Charla: « Data Science & Cloud Computing »

Raúl Rodríguez Pérez[1]

[1] Universidad de Granada, Granada 18010, España
raulrguez@correo.ugr.es

Resumen. El documento en cuestión trata, en primer lugar, del resumen de la conferencia ofrecida por 2 de los trabajadores de la empresa ElastaCloud el día 5 de marzo de 2021. En dicho resumen, se va comentando paso a paso todos los conceptos que desarrollaron los ponentes, como pueden ser, el Data Science, Cloud thecnologies, qué metodología de trabajo emplean en la empresa… Pero principalmente, debemos destacar que la temática de la charla no era otra que explicar los diferentes roles de trabajadores que poseía la empresa. A continuación, se realiza una valoración personal de la charla, aportando tanto mis opiniones respecto a los temas abordados, como sugerencias y/o aportaciones a la charla, es decir, si considero que la información aportada es completa, o añadiría algún punto más. Y finalmente, pongo a disposición todas las referencias a enlaces o citas que he ido recopilando a la hora de realizar el documento.

Palabras claves: Data Science, Cloud Computing, Business Intelligence engineer, ElastaCloud, Agile Software Development.

1 Resumen de la charla

1.1 Introducción

El tema principal de la conferencia ofrecida por el Máster en Ciencia de Datos e Ingeniero de Computadores, se basaba en explicar los tres distintos tipos de roles que posee la empresa **ElastaCloud** [1]. Dichos roles serían comentados de primera mano por trabajadores de la empresa en cuestión; por un lado Antonio Velasco Fernández (data Scientist), y por el otro Amanda Fernández Piedra (Cloud Data & BI Engineer).

1.2 Presentación ElastaCloud

En primer lugar, tras una pequeña presentación de los ponentes, el primer punto de la charla fue una breve explicación sobre ElastaCloud. Dicha empresa se dedica principalmente a proyectos de Big Data, siendo especialistas en Arquitectura Cloud (concretamente Microsoft Azure [2]) y Data Science. Como aspectos a destacar, tenemos que los fundadores de la empresa formaron parte del equipo de desarrollo de la plataforma Azure de Microsoft. Además, también nos comentaron que la empresa es partner de Microsoft UK, siendo este el país donde residen la mayoría de sus clientes. En relación a esto, Amanda procedió a explicarnos cuales eran sus principales clientes.

De entre todos ellos podemos destacar a empresas como ‘BP’ o ‘Drax’, por parte del sector de energías renovables, ‘NHS’ en el sector de salud, y ‘Porterbrook’ y ‘Murphy’ como representantes del sector de las infraestructuras de trenes.

1.3 Roles de ElastaCloud: Data engineer y Business Intelligence engineer

A continuación, como segundo punto de la charla, Amanda nos explicaría dos de los tres perfiles de trabajadores de ElastaCloud: Por un lado el ingeniero de datos o ‘**Data engineer**’, y por otro lado el ‘**Business Intelligence engineer**’. Comenzaría su explicación con una pregunta que, nosotros como estudiantes de la asignatura de desarrollo de sistemas distribuidos, tendríamos que tener clara su respuesta. La pregunta en cuestión era: ¿Cual es el problema principal de la ciencia de datos?, siendo la respuesta a la misma; la cantidad de datos que genera cada usuario. Ya que hoy en día dicha cantidad va en aumento, dejando inutilizadas a las herramientas tradicionales de análisis de datos. He aquí el principal motivo por el cual se están empezando a usar con mayor frecuencia tecnologías en la nube. Ya que, como hemos visto en el tema 1 de DSD, estas proveen escalabilidad, redundancia y bajo demanda. Seguidamente, la ponente nos mencionaría las tres principales plataformas que aportan servicios a la nube (Microsoft Azure, Google Cloud, Amazon Web Services), señalando brevemente gran parte de la infraestructuras que nos proveen. Destacando de entre todas ellas, las capacidades SaaS [3] (con Maquinas virtuales, redes, etc), la seguridad, las herramientas de desarrollo y los entornos integrados para el testeo y la producción.

Acto seguido, se explicó brevemente los cuatro puntos mas importante sobre el Big data; Volume - el volumen de los datos, Variety - la variedad de los datos entendida como la cantidad de formas diferentes en las que llegan los datos y la misión de sintetizar y quedarse con la información útil de cada dato. Velocity - la velocidad para realizar el análisis de datos (como comentamos anteriormente esto punto se ha visto enormemente beneficiado por el uso de Cloud Computing). Y por último, Veracity - la veracidad de los datos, saber con seguridad que lo que se ha procesado no contiene errores.

‘Cloud Data Pipeline’. Tras esto, la ponente prosiguió la charla, y llegó a la parte más interesante de la misma en mi opinión. En dicha parte se explicó la infraestructura de una plataforma de datos en el entorno de Azure. Es decir, conoceríamos las tecnologías y el proceso que se realiza desde que un cliente hace la petición de un proyecto, hasta que se finaliza el mismo. En primer lugar, dentro de lo que la ponente denominó como "Cloud Data Pipeline", se encontrarían todos los datos que son aportados por los clientes (estos pueden ser de muchos tipos como datos de fichero, base de datos, etc). Tras esto, dichos datos empezarían su transformación por el "Azure Data Factory" (aquí comienza la labor del data engineer), una herramienta que permite organizar y orquestar los pasos para procesar los datos. La primera fase de dicha herramienta es la capa de "Landing", que se encarga de básicamente copiar los datos de un sitio a otro. A continuación, se emplea la herramienta "Databricks" [4], una herramienta distribuida que permite procesar todos los datos (en ElastaCloud programan

los procesos mayoritariamente en Python). Una vez se han procesado los datos, estos pasan a la capa de "staging", en los que se transforman un poco los datos según los requisitos del cliente. Y por último, pasamos los datos al equipo de **Business Intelligence,** los cuales se encargan de crear un modelo muy eficiente para que el usuario final sea capaz de tomar decisiones a partir de los datos.

Una vez explicado esto, la ponente recalcó en repetidas ocasiones que estas no son las únicas tecnologías que emplean, ya que en ElastaCloud siempre están al tanto de nuevos lanzamientos por si acaso se podrían beneficiar de nuevas tecnologías emergentes. Sobretodo comentó que al conocer un mayor número de tecnologías, pueden realizar una mejor toma de decisiones para abordar los problemas específicos de los clientes.

1.4 Agile Software Development

Para finalizar, la ponente comentó brevemente sobre el método con el cual trabajan en ElastaCloud. Dicho método se basa en la metodología de Desarrollo Ágil (**Agile Software Development** [5]), la cual se basa en que sus entregas de desarrollo con el cliente tiene un plazo determinado. Cuando finaliza este plazo, entre cada entrega, se hace una revisión de lo que se ha desarrollado, además de un análisis retrospectivo para saber que se ha podido realizar mejor. Además Amanda recalca que hay 100% de comunicación con el cliente durante todo el transcurso del desarrollo, es decir, el cliente sabe en todo momento lo que el equipo de desarrollo está realizando. Así mismo, Amanda también destacó la importancia de saber Inglés, ya no solo para su caso en la que la mayoría de sus clientes son angloparlante, sino porque según comentaba la ponente la mayoría de las documentaciones sobre cualquier tecnología están en dicho idioma. Por lo que es vital tener una buena compresión del idioma para realizar no tan solo un mejor trabajo, sino una mejor adaptación cuando se requiera aprender una nueva tecnología.

1.5 Roles de ElastaCloud: Data Scientist

Tras esto, Amanda dio paso a su compañero Antonio para que nos hiciera una pequeña introducción a Data Science. Esta trata de unificar la estadística, el análisis de datos y el aprendizaje máquina teniendo como propósito la extracción de un valor o un conocimiento que no es obvio a primera vista, para realizar una mejor toma de decisiones y/o automatizarlas.

Luego de dar la definición de Data Science, Antonio explicó algunas aptitudes básicas que debe poseer un científico de datos. De entre todas ellas quiero destacar la versatilidad para poder adaptarse a los diferentes campos sobre los que vas a trabajar (por ejemplo si tienes un proyecto sobre energías renovables, saber manejar los términos y datos que se van a emplear). Además también recalcó la importancia de poseer un mínimo conocimiento no solo en estadística, sino en diversas tecnologías y algunos lenguajes de programación.

A continuación, el ponente nos mostró el ciclo de vida del Data Science con un esquema. Lo más destacable de esta parte fue la explicación de que dicho ciclo es recur-

sivo, es decir, a medida que se avanza por las distintas faces se pueden encontrar problemas o entender situaciones que hagan retroceder de fase para asentar bien las bases del proyecto.

Para finalizar, el ponente explicó un caso de estudio, el cual consistía en predecir respuestas a una campaña de Marketing. No voy a entrar en mucho detalle en esta parte de la charla, ya que se centra en aspectos mucho más relacionados con los estudiantes de economía y estadística y menos en lo que nos concierne a nosotros como son los sistemas distribuidos.

1.6 Preguntas y respuestas

Finalmente, al acabar la charla hubo un turno de preguntas y respuestas. Estas en su mayoría fueron relacionadas con el caso de estudio, siendo preguntas más técnicas sobre el uso de ciertos aspectos estadísticos y económicos. Aun así hubo una pregunta en concreto que me llamó la atención, ya que le preguntaban a Amanda si podía explicar un poco más sobre la herramienta de 'Databricks'. Esto era una duda que me había quedado a mi, ya que me interesaba el funcionamiento concreto de dicha herramienta. La ponente explicó que esta era un servicio en sí mismo que se enlazaba con Azure, proveyendo toda la estructura (en estos casos los clústeres) para que puedas ejecutar tú mismo el código que procesa los datos de forma completamente distribuida.

2 Valoración personal

Como conclusión final a este documento, quiero dar mi opinión acerca de varios temas desarrollados en la charla. En primer lugar, y como ya sabíamos con anterioridad, los sistemas distribuidos y el Cloud Computing son el futuro de una gran cantidad de empresas. No tan solo por los aspectos que comentaba Amanda de que las tecnologías antiguas se quedaban obsoletas por el aumento masivo de datos que generábamos, sino por que, estas tecnologías nos aportan grandes ventajas como pueden ser: capacidad de crecimiento (al ser escalables), un alto rendimiento al realizar una programación paralela, mayor tolerancia a fallos que los sistemas no distribuidos. Además de que por la parte del Cloud Computing, también encontramos numerosas ventajas como; las reducciones de coste al prescindir de inversiones en infraestructuras TI propias, el aumento de la movilidad, al poder acceder desde cualquier dispositivo y lugar, una capacidad de almacenamiento casi ilimitada, etc. A lo largo de mi carrera como estudiante de informática diversos profesores ya nos lo habían comentado, pero que esta vez sea una trabajadora desde la misma empresa quien nos comente como ha sido el paso y las ventajas que posee emplear dichas tecnologías, hace que uno se de realmente cuenta de la importancia que tiene conseguir una buena base para el futuro laboral.

Por otro lado, en la conferencia se trataron dos temas que fueron los que más llamaron mi atención. Uno de ellos fue las herramientas para procesar datos como, por ejemplo, 'Databricks'. En la charla no se profundizó apenas en dichas herramientas, simplemente Amanda nos comentó cual era su objetivo. Y no fue hasta que en el

turno de preguntas y respuestas que le preguntaron si podía explicar más profundamente la herramienta. Con su explicación y lo que he investigado por mi cuenta sobre 'Databricks', he llegado a la conclusión de que es un "entorno de trabajo" ideal para la colaboración de dos de los roles que se explican en la charla (Data Scientist y Data engineer), sobretodo para estos últimos, ya que les abre un mundo de posibilidades para planificas y/o ejecutar los diversos clusters que han desarrollados los Data Scientist.

El otro tema que llamó mi atención fue la metodología de trabajo que empleaban en ElastaCloud. Me puse a investigar sobre si era una metodología muy antigua o reciente, y en que ámbito, o mejor dicho, cual era el tipo de empresas que mejor le convendría realizar dicha metodología. Ya que simplemente la ponente solo explicó en que consistía y como lo llevaban a cabo en ElastaCloud. En primer lugar encontré que el desarrollo ágil de software comenzó sobre los años 1990, a raíz de las protestas del desarrollo contra el método en cascada. Además me di cuenta de que muchísimas empresas muy importantes empleaban esta metodología como pueden ser Apple, PayPal, Facebook, Zara, etc. Me hubiera gustado que en la charla se hubiera mencionado un poco más de este tema, pero al investigarlo por mi cuenta también me enteré y descubrí mucho mejor en que consistía y su importancia.

Para finalizar también quiero destacar como hicieron los ponentes, la importancia una vez más de aprender y usar con fluidez y agilidad el idioma universal del mundo, el inglés. Durante toda la charla hicieron pequeñas referencias a la importancia del idioma no solo para el ámbito laboral sino para un trabajo de comunicación básica con el cliente.

Referencias

1. ElastaCloud Homepage, https://elastacloud.com/ 2021/3/7.
2. Microsoft Azure Homepage, https://azure.microsoft.com/es-es/ 2021/3/7.
3. Definition of SaaS, https://www.infoworld.com/article/3226386/what-is-saas-software-as-a-service-defined.html 2021/3/7.
4. Información de investigación sobre Databricks: 2021/5/7
   - https://databricks.com/
   - https://blogs.encamina.com/por-una-nube-sostenible/que-es-azure-databricks/
   - https://aprenderbigdata.com/databricks/
5. Información de investigación sobre Agile Software Development: 2021/5/7
   - https://www.cprime.com/resources/what-is-agile-what-is-scrum/
   - https://es.wikipedia.org/wiki/Desarrollo_%C3%A1gil_de_software
   - https://www.infoworld.com/article/3237508/what-is-agile-methodology-modern-software-development-explained.html
   - https://www.iebschool.com/blog/metodologia-agil-agile-scrum/