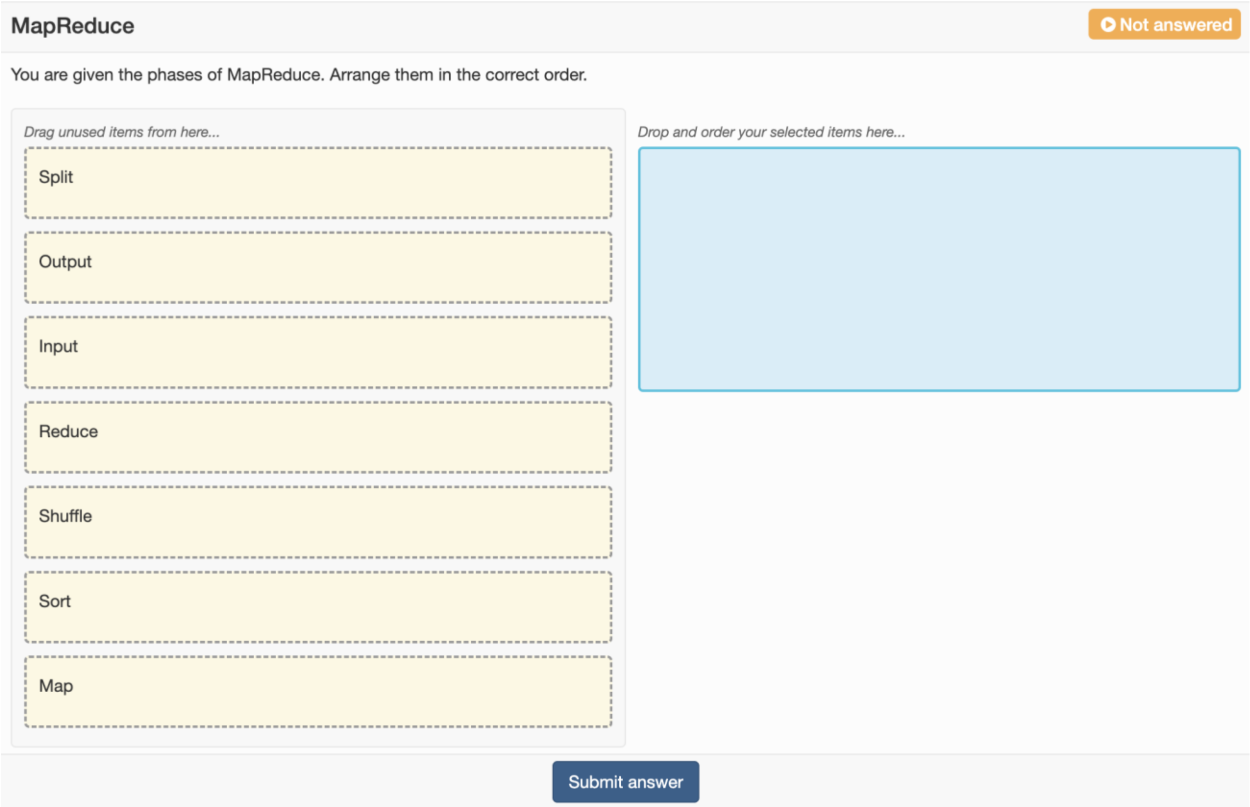
Big Data Tutorial Assignments 4 and 5

Marina Ernst
marinaernst@uni-koblenz.de

Institute for Web Science and Technologies
Universität Koblenz

› Assignment 4

Recall



MapReduce

You are given the phases of MapReduce. Arrange them in the correct order.

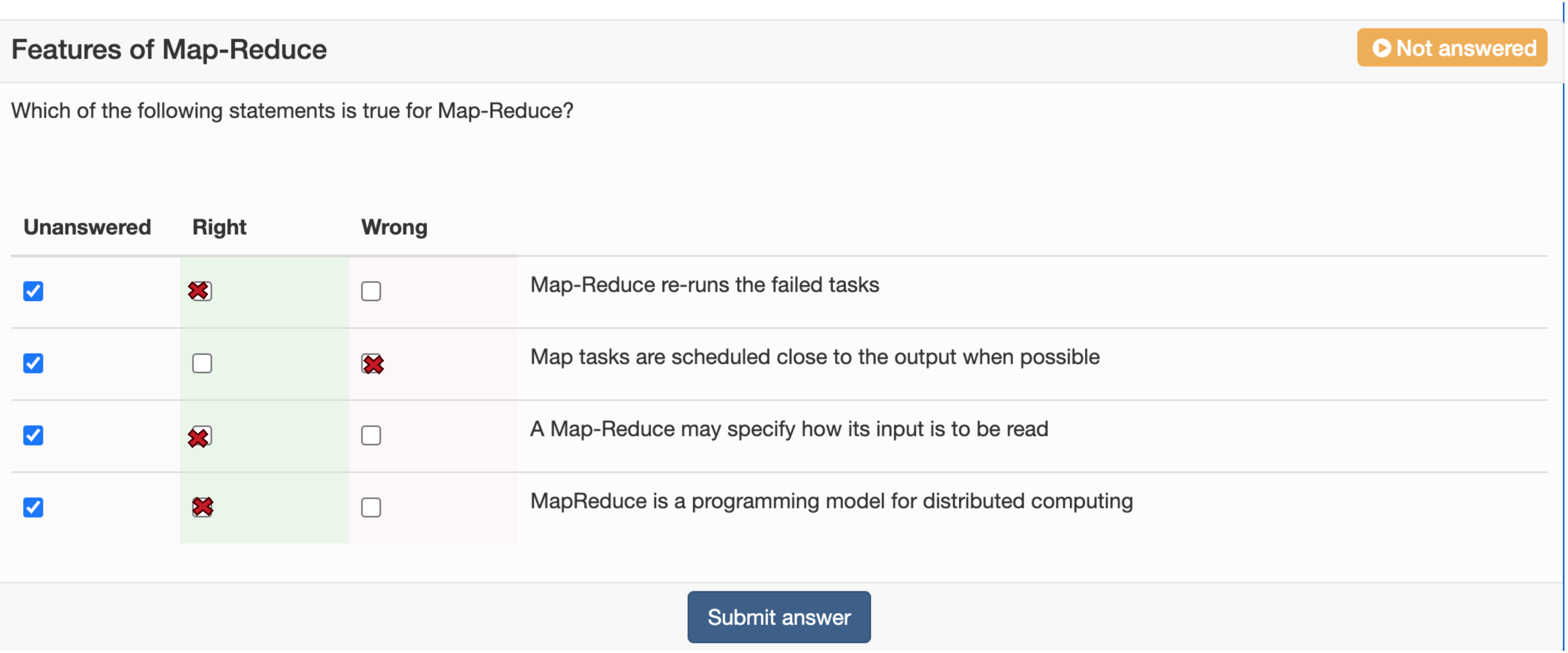
Drag unused items from here...

- Split
- Output
- Input
- Reduce
- Shuffle
- Sort
- Map

Drop and order your selected items here...

Input > Split> Map > Sort > Shuffle > Reduce > Output

Recall



Features of Map-Reduce

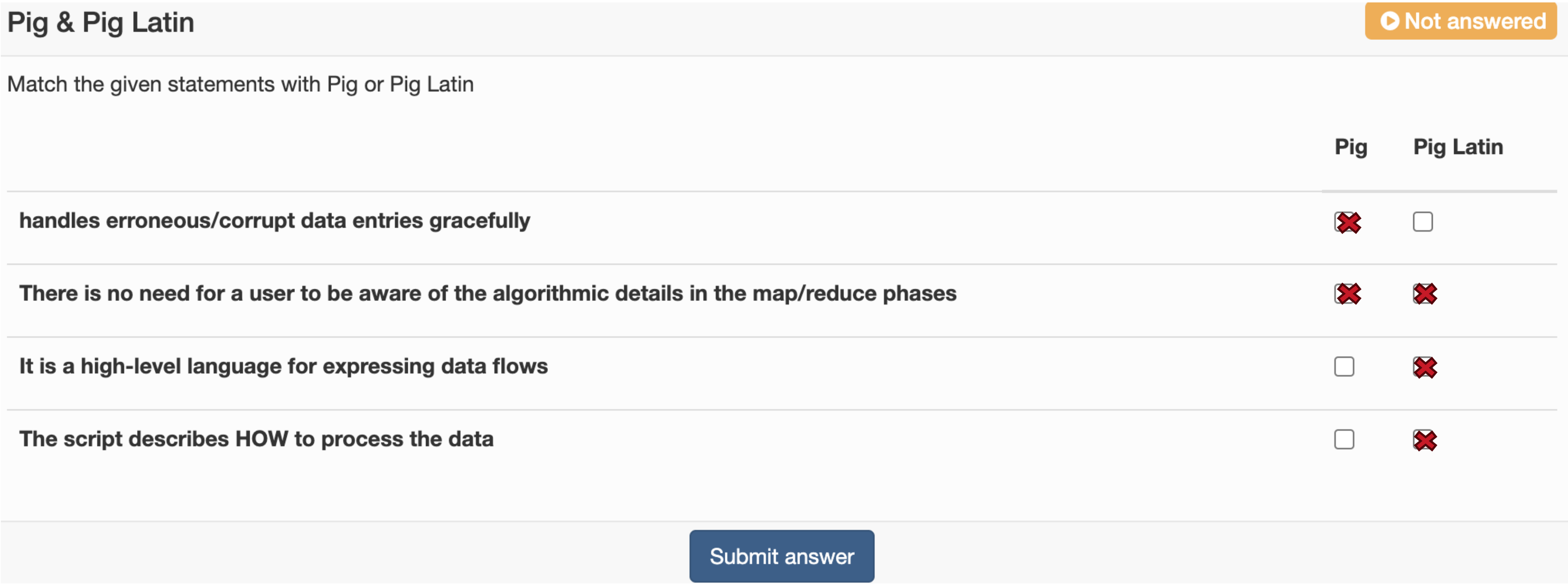
Not answered

Which of the following statements is true for Map-Reduce?

Unanswered	Right	Wrong	
☑	☒	□	Map-Reduce re-runs the failed tasks
☑	□	☒	Map tasks are scheduled close to the output when possible
☑	☒	□	A Map-Reduce may specify how its input is to be read
☑	☒	□	MapReduce is a programming model for distributed computing

Submit answer

Recall



Pig & Pig Latin

Not answered

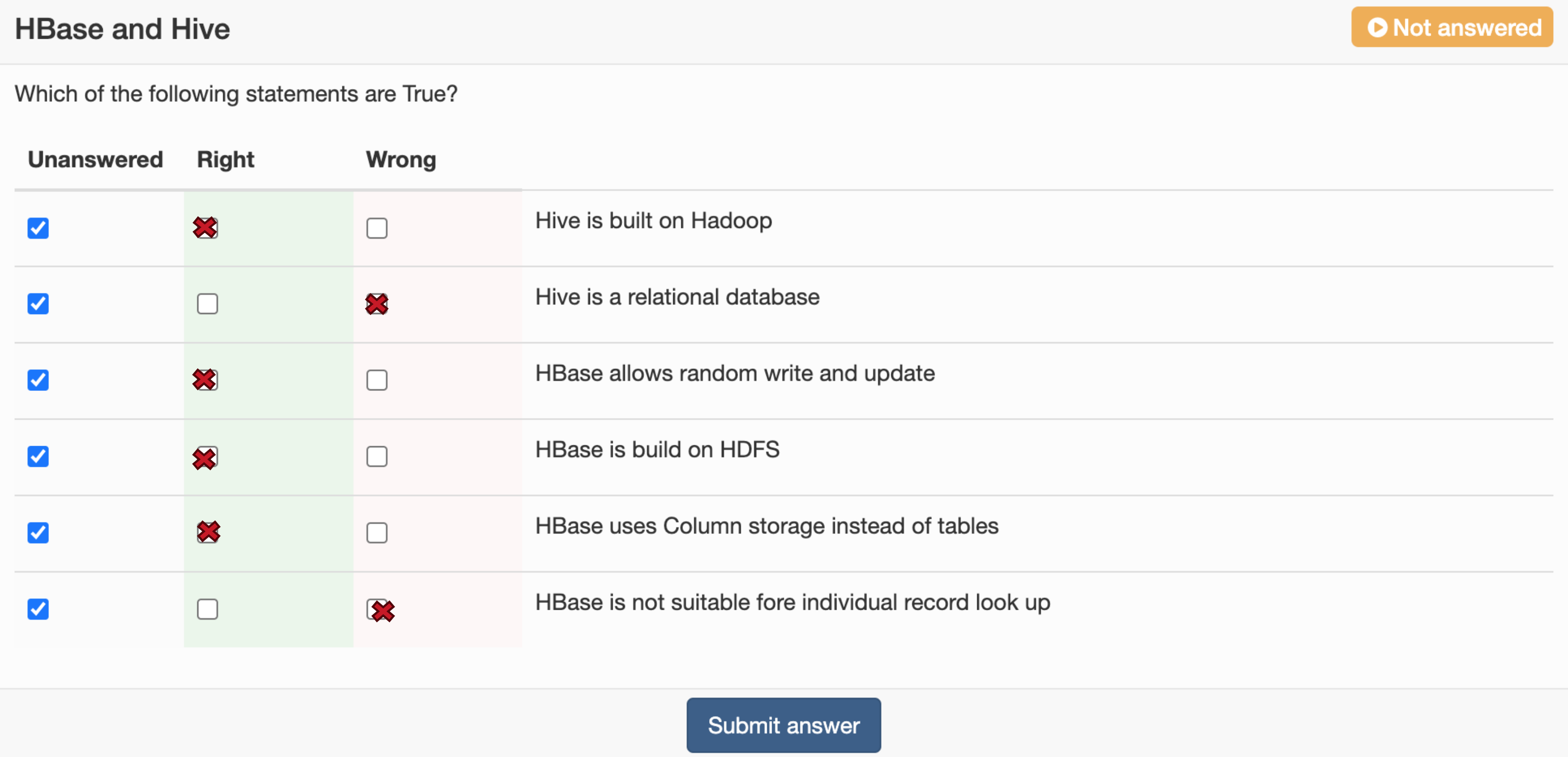
Match the given statements with Pig or Pig Latin

	Pig	Pig Latin
handles erroneous/corrupt data entries gracefully	☒	□
There is no need for a user to be aware of the algorithmic details in the map/reduce phases	☒	☒
It is a high-level language for expressing data flows	□	☒
The script describes HOW to process the data	□	☒

Pig is an interactive, or script-based, execution **environment** supporting **Pig Latin**, a **language** used to express data flows.

https://www.dummies.com/article/technology/information-technology/data-science/big-data/hadoop-pig-and-pig-latin-for-big-data-167026/

Recall



HBase and Hive

Not answered

Which of the following statements are True?

Unanswered	Right	Wrong	
☑	☒	□	Hive is built on Hadoop
☑	□	☒	Hive is a relational database
☑	☒	□	HBase allows random write and update
☑	☒	□	HBase is build on HDFS
☑	☒	□	HBase uses Column storage instead of tables
☑	□	☒	HBase is not suitable fore individual record look up

Knowledge Questions

ML in Big Scale

Imagine the situation when you need to perform k-means clustering on a large chunk of traffic data. To do that, you employ scikit-learn, and it works initially. Then more records are added to your data, and you are facing time-out.

What can you do to overcome this problem? Explain your solution.

Option 1: Use the ML in Spark including sckitlearn

Knowledge Questions

Apache Mahout

Explain why Mahout was rebuilt on top of Samsara.

Solution is based on the Section 3 of Apache Mahout: Machine Learning on Distributed Dataflow Systems.
Those are 2 main reasons that should be mentioned in one way or another:

1. The MapReduce paradigm was suboptimal for the distributed execution of ML algorithms, both for reasons of usability and performance.
2. The available programming abstractions typically rely on partitioned, unordered bags; this is a mismatch for ML applications that mostly operate on tensors, matrices and vectors.

Knowledge Questions

The Future of Hadoop

Based what you already know about Hadoop and MapReduce what is your opinion on the future of those frameworks? Which potential challenges are there for Hadoop MapReduce?

Use further reading and external sources to support your point of view. (Don't forget the citation). When you talk about challenges, pay attention to the "MapReduce: an infrastructure review and research insights" paper from further reading. You do not have to copy all the challenges from there - you may use some of them to support your opinion.

Knowledge Questions

The Future of Hadoop

Possible Scenario 1:

MapReduce and Hadoop are going to be used less.
Statistically Hadoop is loosing it's popularity (example - google trends)
Possible reasons why:

- Fast-growing Cloud Vendors and Services (better suited for modern BI and ML problems)
- Limited support of Hadoop
- Complexity of the ecosystem

Knowledge Questions

The Future of Hadoop

Possible Scenario 2:

It would stay afloat for quite some time. the momentum was just shifted to SPARK. The many companies still using it (Amazon, Wall Street financial trading companies. etc)

Knowledge Questions

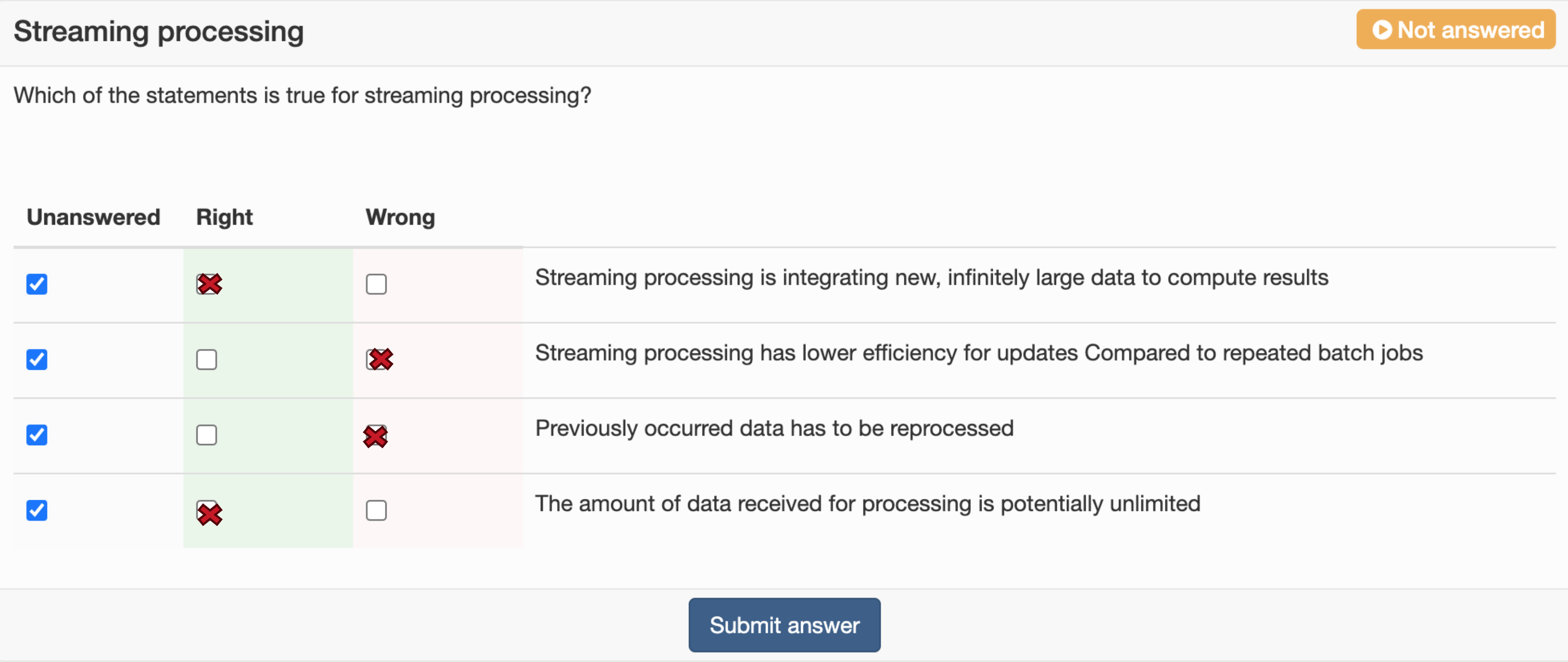
The Future of Hadoop

Hadoop and MapReduce challenges:

1. Improving Performance
2. Decoupling and keeping the scalability
3. Network overhead
4. Appropriate parameter setting
5. …

› Assignment 5

Recall



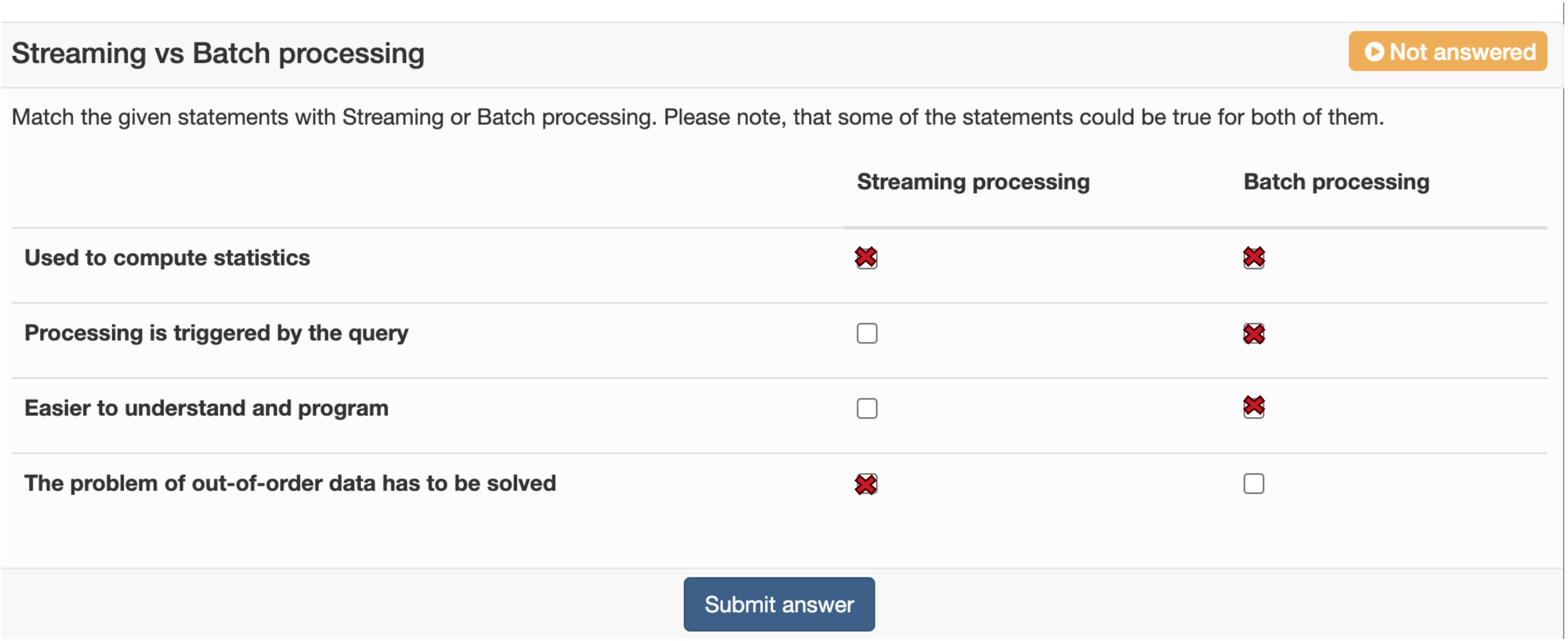
Streaming processing

▶ Not answered

Which of the statements is true for streaming processing?

Unanswered	Right	Wrong	
☑	☒	□	Streaming processing is integrating new, infinitely large data to compute results
☑	□	☒	Streaming processing has lower efficiency for updates Compared to repeated batch jobs
☑	□	☒	Previously occurred data has to be reprocessed
☑	☒	□	The amount of data received for processing is potentially unlimited

Recall



Streaming vs Batch processing

Match the given statements with Streaming or Batch processing. Please note, that some of the statements could be true for both of them.

	Streaming processing	Batch processing
Used to compute statistics	☒	☒
Processing is triggered by the query	☐	☒
Easier to understand and program	☐	☒
The problem of out-of-order data has to be solved	☒	☐

Recall

Sliding window

In your own words, explain what a sliding window of data is.

Why use Sliding Window?

1. Keep track of the reception and order of the data.
2. Identify duplicated or missing information

What is Sliding Window?

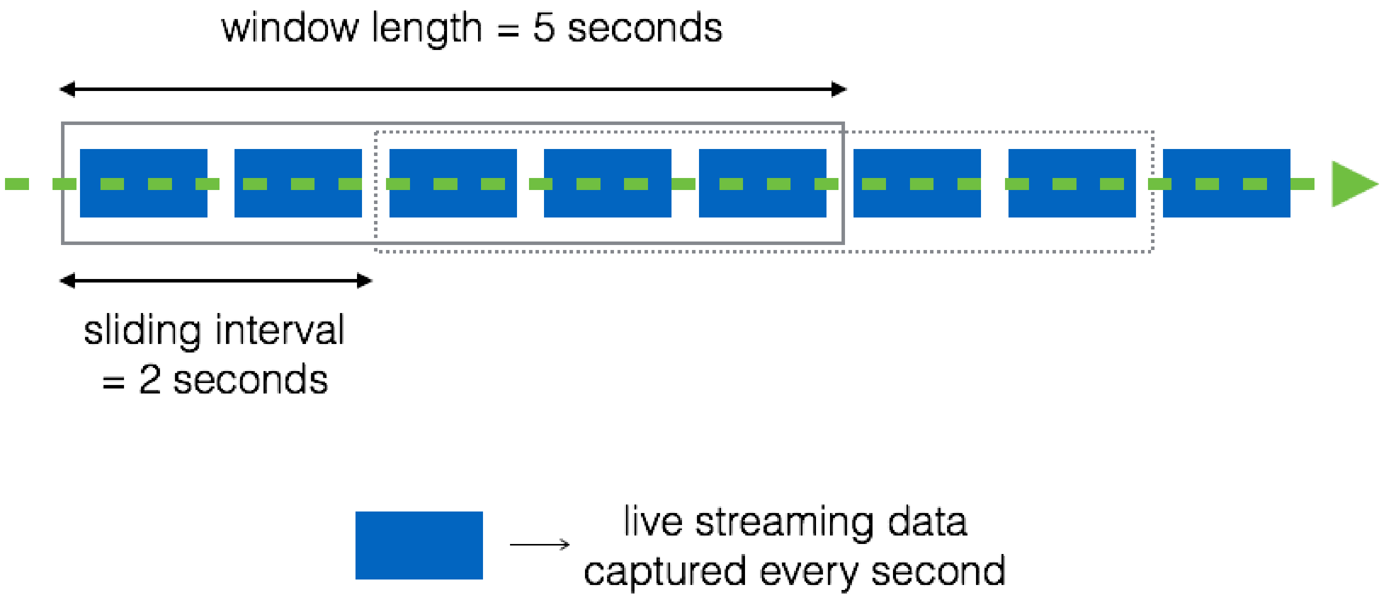
Windowing is one of the most frequently used processing methods for streams of data. An unbounded stream of data (events) is split into finite sets, or windows , based on specified criteria, such as time.

A sliding window is an overlapping window. A sliding window is defined with a window interval and a sliding offset.

Recall

Sliding window

In your own words, explain what a sliding window of data is.



https://prateekvjoshi.files.wordpress.com/2015/11/2-windowed-processing.png

Knowledge Questions

Streaming Data Processing Example

Using external sources give an example of a use case for streaming data processing.

Explain the reason why streaming is used there and which approach (from the discussion in the lecture) is used. Which technology is employed (eg. Apache Storm, Apache Flink, etc.)? Which streaming data sources are used (Kafka, kinesis, etc)?
Do not repeat the examples from the lecture.

Knowledge Questions

Streaming Data Processing Example

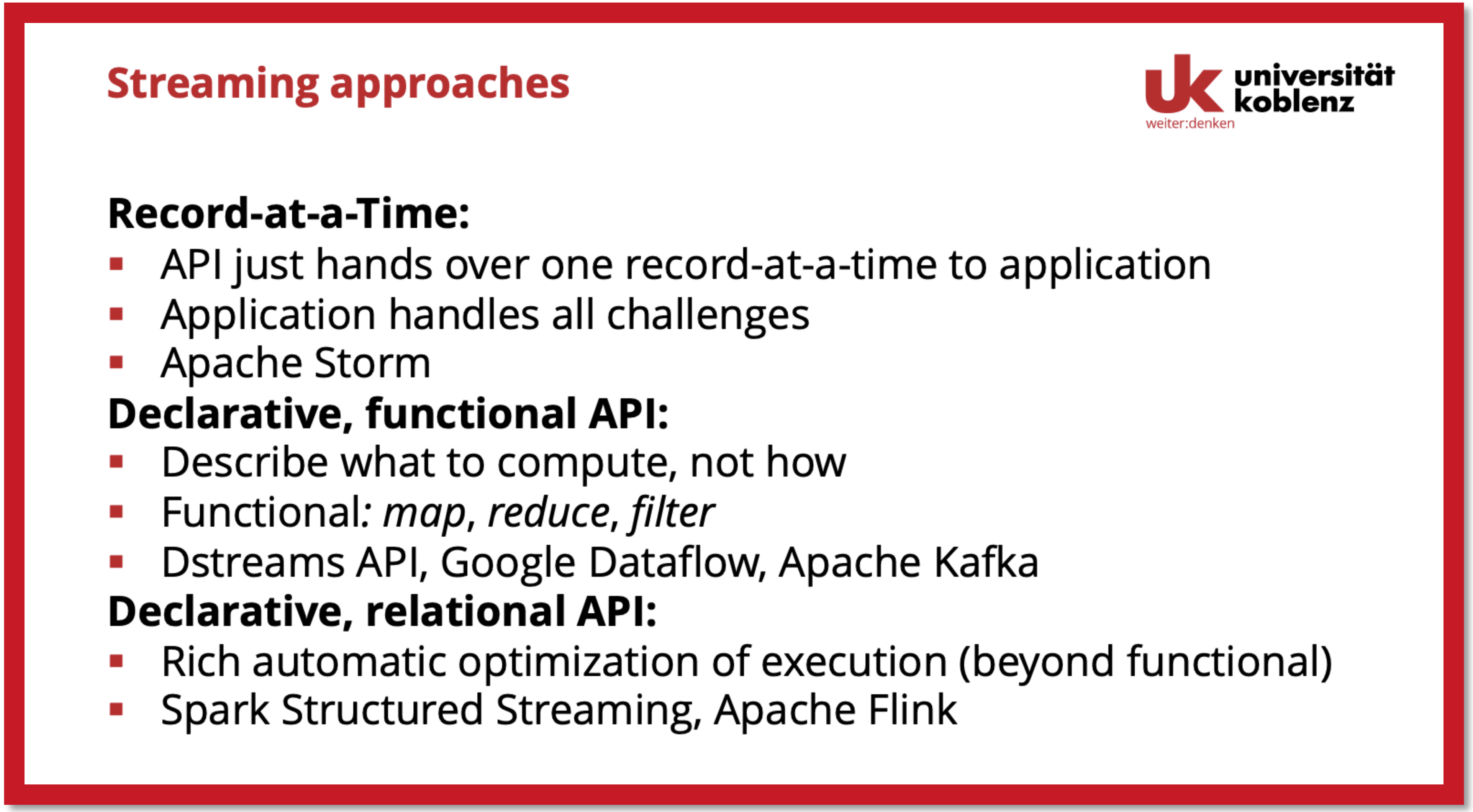
To continue its business operations Zalando is equipped with Saiki which is a platform for data integration and distribution. Saiki helps with the data ingestion and distributes the data for analytical system to access. They needed a framework for real time data processing.

Approach:
Zalando's event logs were registered on **Apache Kafka**. Then Saiki used Amazon S3 for storing and from there it is made available to analytical systems. The usage of Data Lake provides security, cost reduction and it is then accessed by Oracle and other systems.

Technology: To assess and choose better candidate for stream processing, the company created Proof of Concepts (POC) for **Apache Spark and Flink**. They have requirement for high performance and low latency and after their tests they concluded Flink is more suitable than Spark. Another factor for choosing Flink was their developer community. Saiki found the developer community of Flink is eager to improve the product and really fix bugs reported by users.

Source:
https://engineering.zalando.com/posts/2016/03/apache-showdown-flink-vs.-spark.html

Knowledge Questions



Streaming approaches

Record-at-a-Time:

- API just hands over one record-at-a-time to application
- Application handles all challenges
- Apache Storm

Declarative, functional API:

- Describe what to compute, not how
- Functional: *map*, *reduce*, *filter*
- Dstreams API, Google Dataflow, Apache Kafka

Declarative, relational API:

- Rich automatic optimization of execution (beyond functional)
- Spark Structured Streaming, Apache Flink

Knowledge Questions

Open issues in smart grid

Now your task is to give and explain one issue with a Big Data application that is specific to the smart grid.
Please use the further reading materials to fullfill this task.

Answer:

- Lack of standard data format for the information software and data base structures
- Interoperability of different information and communication systems deployed in the smart grids
- Isolated storage of data in various systems
- Most smart grid generated data are confidential or related with privacy issues
- Lack of strategic vision
- Complexity requiring muti-disciplinary research and development

Knowledge Questions

Stream Processing Frameworks

Study the further reading material on Stream Processing Frameworks.
Based on that research, explain when to use Flink and when Structured Streaming.
Why?

Apache Flink: for simple stateless operations such as ingest and parse (low latency), for joining phase, when the latency and throughput are of equal importance.

Structured Streaming: when throughput the key measure, it offers a high-level, intuitive API that gives very high throughput with minimal tuning.

Q&A

That's all, folks! Happy coding!