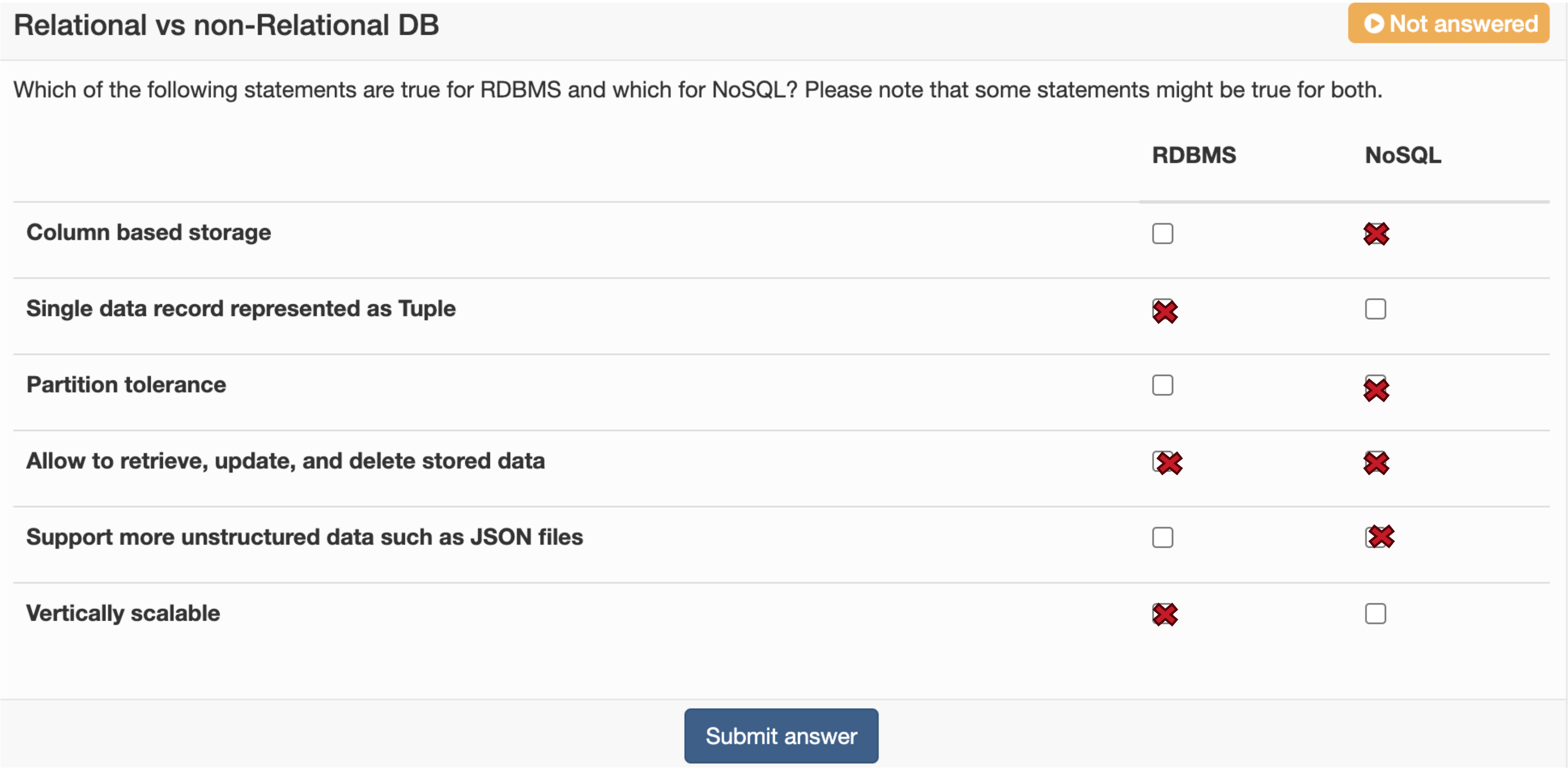
Big Data Tutorial Assignment 3

Marina Ernst
marinaernst@uni-koblenz.de

Institute for Web Science and Technologies
Universität Koblenz

› Assignment 3

Recall

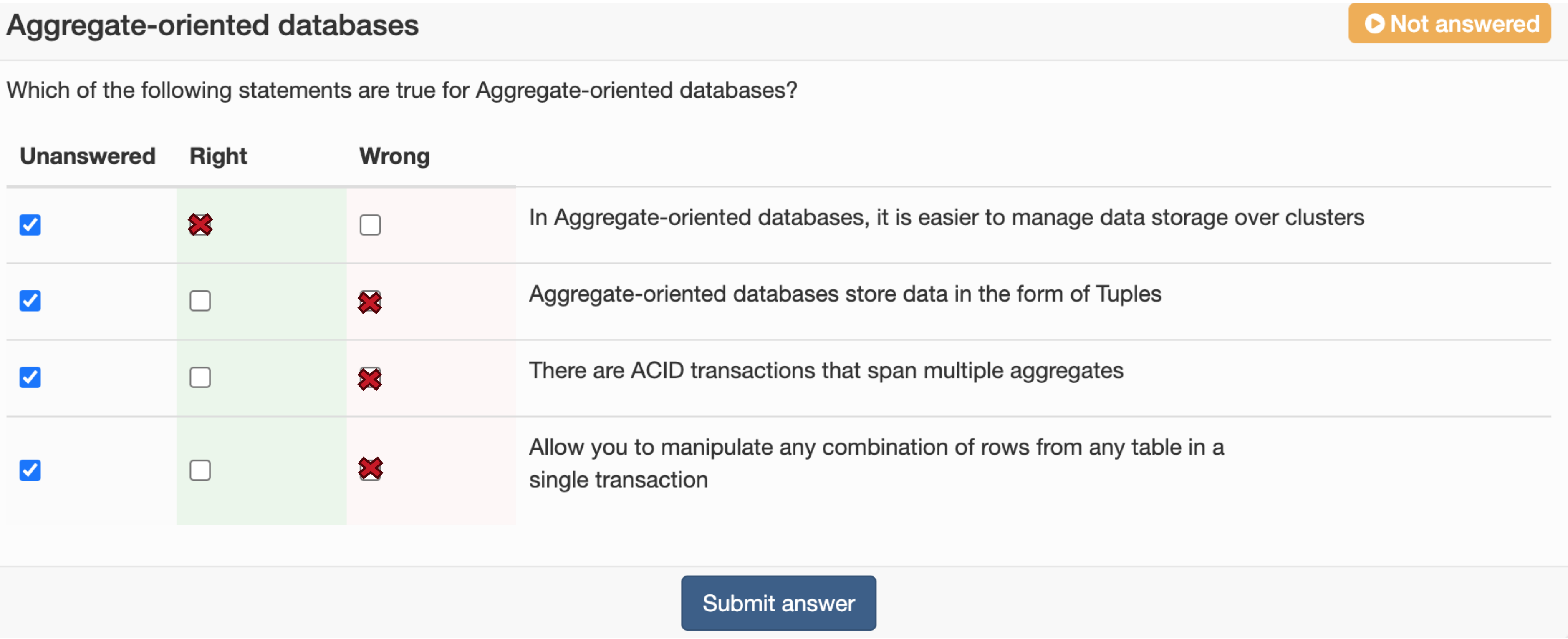


Relational vs non-Relational DB

Which of the following statements are true for RDBMS and which for NoSQL? Please note that some statements might be true for both.

	RDBMS	NoSQL
Column based storage	□	☒
Single data record represented as Tuple	☒	□
Partition tolerance	□	☒
Allow to retrieve, update, and delete stored data	☒	☒
Support more unstructured data such as JSON files	□	☒
Vertically scalable	☒	□

Recall

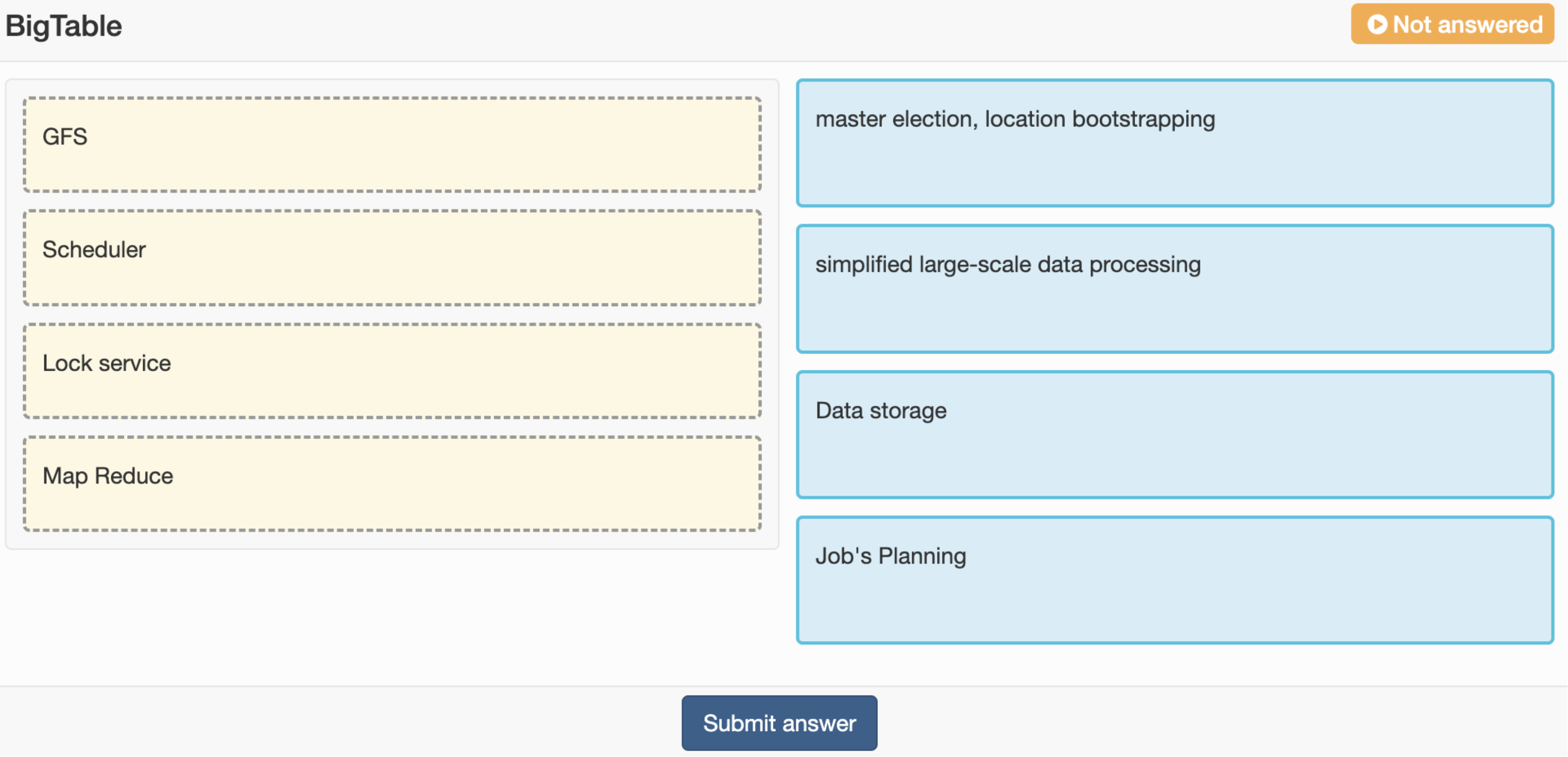


Aggregate-oriented databases

Which of the following statements are true for Aggregate-oriented databases?

Unanswered	Right	Wrong	
☑	☒	□	In Aggregate-oriented databases, it is easier to manage data storage over clusters
☑	□	☒	Aggregate-oriented databases store data in the form of Tuples
☑	□	☒	There are ACID transactions that span multiple aggregates
☑	□	☒	Allow you to manipulate any combination of rows from any table in a single transaction

Recall



BigTable

GFS

Scheduler

Lock service

Map Reduce

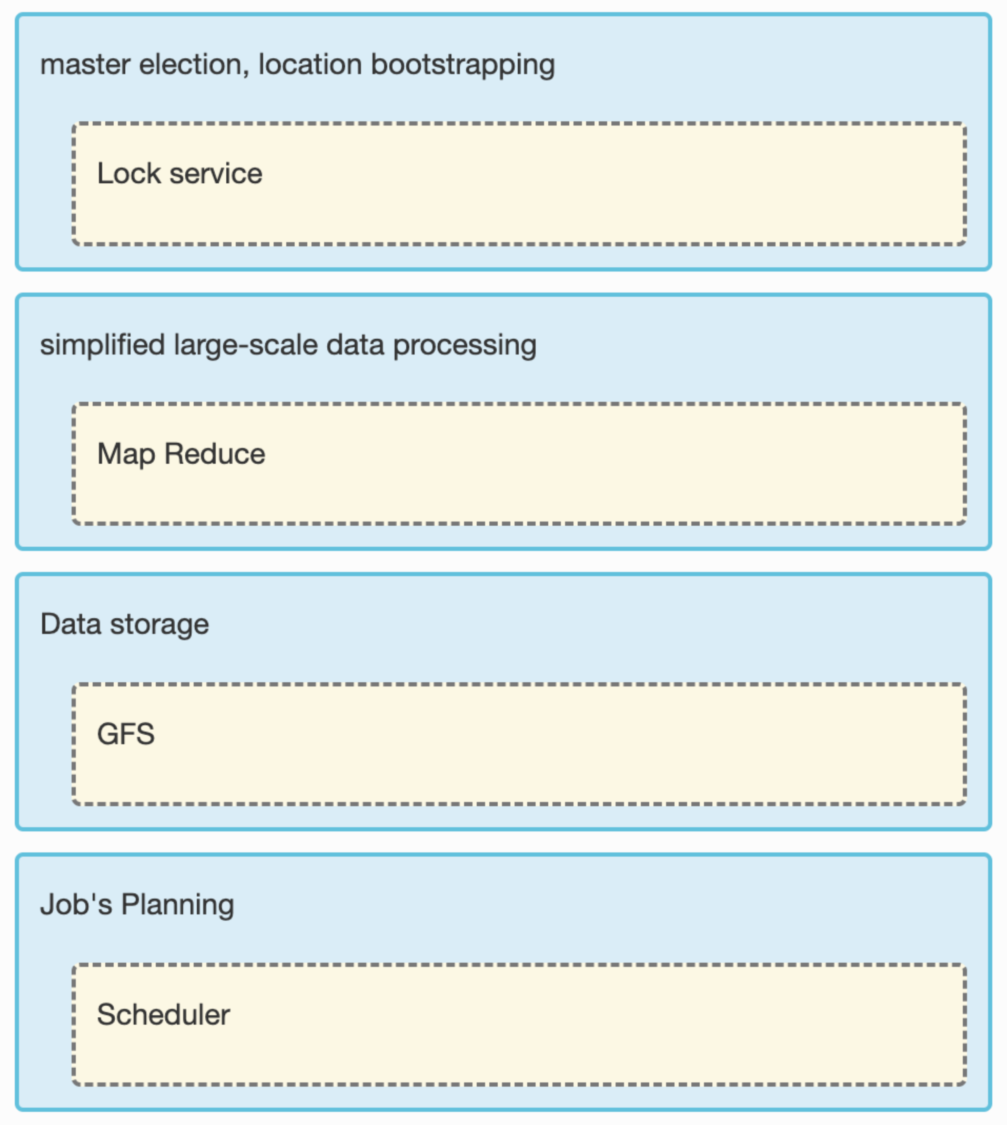
master election, location bootstrapping

simplified large-scale data processing

Data storage

Job's Planning

Recall



master election, location bootstrapping

Lock service

simplified large-scale data processing

Map Reduce

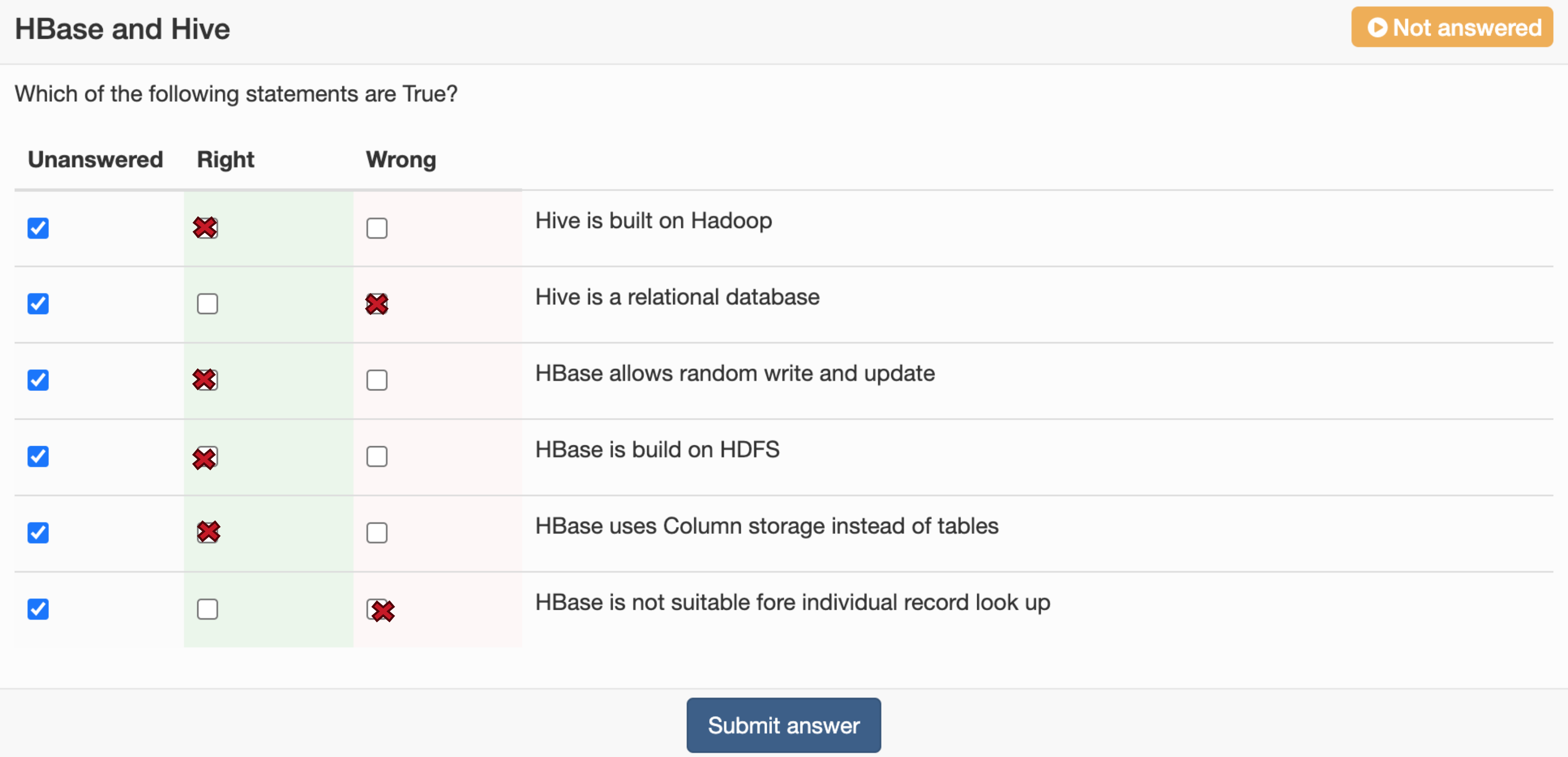
Data storage

GFS

Job's Planning

Scheduler

Recall



HBase and Hive

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

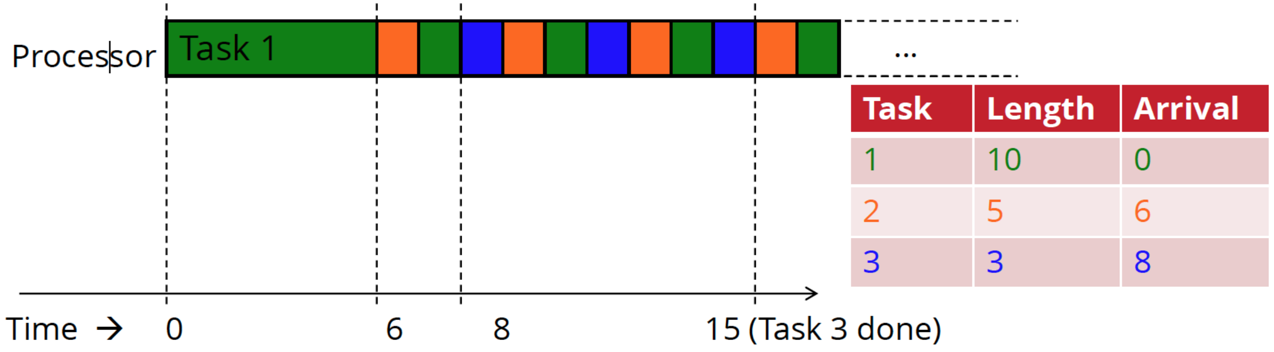
Recall

Round-Robin Scheduling

In your own words, explain how Round-Robin Scheduling works and when to use it.

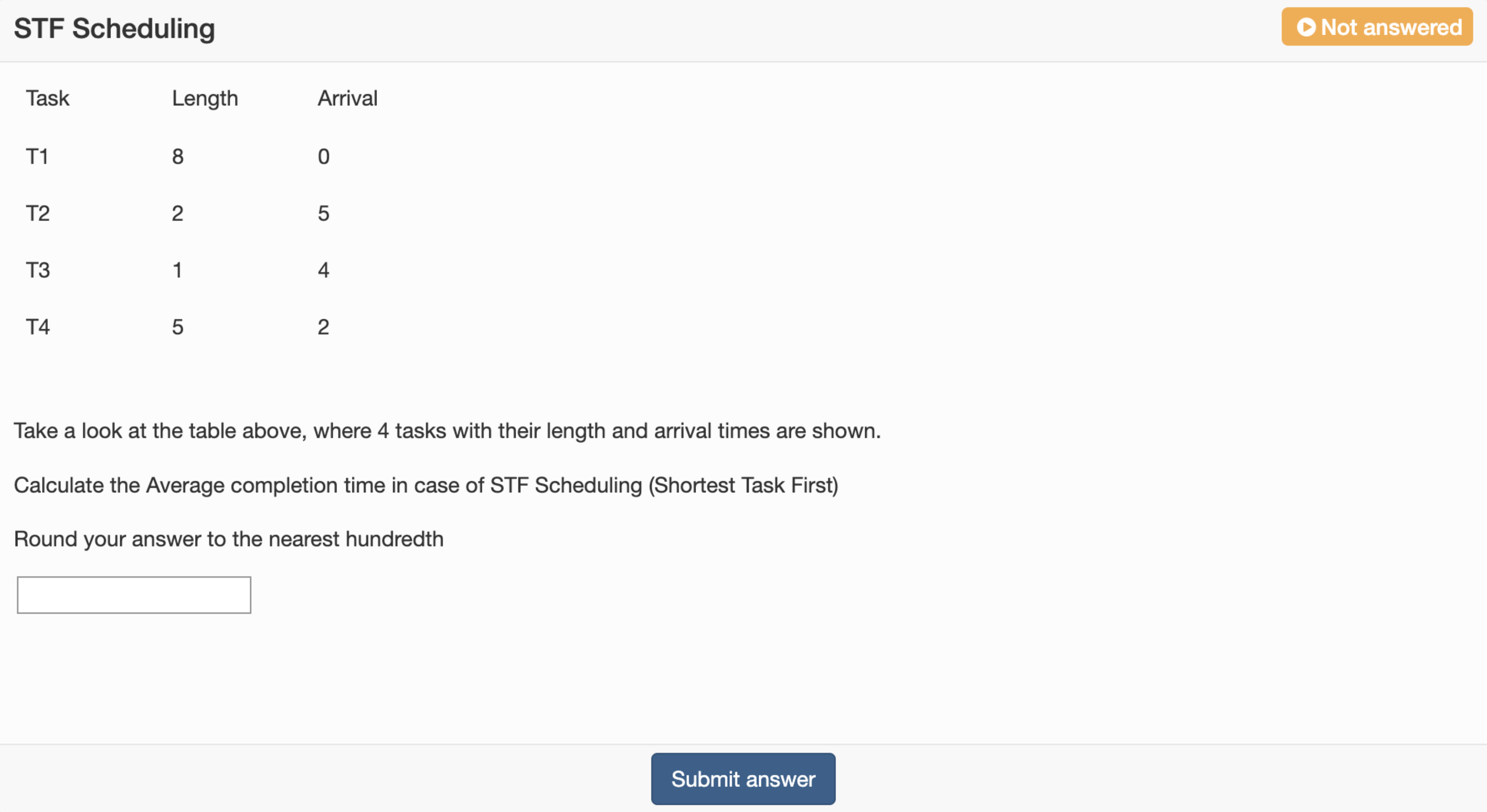
Each task gets an equal share of the CPU time.

Preferable for Interactive applications and when user needs quick responses from system



Task	Length	Arrival
1	10	0
2	5	6
3	3	8

Recall



STF Scheduling

Task	Length	Arrival
T1	8	0
T2	2	5
T3	1	4
T4	5	2

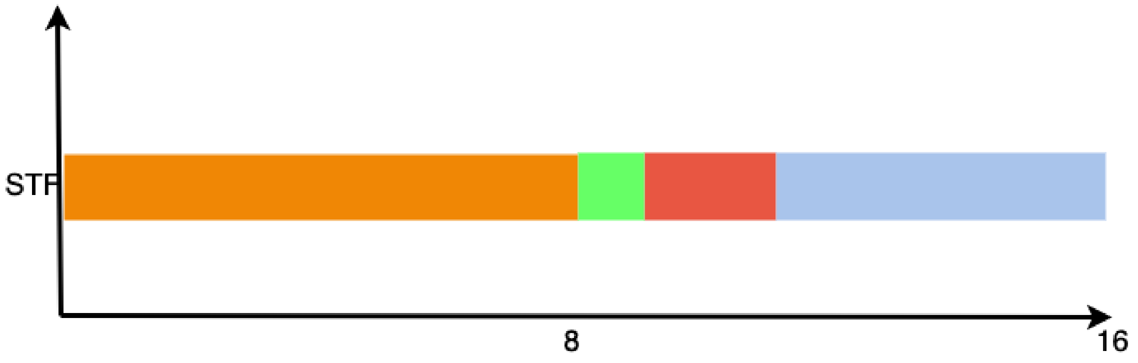
Take a look at the table above, where 4 tasks with their length and arrival times are shown.

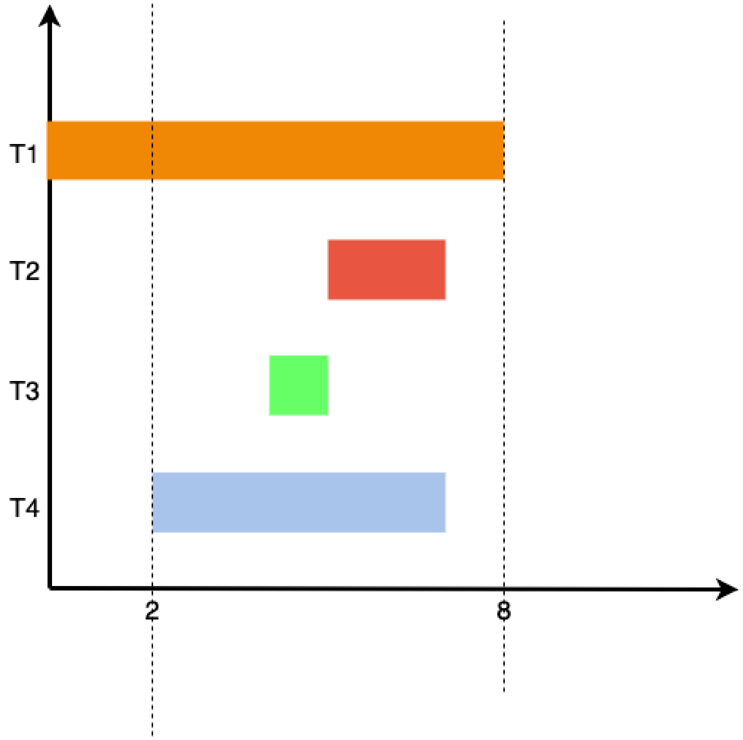
Calculate the Average completion time in case of STF Scheduling (Shortest Task First)

Round your answer to the nearest hundredth

Recall

Task	Length	Arrival
T1	8	0
T2	2	5
T3	1	4
T4	5	2





Average completion time (T1+T2+T3+T4)/4

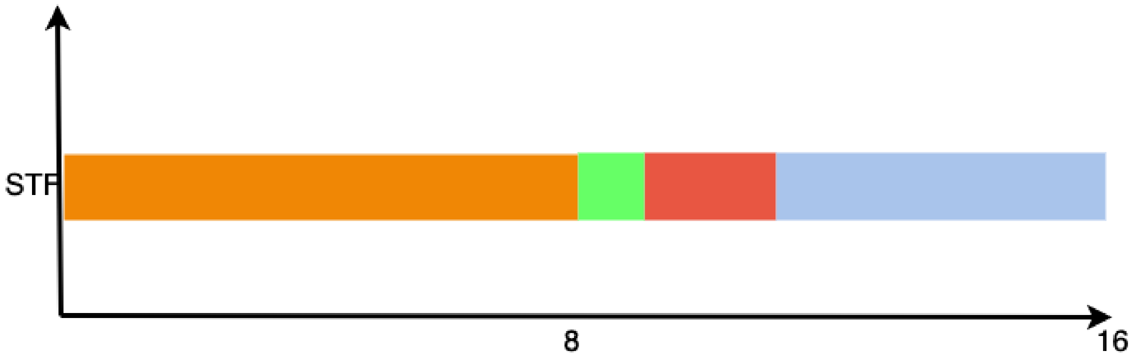
completion time = processing time + total waiting time =

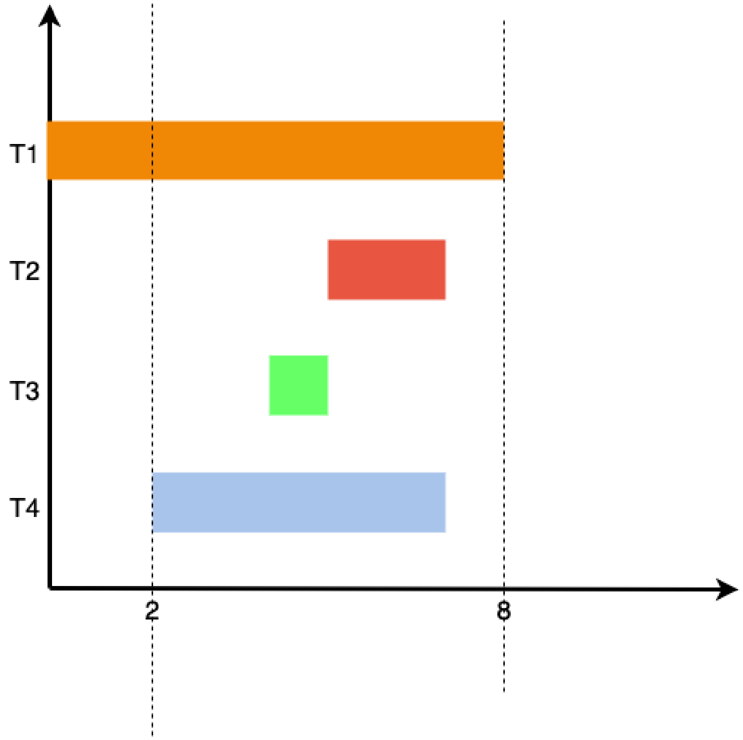
T1=8
T3=9
T2=11
T4=16

(8+9+11+16)/4=44/4=11

Recall

Task	Length	Arrival
T1	8	0
T2	2	5
T3	1	4
T4	5	2





Turnaround time = completion time – arrival time

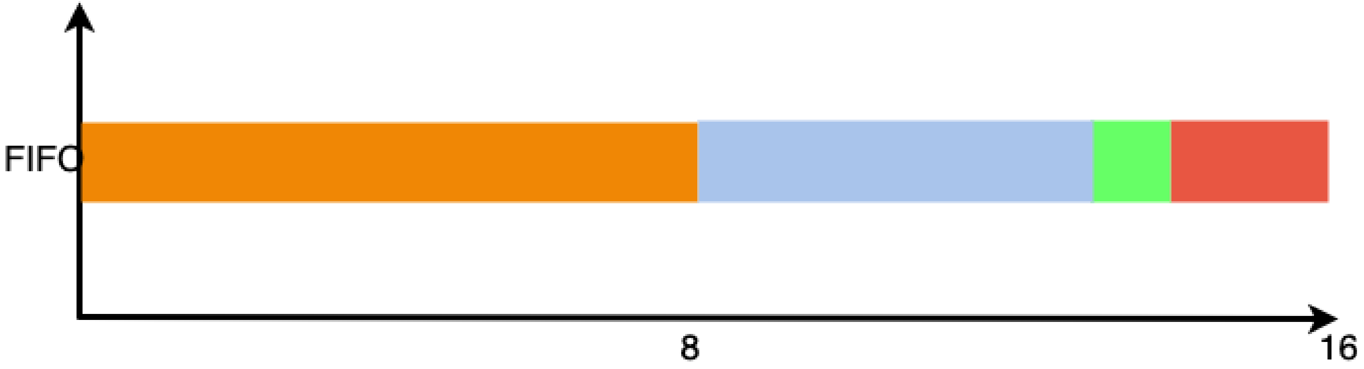
Time between a moment a request is made and when the process is complete

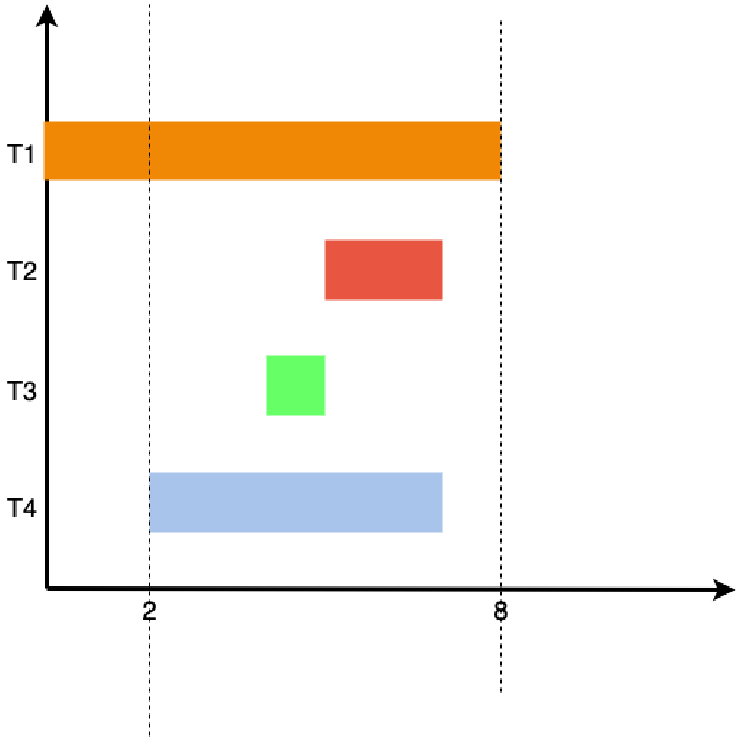
T1=8-0=8
T3=9-4=5
T2=11-5=6
T4=16-2=14

(8+5+6+14)/4=33/4=8.25

Recall

Task	Length	Arrival
T1	8	0
T2	2	5
T3	1	4
T4	5	2





Let's calculate FIFO avg. compilation time
Average completion time (T1+T2+T3+T4)/4
completion time = finishing time – arrival time

T1=8
T4=13
T3=14
T2=16

(8+13+14+16)/4=12.75

Knowledge Questions

Bloom filter

In the paper dedicate to HBase application on Facebook, the Bloom filter is mentioned. Explain what is the Bloom filter and how it works in general. Use the further reading materials and external souses.

A Bloom filter is a bit array. There are also *k* different hash functions, each of which maps a set element to 1 of the bit positions.

- To add an element, feed it to the hash functions to get *k* bit positions, and set the bits at these positions to 1.
- To test if an element is in the set, feed it to the hash functions to get *k* bit positions.
- If any of the bits at these positions is 0, the element is NOT in the set.
- If all are 1, then the element **may be** in the set.

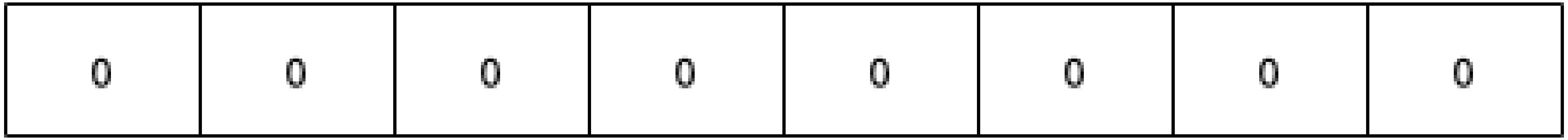
https://yourbasic.org/algorithms/bloom-filter/

Knowledge Questions

Bloom filter

We have a bloom filter with 2 hash functions (H1 and H2). The filter helps us find out if the food an allergen or not. 2 foods are given as allergen: orange and strawberry.

H1(orange)=0
H2(orange)=2
H1(strawberry)=2
H2(strawberry)=6



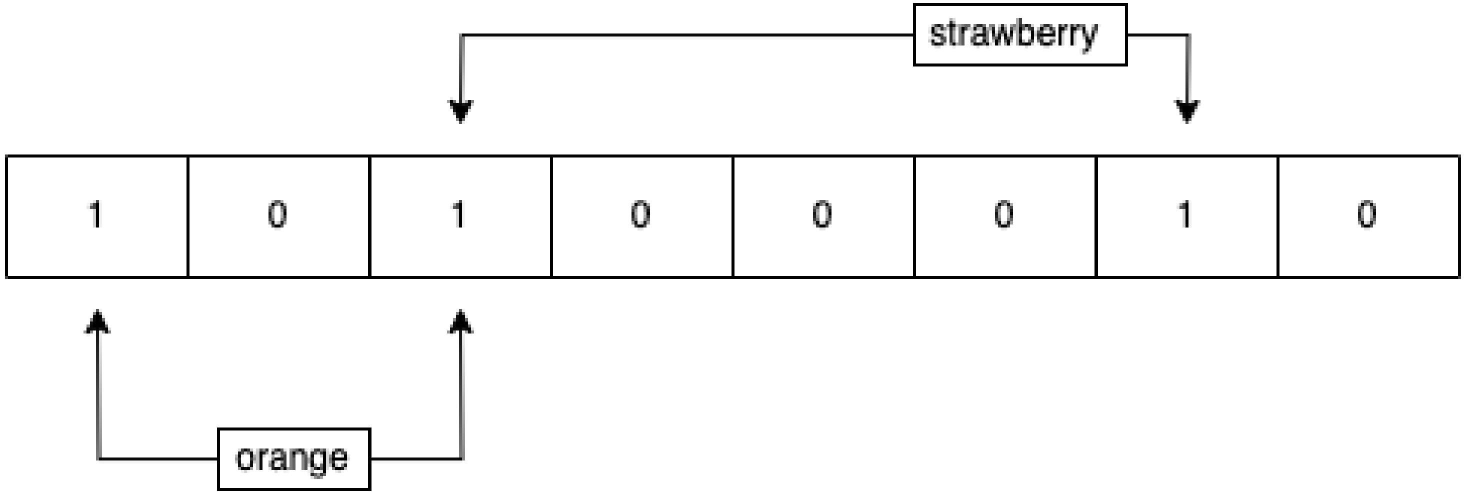
0	0	0	0	0	0	0	0

Knowledge Questions

Bloom filter

We have a bloom filter with 2 hash functions (H1 and H2). The filter helps us find out if the food an allergen or not. 2 foods are given as allergen: orange and strawberry.

H1(orange)=0
H2(orange)=2
H1(strawberry)=2
H2(strawberry)=6



strawberry

1	0	1	0	0	0	1	0

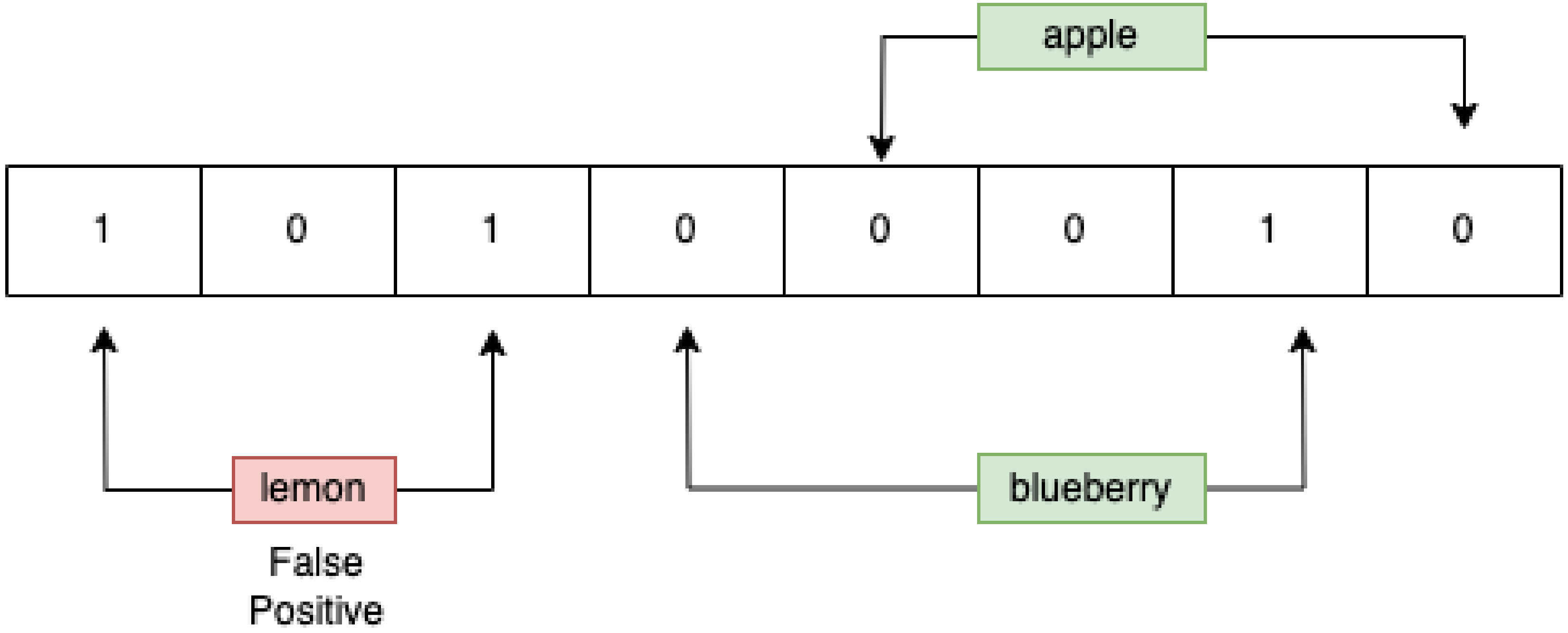
orange

Knowledge Questions

Bloom filter

We have a bloom filter with 2 hash functions (H1 and H2). The filter helps us find out if the food an allergen or not. 2 foods are given as allergen: orange and strawberry.

Let's see if apple, blueberry and lemon are allergens



Knowledge Questions

HFile V2

Based on further reading about HBase explain why the HFile V2 format was needed?

Solution:

Growing size of index started to become a problem, index the days were growing rapidly. As a temporary solution the data block size was doubles. And at the same time the work on a new format HFile V2 started.
In HFile V2 index would be a multi-level data structure and monolithic bloom filter is split into smaller blooms, each corresponding to range of key in HFile. The bloom could and index blocks could be loaded on demand and cached. That made the process faster with bigger data (especially since in practice accessed index blocks are often cached)

Knowledge Questions

NoAM

Base on further reading materials define NoAM.

NoAM (NoSQL Abstract Model) - high-level data model for NoSQL databases the NoAM data model exploits the commonalities of the data modeling elements available in the various NoSQL systems and introduces abstractions to balance their differences and variations.

NoAM data model is defined as follows:

- A NoAM database is a set of collections. Each collection has a distinct name.
- A collection is a set of blocks. Each block in a collection is identified by a block key, which is unique within that collection.
- A block is a non-empty set of entries. Each entry is a pair (ek, ev) where ek is the entry key (which is unique within its block) and ev is its value

Q&A

› That's all, folks! Happy coding!