Department of Computer science

For first-year Master: Internet of Things & CybersecurityB

Big Data: Exam- the answer

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv

الاسم: .... اللقب: .... رقم التسجيل: ....

EX 01 : (7.5 p)

Mark √ for the correct answer

What is Big Data?	
A type of database	
Data analysis technique	
Large amounts of data	✓

Which of the following is not a characteristic of Big Data?	
Variety	
Volatility	✓
Veracity	

Which of the following tools is used for distributed storage and processing of large data sets?	
SQL	
Hadoop	✓
Oracle	

What is Hadoop Distributed File System (HDFS) designed for?	
Real-time processing	
Storing small files	
Handling large data sets with fault tolerance	✓

Which component is responsible for resource management?	
MapReduce	
YARN	✓
Hive	

What does the term 'MapReduce' refer to ?	
A data storage technique	
A programming model for data processin	✓
A database management system	

In the context of Big Data, what does 'velocity' refer to?	
The speed at which data is generated and processed	✓
The type of data	
The amount of data	

The key attribute of business intelligence aspect of Big data is :	
Volume	✓
Visibility	
Variety	

In the context of CAP Theorem, what are the characteristics suitable for RDBMS?	✓
CP	
CA	
AP	

In the context of CAP Theorem, what are the characteristics suitable for NoSQL?	
CP	
CA	
AP	✓

In the context of CAP Theorem ,what are the characteristics suitable for Small Data sets?	
CP	
CA	✓
AP	

The key attribute of statistical aspect of Big data is :	
Variety	
Value	
Veracity	✓

In HDFS, Client can find location of blocks from :	
DataNode	
NameNode	✓
Node Manager	

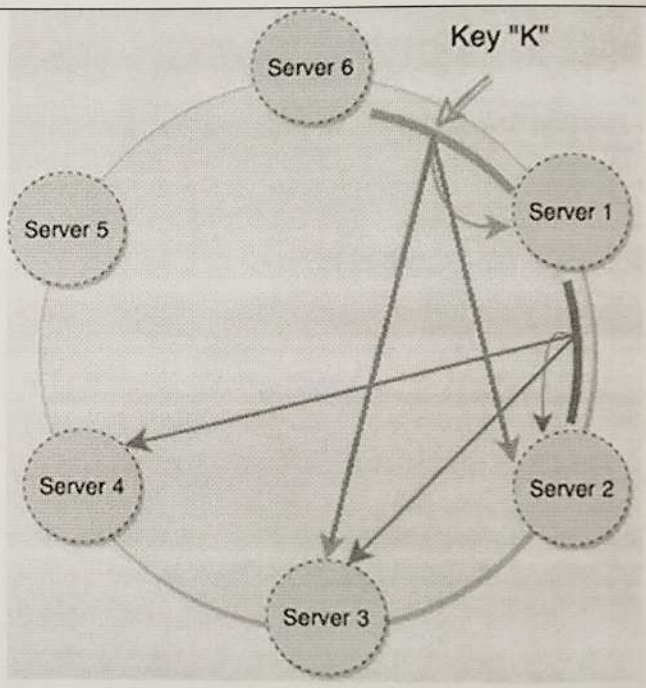
Semi-structured data	
HTML documents	✓
Invoices	✓
Word document	

In the context of Meta Analyze, the Funnel Plot is used for the purpose of :	
Identify publication bias	✓
Homogeneity evaluation	✓
Understanding the effect of study size	✓

EX 02 : (6 p) : Make the following comparison:

	Traditional RDBMS	Hadoop MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

EX 03 : (6.5 p) : Explain the following figure:



As we saw above, Consistent Hashing helps with efficiently partitioning and replicating data; therefore, any distributed system that needs to scale up or down or wants to achieve high availability through data replication can utilize Consistent Hashing.

تساعد عملية التجزئة المتسقة في تقسيم البيانات ونسخها بكفاءة؛ لذلك، فإن أي نظام موزع يحتاج إلى التوسيع أو التخفيض أو يريد تحقيق توفر عالٍ من خلال تكرار البيانات يمكنه استخدام التجزئة المتسقة.

To ensure highly available and durability, Consistent Hashing replicates each data item on multiple N nodes in the system where the value N is equivalent to the replication factor.

لضمان التوفر العالي والمتانة، تقوم عملية التجزئة المتسقة بتكرار كل عنصر بيانات على عقد N متعددة في النظام حيث تكون القيمة N مكافئة لعامل النسخ المتماثل.

The replication factor is the number of nodes that will receive the copy of the same data. For example, a replication factor of two means there are two copies of each data item, where each copy is stored on a different node.

عامل النسخ هو عدد العقد التي ستتلقى نسخة من نفس البيانات. على سبيل المثال، يعني عامل النسخ المتماثل وجود نسختين من كل عنصر بيانات، حيث يتم تخزين كل نسخة على عقدة مختلفة.

Each key is assigned to a coordinator node (generally the first node that falls in the hash range), which first stores the data locally and then replicates it to clockwise successor nodes on the ring.

يتم تعيين كل مفتاح إلى عقدة منسقة (عادةً العقدة الأولى التي تقع في نطاق التجزئة)، والتي تقوم أولاً بتخزين البيانات محليًا ثم تكرارها إلى العقد اللاحقة في اتجاه عقارب الساعة على الحلقة.

In a distributed system, any server responsible for a huge partition of data can become a bottleneck for the system (عنق الزجاجة للنظام). A large share of data storage and retrieval requests will go to that node which can effectively bring the performance of the whole system down. Such loaded servers are called hotspots.

في النظام الموزع، يمكن لأي خادم مسؤول عن قسم ضخم من البيانات أن يصبح عنق الزجاجة للنظام (عنق الزجاجة الكاملة). ستذهب حصة كبيرة من طلبات تخزين البيانات واسترجاعها إلى تلك العقدة مما قد يؤدي إلى انخفاض أداء النظام بأكمله بشكل فعال. تسمى هذه الخوادم المحملة بالنقاط الساخنة. يمكننا تعيين عدد كبير من النطاقات الفرعية لخادم قوي وعدد أقل من النطاقات الفرعية لخادم أقل قوة