Fundamentals of Data Engineering 1st edition Reis Test Bank

SKU: 9781098108304-TEST-BANK

Fundamentals of Data Engineering

ISBN 9781098108304 | Chapter 1: Data Engineering Described

Total Questions: 150

Multiple Choice

Q1.
According to the chapter's definition, data engineering primarily focuses on turning raw data into what outcome?

A) High-quality, consistent information for downstream use cases ✓ Correct
B) End-user applications for direct customer interaction
C) Manual reports created only by business users
D) Standalone machine learning models without pipelines

Rationale: The chapter defines data engineering as creating systems and processes that transform raw data into high-quality, consistent information for downstream use.

Q2.
Which stage is part of the data engineering lifecycle described in the chapter?

A) Monetization
B) Ingestion ✓ Correct
C) Visualization governance
D) Procurement

Rationale: Ingestion is explicitly listed as one of the lifecycle stages.

Q3.
Which of the following is identified as an undercurrent across the entire data engineering lifecycle?

A) Graphic design
B) Human resources
C) DataOps ✓ Correct
D) Brand management

Rationale: DataOps is one of the lifecycle undercurrents named in the chapter.

Q4.
Which sequence correctly reflects the lifecycle stages named in the chapter?

A) Generation, Storage, Ingestion, Transformation, Serving ✓ Correct
B) Storage, Procurement, Transformation, Analysis, Archiving
C) Generation, Cleansing, Modeling, Reporting, Monetization
D) Collection, Governance, Warehousing, Visualization, AI

Rationale: The chapter lists the stages as Generation, Storage, Ingestion, Transformation, and Serving.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q5.

Bill Inmon is credited in the chapter with officially coining which term in 1990?

A) Data lake

B) Data warehouse ✓ Correct

C) Data mesh

D) Data fabric

Rationale: The chapter states that Bill Inmon officially coined the term data warehouse in 1990.

Q6.

Which technology did Oracle help popularize after IBM developed the relational database and SQL?

A) Relational database technology ✓ Correct

B) Graph neural networks

C) Blockchain storage

D) Container orchestration

Rationale: The text notes that Oracle popularized relational database technology after IBM developed it and SQL.

Q7.

Which pair developed widely used eponymous data-modeling approaches for data warehousing?

A) Google and Amazon

B) Ralph Kimball and Bill Inmon ✓ Correct

C) Monica Rogati and Lewis Gavin

D) Jesse Anderson and Maxime Beauchemin

Rationale: Ralph Kimball and Bill Inmon are identified as creators of influential data-modeling techniques.

Q8.

What does MPP stand for in the context of analytics databases?

A) Managed Pipeline Processing

B) Massively Parallel Processing ✓ Correct

C) Modular Platform Provisioning

D) Metadata Policy Planning

Rationale: The chapter defines MPP as massively parallel processing.

Q9.

Which role is described as a precursor to today's data engineering discipline?

A) Clinical informaticist

B) Data warehouse engineer ✓ Correct

C) User experience designer

D) Security auditor

Rationale: The chapter states that data warehouse and BI engineering were precursors to modern data engineering.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q10.

According to the Oxford English Dictionary definition cited in the chapter, big data refers to what?

A) Any data stored in the cloud
B) Data generated only by social media platforms
C) Extremely large data sets analyzed computationally for patterns and associations ✓ Correct
D) Structured data kept in relational databases only

Rationale: The chapter quotes the OED definition describing big data as extremely large data sets analyzed computationally for patterns and associations.

Q11.

Which three V's are used in the chapter to describe big data?

A) Validity, value, verification
B) Velocity, variety, volume ✓ Correct
C) Volume, visibility, versioning
D) Variance, vectorization, volatility

Rationale: The chapter identifies velocity, variety, and volume as the three V's of big data.

Q12.

Google published a paper on which system in 2003?

A) Apache Spark
B) Google File System ✓ Correct
C) Amazon S3
D) Apache Kafka

Rationale: The chapter notes that Google published a paper on the Google File System in 2003.

Q13.

Which data-processing paradigm did Google publish a paper on in 2004?

A) MapReduce ✓ Correct
B) Lambda architecture
C) OLAP cubes
D) Entity resolution

Rationale: The chapter states that Google published a paper on MapReduce in 2004.

Q14.

Apache Hadoop was developed at Yahoo and later open sourced in which year?

A) 2001
B) 2004
C) 2006 ✓ Correct
D) 2010

Rationale: The chapter specifies that Yahoo developed and later open sourced Apache Hadoop in 2006.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q15.
Which company is described as creating the first popular public cloud through AWS?

A) Microsoft
B) Google
C) IBM
D) Amazon ✓ Correct

Rationale: The chapter identifies Amazon Web Services as the first popular public cloud.

Q16.
What does EC2 stand for in AWS terminology?

A) Elastic Compute Cloud ✓ Correct
B) Enterprise Cluster Core
C) External Compute Channel
D) Elastic Catalog Cluster

Rationale: The chapter explicitly expands EC2 as Amazon Elastic Compute Cloud.

Q17.
What does S3 stand for in AWS terminology?

A) Scalable Streaming Service
B) Simple Storage Service ✓ Correct
C) Secure Schema Store
D) System Storage Stack

Rationale: The chapter explicitly expands S3 as Amazon Simple Storage Service.

Q18.
During the big data engineering era, which approach was described as being in vogue?

A) GUI-first administration
B) Code-first engineering ✓ Correct
C) Paper-based workflow design
D) Spreadsheet-centric deployment

Rationale: The chapter states that code-first engineering became fashionable with the rise of MapReduce.

Q19.
Which of the following is listed as part of the Hadoop ecosystem in the chapter?

A) YARN ✓ Correct
B) TensorFlow Serving
C) Power BI
D) Salesforce CRM

Rationale: YARN is specifically named as part of the Hadoop ecosystem.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q20.

According to the chapter, the term “big data” is now largely considered what?

A) A mandatory regulatory category
B) A relic describing a particular time and approach ✓ Correct
C) A synonym for cloud-native AI
D) A replacement for data engineering

Rationale: The chapter says the term big data is essentially a relic tied to a specific era and approach.

Q21.

In the 2020s, the data engineer discussed in the chapter is described more precisely as a what?

A) Visualization engineer
B) Data lifecycle engineer ✓ Correct
C) Database operator
D) Research statistician

Rationale: The chapter explicitly describes the modern data engineer more precisely as a data lifecycle engineer.

Q22.

Which pair of privacy-related acronyms are data engineers said to be increasingly conversant in?

A) HL7 and FHIR
B) TCP and IP
C) CCPA and GDPR ✓ Correct
D) OLTP and OLAP

Rationale: The chapter notes that data engineers are now conversant in CCPA and GDPR.

Q23.

How does the chapter position data engineering relative to data science?

A) Data engineering is always a subdiscipline of data science
B) Data engineering is separate from but complementary to data science ✓ Correct
C) Data engineering replaces data science in mature organizations
D) Data engineering is unrelated to analytics

Rationale: The authors argue that data engineering is distinct from, yet complementary to, data science and analytics.

Q24.

According to the chapter, data engineering sits in what relationship to data science?

A) Downstream from data science
B) Parallel but isolated from data science
C) Upstream from data science ✓ Correct
D) Outside the data value chain

Rationale: The chapter explicitly states that data engineering sits upstream from data science.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q25.
Monica Rogati's hierarchy is used in the chapter to show that companies need to build what before tackling AI and ML?

A) A social media presence
B) A solid data foundation ✓ Correct
C) A blockchain layer
D) A proprietary query language

Rationale: The chapter uses Rogati's hierarchy to emphasize building a strong data foundation first.

Q26.
The chapter estimates that many data scientists spend approximately what proportion of their time in the lower parts of the hierarchy gathering, cleaning, and processing data?

A) 10% to 20%
B) 30% to 40%
C) 50% to 60%
D) 70% to 80% ✓ Correct

Rationale: The text states that an estimated 70% to 80% of data scientists' time is spent in those lower-level tasks.

Q27.
Which of the following is listed as an axis that data engineers must constantly optimize?

A) Celebrity endorsement
B) Agility ✓ Correct
C) Office location
D) Patent volume

Rationale: Agility is one of the balancing axes named in the chapter.

Q28.
Which activity does the chapter say a data engineer typically does not directly do?

A) Manage data pipelines
B) Prepare data for analysis
C) Build ML models ✓ Correct
D) Design data architectures

Rationale: The chapter states that data engineers typically do not directly build machine learning models.

Q29.
What is data maturity according to the chapter?

A) The age of an organization's databases
B) The progression toward higher data utilization, capabilities, and integration ✓ Correct
C) The number of data engineers employed by a company
D) The extent of cloud spending on analytics tools

Rationale: The chapter defines data maturity as progression toward greater data utilization, capability, and integration.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q30.

How many stages are in the chapter's simplified data maturity model?

A) Two
B) Three ✓ Correct
C) Four
D) Five

Rationale: The simplified model contains three stages.

Q31.

Which is the first stage in the simplified data maturity model?

A) Leading with data
B) Scaling with data
C) Starting with data ✓ Correct
D) Optimizing with data

Rationale: The first stage named in the model is starting with data.

Q32.

In the "starting with data" stage, a data engineer is most likely to be what?

A) A narrowly focused specialist only
B) A generalist playing several roles ✓ Correct
C) A chief data officer
D) An AI researcher with no operational duties

Rationale: The chapter explains that early-stage data engineers are usually generalists who may play multiple roles.

Q33.

Which recommendation is given for organizations in the "starting with data" stage?

A) Prioritize ML before building a data foundation
B) Avoid talking to business stakeholders
C) Use off-the-shelf, turnkey solutions where possible ✓ Correct
D) Stand up complex clusters for all workloads

Rationale: The chapter advises avoiding unnecessary complexity and using turnkey solutions wherever possible.

Q34.

In the "scaling with data" stage, data engineering roles tend to move from what to what?

A) Analysts to executives
B) Generalists to specialists ✓ Correct
C) Architects to auditors
D) Managers to interns

Rationale: The chapter states that roles move from generalists to specialists during scaling.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q35.

Which practice is specifically identified as a goal in the “scaling with data” stage?

A) Eliminating all governance controls
B) Adopting DevOps and DataOps practices ✓ Correct
C) Replacing all managed services with custom code
D) Restricting analytics to executives only

Rationale: Adopting DevOps and DataOps practices is listed as a stage 2 goal.

Q36.

At the “leading with data” stage, what is described as a significant danger?

A) Using SQL for transformations
B) Complacency ✓ Correct
C) Hiring analysts
D) Documenting metadata

Rationale: The chapter warns that complacency is a major danger at stage 3.

Q37.

Which of the following is highlighted in the “leading with data” stage?

A) Exclusive reliance on ad hoc reporting
B) Automation for seamless introduction and usage of new data ✓ Correct
C) Avoidance of collaboration with ML engineers
D) Elimination of data catalogs

Rationale: Stage 3 includes creating automation for seamless onboarding and use of new data.

Q38.

According to the chapter, what kind of formal educational path currently exists for becoming a data engineer?

A) A universally standardized university curriculum
B) A well-established licensing pathway
C) No standard formal path ✓ Correct
D) A mandatory apprenticeship model

Rationale: The chapter states that universities do not have a standard data engineering path and common curricula do not yet exist.

Q39.

Which adjacent field is identified as making transition into data engineering easier?

A) Graphic illustration
B) Software engineering ✓ Correct
C) Retail merchandising
D) Public relations

Rationale: Software engineering is listed as one of the adjacent fields that can ease the transition into data engineering.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q40.

Which statement best reflects the chapter's view of Agile, DevOps, and DataOps?

A) They are fundamentally cultural, not merely technological ✓ Correct
B) They are obsolete in modern data teams
C) They apply only to software engineering, not data
D) They are solved primarily by selecting the right vendor tool

Rationale: The chapter emphasizes that Agile, DevOps, and DataOps are fundamentally cultural practices.

Q41.

According to the chapter, should a data engineer know how to code?

A) No, because managed services eliminate coding needs
B) Only if working in research laboratories
C) Yes, production-grade software engineering is expected ✓ Correct
D) Only SQL is necessary; no other coding is needed

Rationale: The chapter clearly states that data engineers should have production-grade software engineering skills.

Q42.

Which language is described as the lingua franca of data?

A) Go
B) SQL ✓ Correct
C) Rust
D) R

Rationale: The chapter says SQL has reemerged as the lingua franca of data.

Q43.

Which language is described as the bridge between data engineering and data science?

A) Python ✓ Correct
B) COBOL
C) PHP
D) MATLAB

Rationale: Python is explicitly described as the bridge language between data engineering and data science.

Q44.

Which category best describes Java and Scala in the chapter's programming language framework?

A) Visualization languages
B) Primary data engineering languages ✓ Correct
C) Legacy-only languages
D) Business-user scripting languages

Rationale: The chapter lists JVM languages such as Java and Scala among the primary data engineering languages.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q45.

Which command-line language is identified as important for Linux-based productivity and scripting?

A) Ruby
B) bash ✓ Correct
C) Perl
D) Lua

Rationale: The chapter notes that bash is the command-line interface for Linux and valuable for scripting and workflow productivity.

Q46.

What does the “A” stand for in type A data engineer?

A) Analytics
B) Automation
C) Abstraction ✓ Correct
D) Architecture

Rationale: The chapter defines type A data engineers as those where A stands for abstraction.

Q47.

What does the “B” stand for in type B data engineer?

A) Build ✓ Correct
B) Business
C) Batch
D) Benchmark

Rationale: The chapter defines type B data engineers as those where B stands for build.

Q48.

An external-facing data engineer most commonly supports systems associated with which of the following?

A) Internal HR policy reviews
B) Customer-facing applications such as ecommerce platforms ✓ Correct
C) Academic peer review workflows
D) Paper-based archival processes

Rationale: The chapter describes external-facing data engineers as supporting applications like social media, IoT, and ecommerce platforms.

Q49.

Which role is described as designing the blueprint for organizational data management?

A) Data architect ✓ Correct
B) Data analyst
C) Project manager
D) AI researcher

Rationale: The chapter states that data architects design the blueprint for organizational data management.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q50.

Which C-suite role is responsible for a company's data assets and strategy according to the chapter?

A) Chief operating officer

B) Chief data officer ✓ Correct

C) Chief marketing officer

D) Chief financial officer

Rationale: The chapter identifies the chief data officer as responsible for the company's data assets and strategy.

Q51.

A hospital analytics team wants to reduce the time data scientists spend manually cleaning encounter data before building readmission models. Which data engineering action best addresses this problem according to the chapter?

A) Have data scientists continue preparing local extracts on their workstations

B) Automate collection, cleaning, and preparation of source data into reliable pipelines ✓ Correct

C) Ask executives to prioritize model tuning over data quality work

D) Replace analysts with machine learning engineers

Rationale: The chapter states that data engineers should automate the lower layers of data preparation so data scientists can focus on analytics and ML.

Q52.

A startup with only two people handling data has vague goals, many ad hoc requests, and no formal architecture. Which priority is most appropriate for its data engineer?

A) Launch a custom MLOps platform immediately

B) Build a solid data foundation aligned to business goals ✓ Correct

C) Adopt multiple bleeding-edge distributed frameworks at once

D) Specialize team members into narrow lifecycle roles

Rationale: In the starting-with-data stage, the data engineer should define architecture, audit data, and build a foundation for future use cases.

Q53.

A retail company with a few gigabytes of daily sales data is considering deploying and self-managing a large Hadoop cluster because competitors mention big data. What is the best recommendation?

A) Proceed because Hadoop is required for any modern analytics effort

B) Delay all analytics until petabyte scale is reached

C) Prefer simpler off-the-shelf or managed solutions unless custom complexity creates advantage ✓ Correct

D) Move all transformation work from SQL to MapReduce code

Rationale: The chapter warns against unnecessary technical complexity and recommends avoiding undifferentiated heavy lifting.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q54.

A data engineer is choosing between a managed cloud warehouse and a custom-built distributed processing stack for a standard BI workload. Which decision principle from the chapter is most relevant?

A) Prefer the tool with the most social media attention

B) Optimize for simplicity, cost-effectiveness, and business value ✓ Correct

C) Select the option requiring the most low-level infrastructure work

D) Choose the architecture that maximizes technical debt for speed

Rationale: Modern data engineers are expected to balance cost, simplicity, agility, and value rather than chase complexity.

Q55.

An organization has formal data practices and now needs robust pipelines that can support future ML use cases. Which data maturity stage best fits this scenario?

A) Starting with data

B) Scaling with data ✓ Correct

C) Leading with data

D) Pre-data warehousing

Rationale: Stage 2, scaling with data, emphasizes formal practices, scalable architectures, DataOps, and systems that support ML.

Q56.

A mature insurer has automated pipelines, self-service analytics, and seamless onboarding of new data sources. Leadership now wants better lineage, catalogs, and metadata management. This request most closely aligns with which stage?

A) Starting with data

B) Scaling with data

C) Leading with data ✓ Correct

D) Big data engineering only

Rationale: Stage 3 focuses on enterprisey capabilities such as governance, quality, catalogs, lineage, and metadata systems.

Q57.

A data engineer is asked to build a dashboard, define KPIs, train a recommendation model, and maintain ingestion pipelines. Based on the chapter, which task is least central to the typical data engineer role?

A) Maintaining ingestion pipelines

B) Preparing curated datasets for downstream use

C) Training the recommendation model directly ✓ Correct

D) Managing data systems across the lifecycle

Rationale: The chapter notes that data engineers typically do not directly build ML models, though they should understand the area.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q58.

A healthcare company is selecting a language for large-scale warehouse queries and transformations because analysts, engineers, and some streaming tools all need a common interface. Which language best fits the chapter's guidance?

A) SQL ✓ Correct
B) R
C) C++
D) Julia

Rationale: SQL is described as the lingua franca of modern data work and is broadly applicable across warehouses, lakes, and streaming systems.

Q59.

A data engineer needs a language that bridges orchestration code, data engineering tools, and data science frameworks in one environment. Which choice is most aligned with the chapter?

A) COBOL
B) Python ✓ Correct
C) MATLAB
D) PHP

Rationale: Python is described as the bridge language between data engineering and data science and is widely supported by modern tools.

Q60.

During a cloud migration, a team needs to inspect files, automate command-line tasks, and call operating system commands from orchestration workflows. Which skill is most directly useful?

A) Bash or an equivalent shell interface ✓ Correct
B) Only spreadsheet macros
C) Exclusive use of BI dashboards
D) Manual GUI-only administration

Rationale: The chapter emphasizes bash or equivalent CLI skills for scripting, OS operations, and pipeline support.

Q61.

A company has one engineer who mainly assembles managed services, SaaS tools, and standard products rather than building custom infrastructure. This engineer is best described as which type?

A) Type A data engineer ✓ Correct
B) Type B data engineer
C) Type A data scientist
D) Chief analytics officer

Rationale: Type A data engineers focus on abstraction and avoid undifferentiated heavy lifting by using off-the-shelf tools.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q62.

A streaming media company develops proprietary data tooling because its recommendation and event-processing needs are mission-critical and unique. Which role description best fits the engineer leading this work?

A) Type A data engineer
B) Type B data engineer ✓ Correct
C) Business analyst
D) Project manager

Rationale: Type B data engineers build custom tools and systems when doing so supports core competency and competitive advantage.

Q63.

A data engineer joins a company where business users submit mostly ad hoc requests and enthusiasm for data may fade without visible success. Which tactic is most appropriate?

A) Avoid stakeholder interaction until the architecture is complete
B) Pursue quick wins while planning to manage resulting technical debt ✓ Correct
C) Refuse all short-term requests to preserve purity of design
D) Delay delivery until an enterprise-wide governance office is formed

Rationale: The chapter recommends quick wins in early maturity stages but cautions that they often create technical debt that must be addressed.

Q64.

A stage 2 organization wants to scale faster, but engineers argue the main bottleneck is the number of cluster nodes. According to the chapter, what is the more likely bottleneck?

A) The data engineering team's throughput ✓ Correct
B) The number of BI dashboards
C) The age of the company
D) The quantity of ad hoc spreadsheets

Rationale: The chapter explicitly states that the main bottleneck for scaling is often the data engineering team rather than infrastructure capacity.

Q65.

A stage 3 company begins funding several technically impressive internal data projects that have no clear business use. What risk from the chapter does this illustrate?

A) Underinvestment in SQL
B) Complacency and technology distraction through hobby projects ✓ Correct
C) Insufficient command-line literacy
D) Failure to use any managed services

Rationale: At high maturity, the chapter warns against complacency and expensive technology distractions that do not deliver business value.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q66.

A data engineer is meeting with application developers before a new patient portal launches. What is the most important reason for this early collaboration?

A) To ensure the portal uses only open-source software

B) To understand source data volume, frequency, format, and compliance needs ✓ Correct

C) To transfer dashboard development to software engineers

D) To eliminate the need for data architecture

Rationale: The chapter stresses close work with software engineers to understand source systems and factors affecting the lifecycle, including security and regulation.

Q67.

A health app exposes customer-level usage data through an external query interface. Which concern is more prominent here than in an internal-only warehouse?

A) Multitenant query security and limits on user queries ✓ Correct

B) Whether analysts prefer spreadsheets

C) How to define internal KPIs

D) Whether SQL can parse JSON

Rationale: External-facing systems require special attention to concurrency, tight query limits, and multitenant security.

Q68.

A company primarily needs pipelines and warehouses for internal BI dashboards, reports, and data science work. Which orientation best describes the responsible data engineer?

A) Internal-facing ✓ Correct

B) External-facing

C) Research-facing

D) Investor-facing

Rationale: Internal-facing data engineers focus on pipelines and platforms that serve internal stakeholders and business processes.

Q69.

A data engineer is asked whether data engineering is merely a subdiscipline of data science. Which response best reflects the chapter's position?

A) Yes, because all data work belongs to data science

B) No, data engineering is separate but complementary and sits upstream from data science ✓ Correct

C) No, because data science sits upstream from engineering

D) Yes, but only in cloud environments

Rationale: The chapter argues that data engineering is distinct from data science and provides upstream inputs for downstream analytical work.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q70.

A chief executive asks why the organization should invest in ingestion, storage, and transformation before expanding AI efforts. Which explanation best aligns with the Data Science Hierarchy of Needs?

A) AI eliminates the need for foundational data work

B) Reliable lower-layer data capabilities are prerequisites for successful analytics and ML ✓ Correct

C) Model performance depends mainly on executive sponsorship

D) Data scientists should own all infrastructure layers themselves

Rationale: The hierarchy emphasizes that collection, movement, storage, and processing form the foundation required before advanced ML can succeed.

Q71.

A data scientist at a clinic spends most of the week manually joining extracts and cleaning records on a laptop. What does the chapter suggest this often indicates?

A) An ideal data science workflow

B) Immature data engineering and data science practices ✓ Correct

C) That machine learning is unnecessary

D) That only bigger laptops are needed

Rationale: The chapter notes that excessive manual prep by data scientists often reflects immature practices and insufficient automation.

Q72.

A university asks how to create a standard undergraduate pipeline for producing data engineers. Which statement best matches the chapter?

A) Most universities already follow a standard data engineering curriculum

B) Data engineering has a well-established licensure pathway

C) There is little formal standardized training, so self-study remains important ✓ Correct

D) Only doctoral programs can prepare data engineers

Rationale: The chapter states that formal training pathways are limited and self-study is a major part of entering the field.

Q73.

A database administrator wants to transition into data engineering. Based on the chapter, how should this background be viewed?

A) Poor preparation because database work is unrelated to data roles

B) Helpful because adjacent data-aware fields provide relevant context and skills ✓ Correct

C) Useful only for external-facing mobile applications

D) Insufficient unless paired with AI research experience

Rationale: The chapter identifies adjacent roles such as database administration as strong starting points for moving into data engineering.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q74.

A manager asks whether a modern data engineer still needs to code now that many services are managed. What is the best answer?

A) No, coding is obsolete in data engineering

B) Yes, production-grade software engineering remains essential despite higher abstraction ✓ Correct

C) Only SQL is needed; no other code matters

D) Coding matters only for on-premises Hadoop clusters

Rationale: The chapter explicitly states that data engineers should have production-grade software engineering skills even with managed services.

Q75.

A team must decide whether to implement a complex natural language text-stemming pipeline in SQL or native Spark code. Which choice best reflects the chapter's guidance?

A) Use SQL whenever possible, even if it is awkward and inefficient

B) Use native Spark when SQL is not the right tool for the job ✓ Correct

C) Avoid both and require manual preprocessing

D) Choose the option with the longest codebase

Rationale: The chapter emphasizes strong SQL proficiency while also recognizing when another tool, such as native Spark, is superior.

Q76.

A company wants to adopt Agile, DevOps, and DataOps by purchasing a single new platform. Which response best reflects the chapter's view?

A) That will work because these are mainly software procurement issues

B) These practices are fundamentally cultural and require organizational buy-in ✓ Correct

C) Only the data team needs to participate

D) They are relevant only in stage 3 organizations

Rationale: The chapter stresses that Agile, DevOps, and DataOps are cultural foundations, not problems solved by technology alone.

Q77.

A data engineer proposes a costly architecture without estimating time to value, ownership costs, or opportunity cost. Which business responsibility is being neglected?

A) Cost control ✓ Correct

B) Metadata management

C) External query concurrency

D) Model explainability

Rationale: The chapter highlights controlling costs, including time to value, total cost of ownership, and opportunity cost, as a core responsibility.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q78.
During vendor selection, a team favors a popular tool solely because several Silicon Valley firms use it. Which chapter-based critique is most appropriate?

A) This is acceptable because social proof is the best architecture criterion
B) Technology choices should be driven by customer and business value, not hype ✓ Correct
C) Only open-source projects should ever be considered
D) The company should first become a public cloud provider

Rationale: The chapter warns against adopting bleeding-edge tools based on social proof rather than practical value.

Q79.
A hospital system wants someone to design the blueprint for organizational data management, bridge technical and nontechnical stakeholders, and guide cloud migration strategy. Which role most closely matches this need?

A) Data architect ✓ Correct
B) Data analyst
C) Chief algorithms officer
D) Project manager

Rationale: The chapter describes data architects as blueprint designers and strategic bridges for organizational data management and major initiatives.

Q80.
In a smaller company without a formal data architect, who may need to assume many architecture responsibilities?

A) Only the CEO
B) The data engineer ✓ Correct
C) Only the data analyst
D) The external customer

Rationale: The chapter notes that depending on company size and maturity, a data engineer may overlap with or assume data architect responsibilities.

Q81.
A data engineer is building observability dashboards for pipeline health and coordinating with operations teams. Which upstream role is especially relevant because it both produces and may consume operational data?

A) DevOps engineers or SREs ✓ Correct
B) Product managers
C) Chief analytics officers
D) Business analysts only

Rationale: DevOps and SREs produce monitoring data and may also consume it downstream through dashboards and operational coordination.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q82.
A data analyst notices recurring inconsistencies in payer codes and collaborates with engineering to fix upstream transformations. Why is this collaboration valuable?

A) Analysts are usually the primary builders of production ML systems
B) Analysts possess domain expertise in definitions, characteristics, and quality issues ✓ Correct
C) Analysts own all cloud infrastructure decisions
D) Analysts replace the need for data governance

Rationale: The chapter emphasizes that analysts' subject-matter expertise is especially helpful for improving data quality.

Q83.
A machine learning engineer provisions scalable cloud resources for model training and deployment while also understanding TensorFlow infrastructure needs. How does the chapter position this role relative to data engineering?

A) As unrelated to data engineering
B) As overlapping with data engineering and data science ✓ Correct
C) As identical to a project manager
D) As a purely nontechnical leadership role

Rationale: The chapter describes ML engineering as overlapping with both data engineering and data science, especially in production systems.

Q84.
A chief information officer is leading an internal ERP modernization and wants input on data systems and cloud migration. Why would collaboration with data engineering leadership be appropriate?

A) Because the CIO is responsible for internal IT strategy and major information initiatives ✓ Correct
B) Because the CIO primarily manages external mobile app monetization
C) Because the CIO replaces all data analysts
D) Because the CIO only handles marketing analytics

Rationale: The chapter defines the CIO as the senior executive for internal IT, major initiatives, and information strategy.

Q85.
A chief technology officer asks the data team to help plan event capture from a consumer mobile app. Why is this request aligned with the CTO role described in the chapter?

A) The CTO focuses on internal payroll systems only
B) The CTO owns external-facing application strategy and architectures that generate key source data ✓ Correct
C) The CTO is mainly responsible for spreadsheet governance
D) The CTO typically replaces the chief data officer in privacy matters only

Rationale: The chapter describes the CTO as outward-facing and responsible for technologies such as mobile and web apps that create important data sources.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q86.

An executive is responsible for data assets, privacy, master data management, and overall data strategy. Which title best fits this description?

A) Chief data officer ✓ Correct

B) Chief operating officer

C) Chief algorithms officer

D) Chief project officer

Rationale: The chapter identifies the chief data officer as the executive focused on data assets, strategy, privacy, and related core functions.

Q87.

A company separates responsibility for delivering data technology from responsibility for analytics-driven decision-making. Which executive would most likely own the latter area?

A) Chief analytics officer ✓ Correct

B) Chief information officer

C) Chief technology officer

D) Chief reliability officer

Rationale: The chapter states that the chief analytics officer is responsible for analytics, strategy, and decision making for the business.

Q88.

A highly technical executive with deep machine learning research experience is setting the company's model R&D agenda and building research teams. Which role best matches this description?

A) Chief algorithms officer ✓ Correct

B) Chief information officer

C) Project manager

D) Data architect

Rationale: The chapter describes the chief algorithms officer as a technical C-suite role focused specifically on data science and ML research leadership.

Q89.

A multiyear migration from on-premises warehouses to a cloud data platform needs someone to filter requests, prioritize deliverables, and keep sprints on track. Which collaborator is most appropriate?

A) Project manager ✓ Correct

B) AI researcher

C) Data analyst

D) Chief executive officer

Rationale: Project managers are described as directing traffic, prioritizing deliverables, and coordinating long-running initiatives.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q90.

A company is launching a reusable patient-risk scoring dataset as a formal internal offering with ongoing enhancements and stakeholder priorities. Which nonengineering role is most directly associated with this data product?

A) Product manager ✓ Correct
B) Site-reliability engineer
C) Database auditor
D) Accounts payable manager

Rationale: The chapter notes that product managers increasingly oversee data products and balance business and customer needs with technology work.

Q91.

A data team is technically strong but repeatedly fails to gain adoption because it rarely speaks with clinicians, managers, or finance staff. Which chapter lesson is most relevant?

A) Success is primarily determined by having the newest tools
B) Communication across technical and nontechnical groups is critical to success ✓ Correct
C) Dashboards should be avoided until stage 3
D) Only external-facing engineers need stakeholder skills

Rationale: The chapter repeatedly emphasizes that communication and organizational understanding are often more decisive than technology alone.

Q92.

A health system wants to know whether its century-long history and large revenue automatically mean it has high data maturity. Which response is most accurate?

A) Yes, age and revenue are the main determinants of maturity
B) No, maturity depends more on how data is leveraged as a competitive advantage ✓ Correct
C) Yes, because older firms always have better governance
D) No, because only startups can be data mature

Rationale: The chapter explicitly states that data maturity does not simply depend on company age or revenue.

Q93.

A new data engineer wants to maximize long-term effectiveness in a fast-changing field. Which learning strategy best fits the chapter?

A) Ignore fundamentals and chase every new tool
B) Focus on fundamentals while tracking relevant new developments ✓ Correct
C) Specialize only in one deprecated framework
D) Wait for universities to standardize the field before learning more

Rationale: The chapter advises focusing on enduring fundamentals while staying aware of new technologies and paradigms.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q94.

A team debates whether modern data engineering should still center on monolithic frameworks or on modular interoperation among many tools. Which statement best reflects the chapter's view of the 2020s?

A) Data engineering is increasingly about connecting modular technologies to serve business goals ✓ Correct

B) All useful data work must return to self-managed Hadoop clusters

C) Only single-vendor monoliths can support the lifecycle

D) Interoperability is less important than ever

Rationale: The chapter characterizes current practice as modular, decentralized, and focused on interoperation across the lifecycle.

Q95.

An organization is rebuilding its data practices around privacy, anonymization, governance, and compliance with laws such as GDPR and CCPA. These concerns are best understood as what within the lifecycle model?

A) A separate optional phase after serving

B) Undercurrents spanning the entire lifecycle ✓ Correct

C) Tasks exclusive to data analysts

D) Issues limited to external-facing mobile apps

Rationale: Security and data management are described as undercurrents that cut across all lifecycle stages.

Q96.

A company historically emphasized cluster administration but now wants engineers to spend more time on governance, orchestration, and architecture. What broader shift does this illustrate?

A) A move from data lifecycle engineering back to pure BI reporting

B) A move from low-level big data maintenance toward higher-value lifecycle management ✓ Correct

C) A rejection of all managed services

D) A return to paper-based records management

Rationale: The chapter describes a shift away from low-level framework maintenance toward higher-value concerns across the data lifecycle.

Q97.

A team is deciding how to structure data engineering support. One proposal is a centralized service team handling requests from across the company; another is embedding engineers with specific products. Which chapter concept does this reflect?

A) The distinction between services and cross-functional models ✓ Correct

B) The difference between SQL and Python

C) The three V's of big data

D) The separation of storage and serving

Rationale: The chapter explains that data engineers often work either as centralized service teams or within cross-functional assignments.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q98.

A company wants to improve patient-facing analytics features that query a shared multitenant dataset while also supporting internal BI. Which recommendation best synthesizes the chapter's guidance?

A) Use one unrestricted query pattern for both audiences to simplify governance

B) Design for blended internal and external responsibilities with stricter external query controls and security ✓ Correct

C) Separate all engineers from software teams to reduce coordination overhead

D) Prioritize model building over source-system understanding

Rationale: The chapter notes that internal and external responsibilities are often blended, but external-facing systems require tighter concurrency and security controls.

Q99.

A data engineer in an early-stage company is tempted to build many custom components to demonstrate technical sophistication. Which recommendation from the chapter is strongest?

A) Build custom solutions only where they create competitive advantage ✓ Correct

B) Custom code is always preferable to managed services

C) Technical sophistication should take priority over business traction

D) Avoid any off-the-shelf tools during early maturity

Rationale: The chapter advises using turnkey solutions where possible and reserving custom builds for true competitive advantage.

Q100.

A chief executive asks for a concise description of data engineering's core mission. Which answer best matches the chapter's definition?

A) Designing dashboards and KPIs for executive reporting

B) Developing and maintaining systems that turn raw data into high-quality, consistent information for downstream use ✓ Correct

C) Researching novel neural network architectures

D) Replacing software engineering with data science

Rationale: The chapter defines data engineering as building and maintaining systems and processes that convert raw data into reliable information for analysis and machine learning.

Q101.

A company plans to modernize its analytics platform and is debating whether to define data engineering primarily by specific tools such as Hadoop and Spark. Based on the chapter’s definition, which framing is most defensible?

A) Data engineering is best defined by the dominant processing framework in use

B) Data engineering is the lifecycle management of systems and processes that convert raw data into high-quality information for downstream use ✓ Correct

C) Data engineering is a subdiscipline of data science focused on model feature preparation

D) Data engineering is synonymous with database administration and ETL maintenance

Rationale: The chapter defines data engineering holistically as developing, implementing, and maintaining systems and processes that transform raw data into high-quality information for downstream use cases.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q102.

An executive team wants to evaluate a proposed data platform solely by benchmark speed and storage scale. Which critique best reflects the lifecycle perspective described in the chapter?

A) Platform evaluation should focus first on whether the technology is popular in the market

B) The key question is whether the platform minimizes SQL usage across the organization

C) Technology should be assessed in terms of how it supports data through the lifecycle and end business goals, not as an end in itself ✓ Correct

D) The platform should be selected by the data science team because they are the primary data consumers

Rationale: The lifecycle shifts attention away from technology fixation and toward the data itself and the end goals it must serve.

Q103.

A startup is considering a large investment in a custom distributed processing engine to handle a few gigabytes of daily data. Which recommendation is most aligned with the chapter's evaluation of past big data practices?

A) Proceed, because all data platforms should be designed for petabyte scale from day one

B) Proceed, because custom systems are always the fastest way to achieve maturity

C) Avoid the investment unless it creates competitive advantage, because adopting complex big data tools for small data problems is often wasteful ✓ Correct

D) Avoid SQL entirely and require engineers to write only low-level distributed code

Rationale: The chapter warns against using big data tooling for small data problems and advises avoiding undifferentiated heavy lifting unless custom work yields competitive advantage.

Q104.

A data leader claims that the decline of the term "big data" means large-scale processing is no longer important. Which interpretation is most accurate?

A) The need for large-scale processing disappeared when Hadoop adoption fell

B) The term faded mainly because processing large data became more accessible through simplification and abstraction ✓ Correct

C) Large-scale processing is now exclusively the responsibility of data scientists

D) The term disappeared because relational databases replaced all distributed systems

Rationale: The chapter explains that big data lost steam largely because tooling became simpler and more accessible, not because large-scale data challenges vanished.

Q105.

A data engineer is asked to justify why privacy and governance discussions now consume more attention than low-level cluster tuning. Which explanation best synthesizes the chapter's view of the 2020s role shift?

A) Modern tools have reduced some infrastructure burden, allowing engineers to focus higher in the value chain on governance, security, and lifecycle management ✓ Correct

B) Data engineers no longer need technical skills because vendors manage everything

C) Governance replaced architecture as the only relevant competency in data work

D) Cluster tuning is obsolete because all systems now have identical performance characteristics

Rationale: The chapter describes a shift from low-level framework maintenance toward higher-value concerns such as security, data management, DataOps, and lifecycle management.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q106.

A chief data officer argues that data engineering should report under data science because machine learning creates the most visible value. Which response best matches the chapter's conceptual model?

A) Agreed, because data engineering is a minor branch of analytics

B) Agreed, because data scientists own the full hierarchy of data needs

C) Disagree, because data engineering is distinct from data science and sits upstream by providing the inputs used downstream ✓ Correct

D) Disagree, because data engineering should be isolated from all analytical functions

Rationale: The chapter explicitly treats data engineering as separate from but complementary to data science, with an upstream role in supplying usable data.

Q107.

A firm complains that its data scientists spend most of their time cleaning and gathering data rather than experimenting with models. According to the chapter, what organizational conclusion is most justified?

A) The data scientists are underperforming because model tuning should always begin immediately

B) The company likely lacks a sufficiently strong data engineering foundation to automate lower-level data work ✓ Correct

C) The best solution is to prohibit data scientists from accessing raw data

D) The company should replace SQL with spreadsheets for faster preparation

Rationale: The chapter argues that excessive time spent on gathering and cleaning data often reflects immature data engineering and insufficient foundational support.

Q108.

A hiring committee wants one role to build dashboards, define KPIs, create ML models, and engineer pipelines. Which evaluation best reflects the chapter's role boundaries?

A) This is appropriate because data engineers typically own all downstream analytical outputs

B) This is inappropriate because data engineers should understand these functions but do not typically directly perform all of them ✓ Correct

C) This is inappropriate only in highly mature companies

D) This is appropriate if the company uses cloud services

Rationale: The chapter notes that data engineers should understand analytics and ML needs but typically do not directly build models, dashboards, KPIs, or software applications.

Q109.

A manager says company age and revenue are the best proxies for data maturity. Which rebuttal is most consistent with the chapter?

A) Data maturity is determined primarily by annual IT spending

B) Data maturity depends mainly on the number of data engineers employed

C) Data maturity is better understood as how effectively data is utilized and integrated as a competitive advantage ✓ Correct

D) Data maturity is equivalent to cloud adoption level

Rationale: The chapter states that data maturity is about progression in data utilization, capabilities, and integration, not simply age or revenue.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q110.

A small organization with loosely defined goals asks whether it should prioritize machine learning before building pipelines and architecture. Which recommendation best aligns with the chapter's analysis of stage 1 maturity?

A) Prioritize ML first because early wins are impossible without advanced models

B) Delay all reporting until a full governance program is complete

C) Focus first on stakeholder buy-in, architecture definition, data auditing, and a solid foundation before major ML efforts ✓ Correct

D) Adopt a large Hadoop cluster to accelerate experimentation

Rationale: For organizations starting with data, the chapter recommends establishing buy-in, architecture, data auditing, and foundational systems before pursuing substantial ML initiatives.

Q111.

A stage 1 company delivers several quick dashboard wins but accumulates brittle pipelines and inconsistent definitions. What is the most accurate evaluation?

A) This is acceptable only if there is a plan to reduce the resulting technical debt ✓ Correct

B) This is ideal because technical debt improves agility permanently

C) This indicates the company has already reached stage 3 maturity

D) This proves the organization should abandon quick wins entirely

Rationale: The chapter acknowledges that quick wins can create technical debt and stresses having a plan to reduce that debt.

Q112.

During stage 2 growth, a team wants to adopt a fashionable technology because leading Silicon Valley firms use it. What decision principle from the chapter should govern the choice?

A) Adopt it if it is open source and code-first

B) Adopt it if it reduces SQL dependence

C) Adopt it only if it delivers clear value to customers and supports business goals ✓ Correct

D) Adopt it because social proof reduces architectural risk

Rationale: The chapter warns against adopting bleeding-edge technologies based on social proof and emphasizes value-driven decisions.

Q113.

A scaling company is struggling to meet growing data demand and assumes the main bottleneck is insufficient storage nodes. Which assessment best fits the chapter?

A) The primary bottleneck in scaling is often the data engineering team's throughput rather than raw infrastructure ✓ Correct

B) Storage is always the dominant bottleneck in stage 2 organizations

C) The issue is that analysts should own pipeline operations

D) The solution is to move all workloads to spreadsheets

Rationale: The chapter explicitly states that the main bottleneck for scaling is usually not cluster nodes or storage but the data engineering team.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q114.

A mature enterprise at stage 3 wants to invest heavily in a custom metadata and governance platform. Which condition most strongly justifies this according to the chapter?

A) The project uses the newest vendor-neutral buzzwords

B) The platform directly strengthens data as a competitive advantage and supports enterprisey needs such as governance and quality ✓ Correct

C) The project eliminates the need for collaboration with analysts

D) The project replaces all managed services with self-hosted systems

Rationale: Stage 3 organizations are encouraged to build custom tools when they leverage data as a competitive advantage and support management, governance, and quality.

Q115.

A stage 3 company has stable self-service analytics and automated pipelines, so leadership reduces investment in maintenance and improvement. What risk does the chapter highlight most directly?

A) The company will lose all SQL capability immediately

B) The company may become complacent and regress to a lower maturity stage ✓ Correct

C) The company will necessarily need to abandon governance

D) The company will no longer need data engineers

Rationale: The chapter identifies complacency at stage 3 as a major danger that can cause organizations to fall back to a lower maturity level.

Q116.

A career changer from software engineering asks why the transition into data engineering is often smoother from adjacent fields. Which rationale is most accurate?

A) Adjacent fields provide data awareness, organizational context, and transferable technical skills relevant to data engineering problems ✓ Correct

B) Adjacent fields eliminate the need for self-study in data engineering

C) Only software engineers can become data engineers successfully

D) Adjacent fields are useful only because they require less communication with stakeholders

Rationale: The chapter notes that adjacent fields such as software engineering, ETL, DBA, data science, and analysis provide data awareness and relevant technical context.

Q117.

A technically strong engineer dismisses communication and requirement gathering as secondary to coding ability. Which evaluation best reflects the chapter's business responsibilities for data engineers?

A) This is appropriate in highly technical organizations

B) This is risky because success often depends more on communication, requirements, and organizational navigation than on technology alone ✓ Correct

C) This is acceptable if the engineer knows Python and SQL well

D) This is ideal for external-facing data engineering only

Rationale: The chapter emphasizes that communication, scoping requirements, and understanding the organization are critical and that success or failure is rarely just a technology issue.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q118.

A team treats Agile, DevOps, and DataOps as purely tool selection problems. Which critique is most consistent with the chapter?

A) They are mainly procurement decisions

B) They are fundamentally cultural practices requiring organization-wide buy-in, not merely technology adoption ✓ Correct

C) They apply only to software engineering, not data engineering

D) They are relevant only after stage 3 maturity is reached

Rationale: The chapter explicitly characterizes Agile, DevOps, and DataOps as fundamentally cultural rather than purely technological.

Q119.

A director asks how a data engineer should think about cost management when selecting a platform. Which answer best reflects the chapter's guidance?

A) Choose the cheapest option regardless of time to value

B) Optimize for low monthly cost only, because opportunity cost is too abstract to measure

C) Balance time to value, total cost of ownership, and opportunity cost while monitoring costs continuously ✓ Correct

D) Ignore cost until the platform is in production

Rationale: The chapter advises data engineers to control costs by optimizing time to value, total cost of ownership, and opportunity cost and by monitoring costs.

Q120.

A senior engineer argues that managed services have made software engineering skill unnecessary in data engineering. Which response best matches the chapter's evaluation?

A) Correct, because orchestration frameworks eliminate the need for production-grade code

B) Correct, because SQL is now the only required skill

C) Incorrect, because abstraction reduces some low-level work but production-grade software engineering remains essential ✓ Correct

D) Incorrect, because all data engineers must now write distributed storage engines from scratch

Rationale: The chapter states that managed services reduce low-level effort, but data engineers still need production-grade software engineering skills.

Q121.

A team must choose one language to prioritize for broad interoperability across databases, data lakes, and many modern processing engines. Which choice is most defensible based on the chapter?

A) SQL, because it has reemerged as the common interface and lingua franca of data ✓ Correct

B) C++, because it is the default language of managed cloud services

C) R, because it is the primary production language for data architecture

D) JavaScript, because it replaced SQL in analytics workflows

Rationale: The chapter describes SQL as the most common interface for databases and data lakes and the reemergent lingua franca of data.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q122.

A data engineer proposes writing a highly complex SQL query to perform raw text stemming and tokenization in an NLP pipeline. What judgment best reflects the chapter's view of tool selection?

A) This is ideal because SQL should always be used if possible

B) This is acceptable only if Python is unavailable in the company

C) This is poor tool choice because a proficient engineer should recognize when a non-SQL alternative such as native Spark is superior ✓ Correct

D) This is required for production-grade engineering

Rationale: The chapter explicitly notes that while SQL is powerful, proficient engineers recognize when another tool, such as native Spark, is a better fit.

Q123.

A candidate is highly proficient in SQL and Python but has no command-line experience. Which concern is most justified from the chapter's perspective?

A) Lack of bash or equivalent CLI skills may reduce productivity in scripting, OS operations, and pipeline workflows ✓ Correct

B) Command-line skills are obsolete in modern data engineering

C) CLI knowledge matters only for ML engineers

D) Without bash, the candidate cannot learn SQL management frameworks

Rationale: The chapter highlights bash or PowerShell proficiency as important for scripting, OS operations, and efficient workflows.

Q124.

A training program focuses almost entirely on the newest vendor tools and neglects fundamentals. Which evaluation best matches the chapter's advice on keeping pace?

A) This is ideal because fundamentals change too slowly to matter

B) This is incomplete because engineers should anchor in fundamentals while tracking new developments for lifecycle relevance ✓ Correct

C) This is sufficient if the tools are cloud-native

D) This is appropriate only for type B engineers

Rationale: The chapter advises focusing on fundamentals to understand what will not change while staying current on relevant new developments.

Q125.

A company wants to minimize engineering effort by assembling managed services and off-the-shelf tools rather than building custom infrastructure. Which data engineer profile is most closely aligned with this strategy?

A) Type A, because the emphasis is abstraction and avoiding undifferentiated heavy lifting ✓ Correct

B) Type A, because the emphasis is advanced ML experimentation

C) Type B, because the emphasis is custom systems from the outset

D) Type B, because the emphasis is replacing managed services with proprietary frameworks

Rationale: Type A data engineers focus on abstraction and using off-the-shelf products and managed services to avoid unnecessary heavy lifting.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q126.

A business has reached stage 2 maturity and now needs specialized systems tightly tied to its core competitive advantage. Which staffing inference is most consistent with the chapter?

A) The company should avoid any specialization and keep only generalists

B) Type B data engineering capabilities become more relevant as custom systems are needed ✓ Correct

C) The company should replace data engineers with project managers

D) Type A engineers are no longer useful in any capacity

Rationale: The chapter associates type B data engineers with building custom tools and systems, especially in stage 2 and 3 organizations.

Q127.

A team is debating whether type A and type B data engineers are mutually exclusive organizational categories. Which interpretation best reflects the chapter?

A) They are rigid and mutually exclusive career tracks

B) They are simplistic distinctions, and the same person or company may involve both profiles ✓ Correct

C) They correspond exactly to internal-facing and external-facing roles

D) They map directly onto data scientist type A and B roles without modification

Rationale: The chapter states that type A and type B distinctions are simplifications and may coexist within the same company or person.

Q128.

An ecommerce firm needs a data engineer to support customer-facing query capabilities over multitenant data. Which design concern deserves heightened attention compared with internal-facing systems?

A) Minimizing the number of business stakeholders

B) Strict limits on user queries and stronger security controls due to concurrency and multitenancy risks ✓ Correct

C) Eliminating orchestration to reduce latency

D) Replacing all SQL interfaces with spreadsheets

Rationale: The chapter notes that external-facing systems often require tighter query limits and more complex security, especially with multitenant data and high concurrency.

Q129.

A manager says internal-facing data engineering can be planned independently from any external-facing data concerns. Which response is most consistent with the chapter?

A) Correct, because internal and external data systems are unrelated

B) Correct, because external-facing systems do not depend on internal data preparation

C) Incorrect, because responsibilities are often blended and internal-facing data is usually a prerequisite to external-facing data ✓ Correct

D) Incorrect only in stage 1 companies

Rationale: The chapter explains that internal- and external-facing responsibilities are often blended and that internal-facing data is usually a prerequisite for external-facing data.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q130.

A new data engineer wants to understand why collaboration with data architects matters even in cloud-first environments. Which explanation best synthesizes the chapter's position?

A) Cloud environments remove the need for architecture, so collaboration is optional

B) Architects still guide overall strategy, governance, and major initiatives, even though cloud implementation has made some decisions more fluid ✓ Correct

C) Architects are relevant only for on-premises procurement cycles

D) Architects primarily maintain dashboards for executives

Rationale: The chapter says cloud has shifted boundaries, but data architects remain influential in strategy, governance, and major architectural initiatives.

Q131.

A data engineering team is repeatedly surprised by schema changes and missing event fields from application systems. Which upstream collaboration failure is most evident?

A) Insufficient coordination with software engineers about source-system data generation characteristics ✓ Correct

B) Too much collaboration with data analysts

C) Overreliance on product managers for sprint planning

D) Excessive focus on downstream self-service tools

Rationale: The chapter stresses that data engineers must work closely with software engineers to understand application-generated data volume, frequency, format, and related constraints.

Q132.

An operations team produces monitoring metrics and consumes reliability dashboards built from the same data platform. How should its relationship to data engineering be interpreted?

A) Strictly downstream only

B) Strictly upstream only

C) Both upstream and downstream, depending on whether it is producing operational data or consuming data products ✓ Correct

D) Outside the data engineering lifecycle entirely

Rationale: The chapter notes that DevOps and SRE roles may be upstream as data producers and also downstream as data consumers or collaborators.

Q133.

A data science group develops models locally on single workstations, downsamples data heavily, and struggles to deploy to production. Which intervention best reflects the chapter's rationale for data engineering support?

A) Increase model complexity to compensate for smaller samples

B) Provide scalable, automated data and production infrastructure so data science work is not constrained by local tooling ✓ Correct

C) Move all modeling tasks to spreadsheets

D) Require data scientists to manage source-system schemas directly

Rationale: The chapter argues that data engineering should provide automation and scale so data scientists are not bottlenecked by local environments and manual preparation.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q134.

A data analyst identifies recurring semantic inconsistencies in a sales dataset and wants to improve trust in executive reporting. Which collaborative role is most supported by the chapter?

A) The analyst should avoid involvement because data quality is solely an engineering issue

B) The analyst's domain expertise is valuable for improving data quality in collaboration with data engineers ✓ Correct

C) Only the CIO should resolve data definition issues

D) The issue should be deferred until a machine learning team is formed

Rationale: The chapter highlights data analysts' subject-matter expertise as especially valuable in improving data quality alongside data engineers.

Q135.

A company is confused about boundaries among data engineering, data science, and ML engineering in production systems. Which distinction is most accurate according to the chapter?

A) ML engineers focus solely on dashboard design and KPI reporting

B) Data engineers and ML engineers never overlap in infrastructure responsibilities

C) ML engineers emphasize training and operating ML systems at scale, while boundaries with data engineering and data science can be blurry ✓ Correct

D) Data scientists are primarily responsible for cloud migrations

Rationale: The chapter states that ML engineers build and operate ML infrastructure at scale and that their boundaries with data engineering and data science are blurry.

Q136.

An AI research lab with abundant funding wants to maximize researcher productivity. Based on the chapter, which organizational support model is most consistent?

A) Researchers should each self-manage all engineering and infrastructure tasks

B) Highly specialized researchers often work with supporting engineering teams that facilitate their work ✓ Correct

C) Researchers should report only to project managers and not engineers

D) Researchers should avoid production concerns entirely, even in applied settings

Rationale: The chapter notes that well-funded AI research organizations typically pair specialized researchers with supporting engineering teams.

Q137.

A CEO asks the data engineering lead for input on a potential cloud migration. Why is this request consistent with the chapter's view of the role?

A) Because data engineers mainly own board governance and investor relations

B) Because data engineers maintain visibility into available data and are often heavily involved in major architectural changes such as cloud migrations ✓ Correct

C) Because CEOs are expected to choose low-level frameworks directly

D) Because cloud migration is primarily a data science modeling issue

Rationale: The chapter states that data engineers are often heavily involved in major architectural initiatives, including cloud migrations and new data systems.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q138.

An internal executive role must align IT strategy with business processes and shape data culture in a less mature organization. Which role best fits this description?

A) Chief technology officer
B) Chief algorithms officer
C) Chief information officer ✓ Correct
D) Chief analytics officer

Rationale: The chapter describes the CIO as the senior internal-facing executive responsible for IT, business processes, and often shaping data culture in less mature organizations.

Q139.

A company's senior technology leader is primarily responsible for outward-facing application strategy and architectures that generate critical source data. Which role is this?

A) Chief technology officer ✓ Correct
B) Chief information officer
C) Chief data officer
D) Chief analytics officer

Rationale: The chapter defines the CTO as the outward-facing executive owning technology strategy and architectures for external-facing applications.

Q140.

An executive role is responsible for the company's data assets and strategy, with emphasis on business utility, privacy, and functions such as master data management. Which role best matches?

A) Chief analytics officer
B) Chief data officer ✓ Correct
C) Chief technology officer
D) Chief operating officer

Rationale: The chapter identifies the CDO as responsible for data assets and strategy, including privacy and master data management.

Q141.

A company has both a CDO and a CAO. Which division of responsibility is most consistent with the chapter?

A) The CDO focuses on analytics and decision making, while the CAO focuses on storage procurement
B) The CDO focuses on technology and organization for delivering data, while the CAO focuses on analytics, strategy, and decision making ✓ Correct
C) The CDO owns external applications, while the CAO owns internal IT
D) The roles are identical and differ only by title preference

Rationale: The chapter distinguishes the CDO as focused on delivering data and the CAO as focused on analytics, strategy, and business decision making.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q142.

A company appoints a highly technical executive with deep ML research experience to set research agendas and provide technical leadership for algorithmic initiatives. Which role is being described?

A) Chief analytics officer
B) Chief information officer
C) Chief algorithms officer ✓ Correct
D) Chief data officer

Rationale: The chapter describes the chief algorithms officer as a highly technical C-suite role focused specifically on data science and ML research and leadership.

Q143.

A multiyear warehouse migration has many stakeholders, competing requests, sprint planning needs, and changing business priorities. Which role is best positioned to filter requests and prioritize deliverables to keep the initiative on track?

A) Project manager ✓ Correct
B) Data analyst
C) AI researcher
D) Chief algorithms officer

Rationale: The chapter describes project managers as directing traffic, prioritizing deliverables, and balancing business demands during major initiatives.

Q144.

A team is building a reusable customer-360 dataset intended to serve multiple business users as a maintained deliverable. Which management role is most directly associated with ownership of this kind of data product?

A) Project manager
B) Product manager ✓ Correct
C) Site-reliability engineer
D) Chief information officer

Rationale: The chapter states that product managers oversee product development and that, in this context, such deliverables are data products.

Q145.

A data engineering leader wants to remind the team why technical excellence alone is insufficient. Which statement best captures the chapter's final organizational message?

A) Engineers are hired mainly to code independently with minimal stakeholder interaction
B) Engineers should understand the problems, tools, and people they serve rather than working in isolation ✓ Correct
C) Only executives need to understand business context
D) Technical mastery eliminates the need for organizational awareness

Rationale: The chapter concludes that engineers must deeply understand the problems, technologies, and people involved, not just code in isolation.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q146.
A healthcare startup with a tiny data team asks whether it should build custom pipelines and infrastructure whenever possible to demonstrate technical sophistication. Which recommendation best reflects the chapter?

A) Yes, because custom engineering is the fastest path to maturity in early-stage organizations
B) Yes, because off-the-shelf tools prevent competitive differentiation
C) No, because early-stage teams should avoid undifferentiated heavy lifting and build custom only when it creates competitive advantage ✓ Correct
D) No, because stage 1 organizations should not build any pipelines at all

Rationale: The chapter advises early-stage organizations to avoid unnecessary complexity and reserve custom building for true competitive advantage.

Q147.
A hospital analytics group wants to improve model deployment success for its data scientists. The scientists currently spend most of their time wrangling extracts from source systems. Which strategic decision is most aligned with the chapter?

A) Hire more dashboard developers and postpone data engineering investment
B) Strengthen data engineering to automate collection, cleaning, processing, and scalable delivery of data for downstream modeling ✓ Correct
C) Require data scientists to become database administrators before building models
D) Shift all modeling to ad hoc spreadsheets to reduce engineering overhead

Rationale: The chapter presents data engineering as the foundation that automates lower-level data work and enables data science success in production.

Q148.
A mature payer organization is deciding between a flashy custom platform initiative and incremental investments in lineage, cataloging, governance, and quality controls. Which recommendation best fits the chapter's stage 3 priorities?

A) Favor the flashy platform because mature organizations should pursue expensive hobby projects
B) Favor the governance-oriented investments because stage 3 emphasizes enterprisey data management and broad dissemination of trustworthy data ✓ Correct
C) Avoid metadata tools because they are only useful in stage 1
D) Replace self-service analytics with centralized manual reporting

Rationale: The chapter states that stage 3 organizations should emphasize governance, quality, metadata, lineage, catalogs, and broad organizational dissemination of data.

Q149.
A university asks whether a formal standard path currently exists for preparing data engineers. Which conclusion is most faithful to the chapter?

A) Yes, most universities now offer a standard data engineering curriculum
B) Yes, the field is fully standardized through boot camps and certifications
C) No, the field lacks a standard formal path, so entrants should expect significant self-study ✓ Correct
D) No, because data engineering can only be learned through database administration roles

Rationale: The chapter notes that there is little formal training and no standard curriculum, making self-study essential.

Fundamentals of Data Engineering 1st edition Reis Test Bank

Q150.

A data platform committee must choose between two proposals: one optimizes for maximal technical sophistication, and the other emphasizes simplicity, interoperability, cost control, and agility across the lifecycle. Which proposal better reflects the chapter's synthesis of modern data engineering?

A) The first, because sophistication is the primary marker of engineering quality

B) The first, because modern data engineers should prioritize monolithic frameworks

C) The second, because modern data engineering emphasizes balancing cost, agility, scalability, simplicity, reuse, and interoperability to deliver business value ✓ Correct

D) Neither, because data engineering should avoid business considerations

Rationale: The chapter frames modern data engineering as a balancing act across cost, agility, scalability, simplicity, reuse, and interoperability in service of business value.

Fundamentals of Data Engineering 1st edition Reis Test Bank