Partly Cloudy with a Chance of Zarr

A Virtualized Approach to Zarr Stores from ECMWF's Fields Database

FOSDEM'26

Tobias Kremer
tobias.kremer@ecmwf.int

ECMWF

WARM WORLD

SPONSORED BY THE
Federal Ministry of Research, Technology and Space

Who Am I?

Tobias Kremer

- Research Software Engineer @ ECMWF
- Former cloud backend engineer
- Working in WarmWorld Easier with colleagues from
  - Forschungszentrum Jülich (JSC)
  - Deutsches Klimarechenzentrum (DKRZ)
  - Max-Planck-Institut für Meteorologie (MPI-M)
  - Climate Service Center Germany (GERICS)
  - Karlsruher Institut für Technologie (KIT)
  - Universität zu Köln
- ECMWF – MARS ecosystem
  - Fields Database (FDB)
  - Python Interface for the FDB (PyFDB)
  - Zarr Interface for the FDB (ZFDB)

What is ECMWF?

Established in 1975, Intergovernmental Organisation

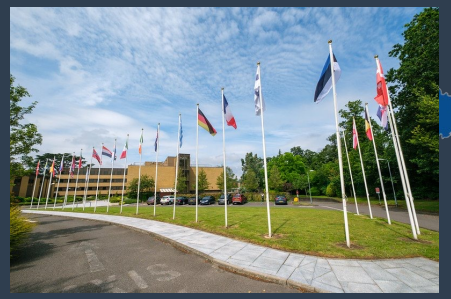
- 23 Member States | 12 Cooperating States
- 500+ staff

24/7 operational service

- Operational NWP – 4x forecasts / day
- Supporting NWS (coupled models) and businesses

Research institution

- Experiments to continuously improve our models
- Reforecasts and Climate Reanalysis



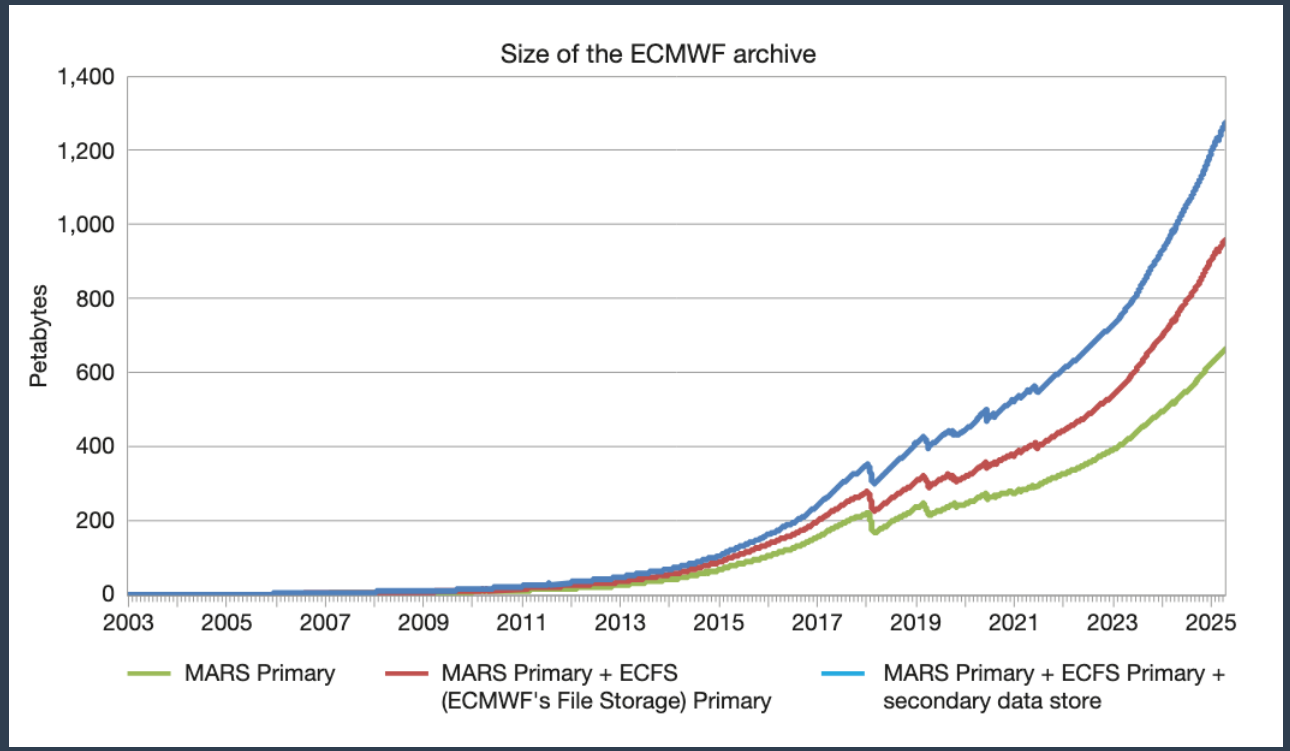
Reading, GB

Bonn, DE

Bologna, IT

Why are we talking about Zarr?

- Explosion of scientific data
  - ECMWF produces ~360 TiB of forecast data per day (projection 2027: 1 PiB)
  - Need for scalabitility and open formats
- Increasing pool of target users
  - Social Science
  - Geologists
  - …
- Many new use cases in the Python / Machine Learning domain



https://www.ecmwf.int/sites/default/files/elibrary/072025/81670-ecmwfs-improving-data-services-in-the-era-of-cloud-computing-and-machine-learning.pdf

Why are we talking about Zarr?

Traditional HPC use-cases:

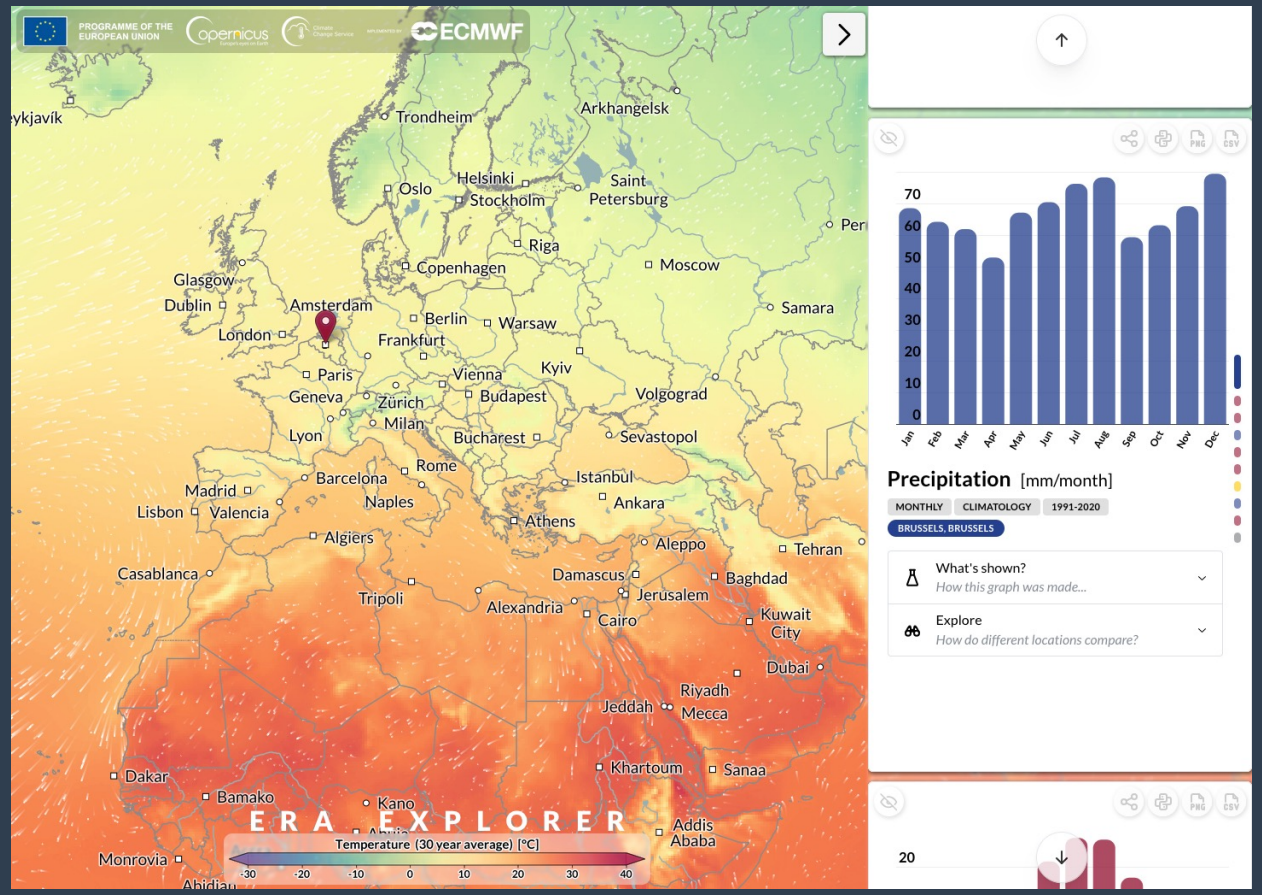
- NWP workflows writing/reading on parallel FS
- Analysis of forecasts in weather and climate

Shift in interest in recent years:

- Hybrid workflows (HPC + cloud)
- Semantic data storage on the cloud
- Interactive analysis / ML training on large datasets
- **A**nalysis-**R**eady **C**loud-**O**ptimized Datasets

Can we give flexible access to data?

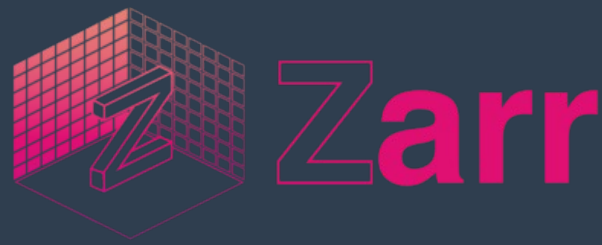
(even without prior domain knowledge?)



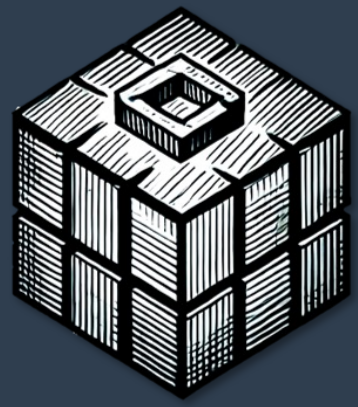
https://era-explorer.climate.copernicus.eu/?lat=50.86&lng=4.35&plot=0

Goal of the talk

- Show how Zarr fits into HPC and open-source ecosystems
- Bridge the gap between classical HPC and modern cloud based solutions



?



FDB

What is the FDB?

Short:

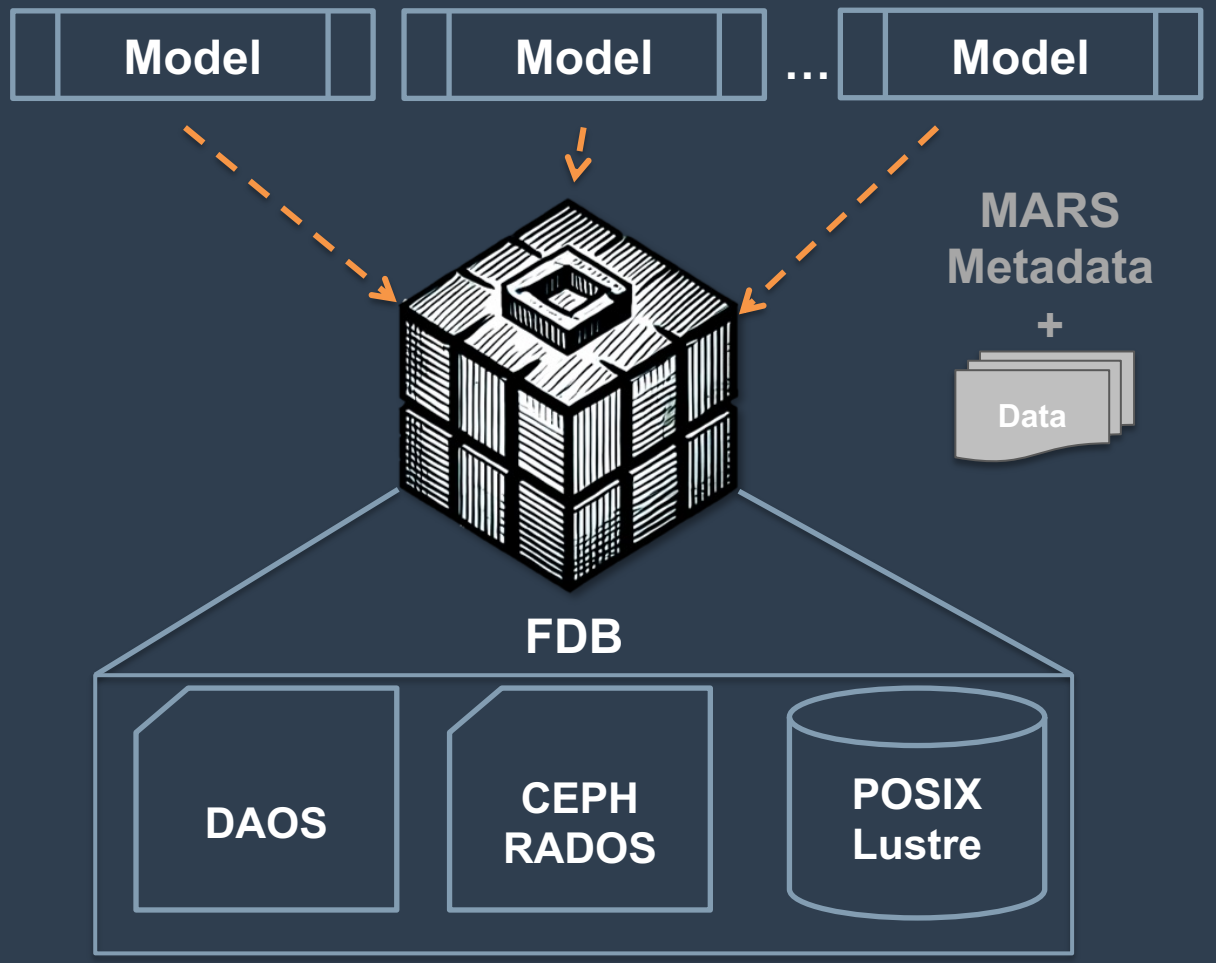
Domain specific high-performance object store
(Transactional, No explicit synchronization, No MPI)

Key-value store

- Keys are scientific meta-data (MARS Metadata)
- Values are byte streams (GRIB, ODB)

Based on data semantics e.g:

```
param=temperature/humidity,
steps=0/to/240/by/3,
date=19990101/to/20251231,
```



What is the FDB?

Short:

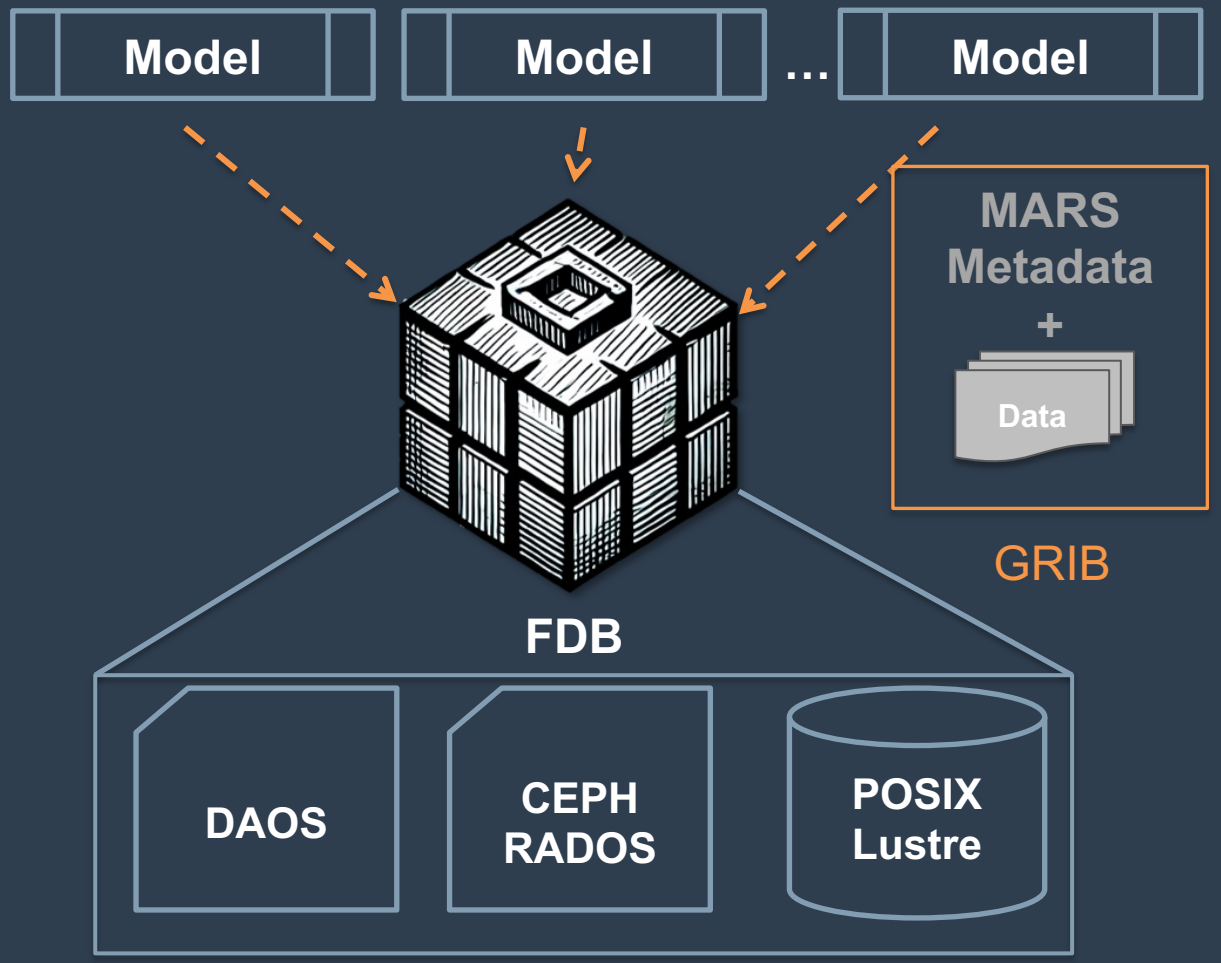
Domain specific high-performance object store
(Transactional, No explicit synchronization, No MPI)

Key-value store

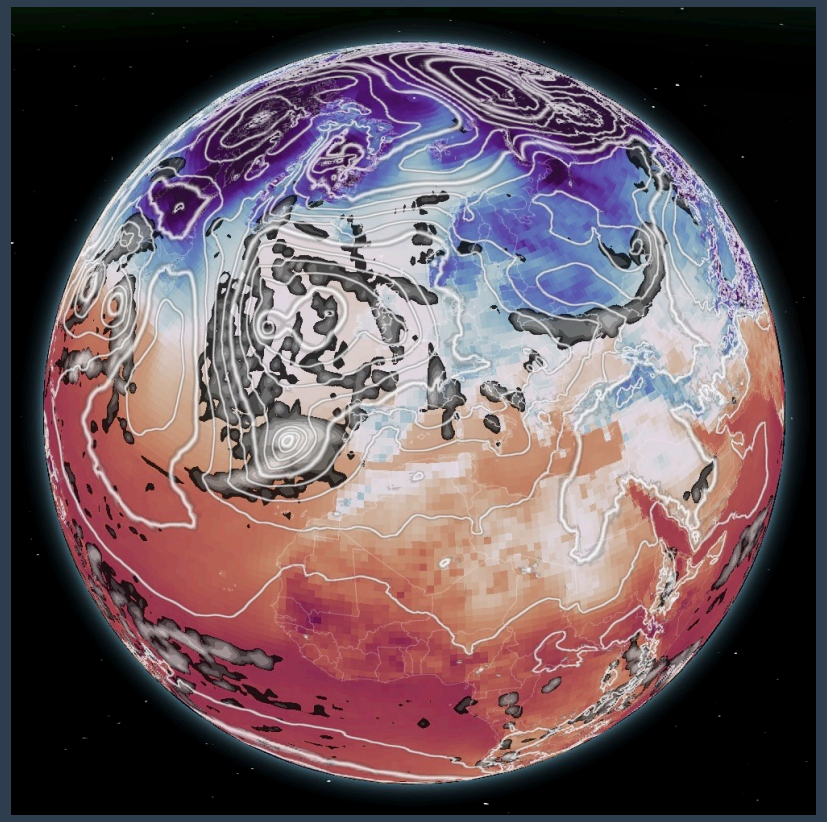
- Keys are scientific meta-data (MARS Metadata)
- Values are byte streams (GRIB, ODB)

Based on data semantics e.g:

```
param=temperature/humidity,
steps=0/to/240/by/3,
date=19990101/to/20251231,
```



What is GRIB?



GRIB - **G**eneral **R**egularly-distributed **I**nformation in **B**inary form

- Stores gridded fields e.g., temperature, pressure
- Standardized by WMO for meteorological data exchange
- Binary, compact, and efficient for large NWP datasets
- Optimized for archival workflows:
    - Sequential access
    - Minimal overhead for operational pipelines
    - Self-describing
- But:
    - Complex structure (requires external tables)
    - Not cloud-native

Grib File Structure on Disk

Section 0 Indicator	Section 1 Identification	Section 2 Local Use	Section 3 Grid Definition	Section 4 Product Definition	Section 5 Data Representation	Section 6 Bitmap	Section 7 Data	Section 8 End

What is Zarr?

Short:

A Library and a ***Format***

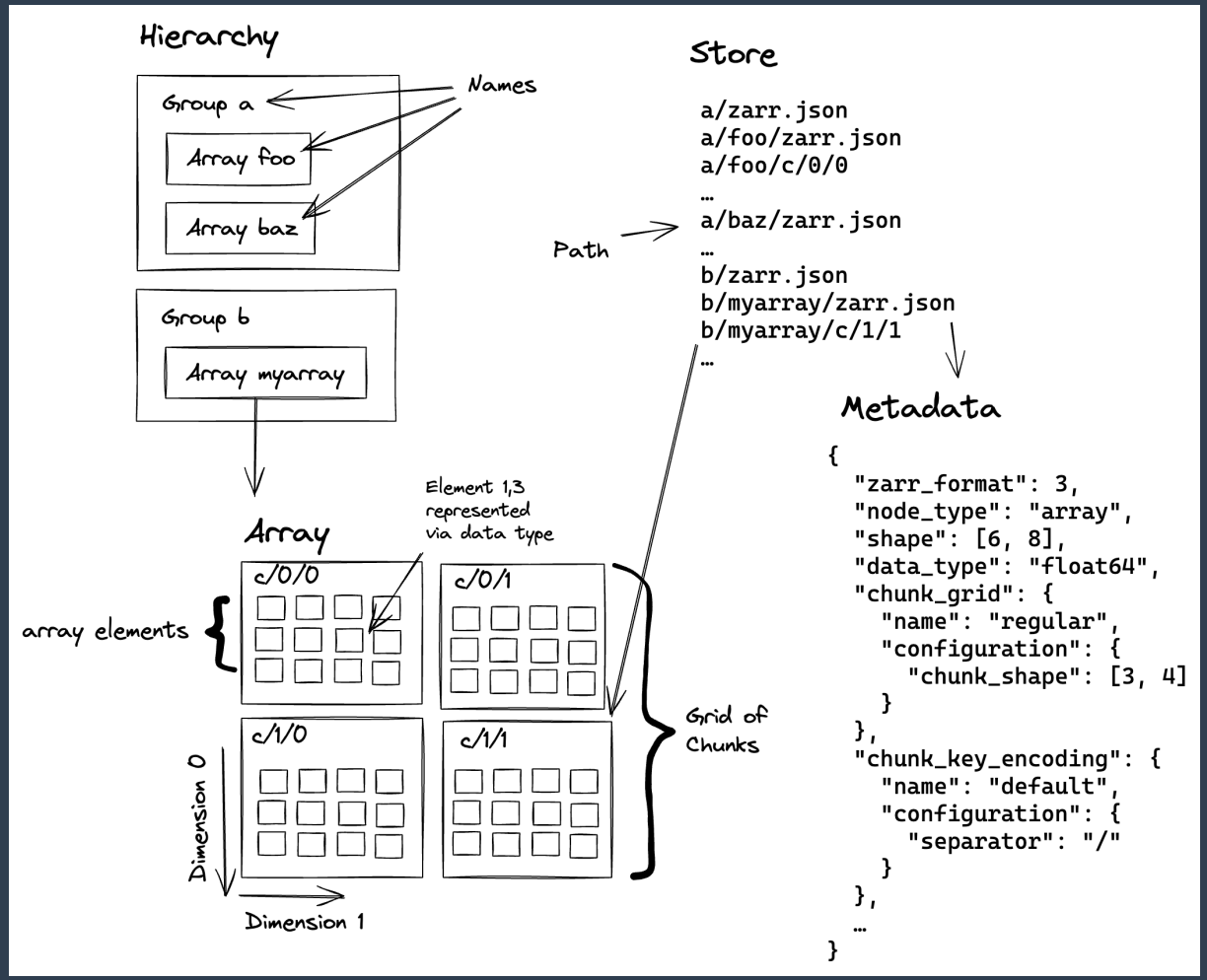
(An Interface and a behavioural contract)

Definition:

Chunked, compressed, N-dimensional arrays

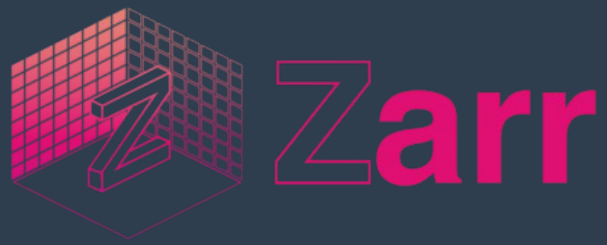
Key features:

- Storage agnostic
- Language interoperability
- Chunking for performance
- Maps onto directories and files:
  - Groups map to Directories
  - Arrays map to Set of Files
  - Chunk map to File

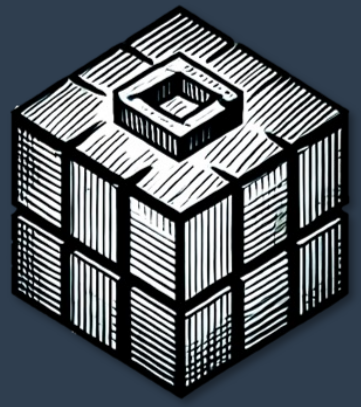


https://zarr-specs.readthedocs.io/en/latest/_images/terminology-hierarchy.excalidraw.png

How to bridge the gap between GRIB and Zarr



?

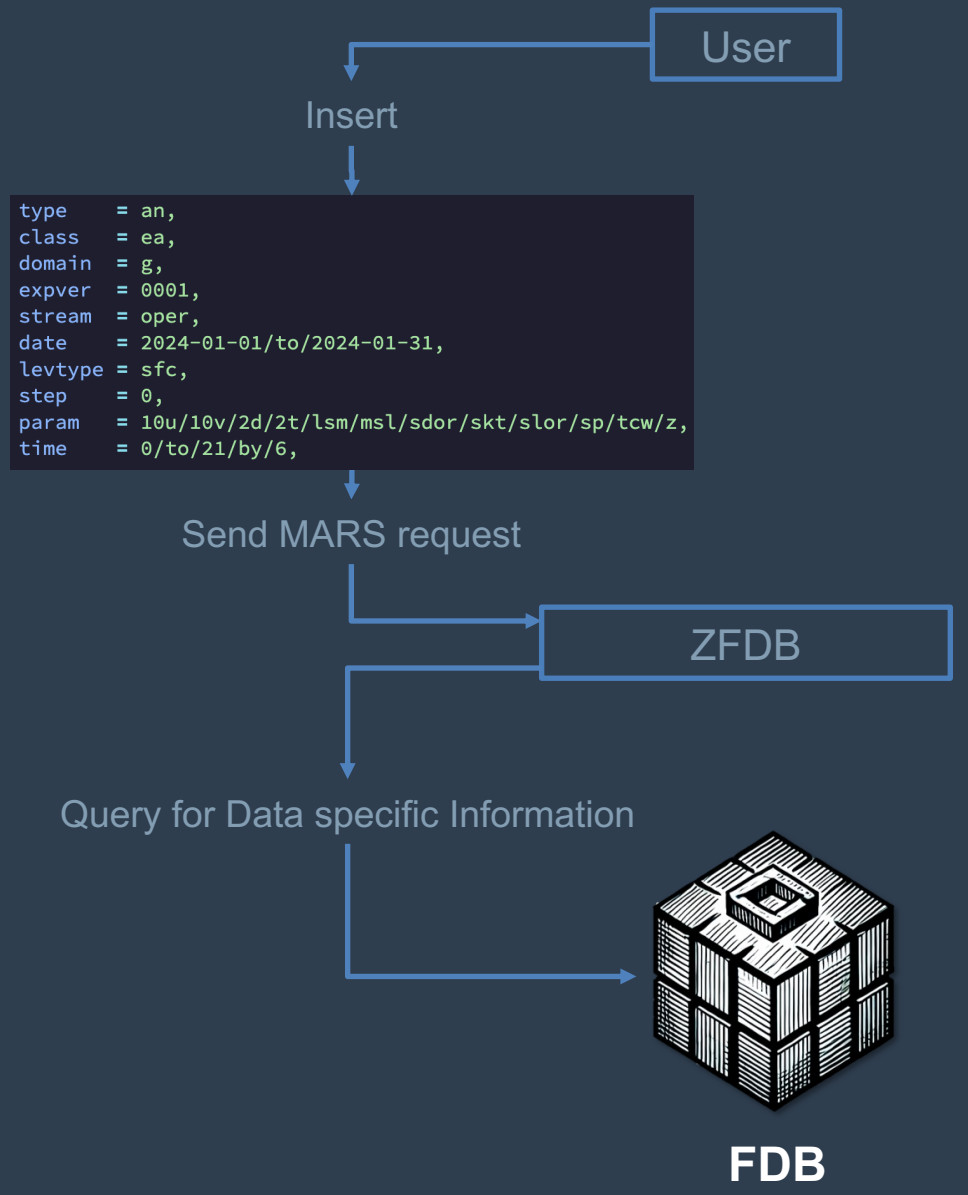


FDB

Local

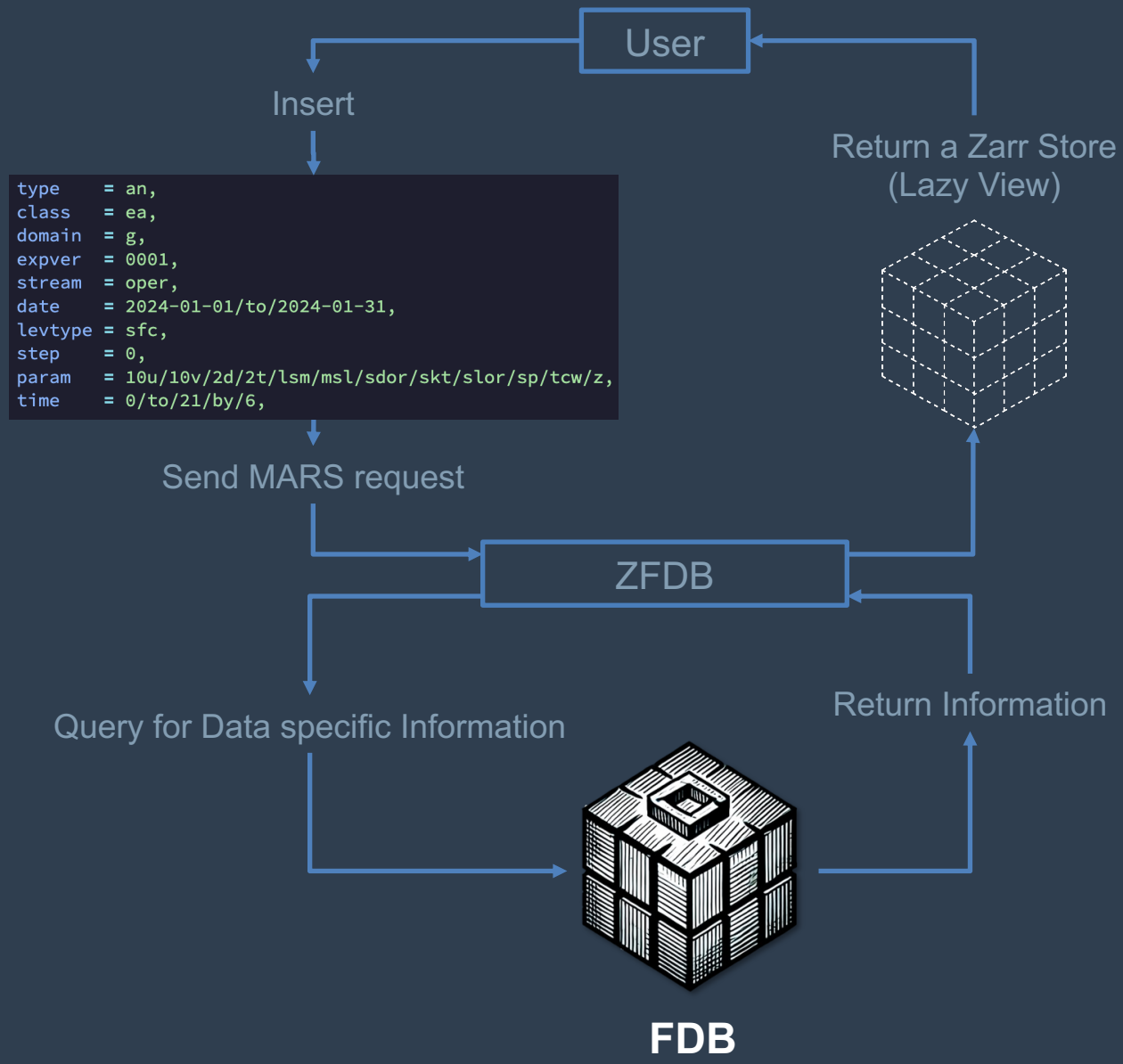
Local - Creating a Lazy View

- Define a view onto GRIB data stored in a FDB
- Retrieve a Zarr-Store
- Zarr's chunking maps to FDB chunking
- View is cached and reused

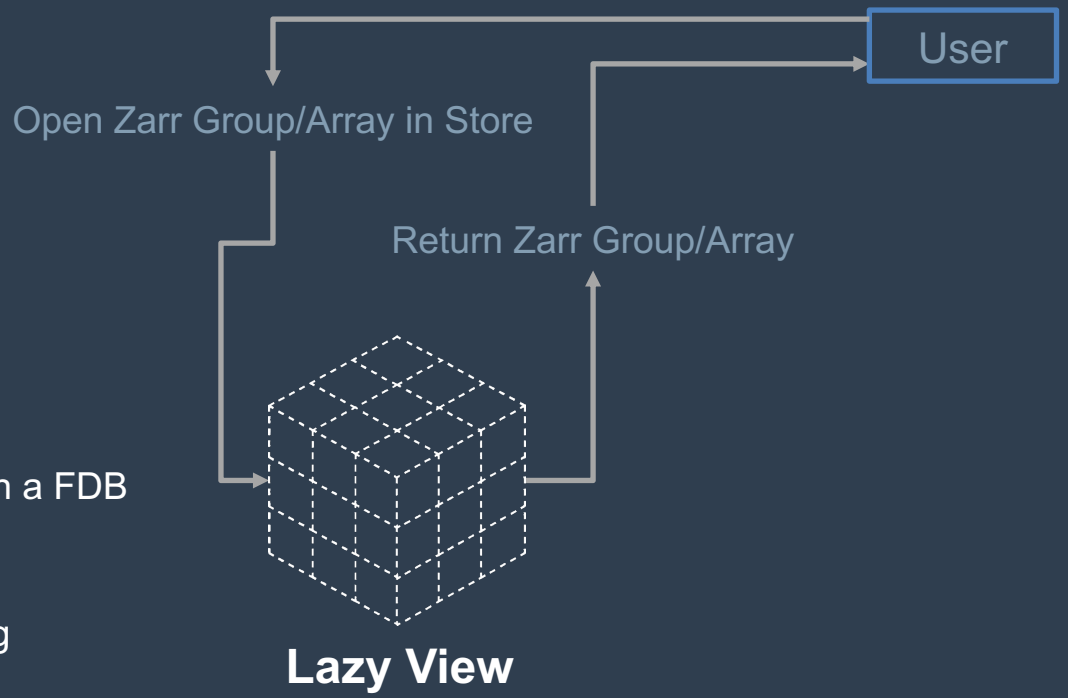


Local - Creating a Lazy View

- Define a view onto GRIB data stored in a FDB
- Retrieve a Zarr-Store
- Zarr's chunking maps to FDB chunking
- View is cached and reused



Local - Query View / Zarr Chunk

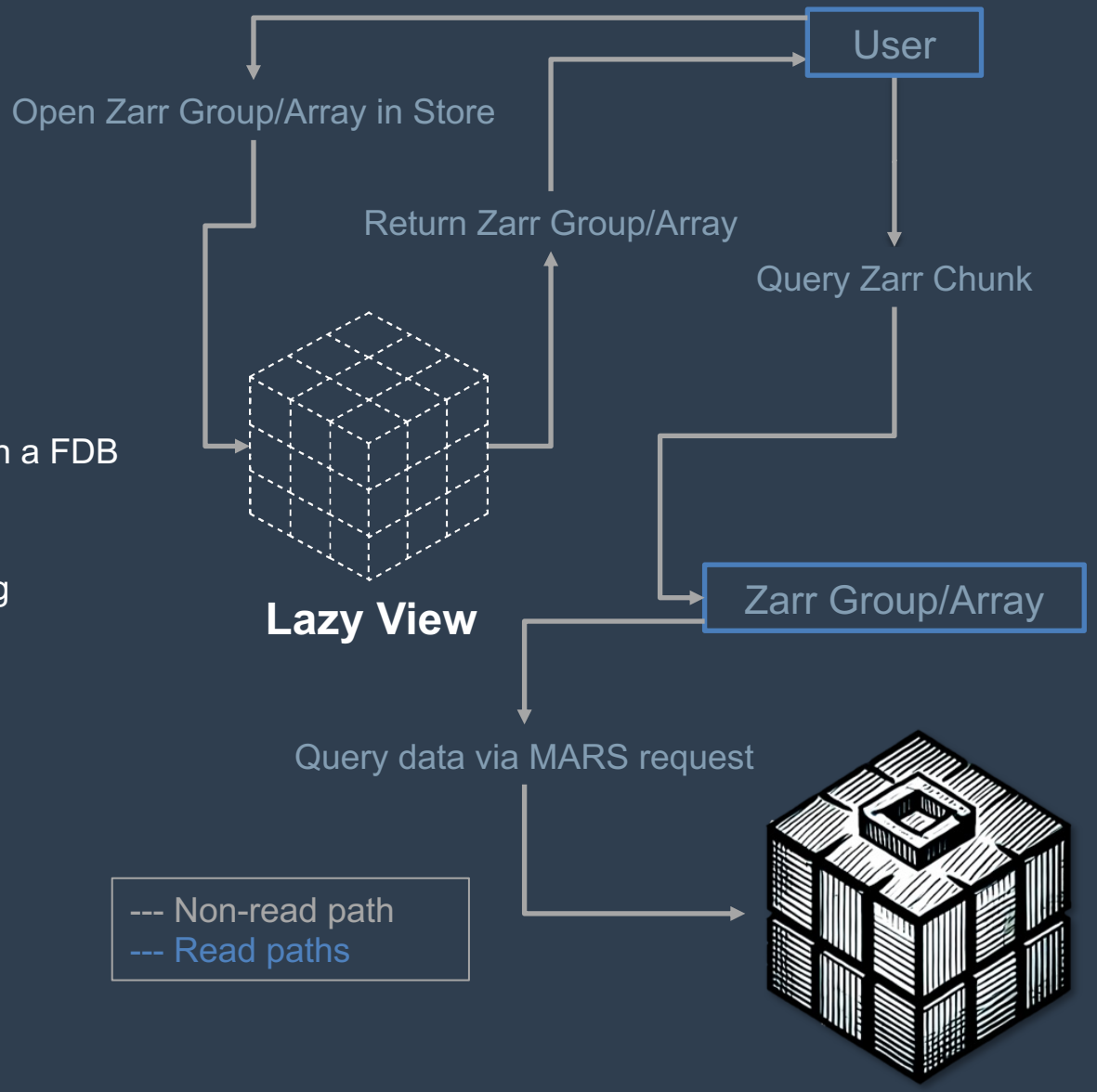


- Define a view onto GRIB data stored in a FDB
- Retrieve a Zarr-Store
- Zarr's chunking maps to FDB chunking
- View is cached and reused

--- Non-read path
--- Read paths

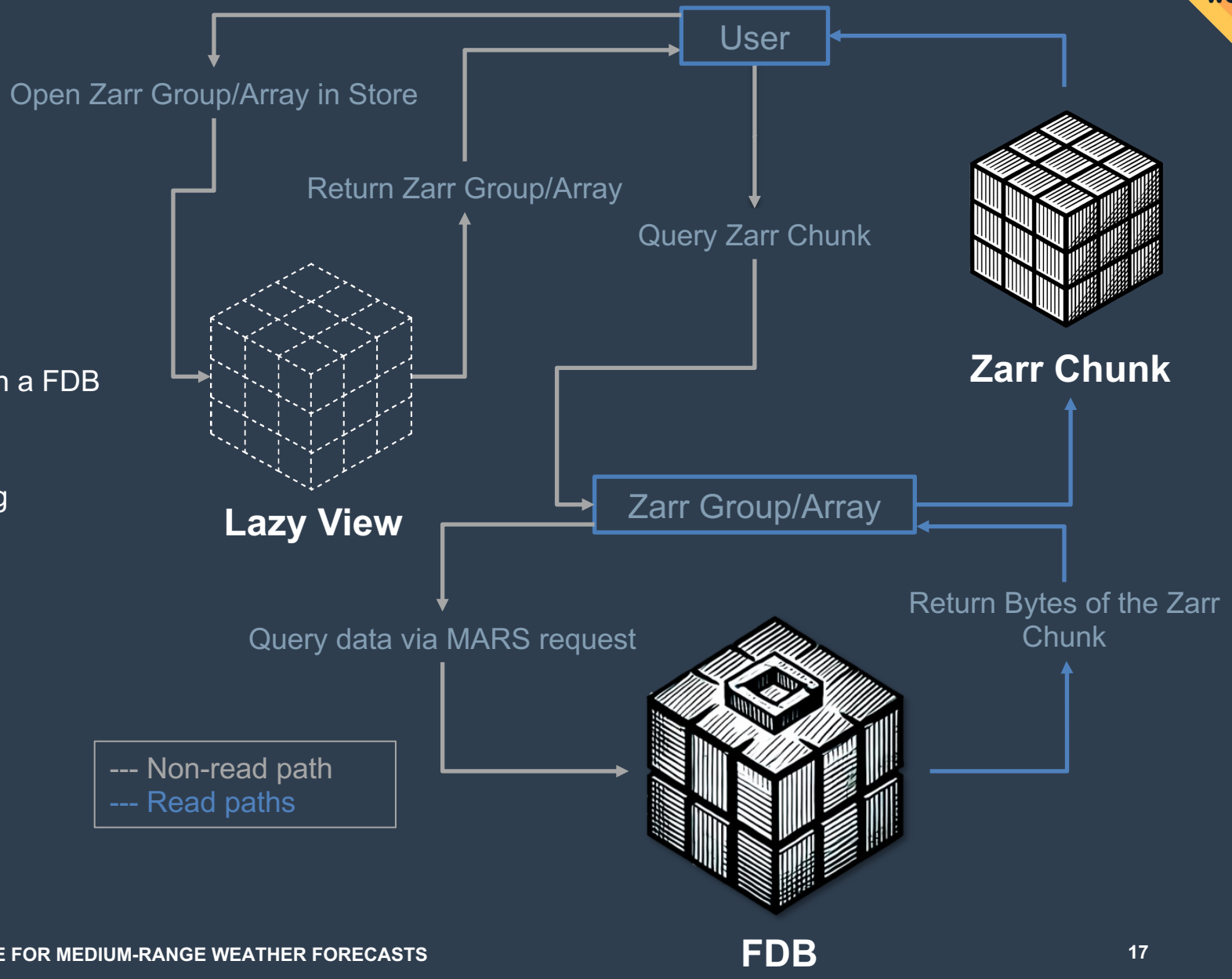
Local - Query View / Zarr Chunk

- Define a view onto GRIB data stored in a FDB
- Retrieve a Zarr-Store
- Zarr's chunking maps to FDB chunking
- View is cached and reused



Local - Query View / Zarr Chunk

- Define a view onto GRIB data stored in a FDB
- Retrieve a Zarr-Store
- Zarr's chunking maps to FDB chunking
- View is cached and reused



```python
def store_creation(fdb_config_path: pathlib.Path) -> zarr.ABC.store.Store
  builder = SimpleStoreBuilder(fdb_config_path)
  builder.add_part(
    "class=ea,type=an,stream=oper,date=2024-01-01/to/2024-01-31,time=0/to/21/by/6,step=0,domain=g,expver=0001,"
    "param=10u/10v/2d/2t/lsm/msl/sdor/skt/slor/sp/tcw/z,"
    "levtype=sfc,"
    [
      AxisDefinition(["date", "time"], True),
      AxisDefinition(["param"], True)
    ],
    ExtractorType.GRIB,
  )
  builder.add_part(
    "class=ea,type=an,stream=oper,date=2024-01-01/to/2024-01-31,time=0/to/21/by/6,step=0,domain=g,expver=0001,"
    "param=q/t/u/v/w/vo/d,"
    "levtype=ml,"
    "levelist=48/60/68/74/79/83/90/96/101/105/114/120/133,"
    [
      AxisDefinition(["date", "time"], True),
      AxisDefinition(["param", "levelist"], True)
    ],
    ExtractorType.GRIB,
  )
  builder.extendOnAxis(1)
  store = builder.build()
  return store
```

results in

Zarr

Store with Numpy Array like Zarr Cube
Size = (32 * 4, 12 + 7 * 13, 40320)

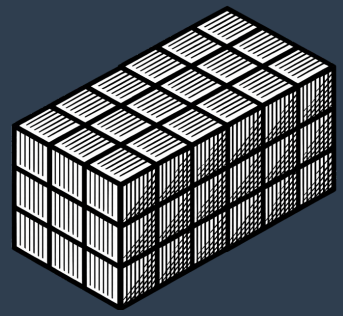
[#Datetime, #Parameter_Levelist, #Implicit Field Entries]

Lazy View

+

Lazy View

results in

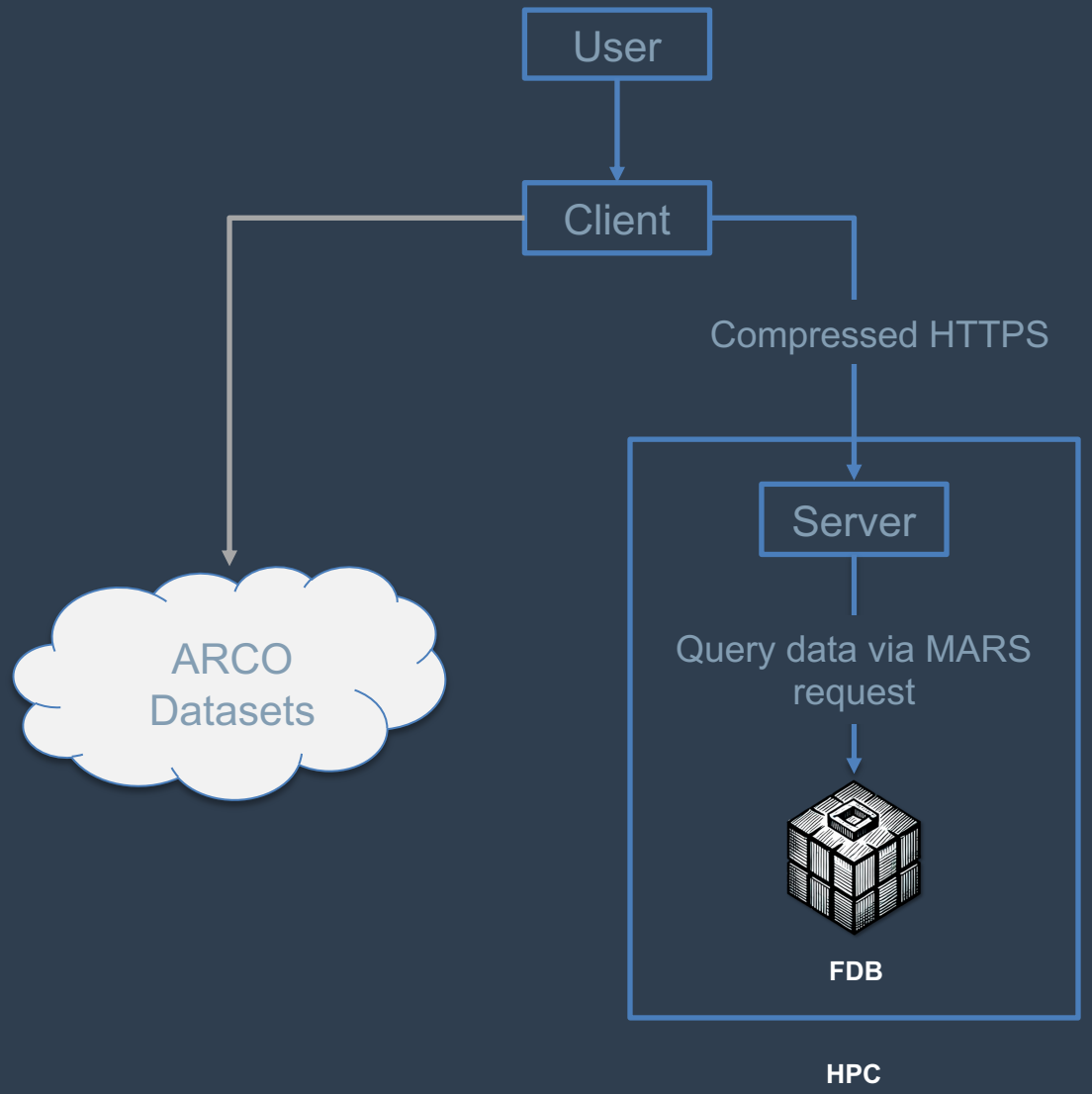


Zarr Array

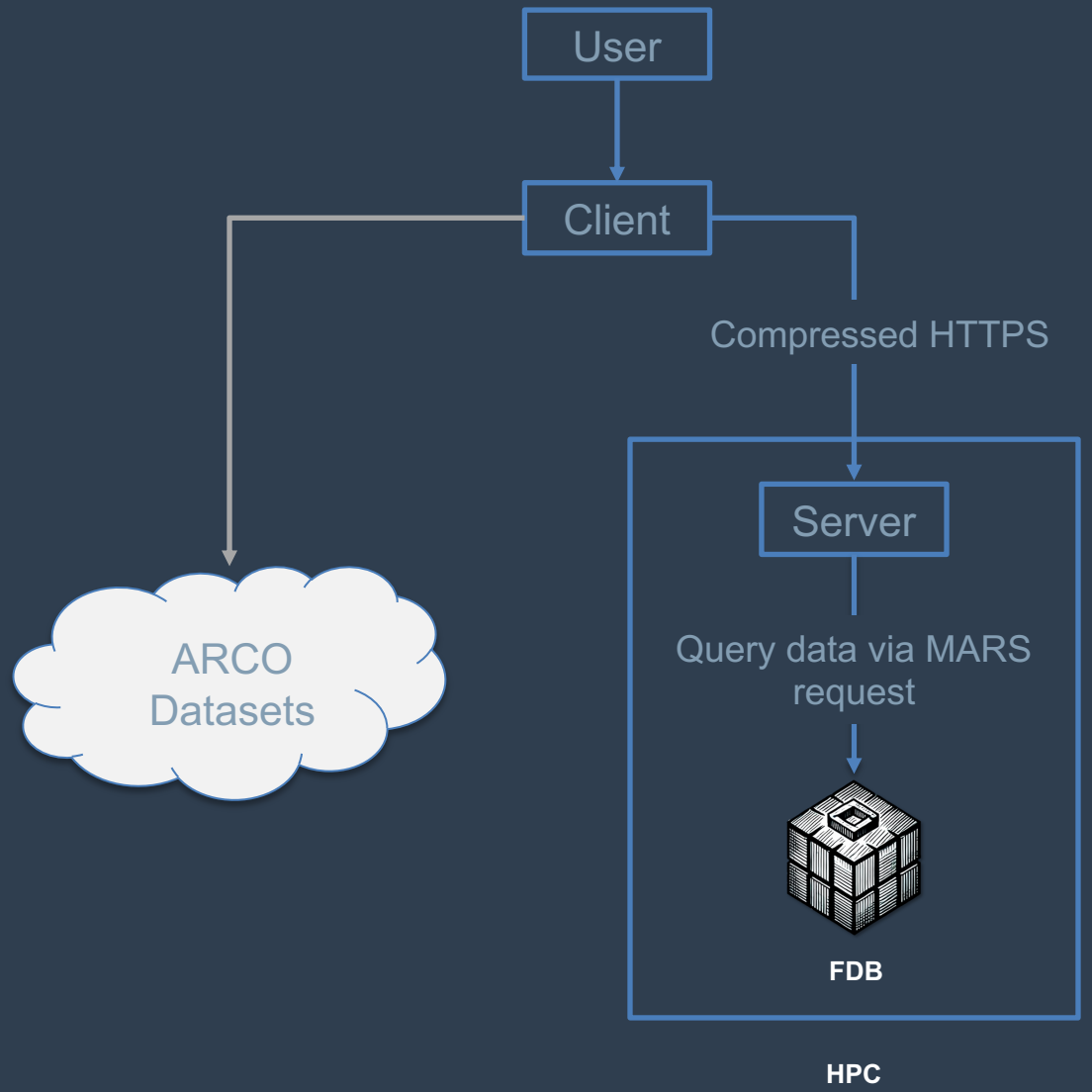
ECMWF EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS

Remote

Remote - Zarr via HTTPS

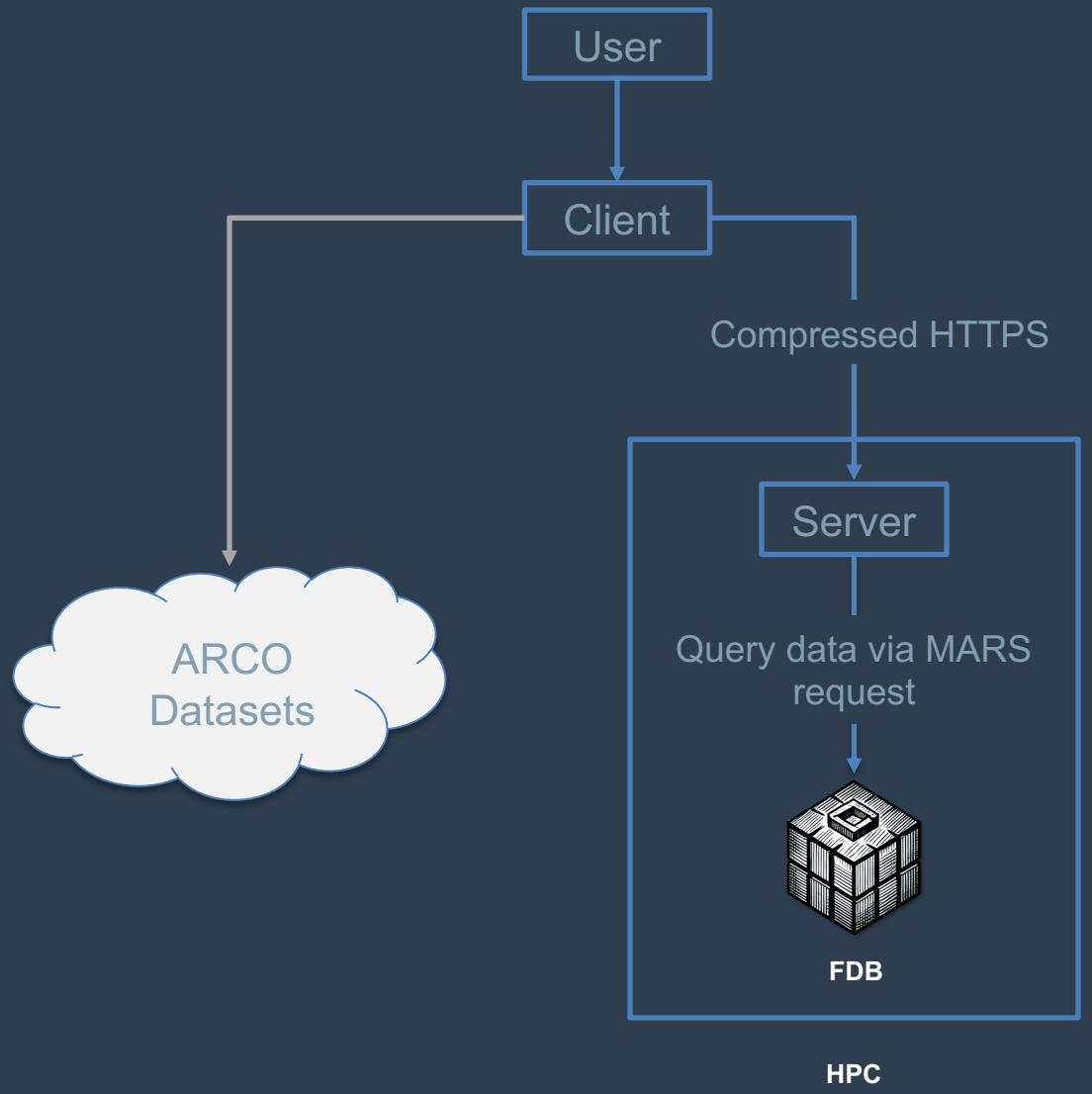


Remote - Zarr via HTTPS



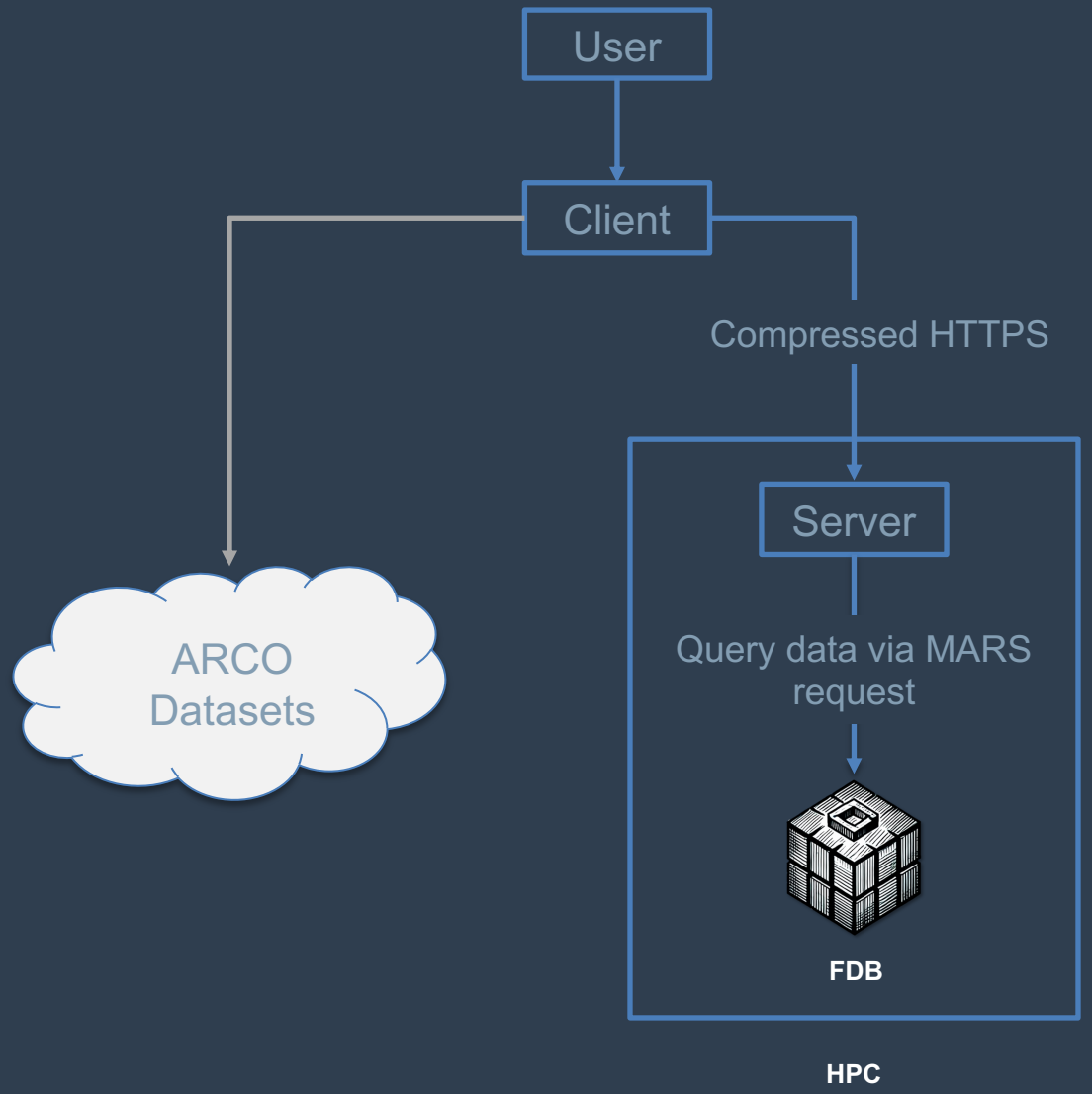
1. Client
   - Triggers creation of a virtual view

Remote - Zarr via HTTPS



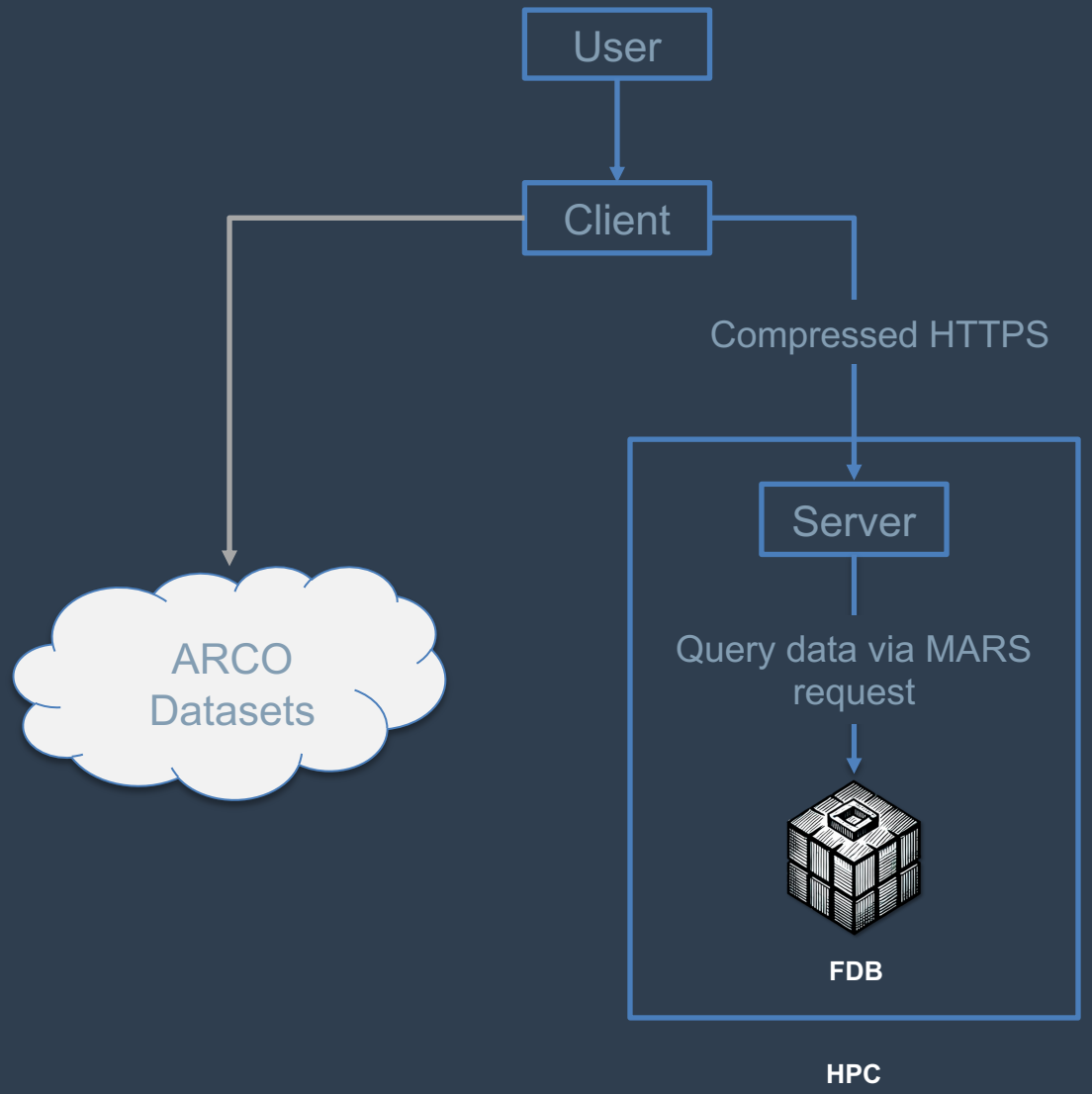
1. Client
   - Triggers creation of a virtual view
2. Server
   - Creates virtual view of FDB data

Remote - Zarr via HTTPS



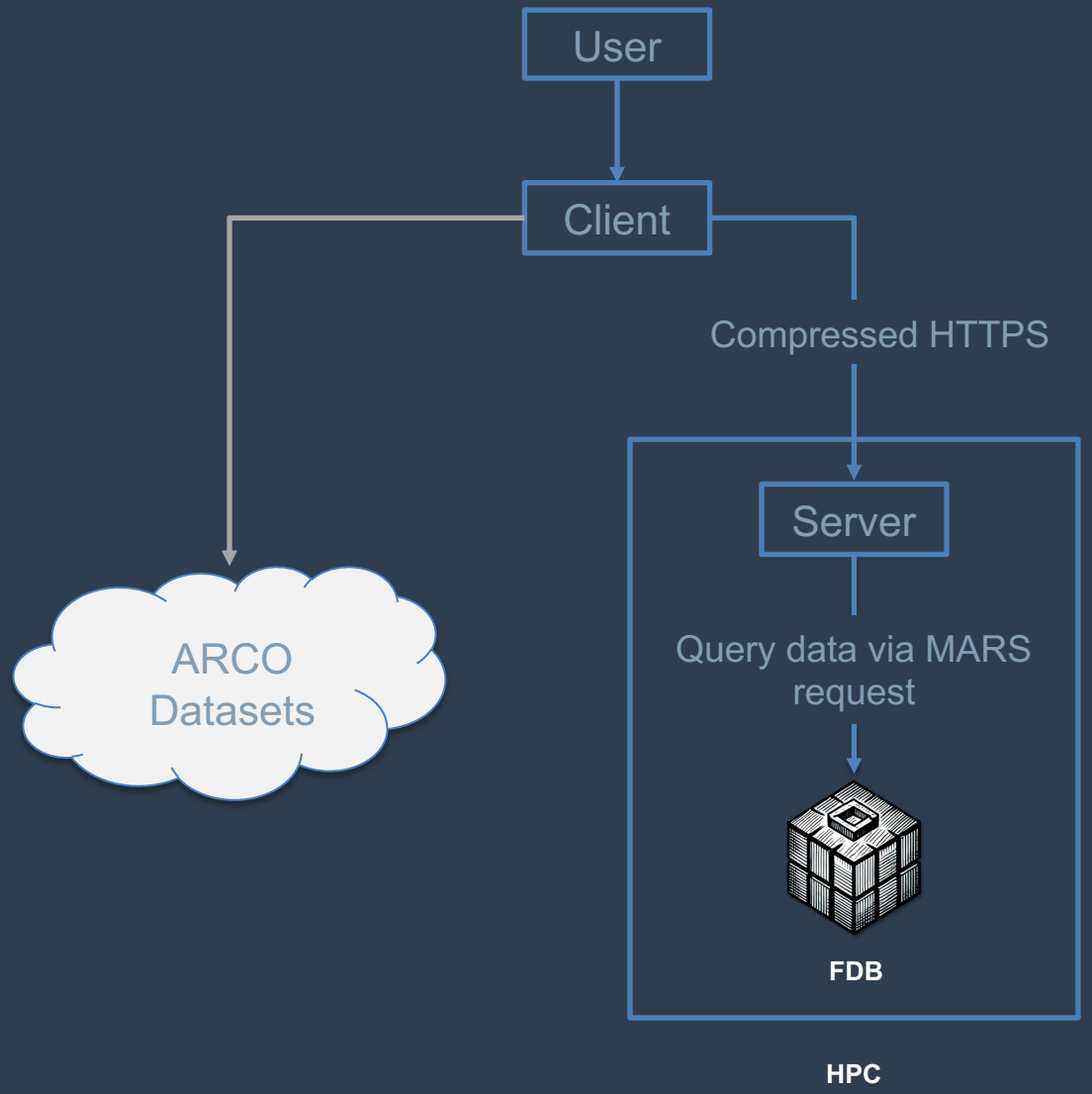
1. Client
   - Triggers creation of a virtual view
2. Server
   - Creates virtual view of FDB data
3. Client
   - Opens virtual view via FSSpecStore
   - Send chunk request to Server

Remote - Zarr via HTTPS



1. Client
   - Triggers creation of a virtual view
2. Server
   - Creates virtual view of FDB data
3. Client
   - Opens virtual view via FSSpecStore
   - Send chunk request to Server
4. Server
   - Queries Zarr chunk data from FDB on HPC
   - Chunk is send via compressed HTTPS to Client

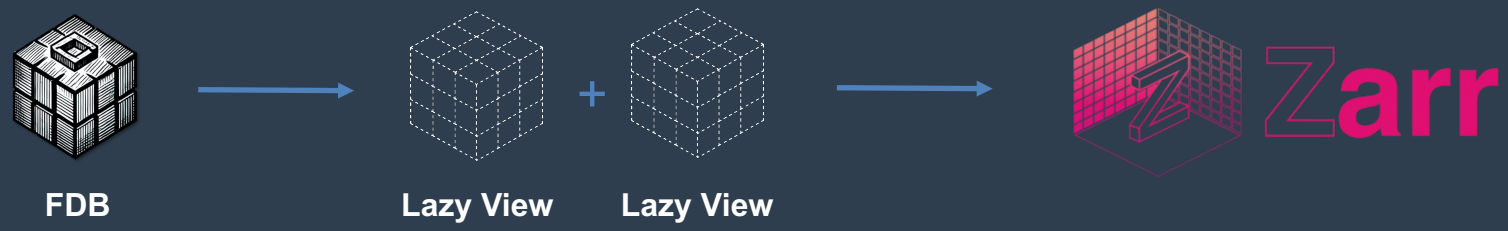
Remote - Zarr via HTTPS



1. Client
   - Triggers creation of a virtual view
2. Server
   - Creates virtual view of FDB data
3. Client
   - Opens virtual view via FSSpecStore
   - Send chunk request to Server
4. Server
   - Queries Zarr chunk data from FDB on HPC
   - Chunk is send via compressed HTTPS to Client
5. Client
   - Return the data as part of the Array access

Summary / Acknowledgments

Mapping between metadata driven GRIB and Zarr via lazy views to support shifting user requirements towards cloud use-cases



Special Thanks

- Kai Kratz, Simon Smart, Emanuele Danovaro @ECMWF
- Carsten Hinz @JSC
- Colleagues @WarmWorld Easier
- Data Mangagment and Service Team @ECMWF

The project on which this presentation is based was funded by the German Federal Ministry of Research, Technology and Space under the funding code 01LK2204E. The responsibility for the content of this publication lies with the author.

SPONSORED BY THE
Federal Ministry of Research, Technology and Space

Link Collection

WarmWorld	https://warmworld.de/
WarmWorld Easier	https://esm-data.fz-juelich.de/wweasier/docu/
FDB	https://github.com/ecmwf/fdb
FDB Documentation	https://fields-database.readthedocs.io/en/latest/
Zarr Python	https://github.com/zarr-developers/zarr-python
Zarr Documentation	https://zarr.readthedocs.io/en/stable/
ERA5 Explorer	https://era-explorer.climate.copernicus.eu