**DISTRIBUTED DATABASES FINAL EXAM**

1. Linguistic matching approaches, as the name implies, use element names and other textual information (such as textual descriptions/annotations in schema definitions) to perform matches among elements. **True**

2. Hybrid algorithms combine multiple matchers within one algorithm. **True**

3. Binary approaches are a special case of iterative n-ary. **True**

4. A number of issues affect the particular matching algorithm are: **All of given (Element versus structure level matching; Matching cardinality, Schema versus instance matching)**

5. Learning-based approaches formulate the problem as one of classification where concepts from various schemas are classified into classes according to their similarity. **True**

6. Communication autonomy means that a component DBMS communicates with others at its own discretion, and, in particular, it may terminate its services at any time. **True**

7. Query translation and execution is performed by the wrappers using the component DBMSs. **True**

8. Customized Approach **uses previous knowledge about the component DBMSs, as well as their external characteristics, to subjectively determine the cost information.**

9. GFS organizes files as a tree of directories and identifies them by pathnames. **True**

10. Non-monotonic queries may produce results that cease to be valid as new data is added and existing data changed (or deleted). **True**

11. A monotonic query is one whose results can be updated incrementally. **True**

12. Google applications have the following characteristics: **All of given (their files are very large, typically several gigabytes, containing many objects such as web documents, high throughput is more important than low latency, workloads consist mainly of read and append operations, while random updates are rare. Read operations consist of large reads of bulk data and small random reads.)**

13. Object storage manages data as objects. **True**

14. Graph databases represent and store data directly as graphs which allows easy expression and fast processing of graph-like queries. **True**

15. Neo4J – architecture **[provides extensive support for high availability through full replication both at the cluster level and across data centers]**

16. NoSQL means **Not only Structured Query Language**

17. Which are the main categories of NoSQL systems? **all of the above (key-value, graph, document, wide column)**

18. NewSQL's role is to: **all of the above (seeks to combine the scalability of NoSQL systems with the strong consistency and usability of relational DBMSs.; systems provide scalability, as well as availability, flexible schemas, practical APIs for programming complex data-intensive applications)**

19. HadoopDB provide the best of both parallel DBMS (high performance data analysis over structured data) and MapReduce-based systems (scalability, fault-tolerance, and flexibility to handle unstructured data) with an SQL-like language (HiveQL) and a relational data model **True**

20. The web consists of “pages” that are connected by hyperlinks, and this structure can be modeled as a directed graph that reflects the hyperlink structure **True**

21. In every search engine the\_\_\_\_\_\_\_\_\_\_\_ plays one of the most crucial roles. **crawlers**

22. [[3]] have been a major focus of database technology. **both (declarative querying, efficient execution of queries)**

23. OEM (Object Exchange Model) is a self-describing semistructured data model. **True**

24. RDF is the data model on top of HTML and forms a fundamental building block of the semantic web **False**

25. The types of freshness constraints that can be specified are: **All of given (Drift constraints on multiple data items, Time-bound constraints, Value-bound constraints)**

26. The purposes of replication are: **Both of given (System availability and performance, Scalability and application requirements)**

27. Distributed DBMSs may remove single points of failure by replicating data, so that data items are accessible from multiple sites. Consequently, even when some sites are down, data may be accessible from other sites. **True**

28. Eager techniques typically enforce strong mutual consistency criteria. **True**

29. Transaction consistency requires that the global execution history be serializable. **True**

30. Lazy distributed replication protocols are the most complex ones owing to the fact that updates can occur on any replica and they are propagated to the other replicas lazily. **True**

31. Rewriting a query expressed on the global schema into an equivalent query on the views describing the local relations is difficult for three reasons. **All of given (There may be many more views than global relations, Views may contain complex predicates to reflect the content of the local relations, No direct correspondence between the terms in GS)**

32. Logical data integration and the resulting systems are known by a variety of names: **Both of given (data integration, information integration)**

33. In the second step of bottom-up design, the intermediate schemas are used to generate a GCS. The schema generation process consists of the following order of steps: 1. Schema matching to determine the syntactic and semantic correspondences among the translated LCS elements or between individual LCS elements and the predefined GCS elements 2. Schema mapping that determines how to map the elements of each LCS to the other elements of the GCS 3.Integration of the common schema elements into a global conceptual (mediated) schema if one has not yet been defined **1, 3, 2**

34. The bottom-up design involves both the generation of the GCS and the mapping of individual LCSs to this GCS. **GCS, LCSs, GCS**

35. In the black-box approach, the cost functions are expressed logically (e.g., aggregate CPU and I/O costs, selectivity factors), rather than on the basis of physical characteristics (e.g., relation cardinalities, number of pages, number of distinct values for each column). **True**

36. Logical data integration and the resulting systems are known by a variety of names; data integration and information integration. **True**

37. In most cases, the rule base contains both intra and inter schema rules. **True**

38. Integration methodologies can be classified as binary or n-ary mechanisms based on the manner in which the local schemas are handled in the first phase. **True**

39. Main advantages of shared-disk are: **All of given (good load balance, simple and cheap administration, high availability)**

40. Intra-operation means: **The same operation on different data**

41. A parallel query optimizer has following component/s: **All of given (a search space, a cost model, a search strategy)**

42. Shared-nothing and shared-disk are used in tightly coupled multiprocessors, while Shared-memory is used in clusters. **False**

43. NoSQL DBMSs and big data systems typically use shared nothing. **True**

44. With UMA, the physical memory is shared by all processors, so access to memory is in constant time. **True**

45. Potential Benefits of P2P Systems is/are: **All of given (High availability through massive replication, Dynamic self-organization and Load balancing, Scale up to very large numbers of peers and Parallel processing)**

46. The potential advantages of using a blockchain-based cryptocurrency are the following: **All of given (Fewer risks for merchants, Trust through the blockchain, without any central authority, Security and control, Low transaction)**

47. Each peer in Unstructured P2P networks unknowns its neighbors, but does not know the resources that they have. **False**

48. Early work on distributed DBMSs had primarily focused on P2P architectures where there was no differentiation between the functionality of each site in the system. **True**

49. There are two alternatives to maintaining indices in P2P system: centralized and distributed. **True**

50. Blockchain is a recent P2P infrastructure that can record transactions between two parties efficiently and safely **True**

51. There are two alternatives to maintaining indices in P2P system: centralized and distributed. In distributed, one peer stores the metadata for the entire P2P system. **False**  
52. GFS organizes files as a tree of directories and identifies them by pathnames. **True**

53. Hadoop processing engine has two types of nodes, the master node and the worker nodes. **True**

54. Big data applications or systems, are typically characterized by the “four Vs”. **True**

55. In Single Master with Limited Transparency, the update transactions are submitted and executed directly at the master site (as in the eager single master); once the update transaction commits, the refresh transaction is sent to the slaves. 1. an update transaction is first applied to the master replica 2. the refresh transaction is sent to the slaves 3. the transaction is committed at the master, and then; The sequence of execution steps is as follows: **1, 3, 2**

56. Weak mutual consistency criteria do not require the values of replicas of a data item to be identical when an update transaction terminates. **True**

57. In eager centralized replica control, a master site controls the operations on a data item. **True**

58. The autonomy of the component DBMSs poses problems. DBMS autonomy can be defined along the following main dimension/s: **All of given (execution, communication, design)**

59. N-ary integration mechanisms integrate more than two schemas at each iteration. **True**

60. The bucket algorithm considers each predicate of the query independently to select only the views that are relevant to that predicate. **True**

61. Linear scale-up refers to a sustained performance for a linear increase in **Both of given (database size, load and number of nodes)**

62. The parallel database system should demonstrate extensibility advantages: **Both of given (linear scale-up, linear speed-up)**

63. The objectives of parallel database systems are similar to those of distributed DBMSs (performance, availability, extensibility), but have somewhat different focus due to the tighter coupling of computing/storage nodes **True**

64. The term UMA reflects the fact that accesses to the (virtually) shared-memory have a different cost depending on whether the physical memory is local or remote to the processor. **False**

65. Linear scale-up means sustained performance for a linear increase of database size and load. **True**

65. The main types of complex queries which are useful in P2P systems are: **All of given (join queries, top-k queries, range queries)**

66. An \_\_\_\_\_\_\_\_\_\_\_\_\_\_ peer should be able to join or leave the system at any time without restriction. It should also be able to control the data it stores and which other peers can store its data (e.g., some other trusted peers). **autonomous**

67. To increase data availability and access performance, P2P systems replicate data. **True**

68. Mapping Based on Machine Learning approach is generally used when the shared data is defined based on ontologies and taxonomies as proposed for the semantic web. **True**

69. In a super peer network, a requesting peer sends the request, which can be expressed in a high-level language, to its responsible super peer. **True**

70. There are two main approaches for supporting range queries in structured P2P systems. **True**

71. MapReduce does not require the existence of a schema and does not provide a high-level language such as SQL. **True**

72. Persistent queries may be monotonic or non monotonic. **True**

73. Big data is a major aspect of data science, which combines various disciplines such as data management, data analysis and statistics, machine learning, and others to produce new knowledge from data. **True**

74. Design autonomy may restrict the availability and accuracy of cost information that is needed for query optimization. **True**

75. Mapping maintenance creation is the process of creating explicit queries that map data from a local database to the global one. **False**

76. The main problem/s of query optimization in multidatabase systems is/are: **All of given (heterogeneous cost modeling, heterogeneous query optimization (to deal with different capabilities of component DBMSs), adaptive query processing (to deal with strong variations in the environment—failures, unpredictable delays, etc.))**

77. Linguistic matchers that operate at the schema element-level typically deal with the names of the schema elements and handle cases such as: **all of given (hypernyms, synonyms, homonyms)**

78. Mapping creation is the process of detection and correction of mapping inconsistencies resulting from schema evolution. **False**

79. Query classification can be performed according to query characteristics (e.g., unary operation queries, two-way join queries), characteristics of the operand relations (e.g., cardinality, number of attributes, information on indexed attributes), and characteristics of the underlying component DBMSs (e.g., the access methods that are supported and the policies for choosing access methods). **True**

80. If the GCS is defined upfront, the relationship between the GCS and the LCSs can be of two fundamental types: local-as-view and global-as-view. **True**

81. Extract–transform–load (ETL) tools enables extraction of data from sources. **True**

82. Linguistic techniques can be applied in both schema-based approaches and instance-based ones. **True**  
  
83. Database integration can be either physical or logical. **True**

84. In the second step of bottom-up design, the intermediate schemas are used to generate a GCS. **True**

85. Inter-query means: **Different queries on the same data**

86. As in a parallel DBMS, the database cluster middleware has several software layers: transaction load balancer, replication manager, query processor, and fault tolerance manager. **True**

87. The objective of parallel query processing is to transform queries into execution plans that can be efficiently executed in parallel. **True**

88. Inter-operation means: **Different operations of the same query on different data**

89. Extensibility is the ability to expand the system smoothly by adding processing and storage power to the system. **True**

90. Parallel query optimization exhibits similarities with distributed query processing. **True**

91. Query processor **receives and manages client queries, such as compile query, execute query, and start transaction.**

92. Shared-nothing clusters are widely used in practice, typically using NUMA nodes, because they can provide the best cost/performance ratio and scale up to very large configurations (thousands of nodes). **True**

93. A general solution to load balancing is the execution model called Dynamic Processing (DP). **True**

94. Multicore processors are also based on SMP, with multiple processing cores and shared-memory on a single chip. **True**

95. Data processor **manages the database’s data and system data (system log, etc.) and provides all the low-level functions needed to execute queries in parallel, i.e., database operator execution, parallel transaction support, cache management, etc.**

96. There are two main technologies to share disks in a cluster: network-attached storage (NAS) and storage-area network (SAN). **True**

97. A parallel computer, or multiprocessor, is a form of distributed system made of a number of nodes (processors, memories, and disks) connected by a very fast network within one or more cabinets in the same room. **True**

98. The infrastructure of all P2P systems is a P2P network, which is built on top of a physical network (usually the Internet). **True**

99. The key aspect of the proposed peer reference architecture is the separation of the functionality into three main components. **True**

100. The approaches which are used by P2P systems for defining and creating the mappings between peers’ schemas can be classified as follows: **All of given (schema mapping using information retrieval (IR) techniques, mapping based on machine learning techniques, pairwise schema mapping, common agreement mapping)**

101. There are two alternatives to maintaining indices in P2P system: centralized and distributed. In centralized each peer maintains metadata for resources that it holds **False**

102.OceanStore is a data management system designed to provide continuous access to persistent information. **True**

103. The main difference/s od P2P systems is/are: **All of given (the inherent heterogeneity of every aspect of the sites and their autonomy, the massive distribution in current systems, the considerable volatility of these systems)**

104. Super peer P2P systems are hybrid between pure P2P systems and the traditional client–server architectures. **True**

105. In Pairwise Schema Mapping, each user defines the mapping between the local schema and the schema of any other peer that contains data that are of interest. **True**

106. The type of index supported by a P2P system (centralized or distributed) impacts how resources are searched. **True**

107. Unstructured P2P networks refer to those with no restriction on data placement in the overlay topology. **True**

108. In systems that maintain a centralized index, the process involves consulting the central peer to find the location of the resource, followed by directly contacting the peer where the resource is located. **True**

109. P2P networks provide basic techniques for routing queries to relevant peers and this is sufficient for supporting simple, exact-match queries. **True**

110. OceanStore allows concurrent updates on replicated objects and relies on reconciliation to assure data consistency. **True**

111. Many of the popular aggregation functions such as Min, Max, and Average are monotonic. **True**

112. In modern P2P systems, the sites are (quite often) people’s individual machines and they join and leave the P2P system at will, creating considerable hardship in the management of data. **True**

113. Top-k queries have been used in many domains such as: **All of given (multimedia databases, information retrieval, network and system monitoring )**

114. In more advanced P2P systems where replicas can be updated, there is a no need for proper replica management techniques. **False**

115. The reduce function is applied to each record in the input dataset to compute zero or more intermediate (key,value) pairs. **False**

116. Big data management relies on a distributed storage layer, whereby data is typically stored in files or objects distributed over the nodes of a shared-nothing cluster. **True**

117. Database queries may share parts that are identical and techniques for optimizing a batch of queries have long been of interest and this is referred to as multiquery optimization. **True**

118. Map and reduce can be automatically processed in parallel, on different data partitions using many cluster nodes. **True**

119. MapReduce is a simplified parallel data processing approach for execution on a computer cluster. **True**

120. The issues related to consistency of a replicated database are: **Both of given (mutual consistency, transaction consistency)**

121. Mutual consistency requires that the global execution history be serializable. **False**

122. Lazy propagation is used in those applications for which strong mutual consistency may be unnecessary and too restrictive. **True**

123. The critical issue is to ensure that concurrent conflicting Write operations initiated at different sites are executed in the same order at every site where they execute together (of course, the local executions at each site also have to be serializable). **True**

124. Centralized update propagation techniques require that updates are first applied at a master copy and then propagated to other copies (which are called slaves). **True**

125. Dynamic Approach **monitors the runtime behavior of component DBMSs, and dynamically collects the cost information**

126. In global-as-view systems (GAV), the GCS definition exists, and each LCS is treated as a view definition over it. **False**

127. Client manager **manages the connections and disconnections between the client processes, which run on different servers, e.g., application servers, and the query processors**

128. The objective of NUMA is to provide a shared-memory programming model and all its benefits, in a scalable architecture with distributed memory. **True**

129. Tapestry is an extensible P2P system that provides decentralized object location and routing on top of a structured overlay network. **True**

130. Google File System (GFS) has been developed by Google for its internal use and is used by many Google applications and systems. **True**

131. Distributed storage layer typically provides two solutions to store data, objects, or files, distributed over cluster nodes. **True**

132. There are multiple masters; each master execution is similar to lazy single master in the way it handles transactions. This is the case of: **Lazy Primary Copy/Limited Transparency**

133. Mutual consistency deals with the convergence of the values of physical data items corresponding to one logical data item. **True**

134. Replication may be dictated by the applications, which may wish to maintain multiple data copies as part of their operational specifications. **True**

135. Weak mutual consistency criteria require that all copies of a data item have the same value at the end of the execution of an update transaction. **False**

136. The alternative approaches exist for determining the cost of executing queries at component DBMSs are: **All of given (Customized Approach, Black-Box Approach, Dynamic Approach)**

137. Query processing in a multidatabase system is more complex than in a distributed DBMS for the following reason/s: **All of given (Different data models and query languages, Different runtime performance and unpredictable behavior, Different computing capabilities, Different processing cost and optimization capabilities)**

138. Super peer networks exploit and take advantage of peers’ different capabilities in terms of: \_\_\_\_\_\_\_\_\_\_\_\_\_\_ as super peers take on a large portion of the entire network load. **All of given (CPU power, bandwidth, storage capacity)**

139. DHT-based systems provide some APIs: **Both of given (get(key), put(key, data))**

140. The most popular indexing and data location mechanism that is used in structured P2P networks is a distributed hash table (DHT). **True**

141. Blockchain was invented for bitcoin to solve the double spending problem of previous digital currencies without the need of a trusted, central authority. **True**

142. One possible approach for processing top-k queries in unstructured systems is to route the query to all the peers, retrieve all available answers, score them using the scoring function, and return to the user the k highest scored answers. **True**

143. The primary disadvantage is that, as in any centralized algorithm, if there is one central site that hosts all of the masters, this site can be overloaded and can become a bottleneck. **True**

144. There are three issues related to consistency of a replicated database. **False**

145. Strong mutual consistency criteria do not require the values of replicas of a data item to be identical when an update transaction terminates. **False**

146. Distributed techniques apply the update on the local copy at the site where the update transaction originates, and then the updates are propagated to the other replica sites. **True**

147. The execution autonomy of multidatabase systems makes it difficult to apply some of the query optimization strategies we discussed in previous chapters. **True**

148. A “complete” matching algorithm or methodology usually needs to make use of more than one individual matcher. **True**

149. Mapping Based on Machine Learning uses machine learning techniques to automatically extract the mappings between the shared schemas. **True**

150. An efficient algorithm for top-k query processing in centralized and distributed systems is the Threshold Algorithm (TA). **True**

151. The map function is applied to all the values that share the same unique key in order to compute a combined result. **False**

152. The site that hosts the master copy is similarly called the slave site, while the sites that host the slave copies for that data item are called master sites. **False**

153. The disadvantages are that the replicas are not mutually consistent and some replicas may be out-of - date, and consequently, a local read may read stale data and does not guarantee to return the up-to-date value. **True**

154. The main disadvantage of eager update propagation is that a transaction has to update all the copies before it can terminate. **True**

155. Mutual consistency refers to the replicas converging to the same value. **True**

156. Black-Box Approach **treats each component DBMS as a black box, running some test queries on it, and from these determines the necessary cost information**

157. There are two issues related to schema mapping that we will study: mapping creation and mapping maintenance. **True**

158. In local-as-view (LAV) systems, the GCS is defined as a set of views over the LCSs. **False**

159. Binary integration methodologies involve the manipulation of two schemas at a time. **True**

160. In a parallel system, accommodating increasing database sizes or increasing performance demands (e.g., throughput) should be easier. **True**

161. The \_\_\_\_\_\_\_\_\_\_\_\_ use of the P2P system resources (bandwidth, computing power, storage) should result in lower cost, and, thus, higher throughput of queries, i.e., a higher number of queries can be processed by the P2P system in a given time interval. **efficient**

162. Super peer P2P systems are similar to client–server architectures in that not all peers are equal. **True**

163. Efficiency and quality of service should be maintained despite the failures of peers. **True**

164. In Common Agreement Mapping approach, the peers that have a common interest agree on a common schema description for data sharing. **True**

165. Quality of service in P2P databases refers to the user-perceived efficiency of the system, such as completeness of query results, data consistency, data availability, and query response time. **True**

166. The primary advantage of lazy update propagation techniques is that they generally have lower response times for update transactions, since an update transaction can commit as soon as it has updated one copy. **True**

167. Strong mutual consistency criteria require that all copies of a data item have the same value at the end of the execution of an update transaction. **True**

168. These are called distributed techniques since different transactions can update different copies of the same data item located at different sites. **True**

169. The bottom-up design involves both the generation of the GCS and the mapping of individual LCSs to this GCS. **True**

170. Bottom-up design occurs in two general steps: schema translation (or simply translation) and schema generation. **True**

171. Linear scale-up refers to a sustained performance for a linear increase in both database size, load and number of nodes. **True**

172. The main advantages of super peer networks are **Both of given (efficiency, quality of service)**

173. Unstructured networks are the earliest examples of P2P systems whose core functionality remains file sharing. **True**

174. In systems that maintain a distributed index, there are a number of search alternatives. **True**

175. Transaction consistency refers to the replicas converging to the same value. **False**

176. In global-as-view systems (GAV), the GCS definition exists, and each LCS is treated as a view definition over it. **False**

177. Explain the two popular approaches to lightweight web data integration web portals and mashups that aggregate web and other data on specific topics such as travel, hotel bookings, etc:  **Web portals are websites that aggregate data from various sources on specific topics like travel or hotel bookings. They provide a centralized platform where users can access comprehensive information from different websites in one place. On the other hand, mashups combine data from multiple sources, including web APIs, to create new integrated applications or services. They allow users to interact with and manipulate data from different sources, providing personalized and customized experiences based on their specific needs.**

188. What is and represents Metasearching? **Metasearching is an approach for querying the hidden web. It involves selecting relevant databases, translating the user's query for each database, and merging the results. This ensures a comprehensive search experience by accessing multiple sources and providing the user with relevant information.**

189. State several reasons that have motivated the need for NoSQL systems

**NoSQL systems are databases that can store a lot of data, work fast, and change easily. They are good for different reasons. They can grow bigger by adding more computers. They can work faster by using simple data structures and avoiding complex operations. They can store different kinds of data without needing a fixed structure. They can run on cheap hardware and save money on hardware and software licenses.**