	Data bases	Files/csv	streaming	API
data source				

	incremental, full, bulk lambda/glue/emr	incremental, whole file transfer and loading file/stream	real-time streaming/processing AWS kinesis	get data from API
Ingestion				

Lakehouse Storage

Data Lake

Processed Zone

catalog

Shared Catalog (Cloudera/Glue catalog/Informatica)

Processing

SQL bases ELT   Big data processing   Near Real-time ETL

(AWS Glue, EMR, Lambda, Spark, kafka, Hive, Bige, HBase)

Data Warehouse (Redshift)

Consumption

Machine Learning   Business intelligence   Reporting   Interactive Queries

the diagram above represents a quick view of a consolidated Lakehouse solution that ingests cleanses, processes, and loads data at different freshness levels for structured, semi-structured, and unstructured data:

this includes all sources of data even audio/video kinds of data that could be processed in ML for example.

structured and unstructured data could be ingested into the AWS S3 storage at any point in patch or streaming.

at the lake house storage, we will have native ingestion between the data lake and warehouse, processed/cured zone, and even snaps.

all the metadata that consists of the data schema and data locations are stored in the data catalog.

this architecture is unified for all purposes of analysis with better performance and cost-effectiveness. it can be built in an Agile approach.

Thanks
Abdalrhman
abdalrhman.alryah@linux.com. | 00971562523380