnelles, et enfin, nous réalisons l'expérimentation.

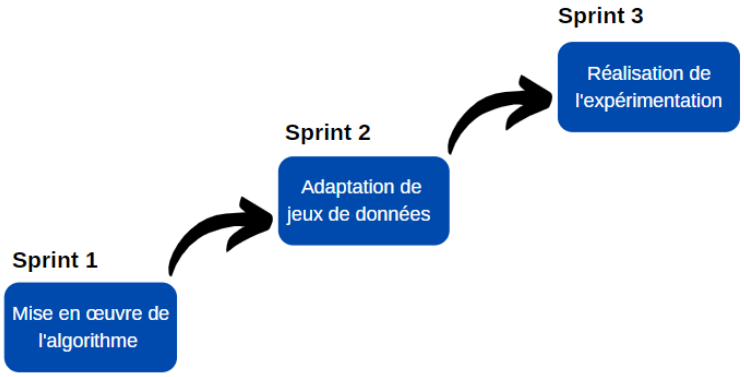


Figure 2 Sprints d'Expérimentations

Il est alors possible de suivre le processus de développement logiciel agile, qui comprend l'analyse, la collecte des besoins, la conception, la vérification et le développement, et qui se termine par la livraison dans chaque sprint.

Au cours de la réalisation des expérimentations, nous avons également organisé quelques activités agiles pour m'aider à réaliser les missions plus efficacement :

- Daily meeting

Le daily meeting est un bon moyen pour l'équipe de développement de suivre l'avancement des tâches, de coordonner l'avancement des tâches et de concilier les tâches en cours avec les difficultés rencontrées chaque jour.

Tous les matins, M.Yuzhao YANG et moi faisons un daily meeting en discutant de l'avancement de mes tâches, les tâches prévues ainsi que mes difficultés durant la compréhension d'articles scientifiques ou la programmation. Grâce aux daily meetings, j'ai pu m'adapter plus rapidement au contexte de l'imputation de données, comprendre l'objectif d'articles scientifiques et améliorer ma vitesse de la programmation.

- Rétrospective

  Une rétrospective est l'occasion pour une équipe de réfléchir au passé et d'apprendre du passé lors d'une réunion. L'objectif principal est d'inspecter la situation et de s'adapter à la réalité. En résumé, l'objectif de la rétrospective est d'aider les équipes à s'améliorer en permanence afin d'être plus efficaces à l'avenir. La rétrospective est donc un élément important du principe de méthode agile.

  Tous les quinze jours au cours de l'expérimentation, l'autre stagiaire, Mme. Yuni CHEN et moi-même avons des réunions de la rétrospective avec M. Yuzhao YANG et son directeur M. Franck RAVAT. Lors de ces réunions, nous faisons le point sur ce que nous avons fait, ce que nous faisons et les problèmes que nous avons rencontrés. M. Franck RAVAT nous pose des questions pour nous aider à approfondir notre réflexion et nous donne un retour et il nous donne aussi des conseils sur la manière de mieux faire notre travail.

2.4.2. Application web

Pendant le développement de l'application web, nous avons appliqué Scrum, un cadre de développement agile, pour gérer notre projet. Une équipe Scrum se compose généralement de trois rôles : l'équipe de développement, le Scrum master et le Product Owner (PO). Dans ce projet, Mme. Yuni CHEN et moi sommes l'équipe de développement et les Scrum Master chargés de développer l'application et d'organiser les réunions, tandis que M. Yuzhao YANG est notre PO chargé de proposer les besoins et de valider les fonctionnalités de développement. Figure 3 présente notre organigramme pendant le développement.

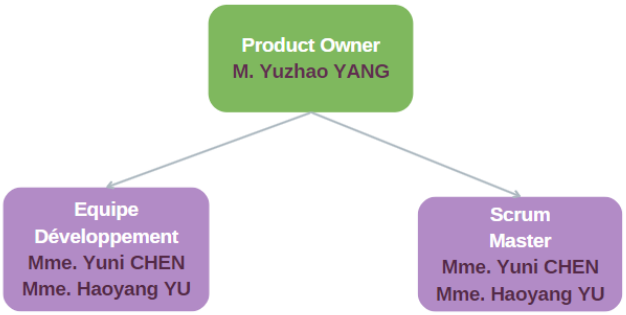


Figure 3 Organigramme de l'Equipe Scrum

Avant de commencer le développement, Mme. Yuni CHEN et moi avons analysé le projet et l'environnement technique. De plus, nous avons identifié les activités agiles qui seraient utilisées durant cette période, comme les réunions quotidiennes et la programmation par paire. Ensuite, dans chaque sprint d'une semaine, nous avons effectué le même déroulement que dans la Figure 3 déroulement de sprint :

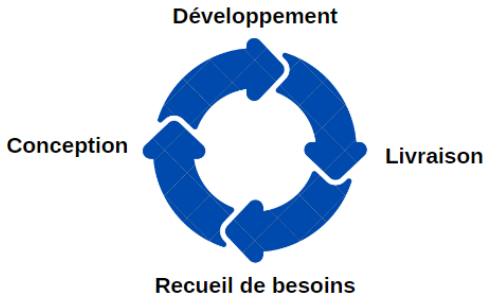


Figure 4 Déroulement de Sprint

- Recueil des besoins/ Livraison

Chaque sprint commence par le recueil des besoins des utilisateurs, ce qui permet à l'équipe de développement de communiquer avec le PO de façon régulière. En conséquence, PO peut être impliqué dans le développement et voir comment il progresse. Et grâce à la communication, l'équipe de développement peut identifier les changements dans les besoins à temps pour éviter les incohérences entre le produit réel et les besoins.

La livraison du sprint en cours a généralement lieu en même temps que le recueil des besoins pour le nouveau sprint, car le PO propose généralement des changements à la livraison pour la rendre plus conforme à sa vision de l'application. Mme. Yuni CHEN et moi organisons avec M. Yuzhao YANG à la fin de chaque sprint pour lui montrer les fonctionnalités que nous avons réalisées et celles que nous sommes en train de programmer. Ensuite, M. Yuzhao YANG fait des suggestions en fonction des fonctionnalités que nous avons réalisées. Cette réunion porte donc non seulement sur la livraison, mais aussi sur le recueil des besoins.

- Conception

Après avoir recueilli les besoins, nous concevons ou modifions la modélisation en fonction des besoins pour la phase suivante du développement. Au cours de ce processus, Mme. Yuni CHEN et moi travaillons indépendamment, puis nous discutons et affinons le design final.

- Développement

La programmation en binôme est une méthode agile de travail importante qui permet de développer un meilleur code plus rapidement en réduisant les risques et en

partageant les connaissances techniques dans toute l'organisation. Avec la programmation en binôme, deux développeurs travaillent sur un seul ordinateur et collaborent sur la même conception, algorithme, code ou test.

Au cours du développement de l'application, nous programmons des fonctionnalités séparément, mais lorsque nous devons modifier l'architecture de l'application, la structure du code ou le code qui s'applique à l'ensemble de l’application, Mme. Yuni CHEN et moi programmons en binôme. Une personne est chargée de corriger le code, tandis que l'autre assure le suivi des corrections et détecte les erreurs dans le code, comme les noms de variables et de méthodes. Si nous travaillons longtemps, nous échangeons les rôles toutes les demi-journées. Parfois, nous travaillons ensemble et trouvons une meilleure solution à un problème. De cette manière, nous pouvons aussi facilement éviter d'éventuelles petites erreurs et réduire le temps nécessaire à la résolution des problèmes. De plus, nous révisons également le code ensemble avant la livraison ou après la réalisation d'une fonctionnalité afin de garantir la qualité.

3. Expérimentations

J’ai effectué des expérimentations pour valider l’algorithme d’imputation de données dimensionnelles manquantes proposé par M. Yuzhao YANG en le comparant avec d’autres algorithmes d’imputation existants. La validation s’effectue en termes de l’efficacité, l'efficience et le respect de la strictness de hiérarchie (i.e. si les hiérarchies strictes restent strictes après l’imputation).

Cette partie de missions de stage se déroulent en quatre étapes dans la Figure 5, à savoir l'analyse d'algorithmes existants, la conception de l'expérimentation, la recherche du jeu de données et l’implémentation des expérimentations. Je décris ces quatre étapes en détail dans cette section.

Figure 5 Déroulement d’Expérimentations

3.1. Analyse de l’existant

L'analyse de l’existant est une étape importante pour m'aider à mieux comprendre le contexte et la direction de l'expérimentation. Elle sert de préparation à la mise en œuvre suivante de l'expérimentation. Je l'ai analysé en deux parties, la recherche et la technique.

3.1.1. Recherche

Au cours de cette analyse de recherche, j'ai lu cinq articles sur l'imputation de données et j'ai appris les différentes méthodes d'apprentissage automatique et les modèles utilisés, tels que le KNN, Bayes et régression linéaire, etc. Cette première analyse m'a permis de mieux comprendre les différentes méthodes existantes sur l’imputation de données manquantes.

Après avoir lu et compris chaque article, j'ai discuté avec M. Yuzhao Yang de la manière dont les algorithmes proposés dans les articles étaient mis en œuvre et s'ils pouvaient être utilisés dans nos expérimentations.

Enfin, les algorithmes de trois des articles ont été choisis pour des expérimentations de comparaison. L'un des deux autres articles a été abandonné, car le modèle d’algorithmes n'était pas applicable aux données multidimensionnelles, tandis que l'autre article était similaire à l'un des modèles déjà choisi. Les algorithmes comparatifs utilisés dans les expérimentations sont présentés ci-dessous.

- **KNN** [11] :

  Cette approche utilise un algorithme K plus proches voisins (KNN) de base pour classer les tuples de données manquantes, puis génère des valeurs de remplacement pour les données manquantes en fonction des K plus proches classes.

  Tout d'abord, les différentes classes possibles de données manquantes sont identifiées, ensuite les caractéristiques qui déterminent les différentes classes et les poids des différentes caractéristiques sont déterminés. L'étape suivante consiste à calculer la distance Euclidienne entre les caractéristiques du tuple de données manquantes et les caractéristiques des autres tuples, en prenant le minimum des k premières valeurs de distance. La dernière donnée manquante est remplacée par la catégorie qui apparaît le plus fréquemment dans ces K lignes de données.

- **NB** [12] :

  Il s'agit d'une approche d'imputation basée sur l'apprentissage automatique et fondée sur l'algorithme de Bayes naïf.

  L'idée principale de cette approche est d'utiliser l'attribut des données manquantes dans le tuple de données comme une classe et les différentes valeurs d'attribut comme des valeurs de classe. Les attributs existants dans le tuple de données sont utilisés comme valeurs caractéristiques pour construire un classificateur de Naïve Bayes et la valeur le plus probable pour les données manquantes sont calculées.

- **MIBOS** [13] :

  Il s'agit d'une méthode d'imputation basée sur des statistiques.

  Cette méthode remplace les valeurs manquantes en comptant le nombre d'attributs pour lesquels le tuple de données avec des données manquantes a la même valeur que les autres tuples.

  Par exemple, l'instance x1 possède quatre attributs a1, a2, a3, a4, où a1 est la valeur manquante. L'instance x2 a les attributs a2, a3, a4, où a2 de x2 est égal à a2 de x1 et a3 et a4 ne sont pas égaux. Ainsi, x2 obtient un score de 1. Pour ce faire, nous trouvons l'instance ayant le score le plus élevé et nous remplaçons a1 de x1 par son a1.

3.1.2. Technique

Avant que je ne commence mon stage, M. Yuzhao YANG avait déjà mis en œuvre l'algorithme proposé pour le processus. En tant qu'expérimentations comparatives, elles étaient nécessaires de s'assurer que l'environnement objectif était aussi identique que possible. J'ai donc utilisé le même langage de programmation Python et le même système de la base de données ORACLE que M. Yuzhao YANG.

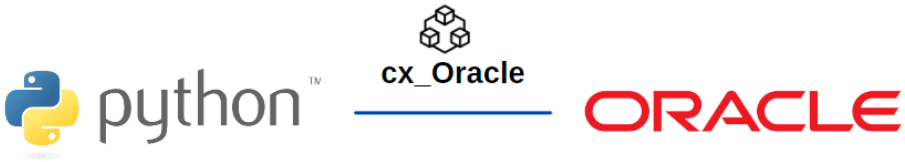


Figure 6 Environnement Technique

Python fournit plusieurs librairies puissantes et complètes pour un traitement des données et un travail d'apprentissage automatique plus efficaces. Oracle est une base de données relationnelle classique qui permet de stocker des données au format R-OLAP. J'ai utilisé Python pour implémenter les algorithmes permettant de compléter les données manquantes dans un entrepôt de données d'Oracle via le module d'extension cx_oracle.

3.2. Conception

La conception expérimentale est une étape importante dans la réalisation d'une expérimentation, car des différences dans la conception expérimentale sont susceptibles de produire des résultats opposés. Pour garantir la rigueur de l'expérience, nous l'avons conçue à la fois en termes d'indicateurs et de stratégies.

3.2.1. Indicateurs

L'objectif des expérimentations est de valider l'efficacité et l'efficience de notre algorithme et le respect de la strictness de hiérarchie avec les données imputées. Afin d'atteindre cet objectif, nous avons proposé cinq indicateurs.

Les trois indicateurs suivants sont utilisés pour valider l'efficacité de l'algorithme. {Imputed} est l'ensemble de toutes les valeurs imputées et {True} est l'ensemble des valeurs imputées correctes.

- **Recall** = {Imputed} ∩ {True} / {True}
- **Precision** = {Imputed} ∩ {True} / {Imputed}
- **F-score** = 2 × P recision × Recall / Precision + Recall

Nous avons aussi utilisé l'indicateur de temps d'imputation pour valider l'efficacité de différents algorithmes.

- **Run time :** Temps d'exécution de l'algorithme pour toutes les valeurs manquantes.

Enfin, nous avons défini un indicateur pour valider la strictness de hiérarchie. Strictness Degree est utilisé pour calculer si un paramètre dans une hiérarchie se conforme à une relation stricte avec ses attributs, ou si un paramètre se conforme à une relation stricte avec le paramètre au niveau élevé de granularité.

- **Strictness Degree (SD)** :

$$SD(p,a) = \frac{N_r(p,a)}{N_d(p)}$$

Cette formule est le degré de strictness du calcul entre un paramètre *p* du et un attribut *a*, qui peut être soit son niveau de granularité supérieur, soit son attribut faible correspondant. Tout d'abord, on calcule le nombre de valeurs différentes du paramètre *p*, ce qui correspond à $N_d(p)$ dans la formule. Ensuite, selon le principe des relations strictes dans les modèles de données multidimensionnelles, pour chaque valeur différente de *p*, la valeur correspondante de *a* doit être unique. En d'autres termes, il devrait y avoir $N_d(p)$ relations strictes. Nous calculons ensuite le nombre de paramètres *p* pour lesquelles il n'y a pas de relation conflictuelle $N_r(p,a)$. $N_d(p)$ et $N_r(p,a)$ sont divisés pour donner l'indicateur de rigueur final.

$$SD(P,A,M) = \frac{\sum_{p_i \in P} \sum_{a_i \in M[p_i]} N_r(p_i,a_i)}{\sum_{p_i \in P} N_d(p_i)|M[p_i]|}$$

Le degré de strictness peut également être calculé pour plusieurs paramètres $P = \{p1, ...pn\}$ par rapport à plusieurs attributs $A = \{a1, ...an\}$. Ici, nous avons aussi une mapping $M = P \rightarrow A$ indiquant pour chaque paramètre, le degré de strictness est calculé par rapport à quels attributs. Ainsi, le nombre de relations peut être obtenu par l'ajout de multiplication du nombre de valeurs distinctes de chaque paramètre avec le nombre de ses attributs faibles correspondants. Nous vérifions également ces relations et obtenons le nombre de relations qui ne portent pas de conflits $Nr(pi, ai)$ pour chaque paramètre $pi \in P$ et chacun de ses attributs faibles correspondants $ai \in M[pi]$. Alors nous pouvons obtenir le nombre de toutes les relations sans conflit. Le degré de strictness dans ce cas peut être calculé comme le formulaire ci-dessus.

3.2.2. Stratégie

Tout d'abord, nous avons apporté quelques modifications aux trois méthodes de recherche décrites ci-dessus afin de mener l'expérience de manière plus précise et complète. D'ailleurs, nous avons ajouté une approche expérimentale de base où, si l'élément manquant est numérique, la valeur moyenne de cet attribut est remplacée. Si la valeur manquante est textuelle, le nombre maximum d'occurrences de cet attribut est utilisé pour remplacer les données manquantes de l'attribut. Voici la version définitive de la conception de processus expérimental pour nos quatre méthodes de l'imputation.

- **KNN**

Pour l'approche KNN, comme l'algorithme original implique seulement des attributs numériques et fournit des poids entre les attributs. Cependant, dans le contexte de notre sujet, les données manquantes peuvent être textuelles ou numériques et il n'est pas possible de savoir les poids des différents attributs. Par conséquent, nous avons apporté trois améliorations à l'approche originale de l'article.

1. Utiliser la méthode Levenshtein pour calculer la distance d'un attribut de texte.
2. Normaliser les distances entre les valeurs des attributs numériques.
3. L'attribut numérique imputé par la moyenne pondérée des K plus proches voisins.

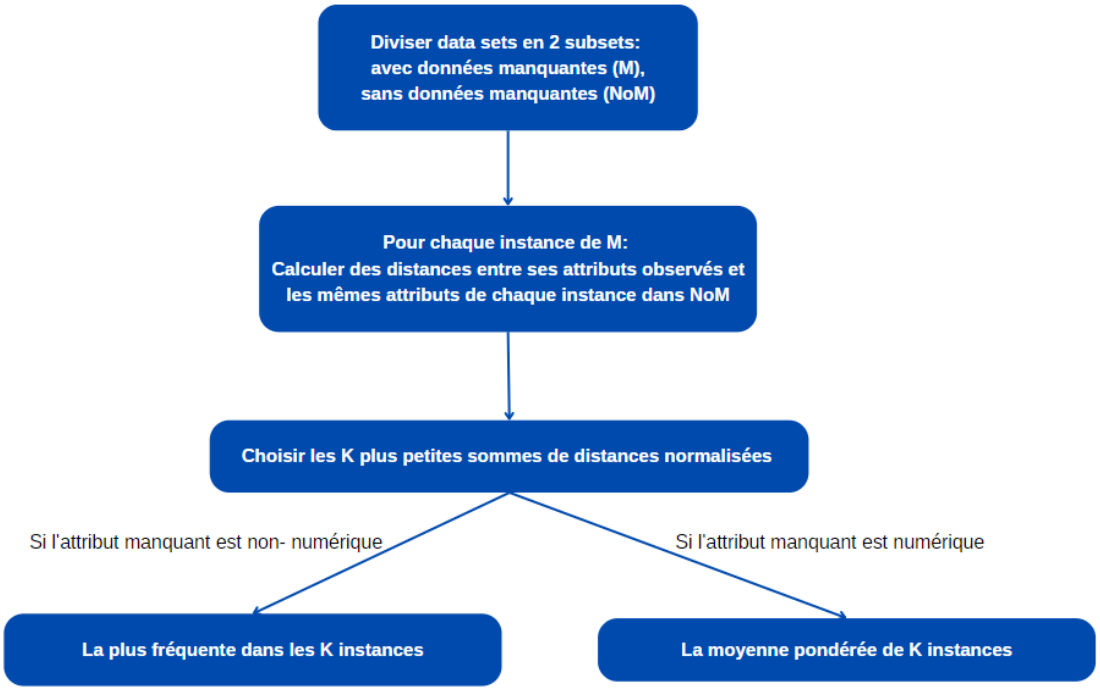


Figure 7 Processus de l'approche KNN

Pour choisir la valeur K la plus optimisée, j'ai testé 20 fois les différentes valeurs de K en la variant de 1 à 10. J'ai ensuite choisi la valeur qui a la meilleure efficacité comme la valeur K finale. Le processus de l'approche KNN ajustée est présenté dans la Figure 7.

- **NB**

L'approche naïve bayésienne dans l'article consiste à classer d'abord tous les ensembles de données en fonction des différentes valeurs de classe. Ensuite, chaque classe est divisée en deux groupes : les données manquantes et les données complètes. Enfin, un classificateur naïf bayésien est construit en utilisant les données complètes de ces deux groupes pour imputer les données manquantes.

Les données de nos jeux de données n'ont aucune restriction de classe, et la plupart des données n'ont aucun attribut de classe. Par conséquent, pour tenir compte de nos données expérimentales, nous avons éliminé l'étape initiale de classification de l'ensemble de données par classe.

La Figure 8 montre le processus d'imputation de l'approche naïve bayésienne.

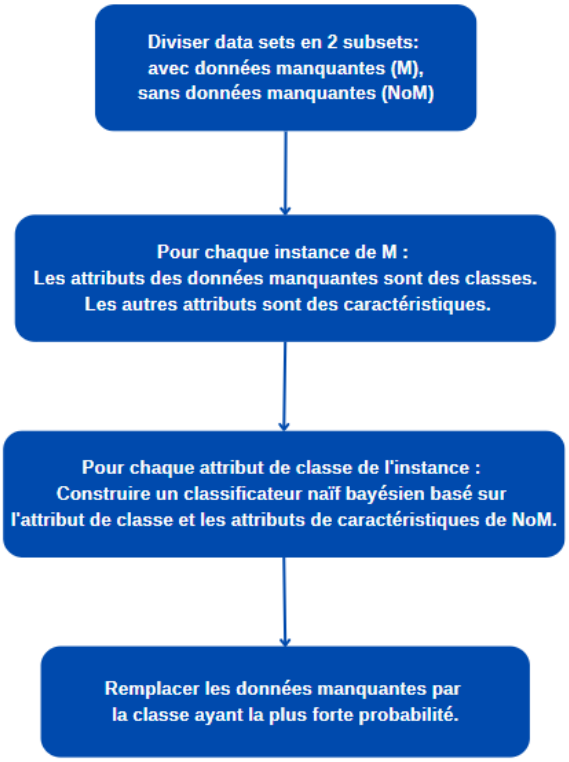


Figure 8 Processus de l'approche Naïve Bayes

- **MIBOS**

En fait, dans la méthode MIBOS, les données imputées en premier sont également utilisées pour participer aux données imputées plus tard. Ainsi, il existe deux versions

de la méthode dans l'article, une version de base dans laquelle les données sont imputées dans l'ordre dans lequel elles sont stockées dans l'ensemble de données. L'autre version étendue consiste à trier l'ensemble des données par ordre décroissant en fonction de la quantité de données manquantes. En haut se trouvent les tuples avec le plus de données manquantes et en bas les tuples avec le moins de données manquantes.

Comme les données manquantes sont imputées par davantage de données originales pour rendre les résultats plus précis, nous avons choisi la version étendue comme le montre la Figure 9 pour l'expérimentation.

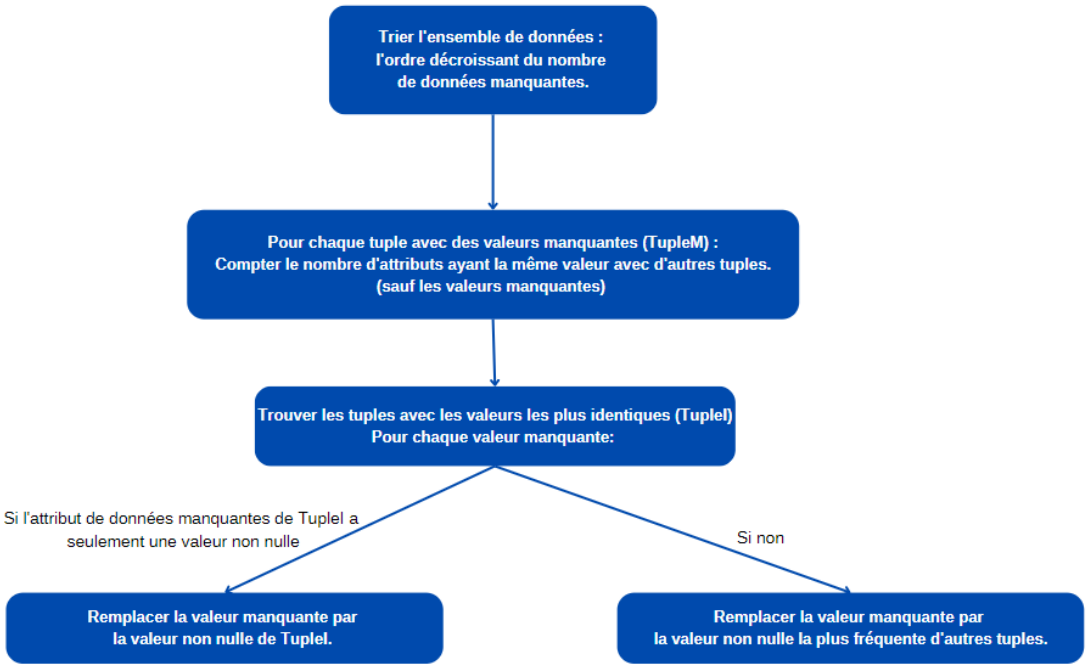


Figure 9 Processus de l'approche MIBOS

- **BASIC**

Comme mentionné ci-dessus, il s'agit de l'expérimentation de base que nous avons ajoutée. La principale méthode consiste à remplacer les données manquantes par la moyenne ou la fréquence la plus élevée des données. Le processus est montré à la Figure 10.

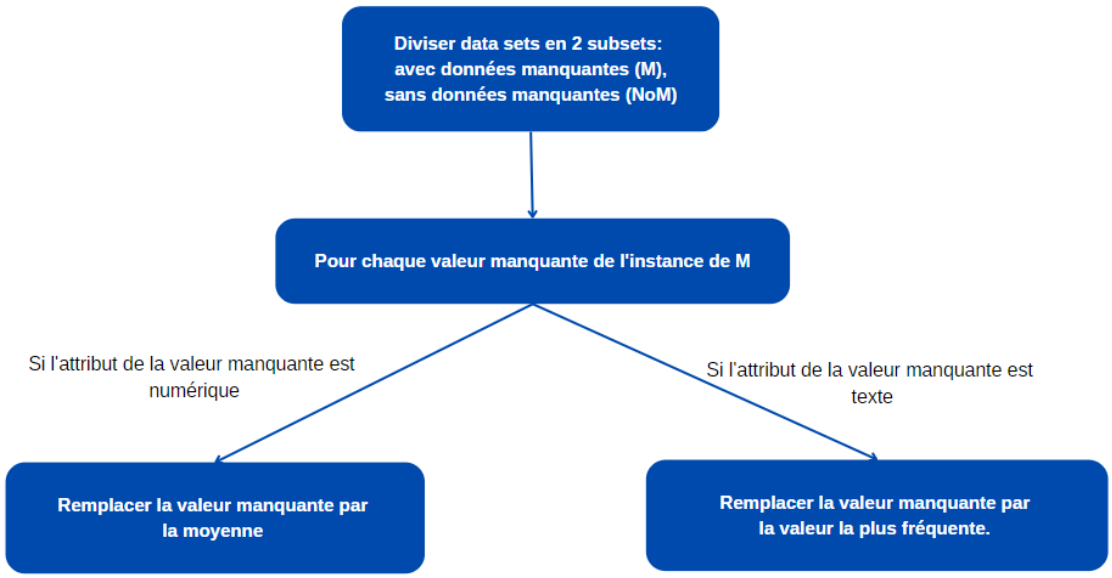


Figure 10 Processus de l'approche BASIC

De plus, afin d'analyser l'effet de l'imputation des données pour différents degrés de données manquantes, nous avons conçu différents taux de données manquantes (1%, 5%, 10%, 20%, 30%, 40%) pour les attributs catégoriels. Dans le but de générer un certain pourcentage de données manquantes d'un attribut, nous trions aléatoirement toutes les instances et supprimons les données pour l'attribut pour le premier certain pourcentage d'instances. Pour chaque taux de données manquantes dans chaque ensemble de données, nous avons effectué 20 fois et obtenu une valeur moyenne.

Enfin, deux stratégies de données manquantes ont été mises en œuvre :

- Mono-attribut : appliquer chaque taux manquant à chaque attribut individuel.
- Multi-attributs : appliquer chaque taux manquant à tous les attributs catégoriels et obtenir une valeur moyenne pour chaque taux manquant.

Car en réalité, les données manquantes peuvent n'exister que dans un seul attribut. Nous pouvons remplacer une valeur manquante dans un attribut en nous basant sur une instance qui ne contient pas la valeur manquante. Des données manquantes peuvent également être présentes dans plusieurs attributs. Cela signifie que nous pouvons remplacer les valeurs manquantes sur la base des instances qui ont des valeurs manquantes, ce qui peut avoir un impact sur la précision des résultats expérimentaux.

3.3. Jeux de données

Pour les jeux de données appliqués dans l'expérimentation, nous avons besoin des jeux de données qui peuvent être modélisés en modèles multidimensionnels qui ont au moins trois niveaux ou plus dans une hiérarchie. J'ai donc trouvé les quatre jeux de données multidimensionnelles suivants.

- **ADVENTURE**

  Il s'agit d'un fabricant multinational fictif de vélos appelé "Adventure Works Cycles". Il comporte deux dimensions, produits et vendeurs, et une hiérarchie < H_produit >, comme le montre la Figure 11.

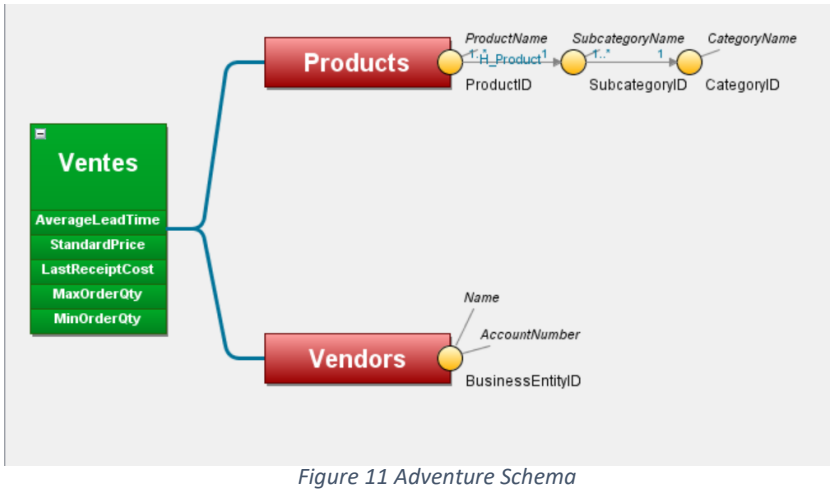


Figure 11 Adventure Schema

- **F1**

  La Figure 12 est un jeu de données qui contient les informations de directeurs et concours sur les courses de Formule 1 d'années 1950 à aujourd'hui. Il a deux dimensions et une hiérarchie <H_RaceLocation>.

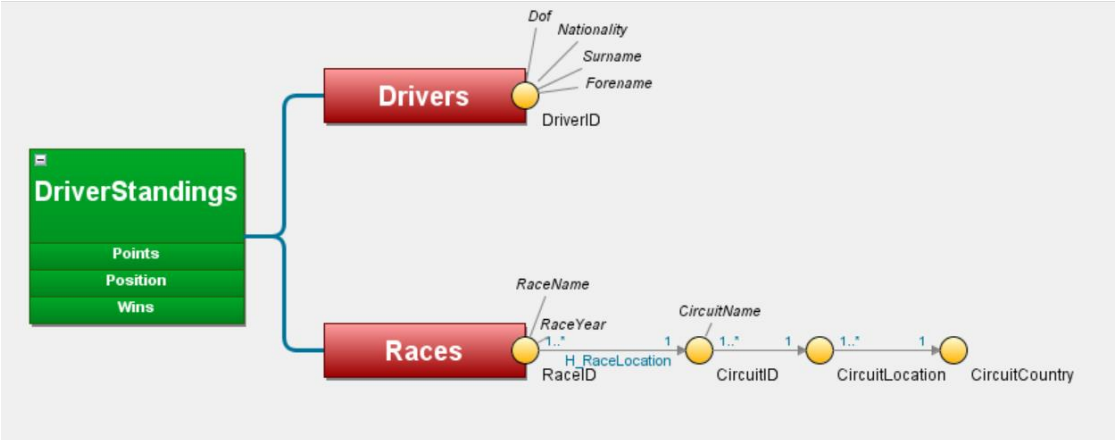


Figure 12 F1 Schema

- **GoSales**

  La Figure 13 présente le jeu de données d'IBM contenant des informations sur les ventes quotidiennes, les méthodes, les détaillants et les produits d'une chaîne de magasins d'équipement de plein air "Great Outdoors" (GO). Ce jeu de données se compose de trois dimensions et de quatre hiérarchies, dont deux ont trois ou supérieur à trois niveaux.

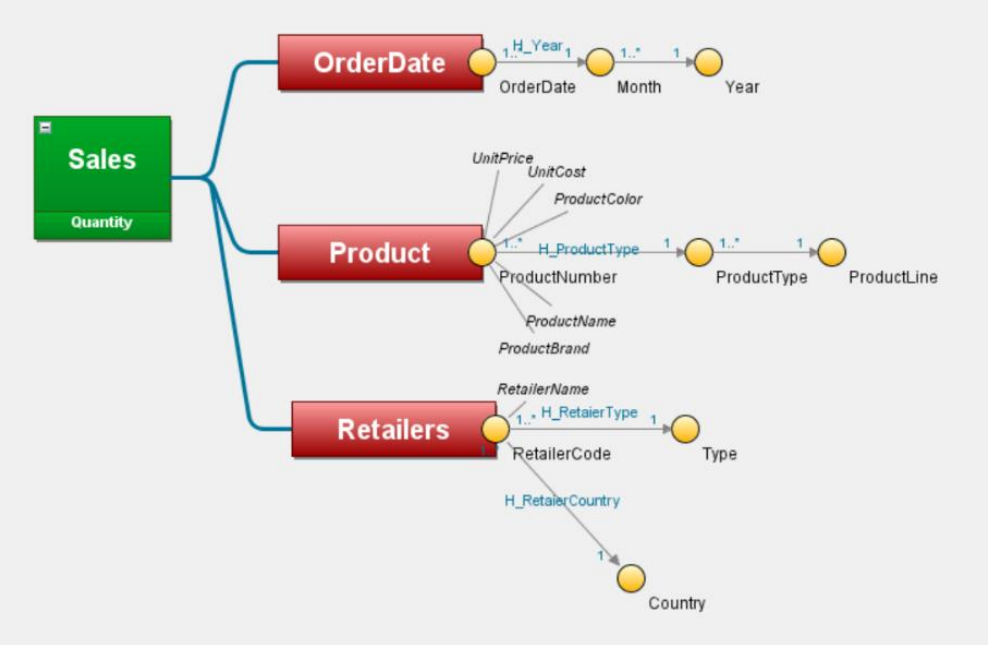


Figure 13 GoSales Schema

- **Organisations**

  La Figure 14 montre un jeu de données géographiques de l'Université de Göttingen qui décrit des informations d'organisations sur 185 pays. Il contient une dimension et une hiérarchie.

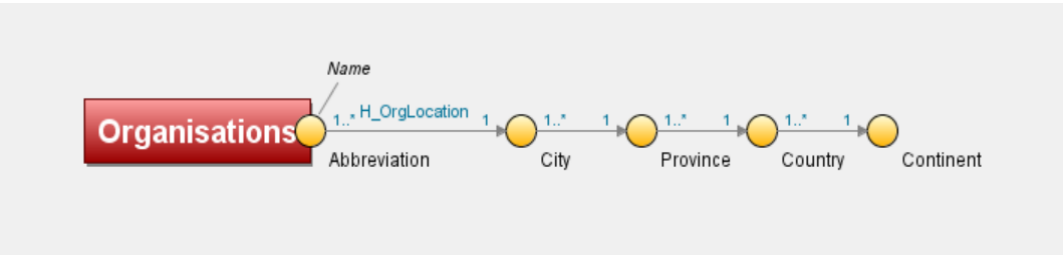


Figure 14 Organisations Schema

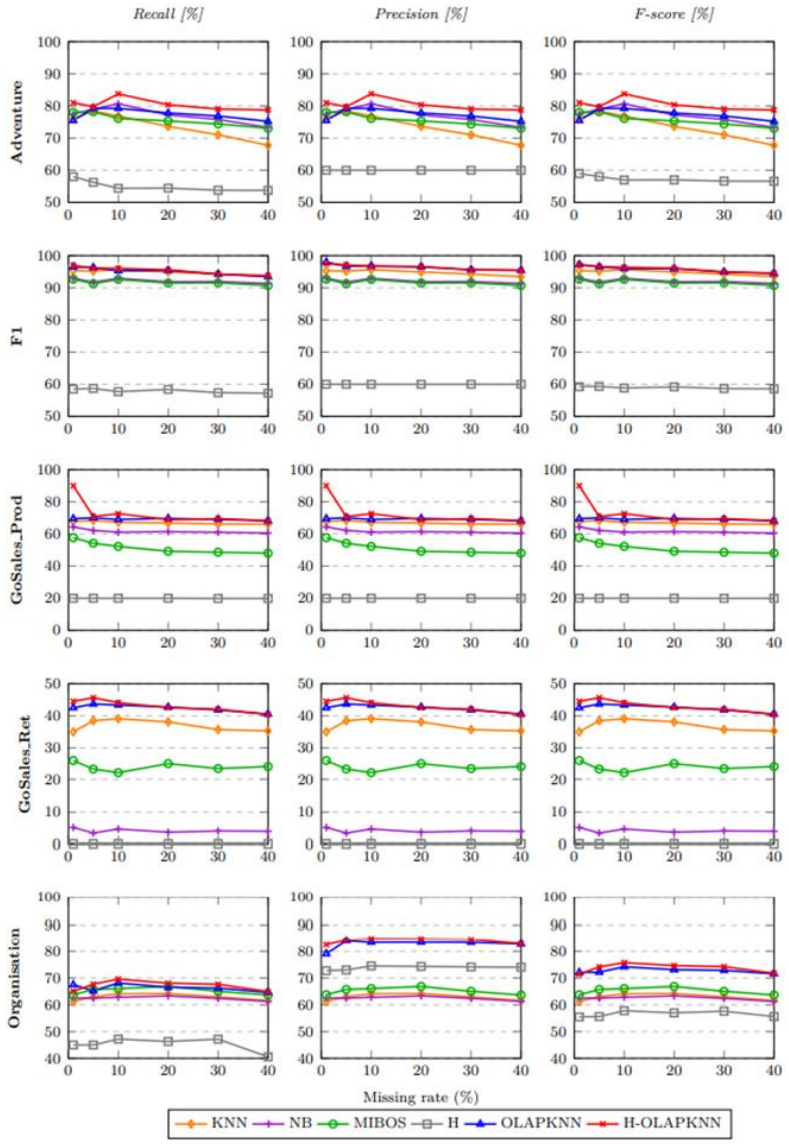
3.4. Résultat

À cette phase, j'ai fait des expérimentations pour imputer les données manquantes en utilisant quatre méthodes différentes ci-dessus mentionnées pour chaque jeu de données.

Après avoir terminé les expérimentations, les résultats des expérimentations BASIC étaient nettement inférieurs aux résultats des autres méthodes de recherche et n'étaient pas informatifs. Par conséquent, les résultats comparatifs suivants n'incluent pas les expérimentations BASIC. Ensuite, j'ai traité les résultats et les ai rendus à M. Yuzhao YANG. Nous avons finalement obtenu des graphiques pour visualiser les résultats. Dans les graphiques, H-OLAPKNN est l'algorithme que M. Yuzhao YANG a proposé. Elle est un algorithme hybride, H et OLAPKNN représentent les deux algorithmes composants de H-OLAPKNN.

- **Résultats sur l'efficacité**

Les Figure 15 et Figure 16 montrent respectivement les résultats dans les expérimentations de mono-attribut et multi-attributs sur l'efficacité contenant le rappel, la précision et le F-score de chaque jeu de données pour chaque taux de valeurs manquantes en utilisant chaque algorithme. Nous constatons que les indicateurs d'efficacité baissent avec l'augmentation de taux de valeurs manquants et l'algorithme H-OLAPKNN que M. Yuzhao YANG a proposé a généralement la meilleure efficacité dans le cas de mono-attribut et aussi multi-attributs. M. Yuzhao YANG a ensuite fait



des analyses plus profondes basées sur ces résultats et les a expliquées dans sa thèse.

Figure 15 Résultat d'Expérimentations-1

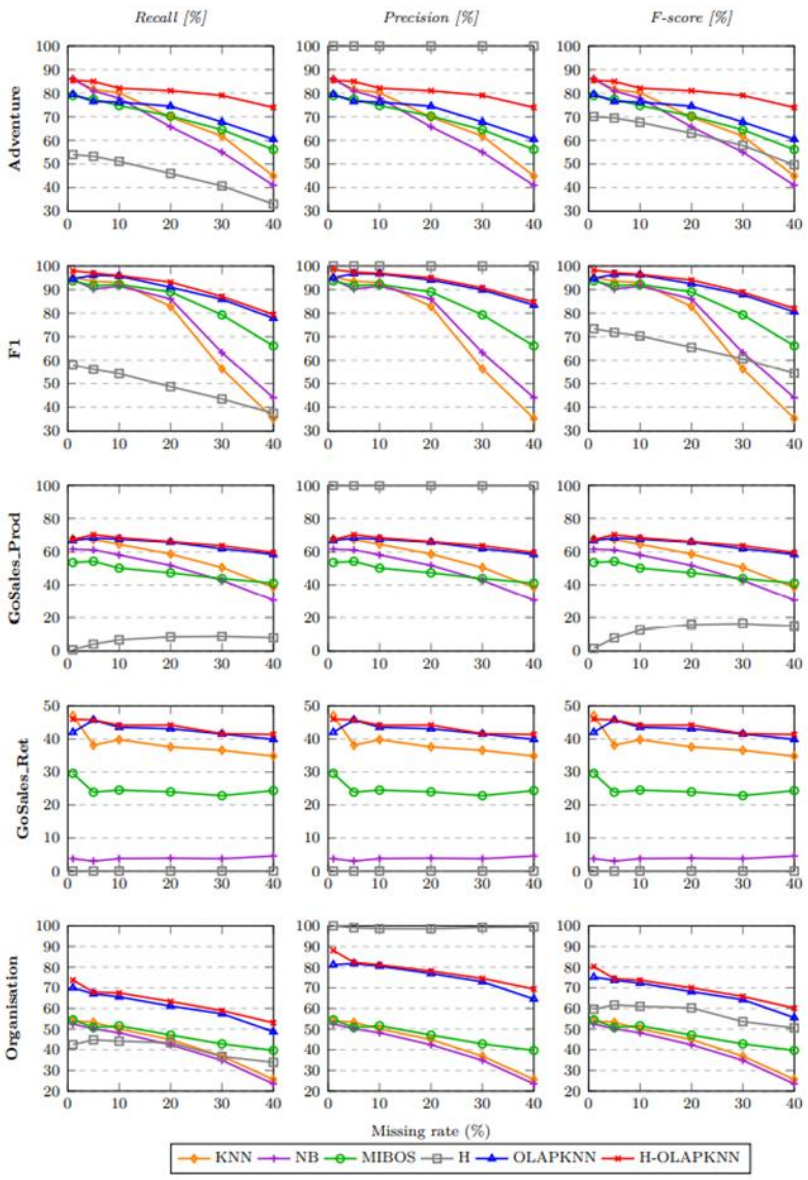


Figure 16 Résultat d'Expérimentations-2

- **Résultat sur le temps d'exécution**

Les Figure 17 et Figure 18 montrent respectivement les résultats dans les expérimentations de mono-attribut et multi-attributs sur l'efficience contenant le temps d'exécution de chaque jeu de données pour chaque taux de valeurs manquantes en utilisant chaque algorithme. Nous constatons que dans la plupart du temps, le temps d'exécution augmente avec l'augmentation de taux de valeurs manquantes. Pourtant, dans le cas de multi-sources, pour les algorithmes MIBOS et KNN, il y a une exception. Le temps d'exécution augmente d'abord et puis baisse avec l'augmentation de taux de valeurs manquantes. J'ai analysé avec M. Yuzhao YANG et nous avons trouvé la

raison pour laquelle il y a cette exception. Ces deux algorithmes cherchent des données remplacées dans les tuples contenants pas de données manquantes. Il y a moins de telles tuples quand le taux de valeurs manquantes est élevé, ces algorithmes mettent donc moins de temps pour la recherche de données remplacées et consomment donc moins de temps d'exécution. Nous constatons aussi que l'algorithme H-OLAPKNN que M. Yuzhao YANG a proposé a généralement la meilleure efficience dans le cas de mono-attribut et aussi multi-attributs. M. Yuzhao YANG a ensuite fait des analyses plus profondes basées sur ces résultats et les a expliquées dans sa thèse.

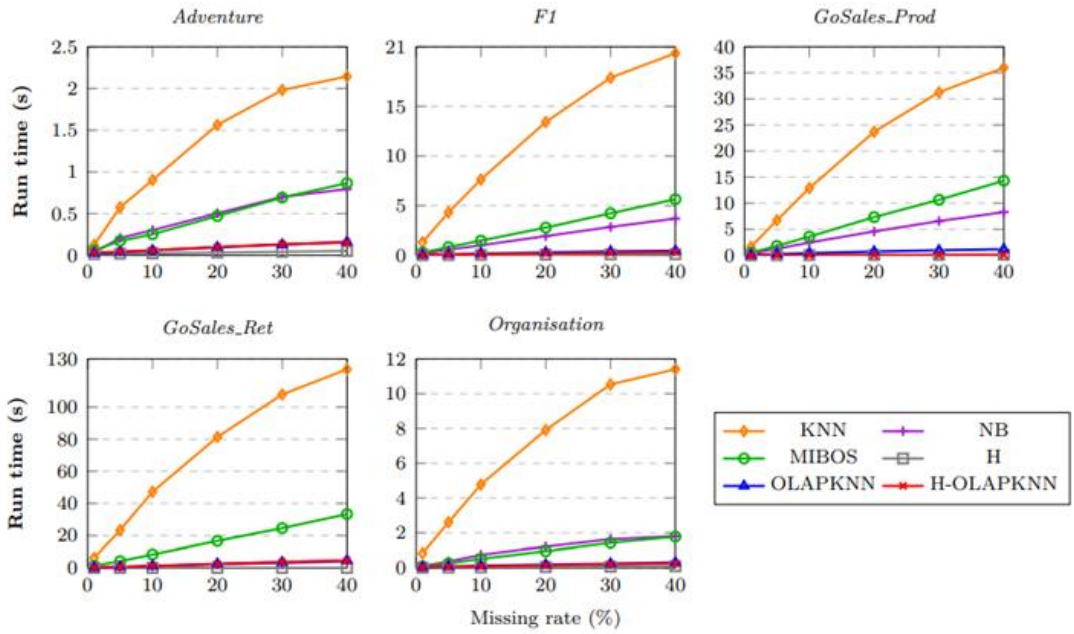


Figure 17 Résultat d'Expérimentations-3

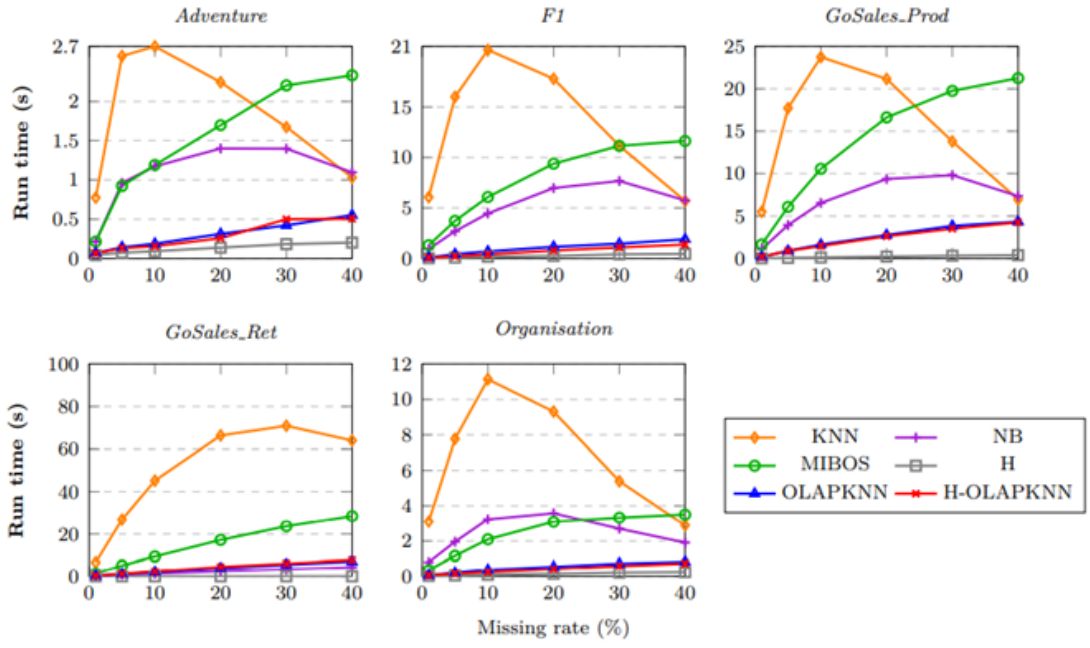


Figure 18 Résultat d'Expérimentations-4

- **Résultats sur le strictness de hiérarchies**

Les Figure 19 et 20 montrent respectivement les résultats dans les expérimentations de mono-attribut et multi-attributs sur le strictness de hiérarchies contenant le degré de strictness de chaque jeu de données pour chaque taux de valeurs manquantes en utilisant chaque algorithme. Nous constatons que l'algorithme H-OLAPKNN que M. Yuzhao YANG a proposé respecte le strictness de hiérarchie et a toujours un degré de strictness de 100% sauf pour le jeu de données Organisation où il y a une hiérarchie non-strict. Le degré de strictness des autres algorithmes de comparaison baisse en augmentant le taux de valeurs manquantes. M. Yuzhao YANG a ensuite fait des analyses plus profondes basées sur ces résultats et les a expliquées dans sa thèse.

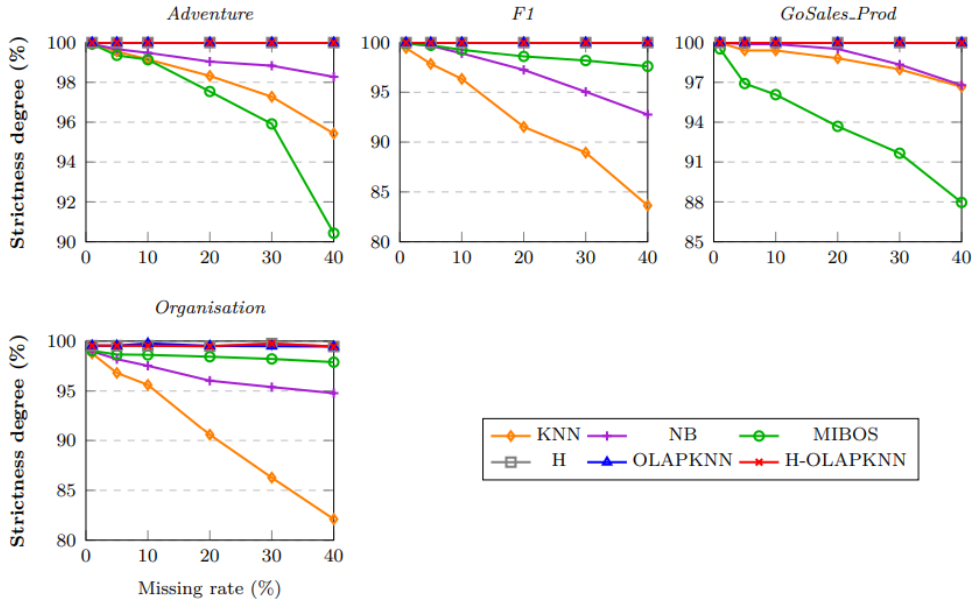


Figure 19 Résultat d'Expérimentations-5

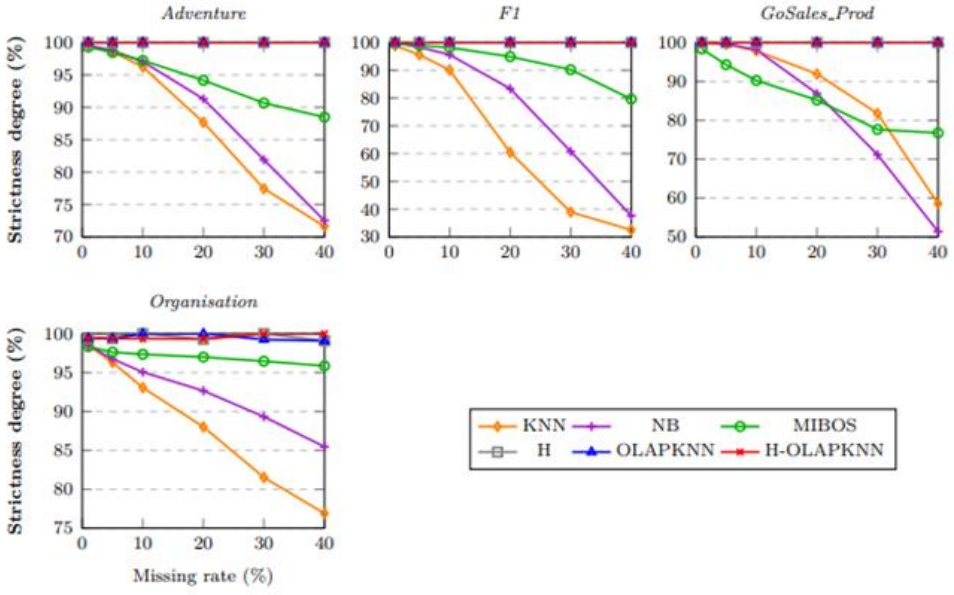


Figure 20 Résultat d'Expérimentations-6

4. Application web

L'objectif de cette partie de démarches est de développer une application web pour implémenter la solution complète de l'entreposage automatique de données tabulaires. L'application applique l'algorithme proposé par M. Yuzhao YANG pour faciliter l'utilisation par les utilisateurs non experts. Elle permets aussi aux utilisateurs experts de faire des modifications le plus possible.

Cette partie de démarches se déroule en quatre étapes : analyse de l'existant, recueil des besoins, conception et développement. L'ensemble du processus est réalisé de manière agile, incrémentale et itérative, chaque phase de sprint étant constituée de ces quatre étapes. Je les présente en détail ci-dessous.

4.1. Analyse de l'existant

Pour cette mission de développement, j'ai travaillé avec M. Yuni CHEN et M. Yuzhao YANG pour analyser d'abord le contexte de notre développement. Il y a deux sections principales : générale et technique. La section générale définit la répartition des tâches et les principales fonctions de l'application, la section technique définit le langage de programmation à utiliser.

4.1.1. Générale

Le processus d'automatisation proposé par M. Yuzhao YANG, comme celui mentionné ci-dessus, se divise en trois parties : la conception et l'implémentation d'entrepôts de données, la fusion d'entrepôts de données et l'imputation de données.

Nous avons divisé le travail en trois parties selon le processus, comme M. Yuni CHEN a terminé son stage plus tôt, elle était chargée de développer la première partie et j'étais chargée de développer la deuxième et la troisième partie.

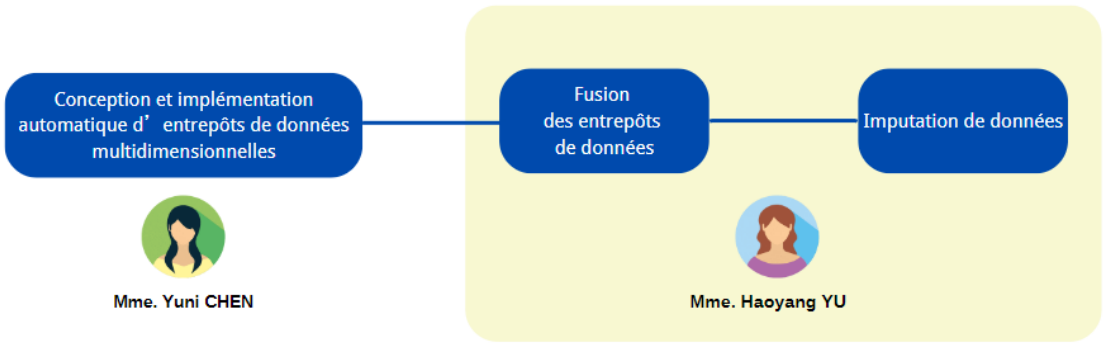


Figure 21 Répartition

4.1.2. Technique

L'analyse de l'existant en termes de technologie est une étape nécessaire avant de choisir différents langages de programmation et frameworks de développement et peut réduire le temps nécessaire aux modifications et à l'intégration ultérieures.

Comme le reste de l'équipe du projet BI4PEOPLE avait déjà commencé à développer d'autres parties du projet en Node.js avant nous, nous avons également choisi Node.js pour faciliter l'intégration de tous les autres projets à l'avenir.

Node.js est un environnement d'exécution qui fait de JavaScript un langage côté serveur. Node.js contient ainsi une série de modules intégrés. L'un de ces modules qui s'appelle child process permet d'exécuter directement du code python, ce qui nous permet de référencer directement le code algorithmique de M. Yuzhao YANG dans le développement de notre application. De plus, comme Node.js dispose d'une série de modules, les applications développées en JavaScript peuvent fonctionner comme des serveurs autonomes sans devoir utiliser un serveur web, tel que le serveur HTTP Apache.

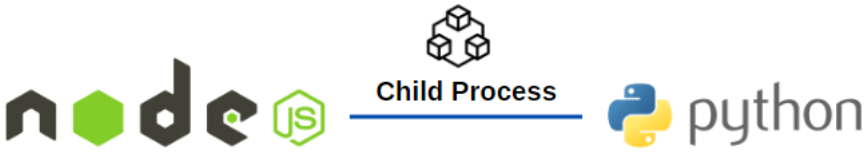


Figure 22 Node.js avec Python

4.2. Recueil de besoins

Après avoir analysé l'existant, M. Yuzhao YANG a présenté ses besoins en tant que PO. Il convient de noter que ces besoins ont évolué pendant le développement. Ci-dessous le tableau final des user stories de PO identifiés à la fin du développement.

Tableau 1 User Stories

N° US	En tant que…	Je veux…	Afin de…	Priorité
1	expert/ non-expert	Remettre le fichier csv.	Enregistrer les données dans un entrepôt de données.	1
2	expert/ non-expert	Transformer des données du fichier en schéma multidimensionnel.	Enregistrer les données dans un entrepôt de données.	1

3	expert/ non-expert	Consulter les résultats de détections.	Gérer les changements d'entrepôt de données.	1
4	expert/ non-expert	Modifier les résultats de détections.	Gérer les changements de données à enregistrer dans un entrepôt de données.	2
5	expert/ non-expert	Créer un entrepôt de données.	Enregistrer les données dans un entrepôt de données.	1
6	expert/ non-expert	Consulter les données d'entrepôt	Choisir l'entrepôt de données à fusionner	1
7	expert/ non-expert	Modifier les résultats de la fusion de l'entrepôt de données.	Gérer les changements de données à enregistrer dans un entrepôt de données.	2
8	expert/ non-expert	Choisir un entrepôt de données.	Imputer un entrepôt de données.	1
9	expert/ non-expert	Consulter le nombre de attributs manquants	Gérer les attributs à imputer.	1
10	expert/ non-expert	Choisir l'algorithme à imputer	Gérer la méthode à imputer.	1
11	expert/ non-expert	Consulter les résultats d'imputation (taux d'imputation)	Faire le bilan de résultats d'imputations	1

12	expert/ non-expert	Modifier la hiérarchie de date	Gérer les données de l'entrepôt de données à enregistrer.	3
13	expert	Consulter les résultats de la fusion en mode analyse forme.	Avoir une image plus complète des résultats de la fusion	3
14	expert	Modifier le schéma de l'entrepôt de données à enregistrer	Gérer le schéma du fichier.	3
15	non-expert	Avoir les explications pour les noms propres	Comprendre mieux l'application	3

4.3. Conception

Après discussion, nous avons décidé de concevoir l'application pour avoir trois fonctionnalités principales correspondant au processus d'entreposage automatique. Dans le même temps, afin de répondre aux besoins de M. Yuzhao YANG, nous avons conçu le logiciel selon un modèle de conception de logiciels divisé en quatre parties.

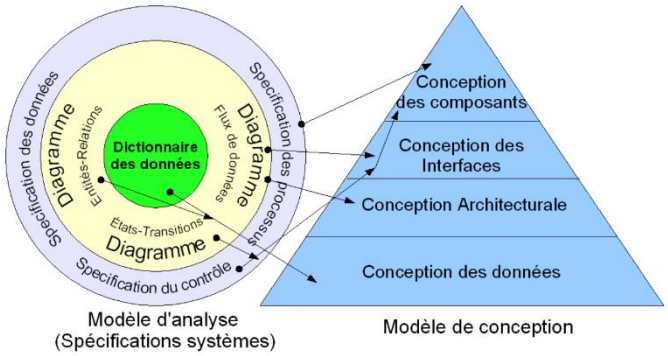


Figure 23 Modèles d'analyse et d'architecture centrées sur les données avec traçabilité

4.3.1. Conception de Données

Pour la conception de données, nous avons choisi d'utiliser le langage UML pour la modélisation.

UML (Unified Modelling Language) est un langage de modélisation standardisé qui consiste en un ensemble de diagrammes. L'utilisation d'UML pour aider les développeurs de systèmes à clarifier, afficher, construire et documenter les résultats des logiciels, UML est une partie très importante du développement de logiciels orientés objet.

Nous avons choisi d'utiliser UML car (i) il nous fournit un langage de modélisation visuelle expressif et facilement accessible qui nous permet de développer et d'échanger des modèles significatifs ; (ii) il est utilisé indépendamment de tout langage de programmation et processus de développement particulier ; (iii) il se prête à la programmation orientée objet, puisque nos applications sont développées dans un langage orienté objet.

Figure 24 montre le modèle UML que nous avons utilisé pour développer notre application, qui illustre les relations entre les données que nous enregistrer dans l'entrepôt de données.

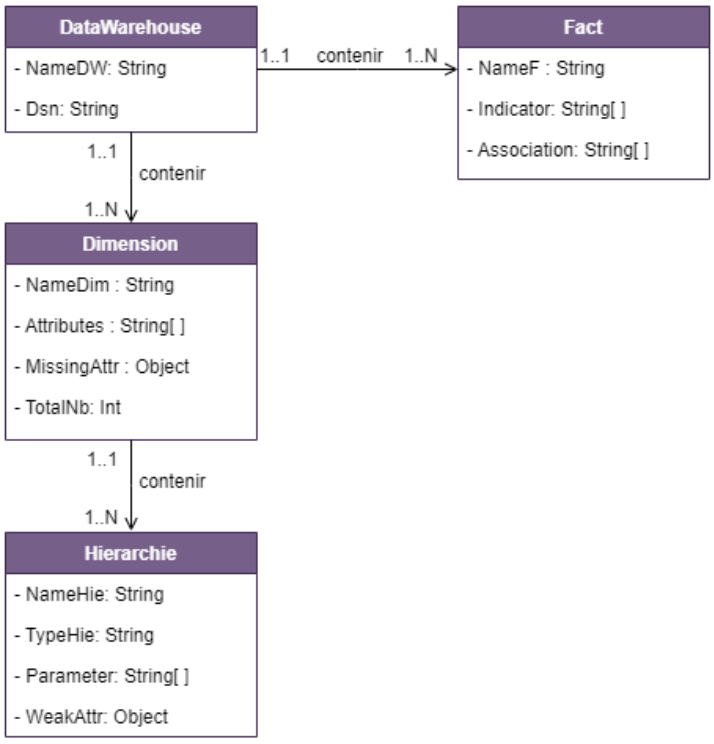


Figure 24 UML

4.3.2. Conception Architecturale

Figure 25 montre l'architecture informatique, qui peut être divisée en architecture matérielle et architecture logicielle. L'architecture matérielle décrit la disposition interne des composants électroniques et leurs interactions. L'architecture logicielle décrit les

différents éléments d'un ou de plusieurs systèmes informatiques, leurs interrelations et leurs interactions, en termes symboliques et schématiques. Le modèle d'architecture logicielle généré pendant la phase de conception décrit le "comment", la manière dont le logiciel est conçu pour répondre aux spécifications et aux besoins.

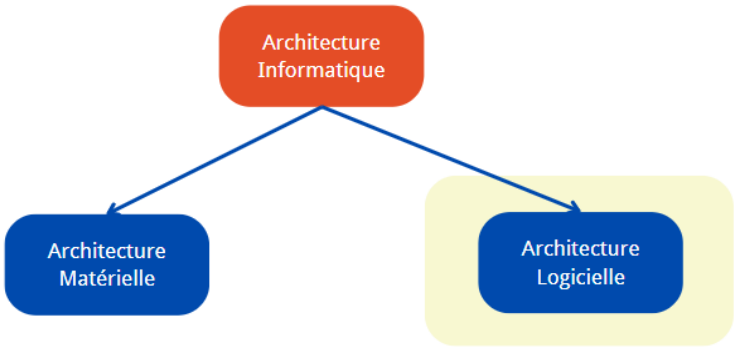


Figure 25 Catégorie de l'Architecure Informatique

Dans cette phase de conception, nous avons uniquement travaillé sur la conception logicielle. En fait, durant le développement, Mme. Yuni CHEN et moi avons utilisé les ordinateurs de l'université avec le même modèle. Et notre application ne nécessite pas d'autre support matériel pour le moment.

Pour l'architecture logicielle, nous avons choisi le modèle MVC du modèle d'architecture. En raison de la contrainte de temps et de la nécessité d'apprendre Node.js, nous avons choisi le modèle MVC le plus courant et le plus familier.

Figure 26[3] montre le modèle MVC. Le Modèle-Vue-Contrôleur (MVC) divise la logique du programme pertinent en trois parties. Ceci est fait pour séparer la représentation interne de données de la façon dont elle est utilisée. Le MVC permet d'utiliser différentes représentations du même programme. Le contrôleur existe pour assurer que Modèle et Vue sont synchronisés, Vue doit être mise à jour une fois que Modèle a changé. Il régule mieux la correspondance entre Modèle et Vue.

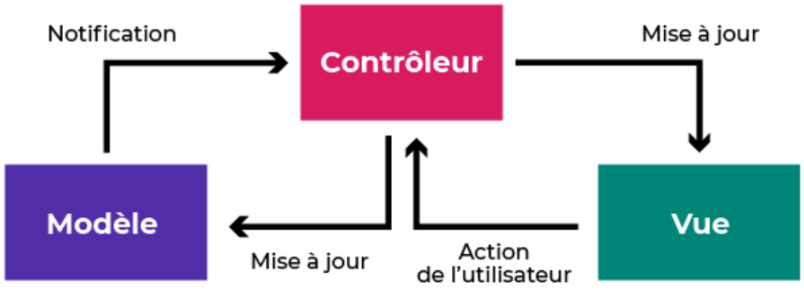


Figure 26 Modèle MVC

[3] https://openclassrooms.com/fr/courses/7160741-ecrivez-du-code-python-maintenable/7188702-structurez-une-application-avec-le-pattern-d-architecture-mvc

4.3.3. Conception des Interfaces

Selon la répartition du travail, Mme. Yuni CHEN et moi avons conçu les maquettes séparément. J'ai conçu la deuxième partie et la troisième du modèle, qui comprenait la fonction de fusion de l'entrepôt de données et la fonction d'imputer des informations manquantes. Une fois la conception initiale terminée, nous en avons discuté et vérifié ensemble, puis nous les avons validées lors d'une réunion avec M. Yuzhao YANG.

Compte tenu du temps restant pour le développement, nous avons décidé de réaliser d'abord les backlog des priorités 1 et 2. Par conséquent, lors de la conception des interfaces, nous n'avons pas pris en compte les user stories de la priorité 3.

Je présente ci-dessous la conception finale validée.

- Entreposage automatique

  Nous avons divisé cette partie de la fonction en quatre étapes. En d'autres termes, elle correspond également à quatre pages.

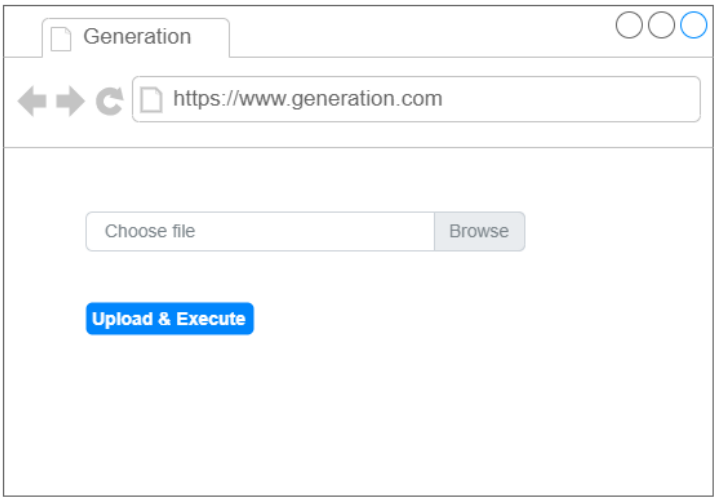


Figure 27 Maquette de Gérération-1

La Figure 27 montre la première étape, où l'utilisateur peut sélectionner et remettre le fichier csv à stocker dans l'entrepôt de données en cliquant button < Upload & Execute >. Aucune information supplémentaire n'est requise de la part de l'utilisateur.

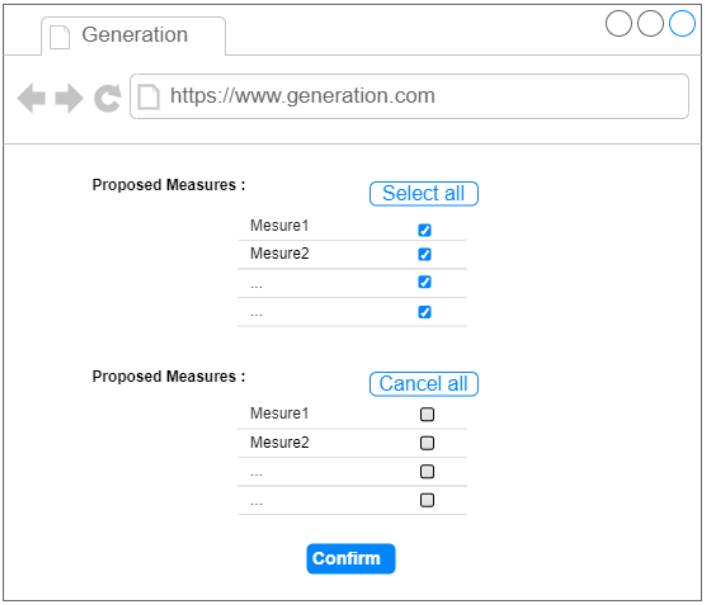


Figure 28 Maquette de Génération-2

La Figure 28 montre la deuxième page de la fonctionnalité qui présente les résultats de mesures de la transformation du fichier remis. Les résultats sont affichés dans deux tableaux. L'utilisateur peut ensuite sélectionner les mesures qu'il veut garder. Une fois que l'utilisateur a terminé la sélection et l'a confirmée, l'application détecte automatiquement les hiérarchies éventuelles.

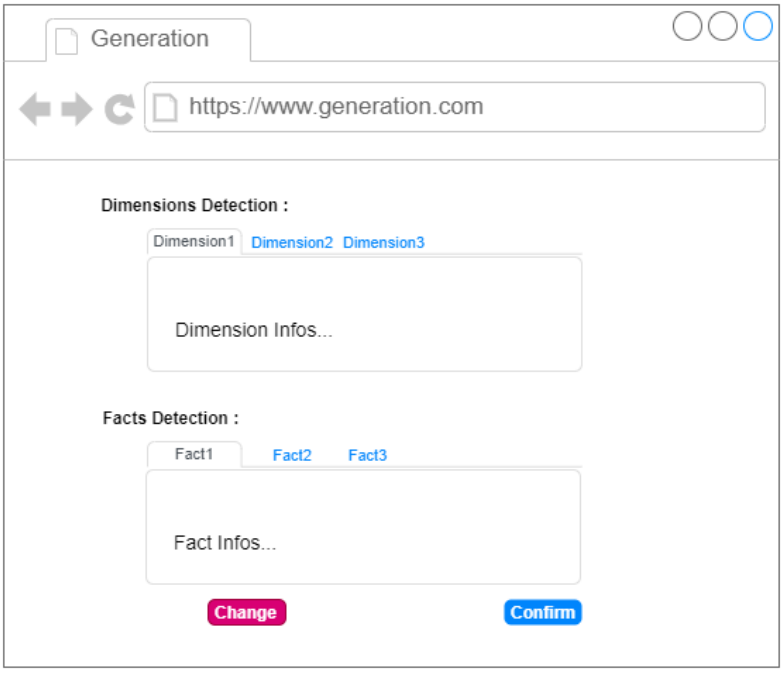


Figure 29 Maquette de Génération-3

La Figure 29 montre le résultat final de la transformation, qui contient les dimensions et les faits de la transformation finale d'entrepôt de données multidimensionnel. L'utilisateur peut modifier les informations du résultat en cliquant sur le bouton.

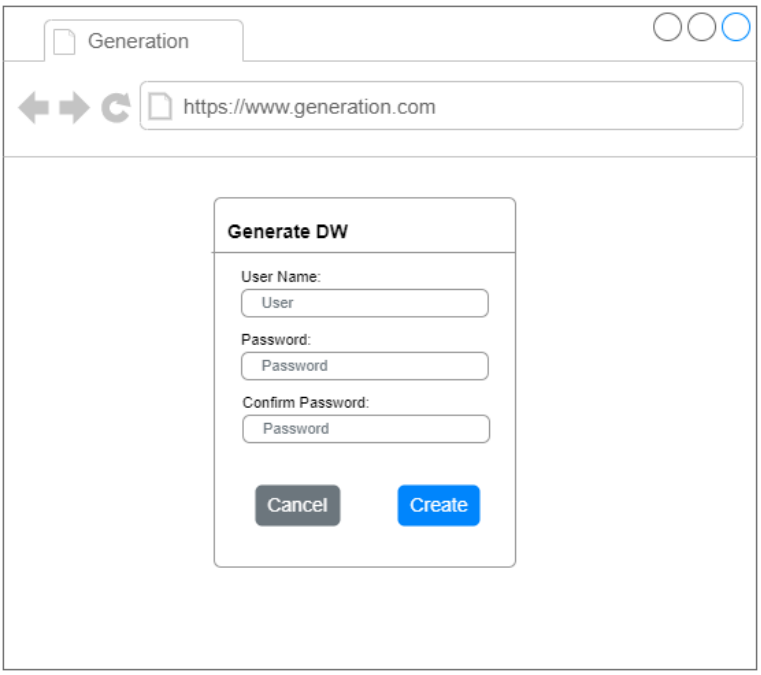


Figure 30 Maquette de Génératio-4

La Figure 30 présente l'étape finale de la première fonction. Une fois que l'utilisateur a confirmé les informations de dimensions et de faits , il devra définir le nom d'utilisateur et le mot de passe pour le nouvel entrepôt de données stocké.

- Fusion d'entrepôts de données

  Nous avons conçu cette fonction en 3 pages, comprenant la sélection de l'entrepôt de données, la confirmation de l'enregistrement du nouvel entrepôt de données et l'édition de l'entrepôt de données fusionné.

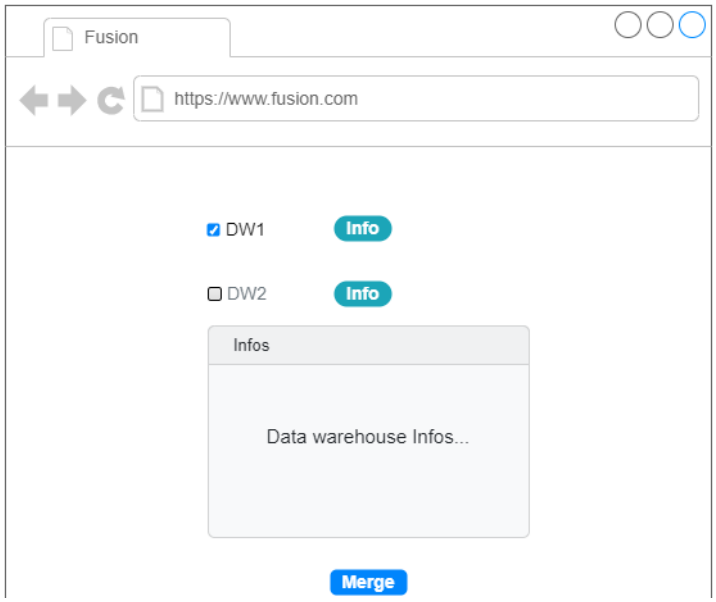


Figure 31 Maquette de Fusion -1

La Figure 31 est la page de sélection de la base de données, qui permet à l'utilisateur de sélectionner l'objet de fusion en fonction des informations de l'entrepôt de données. L'utilisateur peut consulter les informations de l'entrepôt de données en cliquant sur le bouton < Info > et passer à l'étape suivante en cliquant sur < Merge >.

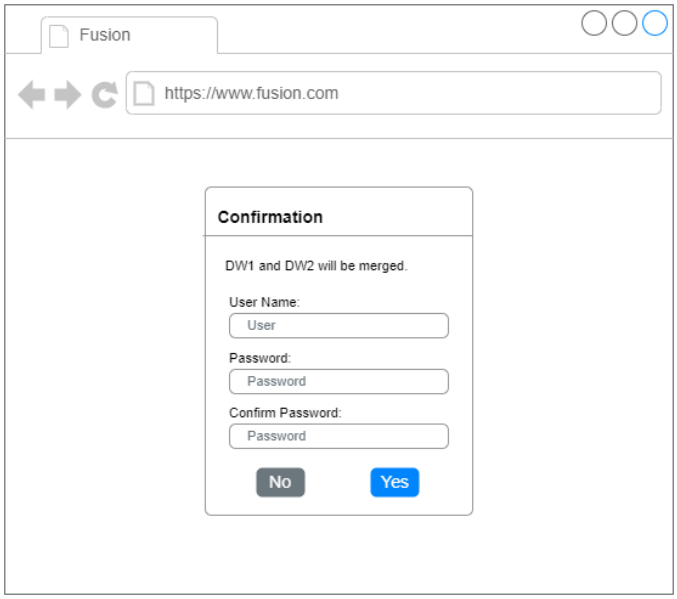


Figure 32 Maquette de Fusion-2

Une fois que l'utilisateur a confirmé les deux entrepôts de données à fusionner, une fenêtre apparaît, comme le montre la Figure 32, dans laquelle l'utilisateur saisit le nom et le mot de passe pour enregistrer les résultats dans le nouvel entrepôt de données.

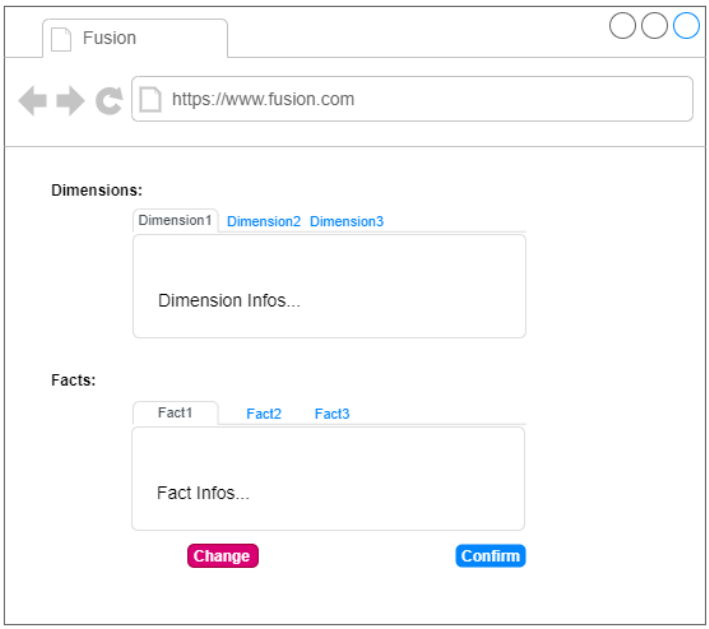


Figure 33 Maquette de Fusion-3

Les résultats de l'entrepôt de données fusionné sont affichés dans la Figure 33, les dimensions et les faits étant présentés séparément. Cela permet à l'utilisateur de voir les résultats fusionnés d'une manière plus visuelle. Les utilisateurs peuvent également modifier et enregistrer leurs résultats en cliquant sur les boutons < Change > ou < Confirm >.

- Imputation de données manquantes

Nous avons conçu les quatre pages suivantes pour la fonction d'imputation de données : sélectionner la base de données, sélectionner les attributs à imputer, sélectionner la méthode d'imputation et enfin afficher le résultat.

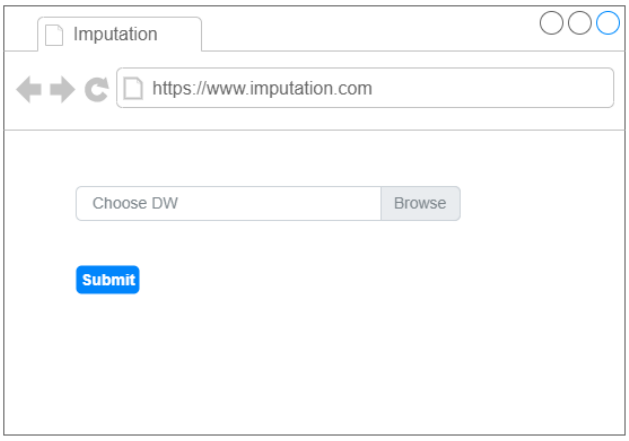


Figure 34 Maquette d'imputation-1

Tout d'abord, l'utilisateur peut sélectionner l'entrepôt de données à imputer à l'aide de la liste déroulante, comme le montre la Figure 34, puis cliquer sur le bouton <Submit> pour passer à la page suivante.

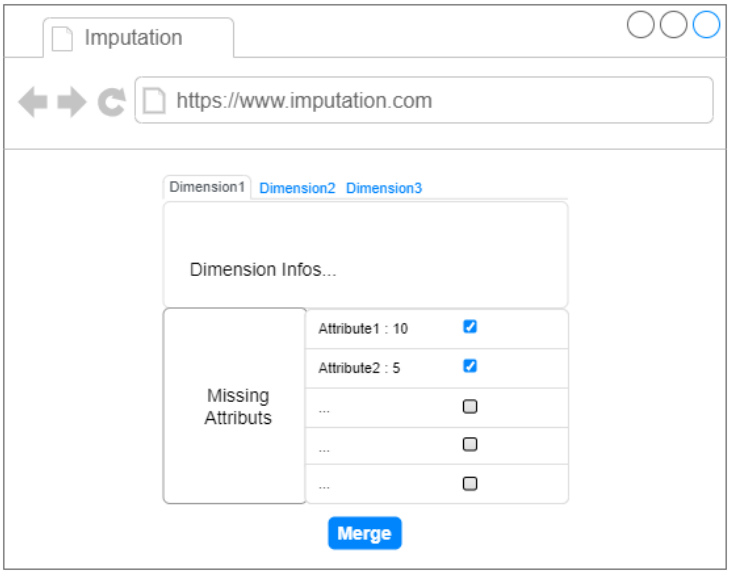


Figure 35 Maquette d'imputation-2

La Figure 35 présente les informations de l'entrepôt de données choisi. Contrairement aux deux fonctions précédentes, elle montre non seulement les informations de dimensions, mais aussi les attributs avec des données manquantes et la quantité de données manquantes. Les utilisateurs peuvent sélectionner les attributs qu'ils souhaitent imputer. Cliquer sur le bouton < Merge > pour passer à la page suivante.

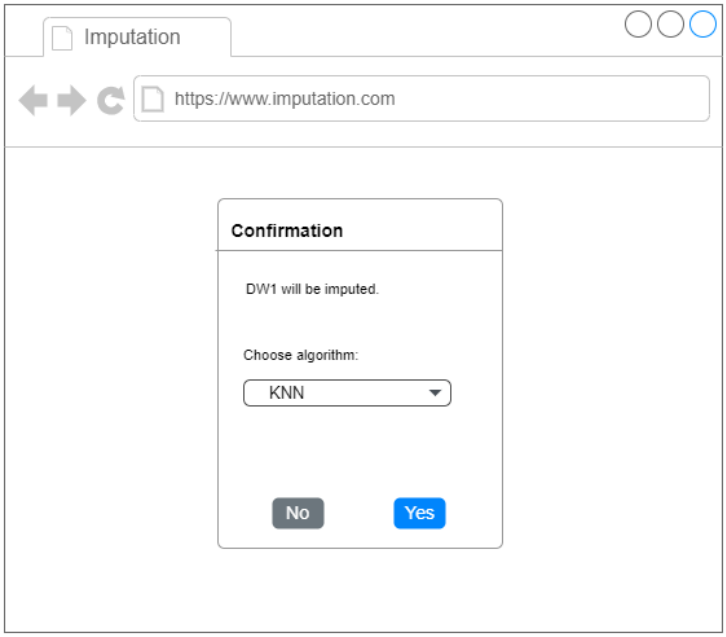


Figure 36 Maquette d'imputation-3

La Figure 36 est la confirmation d'imputation qui permet à l'utilisateur de sélectionner l'algorithme proposé par M. Yuzhao YANG à appliquer aux données manquantes. L'algorithme impute les données manquantes après que l'utilisateur a cliqué sur le bouton <Yes>.

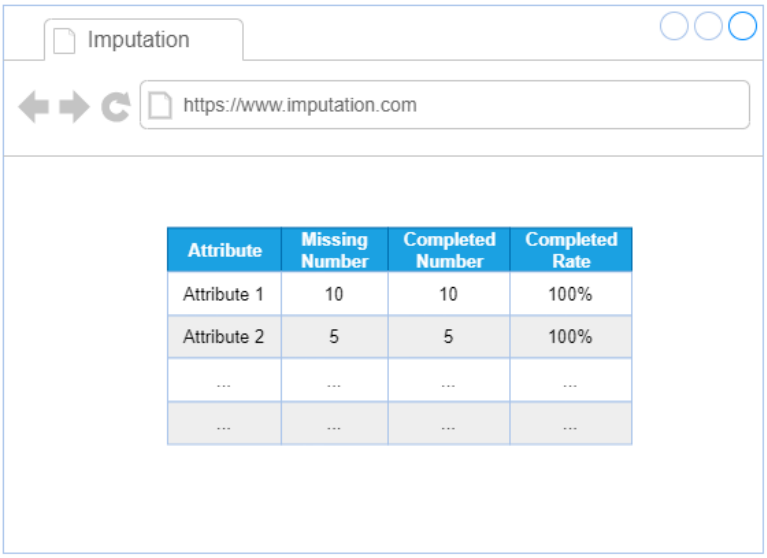


Figure 37 Maquette d'imputation-4

A la fin de la fonction, les résultats d'imputation sont présentés dans la Figure 37. Les trois principaux indicateurs sont le nombre de données manquantes, le nombre d'imputations et le taux d'imputation. Ils sont calculés pour chaque attribut imputé.

4.3.4. Conception des Composants

En architecture logicielle, un composant logiciel est un élément constitutif d'un logiciel destiné à être incorporé en tant que pièce détachée dans des applications. Les paquets, les bibliothèques logicielles, les exécutables, les fichiers, les bases de données ou encore des éléments de configuration (paramètres, scripts, fichiers de commandes) sont des composants logiciels.

Comme il y a tellement de composants différents, je vais les présenter en deux parties : composants du projet et composants de l'application.

- Composants du projet :

  Afin de rendre le programme plus modulaire et d'augmenter sa réutilisabilité. Non seulement nous utilisons le modèle MVC, mais nous séparons également les contrôleurs des routes. Le routage permet d'envoyer les requêtes HTTP vers le contrôleur approprié, de sorte que les fonctionnalités du contrôleur puissent être réutilisées. Ainsi, l'ensemble du flux de données est montré comme la Figure 38 suivante.

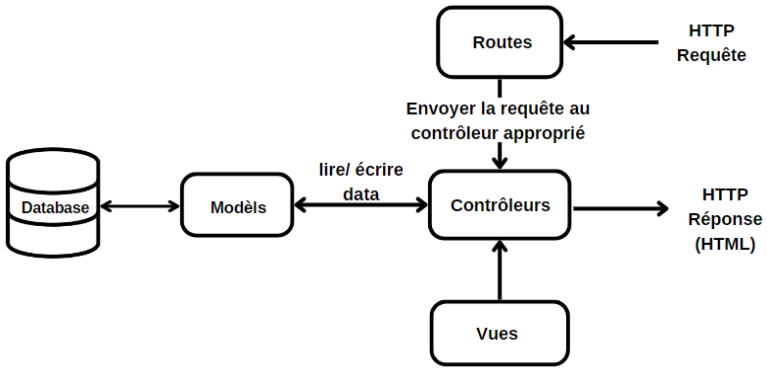


Figure 38 Flux de données

Nous avons conçu les dossiers en fonction des différents modules de la Figure 38 ci-dessus.

- Routes

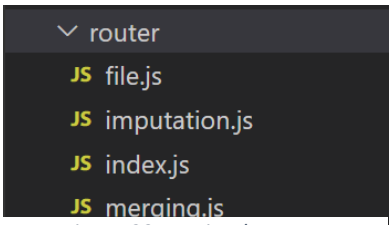


Figure 39 Dossier de Routes

Ce dossier contient tous les routeurs correspondant aux différentes routes.

- Contrôleurs

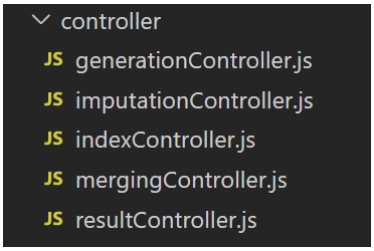


Figure 40 Dossier de Contrôleurs

Le dossier du contrôleur contient toutes les fonctions qui doivent être appelées pour les routes.

- Modèles

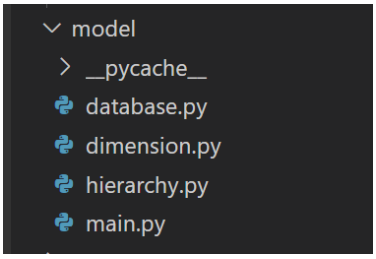


Figure 41 Dossier de Modèles

Le dossier de la modèle comprend l'algorithme proposé par M. Yuzhao YANG la connexion à la base de données en Python.

- Vues

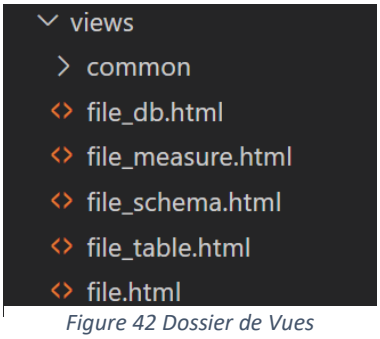


Figure 42 Dossier de Vues

Le dossier Views est constitué des fichiers de HTML que les interfaces présentent finalement à l'utilisateur.

L'ensemble du dossier de notre projet est présenté dans la Figure 43 architecture du projet.

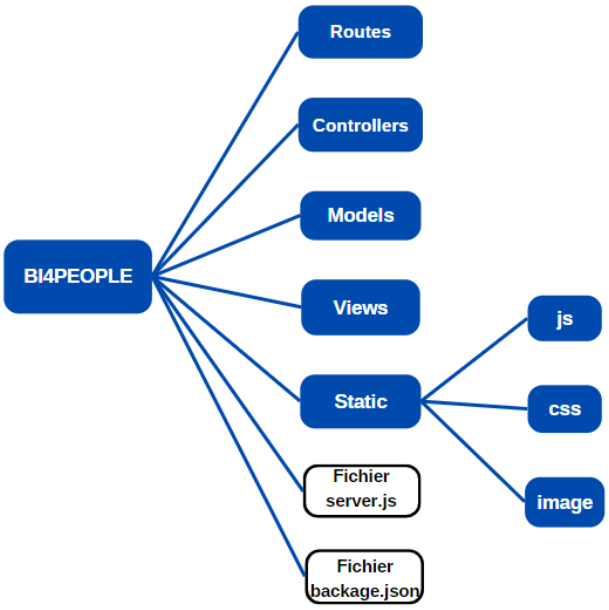


Figure 43 Architecture du Projet

- Composants de l'application :

  La Figure 44 présente les modules, les frameworks et les templates que nous avons choisi en l'architecture de l'application.

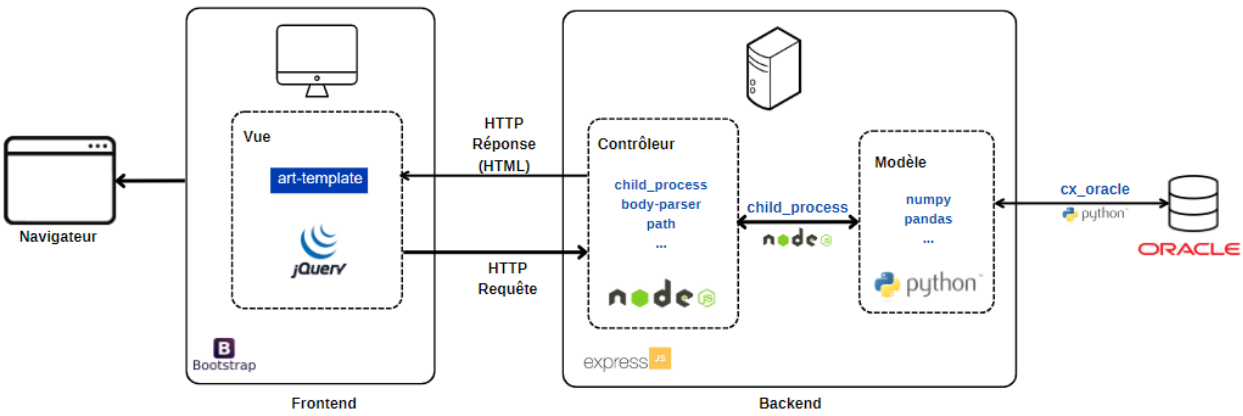


Figure 44 Architecture de l'Application

  - Frontend

Pour la conception du frontend, nous avons d'abord choisi le framework Bootstrap. Il s'agit d'un cadre frontal pour le développement rapide

d'applications web et de sites web. Il possède presque tous les composants nécessaires aux projets web, offrant un large choix de styles. Et nous n'avons pas à nous soucier de la compatibilité des navigateurs avec Bootstrap, il fonctionne sur tous les types de navigateurs.

De plus, pour le moteur de template, nous avons choisi art-template. Il utilise la technologie de prédéclaration d'étendue pour optimiser la vitesse de rendu des modèles. Cela permet la performance de l'application d'être proche de la limite de JavaScript. Donc l'utilisation de l'art-template permet de gagner beaucoup de temps, surtout lorsqu'il s'agit de traiter de grandes quantités de données. La Figure 45[4] montre comment il se compare à d'autres moteurs de template.

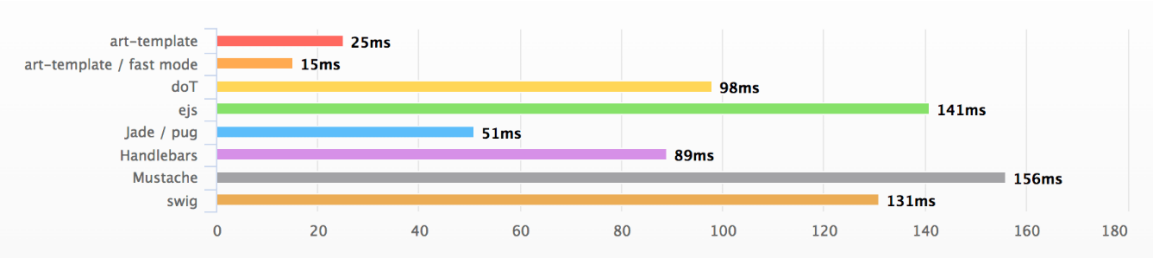


Figure 45 Vitesse de l'Art-template

Enfin, nous avons ajouté la bibliothèque jQuery afin de faciliter l'animation des pages et la mise en œuvre de l'ajax. Elle est une bibliothèque JavaScript rapide, petite et riche en fonctionnalités. Elle facilite la navigation et la manipulation des documents HTML, la gestion des événements, les animations et Ajax.

- Backend

Pour la conception du backend, nous avons d'abord choisi le framework Express. Il est un framework d'application web Node.js minimaliste et flexible qui offre un ensemble de fonctionnalités puissantes pour les applications web et mobiles. Il prend aussi en charge les détails essentiels du backend tels que les sessions, le traitement des erreurs et le routage.

Nous pouvons ensuite développer des contrôleurs en utilisant les différents modules fournis par Node.js. Par exemple, l'un des modules qui s'appelle 'child_process' peut exécuter directement un fichier python et obtenir le résultat. Nous l'utilisons pour connecter le fichier Python du modèle au contrôleur.

Enfin, les bibliothèques de Python peuvent accélérer l'exécution de nos algorithmes. L'un d'entre eux qui s'appelle 'cx_oracle' peut utiliser pour connecter Python à notre entrepôt de données Oracle et modifier ses données.

[4] https://aui.github.io/art-template/

4.4. Développement

J'ai réalisé la plupart des travaux de développement jusqu'à la remise du rapport et je vais détailler les fonctionnalités et les pages qui ont été mises en œuvre.

4.4.1. Implémentation de données

Notre application utilise deux méthodes principales pour stocker et modifier les données générées par l'application. La relation entre eux est illustrée à la Figure 46.

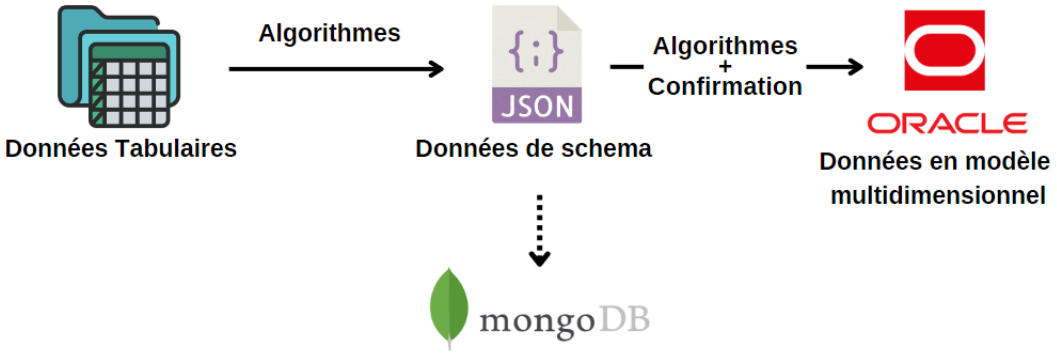


Figure 46 Implémentation de données

- JSON

  Notre application transforme les données tabulaires en données multidimensionnelles pour les enregistrer dans Oracle. Au cours de ce processus, il est nécessaire de visualiser les données dans un schéma multidimensionnel.

  Nous utilisons actuellement les fichiers de JSON pour y parvenir. Le schéma résultant de chaque transformation de données est stocké en JSON. Le fichier de JSON peut être visualisé par l'utilisateur et qui peut être consulté par l'utilisateur. Ce n'est qu'une fois que l'utilisateur a validé les données que nous les enregistrons dans l'entrepôt de données.

  Jusqu'à présent, nous avons stocké le fichier JSON dans le dossier du projet. Dans le futur, si nous avons le temps, nous le stockerons dans la base de données MongoDB. Parce qu'il s'agit d'un système de gestion des données orienté vers les documents.

- Oracle

  Dans la base de données Oracle, nous enregistrons toutes les créations, les modifications et les imputations de l'entrepôt de données pendant notre processus. Comme application initiale, nous avons implémenté deux jeux de données utilisés par M. Yuzhao YANG dans ses recherches.

4.4.2. Implémentation de l'application

Au moment de la rédaction de cet article, nous avons terminé la plupart des fonctionnalités de l'application et je vais couvrir chacune des trois principales fonctionnalités dans cette section.

L'interface de notre application est divisée en deux parties principales, comme le montre la Figure 47. À gauche, une barre latérale permet de sélectionner les fonctions. En plus des trois fonctions principales, nous avons développé deux versions pour les experts et les non-experts. Pour les pages destinées aux non-experts, nous avons enlevé les mots experts et ajouté des notes explicatives. À droite, les pages interactives des trois fonctions sont affichées.

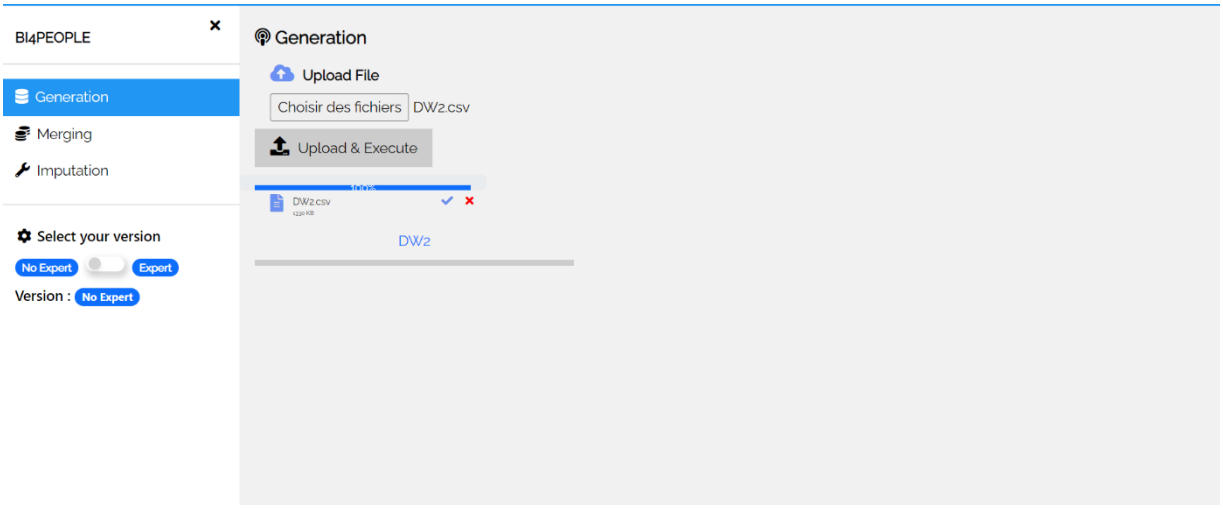
Figure 47 Interface d'Index

- Conception et implémentation automatique d'entrepôts de données multidimensionnelles

  Pour la première fonction, il s'agit de conception et implémentation automatique d'entrepôts de données multidimensionnelles. Nous avons réalisé toutes les fonctions prévues dans la maquette et ajouté des fonctions de priorité 3, sélection les attributs de hiérarchies des dates et modification le schéma transformé en mode expert. La mise en œuvre de cette fonction est décrite dans les étapes suivantes :

1. Sélection de fichiers

L'utilisateur clique d'abord sur le bouton <Choisir des fichiers> de la Figure 48 pour sélectionner le fichier CSV qu'il souhaite enregistrer dans l'entrepôt de données. Ensuite, cliquez sur le bouton < Upload & Execute > pour exécuter.



La barre de progression située sous le bouton indique 100 % lorsque l'exécution est terminée.

Figure 48 Interface d'Entreposage Automatique-1

2. Détection de mesures

En cliquant sur DW2 dans la Figure 48, Figure 49 ci-dessous apparaît, montrant les informations détaillées de la conception d'entrepôt de données multidimensionnelles. Tout d'abord, sélectionner les mesures qui seront enregistrées dans le tableau fait. Ces mesures sont indiquées comme des indicateurs dans les pages aux non-experts.

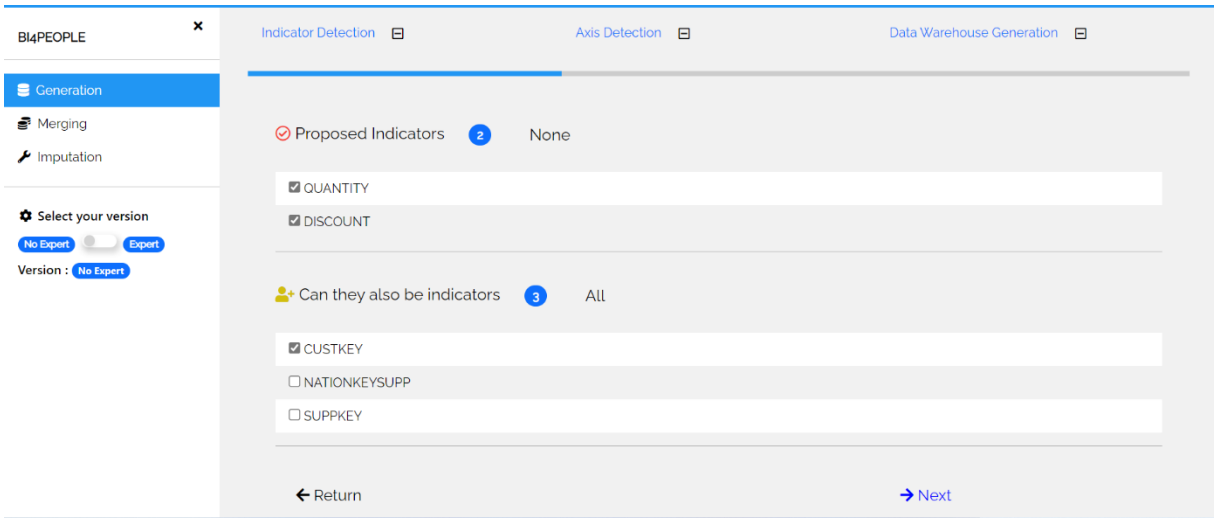


Figure 49 Interface d'Entreposage Automatique-2

3. Détection de dimensions et faits

L'utilisateur clique sur le bouton < Next >de la Figure 49 pour passer à l'interface suivant, Figure 50, qui est la détection des dimensions et faits.

Compte tenu de la spécificité et de la diversité de la dimension date, notre application permet à l'utilisateur de sélectionner les paramètres de la dimension date en fonction de ses besoins.

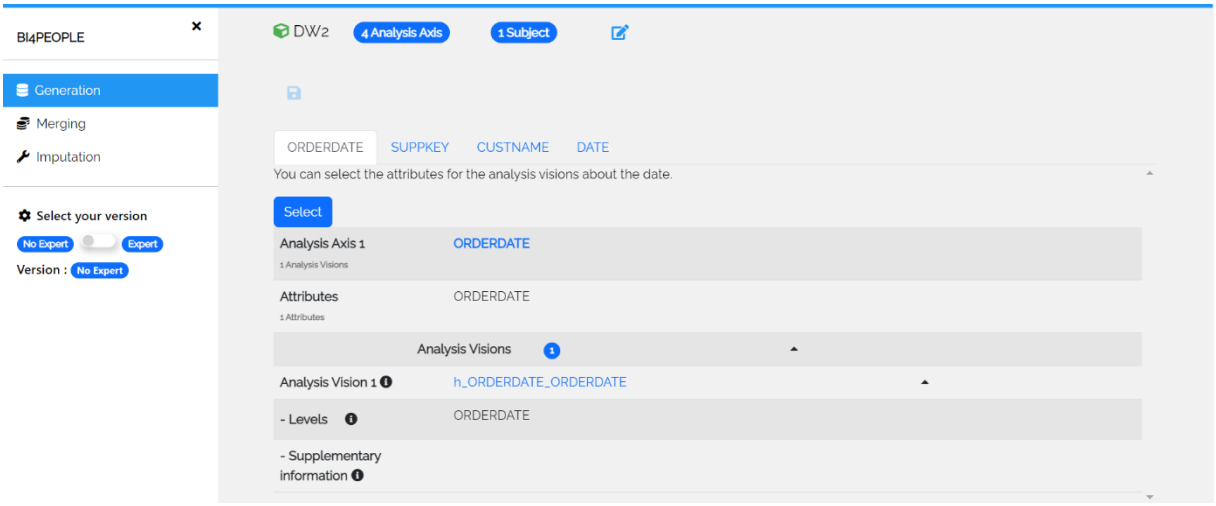


Figure 50 Interface d'Entreposage Automatique-3

Pour l'interface de la détection, il existe deux différences entre les pages experts et non-experts. Le mode non-expert permet à l'utilisateur de passer la souris sur une icône pour voir l'explication du mot, comme le montre la Figure 50. Le mode expert permet à l'utilisateur de modifier le schéma des résultats de la détection. Les utilisateurs peuvent faire des modifications en cliquant sur le bouton < Edit Schema > dans la Figure 51.

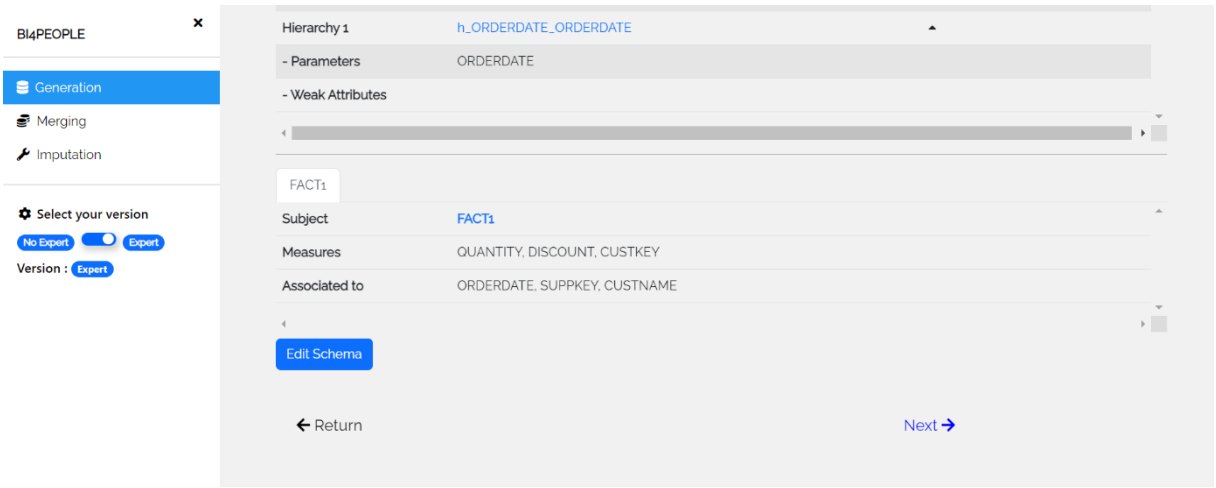


Figure 51 Interface d'Entreposage Automatique-3

4. Enregistrement de l’entrepôt de données

Enfin, l'utilisateur clique sur le bouton < Next > pour faire apparaître la fenêtre de la Figure 52 afin de saisir les informations requises pour que l'entrepôt de données soit enregistré avec succès.

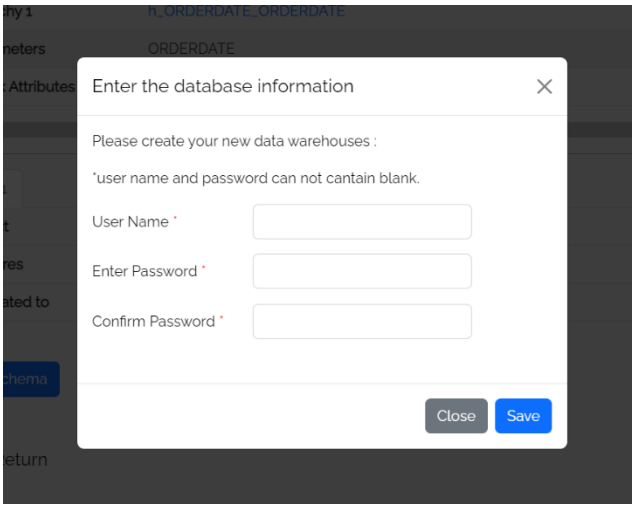


Figure 52 Interface d'Entreposage Automatique-6

- Fusion des entrepôts de données

En ce qui concerne les fonctions de fusion, nous avons mis en œuvre toutes les fonctions conçues dans les maquettes, ainsi que la fonction permettant aux experts de consulter des résultats de fusion plus détaillés. Dans ce qui suit, je vais présenter en détail l'implémentation de cette fonction.

1. Sélection des entrepôts de données

La Figure 53 montre la page de sélection des entrepôts de données pour fusionner.

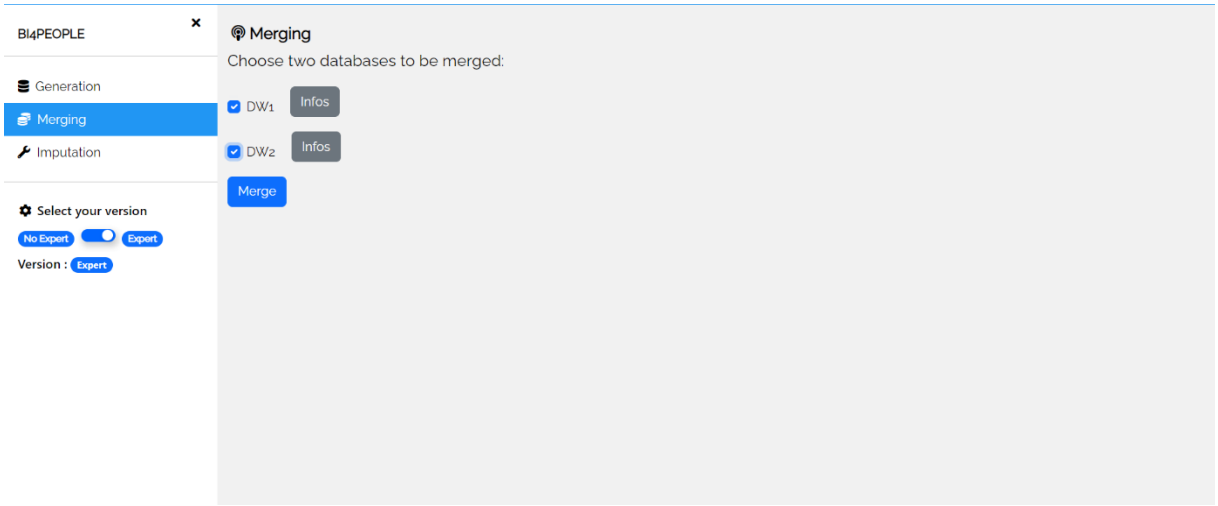


Figure 53 Fusion des Entrepôts de données-1

Cliquer sur le bouton < Infos > permettre à l'utilisateur de visualiser les informations de dimensions et de faits dans l'entrepôt de données, comme le montre la Figure 54.Une fois que l'utilisateur a terminé sa sélection, cliquez sur le bouton < Merge > pour passer à l'étape suivante.

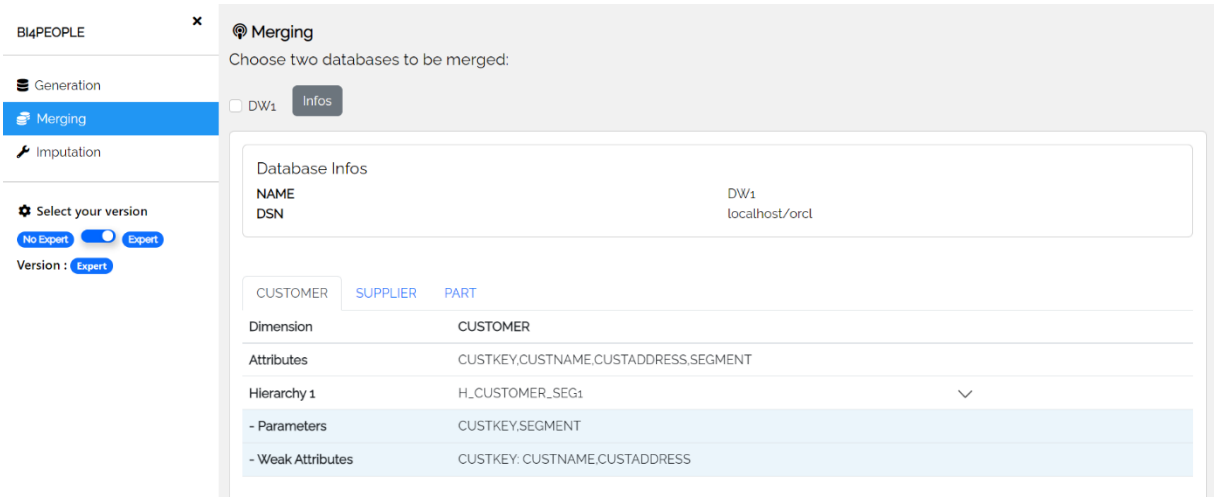


Figure 54 Fusion des entrepôts de données-2

2. Enregistrement de l'entrepôt de données

L'utilisateur clique sur le bouton < Merge > et la fenêtre pop-up de la Figure 55 apparaît, permettant d'enregistrer les données fusionnées dans un nouvel entrepôt de données.

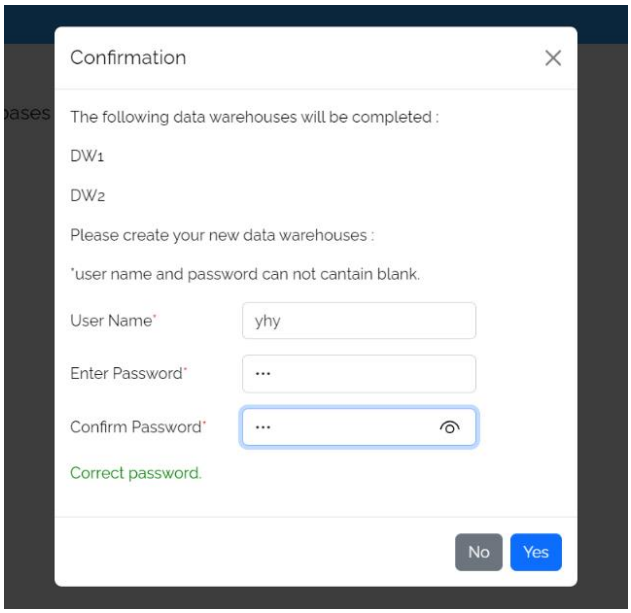


Figure 55 Fusion des entrepôts de données-3

3. Consultation et modification de résultats

La Figure 56 et Figure 56 montrent les résultats de la fusion de l'entrepôt de données. Comme pour la fonction d'entreposage automatique, l'utilisateur peut cliquer sur le bouton < Modifier > pour modifier les noms des dimensions et des hiérarchies.

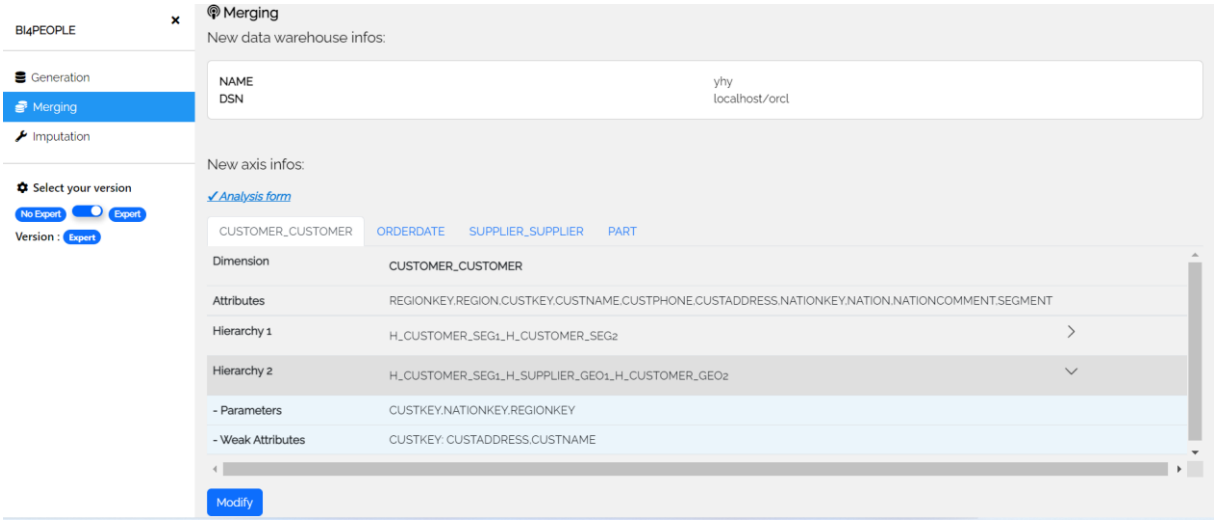


Figure 56 Fusion des entrepôts de données-4

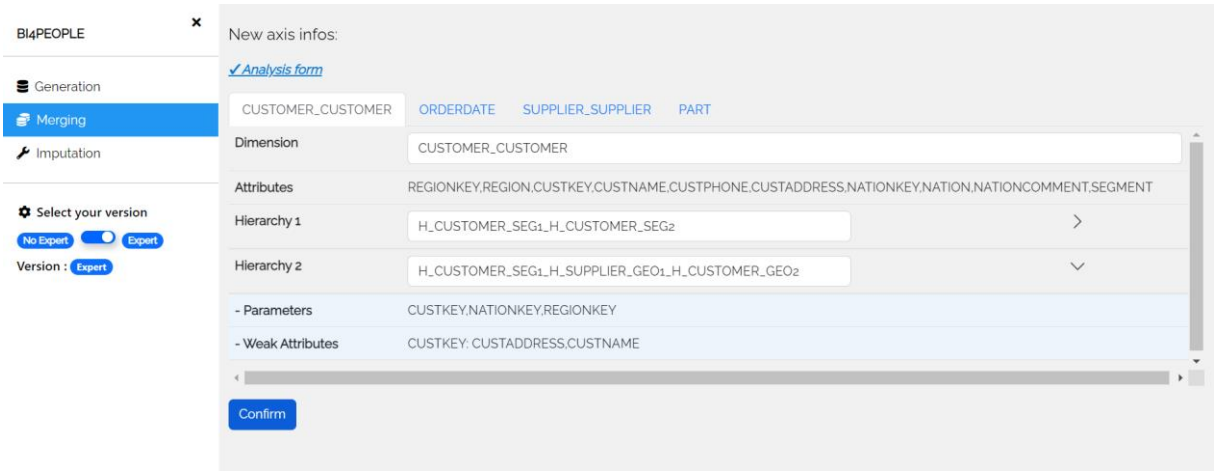


Figure 57 Fusion des entrepôts de données-5

En mode expert, l'utilisateur peut consulter des résultats de fusion plus détaillés en cliquant sur le bouton < Analysis form>. Le tableau de dimensions de l'exemple de Figure 58 comporte trois niveaux, soit un de plus qu'en mode non-expert.

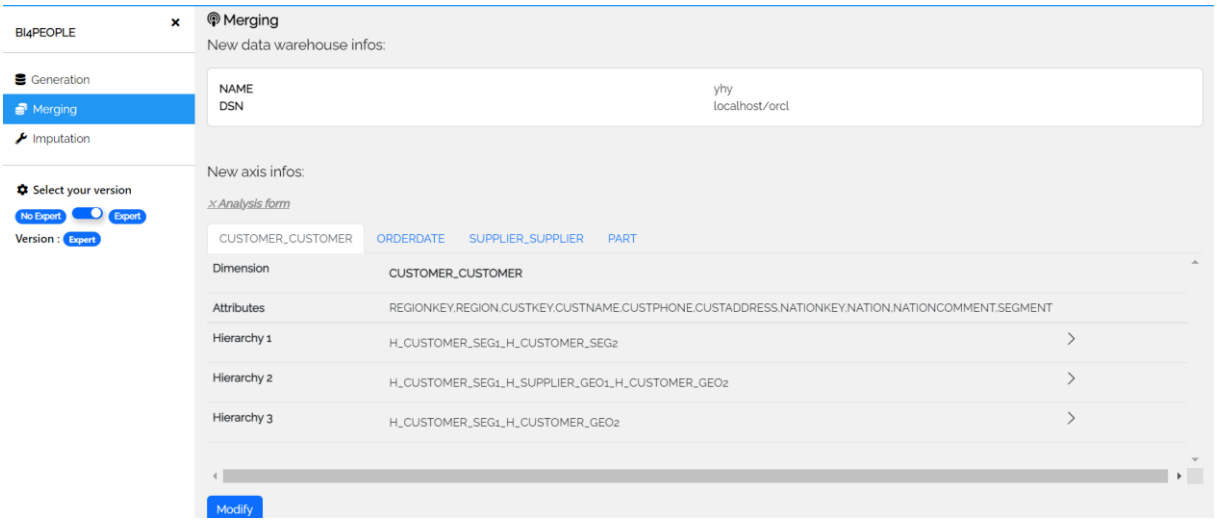


Figure 58 Fusion des entrepôts de données-6

- Imputations de données

  Nous avons mis en place la fonction d'imputation de données basée sur les maquettes, qui est présente ci-dessous.

  1. Sélection de l'entrepôt de données

  La liste déroulante de la Figure 59 permet à l'utilisateur de sélectionner l'entrepôt de données qu'il souhaite imputer. Les entrepôts de données qui ont été fusionnées ou générées seront également dans la liste.

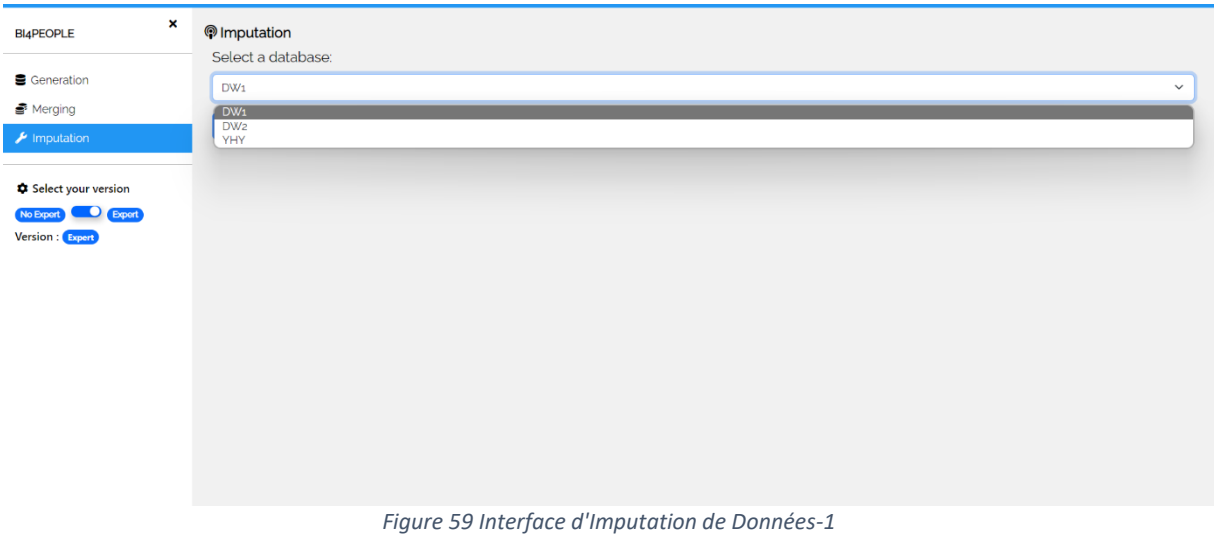


Figure 59 Interface d'Imputation de Données-1

2. Sélection d'attributs à imputer

Après avoir sélectionné l'entrepôt de données, l'utilisateur est redirigé vers la page montrée à la Figure 60, qui présente des informations sur les différentes dimensions et les attributs des données manquantes. Les utilisateurs peuvent sélectionner les attributs d'une ou plusieurs dimensions qu'ils souhaitent imputer.

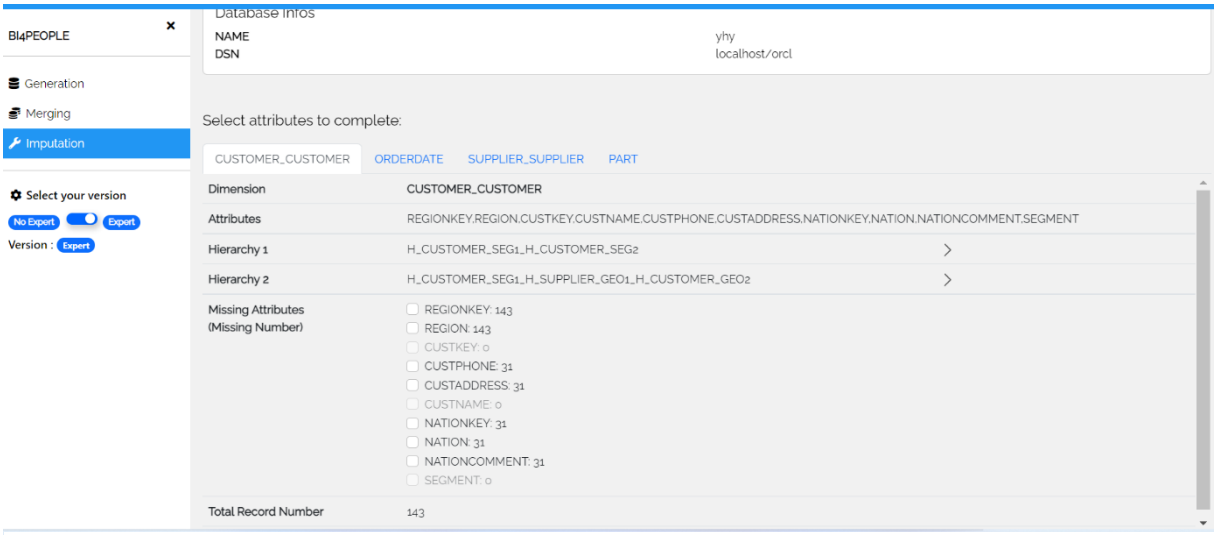


Figure 60 Interface d'Imputation de Données-2

3. Choix d'algorithmes

Après avoir confirmé les attributs à imputer, l'utilisateur peut choisir la méthode d'imputation. La Figure 62 montre la page pour les experts sélectionnent l'algorithme, les deux options correspondant aux deux méthodes d'imputation

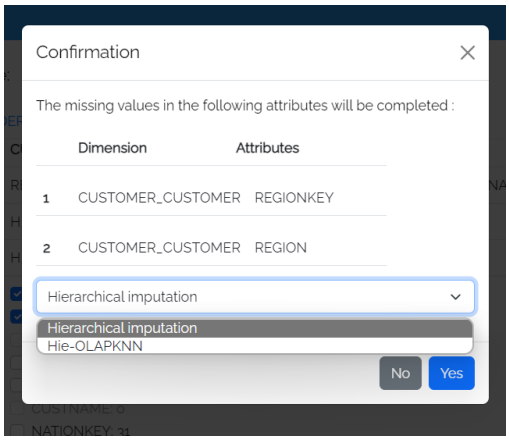


Figure 61 Interface d'Imputation de Données-3

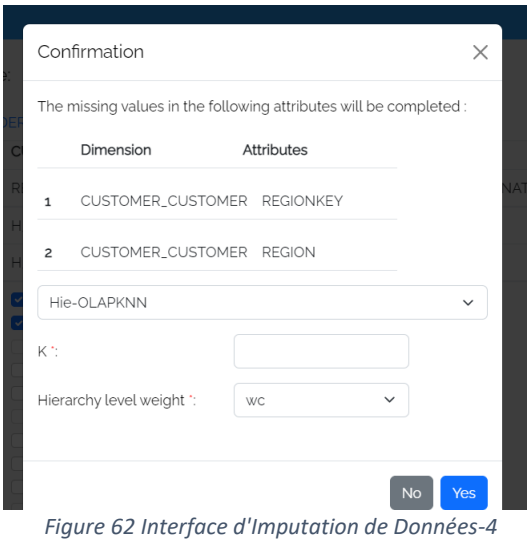


Figure 62 Interface d'Imputation de Données-4

différentes proposées par M. Yuzhao YANG. Lorsque l'utilisateur expert sélectionne la méthode < Hie-OLAPKNN >, la valeur K et la méthode de pondération de hiérarchie doivent être saisies, comme dans la Figure 61.

La Figure 63 montre l'interface pour un utilisateur non-expert, où nous avons remplacé le nom professionnel de l'algorithme par un texte facile à comprendre. De plus, l'utilisateur n'a pas besoin de saisir la valeur de K et la pondération.

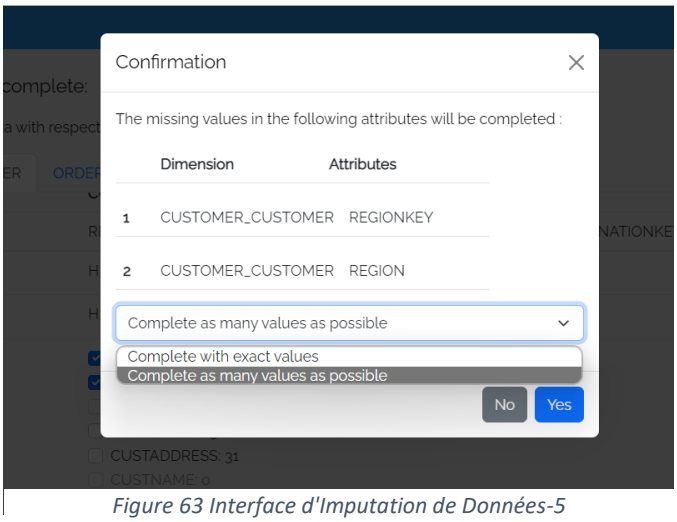


Figure 63 Interface d'Imputation de Données-5

4. Consultation de résultats

Enfin, nous montrerons à l'utilisateur le résultat final, comme le présent la Figure 64. Considérant que les utilisateurs peuvent sélectionner des attributs sur plus d'une dimension, nous avons ajouté des informations sur les dimensions. L'utilisateur peut donc voir combien d'attributs étaient auparavant manquants pour un attribut d'une dimension dans l'entrepôt de données, combien de données ont été imputées, et le taux d'imputation.

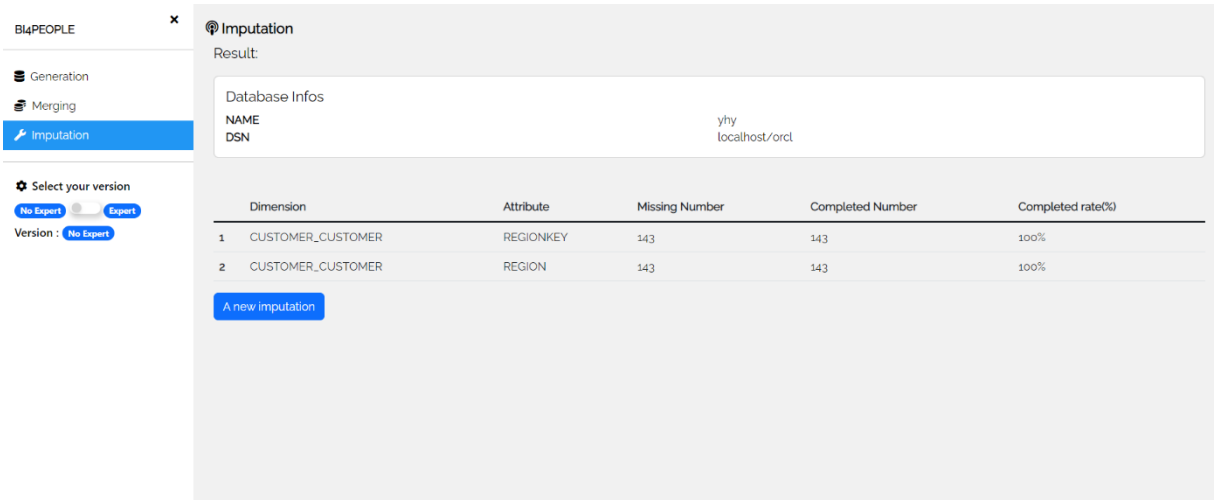


Figure 64 Interface d'Imputation de Données-5

5. Bilan

Ce stage a été pour moi la première expérience professionnelle dans le domaine de l'informatique. Concernant ce rapport, je vais résumer ce stage de six mois en deux parties en termes d'apports pour l'IRIT et d'apports personnels. Les apports personnels seront divisés en trois sections : technique, professionnel et personnel.

5.1. Apports pour l'IRIT

Ma principale contribution à l'IRIT est le développement d'une application web pour présenter les résultats du projet M. Yuzhao YANG. La fonctionnalité de l'application web fait également partie d'un système de BI qui sera développé dans le cadre du projet collaboratif BI4PEOPLE de l'IRIT. L'application peut donc être utilisée comme base pour le développement futur du système. À long terme, l'objectif est d'intégrer le développement réalisé dans un système de BI permettant la transformation de données tabulaires en données multidimensionnelles et leur stockage dans un entrepôt de données.

Deuxièmement, les expérimentations que j'ai faites ont aussi aidé M. Yuzhao YANG à valider l'algorithme de recherche qu'il a proposée et à enrichir les données expérimentales pour sa thèse.

Les résultats de ces deux parties ont permis à M. YuZhao YANG de terminer sa thèse plus rapidement.

5.2. Apports personnels et les difficultés

5.2.1. Technique

En termes de technique, j'ai pu apprendre un grand nombre de nouvelles connaissances techniques grâce à ce stage.

Tout d'abord, en ce qui concerne les langages de programmation, mes compétences en programmation python et JavaScript se sont améliorées.

- Python

  Avant ce stage, je n'utilisais Python que pour mettre en œuvre des algorithmiques de base telles que les boucles <for> et les conditions <if>. Pendant mon stage, j'ai utilisé différentes bibliothèques de Python pour effectuer plus rapidement des calculs sur des jeux de données. La programmation orientée objet a été également utilisée pour réaliser les expérimentations. Ce faisant, j'ai découvert la puissance et la richesse de Python.

- JavaScript

  J'ai dû apprendre et utiliser le langage JavaScript, car notre projet spécifiait l'utilisation de Node.js comme environnement d'exécution pour le développement d'applications web. Nous avons par ailleurs utilisé la bibliothèque jQuery pour animer de pages web, qui était aussi basée sur JavaScript. Dans le processus d'apprentissage et de pratique continus, j'ai amélioré mes compétences en programmation dans le langage JavaScript.

En outre, j'ai acquis une meilleure compréhension de connaissances techniques d'applications web.

Pendant les cours de M2, nous avons appris la structure et les connaissances des applications web et développé des applications simples. Pour apprendre Node.js, il est important de comprendre ce qu'est Node.js et comment construire des applications avec. J'ai appris en révisant les cours de programmation web, en regardant des vidéos sur YouTube et en consultant la documentation officielle.

Grâce aux études, j'ai trouvé que la logique sous-jacente est la même, quel que soit le langage ou l'environnement d'exécution. Une application web doit envoyer une requête depuis le frontend, la recevoir du backend, la traiter et enfin la renvoyer au frontend. Ce n'est qu'alors que l'utilisateur peut recevoir correctement les informations de la page web.

5.2.2. Professionnel

D'un point de vue professionnel, le stage m'a permis de mieux comprendre le domaine de recherche et d'avoir une plus grande admiration pour les personnes qui la font. Les apports professionnels se répartissent dans les trois parties suivantes : compréhension de l'article scientifique, recherche de jeux de données, communication.

- Compréhension de l'article scientifique

La lecture de différents d'articles a amélioré ma capacité à comprendre les articles scientifiques sur le domaine de données.

Quand j'ai commencé à comprendre l'article sur l'imputation de données, c'était difficile pour moi. Je n'ai pas compris le contexte, les formules et les résultats expérimentaux. Grâce à la patience de M. Yuzhao YANG qui m'a expliqué et analysé les algorithmes de la thèse, j'ai progressivement compris les articles.

Au fur et à mesure que la quantité d'article scientifique lue augmente, sa structure générale devient plus claire. J'ai trouvé mon propre ordre de lecture de l'article en fonction de l'objectif de mon travail expérimental.

Tout d'abord, je lis le résumé et le titre de l'article. Il est important de rechercher les méthodes d'apprentissage automatique qu'il écrit ou tout autre terme que je ne

comprends pas et de les comprendre. Parce qu'ils sont la partie fondamentale de l'article. J'ai passé ensuite l'introduction et les sections relatives au contexte et lit la section sur l'algorithme proposé et ses notes explicatives. Cette partie est le cœur de l'algorithme dans l'article et la partie que j'ai essayé de comprendre et d'implémenter en code. Le reste de l'article peut être lu, mais cela ne prend pas beaucoup de temps.

- Recherche de jeux de données

J'ai amélioré mes compétences de recherche de jeux de données pour l'expérimentation.

Lorsque j'ai commencé à trouver des jeux de données, j'ai d'abord cherché des sites web proposant différents ensembles de données. Ensuite, j'ai recherché sur chaque site les mots-clés auxquels je pouvais penser sur le sujet, je les ai ouverts ou téléchargés et j'ai vérifié si les données correspondaient aux demandes. Mais finalement, j'ai trouvé très peu de jeux de données.

Il m'est alors venu à l'esprit que je pourrais inverser le processus et rechercher simplement les demandes de données. Dans nos expérimentations, notre principale demande pour le jeu de données était que les données correspondent à un modèle multidimensionnel. Enfin, dans les résultats de la recherche, j'ai trouvé un site web qui recueillait des données relationnelles. Finalement, quatre d'entre eux ont été sélectionnés pour l'expérimentation.

Le processus de recherche dans les jeux de données m'a appris à ne pas m'en tenir à une seule approche, mais à sortir des sentiers battus et à essayer différentes approches du problème.

- Communication

Pendant mon stage, j'ai appris à communiquer et à demander des conseils lorsque je rencontre les difficultés.

C'est M. Franck RAVAT qui me l'a suggéré lors d'une réunion avec lui. À ce moment-là, j'étais bloqué sur un problème technique durant deux ou trois jours. Lors de la réunion, j'ai mentionné mon problème et M. Franck RAVAT a suggéré que je peux demander aux autres dans le bureau. Après la réunion, j'ai immédiatement cherché d'autre doctorat pour résoudre le problème et discuté avec M. Yuzhao YANG, et j'ai finalement résolu le problème le jour même.

Après cela, je me suis donné un délai d'un jour maximum pour réfléchir à une question par moi-même. Si cela prend plus d'un jour, je trouverai quelqu'un d'autre à qui poser la question et discuter de la difficulté. Je remercie également le sympathique de tous les membres du bureau et la patience avec laquelle ils m'aident lorsque je pose des questions. Enfin, je ne suis pas resté bloqué sur un seul problème pendant des jours.

5.2.3. Personnel

D'un point de vue personnel, ce stage m'a permis de ne plus avoir peur d'apprendre de nouvelles technologies et connaissances, et d'améliorer ma capacité d'autonomie. En repensant à ces six derniers mois, je ne m'attendais pas à découvrir autant de nouvelles connaissances.

Apprendre Node.js a été difficile au début, principalement parce que je ne savais pas comment faire à commencer. La documentation et les vidéos d'instruction en ligne sont si longues qu'il est impossible de les lire et de les apprendre toutes avant de développer. Il n'est pas non plus possible de demander à quelqu'un de me l'expliquer directement dès le début.

Depuis, j'ai réduit mes objectifs, passant de l'apprentissage de Node.js au développement d'applications web avec Node.js. J'ai cherché ensuite des vidéos d'instruction pertinentes, qui ne sont généralement pas très longues. Après avoir suivi la vidéo pour créer une application web simple, j'ai appris à connaître les composants de base et le fonctionnement sur l'application web de Node.js. J'ai ensuite commencé à développer notre application web et à chercher des solutions lorsque je rencontrais des problèmes spécifiques.

Cette méthode consistant à restreindre mes objectifs d'apprentissage ne me permettra pas de tout apprendre, mais c'est un bon moyen de résoudre un problème ou d'apprendre une compétence par moi-même. Cela m'a aussi donné un sentiment d'accomplissement parce que je pouvais voir les résultats de ma pratique très rapidement en cours de route.

6. Conclusion

Ce stage est une opportunité importante pour moi, qui me permet d'acquérir une expérience riche et précieuse afin de faciliter la transition entre ma carrière universitaire et ma carrière professionnelle.

Au cours de ce stage, j'ai participé à un projet de recherche, en concevant et en réalisant des expérimentations scientifiques pour la première fois. J'ai également appris et développé une application web en utilisant l'environnement d'exécution Node.js.

Ce stage a amélioré mes compétences professionnelles de communication et d'autonomie. Il a aussi amélioré mes compétences techniques, mes compétences en programmation en python et en JavaScript. Ces compétences améliorées pourraient m'aider à trouver un emploi dans le domaine informatique à l'avenir.

Enfin, je tiens à exprimer que je suis très heureuse de participer à ce stage et je tiens à remercier toutes les personnes qui m'ont aidé pendant mon stage et grâce auxquelles mon stage a été effectué avec succès.

Référence

[1] Yang, Y., Abdelhedi, F., Darmont, J., Ravat, F., & Teste, O. (2021). Internal Data Imputation in Data Warehouse Dimensions. In International Conference on Database and Expert Systems Applications (pp. 237-244). Springer, Cham.

[2] Yang, Y., Abdelhédi, F., Darmont, J., Ravat, F., & Teste, O. (2022). Automatic Machine Learning-Based OLAP Measure Detection for Tabular Data. In International Conference on Big Data Analytics and Knowledge Discovery (pp. 173-188). Springer, Cham.

[3] Yang, Y., Darmont, J., Ravat, F., & Teste, O. (2021). An Automatic Schema-Instance Approach for Merging Multidimensional Data Warehouses. In 25th International Database Engineering & Applications Symposium (pp. 232-241).

[4] Yang, Y., Darmont, J., Ravat, F., & Teste, O. (2022). Dimensional Data KNN-Based Imputation. In European Conference on Advances in Databases and Information Systems (pp. 315-329). Springer, Cham.

[5] Yang, Y., Darmont, J., Ravat, F., Teste, O.: Automatic Integration Issues of Tabular Data for On-Line Analysis Processing. In: 16e journées EDA Business Intelligence & Big Data (EDA 2020). vol. B-16, pp. 5–18 (2020)

[6] F Ravat, O Teste, R Tournier, G Zurfluh. (2011). Multidimensional Database Design from Document-Centric XML Documents. In: Cuzzocrea, A., Dayal, U. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2011. Lecture Notes in Computer Science, vol 6862. Springer, Berlin, Heidelberg.

[7] Ravat, F., Teste, O., Tournier, R., Zurfluh, G. (2009). Designing and Implementing OLAP Systems from XML Documents. In: Kozielski, S., Wrembel, R. (eds) New Trends in Data Warehousing and Data Analysis. Annals of Information Systems, vol 3. Springer, Boston, MA.

[8] Watson, H. J. and Wixom, B. H. (2007). The current state of business intelligence. Computer, 40(9):96–99.

[9] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.

[10] Roman, D., Dimitrov, M., Nikolov, N., Putlier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Simov, A., and Petkov, Y. (2016). Datagraft: Simplifying open data publishing. In European Semantic Web Conference, pages 101–106. Springer.

[11] Domeniconi, C. and Yan, B. (2004). Nearest neighbor ensemble. In ICPR, volume 1.

[12] Garcia, A. J. and Hruschka, E. R. (2005). Naive bayes as an imputation tool for classification problems. In Fifth International Conference on Hybrid Intelligent Systems (HIS'05), pages 3–pp. IEEE.

[13] Wu, S., Feng, X., Han, Y., and Wang, Q. (2012). Missing categorical data imputation approach based on similarity. In 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 2827–2832. IEEE.