Developing a Next Generation Platform for Geodetic, Seismological and other Geophysical Data Sets and Services

Chad Trabant[1], Henry Berglund[2], Jerry Carter[1], Dave Mencin[2], and the UNAVCO and IRIS teams[1,2]

1. IRIS Data Services
2. UNAVCO Geodetic Data Services

EGU22 - 8905 ESSI2.3

Context and purpose

Operated by IRIS and UNAVCO respectively, the SAGE and GAGE facilities manage the US national Earth Science seismological and geodetic facilities on behalf of the US National Science Foundation.

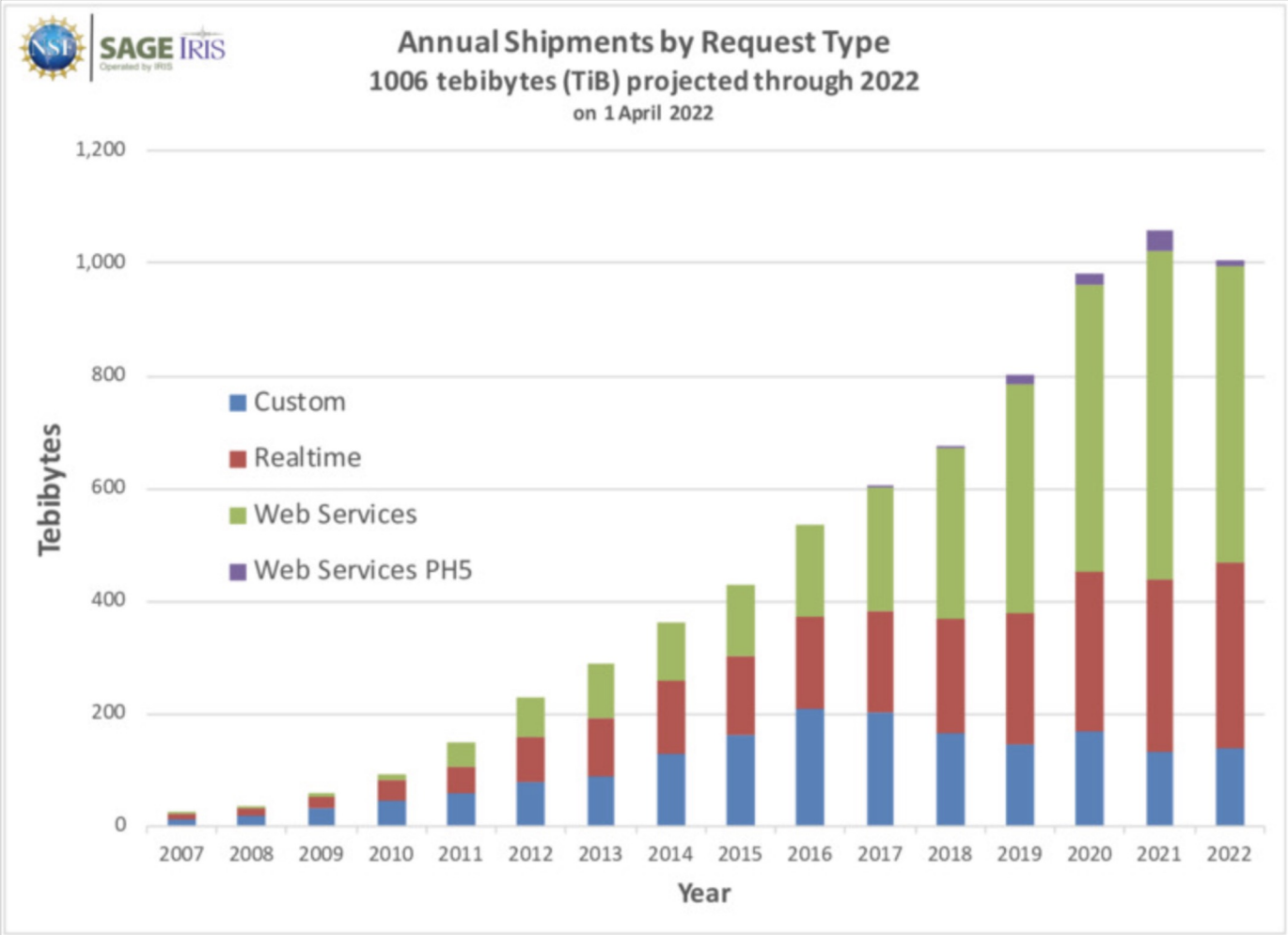
The data service components of each facility operate data centers supporting large national and international audiences of researchers. Both facilities operate physical data centers, with a small number of capabilities supported by cloud-like compute centers.

UNAVCO and IRIS are merging: UNAVCO + IRIS **= EarthScope**

Purpose: Design and build a unified platform for ingestion, archiving, curation, high-level product generation, quality measurements, and distribution in a cloud environment for the operation of a combined data services facility.

Current scale

- Combined repositories contain a few petabytes of data, growing with continuous collection and new sources
- 100s of thousands of individual data channels
- Millions of requests per day, distributing more than petabyte per year



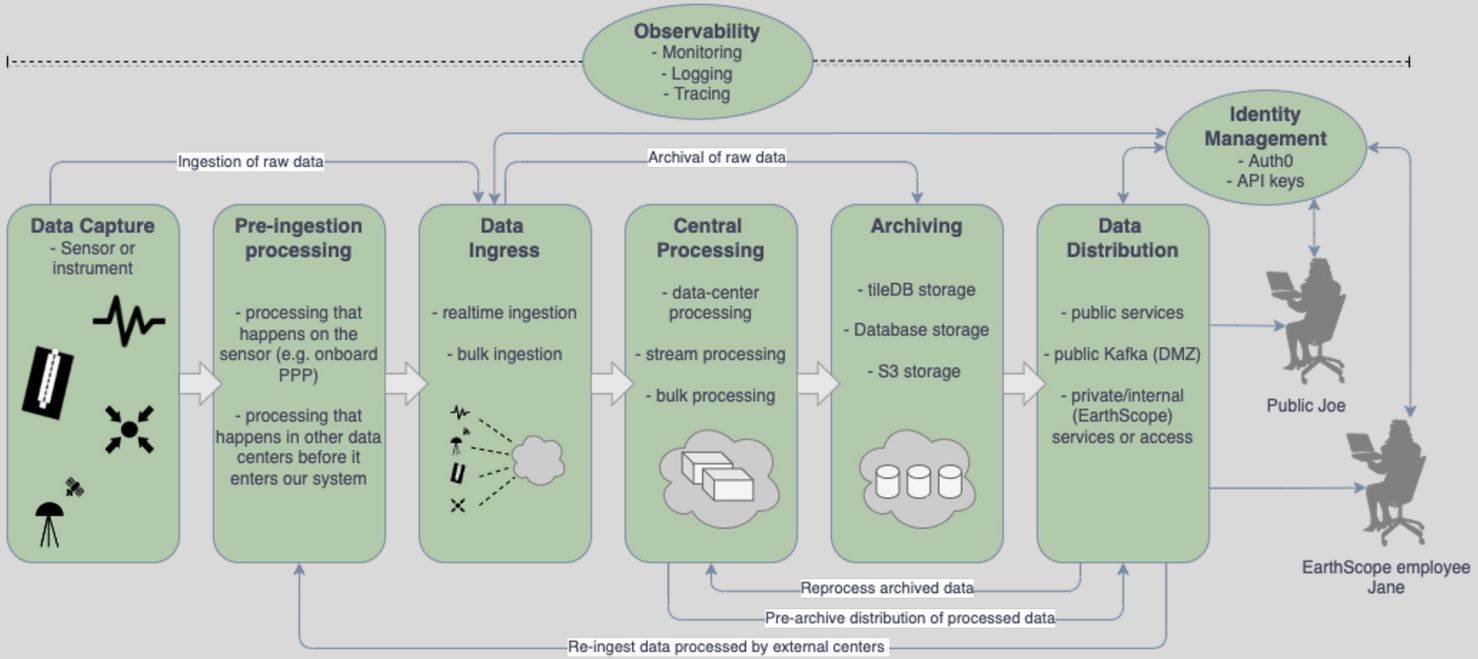
Annual data distribution volumes from IRIS Data Services, one of the two facilities

In the coming years, we project a significant increase in data input and output.

We also anticipate that the facility will handle an increasing variety of data types.

Guiding principles

- **Start from a “new system” perspective, informed by uses and needs of current facilities**
- Support primary access mechanisms currently offered at existing facilities
- Agile design and development methodology
- Common data containers, software and COTS solutions whenever possible
- **Include capabilities to support FAIR data principles at a fundamental level**
- Common identity management system and strategy



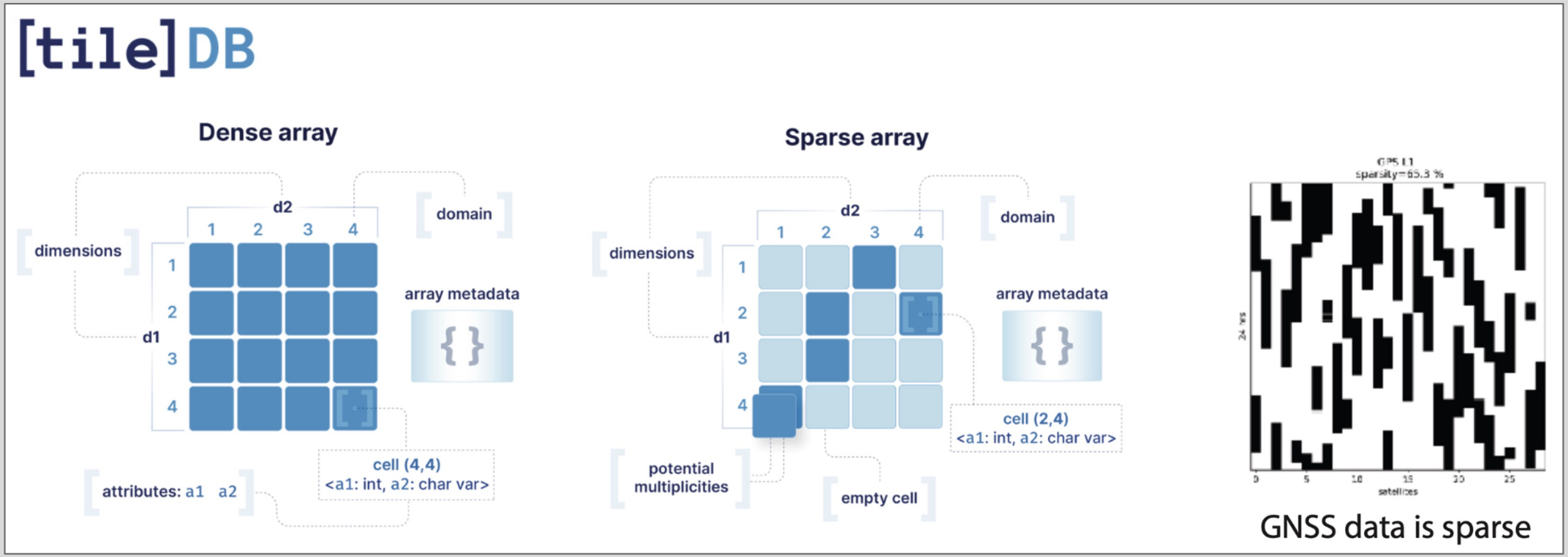
The conceptual model for the Common Cloud Platform

Current prototyping work

- **Core platform, the foundation of the system**
  - Infrastructure as Code, container orchestration, adaptation to cloud technologies, etc.
- **Metadata subsystem**
  - Generalized, initially focussed on StationXML and GeodesyML
  - FAIR data principles in design, JSON-LD
- **Generalized data container**
- **Unified identity management system**

General data containers

- Most data are 1 or more dimensions of arrays
- TileDB fits many of our needs: open specification/source, versioning, sparse arrays, cloud optimized



Data are Analysis Ready and Cloud Optimized (ARCO)

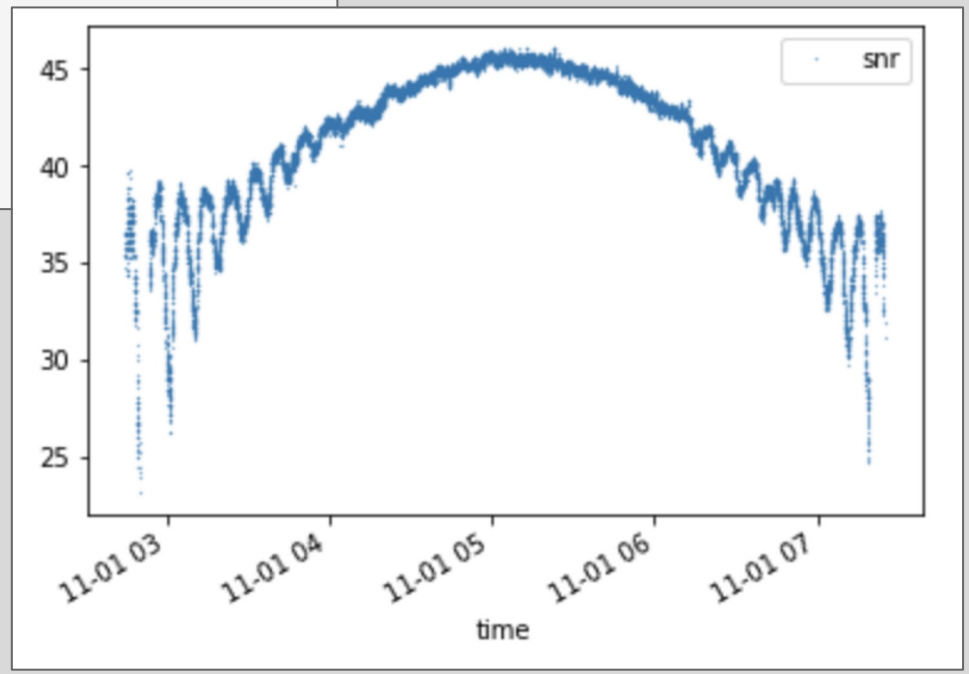
```
epoch = datetime.datetime.utcfromtimestamp(0)

def unix_time_second_micro(dt):
    return (dt - epoch).total_seconds() * 1000000

s = np.int64(unix_time_second_micro(datetime.datetime.fromisoformat('2021-11-01T02:00:00')))
e = np.int64(unix_time_second_micro(datetime.datetime.fromisoformat('2021-11-01T08:00:00')))


# Open the array in read mode and read the whole array
with tiledb.open("s3://unavco/tiledb/P041_1Hz.tdb/", mode='r', config=tiledb.Config.load("./config.txt")) as A:
    %time df = A.df[s:e, 0, 1, '1C']
    %time df = df.sort_values(by=['time','sys','sat','obs'])
    %time df['time'] = pd.to_datetime(df['time'], unit='us')
    %time df.plot(x='time', y='snr', marker='.', markersize=0.5, linestyle='')

%time df.obs.unique()
```



A Python example showing a four dimensional slice to query and plot GPS PRN 01 S1C signal-to-noise.

Expected implications

A large project, re-imagining the data management platform for SAGE and GAGE allowing for:

- **Integrated, multi-domain data access from a single facility**
- **Seamless transition for many users**
- **Services and capacity not possible with the current systems**
- **Potential for supporting research data processing in the same cloud environment as the data, avoiding over-the-internet transfer**
- Design with cost effective operation and growth
- Opportunity to develop a culture between facility staff
- Opportunity to address legacy technical debt

Timeline:

- **Initial system: early 2023**
- **User testing: mid-2023 - 2024**
- **Fully operational: 2024**

Thanks!

GAGE Geodetic Facility for the Advancement of Geoscience

UNAVCO

SAGE Seismological Facility for the Advancement of Geoscience

IRIS