Lección 5.7 Soluciones en la nube:

Hasta el momento, en este módulo hemos estudiado diferentes arquitecturas las cuales aplican técnicas de almacenamiento y procesamiento distribuido, como HDFS y Map Reduce, para afrontar el procesamiento de Big Data. Como podemos suponer, un arquitectura distribuida requiere de una compleja y costosa infraestructura hardware subyacente. Además, de los costes que puedan suponer tales recursos, la dificultad de la configuración de, por ejemplo, un clúster Hadoop es bastante elevada. Evidentemente, esto limita el acceso al Big Data a las PYMES, emprendedores o investigadores que cuenta con un bajo presupuesto y un equipo IT reducido. Por suerte, para ofrecer soluciones a estos y otros problemas del Big Data han surgido las soluciones Big Data en la nube.

1. Big Data en la nube

- Computación en la nube (Cloud Computing): Aplicaciones, servicios y almacenamiento que se ejecutan de forma distribuida en otras computadoras distintas a la nuestra, y los que se accede a través de internet.
    - Sin necesidad de que dispongamos de estos recursos de forma local (en nuestra computadora)
- En los últimos años, han surgido numerosas soluciones para el procesamiento y análisis de Big Data en la nube
    - Al igual que las soluciones locales, permiten afrontar las características del Big Data (5v's)
    - Ejemplos: Windows Azure, Amazon Web Services, Google Cloud Platform

2. Tipos de soluciones en la nube

Este tipo de soluciones para Big Data pueden clasificarse en tres tipos, según el concepto de computación en la nube:

- **Infraestructura como Servicio (IaaS)**: Alquilar procesamiento, almacenamiento, capacidad de red y otros recursos computacionales
    - Ej. Amazon Simple Storage Service (S3) (Almacenamiento) y Amazon Elastic Compute Cloud (EC2) (Capacidad de computación)

- **Plataforma como Servicio (PaaS, Platform As Service)**: Permiten desplegar aplicaciones creadas por los clientes a la nube
    - Ej. HDInsight (Windows Azure) y Amazon Elastic MapReduce (EMR) proporcionan distribuciones Hadoop para su uso en la nube.
- **Software como Servicio (SaaS, Software As Service)**: Uso de la aplicación del proveedor sobre la red
    - Ej. Google Big Query permite el análisis de Big Data en la nube y otras apps de Google permiten la generación de visualizaciones para BI.

3. Ventajas

El uso de soluciones en la nube tiene algunas ventajas importantes frente a las soluciones que requieren de una infraestructura local. Una de las más importantes es la optimización de recursos y , por tanto de, costes.

- Optimización de recursos y costes
    - “Pay-as-you-go” Costes en función de las necesidades de recursos actuales del proyecto.
    - Reduce al mínimo necesidades de personal IT para la instalación, mantenimiento y configuración de la infraestructura Big Data (ej. clúster Hadoop)
- No es necesario preocuparnos por contar con el hardware/infraestructura correcta al comenzar un proyecto de Big Data. Es fácil escalar hacia arriba o hacia abajo, según las necesidades de nuestro proyecto
    - Muy apropiado para el desarrollo de prototipos o POC (Prueba de concepto, Proof Of Concept)
- La nube permite el aprovechamiento del Big Data a aquellos que no cuentan con la experiencia o los recursos necesarios
    - Ej. Emprendedores, estudiantes, investigadores, PYMES

4. Desventajas

Sin embargo, no todo son ventajas en el uso de soluciones Big Data en la nube...

- Tiempo de subida y descarga de los conjuntos de datos a la nube
- Dificultad para controlar la ejecución distribuida de los procesos MapReduce sobre el clúster HDFS
    - Puede afectar al rendimiento
- Seguridad y privacidad de los datos que se almacenan en la nube

5. Principales soluciones Big Data en la nube

Algunas de las soluciones Big Data en la nube más conocidas son

- **Amazon Web Services**
  - Amazon Elactic Compute Cloud (S2) - Ofrece capacidad de computación escalable y SO's Windows o Linux preinstalado
  - Amazon Elactic MapReduce (EMR) - Plataforma Hadoop en la nube que soporta trabajos MapReduce, MapReduce Streaming, Hive y Pig
  - Amazon Simple Storage (S3) - Almacenamiento de bajo coste, escalable, confiable, seguro y rápido en la nube.
- **Windows Azure**
  - HDInsight: Implementación de Microsoft de Apache Hadoop en la nube, que permite escalar desde 4 hasta 32 nodos del clúster y de forma transparente al usuario
  - Azure (ASV) Blob Storage: Almacenamiento en la nube de alto rendimiento, confiable, escalable y seguro. Soporta almacenamiento basado en tablas y el uso de colas.
- **Google Big Query**
  - Almacenamiento y análisis de Big Data en la nube usando una BD NoSQL de tipo columnar desarrollada por Google, la cual ofrece un dialecto de SQL para las consultas.