KỸ SƯ DỮ LIỆU

LÊ NGỌC KHÁNH

MỤC TIÊU NGHỀ NGHIỆP

Tôi mong muốn nâng cao kỹ năng xử lý dữ liệu streaming sử dụng Kafka và Spark Structured Streaming để phục vụ các bài toán real-time như giám sát hành vi người dùng hoặc phát hiện gian lận.

THÔNG TIN CÁ NHÂN

19/04/1981
Hà Nội
thaolinh252512@gmail.com
0765857296
www.website.com

HỌC VẤN

- Khoa học dữ liệu tại Đại học Công nghệ Thông tin – ĐHQG TP.HCM - Cơ sở dữ liệu và khai phá dữ liệu tại Đại học Khoa học Tự nhiên – ĐHQG Hà Nội

KỸ NĂNG

- Google Cloud Platform (GCP)

- AWS (S3, EMR, Glue)

- Apache Kafka

- NoSQL (MongoDB, Cassandra)

KINH NGHIỆM LÀM VIỆC

- **DATA ENGINEER** TẠI CÔNG TY DATAFLOW VIỆT NAM (2021-2023)

+ THIẾT KẾ VÀ TRIỂN KHAI PIPELINE ETL SỬ DỤNG APACHE AIRFLOW

+ TÍCH HỢP DỮ LIỆU TỪ NHIỀU NGUỒN NHƯ MYSQL, GOOGLE ANALYTICS VÀ API BÊN THỨ BA

+ TỐI ƯU HOÁ QUY TRÌNH XỬ LÝ BẰNG SPARK TRÊN GCP DATAPROC

- **BIG DATA ENGINEER** TẠI LOGIAI (2020-2022)

+ XÂY DỰNG PIPELINE THU THẬP VÀ XỬ LÝ DỮ LIỆU VẬN CHUYỂN TỪ THIẾT BỊ IOT

SỞ THÍCH

+ SỬ DỤNG SPARK STREAMING ĐỂ XỬ LÝ DỮ LIỆU REAL-TIME

- Thể thao

- Học ngoại ngữ

+ KẾT NỐI HỆ THỐNG VÀO ELASTICSEARCH VÀ PHÁT TRIỂN DASHBOARD GIÁM SÁT

- Tham gia hội thảo công nghệ

- Viết blog kỹ thuật

NGƯỜI GIỚI THIỆU

- **JUNIOR DATA ENGINEER** TẠI RETAILTECH (2020-2021)

- Bà Phạm Thanh Hương (Data Operations Manager - Công ty ReportPro) - huong.pham@reportpro.vn - 0912111222

+ PHỐI HỢP XÂY DỰNG HỆ THỐNG PHÂN TÍCH TỒN KHO THEO THỜI GIAN THỰC

- Ông Phan Thành Tâm (Big Data Architect - Công ty LogiData) - tam.phan@logidata.vn - 0955111222

+ XỬ LÝ DỮ LIỆU STREAMING TỪ KAFKA VÀ LƯU TRỮ VÀO BIGQUERY

+ XÂY DỰNG CÁC VIEW LOGIC TRÊN DBT PHỤC VỤ BI DASHBOARD

DANH HIỆU VÀ GIẢI THƯỞNG

- **2021** - Kỹ sư dữ liệu xuất sắc quý I tại Công ty DataBridge

- **2021** - Bằng khen về bảo mật và chất lượng dữ liệu – Công ty InfosecData

- **2021** - Giải thưởng 'Dự án phân tích dữ liệu xuất sắc' tại FinData

- **2022** - Top 3 nhân viên cải tiến hệ thống ETL tại SmartRetail

CHỨNG CHỈ

- **2023** - Designing Data-Intensive Applications – O'Reilly Certification Program

- **2021** - SQL for Data Engineering – Datacamp

- **2022** - ETL and Data Pipelines with Shell, Airflow and Kafka – Coursera

- **2023** - Modern Data Engineering with dbt – dbt Labs

HOẠT ĐỘNG

- Thành viên nhóm kỹ thuật dữ liệu tại Dự án E-Government (2023)

+ Thiết kế hệ thống thu thập dữ liệu hành chính từ nhiều bộ ngành.

+ Thiết lập hệ thống kiểm tra chất lượng dữ liệu tự động.

+ Triển khai pipeline đồng bộ dữ liệu hằng ngày với độ trễ thấp.

- Mentor khóa học nền tảng kỹ sư dữ liệu tại Trung tâm Đào tạo CloudTech (2023)

+ Hướng dẫn học viên triển khai hệ thống ingest dữ liệu bằng Kafka.

+ Đánh giá bài tập về xử lý dữ liệu song song với Spark.

+ Tư vấn về định hướng nghề nghiệp cho sinh viên muốn theo ngành data engineering.

- Diễn giả hội thảo 'Big Data Architecture' tại Data Talks Vietnam (2022)

+ Trình bày kiến trúc hệ thống thu thập và xử lý dữ liệu đa

nguồn.

+ Phân tích ưu nhược điểm của Data Warehouse vs Data Lakehouse.

+ Giới thiệu các công cụ phổ biến như Airflow, dbt, Snowflake.

DỰ ÁN

- Realtime Analytics cho hệ thống bán lẻ toàn quốc (Big Data Engineer, Retail360) 2023

Xây dựng hệ thống xử lý dữ liệu bán hàng theo thời gian thực để hỗ trợ ra quyết định tức thì.

+ Sử dụng Kafka để thu thập dữ liệu từ các chi nhánh toàn quốc

+ Xử lý dữ liệu streaming bằng Spark Structured Streaming

+ Gửi dữ liệu về Redshift và hiển thị trên dashboard BI

- Data Lake cho hệ thống quản trị khách hàng (CRM) (Data Engineer, CRMPro) 2022

Xây dựng nền tảng lưu trữ dữ liệu tập trung phục vụ phân tích hành vi khách hàng.

+ Tạo pipeline ingestion từ Salesforce, Google Ads và Facebook API

+ Lưu trữ dữ liệu theo mô hình phân vùng S3 Data Lake

+ Sử dụng Airflow để lập lịch và monitor luồng dữ liệu hàng ngày

- Kiến trúc dữ liệu cho nền tảng giáo dục trực tuyến (Cloud Data Engineer, LearnHub) 2022

Chuẩn hoá kiến trúc lưu trữ và xử lý dữ liệu học viên để phục vụ dashboard học tập.

+ Thiết kế hệ thống lưu trữ với BigQuery và Data Studio

+ Xây dựng Dataflow jobs để xử lý dữ liệu sự kiện học tập

+ Tạo luồng dữ liệu từ Firebase tới Google Cloud