# **BÀI 1: KHÁI NIỆM VỀ BIGDATA**

## **Định nghĩa Big Data**

- Big Data (Dữ liệu lớn) là tập hợp dữ liệu có khối lượng rất lớn, tốc độ tạo dữ liệu nhanh, và có sự đa dạng cao.

- Không thể xử lý bằng các công cụ truyền thống (ví dụ: Excel, cơ sở dữ liệu quan hệ).

## **Đặc điểm của Big Data – 5V**

Volume: Khối lượng dữ liệu lớn (tính bằng terabyte, petabyte).

Velocity: Tốc độ tạo và xử lý dữ liệu nhanh (gần như theo thời gian thực).

Variety: Đa dạng về loại dữ liệu (dữ liệu văn bản, hình ảnh, video, log...).

Veracity: Tính xác thực và chất lượng dữ liệu.

Value: Giá trị tiềm năng mà dữ liệu mang lại.

## **Ứng dụng của Big Data:**

Thương mại điện tử: Gợi ý sản phẩm, cá nhân hóa trải nghiệm, quản lý kho hiệu quả.  
Ví dụ: Amazon, Shopee.

Mạng xã hội: Phân tích xu hướng, cảm xúc người dùng, kiểm duyệt nội dung.  
Ví dụ: Facebook, Twitter.

Y tế: Hỗ trợ chẩn đoán bệnh, cá nhân hóa điều trị, dự đoán dịch bệnh.  
Ví dụ: IBM Watson Health.

Giao thông – Logistics: Tối ưu tuyến đường, theo dõi vận chuyển, phân bổ tài xế.  
Ví dụ: Grab, Uber.

Tài chính - Ngân hàng: Phát hiện gian lận, chấm điểm tín dụng, hỗ trợ đầu tư.  
Ví dụ: Fintech, ngân hàng số.

Sản xuất công nghiệp: Bảo trì dự đoán, tối ưu dây chuyền, kiểm tra chất lượng.  
Ví dụ: GE, Siemens.

Giáo dục: Cá nhân hóa giảng dạy, phân tích tiến độ học tập, phát hiện học sinh gặp khó khăn.  
Ví dụ: Coursera, edX.

Chính phủ - đô thị thông minh: Giám sát giao thông, phân tích dân số, cảnh báo thiên tai.  
Ví dụ: Singapore Smart Nation.

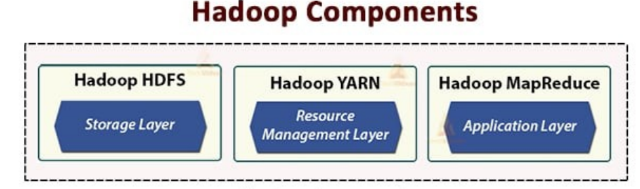
# **Bài 2: Giới thiệu về Hadoop**

## **1. Hadoop là gì?**

Là một nền tảng mã nguồn mở dùng để lưu trữ và xử lý Big Data.

Do Apache phát triển.

## **2. Thành phần chính của Hadoop**



HDFS (Hadoop Distributed File System): Hệ thống lưu trữ phân tán.

MapReduce: Mô hình xử lý dữ liệu song song.

YARN (Yet Another Resource Negotiator): Quản lý tài nguyên và lịch trình job.

Common: Các thư viện dùng chung.

## **3. Ưu điểm của Hadoop**

Xử lý dữ liệu lớn hiệu quả.

Mở rộng quy mô dễ dàng.

Tương đối rẻ vì sử dụng phần cứng thông thường.

---

### 📙 Bài 3: Quản lý Big Data và Big Data Processing

1. Quản lý Big Data:

Cần có hệ thống lưu trữ phân tán.

Quản lý truy cập, bảo mật và quyền người dùng.

Phân tích log, kiểm tra dữ liệu thường xuyên.

2. Big Data Processing (Xử lý dữ liệu lớn):

Batch Processing: Xử lý theo lô (dùng MapReduce, Spark...).

Stream Processing: Xử lý theo thời gian thực (dùng Apache Storm, Kafka, Spark Streaming...).

3. Các công cụ phổ biến:

Apache Spark, Apache Flink, Apache Storm, Hive, Pig...

---

Bạn có muốn mình tạo file Word hoặc PDF từ phần ghi chép này không?

# Bài 4: Giới thiệu về Spark

# Bài 5: Spark RDD

# Bài 6: SparkSQL, Data-frames và Datasets

# Bài 7: SparkSQL và SparkSQL Table

# Bài 8: Data Transformation với Spark

# Bài 9: Data Aggregations và Join trên Spark

# Bài 10: Spark Streaming

# Bài 11: Đọc dữ liệu với Kafka Source và các phép Join với Stream

# Bài 12: Streaming Windowing và Aggregates

# Bài 13: Giám sát & Điều chỉnh

# Assingment 1 – Phân tích hành vi và thói quen của người dùng Stack Overflow

# Phần 3: Apache Airflow

# Bài 14: Giới thiệu về Apache Airflow

# Bài 15: Tạo Data Pipeline với Airflow

# Bài 16: Chạy Data Pipeline song song

# Bài 17: Các khái niệm nâng cao trong Airflow

# NGHIÊN CỨU KHOA HỌC

## 📌 **Đề tài gợi ý:**

### **"Đảm bảo bảo mật và quyền riêng tư trong xử lý dữ liệu lớn: Thực trạng, thách thức và giải pháp"**

## 🎯 **1. Mục tiêu nghiên cứu**

* Phân tích các **rủi ro bảo mật và vi phạm quyền riêng tư** trong hệ thống Big Data.
* Khảo sát các **cơ chế bảo mật và kỹ thuật bảo vệ dữ liệu** đang được áp dụng (ví dụ: mã hóa, phân quyền, ẩn danh dữ liệu).
* Đề xuất mô hình/kỹ thuật **cải thiện bảo mật và bảo vệ quyền riêng tư** trong một hệ thống xử lý dữ liệu lớn cụ thể (ví dụ: hệ thống Spark, Hadoop, hoặc dữ liệu mạng xã hội...).

## ❓ **2. Câu hỏi nghiên cứu**

* Những **lỗ hổng bảo mật phổ biến** trong hệ thống xử lý Big Data là gì?
* Các kỹ thuật hiện có (mã hóa, kiểm soát truy cập, token hóa...) **hiệu quả như thế nào**?
* Làm thế nào để cân bằng giữa **khả năng khai thác dữ liệu** và **bảo vệ quyền riêng tư người dùng**?

## 🧪 **3. Phương pháp nghiên cứu**

* **Khảo sát tài liệu và nghiên cứu hiện tại** về bảo mật Big Data (IEEE, ACM, Springer,...)
* Phân tích một hệ thống Big Data cụ thể (VD: Hadoop, Spark) để xác định điểm yếu bảo mật
* Mô phỏng một mô hình cải tiến: ví dụ triển khai **mã hóa dữ liệu đầu vào + kiểm soát truy cập người dùng**
* Đánh giá hiệu năng hệ thống với và không có bảo mật (so sánh độ trễ, tốc độ xử lý,...)

## 🔐 **4. Một số khái niệm/công nghệ liên quan**

| **Vấn đề** | **Kỹ thuật/Công nghệ** |
| --- | --- |
| **Xác thực và phân quyền** | Kerberos, LDAP, Role-based Access Control (RBAC) |
| **Mã hóa dữ liệu** | AES, RSA, Homomorphic Encryption |
| **Ẩn danh dữ liệu** | k-anonymity, differential privacy |
| **Ghi log và giám sát truy cập** | Apache Ranger, Apache Knox |
| **Bảo vệ trong xử lý phân tán** | TLS/SSL cho Spark, HDFS encryption zone |
| **Luật và quy định** | GDPR, CCPA, ISO/IEC 27001 |

## 📊 **5. Ví dụ hướng triển khai cụ thể**

**Tiểu đề tài**: Ứng dụng kỹ thuật k-anonymity và mã hóa AES trong bảo vệ dữ liệu cá nhân khi phân tích hồ sơ bệnh nhân bằng Apache Spark

* Thu thập dữ liệu mẫu từ Kaggle (VD: Hospital Inpatient Discharges)
* Làm sạch và xử lý dữ liệu bằng Spark
* Ứng dụng kỹ thuật **ẩn danh hóa** tên, địa chỉ, ID (k-anonymity)
* Mã hóa các trường nhạy cảm bằng AES trước khi lưu
* Đánh giá thời gian xử lý, chi phí tính toán, mức độ bảo vệ

## 📚 **6. Tài liệu tham khảo**

* Zikopoulos & Eaton (2011) – Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data
* IEEE Access & ACM Digital Library – chủ đề: Big Data Privacy and Security
* NIST SP 800-53 – Tiêu chuẩn bảo mật thông tin của Mỹ
* OWASP Big Data Security Project

## ✅ **7. Kết luận**

Đây là một đề tài nghiên cứu **có tính ứng dụng cao**, giúp sinh viên hiểu rõ cách triển khai hệ thống Big Data **vừa hiệu quả vừa an toàn**, đồng thời nâng cao ý thức về **đạo đức dữ liệu và quyền riêng tư**.

Nếu bạn muốn, mình có thể giúp bạn **viết đề cương nghiên cứu khoa học chi tiết**, bao gồm cả **mục tiêu**, **giả thuyết**, **kế hoạch thực hiện**, và **tài liệu tham khảo** cụ thể. Bạn có muốn làm tiếp phần đó không?

# 