Informe académico del curso manejo de grandes volúmenes de información con Spark, Scala y AWS

Carlos Andrés Caro Sáenz
Sebastián Castro Ochoa

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Maria Bernarda Salazar Sánchez, Doctora (Ph.D.) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2025

Cita	(Caro Saénz & Castro Ochoa, 2025)
Referencia **Estilo APA 7 (2020)**	Caro Saénz, C. A., & Castro Ochoa, S. (2025). *Informe académico del curso manejo de grandes volúmenes de información con Spark, Scala y AWS* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Danny Alexandro Múnera Ramírez
Coordinadora del Programa: Maria Bernarda Salazar Sánchez

Dedicatoria

Dedicado a mi esposa y familia quienes siempre han confiado en mí.

Carlos Caro

Dedicado al Sebastián que con mucha duda realizó este proceso. Se cumplió.

Sebastián Castro

Agradecimientos

Agradecimientos a Dios, mi esposa por apoyarme siempre, a mi compañero Sebastián por todo el compromiso y ganas en esta etapa, a nuestra asesora Maria B. Salazar por todo el apoyo y orientación en este informe.

Carlos Caro

A mi familia por el apoyo incondicional, al laboratorio clínico hematologico por hacer posible este propósito, a los profesores por compartir su conocimiento y a mi compañero Carlos por hacer más ameno este camino.

Sebastián Castro

Tabla de contenido

Siglas, acrónimos y abreviaturas

PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia
IA	Inteligencia Artificial
AWS	Amazon Web Services
HDFS	Hadoop Distributed File System
EC2	Elastic Compute Cloud
S3	Simple Storage Service
EMR	Elastic MapReduce
DAG	Directed Acyclic Graph
API	Application Programming Interface
RDD	Resilient Distributed Dataframe
SQL	Structured Query Language
CSV	Comma Separated Values

Resumen

Este trabajo presenta un análisis descriptivo del curso “Manejo de grandes volúmenes de información con Spark, Scala y AWS”, enfocado en las herramientas fundamentales para la gestión de datos a gran escala. Se exploran los principales paradigmas de procesamiento distribuido, los lenguajes y frameworks más utilizados como PySpark, Scala y scrapy, así como las técnicas comunes para la extracción y transformación de datos. Además, se analizan las ventajas del uso de servicios en la nube para la creación y gestión de clústeres de procesamiento Big Data, con especial énfasis en la integración de soluciones ofrecidas por Amazon Web Services (AWS), incluyendo EC2, S3 y EMR.

Este enfoque permite comprender cómo estas tecnologías se complementan para construir soluciones escalables, eficientes y adaptadas a las necesidades actuales del análisis de datos.

Palabras clave — Big data, Hadoop, Spark, Scala, AWS.

1. Introducción

En un mundo impulsado por los datos, la capacidad de gestionar y analizar grandes volúmenes de información se ha convertido en una habilidad esencial para enfrentar los retos del siglo XXI, ya que estrategias tradicionales de tratamiento de datos con Python, por ejemplo, mediante el uso de librerías como pandas y numpy no logran ofrecer rendimientos óptimos a medida que aumenta la cantidad de información a procesar, es por esto que se da lugar al uso de herramientas con mayor escalabilidad basadas en estrategias de procesamiento distribuido, el cual propone un cambio de procesamiento secuencial o lineal, por uno paralelo donde se pueda aprovechar las ventajas del escalamiento horizontal o bien aumento del número de máquinas para el cómputo.

En este informe se encuentra plasmado el análisis académico realizado entorno al curso 50 hrs big data mastery: pyspark, Amazon Web Services (AWS), scala & data scraping ofrecido por la plataforma Udemy (Sciences, 2024). Esta posibilidad de formación virtual, refuerza conceptos importantes de ingeniería de datos que permiten resolver problemas utilizando tecnologías como Hadoop (Changxi Ma, 2023), el cual brinda un ecosistema de herramientas adaptadas a los volúmenes de información manejados, en especial mediante el uso de Spark (Marjan Asgari, 2022); lenguajes de programación optimizados para estas tareas como Scala (Chen Li, 2025), el cual se encuentra muy bien integrado para personalizar y sacar mayor provecho a Spark; recursos de computación en la nube ofrecidos AWS (Shengying Yang, 2023); y habilidades de gran utilidad como el “scrapy” (Priam Pillai, 2020) para extracción de información.

El objetivo de este informe es la descripción de los conceptos relevantes para el manejo de grandes volúmenes de información a través de las tecnologías del ecosistema Hadoop anteriormente mencionadas, de modo que se logrará abordar asuntos como servicios adicionales de Hadoop, aplicaciones de Spark y características principales de Scala para el manejo de información e implementación clusters sobre servicios de AWS.

En general, considerando la tendencia del mundo de los datos y el auge de los modelos de inteligencia artificial (IA) generativa, la gestión de grandes volúmenes de información se convierte sin duda en una tarea fundamental para potenciar el negocio.

2. Justificación Académica

El curso inicia con una introducción a los principios del manejo de grandes volúmenes de datos (Big Data) y su procesamiento distribuido, destacando el rol fundamental del ecosistema Hadoop. Se aborda la forma en que HDFS (Hadoop Distributed File System) permite almacenar datos a gran escala de forma eficiente y tolerante a fallos, preparando el terreno para herramientas más ágiles como Apache Spark.

Spark: Potencia y velocidad en el procesamiento
Apache Spark se presenta como la evolución natural del procesamiento distribuido. Gracias a su arquitectura en memoria, permite transformar y analizar datos a gran velocidad. El curso introduce a PySpark como la interfaz en Python para interactuar con Spark, enseñando transformaciones, acciones y optimización de tareas en clústeres distribuidos.

Scala: Precisión y rendimiento en la analítica
El lenguaje Scala se explora por su estrecha integración con Spark, ofreciendo una sintaxis concisa y orientada a objetos que mejora la expresividad del código. Esta combinación resulta ideal para crear pipelines de procesamiento robustos y personalizados en entornos de Big Data.

Scraping de datos y AWS: de la recolección al despliegue
Una sección esencial del curso está dedicada a Scrapy, un framework en Python que facilita la extracción de información web, permitiendo construir datasets personalizados para análisis posteriores. Además, se muestra cómo desplegar soluciones Big Data en la nube utilizando servicios de AWS como EC2 (máquinas virtuales), S3 (almacenamiento) y EMR (Elastic MapReduce), brindando escalabilidad y eficiencia a los proyectos.

El curso ofrece una formación sólida y actualizada en tecnologías clave del entorno Big Data, y su integración en un contexto académico potencia la comprensión de los conceptos mediante la práctica. Aprender Spark con Scala, automatizar la recolección de datos y desplegar modelos en la nube con AWS son habilidades que hoy marcan la diferencia en el campo de la analítica avanzada.
.

3. Descripción Técnica

Para el desarrollo del curso se abordaron diferentes temáticas que serán descritas a continuación, cada temática se revisó mediante enfoques teórico-prácticos que permitieron comprender el concepto y su implementación según corresponde, esencialmente desde entornos locales de ejecución mediante el uso de Visual Studio Code como entorno de desarrollo.

3.1. Fundamentos de Big Data y Ecosistema Hadoop

3.1.1. ¿Qué es Big Data?

El crecimiento exponencial en la generación de datos durante las últimas décadas ha potenciado un cambio sustancial en la forma en que se almacenan, procesan y analizan grandes volúmenes de información. A todo ese crecimiento se le denomina Big Data, y no solo implica el manejo de datos en gran cantidad, sino también su variedad, velocidad y veracidad, lo que también se conoce como las "V" del Big Data.

En este nuevo paradigma, las herramientas tradicionales como pandas o numpy se ven limitadas ante flujos masivos de información, y es allí donde las tecnologías del ecosistema Hadoop ganan relevancia. Estas tecnologías permiten implementar esquemas de procesamiento distribuido, donde múltiples nodos trabajan en paralelo para realizar tareas de lectura, transformación y análisis de datos, optimizando tanto el tiempo de respuesta como la escalabilidad del sistema.

3.1.2. HDFS y el procesamiento distribuido

Dentro de este ecosistema, uno de los componentes esenciales es el Hadoop Distributed File System (HDFS), un sistema de archivos diseñado específicamente para trabajar con volúmenes masivos de datos en entornos distribuidos. A diferencia de los sistemas de archivos tradicionales HDFS divide los archivos en bloques y los almacena de forma redundante en diferentes nodos del clúster, lo cual brinda una alta tolerancia a fallos y disponibilidad de la información, ya que esta se encuentra distribuida, y en algunos casos, replicadas en varios nodos según se configure el sistema.

El procesamiento distribuido no solo reduce cuellos de botella, sino que permite aprovechar las ventajas del escalamiento horizontal, es decir, en lugar de centrarse en una sola máquina con mayor cantidad de recursos se pueden añadir más nodos al sistema, lo que mejora la capacidad de cómputo y almacenamiento del sistema en conjunto. Esta lógica de trabajo se exploró de manera práctica en el curso, evidenciando cómo el diseño de HDFS facilita el trabajo posterior con motores de procesamiento como Spark.

3.2. Apache Spark como motor de procesamiento

Apache Spark surge como una evolución natural de los modelos de procesamiento distribuido, ofreciendo una alternativa moderna al modelo de MapReduce. Uno de sus mayores aportes radica en su capacidad para trabajar en memoria, evitando así la

escritura constante en disco, lo cual proporciona importantes mejoras de rendimiento, especialmente en flujos de datos iterativos o aplicaciones de machine Learning, sin embargo, Spark puede no ser tan eficiente como otras herramientas diseñadas específicamente para ese fin.

3.2.1. Ventajas frente a MapReduce

Si bien MapReduce fue una innovación clave para el manejo de Big Data, su modelo secuencial y su fuerte dependencia de la lectura y escritura en disco constituyeron ciertas limitaciones. Spark, por el contrario, introduce el concepto de arquitectura basada en DAGs (grafos acíclicos dirigidos) que permite representar y optimizar flujos de datos de manera más eficiente. Este enfoque reduce la latencia y mejora la reutilización de resultados intermedios.

Durante el curso, se profundizó en estas diferencias a través de ejercicios prácticos, demostrando cómo tareas que tomarían varios minutos en MapReduce pueden resolverse en segundos con Spark, gracias a su ejecución en memoria y su modelo de programación más intuitivo.

3.2.2. Transformaciones y acciones en PySpark

Uno de los módulos centrales del curso se enfocó en PySpark, la API de Apache Spark para Python, que combina la potencia de Spark con la familiaridad del lenguaje de programación Python. En este contexto, se abordaron conceptos fundamentales como las transformaciones (filter(), map(), select()) y las acciones (count(), collect(), show()), entendidas como operaciones perezosas y ejecutoras respectivamente. Este modelo de ejecución diferida permite optimizar el plan de ejecución de los procesos antes de llevarlos a cabo, lo cual es una diferencia clave respecto a otras librerías de Python.

En otras palabras, las operaciones de transformación se convierten en indicaciones que son procesadas por Spark para crear un plan de ejecución optimizado, y que se aplican cuando se realizan acciones como count, collect o show. De esta forma se logra una mejora sustancial en comparación con librerías como Pandas que realiza las transformaciones de inmediato. En la sección correspondiente del curso, se realizaron ejercicios que involucraban la manipulación de grandes volúmenes de datos en estructuras como Resilient Distributed Dataframes (RDDs) y DataFrames, lo que permitió tener una vista práctica del concepto y las ventajas que ofrece.

3.3. Lenguaje Scala y su integración con Spark

3.3.1. Características clave de Scala

Scala es un lenguaje de programación moderno que combina elementos de la programación orientada a objetos y funcional, lo que lo convierte en una herramienta poderosa y versátil para el desarrollo de soluciones de procesamiento de datos. Aunque no tiene la misma popularidad que Python o Java, Scala se ha posicionado como un lenguaje de referencia en entornos de Big Data gracias a su integración nativa con Apache Spark.

Entre sus características clave se destacan su tipado fuerte y estático, su capacidad de inferencia de tipos, su sintaxis y su interoperabilidad con Java, lo que permite aprovechar bibliotecas ya existentes en Java. Estos aspectos hacen de Scala una opción robusta para desarrollar soluciones eficientes y de baja latencia, especialmente cuando se requiere optimización en el manejo de datos distribuidos.

Durante el desarrollo del curso, Scala fue introducido desde un enfoque práctico, facilitando su aprendizaje a través de ejercicios que incluían desde funciones básicas hasta expresiones lambda y colecciones inmutables, lo cual permitió establecer una base sólida para su uso posterior en Spark.

3.3.2. Casos de uso combinando Scala y Spark

La sinergia entre Scala y Spark va más allá de la compatibilidad técnica, Spark fue escrito originalmente en Scala, lo que garantiza que muchas de las nuevas funcionalidades o mejoras estén disponibles primero para este lenguaje. En consecuencia, trabajar con Spark desde Scala ofrece ventajas como mejor rendimiento en la ejecución y acceso a una documentación más completa en ciertos temas avanzados.

En el curso se desarrollaron casos de uso que involucraban procesamiento de logs, análisis de archivos CSV de gran tamaño y transformaciones complejas usando APIs de Spark SQL y RDDs. Este enfoque permitió contrastar la experiencia de uso entre PySpark y Scala, y evidenciar cómo el dominio de ambos lenguajes abre mayores oportunidades laborales en proyectos de ingeniería de datos, donde el rendimiento y la optimización del código son factores críticos.

3.4. Computación en la nube con AWS

La evolución hacia la computación en la nube ha transformado la forma en que las empresas gestionan sus infraestructuras de datos. Amazon Web Services (AWS) se ha posicionado como una de las plataformas más fuertes en este campo, ofreciendo un amplio catálogo de servicios para el almacenamiento, procesamiento y análisis de grandes volúmenes de información.

Uso de EC2, S3 y EMR para procesamiento Big Data

Durante el curso, se abordó el uso de servicios clave de AWS como EC2 (Elastic Compute Cloud), S3 (Simple Storage Service) y EMR (Elastic MapReduce). Donde se pudo comprender cómo cada uno cumple un rol fundamental en la arquitectura de Big Data moderna

- **EC2** permite la creación de instancias virtuales que funcionan como servidores escalables para desplegar aplicaciones, incluyendo clústeres de Spark personalizados.
- **S3** proporciona almacenamiento distribuido y altamente disponible para guardar datasets de entrada, resultados de análisis y copias de seguridad.
- **EMR** facilita el despliegue automatizado de clústeres de Hadoop y Spark, permitiendo ejecutar cargas de trabajo de Big Data sin necesidad de configuración manual compleja.

La combinación de estos servicios permite implementar pipelines de procesamiento robustos, donde el escalamiento horizontal puede ser automatizado e integrado con otros servicios de Nube aplicables a la solución. La flexibilidad que aporta el uso de servicios de nube no solo reduce

costos, sino que permite que los equipos de datos se concentren en el desarrollo de soluciones y no tanto en la gestión de infraestructura, liberando así carga operativa y administrativa importante para el enfoque en otros temas.

4. Conclusiones

La realización de este curso nació de la necesidad de adoptar los conocimientos básicos necesarios sobre las tecnologías de procesamiento distribuido, con el fin de enfrentar los retos asociados al manejo de grandes volúmenes de información. A partir del análisis del curso se consolidaron conocimientos clave sobre herramientas como Hadoop, Spark, Scala, Web Scrapy y AWS, destacando su aplicabilidad real mediante ejercicios prácticos. Esta formación refuerza competencias fundamentales en la ingeniería de datos moderna, y aporta una base sólida para el diseño de soluciones escalables, eficientes y alineadas con las demandas actuales del sector tecnológico. En general, estos conocimientos aportan gran versatilidad y amplitud en los campos de acción al ingeniero especializado en datos, por lo que la profundización de estos temas aporta ventajas laborales para empleabilidad.

Referencias

Changxi Ma, M. Z. (2023). An overview of Hadoop applications in transportation big data. *Journal of Traffic and Transportation Engineering (English Edition)*, Volume 10, Issue 5.

Chen Li, Y. Z. (2025). Mining area skyline objects from map-based big data using Apache Spark framework. *Array*, Volumen 25.

Marjan Asgari, W. Y. (2022). Spatiotemporal data partitioning for distributed random forest algorithm: Air quality prediction using imbalanced big spatiotemporal data on spark distributed framework. *Environmental Technology & Innovation*, Volumen 27.

Priam Pillai, D. A. (2020). Understanding the requirements Of the Indian IT industry using web scrapping. *Procedia Computer Science*, Volume 172.

Sciences, A. (2024). *50 Hrs Big Data Mastery: PySpark, AWS, Scala & Data Scraping*. Obtenido de Udemy: https://www.udemy.com/course/big-data-with-scalasparkpyspark-and-aws-a-z-50-hours/

Shengying Yang, W. J. (2023). Optimized hadoop map reduce system for strong analytics of cloud big product data on amazon web service. *Information Processing & Management*, Volume 60, Issue 3.