Adv of using cloud:-

→ any amount of data

→ no CPU, memory limitation

→ CPU → GPU , hardware expansion.

Ingest → Analyze → Explore

Massive datasets [social media channels, internal company sources
web applications]

Messy, poorly documented.

On server/laptop

limited by resources → hardware
CPU
memory
buy additional resources.

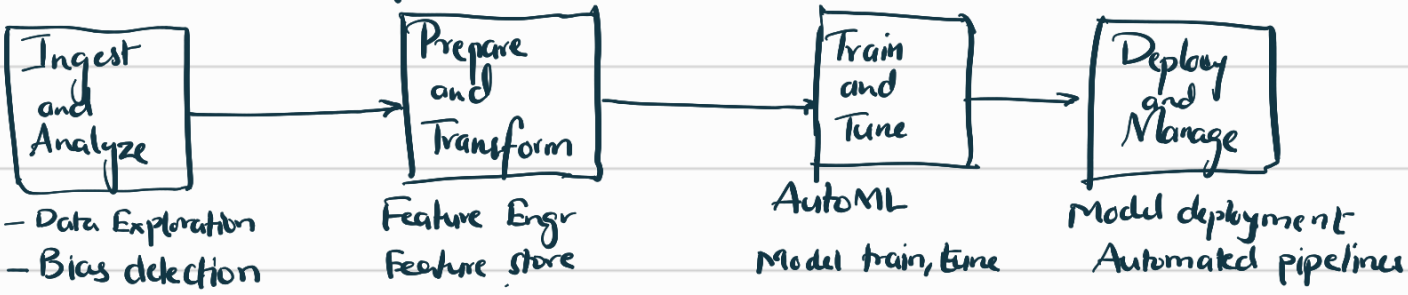
Use Cloud :-

Scale Up:- Switch to GPU or use instance of multiple CPUs.

Scale Out:- Distributed training on multiple CPU instances.

} Store and process any amounts of data.

Machine Learning Workflow :-

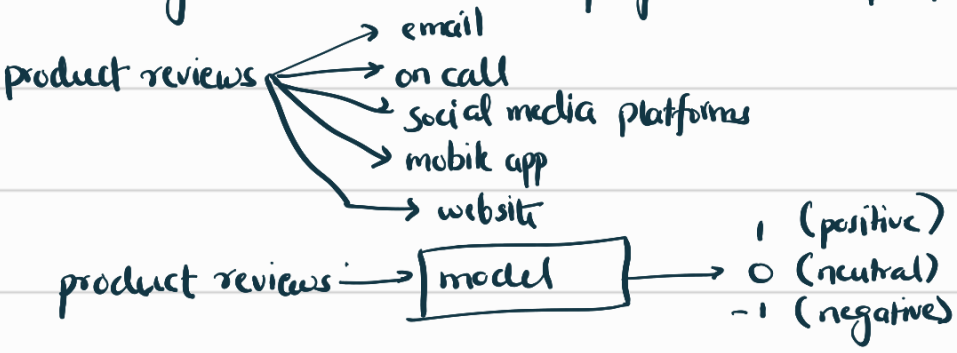


Amazon S3, (storage)	ASM Data wrangler	ASM autopilot	ASM endpoints
Amazon Athena (SQL)	ASM processing jobs	ASM training & debugger	ASM Batch transform
Amazon Glue (catalog)	ASM feature store	ASM hyperparameter tuning	ASM pipelines
ASM Data wrangler, clarify			

Use Case [Text Analysis]

Multi-Class Classification.

Assume working on a ecommerce company with multiple products



Ingestion and Exploration:-

Ingest data into data lakes (centralized and secure repository that can store can store, discover and share data at any scale and any type)

- structured (csv, excel)
- semi structured (xml, json)
- unstructured (img, audio)
- streaming data (app which gives continuous log files) / social media cont. feed

Datalake to be governed.

- ways to discover and catalog new data
- secure and control access

Data lakes build on object storage (S3) Simple Storage Service.

File storage [stores and manages data as individual files in a hierarchical order]

Block storage [stores and manages data as chunks called blocks]
block → unique identifier but no metadata

object storage [data is stored and managed as objects which consists of data, metadata (last modified etc.), identifier]

Object storage is helpful for storing growing data of any type.

Durable, Available, exabyte scale

Secure, Compliant, Auditable

AWS Data Wrangler :-

- opensource python library.
- connects pandas DF with AWS services.
- load / unload data from
  - datalakes
  - databases
  - data warehouses.

AWS Glue :-

Data Catalog [what data is stored in the S3 bucket / datalake]

Name reviews

Database dsoaws-deep-learning

Location s3://<bucket>/<>

Classification / type csv

→ Creating a reference to data (S3 to table mapping)

→ metadata / schema stored in tables

→ No data is moved. Only metadata is present in Glue database.

→ Automatically setup using glue crawlers to
- infer the schema
- update the data catalog.

Amazon Athena :-

- Query the data in S3 using Athena
- using SQL
- serverless (no infrastructure to set up)
- schema lookup in AWS Glue Data catalog

Pros :-

Athena can deal with
- highly complex analytical queries
- GB → TB → PB
- need not worry about compute, memory resources as athena will scale out and split query into simpler queries and runs in parallel

- based on presto (open source distributed SQL engine to run queries on any data sizes)
- no infrastructure setup / data movement is required.

Lab:-

boto3 :- AWS SDK for python used to create, config, manage AWS services

sagemaker :- python SDK with high level abstractions for working with sagemaker

awswrangler :- python opensource that extends the power of pandas to AWS connecting dataframes to amazon related services.

athena :- can be used to interact the data in S3 bucket with standard SQL.

write SQL → Amazon Glue → S3

dataframe is returned.