TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN TP.HCM
Khoa Hệ Thống Thông Tin
Chương trình đào tạo: Kỹ sư Hệ thống Thông tin

Bài Thực hành số 3

Tên môn học: ***Dữ liệu lớn***

I. Tóm tắt bài thực hành

1. Yêu cầu lý thuyết

Sinh viên đã được trang bị kiến thức:

- Cấu trúc hệ thống phân tán và framework lập trình Apache Spark
- Đối tượng RDD (Resilient Distributed Dataset) trong Apache Spark
- …

2. Nội dung

❖ **Ôn tập lại những kiến thức cần thiết**

- Thành thạo ngôn ngữ lập trình Python.
- Xem lại cấu trúc hệ thống phân tán và framework lập trình Apache Spark từ những kiến thức được học trên lớp lý thuyết và thực hành.
- Xem lại kiến thức về đối tượng RDD (Resilient Distributed Dataset) trong Apache Spark đã được học trên lớp lý thuyết.

❖ **Lập trình Python với Apache Spark thông qua PySpark**

- Tham khảo tại đường dẫn: http://spark.apache.org/docs/latest/rdd-programming-guide.html

❖ **Sử dụng lệnh spark-submit để triển khai ứng dụng**

- Tham khảo tại đường dẫn: http://spark.apache.org/docs/latest/submitting-applications.html

3. Kết quả cần đạt

✓ Sinh viên cần nắm rõ phương pháp lập trình bằng ngôn ngữ Python trên framework Apache Spark thông qua việc thao tác với RDD sử dụng PySpark.

✓ Sinh viên biết được cách sử dụng lệnh spark-submit để triển khai ứng dụng dữ liệu lớn. Hiểu và có thể hiệu chỉnh được các tham số khi triển khai.

II. Yêu cầu bài làm sinh viên

Nội dung thực hành buổi 03 được thực hiện theo từng cá nhân. Sinh viên upload một tập tin **<MSSV>.zip** hoặc **<MSSV>.rar** nén bên trong là các tập tin sau

- baitap01.py | ipynb
- baitap02.py | ipynb
- …

Lưu ý: Bài nộp không theo đúng quy định này sẽ không được tính.

III. Bài tập

Sau khi cài đặt xong Apache Spark ở buổi học trước, sinh viên cần khởi động hệ thống phân tán và thực hiện theo các hướng dẫn của giảng viên thực hành. Thực hiện viết chương trình sau:

- Bài tập 1: Tính **trung bình cộng**, **trung bình nhân** của một dãy **số thực**.
- Bài tập 2: Tính **phương sai** và **độ lệch chuẩn** của một dãy **số thực**.
- Bài tập 3: Tính **tổng** của một dãy **phân số** đọc từ **tập tin** (không sử dụng thư viện phân số có sẵn).

IV. Phụ lục

1. Cài đặt Apache Spark để sử dụng PySpark trên Jupyter Notebook

Mở tập tin `conf/spark-env.sh` thêm vào các biến môi trường cho Apache Spark như sau:

```
PYSPARK_PYTHON=python3
PYSPARK_DRIVER_PYTHON=jupyter
PYSPARK_DRIVER_PYTHON_OPTS=notebook
```

~ HẾT ~