Acerca de la Ingeniería de Datos para Aprendizaje Automático

Dr. Leandro Nahabedian. Profesor Visitante-DC-FCEN. Doctor en Ciencias de la Computación. con la colaboración del Dr. Ricardo Oscar Rodríguez (Profesor regular Asociado dedicación Exclusiva)

Programa:

Que al final del curso el alumno pueda desenvolverse en tres responsabilidades utilizando cualquier batería de componentes que hoy en día nos ofrecen los distintos proveedores *cloud*: Amazon Web Services (AWS), Google Cloud Platform (GCP) y Microsoft Azure.

La idea no es hacer foco en cada una de ellas, sino que abstraernos de las mismas para presentarles una batería de componentes única que solo posee las características que son compartidas entre los tres proveedores.

Temario:

- Clase 1: Presentación. ¿Cuál es el rol del *data engineer*? Proveedores *cloud*: AWS GCP Azure. Generalidades de los tres proveedores. Revisión general de todos los componentes. Almacenamiento
- Clase 2: Conceptos básicos de almacenamiento de datos. Tipos de datos a almacenar. Base de datos relacionales vs no-relacionales. Buckets. Graph databases. Caching.
- Clase 3: Datos en real-time. Procesamiento de datos en *streaming*. Arquitectura de *streaming* con *retries*. Pub/Sub vs Stream Analytics vs Kinesis vs Kafka Procesado
- Clase 4: ETL vs ELT. Orquestadores (Airflow). *Stateless apps*.
- Clase 5: Spark (pySpark). Databricks. Soluciones Dockerizadas Arquitectura
- Clase 6: Misc: Infrastructure as code. CI/CD. Monitoreo y Reportes.
- Clase 7: Soluciones a alto nivel utilizando los componentes vistos. Soluciones *Batch*. Soluciones Real-time. Arquitectura Lambda. Repaso para el examen.
- Clase 8: Examen

Bibliografía:

La bibliografía es principalmente la documentación de estos tres proveedores *cloud* y sus *learning paths*: AWS, GCP y Azure. Como bibliografía adicional se destacan algunos papers que se listan a continuación:

- Polyzotis, Neoklis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. “Data validation for machine learning.“ Proceedings of Machine Learning and Systems 1 (2019).
- Polyzotis, Neoklis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. “Data Management Challenges in Production Machine Learning.“ In Proceedings of the 2017 ACM International Conference on Management of Data (2017).
- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. “Everyone wants to do the model work, not the data work: Data Cascades in High-Stakes AI“. In Proceedings of the Conference on Human Factors in Computing Systems, (2021).