AWS :- AWS - ML Date : 02/01/2022.

→ AWS ML OVERVIEW :—

→ How to Select ML technique to Solve given business Problems

→ Which AWS Services

→ How to design, implement, Scale, Secure.

→ Most Cost effective solutions.

→ 65 questions Scenario → 3 hours

75-1. pass exam → $ 300 US dollars.

1. Data Engineering — 20-1.
2. EDA — 24-1.
3. Modeling — 36-1.
4. ML implementation & operation — 20-1.

100-1.

DE:

1. Create Data Repository
2. Data ingestion
3. Data transformation

EDA:-

1. Prepare data for Modeling
2. Perform feature engineering
3. Analyse & Visualize data from ML

Modeling:

1. Frame Business Problems
2. Select appropriate ML model
2. Train
4. Perform hyper parameter
5. Evaluate ML

ML implementation:

→ Build ML, performance, availability, scalability, resiliency, and fault tolerance

→ Recomanded and Implement the apporte ML Servius and features for a given problem

→ Apply Basic AWS Security practices to ML Solutions

→ Deploy ML Solutions.

Important Concepts:-

1. Data storage
2. EDA — AWS Glue, Data pipeline, AWS Data migrations, AWS Batch/Step
3. Model
4. MLops.

Kinesis overviews. } Amazon Athena features.
Kinesis

Quicksight features

QuickRight ML

EMR
↳ Hadoop
Apache spark
features

ML overview

→ Features extracted Manually in ML and in case of DL it's done automatically

AI

ML, robotics and Computer Vision → ML
- Supervised → Classification, Regression
- Un supervised → clustering
- Reinforcement

** AWS **

→ Cloud platforms

→ 165 Fully featured services (40 of them are not offered anywhere)

Compute, Storage, Database } data + model → Compute

Data + Model + Compute

Compute → EC2, ECR, ECS.

Data Lingo:-

Data
- Unlabeled Data
- Labeled Data

imagenet → 21,841 classes.

→ Aws Ground truth :-

Data types

Quantitive	Qualitative	Ordinal
Numerical	Categorical	mix between numerical and Categorical Data.
↳ discrete or Continuous	↳ Ex: race, sex, educational	
	You can't multiply them together for example.	

→ Database vs datalake vs Warehouse

Database		Warehouse
Structured ~~Data~~ with defined Schema.		Data bases are put together to form a Datawarehouse

Data lake:-

→ Data lake is a centralized repository for structured and unstructured data storage.

→ Data lakes could be used to store raw data as is without any structure (schema)

→ there is no need to perform any ETL

→ processing could be done on export to. schema is defined.

AWS key storage 99.99999%

storage services allows extremly durable entreprises / individual

1. Amazon S3 → to store and protect any amount of data
2. Aurora RDS → Relational Database, fully managed. mysql, postgre SQL.
3. Red Shift → data warehousing service.
4. DynamoDB. →

→ Redshift uses a unique data warehouse architect that relies on Massively parellel processing (MPP)

→ Redshift uses Machine learning to optimise performance

DOMAIN #1

1. DATA ENGINEERING :- Code :- V135%.

Amazon S3: Simple storage Service.

1. Scalability
2. Data availability
3. Security
4. performance.

Data is Storage in Buckets, Global unique name.
You can store intimate amount of data in bucket in which each object can contains upto 5TB.

Object tages : Confidential / PHI / project.

Here an example of an object tag

1. project → ML

Classification → Confidential

PHI → True.

Aws S3 Data lake :-

× gigabytes → petabytes

Aws S3 common formats: CSV, parquet, ORC, Avro, Probobut. and Json.

→ Amazon S3 hosts that data the Machine learning such as sagemaker will use for model training and testing.

training Data → sagemaker.
trained Mode

Aws S3 Amazon sagem

→ Compute and Data processing decoupled

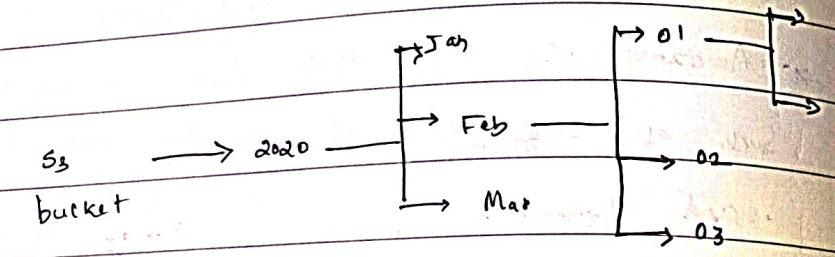
→ Centralized Data Architecture.

→ Integration with Clueless / Serverless Aws Services.

→ Standradized APIs. → Hadoop, Sparks.

AWS S3 data Partitioning:-

→ Data Partitioning is critical when querying data because it could dramatically reduce the cost required for scanning



Amazon Storage tiers.

1. S3 standrad → Works with general purpose and frequstly used / Accesd
2. S3 intelligent : Works well with data has Varying Acess.
3. S3 Standrad - infrequent Acess.
4. S3 One zone - infrequent Acess
5. S3 Glacier.

General → infrequent Access → Glacier.

Life cycles:-

transition actions: Policy that governs when a object transition from one storage class to another.

→ Glacier:- Data need to be retained for Compliance purpose:
Health care records.

→ XML files used for life cycle configurations

Amazon S3 encryption.

→ Data security we need to encryption.

Amazon S3 - managed keys (SSE-S3)

Customer master keys - stored in Aws key management services.

This allows for extra security.

SSE-C :- Manage their own keys to perform data encryption.

→ For ML → below encryption used.
1. SSE-S3 and Aws KMS.

Amazon S3 Security :-

→ Security of the cloud → Aws

→ Security in the Cloud → User

Amazon S3 ensure data protection and durability by

1. Adding redundency storage over multiple devices across many facilities / regions
2. S3 ensure that any corrupted files are detected and repaired.
3. Using Versiong : Which allows users to retrive several versions of the same object.

server side encryption

intransit data encryption

Amazon Macie | V. Important

Perform auditing to amzon S3 bucket

Macie is security service to automatically find sensitive data and alert users for anomolies.

→ "block public Acess" → by default :, you

Logging & Monitoring

1. Cloud watch alarms : Certain thresold is exceeded for multiple number of cycles

2. Cloud trail logs : tracks activity made by users, roles, Ip address, record of past requests, timing of the request.

3. S3 Acess logs :-

Networking - VPC End point gateway :

Amazon Virtual private cloud allows user to create Aws resource inside a Virtual network.

This means that traffic will never leave or go through the public internet and will stay inside the VPC for maximum security

→ Vpc end point will route request to Amazon S3 and back to the VPC.

Tagging :-

You can use tagging bucket policies to ensure security.

" Sensitive = TRUE."

	Data Warehouse	Database
Used for ?	Used for data analytics	Transaction proces
Sources of data ?	· Data gathered in multiple sources.	Data collected as - is from one sour
Data Writing frequency ?	Bulk Write Per fixed Schedule	Many write operati as data become
Storage optimized for ?	optimized for high ~~frequency~~ Speed query in column format	write optimized in single-row

→ Elastic search :-

↳ ingestion, enrichment, storage, analysis and Visualization

→ Tools are known as ELK stack

→ fully Managed Service that allows that allows for Elastic search deploy ment easily and Securly

→ Click Stream analytics, data indexing

* AWS - GLUE * V. Important Code:- !2221.

AWS- GLUE → ETL

Data generated by AWS Glue could be fed into analytics tools such as Amazon Quick Sight.

1. Setup Aws Glue and direct it to your Stored data on AWS
2. Aws Glue extracts metadata
3. once catloged, you can query and perform data ETL.

→ aws Glue → Integrated S3, quicksight, Redshift.

→ aws Glue → Crawls data storage, extract data Schemas and transformations

→ Aws glue generates data transformations Codes - on its own.

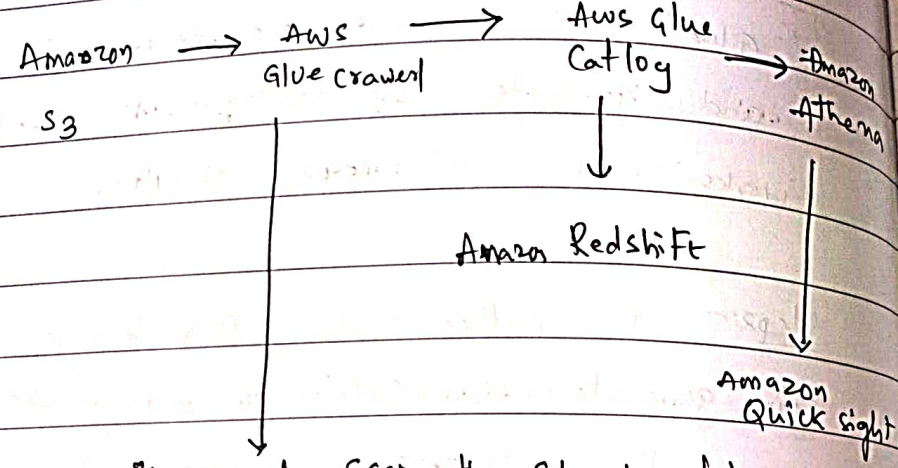
***

Step1:- After identifying the source data. Aws Glue will Crawl the data Source (S3) and generate the Catalogue. Aws Glue works with all kinds of data.

Step2:- In python/scala prog language, aws glue generate ETL Code to extract the data from the source and transform it to Match target schema and load it into the target. Users can then test the code in IDE/note book

Step3:- Aws Glue facilitates running Scheduling Multiple ETL Jobs. Aws Gl efficiently and automatically scale based on demand.

Aws Glue :-



it is used scan the stored data and infer the schema and partitions.

↓ once it scan it generates the table in data catlog.

→ you can run the Crawles on demand

→ Machine Learng transformation:

* Find Matches : Could be used to find matching data which enables you to find related Products, customers

V. Important :-

Aws DATA pipeline :-

* Aws Data pipeline is an orchastration Service that allows Aws user to transfer data reliably and Securly between various Aws compute and Storage Services.

* you can transfer the data to S3, RDS, Dynamo DB, Amazon EMR.

* Aws guarantee resource availability and handles failures.

Example :-

→ Data in dyanamo DB is copied to Amazon S3(Json) Continuously as part of Data pipeline

→ Json files are then converted into CSV to be used by Amazon sagemaker.

→ Amazon sagemaker consumes data, trains the model, Update end point for inference

4. Data in csv format is being queried by amazon Athena

5. Data pipeline manages the workflow as per Customer requirements.

AWS GLUE :-

→ Glue is an ETL Service that runs on a Serveless apache Spark Environment.

→ Glue Contains a data Catalogue for ETL that could with Athena and Redshift Spectrum.

→ Aws Glue ETL Job uses Scala or Python.

AWS Data Pipeline :-

→ AWS pipeline a Managed orchestration Service

→ Aws data pipeline offers more

AWS DATA MIGRATIONS (DMS)

oracle to Aurora

allows for Database Migration

→

RDS → AWS DMS → Amazon S3

ONPREMISE → AWS DMS → Amazon S3

AWS DMS → Safely and Cheaply

AWS BATCH :-

→ Batch Computation Job on AWS.

→ Aws Batch performs all the Scheduling and execution of the batch using Amazon EC2 and Spot instance

→ Cloud Watch, Events.

ETL → Aws Glue

Batch → AWS

4. Data in csv format is being queried by amazon Athena

5. Data pipeline manages the workflow as per Customer requirements.

AWS GLUE :-

→ Glue is an ETL Service that runs on a Serveless apache Spark environment.

→ Glue Contains a data Catalogue for ETL that could with Athena and Redshift Speeturm.

→ Aws Glue ETL Job uses Scala or Python.

AWS Data Pipeline :-

→ AWS pipeline is a Managed orchestration Service

→ Aws data pipeline offers more flexibility over EC2 Instance.

AWS DATA MIGARATIONS (DMS)

oracle to Aurora

allows for Database Migration

→

RDS → AWS Dms → Amazon S3

ONPRMISE → AWS Dms → Amazon S3

AWS DMS → Safely and Cheeply

Aws BATCH :-

→ Batch Computation Job on AWS.

→ Aws Batch performs all the Sechedulag and execution of the batch using Amazon EC2 and Spot Instance

→ Cloud Watch, Events.

ETL → Aws Glue

Batch → AWS Batch →

AWS STEP Functions

→ AWS Step function for Creating Serverless workflows

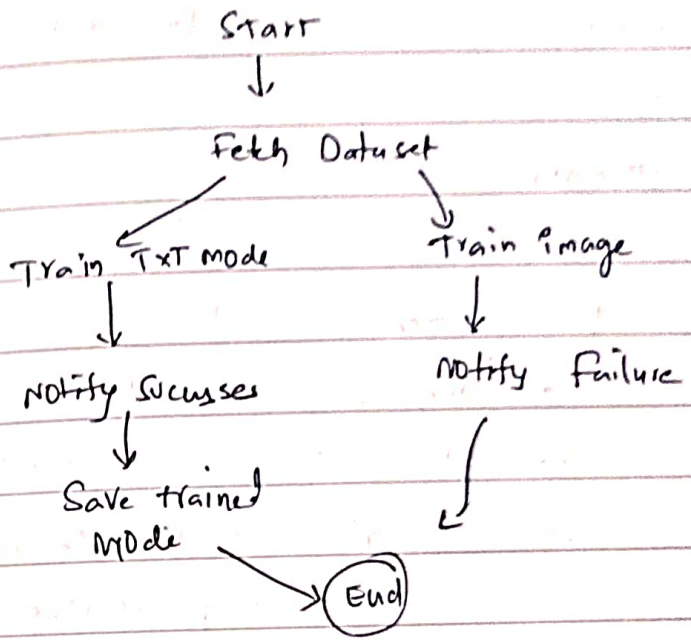
→ output from one step is fed as in input to the next step

→ Aws step function converts workflows into state of machine diagrams that easy to debug and understand.

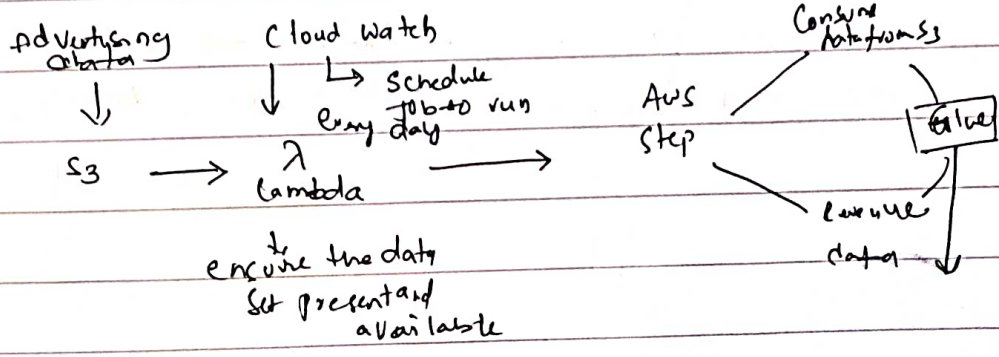
→ Aws step function allows for performing resilient workflow automation fast with out writing code.

→ it allows for advanced error handling and retrying mechanism.

AWS STEP Functions:-



→ Aws Step functions:-



V. Important-

DATA STREAMING

→ Real time

→ Scalable

→ Fully Managed.

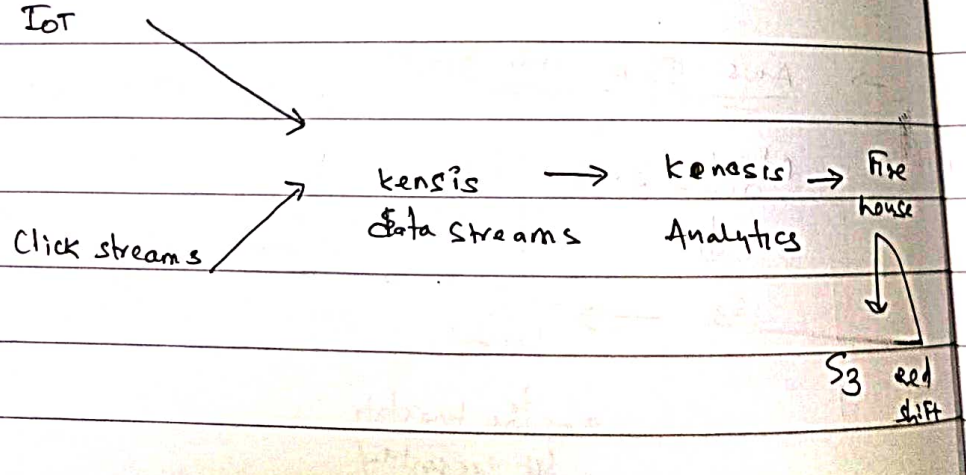
" Kinesis → Apache kafka."

→ Kensis Video streams

Data Streams → Real time

Data Firehose → Near real time

Data Analytics → Data Analytics and Processing Service.

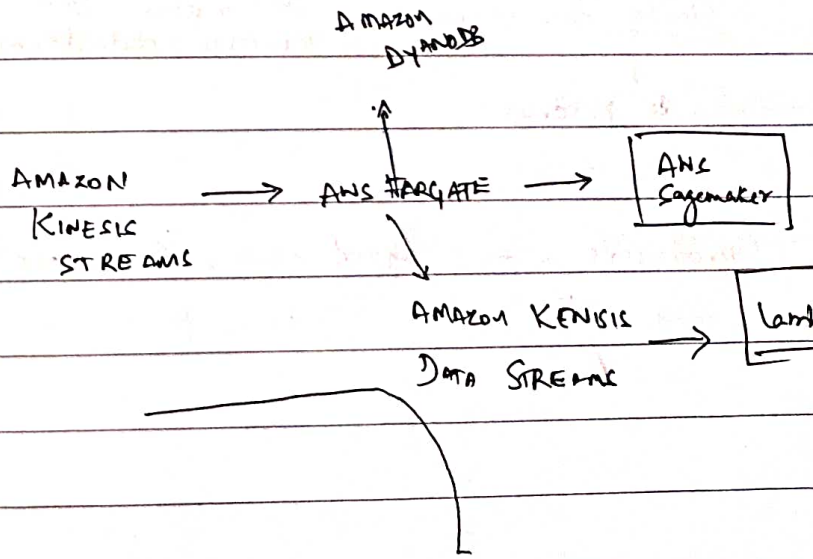


Video stream Sdk → Kenesis Video stream → Amazon Rekognition, mxnet, tensorflow, HLS-Based, Video playback

→ Amazon Kinesis to develop Smart cities by reducing crime rates and Solving

→ Smart preventive maintaince.

→



Kinesis Data streams:

Data producers

Data Consumers

Data Streams.

Shard → base throughput unit

KPL

The KCL allows for data writing to a kinesis data streams as producer.

KCL

KCL allows data Consumption from kinesis datastream

Kinesis SDK API

producers → Shard → Consumers

KINESIS Firehouse

↳ near-real time analytics

Ingested, transformed, encrypted and loaded out to Amazon S3, Redshift, Splunk.

60 seconds

Lambda

Kinesis Data Streams	Kinesis firehose	Amazon S3 Redshift Elasticsearch Splunk.
Device		

	Kinesis Data stream	Firehouse
1.	Real-time	Near-Real time
2.	Real time	60 seconds
3.	from 1 to 7 day	no data storage
4.	Required Shards	Auto scaling

AWS Kinesis Data Analytics

Real time, automatically scales based on incoming data throughput.

zero setup upfront cost, pay per use model

Step1:- Setup Source

Step2:- Develop Queries

Step3:- Setup the destination for processed Data.

Case #1:

Streaming ETL for IOT:

→ Amazon Kinesis Data Analytics could be used to transform and filter Streamy data from IOT, device such as Sensors then send real time ~~data~~ alerts when Variable exceed Certain limit.

KINESIS Summary:

Kinesis Data → For real-time streaming of data and gain real time insights

Kinesis firehouse → for streamy data in near real time and storing them into S3.

Kinesis Video → for streaming live video in real time

Kinesis Analytics → To run SQL queries on Streamy data and generate graphs and gain metrics

V.I

Domain 2 :-

EDA :- 24 %.

Code : A1x 2 =#

Code :- A#r4E

→ Amazon Athena :-

↳ Cost - effective query Serice.

S3 → SQL queries

Athena Very fast, Serverless.

it is Very cost effective, you only pay per queries you choose to run.

→ Define schema

→ use SQL, start running queries to accs data.

Amazon glueng :-

1. Durability & high availability
2. Very Secure
3. Seamless integration with other aws services (Glue).

→ Athena Security :-

→ Aws identy and Acss. Management policies.

→ Transport layer security encrypts in transit data between Athena and Amazon S3.

→ Athena Support- S3 and Apache orc and parquet.

Athena	vs	RedShift
Send queries directly to S3		Send queries directly to S3
Designed for easy adhoc queries into S3		Designed for users of Redshift
Doesn't Require redshift clusters		you will require redshift Clusters.

→ Aws QuickSight :- BI tools, visualize

1. Cost effective
2. High Scalable, Scale from 10 to 10,000 users
3. Allow for easy Data Analytics Embedding
4. End to End business intelligent Solutions.

S3

Glue Catlog → Quicks Sraght

Parquet

Amazon QuickSight: SPICE

→ fast, highly - available

→ import the data into SPICE for a great improvements.

→ Use Case:-

→ packaged data products

→ improve apps by offering Analytics

→ Integrate data into workflows.

↳ his dashboards/reports into portals and Company Sites.

Embedding Analytics:-

→ Provides a Complete Analytics Experiences

→ Ricin Apps

V. Imp

Amazon quick Sight : ML insights:-

→ ML insight features

1. Find data insights:

→ Quicksights ML offers an anomaly detection tool.

→ the tool can perform anomaly detection and notify management.

2. Forecasting:

→ Quick sight ML tool can perform accurate forecasting to predict critical business metrics.

3. Generate Auto-Narratives:

powered by NLP, Quicksight's ML can offers narrative and tell a story!.

→ Quicksight Security:-

Elastic Map Reduce (EMR)

Big Data platform.

Apache Spark, Hive, Hbase.

→ EMR allows more dynamic Scalability using Amazon EC2

→ Storage of Amazon S3.

Load the data → S3 → EMR → S3

Compute

1. Ease to use.
2. Cost Effective
3. Elastic. → Auto Scaling
4. Enhanced Reliability
5. Improved Security
6. Increased flexibility.

Spot instance:-

EMRFS:- EMR:- → allows direct acess to Hadoop Environment

allows for flexible, lower-level acess to tools beyond spark.

(***) Domin 2: Exploratory Data analysis

→ ML pipelines :-

Problem defination → Data source → Data preparation
→ Model training → Model evolution →
Model deployment → Model Monitoring

→ Data source → Kinesis (Real time data)
→ Data preparation → Analysis and Visualization
pre processing and Feature engineering
{ QuickSight, Aws Sagemaker } Ground thruth.
→ Model train :- Model evolutions
↳ Deployment → Monitoring
Aws Sagemaker Hmp
operationalization Cloud watch.

→ 70 to 80%. Data Sourcing + Data preprocessing

(**) Demo :-

→ Aws Sagemaker :
→ Three Dimension → 3,
statistical :-

→ {Demo Statistical Data Analysis using Aws :- }

→ 75% Max/. →

* Skewness :- less than 1 → highest Skewness
Kurtosis → larger Than Zero, "z"

→ Aws {} Aws Segemaker :-
SPICE → Quick sight ←
Manage Data → New data
↳ upload file as csv.
→ Visualize data
→ sales price →

→ Data preparation × Using AWS.

imbalance Data Different scale inconsists formats

Difficult preceptation Missing Data Outlier

Higher Dimensions Highly Correllated features

Malformed

Distrubution.

→ User/system entry errors

→ Heterogenous data sources → scaling

→ Hard facts → Not enough Data.

Imbalance Data:-

→ affect Classification problems.

98% → legit } → Legit

2% → fradulant } → Legit

→ Legit

ML under security threat.

Possible - Solution

under sample majority class

Over sampling minority class

Synthetic Data → SMOTE.

Scale of features:-

→ Inconsistant scale. inconsist formats.

Currency $ Vs € : Always

Solution:- Levenshtein distance ↓

→ Categorical Data:-

→ Label encoding → unique number, not recomd for larger categories

→ One-hot encoding → Sparsecity.

→ Missing Data problem → optional fields.

→ failure of input systems

Drop them :-

Simple

Can lose Critical information

→ Ignoring Missing Values

KNN Wants Values.

Classification trees.

Impute Missing Values:

mean, mode, Predicting

Outliers:

malcom Dat entry

Data processing

extremly rare boy - representative.

Conditions.

they upsets statistical Charatistic.

Solution for outliers

finding outliers ⟶ Z-score.

IQR options

Box and Scatterplot - — Remove

— correction.

⇒ High Dimensolity :-

Curse of dimensolity

→ Very bad

→ challanging to vizulize

→ increase risk of over fitting

→ feature engineering.

Feature Selection. Dimensionality Reduction

Reducing dimension of data

PCA :-

objective of pca is to reduce from n-dimension dataset to K-dimension dataset, by finding K vectors onto which to project the data so to minimize the projection error.

→ Highly Correlated features:-

voilation of regression.

Find highly Correlleted features, remove.

Variance Influance factors, VIF

$X_j = c_1 x_1 + c_2 x_2 + .. + x_n.$

$V.I.P = \frac{1}{1-R^2}$

$R^2 = 1$, it will infinity.

VIF = 1 not Correllated

VIF = 1 to 5 = moderately correlated

VIF > 5, highley correlated, Remove the feature.

Data ingestion and Transformation

→ Data ingestion

Kinesis datastream, firehouse, data Analytics

↓

Massive Scale

durability 7days

↓

Manual. Shard

Configurations.

Use case:-

~~Data Storing~~

Data Storage of streaming

data is needed

Replay of streaming data is

need

→ Real-time processing is a

requirement.

Kinesis firehouse:

fully managed service. doesn't

required manual configuration for scale

Use Cases:-

→ its near real time...

→ Data will be delivered into S3, RedShift,

or Elasticsearch

→ Amount of Streaming data is some what

unpredictable.

Kinesis Video Stream:-

→ No manual configuration required

→ To integrate video data with

facial recognition in Amazon Rekognition

Kinesis Data Analytics:-

→ Enables queries to run against

the input data

Data transformation:

Amazon EMR: Big Data Platforms
- Spark
- Hbase, Hive, Flink, Presto.

AWS Glue:- Fully Managed ETL, Serverless
Glue Crawler → metadata, Schema,
AWS Glue Catlog,
S3, dynamo DB., RDS, Redshift

AWS Data Firehouse Data transform:-
provides the ability to invoke
Lambda function on incoming data
Doesn't require Configuration for scaling

→ 5minutes invocation time, payload
limit of 6MB.

AWS AI Services:-

→ Language Analysis
→ Dataset Analysis → Provides Sub-Services that can be utilized, categorizes documents by topic
→ Sentiment Analysis

* Amazon Comprehend :- extracts insights from text
* Amazon Lex :- Chatbot, for Customer Service.
* Amazon Polly :- ~~text~~ ~~Speech~~ ~~to~~ text to speech.
* Amazon Textract :- data from Scanned documents
* Amazon Translate :- NMT → different languages
* Amazon transcribe :- Speech to text from audio.

Support both batch and real.

→ Auto making generating
→ Includes a Sub-service that supports medical use case.
→

:- Data set Analysis AI Services :-

1. Amazon forecast :
   - → Deep AR, historical time series data
   - → Supply Chain need, demand.
   - → product sales data
2. Amazon Fraud detector :-
   - → Fraud payment fraudulent payment transactions
   - → Identifying fraudent account creations
   - → Identifying fradulent actions taken in a legimate account.
3. Amazon personalize :-
   - → Recommendation Engine.
   - → Content or product recommendations.
   - → Enable Personalise promotions.

Computer vision :- Rekognition

→ Provides auto maked analysis of image and video data.

→ automated for Batch and Real-time

→ Include text extraction, object detection, and face / emotion detection

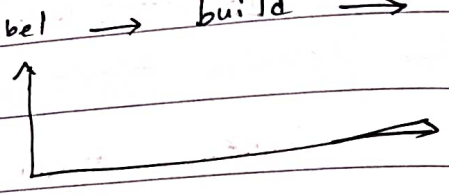
use cases :-

→ Enabling search across large library of digital content.

→ facial Authentication System

→ image moderation.

1. Sage maker proce...



Sagemaker ground :- Label the data; can coverage both human and automated Classification.

→ Utilizing a provided workforce.

→ third party through Amazon Mechanical truck.

→ Enabling an automated labeling workflow

→ Amazon Sagemaker notebooks :-

integrate storage volumes that Can persist notebook instance.

→ Sagemaker Studio :- integrated development for ML.

Provides support for building Jupyterlab, deeply integrated to Sagmaker

→ Auto pilot :- building, training and tuning of machine learning model.

– classification and Regression.

→ Model tuning :-

→ Sagemaker Experiments :-

↳ Autopilot

Sagemaker Neo :- Provides model optim and Cross-platform compilation

→ Enables model execution in th Cloud and on the edge.

→ Hardware : intel, Nvidia. ARM

** * Amazon Augmented AI (A2I):-

→ integrates human review into inference process.

→ provides a configurable workflow for situation where accuracy is critical.

→ Directly integrates into AWS textract, Rekognition.

** → Compute Services for ML:-

→ Amazon-EC2:- Deep learning AMI.

→ Conda AMI.

→ Base AMI.

→ Container Services:- Training or inference can be leveraged on any of the AWS Container Services.

AWS-deep Learning Containers are Pre-configured for specific frameworks.

→ AWS Deep learning Containers are available from ECR

→ Amazon EMR:-

Jupyter notebook, Multiple frame works.

- Tensorflow,
- Mxnet.

→ on-premise servers:-

You can build pre-built sagemaker Docker image.

Use cases:-

→ Training a model on your hardware and uploading for use in sagemaker

→ Pulling a model from sagemaker for inference on your hardware

→ Deploy

Deployment Approaches:-

1. Hosting services:-
   Deploy an inference end point to integrate inference to your workflow or applications

2. Batch transform:- perform an inference job on an aws entire dataset stored in Amazon S3.

** Hosting Services:-

→ Need to deploy inference to Secure Api endpoint

→ Will be integrate inference into an application on workflow

→ Don't have all of the data you will need to perform inferen on in the future.

** Batch transform:-

→ Donot need an exposed endpoint for inference

→ Already have the entire dataset on which you want to perform inferen

→ Need to process an incoming dataset to remove records prior to inference.

→ post-training Validation of model fit

→ Chaining Multiple models

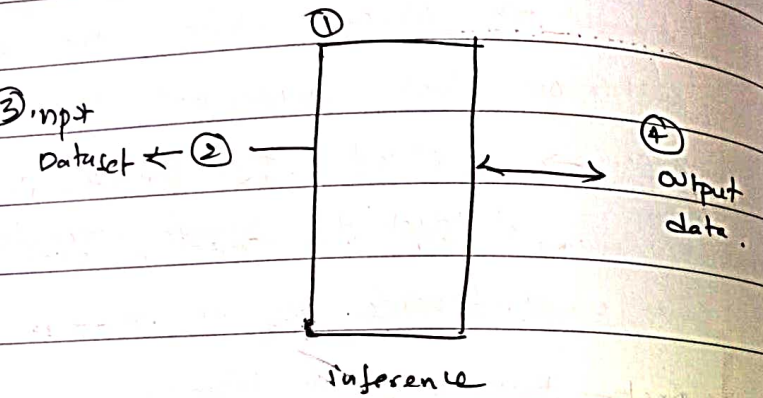
→ predictions will be served outside of Sagemak

→

a) Train and Store mode
b) Load data set for inference into Amazon S3.
c) Specify Configuration for Batch transformation Job.
d) Execute job on dataset.

⇒ Results are stored in S3.

Batch transform Configuration:



→ infrastructure can be configured with number of instances and container

→ they strategy ~~dicted~~ dictates how data to be pulled from the input source.

⇒

Transform data source input defines the sour S3

→ The Assemble with parameter dictates how ~~how~~ the data will be output

→ the transform output data location specifies the output S3 bucket.

⇒

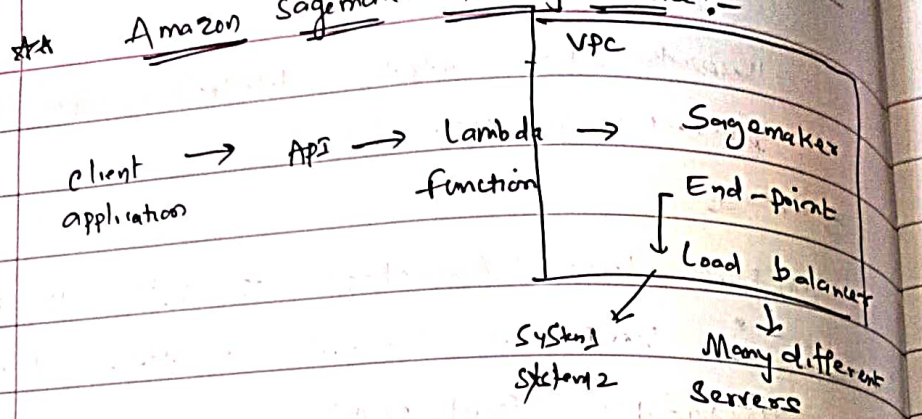
Infrastructure Approches

1. Managed: Sagemaker will launch infrastructure using sagemaker container using the specified number of instances.

2. Custom Container:-

You can specify a docker container image to use for batch transform inference process.

Amazon Sagemaker Hosting Service :-



Deploy :-

* Create model in Sagemaker → Create an Endpoint Configurations

↓

Create an End point from Configuration

→ Endpoints are Secured with HTTPS.

→ End point can auto Scaled based on the defined Minimum and Maximum

→ End-point Can be updated with out downtime.

→ Supports A/B testing through multiple Production Variants.

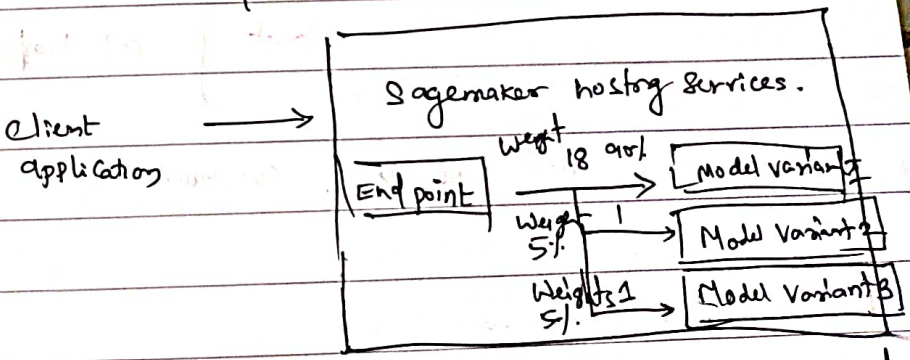
→ ~~A/B testing:~~ Multiple Production Variants:

Production Variants are defined in the end point configuration.

→ The 'intial Variant weight parameter determines the traffic to the Variant

→ Weights per Variants can be updated by Updating the end point Configuration.

Model production variants



gather the data, which model is used

V. Important

Deploying a Model using Hosting Services

Steps in Sagemaker.

→ Training of model
→ Save train Model
→ Creating End-point Configuration
→ Deploying End-Point
→ implementing Multiple Production Variants in an end point.

1. Create a Model → Registring.
2. Create End point ~~Creation~~ configaration
   - Instance type
   - instant ance count
   - Model name

   } all traffic.
3. Creat End point creations. ←

→ implementing Multiple production Variants:-

A/B testing :-

Variants 1 } → initial Weights 1 → 50%.
Variants 2 } 1 → 50%.

↓ Production ready Codes.

Model 1 working very high accuracy

Amazon elastic inference
↓
GPU-(EI)

Multi model End point

End point can
Serve Multiple distinct models
↳ Support multiple production variants on Multiple models.

Does not support ground → Ec2 instance
inference → Complete
Can Reduce.

Inference pipelines.

Support two to five containers.

Http → requests

work flow without erro

(#) securing Sagemaker implementation

→ Security and Compliance is Shared Responsibity between Customer and AWS

→ Revieing the shared Responsibilty model in terms of Sagemaker

→ Eaming high-level Security and compliance considerations

→ Exploring VPC configurations for Sagemaker

→ Controlling access to resources and data using Iam.

→ Revieing how to secure sagemaker notebooks

AWS Responsibity	Customer Responsibity
→ Auss & training for Amzn Employee	
→ Global data Centres and underlying network.	
→	

→ Dance like nobody is watching. Encrypt like every one is.

* Data at Rest
S3, encryption
S3, key
use kms key for encryptions.

* Data Intransit.

→ Sagemaker verified that its data outputs are encrypted at rest.

→ Specific Nitro-based instance are encrypted but not with KMS.

→ Data volume in use

1. Notebook instance
2. processing Jobs
3. Training jobs
4. Batch transform jobs
5. Hosting
6. Hyper

(*) Data in Transit:

within sagemaker inter-network data is encrypted in transit with TLS 1.2

* TLS 1.2 *

→ Some areas are not encrypted by default.

- Communication between control plane ad traing Job Instance
- Node Communication in distributed Processing jobs.
- → Node Communication in distributed training Jobs.

→ Sage maker Data Security:-

(Imp) Amazon Macie:- Fully Managed Service focused on data Security

→ utilizes Machine learning to discover and report on data stored in Amazon S3.

→ Alerts Organization on several Conditions

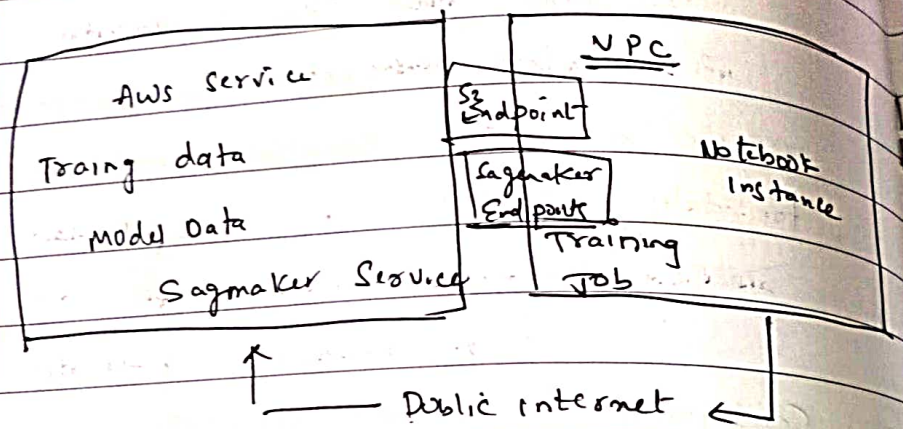
1. Usuall data acess patterns
2. Configuration error for Sensitive data.

Enables automated Compliance Checking for data Storage.

(**) → Configuring for VPC for ML:-

→ To properly secure the infrastructure as well as data the infrastructure can acess, infrastructure should be launched inside of a VPC. Efforts should be taken to minimize or eliminate data travelling over the public internet.

→ Using VPC End Points for AWS S3.



TSL 1.2.

VPC Endpoints:-

1. Gateway endpoints (S3, dyanodb)
2. inference Endpoints (all other Services)

inference end points leverage Aws private links.

VPC's with no public internet need no do not require a NAT or Internet gateway

→ Controlling Access with IAM.

→ Authentication:- Validates that user is who they say that are by validating the submitted Credentials.

→ Authorization:- Controls what a specific validated user can do on the platform based on their configured permissions.

→ Multi-factor Authentication:-
Provides additional security in the authentication process be leveraging either a physical or virtual device that generates token for login to AWS.

→ IAM identities
- User
- Roles: assumes permission for task
- Groups → Allows to manage permissions for a group of iam users

* Policies in AWS IAM :-

→ A Json document that defines permision for AWS IAM identity

→ Defines both the AWS Services that identity can acess and what actions can be taken on that Service

→ Can be either Customer manged or managed by aws

→ You should define least Privilage acess for each user using Iam policies.

(***) Securing Sagemaker notebook :-

→ Sagemaker Notebooks are internet enabled by default.

→ internet acess can be eliminated by launching notebook in a VPC

→ if not launched in Vpc instance is launched in sagemaker-managed Vpc.

Soc 1,2,3 → Sagemakers.

(***) Implementing Highly availabelty and fault tolerance :-

Reliabilty onaws :-

Falut tolerance :-

Being dble to support the failure of Components with in the your architeature.

High availabilty :-

keeping your entire Solution running in the expected manner despite issues that may occur.

Scaling Sagemaker End points :-

→ Vertical Scaling :-

You scale up your instance type to a larger instance type with additional resources

→ Horizontal Scaling :- You ~~scale~~ scale out and additional instances to handle the demand of your applications.

→ * Standrad instances :- t2, t3, m4, m5, m5d.

→ * Memory optimized r5, r5d

→ * Elastic inference :- (eia1, eia2)

→ * Compute optimized (C4, C5, C5d)

* Accelerated Computing (P2, P3, g4dn, inf1)

→ Sagemaker autoScaling :-

Can be Configured, Cloud watch metrics to define Scaling Policy

→ An initial Metric to utilize for scaling is Sagemaker VariantInvocations Per Instance

⇒ the average times per minute that each instance invoked for a variant.

→ Additional metrics Could be utilized.

- cpu utilization.

→ Configuring autoscaling for sagemaker End Point :-

* Deployment Methodologies :-

Continouse Integration :- a development practice that requires developers integrate code into Shared repository for several times in day.