Implementación de un Ambiente de Ciencia de Datos a través de la Librería Open Source Rapids

José Guarnizo1

1Departamento de Ciencias de la Computación  
Universidad Técnica Particular de Loja (Ecuador)  
jaguarnizo4@utpl.edu.ec

**Resumen.** La librería Open Source Rapids ejecuta canalizaciones de datos y análisis de extremo a extremo utilizando las principales características del poder de la GPU de NVIDIA, la librería maneja grandes volúmenes de información en tiempo real y trabaja con computación acelerada, su principal enfoque es realizar una analítica de la información, utilizando el proceso de extracción, transformación y carga e implementar un conjunto de algoritmos Machine Learning, se da realce a sus principales funcionalidades, características y todo su conjunto de bibliotecas que tiene la librería, además, se utiliza la Plataforma Cloud BlazingSQL para la implementación, donde resalta el proceso de construcción de un Ambiente de Ciencia de Datos utilizando Rapids para los procesos de ETL, Machine Learning y Visualización, se implementó una contraparte utilizando librerías que aún no tiene soporte para GPU como Pandas para el proceso ETL, para el proceso de Machine Learning con Scikit-Learn y para la visualización con Matplotlib, la comparación con las librerías tradicionales permite resaltar la aceleración y rapidez que tiene Rapids cuando trabaja con grandes volúmenes de información.

**Palabras clave:** Ciencia de Datos. Rapids. Machine Learning.

# Introducción

La cantidad de datos producidos en la actualidad son de gran volumen, es creciente en instituciones públicas y privadas, producen millones de datos al día, empresas como: bancos, entidades del estados, negocios independientes, entre otros, aportan considerables cantidades de información que se encuentran almacenadas en Bases de Datos, información almacenada en la nube y otros medios con la finalidad de almacenarlas y respaldarlas, lo que causa que no tengan un adecuado tratamiento en cuanto a su procesamiento, implementación y utilización para generar valor y rapidez.

La implementación de un ambiente de un ambiente de ciencia de datos a través de la Librería Open Source Rapids, se enfoca en analizar un conjunto de datos abiertos utilizando las principales características de la librería.

Mediante un análisis completo de la librería Rapids permitirá manejar grandes volúmenes de información en tiempo real, además, poder determinar y hacer uso de las características mas importantes de la librería que permiten: ocupar el máximo de recursos disponibles, tiempos mínimos de cada recurso y capacidad para el análisis, procesamiento e implementación de algoritmos Machine Learning para que los datos generen valor y sean entendibles.

# Librería Open Source Rapids

La librería de código abierto Rapids brinda la capacidad de ejecutar canalizaciones de datos y análisis de extremo a extremo en GPU (Unidad de Procesamiento Gráfico), se enfoca en tareas comunes de preparación de datos y ciencia de datos.

Rapids permite un procesamiento y capacitación enormemente acelerados en tamaños de conjuntos de datos mucho más grandes, brinda a los científicos de datos un enorme salto de rendimiento para resolver desafíos comerciales más complejos como predecir fraudes con tarjetas de crédito, pronosticar inventarios de venta minorista y comprender el comportamiento de los clientes, etc.

El principal objetivo de la librería Rapids no es solo acelerar las partes individuales del flujo de trabajo típico de la ciencia de datos, sino acelerar el flujo de trabajo completo de extremo a extremo, acelera su cadena de herramientas de ciencia de datos de Python con cambios mínimos de código, aumenta la precisión del modelo de aprendizaje automático para aprender, es personalizable, extensible, interoperable, el software de código abierto es compatible con NVIDIA y mejora drásticamente su productividad con ciencia de datos casi interactiva.

A continuación, se ira detallando las principales bibliotecas que tiene Rapids para el ambiente de Ciencia de Datos como: ETL, Machine Learning y Visualización de los datos en forma gráfica.

## Biblioteca Cudf

Cudf es la biblioteca de manipulación de DataFrames basada en Apache Arrow, está enfocada en la preparación de datos, acelera la carga, el filtrado y la manipulación de datos, contiene la definición del modelo de memoria gdf, y las funciones de procesamiento de datos pueden operar en esas respectivas columnas de gdf\_ [1].

Los científicos de datos pueden interactuar con la biblioteca casi exactamente como lo harían con la librería Pandas para las respectivas transformaciones de datos, la limpieza y el análisis respectivo que se requiere a los datos a evaluar, este proceso es conocido como ETL (Extracción, Transformación y Carga).

Cudf está conformado por dos aspectos principales LibGDF y PyGDF, esto hace que la biblioteca trabaje de manera eficiente para el procesamiento y análisis de los datos.

Estructura de la bilioteca cudf

Se enfoca en puntos primordiales como libgdf y pygdf; libgdf es una biblioteca de bajo nivel que contiene implementaciones de funciones y API para C y C++; pygdf es una biblioteca de Python para manipular DataFrames de GPU, permite crear gdf a partir de matrices Numpy, Pandas y PyArrow.

**Fig. 1.** Estructura de la biblioteca Cudf [2].

Cudf contiene otras funciones que permiten operaciones adicionales con los usuarios que están familiarizados con Pandas, Numpy o SQL*,* la biblioteca cudf contiene funciones de ingreso y lectura de datos, estos permiten poder leer archivos csv, json, txt, Apache Parquet, entre otros, directamente en un gdf.

## Biblioteca Cuml

E

## Las tablas

Al igual que las figuras, las tablas deberán ir numeradas correlativamente, pero el título descriptivo de la tabla figurará encima de ésta. Por ejemplo, en la tabla 1 se muestra el número de créditos de cada asignatura de un plan de estudios.

**Tabla 1**. Número de créditos por asignatura del plan de estudios.

|  |  |
| --- | --- |
| **Asignatura** | **Créditos** |
| Arquitectura e Ingeniería de Computadores | 9 |
| Compiladores I | 4,5 |
| Ingeniería del Software I | 7,5 |
| Redes de Computadores | 4,5 |
| Administración de Empresas | 4,5 |
| Métodos Cuantitativos de Organización I | 4,5 |

Las tablas también deberán llevar asociado un texto alternativo para garantizar la accesibilidad del documento. Esto se hace en propiedades de tabla, texto alternativo.

# Los resultados

Este es el apartado en el que mostrar los frutos del trabajo de investigación que estamos describiendo. Dependiendo del tipo de trabajo, este apartado tomara diferentes nombres. Si se trata de un trabajo de investigación cuantitativa, se llamará “resultados” o bien “resultados y discusión” puesto que en él se mostrarán los resultados obtenidos y la interpretación que, de los mismos y de sus consecuencias, hacen los autores, así como las posibles discordancias con otros trabajos anteriores publicados.

Si se trata de una investigación cualitativa, el desarrollo de un prototipo, una innovación introducida en un proceso, etc., el título de este apartado podrá ser diferente, pero en todo caso será aquí donde se muestre el avance que supone el trabajo realizado, en relación con otros trabajos anteriores, exponiendo el alcance de lo conseguido, sus fortalezas y debilidades y las posibles consecuencias teóricas y prácticas que se derivan del mismo [5].

El principal objetivo de un congreso es la transferencia de conocimiento entre especialistas en una materia. Por ello es importante que el artículo aporte información suficiente para que las ideas, métodos, técnicas y resultados de la investigación puedan servir a otros especialistas en sus trabajos y también para que se pueda reproducir la experiencia o adaptarla a otros contextos diferentes.

# La presentación

El artículo presentado será objeto de un proceso de revisión por expertos en la materia tratada en el mismo, que harán llegar sus comentarios y observaciones a los autores. En muchas ocasiones pedirán que se lleven a cabo determinadas modificaciones, como condición para ser aceptado, por lo que es normal el tener e redactar y enviar una versión final.

Llegado el momento de la celebración del congreso, el autor o uno de los autores, si son varios, deberá exponerlo públicamente a los asistentes. Por ello, es importante planificar y preparar cuidadosamente una presentación del trabajo, apoyada normalmente en la proyección de diapositivas en una pantalla.

En este punto se deben tener en cuenta algunos factores que pueden influir en la preparación de la presentación:

* La audiencia. Cuántas personas estarán presentes en la exposición y cuál es su nivel de conocimientos en la materia del trabajo que vamos a exponer.
* El tiempo disponible. Tanto para la exposición como para el turno de preguntas y debate posterior. Como regla general, se planificará una diapositiva por cada minuto.
* Los medios técnicos. Ordenador, proyector, control remoto de la presentación, puntero láser, micrófono, acceso a Internet, etc.

El orador debe ser capaz de describir con rigor la naturaleza del trabajo que ha realizado, pero tratando, al mismo tiempo, de transmitir un sentimiento de entusiasmo y convencimiento a la audiencia.

La presentación debe constar de una introducción, en la que se expongan los objetivos del trabajo y las premisas de las que se ha partido, un nudo, donde se explique cómo se ha desarrollado el trabajo, los pasos que se han seguido, etc., y unas conclusiones se resalten las principales aportaciones que se han hecho.

Las diapositivas de apoyo deben tener un diseño claro y sencillo y contener solamente párrafos simples y figuras y esquemas representativos. Su objetivo es retener la atención de la audiencia y servir de guía al presentador para seguir el discurso narrativo planificado. Los fondos oscuros y recargados, las animaciones excesivas y diapositivas recargadas de textos o de datos, distraen la atención de los asistentes. Los detalles del trabajo que se está exponiendo lo pueden encontrar los interesados en el artículo escrito.

La exposición debe hacerse de frente al público, hablando para todos, sin movimientos excesivos, con un tono de voz claro y confiado, sin usar muletillas y transmitiendo en todo momento seguridad en sí mismo y dominio de todos los detalles que conciernen al desarrollo del trabajo que se está exponiendo.

# Conclusiones

Es el momento de poner de relieve la importancia de los resultados obtenidos y la contribución al conocimiento del trabajo presentado.

En este apartado no se trata de resumir la investigación expuesta en los apartados anteriores, sino de exponer las consecuencias teóricas del trabajo y sus posibles aplicaciones prácticas.

En muchas ocasiones, el artículo presentado muestra los resultados obtenidos en una investigación enmarcada en un proyecto de más envergadura, por lo que es común llamar a este apartado “conclusiones y trabajos futuros”. Se describen entonces los pasos a seguir y los futuros trabajos a realizar a partir de los resultados obtenidos o de los productos desarrollados en éste.

También en las conclusiones pueden resaltarse los resultados obtenidos, como fuente para nuevas hipótesis y sugerir nuevas líneas de investigación.

# Referencias

Las referencias no deben confundirse con una bibliografía general sobre el tema tratado en el artículo. Se trata de una relación de las fuentes de información que se han utilizado como apoyo de las afirmaciones que se hacen en el texto y que aparecen señaladas como citas en el mismo. Solo deben citarse textos que hayan sido publicados y cada referencia debe contener todos los datos que permitan al lector llegar hasta el texto completo.

Las referencias deben incluir el autor o autores, año de publicación, título, editorial, número de la revista en su caso, etc. Si solo están publicados en Internet, y debido a la volatilidad de la información, además de la URL debe especificarse la fecha en la que se hizo la consulta.

Existen estilos, como APA [6] sobre cómo escribir referencias de libros, artículos, tesis, etc.

La lista de referencias puede estar numerada en orden de aparición de las citas en el texto o bien ordenada alfabéticamente. Para artículos de corta extensión, como en este caso, es muy común numerar las citas.

1. Kallestinova, E.D. (2011). *How to Write Your First Research Paper*. Yale Journal of Biology and Medicine. 2011 Sep; 84(3): 181–190. [Publicado online](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178846/).
2. Moreno, F.; Marthe, N.; Rebolledo, L.A. (2010). *Cómo escribir textos académicos según normas internacionales: APA, IEEE, MLA, Vancouver e ICONTEC*. Ediciones Uninorte, Barranquilla (Colombia). 2010.
3. Springer. (2017). *Lecture Notes in Computer Science (LNCS)*. *Information for Authors*. <http://www.springer.com/gp/computer-science/lncs>. (Consultado el 22 de marzo de 2017).
4. Naciones Unidas (2015). *Objetivos de Desarrollo Sostenible: 17 objetivos para cambiar nuestro mundo*. <http://www.un.org/sustainabledevelopment/es/la-agenda-de-desarrollo-sostenible/> (Consultado el 22 de marzo de 2017).
5. García-Colmenarejo, A.I.; Rodríguez Rivollier, M.; Rodríguez-Sedano, M. y Madrid Liras, S. (2015). *Modelo de artículo académico-profesional para autores: Guía de Estilo y criterios APA de publicación*. Revista de Mediación Volumen 8 Nº 2. ISSN: 2340-9754. <http://revistademediacion.com>. (Consultado el 22 de marzo de 2017).
6. Wikipedia. *Estilo APA*. <https://es.wikipedia.org/wiki/Estilo_APA>. (Consultado el 22 de marzo de 2017).