9th International Interdisciplinary Congress on Renewable Energies, Industrial Maintenance, Mechatronics and Informatics

Booklets

RENIECYT
Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas
1702902 CONAHCYT

RENIECYT - LATINDEX - Research Gate - DULCINEA - CLASE - Sudoc - HISPANA - SHERPA UNIVERSIA - Google Scholar DOI - REDIB - Mendeley - DIALNET - ROAD - ORCID - V|LEX

Title: Implementation of a big data ecosystem for interdisciplinary analysis in academic projects: Collaboration Between Engineering and Communication in the Study of the 2024 Mexican Elections

Authors: Mendoza-González, Omar, Hernández-Cabrera, Jesús, Sánchez-Hernández, Miguel Ángel

Editorial label ECORFAN: 607-8695
BCIERMMI Control Number: 2024-01
BCIERMMI Classification (2024): 241024-0001
RNA: 03-2010-032610115700-14
Pages: 13

CONAHCYT classification:
Area: Engineering
Field: Technological sciences
Discipline: Computer technology
Subdiscipline: Data banks

ECORFAN-México, S.C.
Park Pedregal Business. 3580, Anillo Perif., San Jerónimo Aculco, Álvaro Obregón, 01900 Ciudad de México, CDMX,
Phone: +52 1 55 6159 2296
Skype: ecorfan-mexico.s.c.
E-mail: contacto@ecorfan.org
Facebook: ECORFAN-México S. C.
Twitter: @EcorfanC

www.ecorfan.org

Holdings
Mexico Colombia Guatemala
Bolivia Cameroon Democratic
Spain El Salvador Republic
Ecuador Taiwan of Congo
Peru Paraguay Nicaragua

PRESENTATION CONTENT

Introducción

Objetivos

Metodología

Contribución

Resultados

Conclusiones

Referencias

Introducción

- **Big Data y su relevancia**
  - En un contexto donde los datos son cada vez más esenciales, el manejo y procesamiento de grandes volúmenes de información se ha vuelto una competencia crucial, especialmente en la Ingeniería en Computación.
- **Objetivo del proyecto**
  - Implementar un ecosistema de Big Data en la FES Aragón, UNAM, para formar a los estudiantes en tecnologías de vanguardia y aplicar estas herramientas en proyectos académicos interdisciplinarios.
- **Colaboración interdisciplinaria**
  - Este ecosistema integra la capacidad técnica de la Ingeniería en Computación con el análisis cualitativo de la Comunicación, orientado al estudio de las elecciones presidenciales de México 2024.
- **Aplicación práctica**
  - Los estudiantes analizan cómo los candidatos construyen su imagen en redes sociodigitales, combinando el análisis de datos masivos con el estudio sociopolítico.

Objetivos

- **Desarrollar y aplicar** un ecosistema de Big Data en un contexto académico en la FES Aragón, UNAM.
- **Facilitar la colaboración** entre las carreras de Ingeniería en computación y Comunicación y Periodismo para analizar la comunicación política en redes sociodigitales durante las elecciones de México 2024.
- **Comparar la eficiencia** de herramientas de procesamiento de datos convencionales como Python frente a plataformas distribuidas como Hadoop y Spark.

Metodología

- **1. Configuración del Entorno de Big Data**
  - Se implementó un clúster de **Hadoop** virtualizado sobre una **estación de trabajo Dell Precision 7920**
  - El clúster incluyó un **nodo maestro** y **cuatro nodos esclavos**, operando sobre **Linux Raspbian**, con una conexión en red interna eficiente.
- **2. Lectura y Selección de Datos**
  - Datos brutos de **Twitter, Facebook e Instagram** almacenados en **CSV** y **Parquet**.
  - **Apache Spark** fue usado para seleccionar las columnas relevantes:
    - Identificadores de archivo
    - Tiempos de creación de anuncios
    - Impresiones, gastos y distribución por región.

Metodología

- **3. Aplicación del Módulo de Limpieza (moduloLimp)**
  - Funciones clave:
    - **imspend_clean:** Limpieza de datos de impresiones y gastos.
    - **delivery_clean:** Procesamiento de datos en formato JSON a texto plano.
    - **char_map:** Corrección de caracteres no UTF-8.
- **4. Manipulación de Datos**
  - **Cálculo de promedios:** Impresiones y gastos divididos en rangos, cálculo de promedios para análisis detallado.

Metodología

- **5. Análisis y Visualización:**
  - Se utilizó **Spark** para procesamiento de datos y **Matplotlib** para visualización:
    - Filtrado de publicaciones relevantes.
    - Visualización de actividad en redes.
    - Análisis de publicaciones duplicadas.

Metodología

Implementación de un Ecosistema de Big Data para el Análisis Interdisciplinario en Proyectos Académicos: Colaboración entre Ingeniería y Comunicación en el Estudio de las Elecciones en México 2024

Objetivos	Metodología	Contribución
Desarrollar y aplicar un ecosistema de Big Data en un contexto académico en la FES Aragón, UNAM.	**Configuración del Entorno:** Implementación de un clúster de Hadoop virtualizado con múltiples nodos para el procesamiento de datos a gran escala.	**Demostrar el valor** de la colaboración interdisciplinaria entre ingeniería en computación y comunicación para la investigación social y política.
Facilitar la colaboración para analizar la comunicación política en redes sociodigitales durante las elecciones de México 2024.	**Procesamiento de Datos:** Ingesta, limpieza y estructuración de grandes volúmenes de datos provenientes de redes sociodigitales.	**Establecer puntos de referencia** para la eficiencia de diferentes herramientas de procesamiento de datos, destacando la superioridad de Hadoop y Spark para tareas a gran escala.
Comparar la eficiencia de herramientas de procesamiento de datos convencionales como Python frente a plataformas distribuidas como Hadoop y Spark.	**Herramientas de Análisis:** Uso de Spark para el procesamiento de datos y Matplotlib para la visualización; comparación del rendimiento y escalabilidad entre Python y Hadoop/Spark.	**Proporcionar un marco** para futuros proyectos académicos interdisciplinarios que utilicen tecnologías de Big Data.

Contribución

- **Demostrar el valor** de la colaboración interdisciplinaria entre ingeniería en computación y la licenciatura en comunicación y periodismo para la investigación social y política.

- **Establecer puntos de referencia** para la eficiencia de diferentes herramientas de procesamiento de datos, destacando la superioridad de Hadoop y Spark para tareas a gran escala.

- **Proporcionar un marco** para futuros proyectos académicos interdisciplinarios que utilicen tecnologías de Big Data.

Resultados

- **Colaboración Interdisciplinaria y Análisis de Comunicación Política**
  - **Transformación de datos crudos** en información estructurada permitió un análisis profundo de la **comunicación política**.
  - Se evidenció cómo los **candidatos presidenciales de México 2024** utilizaron **estrategias retóricas** (ethos, pathos, logos) para construir su imagen pública.
  - Las configuraciones de **Big Data** resultaron eficaces, proporcionando un entorno flexible y escalable para el manejo de grandes volúmenes de información.
  - Esta **metodología interdisciplinaria** enriqueció el análisis cuantitativo y cualitativo, combinando habilidades técnicas y comunicativas.

Resultados

Resultados de la Limpieza y Estructuración de Datos

Proceso	Porcentaje de Texto Eliminado	Porcentaje de Texto Conservado
Limpieza de delivery_by_region	60% - 66%	40% - 44%
Corrección de Texto (ad_creative_bodies)	3% - 16%	86% - 97%

Rendimiento de Clústeres de Big Data

Nodos	Comienzo	Lanzamiento	Finalización	Tiempo	Memoria usada	Cluster utilizado	Datos procesados
4	13:38:19	13:38:22	13:41:04	162.5s	2048 MB	16.70%	500 MB

Resultados

- **Rendimiento del Clúster Hadoop**
  - **Comparación de tiempos de procesamiento:**
    - Dataset de **145 KB en Python**: 20 segundos.
    - Estimación para **500 MB en Python**: 19 horas y 9 minutos.
    - **500 MB en Hadoop**: 2 minutos y 43 segundos.
  - **Diferencia de rendimiento:**
    - Python en un entorno estándar toma casi un día, mientras que **Hadoop** procesa la misma cantidad en **menos de 3 minutos**, demostrando la eficacia de los entornos distribuidos.
  - **Optimización de la infraestructura** para futuros proyectos con base en las mediciones detalladas del rendimiento del clúster.

Conclusiones

- El proyecto **PAPIIT** demostró el valor del **Big Data** en la investigación interdisciplinaria, especialmente en el análisis de fenómenos sociopolíticos complejos como las **campañas electorales**.
- La colaboración entre **Ingeniería en Computación** y la **Licenciatura en Comunicación y Periodismo** en la **FES Aragón, UNAM**, generó resultados significativos, fortaleciendo la capacidad de abordar desafíos contemporáneos con **tecnologías avanzadas**.
- **Comparación de metodologías:** Python mostró limitaciones de escalabilidad para grandes volúmenes de datos, mientras que **Hadoop** y **Spark** demostraron ser más eficientes y rápidos.
- Este enfoque establece un precedente importante para futuras investigaciones, resaltando la necesidad de utilizar **herramientas de Big Data** para el análisis a gran escala.

Referencias

Referencias

Chang, F., Dean, J., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. (2008). Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS), 26(2), 1-26. https://doi.org/10.1145/1365815.1365816

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 137-150. https://doi.org/10.1145/1327452.1327492

Ghemawat, S., Gobioff, H., & Leung, S. (2003). The Google file system. In Proceedings of the 19th ACM Symposium on Operating Systems Principles (pp. 29-43). ACM. https://doi.org/10.1145/945445.945450

Karau, H., Warren, R., & Wendell, P. (2017). High performance Spark: Best practices for scaling and optimizing Apache Spark (1st ed.). O'Reilly Media.

Moreno, J., Fernandez, E. B., Serrano, M. A., & Fernández-Medina, E. (2019). Secure development of big data ecosystems. IEEE Access, 7, 96604-96619. https://doi.org/10.1109/ACCESS.2019.2929330

Sammer, E. (2021). Hadoop operations: A guide for developers and administrators (2nd ed.). O'Reilly Media.

Saroha, M., & Sharma, A. (2019). Big data and Hadoop ecosystem: A review. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1040-1044). IEEE. https://doi.org/10.1109/ICSSIT46314.2019.8987848

White, T. (2022). Hadoop: The definitive guide (4th ed.). O'Reilly Media.

Zaharia, M., & Wenchen, F. (2020). Learning Spark: Lightning-fast data analytics (2nd ed.). O'Reilly Media.

ECORFAN®