**ARQUITECTURA DE DATOS STREAMING – CAZACORRUPTOS ANALYTICS**

En la actualidad y luego de la revisión de diferentes fuentes de datos se ha identificado el siguiente problema:

Los datos que se encuentran publicados en los portales públicos contra la corrupción se encuentran con diferentes formatos, desactualizados, en diferentes repositorios, no existen las URLS y existe información que se podría considerar como open data como las consultas públicas.

Causas:

No existe una política que exija que los datos se conviertan en una fuente de datos únicos.

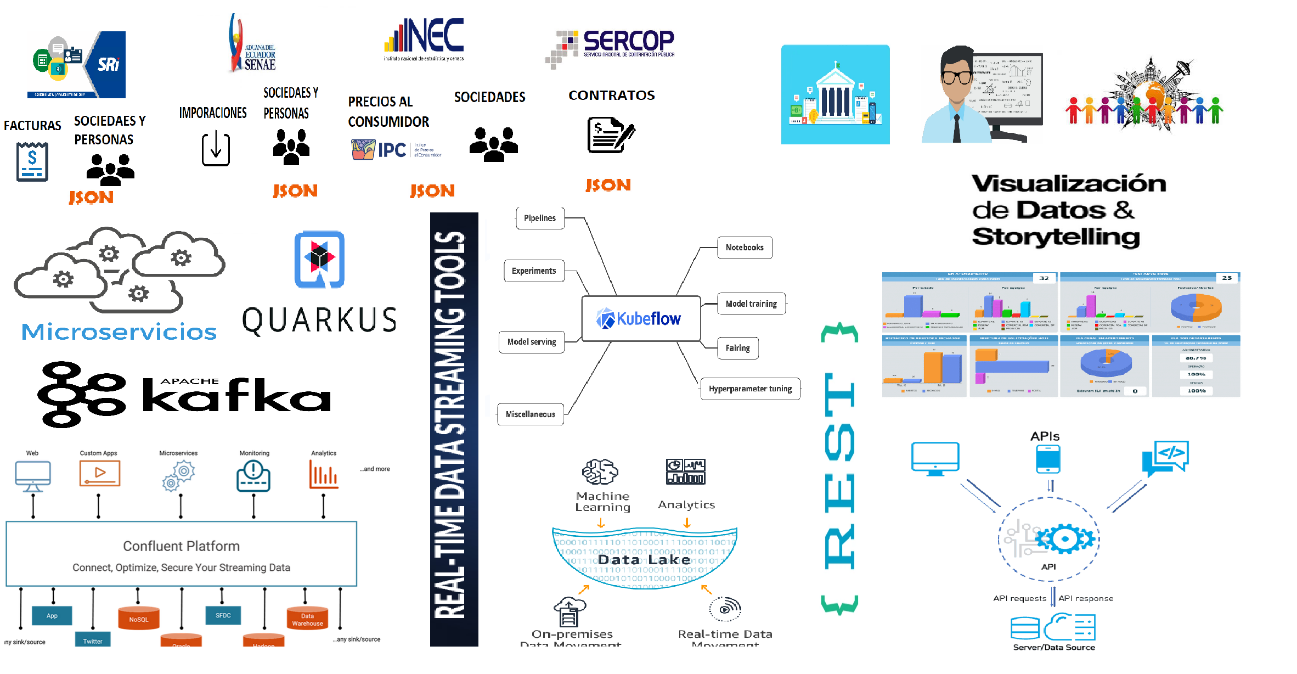
Los servicios en línea de las diferentes instituciones suelen cambiar sus URLS o la información que publican en una fecha para posteriores publicaciones la eliminan de sus reportes.

Los repositorios pueden solicitar el loggeo del usuario, incluso si la información tiene carácter de público.

No existe un verdadero análisis de que información se puede considera pública.

Debe existir liberación de fuentes de información que pueden dar valor a las consultas para combatir la corrupción.

Diagrama de componentes:



Descripción de componentes:

El uso de streaming de datos permitirá el uso de datos en línea como fuente de información actualizada para la creación en ingesta de información para la creación futura de procesos de extracción y procesamiento de información, esta solución permitía tener la información actualizada y disponible para generar consultas dinámicas de esta información.

Mircro servicios: Se utilizar Micro servicios basados en QUARKUS utilizando JSON para ahorrar costos de servicio.

Data Streaming; Se utilizará para la presentación de crear consulta tipo SQL en línea en función de los montos, fecha periodos de los contratos.

Kafka y Confluent: En sus versiones Open Source permitirán la ingesta de datos en colas con escalamiento múltiple y brokers de escucha. Permitirá tener la información almacenada y crear tablas dinámicas de consulta en línea.

Analítica de datos: Para la parte de analítica de datos se hará uso de data lakes y consultas directas sobre los datos en streaming.

Spark: Permite la generación de repositorios tipo Data lakes en función de los datos que se encuentren registrados en Kafka. Se utilizara también como ETL para procesos de merge de información. Permitirá convertir los archivos a CVS. Realizar limpieza y filtrado de datos.

Bases de datos: Existirán diversos repositorios de bases de datos que ayuden a dar soporte y conservar cierta información relevante para la construcción e modelos de aprendiza automático o basados en Deep learning.

Contenedores: Los contenedores serán aplicaciones basadas en Kubernetes para crear mecanismo de auto escalamiento en función de la necesidad de incorporación de nuevos registros.

KubeFlow: Herramienta utilizada para generar modelos predictivos tanto en Machine Learning como Deerp Learning para la creación mantenimiento y despliegue de modelos de machine learning para la extracción de características y la presentación de información en función de los datos.

Rest: Mecanismo de publicación de los pods o terminales de modelos Machine Learning, archivos CVS y Fuentes filtradas de datos.

JSON: Estándar de comunicación entre servicios.