Data @ MSR

Johannes Härtel – johannes.hartel@vub.be

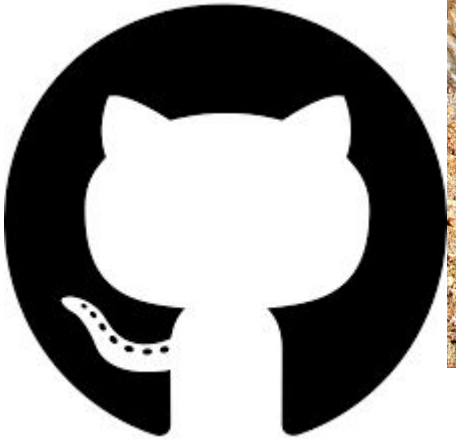
Typical (Research) Questions

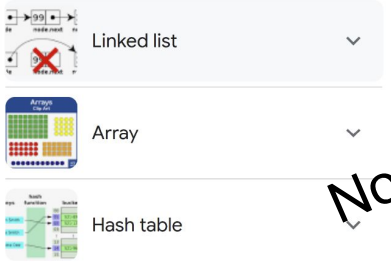
1. What is the average **number** of comments in Java files?
2. What is the Java **file** with the lowest **number** of comments?
3. What is the Java **package** with the lowest **number** of comments?
4. What is the Java **package** with the lowest **fraction** of comments?

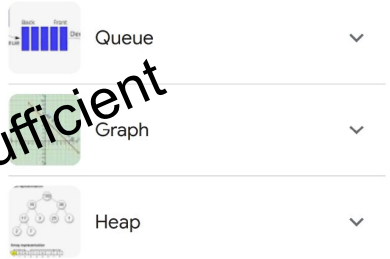
Typical Workflow

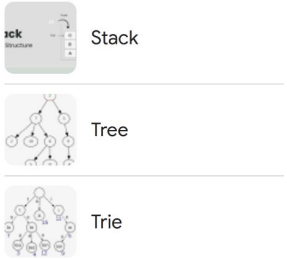
1. Cone repository
2. Extract **data** on the comments.
3. Transform the **data** until it answers your question.

The “**data** structures” used here are somewhat special.









Not sufficient

Johannes Härtel – johannes.hartel@vub.be

Available Libraries:

Scala, Python, Java, R …

*

Python

Python

Python

Python & JS

R

(Conservatively, these are no classical data structures. However, they all focus on representing and processing data.)

Knowing what to use or to avoid

Agenda

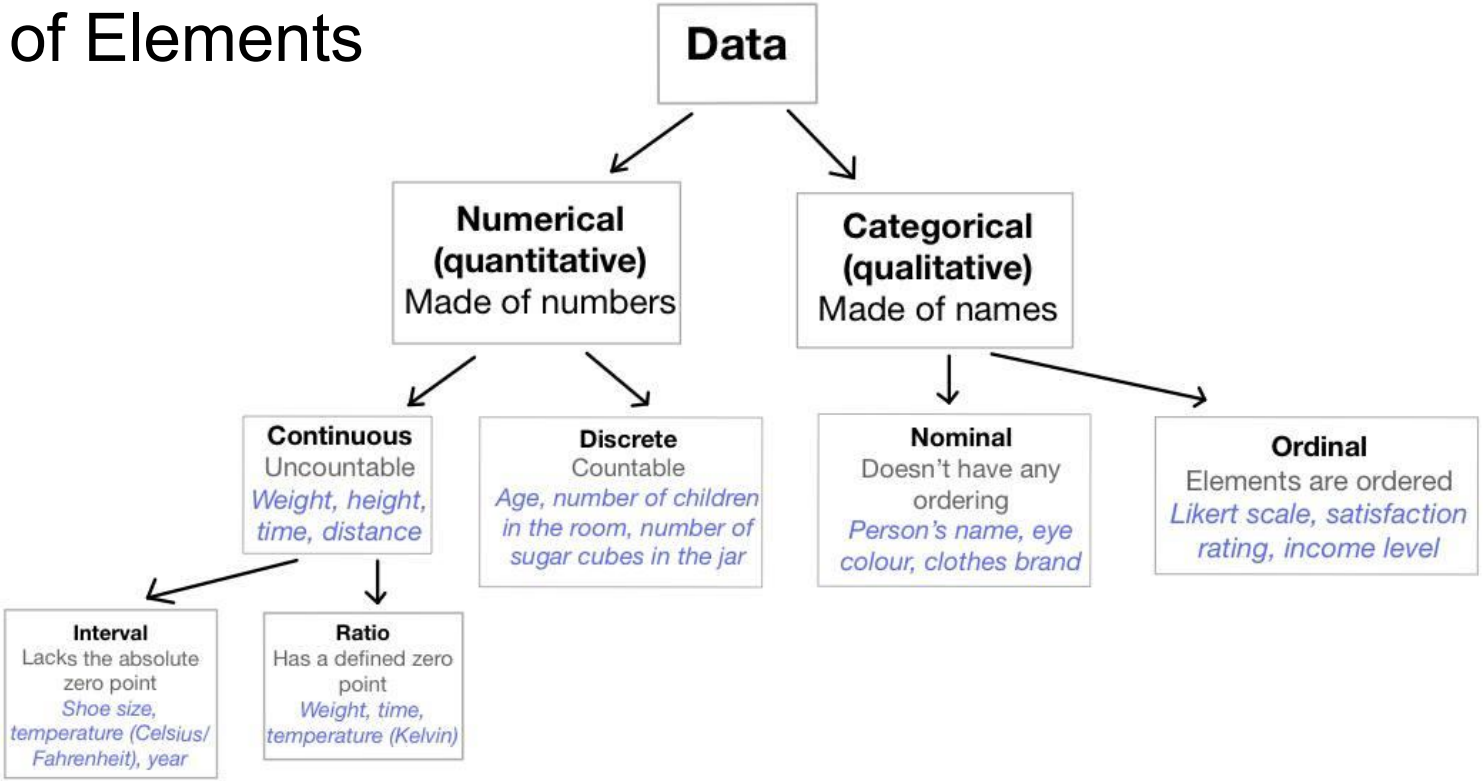
1. Types of Elements
2. Structure of the Collections
3. Functionality
4. Interoperability

Types of Elements

Johannes Härtel – johannes.hartel@vub.be

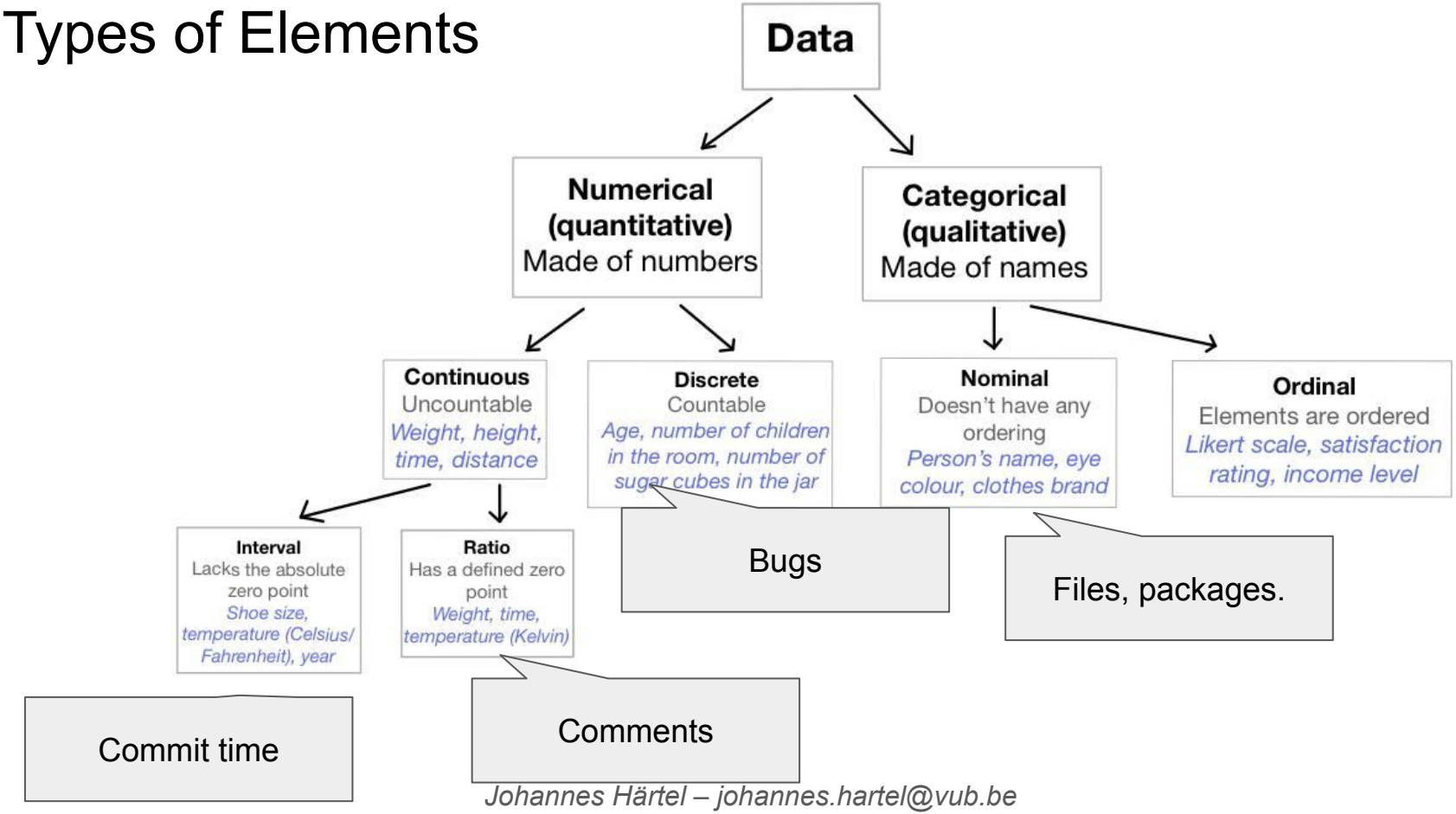
Source: https://hyperskill.org/learn/step/24686

Types of Elements



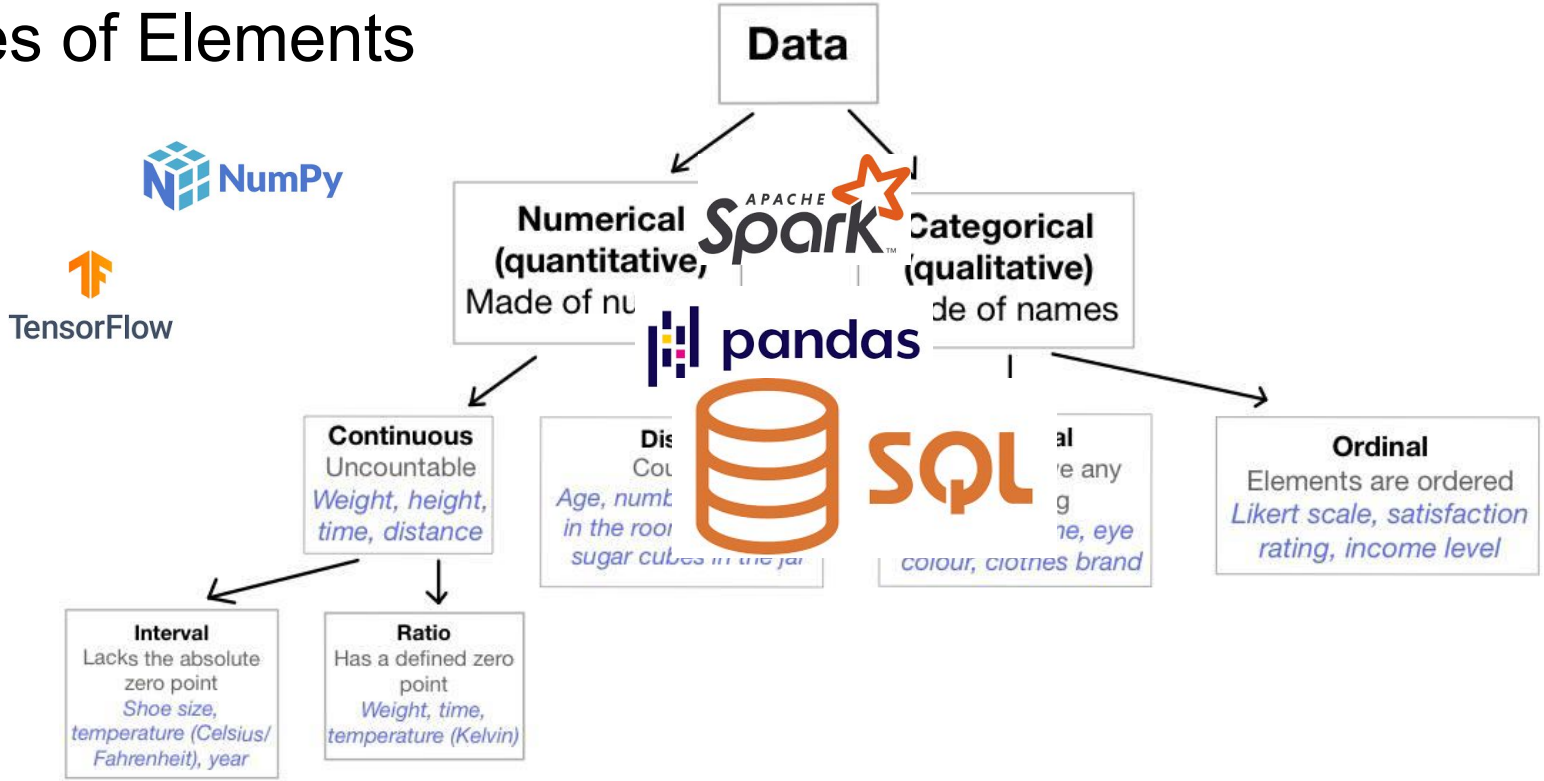
Source: https://hyperskill.org/learn/step/24686

Types of Elements



Source: https://hyperskill.org/learn/step/24686

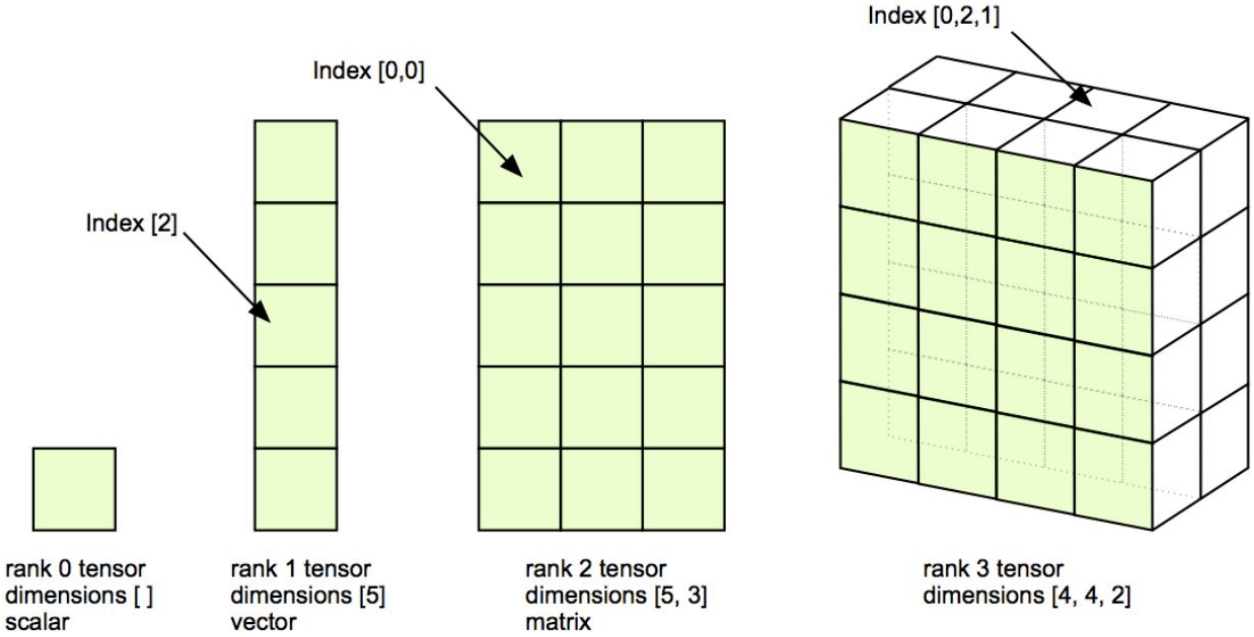
Types of Elements



Structure of the Collections

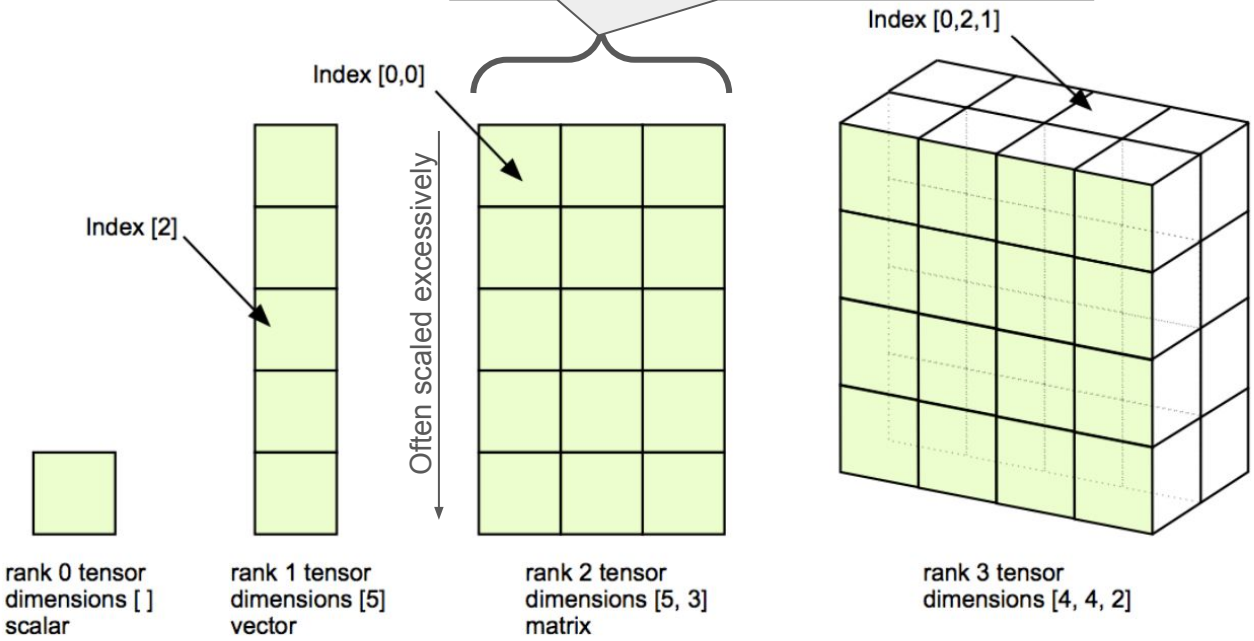
Structure of the Collection

We always have some sort of "**multidimensional array**".

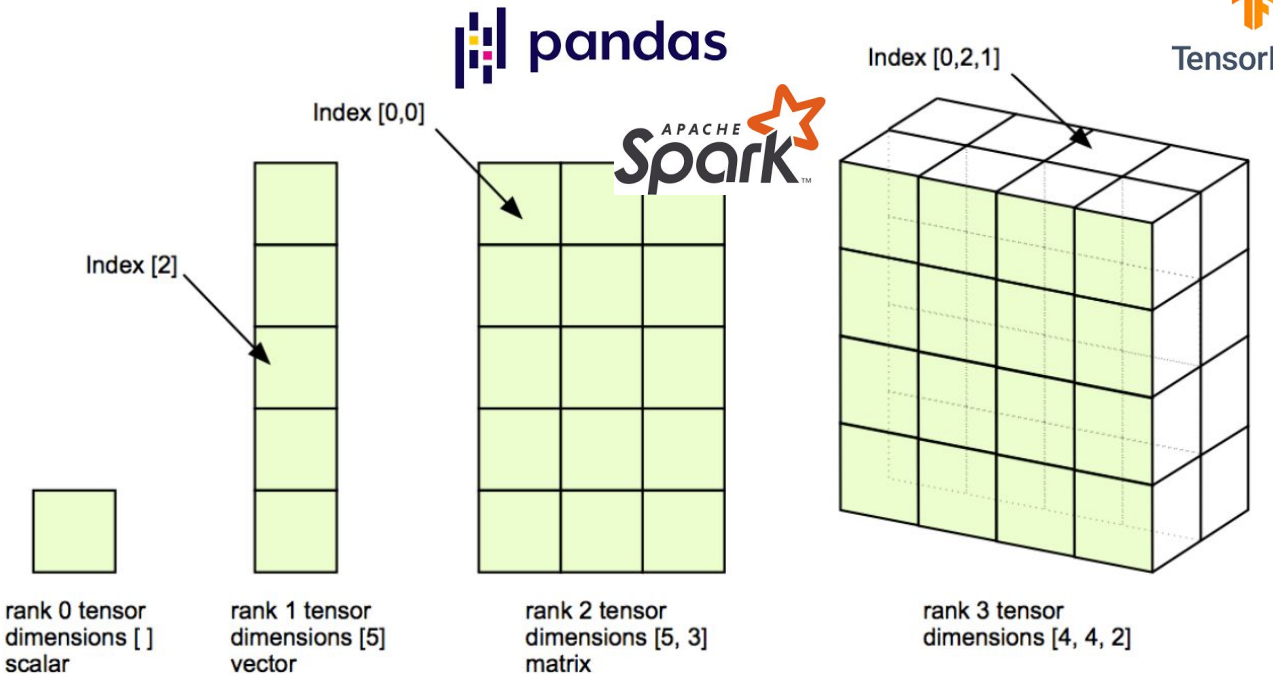


Structure of the Colle

People tend to have **different types** on this dimension. They call such a structure a “table”.



Structure of the Collection



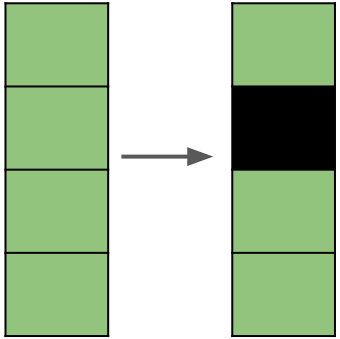
Functionality

Johannes Härtel – johannes.hartel@vub.be

Functionality

(examples)

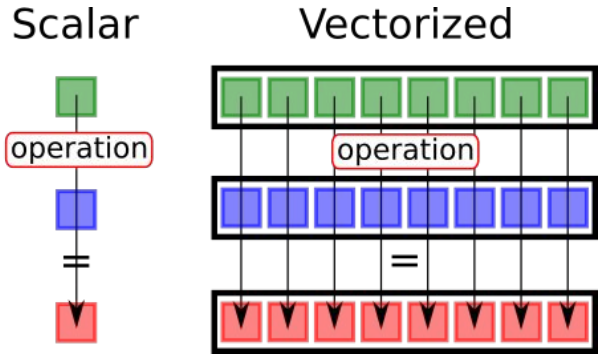
assign



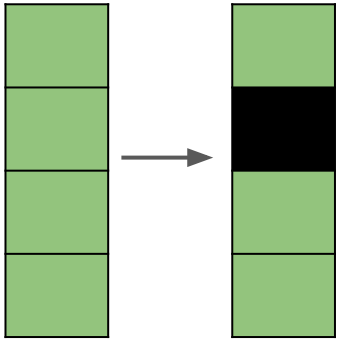
Functionality

(examples)

apply binary op.



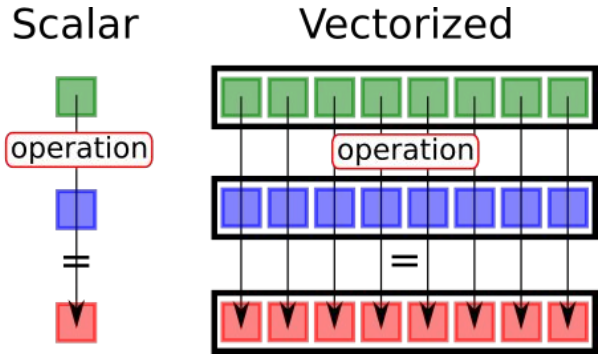
assign



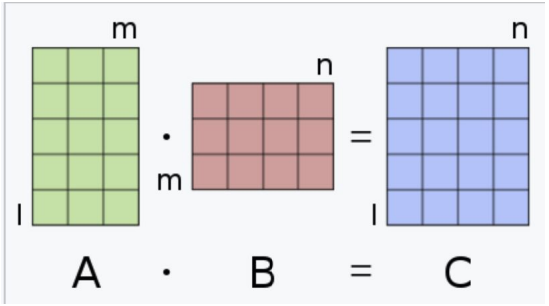
Functionality

(examples)

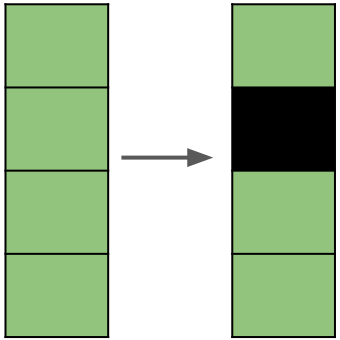
apply binary op.



matrix multiplication



assign



Johannes Härtel – johannes.hartel@vub.be

Functionality

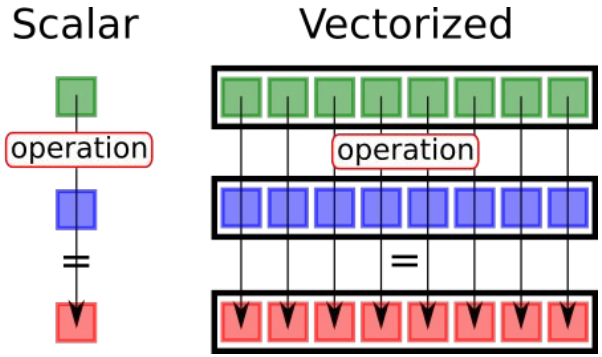
(examples)

group by

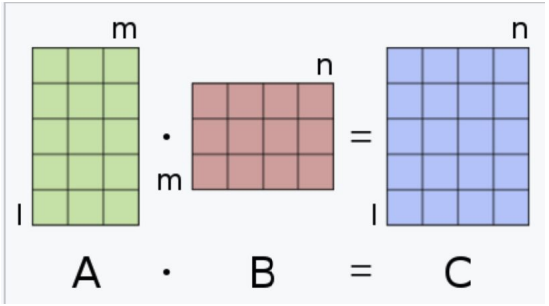
title	genre	price
book 1	adventure	11.90
book 2	fantasy	8.49
book 3	romance	9.99
book 4	adventure	9.99
book 5	fantasy	7.99
book 6	romance	5.88

genre	avg_price	
adventure	(11.90 + 9.99)/2	10.945
fantasy	(8.49 + 7.99)/2	8.24
romance	(9.99 + 5.88)/2	7.935

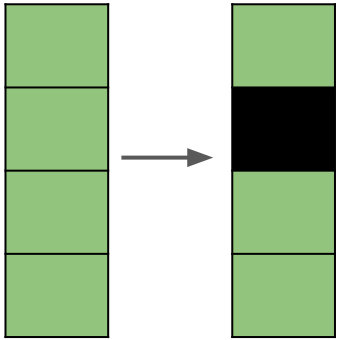
apply binary op.



matrix multiplication



assign



Johannes Härtel – johannes.hartel@vub.be

Functionality

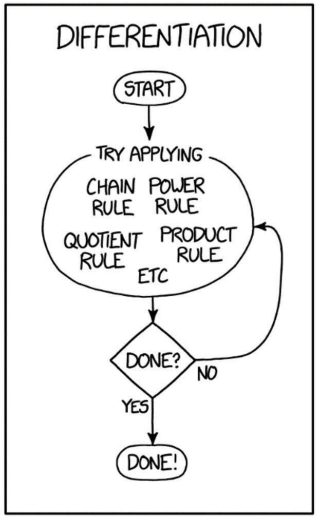
(examples)

group by

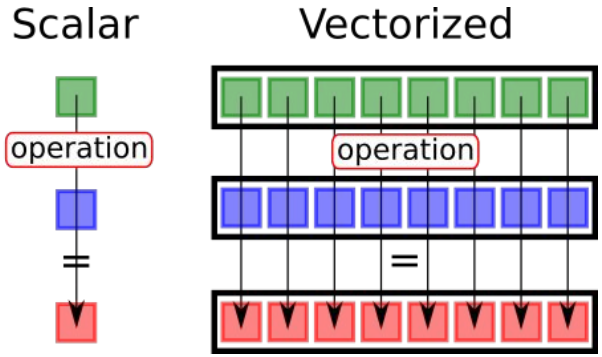
title	genre	price
book 1	adventure	11.90
book 2	fantasy	8.49
book 3	romance	9.99
book 4	adventure	9.99
book 5	fantasy	7.99
book 6	romance	5.88

genre	avg_price	
adventure	(11.90 + 9.99)/2	10.945
fantasy	(8.49 + 7.99)/2	8.24
romance	(9.99 + 5.88)/2	7.935

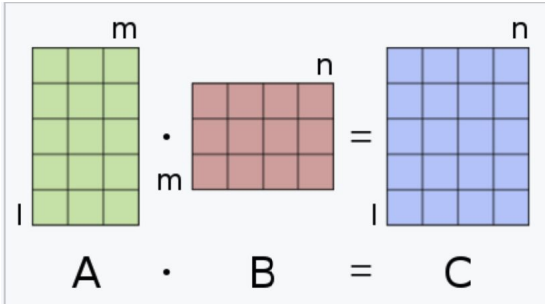
automatic differentiation



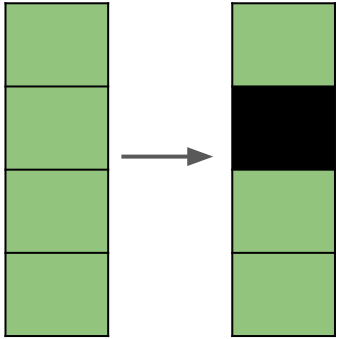
apply binary op.



matrix multiplication



assign



Johannes Härtel – johannes.hartel@vub.be

Execution

Execution

Runs on GPU

Execution

runs native in C++

Runs on GPU

Execution

distributed

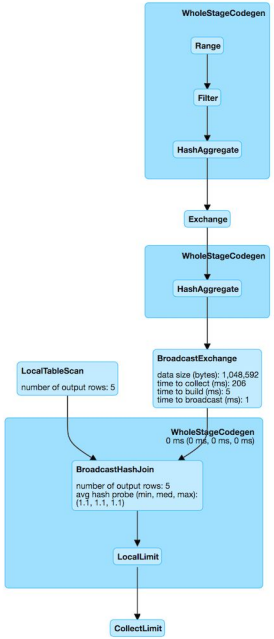
runs native in C++

Runs on GPU

Johannes Härtel – johannes.hartel@vub.be

Execution

data dependency



distributed

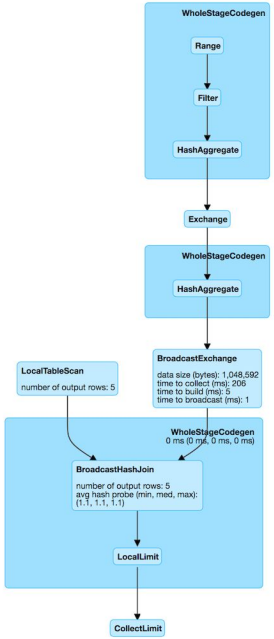
runs native in C++

Runs on GPU

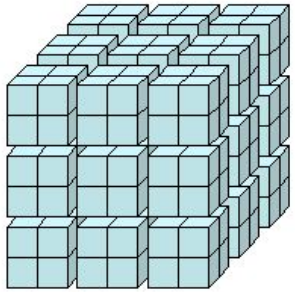
Johannes Härtel – johannes.hartel@vub.be

Execution

data dependency



chunked



distributed

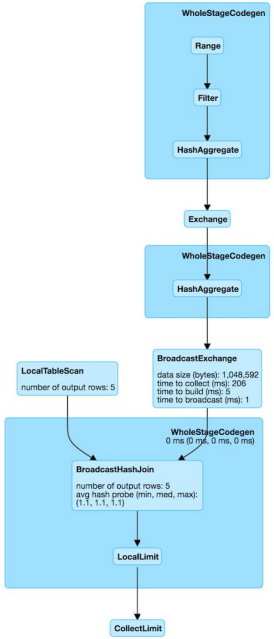
runs native in C++

Runs on GPU

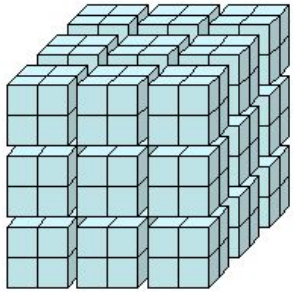
Johannes Härtel – johannes.hartel@vub.be

Execution

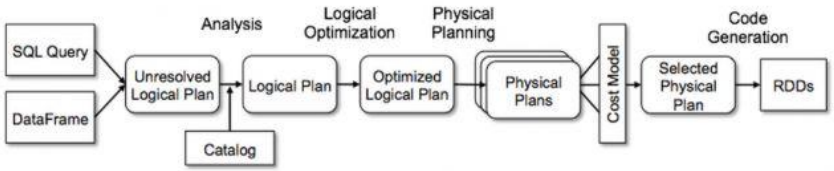
data dependency



chunked



optimized



distributed

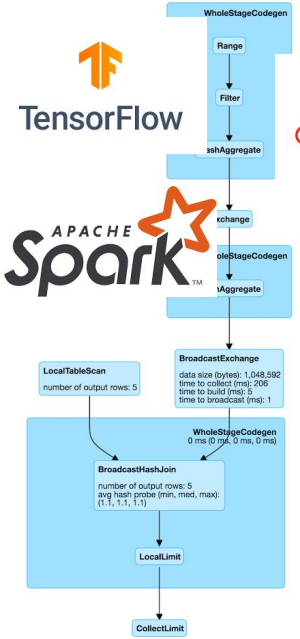
runs native in C++

Runs on GPU

Johannes Härtel – johannes.hartel@vub.be

Execution

data dependency

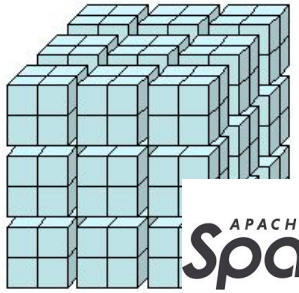


TensorFlow

PyTorch

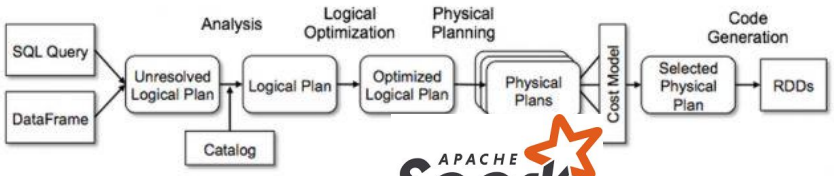
Spark

chunked



Spark

optimized



Spark

distributed

Spark

runs native in C++

pandas

C++

NumPy

Runs on GPU

PyTorch

TensorFlow

Johannes Härtel – johannes.hartel@vub.be

Interoperability

parquet

CSV

Interoperability

parquet

converters 1

csv

pandas.DataFrame.to_numpy

```
DataFrame.to_numpy(dtype=None, copy=False, na_value=_NoDefault.no_default) #
```

Convert the DataFrame to a NumPy array. [source]

By default, the dtype of the returned array will be the common NumPy dtype of all types in the DataFrame. For example, if the dtypes are `float16` and `float32`, the results dtype will be `float32`. This may require copying data and coercing values, which may be expensive.

Interoperability

parquet

converters 1

csv

pandas.DataFrame.to_numpy

`DataFrame.to_numpy(dtype=None, copy=False, na_value=_NoDefault.no_default)` #

Convert the DataFrame to a NumPy array. [source]

By default, the dtype of the returned array will be the common NumPy dtype of all types in the DataFrame. For example, if the dtypes are `float16` and `float32`, the results dtype will be `float32`. This may require copying data and coercing values, which may be expensive.

converters 2

Caution: when constructing a tensor from a numpy array or pandas dataframe the underlying buffer may be re-used:

```
a = np.array([1, 2, 3])
b = tf.constant(a)
a[0] = 4
print(b)  # tf.Tensor([4 2 3], shape=(3,), dtype=int64)
```

Johannes Härtel – johannes.hartel@vub.be

*

Scala, Python, Java, R …

Python

Python

We focus on pandas. The other might be more interesting later.

Python

Python & JS

R

(Conservatively, these are no classical data structures. However, they all focus on representing and processing data.)

Typical (Research) Questions

1. What is the average **number** of comments in Java files?
2. What is the Java **file** with the lowest **number** of comments?
3. What is the Java **package** with the lowest **number** of comments?
4. What is the Java **package** with the lowest **fraction** of comments?

Demo