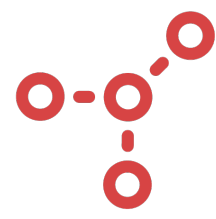




TetraScience Allotrope Conversion As a Service (ACaaS)

An End to End Enterprise Solution Leveraging
Allotrope and HDF5 for Life Sciences

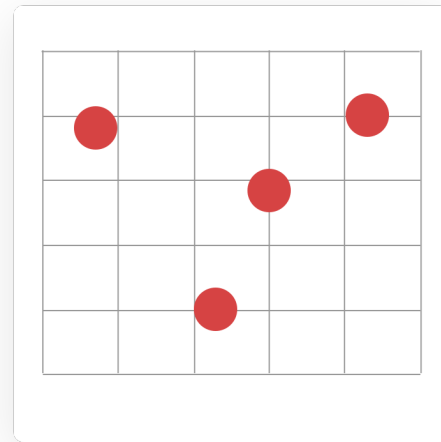
ACaaS, pronounced as “æ•kas”



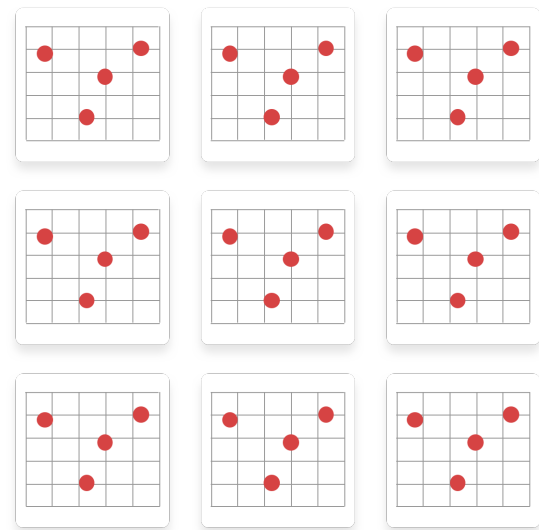
Allotrope Conversion as a Service

Overview

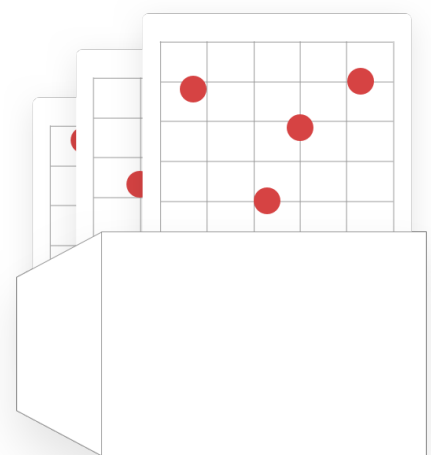
Allotrope
Data
Format
(ADF)



Data Description
(Semantic Model, metadata definitions)



Data Cube
(multi-dimensional matrix)



Data Package
(vendor proprietary file)

HDF5

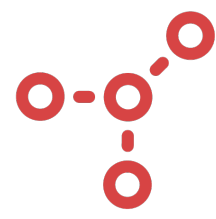
What is Allotrope Data File (ADF) and HDF5? Why are they great?

Allotrope Data File (ADF) is Data Format built on top of HDF5.

It essentially is a container that contains everything (well, almost everything) you probably would like to know and keep about instrument experiment results.

For example, the raw instrument files, the experimental results (such as cell density, chromatograms and etc.), the necessary contexts (such as information about the instrument, experimental method, user, time and etc.)

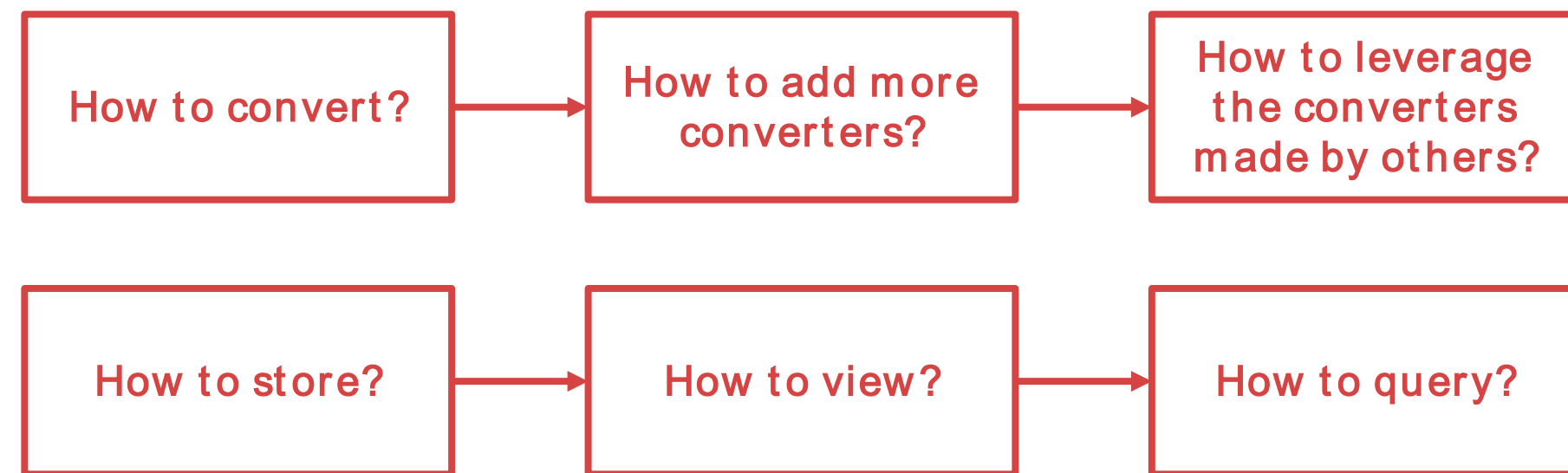
On top of HDF5, Allotrope provides a graph based data model and ontology for key Life Sciences concepts. It leverages triple and RDF to capture scientific context and experimental results and is called "Data Description"



Allotrope Conversion as a Service

Challenges in Adopting Allotrope

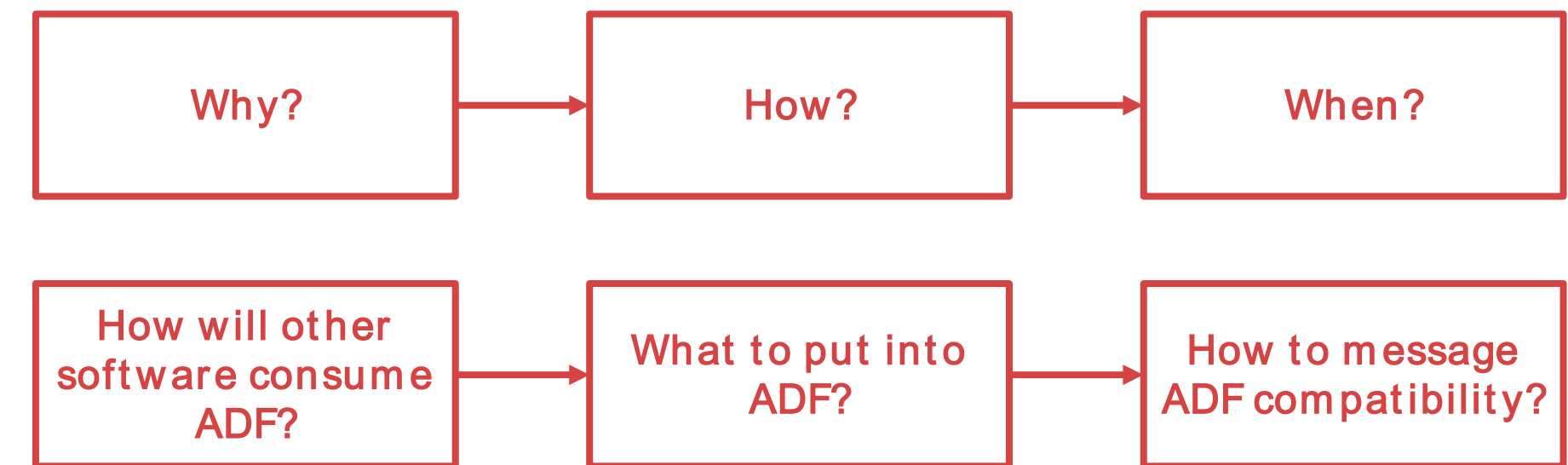
Questions that a **Life Science company** faces in adopting Allotrope



Life Sciences companies

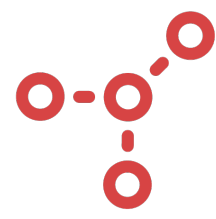
- ADF needs to be adapted for Data Science and Cloud (also observation made by HDF Group)
- Steep learning curve to simply get started
- Difficult to reuse community effort
- Data model and conversion can easily diverge
- Can be slow to deliver business value

Questions that an **Instrument manufacturer** faces in adopting Allotrope



Instrument manufacturers

- Steep learning curve to get started
- Large upfront investment
- Waiting for pharmaceutical companies to require and purchase ADF compatibility
- May not benefit their core business

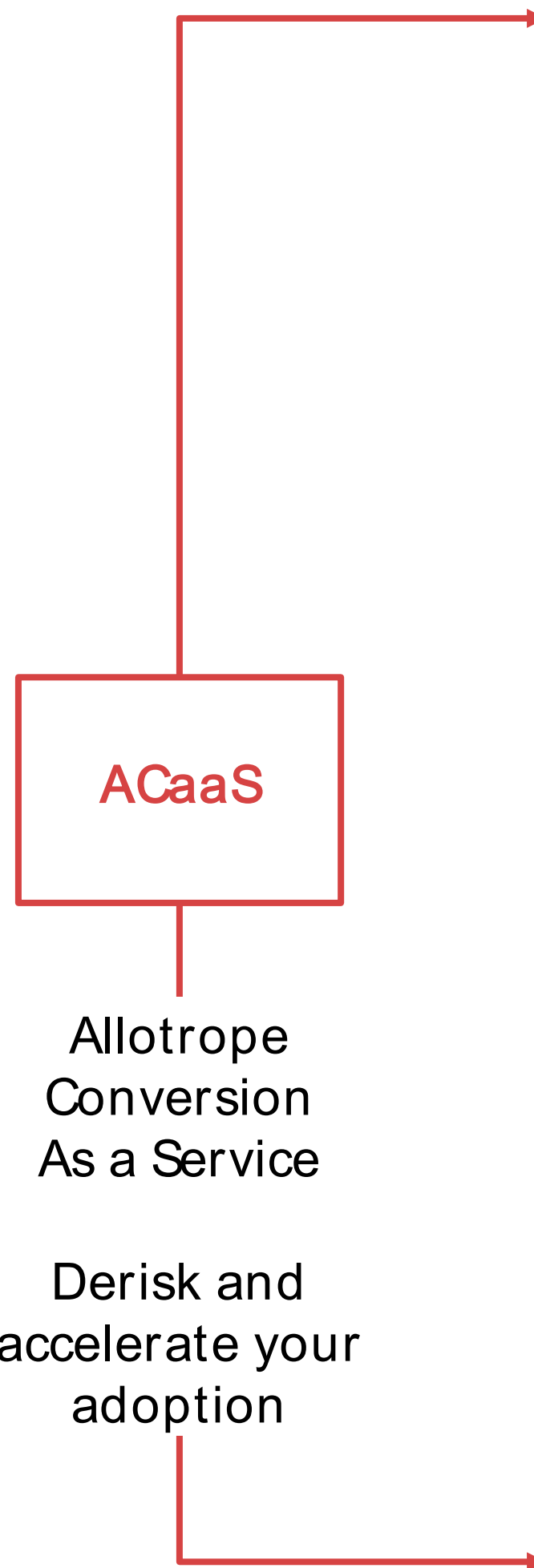


Allotrope Conversion as a Service

Observation and Proposal

Observation of current status

- A lot has been achieved, but **Allotrope is still evolving**. New decisions, methodologies and ideas are being discussed and needs experimentation
- **Chicken vs egg**, there is a dilemma or deadlock around adoption between the pharma and manufacturers.
- There is **gap in Data Science and Cloud** related use cases



One set of tools to generate LOTs ADF Files that covers LOTs of data models via automation. Cloud/docker based, designed for Data Science and no barrier to use.

One set of tools for conversion

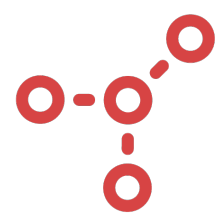
Using the data model differently will only will jeopardize Data Standardization. The conversion should be configurable to enable a guided exploration of different patterns and valid comparison among different approaches.

Automation

Generate LOTS of ADF files and cover LOTS of data models to drive adoption momentum for both the Allotrope Member companies and Allotrope Partner Network companies.

Open source community version (within Allotrope Foundation)

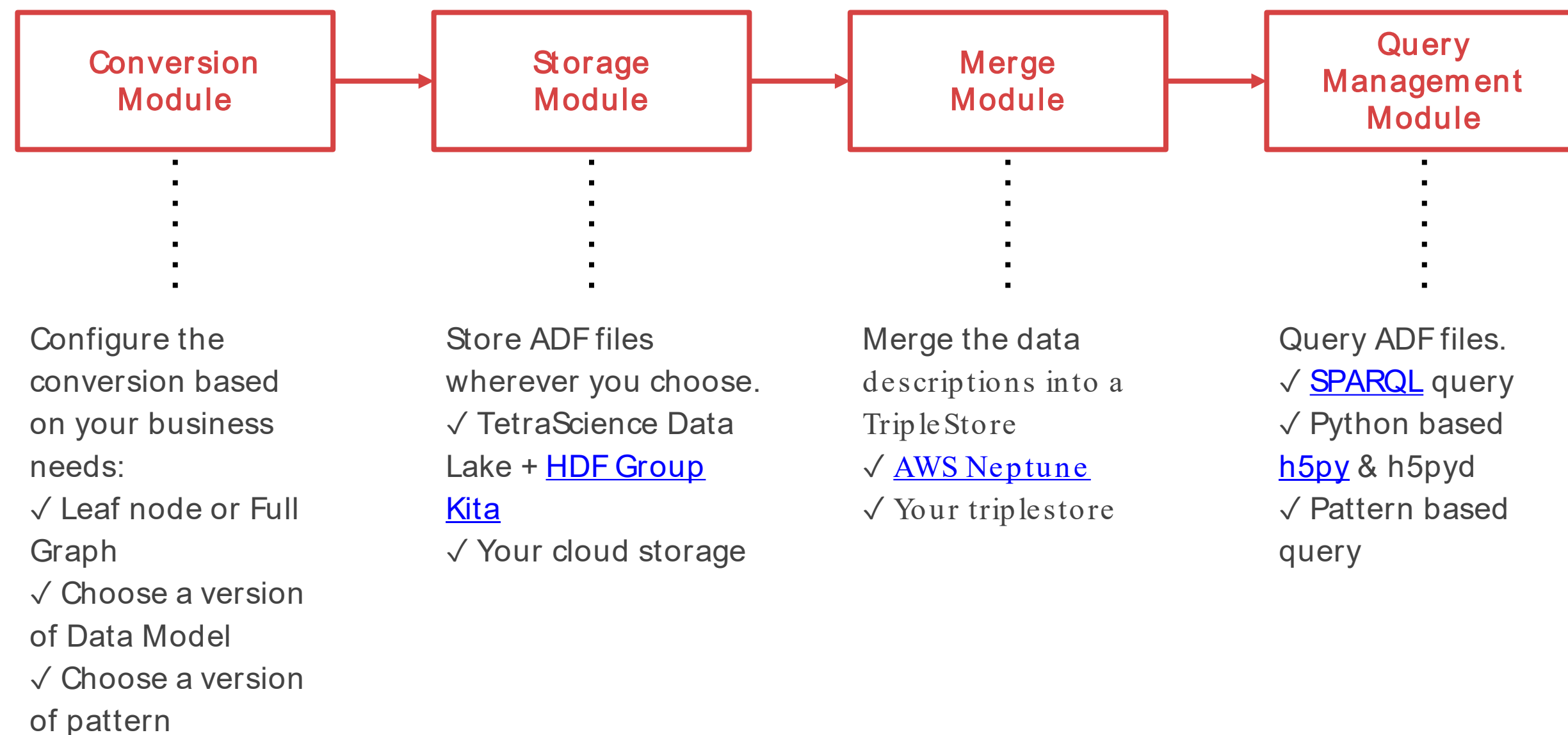
Dramatically lower the barrier for adoption, community contribution. Life Sciences companies just need to focus on building data model!



Allotrope Conversion as a Service

Convert, Store, Merge and Query

ACaaS



End to End

Convert, Store, Merge and Query, ACaaS supports the entire life cycle of Allotrope File, such that your team can focus on performing experiments, building data models and performing analysis.

Configurability

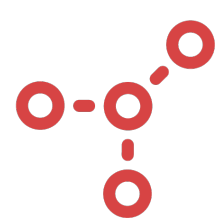
We understand that there are different ways you would like to convert and store our Allotrope files, now you can configure based on your need.

Cloud Native

Built natively on the cloud and compatible with cloud services such that it scales without overhead from your team.

Open Source

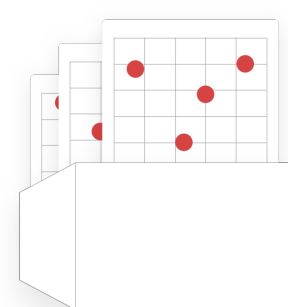
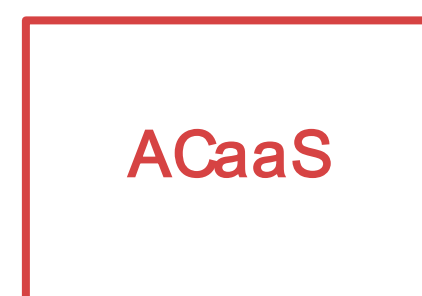
Built on top of TetraScience Data Integration Platform and leverages its Data Connectors, Data Lake and Data Pipelines.



Allotrope Conversion as a Service

Configurable Conversion

Intermediate
Data Schema (IDS)
JSON



```
{
  "sample": {
    "name": "1234-abcd"
  },
  "peak": {
    "number": 1,
    "retention_time": {
      "value": 101,
      "unit": "Milliliter"
    }
  }
}
```

Input should be simple

Use JSON and Excel for discussion, reach consensus and build basic data model. Then have ACaaS introduce the semantics.

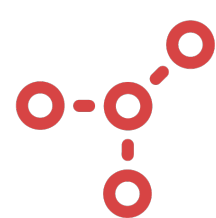
Configure and document the conversion

There are various ways that you may want to convert your instrument data into Allotrope Files. For example, you may choose to use *leaf node pattern*, or *full graph model*, you may choose to use a particular version of the data model and you may choose to apply a particular *pattern*.

TetraScience ACaaS provides a dedicated place that the conversion happens, such that it is clear what pattern is used. TetraScience ACaaS, continues to include the data models finalized by the SMEs from the community.

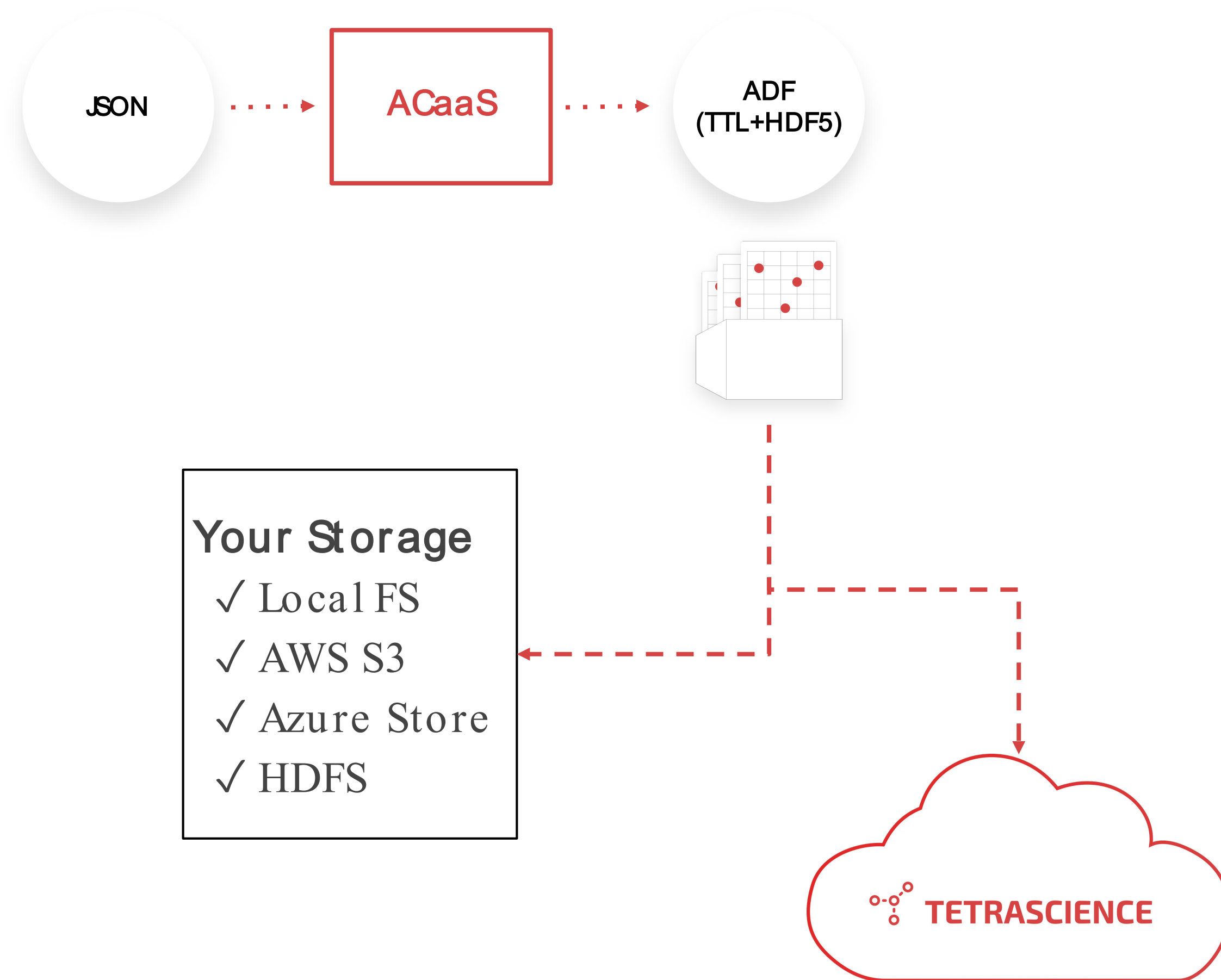
Web API driven

You can send your instrument data in JSON format through web API; ACaaS will convert it for you and then return the ADF file back to you. Thus you can simply use ACaaS as a web service.



Allotrope Conversion as a Service

Store and Find Your ADF Files

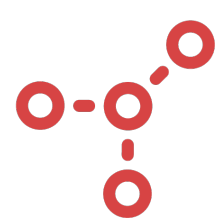


Flexible Storage Options

You can store your ADF files flexibly, for example, local file system, your S3 bucket, your Azure object store, your HDFS or by default, TetraScience Data Lake within our Data Integration Platform, which is built on AWS S3.

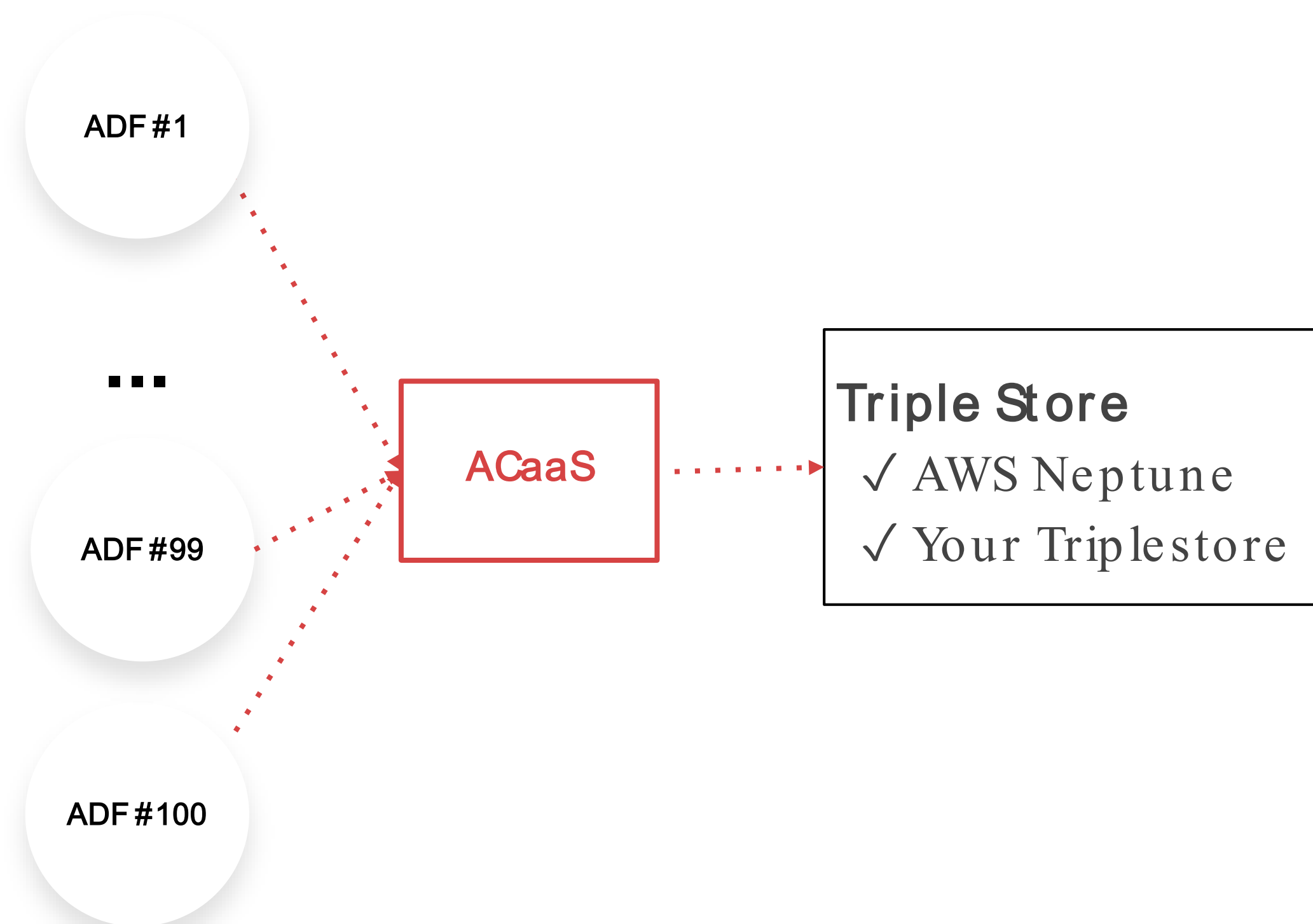
Be able to find your ADF files

Storing your ADF files in the TetraScience Data Lake allows you to search and find your ADF files very easily, using information related to instrument RAW file name, information within the IDS JSON, such as method name, time and etc.



Allotrope Conversion as a Service

Merge Data Description and Graphs



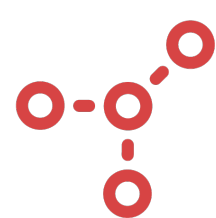
Merge the triples and leverage the power of graph

Data Descriptions follow the semantic web philosophy and uses RDF. To leverage the power and graph based data storage and triple store, it is crucial to merge all the data descriptions from each Allotrope Data File into one single graph.

After each Allotrope file is created, TetraScience will extract the triples and load them into AWS Neptune.

Alternatively, you can also specify your own Triplestore or other Triplestores of interest.

ACaaS Community Version stores the TTL into a local Jena Fuseki server.



Allotrope Conversion as a Service

Slice and Query Your Data Cubes

```
import h5pyd as h5pyd
import matplotlib.pyplot as plt

# define the ADF/HDF5 file
f = h5pyd.File("hdf5://project-a/injection-1.adf")

# explore the ADF/HDF5 file
dset = f['data-cubes/1/measures/2']

print(dset.shape)
# (18000,)

plt.plot(dset[:])
```

Data Cube and HDF5

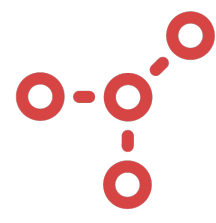
Allotrope Data Cubes are native HDF5 datasets and well suited for storing large, high dimensional data sets, such as images, chromatograms, mass spectrum and etc.

HDF5 Attributes

Besides adding the metadata about the Data Cube into the Data Description. TetraScience ACaaS also uses HDF5's native [attributes](#) for the datasets, such that you can read the ADF file as a pure HDF5 file, without the need to learn SPARQL query.

HDF Kita

TetraScience ACaaS uses HDF Group's [Kita](#). You can use Python to slice and retrieve your Data Cubes. Leveraging the vast amount of data science and software packages provided by Python, scientists and data scientists can now unlock the power of their ADF/HDF5 files and the Data Cubes.



Allotrope Conversion as a Service APN Forum

ACaaS

We would love your feedback!

Come to discuss with the TetraScience team!

APN Forum #1

2:30 PM – 2:55 PM

Wednesday, October 9, 2019, Room 42-1E

APN Forum #2

2:00 PM – 2:15 PM

Thursday, October 10, 2019, Room 42-1E

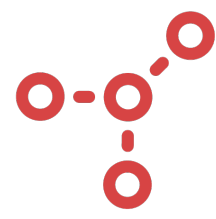


Allotrope Conversion as a Service

APN Forum

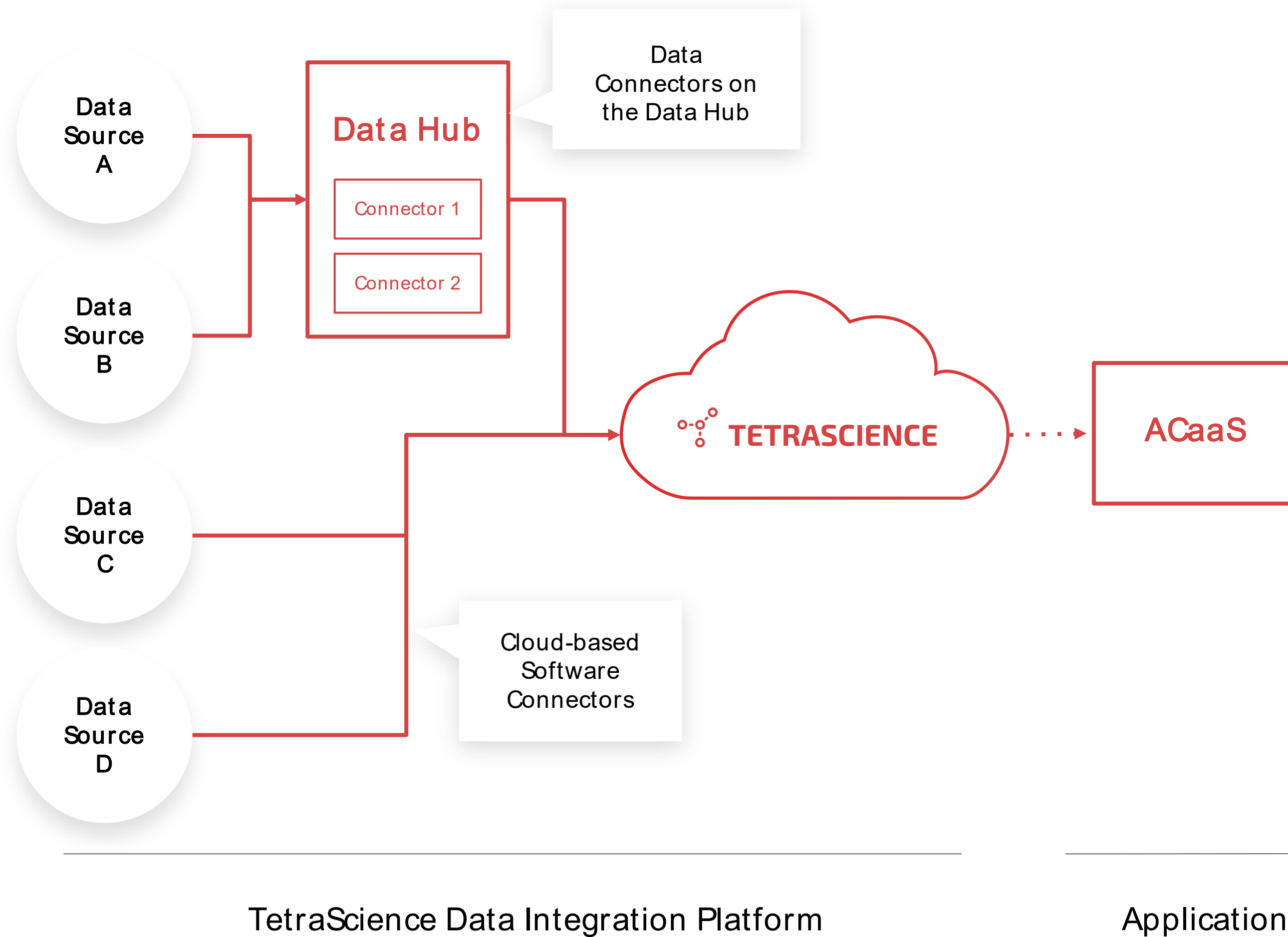
ACaaS

Thank you!



Allotrope Conversion as a Service

An Application On Top of TetraScience Data Integration Platform



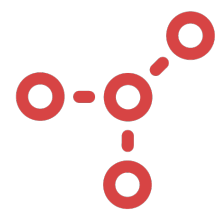
What is ACaaS?

ACaaS is an application built on top of the TetraScience Data Integration Platform.

ACaaS can be enabled very easily through your admin configuration.

ACaaS takes the structured IDS (Intermediate Data Schema) JSON from TetraScience Data Integration Platform and then converts the data into ADF.

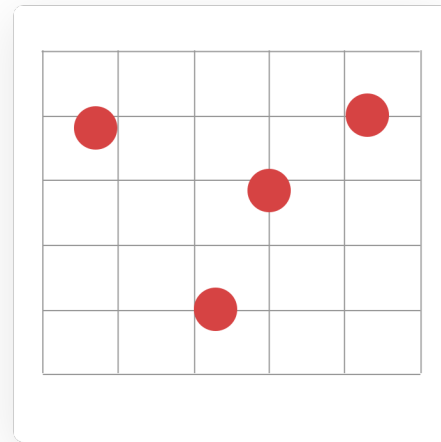
ACaaS Community Version is open source within Allotrope Foundation and can be used **without any dependency**.



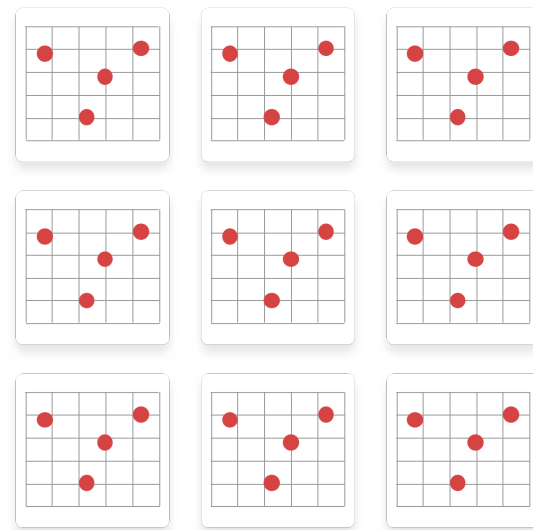
Allotrope Conversion as a Service

“TetraScience ADF Lite”

Allotrope
Data
Format
(ADF)

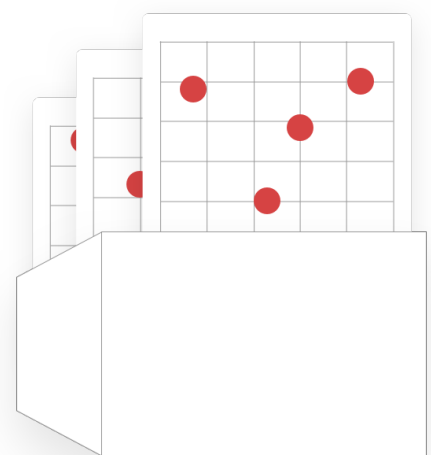


(Data Description)



Data Cube

(cube descriptions
save as HDF5
attributes)



Data Package

(Intermediate Data
Schema, or IDS JSON)

HDF5

Relax the requirements on Data Description -- namely

- do not worry too much that we have to get it right when the file is generated
- do not worry too much if we still need to figure out how to query within graph

Add Data cube descriptions to HDF5 data set attributes

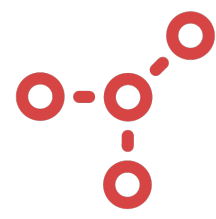
- ADF saves descriptions about the data cubes to the Data Description layer using semantic graphs; in ADF Lite, we leverage HDF5's native [attributes](#) that are attached to HDF5 “groups” and “objects”.

Add the IDS JSON to the Data Package, such that the JSON file can be extracted and used in other web and big data tools, for example:

- Web API, Elasticsearch
- Python/R based data science packages
- Spark, Presto

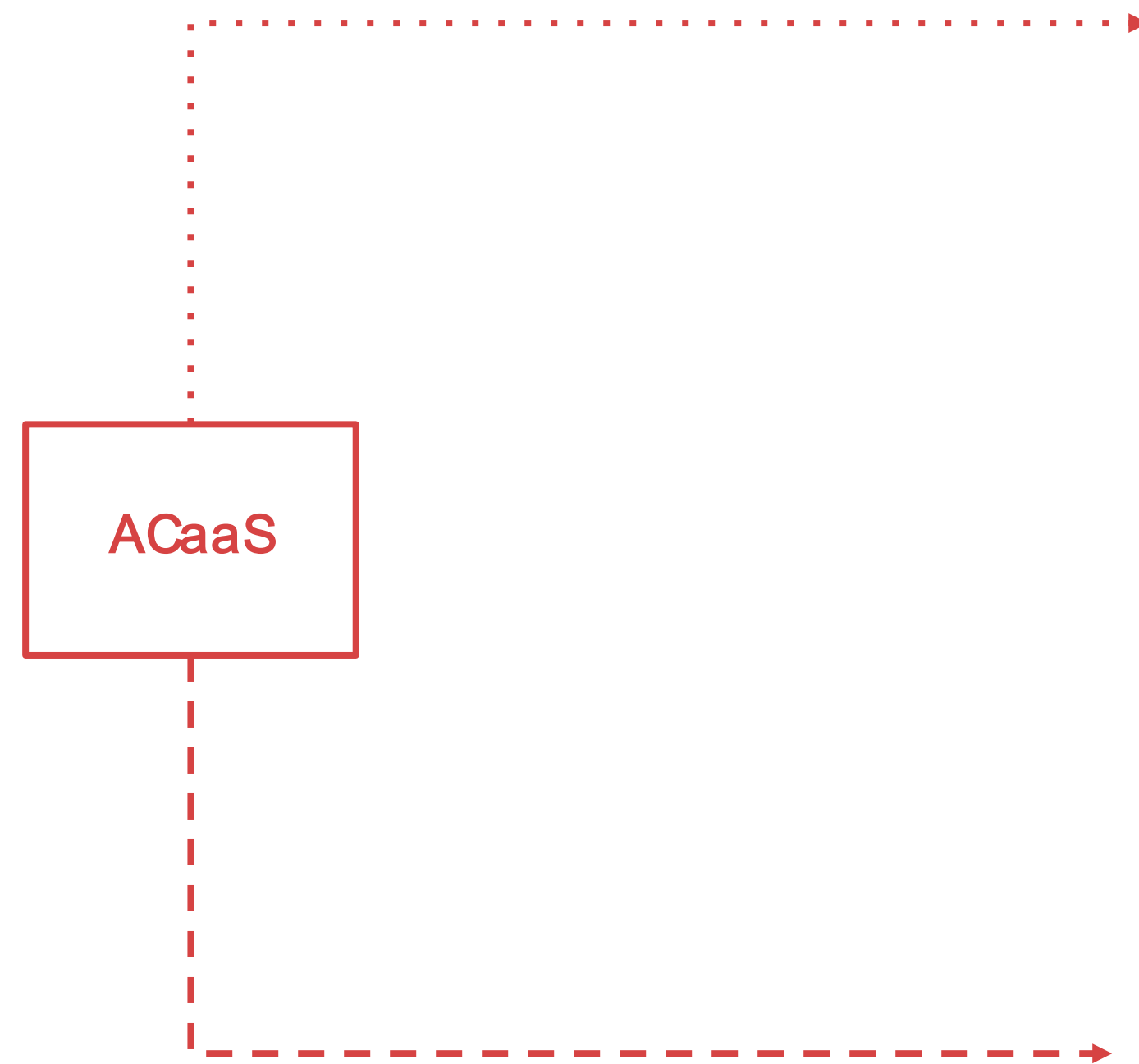
Goal:

- Leverage ADF as a container and drive business value
- Provide more time for graph model and Data Description to mature



Allotrope Conversion as a Service

Our Plan



Community Edition

- Open source
- Docker based
- RESTful API
- Default storage and triple store
- HDF group HSDS

Initial version already available at

<https://allotrope.tetrascience.com>

<https://gitlab.com/allotrope-open-source/ts-lib-allotrope-java>

Enterprise Edition

- Support more storage and triple store
- Supported and managed by TetraScience
- HDF group Kita
- Optimized for enterprise scale and uptime
- Advanced features