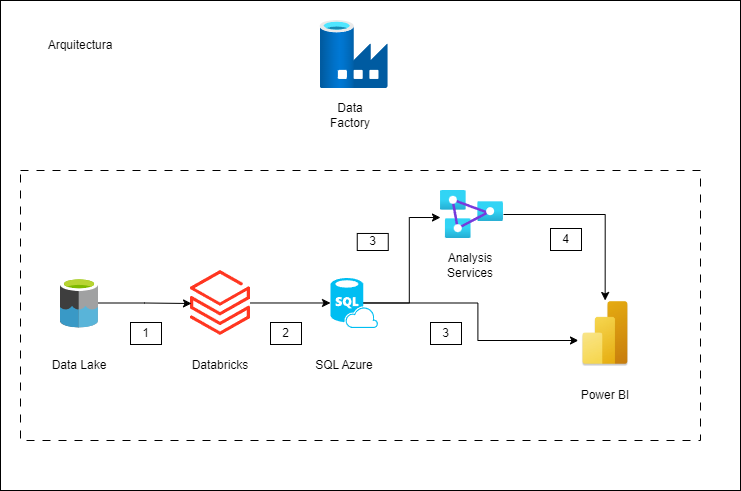
1. Alcance del proyecto y captura de datos
   1. Datos de quejas del consumidor (<https://catalog.data.gov/dataset/consumer-complaint-database>)
   2. Casos de uso
      1. Tablas de análisis para la identificación de compañías con más quejas, del público que se esta viendo afectado, del tiempo de respuesta de la compañía hacia el consumidor, el medio por el cual los consumidores levantan sus quejas, si el consumidor encuentra esta respuesta como satisfactoria o no de la queja.
2. Se hace un análisis exploratorio de los datos en donde se encuentran problemas en la lectura de los datos con errores de formato para dos columnas ( ‘Date received’, ‘Date sent’) y con 9 columnas las cuales cuentan con vacíos.
3. Definir el modelo de datos
   1. Se realiza un modelo estrella con el fin de optimizar la tabla debido a que la mayoría de sus columnas son categóricas lo cual hace que el peso de la tabla principal se reduzca.

Diagrama

Descripción generada automáticamente

* 1. Se Agrega imagen de la arquitectura sugerida para la construcción de las ETL’s donde se hará la limpieza y transformación de la tabla.
     1. Se va a almacenar la información pura en el data lake.
     2. Esta información se va a transformar en databricks con lenguaje Python
     3. Se hará limpieza de datos categóricos la cual consiste en eliminar espacios, estandarizar la escritura todo a minúscula y retirar caracteres especiales.
     4. Se almacenará la información en sql de azure donde están las diferentes tablas relacionadas con el modelo
     5. En el analysis services se hará la construcción del modelo para poder generar luego el reporte que ira conectada a este modelo y se puedan relacionar las tablas de manera correcta.



* 1. La actualización de la información se debe hacer diaria debido a que las quejas se reciben en cualquier momento del día con el fin de poder ser enviadas oportunamente a las empresas para sus respuestas.

1. Completar la redacción del proyecto
   1. ¿Cuál es el objetivo del proyecto?
      1. El objetivo del proyecto es realizar la limpieza y transformación de la data para poder responder las preguntas planteadas en el caso de uso.
   2. ¿Qué preguntas quieres hacer?
      1. ¿Cuáles son las compañías con más quejas?
      2. ¿Cuál es el medio mas utilizado por los usuarios para crear las quejas?
      3. ¿las compañías están respondiendo a estas quejas?
      4. ¿De las respuestas recibidas el consumidor está quedando satisfecho?
   3. ¿Por qué eligió el modelo que eligió?
      1. Se realiza un modelo estrella con el fin de optimizar la tabla debido a que la mayoría de sus columnas son categóricas lo cual hace que el peso de la tabla principal se reduzca y que el modelo sea flexible a la hora de crear los reportes.
   4. Incluya una descripción de como abordaría el problema de manera diferente en los siguientes escenarios:
      1. Si los datos se incrementaran en 100x
         1. Se debería hacer una limpieza de información solo a las variables necesarias para el análisis y depurar mas la tabla con el fin de no llevar información que no se vaya a utilizar. Podrían crearse data mart con el fin de crear tablas optimizadas solo para solucionar problemas puntuales reduciendo el tiempo de procesamiento.
      2. Si las tuberías se ejecutaran diariamente en una ventana de tiempo especifica.
         1. Se debe tener en cuenta tiempos de tolerancia debido a que el proceso dependiendo de el volumen de información se puede demorar mas o menos. Con el fin de evitar que no se pierda información y se corra completo se debe plantear el tiempo necesario y los que los servicios adquiridos sean suficientes para mover la información.
      3. Si la base de datos necesitara ser accedido por mas de 100 usuarios funcionales.
         1. Se les darían permisos específicos a cada usuario con el fin de que acceda solo a la información que le competa para evitar la saturación de la base de datos.
      4. Si se requiere hacer analítica en tiempo real, ¿Cuáles componentes cambiarias en su arquitectura propuesta?
         1. Aunque el procesamiento de información en tiempo real no es sostenible en el tiempo debido al gran volumen de datos procesados por segundo, consideraría que es mejor correr en ciertos horarios fijos el proceso para no ver castigada la base de datos y que los costos sea prudentes con lo que llegue a necesitar el negocio.

