hướng dẫn cài đặt spark cluster trên 1 cụm 3 vps , mỗi vps có cấu hình 2 cpu và 2gb ram , như sau : 172.30.2.98 : làm master node và worker 1 172.30.2.58 : làm slave node và là worker 2 172.30.2.90 : làm slave node và là worker 3

**ChatGPT said:**

ChatGPT

Cài đặt Apache Spark trên một cụm 3 VPS như bạn mô tả bao gồm các bước như sau:

**1. Chuẩn bị môi trường**

**Cài đặt Java**

Apache Spark yêu cầu Java, vì vậy hãy cài đặt OpenJDK trên cả 3 VPS.

bash

Copy code

sudo apt update

sudo apt install -y openjdk-11-jdk

java -version

**Cài đặt Scala**

Scala là cần thiết để Spark hoạt động.

bash

Copy code

sudo apt install -y scala

scala -version

**Cài đặt Python (tùy chọn, nếu dùng PySpark)**

Cài Python 3 và pip.

bash

Copy code

sudo apt install -y python3 python3-pip

**2. Tải và cài đặt Apache Spark**

**Tải Spark trên cả 3 máy**

Tải Apache Spark từ trang chủ hoặc sử dụng lệnh dưới đây để tải bản mới nhất. Thay URL phù hợp với phiên bản bạn chọn.

bash

Copy code

wget https://downloads.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

**Giải nén và thiết lập Spark**

bash

Copy code

tar xvf spark-3.5.0-bin-hadoop3.tgz

sudo mv spark-3.5.0-bin-hadoop3 /opt/spark

**Thiết lập biến môi trường**

Thêm các dòng sau vào file ~/.bashrc trên cả 3 máy:

bash

Copy code

export SPARK\_HOME=/opt/spark

export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

Áp dụng thay đổi:

bash

Copy code

source ~/.bashrc

**3. Cấu hình Spark**

**Cập nhật file spark-env.sh**

Tạo và chỉnh sửa file spark-env.sh trên **cả 3 máy**:

bash

Copy code

cp $SPARK\_HOME/conf/spark-env.sh.template $SPARK\_HOME/conf/spark-env.sh

nano $SPARK\_HOME/conf/spark-env.sh

Thêm nội dung sau (chỉnh sửa theo IP cụ thể của từng máy):

* Trên **master node (172.30.2.98):**

bash

Copy code

SPARK\_MASTER\_HOST=172.30.2.98

SPARK\_WORKER\_MEMORY=1G

SPARK\_WORKER\_CORES=1

SPARK\_WORKER\_INSTANCES=1

* Trên **worker node (172.30.2.58 và 172.30.2.90):**

bash

Copy code

SPARK\_WORKER\_MEMORY=1G

SPARK\_WORKER\_CORES=1

SPARK\_WORKER\_INSTANCES=1

**Cập nhật file slaves trên master node**

Tạo hoặc chỉnh sửa file $SPARK\_HOME/conf/slaves trên **master node (172.30.2.98)**:

bash

Copy code

nano $SPARK\_HOME/conf/slaves

Thêm các dòng sau:

Copy code

172.30.2.98

172.30.2.58

172.30.2.90

**4. Cấu hình SSH không mật khẩu**

Thiết lập SSH không mật khẩu giữa các node để Spark có thể quản lý cluster:

**Tạo SSH key trên master node**

bash

Copy code

ssh-keygen -t rsa -P "" -f ~/.ssh/id\_rsa

**Copy key tới các node (bao gồm chính nó)**

bash

Copy code

ssh-copy-id user@172.30.2.98

ssh-copy-id user@172.30.2.58

ssh-copy-id user@172.30.2.90

**5. Khởi động Spark Cluster**

**Khởi động Master Node**

Trên máy **master node (172.30.2.98)**:

bash

Copy code

start-master.sh

Master sẽ khởi động và cung cấp giao diện web tại:  
http://172.30.2.98:8080

**Khởi động Worker Nodes**

Trên **tất cả các node** (bao gồm master node vì nó cũng làm worker):

bash

Copy code

start-worker.sh spark://172.30.2.98:7077

**6. Kiểm tra Spark Cluster**

Truy cập giao diện web Spark Master:  
http://172.30.2.98:8080

Bạn sẽ thấy cả 3 worker nodes được liệt kê tại đây.

**7. Chạy thử ứng dụng**

Bạn có thể chạy một job thử nghiệm như sau:

bash

Copy code

spark-submit --master spark://172.30.2.98:7077 --class org.apache.spark.examples.SparkPi $SPARK\_HOME/examples/jars/spark-examples\_2.12-3.5.0.jar 100

**Lưu ý**

* Tùy theo ứng dụng thực tế, bạn có thể cần điều chỉnh cấu hình trong file spark-defaults.conf và log4j.properties.
* Đảm bảo mở cổng 7077 (Spark Master) và 8080 (Web UI) trên firewall nếu có.

Chúc bạn thành công! 🎉