جـــامعة الفـــيوم
Fayoum University

اسم الطالب: حسن سالم حسين سعد **الدرجة: الماجيستير**

عنوان الرسالة:تطبيقات الحوسبة السحابية لتحليل البيانات الكبيرة الحجم

المشرفون: ١- أ. د./ عمرو محمد رفعت **٢ –أ.م.د/ رانيا احمد عبد العظيم ابو السعود**

قسم: الهندسة الكهربية **تاريخ منح الدرجة من مجلس الكلية: ٣٠ /١٠/ ٢٠١٨**

ملخصالبحث

الثورة الأخيرة في تقنيات التسلسل قد أدى النمو الهائل في البيانات تسلسل الحمض النووي. كنتيجة لذلك، معظم الأدوات المعلوماتية الحيوية القائمة عفا عليها الزمن لأنها لا تفرق مع البيانات. للتعامل مع "التحميل الزائد للبيانات" ، نقدم هنا تحليل ودراسة "PIG" -tool ، الذي سيتم استخدامه في مقارنات متواليات بيولوجية بين تسلسلات "protein & DNA".. الغرض من المقارنة التسلسلية هو اكتشاف التشابهات بين المتواليات البيولوجية المختلفة ثم إنشاء شجرة التجميع. على سبيل المثال ، فإن جينوم الإنسان والفأر متطابقان بنسبة ٨٥٪. "تسلسل الجيل التالي" ينتج مجموعات ضخمة من السلاسل ليتم تحليلها. تتحدى مجموعة البيانات الضخمة هذه أدوات التحليل التقليدية وتتطلب بشكل متزايد حلولاً جديدة تتلاءم مع منصات البيانات الضخمة. يقدم إطار MapReduceللبرمجيات حلاً قابلاً للتطبيق لتحليل التسلسل على نطاق واسع من حيث الكفاءة والقابلية للتطوير. تم تصميم Hadoop كمشروع مفتوح المصدر من إطار MapReduce لتشغيل التطبيقات على مجموعات كبيرة الحجم مبنية على الأجهزة السلعية. نظام الملفات الموزعة (HDFS Hadoop) و Hadoop MapReduce هما مكونان مهمان من هيكل Hadoop. يوفر HDFS مساحة تخزين قابلة للتوسع ، متسامحة مع الأخطاء وموزعة ، بينما MapReduce هو المفهوم الأساسي لإطار Hadoop ويوفر حل معالجة البيانات على نطاق واسع عبر مئات أو آلاف العقد في مجموعة Hadoop. تم بناء مجموعة بيئة الحوسبة السحابية بجامعة الفيوم على مجموعة علمية لينكس لتحليل البيانات الضخمة (SLBD). SLBD تدير برمجيات المصدر المفتوح مع القدرة الحسابية الكبيرة والبنية التحتية العنقودية عالية الأداء. يحتوي SLBD المكون من مجموعة واحدة على أجهزة كمبيوتر متطابقة ذات مستوى سلعي مترابطة عبر شبكة محلية صغيرة.
يستخدم Cloudera Manager لتكوين وإدارة كومة Apache Hadoop. Hadoop هو إطار يسمح تخزين ومعالجة البيانات الكبيرة عبرعنقود من أجهزة الكمبيوتر باستخدام خوارزمية MapReduce. تقوم خوارزمية MapReduce بتقسيم المهمة إلى مهام أصغر يتم تعيينها إلى عقد الشبكة. يتيح نظام التجميع SLBD معالجة سريعة وفعالة لكميات كبيرة من البيانات الناتجة عن تطبيقات مختلفة. كما يوفر SLBDأداء عاليًا وإنتاجًا عاليًا وتوفرًا عاليًا وقابلية للتوسعة وقابلية للتوسعة.
استخدم النظام المقترح "أداة PIG" التي يوفرها نظام SLBD لكتابة نصوص "PIG" لإجراء مقارنة بين تسلسلات "DNA". يتمتع PIG بميزة كبيرة: إن برمجة PIG تقلل إلى حد بعيد من وقت التطوير لتطبيقات المعلوماتية الحيوية المتوازية. تعتبر مقارنة التسلسل مهمة أساسية في البيولوجيا الجزيئية الحسابية التي تهدف إلى اكتشاف علاقات التشابه بين التسلسل الجزيئي. تتم دراسة العديد من طرق المقارنة الحالية المعتمدة على ترددات الكلمة (k-mers). يستخدم النظام المقترح هذه الطرق لبناء أشجار تطورية من جينوم الميتوكوندريا من ١١ من الفقاريات المعروفة بالشجرة الحقيقية. ويبين تحليلنا أن مسافة الارتباط بالنسبة إلى K = 2) dimers) تنتج أفضل الأشجار.

قسمت الرسالة إلى سبعة فصول وفيما يلي موجزاً عن محتويات كل فصل :

الفصل الاول: يتناول المقدمة والمشكلة التى ترتكز حولها هذه الدراسة والهدف منها

الفصل الثانى : يستعرض الأبحاث السابقة في مجال Parallel Computing مثل MPI و GPU وايضا الفروق بينهم وبين ال hadoop وايضا تقدمه بسيطه عن مكونات وعناصر ال hadoop . ويقدم هذا الفصل ايضا مقدمه عن الDNA ومكوناته

.

جـــــامعة الفـــيوم
Fayoum University

اسم الطالب: حسن سالم حسين سعد **الدرجة: الماجيستير**

عنوان الرسالة:تطبيقات الحوسبة السحابية لتحليل البيانات الكبيرة الحجم

المشرفون: ١- أ. د./ عمرو محمد رفعت **٢ –أ.م.د/ رانيا احمد عبد العظيم ابو السعود**

قســم: الهندسة الكهربية **تاريخ منح الدرجة من مجلس الكلية: ٣٠ /١٠/ ٢٠١٨**

الفصل الثالث : يتناول الفصل الثالث شرح تفصيلى عن كيفية بناء SLBD Cluster وايضا شرح تفصيلي لكل مرحلة من مراحل بناء هذا النظام .وختاما يتناول ايضا مرحلة اختبار هذا النظام عن طريقة تشغيل Sample Code بواسطة لغة PIG Latin. . وهذا النظام هو الذي سيستخدم في بناء التطبيق المقترح.

الفصل الرابع : يستعرض الفصل الرابع دراسة وشرحا تفصيليا للPIG وللغة الPIG Latin - وكيفية كتابت الاوامر وطرق تنفيذها وذلك عن طريق امثلة بسيطةSample Code .

الفصل الخامس : يتناول هذا الفصل كيفية بناء التطبيق المقترح وايضا وضعه علي الSLBD . وهذا التطبيق سيقوم بعمل المقارنة بين سلاسل ال DNA . وكذلك شرح الطريقة التي عن طريقاء يتم ايجاد التشابه بين هذة السلاسل مع شرح مبسطلكيفية عمل هذا التطبيق.

الفصل السادس : يتناول الفصل السادس النتائج التي ظهرت من عملة تشغيل هذا التطبيق ومقارنة هذة النتائجبالنموذج المعروف مسبقا. وايضا مقارنةاداء هذا التطبيق مع تطبيق اخر مثل الKAnalyze. وايضا قياس اداء هذا التطبيق عند وضعهعلها جهاز واحد وايضا عند وضعه علي عدة اجهزة.

الفصل السابع : وهو الفصل الأخير الذى يضم ملخص لنتائج البحث والخطط المستقبلية المقترحة بشأنه.