Kevin Pietersen and Antoine Bagula

Contributors

Zaheed Gaffoor and Nebo Jovanovic

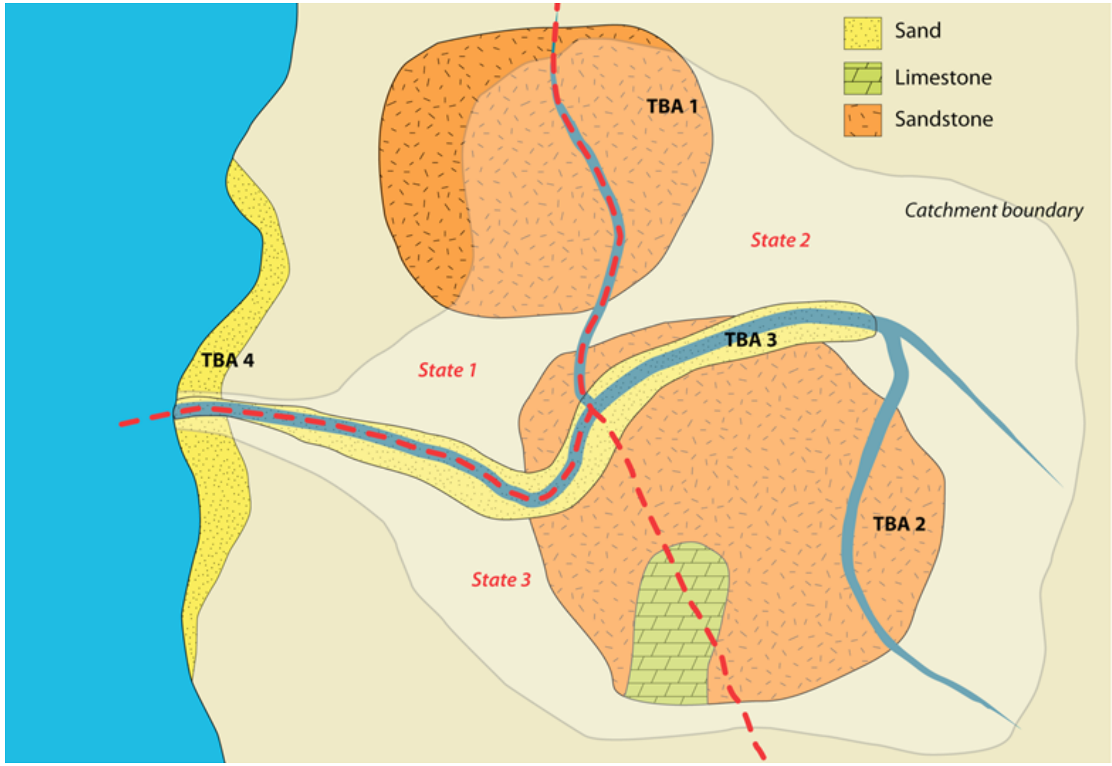
Big Data Infrastructure for Transboundary Aquifer Systems Analytics

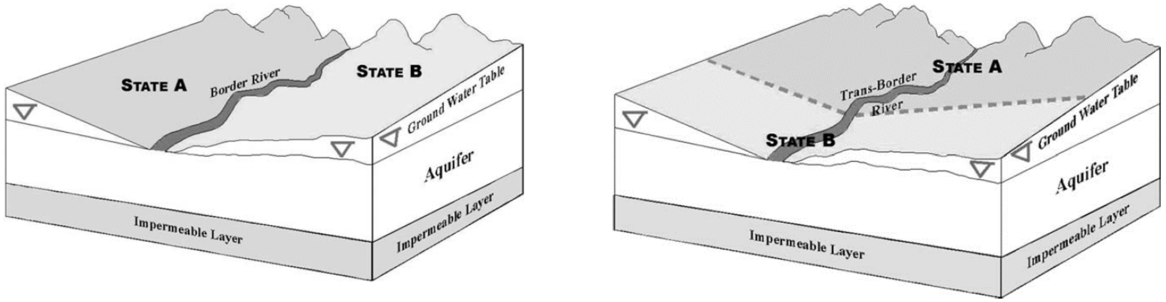
CSIR
our future through science

RESPICE PROSPICE
UNIVERSITY of the WESTERN CAPE

Transboundary Aquifers (TBAs)

(IGRAC, 2016)

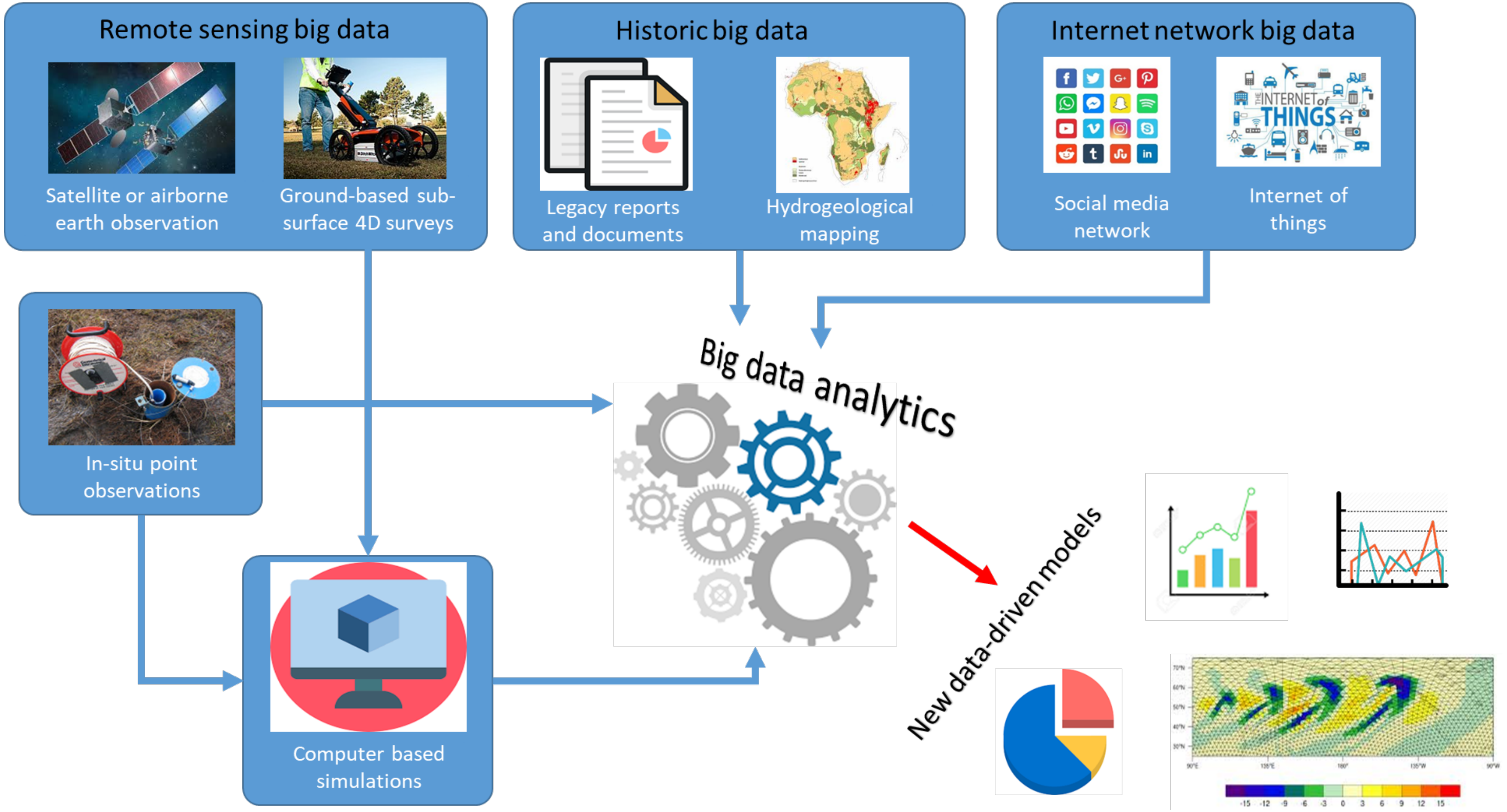




Transboundary Aquifers: Conceptual Models for Development of International Law

by Yoram Eckstein[1] and Gabriel E. Eckstein[2]

Big data in groundwater



Traditional sources of groundwater data

Source	Definition	Constraints
In-situ monitoring programs	• Manual or sensor based observations • Structured data • Stored in spreadsheets • Online repositories (eg. RIMS, NGA)	• Poor temporal and spatial coverage • Cost of installation of piezometers and boreholes • Many offline databases
Historic reports and maps	• Information and data present in reports • Unstructured • Textual • Hardcopy or digital archives	• Data in non-readable machine format
Geophysical surveys	• Geophysical natural or artificial field observations (eg. Electric-magnetic, gravitational) • 1D, 2D, 3D arrays • Structured	• Limited coverage in SADC • Mostly performed during groundwater exploration (once-off)

Remote sensing data

- Numerous earth observation missions
- Some dedicated to hydrological related sciences
- Near real-time
- Global coverage
- Data generated daily from one mission can be in 458 GB
- NASA generates 1,73 GB data every second from remote sensing

Remote sensing mission	Hydrological component	Spatial resolution	Temporal resolution	Launch year
Global precipitation measurement (GPM)	Precipitation	5km	3 hours	2014
Tropical Rainfall Measuring Mission (TRMM)	Precipitation			1997
Terra/MODIS	Evapotranspiration	250m	1 day	1999
Aqua/MODIS	Evapotranspiration	250m	2 day	2002
Soil moisture and ocean salinity (SMOS)	Soil moisture	36km	3 days	2009
Soil moisture active and passive (SMAP)	Soil moisture	36km	3 days	2015
Gravity recovery and climate experiment (GRACE)	Terrestrial water storage	110km-220km	30 days	2002
GRACE-FO	Terrestrial water storage	110km-220km	30 days	2017
Landsat mission	Evapotranspiration/Vegetation/Land Cover	various	various	1972
Sentinel mission	Soil moisture/ Vegetation/Land Cover/Temp	various	various	2014

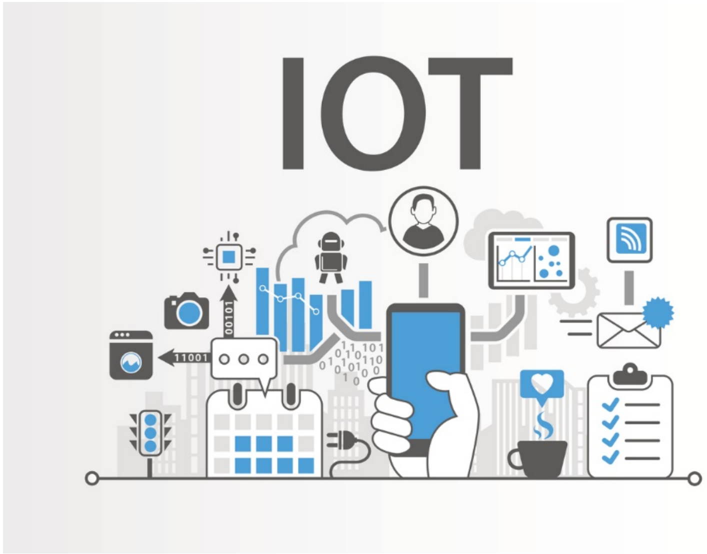
Simulated groundwater data

Synthesised datasets based on a combination of in-situ observations, satellite imagery, and model output

Atmospheric models	Land-surface models	Reanalysis
• Complex numerical models used to simulate weather and climate patterns • Supercomputers necessary • Lots of data processed • Lots of data generated • Structured data • Eg. GCM	• Complex numerical models of land-and shallow subsurface fluxes (energy, biological, water) • Data assimilation techniques used • Processing of hydrological data • Structured data • Eg. LDAS	• Historical datasets reanalysed by combining satellite data and model outputs to improve data coverage and accuracy • By-products of atmospheric and land-surface models • Structured data • Eg. ERA5
• Many datasets are readily available (free or paid)		

Internet network groundwater data

- Hydrologically relevant information present on social media post, blogs, vlogs, webpages, emails, podcasts etc.
- Mostly textual
- Plenty of videos, images and audio
- Highly unstructured (sometimes semi-structured)

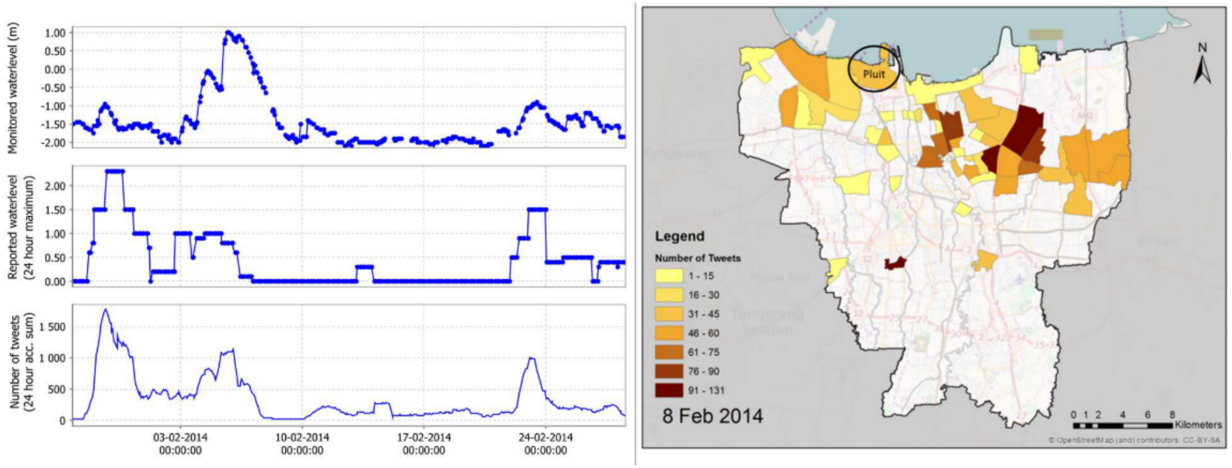


- Data collected and transmitted by connected devices
- Environmental data streaming
- Virtual citizen sciences
- Mostly structured

IoT and social media Big Data applications in hydrology

Lampos and Cristianini (2012) mining and predicting rainfall rates from twitter phrases

Eilander et al. (2016) mining and predicting flood levels from twitter posts



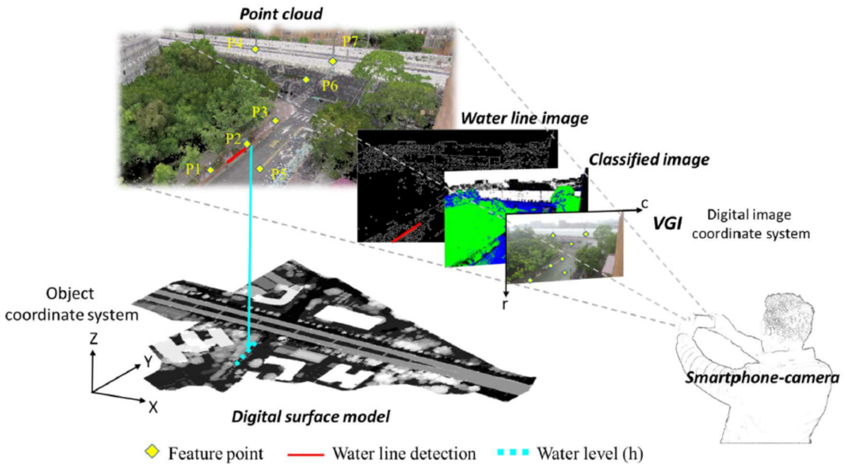
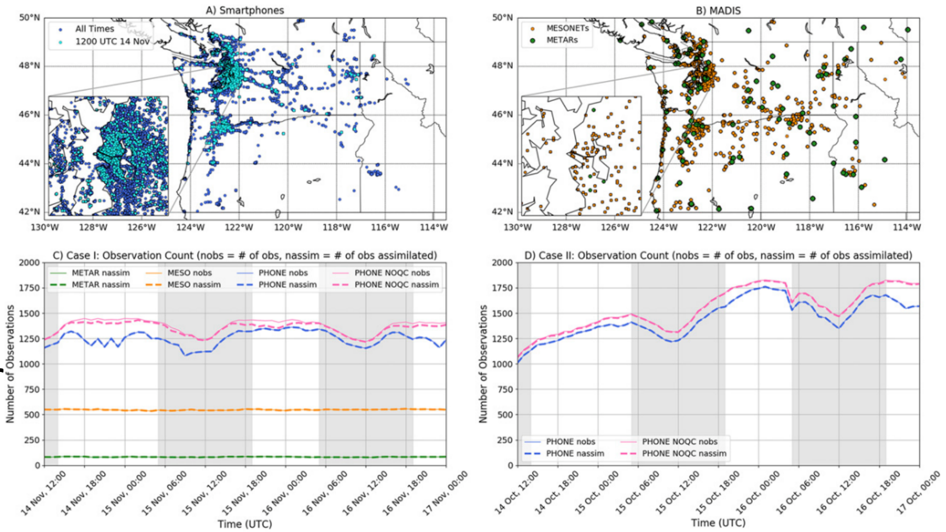


Figure 1. The geometry of VGI water level calculation.

Lin et al. (2020) calculating flood level in urban areas using image analysis

McNicholas and Mass (2018) improving weather modelling through data from smartphones



Challenges

Distributed data storage infrastructure (even within organization)

Eg. NASA DAAC

Difference protocols and user interface to extract data

Large datasets that are difficult to move (petabytes)

Data products are numerous, and technically challenging to navigate

Detailed inventory of all the relevant data products, including meta-data

Computing resources needed to perform functions generally include parallel processing

System requirements

Connect data sources and data products in one central locations (data ingestion)	Not necessarily moving data, but a central location to explore data Requests made to data source as needed
Data curation mechanism	Integrate local and regional datasets Uniform spatial and temporal reference system Quality control features
Data extraction mechanism	Sub-setting Temporal lookups and spatial lookups
Data visualization tools	Graphs, maps, GIS etc
Built-in analytics	Transform data into information In order to inform decision support systems, or early warning systems etc.

Big Data Architectures

01

A big data architecture is designed to handle the ingestion, processing, and analysis of data that is too large or complex for traditional database systems

02

The threshold at which organizations enter the big data realm differs, depending on the capabilities of the users and their tools. For some, it can mean hundreds of gigabytes of data, while for others it means hundreds of terabytes

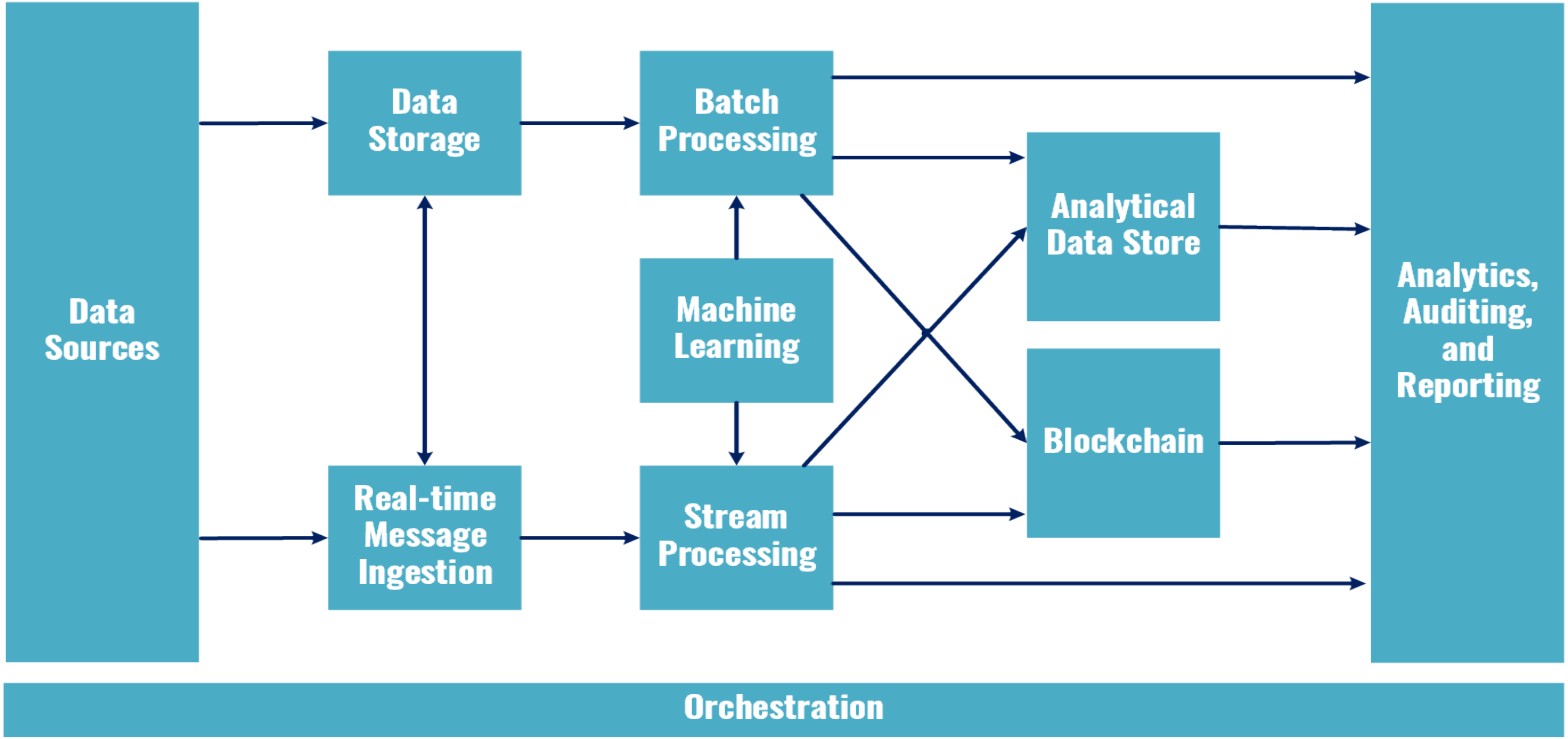
03

As tools for working with big data sets advance, so does the meaning of big data. More and more, this term relates to the value you can extract from your data sets through advanced analytics, rather than strictly the size of the data, although in these cases they tend to be quite large

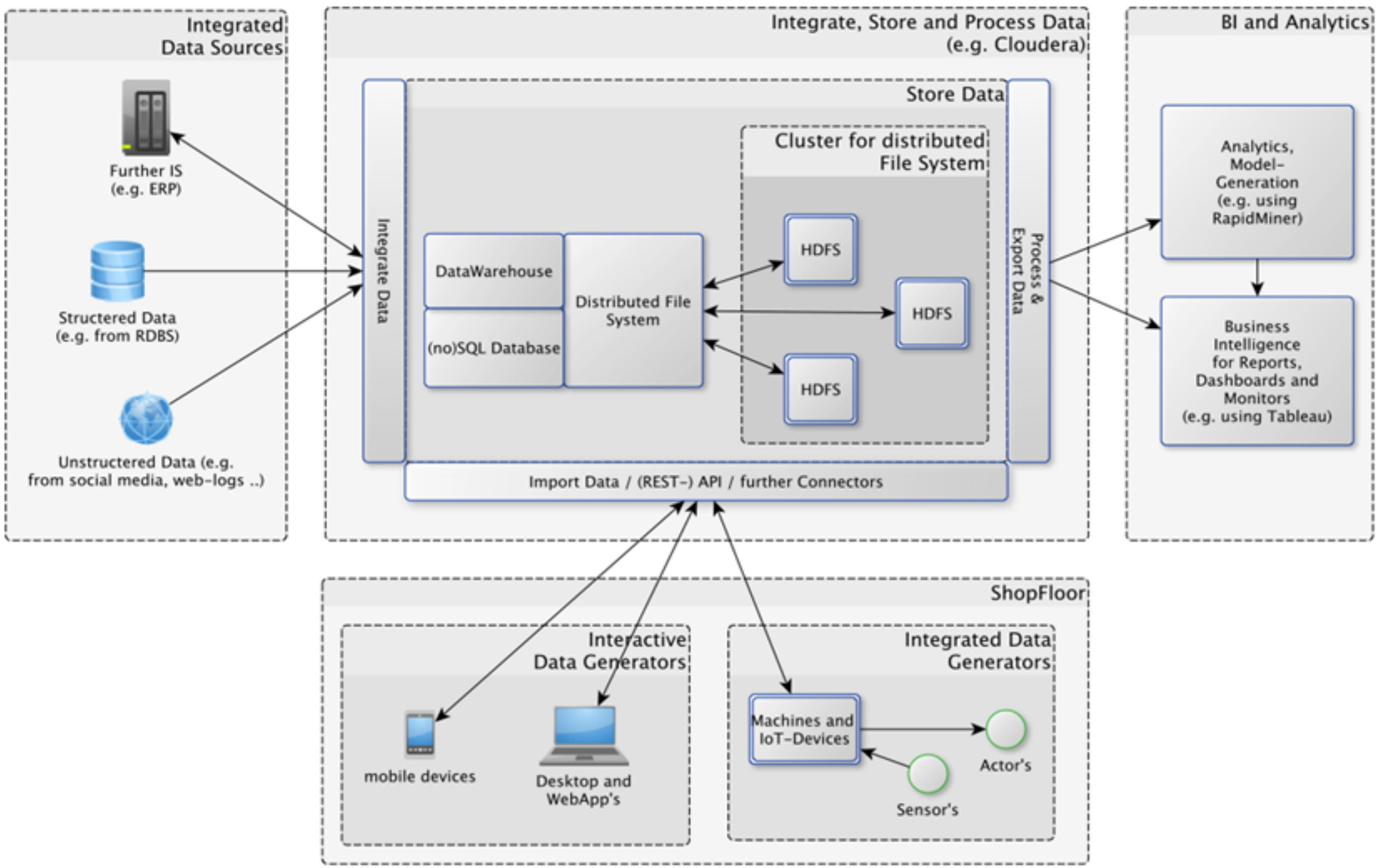
Big Data Architectures

Big data solutions typically involve one or more of the following types of workload:	• Batch processing of big data sources at rest • Real-time processing of big data in motion • Interactive exploration of big data • Predictive analytics and machine learning
Consider big data architectures when you need to:	• Store and process data in volumes too large for a traditional database • Transform unstructured data for analysis and reporting • Capture, process, and analyse unbounded streams of data in real time, or with low latency
Some of the most known architectures include but are not limited to:	• Lambda architecture • Kappa architecture • Internet-of-Things logical architecture

Main components



Our experimental architecture



Big Data processing

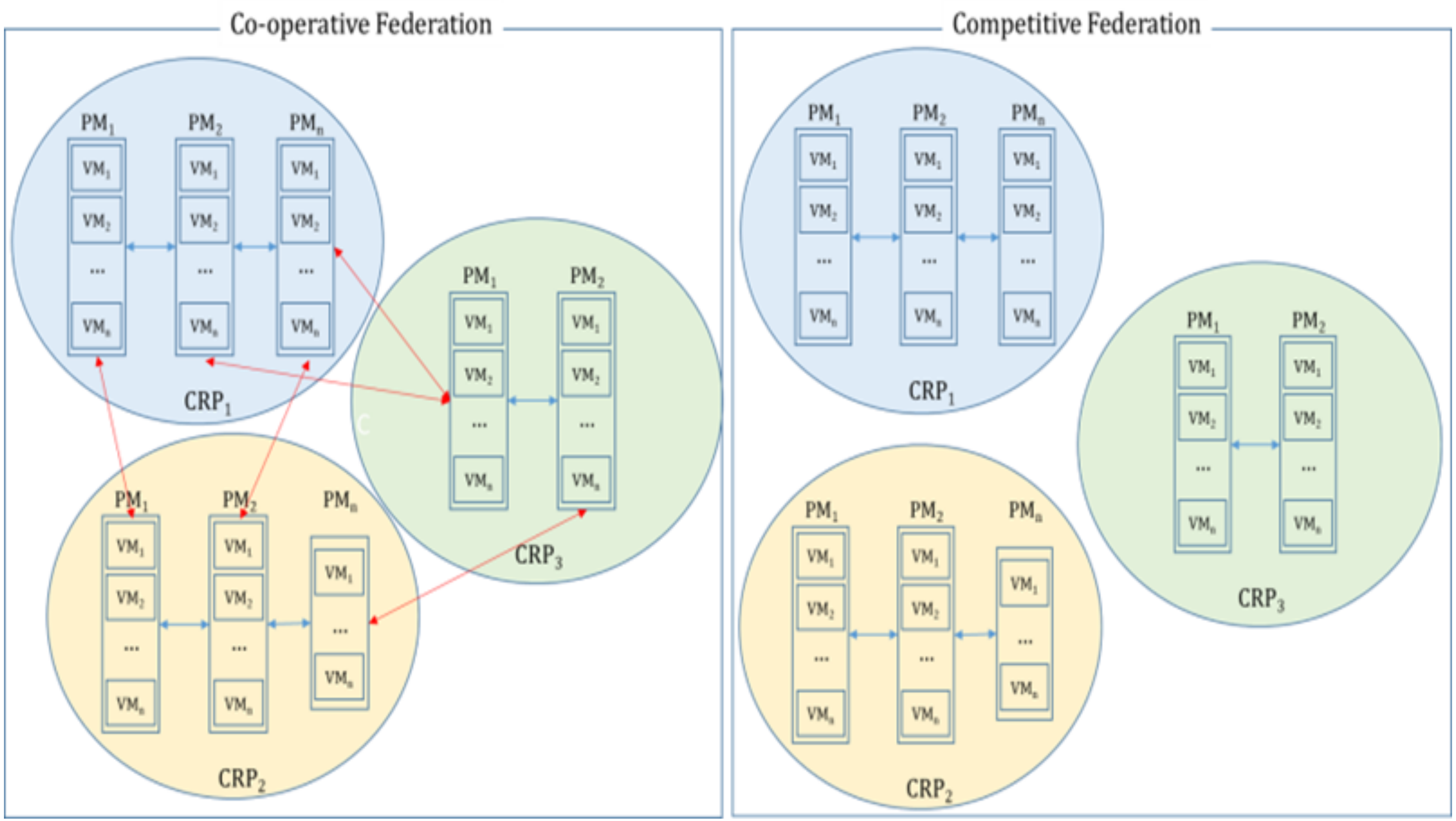
- **Most African countries are either under-developed or developing, hence cannot afford their own dedicated HPCs**
- **A potential solution is Cloud Federation**
- **Cloud Federation is a collaborative model between Cloud Service Providers (across countries)**
- **Federated Clouds allows for remote execution of tasks on computing resources flexibly and cost efficiently**

The need for a collaborative model using cloud federation

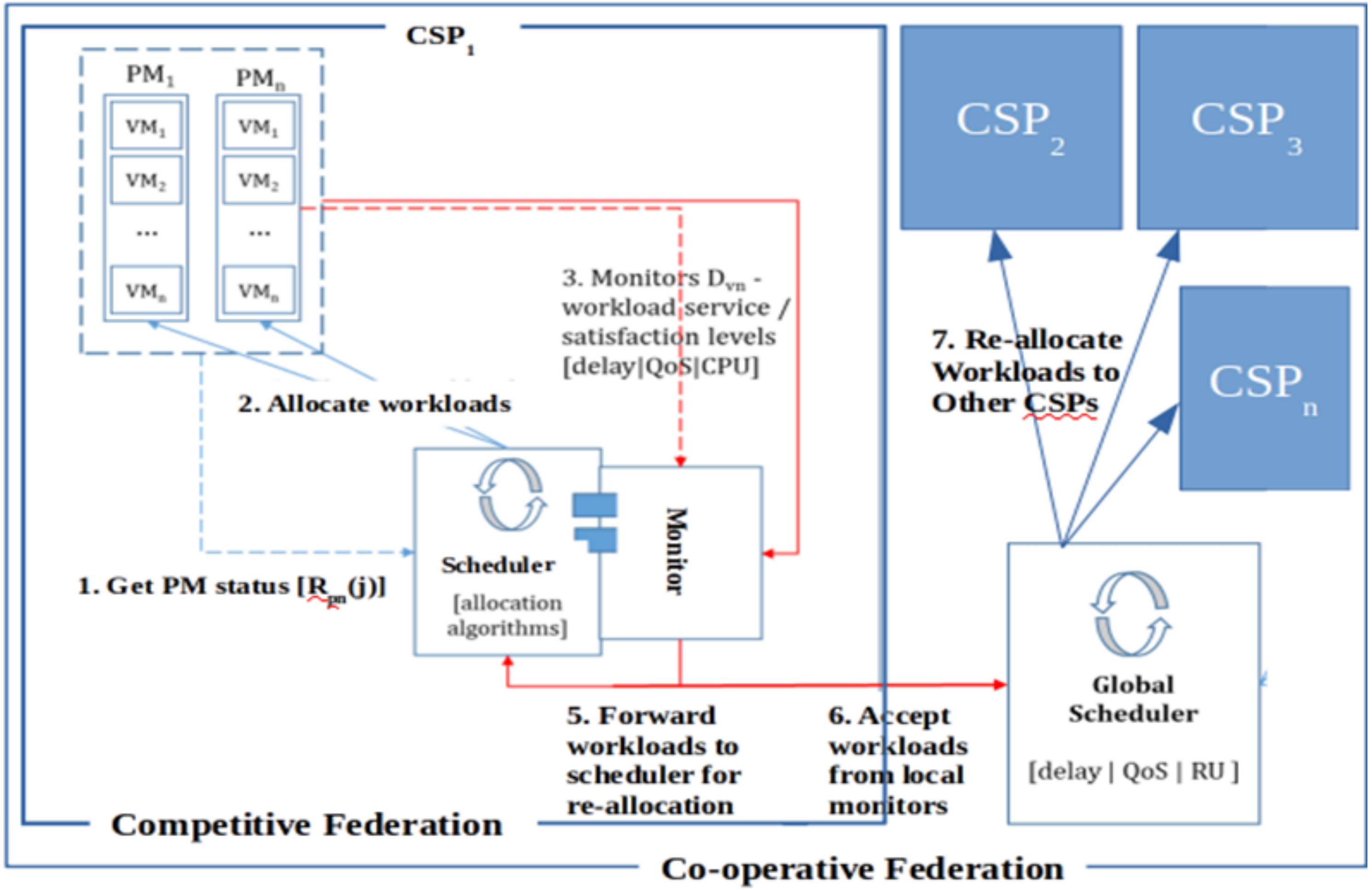
Federation models

- The federation can be done through three models:
  - Cooperative federation model: CSPs work together forming a single virtualized resource pool.
  - Competitive federation model: CSPs work independently.
  - Hybrid federation model: CSPs work independently when under resource-constrained conditions and cooperatively when resources are available.

Big Data processing



Allocation schemes



- Greedy heuristics such as Bin packing and knapsack algorithm.
- Stable marriage & roommate algorithm
- Meta-heuristics such as Genetic Algorithm (GA) & Particle Swam Optimization (PSO)

A Docker -Based Implementation

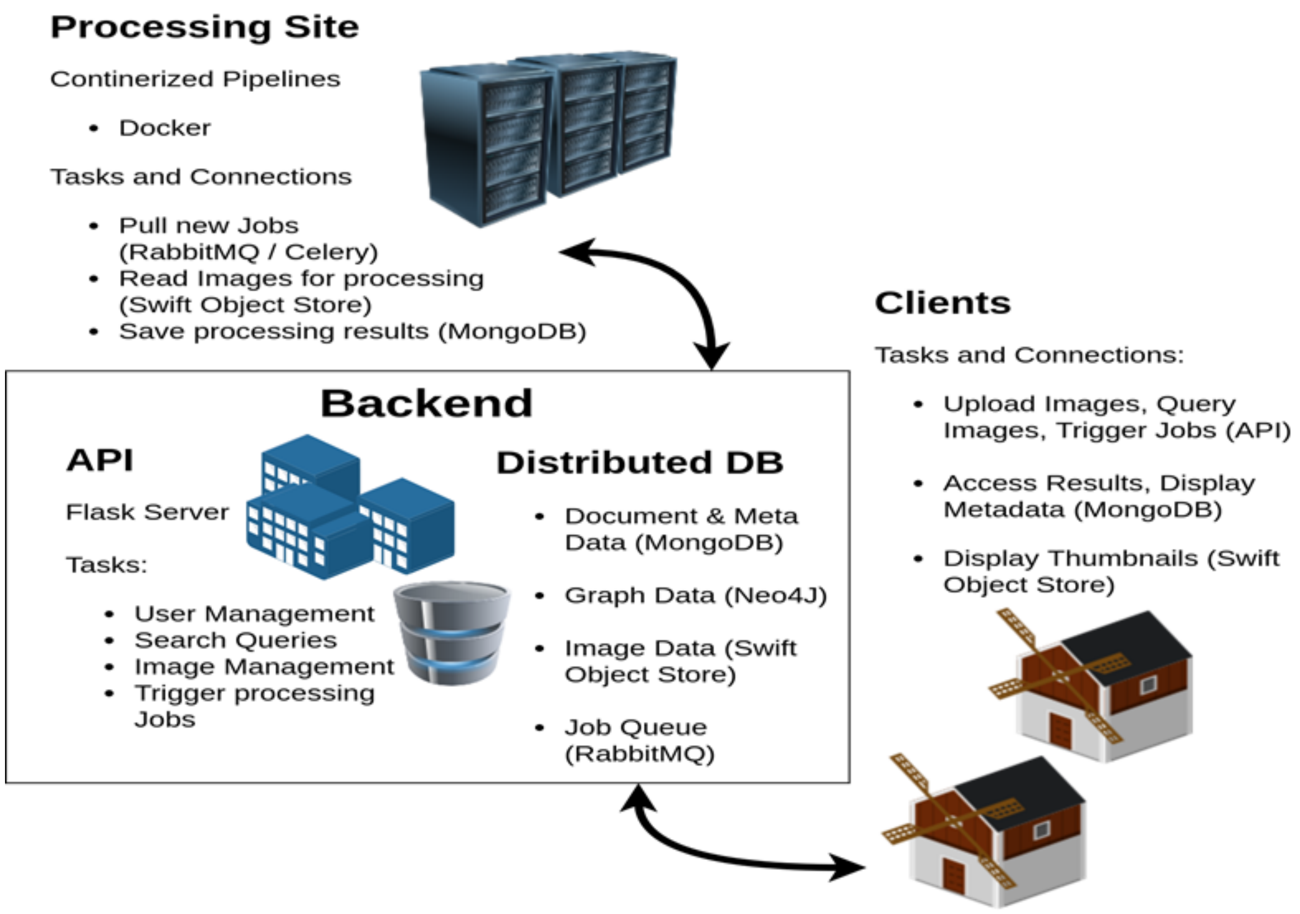
- Install any services and solve contradicting environment requirements on the same hardware by containerizing
- Lightweight distribution compared to other virtualization techniques using virtual machines
- Use of an industry Standard
- Easily share applications with someone else for testing
- Easily to deploy an application to another hardware
- Have an integrated versioning system for required libraries and underlying OS changes

The advantages of a docker-based implementation

ImageFlow

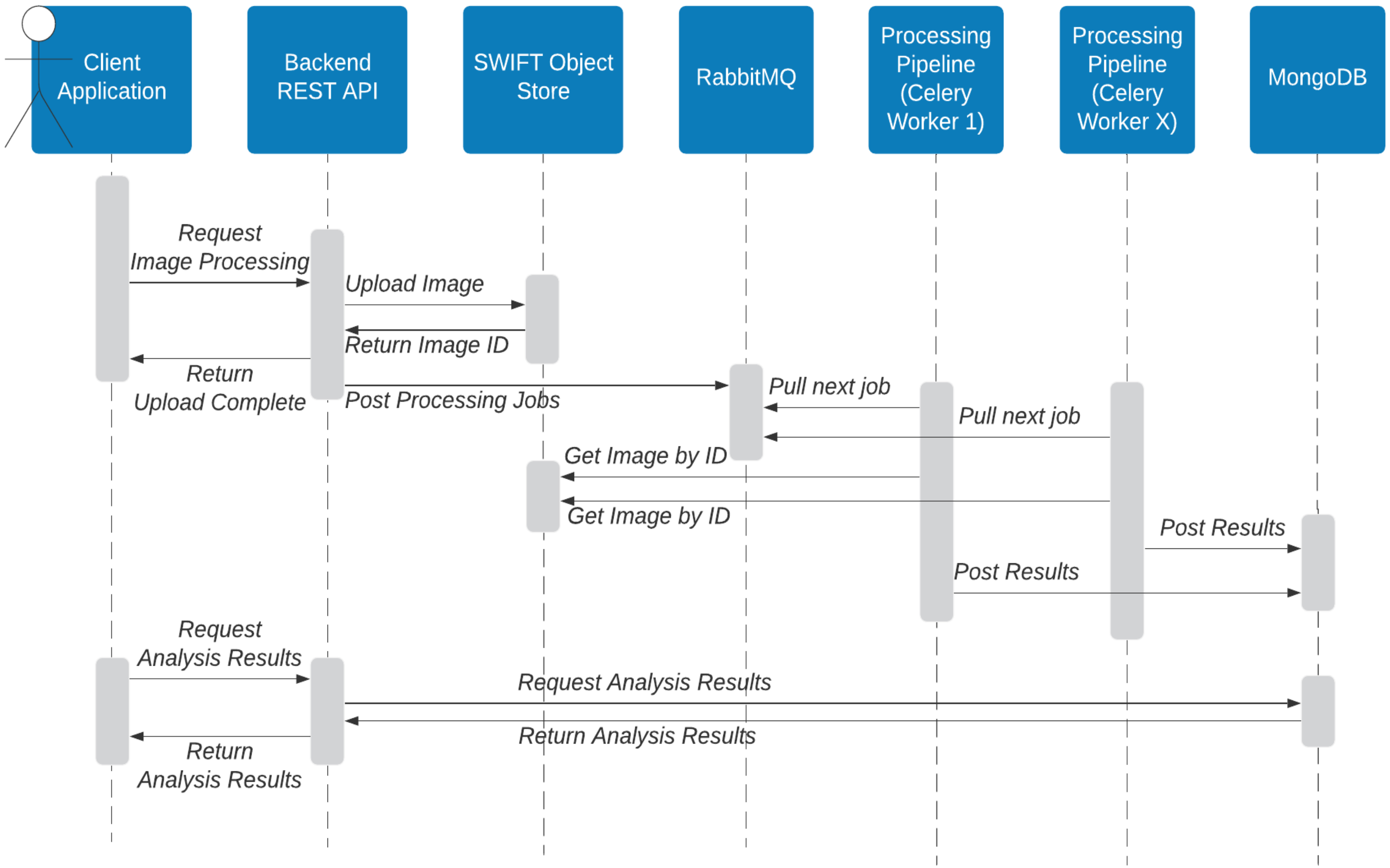
- ImageFlow stores your images on your company server, at home, at one of the big cloud providers or a local data centre you trust
  - Modular. Host components wherever you want. Even separately
  - Open standards and expandability. Link any computing task
  - 100% Open Source & community focused

ImageFlow Architecture

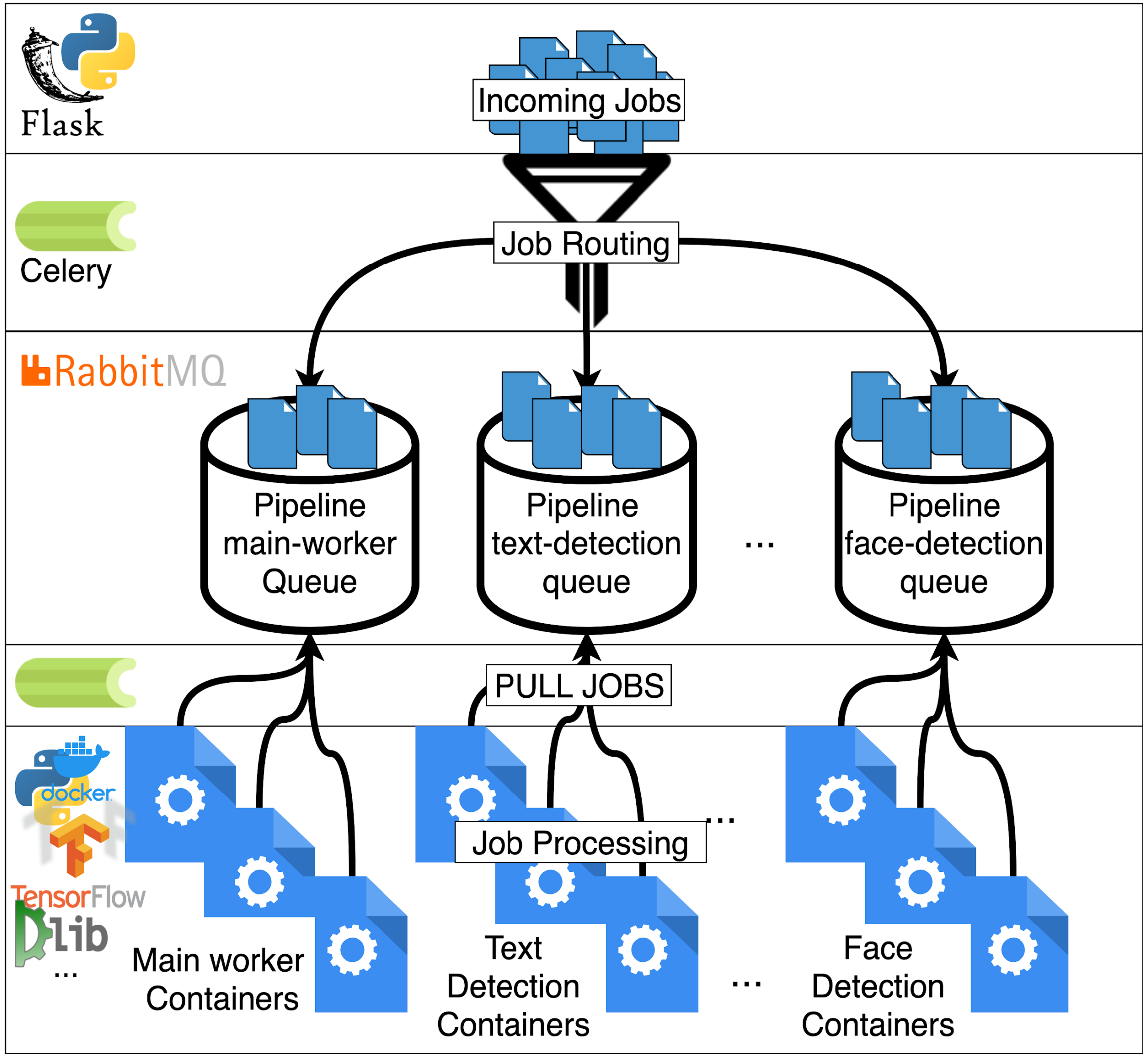


- **Docker-Based Microservices, where each pipeline and element is self-contained.**
- **Scheduling Routed Job Queues for 1+ mio. Jobs/sec**

ImageFlow processing pipeline



ImageFlow scheduling capability



- Assume Each container takes 1 minute per job
- Celery job routing starts topping out at 1 million jobs / sec. = 60 million jobs / min.
- Algorithmic Scalability hits the end when your datacentre supports processing more than **60 million images per minute**

Thank you