UNIVERSIDAD CATÓLICA de Colombia	**Facultad de Ingeniería**	**Ingeniería de Sistemas y Computación**

Asignatura: Análisis Inteligente Utilizando Big Data		**Fecha:** Septiembre – 2018
Nombre de la Practica: Análisis de Datos y Transformaciones		
Unidad Temática: Datos Abiertos y ETL		
Contenido Programático: Modelamiento		
Objetivo de la Practica: Documentar Datasets de problemas reales y construir procesos de ETL y ELT		
Fecha de Entrega: 13 de Septiembre de 2018	**Corte:** Primero	**Puntos:** 25

Enunciado y Problemática:

Dentro de los contenidos explicados en clase se deberá realizar el proceso inicial de un proyecto de Data Science teniendo en cuenta los siguientes datasets:

- Grupo 1: Datos Completos del Icfes 2016
- Grupo 2: Datos Completos del Icfes 2015
- Grupo 3: Datos Compilados generales del Icfes 2011-2016-2
- Grupo 4: Datos de Cuentas de Twitter

Todos los grupos deberán entregar los siguientes productos

- **Documento de Diccionario de Datos:** Dentro de este documento (ver formato) se deberán consignar toda la información requerida para cada uno de los datasets asignados, teniendo en cuenta también lo lineamientos explicados en clase y los propios del documento. Este proceso aplica por igual para todos los sets de datos expuestos.
- **Script de transformación de Datos y Carga:** El grupo deberá crear un script que tenga en cuenta procesos de extracción de datos, limpieza de los mismos y carga de estos a una base de datos según la naturaleza de los mismos datos. El script deberá desarrollarse en los lenguajes pyhton o java exclusivamente, según la habilidad de cada uno de los integrantes del grupo.
- **Reto asignado según la naturaleza del problema:** Cada dataset deberá tener en cuenta lineamientos específicos para su proceso, según indicaciones que se consignarán más delante de este documento

Definición de Problemática por cada dataset:

- **Grupos 1 y 2:** Para este tipo de datos, los estudiantes deberán tener en cuenta inicialmente los siguientes lineamientos:
  - **ETL:** Dentro del proceso los estudiantes deberán recibir un archivo de texto tal cual y como se presenta en el link de Google Drive y deberán crear un proceso en batch automatizado que lea el archivo(simulando que es un equipo servidor), cargarlo en memoria, realizar la limpieza de datos con el fin de realizar una carga en una base de datos MySQL, la carga en la base de datos se realizará en una (1) tabla por set de datos, y el nombre de la tabla será el nombre del archivo en cuestión.

- **Reto:** El reto para estos dos grupos esta dado en el tamaño y performance del proceso del paso del cliente al servidor, el único lineamiento es tratar de hacer la carga del archivo (simulando que es un ambiente desplegado en internet como método de conexión, no local) se realice lo más rápido posible.

- **Grupo 3:** Para este tipo de datos, los estudiantes deberán tener en cuenta inicialmente los siguientes lineamientos:
  - **ETL:** Se deberá recibir el archivo xls, tal cual y como se presenta en el link de Google Drive y deberán crear un proceso en batch automatizado que lea el archivo, cargarlo en memoria, realizar la limpieza de datos con el fin de realizar una carga en una base de datos MySQL, la carga en la base de datos se realizará en una (1) tabla hoja del archivo, por set de datos, y el nombre de la base de datos será el nombre del archivo en cuestión.
  - **Reto:** Se tomará como reto el mismo archivo ya que se encuentra en formato XLS y los datasets en diferentes “Hojas” dentro del mismo archivo
- **Grupo 4:** Para este tipo de datos, los estudiantes deberán tener en cuenta inicialmente los siguientes lineamientos:
  - **ETL:** Se deberá recibir un archivo en formato JSON el cual representa extracciones de cuentas de twitter, leer dicho archivo y guardarlo en una base de datos NoSQL, específicamente el motor MongoDB.
  - **Reto:** Consiste en que se entregarán un número indeterminado de archivos creados de la misma manera y el script en batch automatizado deberá leer todos los archivos e insertarlos en “colecciones” aparte cada colección deberá ser llamada según el nombre del archivo en cuestión.

Mínimos de Entrega:

1. Documento de Diccionario de datos
   - Documentación de los conjuntos de datos según el dataset elegido.
   - Descripción del proceso de ETL o ELT
     - Entradas
     - Procesos Internos
     - Salidas
     - Representación Gráfica.
     - Listado de tecnologías, librerías, frameworks en detalle
2. Implementación de Algoritmo de ETL o ELT según corresponda a las indicaciones generadas en la sección anterior.
   - Entregable como Adjunto.

Criterios de Evaluación:

- ✓ Documentación de los Set de datos (7 pts)
- ✓ Documentación del Proceso de Extracción (7 pts)
- ✓ Proceso de extracción probado en clase (7 pts)
- ✓ Presentación en clase. (4 pts)

NOTA: Cualquier entregable adicional será considerado como adicional y será calificado como tal con puntuación adicional a la expresada en este documento