1. Common AWS Data Pipeline Architecture

Typical pipeline:

1. Data source (DB/API/logs)
2. Ingestion (Kinesis / Glue / DMS)
3. Storage (S3 Data Lake)
4. Processing (Glue / EMR / Lambda)
5. Analytics (Athena / Redshift)
6. Visualization (QuickSight)

---

2. S3 (Data Lake Backbone)

- Object storage
- Scalable + cheap
- Used for Data Lake

Best practices:

- Partitioning: `year=2026/month=01/day=31/`

- Use Parquet for analytics
- Enable versioning + encryption

---

3. AWS Glue

What is Glue?

Serverless ETL service.

Key components:

- Crawlers (infer schema)
- Data Catalog
- Glue Jobs (Spark ETL)

Used for:

- ETL from S3 → Redshift
- Transform raw data to cleaned data

---

4. Amazon Redshift

Data warehouse for analytics.

- Columnar storage
- Optimized for OLAP queries
- Works with BI tools

---

5. Athena

- Query data directly in S3 using SQL
- Pay per query

Best for:

- quick analytics on Data Lake

---

6. Kinesis

Streaming ingestion service.

Used for:

- real-time logs
-

event-driven pipelines

---

7. Lambda

Serverless compute.
Used in pipelines for:

- lightweight transformations
- event triggers (S3 upload processing)

---

8. Glue vs EMR (Very Asked)

Feature	Glue	EMR
Setup	Easy	Complex
Server management	No	Yes
Best for	ETL	big data processing
Cost	pay-per-job	pay-per-cluster

---

9. Data Warehouse vs Data Lake

	Data Lake	Data Warehouse
Data type	raw + structured/unstructured	structured
Use	storage + ML	analytics
Example	S3	Redshift

10. AWS Interview Questions

- What is S3 partitioning?
- Explain Glue crawler
- How to build ETL pipeline?
- How to secure S3?
- Redshift vs Athena?

□ PDF #4: ML / AI Interview Questions (PDF Content)

Title Page

ML & AI Interview Questions
Concepts + Metrics + Model Selection
Prepared by: *Amandeep Kaur*
Version: 1.0

---

1. Supervised vs Unsupervised

- Supervised: labeled data (classification/regression)
- Unsupervised: no labels (clustering, anomaly detection)

---

2. Bias vs Variance

- Bias: underfitting
- Variance: overfitting

---

3. Key Metrics

Classification

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

Regression

- MAE
- MSE / RMSE
- R2 Score

---

4. Overfitting Control

- regularization (L1/L2)
- dropout
- early stopping
- cross validation

---

5. Logistic Regression vs SVM vs Random Forest

- Logistic regression: linear, interpretable
- SVM: works well in high-dimensional spaces
-

Random forest: robust, handles non-linear

---

6. Deep Learning Basics

- activation functions: ReLU, sigmoid, tanh
- optimizer: Adam, SGD
- loss: cross entropy, MSE

---

7. Model Deployment Questions

- What is data drift?
- How to monitor model performance?
- What is inference latency?
- Batch vs real-time inference?

---

8. Rapid Fire Questions

- What is gradient descent?

- What is backpropagation?
- What is feature scaling?
- What is PCA?
- Explain confusion matrix