



Flujo de trabajo en ciencia de datos: Fases, roles y oportunidades laborales


Aprender las diferentes fases de los datos y los diferentes empleos que existen en cada una de ellas.

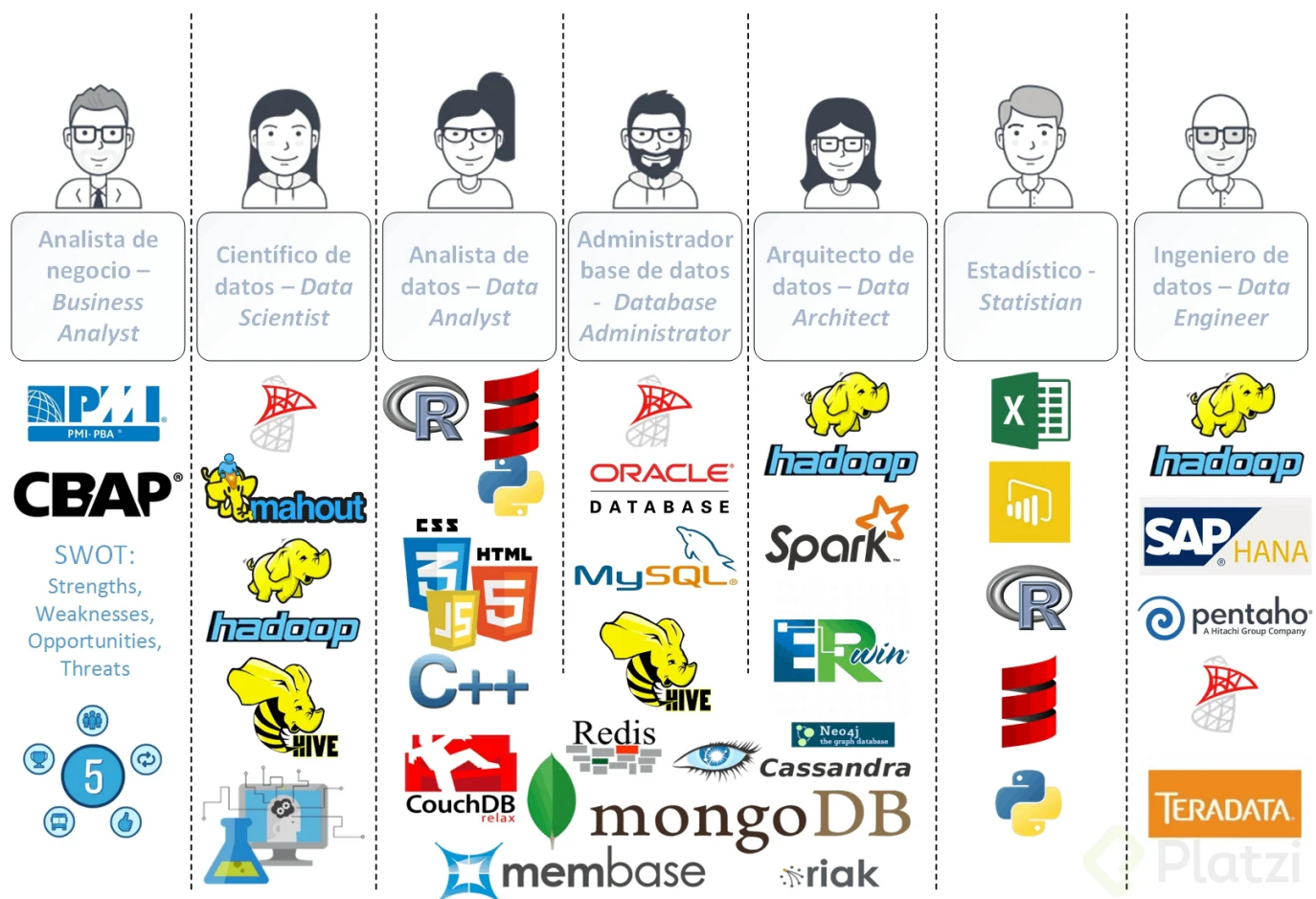
Roles en datos:

 ****Ingeniería de datos:** - ** Base de datos - ETL's / API's - SQL + NoSQL

 ****Analista de Business Intelligence:**** - Extracción y Dashboards - Automatización - SQL y Excel

 ****Data Scientist:**** - Machine Learning - Modelos Estadísticos - R y Python

 ****Data Translator:**** - Data Scientist y Decision Makers - Destiladores de data - Expertos necesidad de negocio



¿Qué es un ETL?

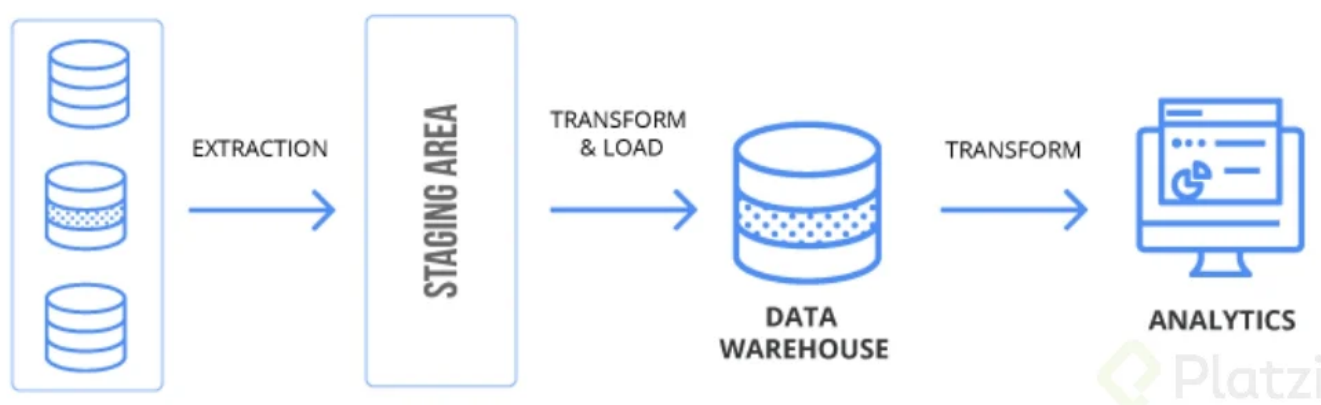
Los datos llegaron para quedarse. Para nadie es un secreto que una buena parte de nuestra economía se basa en datos y los encontramos en cualquier parte, incluso el simple de hecho de estar revisando nuestro smartphone para hablar con otras personas, buscar algo en google, jugar o pasear por las redes sociales. Todo esto genera una gran cantidad de datos minuto a minuto y así en cada dispositivo o cada organización que existe. Piensa en cuántos datos pueden pasar por un banco en un minuto o en una organización que dependa de sensores como la industria petrolera o las startups.

Pero lo interesante acá es el valor que se genera con estos datos gracias a la analítica de datos, visualización, creación de modelos con machine learning, ayuda en toma de decisiones y demás.

Pero ¿todos los datos están listos para ser analizados? La respuesta es NO, la gran mayoría de los datos nacen con una estructura que no es adecuada para el análisis de datos, vienen en formatos desestructurados como JSON o XML, con valores nulos, datos errados, caracteres inválidos, registros duplicados y demás problemas a los que los científicos de datos e ingenieros de datos nos enfrentamos día a día. Pero calma, a pesar de este caos hay una solución: un proceso de ETL.

Extract, Transform, Load (ETL)

ETL por sus siglas en ingles representa **extracción, transformación y carga**. Básicamente consiste en **“Extraer” los datos crudos desde su origen (Source), “Transformarlos” según nuestras necesidades de analítica o la estructura que deseamos y “Cargarlos” a una base de datos orientada a procesos analíticos (Target).**



Extracción:

Acá nace la magia. En este punto es cuando, valga la redundancia, extraemos los datos. Vale la pena resaltar que nuestro origen de datos o “source” puede contener múltiples fuentes de datos. Imagina que requieres extraer la información de una base de datos en PostgreSQL, otra base de datos en Oracle y un archivo CSV que te genera a diario el área de contabilidad. Es muy importante que en este punto tengas en cuenta cuál es el formato y las características de nuestros datos, pues esto nos indicara la mejor manera de extraerlos.

Un ejemplo de esto es si quieres consultar millones de registros de una base de datos, posiblemente hacerlo en un solo “select” estresara el motor de base de datos y tu proceso ETL empezará a tener inconvenientes, lo mejor seria ejecutar este proceso en pequeños lotes incrementales hasta terminar el total de los datos. El éxito de una buena extracción es que cause el menor impacto en el origen de los datos.

Teniendo eso en cuenta puedes hacer tus extracciones de dos formas:

- **Totales:** cada ejecución de extrae en un único llamado la totalidad de tus datos a procesar .
- **Incremental:** el ETL extrae los datos en pequeños lotes múltiples veces. Un ejemplo es un ETL que se ejecuta diariamente, pero solo consulta los datos del día anterior a su ejecución para trabajar con ellos.

Transformación:

En este punto aplicaremos las reglas que tu negocio demanda para realizar un buen proceso de analítica, estas reglas suelen incluir procesos como:

- Filtrar filas por ciertas características.
- Eliminar duplicados.
- Transformar datos (por ejemplo, si el país es Colombia, reemplázalo por 1) México, 2) Perú, 3) Ecuador, etc.).
- Calcular datos nuevos (por ejemplo, con la fecha de nacimiento calcular la edad).
- Agrupar datos (máximo valor, mínimo valor, promedios, conteos, etc.).
- Unir o combinar datos de distintas fuentes.
- Pivotar las tablas.
- Dividir columnas (nombre completo se puede transformar en primer nombre, segundo nombre, primer apellido y segundo apellido).

Y estas solo son algunas de las transformaciones más comunes, este proceso depende mucho de lo que quieras obtener de los datos una vez la cargues al destino o "Target".

Tal vez en este punto te preguntes "¿y en dónde se realizan estas transformaciones? ¿En el "Source" o en el "Target"?". Y la respuesta es: ninguna de las dos. Estas transformaciones se realizan en el área de "Staging", un repositorio temporal para procesar estos datos, funciona por medio de tablas o archivos planos dependiendo de la herramienta de ETL que uses, una vez los datos pasan al destino, este repositorio temporal es eliminado.

Carga:

Este es el proceso final de nuestro ETL. Nuestros datos deben estar transformados y listos en el área de staging, así que ahora debemos proceder a cargarlos a nuestra base de datos de analítica. Esta base de datos comúnmente es un datawarehouse en donde conviven distintos repositorios de datos de una manera no normalizada (como en una base de datos relacional) y con una estructura lista para realizar análisis de datos.

En este punto también debemos tener en cuenta qué motor corre nuestra base de datos de destino, pues dependiendo de esto existen distintas formas más eficientes de cargar la información. Por ejemplo, una base de datos en Redshift es mucho más eficiente haciendo un "Copy" basados en un archivo plano y no realizando tareas de "insert" registro a registro, bases de datos como Oracle son muy buenas haciendo "Bulk Collect" de inserción de datos.

Una vez realizado el cargue de datos tus tablas están listas para ser consultadas en cualquier proceso de analítica y darle valor a tu organización basados en datos.

Sácale provecho

En la economía actual la demanda de datos es muy alta y una base muy importante son los procesos recolección e ingesta de datos y acá un ETL es el rey del baile. Te invito a que aprendas más de este tema pues la clave de un buen proceso de analítica está en el origen de los datos, si basura entra, basura sale, un buen proceso de ETL es fundamental para darle valor a tus datos.

Te recomiendo seguir estos cursos para saber más acerca del tema:

- Curso de Ingeniería de Datos con Python: <https://platzi.com/clases/ingenieria-datos/>
- Fundamentos de Bases de Datos: <https://platzi.com/clases/bd/>
- Curso de Big Data en AWS: <https://platzi.com/clases/big-data/>

Cuéntame en los comentarios si te quedo alguna duda al respecto o como te puedo ayudar. Recuerda: Nunca Pares De Aprender.

RETO EJERCICIO:

Flujo de trabajo en ciencia de datos: fases, roles y oportunidades laborales

¿Cuál es tu rol? Piensa en dónde te proyectas, qué área(s) de la ciencia de datos te llaman la atención, cuáles herramientas te llamaron la atención y qué propósito quieres cumplir como data scientist.

Quiero ser... Data scientist y/o Data Translator (personas del área de negocio)