	Family name:Given name:
1.	(20%) Given N replicas, let's call R the ReadConcern parameter of MongoDB and W the WriteConcern (which indicate respectively the number of copies it reads and writes, before confirming the operation to the user); give the equation involving those variables that corresponds to the eventually consistent configuration.
2.	(20%) Given an empty linear hash with $f(x) = x$ (i.e., we directly apply the module to the keys), and a capacity of four keys per bucket, draw the result of inserting the following keys in the given order: 12, 4, 10, 49, 42, 60, 63, 53, 47, 27, 26, 28, 13, 52.
3.	(20%) Given a file of $3.2GB$ stored in an HDFS cluster of 50 machines, and containing $16 \cdot 10^5$ key-value pairs in a SequenceFile; estimate the execution time of a Spark job containing a single map transformation and an action storing the results in a file. Explicit any assumption you make and consider also the following parameters:
	• Chunk size: $128MB$ (default)
	• Replication factor: 3 (default)
	• Map function (i.e., the parameter of the transformation) execution time: 10^{-3} sec/call (this is the only cost you have to consider)
	• Save action execution time: 0sec (do not consider its cost at all)

4. (20%) Analyze (i.e., briefly give **pros and cons**) the following **JSON design** compared to other equivalent JSON designs from the three perspectives:

```
"BID-PRODUCT":{
1
  "B_ID": int(4), "B_PRICE": int(4), "U_ID": int(4),
2
  "PRODUCT": {
3
4
   "P_ID": int(4), "P_INFO": varchar(100)
5
 }}
 "PRODUCT-SELLER-REGION": {
6
  "P_ID": int(4), "P_INFO": varchar(100),
7
  "USER": {
8
9
   "U_ID": int(4), "U_F_NAME": varchar(20),
   "REGION": {
10
    "R_ID": int(4), "R_NAME": varchar(10)
11
12 }}
 "PRODUCT-COMMENTS": {
13
  "P_ID": int(4), "P_INFO": varchar(100),
14
  "COMMENTS": [{
15
16
   "C_ID": int(4), "C_TITLE": varchar(20), "U_ID": int(4)
17
 }]}
 (a) Read (a.k.a. Query)
   i. Pros
    .....
   ii. Cons
    .....
 (b) Update
   i. Pros
    ii. Cons
    (c) Memory (a.k.a. Space)
   i. Pros
    ......
   ii. Cons
    ......
```

Big Data Management (MIRI)

Final Exam

	Family name:	
	•	
5.	apartments without any secon memory. We want to use Spa	ngoDB collection in a distributed cluster, which contains prices of adary index. Such collection is big enough not to completely fit in ark to compute the standard deviation per neighbourhood. Clearly ion and briefly justify the choice (it is not necessary to provide
	A. Use Spark only to p the result.	ush the query to MongoDB aggregation framework and simply get
	B. Push only some of the rest in Spark.	he operations to MongoDB aggregation framework and the run the
	C. Load the whole colle	ection to an RDD and perform all computations in Spark.