ТЕНДЕНЦІЇ РОЗВИТКУ IT ТЕХНОЛОГІЙ

Згуровський М.З., А.І. Петренко, НТУУ «КПІ»

Сучасні підходи до обробки наукових даних (е-наука) базуються на обробці постійно зростаючих обсягів інформації. Розглянуті тенденції і перспективи розвитку е-науки за умов інформаційного буму, які охоплюють питання забезпечення якості і сумісності даних, використання метаданих і семантики даних, довгострокового збереження даних, інтелектуального оброблення даних, пошуку даних в існуючих джерелах, впливу даних на вибір платформи і її сервісно-орієнтованої архітектури.

Вступ

Прогрес в інформаційних технологіях триває, збільшуючи темп в порівняльні з минулим. Вже сьогодні в світі функціонують декілька суперкомп'ютерів з обчислювальною спроможністю, яка перевершує кілька петафлопс (петафлопс=10^{15} операцій в секунду). Наступне покоління інфраструктури ***е-науки*** буде мати справу з прикладними застосуваннями, для функціонування яких будуть одночасно потрібні мільйони процесорів. До 2020 року очікуються наукові дані обсягом в сотні екзабайт (екзабайт = 10^{18} байт), розподілені в кількох центрах, доступні через різнорідні сховища для аналізу, оброблення і іншої наукової діяльності [1-3].

Тенденції розвитку IT

Можна виділити декілька тенденцій, що характеризують прогрес у розвитку засобів IT:
1.П***остійно наростаючий "інформаційний бум", завдяки чому щороку об'єми наукових даних майже подвоюються.***

Зараз йдеться про одночасну обробку петабайтних наборах даних (петабайт=10^{15} байт), отриманих завдяки підвищеної точності вимірів при спостереженнях та експериментах і досконалості процедур моделювання, що потребує розроблення інтелектуальних методів організації даних для скорочення об'єму пошуку, паралельної обробки і доступу до даних при виконанні пошуку у величезних наборах.

Переконливими прикладами джерела петабайтних наборів даних є Великий Андронний Коллайдер (LHC) , який працює в ЦЕРН і який вироблятиме близько 10 Петабайт необроблених даних в галузі фізики високих енергій у рік; новітній телескопу (Large Synoptic Survey Telescope) з міжнародного проекту Sloan Digital Sky Survey, з допомогою якого досліджується спектр зірок на хвилях за межами видимого спектру; проект GEOSS (Global Earth Observation System of Systems), який базується на використанні супутникових даних для багатьох галузей господарства: прогнозування погоди і можливих врожаїв, спостереження змін клімату і екологічного стану, розповсюдження стихійних лихв (поводів, пожарів, засухи) та багато іншого. До найбільш вагомих з можливих рішень побудови інфраструктури для оброблення даних екзабайтних обсягів можна віднести:

Розподілення: інфраструктура для екзабайтних даних включатиме багато розподілених сховищ, розміщених в різних країнах і доступних на міжнародному рівні. Набори даних зберігатимуться на різних сайтах, можливо, в єдиному форматі замість теперішніх численних стандартів, по суті, по одному для кожної наукової дисципліни. Можливості агрегування і організації паралельного доступу до даних допомагатимуть клієнтським е-науковим застосуванням з різних галузей (наприклад, з проблеми зміни клімату) в отриманні потрібних даних з розподілених сховищ для вивчення і аналізу.

***Копіювання*:** копіювання (реплікація) даних представляє собою головний засіб поліпшення локалізації даних , підвищення їх доступності, зменшення ризиків втрати. Процес копіювання екзабайтних даних повинен враховувати можливі перешкоди при передачі даних (наприклад, мережеві відмови і відмови сховищ) і відновлювати дані в таких випадках. Копіювання даних потрібує подальше дослідження нових алгоритмів, протоколів, схем копіювання, стратегії розміщення сховищ. Буде потрібно дослідити взаємодію між каталогами копій і пристроями збереження даних і проблеми тривалості життя точної копії. Удосконалена контролююча система дозволить перевіряти статус кожної точної копії в термінах наявності, пропускної

спроможності (мережевого часу зчитування/очікування), використання, і т.п.. Нові точні копії повинні автоматично створюватися, виходячи з потреб користувача і інформації про попередній доступ.

Дані про дані (метадані): це описова інформація про дані, яка пояснює вимірювані атрибути, їх імена, одиниці виміру, точність, формат даних та інше. Найбільш важливе те, що метадані включають інформацію про походження даних, описують, як вимірювалися, генерувалися або обчислювалися дані. Використання метаданих спрощує доступ до даних, їх взаємообмін і інтеграцію, забезпечує розуміння даних як інструментальними засобами, так і людьми.

Зберігання і кешування даних: технології зберігання є важливою частиною менеджменту великих об'ємів даних. Технічне забезпечення і програмні компоненти гратимуть вирішальну роль в управлінні великими наборами наукових даних. Паралельний вхід/ вихід будуть критичними з декількох причин: швидкість генерації даних (важлива для зберігання, коли паралельні застосування реалізуються на мільйонах ядер) і швидкість зчитування даних (що здійснюється інструментами аналізу) - це тільки два з можливих прикладів. Потрібен новий ефективний алгоритм кешування для покращення менеджменту даних. Потрібне впровадження нового покоління об'єктно-реляційних систем баз даних, які сприймають будь-який тип даних (будь то звичайне число, масив, рядок символів або складений об'єкт, такий як XML або HTML - документ) як інкапсульований тип, значення якого можуть зберігатися в полі запису. Такі системи баз даних забезпечують потужний асоціативний пошук (пошук за значенням, а не за містом розташування), а також автоматичний паралельний доступ і виконання, що істотно для аналізу петабайтних даних.

Доступ до середовища даних: компоненти середовища екзабайтних даних (сховища, сервіси метаданих, реєстратори, інструментарій аналізу, онтології і т.п.) повинні бути сумісними і взаємодіяти між собою безпосередньо. Уніфіковані інтерфейси повинні приховати різнорідність основних систем і спрощувати доступ до наявних ресурсів. Доступ до основної інфраструктури має бути і повсюдним. Інтегроване середовище, що базується на веб-сервісах, гратиме ключову роль в зміні і поліпшення щоденної діяльності наукових користувачів.

Окрім того, можливості удосконалених систем управління маршрутами виконання сервісів (***workflow-систем***) дозволять дослідникам складати комплексні наукові завдання для розподілених сховищ, обчислювальних ресурсів, баз даних і т.д.. Портали даних наступного покоління забезпечать високий рівень партнерської функціональності в щоденній діяльності дослідників і вчених. Можливості соціальних мереж збільшать рівень обговорень, зворотного зв'язку, обміну науковими результатами і їх розповсюдження серед тематичних груп, науковим команд і т.д.

Пошук і відкриття даних: в екзабайтному середовищі процедури пошуку і відкриття даних продовжать грати вирішальну роль. Будуть використовуватися одночасно декілька джерел метаданих , вносячи свій вклад в опис наборів існуючих даних. Ієрархії метаданих і індексів допомагатимуть в прискоренні пошуку і відкритті даних в такому великому масштабі. Вони будуть основою для створення запитів, їх композиції, відновлення і фільтрації даних.

Ефективні мережеві роботи будуть збирати інформацію про метадані з різних сайтів, відображаючи доступні розподілені набори даних. Автоматичне зчитування метаданих гратиме ключову роль в облегшенні і прискоренні використання метаданих . Інформація про походження даних буде все більш і більш важливішою для ідентифікації, простежуванням і реєстрації історії даних, для етапів оброблення і аналізу даних.

2.Розподілені обчислення змінюють скрізь взаємодію людей і об'єктів з цифровим світом, роблячи персональні пристрої домінуючим засобом інформаційного доступу.

Мова йде про використання природних засобів взаємодії, притаманних людям, не тільки через звичайну мову (як то вже реалізовано в комп'ютері ІБМ «Watson»), але і мову жестів (положення тіла, пильний погляд, ручні рухи) , щоб виказати свою емоцію, настрій, відношення і увагу (Multimodal Human computer interaction).

3.*Постачання знань і цінної інформації через інтелектуальні інфраструктури IT* .

Економічні стимули все більш і більш підтримують постачання інформації через IT або сервіси. Інтелектуальна інфраструктура дозволяє гнучко збалансувати централізацію з

децентралізацію управління, враховуючи обмеження пропускній спроможності локальних мереж, час очікування і різнорідність обчислювальних платформ. В екзамасштабі для аналізу даних будуть потрібні нові математичні підходи, алгоритми і пов'язані з ними паралельні реалізації, здатні завантажити велику кількість доступних процесорів і також забезпечити ефективні доступ і менеджмент даних у файловій системі і на рівні зберігання.
Аналіз даних. При наявності петабайтних наборів даних вимагається нова методологія роботи наукових центрів, яка передбачає переміщення прикладних програм до даних і передачу в наукові центри тільки запитів і отримання відповіді, а не переміщення початкових даних і додатків в локальну систему користувача.

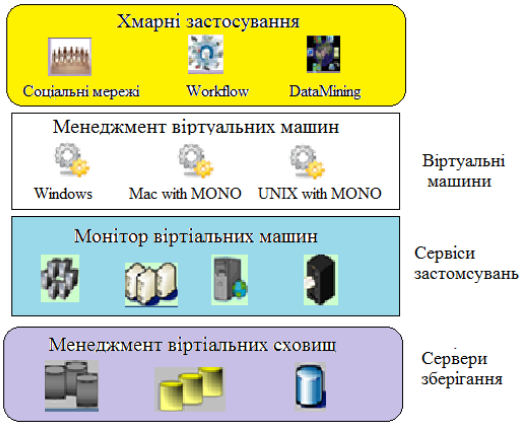

Рис.1. Типовий центр наукових даних

Багаторівнева архітектура типового центру даних, заснованого на хмарі, показана на рис. 1. На її нижніх рівнях розташовані фізичні ресурси центру (сервери зберігання і сервери застосувань), які визначають його потужність. Ці сервери прозоро керуються вище розташованими рівнями віртуалізації сервісів і інструментарію, які допускають спільне їх використання віртуальними серверами. Ці віртуальні сервери ізолюються один від одного, що допомагає досягнути необхідний рівень безпеки.
Хмарні застосування , такі як соціальні мережі, ігрові портали, ділові застосування, доставка змісту , і наукові workflows-системи діють на найвищому рівні архітектури. Використовується нова технологія інтелектуального аналізу даних з метою виявлення прихованих закономірностей у вигляді значущих особливостей, кореляцій, тенденцій і шаблонів. Сучасні системи вилучення знань з «сирих» дозволяють знаходити розчинену в петабайтних сховищах не очевидну, але вельми цінну інформацію [4]. В основу технології Data Mining встановлена концепція шаблонів (pattern), що відображають фрагменти багатоаспектних взаємостосунків в даних. Цими шаблонами є закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у формі, зрозумілій людині. Пошук шаблонів проводиться методами, не обмеженими рамками апріорних припущень про структуру вибірки і вид розподілів значень аналізованих показників.

4.*Створення компонентів програмного забезпечення вищого рівня у вигляді сервісів.*
Сьогодні платформа оброблення даних визначається більше самими даними, а її архітектура орієнтована на сервіси, з яких процедурою композиції можна за бажанням користувача складати прикладні додатки оброблення даних. Серед головних принципів такої сервісно-орієнтованої архітектури (SOA) виділяють наступні: максимальне повторне використання, модульність, здатність до поєднання (композиції), функціональна сумісність, відповідність стандартам. Для роботи з петабайтними наборами даних вимагаються величезні масиви пам'яті і тисячі обчислювальних вузлів, що сьогодні найефективніше забезпечується грід- або хмарними обчислювальними інфраструктурами. Через засновані на хмарних обчисленнях сервіси , які здатні запропонувати високий рівень надійності, масштабованості (через динамічне

постачання ресурсів) і безпеки, можна керувати зберіганням даних, метаданами і науковими доменними онтологіями, розташованими в центрах даних. Висока пропускна спроможність підключення до Інтернет і наявність стандартів дозволить програмним сервісам слугувати структурними блоками при формуванні мережевих застосувань більш високого рівня.

Сумісність (interoperability): із-за великого розміру середовища , різнорідності платформ і складності екзамасштабних систем сумісність грає важливу роль в забезпеченні продуктивної взаємодії всіх задіяних компонентів, сервісів і акторів. Сумісність може бути досягнута шляхом прийняття відповідних стандартів . З іншого боку, процес стандартизації повинен заохочувати розробників звертатися до реальних потреб і призводить до прийняття ефективних і широко погоджених документів. Сумісність робить дійсно "відкритим" екзамасштабне середовище. Стандарти повинні охоплювати проблеми метаданих, протоколів даних і інтерфейсів сховищ

5.*Майбутне IT технологій пов'язане з персональними хмарними обчисленнями*

Хмарні обчислення - модель зручного за вимогою мережевого доступу до розподіленої множини конфігуруємої обчислювальних ресурсів (мереж, серверів VM, сховищ, додатків і сервісів), які можуть швидко вибиратися і змінюватися з мінімальними менеджерськими зусиллями або з мінімальною взаємодією з постачальниками послуг. Хмарна обчислювальна модель обіцяє значну економію витрат, поєднаних з зростаючим застосуванням IT[5]. Вважається переконливим, що уряди і промисловість починають користуватися цією технологію в умовах сьогоденних економічних труднощів.

Висновки

Доступ до інформації взагалі і до наукових даних, зокрема, критичний до безперервного наукового прогресу і технологічного прогресу. Зараз йдеться про обробку петабайтних наборів даних, що потребує розроблення інтелектуальних методів організації даних для скорочення об'єму пошуку, паралельної обробки і доступу до даних при виконанні пошуку у величезних наборах. Для роботи з петабайтними наборами даних вимагаються величезні масиви пам'яті і тисячі обчислювальних вузлів, що сьогодні найефективніше забезпечується грід- або хмарними обчислювальними інфраструктурами. Сьогодні платформа оброблення даних визначається більше самими даними, а її архітектура орієнтована на сервіси, з яких процедурою композиції можна за бажанням користувача складати прикладні додатки оброблення даних

Література

1. Martin Hilbert, Priscila López. The World's Technological Capacity to Store, Communicate, and Compute Information, February 10, 2011- www.sciencexpress.org / 10 February 2011 / Page 5 / 10.1126/science.1200970

2. Accenture Technology Vision 2011 - The Technology Waves That Are Reshaping the Business Landscape- www.accenture.com/us-en/technology/technology-labs/Pages/insight-accenture-technology -vision-2011.aspx

3. Paul Horn.The Future of Information Technology (ppt),University of Colorado,September 14, 2000 – www.cs.colorado.edu/events/lectures/horn/horn.pdf

4. Петренко А.І. Grid і інтелектуальна обробка даних.-\\Системні дослідження і інформаційні технології.-Київ, №4, 2008.-с.97-110.

5.European Commission (2010), The future of Cloud computing: opportunitied for European Cloud computing beyond 2010' – www.cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf