
📘 Domain 1: Data Preparation for Machine Learning

Task 1.1: Ingest and Store Data

Topic 1: Extracting Data from Storage

👶 Explanation (Middle School Level)

Think of your ML data like your schoolwork — scattered across different places:

- Your locker = **Amazon S3** (stores anything: PDFs, images, spreadsheets)
- Your personal USB drive = **Amazon EBS** (only one computer can use it at a time)
- A classroom shared folder = **Amazon EFS** (all students can access it)
- Your teacher's gradebook = **Amazon RDS** (organized rows and columns like Excel)
- An app where you quickly jot down notes = **Amazon DynamoDB** (fast, flexible notes)

Before doing your project (training a model), you need to **gather everything in one place**, just like preparing your homework from multiple sources.

🧪 Example

You're building a model to predict how many students will show up for a field trip:

- Field trip history (CSV files) → stored in **Amazon S3**
- Real-time sign-up app data → **Amazon DynamoDB**
- Class rosters → **Amazon RDS**

- You train the model on an EC2 instance using **Amazon EBS**
- You use **Amazon EFS** to share the updated list between team members on different compute nodes

---

🔍 Storage Services Breakdown

AWS Service	Analogy	Description	Best For
Amazon S3	A giant filing cabinet in the school office	Object storage for all types of files	Training data, documents, backups
Amazon EBS	Your personal USB drive	Fast block storage for EC2	Temporary storage for ML models
Amazon EFS	A shared folder on the school server	File system shared by many EC2s	Shared processing, distributed training
Amazon RDS	Your teacher's digital gradebook	SQL database	Structured tabular data (e.g., users, products)
Amazon DynamoDB	A whiteboard app on your phone	NoSQL key-value store	Fast, flexible app data for ML

⚙️ Advanced AWS Retrieval Tools

Tool	Used With	What It Does

S3 Transfer Acceleration	S3	Speeds up data movement globally using AWS edge locations
EBS Provisioned IOPS	EBS	Guarantees a number of fast input/output operations per second (IOPS)

✅ Pros and ❌ Cons

Service	**✅ Pros**	**❌ Cons**
Amazon S3	✅ Scalable and cheap ✅ Stores any file type	❌ Slower than local disk ❌ Not a file system
Amazon EBS	✅ Very fast I/O ✅ Ideal for EC2 model training	❌ Can only be attached to one instance
Amazon EFS	✅ Shareable across many EC2s ✅ Grows automatically	❌ More expensive ❌ Linux-only
Amazon RDS	✅ SQL-ready, managed ✅ Ideal for structured data	❌ Not for high-speed bulk reads
Amazon DynamoDB	✅ Super fast ✅ Auto-scales with traffic	❌ Not good for complex relationships or large blobs

---

🎯 Use Cases

✅ When to Use

- **S3**: Store training datasets, logs, and large files
- **EBS**: Store temporary model checkpoints during training
- **EFS**: Allow multiple servers to work with the same data
- **RDS**: Run queries on tabular business data (e.g., customers, products)
- **DynamoDB**: Store app data that's constantly changing (like real-time clicks)

❌ When to Avoid

- Avoid **EBS** for shared access — it's tied to one EC2 instance
- Avoid **RDS** if you need very fast reads across regions
- Avoid **DynamoDB** for data that needs advanced SQL joins or analytics

---

📋 10-Question Assessment: Extracting from Storage

Question 1: Which service is best for storing large CSV datasets for training?

A) Amazon DynamoDB

B) Amazon EFS

C) Amazon S3

D) Amazon RDS

Question 2: You want to speed up large S3 file downloads globally. What should you use?

A) EFS Performance Mode

B) S3 Transfer Acceleration

C) CloudFront

D) DynamoDB Accelerator

Question 3: Which service is best for app-level real-time reads/writes?

A) S3

B) RDS

C) EBS

D) DynamoDB

Question 4: What’s a downside of EBS?

A) Slow performance

B) Cannot be shared between EC2 instances

C) Only supports JSON

D) Expensive for small files

Question 5: You need SQL queries for structured tabular data. Which service do you choose?

A) EFS

B) S3

C) RDS

D) DynamoDB

Question 6: Which feature allows EBS to guarantee fast input/output?

A) Transfer Acceleration

B) Performance Boost

C) Provisioned IOPS

D) Fast Read Mode

Question 7: What is a common use for Amazon EFS?

A) Attach to one EC2 instance

B) Share files across many EC2s

C) Store secrets

D) Run SQL queries

Question 8: What is a limitation of DynamoDB?

A) Doesn’t scale

B) Not low-latency

C) Doesn’t support large binary objects well

D) Requires SQL

Question 9: You’re training a model on EC2 and need fast read/write — which storage?

A) S3

B) DynamoDB

C) EBS

D) FSx

Question 10: What is Amazon S3 best known for?

A) SQL analytics

B) NoSQL caching

C) Block storage

D) Object storage

✅ Answer Key

1. C) Amazon S3
2. B) S3 Transfer Acceleration
3. D) DynamoDB
4. B) Cannot be shared between EC2 instances
5. C) RDS
6. C) Provisioned IOPS

7. B) Share files across many EC2s

8. C) Doesn’t support large binary objects well

9. C) EBS

10. D) Object storage