**5. Código y/o Servicio Utilizado para el Procesamiento de la Información y su Visualización**

#### Procesamiento de la Información

Para el procesamiento de datos, se desarrollaron scripts en **PySpark**, lo cual permitió transformar, limpiar y enriquecer los datos obtenidos desde diversas fuentes (estructuradas y no estructuradas). Estos datos fueron inicialmente recolectados mediante scrapers y almacenados como archivos en formato JSON, para posteriormente ser cargados, procesados y almacenados en **MongoDB**.

##### Tecnologías empleadas:

* **PySpark**: Motor de procesamiento distribuido que permite trabajar con grandes volúmenes de datos en paralelo.
* **MongoDB**: Base de datos NoSQL orientada a documentos, ideal para manejar datos semi-estructurados como los provenientes de redes sociales.
* **Conectores Mongo-Spark**: Facilitan la lectura y escritura eficiente entre PySpark y MongoDB.

La combinación de PySpark y MongoDB permite una solución escalable, ágil y flexible. PySpark maneja el procesamiento masivo de datos, mientras MongoDB almacena la información de forma eficiente, lista para ser visualizada o analizada.

##### Flujo de procesamiento ETL:

1. **Extracción**: Datos recolectados desde APIs (Reddit, Twitter, Alpha Vantage) mediante scrapers, almacenados como archivos JSON.
2. **Transformación**: Con PySpark se aplican filtros, conversiones de formato (por ejemplo, timestamps) y selección de campos relevantes.
3. **Carga**: Los datos procesados son almacenados en diferentes colecciones de MongoDB para su posterior análisis.
4. **Visualización**: Metabase consulta estas colecciones para generar dashboards dinámicos.

##### Ejemplo de Código PySpark para Procesamiento de Reddit:



#### Visualización de la Información

Para la visualización de los datos, se utilizó **Metabase**, una herramienta de Business Intelligence de código abierto que permite crear dashboards interactivos conectándose directamente a MongoDB.

##### Configuración:

* **Servidor MongoDB**: 127.0.0.1
* **Puerto**: 27017
* **Base de datos**: bigdata\_db
* **Colecciones utilizadas**:  
  + alpha\_processed (datos financieros estructurados)
  + twitter\_processed (tweets analizados)
  + reddit\_processed (posts relevantes)

##### Tipos de gráficos generados:

* Gráfico de líneas: Evolución diaria del precio de cierre de acciones (Alpha Vantage).
* Gráfico de barras: Tendencias en menciones tecnológicas en Twitter.
* Tablas dinámicas: Comentarios vs. puntuación en Reddit.
* Tarjetas resumen: Métricas clave del comportamiento en redes.

##### Actualización y automatización:

El procesamiento de datos se realiza automáticamente cada 5 minutos mediante tareas programadas, garantizando que los dashboards reflejen información en tiempo real. Metabase se sincroniza con MongoDB en intervalos configurables.

#### Validación y escalabilidad

Durante el desarrollo se realizaron pruebas de integridad y consistencia para garantizar que el filtrado no eliminara datos relevantes. Se validó que los datos procesados conservaran estructura, formatos de fechas y campos clave. La solución es escalable, y se proyecta el uso futuro de herramientas como **Apache Airflow** para orquestación, o **Apache Kafka** para streaming en tiempo real.

#### Conclusión

Este sistema integra herramientas modernas como PySpark, MongoDB y Metabase para ofrecer una solución completa desde la recolección hasta la visualización de datos. Gracias a su estructura escalable y modular, permite visualizar información actualizada minuto a minuto y tomar decisiones informadas basadas en grandes volúmenes de datos.