Interview at UKHO

Data Egnineer

NA Courtman

1st September 2022

Technical Exercise

Scenario

A research team at the UKHO is responsible for compiling oceanographic data. They take data from multiple sources and every month, aggregate information into a report.

The sources of data are:

- Ocean tides data received via API that is saved as .csv files in a folder on a shared drive by a team member every Wednesday. Ocean tides data is saved in a tabular format with columns:
  - Latitude
  - Longitude
  - Tidal height
- Model output for ocean surface current data in netCDF format that is placed in an AWS S3 bucket every two weeks. The data is downloaded manually from S3 and stored in a folder on a shared drive by a team member. This data can be considered to have columns:
  - Latitude
  - Longitude
  - Current direction (vector)
  - Current strength
- Monthly acoustic data delivered by SFTP for ocean mammal distribution detection. This data is copied from the SFTP server and stored in a folder on a shared drive. The following is provided for each recording:
  - Metadata: latitude, longitude, time of recording
  - Audio file

- Relevant research papers and reports in PDF format that are found by team members and kept in a folder on a shared drive. Papers are not georeferenced but refer to their areas of relevance in their titles/text.

Reports cover specific pre-defined areas or additional areas if requested by the consumers of these reports. These additional areas are received by email and included in the next scheduled report.

The team aggregates this raw data using a mix of scripts inherited from previous team members and manual work. No one person is assigned to run the scripts and it is usually done by whoever gets there first.

The team approaches the organization's data science and engineering function asking for help to reduce the burden of this process which takes a significant proportion of their time due to the number of manual steps.

Topic 1: Data ingest pipelines building

The focus of the first phase of your work is ingestion of data into the new data store. Please describe:

How many pipelines you might build? 1 for each data source; tailor pipeline to activity - separate AWS stacks - possibly in a monorepo

- API/SFTP ingest, use lambda to S3
- S3, copy using S3 SNS trigger (topic maybe)
- Shared drive, see if there's an API for the shared drive, or maybe a scheduled scrape

What ingest approaches might you use for the various data? MVP: set up landing zone, and copy to original source (prevent user deletion/moddifcation), processing step in a lambda (see next), publication data store (eg. s3)

What other steps might you add into the pipeline(s)? Validation, deduplication, updating holdings catalog, cleaning

You are also told that it is common for research papers to be duplicated as the same paper can be flagged by multiple members of the team. Acoustic data can sometimes be corrupted.

How might you tackle these issues? Papers can be cross-checked via names/checksums. Audio corrpution checking using CLI tools.

How might you deal with historic data? Establish validity time periods, use archive data storage (update catalog) (eg. S3 -¿ Glacier), big bang ingestion/data slurp using snowball

Estimated interview discussion time: 10 minutes

Topic 2: Data storage and cataloguing

Initially, the team are interested in simply recreating their existing PDF reports from the newly made data store. To do this, the plan is to simply extract all data types received over the last month and then run slightly modified scripts on this extracted data.

They mention that it's important for them and consumers of the reports to be able to quickly find the original data source for data referenced in the reports together with the data's history.

What approach could you use to store each dataset? Store on a semi-public (unlisted URL) data store, ie. S3 buckets. Or run everything behind a auth login (ie. cognito). Do a big-bang ingest of all available data

How might you make the data easy to access/discover and what tools might you use? - data catalog, S3 console build a quick dashboard to list file contents, linking within PDF to original source

Estimated interview discussion time: 5 minutes

Topic 3: Approach to delivering software

You are tasked to write the code that implements the ingest of the Ocean surface currents model dataset. The software you write should read the data from the netCDF format and load the data into a database that would allow geospatial queries to be run against it.

Describe some of the technologies you might employ and why you have chosen them - Ingest data to PostgreSQL, PostGIS; handles structured data well, indexable, can handle server-side geospatial queries (intersections etc). Alternatively, use GeoPandas, store in S3 as netCDF, or Parquet, or Avros, unlikely CSV.

Explain how you would:

Break the task down into processing steps, what steps do you think are required? :

1. Load netCD - Python using netcdf4 or gdal
2. Convert to geopandas dataframe, using lat and lon as geom
3. Load to database gdf.to_postgis()

Ensure the code will successfully process the data? Run validation checks (ie Null/NaN values, acceptable ranges of lat/lon, current direction and strength - no negatives), defensive coding with TDD should cover this (go over in detail)

Make sure the code is easy for others to maintain and understand? Use clean code techniques, use careful variable and function naming, verbosely named tests (highest feasible coverage), enusre linting

Create a continuous integration and deployment capability? Code stored in a git repo, use Github/lab pipelines, or Jenkins. Deploy to test env. Run smoke tests. Enforce linting and tests pass on pre-merge hooks

A senior data engineer reviews your code and leaves you a comment stating that they want you to consider how you might detect and handle the incoming data feed stopping unexpectedly.

Describe how you might detect and handle this? Either timeouts, or only run lambdas on trigger events

Estimated interview discussion time: 10 minutes

Topic 4: Stakeholder engagement

As a data engineer, you are tasked with the delivery of a feature in the overall solution. The end users would like to, as an interim deliverable, speed up the production of one of their PDF reports. You set up a session with the users and stakeholders to discuss what their requirements are for this interim solution

How would you approach the following?

The process for gathering their requirements Send out agenda ahead of time, provide possbile areas that may contain requirements, eg. look and feel of report, receipt process (email, download), speed expectations

The process for refining the requirements Collate requreiments, Gherkin with dev team, collate items needing clarification and re-engage stakeholders, use sizing estimates, break into tasks. Provide time estimates transparently

Important questions you might ask the team about the challenges and problems they obser
:

1. Where are there current pain points
2. What tasks are repetative currently (ie. what can be automated)
3. What is the current process for viewing/storing the generated reports (should it be replicated or improved)

Estimated interview discussion time: 5 minutes

Behaviours

Communicating and Influencing

Communicate in a straightforward, honest and engaging manner, choosing appropriate styles to maximise understanding and impact

Teaching - SA-V ratio

I am able to communicate complicated concepts to non-technical stakeholders. As a Physics teacher I often helped students understand complex topics by using approriate styles based on the student's needs. One student struggled to understand surface-area to volume ratio. I was aware that she was a physical and visual learner, so I demonstrated the concept using both a processor cooling block, and diagrams on the board which she was able to interact with. I knew that she had grasped the understanding as she had a eureka moment.

Encourage the use of different communication methods, including digital resources and highlight the benefits, including ensuring cost effectiveness

HMLR - whiteboarding and standups for SFLAPI migration

I use a wide variety of communication methods depending on what is appropriate. Whilst supporting a development team migrating platforms at HMLR I engaged the team in whiteboarding sessions which facilitated expediant discussions and helped people who thought visually. I also provided regular, short updates on progress which allowed non-technical stakeholders to make key decisions relating to cost and time management.

Ensure communication has a clear purpose and takes into account people's individual needs

Teaching - Lesson structure, variety of activities

I am able to keep my communications concise, clear, and suited to the audience. Whilst teaching I planned my lessons to have a clear goal (often outlined in the starter activity), developed repeatedly through multiple teaching methods (visual, verbal, physical), and reaffirmed during a plenary phase. This always ensured that all my students where able to learn effectively. This is suited to communications during meetings where purpose and summary can be achieved through good structuring.

Share information as appropriate and check understanding

Teaching - Questioning

I am able to share appropirate information and confirm understanding, and this was an important part of teaching. To achieve it I balanced the detail of topics taught to the set of students I was teaching. More able learners I was able to challenge with stretch problems, whilst less able students I was

able to spend more time developing a grasp of the basics. In order to check their understanding I used (amongst other techniques) discussion sessions with questioning techniques that allowed students to demonstrate understanding of a topic. This gave me insight into what students had grasped, and what needed to be revised.

This would be applicable when engaging with stakeholders as I would be able to ask them to (for example) test a piece of code relying upon a login, explain the login process and ask a question like "can you see X status bar" - relying upon them having logged in. This would confirm that my instructions had been clear.

Show positivity and enthusiasm towards work, encouraging others to do the same

UKHO - Interpolation; geospatial fun

Whilst carrying out spike work regarding interpolation of AIS data, I was aware that the work was likely to be unfeasable due to cost and compute constraints. This caused a sense of "pointlesness" surrounding the task. Undestanding that the task was actually quite an exciting problem to solve using GeoPandas and goespatial joins, I was able to complete the work expediantly. This resulted in decisions about whether to take interpolation forward being made much earlier.

Ensure that important messages are communicated with colleagues and stakeholders respectfully, taking into consideration the diversity of interests

UKHO - Interpolation; teaching Pete, providing facts

Having carried out the investigation into interpolation compute time and cost, I understood the importance of communicating my findings quickly to my delivery manager. In the Teams call I had with him I needed to teach him the basics of what interpolation was, and how that related to the compute (touching on spatial joins). Having succesfully had this meeting; this could then be communicated sensitively to stakeholders, some of which had a deep attachment to interpolation. I provided factual data, including images and compute times to the stakeholders which mitigated against discussions that could have turned nasty.

Working Together

Encourage joined up team work within own team and across other groups

Teaching - Welfare of student

I have regularly contributed to and encouraged joined up team work across teams. Whilst teaching I became aware of a student who was displaing some welfare issues in my classroom. I communicated this to her housemistress as

part of the tutoring team, and additionally to the welfare officer at the school. This contributed to a wider picture of the student's situation, and I as her tutor and teacher, was able to put into place best practices surrounding interactions with her. This resulted in her being better engaged in lessons. I could apply this to working at UKHO by being aware of who my team is, and who amongst my team, and stakeholder teams, would be the person of appropriate resonsibility for any particular issues or discussions.

Establish professional relationships with a range of stakeholders

HMLR - Digital Street

Whilst working at HMLR I carried out a piece of work with multiple external parties. As the infrastructure specialist I needed to communicate complex concepts to non-technical stakeholders, as well as coordinating with external developers. I established professional relationships with these people by exclusively using appropriate channels (email, Teams), as well as maintaining a supportive and helpful attitude. This resulted in the project being successfully, and positively completed.

Collaborate with these to share information, resources and support

HMLR - Onboarding DDRS

I am an excellent collaborator as I can naturally share information, resources and support. Whilst setting up a new development team on the infrastructure at HMLR, I realised that their understanding of the infrastructure and deployment pipelines was limited. I decided to put together a series of training sessions, along with documentation to help improve their understanding. As a result, I was able to quickly on-board them and other teams afterwards which saved on wasted time and therefore costs. This is a skill I can apply at UKHO, by putting together documentation about our practices (which I have already started on), and engage in mentoring during the onboarding new team members.

Invest time to develop a common focus and genuine positive team spirit where colleagues feel valued and respect one another

UKHO - Teams call with James over PR

In my current role, I place a high importance on team spirit within my immediate and wider teams. Upon raising a recent merge request, a colleague had some feedback about moving some test setup into a fixture. I took the oppurtunity to have a Teams call, instead of just a textual back and forth, to work through the potential solution. This not only allowed me to learn more effectively, but also enabled us develop our working relationship, and build up a team spirit.

UKHO - Jiri-contractor

Recently, due to leave, my team was reduced to myself, a permanent and a contractor colleague. I stepped up during team meetings to give direction

on task management. During this meeting I became aware that my permanent collegue was uncomfortable. Instead of bringing it up in the meeting, I contacted him directly afterwards. He was able to respectfully explain his discomfort, and we positively worked through the particular issue and were able to reach a point where he was able to successfully achieve tasks.

Put in place support for the wellbeing of individuals within the team, including consideration of your own needs

UKHO - Mansplaining; my reaction

In my current role, I initially found it tricky when a colleague always seemed to be over-explaining concepts to me. I respectfully took onboard their explanations but began to feel that my abilities were being doubted, and indeed started to doubt them myself. Instead of letting this fester, I discussed this with my line manager. This discussion helped me understand my colleagues intentions and behaviours in a new light, and I have been able to gain confidence in myself, as well as develop a closer professional relationship with my colleague.

Make it clear to all team members that bullying, harassment and discrimination are unacceptable

HMLR - Dylan

In my role at HMLR, there was a particular architect whose attitude had a negative effect on many of my developer colleagues. During a discussion with a developer, the architect abruptly joined in and dismissively addressed my colleague, however, he seemed unaware of the upset he was causing. Being aware that my developer colleague was already at their limit I diffused the situation by drawing the architect off into a private meeting room where I carefully helped him understanding how the way he addressed my colleague had unintentionally caused upset. This gave my developer colleague time to cool down and resulted in the architect engaging with the developer more respectfully in future meetings.

Actively seek and consider input of people from diverse backgrounds and perspectives

Teaching - Tutor group interfaith

Whilst tutoring during my teaching post, part of my role was to encourage cross-cultural understanding. One of the simplest ways I did this was by giving students of different faith backgrounds the oppurtunity to give some background and context to their faith in and around important religions observances. This led to a positive attitude of sharing (often food) and understanding between the students of all the different faith groups represented in my tutor group.