

Lưu trữ và xử lý dữ liệu lớn

Hệ thống phân tích nhu cầu tuyển dụng

Giảng viên hướng dẫn: PGS.TS.Nguyễn Bình Minh

Nhóm 11:

Nguyễn Tiến Long - 20180129

Phan Việt Hoàng - 20180086

Phạm Trần Anh - 20180018

Võ Hồng Sang - 20183973

Trường ĐH CNTT&TT-ĐBBKHN

Ngày 4 Tháng 1 Năm 2022

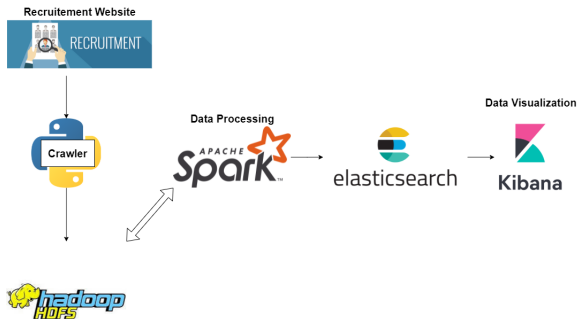
Mục lục

- 1 Trình bày vấn đề
- 2 Tổng quan hệ thống
- 3 Luồng dữ liệu hệ thống
- 4 Demo
- 5 Hạn chế và hướng phát triển tiếp theo

Phân tích nhu cầu tuyển dụng

- Việc nắm bắt được nhu cầu thị trường lao động giúp những người đang tìm việc có được sự chuẩn bị tốt nhất khi đi ứng tuyển xin việc tại các công ty
- Để biết được thị trường lao động đang cần gì, một giải pháp đơn giản mà hiệu quả là thực hiện đánh giá, thống kê những kỹ năng, kiến thức được miêu tả trong các đơn tuyển dụng của các công ty trên các trang mạng tìm việc làm
- Nhóm em đề xuất một hệ thống có khả năng lưu trữ và xử lý dữ liệu lớn là RecruitmentInsight. Hệ thống được triển khai công cụ Docker Compose. Nguồn dữ liệu nhóm lựa chọn để nghiên cứu được thu thập từ trang web TopCV.

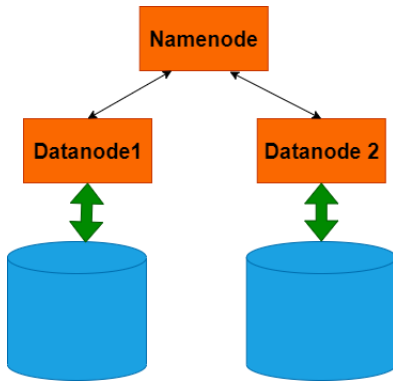
Tổng quan hệ thống



Hình 1: Tổng quan hệ thống

Tổng quan hệ thống

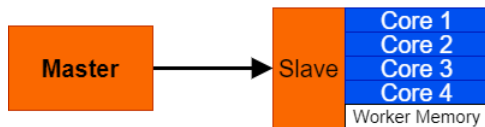
- Hadoop Cluster



Hình 2: Hadoop Cluster

Tổng quan hệ thống

- Spark Cluster



Hình 3: Spark Cluster

- Elasticsearch Cluster



Hình 4: Elasticsearch Cluster

Docker Compose

- Hadoop Cluster

```
namenode:
  image: bde2828/hadoop-namenode:2.0.0-hadoop3.2.1-java8
  container_name: namenode
  restart: always
  ports:
    - 9870:9870
    - 9000:9000
  volumes:
    - hadoop_namenode:/hadoop/dfs/name
  environment:
    - CLUSTER_NAME=test
  env_file:
    - ../hadoop.env

datanode1:
  image: bde2828/hadoop-datanode:2.0.0-hadoop3.2.1-java8
  container_name: datanode1
  restart: always
  volumes:
    - hadoop_datanode1:/hadoop/dfs/data
  environment:
    SERVICE_PRECONDITION: "namenode:9870"
  env_file:
    - ../hadoop.env

datanode2:
  image: bde2828/hadoop-datanode:2.0.0-hadoop3.2.1-java8
  container_name: datanode2
  restart: always
  volumes:
    - hadoop_datanode2:/hadoop/dfs/data
  environment:
    SERVICE_PRECONDITION: "namenode:9870"
  env_file:
    - ../hadoop.env
```

Hình 5: Docker Hadoop Cluster

Docker Compose

- Spark Cluster

```
spark-master:
  image: bde2020/spark-master:3.0.0-hadoop3.2
  container_name: spark-master
  depends_on:
    - namenode
    - datanode
  ports:
    - "8080:8080"
    - "7077:7077"
  environment:
    - INIT_DAEMON_STEP=setup_spark
    - CORE_CONF_fs_defaultFS=hdfs://namenode:9000

spark-worker:
  image: bde2020/spark-worker:3.0.0-hadoop3.2
  container_name: spark-worker
  depends_on:
    - spark-master
  ports:
    - "8081:8081"
  environment:
    - "SPARK_MASTER=spark://spark-master:7077"
    - CORE_CONF_fs_defaultFS=hdfs://namenode:9000
```

Hình 6: Docker Spark Cluster

Docker Compose

- Elasticsearch Cluster

```
elasticsearch:
  image: docker.elastic.co/elasticsearch/elasticsearch:7.15.1
  container_name: elasticsearch
  volumes:
    - esdata:/usr/share/elasticsearch/data
  ports:
    - 9200:9200
    - 9300:9300
  environment:
    - xpack.security.enabled=false
    - "discovery.type=single-node"

kibana:
  image: docker.elastic.co/kibana/kibana:7.15.1
  container_name: kibana
  ports:
    - 5601:5601
  environment:
    ELASTICSEARCH_URL: http://elasticsearch:9200
    ELASTICSEARCH_HOSTS: '["http://elasticsearch:9200"]'
```

Hình 7: Docker Elasticsearch Cluster

Mẫu đơn tuyển dụng

 <https://www.topcv.vn/brand/vienthongmobifone/tuyen-dung/lap-trinh-vien-trien-kh>

Mô tả công việc

Triển khai các dự án hỗ trợ dịch vụ, hệ thống hỗ trợ kinh Doanh, bán hàng của MobiFone

Triển khai dự án trên nền tảng dữ liệu lớn, AI, IOT.

Thực hiện lập trình phát triển mới/nâng cấp/sửa chữa phần mềm theo yêu cầu nghiệp vụ và tài liệu thiết kế từ MobiFone;

Thực hiện sửa lỗi phần mềm trong quá trình phát triển và kiểm thử, phối hợp với đội kiểm thử trong quá trình sửa lỗi;

Thực hiện phối hợp triển khai thử nghiệm và triển khai chính thức phần mềm;

Hỗ trợ hoàn thành tài liệu cài đặt, triển khai và vận hành khai thác

Yêu cầu ứng viên

Tốt nghiệp Đại học chuyên ngành cử nhân/kỹ sư Công nghệ thông tin, Điện tử viễn thông, Toán tin, ... tại các trường Đại học công lập, Đại học dân lập hoặc các trường Đại học nước ngoài có uy tín

Ưu tiên có kinh nghiệm lập trình và triển khai dự án phát triển trên nền tảng Java, PYTHON...

Ưu tiên có sử dụng thành thạo một trong các hệ quản trị cơ sở dữ liệu MySQL, SQL Server, Oracle...

Ưu tiên có kinh nghiệm làm việc với hệ thống Message Queue như Kafka, RabbitMQ, ActiveMQ...

Quyền lợi được hưởng

Môi trường làm việc tốt, năng động, chuyên nghiệp

Nhiều chính sách đãi ngộ: hỗ trợ ăn trưa 1.000.000đ/tháng cùng các khoản thưởng lễ tết, các ngày kỷ niệm thành lập Tổng Công ty, Trung tâm,...

Tham gia các chương trình bảo hiểm sức khỏe tại các Công ty Bảo hiểm lớn, có uy tín (Bảo Việt, PVI,...)

Thường xuyên tham dự các khóa đào tạo phong phú, đa dạng nhằm giúp CBCNV không ngừng phát triển bản thân, nâng cao kỹ năng cũng như chuyên môn nghiệp vụ.

Trải nghiệm các hoạt động văn hoá, thể thao, du xuân, nghỉ mát hè trong và ngoài nước,...

Hình 8: Mẫu đơn tuyển dụng

Thu thập data

```

name : CÔNG TY CỔ PHẦN CÔNG NGHỆ GEEK UP

Mô tả công việc : #Frontend #Backend #MobileTrong các kỳ thực tập, GEEK Up đã áp dụng các công nghệ đang thịnh hànhAndroid (Mobile);ReactJS (Frontend);NodeJS, NestJS, Docker, Amazon Web Services (Backend)...Tham gia project thực hiện 1 product thật trong 10 tuần, xây dựng product từ nhu cầu thực tế, quan tâm đến business requirements, đảm bảo user experience;Tham gia một quy trình phát triển product đầy đủ từ ý tưởng đến sản phẩm trên tay người dùng: Analysis - Design - Implementation - Operations;Liên tục được Coach hướng dẫn, review và unblock trong suốt 10 tuần thực tập;Được tìm hiểu và làm việc theo framework, process, standard được GEEK Up đúc kết trong suốt hành trình hơn 8 năm Phát triển Phần mềm của mình.

Yêu cầu : Là sinh viên năm 2, 3 hoặc 4 ngành Công nghệ thông tinĐịnh hướng phát triển Backend/Frontend/MobileCó thể đọc và sử dụng document chuyên môn bằng tiếng Anh

Quyền lợi : Giải thưởng cá nhânTOP PERFORMER TRI GIÁ 5.000.000 VNĐ dành cho những bạn thể hiện tốt nhất kỳ thực tập.CƠ HỘI TRỞ THÀNH MEMBER CHÍNH THỨCĐược hỗ trợ discount order nước uống tại văn phòng như nhân viên chính thức;Được Coach review và unblock, nhanh chóng phát triển cả kỹ năng chuyên môn về cả chiều rộng và chiều sâu;Quá trình Onboarding được đầu tư kỹ lưỡng, phối hợp cùng đồng đội ở nhiều chuyên môn: Product Design, Product Analysis, Product Operations...và nhiều hơn nữa về tổng quan quy trình làm product;Performance & Contribution được đánh giá & feedback rõ ràng ở cuối kỳ thực tập, bạn sẽ biết được level of expertise hiện tại của mình và sau khi kết thúc chương trình;Môi trường không phân cấp với Flat Structure, tạo cơ hội để bạn bày tỏ quan điểm, được đóng góp, ghi nhận và phát huy tối đa năng lực bản thân;Chương trình thực tập được thiết kế bài bản để bạn từng bước khám phá tiềm năng bản thân & phát triển trong cả kỹ năng chuyên môn lẫn kỹ năng mềm.

Cách thức ứng tuyển : Ứng viên nộp hồ sơ trực tuyến bằng cách bấmƯng tuyển ngaydưới đây.ƯNG TUYỂN NGAYLưu tinHạn nộp hồ sơ: 08/01/2022
  
```

Hình 9: Raw data thu thập từ trên internet

Tạo Dataframe

```
raw_recruit_df.show(5)
```

```
+-----+-----+-----+-----+-----+
|          name|      Mô tả công việc|      Yêu cầu ứng viên|      Quyền lợi|      Cách thức ứng tuyển|
+-----+-----+-----+-----+-----+
|CÔNG Ty Cổ Phần C...|- Làm việc trực t...|- Tốt nghiệp ĐH t...|- Thu nhập: Lên đ...|Ứng viên nộp hồ s...|
|  Viettel Digital|- \tPhát triển các...|- Có kinh nghiệm ...|- \tMức lương thu ...|Ứng viên nộp hồ s...|
|  Viettel Digital|Chịu trách nhiệm ...|Tốt nghiệp đại họ...|Mức lương thu hút...|Ứng viên nộp hồ s...|
|CÔNG TY CỔ PHẦN T...|- Cài đặt hệ điều...|- Tốt nghiệp trun...|- Lương từ 10 tri...|Ứng viên nộp hồ s...|
|  Viettel Digital|Phát triển các ứn...|Tốt nghiệp Đại họ...|Có thể phỏng vấn ...|Ứng viên nộp hồ s...|
+-----+-----+-----+-----+-----+
only showing top 5 rows
```

Hình 10: Dataframe từ rawdata

Extracted Dataframe

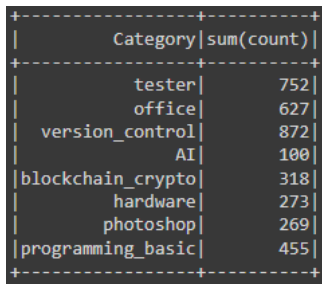
```
extracted_recruit_df.show(5)
```

CompanyName	FrameworkPlatforms	Languages	DesignPatterns	Knowledges	Salaries
Công Ty Cổ Phần C...	[MySQL, Oracle]	[]	[]	[kiếm thử]	[25]
Viettel Digital	[Git, JSON]	[Objective-C, Swi...]	[design pattern]	[UI/UX, Unit Test...]	[5]
Viettel Digital	[Reactjs, Vue, An...]	[css, JavaScript, ...]	[design pattern]	[đồ họa, UI/UX, f...]	[5]
CÔNG TY CỔ PHẦN T...	[.NET]	[]	[]	[Switch, phần cứng]	[0, 10]
Viettel Digital	[Git, JSON]	[XML, Java]	[]	[UI/UX, Unit Test...]	[5, 30]

only showing top 5 rows

Hình 11: Extracted Dataframe

Grouped Knowledge Dataframe

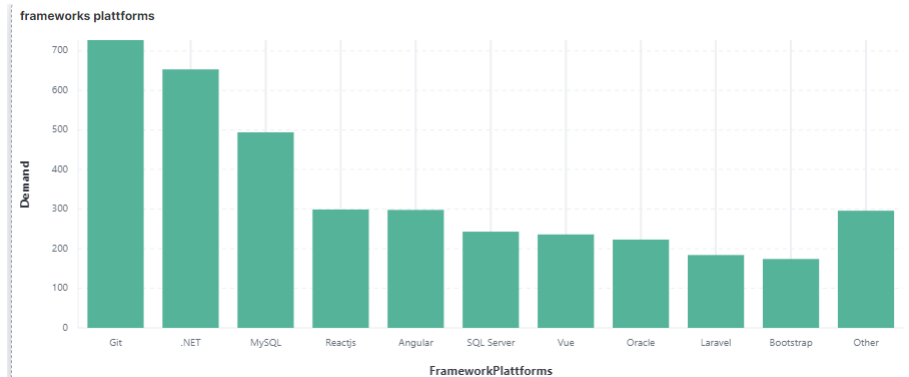


```
+-----+
|      Category | sum(count) |
+-----+
|      tester   |      752   |
|      office   |      627   |
| version_control |      872   |
|           AI  |      100   |
| blockchain_crypto |      318   |
|      hardware |      273   |
|    photoshop  |      269   |
| programming_basic |      455   |
+-----+
```

Category	sum(count)
tester	752
office	627
version_control	872
AI	100
blockchain_crypto	318
hardware	273
photoshop	269
programming_basic	455

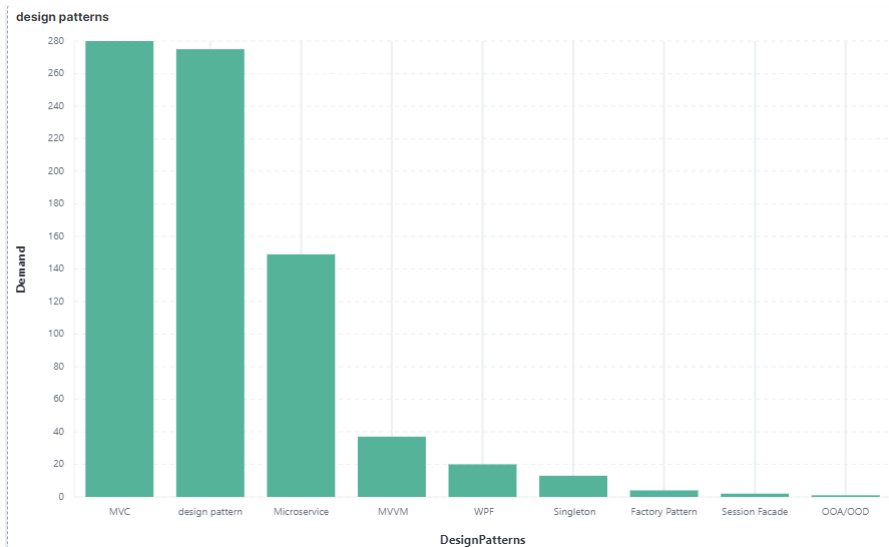
Hình 12: Grouped Knowledge Dataframe

Biểu diễn dữ liệu trên Kibana

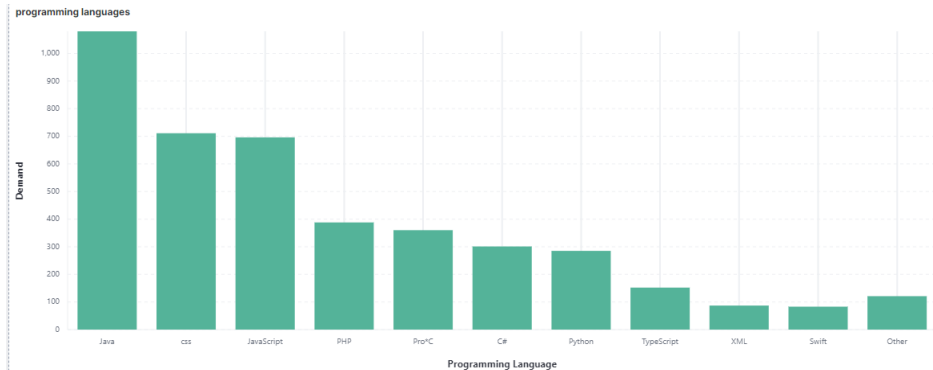


Hình 13: Thống kê Frameworks, Platforms

Biểu diễn dữ liệu trên Kibana

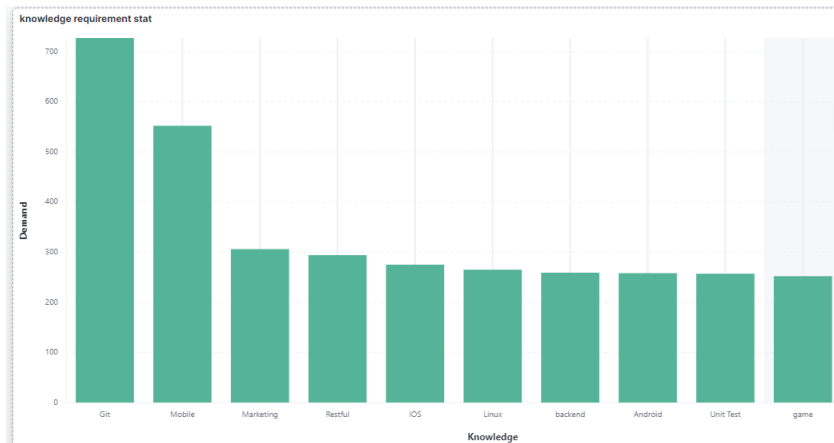


Biểu diễn dữ liệu trên Kibana



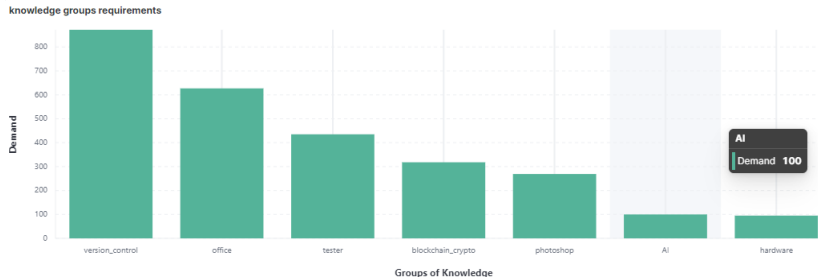
Hình 15: Thống kê ngôn ngữ lập trình

Biểu diễn dữ liệu trên Kibana



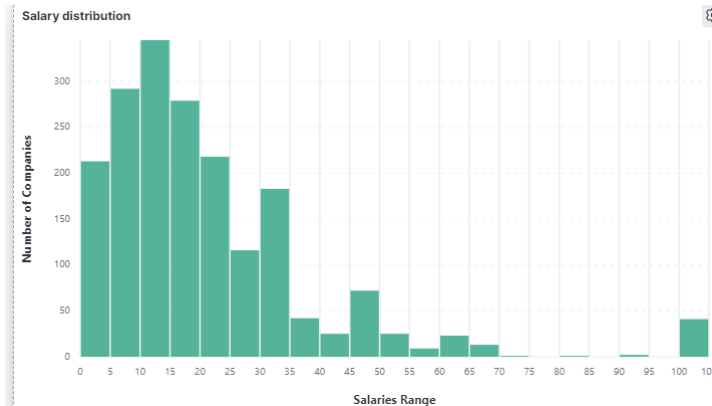
Hình 16: Thống kê kiến thức yêu cầu

Biểu diễn dữ liệu trên Kibana



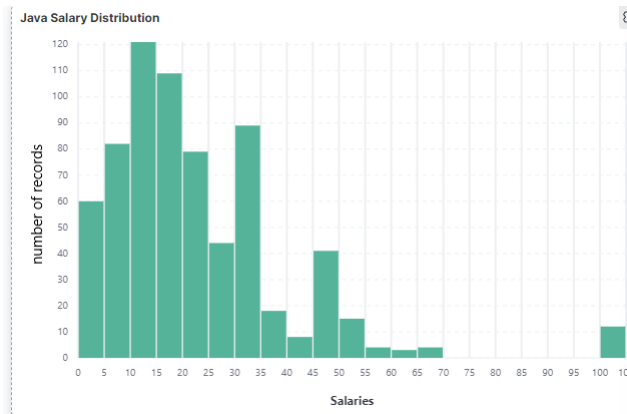
Hình 17: Thống kê nhóm các kiến thức

Biểu diễn dữ liệu trên Kibana



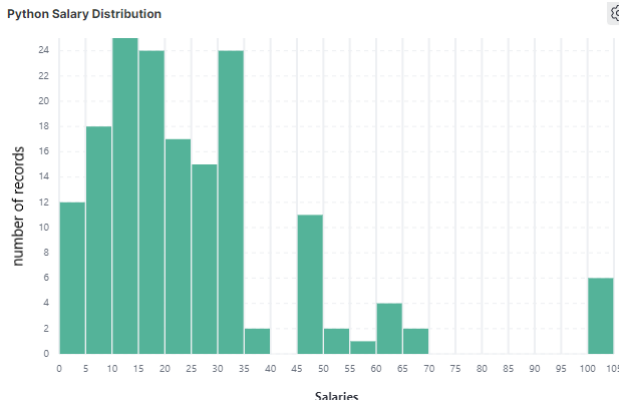
Hình 18: Thống kê lương

Biểu diễn dữ liệu trên Kibana



Hình 19: Thống kê lương offer cho Java

Biểu diễn dữ liệu trên Kibana



Hình 20: Thống kê lương offer cho Python

Hạn chế và hướng phát triển tiếp theo

- Số lượng spark-worker và elasticsearch node còn ít (khiến cho workload lên 1 node là tương đối lớn)
- Sử dụng spark-submit khiến cho việc xử lý dữ liệu không được linh hoạt (có thể khắc phục bằng cách sử dụng spark notebook)
- Khai thác thêm các ý nghĩa dữ liệu khác của đơn tuyển dụng
- Triển khai trên cụm máy tính thật