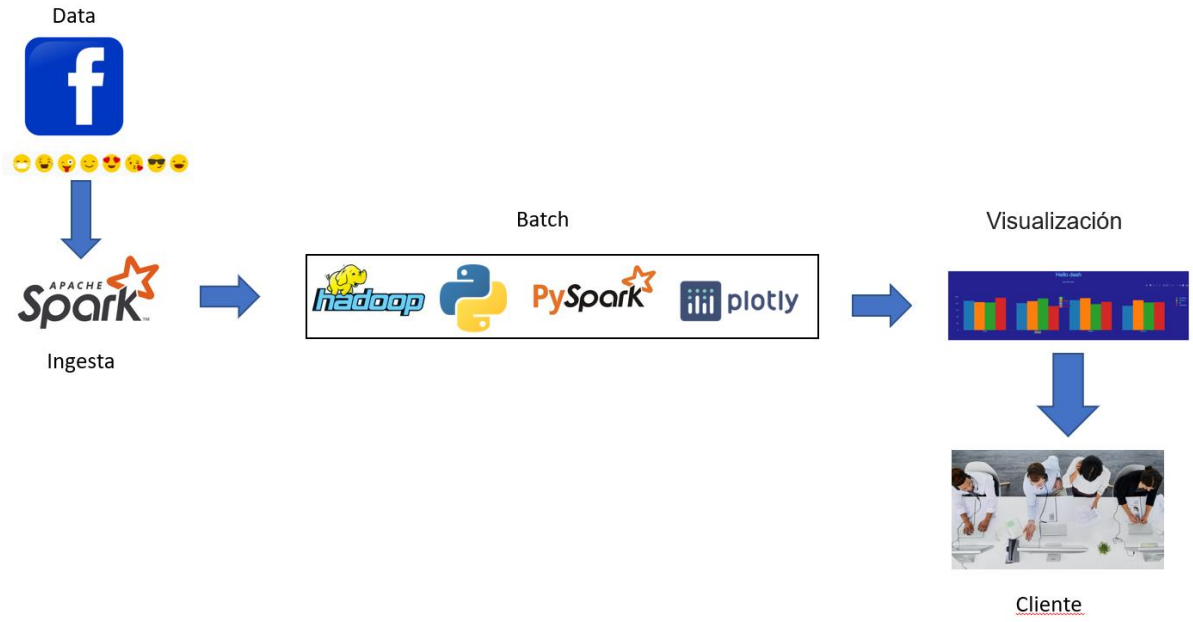
Arquitecturas distribuidas Big Data

El proyecto consistirá en el análisis del público y su interacción con las redes sociales de una empresa. Este proceso de analítica de datos se hace en las consultorías para sus clientes, los cuales tienen páginas en redes sociales y quieren saber sus datos, reacciones o vistas en las publicaciones con respecto al público que interactúa con él. Para ello se deberá de analizar la data de su público y su interacción con la página, entre esta data se encuentran su nombre, edad, sexo y sexo, así como de que país se visita la página.

Para lo anterior se necesitará computación en batch, y la arquitectura lambda, poder estar manipulando estos datos, darles un tratamiento y con ello poder presentar los resultados obtenidos al cliente de una manera clara, precisa y simple con software o herramientas destinadas a ello (power bi, google data studio, tableu, etc) podrá hacer que resuelva sus inquietudes e incluso definir si el público que está teniendo es relevante para él.



En cuestión de herramientas se usará Spark debido a su conexión con Python y en especial por Pyspark para hacer el tratamiento de estos datos, es decir,

la limpieza y modificación de ellos, con el fin de poderlos presentar al cliente ya procesados y tener a su vez la data en limpio, el cual sería un proceso batch.

Se selecciono la arquitectura Lambda por el hecho de poder tener los datos en bruto y procesados, los datos en bruto le serán de utilidad al cliente para verificar estos o examinarlos en procesos internos, mientras que se podrán manejar y modificar estos para graficarlos y tener algo visual que pueda entender de una mejor manera.

- La data se extraerá de facebook con las herramientas que ofrece.
- La ingesta de datos se hace con Sprk debido a que en los siguientes procesos se tiene contemplado Python y su conexión es factible.
- En el proceso batch se usará Python con su freamwork PySpark para poder ver estos datos, modificarlos y tratarlos. Una vez lo anterior se podrán graficar con Plotly y dash.
- La visualización o esta publicación de los datos se llevará a cabo con las librerías mencionadas el final del punto anterior, dándole al cliente los resultados obtenidos en una pagina web libre de etiquetas, es decir, a diferencia de una página tradicional esta no podrá ser buscada en los navegadores, solo las personas que conozcan la liga podrán acceder a ellos.
- Cualquier cambio que solicite el cliente u otro proceso que pueda requerir se podrá hacer gracias a la data en bruto.
- Aunque no se tomo en cuenta el speed layer, esta podría tener relevancia para el cliente, dependiendo de cuál sea su nicho, hay empresas que hacen eventos en vivo, por lo cual si la empresa e suma de estas se deberá de considerar para extraer los datos al momento del en vivo y poderlos analizar con posterioridad para sumarle valor a sus transmisiones y saber el tipo de público que se obtuvo en esta transmisión.
- Las colas deberían de implementarse sobre todo en las respuestas de comentarios, hay hilos en redes sociales los cuales pueden ser no tan informáticos a la hora de analizar, por lo que estos hilos o respuestas

deberían de estar hasta el final de nuestra cola de prioridades, siendo el objetivo el primero comentario.

Un dashboard similar que se podría ver, claro, cambiado los datos y mencionado todas las herramientas a usar, sería algo como lo siguiente

Delimito en la imagen los nombres que puedan asociarse donde trabajo por temas de privacidad.

Tao Izzo Elvira