ĐẠI HỌC KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI

KHOA CNTT

Big data

CHƯƠNG 6: Công nghệ phân tích

- **Bộ công cụ phân tích dữ liệu lớn**
- **Công nghệ Apache Hadoop**
  - Giới thiệu Hadoop
  - Kiến trúc Hadoop
- **Công nghệ Apache Spark**
  - Giới thiệu Spark
  - Thành phần Spark
- **Công nghệ khác**

Công nghệ Apache Hadoop

- **HADOOP là gì?**

Hadoop là một Apache framework nguồn mở viết bằng Java cho phép phát triển các ứng dụng phân tán có cường độ dữ liệu lớn một cách miễn phí. Nó được thiết kế để mở rộng quy mô từ một máy chủ đơn sang hàng ngàn máy tính khác có tính toán và lưu trữ cục bộ.

Hadoop được phát triển dựa trên ý tưởng từ các công bố của Google về mô hình Map-Reduce và hệ thống file phân tán Google File System (GFS). Ccấp cho chúng ta một môi trường song song để thực thi các tác vụ Map-Reduce.

Nhờ có cơ chế streaming mà Hadoop có thể phát triển trên các ứng dụng phân tán bằng cả java lẫn một số ngôn ngữ lập trình khác như C++, Python, Pearl,...

- **Kiến trúc của Hadoop**
  - HDFS (Hadoop Distributed File System): lớp lưu trữ
  - Map-Reduce: lớp xử lý dữ liệu
  - YARN (Yet-Another-Resource-Negotiator): lớp quản lý tài nguyên
- **Tham khảo:** *http://hadoop.apache.org/docs/current/index.html*

Công nghệ Apache Spark

- **SPARK là gì?**

Apache Spark là một framework mã nguồn mở tính toán cụm, được phát triển sơ khởi vào năm 2009 bởi AMPLab. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay.

Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming).

- **Thành phần của Spark**
  - Spark Core,
  - Spark Streaming,
  - Spark SQL,
  - Mllib
  - GraphX

- **Tham khảo:** *https://spark.apache.org/*

Công nghệ khác

Có rất nhiều công nghệ được phát triển cho hệ thống Big Data

- ✔ Apache Storm
- ✔ Apache Cassandra
- ✔ Apache Kafka

Apache Kafka

Apache Kafka là một nền tảng xử lý phân tán trực tuyến hoặc theo sự kiện cho phép các ứng dụng xử lý lượng lớn dữ liệu một cách nhanh chóng. Nó có khả năng xử lý hàng tỷ sự kiện hằng ngày. Đây là một nền tảng dễ dàng mở rộng với khả năng chịu lỗi tốt.

Quá trình xử lý hoạt động trực tuyến cần tạo và đăng ký các luồng dữ liệu tương tự như một hệ thống nhắn tin, lưu trữ các dữ liệu này và sau đó phân tích chúng.

Các tính năng chính của Apache Kafka

- Cho phép mở rộng quy mô dễ dàng và không có rủi ro về thời gian ngừng hoạt động.
- Đáp ứng tốt khi làm việc với khối lượng lớn nhiều luồng dữ liệu.
- Cụm máy chủ Kafka có khả năng chịu lỗi cao.
- Kafka cung cấp thông lượng cao cho cả việc gửi và nhận luồng dữ liệu.

Apache Storm

Apache Storm cũng là một nền tảng mã mở phân tích Dữ liệu lớn có thể xử lý không giới hạn các luồng dữ liệu trực tuyến. Storm hỗ trợ tốt việc xử lý theo thời gian thực, chịu lỗi cao, tương thích với tất cả các ngôn ngữ lập trình và cả giao thức JSON.

Apache Storm rất dễ dùng, dễ dàng mở rộng hệ thống và xử lý tốt các khối dữ liệu lớn – phức tạp.

Các tính năng chính của Apache Storm

- Công cụ phân tích dữ liệu lớn với giao diện đơn giản
- Có thể xử lý 1.000.000 thông tin (100 byte) mỗi giây trên mỗi nút.
- Cho phép sử dụng nhóm các thiết bị (cluster) để thực hiện các tính toán song song.
- Trong trường hợp node bị lỗi, hệ thống sẽ tự động duy trì và chuyển công việc sang nút khác.
- Mỗi đơn vị dữ liệu được xử lý ít nhất một lần.


Cảm ơn các bạn đã
lắng nghe!