République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE SUPERIEURE EN INFORMATIQUE 08 MAI 1945 SIDI BEL ABBES

THESIS

To obtain the diploma of **Master**

Field : **Computer Science**

Specialty : **Système d'Information et Web (SIW)**

Theme

Query Optimizer Using Machine Learning

Presented by: :

Farouk BENSOUKEHAL
Ilias Abderahman SOLTANI
Abdelhadi SOUFI MERZOUG
Yasser KADID

Submission Date : **October, 2025**

In front of the jury composed of:

President:	M^{r} [xx]
Examiner:	M^{r} [xx]
Supervisor:	M^{r} **KECHAR Mohamed**
Co-Supervisor:	M^{r} [xx]

Academic Year 2024/2025

Abstract

Relational databases must turn SQL into efficient execution plans quickly and reliably. Classic optimizers do this with statistics and hand-written rules, which work well on simple, stable data but often struggle with correlations, skew, and changing workloads. This thesis studies when and how machine learning (ML) can help without adding risk or slowing planning.

We organize recent work into five stages of the pipeline: estimating result sizes, predicting and ordering plan costs, choosing join orders, selecting among candidate plans, and adapting during execution. For each stage we compare methods by accuracy, planning overhead, worst-case slowdowns, ability to handle drift, and ease of integrating with existing systems, and we summarize the evidence in compact tables.

The main finding is pragmatic: hybrids that *add* learned signals to the existing optimizer are more dependable than attempts to replace it end-to-end. We provide a simple deployment playbookstart with better result-size estimates, add a lightweight re-ranking/selection step, consider guarded search guidance and modest runtime feedback only with fallbacks and limits, and monitor a few clear metrics. Finally, we outline evaluation and reporting practices that focus on overall optimizer impact rather than isolated model accuracy.

Résumé

Les bases de données relationnelles doivent transformer des requêtes SQL en plans dexécution efficaces, rapidement et de façon fiable. Les optimiseurs classiques sappuient sur des statistiques et des règles écrites à la main; cela fonctionne bien lorsque les données sont simples et stables, mais devient fragile face aux corrélations, aux déséquilibres et aux charges qui évoluent. Ce mémoire examine quand et comment lapprentissage automatique peut aider sans rallonger le temps de planification ni augmenter le risque.

Nous structurons les travaux récents en cinq étapes du pipeline: estimer la taille des résultats, comparer et ordonner les plans, choisir lordre des jointures, sélectionner parmi plusieurs plans proposés, et sadapter pendant lexécution. Pour chaque étape, nous discutons la précision, le surcoût, le risque de contre-performance, la capacité à sadapter aux changements et la facilité dintégration, et nous synthétisons ces éléments dans des tableaux compacts.

La conclusion est pragmatique: les approches hybrides qui *ajoutent* des signaux appris à loptimiseur existant sont plus fiables que les tentatives de le remplacer entièrement. Nous proposons un guide simple de déploiement (mieux estimer, reclasser légèrement les plans, nactiver le guidage de recherche et ladaptation à lexécution quavec des garde-fous) et recommandons une évaluation centrée sur limpact de bout en bout plutôt que sur des métriques de modèle isolées.

الملخص

تحتاج قواعد البيانات العلائقيّة إلى تحويل أوامر SQL إلى خطط تنفيذ فعّالة بسرعة وبموثوقيّة. تعتمد المُحسّنات الكلاسيكيّة على إحصاءات وقواعد ثابتة؛ وهي تعمل جيّدًا عندما تكون البيانات بسيطة وثابتة، لكنها تضعف أمام الارتباطات وعدم التوازن والتغيّر في أنماط الاستخدام. يدرس هذا البحث متى وكيف يمكن لتعلّم الآلة أن يُساعد دون زيادة زمن التخطيط أو تعريض الأداء لخطر التدهور.

نقسّم الأعمال الحديثة إلى خمس مراحل في مسار التخطيط: تقدير أحجام النتائج، مقارنة وترتيب الخطط، اختيار ترتيب التوصيلات، الانتقاء بين الخطط المرشّحة، والتكيّف أثناء التنفيذ. وفي كل مرحلة نناقش الدقّة، والكلفة الإضافيّة، وخطر أسوأ الحالات، والقدرة على التكيّف مع تغيّر البيانات، وسهولة الدمج مع الأنظمة القائمة، مع تلخيص النتائج في جداول موجزة.

تتمثّل الخلاصة العمليّة في أن الحلول الهجينة التي تضيف إشارات متعلَّمة إلى المُحسّن التقليدي أكثر اعتمادًا من محاولات استبداله بالكامل. نقدّم خطوات بسيطة للتطبيق (تحسين تقدير الأحجام أولًا، ثم طبقة خفيفة لإعادة ترتيب الخطط، والنظر في توجيه البحث والتكيّف أثناء التنفيذ مع وجود بدائل آمنة)، ونوصي بأن تركز التقييمات على الأثر الشامل للنظام بدلًا من دقّة النموذج وحدها.