Máster en Business Intelligence & Data Science

Edición 2017/2018

Asignatura: Adquisición de datos
Módulo: Adquisición de datos para toma de decisiones
Profesor(es): José Antonio Rodríguez Díaz, josearodriguezd@gmail.com

OBJETIVOS

El objetivo general del módulo es conocer algunas de las principales herramientas y técnicas utilizadas actualmente para la adquisición e ingesta de datos en entornos con grandes volúmenes de información así como también identificar los casos de uso idóneos para cada una de estas herramientas y las buenas prácticas en su utilización.

En concreto, los objetivos del módulo serán:

1. Conocer las principales herramientas utilizadas para la ingesta de datos.
2. Aprender a migrar datos desde bases de datos relacionales a HDFS y viceversa usando sqoop.
3. Realizar migraciones de datos hacia y desde HDFS haciendo uso de distintos formatos de archivos como Avro, Parquet y de texto con distintos delimitadores.
4. Importar datos en Hive utilizando apache sqoop.
5. Ingesta de datos utilizando apache Flume.
6. Indexación de un sitio Web mediante web crawling.
7. Aprender a identificar cuales herramientas son las idóneas para la adquisición de datos.

REQUISITOS PREVIOS

Es necesario tener una base conceptual del funcionamiento de MapReduce y HDFS (sistema de archivos) así como también de Hive (Data warehouse). Deseable conocer distintos formatos de serialización de datos como pueden ser Parquet, Avro, entre otros.

METODOLOGÍA

En este módulo, el aprendizaje del alumno se llevará cabo principalmente con una enseñanza teórica y práctica, teórica para tener una base conceptual sólida y práctica mediante la realización de ejercicios basados en requisitos funcionales bastantes cercanos a la realidad. Para consolidar el conocimiento se aplicará el aprendizaje basado en casos donde entre todos harán análisis y discusión de casos de uso y de esta manera prepararlos para la toma de decisiones sobre las herramientas a utilizar en determinados casos acorde a los requerimientos.

PROGRAMA

El programa se estructura en los bloques que se describen a continuación.

Bloque 1: "Apache sqoop"

- ¿Qué es Sqoop?
- Sqoop 1 vs Sqoop 2.
- Casos de uso.
- Buenas practicas.
- Ejemplos.

Bloque 2: "Apache flume"

- ¿Qué es Flume?
- Apache flume vs apache sqoop.
- Casos de uso.
- Buenas practicas.
- Ejemplos.

Bloque 3: "Web scrapring"

- ¿Qué es web crawling?
- En que se diferencian web scraping de web crawling.
- Casos de uso.
- Herramientas para hacer web crawling
- Ejemplos con scrapy.

Bloque 4: "Análisis de casos de uso"

- Otras herramientas utilizadas en la adquisición de datos: Apache nifi, apache gobblin y streamsets data collector.
- Análisis de casos de uso.

EVALUACIÓN

La evaluación se realiza mediante Pruebas de Evaluación Continua (PEC). La siguiente tabla las describe. Las fechas de publicación, entrega y publicación de calificaciones se hará disponible a través del Aula Virtual.

Prueba	Tipo (indicar tipo de prueba, práctica, test, etc. Se recomienda ejercicios prácticos o casos prácticos)	Peso
PEC1	Practica individual: ejercicios prácticos usando sqoop	50%
PEC2	Practica individual: ejercicios prácticos usando flume	50%

BIBLIOGRAFÍA OBLIGATORIA

- Documentación oficial de apache sqoop: https://sqoop.apache.org/docs/1.4.7/index.html
- Documentación oficial de apache flume: http://flume.apache.org/documentation.html
- White T. (2015) Hadoop the definitive guide 4th edition. O'Reilly Media Inc.

BIBLIOGRAFÍA RECOMENDADA

- Grover M., Malaska T., Seidman J. & Shapira G. (2015). Hadoop Application Architectures designing real world big data applications. O'Reilly Media Inc.
- Ting K. & Cecho J. J. (2013). Apache Sqoop Cookbook. O'Reilly Media Inc.

PROFESORADO

José Antonio Rodríguez es Ingeniero de Sistemas mención investigación de operaciones graduado en la Universidad de los Andes de Venezuela, actualmente desempeñando el rol de arquitecto Big Data en Atos y con un postgrado en Business Analytics y Big Data cursado en el CIFF. Con más de 10 años de experiencia en el desarrollo de software y con conocimientos en la definición e implementación de arquitecturas Big Data y Cloud para el sector bancario.

LinkedIn: https://www.linkedin.com/in/josedeveloper/
Blog: http://josedeveloper.com
Mail: josearodriguezd@gmail.com