HPDA Exam XXXXX

Matrikel-number	
Course of studies / Studiengang	

Points					
			ANSWER ONE QUESTION		
Question	Q1	Q2	Q3	Q4	TOTAL
Possible marks	15	15	20	20	50
Achieved					
Initials marker					

- ***The exam is 90 minutes.***
- ***You can answer in English or German.***
- ***Make sure to indicate the question answered properly on the answering sheets.***
- ***If you need extra pages, there are some at the end of the exam.***
- ***Wait until the invigilators tell you to start.***

You MUST answer Question 1 and 2 but you may can CHOOSE if you like to answer question 3 and 4. We will take the maximum of Q3 and Q4, if you answer both.

A. **ANSWER Question 1 and 2**

1. (15 marks)

(a) Define the term ACID as well as its FOUR properties.

(5 marks)

(b) Name and define four big data challenges and characteristics with a short sentence each.

(4 marks)

(c) Discuss the relation and impact of big data applications and scientific applications on the scientific method.

(6 marks)

2. (15 marks)

Consider the following properties of a book that need to be managed in a library:

- IBAN, title, author(s), location (of all available copies of this book)

Here, location is information such as “Shelf A, Row 5”.

Users want to perform the following operations:

- retrieve the book information given the IBAN
- find all books written by a specified author

(a) For MongoDB sketch a suitable **document model** and describe how the operations are implemented. Describe in one/two sentences which operation(s) are efficiently performed and why.

(8 marks)

(b) Describe how the IBAN and title could be stored using the NetCDF data model. You may use the CDL.

(3 marks)

(c) Provide an argument why OLAP is not suitable to analyse this data.

(2 marks)

(d) Describe how the authors of all books can be efficiently stored using **a wide columnar model**.

(2 marks)

B. **(ANSWER EITHER Q3 or Q4)**
Each question is worth 20 marks

3. (20 marks)

(a) In a report, a vendor reports performance benchmarks of a computation solution: by using 10 compute nodes, the solution manages to multiply 1.000.000.000 single-precision floating-point values stored in CSV files on a Hadoop file system in 10 seconds.
(1) Develop a performance model suitable to assess the observed performance. State your assumptions about missing information.
(2) Judge the efficiency of the solution by comparing it to your performance expectation.

(8 marks)

(b) A user wants to analyze COVID-19 medical statistics provided in a CSV file in the following format:
country, city, test date, first symptoms date, death date
One line could look like:
Germany, Hamburg, 10/04/2020, 08/04/2020, 25/04/2020

Note that the same city name might occur in different countries, these are different cities.

Sketch a Spark Python program which computes the number of tests conducted in Germany, for which the first symptoms occur in April.
You may use a function like date("DATE string").year() to retrieve individual fields from the date.

(5 marks)

(c) Sketch the Map and Reduce functions (in Pseudocode or a suitable language) for a MapReduce program which computes per city the number of tests that occur in April (from task b).

(7 marks)

4. (20 marks)

A company wants to develop a service that provides a front-end to parallel applications that then run on one of the various cluster systems provided in the company's cloud.
In particular, a user must be able to conveniently:
1) upload/download files in a hierarchical tree,
2) submit a program that then is executed (using the given input/output, specify the number of processes, and runtime arguments to the program),
3) download the generated output.

(a) Sketch the overall hardware architecture of the cloud where this application could run. Include in the figure the distributed memory architecture.

(5 marks)

(b) Design the RESTful API of a software front-end for the service that is able to perform the operations listed above on the hardware architecture.

i) List each individual operation with its semantics, the URL and additional arguments as well as the HTTP verb used. Write a simple example request for each operation. Mention in one sentence one additional requirement for the design that could be relevant in a multi-user environment.

(8 marks)

ii) Discuss on high-level how the design of the front-end could be integrated in a multitier architecture.

(2 marks)

(c) Justify how consistent hashing can be used to provide a **fault-tolerant, scalable** and **performant** storage for the user's files in **this use case**.

(5 marks)