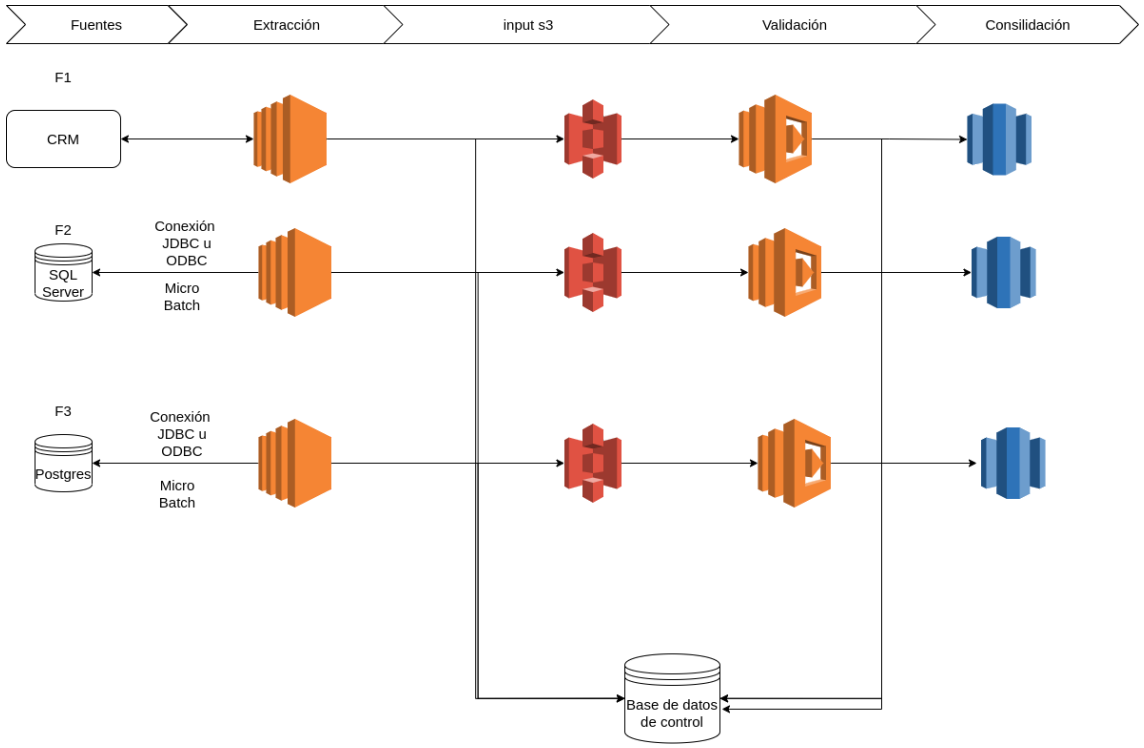
Ejercicio2
propuesta de arquitectura
Para la extracción de las fuentes propuestas se propone la siguiente arquitectura.



se explica cada una de las fases:

Fuentes: para cada una de esta se propone extracciones por micro batch esto es extracciones de pequeños intervalos de información ya sea extracciones de una hora de información o menos

Extracción: para la extracción se propone ya se una o varias instancias de ec2 desde donde se ejecutará por medio de crontab la extracción de data en esta parte también se considerara ocultar data sensible o eliminarla:

- Bases de datos se propone la conexión por JDBC u ODBC para hacer la extracción por medio de consultas SQL esta para tener control de qué tablas consultar y el tamaño del batch a extraer.
- CRM para el CRM en muchos casos el propio CRM tiene una base de datos operativa o este mismo te ofrece algun tipo de exposición a sus datos por ejemplo por API REST

Input S3 : para cada fuente se considera un bucket diferente y en esta parte los datos van estar en texto plano csv se manejan cuatro carpetas por cada bucket:

- **tmp** data cruda y sin ningún tipo de procesamiento.
- **input** si la data es consistente es decir no está corrupta el tamaño de archivos es correcto la data pasará a esta área
- **processing** la data o archivos se copiaran de input a processing en esta área se realizará todo el procesamiento de datos para el procesamiento se proponen dos

herramientas spark o montar tablas externas de athena o redshift para hacer el procesamiento con sentencias SQL

- **Final** en esta área se cargará la data ya procesada y solo cuando se llegue a esta parte se borrara la data de processing e input en esta fase se propone archivos en formato parquet

Validación en la etapa de input en s3 hasta la carpeta final se ejecutarán funciones lambda que se encargará de validar a nivel de metadata que la data llegue de forma correcta a cada área e insertar en la base datos de control el estatus correcto o incorrecto de cada archivo de data

Consolidación ya con la información en área final se montan tablas sobre la data para hacer esta accesible al usuario final para este caso se propone usar redshift.