Este documento se etiqueta como “No compartible” por el hecho de que además de las consignas necesarias para cada parte del laboratorio, algunas de estas partes cuentan con la solución específica para la problemática planteada.

Las soluciones se encuentran en:

- Anexo 1: Lab - Hadoop
- Anexo 2: Lab - Spark
- Anexo 3: Lab - NiFi

Introducción

En el presente documento se les presenta la segunda opción por la que es posible optar para el desarrollo del laboratorio, aquí tendrán a disposición la descripción de la problemática y una descripción breve de los recursos con los cuales cuentan.

Además de las cuestiones a resolver que son específicas para este escenario propuesto, contarán con ejercicios prácticos extras, que serán propios de las unidades de Hadoop y Spark. Estos los pueden encontrar en los anexos correspondientes.

El lab se separa en dos partes:

1. En la sección problemática se detalla una consigna general y una específica para el documento.
2. En los anexos se describirán ejercicios extras que deberán ser completados para las unidades de Hadoop, Spark y NiFi.

Recomendación: realizar primeramente los ejercicios de Hadoop y Spark. ¿Por qué? Porque al atacar a la problemática específica planteada en este documento, tendrán que tener ciertos recursos de los que se hablan en los ejercicios complementarios de Hadoop y Spark. Esto no es obligatorio, pero ciertamente será de ayuda.

Problemática

Consigna general: se debe implementar una solución integral que dé respuesta a las necesidades de información, estas necesidades van desde la ingesta de datos hasta la visualización de los mismos, pasando por etapas de procesamiento y enriquecimiento.

La problemática a resolver se da en un entorno de movimientos de cajeros automáticos y una base de datos, en este escenario contarán con un tópico Kafka ya creado que se encuentra recibiendo mensajes de un productor, lo que deberán realizar es: la ingesta de los datos de los movimientos de los cajeros, el posterior enriquecimiento de los mensajes y luego visualizaciones que permitan entender los datos.

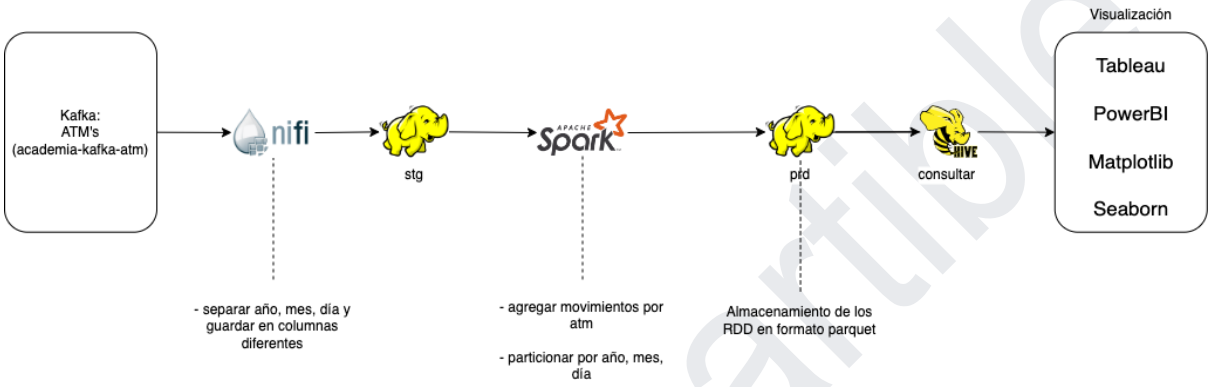
Se les proveerá de archivos en formato CSV que tendrán el siguiente esquema:

- **empleados.csv**: maestro de empleados
  - id_vendedor,sucursal,nombre
- **fact.csv**: tabla de hechos con compras
  - timestamp;sku;vendedor;cantidad
- **locales.csv**: maestro de sucursales
  - id_sucursal;nombre;tipo;
- **producto.csv**: maestro de productos
  - id_producto;familia;nombre;precio_unitario

La estructura de los mensajes que estarán disponibles en Kafka es la siguiente:

```
{
    “id_atm”: <integer>,
    “id_usuario”: <uuid>,
    “id_sucursal”: <locales.id_sucursal>
    “movimiento”: <integer>,
    “fecha”: <timestamp>
}
```

Ejemplo de implementación conceptual:



La implementación específica de la solución es el desafío que enfrentarán.

- Existe un productor Kafka que está agregando datos de manera constante a un tópico: *academia-kafka-atm*.
- La ingesta de estos datos puede hacerse desde NiFi o Spark.
- El enriquecimiento lo pueden realizar mediante NiFi o Spark, ideal que incorporen la herramienta que aún no hayan aplicado.
- Los datos enriquecidos deberán estar disponibles para la consulta mediante SQL a través de Hive.
- Y posterior visualización con la herramienta que mejor manejen, siendo deseable que puedan armar algo con Flask y/o PowerBI.

Cuestiones a tener en cuenta:

- Se cuenta en el entorno corporativo con las diferentes herramientas que permitirán el acceso y la manipulación de los datos:
    - Hadoop
    - Kafka
        - Tópico con mensajes para esta consigna: `academia-kafka-atm`

        - Brokers / Bootstrap servers: `dkafka01:9092,dkafka02:9092,dkafka03:9092`
    - NiFi
        - URL del entorno: https://cdh001-e01.bancogalicia.com.ar:9444/nifi/?processGroupId=018111ac-d70b-1594-a5fb-48ff778256b8&componentIds=
        - Acceso: `legajo (con "L" en mayúscula) y contraseña de Windows`
    - Spark
        - Ver configuración en notebook compartida en el módulo de la academia.