**Page 1**

1. Which of the following is an example of big data utilized in action today?
   * A. The Internet
   * B. Individual, Unconnected Hospital Databases
   * C. Wi-Fi Networks
   * D. Social Media
2. What reasoning was given for the following: why is the "data storage to price ratio" relevant to big data?
   * A. Larger storage means easier accessibility to big data for every user because it allows users to download in bulk.
   * B. It isn't, it was just an arbitrary example of big data usage.
   * C. Companies can't afford to own, maintain, and spend the energy to support large data storage unless the cost is sufficiently low.
   * D. Lower prices mean larger storage becomes easier to access for everyone, creating bigger amounts of data for client-facing services to work with.
3. What is the best description of personalized marketing enabled by big data?
   * A. Being able to obtain and use customer information for groups of consumers and utilize them for marketing needs.
   * B. Being able to use personalized data from every single customer for personalized marketing needs.
   * C. Marketing to each customer on an individual level and suiting to their needs.
4. Of the following, which are some examples of personalized marketing related to big data?
   * A. News outlets gathering information from the internet in order to report them to the public.
   * B. A survey that asks your age and markets to you a specific brand.
   * C. Facebook revealing posts that cater towards similar interests.

**Page 2**

1. What is the workflow for working with big data?
   * A. Theory -> Models -> Precise Advice
   * B. Extrapolation -> Understanding -> Reproducing
   * C. Big Data -> Better Models -> Higher Precision
2. Which is the most compelling reason why mobile advertising is related to big data?
   * A. Mobile advertising in and of itself is always associated with big data.
   * B. Since almost everyone owns a cell/mobile phone, the mobile advertising market is large and thus requires big data to contain all the information.
   * C. Mobile advertising benefits from data integration with location which requires big data.
   * D. Mobile advertising allows massive cellular/mobile texting to a wide audience, thus providing large amounts of data.
3. What are the three types of diverse data sources?
   * A. Machine Data, Organizational Data, and People
   * B. Machine Data, Map Data, and Social Media
   * C. Information Networks, Map Data, and People
   * D. Sensor Data, Organizational Data, and Social Media
4. What is an example of machine data?
   * A. Weather station sensor output.
   * B. Social Media
   * C. Sorted data from Amazon regarding customer info.
5. What is an example of organizational data?
   * A. Disease data from Center for Disease Control.
   * B. Satellite Data
   * C. Social Media

**Page 3**

1. Of the three data sources, which is the hardest to implement and streamline into a model?
   * A. Machine Data
   * B. People
   * C. Organizational Data
2. Which of the following summarizes the process of using data streams?
   * A. Integration -> Personalization -> Precision
   * B. Big Data -> Better Models -> Higher Precision
   * C. Theory -> Models -> Precise Advice
   * D. Extrapolation -> Understanding -> Reproducing
3. Where does the real value of big data often come from?
   * A. Size of the data.
   * B. Having data-enabled decisions and actions from the insights of new data.
   * C. Combining streams of data and analyzing them for new insights.
   * D. Using the three major data sources: Machines, People, and Organizations.
4. What does it mean for a device to be "smart"?
   * A. Having a specific processing speed in order to keep up with the demands of data processing.
   * B. Connect with other devices and have knowledge of the environment.
   * C. Must have a way to interact with the user.
5. What does the term "in situ" mean in the context of big data?
   * A. Bringing the computation to the location of the data.
   * B. In the situation
   * C. The sensors used in airplanes to measure altitude.
   * D. Accelerometers.

**Page 4**

1. Which of the following are reasons mentioned for why data generated by people are hard to process? Choose all that apply.
   * A. Very unstructured data.
   * B. Skilled people to analyze the data are hard to come by.
   * C. The velocity of the data is very high.
   * D. They cannot be modeled and stored.
2. What is the purpose of retrieval and storage; preprocessing; and analysis in order to convert multiple data sources into valuable data?
   * A. Since the multi-layered process is built into the Neo4j database connection.
   * B. Designed to work like the ETL process.
   * C. To enable ETL methods.
   * D. To allow scalable analytical solutions to big data.
3. Which of the following are benefits of organization-generated data? Choose all that apply.
   * A. Better Profit Margins
   * B. Customer Satisfaction
   * C. High Velocity
   * D. Improved Safety
   * E. Higher Sales
4. What are data silos and why are they bad?
   * A. Data produced from an organization that is spread out. Bad because it creates unsynchronized and invisible data.
   * B. A giant centralized database to house all the data production within an organization. Bad because it hinders opportunity for data generation.
   * C. Highly unstructured data. Bad because it does not provide meaningful results for organizations.
   * D. A giant centralized database to house all the data produces within an organization. Bad because it is hard to maintain as highly structured data.

**Page 5**

1. Which of the following are benefits of data integration? Choose all that apply.
   * A. Unify your data system.
   * B. Adds value to big data.
   * C. Reduce data complexity.
   * D. Increase data collaboration.
   * E. Monitoring of data.
   * F. Increase data availability.
2. Amazon has been collecting review data for a particular product. They have realized that almost 90% of the reviews were mostly a 5/5 rating. However, of the 90%, they realized that 50% of them were customers who did not have proof of purchase or customers who did not post serious reviews about the product. Of the following, which is true about the review data collected in this situation?
   * A. Low Volume
   * B. Low Veracity
   * C. High Veracity
   * D. High Volume
   * E. Low Valence
   * F. High Valence
3. As mentioned in the slides, what are the challenges to data with a high valence?
   * A. Complex Data Exploration Algorithms
   * B. Difficult to Integrate
   * C. Reliability of Data

**Page 6**

1. Which of the following are the 6 V's in big data?
   * A. Vision
   * B. Valence
   * C. Variety
   * D. Veracity
   * E. Value
   * F. Velocity
   * G. Volume
2. What is the veracity of big data?
   * A. The abnormality or uncertainties of data.
   * B. The connectedness of data.
   * C. The size of the data.
   * D. The speed at which data is produced.
3. What are the challenges of data with high variety?
   * A. Hard in utilizing group event detection.
   * B. Hard to perform emergent behavior analysis.
   * C. Hard to integrate.
   * D. The quality of data is low.
4. Which of the following is the best way to describe why it is crucial to process data in real-time?
   * A. More accurate.
   * B. Prevents missed opportunities.
   * C. More expensive to batch process.
   * D. Batch processing is an older method that is not as accurate as real-time processing.
5. What are the challenges with big data that has high volume?
   * A. Storage and Accessibility
   * B. Effectiveness and Cost
   * C. Cost, Scalability, and Performance
   * D. Speed Increase in Processing

**Page 7**

1. Which of the following are parts of the 5 P's of data science and what is the additional P introduced in the slides?
   * A. Perception
   * B. Programmability
   * C. Purpose
   * D. Product
   * E. Platforms
   * F. Process
   * G. People
2. Which of the following are part of the four main categories to acquire, access, and retrieve data?
   * A. Remote Data
   * B. Web Services
   * C. Traditional Databases
   * D. Text Files
   * E. NoSQL Storage
3. What are the steps required for data analysis?
   * A. Select Technique, Build Model, Evaluate
   * B. Investigate, Build Model, Evaluate
   * C. Classification, Regression, Analysis
   * D. Regression, Evaluate, Classification
4. Of the following, which is a technique mentioned in the videos for building a model?
   * A. Analysis
   * B. Investigation
   * C. Validation
   * D. Evaluation

**Page 8**

1. What is the first step in finding a right problem to tackle in data science?
   * A. Ask the Right Questions
   * B. Define the Problem
   * C. Define Goals
   * D. Assess the Situation
2. What is the first step in determining a big data strategy?
   * A. Organizational Buy-In
   * B. Build In-House Expertise
   * C. Business Objectives
   * D. Collect Data
3. According to Ilkay, why is exploring data crucial to better modeling? Data exploration... <complete the sentence>
   * A. leads to data understanding which allows an informed analysis of the data.
   * B. enables histograms and others graphs as data visualization.
   * C. enables understanding of general trends, correlations, and outliers.
   * D. enables a description of data which allows visualization.
4. Why is data science mainly about teamwork?
   * A. Data science requires a variety of expertise in different fields.
   * B. Exhibition of curiosity is required.
   * C. Analytic solutions are required.
   * D. Engineering solutions are preferred.
5. What are the ways to address data quality issues?
   * A. Data Wrangling
   * B. Generate best estimates for invalid values.
   * C. Merge duplicate records.
   * D. Remove data with missing values.
   * E. Remove outliers.

**Page 9**

1. What is done to the data in the preparation stage?
   * A. Identify Data Sets and Query Data
   * B. Understand Nature of Data and Preliminary Analysis.
   * C. Retrieve Data
   * D. Build Models
   * E. Select Analytical Techniques
2. Which of the following is the best description of why it is important to learn about the foundations for big data?
   * A. Foundations help you revisit calculus concepts required in the understanding of big data.
   * B. Foundations stand the test of time.
   * C. Foundations is all that is required to show a mastery of big data concepts.
   * D. Foundations allow for the understanding of practical concepts in Hadoop.
3. What is the benefit of a commodity cluster?
   * A. Much faster than a traditional super computer.
   * B. Cost Effective
   * C. Enables fault tolerance
   * D. Prevents network connection failure.
   * E. Prevents individual component failures.
4. What is a way to enable fault tolerance?
   * A. System Wide Restart
   * B. Redundant Data Storage
   * C. Data-Parallel Job Restart
   * D. Distributed Computing
   * E. Better LAN Connection

**Page 10**

1. What are the specific benefit(s) to a distributed file system?
   * A. High Fault Tolerance
   * B. High Concurrency
   * C. Large Storage
   * D. Data Scalability
2. Which of the following are general requirements for a programming language in order to support big data models?
   * A. Optimization of Specific Data Types
   * B. Utilize Map Reduction Methods
   * C. Enable Adding of More Racks
   * D. Handle Fault Tolerance
   * E. Support Big Data Operations
3. What does laaS provide?
   * A. Hardware Only
   * B. Computing Environment
   * C. Software On-Demand
4. What does PaaS provide?
   * A. Hardware Only
   * B. Software On-Demand
   * C. Computing Environment
5. What does SaaS provide?
   * A. Computing Environment
   * B. Software On-Demand
   * C. Hardware Only

**Page 11**

1. What are the two key components of HDFS and what are they used for?
   * A. FASTA for genome sequence and Rasters for geospatial data.
   * B. NameNode for metadata and DataNode for block storage.
   * C. NameNode for block storage and Data Node for metadata.
2. What is the job of the NameNode?
   * A. Coordinate operations and assigns tasks to Data Nodes
   * B. Listens from DataNode for block creation, deletion, and replication.
   * C. For gene sequencing calculations.
3. What is the order of the three steps to Map Reduce?
   * A. Shuffle and Sort -> Map -> Reduce
   * B. Map -> Reduce -> Shuffle and Sort
   * C. Shuffle and Sort -> Reduce -> Map
   * D. Map -> Shuffle and Sort -> Reduce
4. What is a benefit of using pre-built Hadoop images?
   * A. Guaranteed hardware support.
   * B. Less software choices to choose from.
   * C. Quick prototyping, deploying, and validating of projects.
   * D. Quick prototyping, deploying, and guaranteed bug free.
5. What is an example of open-source tools built for Hadoop and what does it do?
   * A. Zookeeper, analyze social graphs.
   * B. Zookeeper, management system for animal named related components.
   * C. Giraph, for processing large-scale graphs.
   * D. Giraph, for SQL-like queries.
   * E. Pig, for real-time and in-memory processing of big data.

**Page 12**

1. What is the difference between low level interfaces and high level interfaces?
   * A. Low level deals with interactivity while high level deals with storage and scheduling.
   * B. Low level deals with storage and scheduling while high level deals with interactivity.
2. Which of the following are problems to look out for when integrating your project with Hadoop?
   * A. Advanced Alogrithms
   * B. Random Data Access
   * C. Task Level Parallelism
   * D. Data Level Parallelism
   * E. Infrastructure Replacement
3. As covered in the slides, which of the following are the major goals of Hadoop?
   * A. Facilitate a Shared Environment
   * B. Enable Scalability
   * C. Provide Value for Data
   * D. Latency Sensitive Tasks
   * E. Optimized for a Variety of Data Types
   * F. Handle Fault Tolerance
4. What is the purpose of YARN?
   * A. Allows various applications to run on the same Hadoop cluster.
   * B. Enables large scale data across clusters.
   * C. Implementation of Map Reduce.

**Page 13**

1. What are the two main components for a data computation framework that were described in the slides?
   * A. Node Manager and Applications Master
   * B. Applications Master and Container
   * C. Resource Manager and Node Manager
   * D. Resource Manager and Container
   * E. Node Manager and Container
2. Download the text to Alice's Adventures in Wonderland from [link] and run word-count on it. How many times does the word Cheshire occur? (Do not include the word 'Cheshire with an apostrophe. The string -->'Cheshire<-- does not count)
   * *(Requires numerical answer)*
3. The set of example MapReduce applications includes wordmedian, which computes the median length of words in a text file. If you run wordmedian using words.txt (the Shakespeare text) as input, what is the median word length? Note that wordmedian prints the median length to the terminal at the end of the MapReduce job; the output file does not contain the median length.
   * *(Requires numerical answer)*
4. (Questions 1-3 pertain to the video lecture "Exploring the Relational Data Model of CSV") What is the approximate population of La Paz county in the state of Arizona for the CENSUS2010POP (column H)? (Choose the best answer.)
   * A. 20000
   * B. 10000
   * C. 25000
   * D. 15000

**Page 14**

1. What county in the state of Wyoming has the smallest estimated population?
   * A. Uinta
   * B. Niobrara
   * C. Platte
   * D. Sweetwater
2. At 2:45 of the video, the Instructor creates a filter for all of the counties in California with a population greater than 1,000,000. However, included in the results is the entire state of California. This anomalous value might skew our analysis if, for example, we wanted to compute the average population of these results. What additional filter might work to resolve this problem?
   * A. Add a filter to detect and remove results which do not include the word "County" in column G.
   * B. Add a filter which finds all counties with population greater than 100,000 AND less than 10,000,000 for column H (CENSUS2010POP).
   * C. Add a filter where the value in column E is greater than 1,000,000.
   * D. None of the above
3. (Questions 4 and 5 pertain to the video "Exploring Sensor Data") How often (in seconds) do the R5 measurements occur?
   * A. 60
   * B. 50
   * C. 40
   * D. 30

**Page 15**

1. What is the field for rain accumulation?
   * A. Sm
   * B. Dn
   * C. Dx
   * D. Rc
2. (Questions 6 and 7 pertain to the video lecture "Exploring the Array Data Model of an Image") What is the (Red, Green, Blue) pixel value for location 500, 2000?
   * A. (134, 145, 46)
   * B. (50, 156, 182)
   * C. (100, 123, 149)
   * D. (163, 118, 79)
3. Is this value likely to be land or ocean?
   * A. Ocean
   * B. Land
4. (Questions 8 and 9 pertain to the video lecture "Exploring the Semistructured Data Model of JSON") Given a tweet, what path would you most likely enter to obtain a count of the number of followers for a user?
   * A. user/followers\_count
   * B. user/statuses\_count
   * C. user/listed\_count
   * D. None of the above
5. Which of the following fields are nested within the 'entities' field (select all that apply)?
   * A. user\_mentions
   * B. urls
   * C. symbols
   * D. views
   * E. events
   * F. tweets

**Page 16**

1. What is a possible pitfall of utilizing Excel as a way to manipulate small databases?
   * A. Excel does not enforce many principles of relational data models.
   * B. Excel does not allow algorithms for data manipulation.
   * C. Excel is a user program and thus cannot run on a server.
2. What does the term "atomic" mean in the context of relational databases?
   * A. A tuple that cannot be reduced.
   * B. One unit of information that cannot be decomposed.
   * C. A column or row of data. Depends on the context.
   * D. Fixed schema of a particular database.
3. What is the Pareto-Optimality problem?
   * A. Find the shortest path from source node to target node.
   * B. Find the best possible path given two or more optimization criteria where neither constraint can be fully optimized simultaneously.
   * C. Find the optimal path that requires going through specific nodes given by the user.
4. What constitutes a community within a graph?
   * A. Many anomalous neighborhoods within the same vicinity.
   * B. A dense amount of edge connections between nodes in a community and a few connections across communities.
   * C. A neighborhood defined by an integer constant K around a specific node. All K+1 nodes belong in another community.
   * D. High density of nodes at a certain location.

**Page 17**

1. Why are trees useful for semi-structured data such as XML and JSON?
   * A. Trees take advantage of the parent-child relationship of the data for easy navigation.
   * B. Computers can easily visualize the data with a tree structure.
   * C. It is not always the case that XML and JSON can be represented as trees.
   * D. They are only useful for XML data as tree-like structure is apparent with tags. While JSON does not contain a tree-like structure as it contains arrays.
2. What is the general purpose of modeling data as vectors?
   * A. Results can be ordered by similarity using vector projection.
   * B. Enables image searching.
   * C. Enables weighting of the query.
   * D. The ability to normalize vectors allowing probability distributions.
3. For the following questions 7, 8, and 9, suppose a registration website creates data with the following fields for each person registered (note: if the user does not input a value, NULL is stored instead): Name, Date, Address, and Account Number. Suppose we collect data month by month. Each month, we would have a batch of data containing the fields listed above. At the end of the year, we want to summarize our registrant activities for the entire year, so we would remove redundancies in our data by removing any records with duplicate account numbers from month to month. What type of operation do we use in this scenario?
   * A. Union
   * B. Join
   * C. Subsetting
   * D. Not an Operation

**Page 18**

1. From the information given in question 7, what are the constraints, if any, which we have placed on the Account Number field for the end of year collection?
   * A. Account should have at most n digits.
   * B. There are no constraints.
   * C. If we had n duplicate Account Numbers then we will remove n-1 duplicate fields.
   * D. Account Number should be unique.
2. Suppose 100 people signup for our system and of the 100 people, 60 of them did not input an address. The system lists the values as NULL for these empty entries in the address field. Would this situation still have structure for our data?
   * A. Yes the data has structure because we have placed a structural constraint on the data, thus the data will always have the originally defined structure.
   * B. No because the majority of data do not have a specific field filled, thus our originally defined structure is lost.
3. What is true between data modeling and the formatting of the data?
   * A. The data does not necessarily need to be formatted in a way that represents the data model. Just so long as it can be extrapolated.
   * B. There is always one specific schema for storing model data that is the best and preferred method for the specific data representation.
   * C. There is a one to one correspondence between formatting data and data modeling. For every model of data, there is only one way to store the data.

**Page 19**

1. What is streaming?
   * A. Using static data stored from a real time source in order to process and guide the application.
   * B. Calculating results using real time data otherwise known as streaming data.
   * C. Utilizing real time data to compute and change the state of an application continuously.
   * D. Using sensors to manipulate the system, such as a smart car being able to drive by itself using sensors to detect road hazards.
2. Of the following, what best describes the properties of working with streaming data?
   * A. Independent computations that do not rely on previous or future data.
   * B. Does not ping the source interactively for a response upon receiving the data.
   * C. Always unbounded in sequence, in other words, data is not guaranteed to be in order.
   * D. Data is always utilized for streaming the application.
   * E. Small time windows for working with data.
   * F. Data manipulation is near real time.
3. What is a characteristic of streaming data?
   * A. The data is finite and requires only finite time and space to process the data.
   * B. The data is unbounded in size and the size determines the time and space of processing the data.
   * C. Data is finite in size and size determines the time and space of processing the data.
   * D. Data is unbounded in size but requires only finite time and space to process it.

**Page 20**

1. What type of algorithm is required for analyzing streaming data?
   * A. Accurate and Consistent
   * B. Fast and Simple
   * C. Fast and Complex
   * D. Accurate and Memory Efficient
2. What is lambda architecture?
   * A. A specific method for processing streaming data using special real time processes.
   * B. A method to process streaming data by utilizing batch processing and real time processing.
   * C. A specific hardware architecture for a server made specifically for processing real time data.
3. Of the following, which best represents the challenge regarding the size and frequency of data?
   * A. There may not be data to produce the notion of size and frequency.
   * B. The size and frequency of the streaming data may be too small.
   * C. The size and frequency of the streaming data may be sporadic.
4. What is the difference between data lakes and data warehouses?
   * A. Data lakes house raw data while data warehouses contain pre-formatted data.
   * B. Data lakes utilize hierarchical systems while data warehouses utilize object storage.
   * C. Data lakes contain only files while data warehouses contain only databases.

**Page 21**

1. What is schema-on-read?
   * A. The process where formatted data is given structure when read.
   * B. Another name for data lakes.
   * C. Data is stored as raw data until it is read by an application where the application assigns structure.
   * D. The process where data is pre-formatted prior to being read but the schema is loaded on read.
2. The desired characteristics of a BDMS include (select all that apply):
   * A. A flexible semi-structured data model
   * B. Support for ACID
   * C. A full query language
   * D. Continuous data ingestion
   * E. Support for common "Big Data" data types
   * F. Narrow range of query sizes
3. Fill in the blank with the best answer: CAP theorem states that all at once within a distributed computer system?
   * A. it is necessary to have consistency, availability, and partition tolerance
   * B. it is necessary to have consistency, accuracy, and partial tolerance
   * C. it is impossible to have consistency, accuracy, and partial tolerance
   * D. it is impossible to have consistency, availability, and partition tolerance
4. What is the purpose of the acronym BASE?
   * A. To impose properties on a BDMS in order to guarantee certain results.
   * B. Enables stricter enforcement of ACID type design.
   * C. The same as ACID.
   * D. To overcome CAP theorem.

**Page 22**

1. What are ziplists in Redis?
   * A. A special type of data type that can store hashes that point to multiple attributes.
   * B. A special type of data type that can store up to 512 mb of image data.
   * C. A compressed list that is stored within the value of the database.
   * D. A look up table that is stored as a value in the database. Look up table points to actual values in memory.
2. What is one of the main features of Aerospike?
   * A. Enables real time data streaming from external sources.
   * B. Support for geospatial data storage and geospatial queries.
   * C. Better equipped for string based search applications.
   * D. Images as values within the database.
3. What database would be best suited for the following scenario: An app development company is trying to implement a cloud based storage system for their new map-based app. The cloud will manage the longitude and latitude of the data in order to track user location.
   * A. Solr
   * B. Redis
   * C. Aerospike
   * D. Vertica
4. What database would be best suited for the following scenario: A big wholesale company is trying to implement a search engine for their products.
   * A. Redis
   * B. Aerospike
   * C. Solr
   * D. Vertica

**Page 23**

1. Which of the following data types are supported by Redis? (select all that apply)
   * A. Strings
   * B. Lists
   * C. Images
   * D. Hashes
   * E. Sorted Sets
   * F. Streaming Video
2. What does it mean for a query language to be declarative?
   * A. The language specifies both the process of how to obtain the data and specifies what data to obtain.
   * B. The language specifies what data to obtain.
   * C. The language specifies the process of how to obtain the data.
   * D. A language specific declaration of data types in order to define the method of data retrieval.
3. Use the following table named "user\_table" to answer the next 2 problems. (Table shown in image) How would you go about querying the entire username column (however many)?
   * A. SELECT username FROM user\_table
   * B. SELECT user\_table FROM user\_name
   * C. SELECT username FROM user\_table WHERE userId=1
   * D. SELECT username FROM userid WHERE \*

**Page 24**

1. How would you go about querying the entire database table (please refer to question 2's table)?
   * A. SELECT username, email FROM userid
   * B. SELECT \* FROM user\_table
   * C. SELECT user\_table FROM \*
   * D. SELECT \* FROM \*WHERE user\_table
2. What is the global indexing table?
   * A. A global table that uses a specific technique called indexing and the table uses an index as the primary key.
   * B. An index table in order to keep track of data records within one machine.
   * C. An index table in order to keep track of a given data type that might exist within multiple machines.
   * D. An index table in order to keep track of a given data type that might exist within one machine.
3. What are the three computing steps of a semi-join?
   * A. Project, Ship, Reduce
   * B. Project, Decompose, Send
   * C. Index, Join, Display
   * D. Query, Join, Display
   * E. None Applicable
4. What is the purpose of a semi-join?
   * A. Increase the efficiency of sending data across multiple machines.
   * B. Another name for join: an operation to combine two tables by column.
   * C. Increase the speed of the join for trade-off of increased data transmission cost.

**Page 25**

1. What is a subquery?
   * A. A query statement within another query.
   * B. A short query than normal.
   * C. An alternative query that acts as a substitute for another query.
2. What is a correlated subquery?
   * A. A type of query that contains a relationship between a variable attribute x and a variable attribute y. The two variables have a dependent relationship causing a correlation.
   * B. A type of query that contains a subquery that requires information from a query one level up.
   * C. A type of query that requires two tables in order to calculate values.
3. What is the purpose of GROUP BY queries?  
   \* A. Enables calculations based on specific columns of the table.  
   \* B. Required before you can use functions like AVG, SUM, MIN, MAX, COUNT.  
   \* C. Enables queries within queries.
4. Consider the following generic statement for questions 10-12: db.<collection>.find(<query filter>, <projection>).<cursor modifier> Which part of the statement would reflect that of the FROM statement in SQL as illustrated in the lecture?  
   \* A. <query filter>  
   \* B. <projection>  
   \* C. <cursor modifier>  
   \* D. <collection>

**Page 26**

1. Which part of the statement would reflect that of the SELECT statement in SQL as illustrated in the lecture?  
   \* A. <collection>  
   \* B. <query filter>  
   \* C. <cursor modifier>  
   \* D. <projection>
2. Which part of the statement would reflect that of the WHERE statement in SQL as illustrated in the lecture?  
   \* A. <projection>  
   \* B. <query filter>  
   \* C. <cursor modifier>  
   \* D. <collection>
3. A sample part of the data structure is as follows: {\_id:1, userIndex: 10, email: "[arealeamil@notreallu.asd](https://www.google.com/url?sa=E&q=mailto%3Aarealeamil%40notreallu.asd)", retainRate:2} What would be the most likely statement that we would need to grab email info for user indexes greater than 24?  
   \* A. db.userIndex.find({email:{

gt:24}}, {\_id:0})  
\* C. db.email.find({userIndex:{

lte:24}}, {email:1, \_id:0})

1. What does it mean to have a \_id:0 within our query statement?  
   \* A. Grab as many objects as possible.  
   \* B. Grab the first object in the results.  
   \* C. Does not have an effect, simple convention left for compatibility issues.  
   \* D. Tell MongoDB not to return a document id.

**Page 27**

1. This quiz encompasses data and content from Week 1 and 2, so we recommend reviewing that material from last week for this quiz as well. What is the highest level that the team has reached in gameclicks? (Hint: use the MAX operation in postgres).  
   \* A. 9  
   \* B. 10
2. C. 8
3. D. 7
4. E. 6

How many user id's (repeats allowed) have reached the highest level as found in the previous question? (Hint: For postgres: you may either use two queries or use a sub-query).

\* A. 106436

\* B. 122757

\* C. 67271

\* D. 98823

\* E. 51294

How many user id's (repeats allowed) reached the highest level in game-clicks and also clicked the highest costing price in buy-clicks? Hint: Refer to question 4 for ideas.

\* A. 66887

\* B. 73226

\* C. 23301

\* D. 32747

What does the following line of code do in postgres? SELECT count(userid) FROM (SELECT buyclicks.userid, teamLevel, price FROM buyclicks JOIN gameclicks on buyclicks.userid = gameclicks.userid) temp WHERE price=3 and teamLevel=5;

\* A. Finds the total number of user ids (repeats allowed) in buy-clicks that have bought items with prices worth $3 and was in a team with level 5 at some point in time.

\* B. Displays the users who have bought items worth $3 and have had a team with level 5.

\* C. This is an invalid line of code, the subquery is not formatted properly.

\* D. Counts the users who exists between both gameclicks and buyclicks files.

Page 28

In the MongoDB data set, what is the username of the twitter account who has a tweet\_followers\_count of exactly 8973882?

\* A. FIFAcom

\* B. SasSpear

\* C. Autocenterit

\* D. CreateImga

What is the main problem with big data information integration?

\* A. Mediated Schema

\* B. Pay-as-you-go model

\* C. Probabilistic Schema Mapping

\* D. Many sources

What would be the two possible solutions associated with "big data" information integration as mentioned in lecture? (Choose 2)

\* A. Probabilistic Schema Mapping

\* B. Customer Transactions

\* C. Pay-as-you-go Model

\* D. Attribute Grouping

\* E. Mediated Schema

What are mediated schemas?

\* A. Schemas created from customer info.

\* B. Schemas created entirely from attribute grouping.

\* C. Schema created from integrating two or more schemas.

\* D. A type of probabilistic schema mapping.

Page 29

In attribute grouping, how would one evaluate if two attributes should go together? (Choose 2)

\* A. Similarity of Attributes

\* B. Customer Interaction

\* C. Probability of Two Attributes Co-occurring

\* D. Integrated Views

\* E. Candidate Designs

What is a data item?

\* A. The real worth of a data value.

\* B. Data found in a mediated schema.

\* C. Data that represents an aspect of a real-world entity.

\* D. Data found in a customer transaction.