Université Claude Bernard Lyon 1

DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **11 avril 2025**

Nom de famille et prénom de l'auteur. e : **Monsieur Samuele LANGHI**

Titre de la thèse **:** L'intégrité des flux de données

Résumé

La qualité des données est essentielle en gestion moderne, garantissant exactitude, cohérence, exhaustivité et fiabilité. Si les bases de données statiques assurent bien l'intégrité, les environnements de flux nécessitent une validation en temps réel pour éviter la propagation d'erreurs. Cette thèse explore l'intégrité des flux de données, définie comme l'adhésion à des contraintes prédéfinies. Dans les applications en temps réel comme la finance et la surveillance énergétique, les anomalies peuvent indiquer des dysfonctionnements ou le bon déclenchement de systèmes automatisés. Il est donc essentiel de détecter et annoter les incohérences plutôt que de modifier ou supprimer les données erronées. La thèse apporte trois contributions majeures : InkStream pour l'évaluation de requêtes continues avec annotations d'incohérences, une approche basée sur des semi-anneaux pour la provenance de requêtes continues et une étude sur l'Event Processing Language (EPL). InkStream propose un cadre d'annotation des incohérences dans les requêtes continues, adaptant la provenance aux flux via des semi-anneaux polynomiaux. L'approche inclut des modifications spécifiques aux fenêtres pour garantir la faisabilité computationnelle. Une innovation clé est le Consistency Graph Summary, optimisant la propagation des annotations par propriétés transitives, réduisant ainsi la complexité. Une algèbre positive pour flux étend les opérations pour intégrer les annotations de cohérence dans les flux relationnels. Les évaluations montrent une amélioration des performances jusqu'à 80% et une surcharge minime. La thèse introduit également le When-Provenance, étendant les modèles de provenance avec des intervalles temporels pour suivre la dérivation des données. Les opérateurs de fenêtre divisent les flux infinis en segments finis. Le choix de la taille optimale des fenêtres est crucial : trop petites, elles omettent des données pertinentes, trop grandes, elles consomment plus de mémoire. Le cadre utilise des semi-anneaux basés sur des intervalles, des arbres de recherche équilibrés et un comptage orthogonal 2D pour une recomputation rapide des requêtes, améliorant ainsi le débogage et l'audit. L'étude d'EPL formalise son modèle de données en se concentrant sur des entités hiérarchiques et objet essentielles au traitement des flux. Une analyse systématique identifie des ambiguïtés et

inefficacités, comme le comportement ambigu de l'opérateur NOT et les redondances de EVERY. Une grammaire modifiée améliore l'optimisation des requêtes. La recherche suit le cadre Macro-Meso-Micro, structurant les questions en niveaux global, intermédiaire et spécifique. Un aspect clé est le problème du Continuous Violation Mapping : comment les violations des contraintes en entrée affectent-elles les résultats des requêtes ? Un diagramme de commutation illustre la propagation des incohérences, démontrant les limites des techniques de réparation traditionnelles et plaidant pour des requêtes conscientes de la cohérence. La thèse examine les travaux existants sur l'intégrité des bases de données, la correction des flux et le suivi de provenance. L'évaluation inclut des études qualitatives et de performances, une démonstration d'InkStream et des cas d'usage en analyse financière et détection de fraude aux avis en ligne. Les conclusions suggèrent d'étendre le modèle de provenance à d'autres contraintes et langages, d'explorer des modèles de données alternatifs et d'étudier des techniques de fenêtrage adaptatif. Cette thèse pose les bases du traitement des flux avec intégrité, introduisant des techniques pour suivre la cohérence sans perturber l'analyse en temps réel. Ces méthodes ont un impact sur des secteurs nécessitant des données de haute qualité, notamment la finance, la santé et l'IoT.

intégrité,flux de données,requête continue,cohérence