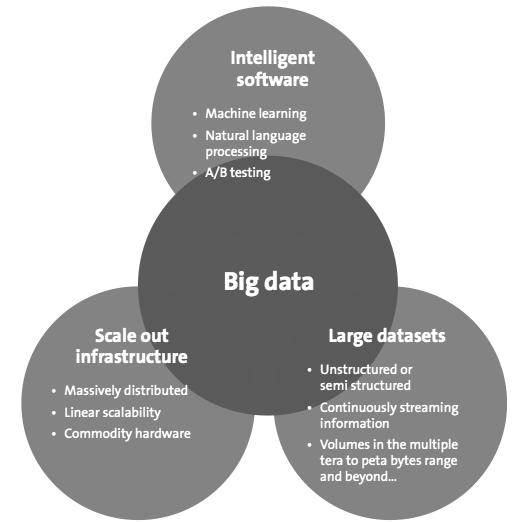
Система кластерних обчислень Apache Hadoop

***Rotenberg Oleksandr***

*Faculty of Information Technology and Computer Engineering, NTUU ‘KPI’*

Великі дані - серія підходів, інструментів і методів обробки структурованих і неструктурованих даних величезних обсягів і значного різноманіття для отримання сприйманих людиною результатів. Як визначальних характеристик для великих даних відзначають «три V»: обсяг (англ. Volume, в сенсі величини фізичного обсягу), швидкість (англ. Velocity в сенсах як швидкості приросту, так і необхідності високошвидкісної обробки і отримання результатів), різноманіття (англ . variety, в сенсі можливості одночасної обробки різних типів структурованих і напівструктурованих даних).

**Apache Hadoop – набір програмних засобів з відкритим вихідним кодом для розробки і виконання розподілених програм обробки Big Data, які працюють на кластерах з сотень і тисяч вузлів. Apache Hadoop успішно використовується в науці та різних галузях промисловості. Наукові застосування включають математику, фізику високих енергій, астрономію, генетику і океанографію. Починаючи як обчислювальна платформа для пошукових систем, Apache Hadoop в даний час використовується для зберігання даних, поведінкового аналізу, крипто аналізу, метеорології, виявленні спаму, обробці природньої мови, генному аналізі, обробці зображень і т.д. Був використаний для реалізації пошукових і тематичних механізмів багатьох високонавантажених веб-сайтів, в тому числі, для Yahoo!, Facebook, Last.fm і eBay.com [1, 2, 3].**

**Також Apache Hadoop використовували для обчислення двохквадрильйонної (1015) цифри числа π, яка виявилася нулем, та для допомоги IBM’s Watson перемогти у телешоу Jeopardy [4, 5].**

**Заснований на ідеях, викладених у двох статтях, виданих корпорацією Google Inc.:**

* + - 1. **The Google File System - пропрієтарна розподілена файлова система GFS, призначена для роботи розподілених додатків, що обробляють велику кількість даних (порядку сотень терабайт) [6].**
      2. **MapReduce: Simplified Data Processing on Large Clusters - модель розподілених обчислень, використовувана для паралельних обчислень над дуже великими (кілька петабайт), наборами даних в комп'ютерних кластерах [7].**

Технологія Big Data, зокрема, гнучка, добре масштабована, і використовує хмарні технології. Ключова роль в обробці і зберіганні великих обсягів даних належить аналітичним інструментам і методам, таким як моніторинг в реальному часі, прогнозного моделювання та використання візуальних панелей даних.

В результаті використання технології Big Data, організація має можливість отримати важливу інформацію протягом декількох секунд, що дозволяють підвищити ефективність економічних рішень, щоб швидше реагувати на зміни в поведінці клієнта і виявити тенденції ринку на самих ранніх етапах теперішній час.

# Джерела

1. Looking at the code behind our three uses of Apache Hadoop –Електронний ресурс. Режим доступу: http://www.facebook.com/notes/facebook-engineering/looking-at-the-code-behindour-three-uses-of-apache-hadoop/468211193919
2. PoweredBy - HadoopWiki–Електронний ресурс. Режим доступу: http://wiki.apache.org/hadoop/PoweredBy
3. Apache Hadoop : The Scalability Update - Електронний ресурс. Режим доступу : <https://www.usenix.org/system/files/login/articles/105470-Shvachko.pdf>
4. Tsz-Wo Sze, “The Two Quadrillionth Bit of Pi Is 0! Distributed Computation of Pi with Apache Hadoop,” arXiv:1008.3171v2 [cs.DC], August 2010
5. Making Watson fast (E. A. Epstein, M. I. Schor, B. S. Iyer, A. Lally,

E. W. Brown,J. Cwiklik ) - Електронний ресурс. Режим доступу : <http://www.andrew.cmu.edu/user/ooo/watson/15%20making%20watson%20fast.pdf>

1. The Google File System –Електронний ресурс. Режим доступу:

http://research.google.com/archive/gfs-sosp2003.pdf

1. MapReduce: Simpli\_ed Data Processing on Large Clusters –Електронний ресурс. Режим доступу: http://research.google.com/archive/mapreduce-osdi04.pdf

Konstantin v. Shvachko