**Evaluación N° 3 BigData Grupo 1**

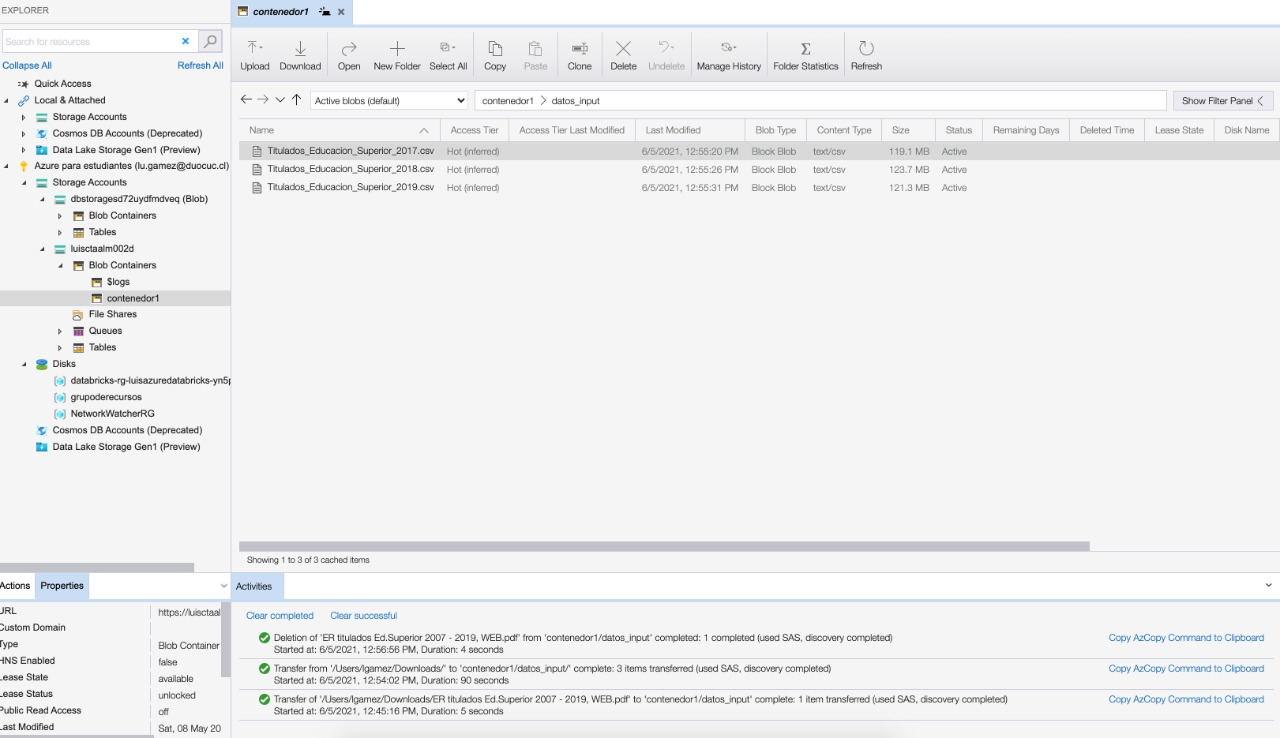
**Instrucciones Generales:**

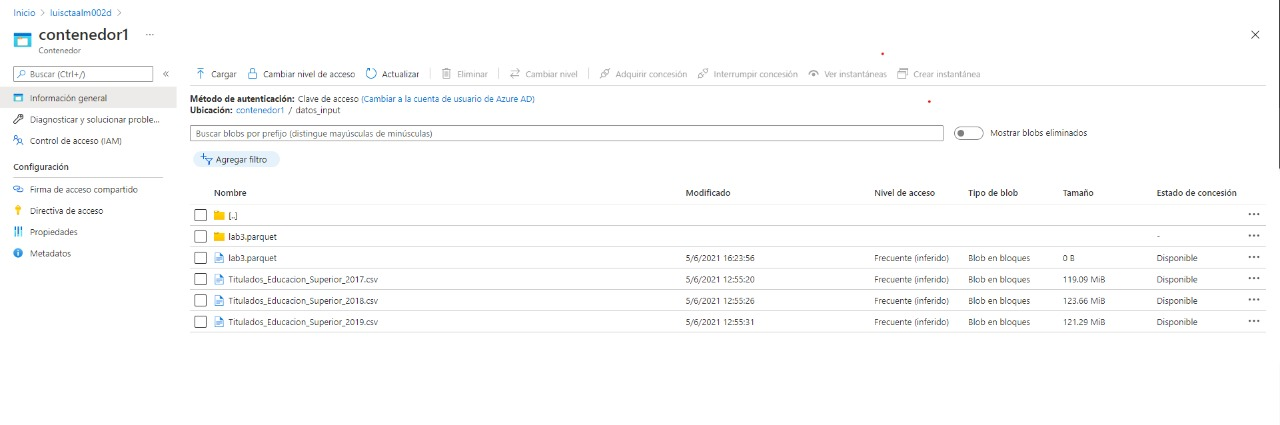
Trabajo grupal donde deberá implementar un proceso de ingesta, transformación y almacenamiento de datos en ambiente Azure Databricks.

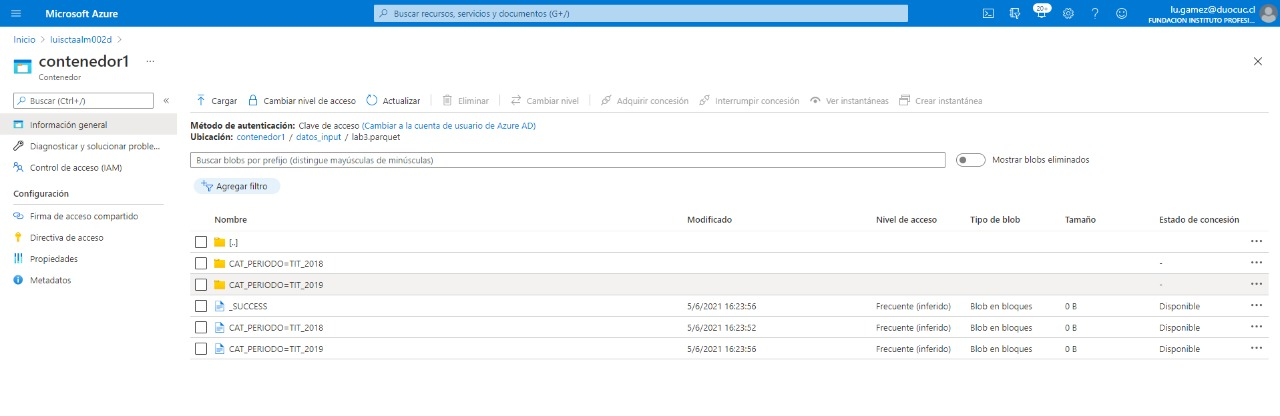
Para el caso indicado deberá:

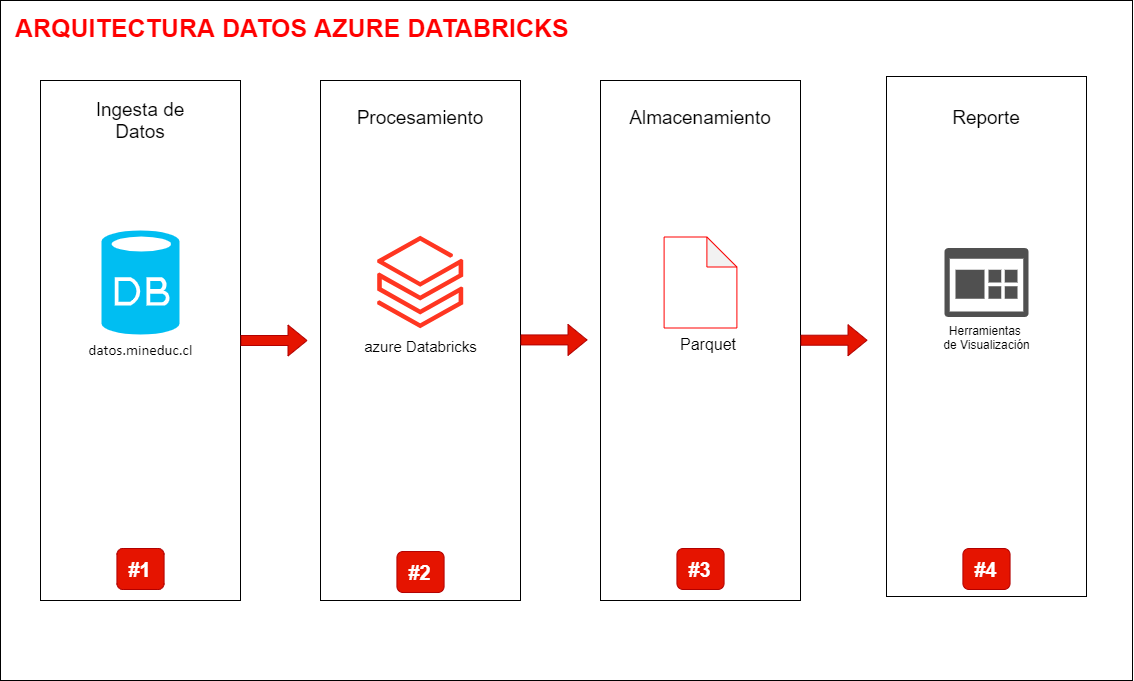
1. Desde el sitio de Mineduc (http://datos.mineduc.cl/dashboards/20207/descarga-bases-de-datos-titulados-de-educacion-superior/) descargar los archivos de los titulados de educación superior para los años del 2017 al 2019 (el formato de los archivos los puede descargar desde este link http://datos.mineduc.cl/datasets/179064-er-titulados-de-educacion-superior.download/.
2. **(1 punto)** Hacer Diagrama de arquitectura de la solución, debe contener
   1. Diagrama de alto nivel, indicando los componentes a utilizar
   2. Diagrama con el detalle de las funciones a implementar
3. **(1 puntos)** Cargar los archivos en una carpeta llamada “datos\_input”, dentro de la cuenta de almacenamiento en Microsoft Azure. Como evidencia, obtener print de pantalla y pegar en este documento
4. **(2 puntos)** Almacenar los datos en un archivo con formato Parquet, particionado por el campo de Año. Como evidencia, realizar print de pantalla del archivo parquet almacenado en la cuenta de almacenamiento, mostrando la estructura de particionamiento. El proceso debe considerar lo siguiente:
   1. (0.5 puntos)Debe crear una función que lea todos los archivos, recibiendo los parámetros correspondientes para su uso.
   2. (1 puntos) Debe crear una función para que guarde los datos en parquet. Esta función debe tener la capacidad de ejecutarse N cantidad de veces, y el resultado sobre la tabla final debe ser el mismo (es decir, ante reprocesos, no debe caerse, no debe duplicar datos, y no puede borrar los otros años cargados).
   3. (0.5 puntos)Función que ejecute el código completo.
5. **(2 punto)** Responder las siguientes preguntas:
   1. (0.5 puntos): Gráfico que muestre la cantidad de graduados por cada institución de educación para el año 2019
   2. (0.5 puntos): Gráfico que muestre la cantidad de graduados por cada carrera para la institución de educación “IP DUOC UC”, comparando los años 2018 y 2019.
   3. (0.5 puntos): Gráfico que muestre la evolución de egresados desde el año 2017 a 2019, comparando las Jornadas diurnas y vespertinas
   4. (0.5 puntos): Gráfico de distribución de las edades, ¿qué características tiene la distribución obtenida para el año 2019?

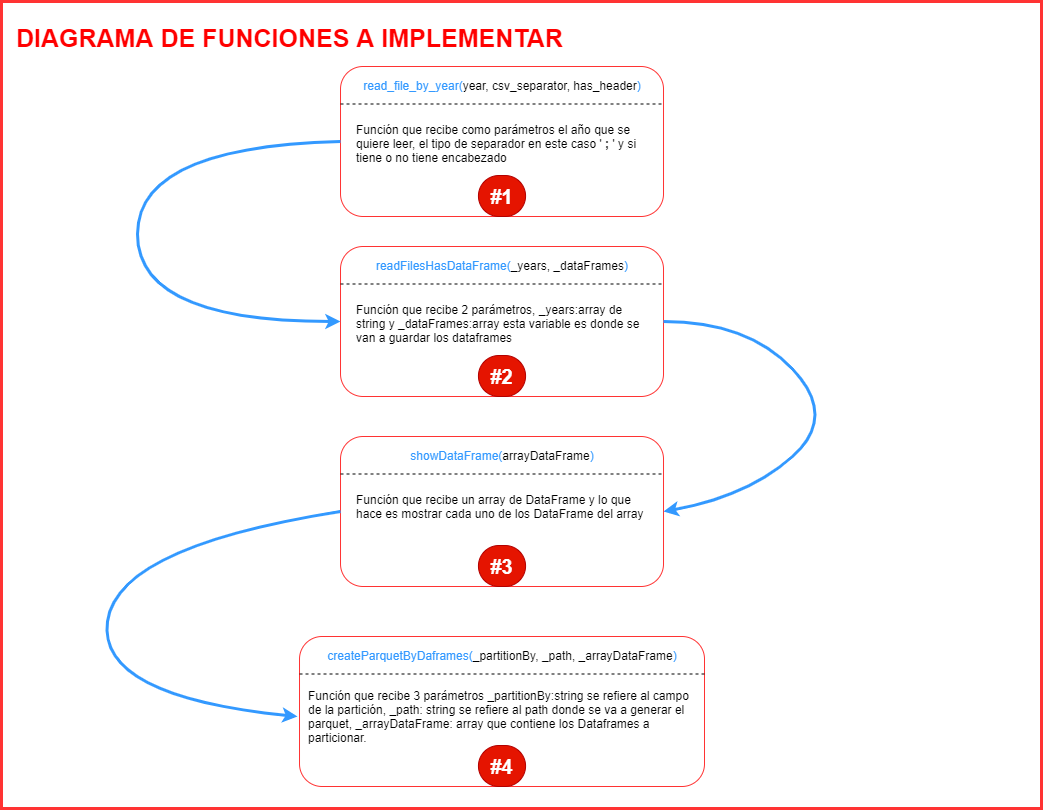
Debe entregar el código del Notebook generado en Azure Databricks exportado en formato HTML, junto con los prints de pantallas en este mismo documento.











**Luis Gamez – Juan Pablo Ureta – Raúl Soissa**