

Data Engineering exam questions

1. Questions for general academic profile

K1. Explain in which engineering scenarios text-based data serialization formats are best suited for representing semi-structured data and justify your choice with features of their syntax. Select a data serialization language for the given scenario.

Scenario example: Define a product document in an e-commerce catalogue with dynamic attributes. Defining this structure requires specifying the unique SKU identifier, name, category, nested dictionary of category-specific features.

K2. Describe the role of markup languages in information systems. Explain the difference between predefined markup languages, metalanguages, and lightweight markup languages. For the given case, choose a language and justify your choice.

Scenario example: Define an email message template notifying of a completed ETL process. Defining this structure requires applying semantic block and formatting elements compliant with email client standards.

K3. Present and discuss a general taxonomy (hierarchy) of programming paradigms, identifying the fundamental criterion for division at the highest level. Characterize selected specific paradigms from each branch of this hierarchy and justify the benefits of combining them in modern multi-paradigm software. For the given case, select a programming language and justify your decision.

Scenario example: A global logistics enterprise needs to model a diverse fleet of transport vehicles—ranging from electric delivery vans to container ships—each requiring specialized route optimization algorithms, payload capacity calculations, and fuel efficiency constraints.

K4. Design the logical and physical data structures for the given scenario, classify them, and give the time complexity of the operations used.

Scenario example: Let C be a prefix code over a finite alphabet A (i.e., for any symbols $x, y \in A$, $C(x)$ is not a prefix of $C(y)$). Propose data structures for representing the code C (e.g., estimate the complexity of the operation that checks whether a string belongs to a given prefix code; consider how the representation of a node's children affects time and memory complexity).

K5. Present and describe the primary types of diagrams used for modelling information systems. Explain which aspects of an information system can be modelled using each specific type of diagram. For the given scenario, indicate which diagram(s) you would use and sketch an initial version of one of them.

Scenario example: How to model the behaviour of an active cruise control unit in a car, which determines whether to accelerate or decelerate and whether to send messages to the driver based on multiple possible signals (activation/deactivation by the driver, pressing the brake, radar detection of approaching a vehicle ahead, current speedometer reading, detection of a sensor error).

K6. For the given scenario, describe the step-by-step process of designing and creating a relational database. What modelling methods will you use in this process? How will you verify the created model?

Scenario example: Global sales analysis: processing and aggregating billions of sales facts combined with extensive dimensions (customer, product, time, geography) to generate multi-dimensional management reports.

K7. For the given case, define the data mining task and provide the steps to solve it.

Scenario example: A telecommunications operator has customer data including contract duration, subscription type, billing amounts, service usage, and customer service contact history. For customers who terminated their contracts in the past, information about their cancellation is also available. Based on these data, the operator wants to identify current customers who are at risk of leaving.

K8. Describe the database types you are familiar with and their applications. For the given scenario, select the optimal database type and justify your decision, taking into account various considerations.

Scenario example: Global sales analysis: processing and aggregating billions of sales facts combined with extensive dimensions (customer, product, time, geography) to generate multi-dimensional management reports.

K9. List the core data quality dimensions, and for the given scenario, explain how in this scenario this dimension is understood and how to achieve this specific aspect of data quality.

Scenario example: Sales data and reporting: A company integrates sales data from regional systems into a daily management report. Some transactions are imported twice, while others are unavailable by the reporting deadline. Regional systems use different product identifiers and date formats. Transaction amounts are expressed in different currencies, but the currency field is sometimes empty. Some recorded amounts differ from those on the corresponding source invoices. Identify the relevant data quality dimensions, explain their meaning in this context, and propose methods to assess and improve them.

K10. Describe the task scheduling problem in an operating system kernel, including its goals and the criteria used to evaluate scheduling algorithms. Name

and briefly characterize the main scheduling algorithms. For the given scenario, select the most appropriate scheduling algorithm and explain in detail how it works and why it is suitable for the scenario.

Scenario examples: Interactive system: A system is used interactively by many users. The scheduler should provide good responsiveness and allow processes to be assigned different priorities. MLFQ (Multi-Level Feedback Queue).

K11. Describe the main functions of an enterprise and how do they influence its efficiency and competitiveness.

K12. How market structure influences company decisions and what is its relevance for business data analysis?

K13. What are the key factors that determine the effectiveness of communication in a modern organization, and how would you explain their importance?

K14. What role do assets play in supporting organizational goals and creating value for stakeholders, and how can their significance be understood from economic, social, and strategic perspectives?

K15. Can modern organizations function effectively without marketing research? Justify your answer.

K16. How can investment projects be evaluated while taking into account their impact on organizational development and the socio-economic environment, and what challenges arise when making investment decisions across different areas of activity?

K17. Discuss the principles of user interface design.

K18. Describe the business process lifecycle.

K19. Describe visualizations of quantitative and qualitative datasets. Discuss their types and applications.

K20. Discuss the cybersecurity risk management: its lifecycle and its place in the enterprise risk management.

2. Questions for “Creating big data solutions” stream (Faculty of ETI)

ETI-S1. Characterize specific methods of processing big files.

ETI-S2. Discuss the core issues concerning reliability and high availability of cloud web services.

ETI-S3. Describe main intelligent agents' properties.

ETI-S4. Describe data models used by Geographic Information Systems.

ETI-S5. For IP protocols describe features, advantages, disadvantages, and necessity to change.

ETI-S6. Characterize the architectures of the large-scale business systems.

ETI-S7. Describe differences and similarities between Apache Hadoop and Apache Spark frameworks.

ETI-S8. Explain the term “race condition”. Provide a code sample with such a condition.

ETI-S9. Describe how elements of agile methodologies serve to achieving project goals. Explain it using an example of the chosen methodology.

3. Questions for “Big data in business management” stream (Faculty of ME)

FME-S1. What is the difference between enterprise IT system integration at the data level (Data-level integration) and integration at the business process level (Process-level integration)?

FME-S2. How can organizations use unstructured data to support decision-making and create business value?

FME-S3. Describe the Enterprise Architecture – its components and role in the organization.

FME-S4. Present and discuss the “Devil’s Quadrangle” as a criterion for evaluating process redesign.

FME-S5. How can complex decision problems be addressed when finding an optimal solution is computationally difficult or infeasible?

FME-S6. Discuss how the domain of Web Intelligence has evolved over the past three decades. In your discussion, highlight the key developments that have shaped its evolution? Using an example, explain why ontologies and vocabularies constitute an important element of the Semantic Web.

FME-S7. Describe one static and one dynamic method to calculate the profitability of AI-based project.

FME-S8. How does digitalisation transform the way organisations create, deliver, and capture value, and what are the implications of these changes for different areas of business activity?

FME-S9. Discuss the impact of macro- and microenvironmental factors on shaping digital marketing plans.

4. Questions for “Intelligent data processing” profile (Faculty of ETI)

ETI-P1. Describe characteristics of storing big data in horizontally scalable systems.

ETI-P2. Discuss how recommendation algorithms work and are evaluated using a selected collaborative algorithm as an example.

ETI-P3. Characterize data clustering methods, including their types, selected algorithms, and areas of application.

ETI-P4. Describe areas and mechanisms for providing safety and security in database management systems.

ETI-P5. Describe areas and mechanisms for providing performance in database management systems.

ETI-P6. Discuss regularization methods for deep neural networks.

ETI-P7. Discuss types of neural layers which can be found in a typical convolutional neural network, like AlexNet.

ETI-P8. Describe a graph data model for data and knowledge management.

5. Questions for “Data Exploration in Management” profile (Faculty of ME)

FME-P1. Discuss the analysis of sector attractiveness and competitive conditions.

FME-P2. Discuss the concept of segmentation in market analysis. Present the most important issues related to it.

FME-P3. Explain the role of visualization in internal and external organizational communication. Present examples of visualization methods and their applications.

FME-P4. Discuss the visualization means and their selection depending on the presented content (data, processes, economic phenomena, etc.)

FME-P5. What phenomena/theories can be used when constructing an investment strategy?

FME-P6. What is the portfolio theory? Characterize the investment strategies that portfolio theory enables you to implement.

FME-P7. Discuss the consequences of using bad quality training data in AI-based business.

FME-P8. List and discuss business models that could be used by AI-based businesses.