

Automatiser le contrôle de
qualité de vos **données** avec la
méthode **2QC+**

2QC+ c'est deux couches de contrôle automatisé dans vos pipelines pour détecter et traiter les problèmes avant qu'ils n'impactent vos décisions.

"La donnée fiable se cultive. Elle n'arrive jamais par hasard."

BY

Kheop**Sys** 



Au mieux, vos décisions valent ce que valent vos données

Dans un contexte où les entreprises ingèrent chaque jour des volumes de données exponentiels, la qualité de ces données est devenue un enjeu critique. Elle conditionne directement la pertinence des décisions, l'efficacité opérationnelle et la compétitivité.

Pourtant, dans la réalité, peu d'entreprises traitent véritablement la qualité de leurs données. Les contrôles, lorsqu'ils existent, sont souvent ponctuels et vite considérés comme acquis, alors même que les systèmes producteurs et les traitements évoluent en permanence.

Face à cette réalité, l'automatisation devient indispensable pour garantir la fiabilité.

Automatisez le contrôle qualité de vos données... et dormez sur vos deux oreilles

Au cours des quinze dernières années, j'ai eu l'opportunité d'accompagner des entreprises de secteurs variés (voyage, automobile, grande distribution, gaming...) dans la conception et le déploiement de produits data stratégiques et fiables.

Ce parcours m'a permis de développer des automatismes et des pratiques adaptées aux enjeux de qualité des données et des traitements associés.

De cette expérience est née **2QC+ (Two-Level Quality Control)**, une méthode structurée qui allie la rigueur de l'expertise métier à la puissance de l'intelligence adaptative.

Contrairement à de nombreuses solutions du marché qui misent sur un discours « **tout machine learning** » ou « **tout IA** », 2QC+ adopte un équilibre pragmatique : des **règles déterministes** robustes, complétées par de **l'intelligence adaptative** (Machine Learning / IA).

Cette approche s'appuie sur deux couches de contrôle automatisé, capables de détecter les problèmes de qualité des données et d'alerter avant que ces derniers ne compromettent les décisions stratégiques.



01

Les enjeux de la qualité des données aujourd'hui

- A. Volume, vitesse et variété croissants.....P.3
- B. Sophistication des usages.....P.3

02

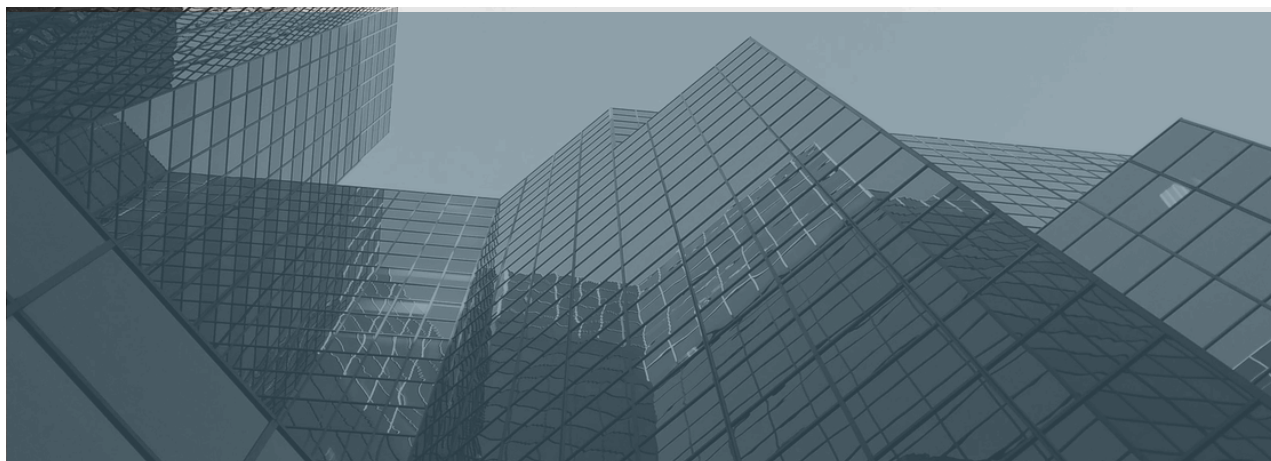
Limites des approches traditionnelles

- A. Contrôles manuels : un modèle dépassé.....P.4
- B. Test basé sur des règles : une approche incomplète.....P.4
- C. Monitoring des métriques : une vision superficielle.....P.4

03

La méthode 2QC+

- A. Comment 2QC+ résout les limites des approches traditionnelles...P7
- B. Approche 2QC+.....P8
- C. Mise en œuvre opérationnelle.....P9
- D. Quickstart avec le framework qc2plus.....P10
- E. Cas d'usage.....P11





Volume, Vitesse et Vitesse croissants

- **Déluge de données** : l'augmentation massive des données rend impossible toute vérification manuelle exhaustive. Sans détection automatique, de nombreuses anomalies passent inaperçues.
- **Multiplication des sources** : systèmes internes, APIs, IoT, données partenaires, open data, réseaux sociaux... Chaque nouvelle source introduit son propre format, sa fréquence et ses règles de gestion.
- **Temps de latence** : les métiers exigent des délais de latence toujours plus courts, parfois en quasi temps réel, ce qui laisse moins de marge pour identifier et corriger les erreurs avant consommation.



Sophistication des usages

- **Complexité croissante des traitements** : enrichissements, agrégations, modèles prédictifs... chaque transformation ajoute un risque potentiel de dégradation de la donnée.
- **Démocratisation de l'analytics** : explosion du nombre d'utilisateurs métiers accédant directement à la donnée, multiplication des tableaux de bord et indicateurs clés, exigences accrues en termes de fraîcheur et de fiabilité.
- **Intelligence artificielle omniprésente** : déploiement massif d'algorithmes de machine learning, sensibilité extrême à la qualité des données d'entraînement, amplification des erreurs par les décisions automatisées.

La mauvaise qualité des données engendre des fausses décisions stratégiques, des retards liés à la multiplication des vérifications manuelles, une perte de confiance dans les outils et un impact réputationnel lié à la défiance des clients et partenaires.

Selon Thomas Redman (MIT Sloan Management Review), ce coût peut représenter **15 à 25 % du chiffre d'affaires** d'une entreprise.



Contrôles manuels : un modèle dépassé

Les contrôles manuels sont rapidement obsolètes, incomplets, coûteux et lents. Ils sont incapables de suivre l'évolution des données. Leur couverture partielle, leur forte dépendance à l'humain et leur faible réactivité laissent passer la majorité des anomalies, avec un impact direct sur la performance et les décisions.

Si un bug génère des doublons sur plusieurs milliers de commandes dans la nuit, il peut s'écouler plusieurs jours avant qu'il soit détecté, et 99 % des transactions affectées passeront inaperçues.



Monitoring des métriques : une vision superficielle

La surveillance d'indicateurs agrégés masque les anomalies locales, ignore le contexte et repose sur des seuils statiques inadaptés aux variations réelles du business. Cette approche unidimensionnelle donne une illusion de normalité, retardant la détection des problèmes structurels.

Si la moyenne quotidienne des transactions reste stable, certaines régions peuvent ne plus recevoir de commandes tandis que d'autres connaissent un pic inhabituel. L'agrégat masque l'anomalie locale et retarde son identification.



Les règles métier : une approche incomplète

Les règles métier seules offrent clarté et intégration facile, mais leur rigidité, leur couverture limitée et leurs faux positifs les rendent inefficaces face à des données dynamiques et complexes. Elles échouent à détecter les patterns subtils, les corrélations et les anomalies multivariées, créant des angles morts critiques.

Une règle "volume de ventes +30 %" peut générer une fausse alerte lors du Black Friday. En croisant avec d'autres KPI — trafic web, ajout au panier, taux de clic — tous en hausse simultanément, on identifie que la variation est justifiée et qu'aucune anomalie n'est à signaler.

Pourquoi 2QC+ ?

Face aux enjeux de qualité des données, les organisations ont besoin d'une approche qui :

- **Automatise la détection sans sacrifier la précision**
- **S'adapte aux évolutions des données et des besoins métier**
- **Détecte les anomalies complexes et multidimensionnelles**
- **Explique les anomalies pour faciliter la résolution**
- **Intègre naturellement les processus existants**
- **Évolue avec la maturité de l'organisation**



2QC+

Philosophie fondamentale

La méthode 2QC+ repose sur le principe que la qualité des données ne peut être automatisée par une approche unique, mais nécessite une stratégie multicouche adaptée au contexte des données (saisonnalité, tendance, ...).



A - Comment 2QC+ résout les limites des approches traditionnelles



Réponse aux limites des contrôles manuels

- Automatisation : détection 7j/7 sans intervention humaine
- Couverture exhaustive : 100% des données contrôlées à chaque pipeline
- Évolution adaptative : les modèles s'ajustent aux nouveaux patterns
- Réactivité immédiate : détection en moins de 15 minutes contre plusieurs jours

Le bug de doublons serait détecté automatiquement en moins de 15 minutes via les règles de Niveau 1 (pattern de duplication), avec isolation immédiate des commandes affectées et alerte aux équipes.



Réponse aux limites du monitoring superficiel

- Granularité fine : gamme produit, région, ...
- Vision multidimensionnelle : analyse des corrélations et interdépendances
- Contextualisation : prise en compte des variations saisonnières, promotions
- Seuils dynamiques : adaptation automatique aux nouveaux patterns

L'anomalie régionale serait immédiatement détectée par le Niveau 2 (détection intelligente) qui analyserait automatiquement les distributions géographiques, identifierait l'absence totale de commandes dans certaines régions malgré un volume global stable, et générerait une alerte explicative.



Réponse aux limites des règles métier

- Réduction des faux positifs : corrélation automatique avec d'autres variables
- Apprentissage continu : les modèles apprennent des variations pour détecter les patterns complexes impossibles à formaliser

Durant le Black Friday, le Niveau 1 détecterait effectivement le volume +30%, mais le Niveau 2 analyserait automatiquement la corrélation avec trafic web (+35%), ajouts panier (+28%), et taux clic (+25%), identifierait la cohérence du pattern, et classerait l'événement comme "variation normale contextuelle" sans générer de fausse alerte.



Niveau 1: contrôles par règles déterministes

Objectif : assurer une couverture exhaustive de tous les cas identifiables par des règles explicites.

Périmètre d'action :

- Contraintes d'intégrité référentielle : cohérence entre tables et clés (ex. les produits vendus existent dans le référentiel produit, chaque contact a une ligne unique, aucune date de transaction dans le futur...).
- Règles métier formalisées : validations spécifiques aux processus (ex. un vol ne peut pas être programmé avec une date d'arrivée avant la date de départ)
- Seuils statistiques contextuels : limites définies selon la distribution historique (ex. alerte si conversion < moyenne 30j - 3σ ou si le volume de sessions web varie de $\pm 30\%$ vs J-1).
- Validations de format et de cohérence : formats, unités, et structures attendues (ex. email valide, code pays ISO, ...)
- Contrôles de fraîcheur : vérification que les données sont disponibles et mises à jour dans les délais attendus (ex. alerte si les ventes de la veille ne sont pas chargées à 9h, ou si un flux de logs temps réel ne s'actualise plus).



Niveau 2: contrôles par l'intelligente adaptative

Objectif : identifier les anomalies complexes tout en éliminant les fausses alertes du niveau 1 grâce à une analyse contextuelle intelligente

Périmètre d'action :

- Analyse des corrélations entre variables : repérer des liens pour contextualiser certaines situations (ex. pendant Black Friday, ventes +30% avec hausse équivalente du trafic et des paniers = variation normale, pas d'alerte).
- Analyse de patterns temporels : intégrer les saisonnalités et cycles naturels (ex. baisse de trafic en août cohérente avec les années précédentes = pas d'anomalie).
- Analyse des distributions par segments : détecter immédiatement les anomalies localisées (ex. absence totale de commandes dans certaines régions malgré un volume global stable = vraie anomalie géographique).

Croiser les analyses du niveau 1 et 2 avant d'alerter permet le filtrage contextuel des fausses alertes (ex. baisse des lead détectée par niveau 1, mais corrélée à la baisse du budget et des impressions = alerte filtrée automatiquement).

2QC+ est conçu pour être simple à mettre en œuvre. En quelques étapes, vous passez d'un environnement sans contrôle qualité structuré à un dispositif complet et opérationnel, illustré ci-après.



2QC+ est implémentée dans le framework **qc2plus**, Open-source, natif Python et adaptable. Il s'installe vite, s'intègre à votre stack (Airflow, Prefect, Dagster, ...) et se configure en YAML, avec un code librement modifiable.

Initialise (1), configure (2) et execute (3) 2QC+

1

```
# install 2QC+ framework
pip install qc2plus

# init 2QC+ project
qc2plus init my_data_quality_project

# go to project folder
cd my_data_quality_project
```

2

```
# profiles.yml
my_data_quality_project:
  target: dev
  outputs:
    dev:
      type: postgresql
      host: localhost
      port: 5432
      user: your_username
      password: your_password
      dbname: your_database
      schema: public
```

2

```
# models/customers.yml
models:
  - name: customers
    description: Customer data quality tests
    qc2plus_tests:
      level1:
        - unique: { column_name: customer_id, severity: critical }
        - not_null: { column_name: email, severity: critical }
        - email_format: { column_name: email, severity: medium }
        - statistical_threshold:
            metric: count
            threshold_type: relative
            threshold_value: 2.0
            severity: medium

      level2:
        correlation_analysis:
          variables: [daily_registrations, daily_activations]
          expected_correlation: 0.8
          threshold: 0.2
        temporal_analysis:
          date_column: created_at
          metrics: [count, avg_revenue]
          seasonality_check: true
        distribution_analysis:
          segments: [country, customer_type]
          metrics: [revenue, orders_count]
```

3

```
# test database connection
qc2plus test-connection --target dev

# run all tests
qc2plus run --target dev

# run specific model
qc2plus run --models customers --target dev

# run only Level 1 tests
qc2plus run --level 1 --target dev

# run with parallel execution
qc2plus run --threads 4 --target dev
```



Self-BI

Dans les organisations où les métiers consomment directement leurs KPI via des outils de BI en self-service, la moindre incohérence détectée par un utilisateur fragilise la confiance et génère des tensions.

Apport de 2QC+ :

- **Proactivité** : l'équipe data connaît déjà les anomalies avant qu'elles ne soient signalées.
- **Communication crédible** : en cas de problème, la réponse n'est plus "*on va vérifier*", mais "*nous avons détecté l'anomalie, voici l'état d'avancement de la résolution*".
- **Confiance renforcée** : le métier se concentre sur l'analyse, et non sur la remise en cause de la donnée.
- **Alignement business-data** : les alertes sont contextualisées par indicateur métier, ce qui facilite la priorisation et le dialogue avec les parties prenantes.

Résultat : Passage d'une posture réactive ("firefighting") à une posture de tiers de confiance, qui crédibilise la fonction data auprès de la direction.

Déploiement ou migration d'un système cœur (CRM, CDP, RCU, ...)

Les déploiements ou migrations sont trop souvent rythmées par des cycles fastidieux de **contrôle-correction-re-contrôle**. Résultat : des délais qui s'allongent, des équipes épuisées, et des anomalies qui subsistent faute d'exhaustivité.

Apport de 2QC+ :

- Contrôle automatisé et systématique de chaque lot déployé ou migré.
- Couverture exhaustive, au-delà de l'échantillonnage manuel.
- Automatisation de la détection, la documentation et explication des anomalies, sans surcharge inutile pour les équipes.
- Pilotage simplifié avec la production de rapports de qualité clairs qui permettent aux comités de pilotage de statuer rapidement sur l'acceptation ou le rejet d'un lot intégré ou migré.

Résultat : temps des contrôles réduit de 80%, déploiement ou migration sécurisé et exhaustif, confiance maintenue entre IT, métiers et direction.



Les modèles Machine Learning en production

Les modèles de scoring, de prévision ou d'attribution performants, sont souvent conçus avec rigueur. Cependant, ils restent vulnérables : une chute de qualité des données d'entrée suffit à déstabiliser l'ensemble, générer des décisions erronées, de la sous-performance, et surtout une perte de confiance du métier.

Apport de 2QC+ :

- **Couverture systématique** : chaque flux alimentant les modèles est contrôlé en continu, pas seulement lors d'audits ponctuels.
- **Réactivité** : les anomalies sont détectées et signalées en minutes, avant qu'elles n'impactent les résultats.
- **Explicabilité** : chaque alerte est documentée et contextualisée, permettant aux équipes de comprendre rapidement la dérive et d'agir.

Résultat : la performance prédictive est maintenue dans le temps, les incidents coûteux sont évités, et la confiance dans les systèmes Data Science reste intacte auprès des métiers et de la direction.

Audit et gouvernance dans un écosystème multi-environnements

Les organisations modernes gèrent un patrimoine data fragmenté (multi-bases, multi-clouds, environnements). Pour les équipes en charge de la gouvernance, scanner l'existant reste une tâche lourde, fastidieuse et largement manuelle.

Apport de 2QC+ :

- **Vision consolidée de la qualité** : chaque environnement est audité automatiquement sur la base de critères homogènes.
- **Détection d'écarts structurels** : incohérences entre bases de référence et bases de consommation, duplications, dérives locales.
- **Aide à la décision** : restitution centralisée dans une interface unique, offrant une vision claire et exploitable par la gouvernance.

Résultat : vision consolidée et documentée de la qualité des données. Les restitutions deviennent claires, factuelles et exploitables pour les stackholders. La gouvernance va au-delà du contrôle basique : elle s'appuie sur un diagnostic approfondi, priorise les risques, et gagne en légitimité dans l'organisation.

2QC+, une méthode née de 15 ans d'expérience terrain



”

La donnée fiable ne s'improvise pas, elle s'organise.

La méthode **2QC+** offre un cadre structuré de gestion de la qualité des données. Elle s'incarne aujourd'hui dans le framework open source **qc2plus**, pensé pour démocratiser et partager les bonnes pratiques de manière collaborative.

Une réponse concrète aux enjeux de la qualité des données

La méthode 2QC+ est le fruit de plusieurs années d'expérience, au cours desquelles j'ai développé des automatismes et des pratiques adaptés aux enjeux de la qualité des données et des traitements qui en dépendent.

Elle offre une approche robuste et pragmatique, en combinant la rigueur de règles déterministes à la flexibilité d'une intelligence contextuelle, garantissant ainsi un contrôle à la fois systématique et pertinent.

Abdoul Raoufou GAMBO
CEO Kheopsys

A propos de Kheopsys

Au travers des missions de conseil et d'expertise, nous accompagnons les organisations à bien exploiter leurs données pour accélérer leur transformation et renforcer leur avantage compétitif.

Exemples :

- Définir une stratégie data claire (organisation, priorisation, feuille de route) pour aligner les initiatives avec les enjeux business.
- Mettre en place une architecture data moderne et sécurisée (data warehouse, CDP, intégrations) afin de fiabiliser et valoriser la donnée.
- Maximiser la performance des campagnes marketing en utilisant des analyses et des algorithmes pertinents pour délivrer le bon message au bon client au bon moment
- Optimiser les parcours clients en identifiant les meilleures opportunités ou points de friction qui influencent la performance (KPIs)
- Acculturer les équipes à la culture data et IA pour renforcer l'adoption des outils et des usages.

Nous intervenons de la stratégie à l'activation opérationnelle, afin de garantir un impact concret et mesurable.



54 rue de Londres, 75008 Paris



+337 56 93 39 09



contact@kheopsys.com



www.kheopsys.com

