Big Data, organisation and analysis

A glimpse of the infrastructures

Steffen M. Noe, Spring 2023

Big Data infrastructure

- The infrastructure consist of hardware
  - storage systems (NAS, tape robots)
  - data transmission systems (fibre optical connections)
  - computing facilities (cloud, server farm,…)
- and of software
  - Hadoop
  - Databases (No-SQL / SQL)
  - Data Warehouse

Let’s start with an example

New York Stock Exchange

- produces 4-5 TB/day
- uses data warehouse technology
- in NYSE case it is IBM Netezza, capable of 2 TB/h load access via standard interfaces SQL/ODBC/JDBC/OLE DB and 4TB/h restore and backup
- Cloud based systems

Let’s have another example

Finnish Met Institute, satellite dat

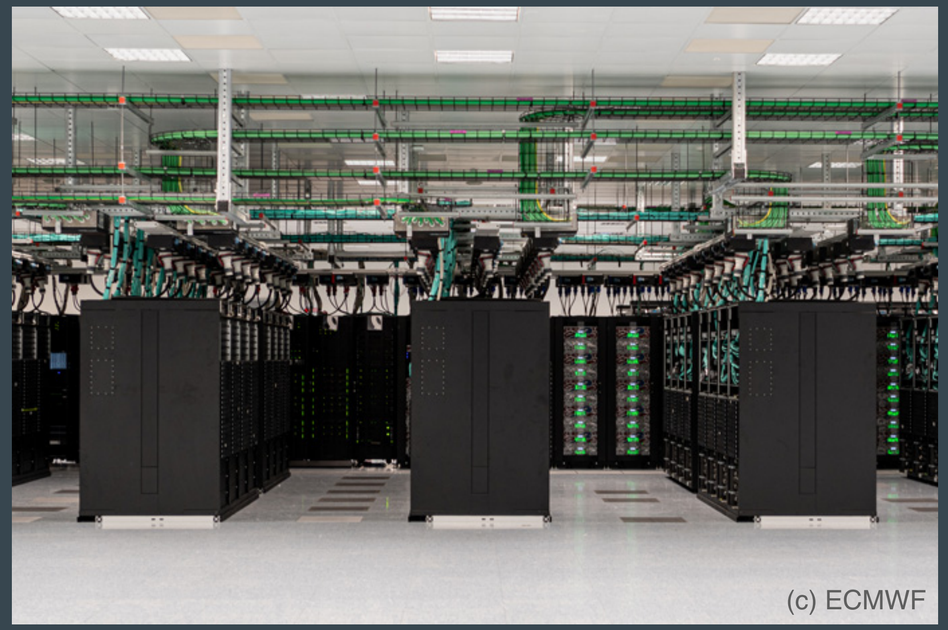
FMI SODANKYLÄ NATIONAL SATELLITE DATA CENTRE

- produces up to 23 TB/day
- has its own high performance data cluster
- transfers data to customers (ESA, Weather Services, ECWMF) via high speed optical fibre
- data transferred to data storages and archives like WEKEO (Sentinel data access)

One more, weather services

ECMWF (European Centre for Medium-Range Weather Forecasts)

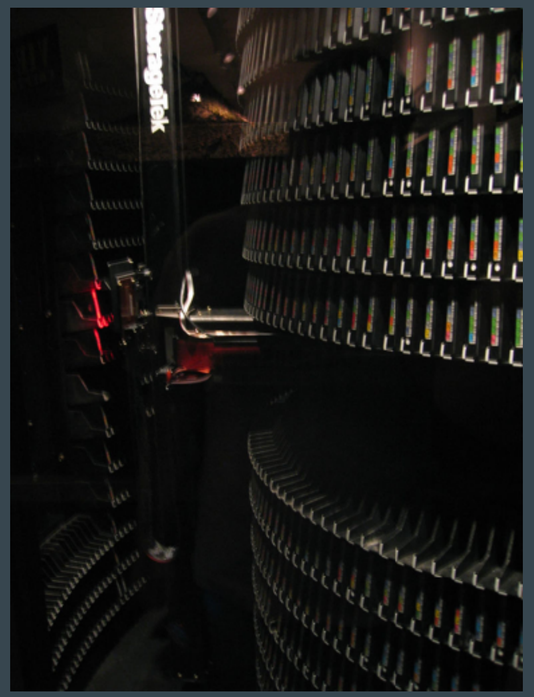
- Multiple data sources
  - Satellite
  - Weather stations
- Large scale data centre and supercomputing facilities
  - one million processors (CPU/GPU/tensor units) over 4 clusters
  - 30 petaflops (10^{15} or a million billion calculations per second)



Some storage types for Big Data

Tape libraries

hardware and access technology
long term storage
slower access

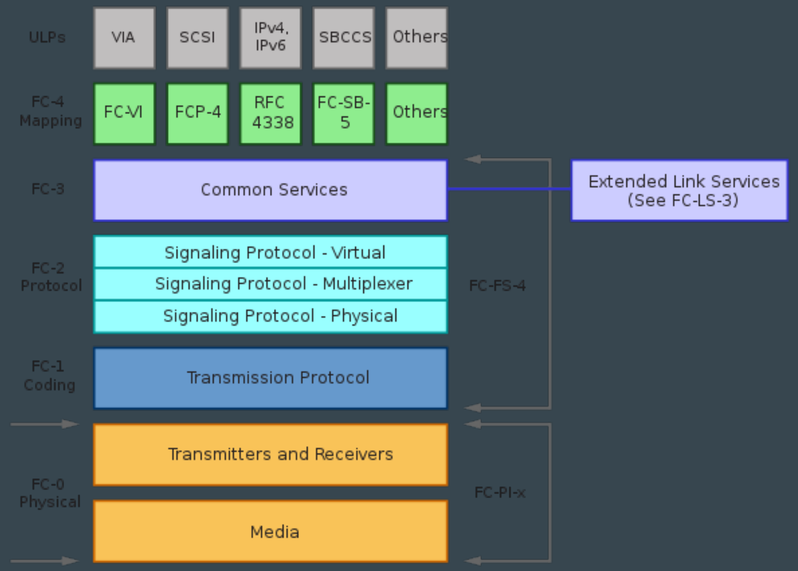


Network attached Storage (NAS)

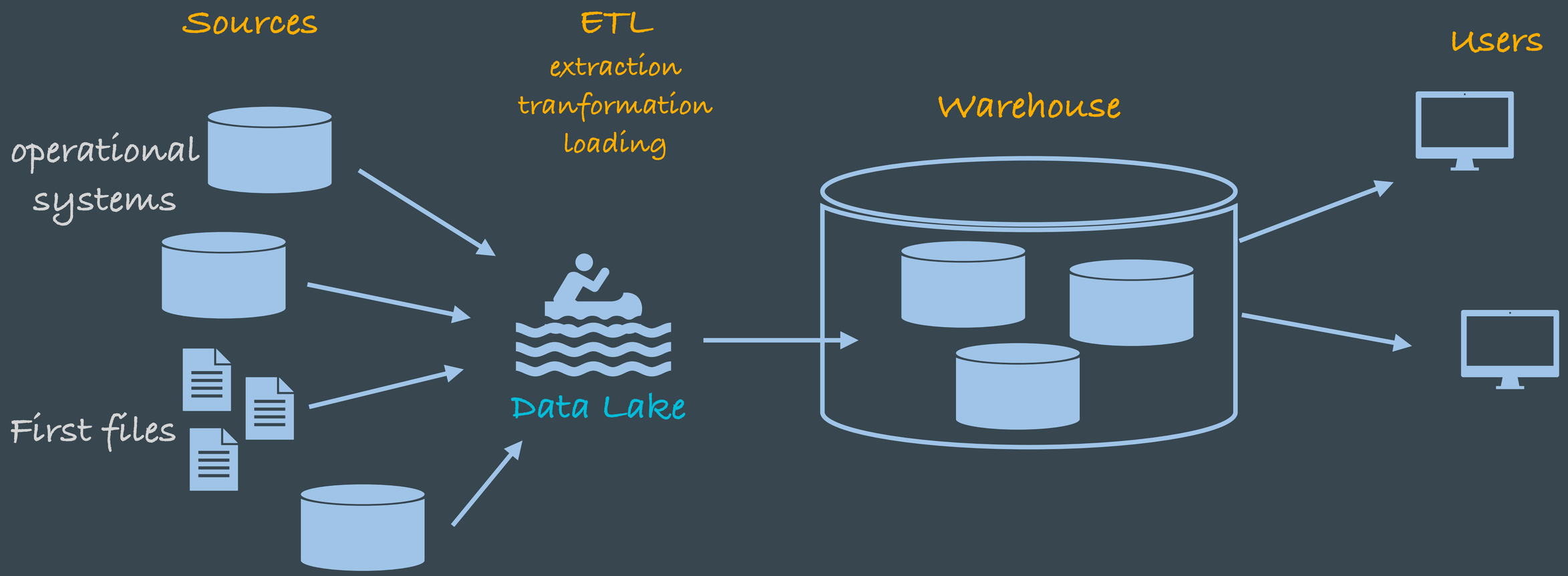
provides hardware and software
optimised to store data on local media
fast access

Storage area network (SAN)

e.g., Fibre channel
its a protocol
distributes access to storage media
where to store remains at the user

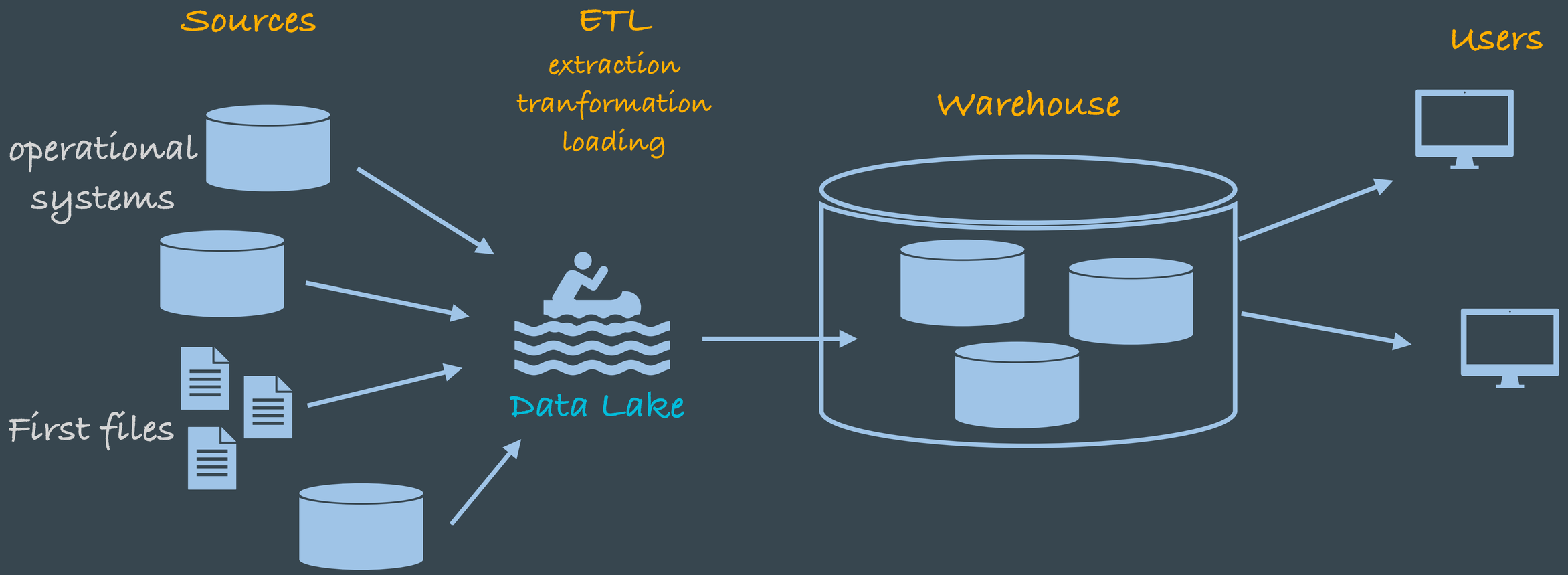


What are data warehouses?



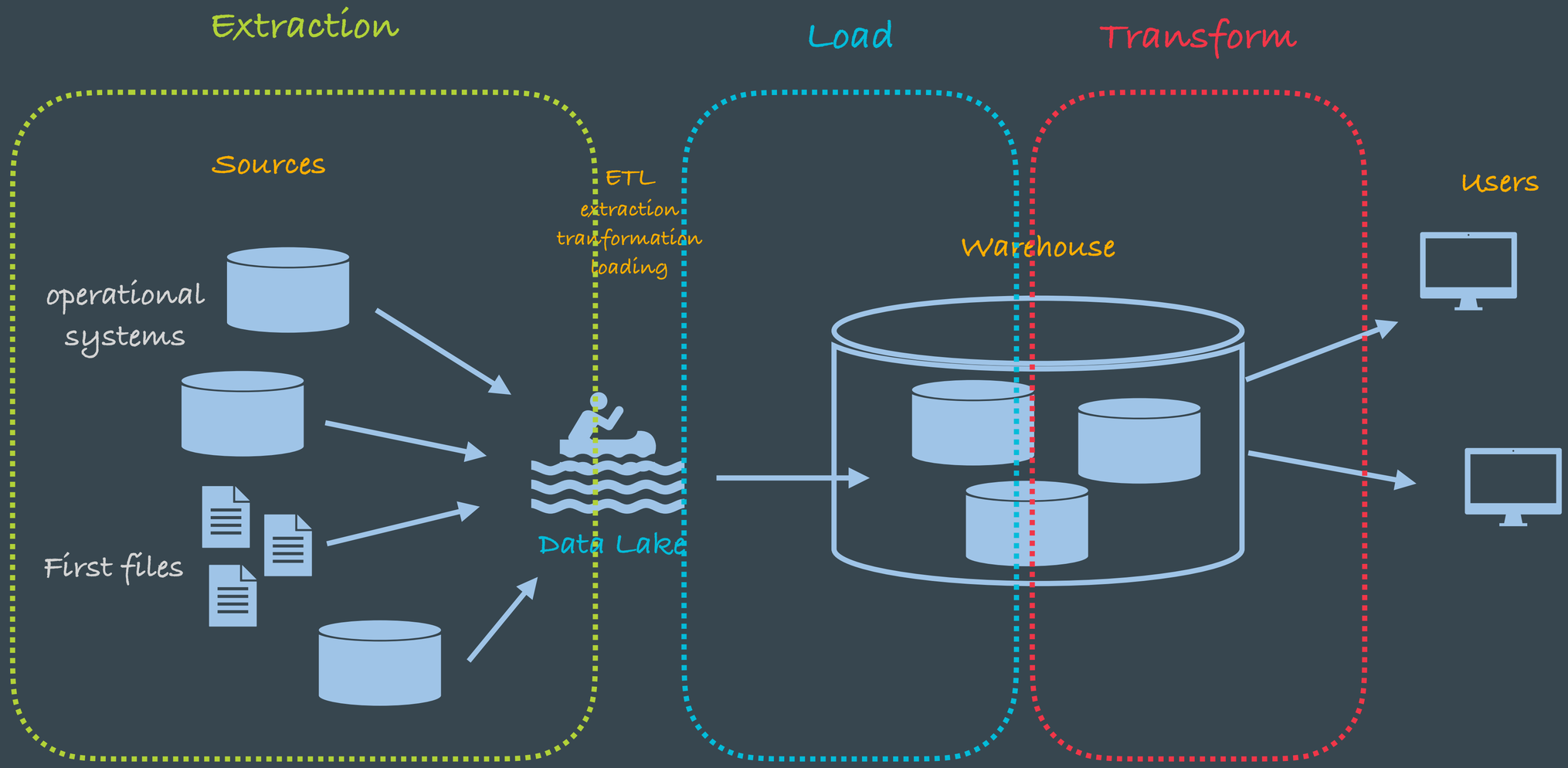
What are data warehouses?

A data warehouse combines many data sources and pool (lake) them into accessible units



What are data warehouses?

A data warehouse combines many data sources and pool (lake) them into accessible units



Key components of a Big Data infrastructure

Data production

Hardware:
Sensor networks, IoT, mobile devices, cameras, scientific sensors, local computers, servers, single-board computer (SBC)

Software:
often self made, provided by sensor manufacturer, OS (filesystem), database

Data extraction

Hardware:
HPC, Server, virtualisation (Docker containers, virtual machines), PC, Laptop

Software:
Python, R, scientific computation, ML, Hadoop, Spark, data joining and filtering, relies on self made scripts

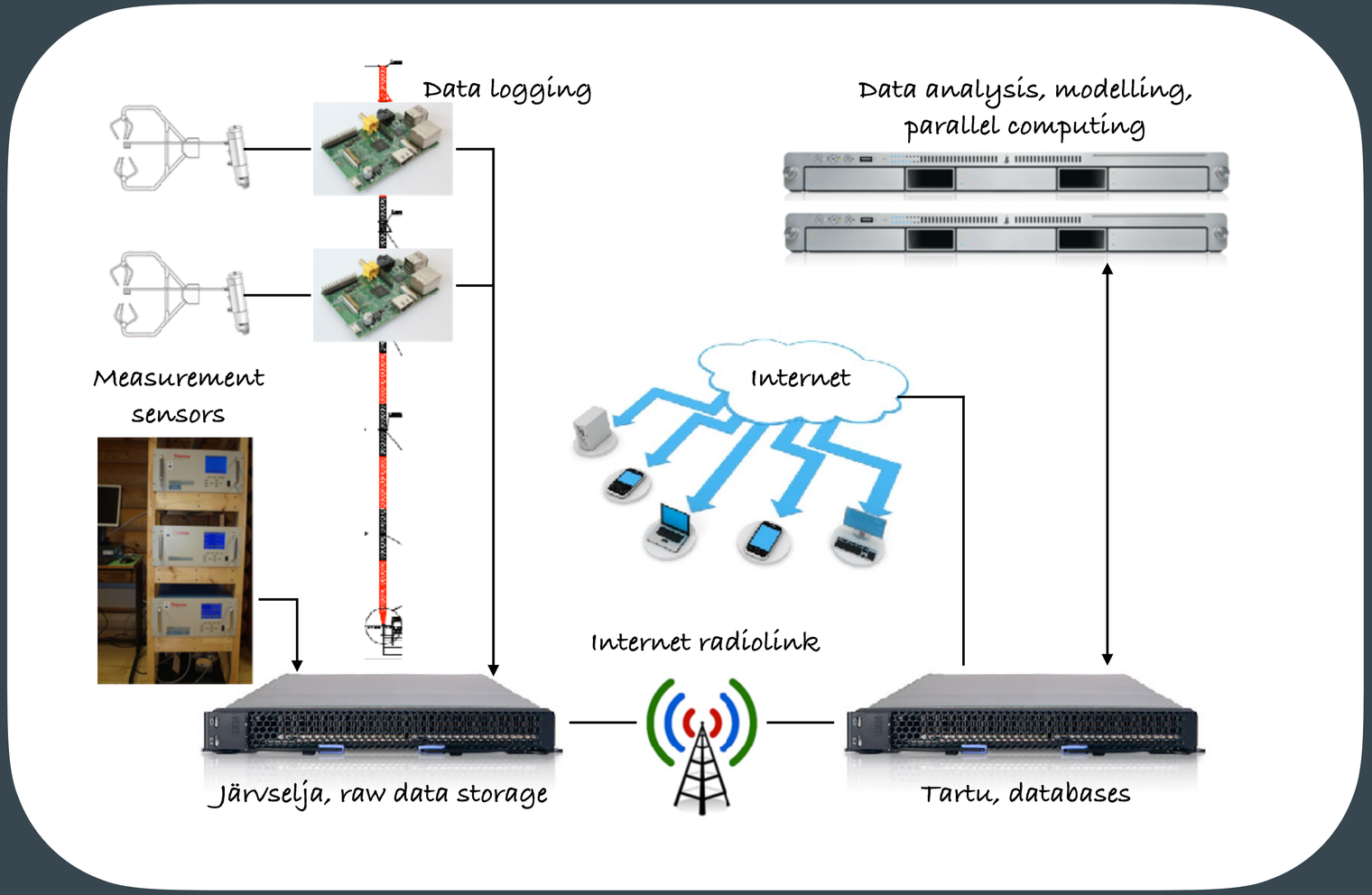
Data provision

Hardware:
Web servers, cloud services, open repositories

Software:
data warehouse, Hadoop, Spark, data pipelines, web interfaces, mostly available as services

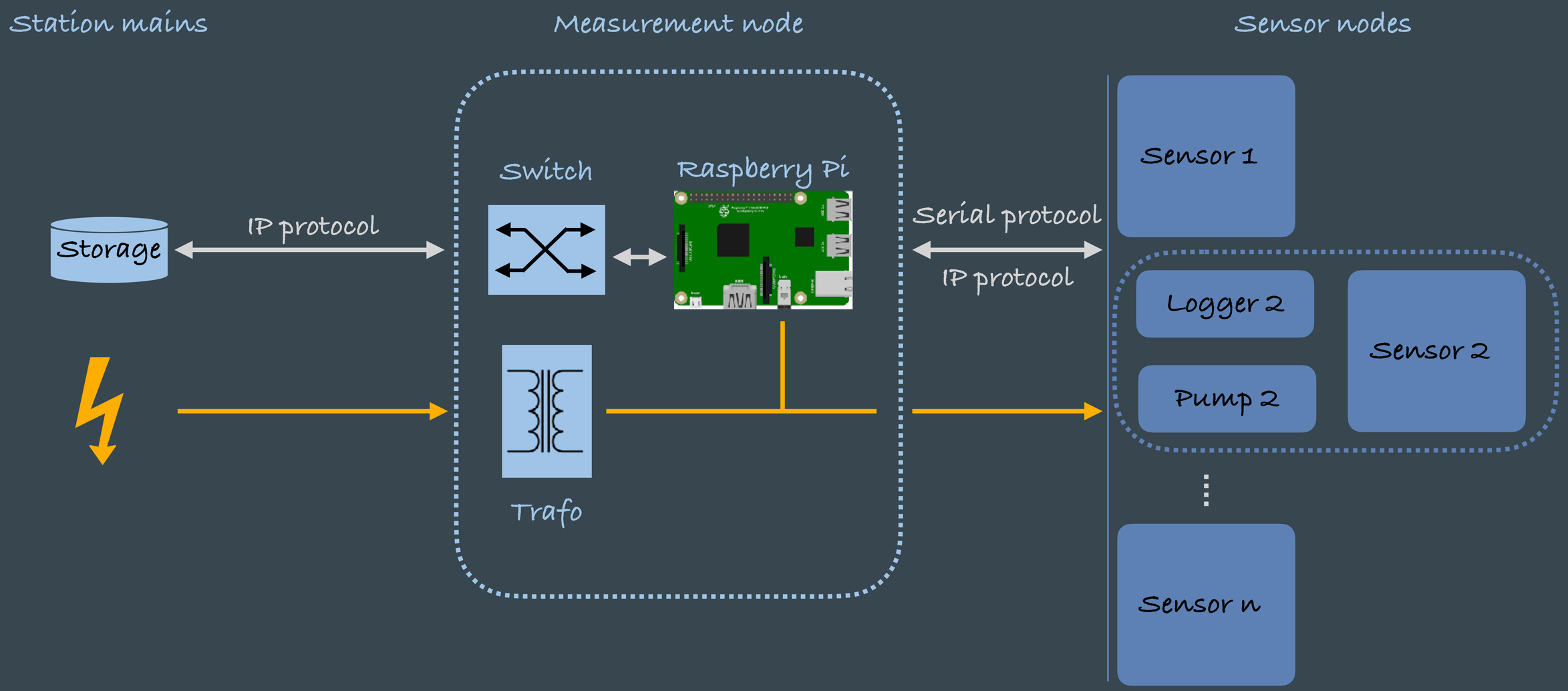
Data extraction and processing at SMEAR Estonia

- Currently we use:
- > 15 Raspberry Pi systems as data entry points
- 2 IBM servers for automated data processing and storage of raw data
- 1 HP workstation in Järvselja (data preprocessing, modelling, virtualisation service, ML training on CPU and GPU)
- 1 Apple Pro workstation in EMU (data mining, virtualisation service, R-Studio server, Jupyter server), Scientist's own PC or Laptop
- 2 NAS servers on site, 1 NAS server in EMU (~ 48 TB)
- Cloud storage in EMU and Tartu University



Automatisation to enable Big Data

SMEAR Estonia as example



Challenge

- Try to find out what are typical data volumes of Earth Observations
- Find some other interesting data volumes? (cars, production, …)