Programa del curso

Semestre 2022-20

Nombre del curso:	Análisis de Información sobre Big Data
Course Name:	Big Data Analytics
Créditos:	4
Profesor:	Claudia L. Jiménez G. Christian Fernando Ariza Porras
Versión PDF	Click Aquí

Propósito

Big Data (Datos Enormes) es el término para referirse al contexto de ***integración y análisis*** de cantidades masivas de ***información*** móvil, web, social y en la nube, ***pertinentes para el usuario*** y relevantes para entender ***el ecosistema de una organización***. El análisis de cantidades enormes de datos que se generan tanto dentro de las organizaciones como fuera de ellas, ha cambiado las ***tecnologías*** y las ***metodologías*** con las cuales se desarrollan soluciones ***basadas en contenidos*** que buscan generar ***valor***, ***diferenciación*** y oportunidad en la ***toma de decisiones***. El propósito del curso es presentar, analizar y utilizar las oportunidades de innovación que ofrece el análisis de grandes cantidades de datos en: la toma de decisiones estratégicas y tácticas de una organización, el desarrollo de aplicaciones en diferentes campos del conocimiento y la selección e integración de infraestructuras que aseguren una alta escalabilidad permitiendo así un crecimiento natural de las soluciones implementadas. ***A nivel estratégico y táctico de una organización***, **Big Data Analytics** busca comprender y aprovechar los datos propios y externos a la empresa con el fin de entender los cambios y las tendencias de mercado, identificar opiniones de segmentos poblacionales relevantes para el negocio e interpretar de flujos de datos provenientes de fuentes sociales para generar análisis de competitividad. ***A nivel de desarrollo de aplicaciones Big Data,*** se generan técnicas y metodologías propias para este tipo de información que además son adaptables a diferentes campos de aplicación, permitiendo así el uso efectivo de los datos en el análisis de una problemática específica. Entre los campos de desarrollo de Big Data se encuentran, entre otros: el análisis de comercio electrónico, el entendimiento en línea de la reacción de clientes frente a un producto o su competencia, la definición y ajuste de políticas públicas, las telecomunicaciones, los videojuegos en línea, las aplicaciones gubernamentales, aplicaciones de salud y ciencia, el análisis del comportamiento urbano y la predicción, prevención y reacción frente a desastres. ***A nivel de infraestructura***, tecnologías como *Hadoop* y *NoSQL* son utilizadas para facilitar la alta escalabilidad necesaria en procesamiento, y almacenamiento de este

tipo de información. El uso de este tipo de tecnologías acompañado de la definición de arquitecturas orientadas a los datos, permite ofrecer sistemas robustos y eficientes generando ventajas competitivas en diferentes perspectivas en el ámbito empresarial así como en los ámbitos científico e investigativo. Finalmente, ***a nivel de información***, suele trabajarse con fuentes estructuradas como no estructuradas, profundamente heterogéneas. La información proviene de fuentes diversas usualmente autónomas, es creciente de forma exponencial y no manipulable de forma efectiva con herramientas tradicionales de gestión de bases de datos. Según IDC, se estima en 1.8 Zetabytes (1.8 * 106 Petabytes) la información generada sólo en 2011, siendo los contenidos los protagonistas. Las fuentes suelen ser blogs, wikis, RSS, email, comunidades participativas como las redes sociales y comunidades virtuales especializadas. Estas se integran con la información propia a las organizaciones y los individuos, de manera ubicua. Se cuenta con ejemplos, desarrollados en el contexto de cursos y proyectos de investigación del grupo, en dominios tan variados como la biología, la medicina, temas financieros, análisis de opinión de productos en el mercado, análisis de imagen de personajes públicos, análisis de comentarios en noticias de prensa y análisis de estado de la comunidad a partir de los streams de redes sociales.

Objetivos

- Identificar las oportunidades de transformación y generación de procesos de generación de valor basadas en el análisis de información, proveniente tanto de fuentes internas como externas a la organización.
- Comprender, definir y evaluar arquitecturas orientadas por datos (Scalable Data-Driven Architectures), en particular aquellas que involucran requerimientos de alta escalabilidad de procesamiento y almacenamiento
- Integrar metodologías y tecnología para el descubrimiento y entendimiento de información basado en fuentes altamente escalables. Ejemplos de ellas son Linked Data, Social Data, Sentiment Analysis, Online Stream Analysis, Web Intelligence.
- Integrar metodologías y tecnología de análisis de información apropiadas para escenarios de datos no estructurados o semiestructurados, en los cuales se enfrenta el proceso de análisis en condiciones de la alta escalabilidad y son necesarios procesos de adaptación y reacción en tiempo real.
- Desarrollar una solución que permita generar valor y diferenciación a partir de procesos de análisis de información sobre Big Data

Metodología

Durante el curso se desarrollan tanto actividades teóricas como prácticas. Los temas y conceptos de base son desarrollados en clase. Para cada capítulo del curso se dispone de amplia bibliografía que complementa los temas, propone ejemplos o casos de estudios y ofrece detalle sobre lo tratado. Las actividades prácticas se desarrollan alrededor de talleres y laboratorios. Son incrementales en los retos técnicos, aunque cada uno se refiera a la práctica de una temática específica. Las actividades prácticas buscan que el estudiante plantee la integración de técnicas y herramientas en el planteamiento de una

solución de generación de valor a través de la gestión de una infraestructura de información altamente escalable, utilizando técnicas de análisis de información adecuadas para el problema planteado. Si bien las actividades prácticas se realizan en grupo, la evaluación de las mismas es individual.

Habilidades y conocimientos previos

- Saber programar. Es deseable tener conocimiento de Phyton, Java o C++ (en ese orden de preferencia). Se espera un nivel intermedio de programación, como mínimo.
- Matemáticas básicas: teoría de conjuntos, álgebra de vectores
- Conceptos básicos de probabilidad y estadística
- Uso y administración básica de Linux
- Conocimiento básico de SMBDR: SQL, modelo relacional
- Conocimiento básico de arquitecturas de software, patrones de desarrollo de software, patrones de arquitectura
- Desarrollo de aplicaciones Web, desarrollo de servicios
- Conocimiento básico de procesamiento de lenguajes

Documentación sobre algunos de estos temas se puede encontrar en: Sistemas Transaccionales Arquitecturas de Software Desarrollo de software en java Programación de Tecnologías Web Adicionalmente, al inicio del curso se pueden sugerir guías en línea o tutoriales que permiten reforzar temas específicos.