EY Open Science Data Challenge 2026

Optimizing Clean Water Supply

优化清洁水供应

Project Documentation / 项目文档

Prepared by Antigravity Assistant

February 14, 2026

Contents

Chapter 1

Introduction / 项目介绍

1.1 Objective / 目标

The **EY Open Science Data Challenge 2026** focuses on a critical global issue: **Optimizing Clean Water Supply**. The primary objective is to develop robust machine learning models capable of predicting water quality across various river locations in South Africa.

EY Open Science Data Challenge 2026 聚焦于一个关键的全球问题：**优化清洁水供应**。主要目标是开发稳健的机器学习模型，预测南非各地河流的水质状况。

Participants must predict three key water quality parameters: 参赛者必须预测三个关键的水质参数：

1. **Total Alkalinity / 总碱度**: Measures the water's ability to neutralize acids. (衡量水体中和酸的能力)
2. **Electrical Conductance / 电导率**: Indicates the concentration of dissolved salts. (指示溶解盐的浓度)
3. **Dissolved Reactive Phosphorus / 溶解性反应磷**: A nutrient often associated with pollution and agricultural runoff. (一种通常与污染和农业径流相关的营养物质)

1.2 Background / 背景

Clean water is essential for life, yet pollution and climate change threaten water sources globally. By leveraging satellite imagery (Landsat), climate data (TerraClimate), and ground-level measurements, this challenge aims to identify key factors influencing water quality and predict future conditions to aid in sustainable water management.

清洁水对生命至关重要，但污染和气候变化威胁着全球水源。通过利用卫星图像 (Landsat)、气候数据 (TerraClimate) 和地面测量数据，本次挑战旨在识别影响水质的关键因素，并预测未来状况，以协助可持续的水资源管理。

1.3 Timeline / 时间表

Milestone / 里程碑	Date / 日期
Enrollment Opens / 报名开始	January 20, 2026
Evaluation Start / 评估开始	March 14, 2026
Finalists Announced / 决赛名单公布	April 1, 2026
Challenge Ends / 挑战结束	May 6, 2026

Snowflake Summit / Snowflake 峰会

Top-performing teams using the Snowflake platform will be invited to the Snowflake Summit in San Francisco, June 1-4, 2026.
使用 Snowflake 平台表现优异的团队将受邀参加 2026 年 6 月 1 日至 4 日在旧金山举行的 Snowflake 峰会。

Chapter 2

Eligibility & Rules / 参赛资格与规则

2.1 Eligibility / 参赛资格

The challenge is open to / 挑战赛面向以下人群开放:

- **University Students / 在校大学生**: Currently enrolled in an accredited institution. (目前在正规院校就读)
- **Early Career Professionals / 早期职业专业人士**: Individuals with less than 5 years of professional experience. (拥有少于 5 年专业经验的个人)

Note / 注意

Participants who do not meet these criteria may still join the challenge but are **not eligible for prizes**.
不符合上述条件的参与者仍可参加挑战，但**没有资格获得奖品**。

2.2 Team Rules / 团队规则

- Teams may consist of up to **3 members**. (团队最多可由 **3 名成员** 组成)
- Each team member must register individually. (每位团队成员必须单独注册)
- Teams can be mixed (students and professionals). (团队可以混合组成，即学生和专业人士)

2.3 Prizes / 奖项

1. **Winner / 冠军**: $5,000
2. **1st Runner-up / 亚军**: $3,000
3. **2nd Runner-up / 季军**: $2,000

2.4 Intellectual Property / 知识产权

Participants retain full ownership of any intellectual property developed during the challenge. However, EY encourages open-sourcing the winning solutions to benefit the broader scientific community.

参赛者保留在挑战赛期间开发的任何知识产权的完全所有权。然而，EY 鼓励开源获奖解决方案，以造福更广泛的科学界。

Chapter 3

Resource Inventory / 资源清单

3.1 Downloaded Resources / 已下载资源

All resources have been successfully downloaded, extracted, and organized in the `resources/` directory. 所有资源均已整理在 `resources/` 目录下。

3.1.1 Documentation / 文档

- **Participant_Guidance.pdf**: Full official guide. (完整官方指南)
- **snowflake_guide.md**: Archived "Getting Started" guide for Snowflake. (归档的 Snowflake 入门指南)
- **challenge_rules_faq.md**: Archived official rules and FAQs. (归档的官方规则和常见问题)

3.1.2 Data / 数据集

Location / 位置: `resources/data/`

- `water_quality_training_dataset.csv`: Historical training data (2011-2015). (历史训练数据)
- `submission_template.csv`: Template for predictions. (预测结果提交模板)

3.1.3 Code: Snowflake Platform / 代码：Snowflake 平台

Location / 位置: `resources/code/snowflake/`

Deep Dive: Snowflake Package / 深度解析：Snowflake 包

The files in this directory are specialized for the Snowflake Cloud Data Platform. 此目录下的文件专为 Snowflake 云数据平台优化。

Core Files / 核心文件:

- **snowflake_setup.sql**:
  - *Purpose*: Sets up network rules to allow your Snowflake environment to talk to the Microsoft Planetary Computer API.
  - *Action*: Must be run first in a Snowflake Worksheet.
- **GETTING_STARTED_NOTEBOOK.ipynb**:
  - *Purpose*: Validates that your environment is correctly configured and can fetch a sample satellite image.
- **BENCHMARK_MODEL_NOTEBOOK_SNOWFLAKE.ipynb**:
  - *Purpose*: An end-to-end example. it loads the training data, features, trains a model (Random Forest/XGBoost), and creates a submission file.

Data Extraction / 数据提取:

- **LANDSAT_DATA_EXTRACTION_NOTEBOOK_SNOWFLAKE.ipynb**:
  - *Purpose*: Queries the Landsat Level-2 satellite data repository. It handles geospatial filtering to match the river locations.
- **TERRACLIMATE_DATA_EXTRACTION_NOTEBOOK_SNOWFLAKE.ipynb**:
  - *Purpose*: Extracts climatological data (precipitation, temperature) which are strong predictors for water quality.

3.1.4 Code: General Platform / 代码：通用平台

Location / 位置: `resources/code/general/`

Deep Dive: General Package / 深度解析：通用包

These notebooks are designed to run in any standard Jupyter environment (Local, Colab, Kaggle). 这些笔记本设计用于在任何标准 Jupyter 环境中运行（本地、Colab、Kaggle）。

- **Benchmark_Model_Notebook.ipynb**:
  - *Content*: Contains a standard Scikit-Learn pipeline. It demonstrates data pre-processing, feature merging, and model training.
- **Landsat_Data_Extraction_Notebook.ipynb**:
  - *method*: Uses the `pystac-client` library to search the Microsoft Planetary Computer catalog for satellite scenes.
- **requirements.txt**:
  - *Critical*: Lists all necessary Python libraries (e.g., `rasterio`, `pystac`, `geopandas`). Run `pip install -r requirements.txt` before starting.

3.1.5 Media / 多媒体

Location / 位置: `resources/media/`

- `Orientation_Session.mp4`: Project overview video. (项目概览视频)
- `How_to_Get_Started.mp4`: Step-by-step startup guide. (逐步启动指南)
- `Tips_for_Success.mp4`: Useful tips. (成功秘诀)

Chapter 4

Setup Guide / 环境配置指南

4.1 Choose Your Path / 选择您的路径

You can participate using either the **Snowflake Platform** (Highly Recommended) or a **General Environment** (Local/Cloud Jupyter).

您可以选择使用 **Snowflake 平台**（强烈推荐）或 **通用环境**（本地/云端 Jupyter）参与。

4.2 Option A: Snowflake (Recommended) / 选项 A：Snowflake（推荐）

1. **Sign Up / 注册**: Use the dedicated 120-day trial link provided in the resources. (使用资源中提供的专用 120 天试用链接)
2. **Setup / 设置**:
   - Log in to your Snowflake account. (登录您的 Snowflake 账户)
   - Open a worksheet and run the content of `resources/code/snowflake/snowflake_setup.sq` (打开工作表并运行 setup.sql 的内容)
   - This script configures external access integrations needed for satellite data. (此脚本配置卫星数据所需的外部访问集成)
3. **Upload / 上传**: Upload the notebooks from `resources/code/snowflake/` to your Snowflake workspace. (上传 snowflake 目录下的笔记本到您的工作区)
4. **Run / 运行**: Open `GETTING_STARTED_NOTEBOOK.ipynb` to verify your setup. (打开 Getting Started 笔记本验证设置)

4.3 Option B: General Environment / 选项 B：通用环境

1. **Environment / 环境**: Ensure you have Python 3.8+ and Jupyter installed. (确保安装了 Python 3.8+ 和 Jupyter)
2. **Dependencies / 依赖**: Install required libraries (pandas, numpy, scikit-learn, rasterio, etc.) using `requirements.txt`. (使用 requirements.txt 安装所需库)
3. **Data / 数据**: Place the `water_quality_training_dataset.csv` in your project data folder. (将训练数据集放入项目数据文件夹)
4. **Run / 运行**: Open `Benchmark_Model_Notebook.ipynb` to start building your baseline model. (打开基准模型笔记本开始构建)

4.4 Submission / 提交

1. Train your model using the training dataset. (使用训练数据集训练模型)
2. Generate predictions for the 200 target points in `submission_template.csv`. (为模板中的 200 个目标点生成预测)
3. Save your results as a CSV file. (保存结果为 CSV 文件)
4. Upload to the contest portal. (上传至竞赛门户)

Chapter 5

Snowflake Deep Dive / Snowflake 深度指南

5.1 Why Snowflake? / 为什么选择 Snowflake?

The challenge organizers have partnered with Snowflake to provide a powerful, cloud-native environment for data engineering and machine learning. Using Snowflake allows for seamless handling of large geospatial datasets (like Landsat) without local storage constraints.

挑战组织者与 Snowflake 合作，为数据工程和机器学习提供强大的云原生环境。使用 Snowflake 可以无缝处理大型地理空间数据集（如 Landsat），而不受本地存储限制的影响。

5.2 Deep Dive: External Access / 深度解析：外部访问

To access satellite data from the Microsoft Planetary Computer, Snowflake needs specific network permissions. The `snowflake_setup.sql` script handles this.

为了访问 Microsoft Planetary Computer 的卫星数据，Snowflake 需要特定的网络权限。`snowflake_setup.sql` 脚本负责处理此问题。

```
-- Example of what the setup script does / 设置脚本示例
CREATE OR REPLACE NETWORK RULE planetary_computer_rule
  MODE = EGRESS
  TYPE = HOST_PORT
  VALUE_LIST = ('planetarycomputer.microsoft.com');

CREATE OR REPLACE EXTERNAL ACCESS INTEGRATION planetary_access
  ALLOWED_NETWORK_RULES = (planetary_computer_rule)
  ENABLED = TRUE;
```

Security Note / 安全提示

Never hardcode API keys or credentials in your notebooks. Use Snowflake Secrets or environment variables if needed.
切勿在笔记本中硬编码 API 密钥或凭据。如有需要，请使用 Snowflake Secrets 或环境变量。

Chapter 6

FAQ / 常见问题

6.1 General Questions / 一般问题

Q: What is the passing threshold? / 及格门槛是多少？
A: You must achieve an R^2 score of at least **0.4** to receive a certificate of completion.

答：您必须达到至少 **0.4** 的 R^2 分数才能获得结业证书。

Q: Can I use other tools or languages? / 我可以使用其他工具或语言吗？
A: Yes, you can use R, Julia, or other languages, but Python is highly recommended and fully supported with starter code.

答：是的，您可以使用 R、Julia 或其他语言，但强烈推荐使用 Python，并提供完整的入门代码支持。

Q: Can I use external data? / 我可以使用外部数据吗？
A: **Yes**, provided the data is free and publicly available to everyone. This ensures reproducibility. Examples include public weather/climate databases, soil maps, and elevation models.

答：**可以**，前提是数据对所有人免费公开可用。这确保了可重复性。例如公共天气/气候数据库、土壤图和高程模型。

Q: How are teams formed? / 团队如何组建？
A: Teams can have up to 3 members. All members must register individually on the platform.

答：团队最多可由 3 名成员组成。所有成员必须在平台上单独注册。

Q: Who owns the code? / 谁拥有代码的所有权？
A: You (the participant) retain ownership of your intellectual property. However, sharing your solution with the community is encouraged after the competition.

答：您（参赛者）保留您的知识产权的所有权。然而，鼓励在比赛结束后与社区分享您的解决方案。