**Nghiên cứu và thử nghiệm kỹ thuật phân tích dữ liệu theo thời gian thực trong lĩnh vực bán hàng.**

1. **Yêu cầu kỹ năng**

Clean code, Business process analysis & modeling, instructional design, development and testing, global project management, monitoring and researching skills.

1. **Mô tả khái quát:**

Phân tích dữ liệu lớn theo thời gian thực (realtime) có nghĩa là dữ liệu lớn được xử lý khi nó đến và người dùng doanh nghiệp có được thông tin chi tiết sử dụng mà không vượt quá khoảng thời gian được phân bổ cho việc ra quyết định hoặc hệ thống phân tích kích hoạt một hành động hoặc một thông báo.

Trọng tâm của đề tài tập trung nghiên cứu về Apache Spark, vốn nổi lên như một khung xử lý dữ liệu lớn quan trọng nhất. Cụ thể sẽ tập trung nghiên cứu về API Spark ML và Spark Streaming cho phép phân tích dữ liệu trong lĩnh vực bán hàng theo thời gian thực/gần thực. Sinh viên cũng nghiên cứu và thử nghiệm các mô hình phân tích dữ liệu trong lĩnh vực bán hàng, kết quả nghiên cứu được cài đặt tích hợp vào hệ thống.

1. **Nội dung chi tiết của đề tài bao gồm:**

* Khảo sát và tìm hiểu chủ đề phân tích dữ liệu bán lẻ trực tuyến. Xác định và phân tích đặc trưng của dữ liệu, các bài toán phân tích SOTA trong lĩnh vực bán lẻ.
* Nghiên cứu và thử nghiệm các kỹ thuật phân tích dữ liệu trong lĩnh vực bán lẻ, cụ thể gồm các bài toán: dự báo doanh số, phân khúc khách hàng, tư vấn giỏ hàng.
* Tìm hiểu về công cụ phân tích và xử lý dữ liệu Apache Spark, các thành phần ứng dụng mở rộng của nó (API Spark ML, Spark Stream, Spark SQL, Spark Core, …) và một số ứng dụng khác trong tương tác và phân tích dữ liệu thời gian thực.
* Tìm hiểu giải pháp lưu trữ NoSQL đảm bảo các tính năng quan trọng: tốc độ đọc và ghi, và khả năng mở rộng quy mô cho khối lượng dữ liệu lớn.
* Ứng dụng các công nghệ đã nghiên cứu cho việc cài đặt hệ thống quản lý và phân tích trực tuyến dữ liệu bán lẻ.
* Hệ thống cài đặt kết hợp sử dụng kết quả mô hình phân tích dữ liệu đã nghiên cứu.
* Kiểm tra, vận hành và đánh giá nhận xét về mặt khả thi hệ thống.

1. **Các kỹ năng và công nghệ sẽ được học khi tham gia dự án:**

* Kỹ năng nghiên cứu, khảo sát thị trường, tìm hiểu tài liệu, bài báo khoa học về một chủ đề.
* Kỹ năng nghiên cứu xây dựng mô hình phân tích dữ liệu kinh doanh.
* Kiến thức, kỹ năng trong xây dựng và triển khai một hệ thống phân tán.
* Kỹ năng khởi tạo và thiết kế các hệ thống có khả năng mở rộng cao có thể chấp nhận, lưu trữ và phân tích khối lượng lớn dữ liệu phi cấu trúc ở chế độ hàng loạt và / hoặc thời gian thực.
* Học tập giải pháp lưu trữ dữ liệu phi quan hệ (NoSQL)
* Học tập và làm quen các ngôn ngữ, framework lập trình web front-end và back-end như Vuejs, Vuetify, Django, …
* Cách sử dụng các mô hình phân tích dữ liệu trong các bài toán thực tế.
* Kỹ năng sắp xếp và quản lý tiến độ dự án, Kỹ năng làm việc nhóm.
* Kỹ năng viết tài liệu báo cáo.

Tài liệu tham khảo

Ecommerce:

* [https://towardsdatascience.com/data-analytics-in-e...](https://towardsdatascience.com/data-analytics-in-e-commerce-retail-7ea42b561c2f?fbclid=IwAR0FC0eiwc3j18XzLvS364fScb5RahsnIHVvV7KF-9C-6YCJgX8axSHJJfQ)
* [https://www.millimetric.ai/.../the-9-biggest-big-data.../](https://www.millimetric.ai/2020/10/13/the-9-biggest-big-data-challenges-for-ecommerce-directors/?fbclid=IwAR2zqnb8Er3KNMF20u9_EQWJad-ki61ou3x14Kf65lGpi6ie4o27u5mWrqw)
* [https://www.webofproceedings.org/.../IWASS.../SS06029.pdf](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.webofproceedings.org%2Fproceedings_series%2FESSP%2FIWASS%25202019%2FSS06029.pdf%3Ffbclid%3DIwAR3yuBWO5zIr5f46b8_8xDj-EUj-vaA85defy9k0Tc847wTcEvOYwcTKAv4&h=AT0yVRYXeejEwUAHhYS3EnIYEy7DGgoOcgU4uiBO_P9mBdczzMHCwqwFqcA5WY52Lmi2_O2SaU20u2Myvba4BxAJAFNezko_qfMeE0_zeuuxTUX0VC9fLDzjs1qVIYA6AZ-Si-eVlv-I-W1CHt2w&__tn__=-UK-R&c%5b0%5d=AT0JXEE8fyiJrNnSCgoBF0Qk2hSOgqGLJ96dQDG65KHn8vgUsAwWfzt0zX_8BLpqxip53mxNxUo3IRGbX_xWePKhBpGFpj1xiL6P6a6UA0MCvub_3t8oWTBHsV40ZIk-AWmhJdmPWZ4ZO2niFDtw-tkH9WiUO8GdnN-cQaU8tvD8fljKN0eJ1fsW8wXSO6iwisK5OWwsiG3KyiUb7TbaMyiM8OBP)

Data analysis tools and challenges:

* [https://www.researchgate.net/.../309254724\_Social...](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F309254724_Social_Networking_Data_Analysis_Tools_Challenges%3Ffbclid%3DIwAR34bIh7MLSY82h2bPayWKcwBkH9asbGvMM-xrCRHda3cds1CjShXu1o8zE&h=AT1uXpq4LkLxtRC-0jFXt2VZgfjQwoP0ii8fchV98z9NMM4RjFuuSqPyymKVR9bvrsMYc-4KHfOlcHhzdv2jdsueAguFnqBTVB4wXsy3BazLKsMFqjcHvfV61UxEk0BUkiPYZ6v0BqjPRZLS92dz&__tn__=-UK-R&c%5b0%5d=AT0JXEE8fyiJrNnSCgoBF0Qk2hSOgqGLJ96dQDG65KHn8vgUsAwWfzt0zX_8BLpqxip53mxNxUo3IRGbX_xWePKhBpGFpj1xiL6P6a6UA0MCvub_3t8oWTBHsV40ZIk-AWmhJdmPWZ4ZO2niFDtw-tkH9WiUO8GdnN-cQaU8tvD8fljKN0eJ1fsW8wXSO6iwisK5OWwsiG3KyiUb7TbaMyiM8OBP)

Apache Spark:

* [https://www.researchgate.net/.../272825265\_Big\_Data...](https://l.facebook.com/l.php?u=https%3A%2F%2Fwww.researchgate.net%2Fpublication%2F272825265_Big_Data_Analysis_Apache_Spark_Perspective%3Ffbclid%3DIwAR0eYOMOnyOd9aDhIJ5oyuDhS4m7OGZeUk6pUfemvHmvTDm-XlVmAJc3Gu8&h=AT0ZEtaAnOavFC6KFW9nYwmttE0d5eQbnK13nL9a8uvawgt_iL_5xR3gfD4HN61OdSJKNstmYfc_jTqhvAcV5_4bzp5IzTWwlwW3Gi6uDwKdItYKM2d_fU34EvdSSNgUrSXS_tVzO7Cz9TEfuwGO&__tn__=-UK-R&c%5b0%5d=AT0JXEE8fyiJrNnSCgoBF0Qk2hSOgqGLJ96dQDG65KHn8vgUsAwWfzt0zX_8BLpqxip53mxNxUo3IRGbX_xWePKhBpGFpj1xiL6P6a6UA0MCvub_3t8oWTBHsV40ZIk-AWmhJdmPWZ4ZO2niFDtw-tkH9WiUO8GdnN-cQaU8tvD8fljKN0eJ1fsW8wXSO6iwisK5OWwsiG3KyiUb7TbaMyiM8OBP)
* [https://www.oreilly.com/.../learn.../9781449359034/ch01.html](https://www.oreilly.com/library/view/learning-spark/9781449359034/ch01.html?fbclid=IwAR0qBt0hmHlJNoG__cHXRZP_2KTS1JJMO3wbmaYveMWEq84_QZ8hhTsgHFs)

Distributed systems:

* [https://www.freecodecamp.org/.../a-thorough.../...](https://www.freecodecamp.org/news/a-thorough-introduction-to-distributed-systems-3b91562c9b3c/?fbclid=IwAR1KL_7DJJ9ZbdrFnJFZ6SosFMz_bNZ9q34LHFCiicxFIVZU7R_9v09CEec)