FACULTAD DE INFORMÁTICA | UNIVERSIDAD NACIONAL DE LA PLATA

CAPTURA Y ALMACENAMIENTO DE INFORMACIÓN

Año 2018

Carrera: ***Especialización en Inteligencia de Datos orientada a Big Data***

Carga Horaria: 64 Hs.

Profesor a Cargo***: Mg. Oscar Bría, Mg. Javier Bazzocco, Ing. María José Basgall***

OBJETIVO

La captura y el almacenamiento de la información son frecuentemente las fases iniciales en un proceso de análisis y representación de datos. En este curso se propone analizar distintos mecanismos capaces de realizar estas tareas de manera eficiente.

El curso tiene dos partes: la primera relacionada con la revisión de distintos mecanismos de captura de la información y la segunda donde se cubren los aspectos relacionados con almacenamiento de información a través de BBDD NoSQL.

MODALIDAD DE EVALUACION

Para aprobar el curso se requiere un 80% de asistencia y la realización de un trabajo final que se definirá una vez completada la exposición de los contenidos teóricos. Dicho proyecto podrá ser realizado en forma individual o grupal y tendrá por objetivo profundizar uno o varios de los conceptos vistos en clase. La entrega consistirá en un reporte técnico de calidad científica producto de un estudio teórico o experimental específico.

PROGRAMA

Parte I – Captura de Datos

- Big Data. Características y desafíos del Big Data. Fuentes de Datos.

- Digitalización de Datos. Tipos de Datos y sus Características: Ópticos: Imágenes, videos y otros. Sonoros: voz, música. Provenientes de Otros Sensores.
- Dispositivos de Captura: Manuales y OCR. Escáneres. Cámaras. Lectores de Marcas: OMR, MICR. Micrófonos. Sensores Varios: Radares, Nucleares, etc.
- Lenguaje de marcado. Metadatos.
- Captura de datos. Extracción de datos de las redes sociales. Obtención de datos históricos y datos en tiempo real. Utilización de APIs y herramientas para capturar datos.
- Captura de información de la web. Web scraping y web crawling. Tipos de web crawlers. Uso de expresiones regulares. Librerías y frameworks específicos.

Parte II – Almacenamiento de Información

- Bases de Datos relacionales y no relacionales o NoSQL. Breve historia de NoSQL. Descripción y tipos de bases de datos NoSQL: orientadas a columnas, documentos, claves-valores, grafos, objetos o híbridas. ACID vs. BASE. Teorema de Brewer. Ventajas y desventajas de NoSQL.
- Orientadas a documentos - MongoDB
  - Historia. Conceptos básicos. Filosofía de diseño. Velocidad, escalabilidad y agilidad. Modelo no relacional. Soporte transaccional. Performance vs. Características. Comparación con SQL
  - Patrones de diseño. Tratamiento de datos en MongoDB. Inserción de datos. Lecturas y consultas. Actualización de datos. Agregación.
  - Gestión de MongoDB. Migración de base de datos. Rendimiento y Sharding. Seguridad
- Orientadas a Columnas - Apache Cassandra: Características principales. Aplicaciones. Modelo de datos. Clustering, replicación. Lenguage de consulta.

BIBLIOGRAFIA

- Ryan Mitchell. *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc., 2015.
- Ricardo Moya. *Scraping en Python (BeautifulSoup), con ejemplos*. URL https://jarroba.com/scraping-python-beautifulsoup-ejemplos/

[último acceso 20-08-2017]

- Mohan, C. History repeats itself: sensible and NonsenSQL aspects of the NoSQL hoopla. Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013.
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., ... & Gruber, R. E. (2008). Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, *26*(2), 4.
- Navin Sabharwal, Shakuntala Gupta Edward. *Big Data NoSQL Architecting MongoDB*. CreateSpace Independent Publishing Platform, 2014.
- Banker, Kyle and Bakkum, Peter and Verch, Shaun and Garrett Doug and Hawkins, Tim. MongoDB in Action. 3rd edition, 2017. Manning Publications.
- Bradshaw, Shannon and Chodorow, Kristina. MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. 3rd Edition
- Carpenter, Jeff and Hewitt, Eben. *Cassandra: The Definitive Guide: Distributed Data at Web Scale*, O'Reilly Media; 2nd edition (July 22, 2016)
- Neeraj, Nishant. "Learning Apache Cassandra: managing fault-tolerant and scalable data." Packt Publishing, 2nd edition (2015).
- Lakshman, A., & Malik, P. (2010). Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, *44*(2), 35-40.
- Wang, Guoxi, and Jianfeng Tang. "The nosql principles and basic application of cassandra model." *Computer Science & Service System (CSSS), 2012 International Conference on*. IEEE, 2012.
- Dede, E., Sendir, B., Kuzlu, P., Hartog, J., & Govindaraju, M. (2013, June). An evaluation of cassandra for hadoop. In *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on* (pp. 494-501). IEEE.