**Data Management Plan (Updated)**

**1. 数据来源 (Data Sources)**

RecipeNLG (full\_dataset.csv)  
来源: https://recipenlg.cs.put.poznan.pl/  
内容: ~2M 条食谱，包含字段 title, ingredients, directions, source, link, NER。

OpenNutrition (opennutrition\_foods.tsv)  
来源: https://www.opennutrition.app  
内容: ~300k 食品营养数据，包含字段 id, name, alternate\_names, description, nutrition\_100g, ingredients, labels。

**2. 数据预处理 (Pre-processing)**

* 将 opennutrition\_foods.tsv 转换为 CSV 格式，方便批量上传。
* 在 full\_dataset.csv 前新增 recipe\_id 字段（从 0 递增），保证每条食谱有唯一主键。
* 统一使用 UTF-8 编码，避免 DynamoDB 导入时出现解析错误。

**3. 存储位置 (Storage)**

* Amazon S3 (Simple Storage Service)
  + Bucket 名称: tp33-data-recipe
  + 用途: 存放原始数据文件（CSV/TSV），作为 DynamoDB 导入的来源，同时作为数据备份。

**4. 数据库架构 (Database Architecture)**

**4.1 DynamoDB 表 – Recipes\_v2**

* 主键 (Primary Key): recipe\_id (String)
* 二级索引 (GSI): gsi\_title\_prefix
* 字段: recipe\_id, title, title\_lc, title\_lc\_first1, ingredients, directions, link, source, NER

用途: 存储并查询所有菜谱信息，支持按标题前缀搜索。

**4.2 DynamoDB 表 – Foods**

* 主键 (Primary Key): id (String)
* 二级索引 (GSI): gsi\_name\_prefix
* 字段: id, name, name\_lc, name\_lc\_first1, alternate\_names, description, nutrition\_100g, ingredients, labels, serving, source

用途: 存储食材与对应的营养信息，支持前缀搜索。

**5. 数据导入 (Data Ingestion)**

* 使用 AWS Console 的 **Import from S3** 功能，将 CSV 文件导入 DynamoDB。
* Recipes 数据 (~2M) 分批导入，Foods 数据 (~300k) 单次导入即可。
* 导入完成后通过 AWS CLI 验证：
* aws dynamodb get-item --table-name Recipes\_v2 --key '{"recipe\_id":{"S":"0"}}'

**6. 数据访问 (Data Access via APIs)**

* **AWS Lambda**: recipes-api, foods-api, nutrition-match-api
* **API Gateway**: 提供 RESTful API，包括：
  + GET /recipes?recipe\_id=...
  + GET /recipes?title\_prefix=...
  + GET /foods?id=...
  + GET /foods?name\_prefix=...
  + POST /match

**7. 数据增强 (Data Enhancement – Iteration 1)**

* 在 recipes-api 增加标签逻辑（habits、categories），支持参数 habit 和 category 的查询与过滤。

**8. 数据安全与访问控制 (Security & Access Control)**

* IAM 用户 & 角色分工明确，Access Key 存放在 ~/.aws/credentials，不硬编码。
* API Gateway 开启 CORS，允许前端安全访问。

**9. 数据备份与恢复 (Backup & Recovery)**

* 原始 CSV 文件存放在 S3，可随时重新导入。
* DynamoDB 可选启用 PITR（Point-in-Time Recovery）。

**10. 数据量控制 (Data Volume Management)**

* Iteration 1: 使用 Python 脚本抽样，生成 ~50k–100k 规模数据集。