*Stack Tecnológico I*

Como bien sabemos, un científico de datos es un profesional dedicado exclusivamente a analizar e interpretar grandes bases de datos. Para ello, debe aprender a utilizar múltiples herramientas que estaremos desarrollando de largo de esta sesión! Empecemos hablando del mundo de las Bases de Datos.

**Bases de Datos**

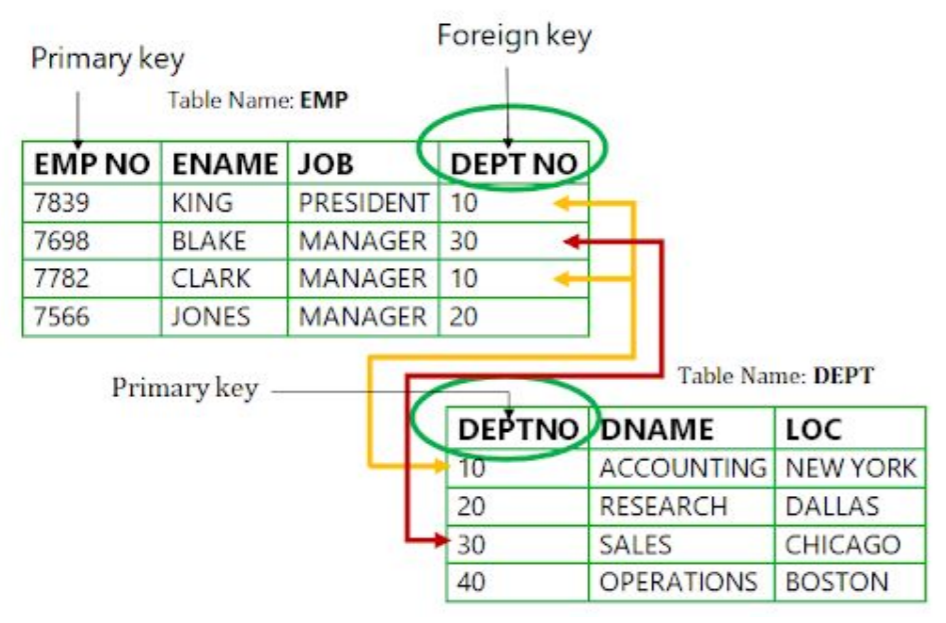
Como ya hemos comentado en sesiones anteriores del curso, una Base de Datos (DB – Database), es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso”.

En términos generales, podemos dividir el mundo de las Bases de Datos, en dos grandes apartados:

* Bases de Datos Relacionales.
* Base de Datos no Relacionales – No SQL.

**Bases de Datos Relacionales**

Responden al Modelo de Datos relacional propuesto por Edward Frank Codd en 1970, tal cual como podemos observar en la siguiente imagen:



Algunos de los Motores de Bases de Datos Relacionales más populares son:

**Microsoft SQL Server** es un sistema de gestión de bases de datos relacionales (RDBMS) que admite una amplia variedad de aplicaciones de procesamiento de transacciones, inteligencia empresarial y análisis en entornos informáticos corporativos.



**MySQL** es un sistema de gestión de bases de datos relacional, desarrollado bajo una licencia dual: Licencia pública general / Licencia Comercial por Oracle Corporation. A su vez también, es una de las bases de datos más populares en general junto a Orablce y SQL Server.



**PostgreSQL** es un sistema gestor de bases de datos relacionales, orientado a objetos, multiplataforma y open source. Está desarrollado desde 1996 por la comunidad de SGBD POSGRES.



**Otras**: Oracle Database, IMB DB2, Access, SQL Cloud y más!.

**Tecnologías de Bases de Datos No Relacionales**

En los últimos años, la cantidad de datos digitales que se genera el mundo se ha multiplicado. las redes sociales y el internet han potenciado muchísimo este contexto.

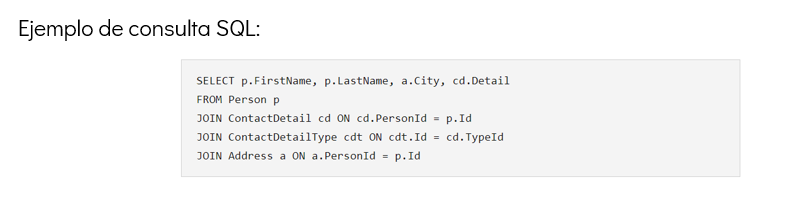
Resulta importante mencionar, que con el surgimiento de las bases de datos relacionales las empresas encontraron el aliado perfecto para cubrir sus necesidades de almacenamiento, disponibilidad, copiado de seguridad y gestión de sus datos. Pero debido a las tendencias actuales en lo que respecta a la gran cantidad de datos, este tipo de sistemas han comenzado a experimentar dificultades técnicas.

Los tipo de datos masivos, reciben el nombre de BigData, y el tipo de tecnología que ha surgido para tratar de poner solución a muchos de estos problemas se conoce como NoSQL.

El término NoSQL fue inicialmente utilizado en el año 1998, y fue para denominar una base de datos relacional que no utilizaba el lenguaje SQL para funcionar. A partir de este acontecimiento, el término fue rescatado en 2009 en unas charlas por defensores de las bases de datos no relacionales.

Esencialmente las Bases de Datos NoSql buscan:

* Evitar la complejidad innecesaria.
* Conseguir un alto rendimiento.
* Escalabilidad horizontal y hardware de bajo coste.
* Transformar la famosa frase: “One size fit’s it all”.





Algunas de las Bases de Datos NoSql más populares son:

**MongoDB** es una base de datos orientada a documentos. Esto quiere decir que en lugar de guardar los datos en registros, guarda los datos en documentos. Estos documentos son almacenados en BSON, que es una representación binaria de JSON.



**Apache Cassandra** se trata de un software NoSQL distribuido y basado en un modelo de almacenamiento de «clave-valor», de código abierto que está escrita en Java. Permite grandes volúmenes de datos en forma distribuida.



**Redis** es un motor de base de datos en memoria, basado en el almacenamiento en tablas de hashes (clave/valor) pero que opcionalmente puede ser usada como una base de datos durable o persistente.



**Neo4j** es una base de datos open-source orientada a grafos escrita en java. Con este tipo de base de datos no sql puedo guardar información en formato de nodos y relacionales.



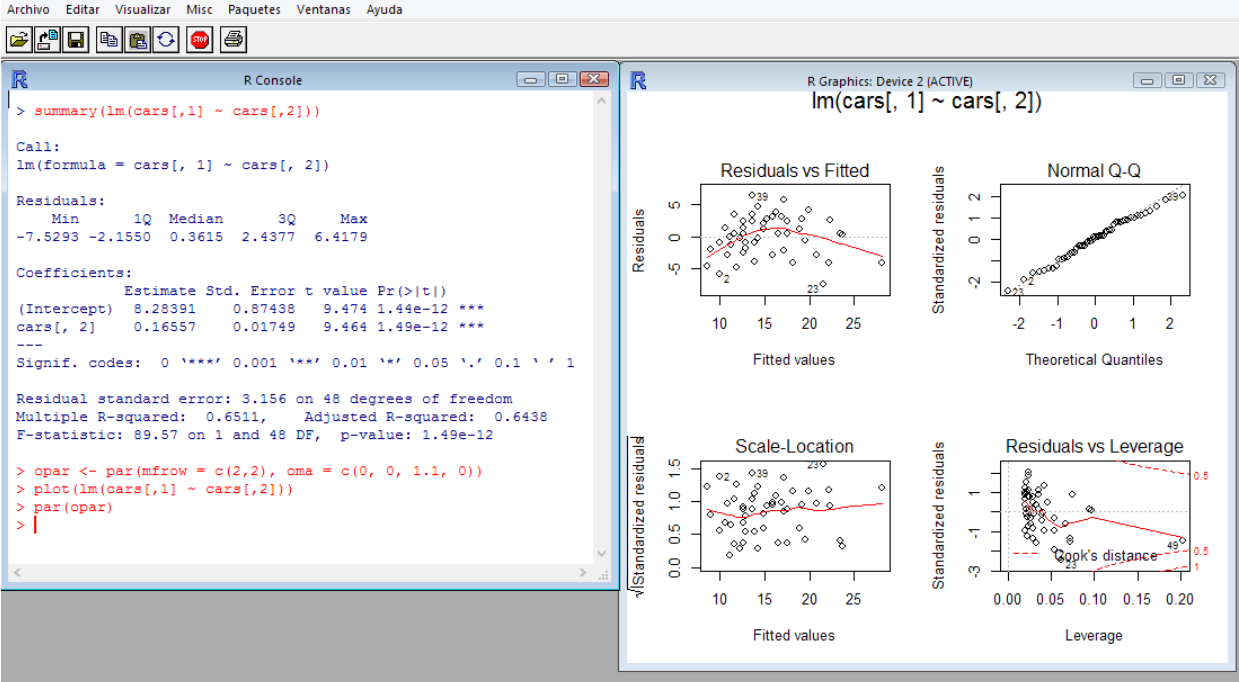
**Otras**: Hbase, CouchDB, No Sql Cloud y más!.

**Lenguajes de Data Science**

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Se trata de uno de los lenguajes de programación más utilizados en investigación científica. R proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, etc), generación de gráficos y es altamente extensible.



**¿Cómo se visualiza R?**



Un poco rústico verdad?

La realidad, es que la interfaz gráfica de R, no es realmente muy atractiva e intuitiva. Es por ello, que para nuestras clases prácticas trabajaremos con RStudio, el IDE de R.

**¿Qué es un IDE?**

Un entorno de desarrollo integrado o entorno de desarrollo interactivo, en inglés Integrated Development Environment, es una aplicación informática que proporciona servicios integrales para facilitarle al desarrollador o programador el desarrollo de software. En este caso en particular RStudio, nos brindará una interfaz mucho más cómoda y amigable para trabajar con R.



**R STUDIO**

Al utilizar RStudio, tendremos múltiples ventajas. Las más importantes a mencionar son:

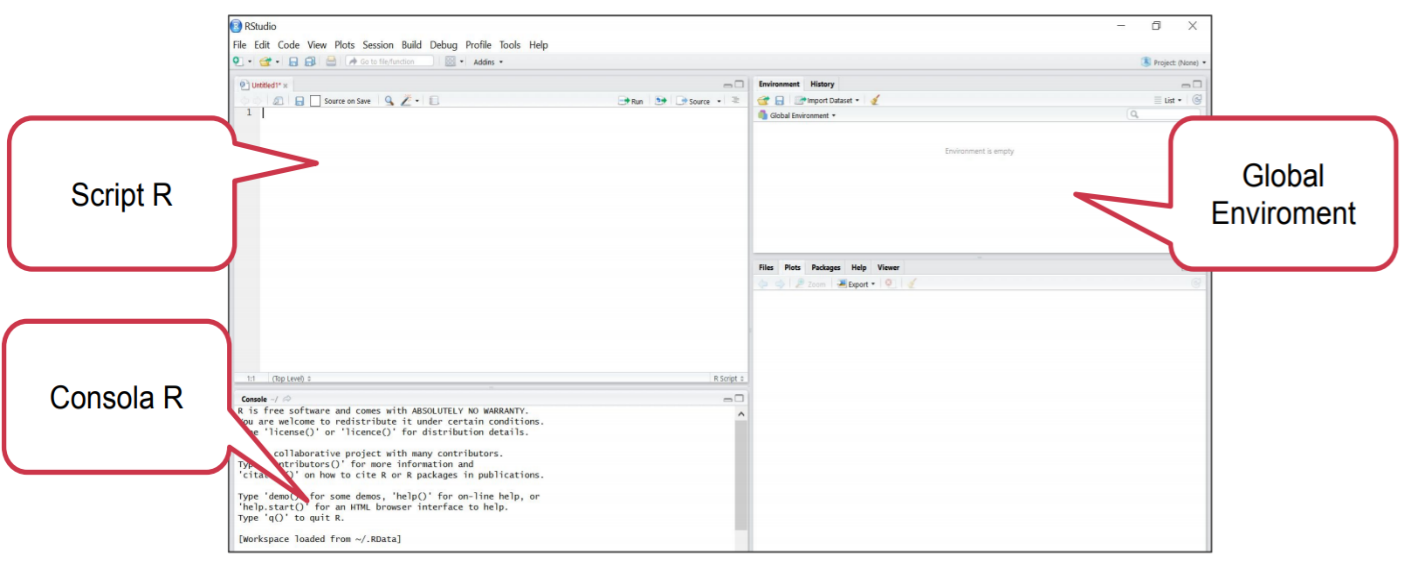
1. Autocompletado.

2. Reconocimiento de sintaxis de programación.

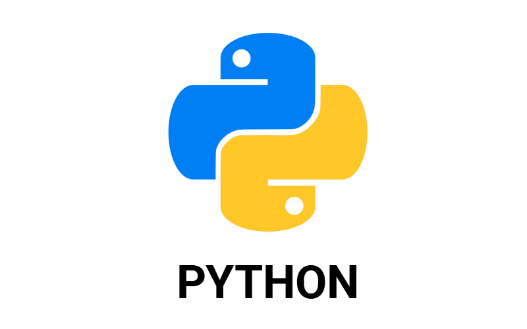
3. Depurador de errores.

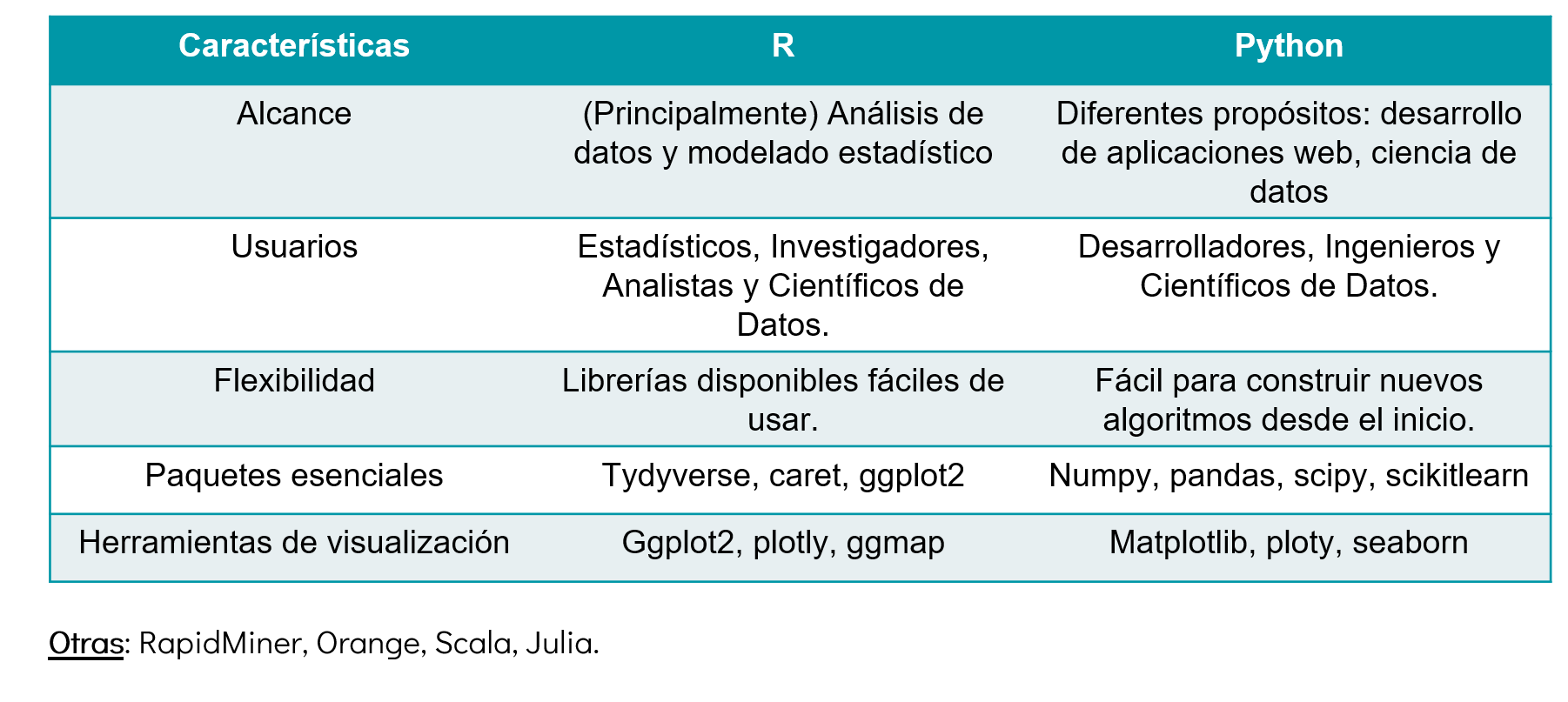
4. Manual de usuarios y ayuda en línea.

¿Cómo se visualiza RSTUDIO?



Python es un lenguaje de programación poderoso y fácil de aprender. Puede ser clasificado como un lenguaje interpretado (ejecuta las instrucciones a medida que las va leyendo) y de alto nivel. Python fue creado a finales de los años 80, por un programador holandés llamado Guido van Rossum, quien sigue siendo aún hoy el líder del desarrollo del lenguaje.





**Herramientas de Visualización de Datos**



Power BI es un conjunto de herramientas y servicios de Inteligencia de Negocios, que permite conectarse a diferentes orígenes de datos, para ser analizados, visualizarlos y compartirlos con toda la organización y clientes. Además, se compone de varias aplicaciones y servicios (Versión desktop, mobile y el servidor).



“Tableau, es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios. La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos”.



Tableau funciona a través de 3 medios principales:

* Escritorio (Tableau Desktop).
* Servidor (Tableau Server).
* En línea (Tableau Online).

Además de estas tres herramientas principales, Tableau integra otras herramientas adicionales para proporcionar una experiencia lo más completa posible a los usuarios:

* Tableau Mobile.
* Tableau Public.
* Tableau Prep.

“QlikView es una herramienta de Business Intelligence, que permite recolectar datos desde diferentes orígenes, basados en ERP, CRM, data warehouses, bases de datos SQL, datos de Excel, etc.”



Google Data Studio es una herramienta de visualización de datos y creación de cuadros de mando. El objetivo de la herramienta, es permitir un análisis de los datos de manera visual de forma que sea más fácil e inmediato obtener resultados.



QuickSight es un servicio rápido de análisis de negocios basado en la nube que facilita la creación de visualizaciones; la realización de análisis ad-hoc y la obtención rápidamente de información de negocios basada en datos.



“IBM Cognos Analytics, es una suite de inteligencia empresarial integrada basada en web de IBM. Proporciona un conjunto de herramientas para informes, análisis, cuadros de mandos y seguimiento de eventos y métricas.”



MicroStrategy es una organización que provee software de reporteo, análisis y monitoreo integrados que permite a las empresas analizar datos almacenados de la empresa y de todas las áreas para tomar mejores decisiones.



Otras herramientas open source para Visualización de Datos pueden ser por ejemplo:

* Shiny de R.
* Flourish en su versión FREE.
* Apache Superset, etc.



Resulta importante mencionar, que existen muchísimas herramientas orientadas a la gestión de comunicación, versionado de datos y trabajo en equipo que tenemos que conocer como Científicos de Datos. Algunas de las más importantes son:

* Slack.
* Trello.
* G - Suite o M365.
* Mural.
* Jamboard.
* Whatsapp o Telegram.
* GitHub, GitLab
* Zoom, Meet, Teams, Jitsi, etc.

