كلية الحاسبات والمعلومات

رسالة ماجستير

مترجم فعال من لغة الاستعلامات المتقدمه إلى خرائط الحد وذلك لتحسين تحليل البيانات الضخمة على بيئة الحوسبة السحابية

إعداد

فوزية رمضان سيد حسان

تحت إشراف

أ.د. إبراهيم فرج

أستاذ بقسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة القاهرة

د. محمد خفاجى

دكتور بقسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة الفيوم

رسالة مقدمة إلى كلية الحاسبات المعلومات

جامعة القاهرة، كجزء من متطلبات الحصول

على رسالة الماجستير فى علوم الحاسب

كلية الحاسبات والمعلومات

جامعة القاهرة

جمهورية مصر العربية

مارس ٢٠١٥

ملخص الرسالة

مع الزيادة المستمرة في حجم البيانات في الأنظمة كبيرة ومع الاحتياجات المتزايدة للتحليل البيانات واسترجاع المعلومات، لقد أصبح اطار عمل MapReduce تقنية فعالة لمعالجة هذة البيانات الضخمة والحاجة إلى مترجم فعال ذو كفاءة ومرونة عالية الجودة من لغه الاستعلامات المتقدمة إلى خرائط هو أمر ضرورى جدا ولأن لغة الاستعلامات SQL لديها قدرة محددوة فى معالجة البيانات. يعتبر HIVE لغة استعلاماتSQL مبنية على اطار عمل MapReduce لمعالجة الاستعلامات فى البيانات الكبيرة. HIVE تدعم الاستفسارات مثل HiveQL بالرغم أن هذة اللغه تدعم الكثير من الاستعلامات الإ أنها تعانى من قصور فى بعض الاستعلامات المتقدمة. وهذا مما دفعنا إلى تطوير مترجم خاص لهذة اللغة لحل هذا القصور .

وعلى الناحية الآخرى ظهر اطار عمل آخر لمعالجة البيانات الضخمة في الأنظمة العنقودية يدعى Flink فهو نموذج فعال لتحليل تلك البيانات.الإ أنه لا يوجد لديه مترجم لاستعلامات SQL فى البيانات الضخمة، وهذا مما دفعنا إلى اقتراح مترجم خاص به يدعى SQL To FLink Translator .العمل في هذه الأطروحة هدفه بناء وتحسين لغه الاستعلامات SQL إلى خرائط الحد عن طريق مساهمتين مختلفتين.

المساهمة الأولى وهى تقديم نظام QRMapper فهو مخطط لإعادة كتابة الاستعلام قد تم تقديمه فى هذة الأطروحة لحل مشكلة التعامل مع الاستعلامات المتقدمة من لغة SQL إلى HiveQL دون أى تغير فى اطار وهيكلة HIVE. ويتكون النظام QRMapper من خمس مراحل رئيسية محلل الإستعلام المدخل SQL، استخراج الاستعلامات الفرعية، تحسين الاستعلام الفرعى، تنفيذ الاستعلام الفرعى والتحويل النهائى للاستعلام. لقد تم تنفيذ نظام QRMapper باستخدام تقنية إعادة صياغة الاستعلام الفرعى وذلك عن طريق تطبيق محسن الاستعلام الفرعى وتحويل الاستعلام الفرعى ثم تطبيق التحويل النهائى للاستعلام على الاستعلام المدخل وذلك بعد الحصول على نتيجة الاستعلام الفرعى .

نظام QRMapper يسمح إمكانية تنفيذ SubQuery وتنفيذ ايضا الاستعلامات المتقدمة من لغة الاستعلامات SQL . يمكنة تنفيذ العمليات المنطقية مثل الاتحاد والتقاطع والنقص.كما أن نظام QRMapper لديه قدره عاليه على تشغيل الاستعلامات المتقدمة مقارنة ببقية الأنظمة المقترحة مثل YSmart, S2mart, QMapper . فنظام QRMapper أيضا لديه أفضل أداء، وهذا مما أثبت فى التجارب العملية باستخدام قواعد البيانات المعيارية .TPC-H

المساهمة الثانية وهى عبارة عن مترجم SQL TO Flink الذى قد تم تطويره فى هذة الأطروحة لحل مشكلة التعامل مع لغه SQL الاستعلامات فى اطار Flink حيث أن Flink لا يدعم أى لغه استعلامات دون أي تغيير في إطار Flink وتعطي إمكانية تنفيذ استعلام SQL على Flink عن طريق تحويل الاستعلام إلى صيغة

خوارزمية لل Flink الذى ينفذ ذلك الاستعلام. ويتكون مترجم SQL TO Flink من ثلاث مراحل رئيسية محلل استعلام الSQL ، استخراج عناصر الاستعلام، توليد كود Flink يعادل استعلام SQL المدخل.

فمترجم SQL TO Flink لديه القدرة على تنفيذ الاستعلامات من لغه SQL التى تحتوى على بعض العناصر مثل FILTER, AGGREGATION, GROUP BY, UNION, JOIN, and DISTINCT. ويعتبر ايضا مترجم SQL TO Flink من أفضل الأنظمة التى لديها القدرة على تطبيق و تشيغل الاستعلامات بسرعة عالية جدا فحين أن الأنظمة الآخرى تنفذ تلك الاستعلامات بمستوى منخفض الأداء، كما أن مترجم SQL TO Flink لديه أفضل أداء عند اختباره على معيار TPC-H.

تشتمل هذه الرسالة علي ستة فصول وهي كالآتي:

الفصل الاول : **المقدمة Introduction**

يوضح الدافع وراء هذا العمل و يحدد المشكلة ومحاولة حلها، ثم يعطي الخطوط العريضة للأطروحة وكيفية تنظيمها.

الفصل الثاني: **الخلفية والدراسات السابقة Background and Related Work**

يقدم نبذة عامة عن البيانات الضخمة، الادوات المستخدمة لمعالجتها، وكذلك مسح للحلول والتقنيات المتعلقة بتقديم الحلول المثلى لمعالجتها.

الفصل الثالث : **: أنظمة لغة الاستعلام المتقدمه إلى خرائط (Proposed SQL–to-MapReduce Translators)**

شرح النظام المقترح الأول فى هذه الاطروحة **(QRMapper: Query Rewritable Mapper)** والذى يقوم بإعادة صياغة الاستعلامات المتقدمة الى خرائط الحد عن طريق تحويلها الى لغه HiveQL، وتفاصيله ومراحل تنفيذه.

ومن ثم شرح النظام المقترح الثانيفى هذه الاطروحة (**SQL To Flink Translato**r) والذى يقوم بترجمه لغه الاستعلام إلى خرائط الحد باستخدام ال Flink، وتفاصيله ومراحل تنفيذه.

الفصل الرابع: **النتائج العملية Experimental Results**

يصف بيئة العمل ووحدة القياس المستخدمة فى تقيم النتائج ويعرض تطبيقات الانظمة عمليا ونتائجها حيث توضح النتائج تفوق النظام المقترح عن الأنظمة السابقة.

الفصل السادس : **الخلاصة و الأعمال المستقبلية Conclusion and Future work**

يتضمن هذا الفصل ملخص الرسالة و عرض الأعمال المستقبلية في موضوع البحث.