# DEALER DATA INTEGRATION DDI

Daniel Fernando Areiza Agudelo

Informe de práctica para optar por el título de Ingeniería de Sistemas

Asesor
Sandra Patricia Zabala Orrego, Especialista en Gerencia

Universidad de Antioquia
Facultad de Ingeniería. Departamento de Sistemas
Ingeniería de Sistemas
Medellín
2023

| Cita | Areiza Agudelo [1] |
|---|---|
| **Referencia** | [1] D. F. Areiza Agudelo, “Dealer Data Integration DDI”, Presencial, Ingeniería de Sistemas, Universidad de Antioquia, Medellín, 2023. |
| Estilo IEEE (2020) | |




Inchcape Colombia [1]

## Dedicatoria

A mi madre quien siempre ha estado presente y apoyándome en mi vida.

**Agradecimientos**

A la empresa Inchcape Colombia por brindarme la oportunidad de trabajar con ellos y permitirme realizar mis prácticas, a mis compañeros de trabajo quienes me ayudaron en este proceso, a la Universidad de Antioquía que me permitió estudiar, a docentes, compañeros y amigos que aportaron en mi formación.

**TABLA DE CONTENIDO**

## LISTA DE FIGURAS

## SIGLAS, ACRÓNIMOS Y ABREVIATURAS

| | |
|---|---|
| **IEEE** | Institute of Electrical and Electronics Engineers. |
| **DDI.** | Dealer Data Integration. |
| **DDC.** | Digital Delivery Centre. |
| **DAP.** | Data Analytics Platform. |
| **PLC.** | Public Limited Company. |
| **PII** | Personal Identifiable Information. |
| **PL** | Pipeline. |
| **DE** | Data Engineer. |
| **ML** | Machine Learning. |
| **BI** | Business Intelligence. |
| **ETL** | Extract, Transform, Load. |
| **ACID** | Atomicidad, Coherencia, Aislamiento y Durabilidad. |
| **AWS** | Amazon Web Services. |
| **UdeA** | Universidad de Antioquia. |

# RESUMEN

El presente informe tiene como propósito ofrecer un análisis detallado del proceso llevado a cabo en el proyecto Dealer Data Integration para la empresa Inchcape PLC. Este proyecto se plantea con el objetivo de centralizar la información de diversos distribuidores en un solo lugar. Inicialmente, la implementación de este proyecto se llevará a cabo en Latinoamérica, comenzando con los distribuidores ubicados en Colombia, y posteriormente se extenderá a otras regiones.

El estado actual de recolección de información plantea desafíos, ya que cada concesionario mantiene sus datos de manera independiente, sin un orden ni un estándar definido. Esta situación genera ineficiencias en el proceso de recopilación y gestión de datos.

Para abordar esta problemática, se ha iniciado el diseño y desarrollo de una Pipeline [12], que desempeñará un papel fundamental en la recopilación y consolidación de la información en un único repositorio. La implementación de esta solución se llevará a cabo utilizando Azure Data Factory, el servicio de ETL en la nube de Azure, y Databricks. Tras rigurosas pruebas de funcionamiento y desempeño, la Pipeline se desplegará para su uso en producción.

La implementación exitosa de esta solución permitirá optimizar la eficiencia en la recopilación y gestión de datos, lo que contribuirá significativamente al logro de los objetivos de Inchcape PLC.

***Palabras clave*** — **Dealer, Data, ETL, Pipeline.**

# ABSTRACT

This report aims to provide a detailed analysis of the process carried out in the Dealer Data Integration project for Inchcape PLC. This project is designed with the goal of centralizing information from various dealers in one place. Initially, the implementation of this project will take place in Latin America, starting with dealers located in Colombia, and it will later expand to other regions.

The current state of information collection presents challenges, as each dealership maintains its data independently, without a defined order or standard. This situation leads to inefficiencies in the data collection and management process.

To address this issue, the design and development of a Pipeline have been initiated, which will play a key role in collecting and consolidating information into a single repository. The implementation of this solution will be carried out using Azure Data Factory, Azure's cloud ETL service, and Databricks. After rigorous testing for functionality and performance, the Pipeline will be deployed for production use.

The successful implementation of this solution will optimize the efficiency of data collection and management, significantly contributing to the achievement of Inchcape PLC's objectives.

***Keywords*** — **Dealer, Data, ETL, Pipeline.**

# I. INTRODUCCIÓN

La empresa Inchcape Colombia, que forma parte del grupo Inchcape PLC, se dedica a la comercialización, distribución y venta de vehículos, así como a ofrecer servicios de postventa y repuestos a nivel internacional. En los últimos 2 años, ha experimentado un gran crecimiento en Latinoamérica.

Para mantenerse a la vanguardia, la empresa ha creado los Digital Delivery Centres (DDCs) con dos sedes, una en Filipinas y otra en Colombia. Estos centros están compuestos por 7 equipos, incluyendo el equipo de Data Analytics Platform (DAP), que a su vez está formado por los equipos de Demand and Product Management, Data Science, Data Engineering, Bussines Intelligence Development, Platform and Enterprise Solutions y Programme Management.

El equipo de Data Engineering (DE) es responsable de reunir y consolidar toda la información de la empresa para asegurar que se tenga una base de datos con la información requerida para cada proyecto en un mismo lugar. Para ello, utilizan un servicio de Data Lakehouse [8], que integra las ventajas de un Data Lake y un Data Warehouse, permitiendo una mayor flexibilidad al combinar las transacciones ACID [11] de los Data Warehouse con la flexibilidad y rentabilidad de los Data Lake. Esto permite aplicar Business Intelligence (BI) y Machine Learning (ML) en los datos.

Para llevar a cabo esta recolección de información, la empresa implementa el uso de pipelines (tuberías de datos), que consisten en una serie de pasos o tareas cuyo objetivo es extraer, transformar y cargar datos de diferentes fuentes a un destino de almacenamiento de datos, en este caso, el almacenamiento se realiza en el Data Lakehouse hospedado en la nube de Microsoft Azure.

# II. OBJETIVOS

## I. OBJETIVO GENERAL

Automatizar y disponibilizar la información proveniente de diferentes fuentes. Con el propósito de integrar en un repositorio de datos la información de los distribuidores en diferentes mercados latinoamericanos, iniciando por Colombia. Esto permitirá capturar y consolidar información estratégica para su uso en diferentes procesos, como planificación, optimización de inventario, marketing, experiencia del cliente, retención y ventas. La información centralizada en DDI se presentará en tableros y tablas de datos, lo que ayudará a reducir la carga operativa, la escritura manual, el riesgo de perder información histórica y las inconsistencias de datos.

## II. OBJETIVOS ESPECÍFICOS

1. Obtener formación y adquirir conocimientos a través de la empresa sobre el manejo de las herramientas Azure Data Factory y Azure Databricks.
2. Aportar a las soluciones para consolidar y disponibilizar la información en el proyecto de DDI.
3. Diseñar y crear la Pipeline principal para DDI.
4. Monitorear el correcto funcionamiento de la pipeline cuando está ya se esté ejecutando.

# III. MARCO TEÓRICO

Para comprender el desarrollo de pipelines en Azure, es importante tener en cuenta que este es un servicio en la nube [10]. Estos servicios son proporcionados por varias plataformas, como Microsoft Azure, AWS y Google Cloud, que ofrecen servicios informáticos como servidores, almacenamiento, bases de datos y software a través de Internet, también conocido como la nube.

Las pipelines en Azure se crean utilizando Azure Data Factory [7], un servicio de integración de datos sin servidor que permite la creación y programación de flujos de trabajo. Estas pipelines son una secuencia de actividades diseñadas para realizar una tarea específica, como recopilar datos de diferentes fuentes de información, transformarlos y disponibilizarlos para su posterior uso, lo que se conoce como ETL [9] de datos.

Azure Databricks [3][6], proporciona un conjunto unificado de herramientas que permiten la creación y modificación de notebooks utilizados en las pipelines para el trabajo de Data Engineering. Esto se logra aprovechando su integración con el almacenamiento y la seguridad de Azure para generar soluciones que se puedan adaptar a cada necesidad.

# IV. METODOLOGÍA

Se inicia el proyecto realizando el planteamiento del problema, así como la necesidad de una solución. De igual manera se indica el contexto de este de una manera más detallada donde se indica que se requiere por parte del equipo DE para el proyecto, esto se lleva a cabo en conjunto con los interesados en el proyecto.

Se procede entonces a diseñar la PL que será usada para llevar a cabo la solución a los requerimientos del proyecto, todo este diseño se realiza usando la herramienta Azure Data Factory [2][4], de Microsoft, la cual hace uso de notebooks creados con la herramienta Databricks con código en PySpark y SQL.

Una vez se finaliza con el primer prototipo de la PL se empiezan a realizar las pruebas necesarias para validar el correcto funcionamiento de esta con una conexión para prueba. Al finalizar estas pruebas de manera exitosa se identifica que es necesario realizar un proceso de cifrado de información que es considerada sensible (PII) y es necesario el desarrollo de una PL adicional para este proceso.

Una vez se finalizó el desarrollo de las PL, se realizaron las pruebas para certificar el correcto funcionamiento de todo el proceso y que se cumpla con lo requerido. Finalmente se realiza la solicitud de despliegue de la PL a producción para que sea ejecutada cada día de manera que esté disponible la información como lo solicito el product owner.

# V. RESULTADOS

**OBJETIVO ESPECÍFICO 1:**

Obtener formación y adquirir conocimientos a través de la empresa sobre el manejo de las herramientas Azure Data Factory y Azure Databricks.

Para el cumplimiento de este objetivo se remitió al uso de cursos en línea [2][3], así como también a documentación en línea [4][5], donde se dedicó un tiempo prudente antes de iniciar con el proyecto de DDI. Durante este tiempo también se recibió a través de reuniones con las partes involucradas en el proyecto una mejor explicación de este y que se esperaba entregar al final.

**OBJETIVO ESPECÍFICO 2:**

Aportar a las soluciones para consolidar y disponibilizar la información en el proyecto de DDI.

Luego de terminado el ciclo de integración al proyecto donde se decide que se buscará un enfoque hibrido entre la metodología ágil Kanban con elementos de gestión, donde se decide tener una comunicación constante y revisión del progreso del proyecto; este se dará con una periodicidad no superior a dos semanas, con el fin de llevar un control del progreso de tareas y el avance del proyecto en general.

**OBJETIVO ESPECÍFICO 3:**

Diseñar y crear la Pipeline principal para DDI [Fig. 1].

Para esto tenemos primero que una Pipeline es una serie de paso o tareas que se usan para extraer, transformar y cargar datos que provienen de varias fuentes y se consolidan en un solo destino para almacenarlos.

Durante el diseño y el desarrollo de la PL se procede a usar la herramienta Data Factory donde se crea el esquema [Fig. 1] inicial para el funcionamiento de esta, el cual consiste en ejecutar una serie de pasos en un orden definido de izquierda a derecha.

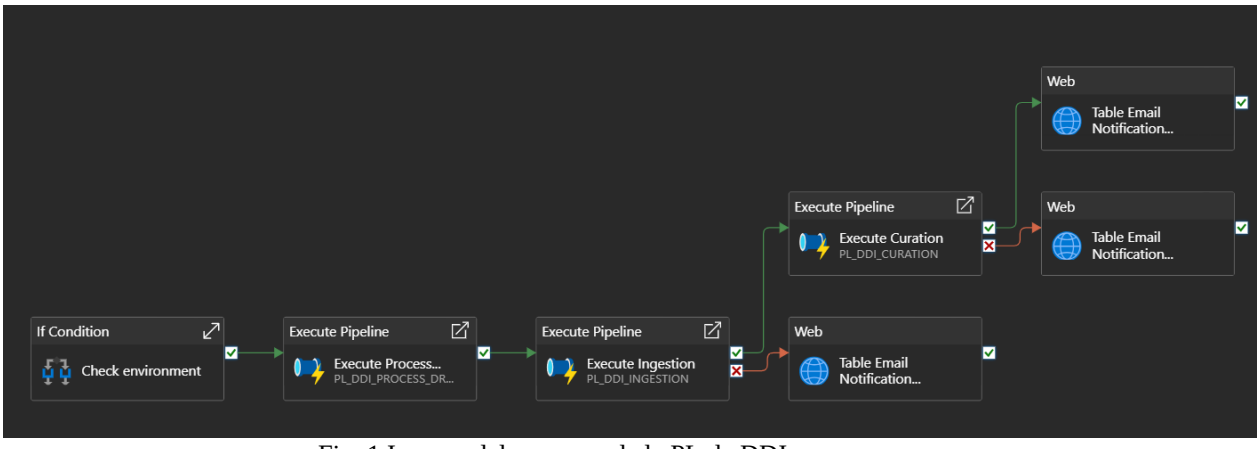


Fig. 1 Imagen del esquema de la PL de DDI

Cada una de estas fases está llamando notebooks [Fig. 2] que son ejecutados en Databricks y esos son los que contienen las indicaciones o instrucciones específicas de lo que se debe realizar en cada uno de los pasos.

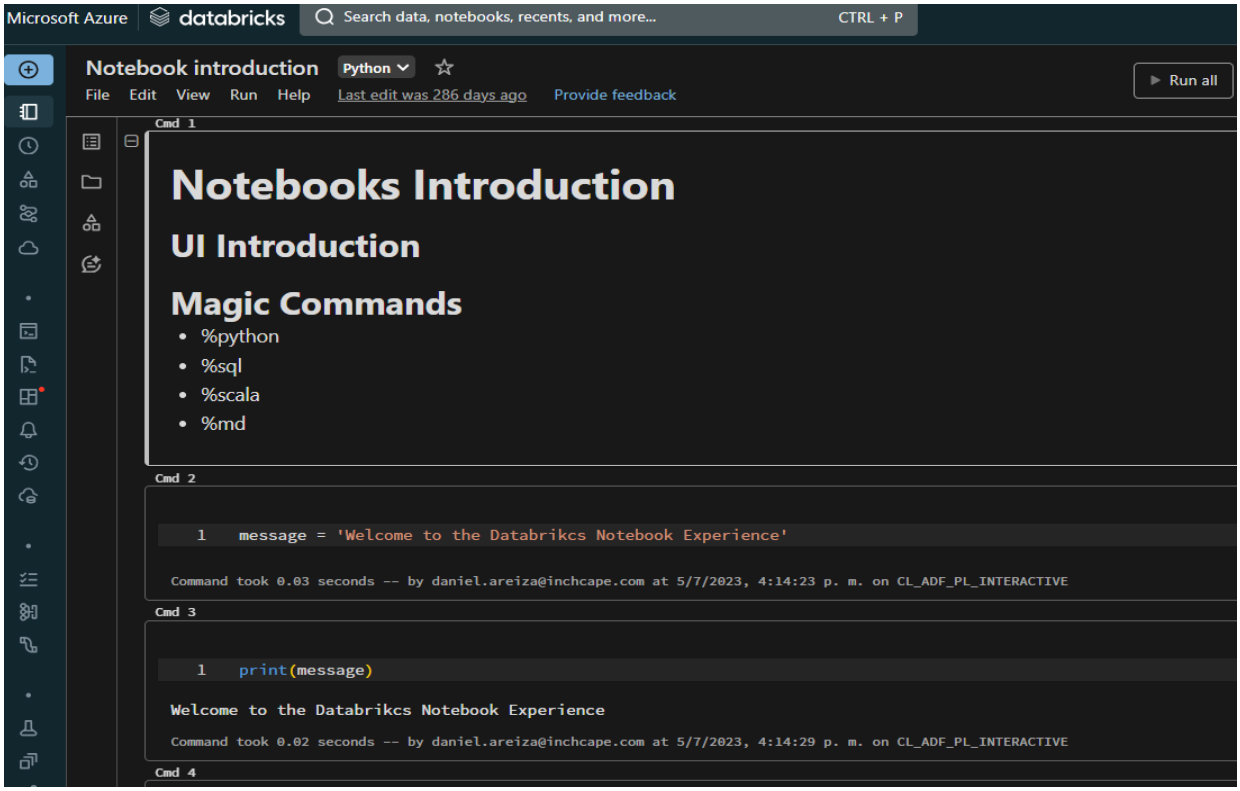


Fig. 2 Imagen notebook Databricks

Para poder acceder a cada uno de estos notebooks se hace necesario definir los parámetros [Fig. 3] que indican el lugar donde estos se encuentran alojados (Default Value), para que de esta manera cada uno de los pasos de la PL sean ejecutados de la manera requerida.

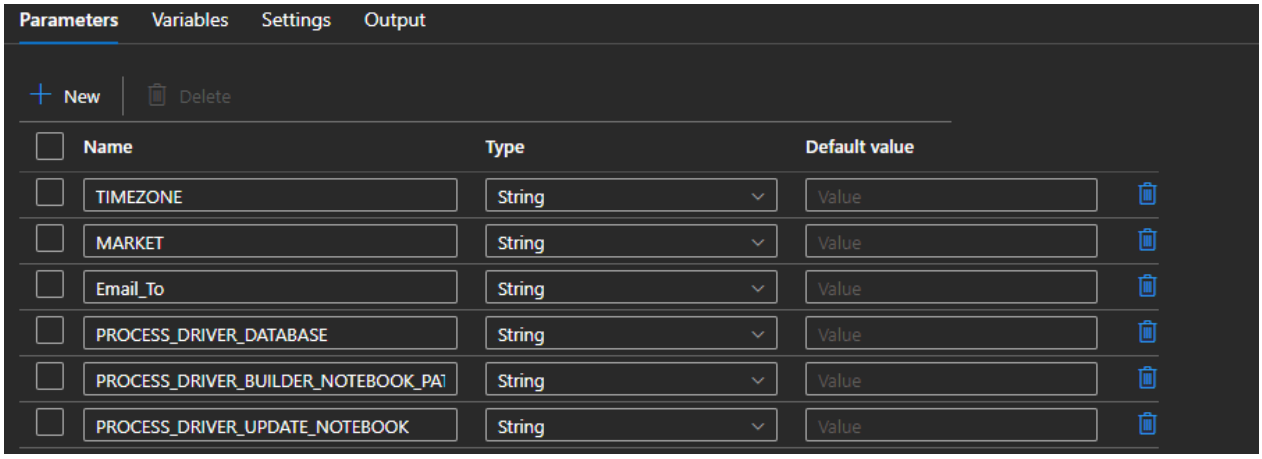


Fig. 3 Parámetros

Dentro de estos pasos tenemos el Process Driver Builder que es donde podemos definir los argumentos de entrada y salida, crear y configurar tareas, permite establecer y configurar variables de entorno y los recursos necesarios.

El proceso de ingesta es el paso donde se obtienen los datos de diferentes fuentes. Estos comprenden extracción de datos de bases de datos, archivos, sistemas de mensajería, aplicaciones web, entre otros, es también en este punto donde se realiza la configuración de las conexiones a las fuentes de datos y los datos a leer.

Luego de ser necesario se pasa al proceso de Transformación, (No fue necesario para DDI) donde los datos se procesan y se transforman para crear información útil y significativa, se realizan cálculos con los datos disponibles y de acuerdo con lo requerido. Finalmente se pasa al proceso de Curado, donde Los datos se limpian, validan y preparan para el análisis posterior, en este punto es donde se realiza lo siguiente:

- Eliminación de valores nulos o duplicados,
- Conversión de formatos de datos y la validación de datos con reglas de negocio.
- Combinación de datos de diferentes fuentes o la creación de estructuras de datos más complejas.

**OBJETIVO ESPECÍFICO 4:**
Monitorear el correcto funcionamiento de la pipeline cuando está ya se esté ejecutando.

Para llevar a cabo este monitoreo [Fig. 4] de una manera adecuada se hace uso de los Status logs, estos son registros que se generan para registrar el estado de las actividades y sus resultados, esto nos brinda información sobre el éxito o fracaso de alguna actividad durante alguno de los pasos de la PL, nos da información de los errores o excepciones que se producen y de esta manera se puede realizar un Monitoreo y diagnóstico de problemas de una manera efectiva.

Finalmente se entrega la PL funcionando luego de haber realizado ejecuciones manuales de esta y un monitoreo para validar que no tenga errores para crear el o los Trigger/Disparador necesarios, que será el encargado de iniciar la ejecución de la PL, este puede ser configurado para ejecutarse sea porque se cumplió con una condición de un evento o porque se le dio una programación específica para que inicie.

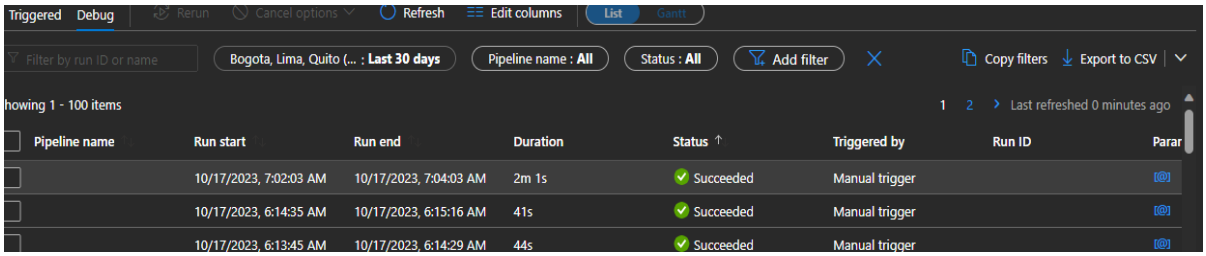

Fig. 4 Monitoreo

# I. CONCLUSIONES

Es importante al comienzo de un proyecto, tener conocimiento de las herramientas que se necesitan para su desarrollo y el alcance de este. De esta manera, cuando sea necesario realizar algún cambio, se tendrá claro hasta dónde se tiene permitido llegar.

Es recomendable realizar una distribución de las tareas que se van a desarrollar. Para lograrlo, es aconsejable hacer uso de una metodología ágil. En este caso, se aplica un enfoque híbrido entre Kanban y elementos de gestión. Esto permite mantener un mejor control del estado del proyecto, ya que se sabe qué se ha completado y qué está pendiente. Esta metodología contribuye a desarrollar el alcance del proyecto de la mejor manera.

Asimismo, la comunicación con las partes implicadas en el proyecto es esencial. Esto permite asegurarse de que se está cumpliendo con lo requerido y de que no se agregarán nuevos requerimientos sin haberlos discutido previamente con los implicados. Esto es crucial para evitar incumplir los objetivos estipulados desde el principio.

Es fundamental llevar a cabo pruebas a lo largo de todo el proceso de desarrollo del proyecto. Esto no solo ayuda a encontrar posibles fallas, sino que también brinda retroalimentación para todos los involucrados. Si se presentan retos, es preferible afrontarlos en estas etapas tempranas en lugar de esperar hasta llegar a un punto muy avanzado del proyecto.

# REFERENCIAS


[1] Bienvenido a Inchcape Colombia - Inchcape Colombia (ES). (2022, 15 diciembre). Inchcape Colombia (ES). https://www.inchcape.com/es-co/.

[2] Ramesh Retnasamy. (s.f.). Azure Data Factory for Data Engineers [En línea]. Disponible en: https://bit.ly/3tvlD2N.

[3] Ramesh Retnasamy. (s.f.). Azure Databricks for Data Engineers [En línea]. Disponible en: https://bit.ly/3FbF0Az.

[4] Microsoft. (s. f.). Azure Data Factory Documentation - Azure Data Factory. Microsoft Learn. [En línea]. Disponible en: https://bit.ly/3FfPizm.

[5] Microsoft. (s. f.). Introducción a Azure Data Factory - Azure Data Factory [En línea]. Disponible en: https://bit.ly/45nVZKl.

[6] Microsoft. (s. f.). Azure Databricks documentation. Microsoft Learn [En línea]. Disponible en: https://bit.ly/3QbJGwy.

[7] Microsoft. (s. f.). Azure Data Factory: servicio de integración de datos. [En línea]. (s. f.). Disponible en: https://bit.ly/46JzikR.

[8] Microsoft. (s.f.). Azure Databricks: Lakehouse. [En línea]. Disponible en: https://bit.ly/45uBR9F.

[9] Microsoft. (s.f.). Extracción, transformación y carga de datos (ETL) [En línea]. Disponible en: https://bit.ly/3PVKXXd.

[10] Microsoft Azure. (s.f.). ¿Qué es la computación en la nube? [En línea]. Disponible en: https://bit.ly/3FfaEwW.

[11] Microsoft Azure. (s.f.). ¿Cuáles son las garantías ACID en Azure Databricks? [En línea]. Disponible en: https://bit.ly/46YeuXb.

[12] Microsoft Azure. (s.f.). Azure Pipelines documentation [En línea]. Disponible en: https://bit.ly/46RRAke.