Microsoft Challenge for IE students

AI Capstone 2025

Microsoft

ie UNIVERSITY

Case Study - Microsoft

The Context

You are part of the innovation team of a very famous **climate laboratory** in the US. Your group oversees **"all data things"** (research, science, engineering, storytelling, design, etc.).

The innovation department is looking for new ideas to monetize US climate data. They want to explore efficient ways to extract insights from climate data scattered across US, to show the information in valuable ways to generate solutions and products to sell, as the have been approached by different corporations and government departments. They prefer **easy-to-deploy/maintain solutions, with clear ROI**. As this is about generating new business lines, that's all the information you have. The good thing is that you can lead the way with your own hypotheses, models, and architectures for the type of customers you like. You are the expert(s).

As the data is vast and the access was very difficult (we have been measuring since 1979 in +800,000 sensors, we have +12Bn data points!), your friendly neighbourhood Microsoft folks have been working hard to release **a new and promising dataset**, which may cover (totally or partially) your project needs and made the data access easy for you. You have decided to use it as a good starting point for your exploration, and the goal is to prepare an **internal proof-of-concept** and to present it to your executive team to **obtain organizational and financial support**.

Case Study - Microsoft
The Dataset

gridMET: A Large-Scale dataset for surface meteorological data in US from 1979 to present

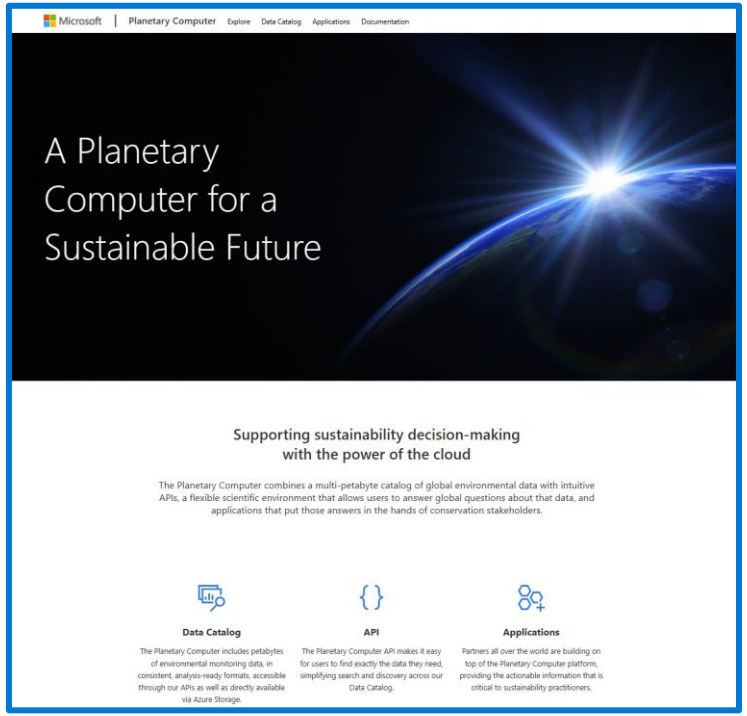
URL: [gridMET | Planetary Computer]

Source: Microsoft Planetary Computer

Sample:

- 12 climate variables (e.g., Wind speed, direction, precipitation…)
- 15341 time points in 585 latitudes and 1386 longitudes

Format: available through STAC API to planetary computer library (example code)*



▸ Dimensions: (**time**: 15341, **lat**: 585, **lon**: 1386, **crs**: 1)

▾ Coordinates: (4)

crs	(crs)	uint16	3
lat	(lat)	float64	49.4 49.36 49.32 … 25.11 25.07
lon	(lon)	float64	-124.8 -124.7 … -67.1 -67.06
time	(time)	datetime64[ns]	1979-01-01 … 2020-12-31

▾ Data variables: (12)

air_temperature	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
burning_index_g	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
dead_fuel_moist…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
dead_fuel_moist…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
mean_vapor_pre…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
potential_evapot…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
precipitation_am…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
relative_humidity	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
specific_humidity	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
surface_downwe…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
wind_from_direc…	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…
wind_speed	(time, lat, lon)	float32	dask.array<chunksize=(30, 585, 1386), meta=n…

▾ Attributes: (19)

Conventions :	CF-1.6
author :	John Abatzoglou - University of Idaho, jabatzoglou@uidaho.edu
coordinate_syste…	EPSG:4326
date :	02 July 2019
geospatial_boun…	POLYGON((-124.7666666333333 49.400000000000000, -124.7666666333333 25.066666666666666, -67.058333300000015 25.066666666666666, -67.058333300000015 49.400000000000000, -124.7666666333333 49.400000000000000))
geospatial_boun…	EPSG:4326
geospatial_lat_m…	49.400000000000000
geospatial_lat_m…	25.066666666666666
geospatial_lat_re…	0.041666666666666
geospatial_lat_u…	decimal_degrees north
geospatial_lon_…	-67.058333300000015
geospatial_lon_…	-124.7666666333333
geospatial_lon_r…	0.041666666666666
geospatial_lon_u…	decimal_degrees east
note1 :	The projection information for this file is: GCS WGS 1984.
note2 :	Citation: Abatzoglou, J.T., 2013, Development of gridded surface meteorological data for ecological applications and modeling, International Journal of Climatology, DOI: 10.1002/joc.3413
note3 :	Data in slices after last_permanent_slice (1-based) are considered provisional and subject to change with subsequent updates
note4 :	Data in slices after last_provisional_slice (1-based) are considered early and subject to change with subsequent updates
note5 :	Days correspond approximately to calendar days ending at midnight, Mountain Standard Time (7 UTC the next calendar day)

Tip: install dependencies: pystac-client planetary_computer numpy pandas xarray zarr adlfs

Case Study - Microsoft

The Challenge

A multi-disciplinary data mandate that should cover:

1. **Initial data exploration and hypotheses**
   a) Technical value and potential limitations of the provided dataset
   b) Different type of analyses at climate level *(describe up to five potential approaches, implement at least one in section 3 – data science)*
   c) Value of the proposed data analyses for the overall business goals of the organisation

2. **Data engineering / prep**
   a) Internal data tools choice *(your own technology stack)*
   b) Main data transformations
      i. Filtering / aggregations at dataset or feature level
      ii. Documented pipeline process and documentation

3. **Data science *(for one specific case)***
   a) EDA / data profiling
   b) Baseline model +Target model choice
   c) Performance metrics choice (e.g., recall, F1 score, etc.)
   d) Model/data iterations and performance improvement
   e) Relation between perf. metrics and business KPIs + preliminary ROI justification

4. **Production-level considerations**
   a) Envisioned end-to-end solution and technical architecture
   b) Legal and IP aspects related to the dataset
   c) Alternative or complementary datasets + potential methodology to build an “in-house” dataset

Case Study - Microsoft

The Deliverables

1. A **PDF file** with the full report (max 15 pages) covering all previously mentioned topics (page 4)
2. A **public Github repo** with at least one model implementation (.ipynb notebook or .py code), related documentation and results
3. An **executive pitch deck** (max 10 slides) to present your approach to the executive level and get their sponsor


Good luck!
Microsoft
ie
UNIVERSITY