

Academia Ingeniero de Datos

Proyecto Integrador

Lineamientos generales

- **Metodología de trabajo:** será de forma colaborativa en grupos de dos o tres integrantes máximo, en lo posible con perfiles que puedan complementarse desde la experiencia previa.
- **Desarrollo E2E:** se pretende que los alumnos puedan demostrar los conocimientos adquiridos a lo largo de los módulos teórico-prácticos impartidos en la academia
- **Temática:** como se comentó durante la cursada, anexados al presente documento se plantearán una serie de ejemplos prácticos que podrán tomarse como base del proyecto a desarrollar, sin embargo, quedará abierta la posibilidad de que cada equipo aporte su impronta ya sea mediante la modificación de alguna de esas ideas o generando una propuesta totalmente diferente.
- **Documentación:** tanto el plan del proyecto como su progreso deberán dejarse asentados, al menos a grandes rasgos, dentro de un repositorio de GitHub. Lo recomendable es que utilicen su correspondiente archivo “[README](#)” (Ejemplo: <https://gist.github.com/Villanuevand/6386899f70346d4580c723232524d35a>) y que cada pieza de código incluya mediante comentarios todo apoyo que crean pertinente para explicar su funcionamiento y facilitar su mantenimiento.
- **Entrega inicial:** durante la semana del 11 de Julio de 2022, cada equipo deberá decidir la temática a abordar y planificar una posible solución. Tendrán que elaborar como entregable un diagrama con una descripción mínima de los componentes de dicha propuesta y aprovecharán la primera sesión de laboratorio para plantear sus dudas al respecto.
- **Entrega final:** se espera que el proyecto sea un desarrollo que pueda implementarse, que sea operativo y reutilizable, para lograrlo se contará con el soporte de los instructores de la academia y habrá varias instancias de laboratorio para revisión, pero por sobre todo se espera la cooperación entre pares mediante la comunidad generada en Teams. La entrega final 29 de Julio de 2022
- **Encuentro de cierre:** algunos proyectos serán preseleccionados para que el día del encuentro el equipo a cargo pueda exponer su experiencia frente al resto de la audiencia de la academia. Se les permitirá a sus integrantes remitir la invitación con algunos colaboradores del banco cercanos y está la posibilidad de que el encuentro se grabe para la promoción interna y externa de la academia.

Restricciones

A continuación, se listan algunas consideraciones importantes para tener en cuenta:

- Repetimos la importancia de que trabajen en equipo de forma coordinada y que todos los integrantes hagan su aporte para que el proyecto se haga más llevadero y puedan cumplir con las entregas.
- Como se utilizará un repositorio externo, tanto los archivos que se suban como su documentación no deberán contener ninguna referencia que se los vincule al banco (por ejemplo, si van a hacer referencia a alguna ruta dentro del HDFS interno omitirían el primer nivel y podrían usar una variable de entorno que luego reemplazarían en el código con un archivo no sincronizado “{HDFS_ROOT}/d/landing_files/academia_de”).
- Por el mismo motivo que el punto anterior, tendrán que abstenerse de publicar información de carácter sensible y en todo caso deberían encargarse de ofuscar previamente aquellos datos que puedan causar algún tipo de conflicto.
- Como es posible que varios equipos trabajen sobre temáticas similares, si encuentran que el objeto que quieren crear ya existe (tabla, vista, directorio, etc.), tengan la precaución de no modificarlo ni borrarlo ya que pueden estar afectando a otro proyecto, en todo caso crearían el objeto con otro nombre similar.
- No es necesario que todo ejecute desde el servidor de Python (*dpyserver02*), aquellos que tengan acceso a una máquina conectada a la red del banco también podrían utilizarla para el desarrollo de la solución, quedando a su cargo la configuración del entorno y dependencias.
- Cuando trabajen en los servidores compartidos tengan en consideración liberar los recursos una vez dejen de usarlos: cerrar terminales y sesiones abiertas, no dejar procesos activos, ir borrando archivos de prueba, etc.

Metodología de evaluación

Se debe tener presente que la entrega y aceptación del proyecto es condición necesaria, junto con la valoración de los módulos en campus, para considerar concluida satisfactoriamente la academia y poder recibir la certificación avalada por Datalytics y Banco Galicia. (*)

Para realizar la evaluación de los proyectos se definirá un comité integrado por algunos integrantes del equipo de formadores de Datalytics en colaboración con los sponsors referentes por parte del banco.

Se apreciarán cuestiones como:

- Conocimientos fundamentales en Python
- Uso del HDFS
- Modelado de los datos en Hive
- Exploración y enriquecimiento de los datos mediante algún análisis aplicado con Spark
- Inclusión de algún flujo mínimo en Nifi
- Deseable que puedan incorporar Kafka
- Explotación de la información generada con alguna demostración visual aprovechando una o más de las siguientes herramientas:
 - Flask (**): exponer una API y/o generar una SPA (Single-Page Application)
 - Jupyter notebook (**)
 - Tablero en PowerBI: ya sea conectándolo directamente al clúster con Impala o a través de una API

En cuanto a la aprobación del trabajo, no requiere de una nota numérica, se juzgará los resultados en base a las consignas, en caso de cumplir se emitirá la certificación, caso contrario se pasara a una instancia de revisión y ajuste hasta que el proyecto quede definitivamente aprobado.

Este mismo comité será también el encargado de determinar cuáles serán aquellos trabajos (en principio tres, con la posibilidad de sumar alguno más según disponibilidad de tiempos y como se coordine el encuentro de cierre) que hayan sobresalido como para ser presentados en el encuentro de cierre.

*(**) Tanto en Flask como en una Jupyter podrían incluir gráficos estáticos y dinámicos (interactivos).*