

Air Quality Data in India

Bases de Datos Avanzadas





Objetivo

Desarrollar un pipeline ETL automatizado, robusto y escalable para integrar datos de múltiples fuentes confiables, transformarlos en un formato limpio y uniforme (incluyendo la generación de indicadores clave como el Índice de Calidad del Aire - AQI) y cargarlos en un sistema centralizado como SQL Server en Azure para sentar las bases y crear dashboards interactivos en Power BI, permitiendo el análisis de tendencias espaciales y temporales en la calidad del aire

Tecnologías Utilizadas



Databricks

Ejecución del proceso
ETL



Azure

Integrar el
almacenamiento de datos



SQL Server

Almacenamiento y
consultas de datos

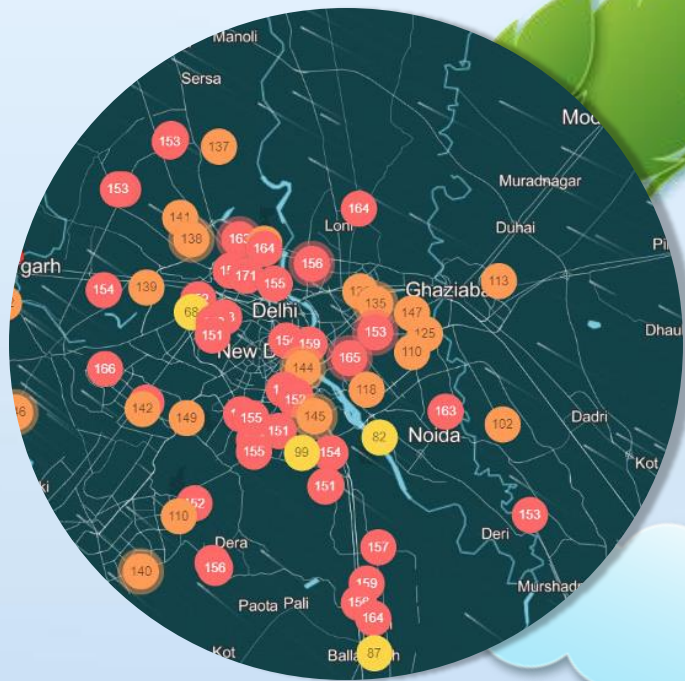


Power BI

Reportes y visualizaciones
interactivas

01

Solución ETL



Extracción

Recopilación y utilización de Databricks para establecer conexiones y poder realizar descargas masivas de datos



Proceso



Recolección

Configuración de scripts automatizados para la recolección de datos en intervalos periódicos.



Validación

De las estructuras de datos y formatos para garantizar la integridad de la información.



Optimización

Uso de pipelines paralelos para optimizar el tiempo de descarga y evitar cuellos de botella.

Transformación

La transformación se enfocó en limpiar y preparar los datos para su análisis.





Proceso



- Eliminación de registros duplicados y valores faltantes mediante técnicas de imputación y filtrado.
- Estandarización de columnas, como conversiones de formatos de fecha y normalización de unidades de medida.
- Creación de nuevas variables, como el Índice de Calidad del Aire (AQI), mediante cálculos basados en concentraciones de contaminantes.
- Implementación de reglas de negocio para garantizar la coherencia de los datos entre diferentes fuentes.
- Uso de Databricks para procesar grandes volúmenes de datos en paralelo, aprovechando la infraestructura escalable de Spark.



Carga

Los datos transformados se cargaron en SQL Server alojado en Azure, lo que permitió su almacenamiento seguro y consultas eficientes.

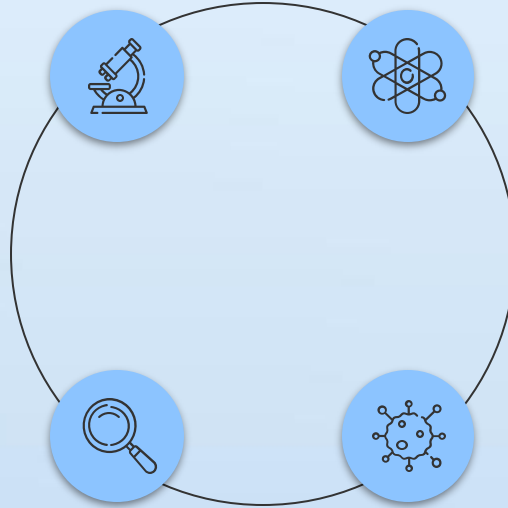


Proceso

Verificación de la integridad de los datos cargados mediante pruebas de validación y conteo de registros

Conexión directa entre SQL Server y Power BI para habilitar la creación de dashboards interactivos en tiempo real

3.



1.

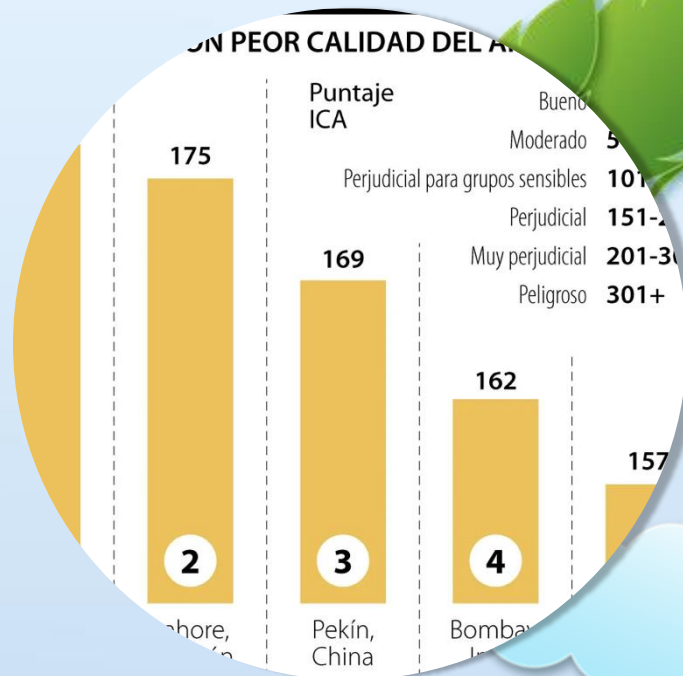
Diseño de tablas optimizadas para el almacenamiento de datos históricos y actuales.

2.

Configuración de integraciones entre Databricks y SQL Server para una carga fluida y automatizada.

02

Resultados





Implementación Técnica Exitosa





Proceso

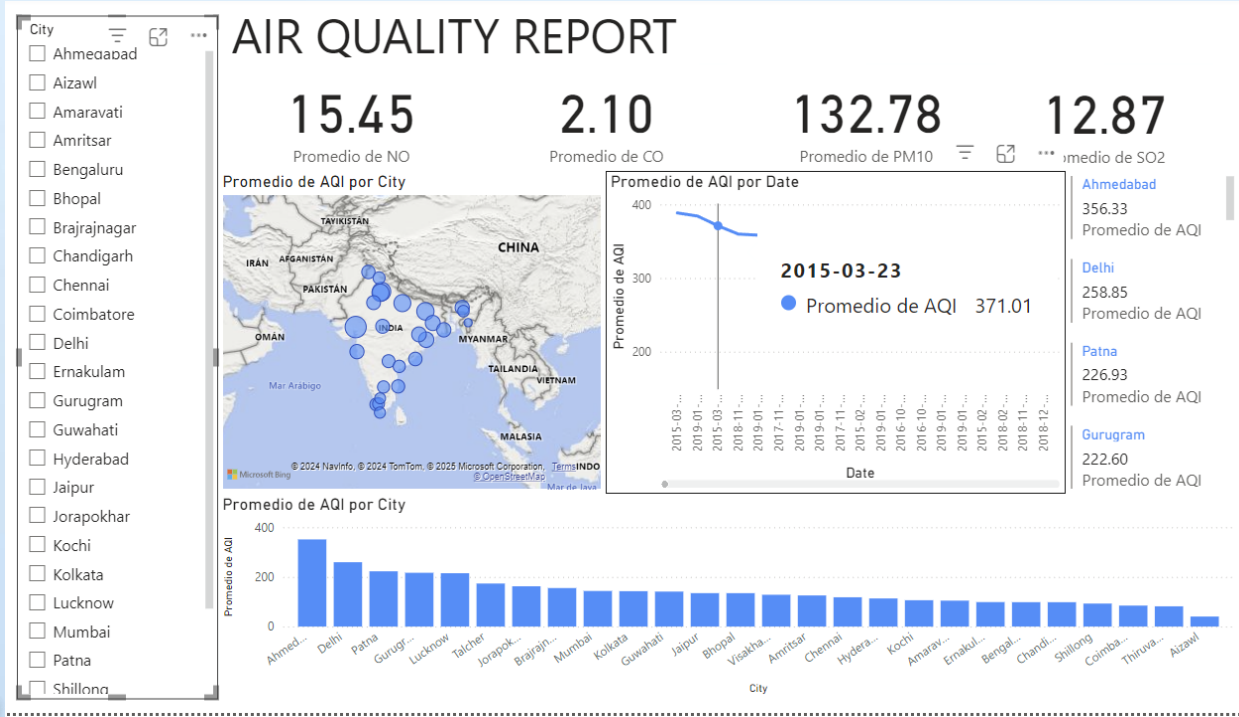
- Se logró implementar un pipeline automatizado capaz de procesar más de 10 millones de registros relacionados con la calidad del aire. Esto incluyó la extracción, limpieza, transformación y carga de datos con un rendimiento eficiente gracias al uso de Databricks y Azure.
- 
- 

Creación de Dashboards

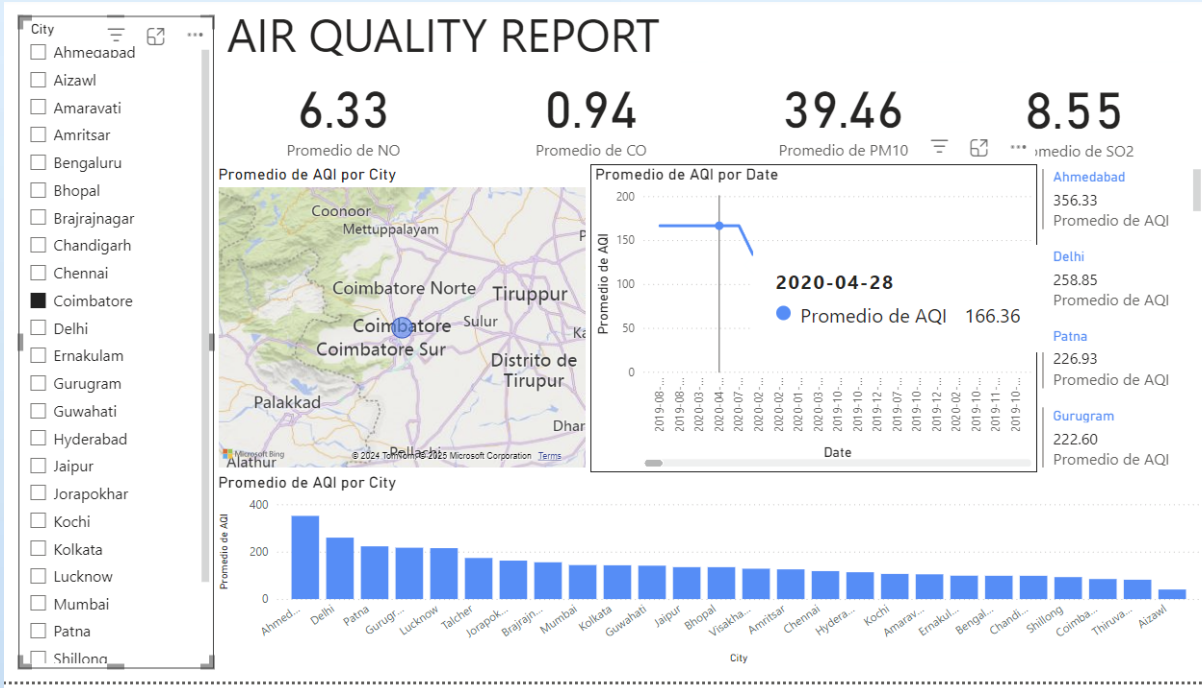
Utilizando Power BI, se generaron visualizaciones dinámicas, facilitando la toma de decisiones



Creación de Dashboards



Creación de Dashboards



Hallazgos Clave del Análisis



Factores que Afectaron





Optimización e Impacto



Se identificó lo siguiente



- La integración de Databricks con Azure permitió reducir el tiempo de procesamiento de datos en un 30%, comparado con métodos tradicionales. Esto garantiza la escalabilidad del sistema para manejar mayores volúmenes de datos en el futuro.
 - Las visualizaciones y análisis generados pueden ser utilizados por tomadores de decisiones y entidades gubernamentales para diseñar políticas más efectivas de mitigación de contaminación, priorizando áreas críticas identificadas en los reportes
- 
- 

03

Conclusiones y Recomendaciones



**El proyecto ETL sobre
calidad del aire en India
demostró ser una solución
eficiente y efectiva para
procesar y analizar
grandes volúmenes de
datos**



Conclusiones



La implementación de Databricks y SQL Server en Azure facilitó la integración, procesamiento y almacenamiento seguro de los datos.



El pipeline ETL automatizado mejoró significativamente la calidad y disponibilidad de la información, proporcionando una base confiable para el análisis y la toma de decisiones.



Las visualizaciones generadas en Power BI permitieron una comprensión clara de las tendencias en los datos, destacando su potencial para influir en la formulación de políticas públicas.

Recomendaciones

Ampliar alcance de datos

Incorporar variables meteorológicas



Automatización

Configurar pipelines más robustos que se actualicen en tiempo real



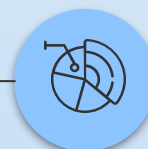
Colaboración Multidisciplinaria

Involucrar a expertos en salud pública y políticas ambientales



Análisis predictivo

Desarrollar modelos de machine learning para predecir la calidad del aire en función de tendencias históricas



¡Gracias!

Apéndices:

- Captura del dashboard creado en Power BI.
- Código fuente del pipeline ETL.
- Ejemplo de consulta SQL utilizada en Hive.

Elaborado por:

- Gutierrez Ramirez Alana Sofia
- Reyes Maldonado Oscar Romario
- Sánchez García Miguel Alexander
- Villagran Salazar Diego

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution

