Résumé structuré : Vérification de la qualité des données

1. Définition et objectifs

La vérification de la qualité des données consiste à évaluer et améliorer la **précision**, **exhaustivité**, **cohérence** et **actualité** des données. Cela vise à :

- Garantir des analyses fiables.
- Soutenir les décisions stratégiques avec des données précises.
- Réduire les coûts liés aux mauvaises données (ex. : estimation IBM 2016 : 3 trillions USD/an aux USA).

2. Problèmes courants de qualité des données

1. **Précision** :
    a. Enregistrements dupliqués.
    b. Fautes de frappe et données mal alignées.
    c. Exemples : valeurs aberrantes ou erreurs lors de la migration de fichiers CSV.
2. **Exhaustivité** :
    a. Données manquantes (ex. : champs vides, placeholders).
    b. Pertes de données dues à des erreurs système.
3. **Cohérence** :
    a. Formats incompatibles (ex. : date en *AAAA-MM-JJ* vs *MM-JJ-AAAA*).
    b. Variations dans les unités (kg vs lbs).
    c. Différences dans les noms (ex. : *John Doe* et *M. John Doe*).
4. **Actualité** :
    a. Données obsolètes (ex. : adresses ou noms non mis à jour).
    b. Besoin de synchronisation avec des bases de données externes.

3. Processus de traitement des données erronées

1. **Identification** :
    a. Règles pour détecter les erreurs (ex. : scripts SQL).
    b. Exemples : données dupliquées, valeurs hors plage, champs manquants.
2. **Correction** :

    a. Automatisation des transformations (ex. : suppression des lignes problématiques).
    b. Application des correctifs lors des chargements de données.
3. **Rapports** :
    a. Signalement des problèmes non résolus pour correction manuelle.
4. **Prévention** :
    a. Validation en amont dans la zone de transit avant l'analyse finale.

4. Outils de gestion de la qualité des données

Quelques solutions :

- **IBM InfoSphere Information Server for Data Quality** :
    - Surveillance et nettoyage continus.
    - Normalisation, association et suivi du lignage des données.
- **Alternatives** : Informatica, Talend, OpenRefine, Microsoft Data Quality Services.

5. Avantages

- Intégration réussie des relations complexes.
- Analyse avancée et apprentissage automatique facilités.
- Confiance accrue dans les décisions stratégiques.