BIGDATA

ThS. Phạm Đức Tú

Phòng NCPT Ứng dụng Viễn thông

Tóm tắt: Ngày nay cùng với việc gia tăng chóng về mặt về số lượng cũng như sự đa dạng dữ liệu dựa trên nền tảng của thông tin số hội tụ toàn cầu như các website thương mại điện tử, mạng xã hội như Facebook, G+...ngày càng làm các nhà kinh doanh mong muốn nhiều hơn các giá trị quý giá đem lại từ các loại dữ liệu đó. Việc đó đồng nghĩa với các nhà khoa học phải đau đầu đối phó với việc lưu trữ, xử lý khối lượng số liệu khổng lồ và đa dạng về chủng loại dữ liệu. Vì vậy nhiều các nhà khoa học đang nghiên cứu các công nghệ, thuật toán để giải quyết bài toán về lưu trữ, xử lý và phân tích các loại dữ liệu lớn(Bigdata) một cách nhanh nhất đáp ứng được yêu cầu của các nhà quản trị kinh tế hoặc phân tích thị trường. Một trong những giải pháp về Bigdata mã nguồn mở rất nổi tiếng đang được rất nhiều các nhà khoa học trên thế giới quan tâm nghiên cứu và hoàn thiện đó chính là Hadoop. Giải pháp Hadoop đem lại rất nhiều tính năng ưu việt trong việc lưu trữ và tính toán xử lý song song trên nhiều máy chủ với số liệu rất lớn trong thời gian rất ngắn. Bài báo này giới thiệu về giải pháp công nghệ tổng quan về Bigdata và giải pháp Hadoop trong vấn đề xử lý dữ liệu lớn.

1. GIỚI THIỆU

Nói về Bigdata thì người ta sẽ quan tâm chính tới ba đặc trưng cơ bản mà một hệ thống Bigdata phải thỏa mãn bao gồm ba yếu tố *3Vs:*

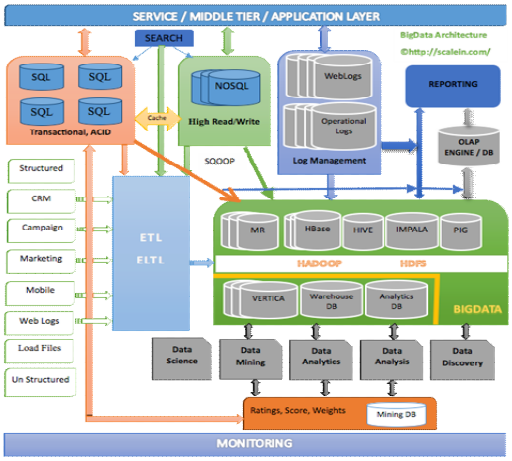
- **Volume** : số lượng dữ liệu cần xử lý cực lớn
- **Velocity** : tốc độ xử lý nhanh (thu thập, xử lý, đáp trả)
- **Variety** : đa dạng loai dữ liệu có cấu trúc và phi cấu trúc

Hệ thống Bigdata không chỉ xử lý các loại dữ liệu truyền thống có cấu trúc mà nó còn lưu trữ và xử lý các loại dữ liệu phi cấu trúc như sau:

- Documents
- Existing relational databases (CRM, ERP, Accounting, Billing)
- E-mails and attachments
- Imaging data (graphs, technical plans)
- Sensor or device data
- Internet search indexing
- Log files
- Social media
- Telephone conversations
- Videos
- Pictures
- Clickstreams (clicks from users on web pages)

2. MÔ HÌNH XỬ LÝ BIGDATA

Big data được xử lý thông qua 4 giai đoạn: thu thập (acquire), tổ chức (organize), phân tích (analyze), quyết định (decide).



Hình 1. Mô hình xử lý Bigdata

- Giai đoạn thu thập: hầu hết đã có giải pháp, ví dụ: Oracle đưa ra NoSQL Database, Google có Google BigTable…
- Giai đoạn tổ chức: có thể lưu trữ dữ liệu ở dạng phân tán, song song… nhưng phổ biến nhất vẫn là Hadoop/MapReduce.

- Giai đoạn phân tích: với các dữ liệu truyền thống, các công ty lớn đều đã có giải pháp. Ví dụ: Oracle có Oracle Data warehousing, IBM có InfoSphere warehouse…
- Giai đoạn quyết định: dựa vào các thông tin được phân tích sẽ đưa ra các quyết định giải pháp kinh doanh kịp thời

3. GIẢI PHÁP HADOOP

- **Hadoop**: bao gồm 2 thành phần chính là MapReduce và HDFS
  - MapReduce: Frameworks cho lập trình (dùng Java) : Đây là mô hình lập trình cho Hadoop. Có hai giai đoạn được gọi là Map và Reduce, một quá trình mà hệ thống thực hiện sắp xếp và chuyển các kết quả đầu ra của Map tới các đầu vào của các bộ Reduce.
  - HDFS (Hadoop Distributed File System): một hệ thống tập tin phân phối thiết kế để chạy trên phần cứng. Nó có nhiều điểm tương đồng với hệ thống tập tin phân tán hiện có. Tuy nhiên, sự khác biệt từ hệ thống tập tin phân phối khác là rất lớn. HDFS chịu lỗi rất tốt và được thiết kế để triển khai trên phần cứng chi phí thấp. HDFS cung cấp khả năng truy cập cao ứng dụng dữ liệu và phù hợp cho các ứng dụng có tập dữ liệu lớn.

– **Ecosystem - hệ sinh thái**

  - **Hadoop Streaming**: Một tiện ích để tạo nên mã MapReduce bằng bất kỳ ngôn ngữ nào: C, Perl, Python, C++, Bash, v.v.
  - **NoSQL:** là 1 dạng CSDL mã nguồn mở có thể lưu trữ nhiều loại dữ liệu khác nhau cả có cấu trúc và phi cấu trúc. NoSQL viết tắt bởi: *None-Relational SQL*, hay có nơi thường gọi là *Not Only SQL*. Một số NoSQL thông dụng như Hbase (a NoSQL tabular store), MongoDB, Cassandra, Oracle NoSQL.

- **Hive và Hue:**
  - Hive giúp code bằng SQL và yêu cầu chuyển thành tác vụ MapReduce
  - Hue là giao diên GUI cho việc viết Hive
- **Pig**: Một môi **trường** lập trình mức cao hơn để viết mã MapReduce
- **Sqoop**: Tool cung **cấp** việc truyền dữ liệu hai chiều giữa Hadoop và cơ sở dữ liệu quan hệ.

4. THÁCH THỨC CỦA BIGDATA

- Khai thác big data đòi hỏi hạ tầng công nghệ lớn
- Khả năng lưu trữ cực lớn
- Nhiều máy tính cỡ lớn để xử lý song song
- Công cụ, phần mềm phân tích chuyên dụng
- Trình độ nhân lực chuyên gia về lĩnh vực Big Data còn rất ít
- Việc phân tích được chi tiết hành vi người tiêu dùng từ mọi góc độ, cũng đồng nghĩa với việc xâm phạm quyền riêng tư cá nhân.

5. ỨNG DỤNG BIGDATA

Bigdata được ứng dụng rất nhiều trong các lĩnh vực sau

– Ngân hàng, chứng khoán
– Viễn thông, Y tế, Giáo dục
– Tài nguyên môi trường, Truyền thông

6. KẾT LUẬN

Việc nghiên cứu và ứng dụng công nghệ vào xử lý Bigdata cần được đầu tư và quan tâm hơn nữa để nâng cao năng lực đội ngũ nghiên cứu và ứng dụng trong các bài toán cần năng lực xử lý dữ liệu lớn cũng như xử lý các loại dữ liệu phi cấu trúc trong tương lai.

7. TÀI LIỆU THAM KHẢO

1. *http://www.hadoop.apache.org*
2. *https://www.ibm.com/developerworks/vn/library/data/2013Q1/dm1209hadoopbigdata/*

Thông tin tác giả: **Phạm Đức Tú**

Năm sinh : 1977

Lý lịch khoa học:

- 1995-1999 : Khoa CNTT, đại học Khoa học tự nhiên, đại học quốc gia Hà nội
- 2011-2013 : Thạc sĩ, khoa học máy tính, Học viện CNBCVT

Hướng NC đang theo đuổi : Nghiên cứu công nghệ trong lĩnh vực xử lý BigData

Email : tupd@ptit.edu.vn ; phamductu@cdit.com.vn