Hackathon FORSA TIC

Thématique 1 : Intelligence Artificielle & Automatisation

Rapport Technique

Tâche 3 : Intégration des Données

Équipe : 511kinderheim

Période : 11–13 décembre 2025

Lieu : La librairie du pôle technologique de Sidi Abdellah, Algérie

13 décembre 2025

Chapitre 1

Contexte et objectif

Le Front Office traite quotidiennement des demandes variées et doit consulter de nombreuses offres commerciales ainsi que des conventions signées, ce qui rend la recherche d'information lente et sujette aux erreurs. L'objectif global du hackathon est de construire une plateforme intelligente tout-en-un, avec un chatbot IA spécialisé comme composant central.

1.1 Positionnement de la Tâche 3

La **Tâche 3 (Intégration des Données)** consiste à intégrer et unifier plusieurs sources (offres, conventions, partenaires et informations issues de documents non structurés) afin d'obtenir une base de données exploitable par la recherche avancée et le chatbot.

1.1.1 Objectifs opérationnels

- Nettoyer et structurer les données issues des documents.
- Créer des liens entre *offres* et *conventions*.
- Réaliser une annotation et une catégorisation automatique (secteur, type d'offre, etc.).

1.2 Étapes de traitement des données

La première étape de l'intégration des données a consisté à traiter les fichiers sources qui étaient dans un format non structuré. Ces fichiers présentaient plusieurs défis :

- **Données non structurées et doublons** : Les fichiers étaient souvent mal formatés, avec des doublons présents sous différents noms de fichiers. Les versions en arabe et en français étaient séparées et avaient des noms différents, ce qui compliquait l'organisation et l'accès aux documents.
- **Fichiers dans des formats erronés** : Certains fichiers étaient au format image ou dans des formats de tableurs incompatibles avec notre système de traitement, ce qui a nécessité une conversion pour pouvoir les exploiter.
- **Incohérences entre les versions arabes et françaises** : Les fichiers des deux langues n'étaient pas toujours alignés et avaient des divergences, notamment en termes de noms de fichiers et de structure des tables.

1.2.1 Résolution des problèmes de formatage et de doublons

Pour résoudre ces problèmes, plusieurs étapes ont été suivies :

- **Réparation et standardisation des formats** : Nous avons d'abord uniformisé les formats de fichiers en les convertissant dans un format standardisé compatible avec notre processus (principalement des fichiers PDF ou DOCX). Les fichiers contenant des tableaux mal formatés ont été réparés pour garantir la lisibilité et la cohérence des données.
- **Gestion des doublons** : Nous avons identifié les doublons en comparant les versions françaises et arabes des fichiers. Parfois, les fichiers étaient des duplicata avec des différences mineures de mise en forme, mais les données étaient identiques. Nous avons donc fusionné ces fichiers sous un même nom normalisé.
- **Organisation des fichiers** : Une fois les doublons supprimés, nous avons réorganisé les fichiers pour garantir que chaque type de document (offres, conventions, guides, etc.) avait une structure homogène. Chaque document a été lié à sa version en arabe et en français avec des noms de fichiers normalisés.

1.2.2 Structuration des données et conversion en fichiers JSON

Après avoir nettoyé les fichiers et organisé les données, nous avons procédé à leur structuration afin de les rendre exploitables par le système. Chaque type de fichier (offres, conventions, dépôts, guides) a été analysé pour en extraire les informations pertinentes. Ce processus a été effectué de la manière suivante :

- **Extraction des informations pertinentes** : Pour chaque fichier, nous avons identifié les informations les plus significatives et pertinentes pour le chatbot, telles que les prix, les partenaires, les types d'offres, la période de validité, les restrictions, etc.
- **Création de fichiers JSON** : Chaque fichier a été converti en un fichier JSON. Ce format permet de structurer les informations sous forme de paires clé-valeur, ce qui est idéal pour la manipulation de données et leur intégration dans la base de données. Le format JSON a également été choisi pour sa facilité d'utilisation avec les systèmes modernes comme MongoDB et pour sa compatibilité avec les API de notre plateforme.
- **Schéma de données défini** : Un schéma de données a été conçu pour chaque type de fichier. Par exemple, pour les offres, le schéma contient des métadonnées (nom de l'offre, prix, partenaire associé, secteur d'activité), ainsi que des informations spécifiques à l'offre (détails, validité, restrictions). Ce schéma garantit que toutes les informations nécessaires sont extraites de manière cohérente et normalisée.

1.2.3 Support multilingue (arabe et français)

Étant donné que les documents sont disponibles en arabe et en français, nous avons mis en place un système multilingue pour garantir que les informations extraites soient disponibles dans les deux langues. Voici les étapes suivies :

- **Alignement des versions arabe et française** : Les informations contenues dans les fichiers en arabe et en français ont été alignées pour garantir qu'elles soient cohérentes entre les deux langues. Chaque fichier JSON contient des entrées pour les deux langues, permettant au chatbot de répondre dans la langue préférée de l'utilisateur.

— **Embedder multilingue** : Nous avons utilisé un embedder multilingue qui permet de représenter efficacement les informations en arabe et en français dans un format standard, facilitant ainsi leur traitement par le système de recommandation et le chatbot.

1.2.4 Stockage des données et gestion via MongoDB

Une fois les données structurées et converties en fichiers JSON, elles ont été stockées dans une base de données MongoDB. Ce choix a été fait pour plusieurs raisons :

— **Accès rapide et flexible** : MongoDB est particulièrement adapté pour le stockage de données non structurées et semi-structurées comme celles que nous avons traitées. Il permet un accès rapide aux données et une flexibilité maximale lors de la récupération des informations.

— **Scalabilité** : MongoDB est hautement scalable, ce qui permet à notre plateforme de gérer un volume croissant de données au fur et à mesure que de nouvelles offres et conventions sont ajoutées.

— **Interaction avec l'UI** : La base de données MongoDB est utilisée par l'interface utilisateur (UI) de la plateforme pour fournir des résultats de recherche en temps réel, ainsi que pour alimenter le chatbot avec les informations nécessaires.

1.2.5 Liaison des données : Liens entre offres et conventions

L'une des étapes les plus importantes du processus a été de lier les données entre elles, notamment les offres et les conventions. Cette étape est cruciale car elle permet d'assurer une recherche efficace et pertinente, tant pour les utilisateurs que pour le chatbot. Voici comment nous avons procédé :

— **Identification des relations entre les offres et les conventions** : Nous avons identifié les points de correspondance entre les offres et les conventions, tels que les partenaires communs, les types d'offres et les catégories. Ces relations ont été utilisées pour lier les documents entre eux.

— **Création de relations dans MongoDB** : Dans la base de données MongoDB, nous avons créé des liens explicites entre les offres et les conventions, en utilisant des identifiants communs (par exemple, l'identifiant du partenaire). Cela permet de lier rapidement une offre à la convention correspondante.

1.2.6 Conclusion

Grâce à l'intégration des données, nous avons réussi à structurer des fichiers auparavant non organisés et à créer un système performant et évolutif pour stocker, gérer et accéder aux données. Les informations extraites, liées et structurées sont désormais prêtes à être utilisées dans le cadre de la recherche avancée et pour alimenter le chatbot IA. Ce processus a permis de faciliter l'interaction avec les données et d'assurer une réponse rapide et précise aux demandes des utilisateurs.

Chapitre 2

Méthodologie Détailée pour l'Appariement des Offres Conventionnelles avec la Base de Données des Offres

2.1 Modèle de Données Utilisé Lors de l'Appariement

2.1.1 Structure de la Base de Données des Offres

Le fichier `offers_pron.JSON` contient plusieurs offres, chacune représentée par :
- **offer_id** : Identifiant unique de l'offre.
- **offer_name** : Le nom de l'offre.
- **Tables** : Chaque offre possède une ou plusieurs tables qui décrivent les détails de l'offre, y compris la technologie, le type d'offre et la tarification. Les tables contiennent généralement :
  - **Technologie** : Le type de service (ADSL, VDSL, Fibre, ONT, etc.).
  - **Type d'Offre** : Catégories telles que ABONNEMENT, NOUVELLE_ACQUISITION, TARIF_PREFERENTIEL.
- **Lignes** : Chaque ligne de la table contient :
  - **Débit** : La vitesse (par exemple 10 Mbps, 100 Mbps, 1 Gbps).
  - **Prix Standard** : Le prix régulier pour cette vitesse particulière.
  - **Prix Préférentiel** : Le prix réduit ou préféré.

2.1.2 Structure de la Convention

Chaque document de convention contient généralement :
- **Service/Technologie** : Des exemples incluent ADSL, Fibre, ONT, ou Idoom Fixe (Téléphonie fixe).
- **Débit** : La vitesse pour les services Internet (par exemple 10 Mbps, 1 Gbps).
- **Prix** :
  - **Prix Standard** : Ce prix est souvent étiqueté comme *tarif actuel*, *tarif grand public* ou *tarif en vigueur* dans la convention.

— **Prix Préférentiel** : Ce prix est souvent étiqueté comme *tarif conventionné*, *tarif après remise*.
— Si le débit est manquant dans la convention, la valeur est définie comme *null* pour `conv_debit`. Dans ce cas, l'appariement se concentre principalement sur la technologie et les prix.

2.2 Aperçu du Processus d'Appariement

2.2.1 MODE A — "APPARIEMENT PAR TABLE" (Lorsque la Convention Contient des Prix Standards)

Étape A1 — Extraction des Lignes de la Convention

Pour chaque convention, nous extrayons les informations suivantes :
— **conv_service** : Le nom du service (ADSL, Fibre, etc.).
— **conv_debit** : Le débit extrait (par exemple 20 Mbps, 50 Mbps, etc.). Si le débit est manquant, il est défini à *null*.
— **conv_standard_price** : Le prix standard extrait (par exemple *tarif actuel*, *tarif grand public*).
— **conv_preferential_price** : Le prix préférentiel extrait (par exemple *tarif conventionné*).

Étape A2 — Sélection des Candidats

L'appariement de la convention avec les offres se fait selon les critères suivants :
— **Correspondance de Technologie** : Si la convention mentionne *Fibre*, l'offre correspondante doit contenir une table *Fibre*.
— **Correspondance Exacte du Débit** : Si la convention a un débit (par exemple 20 Mbps), le système l'apparie avec le débit le plus proche dans les offres. Par exemple, si l'offre propose 20 Mbps et la convention mentionne 25 Mbps, le système priorisera le débit le plus proche dans la même technologie.
— **Correspondance des Prix** :
  — **Priorité 1** : Le prix standard de l'offre doit être aussi proche que possible du `conv_standard_price`.
  — **Priorité 2** : Si une correspondance exacte du prix standard n'est pas trouvée, le système considère alors le prix préférentiel le plus proche.

Étape A3 — Correspondance Lignes à Lignes

Pour chaque ligne de la convention, nous recherchons la ligne correspondante dans l'offre. Ce processus suit un système de priorité :
— **Correspondance de Technologie** : Apparier d'abord en fonction de la technologie (par exemple, *Fibre* vs *Fibre*).
— **Correspondance du Débit** : Si le débit (vitesse) existe à la fois dans la convention et l'offre, il est apparié directement. Si non, on choisit le débit le plus proche en fonction de la différence absolue en Mbps (par exemple, si la convention mentionne 100 Mbps et que l'offre la plus proche est 120 Mbps, c'est une correspondance valide).

— **Correspondance des Prix** : Apparier selon `conv_standard_price` (utiliser le prix le plus proche si une correspondance exacte n'est pas trouvée).

Étape A4 — Calcul des Métriques

Pour chaque ligne appariée, les métriques suivantes sont calculées :
— **delta_price** : La différence absolue entre le `conv_standard_price` et le `best_offer_standard_`
— **discount_amount** : La différence entre `conv_standard_price` et `conv_preferential_price`, indiquant la remise appliquée.
— **discount_percent** : Le pourcentage de réduction basé sur le prix standard.
Les formules pour ces métriques sont les suivantes :
— **delta_price** = $|conv\_standard\_price - best\_offer\_standard\_price|$
— **discount_amount** = conv_standard_price - conv_preferential_price
— **discount_percent** = $\frac{discount\_amount}{conv\_standard\_price} \times 100$

Étape A5 — Score Global

Le score global est calculé selon les critères suivants :
— **total_delta_price** : Somme de `delta_price` pour toutes les lignes appariées.
— **avg_delta_price** : Moyenne des `delta_price` pour les lignes appariées.
— **match_count_exact** : Nombre de correspondances exactes où `delta_price` = 0.
— **match_count_near** : Nombre de correspondances proches où `delta_price` ≤ 5 DA.
Règle de décision pour le MODE A :
— Choisir l'offre avec :
  — Le **minimum total_delta_price** (la plus petite différence de prix totale).
  — Si égalité, choisir l'offre avec le **plus grand match_count_exact** (le plus grand nombre de correspondances exactes).
  — Si égalité, choisir l'offre avec le **plus grand match_count_near**.
  — Si égalité, choisir l'offre avec le **plus grand mapped_rows_count** (l'offre avec le plus grand nombre de lignes appariées).

2.2.2 MODE B — "APPARIEMENT PAR DOMAINE/MOTS" (Lorsque la Convention Ne Contient Pas de Prix Standard)

Étape B1 — Extraction des Mots Clés du Domaine

De la convention, nous extrayons les mots-clés relatifs à la technologie et aux équipements, tels que : ONT, Wi-Fi 6, GPON, XGS-PON, Fibre, ADSL, VDSL, modem, etc.

Étape B2 — Extraction des Prix

Nous collectons tous les prix disponibles dans la convention, en particulier les prix préférentiels.

Étape B3 — Évaluation de Chaque Offre

Chaque offre est évaluée selon les critères suivants :

- **domain_hit_count** : Nombre de mots-clés du domaine de la convention qui correspondent avec le domaine de l'offre.
- **exact_price_hit_count** : Nombre de correspondances exactes de prix entre la convention et les offres.
- **min_price_delta** : La plus petite différence absolue entre un prix dans la convention et un prix dans les offres.

Règle de décision pour le MODE B :

- Choisir l'offre avec :
  - Le **plus grand domain_hit_count** (le plus grand nombre de mots-clés technologiques correspondants).
  - Si égalité, choisir l'offre avec le **plus grand exact_price_hit_count**.
  - Si égalité, choisir l'offre avec la plus petite **min_price_delta**.

2.3 Calcul des Métriques

Les principales métriques utilisées pour évaluer la qualité de l'appariement des offres avec les conventions sont les suivantes :

- **Delta Price** : La différence absolue entre le prix standard de la convention et celui de l'offre. Cela permet de savoir à quel point les prix diffèrent.
- **Discount Amount** et **Discount Percent** : Indiquent combien l'offre est moins chère par rapport au prix standard de la convention, ce qui aide à évaluer l'attractivité de l'offre par rapport à la convention.

Cette méthodologie forme la base de l'appariement des lignes de la convention avec les offres correspondantes, en utilisant à la fois les correspondances exactes et les correspondances proches en fonction des données disponibles dans les conventions et les offres.

Chapitre 3

Partie Deux : Apparier les Conventions aux Offres (Critères et Évaluation)

3.1 Aperçu du Processus d'Appariement

Dans la Partie Un, nous avons détaillé la méthodologie pour appairer les lignes des conventions aux offres en fonction de la technologie, du débit (vitesse), et des prix (standard et préférentiels). Les deux modes d'appariement, **MODE A** (pour les correspondances exactes des prix avec les prix standard) et **MODE B** (pour les correspondances basées sur des mots-clés/domaines, surtout lorsque seuls les prix préférentiels sont fournis), sont appliqués selon les informations disponibles dans la convention.

Maintenant, nous allons détailler l'appariement de chaque convention par rapport aux offres, en clarifiant la façon dont elles ont été comparées et si elles ont correspondu à 100

3.2 Appariement des Conventions aux Offres

En fonction des conventions fournies et de leurs offres correspondantes, nous évaluons chaque cas.

3.2.1 2.1 AT_DG_DCMI_DCGP_CQC_RA_6_0_FR_159_2025 — Nouvelle offre résidentielle Idoom fibre, ADSL et VDSL (V6)

Offres appariées :
— `AT_DG_DCMI_DCGP_CQC_RA_6_0_FR_159_2025` (offre principale)
Conventions appariées (Correspondances exactes : 100%) :
— Convention AT & L'établissement N : Correspondance exacte trouvée pour toutes les lignes.
— Convention AT & L'établissement L : Toutes les technologies correspondent, y compris ADSL, Fibre, VDSL.
— Convention AT & L'établissement M : Correspondance parfaite, mêmes valeurs de débit et prix standard.

— Convention Algérie Télécom & L'établissement AD : Correspondances exactes dans les détails des prix et de la technologie.
— Convention AT & L'établissement O : Correspondance complète sur les services, débits et prix.
— Convention AT & L'établissement R : Même offre avec toutes les conditions remplies (débit, prix, technologie).
— Convention AT & L'établissement T : Correspondance parfaite sur toutes les lignes pour Fibre et ADSL avec débits et prix corrects.
— Convention AT & L'établissement U : Pas de discordance trouvée pour Fibre et VDSL.
— Convention Algérie Télécom & L'établissement X : Toutes les correspondances exactes pour les débits (20 Mbps, 50 Mbps, 100 Mbps).
— Convention Algérie Télécom & L'établissement AC : Pas de différence dans les débits ou les prix.
— Convention AT & L'établissement V : Service et prix identiques pour toutes les lignes.
— Convention AT & L'établissement W : Correspondance complète avec débits et prix standard exacts.
— Convention AT & L'établissement A : Correspondance 100% sur toutes les lignes technologiques.
— Convention AT & L'établissement C : Correspondance exacte pour toutes les lignes Fibre et VDSL.
— Convention AT & L'établissement D : Correspondance complète sans discordance.
— Convention AT & L'établissement F : Offre identique avec les bons appariements de débits.
— Convention AT & L'établissement G : Correspondance totale confirmée, y compris les prix exacts.
— Convention AT & L'établissement H : Correspondance complète avec des prix exacts sur tous les services.
— Convention AT & L'établissement I : Correspondance exacte sur toutes les lignes.
— Convention AT & L'établissement J : Correspondance totale sur toutes les technologies et tranches de prix.
— Convention AT & L'établissement k : Correspondance complète pour les débits et services.

Résumé des correspondances :

100% de correspondance sur toutes les lignes dans 16 conventions (débits, prix standard et préférentiels identiques sur tous les types de technologies).

3.2.2 2.2 AT_DG_DCMI_DCGP_CQC_SR_V3_0_FR_135_2025 — Nouvelle Offre MOOHTARIF (V3)

Offres appariées :

— `AT_DG_DCMI_DCGP_CQC_SR_V3_0_FR_135_2025` (offre principale)

Conventions appariées :

— Convention Algérie Télécom & L'établissement AB : Correspondance parfaite pour la technologie ADSL. 100% de correspondance pour le débit et le prix.
— Convention AT & L'établissement S : Correspondance partielle pour les technologies VDSL et Fibre. Correspondance partielle pour les débits VDSL et Fibre.

— Convention Algérie Télécom & L'établissement Z : Services ADSL et Fibre identifiés, correspondances exactes pour le débit et le prix.
— Convention AT & L'établissement E : Correspondance partielle pour les services et prix VDSL.

Résumé des correspondances :
3 conventions avec 100% de correspondance.
1 convention avec une correspondance partielle.

3.2.3 2.3 AT_DG_DCMI_DCMP_CQC_BI_V1_0_FR_180_2025 — Tarif préférentiel sur ONT Wi-Fi 6

Offre appariée :
— `AT_DG_DCMI_DCMP_CQC_BI_V1_0_FR_180_2025` (offre principale)

Convention appariée :
— Convention AT & L'établissement P : Correspondance sur ONT Wi-Fi 6. Correspondance exacte pour le prix (3500 DA). 100% de correspondance.

Résumé des correspondances :
1 convention avec 100% de correspondance.

3.2.4 2.4 AT_DG_DCMI_DCMP_CQC_BI_4_0_FR_160_2025 — Nouvelle offre résidentielle LOCATAIRE Idoom fibre, ADSL et VDSL

Offre appariée :
— `AT_DG_DCMI_DCMP_CQC_BI_4_0_FR_160_2025` (offre principale)

Conventions appariées :
— Convention AT & L'établissement N : Correspondance parfaite avec les technologies Fibre, ADSL et VDSL.
— Convention AT & L'établissement L : Correspondance pour les services Fibre et VDSL avec les prix exacts.
— Convention AT & L'établissement B : Correspondance exacte pour 20 Mbps, 50 Mbps et 100 Mbps.
— Convention Algérie Télécom & L'établissement X : Correspondance parfaite pour les services Fibre et VDSL.
— Convention AT & L'établissement O : Correspondance exacte pour toutes les lignes avec les débits ADSL et Fibre, ainsi que les prix.

Résumé des correspondances :
100% de correspondance sur toutes les lignes dans 5 conventions.

3.3 Évaluation Générale des Critères d'Appariement

3.3.1 Correspondances Exactes

La majorité des conventions ont trouvé une correspondance exacte, en particulier lorsque les prix standards étaient présents. En **MODE A**, les correspondances étaient basées sur l'alignement exact des débits et des prix.

3.3.2 Correspondances Partielles

Certaines conventions comportaient des tables partielles (seules quelques lignes), ce qui a conduit à une certaine incertitude. En **MODE A**, ces correspondances partielles avaient un `total_delta_price` compris entre 1000 et 2500 DA en fonction de la convention, tandis que les libellés de débit incomplets ou des différences de libellés ont été identifiés comme sources possibles de discordance.

3.3.3 Correspondances par Domaines/Mots

Le **MODE B** a montré une forte correspondance de domaine lorsque la convention utilisait des mots-clés spécifiques comme *ONT Wi-Fi 6*, mais l'offre ne couvrait pas tous les prix attendus (par exemple, l'absence du prix de 2000 DA).

3.4 Résumé de la Partie Deux

3.4.1 Offres Correspondantes

L'offre principale correspond à de nombreuses conventions exactement, en particulier pour les technologies Fibre, ADSL et VDSL.

3.4.2 Précision des Correspondances

Les correspondances exactes étaient fréquentes, surtout lorsque les prix standard et préférentiels étaient inclus dans les conventions.

Les correspondances partielles ont montré des incertitudes, notamment lorsque les conventions utilisaient un autre échelon de prix ou manquaient de débits spécifiques.

3.4.3 Métriques Clés

Les correspondances exactes des débits étaient courantes, avec de nombreuses offres fournissant des prix dans un petit delta par rapport au prix standard de la convention.

Lorsque les prix étaient manquants (par exemple, uniquement les prix préférentiels), le **MODE B** a géré l'appariement en utilisant la correspondance de domaine et la différence de prix la plus proche.