UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

PROYECTO FINAL

DOCENTE:

ING. DANILO JARAMILLO

POR:

Byron Daniel Córdova Romero

Jean Carlos Iñiguez Gutierrez

Daniel Sebastian Ludeña Guaman

FECHA:

25/07/2025

ASIGNATURA: PROGRAMACIÓN AVANZADA

INDICE

Contenido

Introducción 4

Repositorio GitHub: 4

Datos Base: 4

Datos Complementarios 6

Codigos de ofertas laborales de Ecuador: 6

Diseño lógico relacional 6

Diccionario de datos 7

Normalización: 8

Tabla oferta_laboral_ecuador: 8

Tabla localizacion 8

Tabla requisitos: 8

Script SQL 9

Herramientas utilizadas 9

Sistema operativo Ubuntu 9

¿Qué es Zeppelin? 9

Configuración del entorno 10

Instalación de la Máquina Virtual 10

Permisos de usuario de Ubuntu 10

Instalación de mysql 10

Instalación de Zeppelin 11

Instalación de java 11

Instalación de Python 11

Instalación de Spark 12

Instalación de Scala 12

Interprete Zeppelin configuración de mysql, spark , Python 12

Configuración de Spark en Interprete de Zepellin 12

Configuración de mysql en Interprete de Zepellin 13

Configuración de Python en Interprete de Zeppelin 13

Análisis a realizar 13

Visualización de consultas 13

Consulta 1 13

Consulta 2 14

Consulta 3 15

Consulta 4 15

Consulta 5 16
Conclusiones 16
Bibliografía 16

Introducción

El presente informe describe el proceso de análisis de un conjunto de datos relacionados con ofertas laborales, provenientes de un archivo CSV que contiene información detallada sobre vacantes disponibles, requisitos asociados y localización geográfica. Utilizando Apache Spark con la API de Scala, se realizaron diversas consultas estructuradas para responder a preguntas clave como: ¿cuál es el área de estudio más solicitada por provincia y cantón?, o ¿qué regiones demandan más servicios profesionales?

El objetivo principal fue realizar este análisis empleando exclusivamente DataFrames en Spark. Se llevaron a cabo operaciones de transformación como join, groupBy y agg, priorizando buenas prácticas y eficiencia en la manipulación de grandes volúmenes de datos.

El análisis no solo permitió extraer conclusiones relevantes sobre las tendencias laborales en distintas regiones del país, sino que también fortaleció las habilidades en el manejo de datos distribuidos y consultas estructuradas dentro del ecosistema de Apache Spark.

Repositorio GitHub:

Repositorio: Jean-IG23/Proyecto_Final

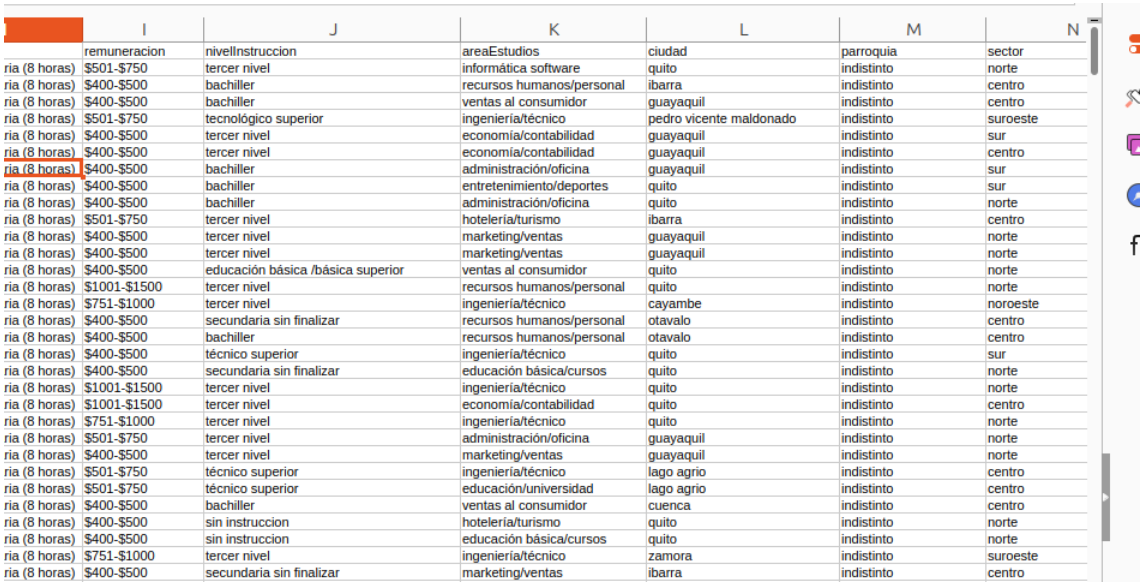
Datos Base:

Los datos utilizados para este análisis provienen de un archivo CSV que contiene información sobre las **ofertas laborales registradas a nivel nacional**, correspondiente al mes de investigación. Este conjunto de datos forma parte de un estudio orientado a comprender mejor la demanda laboral en distintas regiones del país, desagregada por provincia, cantón, área de estudios requerida y tipo de servicio solicitado.
A continuación se presenta una descripción de las variables presentes en los datos:

Id_oferta_laboral	Es el identificador principal de la tabla oferta laboral.
cargo	Se refiere al nombre del puesto ofrecido.
modo	Se refiere al modo de trabajo de la oferta.
fechaPublicado	Se refiere a la fecha en que la oferta laboral fue publicada.
fechaFin	Se refiere a la fecha límite para aplicar a la oferta laboral.
plazas	Se refiere al número de vacantes disponibles.
capacitacion	Se refiere a la información sobre si se ofrece capacitación.
jornadas	Se refiere al tipo de jornada laboral.
remuneracion	Se refiere al salario o remuneración ofrecida para el puesto.

idRe	Es el identificador del requisito asociado a la oferta laboral.
idLo	Es el identificador de la localización asociada a la oferta laboral.
tipo_contrato	Se refiere al tipo de contrato laboral ofrecido.
id(en localizacion)	Es el identificador único de la localización.
ciudad	Se refiere al nombre de la ciudad donde se ubica.
canton	Se refiere al nombre del cantón donde se ubica.
provincia	Se refiere al nombre de la provincia donde se ubica.
region	Se refiere al nombre de la región donde se ubica.
Id(en requisitos)	Es el identificador único del requisito.
nivelInstruccion	Se refiere al nivel de instrucción o educación requerido.
areaEstudios	Se refiere al área de estudios específica requerida
expriencia	Se refiere a la experiencia laboral requerida
nivel_rigor_tecnico	Se refiere al nivel de rigor técnico o especialización requerido

	A	B	C	D	E	F	G	H
1	cargo	modo	fechaPublicado	fechaFin	plazas	experiencia	capacitacion	jornadas
2	desarrollador java jee	tiempo completo	2013-02-22	2013-03-24	1	sin experiencia	0-50 horas	jornada ordinaria (8 horas)
3	guardias de seguridad con experiencia 1 año	tiempo completo	2013-02-23	2013-02-28	20	7-12 meses	0-50 horas	jornada ordinaria (8 horas)
4	vendedor	tiempo completo	2013-02-23	2013-02-26	2	1-3 años	0-50 horas	jornada ordinaria (8 horas)
5	topografo	por obra	2013-02-23	2013-02-28	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
6	asistente contable	tiempo completo	2013-02-25	2013-03-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
7	asistente contable	tiempo completo	2013-02-25	2013-02-26	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
8	mensajero	tiempo completo	2013-02-25	2013-02-26	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
9	guardias de seguridad	tiempo completo	2013-02-25	2013-02-26	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
10	recepcion	tiempo completo	2013-02-26	2013-03-28	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
11	guias turisticas	tiempo completo	2013-02-26	2013-03-01	2	1-3 años	0-50 horas	jornada ordinaria (8 horas)
12	supervisor de promotoria con experiencia	tiempo completo	2013-02-26	2013-02-28	2	1-3 años	0-50 horas	jornada ordinaria (8 horas)
13	director comercial	tiempo completo	2013-02-26	2013-03-01	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
14	vendedor de percha	tiempo completo	2013-02-26	2013-02-28	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
15	trabajador (a) social	tiempo completo	2013-02-26	2013-03-06	1	4-6 años	0-50 horas	jornada ordinaria (8 horas)
16	ingeniero	por obra	2013-02-26	2013-03-04	4	1-3 años	0-50 horas	jornada ordinaria (8 horas)
17	guardias de seguridad con experiencia 1 año	tiempo completo	2013-02-26	2013-03-01	20	1-3 años	0-50 horas	jornada ordinaria (8 horas)
18	supervisor de seguridad con moto o vehiculo	tiempo completo	2013-02-26	2013-03-15	5	1-3 años	0-50 horas	jornada ordinaria (8 horas)
19	vendedor	tiempo completo	2013-02-26	2013-02-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
20	gypsero con experiencia	tiempo completo	2013-02-26	2013-03-15	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
21	ing. en procesos	tiempo completo	2013-02-26	2013-02-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
22	economista	tiempo completo	2013-02-26	2013-02-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
23	ing. civil con experiencia	tiempo completo	2013-02-26	2013-03-15	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
24	sub-administrador	tiempo completo	2013-02-26	2013-02-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
25	secretaria de ventas	tiempo completo	2013-02-26	2013-02-27	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
26	ingeniero electrico junior	tiempo completo	2013-02-27	2013-03-29	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
27	ingeniero electrico junior	tiempo completo	2013-02-27	2013-02-28	1	1-3 años	0-50 horas	jornada ordinaria (8 horas)
28	mesero	tiempo completo	2013-02-27	2013-03-14	1	sin experiencia	0-50 horas	jornada ordinaria (8 horas)
29	supervisor de local	tiempo completo	2013-02-27	2013-02-28	1	practica en empresa	0-50 horas	jornada ordinaria (8 horas)
30	guardias de seguridad	tiempo completo	2013-02-27	2013-03-08	50	sin experiencia	0-50 horas	jornada ordinaria (8 horas)
31	ing civil, ensayos de construccion	tiempo completo	2013-02-27	2013-03-15	1	practica en empresa	0-50 horas	jornada ordinaria (8 horas)
32	impulsadora	eventual	2013-02-27	2013-03-02	1	7-12 meses	0-50 horas	jornada ordinaria (8 horas)
33	impulsadora	eventual	2013-02-27	2013-03-02	1	7-12 meses	0-50 horas	jornada ordinaria (8 horas)



	I	J	K	L	M	N
	remuneracion	nivelInstruccion	areaEstudios	ciudad	parroquia	sector
ria (8 horas)	$501-$750	tercer nivel	informática software	quito	indistinto	norte
ria (8 horas)	$400-$500	bachiller	recursos humanos/personal	ibarra	indistinto	centro
ria (8 horas)	$400-$500	bachiller	ventas al consumidor	guayaquil	indistinto	centro
ria (8 horas)	$501-$750	tecnológico superior	ingeniería/técnico	pedro vicente maldonado	indistinto	suroeste
ria (8 horas)	$400-$500	tercer nivel	economía/contabilidad	guayaquil	indistinto	sur
ria (8 horas)	$400-$500	tercer nivel	economía/contabilidad	guayaquil	indistinto	centro
ria (8 horas)	$400-$500	bachiller	administración/oficina	guayaquil	indistinto	sur
ria (8 horas)	$400-$500	bachiller	entretenimiento/deportes	quito	indistinto	sur
ria (8 horas)	$400-$500	bachiller	administración/oficina	quito	indistinto	norte
ria (8 horas)	$501-$750	tercer nivel	hotelería/turismo	ibarra	indistinto	centro
ria (8 horas)	$400-$500	tercer nivel	marketing/ventas	guayaquil	indistinto	norte
ria (8 horas)	$400-$500	tercer nivel	marketing/ventas	guayaquil	indistinto	norte
ria (8 horas)	$400-$500	educación básica /básica superior	ventas al consumidor	quito	indistinto	norte
ria (8 horas)	$1001-$1500	tercer nivel	recursos humanos/personal	quito	indistinto	norte
ria (8 horas)	$751-$1000	tercer nivel	ingeniería/técnico	cayambe	indistinto	noroeste
ria (8 horas)	$400-$500	secundaria sin finalizar	recursos humanos/personal	otavalo	indistinto	centro
ria (8 horas)	$400-$500	bachiller	recursos humanos/personal	otavalo	indistinto	centro
ria (8 horas)	$400-$500	técnico superior	ingeniería/técnico	quito	indistinto	sur
ria (8 horas)	$400-$500	secundaria sin finalizar	educación básica/cursos	quito	indistinto	norte
ria (8 horas)	$1001-$1500	tercer nivel	ingeniería/técnico	quito	indistinto	norte
ria (8 horas)	$1001-$1500	tercer nivel	economía/contabilidad	quito	indistinto	centro
ria (8 horas)	$751-$1000	tercer nivel	ingeniería/técnico	quito	indistinto	norte
ria (8 horas)	$501-$750	tercer nivel	administración/oficina	guayaquil	indistinto	norte
ria (8 horas)	$400-$500	tercer nivel	marketing/ventas	guayaquil	indistinto	norte
ria (8 horas)	$501-$750	técnico superior	ingeniería/técnico	lago agrio	indistinto	centro
ria (8 horas)	$501-$750	técnico superior	educación/universidad	lago agrio	indistinto	centro
ria (8 horas)	$400-$500	bachiller	ventas al consumidor	cuenca	indistinto	centro
ria (8 horas)	$400-$500	sin instruccion	hotelería/turismo	quito	indistinto	norte
ria (8 horas)	$400-$500	sin instruccion	educación básica/cursos	quito	indistinto	norte
ria (8 horas)	$751-$1000	tercer nivel	ingeniería/técnico	zamora	indistinto	suroeste
ria (8 horas)	$400-$500	secundaria sin finalizar	marketing/ventas	ibarra	indistinto	centro

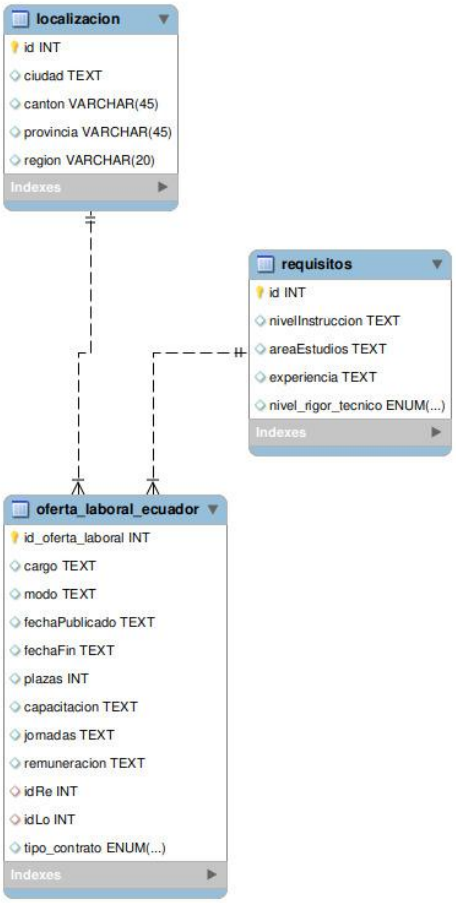
Estos datos, junto con los análisis que se realicen, serán almacenados en un repositorio de GitHub para facilitar su gestión, revisión y colaboración entre investigadores y especialistas interesados en el tema. El repositorio permitirá mantener un registro histórico de los cambios realizados, fomentando así la transparencia y la reproducción de los resultados obtenidos.

Datos Complementarios

Codigos de ofertas laborales de Ecuador:

Datos obtenidos de: [Job Opportunities in Ecuador]

Diseño lógico relacional



Diccionario de datos

ENTIDAD	ATRIBUTOS	TIPO DE DATO	DESCRIPCION
localizacion	id	INT (PK)	Identificador único de la localización.
	ciudad	TEXT	Nombre de la ciudad.
	canton	TEXT	Nombre del cantón.
	provincia	TEXT	Nombre de la provincia.
	region	TEXT	Región del Ecuador.
requisitos	id	INT (PK)	Identificador único de requisitos.
	nivelInstruccion	TEXT	Nivel de educación requerido.
	areaEstudios	TEXT	Área de estudio requerida.
	experiencia	TEXT	Tiempo de experiencia solicitado.
	nivel_rigor_tecnico	ENUM	Grado de exigencia técnica del puesto.

oferta_labora_ecuador	id_oferta_laboral	INT (PK)	Identificador único de la oferta laboral.
	cargo	TEXT	Nombre del cargo ofertado.
	modo	TEXT	Modalidad de trabajo
	fechaPublicado	TEXT	Fecha de publicación de la oferta
	fechaFin	TEXT	Fecha de finalización de la oferta.
	plazas	INT	Número de vacantes disponibles.
	capacitacion	TEXT	Capacitación requerida (si aplica)
	jornadas	TEXT	Tipo de jornada laboral
	remuneracion	TEXT	Sueldo ofrecido.
	idRe	INT (FK)	Clave foránea que referencia a la tabla requisitos.
	idLo	INT (FK)	Clave foránea que referencia a la tabla localizacion.
	tipo_contrato	ENUM	Tipo de contrato ofrecido
RELACION	**TIPO**	**DESCRIPCION**	
oferta_laboral_ecuador-localizacion	Muchos a Uno	Muchas ofertas laborales están ubicada en una localización.	
oferta_labora_ecuador-requisitos	Uno a Muchos	Una oferta laboral tiene un conjunto de requisitos asociados.	

Normalización:

Tabla oferta_laboral_ecuador:

id_oferta_labora	cargo	modo	fechaPublicado	fechaFin	plazas	capacitacion	jornadas	remuneracion	id_requisito	id_localizacion	tipo_contrato
1	desarrollador java jee	tiempo completo	2013-02-22	2013-03-24	1	0-50 horas	jornada ordinaria (8 horas)	$501-$750	1398	1	Indefinido
2	guardias de seguridad con experie...	tiempo completo	2013-02-23	2013-02-28	20	0-50 horas	jornada ordinaria (8 horas)	$400-$500	330	2	Indefinido

Tabla localizacion

id	ciudad	provincia	canton	region
1	quito	pichincha	quito	Sierra
2	ibarra	imbabura	ibarra	Sierra

Tabla requisitos:

id	nivelInstruccion	areaEstudios	experiencia	nivel_rigor_tecnic
1	bachiller	administración/oficina	1-3 años	Medio
2	bachiller	administración/oficina	1-6 meses	Bajo

En este modelo relacional normalizado, se han creado dos tablas adicionales llamadas requisitos y localizacion para eliminar dependencias transitivas y mejorar la organización de los datos en la tabla principal oferta_laboral_ecuador.

La tabla requisitos contiene información detallada sobre los criterios que debe cumplir un postulante, como el nivel de instrucción, área de estudios, experiencia previa y el nivel de rigor técnico requerido. Esta información se vincula con cada oferta laboral mediante una clave foránea (id_requisito), lo que evita la redundancia y permite reutilizar los mismos requisitos en diferentes ofertas.

Por otro lado, la tabla localizacion almacena los datos geográficos relacionados con el lugar donde se ofrece el trabajo, como ciudad, provincia u otras características específicas de la ubicación. Se relaciona con la tabla oferta_laboral_ecuador. a través de la clave foránea (id_localizacion).

Script SQL

Para el Script realizamos la creación de una solo tabla a la cual le agregamos las columnas de datos que utilizamos para este proyecto y se definió el nombre de las variables con su respectivo tipo de dato, para la carga de datos se creó catálogos específicos para las columnas que lo requerían los cuales mediante el comando INSERT van reemplazando los datos de acuerdo con el id de cada columna con el id del catálogo.
Link

Scrip: Proyecto_Final/Script_Final.sql at main · Jean-IG23/Proyecto_Final

Herramientas utilizadas

Sistema operativo Ubuntu

Ubuntu es un sistema operativo libre y gratuito basado en Debian GNU/Linux, que incluye un conjunto completo de software para el uso cotidiano: navegador web, suite ofimática, reproductores multimedia, herramientas de desarrollo, entre otros. Es desarrollado y mantenido por la empresa Canonical Ltd. y cuenta con el apoyo de una gran comunidad de usuarios y desarrolladores.

¿Qué es Zeppelin?

Según Lavalle (2018), Apache Zeppelin es un entorno web interactivo basado en notebooks que facilita el análisis y visualización de datos mediante diversos lenguajes y tecnologías como Spark, Python, Scala, Spark SQL, JDBC, Markdown y Shell, entre otros.

Este entorno presenta múltiples ventajas, entre las que destacan:

- Amplio soporte para intérpretes y plugins que permiten trabajar en diferentes lenguajes de programación.
- Integración intuitiva con diversas fuentes de datos, incluyendo bases de datos no relacionales como Hive o HBase, así como sistemas SQL tradicionales como PostgreSQL o MySQL.

- Funcionalidades colaborativas, que permiten a varios usuarios trabajar simultáneamente en un mismo notebook, con actualizaciones reflejadas en tiempo real.
- Facilidad para importar y exportar notas, lo que mejora la portabilidad del trabajo realizado.
- Visualización clara y directa de los datos, con la posibilidad de exportar los resultados.
- Generación de gráficos dinámicos con múltiples niveles de agregación.
- Es una herramienta gratuita, de código abierto y respaldada por una comunidad activa que contribuye continuamente a su desarrollo.

Configuración del entorno

Instalación de la Máquina Virtual

Permisos de usuario de Ubuntu

Guarda los cambios realizados en el archivo sudoers en nano. Para ello, presiona Ctrl + O, luego presiona Enter para confirmar el nombre del archivo y finalmente, presiona Ctrl + X para salir de nano.

Una vez que hayas completado estos pasos, el usuario "nuevo_usuario" tendrá privilegios de superusuario y podrá usar el comando sudo para ejecutar comandos con privilegios elevados.

Instalación de mysql

PASO	Comando de Consola	Descripción
1	sudo apt update	Actualiza la lista de paquetes disponibles en los repositorios.
2	sudo apt install mysql-server -y	Instala el servidor MySQL y sus dependencias. El -y confirma automáticamente.
3	sudo systemctl start mysql	Inicia el servicio de MySQL.
4	sudo systemctl status mysql	Muestra el estado actual del servicio MySQL (activo, detenido, errores, etc).
5	sudo mysql	Entra al intérprete de comandos de MySQL como usuario root sin

		contraseña (por defecto en Ubuntu).

Instalación de Zeppelin

Paso	Comando de Consola	Descripción
1	sudo apt update	Actualiza la lista de paquetes disponibles.
2	cd /opt	Se mueve al directorio /opt, donde se suelen guardar aplicaciones de terceros.
3	sudo wget https://downloads.apache.org/zeppelin/zeppelin-0.11.0/zeppelin-0.11.0-bin-all.tgz	Descarga Zeppelin (versión 0.11.0) desde el sitio oficial de Apache.
4	sudo tar -xvzf zeppelin-0.11.0-bin-all.tgz	Descomprime el archivo .tgz descargado.
5	cd zeppelin	Entra en la carpeta de Zeppelin.
6	bin/zeppelin-daemon.sh start	Inicia el servidor Zeppelin en segundo plano.
7	bin/zeppelin-daemon.sh status	Verifica que Zeppelin esté en ejecución.
8	xdg-open http://localhost:8080	Abre Zeppelin en el navegador (puerto 8080). También puedes abrirlo manualmente.

Instalación de java

Paso	Comando de Consola	Descripción
1	sudo apt update	Actualiza la lista de paquetes disponibles.
2	sudo apt install openjdk-11-jdk -y	Instala Java JDK 11, necesario para ejecutar Zeppelin.
3	java -version	Verifica que Java se haya instalado correctamente.

Instalación de Python

Paso	Comando de Consola	Descripción
1	sudo apt update	Actualiza los repositorios de paquetes.
2	sudo apt install python3 -y	Instala Python 3 si aún no está instalado.

3	python3 --version	Verifica la versión actual instalada de Python 3.

Instalación de Spark

Paso	Comando de Consola	Descripción
1	sudo apt update	Actualiza los paquetes del sistema.
2	sudo wget https://downloads.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz	Descarga Spark 3.5.0 con soporte para Hadoop 3.
3	sudo tar -xvzf spark-3.5.0-bin-hadoop3.tgz	Descomprime el archivo descargado.
4	sudo mv spark-3.5.0-bin-hadoop3 spark	Renombra la carpeta a simplemente spark.
5	sudo nano ~/.bashrc	Abre el archivo de configuración de la terminal.
6	export SPARK_HOME=/opt/spark export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin	Define variables de entorno para acceder fácilmente a Spark.
7	source ~/.bashrc	Aplica los cambios en las variables de entorno.
8	spark-shell	Inicia la consola interactiva de Spark con Scala (debe abrir sin errores).
9	exit	Sale de la consola interactiva de Spark.

Instalación de Scala

Paso	Comando de Consola	Descripción
1	sudo apt update	Actualiza la lista de paquetes del sistema.
2	sudo apt install scala -y	Instala Scala desde los repositorios oficiales de Ubuntu.
3	scala -version	Verifica que Scala se haya instalado correctamente.

Interprete Zeppelin configuración de mysql, spark , Python

Configuración de Spark en Interprete de Zepellin

Nombre	Valor

SPARK_HOME	/opt/spark
master	local[*]

Configuración de mysql en Interprete de Zepellin

Nombre	Valor
default.driver	com.mysql.cj.jdbc.Driver
default.url	jdbc:mysql://localhost:3306/to_2025
default.user	root
default.password	12345678

Configuración de Python en Interprete de Zeppelin

Nombre	Valor
zeppelin.pyspark.python	/usr/bin/python3
zeppelin.python	/usr/bin/python3
spark.pyspark.python	/usr/bin/python3
spark.pyspark.driver.python	/usr/bin/python3
PYSPARK_PYTHON	/usr/bin/python3

Análisis a realizar

Visualización de consultas

Consulta 1

1. ¿Qué ciudades tienen el mayor promedio de plazas por oferta para trabajos de jornada ordinaria?

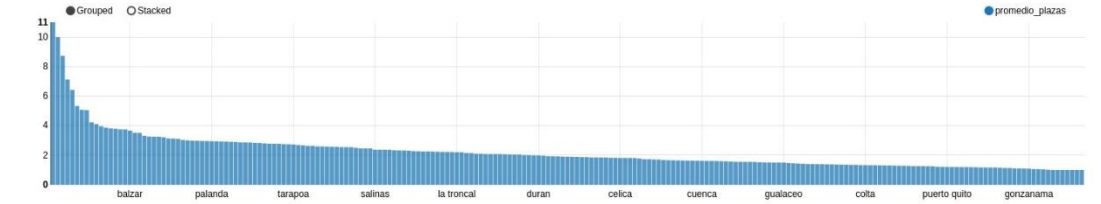
FINISHED

Took 3 sec. Last updated by anonymous at July 22 2025, 11:00:52 AM.

```
val dfJoin = df.join(dfLocalizacion, df("idLo") === dfLocalizacion("id"))

val dfFiltrado = dfJoin.filter(lower(col("jornadas")) === "jornada ordinaria (8 horas)")

val resultado1 = dfFiltrado.groupBy("ciudad")
  .agg(avg("plazas").alias("promedio_plazas"))
  .orderBy(desc("promedio_plazas"))

  z.show(resultado1)
```

SPARK JOB FINISHED

settings




```
dfJoin: org.apache.spark.sql.DataFrame = [id_oferta_laboral: int, cargo: string ... 15 more fields]
dfFiltrado: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id_oferta_laboral: int, cargo: string ... 15 more fields]
resultado1: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ciudad: string, promedio_plazas: double]
```

Took 5 sec. Last updated by anonymous at July 22 2025, 4:34:49 PM. (outdated)

Consulta 2

2. ¿Cuáles son las 10 ciudades con más ofertas sin experiencia?

FINISHED

Took 0 sec. Last updated by anonymous at July 22 2025, 11:14:11 AM.

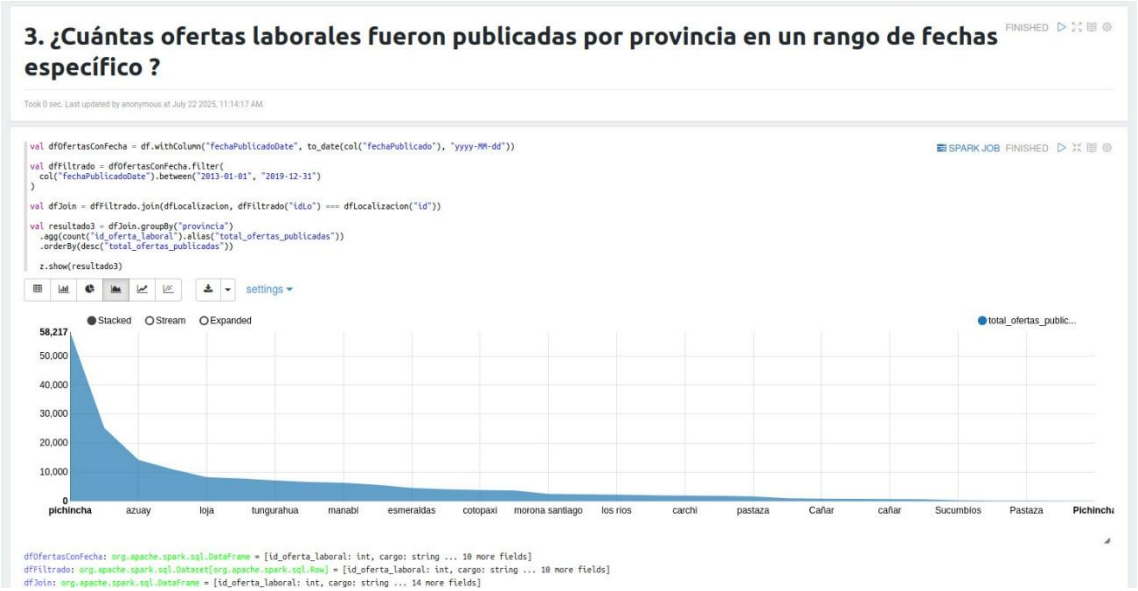
```
val dfJoin = df
  .join(dfRequisitos, df("idRe") === dfRequisitos("id"))
  .join(dfLocalizacion, df("idLo") === dfLocalizacion("id"))

val sinExperiencia = dfJoin.filter(lower(col("experiencia")) === "sin experiencia")

val resultado2 = sinExperiencia.groupBy("ciudad")
  .agg(count("id_oferta_laboral").alias("total_ofertas_sin_experiencia"))
  .orderBy(desc("total_ofertas_sin_experiencia"))
  .limit(10)

  z.show(resultado2)
```

SPARK JOB FINISHED

settings

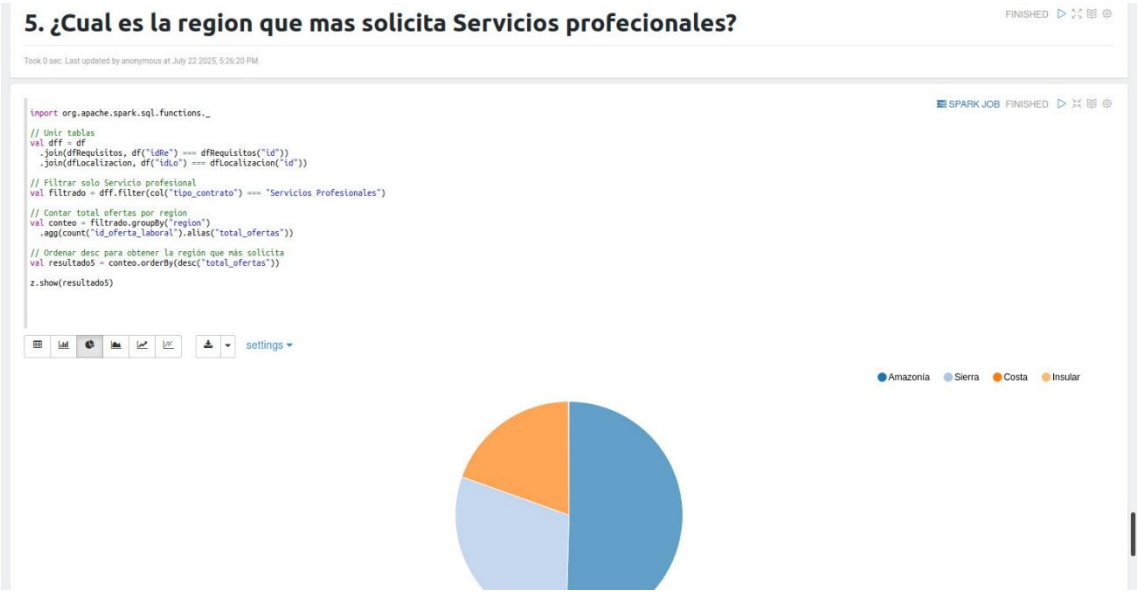

```
dfJoin: org.apache.spark.sql.DataFrame = [id_oferta_laboral: int, cargo: string ... 17 more fields]
sinExperiencia: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id_oferta_laboral: int, cargo: string ... 17 more fields]
resultado2: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [ciudad: string, total_ofertas_sin_experiencia: bigint]
```

Consulta 3



Consulta 4

Consulta 5



Conclusiones

- El uso de Apache Spark mediante la API de Scala permitió procesar grandes volúmenes de datos de forma distribuida y eficiente, facilitando la generación de consultas estructuradas complejas sin necesidad de SQL tradicional.
- Las consultas realizadas revelaron información valiosa sobre la demanda laboral en Ecuador, destacando las diferencias entre regiones en cuanto a servicios profesionales, áreas de estudio y características de vivienda.
- La correcta normalización de las tablas y relaciones permitió una manipulación de datos más limpia y escalable, asegurando integridad referencial y evitando redundancia.
- Apache Zeppelin se consolidó como una herramienta útil para visualizar resultados y presentar de forma dinámica los análisis realizados, lo que mejoró la comprensión de los resultados obtenidos.
- Este proyecto fortaleció competencias clave en programación avanzada, manejo de datos distribuidos, uso de máquinas virtuales, configuración de entornos y visualización de datos, aplicables en escenarios reales del análisis de datos.

Bibliografía

1. Cordero, P. (2020, 29 septiembre). *Como instalar Ubuntu en VirtualBox | Oficina de software libre*. https://osl.ugr.es/2020/09/29/como-instalar-ubuntu-en-virtual-box/
2. Lavalle, A. (2018). Desarrollo de un sistema de Big Data sobre datos abiertos:
3. *Apache Zeppelin | Cloudera*. (2022, 21 noviembre). Cloudera. https://es.cloudera.com/products/open-source/apache-hadoop/apache-

zeppelin.html#:~:text=Zeppelin%20es%20una%20moderna%20plataforma,de%20datos%20cada%20vez%20mayor.

4. Ubuntu. (s.f.). Acerca de Ubuntu. Recuperado el 19 de julio de 2023, de https://ubuntu.com/about

5. wikiHow. (n.d.). *Cómo instalar VirtualBox*. wikiHow. Recuperado de Cómo instalar VirtualBox (con imágenes) - wikiHow

6. OpenAI. (2025). *Respuesta generada por ChatGPT a la consulta "ubuntu sistema operativo definición"* [ChatGPT]. https://chat.openai.com/

7. sparkA (s.p.). Descargar Descargar Spark | Descargar cliente correo electrónico

8. Instituto Nacional de Estadística y Censos (INEC) Ecuador. (s.f.). *Clasificador Geográfico Estadístico del Ecuador*. https://www.ecuadorencifras.gob.ec