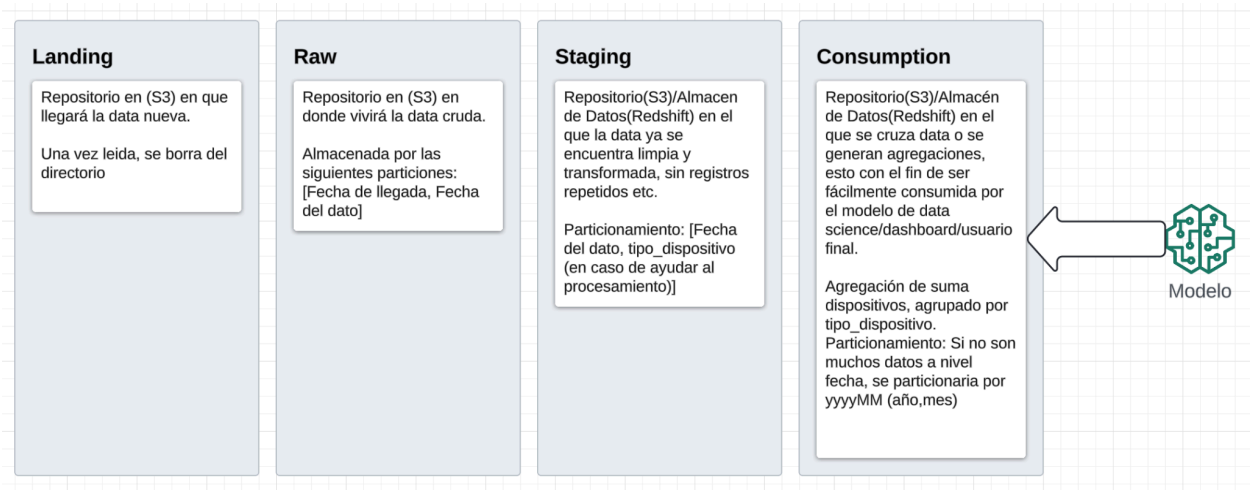
Ejercicio3 - Challenge Kavak Data Engineer

Compañia Té de Pastor

1. Propuesta Técnica

Arquitectura de Datos

Para el procesamiento y limpieza de datos se toma como base el modelo Lake House storage layer, que permitirá tener dividida la data por calidad, puntos de acceso y enriquecimiento de la misma.



Proceso Batch (Caso de data repetida):

Procesamiento de data nueva: Esta se escribirá en su propia partición del día.

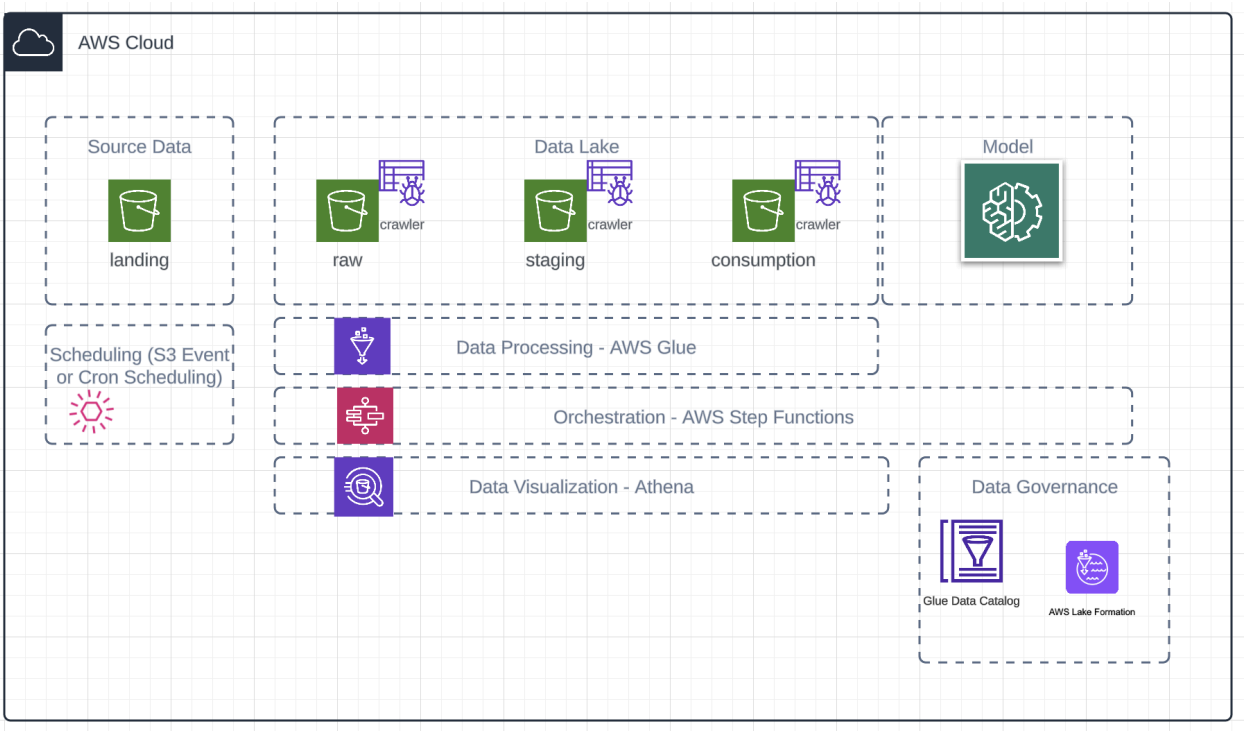
Procesamiento de data anteriores:

- Para el caso de glue (spark) De staging se toma la data de hoy-3, mediante la llave designada se compara contra la historica y esta se sobre-escribe con los registros nuevos.
- Para el caso de Redshift: Maneja operaciones CRUD.

Arquitectura de Infraestructura

Para la arquitectura de infraestructura se toma como proveedor de nube: AWS y se plantean dos propuestas, una totalmente serverless y otra con servicios provisionados.

Propuesta 1 (serverless)



AWS S3: como repositorio de datos

AWS Crawler y AWS Data Catalog: para escaneo de buckets y repositorio de metadata de los archivos

AWS Glue: para transformacion y limpieza de datos

AWS Sage Maker: para correr modelo

AWS Step Functions: Para orquestación del flujo

AWS Athena: Para visualización de datos en caso de ser necesario

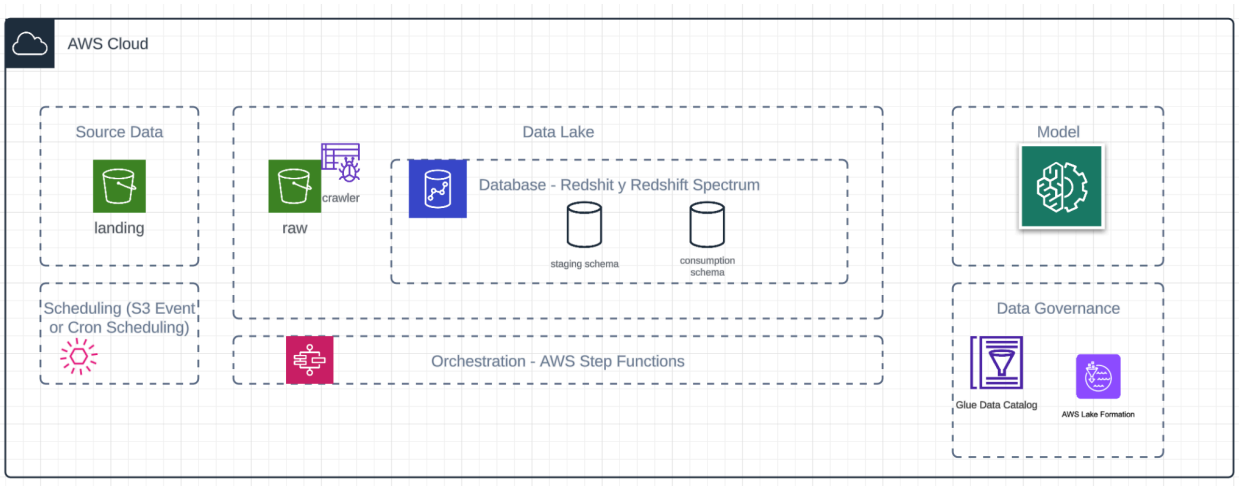
AWS Lake Formation: Para gobernanza y permisos sobre la data

AWS SNS y CloudWatch: Para alertas enviadas desde aws glue en la validacion, limpieza y transformacion

Scheduling: Puede ser mediante un CRON en el step function diariamente a las xx:xx horas o también puede mandarse a ejecutar por un S3 Event y AWS Lambda, cuando detecte el S3: Object Creation.

AWS Code Commit, Code Build y Code Pipeline: Para versionamiento de codigo y despliegue continui

Propuesta 2 (servicios provisionados)



AWS S3: como repositorio de datos de landing y raw

AWS Redshift Spectrum: Para visualizacion de datos en raw

AWS Redshift: Como repositorio de datos de staging y consumption, así como se pueden ejecutar jobs de transformacion y limpieza de datos con el mismo motor de redshift y visualizador de datos.

AWS Step Functions: Para orquestación del flujo

AWS Crawler y AWS Data Catalog: para escaneo de buckets y repositorio de metadata de los archivos

AWS Sage Maker: para correr modelo

AWS Lake Formation: Para gobernanza y permisos sobre la data

AWS SNS y CloudWatch: Para alertas enviadas desde aws glue en la validacion, limpieza y transformacion

AWS Code Commit, Code Build y Code Pipeline: Para versionamiento de codigo y despliegue continui

Scheduling: Puede ser mediante un CRON en el step function diariamente a las xx:xx horas o también puede mandarse a ejecutar por un S3 Event en landing y AWS Lambda, cuando detecte el S3: Object Creation.

2. Planeación a alto nivel

(Dinos qué podríamos lograr con certeza en 2 meses si tuvieras en tu equipo un ingeniero de datos, un ingeniero de aprendizaje de máquina y un site reliability engineer.)

Alcance

Generar el proceso para la ejecución de el modelo predictivo X, el cuál se alimentará de archivos que contienen datos de geolocialización de dispositivos móviles, estos serán depositados diariamente, los cuales serán limpiados y procesados a traves de 3 capas para ser la fuente de datos y entrada de la ejecución del modelo X con los parámetros y salidas W, X, Y, Z.
Los puntos que se estaría entregado en esta fase son:

- El resultado del modelo, transformaciones y limpieza de datos de acuerdo a lo que se definirá en las primeras 3 semanas.
- Ejecución del pipeline productivo diario de forma correcta, completa y en tiempos/performance acordado
- Entrega de infraestructura como código
- Código versionado con integración continua
- Código del procesamiento de datos escalable y modular

Rol	Tarea	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8
DS	Definición de Modelo, Variables y Fuentes								
DE	Definicion de Arquitectura de datos								
DE y SRE	Definición de Arquitectura Técnica								
SRE	Creación de pipeline, ambientes, servicios y permisos								
DE	Creación de flujo de datos (3 capas, orquestador)								
DS	Creación del modelo y entrenamiento								
DS	Análisis de resultado del modelo y ajustes								
SRE, DE	Pruebas de Integracion								
SRE, DE, DS	Despliegue en producción								

SRE - Site Reliability Engineer

DE - Data Engineer

DS - Data Scientist

Consideraciones y suposiciones de la estimación

- Las estimaciones son en un alto nivel, en caso de que la estimación no corresponda en tiempo se revisa con el equipo para acortar el alcance de la tarea; afectado lo menos posible al alcance.
- Se asume que ya se tienen los permisos/credenciales correspondientes a datos y ambientes
- El rol únicamente describe al(los) responsables de la tarea, se asume que existe una colaboración entre roles y stakeholders para lograr finalizar la tarea si así se requiere.
- Se asume que se tiene por lo menos un ambiente de desarrollo y otro de producción
- Los ingenieros involucrados ya tienen experiencia trabajando con las tecnologías a utilizar
- Los ingenieros involucrados ya tienen conocimiento sobre los datos y familiarización con la problematica a resolver
- Los tiempos de procesamiento del flujo completo de datos para la obtención del resultado final se estarán probando e informando a lo largo del proyecto, esto con el fin de aplicar optimizaciones y estimaciones del tamaño de los recursos en los ambientes.

3. Niveles de servicio

(Cuáles serían los niveles de servicio a los que podrías comprometer tu sistema?)

Antes de comprometer algún nivel de servicio, preguntaría en el requerimiento dudas como el impacto en caso de algún fallo, en que horario la información se necesita que esté disponible? para poder priorizar, ; ya que, dependiendo del impacto se ofrecerá una solución que la mitigue; [replicación en AZ, alertas con severidades en caso de fallo, tipo/tamaño de servers, estrategias de ejecución del flujo etc]