Towards the Specification and Generation of Time Series Datasets from Data Lakes

Brian Sal[1], Alfonso de la Vega[1], Patricia López-Martínez[1], Diego García-Saiz[1], Alicia Grande[2], David López[2] and Pablo Sánchez[1]

[1]Software Engineering and Real-Time Systems, Universidad de Cantabria
Santander (Cantabria, Spain)

[2]LIS Data Solutions
Santander (Cantabria, Spain)

Índice

Índice

Context

1. Work developed as part of a project with LIS Data Solutions.
2. LIS Data Solution is a software company based in North Spain focused on data collection and data analysis systems.
3. They wanted to organize all data they store by creating a *datalake*.

Context

1. Work developed as part of a project with LIS Data Solutions.
2. LIS Data Solution is a software company based in North Spain focused on data collection and data analysis systems.
3. They wanted to organize all data they store by creating a *datalake*.

Context

1. Work developed as part of a project with LIS Data Solutions.
2. LIS Data Solution is a software company based in North Spain focused on data collection and data analysis systems.
3. They wanted to organize all data they store by creating a *datalake*.

Índice

Data Lakes

1. Data is considered *the new gold*, so companies preserved them.
2. Nobody knows how these data are going to be analyzed.
3. A *data lake* is a kind of organized (big) hard drive where heterogeneous data is stored in raw format.

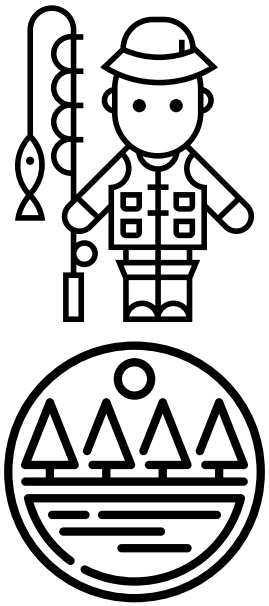
Data Lakes

1. Data is considered *the new gold*, so companies preserved them.
2. Nobody knows how these data are going to be analyzed.
3. A *data lake* is a kind of organized (big) hard drive where heterogeneous data is stored in raw format.

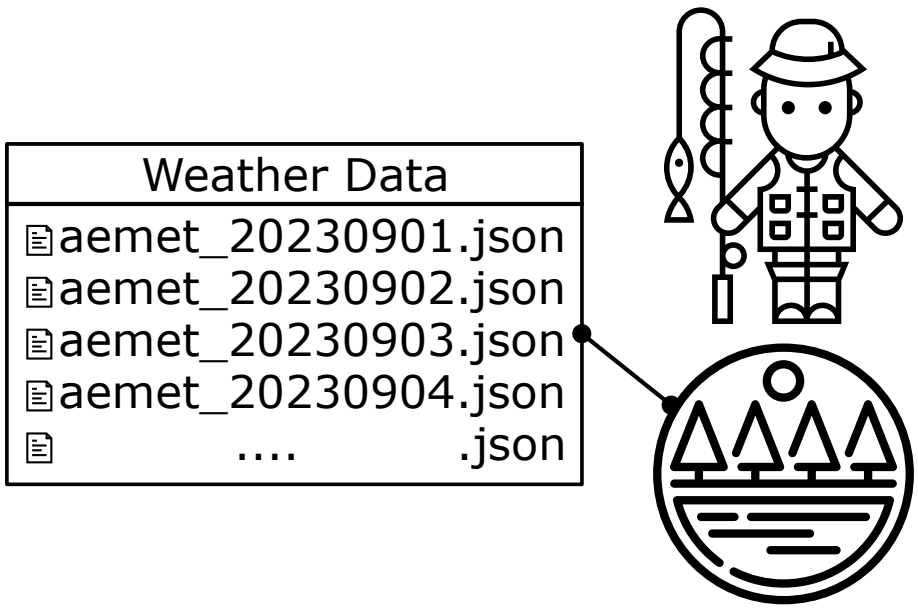
Data Lakes

1. Data is considered *the new gold*, so companies preserved them.
2. Nobody knows how these data are going to be analyzed.
3. A *data lake* is a kind of organized (big) hard drive where heterogeneous data is stored in raw format.

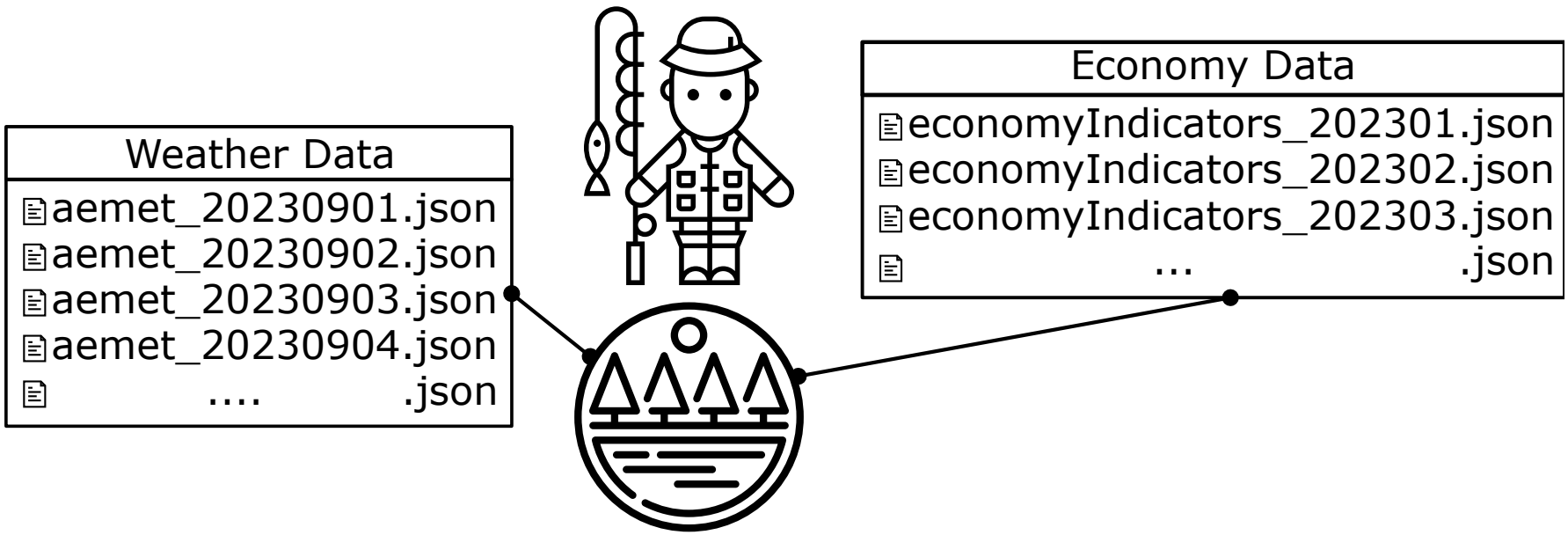
Data Lake Example



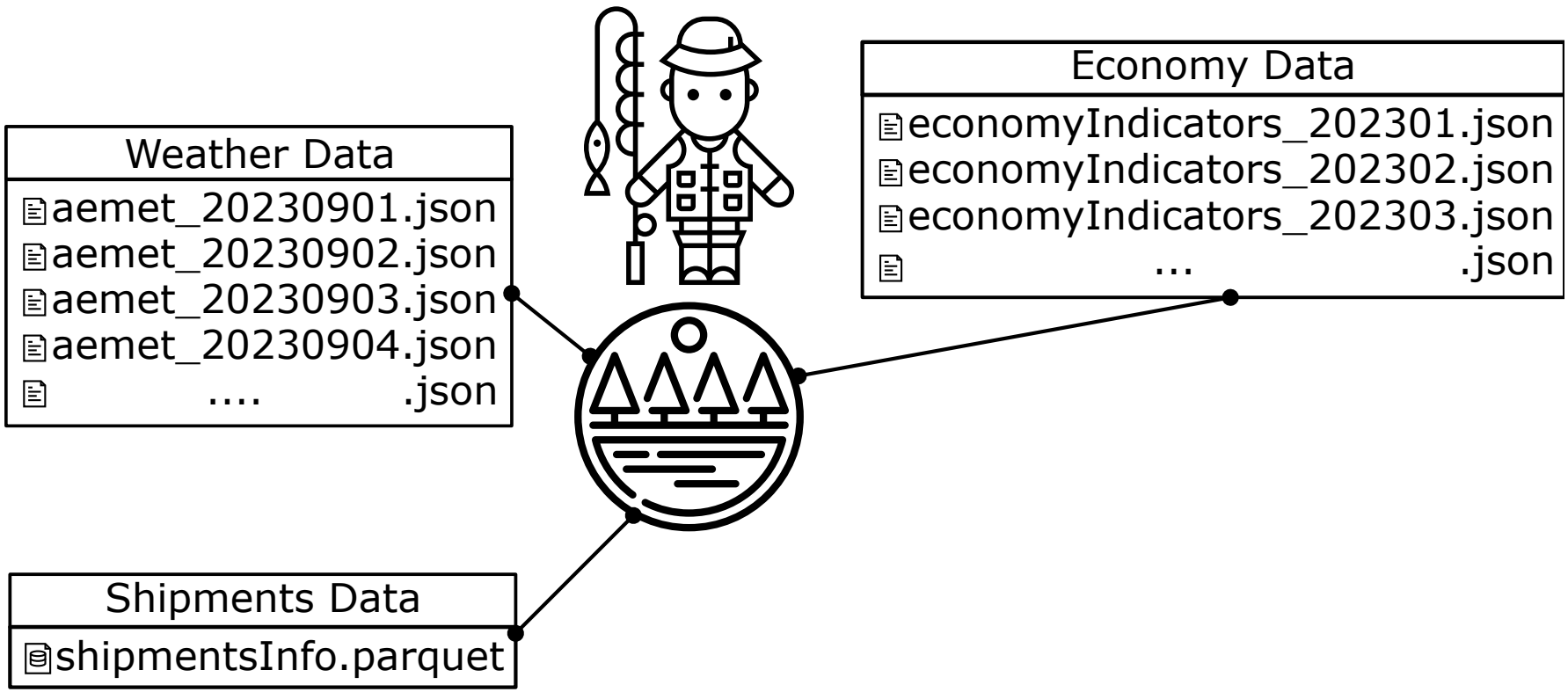
Data Lake Example



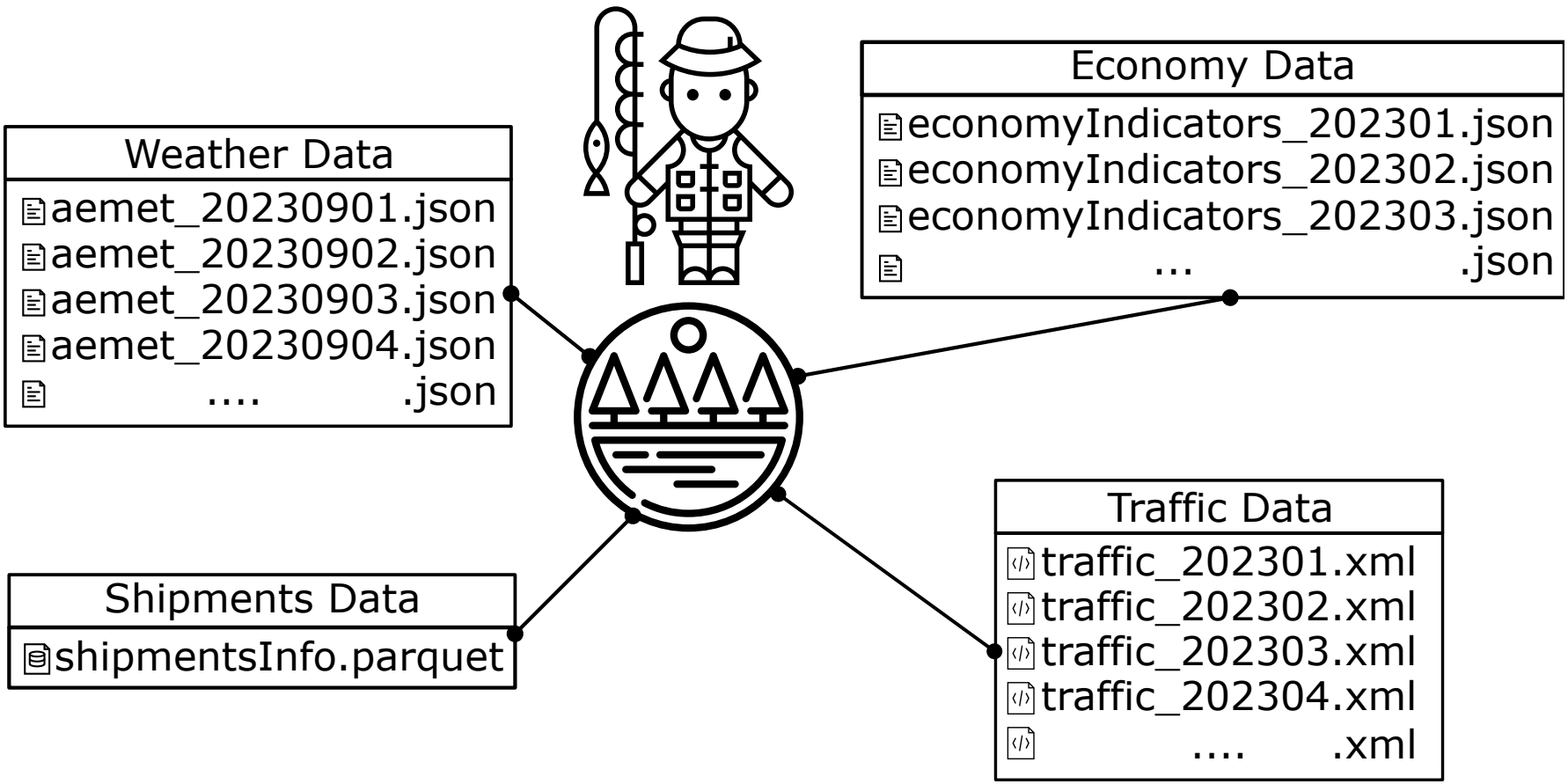
Data Lake Example



Data Lake Example



Data Lake Example



Índice

Índice

Running Example

Goal

Forecast the shipment volume and delivery times for some logistics companies using the data stored in the datalake for a specific Spanish city.

Índice

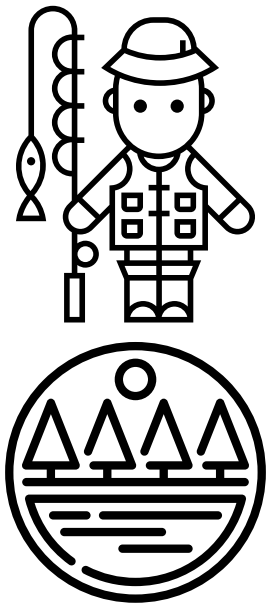
Time Series Datasets

Date	Shipments	Temp. (°C)	Rain (mm)	GDP	Traffic (trucks/h)
2019-02-07	31	13.4	1.7	965.6	1230
2019-02-08	28	10.2	0.5	965.6	1320
2019-02-09	34	11.8	0.8	965.6	1280
...	...	...	...	...	...

Dataset Creation Challenges

1. Find the appropriate set of *blobs* or files.
2. Select the correct subset of *blobs* for each data source.
3. Handle multiple and heterogeneous file formats.
4. Extract data from each file or set of files
5. Harmonize of sampling frequencies.

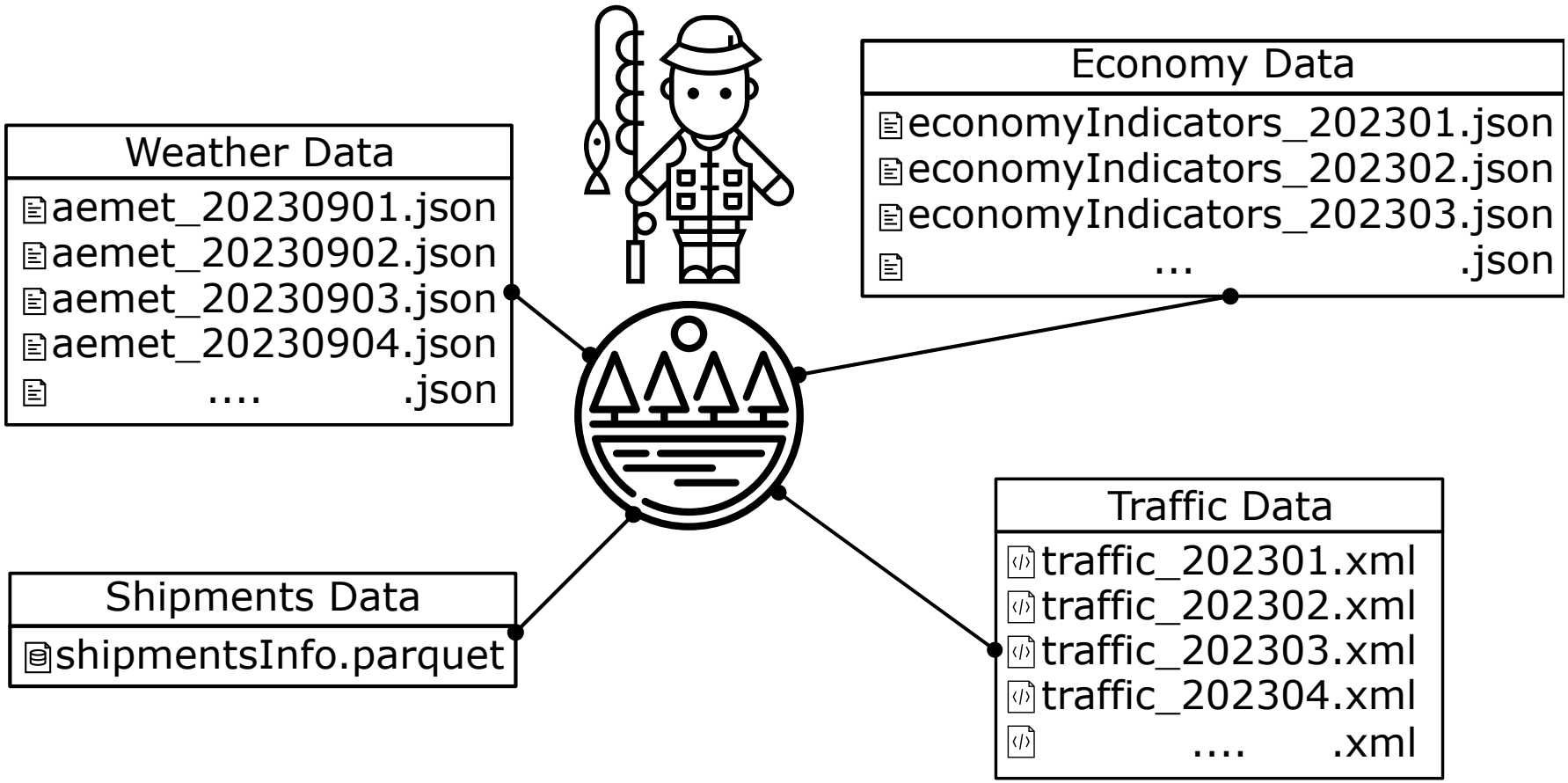
Dataset Creation Challenges



Dataset Creation Challenges

1. Find the appropriate set of *blobs* or files.
2. Select the correct subset of *blobs* for each data source.
3. Handle multiple and heterogeneous file formats.
4. Extract data from each file or set of files
5. Harmonize of sampling frequencies.

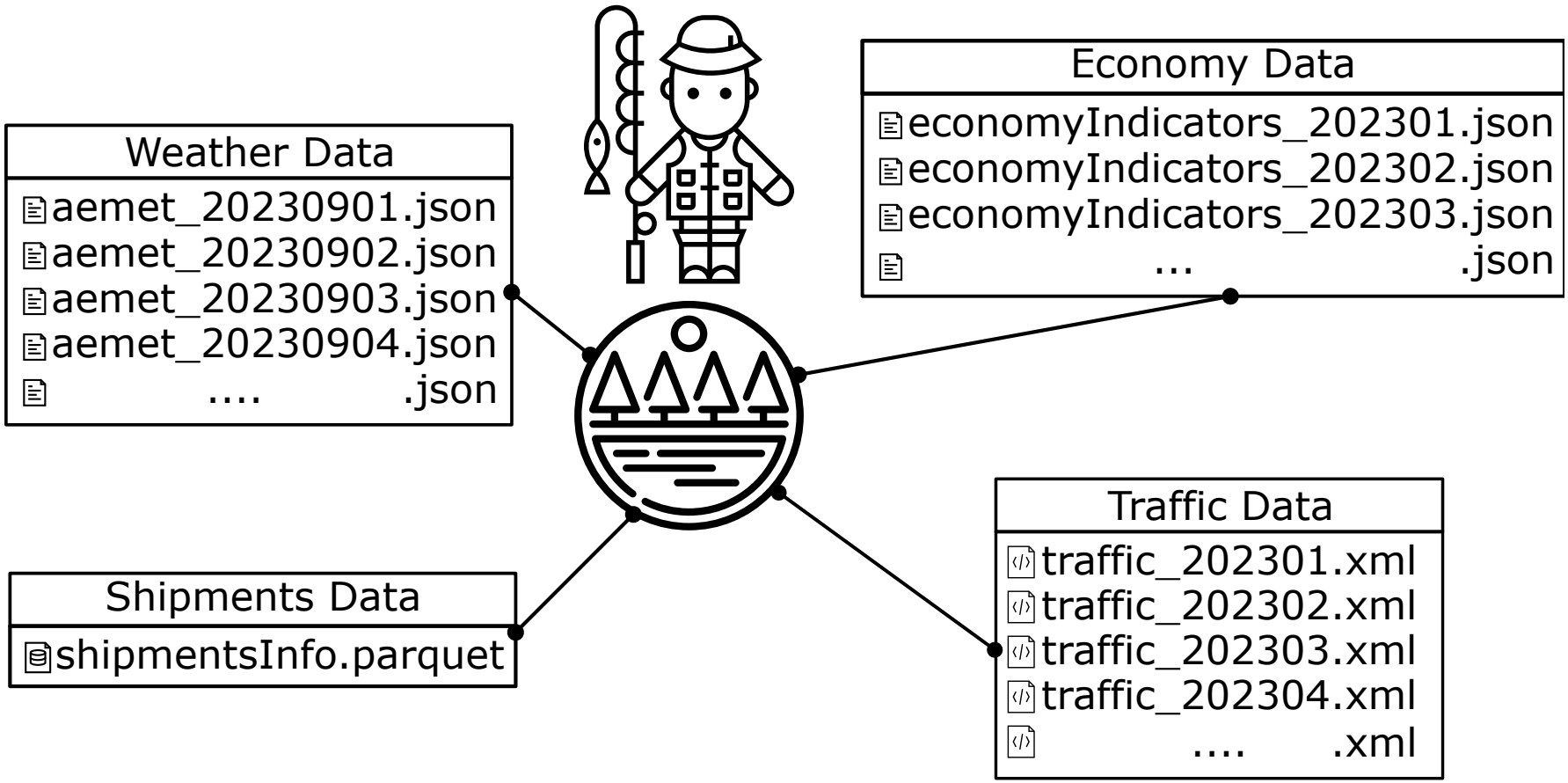
Dataset Creation Challenges



Dataset Creation Challenges

1. Find the appropriate set of *blobs* or files.
2. Select the correct subset of *blobs* for each data source.
3. Handle multiple and heterogeneous file formats.
4. Extract data from each file or set of files
5. Harmonize of sampling frequencies.

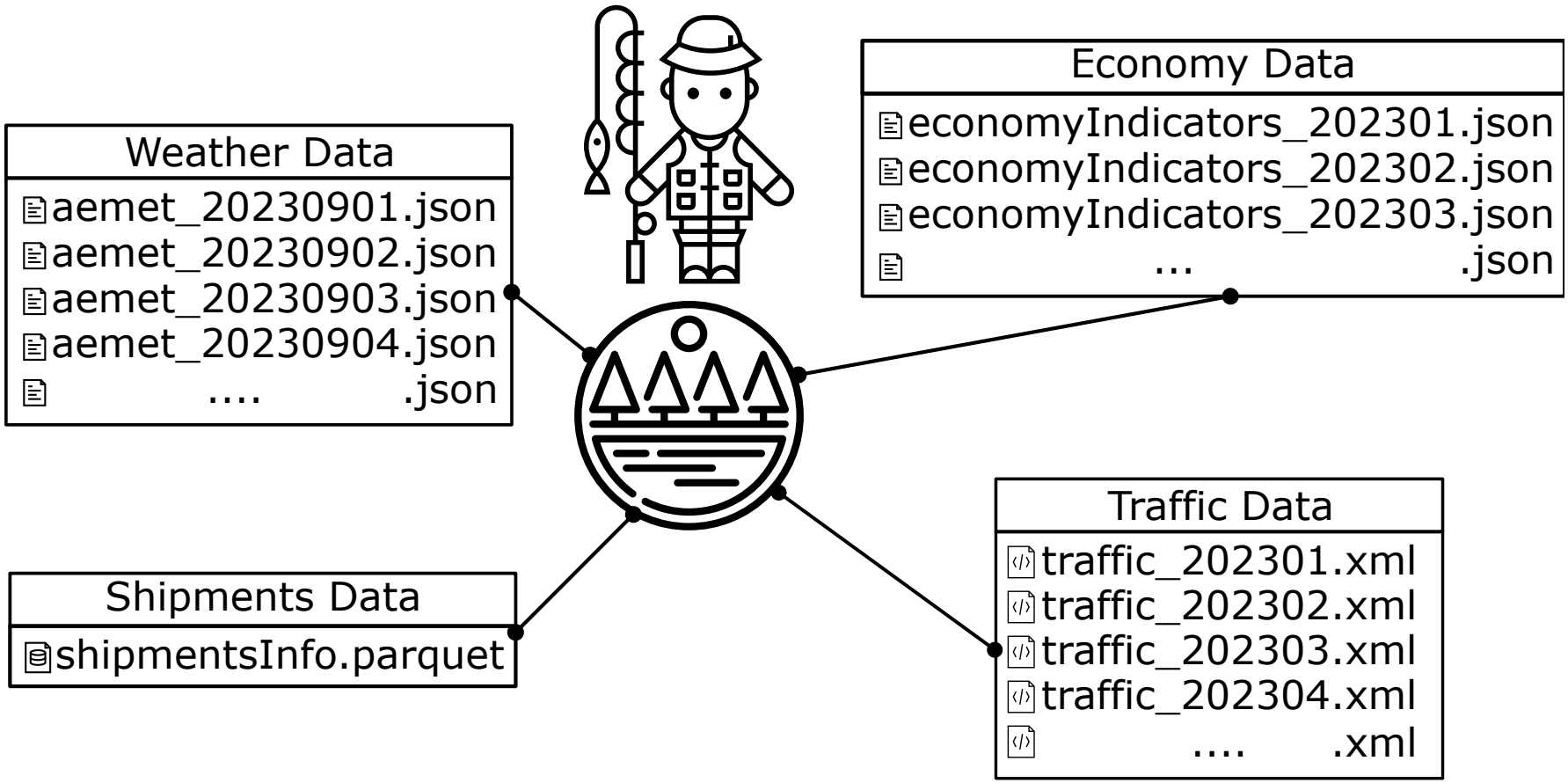
Dataset Creation Challenges



Dataset Creation Challenges

1. Find the appropriate set of *blobs* or files.
2. Select the correct subset of *blobs* for each data source.
3. Handle multiple and heterogeneous file formats.
4. Extract data from each file or set of files
5. Harmonize of sampling frequencies.

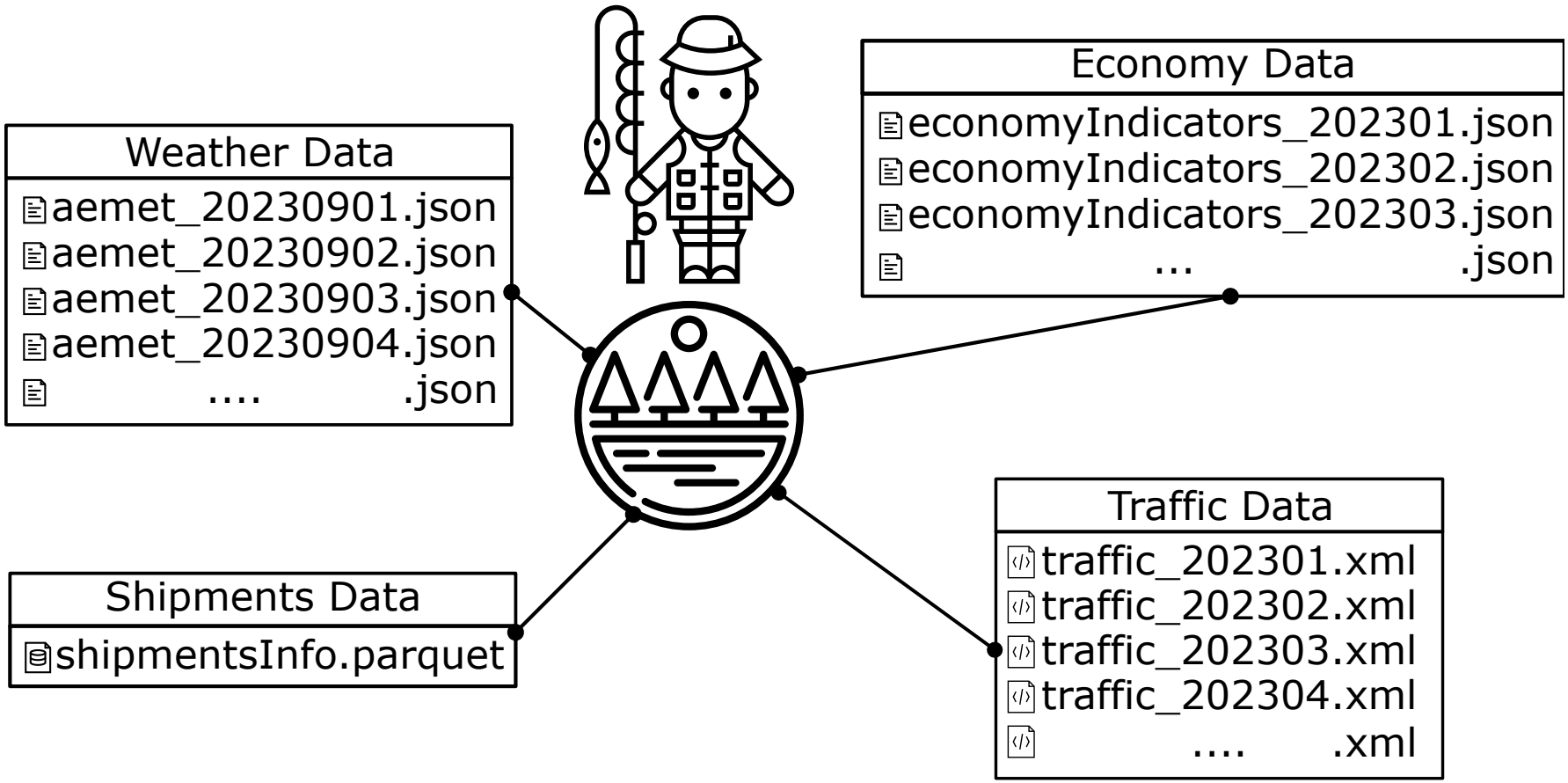
Dataset Creation Challenges



Dataset Creation Challenges

1. Find the appropriate set of *blobs* or files.
2. Select the correct subset of *blobs* for each data source.
3. Handle multiple and heterogeneous file formats.
4. Extract data from each file or set of files
5. Harmonize of sampling frequencies.

Dataset Creation Challenges



Dataset Creation State-of-the-Art

1. Data scientists write large scripts in languages such as R or Pandas.
2. The script for the running example is $\sim$ 500 lines long.
3. It is a time consuming and error-prone process.
4. It hampers the participation of domain experts in this process.

Dataset Creation State-of-the-Art

1. Data scientists write large scripts in languages such as R or Pandas.
2. The script for the running example is $\sim$ 500 lines long.
3. It is a time consuming and error-prone process.
4. It hampers the participation of domain experts in this process.

Dataset Creation State-of-the-Art

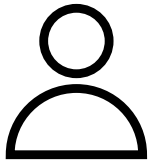
1. Data scientists write large scripts in languages such as R or Pandas.
2. The script for the running example is $\sim$ 500 lines long.
3. It is a time consuming and error-prone process.
4. It hampers the participation of domain experts in this process.

Dataset Creation State-of-the-Art

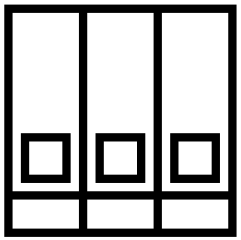
1. Data scientists write large scripts in languages such as R or Pandas.
2. The script for the running example is $\sim$ 500 lines long.
3. It is a time consuming and error-prone process.
4. It hampers the participation of domain experts in this process.

Índice

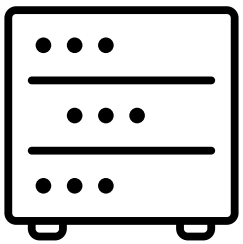
Solution Overview



Data Scientist / Stakeholder

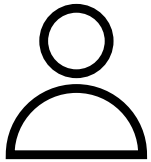


Data Catalog

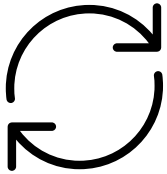


Data Lake

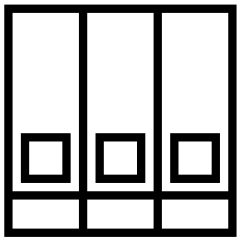
Solution Overview



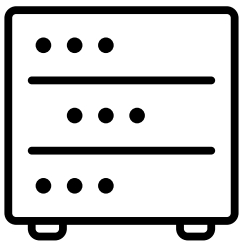
Data Scientist / Stakeholder



Feature Selection

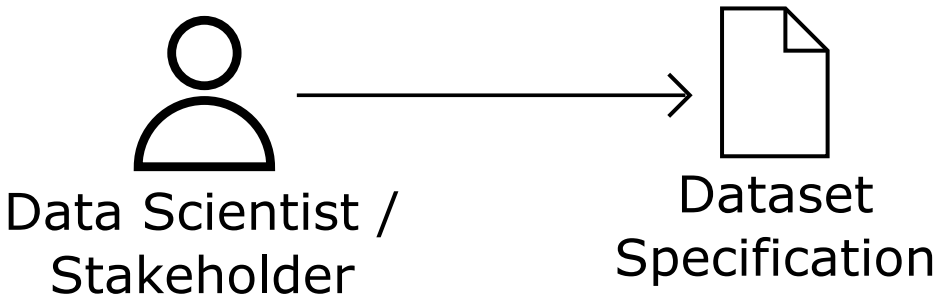


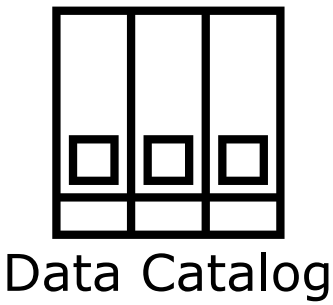
Data Catalog

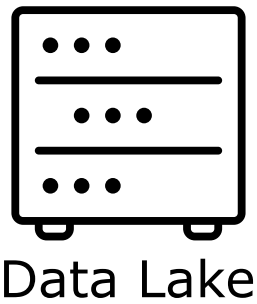


Data Lake

Solution Overview



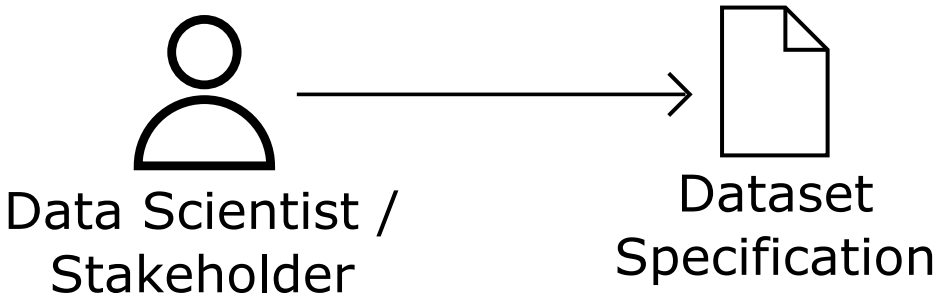


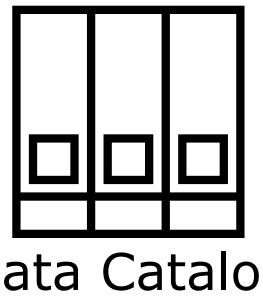


The Hannah Language

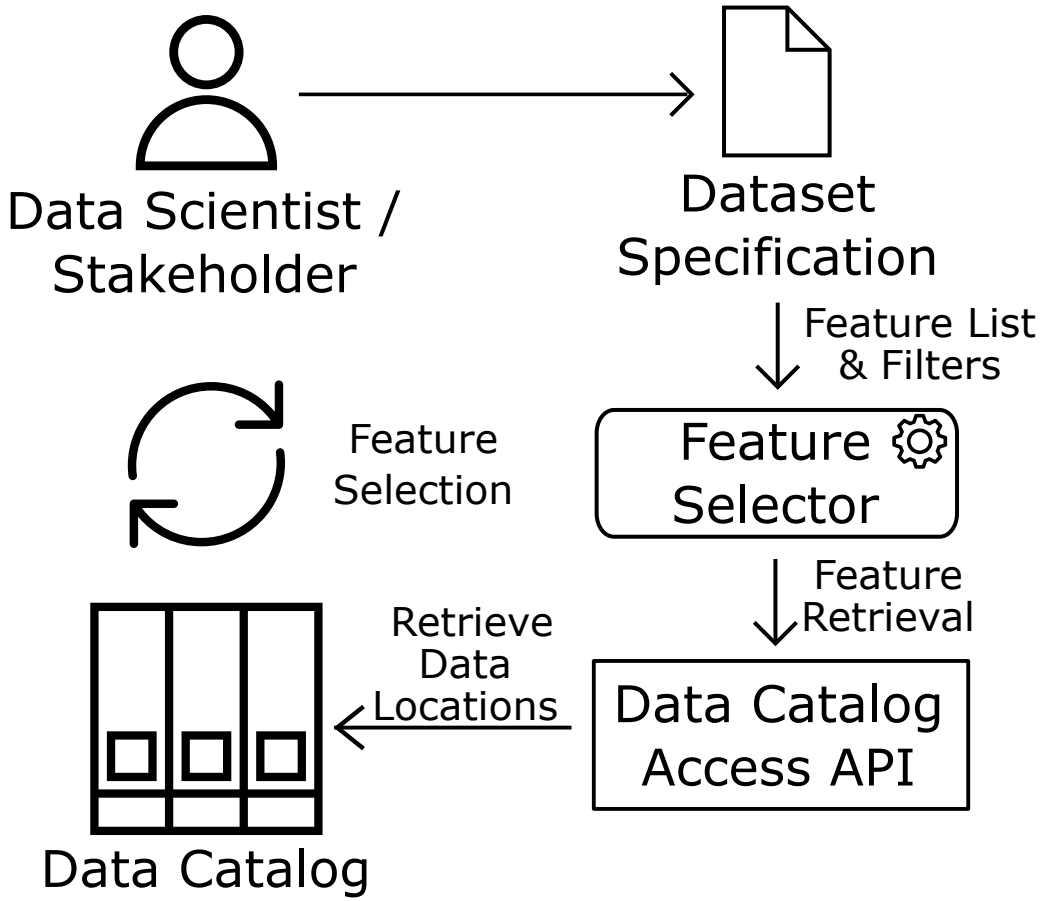
```
1  dataset ShipmentForecasting
2      sampling daily
3      from 2023/01/01 to 2023/06/30
4      with currentDay as index
5      with features
6        from ShipmentInformation {
7            shipments is count(id[orderDate=currentDay])
8        }
9        from WeatherData[city='Santander'] {
10           temperature;
11           rain;
12       }
13       from EconomyData {
14           gdp expanded by linear_interpolation;
15       }
16
```

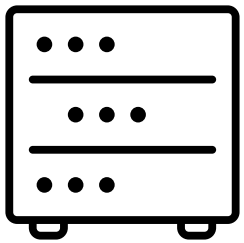
The Hannah Framework



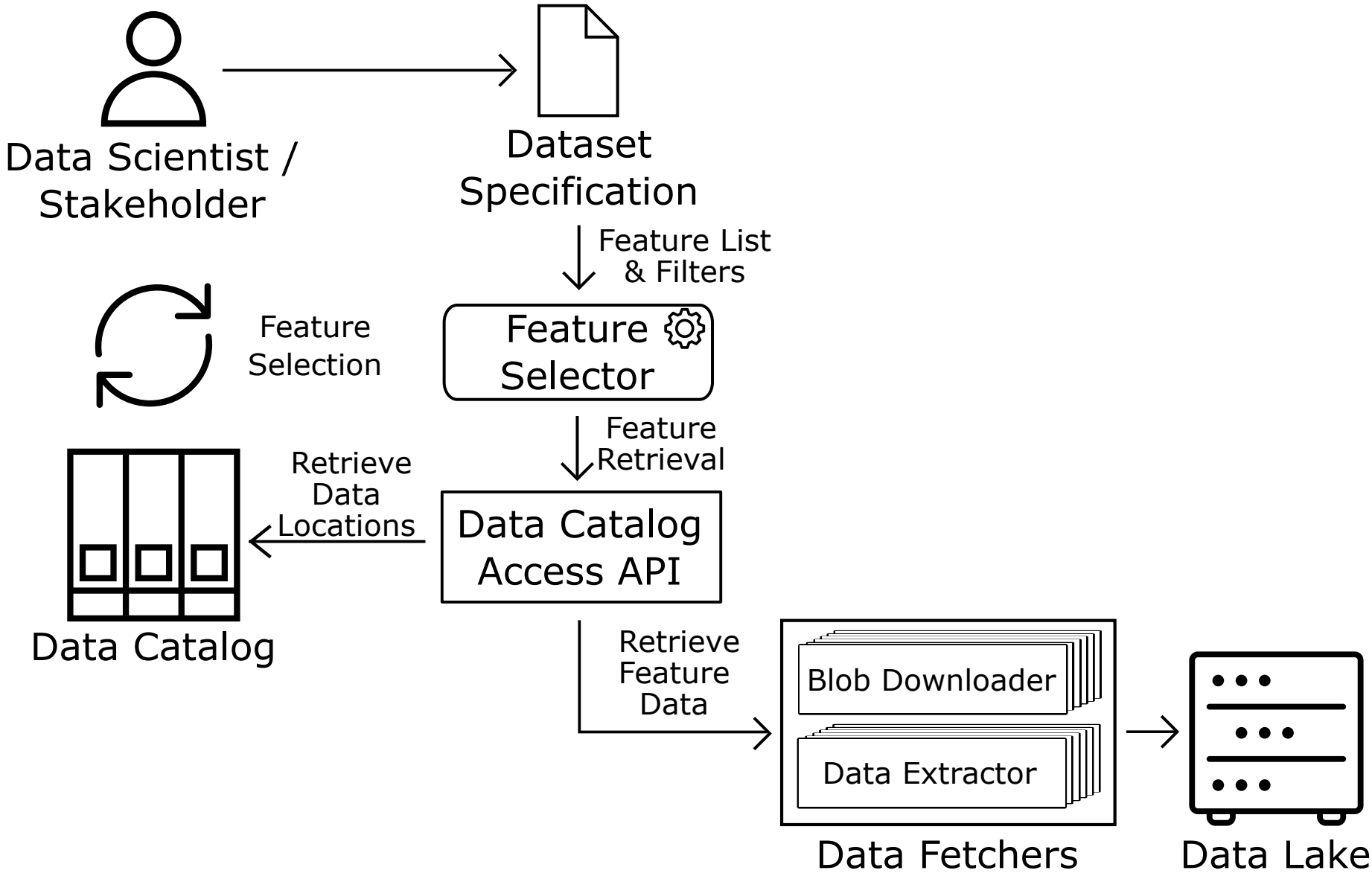


The Hannah Framework

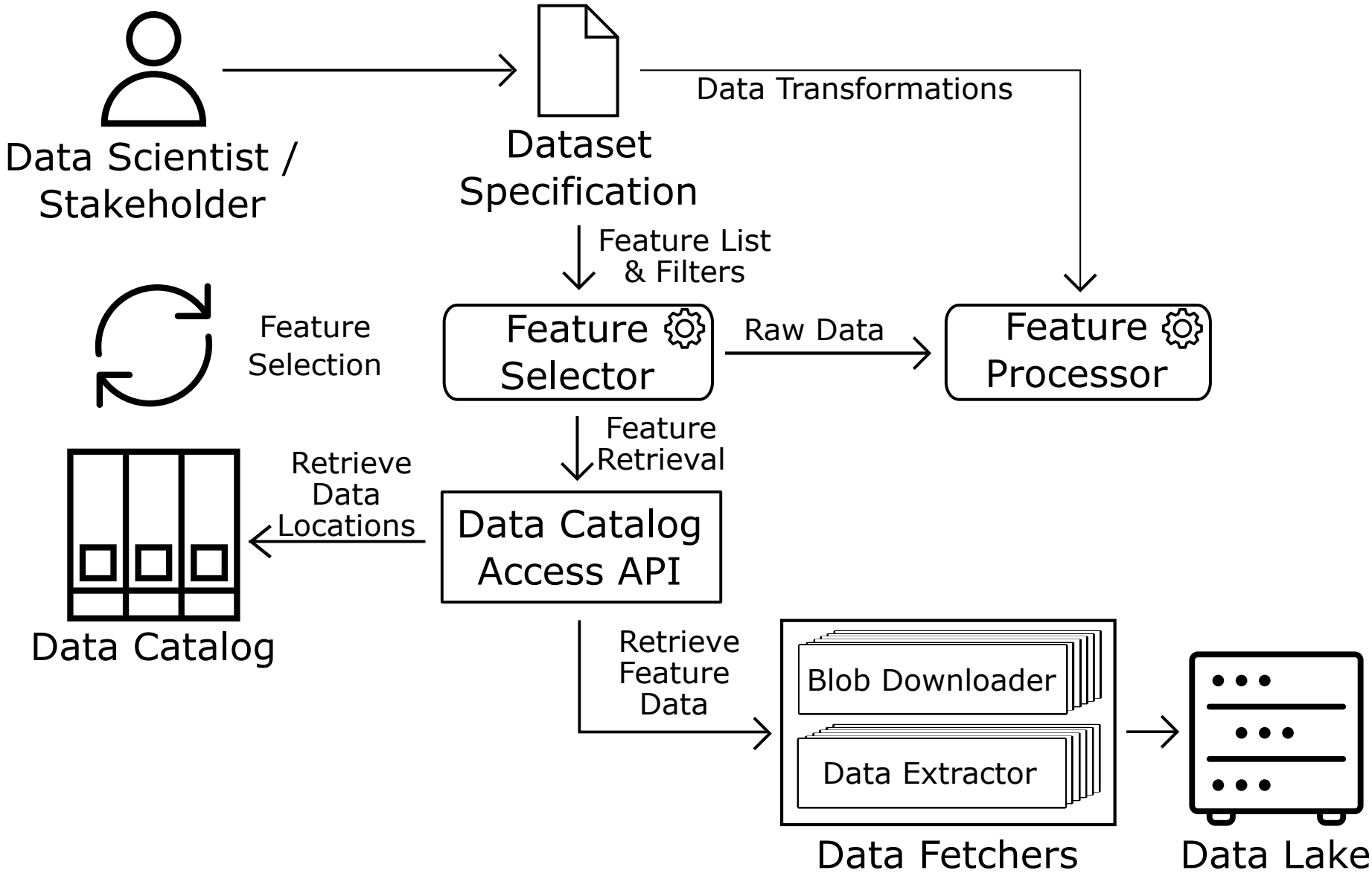




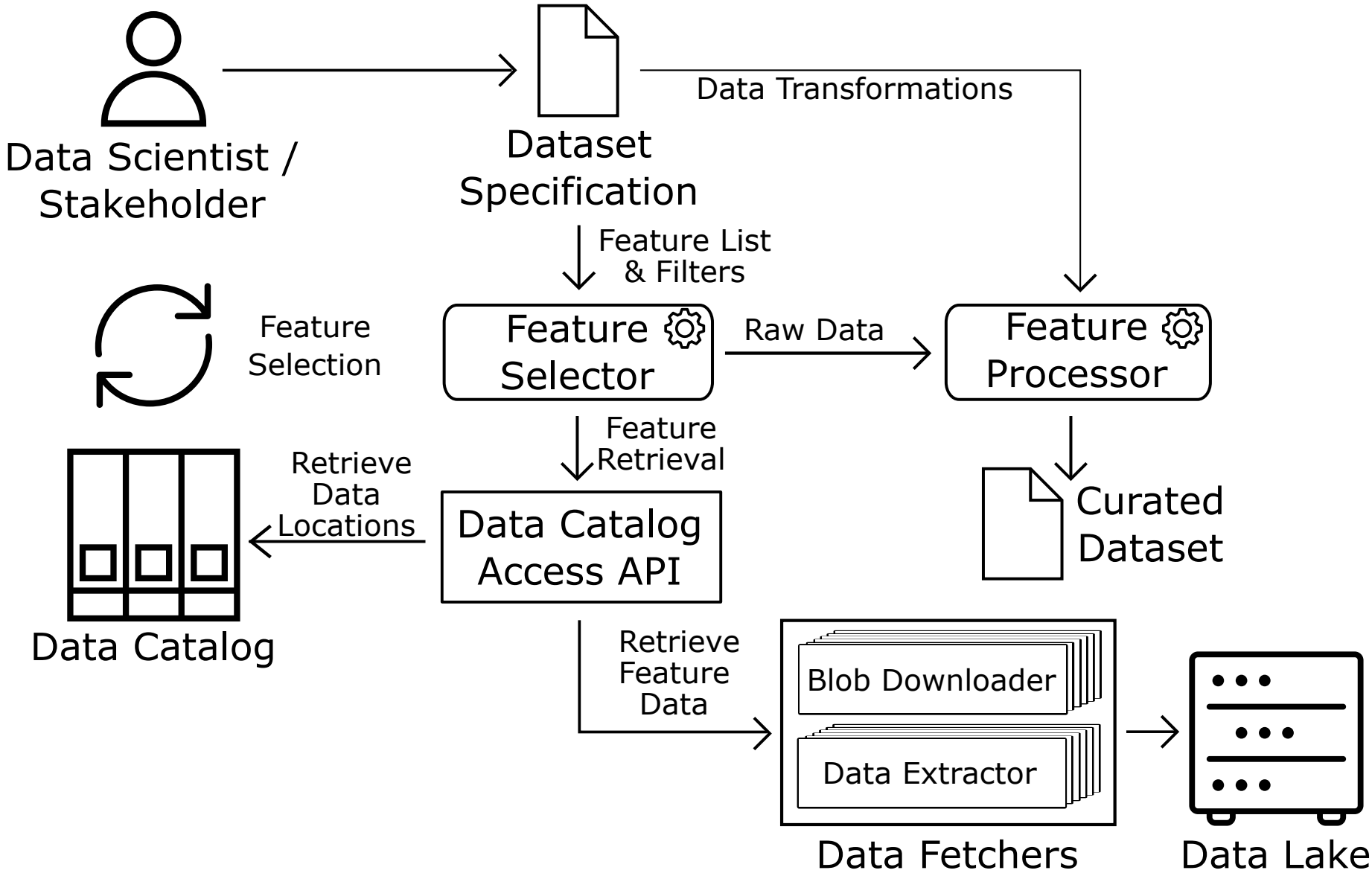
The Hannah Framework



The Hannah Framework



The Hannah Framework



Índice

Current Work Status

1. We have analyzed state-of-the-art and related work.
2. We have sketched a grammar for the Hannah language.
3. Other elements are still future work.

Current Work Status

1. We have analyzed state-of-the-art and related work.
2. We have sketched a grammar for the Hannah language.
3. Other elements are still future work.

Current Work Status

1. We have analyzed state-of-the-art and related work.
2. We have sketched a grammar for the Hannah language.
3. Other elements are still future work.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Summary

1. Many companies are incorporating *data lakes*.
2. Retrieving and transforming data from a data lake to create datasets is time consuming and prone to errors.
3. This work proposes *Hannah*, a high-level declarative language for specifying datasets for time series analysis, plus a framework for automatically processing these specifications.
4. Both of these elements should help to reduce the effort required for the creation of data sets and to involve domain experts in these processes.

Thanks for your attention.

Acknowledgments

- File type icons in slide 7 by SVG Repo.
- Lake and fisherman icons in slide by Scco of SVG Repo.
- Server, Gear Wheels in slide 18 by SVG Repo.
- Cycle Arrows in slide 18 by IconPark of SVG Repo.