Data Engineer Training

Update 01/07/2023

Tài liệu cho khóa train cơ bản cho thực tập parttime trong vòng 3 tháng

Thiết kế 3 ngày * 4 tuần * 3 tháng. Review theo tuần tức tương đương 12 đơn vị bài làm

Lộ trình bao gồm:

- 4 Tuần đầu: Java core cơ bản, cơ sở dữ liệu cơ bản, Maven, OOP, Serialize
- 4 Tuần tiếp theo: Cơ sở lí thuyết cho dữ liệu lớn hadoop, hdfs, yarn, mapreduce, spark
- 4 Tuần cuối: Cài cắm cụm hadoop, xây dựng một luồng data cơ bản.

Các yêu cầu chung:

```
(1) sinh viên nhận task theo tuần để tự tự tìm hiểu
(2) thắc mắc liên hệ người hướng dẫn, không quá 30p/ ngày
(3) thực tập sinh liên hệ người hướng dẫn vào buổi cuối của mỗi tuần để review công việc
(không quá 15 phút/ lần)
(4) xong sớm có thể báo review sớm và chuyển sang phần tiếp theo
(5) trước khi review yêu cầu viết file report(các công việc đã làm, kết quả công việc,
code nếu có, các khó khăn cần giải đáp) bằng markdown đẩy lên github cá nhân sau đó gửi
trước cho người hướng dẫn, mỗi tuần là một file markdown.
```

1. Tuần 1

1.1. Yêu cầu (1): oop

- Mô tả: xây dựng chương trình java bất kì có xử dụng đầy đủ 4 tính chất của oop
- Tham khảo: http://www.w3resource.com/java-tutorial/java-object-oriented-programming.php
- Điều kiên hoàn thành: Từ chương trình đã xây dựng trình bày về đã áp dụng oop như nào, (2) hiểu các khái niệm interface,static,....

1.2. Yêu cầu (2) Đọc ghi file

- viết trương trình java đọc ghi file theo 2 dạng binary và text
- viết trương trình java thao tác với file và thư mục: list các file, đọc nội dung file

2. Tuần 2

2.1. Collection(s)

- Mô tả: viết trương trình java sử dụng các cấu trúc dữ liệu HashMap, HashSet, ArrayList
- Tham khảo (Overview) http://cs.lmu.edu/~ray/notes/collections/ (compare) http://www.codejava.net/java-core/collections/java-collections-framework-summary-table (performance) http://infotechgems.blogspot.com/2011/11/java-collections-performance-time.html
- Điều kiên hoàn thành: Cần nắm được HashMap,HashSet,ArrayList là gì, cách phương thức sử dụng ra sao, so sánh các đặc điểm

2.2. Thuật toán

- tạm thời bỏ qua.

2.3. Serialize

- tìm hiểu liên quan serialize trong java. (viết code ví dụ minh họa bằng java và giải thích code) (phần này có chút liên quan tới trên cơ sở kiến thức từ 1.2)

3. Tuần 3

3.1. Exception

- nắm được cách xử lí exception, hệ thống các exception trong java, lấy ví dụ về một exception bất kì và cách fix nó.
- tham khảo: http://www.journaldev.com/1696/java-exception-handling-tutorial-with-examples-and-best-practices

3.2 Concurrency (optional):

- Yêu cầu: (1) viết 1 luồng chạy ngầm kế thừa Runnable sử dụng java, (2) viết chương trình sử dụng threadpool bằng ngôn ngữ java
- tham khảo: http://winterbe.com/posts/2015/04/07/java8-concurrency-tutorial-thread-executor-examples/ (cái này nó viết bằng syntax của java 8)
- - tìm hiểu lock, atomic integer và concurrent hashmap
- - Seminar: thực tập sinh trong phần này nếu có nhu cầu có thể làm silde thuyết trình về những tìm hiểu của mình với team platform.

3.3. json

- Yêu cầu: Dựa trên Serialize ở java tại mục 2.2, viết code có sử dụng json (yêu cầu có sử dụng maven)
- tham khảo: (1)https://www.tutorialspoint.com/json/json_overview.htm, (2) http://www.tutorialspoint.com/json/json_java_example.htm

- nắm được json là gì, sử dụng java parse json, lấy giá trị, chuyển jsonobject thành string
- sử dụng thư viện gson để parse trực tiếp 1 string sang 1 object tương ứng

4. tuần 4

- Yêu cầu (1): Tự cài đặt một cơ sở dữ liệu trên máy tính (cụ thể là mysql). trình bày chi tiết về các thành phần liên quan
- Yêu cầu (2): Tự lấy ví dụ về 5 câu query không tốt và cách tối ưu nó.
- Yêu cầu (3): Tìm hiểu về các loại db và trình bày lại (ít nhất 3 db thuộc loại sql, 3 db thuộc loại no sql). tham khảo: https://www.ml4devs.com/articles/datastore-choices-sql-vs-nosql-database/

5. Tuần 5-6

- Tìm hiểu về bigdata. tham khảo khóa bigdata trên educative (link khóa học , tài khoản sẽ cấp sau)
- Tham khảo doc: https://docs.google.com/document/d/1DNDmlSyefGOA83d0ugFABpK5BmC_FK9G9pp_SIOWm2s/edit
- Các khái niệm cần nắm vững:

```
(1) Dữ liệu lớn và Hadoop ecosystem, khai niệm cơ bản về hdfs, yarn, spark
(2) Quá trình đọc ghi trong hdfs, khái niệm namenode, datanode, secondary namenode, hdfs block, block
replication,
(3) Các thành phần của yarn, khái niệm về mapreduce
(4) Các thành phần của spark, spark api(action, tranformation),
```

- Các khai niệm nâng cao

```
(1) HA trong hdfs(khái niệm JournalNode, zookeeper)
(2) Khái niệm editlog, stand by namenode, fs image
```

6. Tuần 7

- chủ động liên hệ lịch review với người hướng dẫn trong khoảng 1 tiếng để nhận feedback các phần còn thiếu
- tự tỉm hiểu về các phần còn thiếu trong thời gian còn lại
- cuối tuần này (thứ 7) sẽ tổ chức buổi seminar. mỗi người phải chuẩn bị một slide nói về hadoop-spark và gửi trước 20h thứ 6 tuần này. ai làm tốt có thể trình bày trước nhóm.

7. Tuần 8

- giải đáp các mục còn tồn đọng trong tuần 7.
- Yêu cầu tìm hiểu về Shell linux:

```
- Các command cơ bản: cd, ls, cp, mv, mkdir,cat,head
- Các command liên quan đến quyền : chmod, chown, ls -l
- Các command thực hiện song song:
cat test.txt | wc -l
cat test.txt | grep "a"
cat test.txt | head
echo "aabb"  > test.txt
echo "cc" >> test.txt
- Sử dụng vim: tạo file mới, sửa file, save,vvv
- Quản lý tiến trình: htop, ps aux, kill -9
```

- Yêu cầu tìm hiểu về docker. Điều kiện hoàn thành: viết một chương trình sử dụng rest api cơ bản(GET /ping và response về "pong") bằng flask python. sau đó triển khai service trên docker sao cho đứng bên ngoài máy tính có thể call được service chạy trong docker.

8. Tuần 9,10,11

8.1. Setup môi trường cài đặt cụm HDFS, YARN

- yêu cầu chạy được chương trình word count với hadoop mapreduce
- lưu ý về phiên bản hadoop, spark. Thực tập sinh cần tìm hiểu sự khác nhau giữa các phiên bản
- Cài đặt hadoop 2.x theo mô hình sau (chú ý cài cơ bản k cần HA):
- Link port: Để tránh trùng port các sinh viên dùng port theo công thức: port = port_default + so_thu_tu mình trên sheet. (phần port chỉ định này chưa hoàn thiện, sẽ gửi lại sau khi tới thời gian)

8.2. Cài đặt Spark standalone, spark trên yarn

- yêu cầu chạy được chương trình word count với spark chạy trên yarn (tham khảo hình trên)
- tự sinh dữ liệu người dùng , visualize lên bằng pyspark.
- yêu cầu cụ thể:

```
(1) sinh file parquet trên hệ thống hdfs đã cài đặt: khoảng 1 triệu bản ghi gồm các cột:
tên, ngày sinh, địa chỉ (địa chỉ random từ 1-100), giới tính, số điện thoại.
(2) chạy jupyter notebook
(3) sử dụng pyspark để đọc file trên notebook. chạy spark chế độ standalone
(nâng cao: *có thể chạy trên yarn*)
(4) visualize: thống kê lượng user theo tuổi
(khoảng độ tuổi ví dụ từ 10-20, 20-30 tuổi, ...), giới tính bằng các biểu đồ hợp lí
(5) chuyển các code pyspark về code java và submit với spark chạy trên yarn.
ở bước này không cần visualize mà chỉ cần hiên thị số liệu.
```

9. Tuần 12

- Sinh viên thực tập sau khi hoàn thành hết toàn bộ công việc (có thể hoàn thành trước hạn), có thể lựa chọn tìm hiểu thêm công nghệ hoặc đề nghị làm chung một số task thực tế với team.
- Ví dụ về tìm hiểu thêm công nghệ như kafka, zookeeper, hbase, nifi, airflow