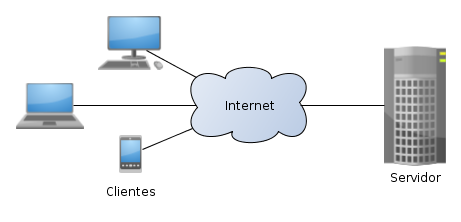
**Stack Tecnológico II**

Cómputo en la Nube

En la actualidad, el Cloud Computing es una tendencia en alza especialmente dentro del mundo de la Analítica de Datos y como Data Scientist resulta realmente muy importante que podamos entender qué es el Cloud, sus características y cuáles son los diferentes servicios que podemos utilizar en la nube para complementar nuestro Stack Tecnológico como Científicos de Datos.

Arquitectura Cliente – Servidor



Puntos clave de la arquitectura:

* Cliente.
* Servidor.

¿Qué es el Cloud Computing?

“El Cloud Computing o Cómputo en la nube, es una nueva tecnología que nos permite acceder mediante un sistema remoto, al software, al procesamiento de datos y al almacenamiento de archivos.”



Tipos de Nubes disponibles:

* Nube pública.
* Nube privada.

Ventajas que aporta el Cloud Computing al Big Data

Existen muchas ventajas que podemos mencionar acerca de los beneficios de implementar Cloud en nuestras organizaciones las más relevantes son las siguientes:

* Abaratamiento de costes.
* Inmediatez.
* Capacidad de proceso.
* Concurrencia.
* Seguridad.
* Importante:

Importante: “El Cloud Computing, permite la aceleración y gestión de procesos computacionales, haciendolos más eficientes”.

Consideraciones importantes para analizar

Ventajas In - House:

* Customización: Con nuestra propia infraestructura de redes y de datos podemos obtener una mayor customización de recursos por ejemplo en base a los requerimientos de la organización.
* Conocimiento: El conocimiento de los sistemas y datos reside únicamente in-house. Esto puede ser preferible para compañías con requerimientos muy específicos de seguridad.
* Sin costos mensuales: Usualmente, los planes de deployment on-premise se estructuran en planes anuales o multi-anuales, potencialmente eliminando la necesidad de costos mensuales.

Desventajas In - House:

* Elasticidad: Falta de elasticidad para el uso de mas recursos en caso de ser necesario dado un pico de demanda.
* Costos de Infraestructura: Los costos de infraestructura para montar un datacenter propio son mucho más superiores que si adquirimos una licencia con algún proveedor de servicios Cloud.
* Personal altamente capacitado: Si quisiéramos tener nuestra infraestructura In – House deberemos contar con personal altamente capacitado en temas de Seguridad e Infraestructura informática.

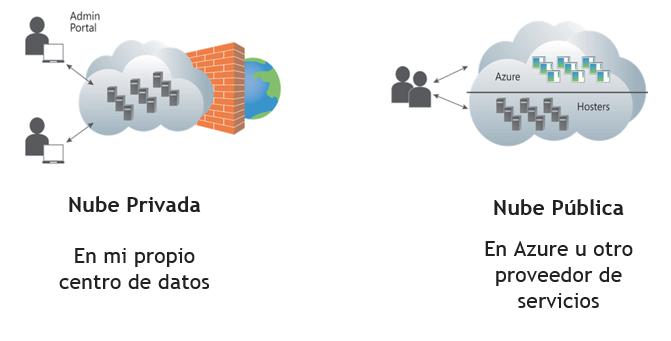
Modelos de Servicios en la Nube:

* Infraestructura como servicio (IaaS).
* Plataforma como servicio (PaaS).
* Software como servicio (SaaS).



Tipos de Nubes

* Nube Privada.
* Nube Pública.

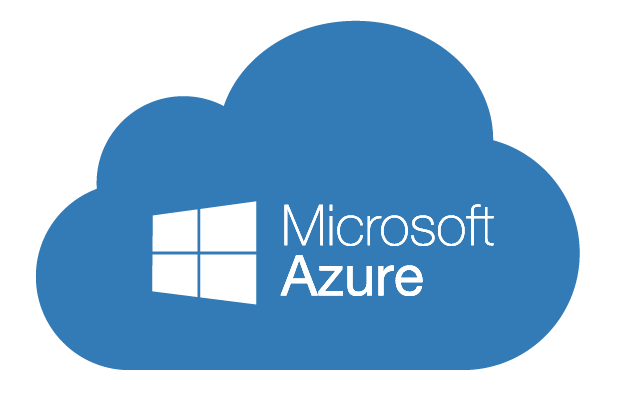


Proveedores de Servicios en la Nube

“Quizás Amazon Web Services, es una de las plataformas de servicios en la nube más conocidas, proporciona una variedad de servicios de infraestructura, tales como almacenamiento, redes, bases de datos, servicios de aplicaciones, potencia de cómputo, mensajería, inteligencia artificial, servicios móviles, seguridad, identidad y conformidad, entre otros. ¿Cuál es su principal ventaja?, sin dudas la experiencia en el mercado y el servicio de atención al cliente que ofrece”.



“Microsoft Azure, proporciona una gama de servicios en la nube, incluidos servicios vinculados con la computación, analítica, almacenamiento y redes. Los usuarios pueden elegir entre estos servicios para desarrollar y escalar nuevas aplicaciones, o ejecutar aplicaciones existentes. ¿Cuál es su principal ventaja?, la rapidez”.



“Gooogle Cloud Platform, también conocida por la nomenclatura GCP, es una suite que contiene diversos servicios que funcionan en la misma infraestructura que utiliza Google de manera interna, por ejemplo con servicios como Youtube o Google Search”. ¿Cuál es su principal ventaja?, la seguridad.



¿Ahora bien, cuál elegir?

Resulta relevante destacar, que las distintas soluciones anteriormente mencionadas, ofrecen una serie de ventajas, similitudes y ciertas diferencias. La elección de una u otra tecnología, deberá realizarse en función de las necesidades computacionales de la organización y acompañado por la asesoría de un profesional capacitado en la temática. Quizás, una buena recomendación antes de tomar una decisión definitiva al respecto, sería tomar una prueba gratuita de cada una de las herramientas, para vivenciar de esa manera la experiencia de utilizar cada una de ellas.



Big Data

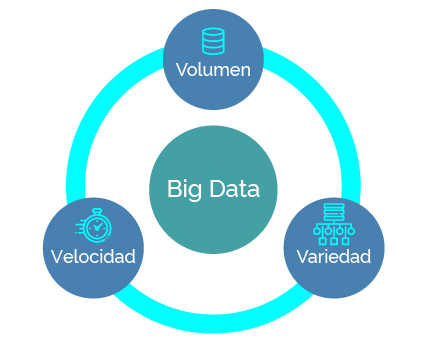
Fundamentos Generales

“Big data es una colección de grandes volúmenes de datos, complejos y muy difíciles de procesar, a través de herramientas de gestión y procesamiento de datos tradicionales”.

Impulsores del Big Data: Sin dudas existen múltiples catalizadores del Big Data, podemos mencionar:

* IoT.
* Redes Sociales.
* El internet.
* Tics.
* Cloud Computing.

Las 3 V del Big Data



Principales Herramientas de Big Data

Hadoop

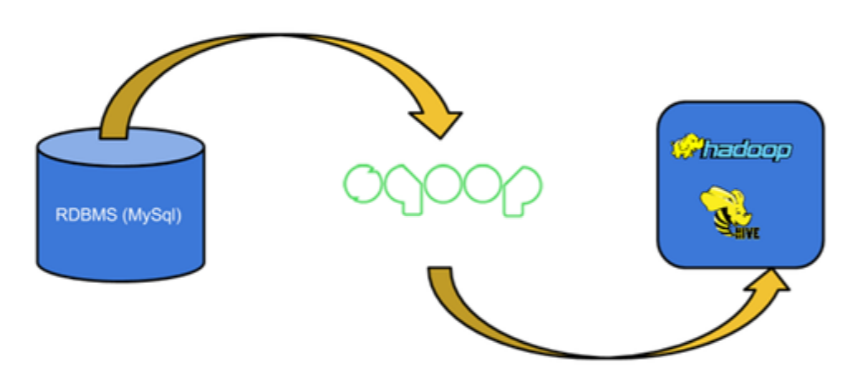
Sin dudas el primer software que se utilizo para Big Data fue Apache Hadoop pero ¿Qué es Hadoop? Es un framework implementado en Java que permite el almacenamiento y procesamiento distribuido de grandes conjuntos de datos estructurados, semi-estructurados y no estructurados. Está diseñado para trabajar en clústers y tiene una alta tolerancia a fallas.

¿Para qué sirve Hadopp?

* Almacenar archivos de manera distribuida.
* Procesar archivos de manera distribuida.

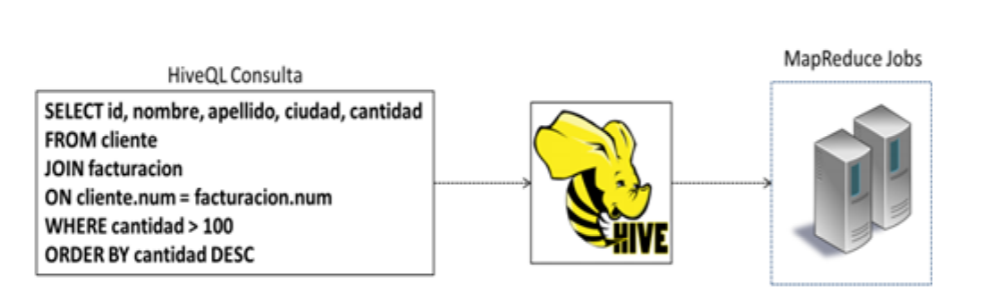
Apache Sqoop

Es una herramienta que está dentro de proyecto Apache, y además forma parte del ecosistema Hadoop. Sqoop nos permite la transferencia de datos entre un RDBMS como MySQL o oracle y HDFS.



Hive

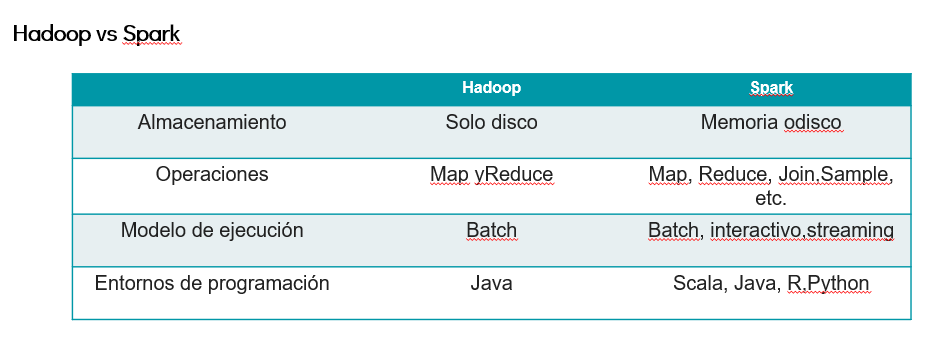
¿Qué es Hive? Es un software que facilita leer, escribir y manejar largos conjunto de datos, que residen en un almacenamiento distribuido, usando SQL. Otro concepto importante para comentar es lo que se conoce como MapReduce, básicamente MR es un modelo de programación utilizado en Hadoop que facilita e trabajo con grandes volúmenes de datos distribuidos a través de múltiples máquinas dentro de un clúster.



Spark

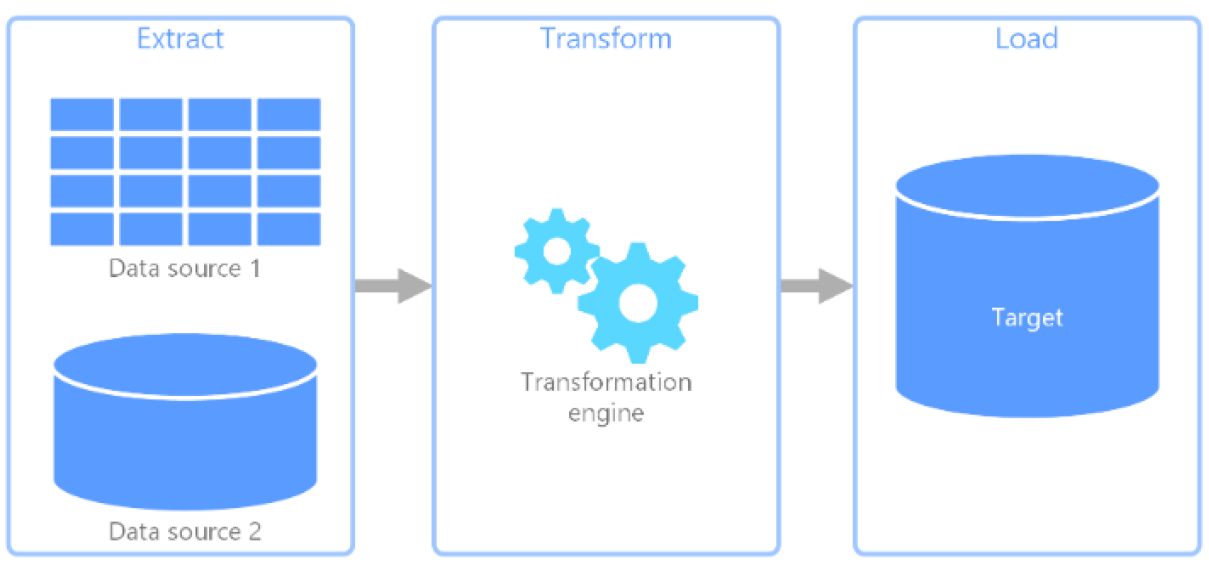
Apache Spark es un framework de propósito general para el procesamiento de datos en un clúster. Por supuesto tiene sus propias características, como ser las siguientes:

* Proyecto de código libre.
* Framework de procesamiento unificado (Batch y Streaming).
* Escalable y tolerante a fallos.
* Potente motor de procesamiento de datos masivos en memoria (in-memory processing) sobre un cluster (Spark está integrado con Hadoop, pero no depende de él).



ETL

Extraer, transformar y cargar (ETL, Extract, Transform, Load) es el proceso de compilación de datos a partir de un número ilimitado de fuentes, su posterior organización y su carga en un repositorio.



Principales Herramientas de ETL – On Premise

“Oracle Data Integrator (ODI), fue lanzada al mercado en el año 2006. Es una plataforma de integración completa de ETL que cubre los requisitos de integración de datos. Además, también ofrece un entorno gráfico para construir, gestionar y mantener procesos de integración de datos en sistemas de inteligencia empresarial.”



“Sin dudas, Pentaho Data Integration (PDI) es la herramienta open source más utilizada y conocida en la actualidad. La suite pentaho, nos permite implementar con gran eficiencia los procesos de extracción, transformación y carga de datos. Además, nos ofrece la posibilidad de realizar informes y analíticas sobre los datos. También, es posible adquirir una licencia de la herramienta para potenciar aún más sus capacidades.”



“Talend, ofrece una variedad de productos para realizar la integración de datos y los procesos de ETL. Open Studio for Data Integration, es open source y realmente muy potente, ya que funciona bien con integraciones para Google Big Query o Redshift. Luego, existen múltiples alternativas de pago, las cuales combinan diversas funcionalidades y herramientas.”



“Creado por la empresa líder en tecnología, la solución SAS Data Integration proporciona una herramienta de diseño visual, que simplifica significativamente la construcción, ejecución y mantenimiento de los procesos de integración de datos empresariales. También, cuenta con una potente interfaz fácil de usar y amigable con el usuario.”



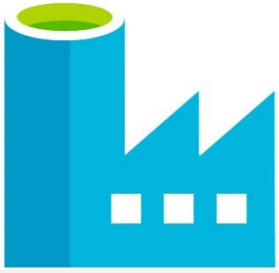
SQL Server Integration Services, también conocida por las siglas – SSIS, es un componente de Microsoft SQL Server utilizado para la extracción, transformación y carga de datos (ETL). Adicionalmente, es una herramienta muy popular, con mucha potencialidad y de las más reconocidas y utilizadas actualmente en el mercado.



Principales Herramientas de ETL – On Cloud

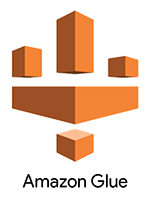
Azure Data Factory:

Es un servicio serverless de ETL el cual nos permite realizar procesos de data integration, data transformation y data load. A su vez ofrece una Interfaz gráfica muy intuitiva para la aplicación y creación de los ETL.



AWS Glue:

Servicio administrado del cloud AWS que nos permite hacer procesos ETL de una manera visual para extraer o volcar datos a AWS.



Data Flow:

Nos permite desarrollar ETL desde cero utilizando un pipeline o flujos de datos, también nos ayuda a realizar la ingesta de la información en tiempo real o en Batch de acuerdo al requerimiento de la empresa.

