การประเมินประสิทธิภาพของระบบการจัดการฐานข้อมูล XML แบบโอเพนซอร์ส

Performance Evaluation on Open Source Native XML DBMSs

ณัฐกานต์ อัมรินทรารักษ์[1*] กานดา สายแก้ว[1**] ศิษเฎศ ทองสิมา[2***] นุวีย์ วิวัฒนวัฒนา[3****]

[1] ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น, 40002, ประเทศไทย

[2] ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ปทุมธานี, 12120, ประเทศไทย

[3] กรมควบคุมมลพิษ 92 ถนนพหลโยธิน เขตพญาไท กรุงเทพฯ 10400, ประเทศไทย

Email: a.nuttakan@gmail.com[*], krunapon@kku.ac.th[**], sissades@biotec.or.th[***], nuwee.w@pcd.go.th[****]

Abstract

Native XML Database Management System (Native XML DBMS) is designed especially to store and manage tree-structured XML documents. This system differs from Relational DBMSs which store data in flat tables. However, different native XML DBMSs have different approaches in implementing storage manager and query engines. Choosing an appropriate Native XML DBMS to manage a large and complex XML data can dramatically reduce processing time.

This paper presents a benchmark result of testing open source native XML DBMSs. The selected open-source native XML databases for evaluation are eXist, Berkeley XML DB and Sedna. The testing queries are taken from XMark which is widely used to test the performance of XML databases. In addition, common queries were executed on real-world genomic data. The analyzed experimental result indicates pros and cons of different Native XML DBMSs.

Key Words: Native XML DBMS, Performance Evaluation, Benchmark

บทคัดย่อ

Native XML Database Management System (Native XML DBMS) คือ ระบบการจัดเก็บ และจัดการข้อมูลลักษณะ XML โดยเฉพาะ มีลักษณะการจัดเก็บข้อมูลในรูปแบบของทรี ซึ่งต่างจาก Relational Database Management System (RDBMS) ที่มีการจัดเก็บข้อมูลในลักษณะตาราง การเลือกใช้งาน Native XML DBMS ที่เหมาะสมกับลักษณะของข้อมูล XML เป็นสิ่งสำคัญอย่างยิ่งโดยเฉพาะกับข้อมูลที่มีขนาดใหญ่และซับซ้อน เพื่อลดเวลาในการประมวลผล

บทความนี้นำเสนอการทดสอบประสิทธิภาพของระบบการจัดการฐานข้อมูลโอเพนซอร์สสำหรับจัดเก็บข้อมูล XML (Open Source Native XML DBMS) กับการประมวลผลข้อมูล XML ที่มีความซับซ้อนดังเช่นฐานข้อมูลเชิงชีววิทยา เพื่อเปรียบเทียบประสิทธิภาพในการประมวลผล ซึ่งประกอบไปด้วย Berkeley DB XML, eXist และ Sedna คำค้นที่ใช้ในการทดสอบคือ คำค้นที่ระบุโดย XMark ซึ่งมักจะถูกใช้เพื่อประกอบการพิจารณาในการเลือกฐานข้อมูลสำหรับจัดเก็บและค้นหาข้อมูล XML นอกจากนี้ยังมีการทดสอบคำค้นที่ใช้งานกับฐานข้อมูลจีโนมจริง ผลการสอบและการวิเคราะห์ทำให้ทราบว่า แต่ละฐานข้อมูลนั้นมีข้อดีข้อด้อยแตกต่างกัน

คำสำคัญ: ระบบการจัดการฐานข้อมูล XML, การประเมินประสิทธิภาพ, การทดสอบ

1. บทนำ

จากคุณสมบัติของภาษา XML [5] ที่มีความยืดหยุ่น สามารถออกแบบรูปแบบข้อมูล ชื่ออิลิ-เมนต์และแอตทริบิวต์ตลอดจนโครงสร้างของข้อมูลได้ และการที่ไม่ขึ้นอยู่กับแพลตฟอร์มใดแพลตฟอร์มหนึ่งทำให้ XML ถูกนำมาใช้งานอย่าง

แพร่หลาย แต่ว่าการจัดการกับข้อมูลที่อยู่ในรูปแบบ XML นั้นยังมีความยุ่งยาก จึงมีผู้พัฒนา Native XML DBMS เพื่อช่วยทำหน้าที่ในการจัดเก็บและจัดการข้อมูลประเภท XML เป็นการเพิ่มประสิทธิภาพในการจัดการข้อมูล XML ซึ่งมีโครงสร้างข้อมูลในรูปแบบต้นไม้ (Tree) เพราะหากใช้ Relational DBMS จะต้องเสียเวลาการแปลงรูปแบบข้อมูลให้เป็นตารางอีกทั้งใช้เวลามากในการหาความสัมพันธ์ระหว่างโหนดต่าง ๆ ดังนั้นการเลือกใช้ฐานข้อมูลที่เหมาะสมกับรูปแบบข้อมูลและการใช้งานจึงมีความสำคัญอย่างยิ่งต่อประสิทธิภาพของการทำงาน

ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) [9] ร่วมกับภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น ได้ทำการพัฒนาเว็บเซอร์วิส [10] ที่ให้บริการข้อมูลจีโนมของสิ่งมีชีวิตต่างๆ อาทิ กุ้งกุลาดำ, ข้าว, สาหร่ายเกลียวทอง และ ความหลากหลายทางพันธุกรรมของมนุษย์ในรูปแบบของ Single Nucleotide Polymorphisms หรือเรียกสั้นๆว่า สนิป แก่นักวิจัยและผู้สนใจเพื่อได้ข้อมูลที่เป็นประโยชน์นั้นก็ได้มีการนำเอา XML มาเป็นมาตรฐานในการแลกเปลี่ยนข้อมูล ซึ่งลักษณะของข้อมูลของฐานข้อมูลชีววิทยาเชิงจีโนมนั้น มีลักษณะข้อมูลที่มีขนาดใหญ่เนื่องจากมีการจัดเก็บรายละเอียดของลำดับเบสดีเอ็นเอและข้อมูลสนับสนุนอื่นๆ เอาไว้ด้วยกัน (ปัจจุบันข้อมูลขนาดประมาณ 20 GB และมีแนวโน้มที่จะเพิ่มขึ้นเรื่อยๆ) การพิจารณาเลือก Native XML DBMS เพื่อมาจัดการกับงานที่มีข้อมูลปริมาณมากเช่นนี้จึงมีความสำคัญอย่างยิ่ง

งานทดสอบประสิทธิภาพในการประมวลผลข้อมูลของระบบจัดการข้อมูลที่ผ่านมามีความแตกต่างกันไป เช่นการทดสอบประสิทธิภาพของฐานข้อมูลประเภท Relational DBMS ซึ่งอาจไม่เหมาะสมกับการจัดเก็บข้อมูล XML และการค้นหาความสัมพันธ์ระหว่างข้อมูลย่อยในข้อมูล XML ดังนั้น งานนี้แตกต่างจากงานอื่นตรงที่เน้นทดสอบเฉพาะ Native XML DBMS ที่เป็นโอเพนซอร์สและสามารถรองรับการค้นหาข้อมูลด้วยภาษา XQuery เหตุผลที่เลือกใช้ฐานข้อมูลโอเพนซอร์สเพราะองค์กรจะได้ประหยัดค่าใช้จ่ายอีกทั้งสามารถดูหรือแก้ไขซอร์สโค้ดได้ด้วย ซึ่งในหัวข้อที่ 2 จะแนะนำให้รู้จักกับ Native XML DBMS ที่เลือกมาทดสอบ ในหัวข้อที่ 3 นำเสนอเครื่องมือในการทดสอบ (XMark) และวิธีการทดสอบได้อธิบายในหัวข้อที่ 4 ผลการทดสอบถูกนำเสนอในหัวข้อที่ 5 และสรุปผลการทดสอบในหัวข้อที่ 6 หัวข้อสุดท้ายจะนำเสนอการ วิเคราะห์ผลการทดสอบและแนวทางการนำไปใช้งาน

2. Native XML DBMSs ที่ใช้ในการทดสอบ

Native XML DBMSs ที่ใช้ในการทดสอบได้แก่ Berkeley DB XML, eXist, และ Sedna XML DBMS เหตุผลที่เลือก Native XML DBMSs 3 ตัวนี้ก็เพราะมีการพัฒนาอย่างต่อเนื่อง มีรูปแบบการใช้งานที่ไม่ขึ้นอยู่กับแพลตฟอร์มใดแพลตฟอร์มหนึ่ง

2.1. Berkeley DB XML [1]

Berkeley DB XML เป็น Native XML DBMS พัฒนาบนพื้นฐานของ Berkeley DB โดยเพิ่ม XML Parser, XML indexes นอกจากนี้มีการจัดเก็บเอกสาร XML เป็นกลุ่มลอจิกเรียกว่า Container สามารถจัดเก็บเอกสารทั้งที่อยู่ในรูปแบบ

Document หรือ Meta data รองรับ DTD และ XML Schema รวมถึงสามารถทำการตรวจสอบความถูกต้องของข้อมูลได้ นอกจากนี้ Berkeley DB XML ยังสนับสนุนการค้นหาข้อมูลโดยใช้ภาษา XQuery และมีไลบรารีซึ่งสามารถเรียกใช้งานได้ในหลายภาษา เช่น Java, Perl, Python และ PHP

2.2. eXist [2]

eXist พัฒนาโดยใช้ภาษา Java จัดเก็บข้อมูลอยู่ในลักษณะไบนารีทรี และไฟล์เอกสารจัดเก็บอยู่ใน Collection รองรับคำสั่ง XQuery/XPath2.0 สามารถใช้งานในลักษณะ Stand-alone, Embedded Database หรือ Servlet Engine สามารถค้นหาพร้อมกันหลาย Collection หรือ พร้อมกันหลายเอกสาร สามารถแก้ไขข้อมูลโดย XUpdate ได้ แต่ต้องอยู่ในรูปแบบการใช้งานแบบ Embedded Database Server เท่านั้น สามารถถูกเรียกใช้งานได้ผ่าน XML-RPC, REST Web Services API, SOAP และ WebDAV มีการสร้าง Index ให้กับอิลิเมนต์และแอตทริบิวต์แบบอัตโนมัติ

2.3. Sedna XML DBMS [3]

Sedna XML DBMS พัฒนาโดยใช้ภาษา Java สามารถรองรับข้อมูลทั้งแบบ Data–Centric และ Document -Centric โดยไม่จำกัดขนาด มีการค้นหาและประมวลผลข้อมูลโดย XQuery จากมาตรฐาน W3C[6] มี Java API ที่สามารถเรียกใช้งานได้ ในแง่การใช้งาน Sedna จะมีความยืดหยุ่นสำหรับนักพัฒนาโปรแกรม โดย Sedna มี API ในหลายภาษา เช่น Java, C, Scheme, PHP, Python, .NET และ OmniMark นอกจากนี้ยังมีโปรโตคอลที่อนุญาตให้มีการพัฒนา API ในภาษาโปรแกรมอื่น

ตารางที่ 1 คุณลักษณะต่างๆของฐานข้อมูลที่เลือกใช้ในการทดสอบ

ฐานข้อมูล	รูปแบบคำสั่ง	รูปแบบผลลัพธ์	รูปแบบการจัดเก็บข้อมูล
eXist	XQuery, XPath	XML	Document-centric
Berkeley DB XML	XQuery, XPath	XML	Data-centric
Sedna	XQuery	XML	Document-centric, Data-centric

จากตารางที่ 1 เห็นได้ว่าการทดสอบนี้จะเลือกเฉพาะฐานข้อมูลชนิดโอเพนซอร์สที่รับคำสั่งในการค้นหาในลักษณะ XQuery และให้ผลลัพธ์เป็น XML

3. เครื่องมือทดสอบประสิทธิภาพการจัดเก็บและค้นหาข้อมูล XML XMark

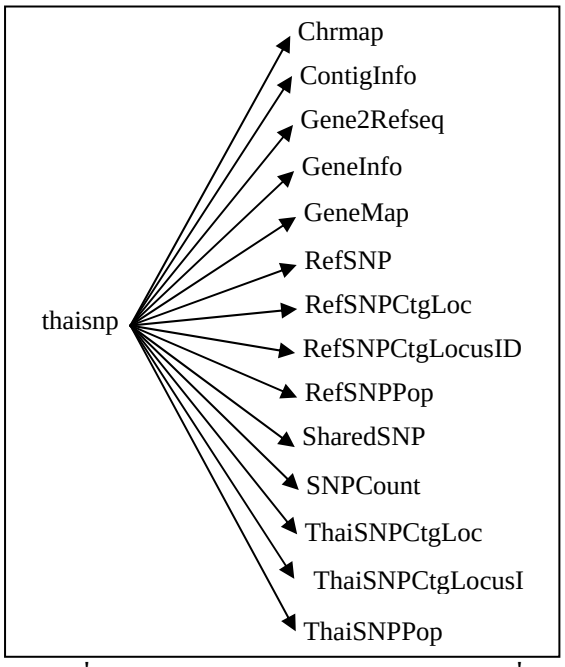
วัตถุประสงค์หลักของ XMark [4] คือ ต้องการสร้างเครื่องมือในการทดสอบประสิทธิภาพสำหรับซอฟต์แวร์ที่ใช้ในการจัดเก็บและประมวลผลข้อมูลในลักษณะ XML เพื่อทดสอบความสามารถในการรับมือกับคำสั่งค้นหาที่หลากหลายแตกต่างกันไป ประกอบด้วยชุดคำสั่งภาษา XQuery 14 กลุ่ม จำนวน 20 คำสั่ง XMark สามารถสร้างข้อมูลที่ใช้ในการทดสอบชุดคำสั่งที่มีโดยสามารถระบุขนาดข้อมูลที่ต้องการได้

4. ข้อมูล ชุดคำสั่ง และเครื่องคอมพิวเตอร์ในการทดสอบ

4.1. ข้อมูลที่ใช้ในการทดสอบ

ข้อมูลที่ใช้ทดสอบ คือข้อมูลสนิปของคนไทย ซึ่งอยู่ภายใต้การพัฒนาของ BIOTEC เก็บข้อมูลด้านชีววิทยาในระดับองค์ประกอบของยีนของคนไทย [7] เหตุผลที่เลือกใช้ข้อมูลไทยสนิป เนื่องจากได้ถูกพัฒนามานาน มีความสมบูรณ์ เป็นข้อมูล

ขนาดใหญ่มีตาราง 62 ตาราง ประกอบไปด้วยข้อมูลจำนวนหลายล้านเรคคอร์ด ซึ่งปัจจุบันใช้งานอยู่ในรูปแบบ Relational Database (RDB) จึงต้องทำการแปลงให้อยู่ในรูปแบบ XML โดยใช้อัลกอริทึมในการแปลงจาก RDB Schema ให้เป็น XML Schema [8] ข้อมูลที่ใช้เลือกมา 14 ตาราง หลัก จากทั้งหมด 62 ตาราง ซึ่งมีโครงสร้างแสดงดังรูปที่ 1 และรูปที่ 2 โดยรูปที่ 1 แสดงอิลิเมนต์ลำดับที่ 1 มีรูทอิลิเมนต์ ชื่อว่า thaisnp ประกอบไปด้วยอิลิเมนต์ลูกจำนวน 14 อิลิเมนต์, รูปที่ 2 แสดงอิลิเมนต์ลำดับที่ 2 แต่ละโหนดแทนด้วยตาราง ในฐานข้อมูลจาก RDB Schema



รูปที่ 1 สกีมาของข้อมูล Thaisnp ในระดับที่ 1

แต่ละอิลิเมนต์ใน XML Schema ประกอบไปด้วยอิลิเมนต์ลูก 2 กลุ่ม กลุ่มแรกแปลงจากคอลัมน์ในตารางนั้นๆ ให้เป็นอิลิเมนต์ (ไม่แสดงในภาพเนื่องจากมีจำนวนมาก) และกลุ่มที่ 2 เป็น complex อิลิเมนต์ที่แปลงจากตารางที่มีความสัมพันธ์(ตามความสัมพันธ์ใน RDB) กับอิลิเมนต์ดังกล่าว

หลังจากนั้นจึงทำการสร้างเอกสาร XML จากสกีมาดังกล่าว เพื่อใช้ในการทดสอบจำนวน 3 ขนาดคือ 10MB 100MB และ 1GB

4.2. ชุดคำสั่งที่ใช้ในการทดสอบ

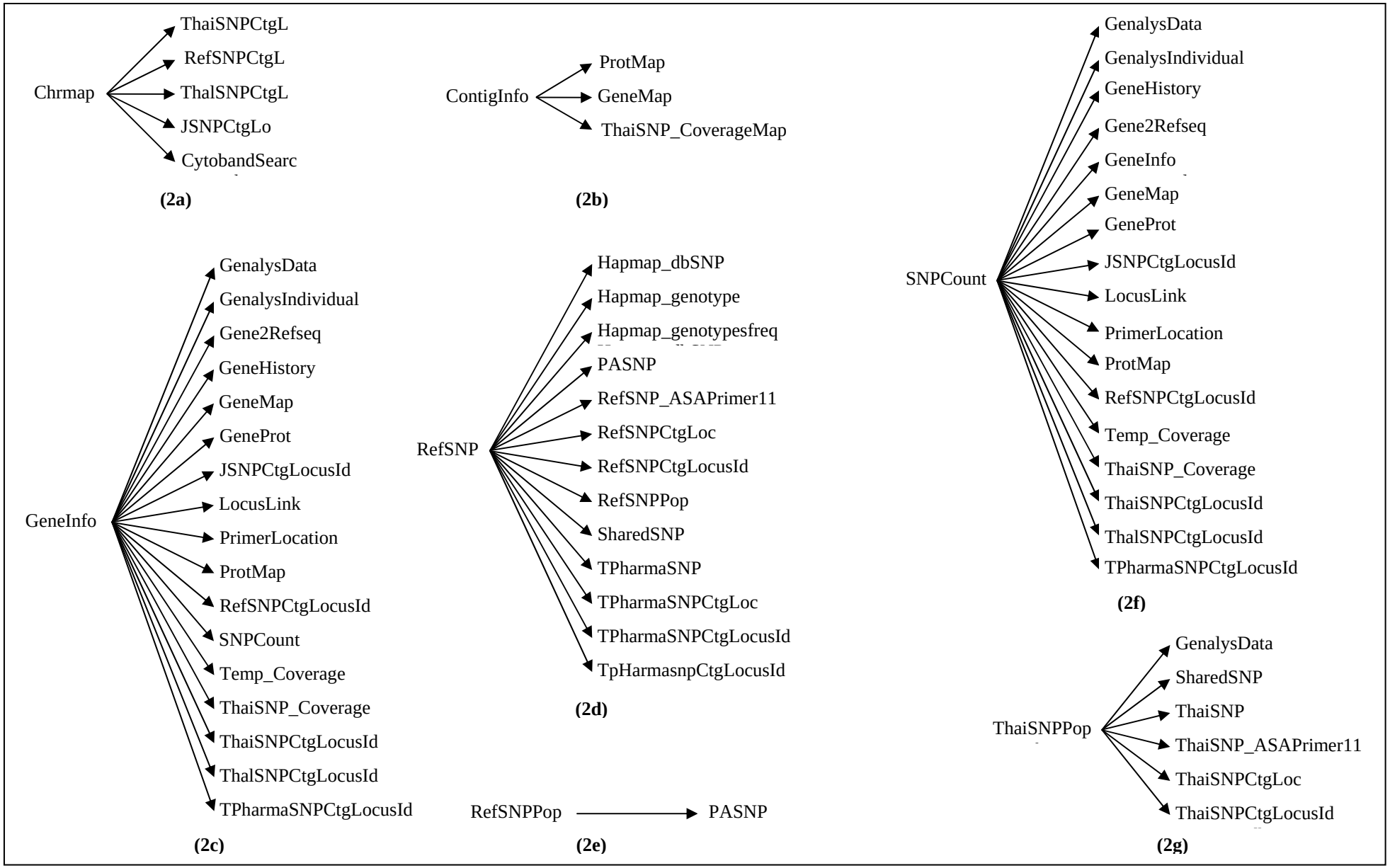
ชุดคำสั่งที่ใช้ในการทดสอบแบ่งเป็นสองกลุ่มใหญ่ๆ คือ 1) กลุ่มคำสั่งมาตรฐานของ XMark เพื่อทดสอบความสามารถต่างๆ ในการประมวลผล 2) กลุ่มคำสั่งที่ใช้งานจริงโดยเลือกรูปแบบคำสั่งที่แตกต่างกัน และทำการแปลงจาก SQL command ให้อยู่ในรูปแบบ XQuery ในการทดสอบนั้น กลุ่มคำสั่งที่ 1 จะทำการทดสอบกับข้อมูลทั้ง 3 ชุด (10MB,100MB,1GB) แต่กลุ่มคำสั่งที่ 2 จะทำการทดสอบกับข้อมูลชุดสุดท้ายเพียงชุดเดียว(1GB) สามารถดูรายละเอียดของชุดคำสั่งในการทดสอบได้จาก [11]

4.3. เครื่องคอมพิวเตอร์ที่ใช้ในการทดสอบ

เครื่องคอมพิวเตอร์ที่ใช้ในการทดสอบคือเครื่องที่มี CPU เป็น Intel Pentium M 1.73 GHz , Memory 1GB, และ HD 80 GB ระบบปฏิบัติการ Windows XP SP2 และทำการทดสอบแบบ Stand-alone ใน Command line โหมด

5. ผลการทดสอบ

ก่อนทำการทดสอบการค้นหาข้อมูลต้องทำการโหลดข้อมูลไปยังฐานข้อมูลก่อนเป็นอันดับแรก ตารางที่ 2 แสดงเวลาในการโหลดข้อมูลขนาดต่างๆ เข้าไปยังฐานข้อมูล ซึ่งผลที่ได้ คือ eXist ใช้เวลาในการโหลดข้อมูลนานที่สุด และเกิดข้อผิดพลาดเนื่องจากหน่วยความจำไม่เพียงพอ (OOM = Out Of Memory) เมื่อข้อมูลมีขนาด 1GB Berkeley DB XML ใช้เวลาน้อยลงมา และ Sedna ใช้เวลาน้อยที่สุด



รูปที่ 2 สกีมาของข้อมูล Thaisnp ในระดับที่ 2

ตารางที่ 2 เวลาในการโหลดข้อมูลเข้าไปยังฐานข้อมูล

ฐานข้อมูล	เวลาในการโหลดข้อมูล (วินาที)		
	10MB	100MB	1GB
eXist	172.9	421.5	OOM
Berkeley DB XML	4.6	44.3	628.3
Sedna	2	13.5	198.9

หลังจากนั้นจึงทำการทดสอบกับชุดคำสั่งต่าง ๆ โดยทำการประมวลผล 5 ครั้งต่อ 1 คำสั่งพร้อมจับเวลา ในผลการทดสอบทั้ง 5 ครั้ง ตัดค่าเวลาน้อยที่สุดและมากที่สุด แล้วหาค่าเฉลี่ยของ 3 ครั้งที่เหลือ

5.1. ผลการทดสอบช่วงที่ 1 การทดสอบกับชุดคำสั่งตามมาตรฐาน XMark

จากตารางที่ 3 OOM หมายถึง ความผิดพลาดเนื่องจากหน่วยความจำไม่เพียงพอ NR = No Result หมายถึง ไม่มีผลทดสอบเนื่องจากไม่สามารถสร้างฐานข้อมูลได้NS = Not Supported หมายถึง ฐานข้อมูลไม่รองรับคำสั่งนั้น

เมื่อทดสอบกับข้อมูลขนาด 10MB eXist ใช้เวลาโดยเฉลี่ยประมาณ 0.37 วินาที แต่สำหรับชุดคำสั่งที่ 4 ที่ทดสอบความสามารถในการประมวลผลโดย Path expression ใช้เวลามากกว่า 15 นาที ส่วน Berkeley DB XML ใช้เวลาเฉลี่ยตั้งแต่ 1.37 วินาที ในขณะที่ Sedna เป็นฐานข้อมูลโอเพนซอร์สที่ใช้เวลาน้อยที่สุดโดยใช้เวลาเฉลี่ยประมาณ 0.26 วินาที แต่ Sedna ไม่สามารถใช้ตอบคำถามคำสั่งที่ 14 ซึ่งเป็นคำสั่งที่ให้หาอิลิเมนต์ที่เกี่ยวข้องและมีคำสำคัญตามที่กำหนดไว้

เมื่อทดสอบกับข้อมูลขนาด 100MB eXist ใช้เวลาโดยเฉลี่ยประมาณตั้งแต่ 2.14 วินาที ส่วนชุดคำสั่งที่ 4 นั้นยังคงใช้เวลามากกว่า 15 นาที เหมือนเมื่อทดสอบกับข้อมูลขนาด 10 MB นอกจากนี้เมื่อทดสอบโดยใช้คำสั่งที่ 20 เกิดความผิดพลาดเนื่องจากหน่วยความจำไม่เพียงพอ ซึ่งคำสั่งที่ 20 เป็นการทดสอบการรวบรวมข้อมูลจำนวนมาก ส่วนผลการทดลองกับ Berkeley DB XML นั้นพบว่า ใช้เวลาเฉลี่ย 31.21 วินาที ในชุดคำสั่งที่ 7, 14 และ 20 นั้นใช้เวลาประมาณ 150-200 วินาที และผลการทดสอบฐานข้อมูล Sedna ใช้เวลาโดยเฉลี่ยตั้งแต่ 1.86 วินาทีซึ่งเป็นเวลาเฉลี่ยที่น้อยที่สุด

สำหรับผลการทดสอบกับข้อมูลขนาด 1 GB นั้นไม่สามารถทดสอบกับฐานข้อมูล eXist ได้เนื่องจากไม่สามารถโหลดข้อมูลไปยังฐานข้อมูล และเมื่อทดสอบกับฐานข้อมูล Berkeley พบปัญหาหน่วยความจำไม่พอในชุดคำสั่งประมาณ 9 ชุด ส่วนชุดคำสั่งที่สามารถรันได้ตามปกตินั้นใช้เวลาเฉลี่ยประมาณ 27.12 วินาที เมื่อทดสอบโดยใช้ Sedna นั้นพบว่าใช้เวลาโดยเฉลี่ย 7.6 วินาที โดยที่ชุดคำสั่งที่ 6 – 9 ใช้เวลาค่อนข้างนาน และ ชุดคำสั่งที่ 20 ใช้เวลามากกว่า 15 นาที

ตารางที่ 3 ผลทดสอบกับชุดคำสั่งมาตรฐาน XMark

คำค้น	เวลาที่ใช้ในการประมวลผล(วินาที)								
	10MB			100MB			1GB		
	eXist	Berkeley	Sedna	eXist	Berkeley	Sedna	eXist	Berkeley	Sedna
Q1	0.057	0.016	3.717	0.151	0.521	3.703	NR	65.203	3.658
Q2	0.413	2.005	0.04	0.214	0.1356	0.047	NR	3.891	0.097
Q3	0.36	0.047	0.029	0.713	0.729	0.062	NR	45.062	0.367
Q4	> 15 min	2.015	0.027	> 15 min	0.0523	0.032	NR	16.922	1.435
Q5	0.133	0.047	0.027	0.207	0.989	0.06	NR	37.094	0.367
Q6	1.322	2.546	0.243	5.114	5.349	0.946	NR	OOM	13.327
Q7	0.094	2.37	0.331	1.099	178.791	2.834	NR	OOM	40.089
Q8	0.275	2.547	0.05	6.281	10.031	0.407	NR	OOM	20.053

ตารางที่ 3 ผลทดสอบกับชุดคำสั่งมาตรฐาน XMark (ต่อ)

คำค้น	เวลาที่ใช้ในการประมวลผล(วินาที)								
	10MB			100MB			1GB		
	eXist	Berkeley	Sedna	eXist	Berkeley	Sedna	eXist	Berkeley	Sedna
Q9	0.198	2.37	0.025	6.14	86.739	0.152	NR	OOM	17.839
Q10	0.063	0.265	0.04	0.417	1.38	0.081	NR	OOM	0.686
Q11	0.135	0.729	0.03	0.937	2.396	0.036	NR	36.453	0.747
Q12	0.021	0.083	0.032	0.146	1.593	0.194	NR	73.078	0.067
Q13	0.031	2.021	0.026	0.125	0.083	0.031	NR	0.891	0.072
Q14	2.677	2.63	NS	12.301	185.187	NS	NR	OOM	NS
Q15	0.552	2.093	0.072	2.339	3.51	0.499	NR	OOM	4.566
Q16	0.031	2.031	0.024	0.12	0.078	0.0287	NR	18.921	0.053
Q17	0.02	2.073	0.026	0.073	0.078	0.027	NR	0.282	0.026
Q18	0.031	0.047	0.029	0.036	0.078	0.026	NR	0.516	0.069
Q19	0.296	0.598	0.117	2.062	9.833	0.96	NR	OOM	9.789
Q20	0.397	0.932	0.033	OOM	136.641	25.27	NR	OOM	>15 min

5.2. การทดสอบช่วงที่ 2 การทดสอบกับชุดคำสั่งที่ใช้งานจริง กับข้อมูลขนาด 1GB

ในการทดสอบกับชุดคำสั่งที่ใช้งานกับข้อมูลจีโนมจริงนั้น จะทำการทดสอบกับข้อมูลที่มีขนาด 1 GB เท่านั้น เพราะในการใช้งานจริงฐานข้อมูลที่เก็บข้อมูลจีโนม จะมีขนาดใหญ่ คือมากกว่า 500 MB ขึ้นไปซึ่งได้ผลการทดสอบดังตารางที่ 4

จากการทดลองพบว่า ไม่สามารถทำการทดสอบฐานข้อมูล eXist ได้เนื่องจากไม่สามารถโหลดข้อมูลไปเก็บไว้ได้ ส่วนการทดลองกับฐานข้อมูล Berkeley DB XML นั้นใช้เวลาในการประมวลผลเฉลี่ย 50.82 วินาที และการทดลองกับฐานข้อมูล Sedna นั้นพบว่าใช้เวลาเฉลี่ยประมาณ 0.05 วินาทีในการประมวลผลและให้คำตอบกลับมา

ตารางที่ 4 ผลทดสอบกับชุดคำสั่งที่ใช้งานจริง

ชุดคำสั่ง	เวลาที่ใช้ในการประมวลผล (วินาที)		
	eXist	Berkeley DB XML	Sedna
Q1	NR	128.276	0.056
Q2	NR	70.641	0.025
Q3	NR	77.646	0.086
Q4	NR	13.771	0.031
Q5	NR	1.594	0.101
Q6	NR	63.214	0.039
Q7	NR	0.599	0.026

6. สรุปผลการทดสอบ

จากการทดสอบสามารถสรุปผลการทดสอบได้ดังนี้

ฐานข้อมูล eXist สามารถประมวลผลได้ดีกับฐานข้อมูลที่มีขนาดน้อยกว่า 1 GB และสามารถประมวลผลข้อมูลมากกว่า 1 เอกสารได้ แต่ไม่สามารถนำไปใช้กับข้อมูลที่มากกว่า 1 GB ได้ดี

ส่วนฐานข้อมูล Berkeley DB XML ประมวลผลกับข้อมูลขนาด น้อยกว่า 100 MB ได้ดี สามารถประมวลผลข้อมูลมากกว่า 1 เอกสารได้ แต่สำหรับฐานข้อมูลขนาด 1GB ไม่สามารถประมวลผลกับชุดคำสั่งที่ใช้ทดสอบความสามารถในการประมวลผลโดย Path expression ชุดคำสั่งที่ใช้ทดสอบความสามารถในการใช้งาน reference ความสามารถในการค้นหาจากทั้งฐานข้อมูล และ การสร้างผลลัพธ์ใหม่จากข้อมูลที่ค้นหามาได้

ฐานข้อมูล Sedna สามารถประมวลผลได้ดีที่สุดในบรรดาฐานข้อมูลที่นำมาได้ในการทดสอบ และสามารถประมวลได้ดีกับข้อมูลหลายขนาด แม้ในบางคำสั่งจะใช้เวลาในการประมวลผลนานพอสมควร แต่อย่างไรก็ดี Sedna ไม่สามารถประมวลผลพร้อมกันหลายเอกสารได้

7. วิเคราะห์ผล และแนวทางการนำไปใช้งาน

จากผลการทดสอบและการใช้งานพบว่าแต่ละฐานข้อมูลมีข้อดีข้อด้อยแตกต่างกันไป eXist เหมาะกับการใช้งานที่ขนาดข้อมูลไม่ใหญ่นักคือไม่เกิน 100MB (ทั้งนี้ขึ้นอยู่กับสมรรถนะของเครื่องที่ให้บริการ) ด้วยเทคโนโลยีของ eXist ที่ใช้ DOM ในการจัดการข้อมูลทำให้ฐานข้อมูลต้องการหน่วยความจำเป็นจำนวนมากในการประมวลผล ข้อดีของ eXist คือสามารถทำการค้นหาแบบหลายเอกสารหรือหลายฐานข้อมูลพร้อมกันได้ ในแง่ของผู้ดูแลระบบสามารถจัดการได้ง่ายโดย Graphic User Interface

Berkeley DB XML มีลักษณะการเก็บข้อมูลโดย ทำการแบ่งเอกสารเอกซ์เอ็มแอลให้เป็นโหนดย่อย และทำการเก็บในลักษณะลอจิกทรี เมื่อเอกสารใหญ่ถูกแบ่งเป็นส่วนประกอบย่อยๆ ทำให้ต้องการหน่วยความจำไม่มากนักในการประมวลผล สามารถรองรับข้อมูลที่มีขนาดใหญ่ได้ในระดับหนึ่ง หากมีการจัดการข้อมูลที่ดีเพราะสามารถทำการประมวลผลได้หลายเอกสารพร้อมกัน

Sedna มีความสามารถในการจัดการข้อมูลได้ดีที่สุดจากการทดสอบนี้ทั้งในขนาดข้อมูลเล็กกลาง และใหญ่ โดยที่ใช้เวลาในการประมวลผลค่อนข้างน้อย เหมาะสำหรับข้อมูลที่เก็บข้อมูลทั้งหมดไว้ในเอกสารเดียว เพราะไม่สามารถประมวลผลพร้อมกันหลายเอกสารได้ อย่างไรก็ดีในการจัดการข้อมูลที่มีขนาดใหญ่จะสร้างความลำบากในการจัดการมากกว่าเนื่องจากในการเก็บข้อมูลขนาดใหญ่นั้น มักจะมีการแบ่งข้อมูลไว้ในหลายเอกสาร

8. กิตติกรรมประกาศ

ขอขอบคุณศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติที่ให้การสนับสนุนการศึกษาและการทำวิจัยนี้

9. เอกสารอ้างอิง

[1] Oracle Berkeley DB XML. Retrieved January 30,2008 from http://www.oracle.com/database/berkeley-db/xml/index.html
[2] Exist – Open source native XML database. Retrieved January 30,2008 from http://exist.sourceforge.net
[3] Sedna XML database Retrieved February 2,2008 from http://modis.ispras.ru/sedna/index.htm
[4] XMark – An XML Benchmark Oroject. Retrieved January 5,2008 from http://www.xml-benchmark.org/
[5] XML Retrieved December 22,2007 from http://www.w3.org/XML/
[6] W3C - World Wide Web Consortium .Retrieved January 10,2008 from http://www.w3.org/
[7] ThaiSNP- Thailand SNP Discovery Project. Retrieved January 22,2008 http://thaisnp.biotec.or.th/
[8] Paolo Pigozzo and Elisa Quintarelli “An algorithm for generating XML Schemas from ER Schemas”, SEBD 2005 p 192 -199
[9] ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) Retrieved January 10,2006 http://www.biotec.or.th/biotechnology-th/
[10] Web services @ W3C Retrieved January 23,2008 from http://www.w3.org/2002/ws/
[11] ชุดคำสั่งในการทดสอบประสิทธิภาพของระบบการจัดการฐานข้อมูลโอเพนซอร์สสำหรับจัดเก็บข้อมูล XML Retrieved February 12, 2008 from http://std.kku.ac.th/4950400827