Universidad de los Andes
Facultad de Ingeniería

Ingeniería de Sistemas y Computación
Maestría
MINE4102 – Análisis de Información sobre Big Data
https://cursos.virtual.uniandes.edu.co/mine4102/
Semestre: 202320

ABET Engineering Accreditation Commission

Programa del Curso

Información general

Profesores	Correo electrónico	Atención a Estudiantes	
Christian Ariza	cf.ariza975@uniandes.edu.co	Atención virtual	El primer canal de comunicación es la sesión sincrónica de encuentro programado en Banner. Los datos de conexión se encuentran en Bloque Neón (BN). En otros horarios, el primer canal de comunicación es foro en BN. Para consultas individuales se coordina cita vía correo electrónico uniandes
Lorena Peñuela	l.penuelac@uniandes.edu.co		

Objetivos pedagógicos

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.

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.

A NIVEL ESTRATÉGICO Y TÁCTICO de una organización, el análisis de Big Data 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 cuenta con técnicas y metodologías propias para el manejo de 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. En particular, la paralelización de procesos es una técnica que permite la escalabilidad de las aplicaciones.

A nivel de INFRAESTRUCTURA, tecnologías se estudia Hadoop y se utilizan, entre otras, Spark y sistemas manejadores de bases de datos NoSQL 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.

El curso aborda las técnicas, herramientas y metodologías de base para el descubrimiento, entendimiento, procesamiento y la generación de valor en el contexto de datos enormes (Big Data). Se espera que el estudiante desarrolle habilidades en el planteamiento y desarrollo de soluciones que permitan ofrecer valor a partir de información relevante y pertinente en un contexto global, en el cual el análisis y toma de decisiones involucra tanto la información propia como externa.

Este objetivo se desarrolla a través de la comprensión de la tecnología, análisis de alternativas, estudio de herramientas, planteamiento metodológico de modelaje de las arquitecturas subyacentes y estudio de técnicas de procesamiento, bajo criterios de calidad de servicio, propias a la gestión de *Big Data* y de contenidos.

De esta forma, se espera desarrollar y reforzar las siguientes competencias:

- **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. **Conocer, utilizar y desarrollar criterio** sobre las posibilidades ofrecidas por la tecnología que permite el manejo de contenidos semiestructurados y no estructurados, de manera que el estudiante pueda **evaluar** su valor y alcance en el contexto de soluciones en los cuales la gestión de documentos y datos no alfanuméricos sea prioritaria.
- **Estudiar** el estado del arte en gestión de contenidos dinámicos, colaborativos y semiestructurados, de forma que el estudiante conozca las alternativas de tecnología existentes, su fundamentación y los problemas abiertos o en investigación
- **Conocer** herramientas de gestión de información que constituyen el estado del arte en la gestión moderna de cantidades enormes de datos.
- **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** 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*
- Generar habilidades de **uso e integración de herramientas**, formas de procesamiento, indexación y modelaje de datos en contextos de alta escalabilidad.

Habilidades y conocimientos previos

Se espera que, desde el inicio del curso, los estudiantes cuenten con habilidades de trabajo en equipo, responsabilidad, buen manejo del idioma (oral y escrito), habilidad lectora (en español e inglés) y comportamiento ético.

En cuanto a los aspectos profesionales técnicos, se espera que el estudiante tenga habilidades intermedias de programación (preferiblemente en Java o en Python), desarrollo Web, conocimiento básico de sistema operativos basados en Unix, bases de probabilidad y estadística, y conocimientos básicos de tecnologías de manejo de bases de datos. Adicionalmente, aunque no es requerido, se recomienda tener conocimiento básico de contenedores, como Docker y Kubernetes.

Plan de temas

- Caracterización de *Big Data – Arquitecturas escalables, frameworks, ecosistemas. Almacenamiento, procesamiento*
- Procesamiento a escala: Map Reduce, Spark
- Almacenamiento a escala: HDFS, repositorios NoSQL, DataLakes, productos representativos
- Estudio de casos
- Procesamiento escalable de flujos de datos y procesamiento en tiempo real
- Procesamiento escalable de datos espaciales

Metodología

Durante el curso se desarrollan tanto actividades teóricas como prácticas, de manera las temáticas de cada uno de los módulos sean complementarias y generen valor, tanto de manera independiente, como integradas en soluciones que apoyan el desarrollo estratégico de las organizaciones.

Las actividades prácticas se desarrollan en grupos de hasta 3 estudiantes. Se programan alrededor de tareas, talleres y laboratorios, que abordan tecnologías, herramientas o problemáticas de alcance puntual y buscan que el estudiante logre una experiencia práctica de base en el tema correspondiente.

El avance en el curso se hace mediante el uso incremental de las herramientas, metodologías y tecnología, en prácticas que utilizan contenidos e infraestructura de información altamente escalable.

Al final del curso se espera que los estudiantes hayan desarrollado **soluciones funcionales completas** en un ambiente de integral de ciclo de vida de información altamente escalable, heterogénea y no estructurada, que evidencie su dominio en el uso las tecnologías, herramientas y conceptos vistos en el curso, aplicados a un contexto específico.

La entrega y evaluación de resultados de los talleres se hace mediante la construcción, publicación y demostración de soluciones Web funcionales.

En cada uno de los módulos se tienen actividades que permiten a los estudiantes:

- Estructurar el conocimiento sobre los temas a desarrollar. Esto se realiza mediante actividades de exploración bibliográfica, presentaciones, discusiones en clase y revisión del estado del arte, a nivel tecnológico y de investigación.
- Entender, construir y publicar soluciones desarrolladas sobre ambientes de información no estructurada y escalable.
- Conocer diversas herramientas de acuerdo con las particularidades de la solución informática planteada y generar habilidades de autoaprendizaje.
- Utilizar la tecnología de *Big Data* y análisis de contenidos para solucionar problemas no convencionales centrados en información.
- Reconocer y gestionar la calidad de la solución planteada, principalmente en las dimensiones de escalabilidad, heterogeneidad y análisis de contenidos.
- Plantear escenarios de generación de valor para el desarrollo estratégico de soluciones innovadoras apoyadas en *Big Data* y en información no estructurada o semiestructurada.

Evaluación y aspectos académicos

Generalidades

- Sesiones sincrónicas: 3 horas semanales, de asistencia obligatoria. Las sesiones serán virtuales pero los exámenes (2) se programarán de forma presencial.
  Durante estas sesiones el profesor lleva bitácora de presencia y participación de los estudiantes. Esta bitácora constituye el registro de asistencia. El estudiante debe asistir al menos al 80% de las sesiones sincrónicas y sesiones de trabajo supervisado para aprobar el curso, de acuerdo con el RGEM.
- Laboratorios: actividades prácticas autónomas, que permiten experimentar con plataformas específicas o reforzar temas presentados. Se realizan en la infraestructura de laboratorios del Departamento. Pueden ser planeadas en espacios de encuentros sincrónicos, individuales o en grupos. El profesor dará las indicaciones pertinentes en cada caso.
- Talleres: el desarrollo de los talleres constituye la práctica fundamental del curso. Se trabaja en grupos de 3 estudiantes. Son actividades prácticas para desarrollar por el estudiante fuera de clase, que refuerzan o permiten experimentar con los temas y conceptos desarrollados. Exigen competencias incrementales, correspondientes a los niveles pedagógicos del curso.
- Tareas: actividades puntuales que se desarrollan bien sea en durante las sesiones sincrónicas o en forma autónoma; pueden ser asignadas de forma individual o en grupo. Pueden ser desarrollos prácticos, hojas de trabajo, preparación de presentaciones orales, etc.
- Durante el curso, a través de las actividades prácticas propuestas**, cada estudiante debe mostrar habilidades de diseño, desarrollo, uso, publicación, evaluación y presentación de sus resultados**.
- Tecnología utilizada: sistema operativo Unix, ambientes de desarrollo de software propios a los temas abordados en el curso, repositorios NoSQL, MapReduce, ambiente Spark, DataLakes, herramientas de análisis de contenidos, herramientas de análisis de datos estructurados y no estructurados, altamente escalables.
- Se realizan exámenes individuales, programados durante sesiones sincrónicas. Las sesiones sincrónicas de examen están programadas de manera presencial. En ningún caso se autorizan evaluaciones de forma virtual o remota. La única excepción podría darse por lineamientos institucionales que indiquen que se requiere, en esas fechas, que todas las actividades académicas se realicen de manera virtual. Esto sería comunicado a los estudiantes de manera oficial por las autoridades de la Universidad.
- Las evaluaciones de los talleres incluyen sustentaciones, en grupo, en momentos programados para ello y diferentes a las sesiones sincrónicas de avance en el curso. Las sustentaciones se realizan de manera presencial, salvo que se indique o autorice hacerlo virtualmente, por razones de fuerza mayor.
  Todas las actividades propuestas que se realicen en grupo conllevan elementos de evaluación individual para cada uno de los integrantes.

- Se realiza seguimiento a la preparación y participación en el curso.
- Los aspectos relacionados con el trabajo en equipo (normas, seguimiento y evaluación individual del trabajo en equipo), aspectos relacionados con la entrega de trabajos prácticos, logística de evaluaciones, reglas de presentación de exámenes y normas de presentación de informes técnicos, hacen parte de las reglas del curso y se encuentran en documentos adjuntos.
- El agendamiento de actividades básicas del curso se anuncia en la primera clase del semestre y está sujeto a cambios, de acuerdo con el desarrollo del semestre.
- No se realiza ni autoriza la grabación de las sesiones presenciales. Este lineamiento es institucional, tanto de las autoridades de la Universidad como propias al Departamento de Ing. de Sistemas y Computación. Las sesiones sincrónicas presenciales no son transmitidas.
- La grabación de sesiones de sustentación, de realizarse, estarán a cargo del profesor o del monitor del curso.
- Durante las sesiones sincrónicas virtuales se espera que los estudiantes mantengan siempre habilitadas su cámara y micrófono, para facilitar su participación, actividad y retroalimentación con sus compañeros y el profesor. En caso de tener dificultades de infraestructura tecnológica, los estudiantes deben desplazarse al Campus para atender la sesión. Allí cuentan con servicio de energía, internet y préstamo de equipos computacionales.

El curso tiene como canales oficiales de comunicación el correo electrónico uniandes, la lista de correo del curso, el sistema de apoyo a la docencia Bloque Neón (https://bloqueneon.uniandes.com) y la página Web del curso.

Evaluación del curso

Los elementos de evaluación son los siguientes:

- Al menos dos exámenes parciales, individuales. En el examen pueden ser incluidos elementos escritos, de audio o de video. Así mismo, pueden ser utilizadas tecnologías de monitoreo en las plataformas digitales de respuesta.
- Trabajos prácticos: incluyen talleres, laboratorios y tareas. Si bien pueden ser realizados en grupo, la evaluación es individual. Se tiene en cuenta tanto el producto entregado como el proceso con el cual se logró. La evaluación se realiza a partir de los entregables del trabajo, la documentación solicitada, el despliegue y la demostración del producto cuando así se indica. Son elementos individuales de evaluación: sustentación, contribución individual al trabajo presentado, autoevaluación y evaluación de sus pares. Son elementos que penalizan la evaluación de un entregable: no marcar, marcar de forma incompleta o que no siga la forma solicitada de hacerlo, entregar en formatos de archivos diferentes a los indicados, no respetar los formatos o buenas prácticas en el desarrollo de software o documentos, responder de forma incompleta cuestionarios o encuestas de desempeño de grupos, identificar de forma incorrecta a los compañeros de grupo cuando así se solicita o no seguir, en general, las indicaciones y pautas de los entregables, descuido en redacción y ortografía en los entregables.
- Trabajo en clase: puede incluir tareas, presentaciones, quices, controles de lectura, construcción colectiva de conocimiento, participación en actividades colaborativas, participación en foros o debates y demás actividades propuestas. Pueden ser evaluadas de forma individual o de grupo.
- No toda actividad propuesta en clase, o tarea asignada, es necesariamente evaluada; cualquiera de ellas puede serlo.
- Es criterio de evaluación el seguimiento de un adecuado proceso de trabajo en equipo. Se consideran la adecuada metodología de trabajo, el equilibrio en las responsabilidades y distribución del trabajo y el cumplimiento de compromisos mutuos. El no cumplimiento de estas características conlleva a penalizaciones en las evaluaciones correspondientes.
- Los elementos de autoevaluación, evaluación de pares y declaración de contribución, cuando existen, son de obligatorio cumplimiento e impactan de forma directa el resultado de la respectiva evaluación.
- Las reglas de juego para los trabajos prácticos, para la presentación de documentos y para la presentación de exámenes son definidas en documentos anexos al programa y se publican en el sitio BN del curso.
- Las reglas específicas a cada entrega (objetivos, alcance, retos, entregables) se entregan con el enunciado correspondiente. Estas reglas son de obligatorio cumplimiento.
- Todas las evaluaciones consideran aspectos de calidad de la entrega, tanto en forma como en contenido. Esto incluye el seguimiento de las instrucciones y tiempos de entrega, seguimiento de estándares y formatos, calidad del manejo del idioma en los escritos y calidad del desarrollo del software. **Estos aspectos impactan de forma directa el resultado de la respectiva evaluación.**
- En todos los casos, es indispensable que el software desarrollado funcione de forma efectiva con información proveniente de los *datasets* indicados, que usualmente serán reales. Así mismo, debe cumplir con los requerimientos y restricciones,

tanto técnicos y como funcionales, indicados en el enunciado. **Debe cumplir, sin excepción, los requerimientos de despliegue de la solución**. **No se reciben entregas de resultados en infraestructura diferente a la prevista en cada enunciado.**

Los porcentajes de las evaluaciones se indican a continuación. La distribución en cada categoría es uniforme, salvo que se informe durante el semestre otra determinación:

Evaluaciones	Porcentajes
Parciales (2)	60%
Talleres (2)	28%
Laboratorios, tareas, hojas de trabajo	9%
Cuestionarios	3%

Si, al finalizar el curso, el estudiante cumple simultáneamente con las siguientes dos condiciones, tiene un bono de 0.25/5.00 en su acumulado total de notas:

- Promedio ponderado de parciales >= 3.4/5.0
- Promedio ponderado de talleres >= 4.0/5.0

Política de aproximación de notas finales

- Las evaluaciones se califican entre 0.00 y 5.00 con dos decimales y no hay aproximaciones.
- Las notas definitivas del curso varían entre 1.50 a 5.00, en intervalos de 0.25. Las notas intermedias de dichos intervalos son aproximadas por el profesor teniendo en cuenta el desempeño global del estudiante y del curso. El valor a partir del cual se aproxima en cada intervalo, de forma ascendente o descendente, es decidida por el profesor y se aplica por igual a todos los estudiantes de cada sección.
- Para aprobar el curso es indispensable lograr 3.00/5.00 en el puntaje ponderado. No existe aproximación automática en la nota definitiva; en particular, no hay aproximación a 3.00 de puntajes menores a esta nota (v.gr., 2.99 no es 3.00).

Recursos

Para el desarrollo del curso se cuenta con recursos de infraestructura basados en máquinas virtuales y clústers de cómputo escalable, provistas por el Departamento. Es factible que se programen actividades prácticas en infraestructuras en nube, de acuerdo con la disponibilidad que ellas ofrezcan.

Bibliografía

La mayor parte de la bibliografía del curso se encuentra disponible en forma electrónica. Si es el caso, debe estar en sesión activa de la biblioteca para lograr el acceso.

También se recomienda hacer uso de las posibilidades de acceso académico a bibliografía seleccionada, que ofrecen ACM y O'Reilly a personas con estatus de estudiante.

La bibliografía correspondiente a cada uno de los temas cubiertos es publicada en BN a medida que avanza el semestre.

Protocolo MAAD

El miembro de la comunidad que sea sujeto, presencie o tenga conocimiento de una conducta de maltrato, acoso, amenaza, discriminación, violencia sexual o de género (MAAD) deberá poner el caso en conocimiento de la Universidad. Ello, con el propósito de que se puedan tomar acciones institucionales para darle manejo al caso, a la luz de lo previsto en el protocolo, velando por el bienestar de las personas afectadas.

Para poner en conocimiento el caso y recibir apoyo, usted puede contactar a:

1. Línea MAAD: lineamaad@uniandes.edu.co
2. Ombudsperson: ombudsperson@uniandes.edu.co
3. Decanatura de Estudiantes: Correo: centrodeapoyo@uniandes.edu.co

4. Red de Estudiantes: PACA (Pares de Acompañamiento contra el Acoso) paca@uniandes.edu.co
5. Consejo Estudiantil Uniandino(CEU) comiteacosoceu@uniandes.edu.co