Data Requirements Template

Use this template to document data requirements for your AI/ML project. Answer the five key questions first, then detail each data source.

Project Overview

Project Name: *[Your project name]*

Use Case: *[Brief description of the business problem]*

Date: *[Date created]*

Step 1: Five Key Questions

1. What is the primary business objective?

What specific outcome do you want to predict or automate? Be specific.

Example (Personalized Health): Predict cardiovascular risk by correlating wearable data patterns with clinical health markers.

Example (Manufacturing): Predict equipment failures 24-48 hours in advance to enable preventive maintenance.

Example (Fraud Detection): Detect fraudulent transactions in real-time with <100ms latency and <1% false positive rate.

Your Answer: _____

2. Who are the end users?

Who will consume the model's outputs? What decisions will they make?

User Type	How They Use Output	Required Format
e.g., Patient	*Review health insights*	*Mobile app dashboard*
e.g., Physician	*Validate AI recommendations*	*EHR integration*
e.g., Operator	*Act on maintenance alerts*	*SMS/Email notification*

Your Answer:

User Type	How They Use Output	Required Format

3. What regulatory frameworks apply?

Check all that apply and note specific requirements.

Framework	Applies?	Key Requirements
HIPAA	☐ Yes ☐ No	PHI protection, BAA required
PCI-DSS	☐ Yes ☐ No	Card data encryption, audit trails
GDPR	☐ Yes ☐ No	Consent, right to erasure, DPO
SOX	☐ Yes ☐ No	Financial audit controls
FERPA	☐ Yes ☐ No	Student record protection
State Privacy Laws	☐ Yes ☐ No	*Specify: CCPA, etc.*

Additional Notes: _____

4. What is the expected data volume?

Estimate the scale for capacity planning.

Data Source	Records/Day	Records/Month	Retention Period
e.g., Wearable data	*1,440 (1/min)*	*43,200*	*24 months*
e.g., Transactions	*50,000*	*1.5M*	*7 years*
e.g., IoT sensors	*86,400 (1/sec)*	*2.6M*	*12 months*

Your Estimates:

Data Source	Records/Day	Records/Month	Retention Period

Total Storage Estimate: _____ GB/month

5. What is the required data freshness?

How current does the data need to be for the use case?

Category	Latency Requirement	Your Need
Real-time	< 1 second	☐
Near real-time	1 second - 5 minutes	☐
Hourly	Refreshed every hour	☐
Daily batch	Overnight processing	☐

Category	Latency Requirement	Your Need
Weekly/Monthly	Periodic updates	☐

Justification: *Why does the use case require this freshness?*

Step 2: Data Sources

Data Source 1: *[Name]*

Source Type: *[API / Database / File Export / Stream]*

Owner/Vendor: *[Who controls this data?]*

Access Method: *[REST API / JDBC / SFTP / Kafka]*

Authentication: *[OAuth / API Key / Credentials]*

Update Frequency: *[Real-time / Hourly / Daily / On-demand]*

Fields

Field Name	PHI	PII	Sensitivity	Data Type	Required

Data Source 2: *[Name]*

Source Type: *[API / Database / File Export / Stream]*

Owner/Vendor: *[Who controls this data?]*

Access Method: *[REST API / JDBC / SFTP / Kafka]*

Authentication: *[OAuth / API Key / Credentials]*

Update Frequency: *[Real-time / Hourly / Daily / On-demand]*

Fields

Field Name	PHI	PII	Sensitivity	Data Type	Required

Data Source 3: *[Name]*

Source Type: *[API / Database / File Export / Stream]*

Owner/Vendor: *[Who controls this data?]*

Access Method: *[REST API / JDBC / SFTP / Kafka]*

Authentication: *[OAuth / API Key / Credentials]*

Update Frequency: *[Real-time / Hourly / Daily / On-demand]*

Fields

Field Name	PHI	PII	Sensitivity	Data Type	Required

Summary Checklist

Before proceeding to schema generation, verify:

- ☐ All five key questions answered with specific details
- ☐ All data sources identified with access methods
- ☐ Every field classified for PHI, PII, and Sensitivity
- ☐ Data types specified for all fields
- ☐ Regulatory requirements documented
- ☐ Data volume estimates completed
- ☐ Data freshness requirements justified

Next Steps

1. **Generate JSON Schema** - Use AI to create validation schemas from field classifications
2. **Create Data Dictionary** - Document detailed field specifications
3. **Define Validation Rules** - Establish data quality checks
4. **Generate Sample Data** - Create synthetic test data