Data Science & Cloud Computing

Claudia Salado Méndez

1 Resumen Conferencia

1.1 Introducción a Cloud Engineering y BI

¿Qué problema hay con la ciencia de datos? - Cada día los usuarios generan una gran cantidad de datos, esto cada vez va a más, y las tecnologías que a lo mejor no tienen tanto tiempo y que son relativamente nuevas se van quedando en desuso porque son imposibles de utilizar en la actualidad o si las usáramos tardarian muchísimo tiempo.

Es por esto que tenemos que dar el salto a la nube, ¿ y por qué la nube es la solución? La respuesta es clara, el utilizar la nube nos proporciona escabilidad, esto quiere decir que tiene una alta capacidad para crecer en grandes magnitudes, es totalmente distribuida, es un servicio totalmente descentralizado y auto organizado que facilita el equilibrio dinámico de las cargas tanto de almacenamiento como de procesamiento, también nos permites poner servicios redundantes y es 100% bajo demanda se fabrica bajo el gusto del cliente siguiendo todas las pautas y requisitos que este mande.

1.2 Plataformas más Usadas

- *Azure*
- *Google Cloud Platform*
- *Amazon Web Services*

Estas plataformas mencionadas anteriormente proveen una buena infraestructura:

- **Capacidades Saas:** máquinas virtuales, redes de almacenamiento, ...
- **Entornos integrados:** tanto como de desarrollo como de testeo incluso de producción.
- **Seguridad:** SDL security Development Lifecycle, permite el manejo de permisos con mucha granularidad
- **Herramientas de desarrollo**: como Azure CLI y Azure DevOps (herramienta para hacer despliegues automatizados)
- **The Enterprise Agreement Advantage:** al ser una empresa puedes tener acuerdos con Microsoft, que benefician a la empresa a progresar como negocio y dar mejores soluciones a los clientes.

1.3 Big Data

Lo más importante del big data son las "4 v's"

La primera es el *volumen de datos* es decir la cantidad de datos que se generan al día Como dato curioso en 2020 se llegaron a crear 40 zettabytes de datos.

La *variedad de los datos*, esto significa que no siempre los datos originales están en la forma con la que vamos a trabajar con ellos, van a llegar en muchos formatos por ejemplo tenemos de tipo sanitario, los que llegan de facebook, los de twitter, etc. Todos tienen formatos distintos. Por lo tanto el deber de un ingeniero de datos es agrupar todos los datos para poder sacar la mayor parte de información posible y que se pueda utilizar, porque como está el dato de origen no se puede utilizar.

La *velocidad de los datos*, velocidad a la que somos capaces de analizar dichos datos, como he explicado al principio con la infraestructura que había antes no se alcanza una buena velocidad, se tardaba muchísimo y por eso se pasó a utilizar la nube así podemos tener máquinas procesando datos cuando queramos

La *veracidad de los datos*, su función es proporcionar seguridad, es muy importante que el cliente u otros usuarios que usen los datos que hemos procesado se fien de que lo hemos hecho correctamente.

1.4 Cloud Data Pipeline

El punto de partida son los datos que el cliente quiere que recojamos pueden ser datos de ficheros, de una base de datos, de aws, … , de cualquier tipo.

El cliente nos da unos requisitos de que es lo que tenemos que hacer con esos datos nosotros lo transformaremos y los llevamos a una "capita" que se llama *landing*, que básicamente es la inserción de los ficheros tal cual vienen. De aquí parte el perfil de data engineer y el de data science, quienes necesitan el dato tal cual le llega,

¿Cómo se orquesta toda la carga de los datos? - con Azure Data Factory.
Esta es una herramienta que lo que permite es organizar todos los pasos que tenemos que lanzar para que se procese un dato, la primera parte sería el *landing* que es una copia de ficheros a otro sitio, después data bridge, es una herramienta que permite procesar todos los datos, es escalable y distribuida.

Con lo cual programando (en python usualmente) con data bridge todos los procesos y ya teniendo todos los datos procesados pasan a la capa de *staging* que es la siguiente en la que ya hemos limpiado y transformado los datos siguiendo los requisitos que nos específico el cliente pasamos en último lugar todos los datos al siguiente equipo qué es *Business Intelligence*, le dejamos todos los datos ya procesados y se encarga de generar un modelo que es super eficiente para que el usuario final sea capaz de ver los datos y tomar decisiones a partir de ellos. todo esto acompañado de gráficas,

reportes y tablas agrupando datos con Power BI para facilitar al cliente la toma de decisiones.

Aunque estos son los pasos más comunes a seguir y las herramientas que más se utilizan en ellos hay que estar pendientes constantemente porque microsoft saca continuamente nuevas herramientas y puede ser que encaje mejor con lo que nos pide un cliente.

Por ejemplo, hace pocos meses Microsoft sacó una nueva herramienta, Azure Synapse Analytics, que es adecuada para la fase de modelado y podría sustituir a la tecnología de una base de datos SQL pero no hay que generalizarlo puede que sea mejor para lo que nos pide este cliente pero no para otros. Esto hace que haya que estar continuamente reinventando y pensando en soluciones nuevas.

1.5 Metodología del Desarrollo Ágil

Las entregas del desarrollo con el cliente es mediante sprint suelen ser entre 7 y 15 días, entre cada sprint se revisa todo lo que se ha desarrollado se hace una pequeña demo del desarrollo que se ha hecho, se revisa si el trabajo se ha hecho bien o no mediante una retrospectiva, se comenta y se mejora para el siguiente. Todo eso teniendo también un Product Backlog, todas las tareas disponibles que el cliente quiere que hagamos, cada sprint se hace una reunión y se decide qué tareas entraran en ese sprint y durante este se van desarrollando poco a poco, hay un 100% de comunicacion con el cliente mientras se desarrolla, el cliente siempre sabe lo que el equipo de desarrollo está haciendo y si hay un problema se soluciona en ese momento.

1.6 Data Science

¿Qué es? - Es un concepto para unificar la estadística, el análisis de los datos, el aprendizaje máquina y sus métodos relacionados con el propósito de que después de procesar los datos y subirlos a la nube, la ciencia de datos va a extraer su valor y una información adicional que no es obvia a primera vista para ayudar a tomar decisiones más inteligentes y/o automatizarlas.

1.7 Data Scientist

Es el encargado de operar con los datos, es imprescindible que tenga conocimientos avanzados en estadística y matemáticas, que conozca también el campo que opera en ese momento, y saber de las tecnologías y lenguajes de programación, es muy importante que sepa adaptarse.

1.8 Ciclo de Vida de un Proyecto de Ciencia de Datos

Hay varias etapas, las cuales son:

- **Entender el negocio:** cuáles son las oportunidades o el problema que queremos solucionar.
- **Entender los datos:** comprender qué tipo de datos tenemos disponibles cuales son las limitaciones de estos, por que datos hay que pagar, cuales son fáciles o difíciles de obtener, si son heterogéneos etc

Estos dos pasos son recursivos cada vez que entendamos más del negocio tendremos que entender más sobre los datos o surgirán otros problemas y viceversa..

- **Preparar los datos y el modelo:** una vez que tengamos claro los datos a utilizar pasamos a prepararlos porque los modelos a aplicar necesitan que los datos estén procesados de una manera.
- **Evaluar los datos:** cuando el modelo está hecho lo evaluamos vemos si es eficaz al resolver el problema y si es satisfactorio pasamos a desplegarlo y si no volvemos a los pasos anteriores así hasta que esté correcto.

2 Comparación con Información Adicional

Para poder comparar la información dada en la conferencia, he buscados otras y también algunos artículos de revistas, algunas de las cosas más importantes que he sacado:

En cuanto a la ciencia de los datos nos da a entender que esta es una habilidad esencial de los ingenieros de software y que al igual que nos dijeron en la conferencia es muy importante tocar campos como la estadística, las matemáticas, etc.
También habla sobre la veracidad, el volumen, complejidad y seguridad de los datos:

'Cuando los datos locales son escasos, mostramos cómo adaptar los datos de otras organizaciones a los problemas locales. Cuando la privacidad se refiere a bloquear el acceso, mostramos cómo privatizar los datos sin dejar de ser capaces de dominarlos. Cuando trabajamos con datos de calidad dudosa, mostramos cómo eliminar la información falsa. Cuando los datos o modelos parecen demasiado complejos, mostramos cómo simplificar los resultados'[1]

Sobre el cloud computing también he encontrado artículos que le dan también importancia a las mismas cosas que se han mencionado en la conferencia, como por ejemplo que el cloud computing surge de la necesidad de aumentar la velocidad de los datos, ya que las arquitecturas anteriores se van quedando obsoletas o difíciles de utilizar. También que todavía es algo muy nuevo y que todavía no se sabe a ciencia cierta cómo se va a desarrollar, aunque algunos ejemplos de la práctica actual ya sugiere que direcciones va a tomar, como por ejemplo: Wordstar para la web, computación empresarial en la nube, infraestructura nublada, sistemas operativos en la nube, etc. [2]

Para hablar un poco sobre el data pipeline, he buscado sobre las arquitecturas DP actuales las cuales manejan el flujo de datos de un proveedor a otro, también proporcionan funcionalidades para usar las operaciones fundamentales de procesamiento (reemplazar y actualizar mensajes, convertir los datos sin procesar a un formato legible, etc). Las tecnologías implementadas permiten al desarrollador agregar múltiples módulos de DP a la configuración para manejar los datos de múltiples fuentes.[3]

3 Valoración/Opinión Personal

En mi opinión y tras haberme informado más de estos temas al buscar nueva información, creo que la conferencia estuvo bastante bien porque tocó los temas más redundantes, y se le dio más importancia a lo mismo que visto en los artículos, lo que ha hecho que al leerlos ya entendiera del tema y pudiera hacer comparaciones mejores.

Es verdad que hay una parte de la conferencia que no le he dado mucha importancia porque aparte de que me costó un poco entenderlo no tenía mucha importancia con esto, fue un ejemplo práctico que se hizo. Tal vez le hubiera dado más importancia a hablar más sobre el cloud computing y el data science más que al caso práctica, que duró bastante tiempo.

En conclusión, la conferencia estuvo muy bien, quitando el altercado que hubo a mi parecer fue amena y cercana, y me ha hecho conocer más estos conceptos que aunque los hemos escuchado bastantes veces en clase pues tampoco se suele profundizar mucho y ahora por fin ya tengo una idea más clara.

Referencias

1. T. Menzies, E. Kocaguneli, F. Peters, B. Turhan y LL Minku, "Ciencia de datos para la ingeniería de software", *35th International Conference on Software Engineering (ICSE) de* 2013, San Francisco, CA, EE. UU., 2013, págs. 1484 -1486, doi: 10.1109 / ICSE.2013.660675
2. Autor: Brian Hayes. Articulo: “Computación en la Nube”, Comunicaciones de la ACM, Volumen 51, Número 7 (2008), Páginas 9-11 https://dl.acm.org/doi/fullHtml/10.1145/1364782.1364786
3. Chinmaya Kumar Dehury, Satish Narayana Srirama, Tek Raj Chhetri, “CCoDaMiC: A framework for Coherent Coordination of Data Migration and Computation platforms”, Future Generation Computer Systems, Volume 109, (2020) Pages 1-16, ISSN 0167-739X, https://doi.org/10.1016/j.future.2020.03.029