TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU LỚN

Giảng viên: TS. Nguyễn Văn Quyết

Nội dung

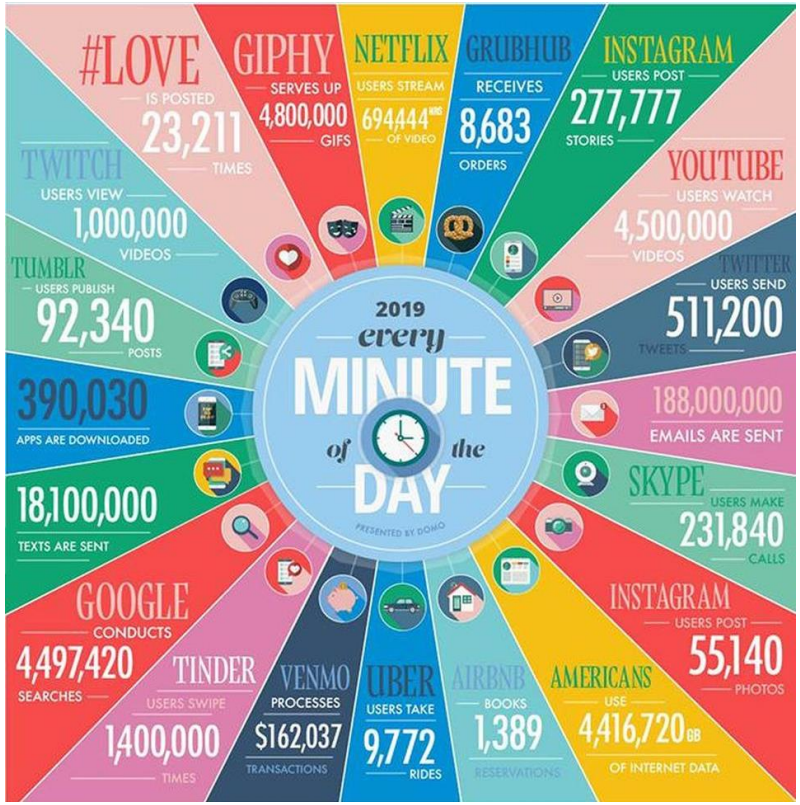
- Giới thiệu về Dữ liệu lớn (Big Data)
- Các kỹ thuật và công cụ cho Phân tích dữ liệu lớn
- Thiết kế các nền tảng Xử lý dữ liệu lớn trong thực tế
- Cài đặt một Big Data Platform
- Hỏi & đáp

Dữ liệu lớn là gì?

- **Dữ liệu lớn** (**Big Data**) là **một tập hợp dữ liệu có kích thước lớn** và **phức tạp** mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được.

Trong 1 phút:
+ 4.5 triệu người xem YouTube
+ 4.5 triệu lượt tìm kiếm trên Google
+ 9.7 nghìn lượt gọi xe Uber

Đến 2025, ước tính **463 exabytes (ET > PT > TB)** dữ liệu sẽ sinh ra / **1 ngày**
~ **212,765,957 DVDs** / **1 ngày**



Dữ liệu lớn đến từ đâu?

"Dữ liệu không bao giờ ngủ"

"Data never sleep"

2020 - Trong 1 phút:

~350K posts trên Instagram (280K – 2019)

~6.6K gói hàng được chuyển đi trên Amazon

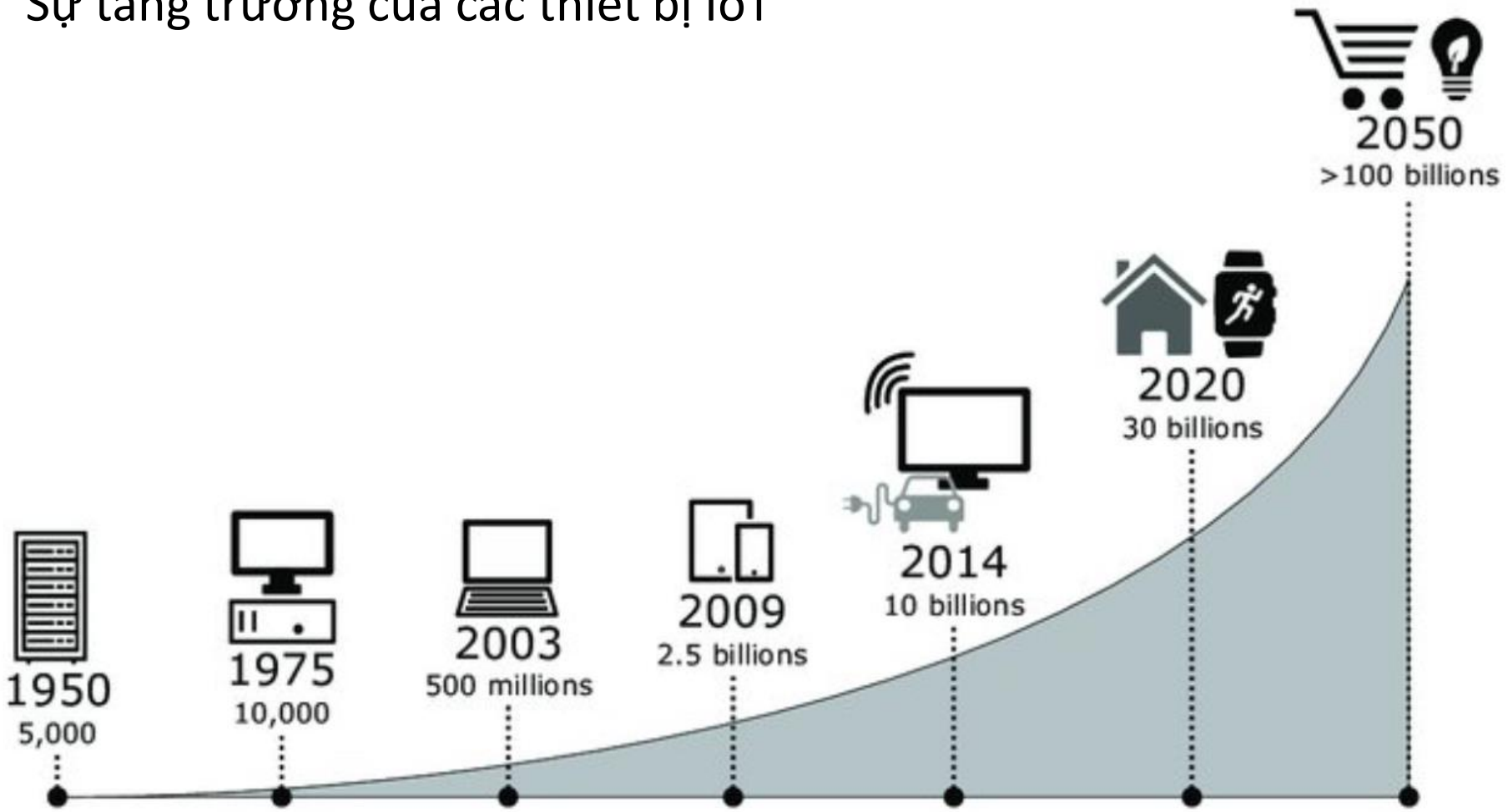
4.57 tỉ người dùng Internet

Tăng 6% so với năm 2019

https://peppytechie.com/2020-data-never-sleeps-version

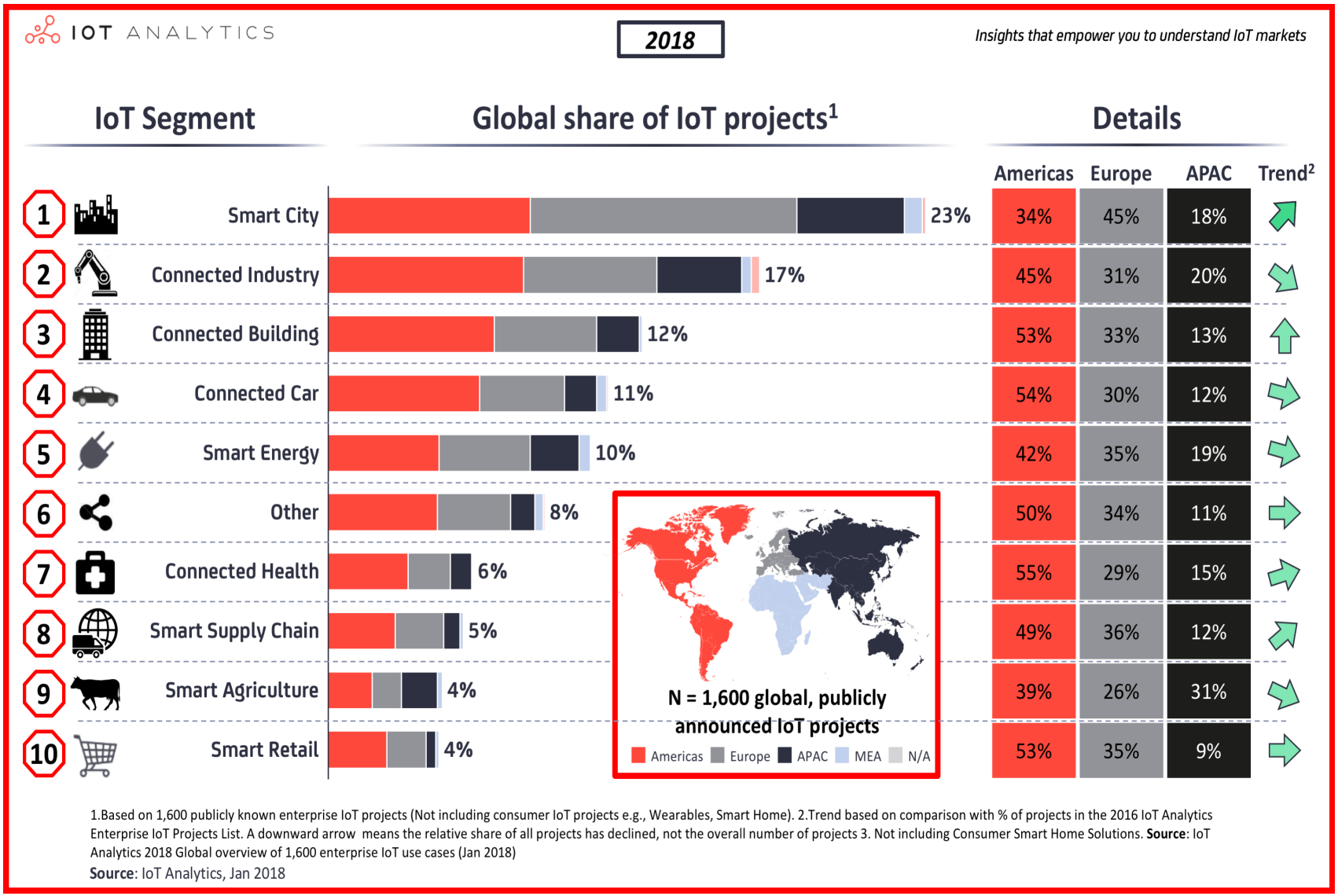
Dữ liệu lớn đến từ đâu?

Sự tăng trưởng của các thiết bị IoT



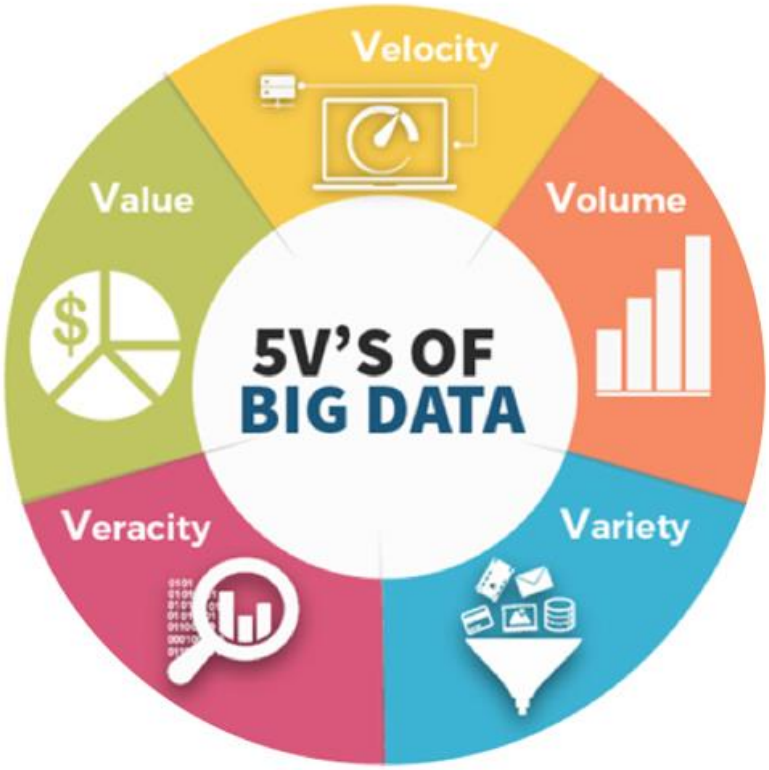
Source: Gartner Website

Dữ liệu lớn đến từ đâu?



Các đặc tính của dữ liệu lớn

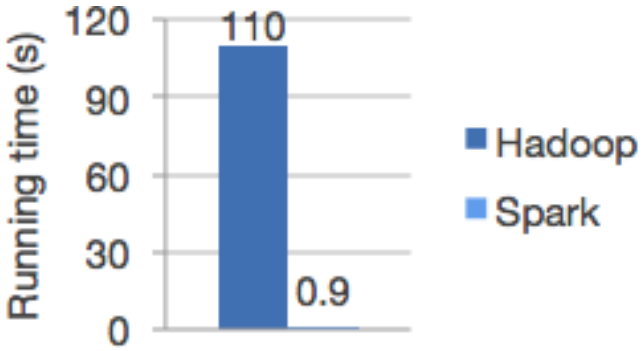
- **The 5V of Big Data**
  - **Volume**: Khối lượng lớn
  - **Velocity**: Tốc độ nhanh
  - **Variety**: Đa dạng
  - **Veracity**: Tính xác thực
  - **Value**: Giá trị
    - Là đặc tính quan trọng nhất của xu hướng sử dụng công nghệ Big Data;
    - Doanh nghiệp phải hoạch định được những giá trị thông tin hữu ích của Big Data cho vấn đề, bài toán hoặc mô hình hoạt động kinh doanh của mình.



➔ Khi xác định được tính chất "**Value**" thì mới nên bắt tay vào Big Data.

Thách thức

- **Truyền dữ liệu**
  - **Phần cứng: Hạ tầng thiết bị**
  - **Phần mềm: Công cụ nào phù hợp với loại dữ liệu gì?**
- **Tốc độ xử lý trong các yêu cầu thời gian thực**
  - **Phần cứng: Hạ tầng thiết bị**
  - **Phần mềm: Thuật toán,…**
- **Nền tảng xử lý dữ liệu**
  - **Hadoop**
  - **Spark**
- **Bảo mật và quyền riêng tư**

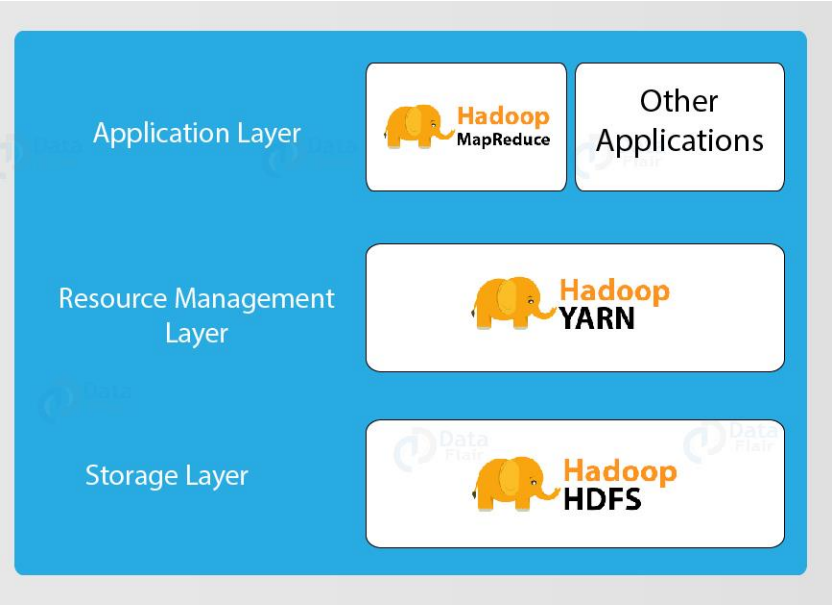


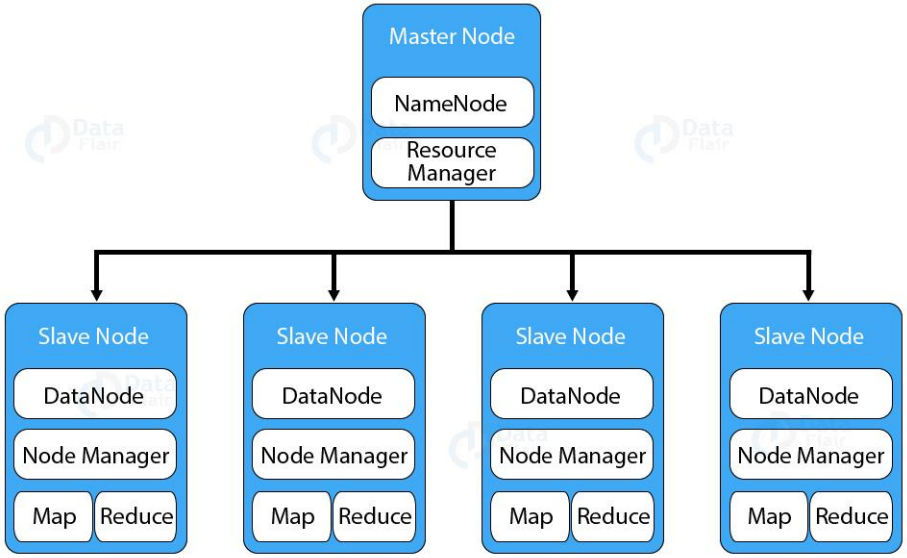
Công cụ phân tích dữ liệu lớn

Giới thiệu về Hadoop

- Hadoop là một framwork giúp lưu trữ và xử lý Big Data áp dụng MapReduce

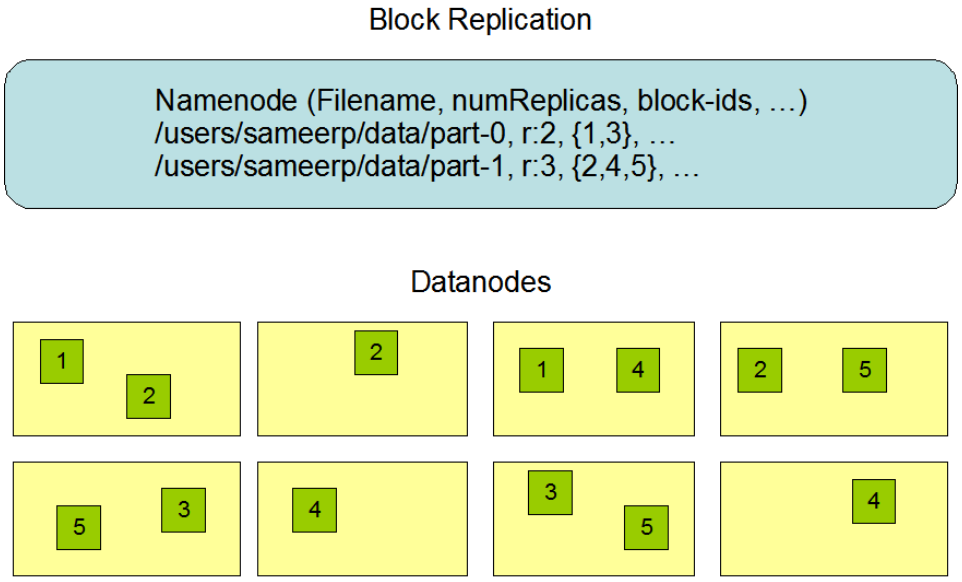
- Kiến trúc một hệ thống Hadoop





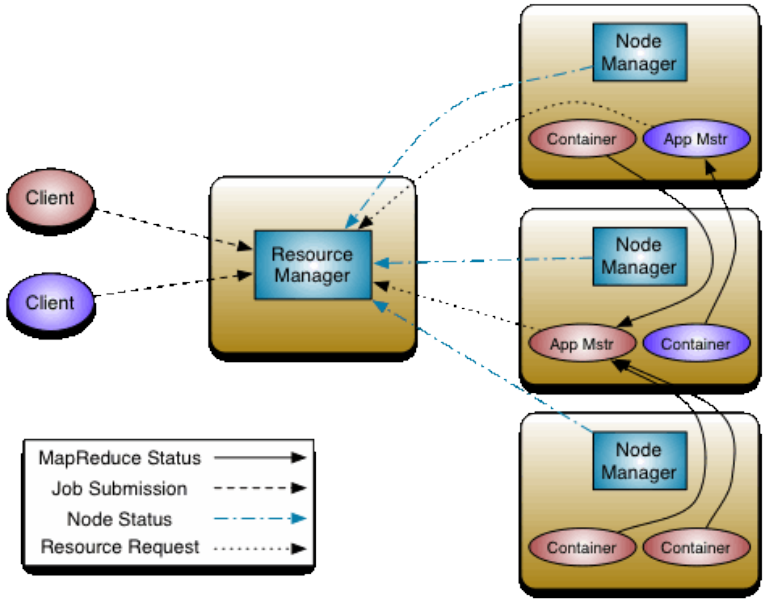
Kỹ thuật và công cụ

- Vai trò các thành phần của Hadoop
  - HDFS (Hadoop Distributed File System)
    - Hệ thống file phân tán cung cấp truy cập thông lượng cao cho ứng dụng khai thác dữ liệu.
    - HDFS là hệ thống tập tin ảo.
    - Một tệp tin lớn trên HDFS được chia thành nhiều mảnh nhỏ, lưu trữ trên nhiều máy chủ khác để tăng sức chịu lỗi và tính sẵn sàng cao.



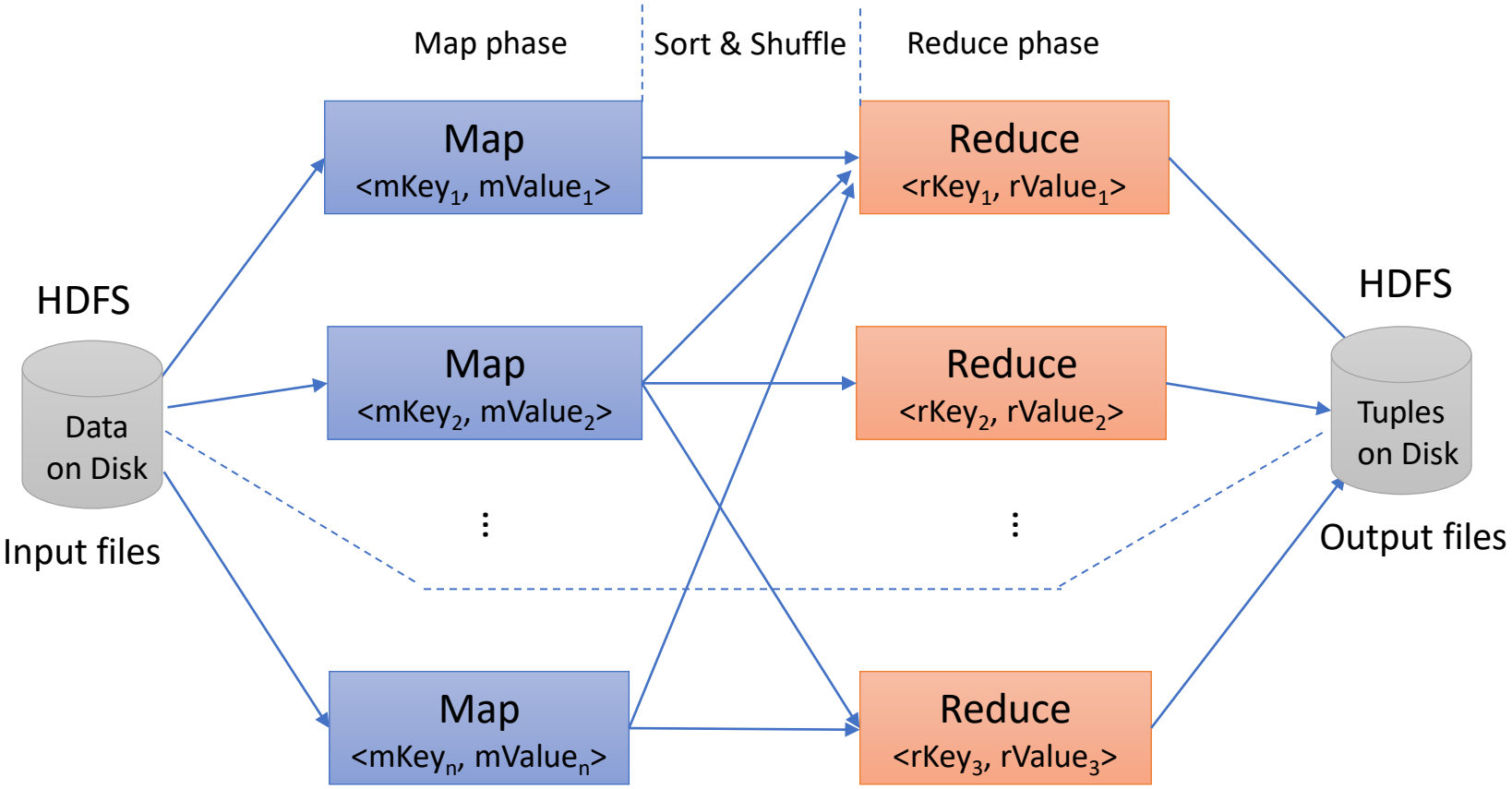
Kỹ thuật và công cụ

- Vai trò các thành phần của Hadoop
  - YARN (Yet-Another-Resource-Negotiator)
    - Quản lý tài nguyên
    - Lập lịch/theo dõi các jobs
    - Thành phần
      - Resource Manager
      - Node Manager
      - Application Master



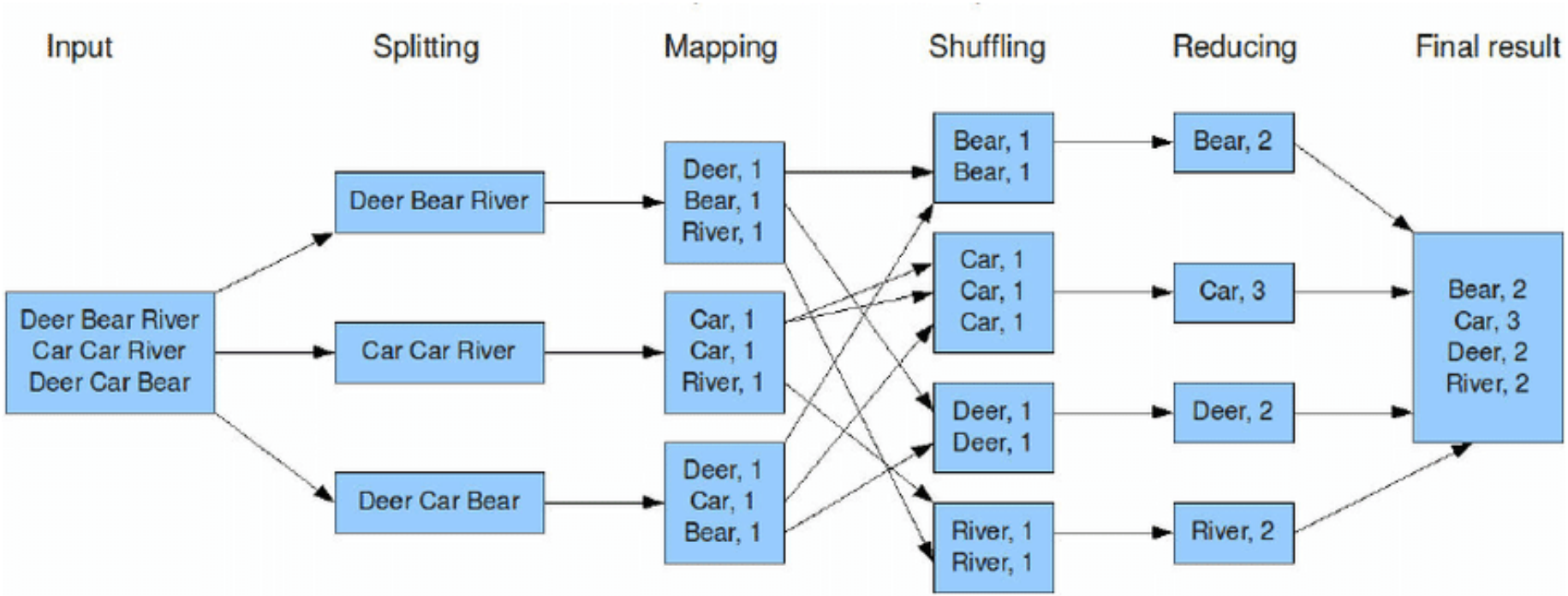
Kỹ thuật và công cụ

- Vai trò các thành phần của Hadoop
  - MapReduce: Mô hình lập trình hỗ trợ xử lý dữ liệu song song phân tán



Kỹ thuật và công cụ

- Vai trò các thành phần của Hadoop
  - MapReduce: Ví dụ 1

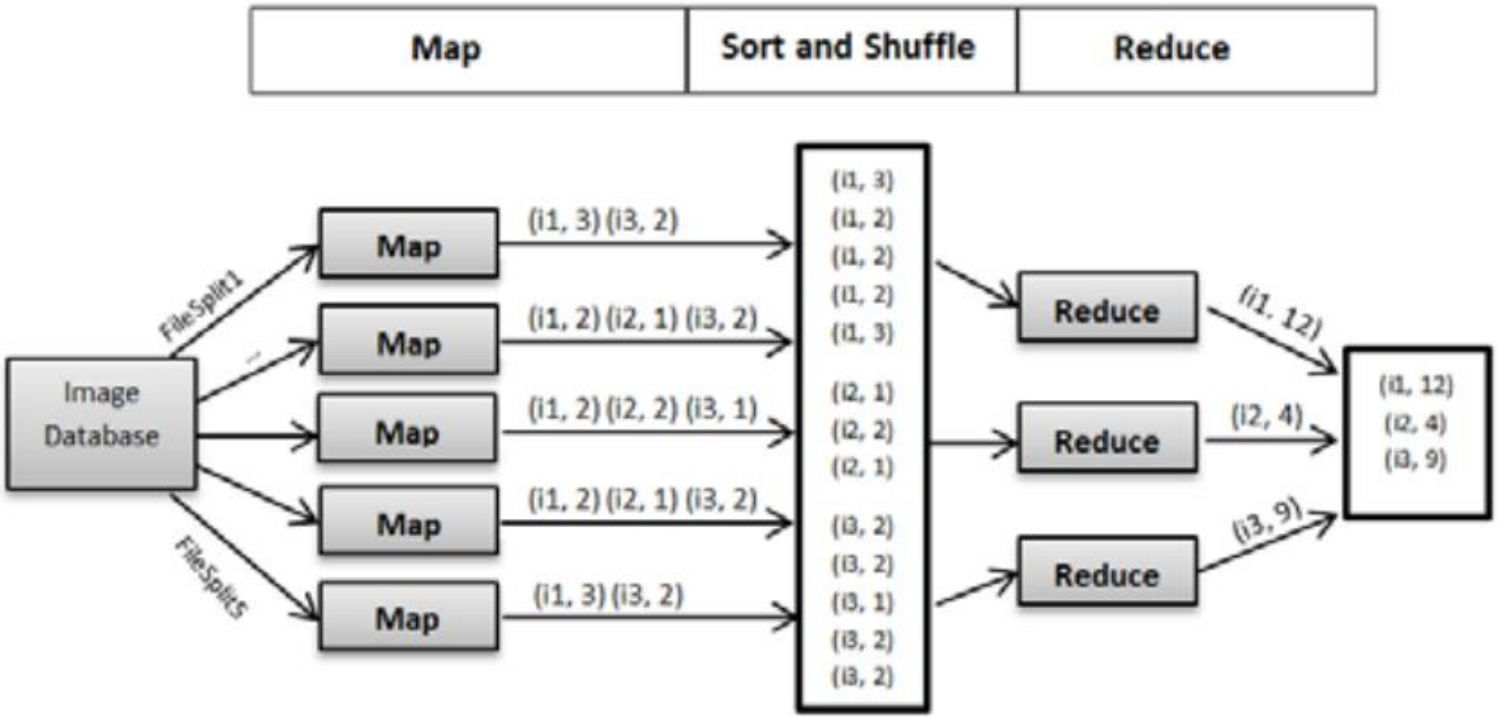


Kỹ thuật và công cụ

- Vai trò các thành phần của Hadoop
  - MapReduce: Ví dụ 2

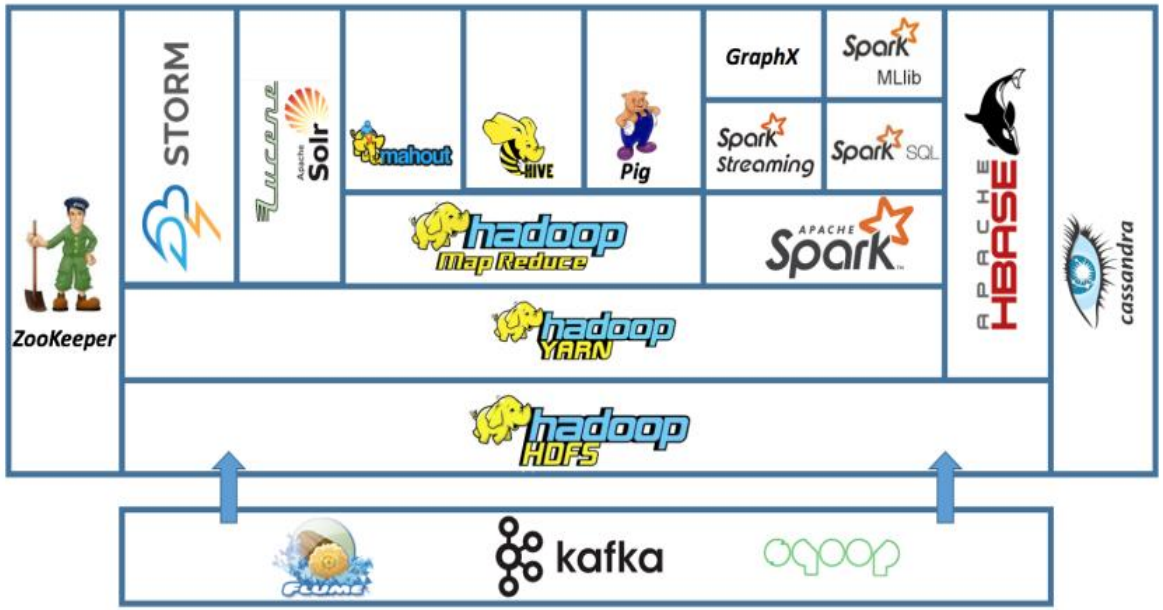
Ảnh 5x5

i3	i1	i1	i1	i3
i1	i3	i2	i3	i1
i1	i2	i3	i2	i1
i1	i3	i2	i3	i1
i3	i1	i1	i1	i3



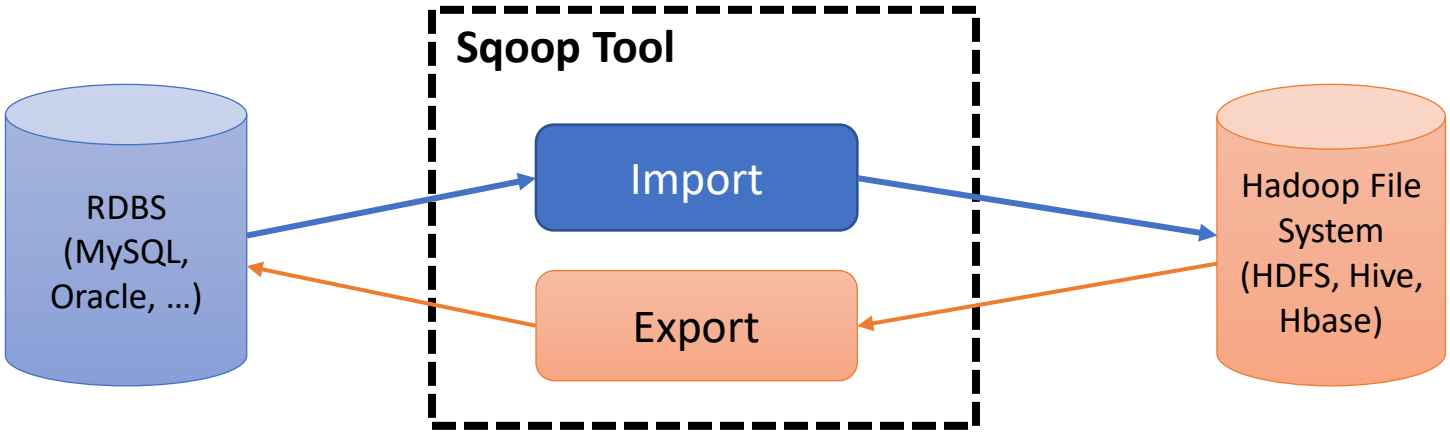
Hệ sinh thái của Hadoop

- Thu thập dữ liệu
  - Sqoop: CSV, SQL, MySQL
  - Flume, Nifi
  - Kafka
- Lưu trữ dữ liệu
  - HDFS
  - Hive
  - Hbase
- Xử lý dữ liệu
  - Hadoop
  - Spark
- Trực quan hóa dữ liệu
  - Zeppellin



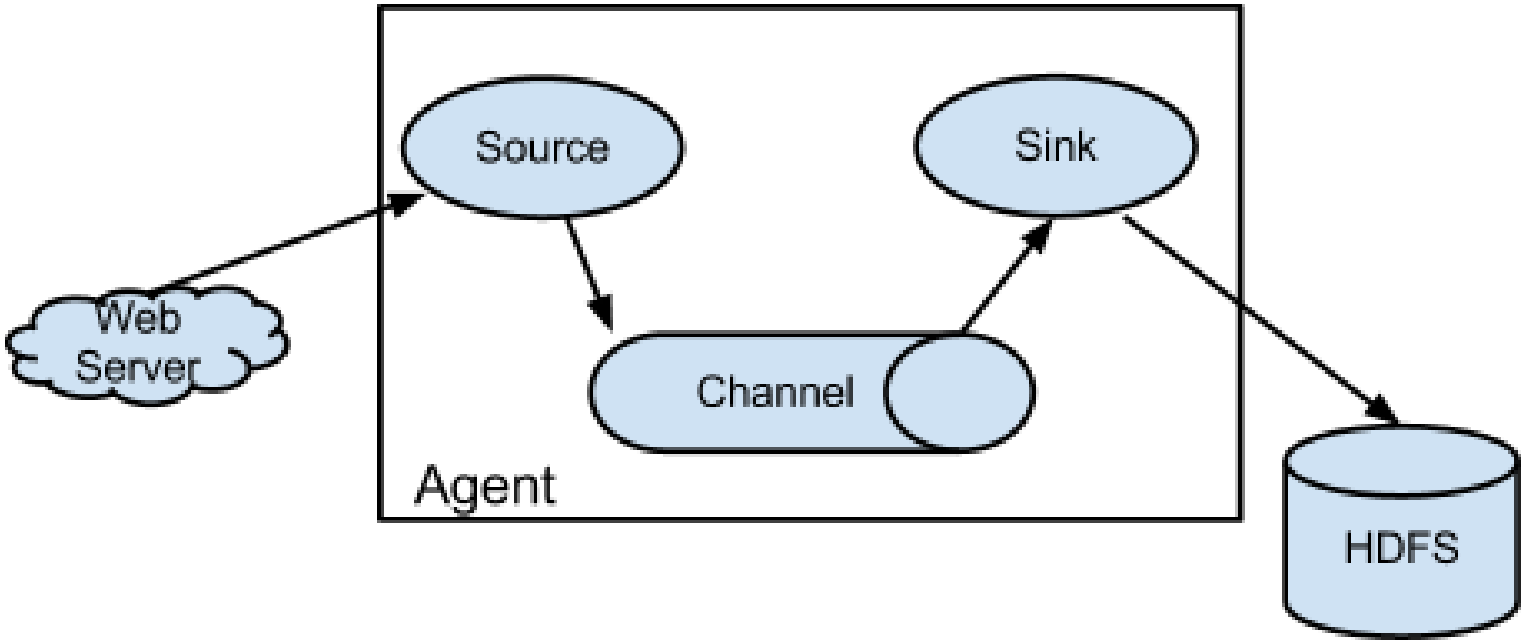
Hệ sinh thái của Hadoop

- Giới thiệu về Sqoop
  - Công cụ hỗ trợ chuyển lượng lớn dữ liệu từ các cơ sở dữ liệu dạng có cấu trúc như RDBS (MySQL, Oracle, …) vào hệ thống Hadoop
  - Quản lý việc chuyển dữ liệu một cách song song từ những nguồn cơ sở dữ liệu khác nhau, xuất ra nhiều định dạng tập tin khác nhau: CSV, Avro, Parquet, SequenceFile;
  - Phiên bản mới nhất Sqoop 1.4.7 (http://sqoop.apache.org/)



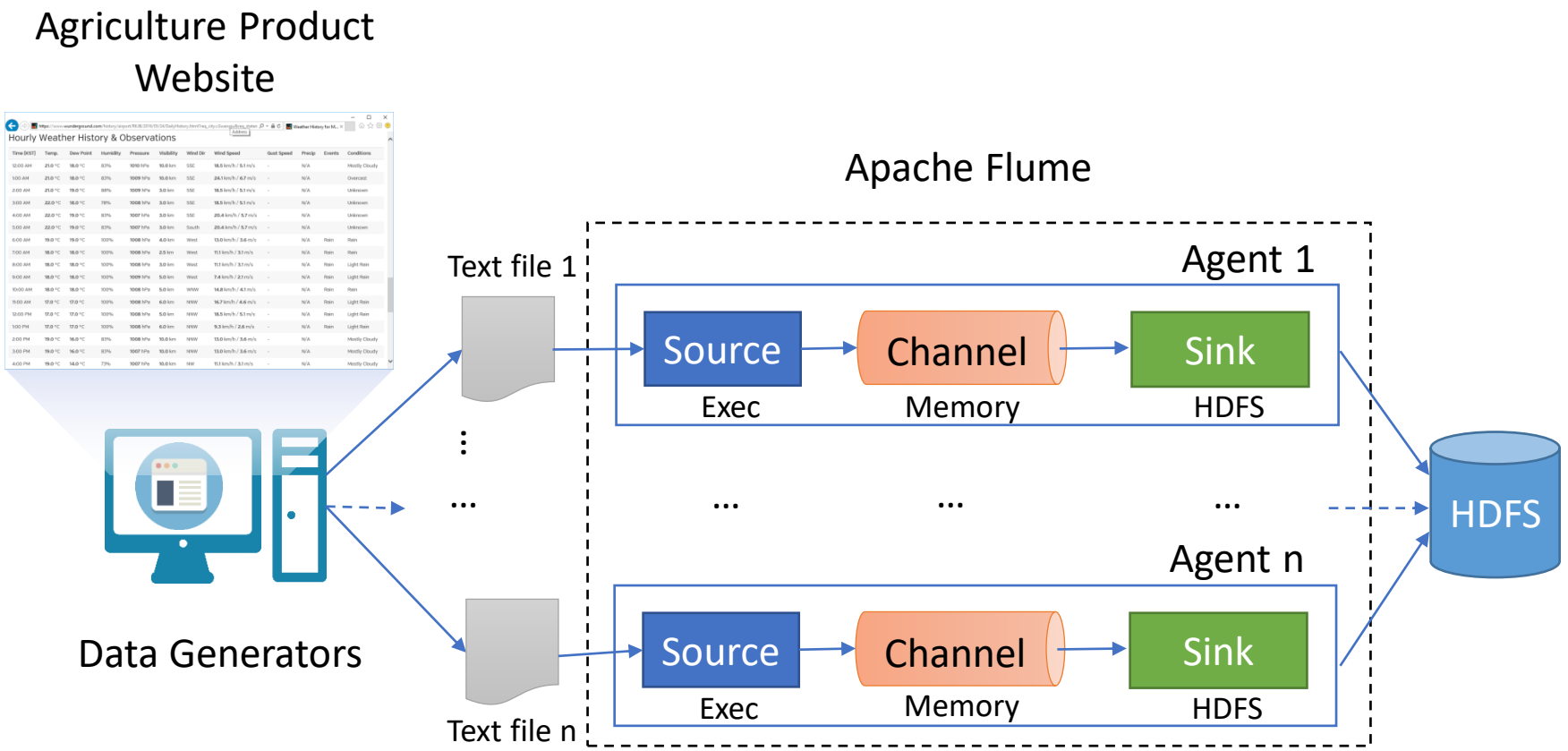
Hệ sinh thái của Hadoop

- Giới thiệu về Flume
  - Công cụ (dịch vụ) thu thập, tổng hợp và di chuyển một lượng lớn dữ liệu sự kiện (streaming event data) trực tuyến một cách hiệu quả.



Hệ sinh thái của Hadoop

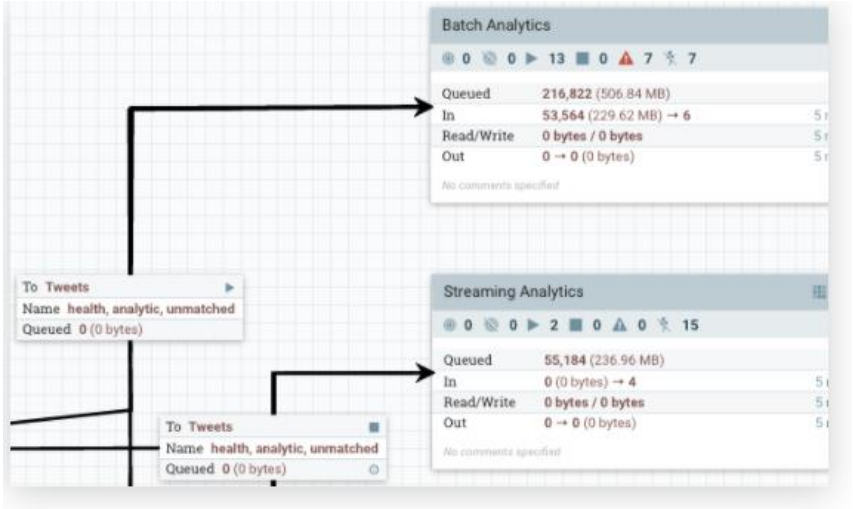
- Giới thiệu về Flume



Flume-based for real-time collecting agriculture data from the web

Hệ sinh thái của Hadoop

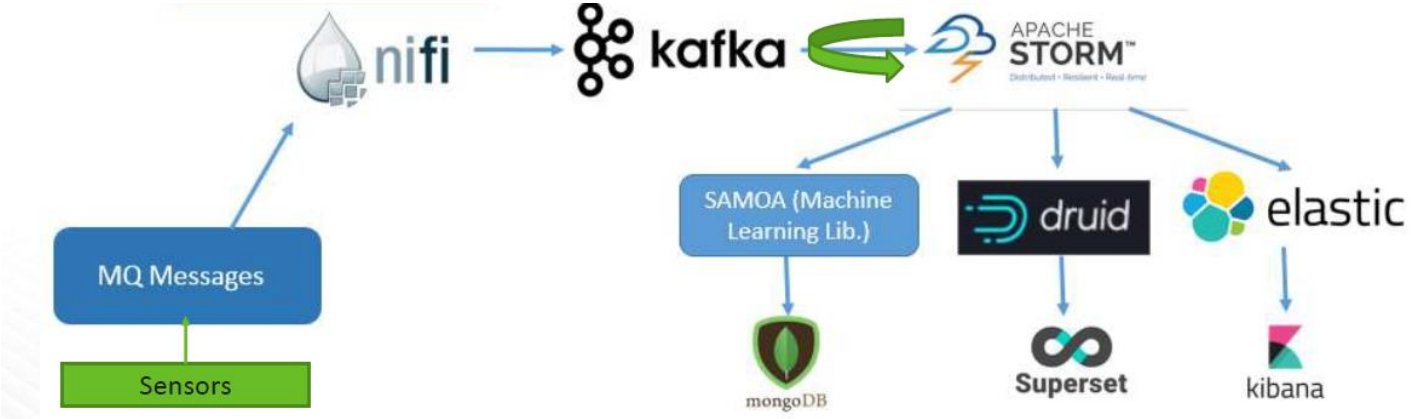
- Giới thiệu về Apache NiFi
  - Công cụ Web hỗ trợ thiết kế, điều khiển luồng dữ liệu giữa các hệ thống
  - Thu thập dữ liệu từ nhiều nguồn
  - Theo dõi luồng dữ liệu từ nguồn bắt đầu đến khi được lưu trữ



Hệ sinh thái của Hadoop

- Giới thiệu về Apache NiFi
  - Các chức năng chính
    - Tự động hóa luồng dữ liệu giữa các hệ thống

      JSON -> Database, FTP-> Hadoop, Kafka -> ElasticSearch, etc…
    - Giao diện sử dụng kéo thả
    - Tập trung vào cấu hình của các khối xử lý (Processor)
    - Dễ dàng mở rộng số máy của một cụm
    - Đảm bảo không có mất mát dữ liệu



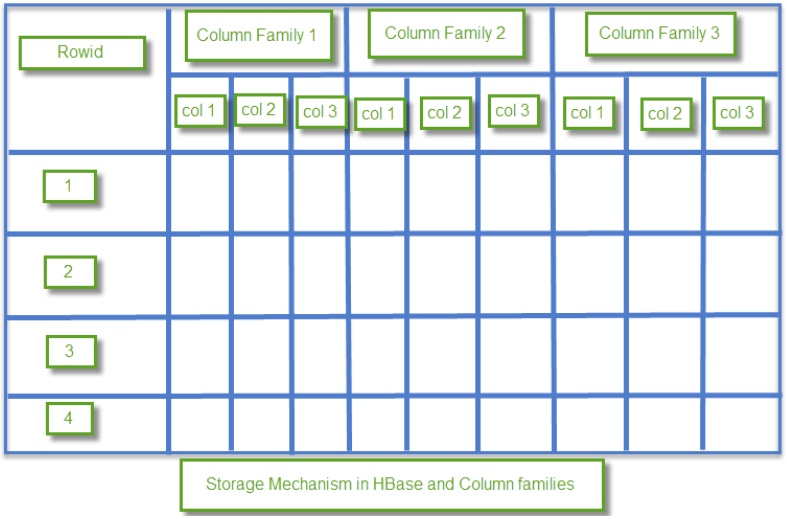
Hệ sinh thái của Hadoop

- Giới thiệu về Apache NiFi
  - Nên sử dụng NiFi
    - Chuyển dữ liệu bảo mật và tin cậy giữa các hệ thống
    - Chuyển dữ liệu từ nguồn tới các nền tảng phân tích
    - Tiền xử lý dữ liệu
    - Thay đổi định dạng dữ liệu
    - Trích xuất dữ liệu
    - Điều hướng

  - Không nên sử dụng NiFi
    - Tính toán phân tán
    - Xử lý các sự kiện phức tạp
    - Thực hiện JOIN, AGGREGATE dữ liệu

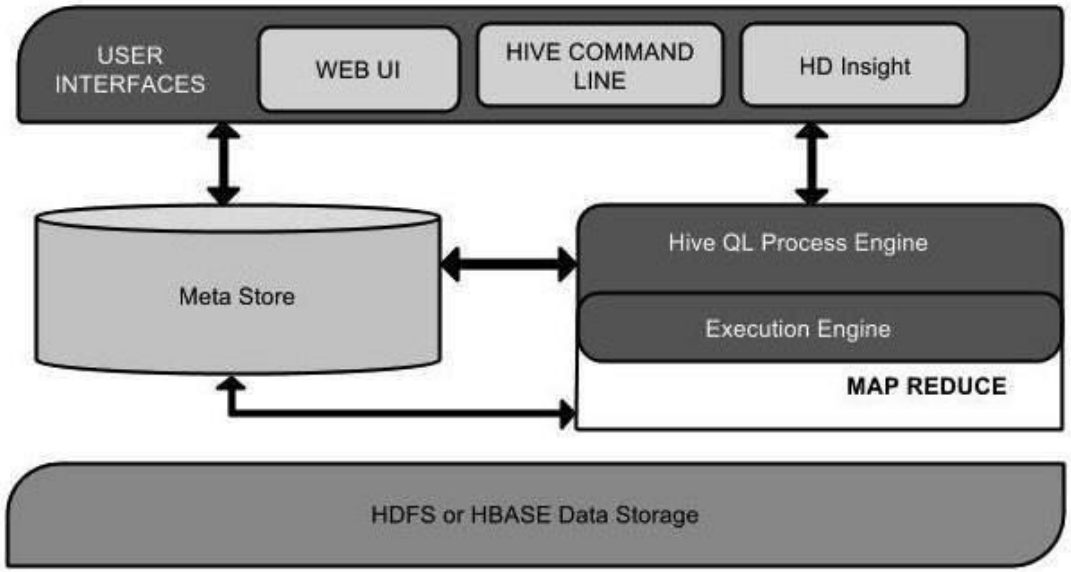
Hệ sinh thái của Hadoop

- Giới thiệu về Hbase
  - Giải pháp lưu trữ dữ liệu lớn, phân tán, hỗ trợ truy cập nhanh và đọc ghi dữ liệu realtime.
  - Khi nào nên chọn:
    - Lưu trữ hàng tỉ bản ghi
    - Tìm kiếm bản ghi trong tập dữ liệu lớn
  - Đặc tính:



Hệ sinh thái của Hadoop

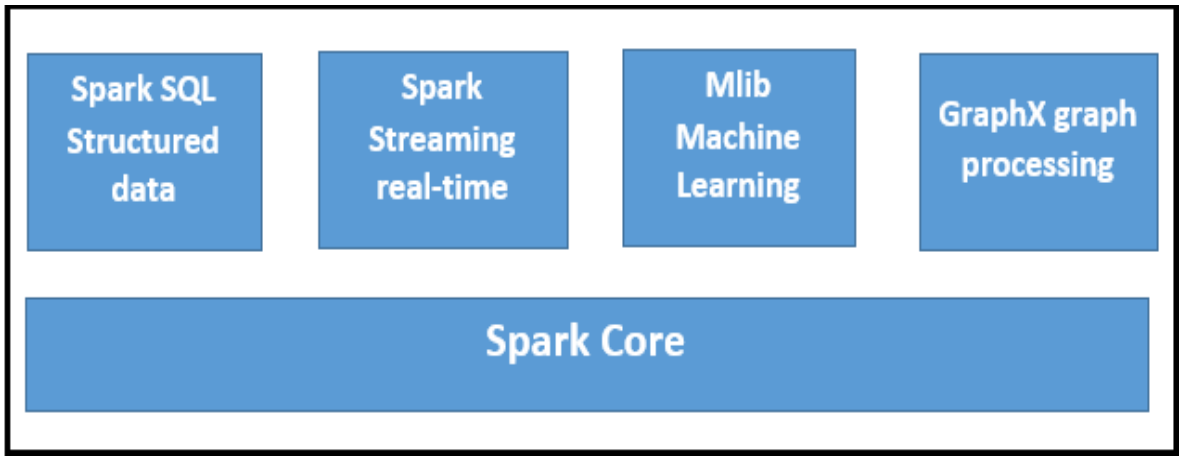
- Giới thiệu về Hive
  - Một công cụ để xử lý dữ liệu có cấu trúc trong Hadoop.
  - Hỗ trợ đọc, ghi và quản lý dữ liệu lớn phân tán sử dụng SQL



Apache Hive

Hệ sinh thái của Hadoop

- Giới thiệu về Spark
  - Một công cụ phân tích hợp nhất để xử lý dữ liệu quy mô lớn.
  - Cung cấp các API hỗ trợ lập trình trong Java, Scala, Python và R
  - Cung cấp thư viện xử lý dữ liệu phong phú: Spark SQL, MLlib cho máy học, GraphX để xử lý đồ thị….

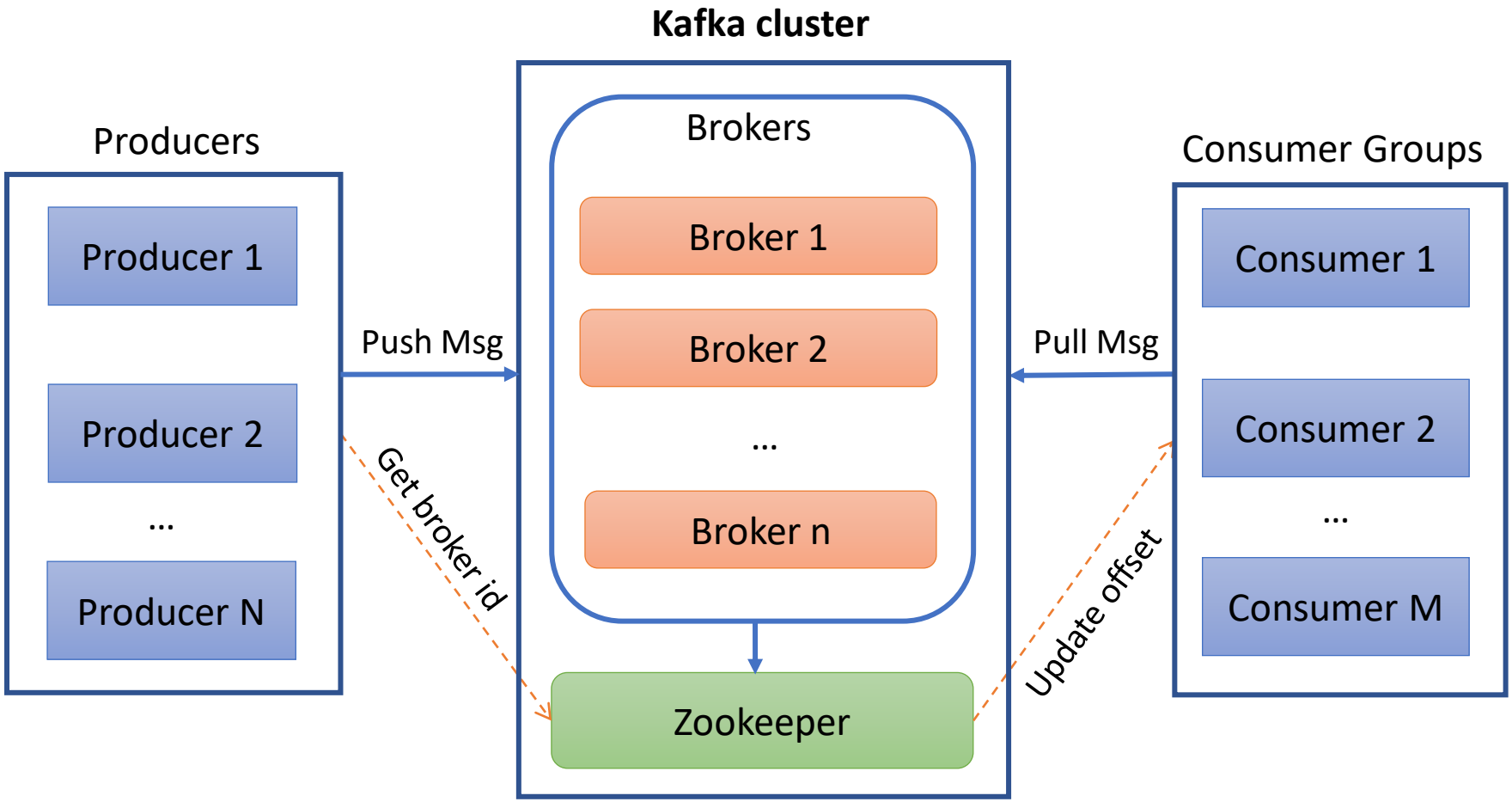


Hệ sinh thái của Hadoop

- Giới thiệu về Kafka
  - Hệ thống message pub/sub phân tán (distributed messaging system).
  - Kafka có khả năng truyền một lượng lớn message theo thời gian thực
  - Hỗ trợ replication (nhân rộng) trong cluster giúp phòng tránh mất dữ liệu.

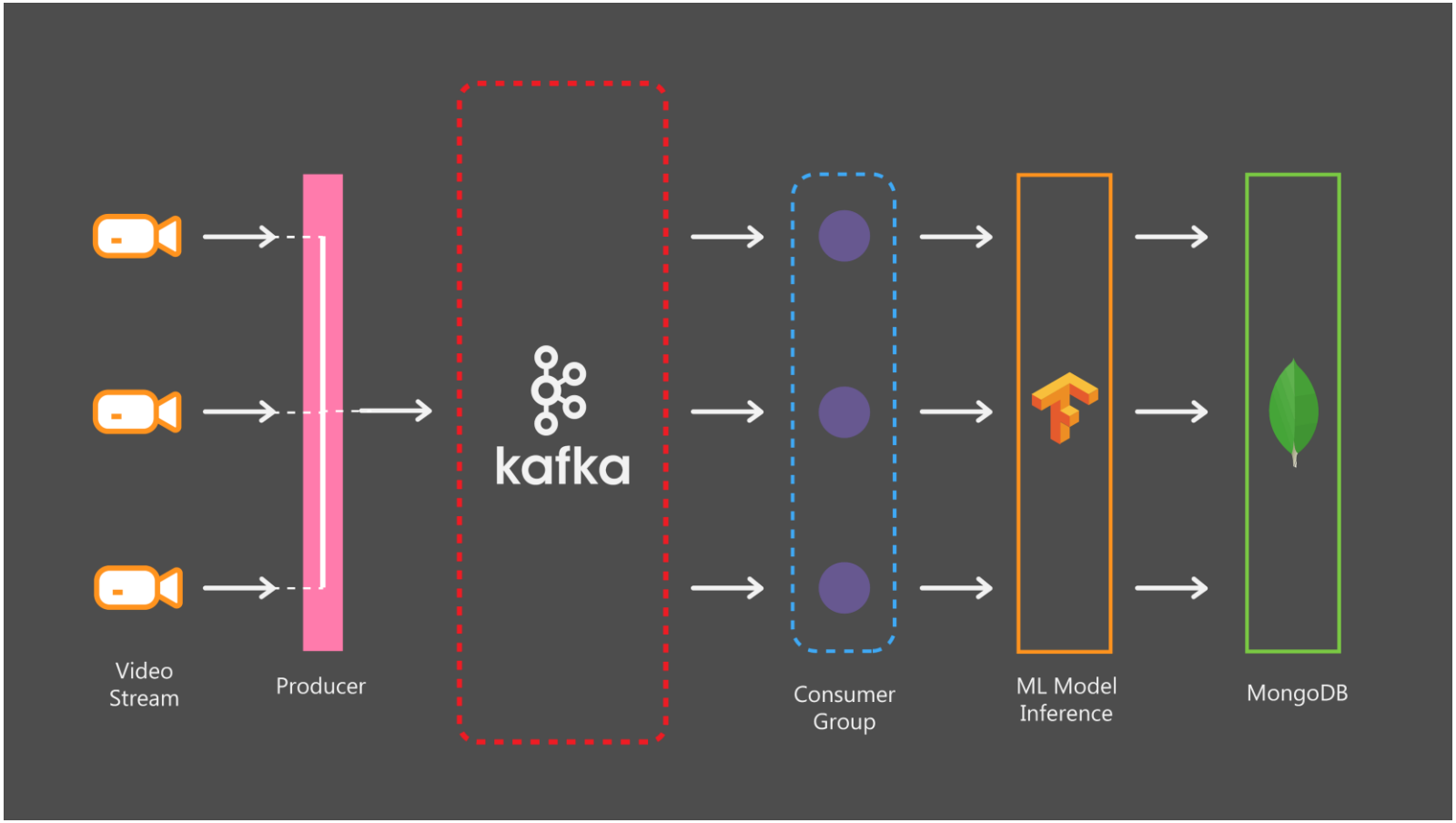
Hệ sinh thái của Hadoop

- Giới thiệu về Kafka



Hệ sinh thái của Hadoop

- Giới thiệu về Kafka



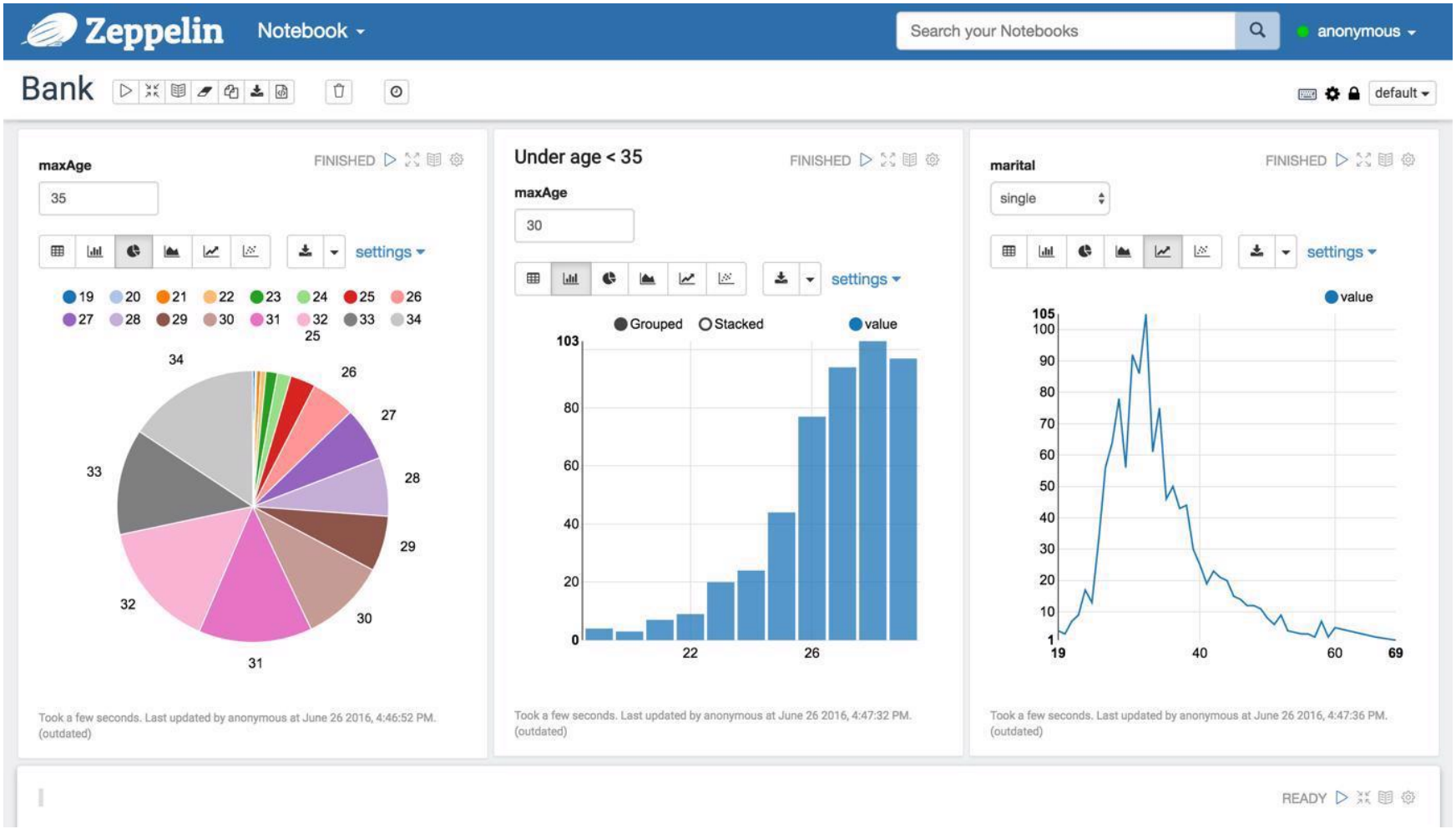
https://github.com/wingedrasengan927

Hệ sinh thái của Hadoop

- Giới thiệu về Apache Zeppelin
  - Apache Zeppelin là một công cụ mã nguồn mở hỗ trợ trực quan hóa dữ liệu lớn trên web
  - Apache Zeppelin được tích hợp với các hệ thống xử lý dữ liệu lớn, phân tán như Apache Spark, Apache Hive, Apache Flink, và nhiều hệ thống khác.
  - Apache Zeppelin cho phép tạo các báo cáo tương tác đẹp, dựa trên dữ liệu với SQL, Scala, R hoặc Python ngay trong trình duyệt.

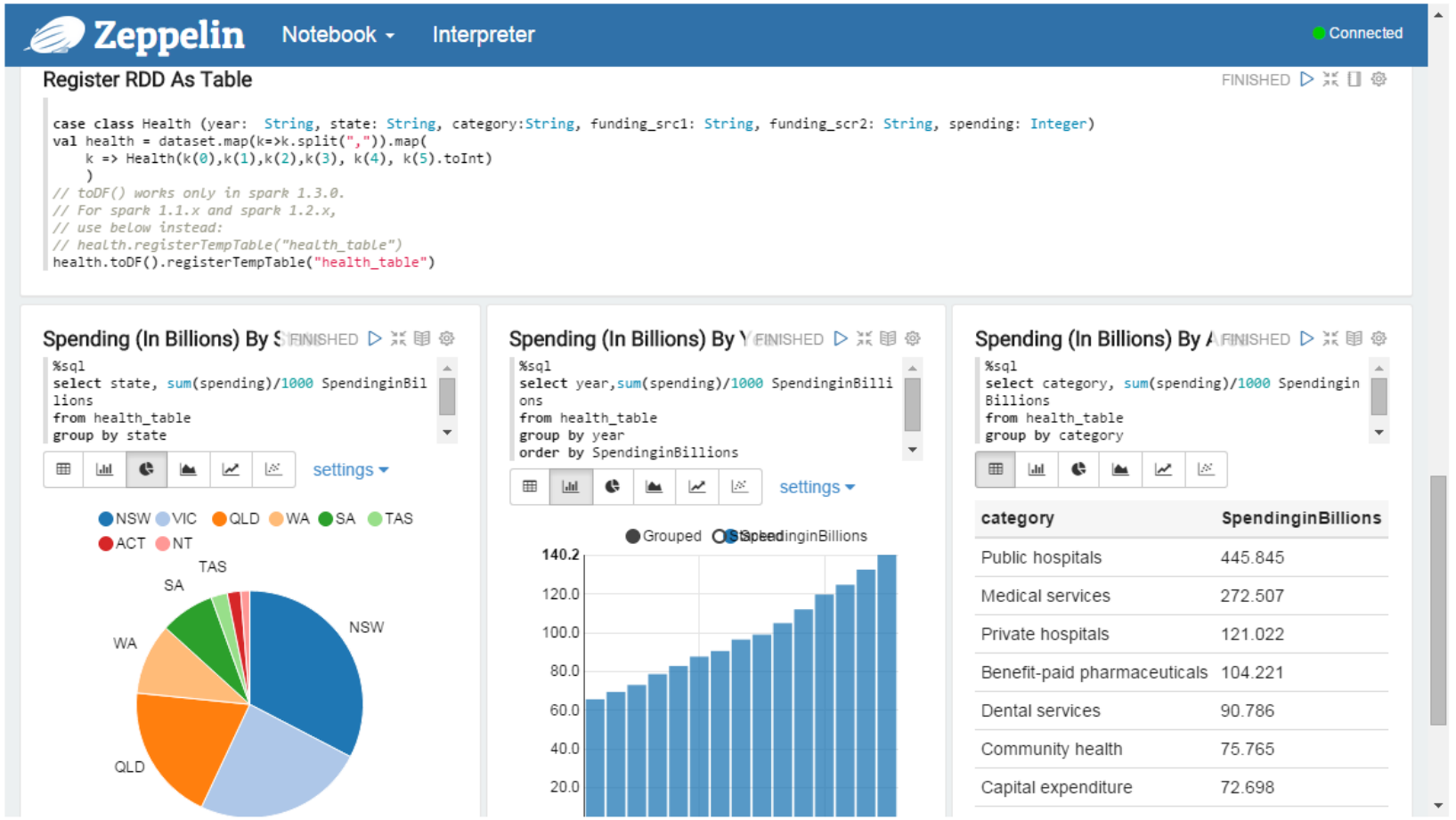
Hệ sinh thái của Hadoop

- Giới thiệu về Apache Zeppelin



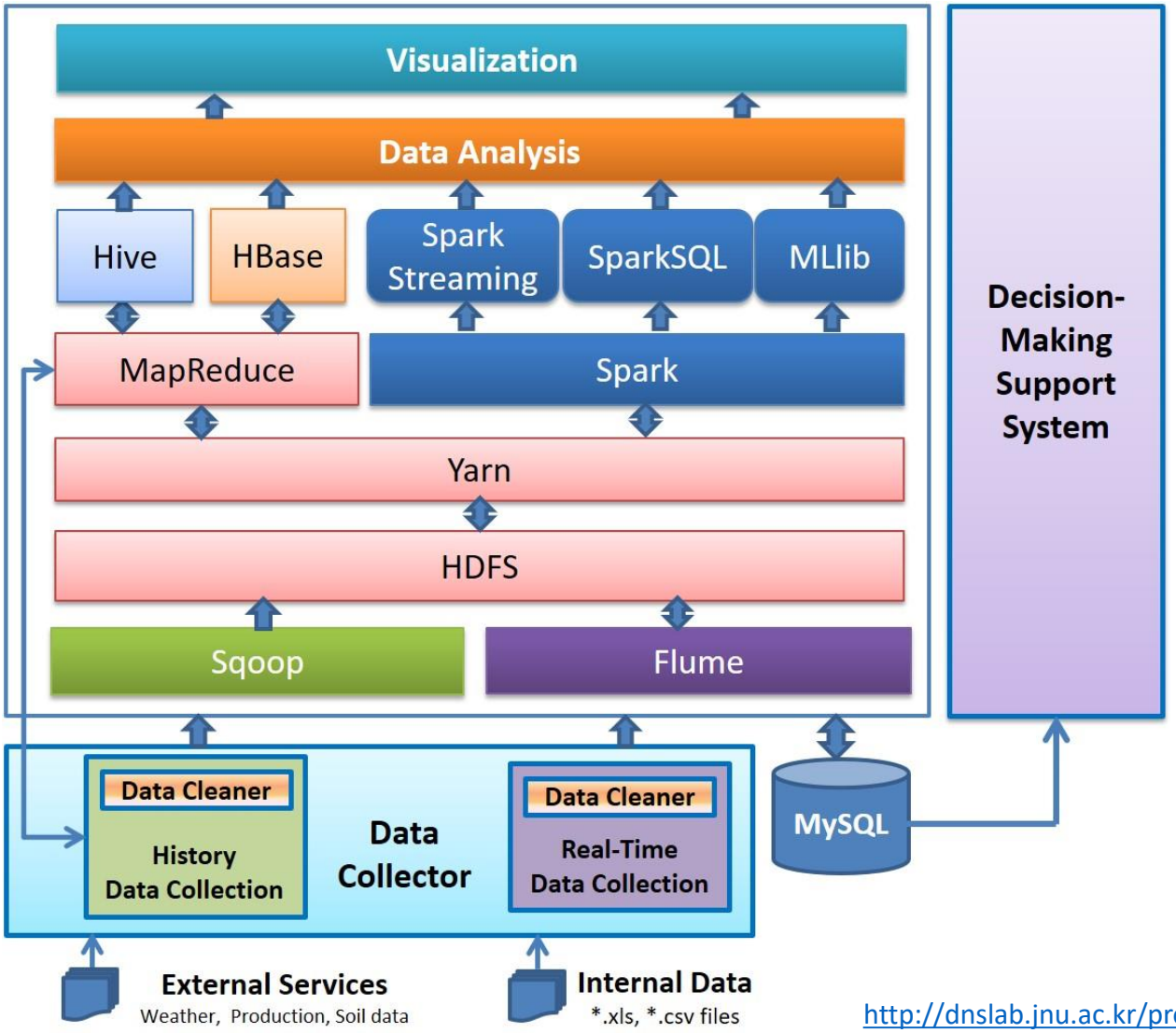
Hệ sinh thái của Hadoop

- Giới thiệu về Apache Zeppelin



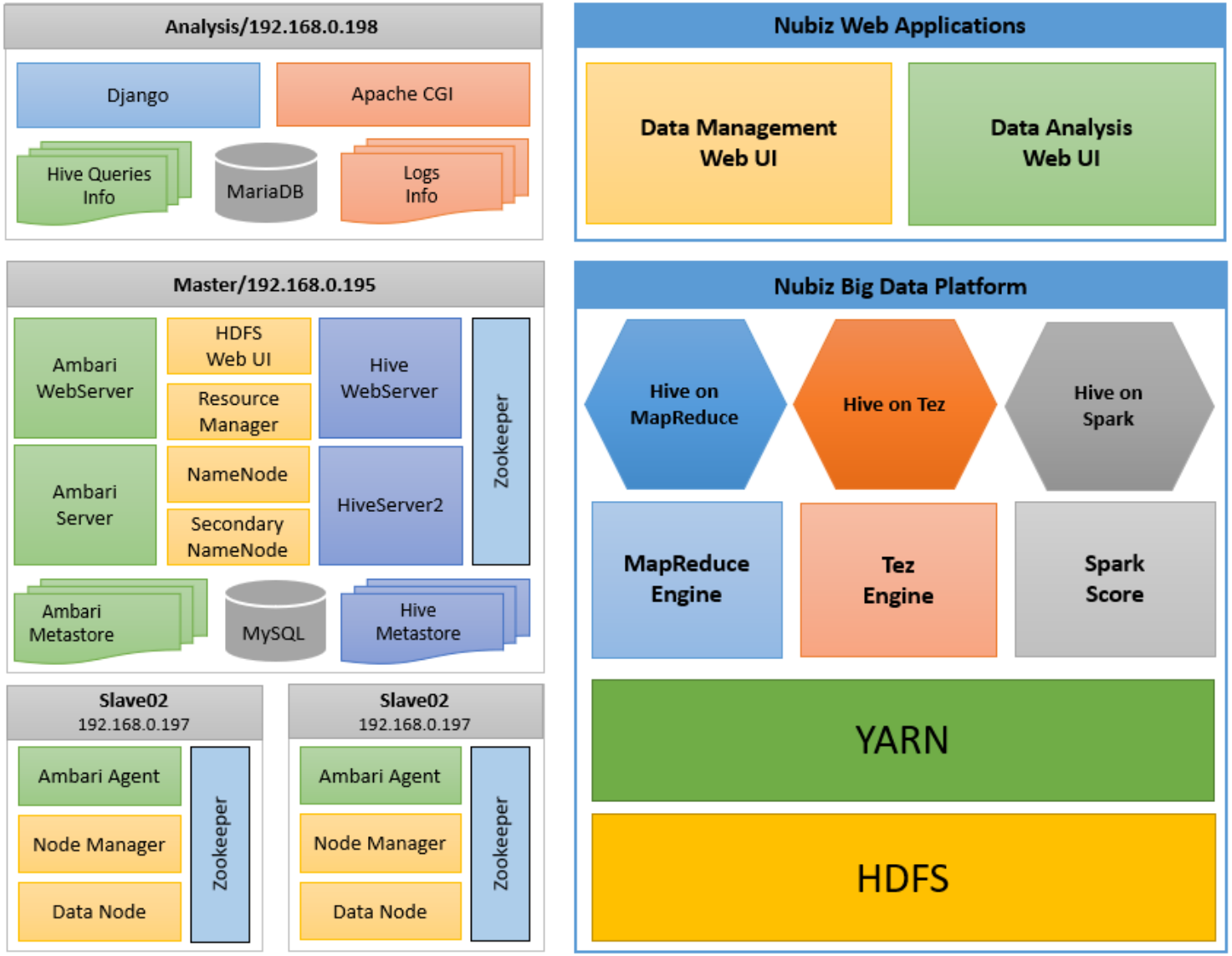
Ví dụ 1: Xây dựng nền tảng xử lý dữ liệu lớn trong nông nghiệp

AgriBigData Platform

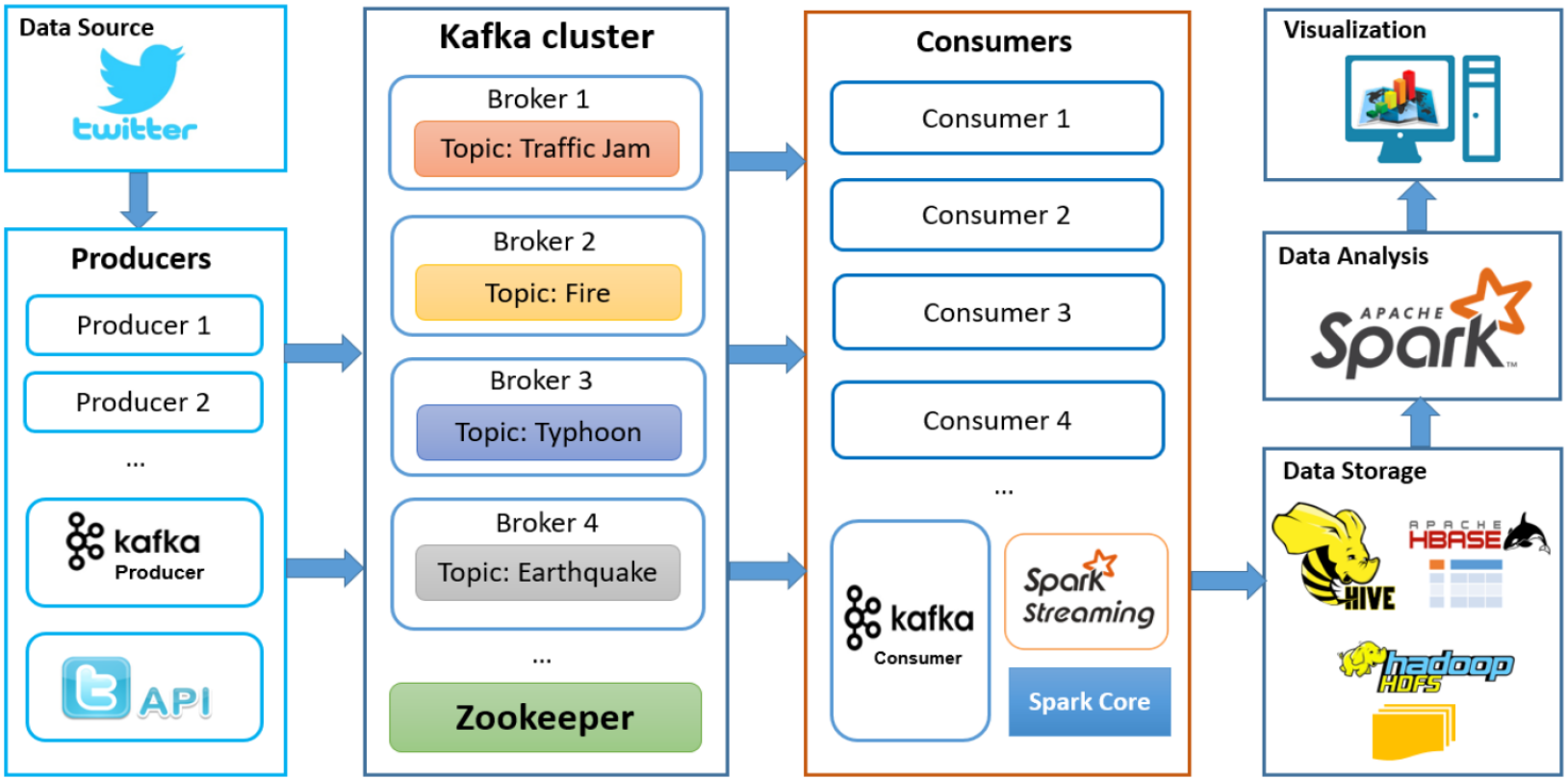


http://dnslab.jnu.ac.kr/projects/agribigdata/

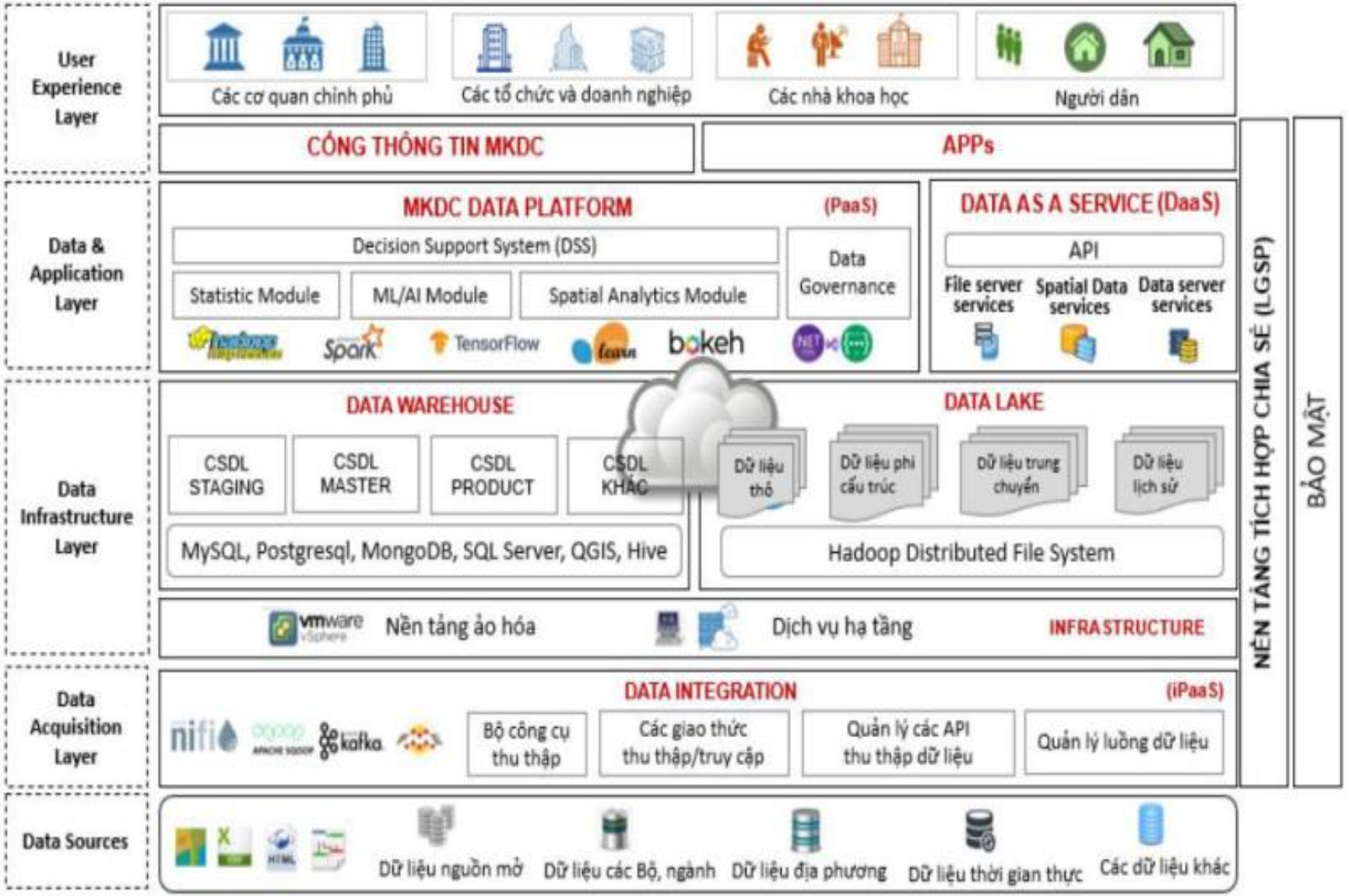
Ví dụ 2: Nubiz Big Data Platform



Ví dụ 3: Disaster Big Data Platform



Ví dụ 4: MKDC Digital Platform



Chuẩn bị môi trường – Phần mềm

- Phần mềm máy ảo: VirtualBox 6.1.16
- Hệ điều hành: Ubuntu 20.04
- Java: Version 1.8
- SSH Server
- Hadoop: Version 3.3.0

Chuẩn bị môi trường – Hệ thống máy tính

STT	Tên máy	Địa chỉ IP	Chức năng
1	master	10.0.2.195	Điều phối
2	slave01	10.0.2.196	Lưu trữ, Tính toán
3	slave02	10.0.2.197	Lưu trữ, Tính toán

Chuẩn bị môi trường – Tài khoản sử dụng hệ thống

- Cách tạo tài khoản:

  $ sudo addgroup **hadoop**

  $ sudo adduser --ingroup **hadoop hduser**

- Thông tin tài khoản:
  - Username: **hduser**
  - Password: **hduser@123**

Bước 1: Cài đặt SSH

- Cài đặt OpenSSH trên tất cả các máy

  $ sudo apt-get install **openssh-server**

- Tạo key trên máy Master

  $ ssh-keygen -t rsa -P ""

- Chép key cho phép truy cập các máy qua SSH
  - Copy trên Master

    $ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
  - Copy vào các Nodes

    $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hduser@slave01

Bước 2: Cài đặt môi trường chạy Java

- $ sudo apt install **openjdk-8-jre-headless**
  - Đường dẫn đầy đủ: **/usr/lib/jvm/java-8-openjdk-amd64**
- $ java **–version**

openjdk version "1.8.0_275"
OpenJDK Runtime Environment (build 1.8.0_275-8u275-b01-0ubuntu1~20.04-b01)
OpenJDK 64-Bit Server VM (build 25.275-b01, mixed mode)

Lưu ý: cài đặt trên tất cả các máy

Bước 3: Cài đặt Hadoop

Download Hadoop

```
$ sudo chown hduser:hadoop /usr/local                    # cấp quyền truy cập
$ cd /usr/local
$ sudo wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz
```

Cài đặt Hadoop

```
$ sudo tar xvf hadoop-3.3.0.tar.gz                       # giải nén
$ sudo mv hadoop-3.3.0 hadoop                            # đổi tên thư mục thành hadoop
$ sudo chown -R hduser:hadoop /usr/local/hadoop          # cấp quyền truy cập
```

Bước 4: Thiết lập biến môi trường

- Thiết lập biến môi trường trong tệp **.bashrc**

$ sudo gedit /home/hduser/.bashrc # mở tệp để sửa

  - Bổ sung các biến môi trường sau vào tệp

```
export HADOOP_HOME="/usr/local/hadoop"
export JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64"
export PATH=$PATH:$HADOOP_HOME/sbin
export PATH=$PATH:$HADOOP_HOME/bin
```

  - Áp dụng

$ source /home/hduser/.bashrc

Bước 5: Cấu hình Hadoop

- Cấu hình biến môi trường trong tệp **hadoop-env.sh**

$ cd /usr/local/hadoop/etc/hadoop # chuyển đến thư mục cấu hình

- Mở tệp **hadoop-env.sh**

$ sudo gedit hadoop-env.sh

- Cấu hình JAVA_HOME: thêm vào tệp

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

Bước 5: Cấu hình Hadoop

- Cấu hình các thuộc tính trong tệp **core-site.xml**

$ sudo gedit **core-site.xml**

- Thêm vào tệp trong cặp thẻ: <configuration></configuration>

```
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://master:9000</value>
</property>
```

Bước 5: Cấu hình Hadoop

- Cấu hình các thuộc tính trong tệp **hdfs-site.xml**

$ sudo gedit **hdfs-site.xml**

- Thêm vào tệp trong cặp thẻ: <configuration></configuration>

```
<property>
        <name>dfs.replication</name>
        <value>2</value>
</property>
<property>
        <name>dfs.namenode.name.dir</name>
        <value>/app/hadoop/hdfs/namenode</value>
</property>
<property>
        <name>dfs.datanode.data.dir</name>
        <value>/app/hadoop/hdfs/datanode</value>
</property>
<property>
        <name>dfs.permissions.enabled</name>
        <value>false</value>
</property>
```

Bước 5: Cấu hình Hadoop

- Tạo thư mục lưu trữ dữ liệu trên máy
  - Trên máy master

```
$ sudo mkdir -p /app/hadoop/hdfs/namenode
$ sudo chown -R hduser:hadoop /app/hadoop/hdfs/namenode
```

  - Trên các máy slave

```
$ sudo mkdir -p /app/hadoop/hdfs/datanode
$ sudo chown -R hduser:hadoop /app/hadoop/hdfs/datanode
```

Bước 5: Cấu hình Hadoop

- Cấu hình các thuộc tính trong tệp **mapred-site.xml**

$ sudo gedit **mapred-site.xml**

- Thêm vào tệp trong cặp thẻ: <configuration></configuration>

```
<property>
        <name>mapreduce.job.tracker</name>
        <value>master:5431</value>
</property>
<property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
</property>
<property>
        <name>yarn.app.mapreduce.am.env</name>
        <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
</property>
<property>
        <name>mapreduce.map.env</name>
        <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
</property>
<property>
        <name>mapreduce.reduce.env</name>
        <value>HADOOP_MAPRED_HOME=/usr/local/hadoop</value>
</property>
```

Bước 5: Cấu hình Hadoop

- Cấu hình các thuộc tính trong tệp **yarn-site.xml**

$ sudo gedit **yarn-site.xml**

- Thêm vào tệp trong cặp thẻ: <configuration></configuration>

```
<property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
</property>
<property>
        <name>yarn.resourcemanager.webapp.address</name>
        <value>master:8088</value>
</property>
<property>
<name>yarn.resourcemanager.scheduler.address</name>
        <value>master:8030</value>
</property>
<property>
        <name>yarn.resourcemanager.resource-tracker.address</name>
        <value>master:8031</value>
</property>
<property>
        <name>yarn.resourcemanager.address</name>
<value>master:8032</value>
</property>
```

Bước 5: Cấu hình Hadoop

- Cấu hình các thuộc tính trong tệp **yarn-site.xml**
  - Thêm vào tệp trong cặp thẻ: <configuration></configuration> (tiếp)

```
<property>
        <name>yarn.resourcemanager.admin.address</name>
        <value>master:8033</value>
</property>
<property>
        <name>yarn.scheduler.capacity.root.support.user-limit-factor</name>
        <value>2</value>
</property>
<property>
        <name>yarn.nodemanager.disk-health-checker.min-healthy-disks</name>
        <value>0.0</value>
</property>
<property>
        <name>yarn.nodemanager.disk-health-checker.max-disk-utilization-per-
disk-percentage</name>
        <value>100.0</value>
</property>
```

Bước 5: Cấu hình Hadoop

- Cấu hình thông tin master
  - Cấu hình tệp master

$ sudo gedit master

  - Thêm tên máy vào tệp:

master

- Cấu hình thông tin slaves
  - Cấu hình tệp workers

$ sudo gedit workers

  - Thêm tên các máy slave vào tệp:

slave01

slave02

Bước 6: Format & Run/Stop Hadoop

Thực hiện trên máy Master

- Format hệ thống

  $ hdfs namenode -format

- Run hệ thống Hadoop

  $ sbin/start-all.sh

- Run Job History

  $ sbin/mr-jobhistory-daemon.sh start historyserver

- Stop hệ thống Hadoop

  $ sbin/stop-all.sh

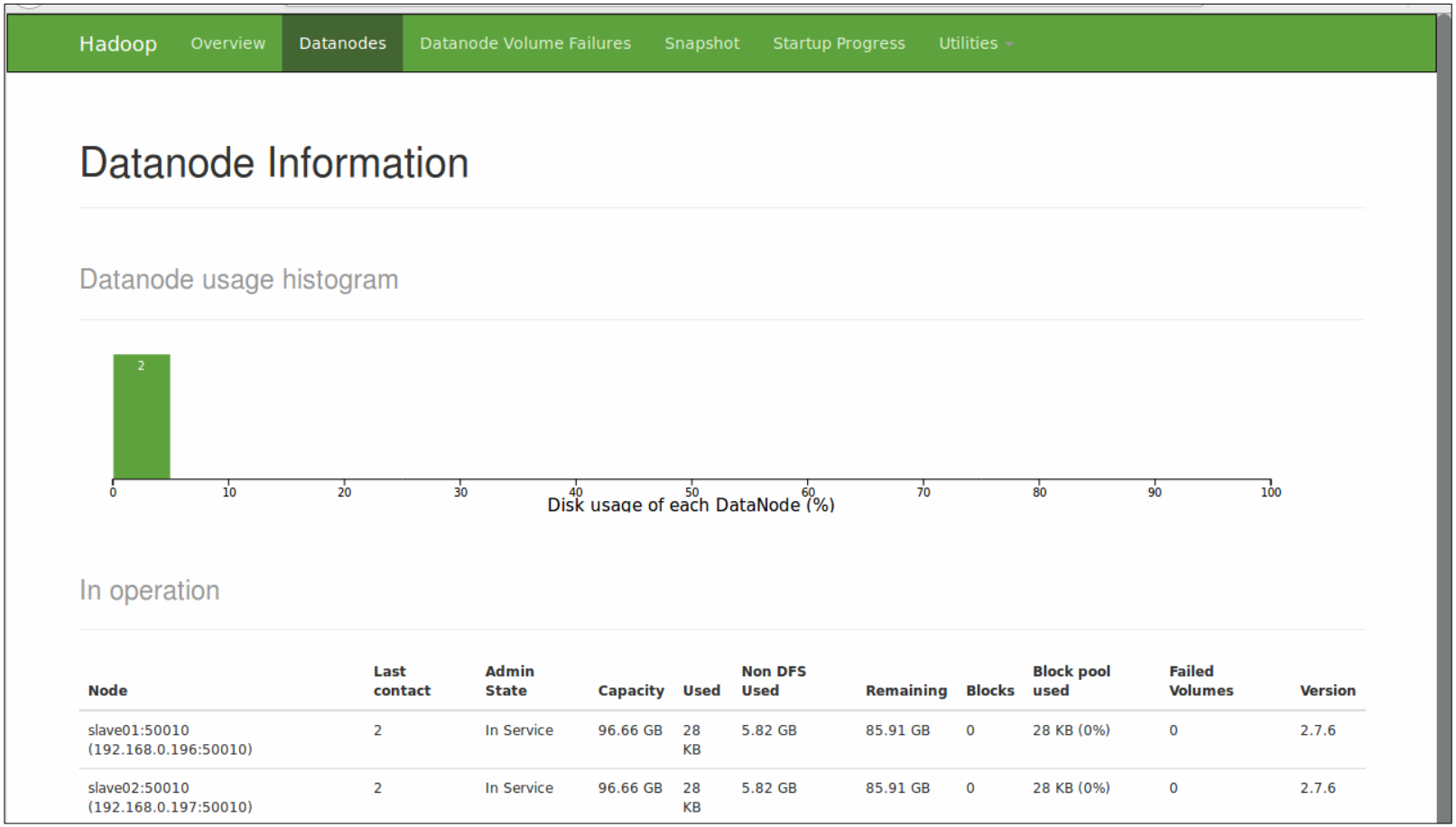
Bước 7: Kiểm tra các giao diện sử dụng hệ thống Hadoop

Truy cập vào các địa chỉ sau trên máy Master

- http://master:9870/ – Giao diện quản lý của NameNode
- http://master:8088/ – Giao diện quản lý của YARN ResourceManager
- http://master:19888/ –Giao diện quản lý của MapReduce JobHistory Server

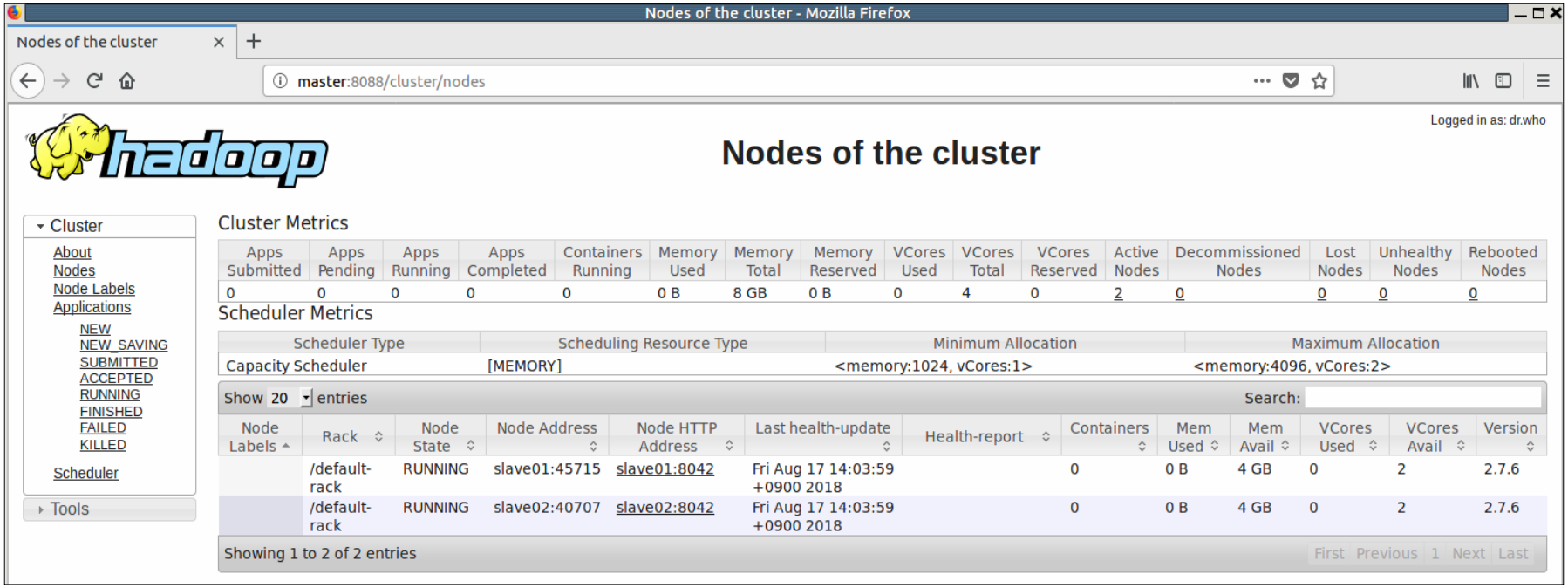
Bước 7: Kiểm tra các giao diện sử dụng hệ thống Hadoop

- http://master:9870/ – Giao diện quản lý của NameNode



Bước 7: Kiểm tra các giao diện sử dụng hệ thống Hadoop

- http://master:8088/ – Giao diện quản lý của YARN ResourceManager



Bước 7: Kiểm tra các giao diện sử dụng hệ thống Hadoop

- http://master:19888/ –Giao diện quản lý của MapReduce JobHistory Server

Trân trọng cảm ơn!

Q&A