Análisis de las Fases de EDA y Procesamiento en Big Data

Asistente Experto en Big Data

29 de noviembre de 2025

Resumen

Este documento sintetiza la conversación mantenida con un asistente de IA especializado en Big Data, donde se detallan las tareas, objetivos y herramientas clave de dos fases fundamentales en el ciclo de vida de la ciencia de datos: el Análisis Exploratorio de Datos (EDA) y la fase de Procesamiento de Datos. El contenido está orientado a servir como guía de referencia para profesionales del sector.
https://medium.com/inside-intelligence/flujo-completo-de-machine-learning-95a1c8219296

Índice

1. Introducción **1**

2. Fase 1: Análisis Exploratorio de Datos (EDA) **2**
2.1. Objetivo Principal 2
2.2. Tareas Clave Realizadas 2
2.3. Herramientas Comunes 4
2.4. Consideración en Big Data 4

3. Fase 2: Procesamiento de Datos **4**
3.1. Objetivo Principal 4
3.2. Relación con el EDA 4
3.3. Tareas Clave Realizadas 4
3.4. Flujo de Trabajo en Procesamiento de Datos 6
3.5. La Diferencia en Big Data: Escala y Herramientas 6

4. Conclusión **6**

5. Ciencia de datos **6**

1. Introducción

Este documento recopila y estructura la información generada durante una consulta sobre las fases de trabajo en proyectos de Big Data. La consulta se centró específicamente en el

Análisis Exploratorio de Datos (EDA) y la fase de Procesamiento de Datos, que son etapas consecutivas y críticas para el éxito de cualquier iniciativa de data science o analytics.

2. Fase 1: Análisis Exploratorio de Datos (EDA)

2.1. Objetivo Principal

El objetivo principal del EDA es entender la estructura, calidad y relaciones inherentes en los datos para guiar el análisis posterior y la selección de modelos. Se trata de una fase de **diagnóstico** donde se identifican patrones, anomalías y se formulan hipótesis iniciales.

2.2. Tareas Clave Realizadas

A continuación, se enumeran las tareas fundamentales ejecutadas durante esta fase:

1. **Recolección y Carga de Datos:**
   - Cargar los *datasets* desde diversas fuentes (BBDD, archivos CSV, JSON, APIs, data lakes como HDFS).
   - En Big Data, esto implica el uso de herramientas como **Spark, Hive, Impala** o frameworks distribuidos para manejar volúmenes masivos.
   - WebScraping
2. **Descripción General del Dataset:**
   - Obtener dimensiones (número de filas y columnas).
   - Verificar tipos de datos (numéricos, categóricos, fechas, etc.).
   - Generar estadísticas descriptivas básicas (media, mediana, desviación estándar, mínimos, máximos, percentiles) para variables numéricas. (Realizar ejercicios con estos cálculos esdísticos)
3. **Limpieza y Preprocesamiento Inicial:**
   - Identificar y manejar valores missing (nulos).
   - Detectar y tratar valores duplicados.
   - Corregir tipos de datos incorrectos (e.g., fechas representadas como *strings*).
4. **Análisis de Variables Individuales (Univariado):**
   - Para variables numéricas: Histogramas, boxplots, gráficos de densidad.
   - Para variables categóricas: Conteo de frecuencias, gráficos de barras, moda.
   - Identificar sesgos, *outliers* o distribuciones anómalas.

5. **Análisis de Relaciones entre Variables (Bivariado/Multivariado):**
    - Entre dos variables numéricas: Gráficos de dispersión (scatter plots), análisis de correlaciones (Pearson, Spearman).
    - Entre una variable numérica y una categórica: Boxplots por categoría, ANOVA.
    - Entre dos variables categóricas: Tablas de contingencia, gráficos de barras apiladas, prueba chi-cuadrado.
    - Elaborar matrices de correlación para variables numéricas.
6. **Detección y Tratamiento de Outliers:**
    - Identificar *outliers* usando métodos estadísticos (rango intercuartílico - IQR, puntuaciones Z).
    - Decidir si eliminarlos, transformarlos o tratarlos en función del contexto del negocio.
7. **Transformación de Variables:**
    - Aplicar normalización (Min-Max, Z-score) o estandarización.
    - Usar transformaciones (log, sqrt) para corregir sesgos en distribuciones.
    - Codificación de variables categóricas (one-hot encoding, label encoding).
    - Creación de nuevas características (*feature engineering*) si es necesario.
8. **Análisis de Componentes Principales (PCA) o Reducción de Dimensionalidad (Opcional):**
    - En *datasets* con muchas variables, usar PCA para visualizar patrones en 2D/3D.
    - Identificar variables que explican la mayor varianza.
9. **Visualización de Datos:**
    - Utilizar librerías como **Matplotlib, Seaborn, Plotly** (en Python) o herramientas como **Tableau, Power BI**.
    - Crear *dashboards* interactivos si es necesario.
    - Generar visualizaciones clave: histogramas, boxplots, scatter matrices, heatmaps de correlación.
10. **Formulación de Hipótesis y Insights Preliminares:**
    - Documentar observaciones: tendencias, patrones, relaciones inesperadas.
    - Plantear hipótesis para validar en fases posteriores.

11. **Validación de Calidad de Datos:**
    - Asegurar que los datos cumplen con las expectativas y reglas del negocio.
    - Verificar consistencia (e.g., fechas futuras en campos de nacimiento).
12. **Documentación y Reporting:**
    - Generar un reporte con hallazgos (e.g., Jupyter Notebook, R Markdown).
    - Incluir visualizaciones, código y explicaciones.

2.3. Herramientas Comunes

- **Lenguajes:** Python (Pandas, NumPy, SciPy), R, Scala.
- **Frameworks Distribuidos:** Apache Spark (con PySpark o SparkR) para *datasets* muy grandes.
- **Visualización:** Matplotlib, Seaborn, Plotly, Tableau, Power BI.
- **Entornos:** Jupyter Notebook, Zeppelin, Databricks, Google Colab.

2.4. Consideración en Big Data

Dado el volumen, variedad y velocidad de los datos en Big Data, el EDA a menudo se realiza en **entornos distribuidos** usando técnicas de muestreo (si el *dataset* es demasiado grande) o computación en *cluster* para agilizar los cálculos.

3. Fase 2: Procesamiento de Datos

3.1. Objetivo Principal

El objetivo de esta fase es transformar los datos crudos en un formato limpio, estructurado y listo para el modelado. Es donde se **implementan las decisiones tomadas durante el EDA** para forjar un conjunto de datos de alta calidad.

3.2. Relación con el EDA

- **EDA:** Es el **diagnóstico**. Se descubren los problemas (valores nulos, *outliers*, etc.).
- **Procesamiento:** Es la **cirugía y el tratamiento**. Se aplican las soluciones para arreglar los problemas encontrados.

3.3. Tareas Clave Realizadas

1. **Limpieza de Datos (Data Cleaning) - Avanzada:**
    - **Manejo de Valores Faltantes:**

        - **Eliminación:** Quitar filas/columnas con muchos nulos.
        - **Imputación:** Rellenar valores usando mediana/media/moda, o algoritmos predictivos (KNN).
        - **Asignación de "Desconocido":** Para variables categóricas.
    - **Manejo de Outliers:**
        - **Transformación:** Aplicar log, sqrt, etc.
        - **Discretización (Binning):** Convertir variables numéricas en categóricas.
        - **Capping:** Limitar valores a un percentil superior/inferior.
        - **Eliminación:** Solo si son errores confirmados.
2. **Transformación y Normalización:**
    - **Codificación de Variables Categóricas:**
        - **One-Hot Encoding:** Ideal para la mayoría de algoritmos.
        - **Label Encoding:** Usado principalmente para árboles de decisión.
        - **Target Encoding:** Poderoso pero con riesgo de *overfitting*.
    - **Escalado/Normalización:**
        - **Normalización (Min-Max):** Escala a un rango fijo [0, 1]. Crucial para SVM y Redes Neuronales.
        - **Estandarización (Z-score):** Media=0, Desviación Estándar=1. Ideal para regresión, logística, k-means.
3. **Ingeniería de Características (Feature Engineering):**
    - **Creación de Features:**
        - De **fechas:** Día de la semana, mes, hora, etc.
        - De **texto:** Longitud, número de palabras, sentimiento (NLP).
        - **Agregaciones:** (e.g., "gasto promedio del último mes").
        - **Interacciones:** Multiplicación/división entre features (e.g., BMI = peso/altura^2).
4. **Reducción de la Dimensionalidad:**
    - **Selección de Características:**
        - Eliminar variables con **varianza baja**.

- Eliminar variables **altamente correlacionadas**.
- Usar métodos estadísticos (Chi-cuadrado) o basados en modelos (Importancia en Random Forest).

- **Extracción de Características:**
  - **Análisis de Componentes Principales (PCA):** Crear nuevas variables no correlacionadas que capturan la varianza máxima.

5. **División de los Datos (Data Splitting):**
   - **Conjunto de Entrenamiento (Training Set):** (70-80 %) Para enseñar al modelo.
   - **Conjunto de Validación (Validation Set):** (10-15 %) Para ajustar hiperparámetros.
   - **Conjunto de Prueba (Test Set):** (10-15 %) Usado **UNA sola vez** para evaluación final. Simula el mundo real.

3.4. Flujo de Trabajo en Procesamiento de Datos

(Diagrama de Flujo: Datos Crudos ->Limpieza ->Transformación ->Ingeniería ->Reducción ->División ->Datos Listos)

3.5. La Diferencia en Big Data: Escala y Herramientas

En Big Data, estas tareas no se hacen en una sola máquina. Se implementan como **pipelines de datos automatizados y distribuidos** usando frameworks como:

- **Apache Spark (Spark SQL, MLlib):** Estándar para procesamiento distribuido a gran escala.
- **Apache Beam / Google Dataflow:** Para pipelines por lotes (*batch*) y en *streaming*.
- **TFX (TensorFlow Extended):** Para pipelines de ML en producción.

4. Conclusión

La fase de **EDA** es fundamental para diagnosticar el estado de los datos y formular hipótesis. La fase de **Procesamiento** es donde se aplica el tratamiento para convertir los datos crudos en un insumo de alta calidad listo para el modelado. La calidad del trabajo en estas fases es often más determinante para el éxito de un proyecto que la elección del algoritmo de modelado en sí mismo. En el contexto de Big Data, la capacidad de ejecutar estas tareas de manera distribuida y eficiente es un requisito indispensable.

5. Ciencia de datos