Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Designing Data-Intensive Applications

ISBN 9781098119065 | Chapter 1: Trade-Offs in Data Systems Architecture

Total Questions: 150

Multiple Choice

Q1.
In the chapter, an application is described as data-intensive when which factor is a primary development challenge?

A) Parallelizing a very large numerical computation
B) Managing data storage, processing, consistency, and availability ✓ Correct
C) Rendering complex user interfaces on client devices
D) Compiling code for multiple operating systems

Rationale: Data-intensive applications are defined by data management challenges such as storage, consistency, failures, and availability.

Q2.
For a web application, what term refers to the server-side code that handles user requests?

A) Frontend
B) Client cache
C) Backend ✓ Correct
D) Browser runtime

Rationale: The backend is the server-side code responsible for handling user requests.

Q3.
Which protocol is identified as a common way for a backend service to be reachable?

A) SMTP
B) HTTP ✓ Correct
C) FTP
D) SNMP

Rationale: The chapter states that backend services are often reachable via HTTP.

Q4.
What does it mean when application code is described as stateless?

A) It stores all data only in RAM
B) It encrypts every request before processing
C) It forgets request-specific information after handling a request ✓ Correct
D) It never communicates with a database

Rationale: Stateless application code does not retain information about a request once that request has been handled.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q5.

Which group is primarily associated with generating reports to help management make better decisions?

A) Frontend developers
B) Business analysts ✓ Correct
C) Network administrators
D) Compliance auditors

Rationale: Business analysts generate reports for decision support and business intelligence.

Q6.

Operational systems are best described as systems in which data is primarily what?

A) Archived for long-term legal retention
B) Created and modified by application code serving users ✓ Correct
C) Stored only as files without schema
D) Queried only through ad-hoc analyst SQL

Rationale: Operational systems are where application code reads and modifies data created through user interactions.

Q7.

Analytical systems typically contain which kind of data relative to operational systems?

A) A read-only copy optimized for analytics ✓ Correct
B) Only encrypted backups for disaster recovery
C) Only unstructured logs with no transformations
D) The authoritative source for all user updates

Rationale: Analytical systems usually contain a read-only copy of data from operational systems, optimized for analysis.

Q8.

What term is used in the chapter to refer loosely to low-latency reads and writes?

A) Checkpoint
B) Transaction ✓ Correct
C) Snapshot
D) Batch job

Rationale: In this chapter, transaction is used loosely to mean low-latency reads and writes.

Q9.

Looking up a small number of records by a key is called what?

A) Vectorization
B) Feature engineering
C) Point query ✓ Correct
D) Bulk import

Rationale: A point query fetches individual records by key.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q10.

The acronym OLTP stands for what?

A) Online logical transfer protocol
B) Operational load tracking process
C) Online transaction processing ✓ Correct
D) Optimized latency transaction platform

Rationale: OLTP is the standard abbreviation for online transaction processing.

Q11.

The acronym OLAP stands for what?

A) Online analytical processing ✓ Correct
B) Operational analytics platform
C) Open application protocol
D) Online archival processing

Rationale: OLAP stands for online analytical processing.

Q12.

Which read pattern is most characteristic of OLTP systems?

A) Aggregate scans over many records
B) Point queries fetching individual records by key ✓ Correct
C) Exploratory scans of historical event logs only
D) Large-scale matrix transformations

Rationale: OLTP systems are characterized by point queries that fetch individual records by key.

Q13.

Which write pattern is most characteristic of analytical systems in the chapter's comparison table?

A) Frequent create, update, and delete of individual records
B) Bulk import through ETL or ingestion from an event stream ✓ Correct
C) Direct user edits to historical records from dashboards
D) Only synchronous writes from mobile apps

Rationale: Analytical systems commonly receive data through bulk import or event streams rather than individual record updates.

Q14.

In the chapter's comparison, OLAP queries are typically described as what type?

A) Fixed and predefined by application code
B) Arbitrary and ad hoc for exploration ✓ Correct
C) Restricted to key-value lookups
D) Limited to maintenance operations

Rationale: Analytical systems support arbitrary, ad-hoc exploratory queries by analysts.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q15.

Which statement best characterizes the data represented in OLTP systems?

A) Historical events over long periods
B) The latest state at the current point in time ✓ Correct
C) Only denormalized aggregates
D) Only immutable append-only logs

Rationale: OLTP systems generally represent the latest current state of the data.

Q16.

Which statement best characterizes the data represented in OLAP systems?

A) Only current account balances
B) Only authorization checks for transactions
C) History of events that happened over time ✓ Correct
D) Only data entered manually by analysts

Rationale: OLAP systems commonly represent historical events over time for analysis.

Q17.

Systems such as Pinot, Druid, and ClickHouse are cited as examples of platforms for which type of use?

A) Offline backup archiving
B) Product analytics or real-time analytics ✓ Correct
C) Client-side mobile synchronization
D) Static website hosting

Rationale: The chapter identifies Pinot, Druid, and ClickHouse as systems for product or real-time analytics.

Q18.

Compared with traditional OLAP systems, product analytics systems are especially optimized for what?

A) Tape-based archival retention
B) Low-latency query responses on real-time ingested data ✓ Correct
C) Manual schema design without ingestion
D) Batch-only monthly reporting

Rationale: Product analytics systems ingest data in real time and optimize for low-latency queries.

Q19.

What is the name for a separate database used for analytics instead of querying OLTP systems directly?

A) Data warehouse ✓ Correct
B) Message queue
C) Object store
D) Service mesh

Rationale: A data warehouse is a separate database used for analytical workloads.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q20.
What term describes the problem in which relevant data is spread across multiple operational systems and is difficult to combine?

A) Data silos ✓ Correct
B) Hot partitioning
C) Write skew
D) Schema drift

Rationale: The chapter names this problem data silos.

Q21.
ETL stands for what sequence of steps?

A) Extract, transform, load ✓ Correct
B) Encode, transfer, log
C) Evaluate, test, launch
D) Encrypt, tokenize, lock

Rationale: ETL is the process of extract, transform, and load.

Q22.
When transformation is performed after loading into the warehouse, the process is called what?

A) TTL
B) ELT ✓ Correct
C) TLE
D) LTE

Rationale: If load occurs before transform, the process is called ELT.

Q23.
Specialist services such as Fivetran, Singer, and Airbyte are described as helping primarily with what task?

A) GPU scheduling for ML training
B) Tracing calls across microservices
C) ETL from external SaaS APIs into a warehouse ✓ Correct
D) Encrypting local browser storage

Rationale: These tools are presented as data connector services for bringing SaaS API data into a warehouse.

Q24.
What does HTAP stand for?

A) High-throughput application protocol
B) Hybrid transactional/analytical processing ✓ Correct
C) Hosted transaction and persistence
D) Hierarchical table access plan

Rationale: HTAP stands for hybrid transactional/analytical processing.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q25.
According to the chapter, HTAP does not replace data warehouses primarily because enterprises commonly have what pattern?

A) One operational database and many warehouses
B) Many operational databases and typically a single warehouse ✓ Correct
C) No operational databases and only streams
D) Only file-based analytics with no SQL systems

Rationale: The chapter notes that enterprises often have many operational databases but usually one data warehouse for combined analysis.

Q26.
What is a data lake?

A) A centralized file-based repository holding any potentially useful analytical data ✓ Correct
B) A transactional database optimized for point queries
C) A cache layer in front of an OLTP database
D) A compliance archive containing only deleted records

Rationale: A data lake is defined as a centralized repository of files containing any data that may be useful for analysis.

Q27.
Which characteristic distinguishes a data lake from a traditional data warehouse in the chapter?

A) A data lake supports only SQL queries
B) A data lake imposes no particular file format, data model, or schema ✓ Correct
C) A data lake stores only relational tables
D) A data lake cannot contain raw operational data

Rationale: Unlike a warehouse, a data lake does not impose a single schema, file format, or data model.

Q28.
The phrase 'raw data is better' is referred to by what label?

A) The lakehouse principle
B) The normalization rule
C) The sushi principle ✓ Correct
D) The star schema doctrine

Rationale: The chapter calls this idea the sushi principle.

Q29.
Which regulation is specifically cited as giving many European residents greater control over personal data since 2018?

A) HIPAA
B) SOX
C) GDPR ✓ Correct
D) FERPA

Rationale: The GDPR is identified as the regulation granting greater control over personal data in Europe.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q30.
What is reverse ETL in the context of the chapter?

A) Deleting source data after warehouse loading
B) Moving outputs from analytical systems back into operational systems ✓ Correct
C) Converting OLTP schemas into normalized form
D) Replacing SQL with Python notebooks

Rationale: Reverse ETL refers to making outputs of analytical systems available to operational systems.

Q31.
A system of record is best defined as which of the following?

A) A temporary store for denormalized query results
B) A canonical and authoritative source of truth for data ✓ Correct
C) A read-only copy of historical events for analysts
D) A backup replica used only during outages

Rationale: A system of record holds the authoritative, canonical version of data.

Q32.
Which of the following is a classic example of a derived data system?

A) A cache ✓ Correct
B) A user's original order entry form
C) A canonical employee master record
D) A legal consent document repository

Rationale: The chapter explicitly identifies a cache as a classic example of derived data.

Q33.
If derived data is lost, what property distinguishes it from a system of record?

A) It cannot be reconstructed without user re-entry
B) It can be recreated from the original source data ✓ Correct
C) It becomes the new canonical truth automatically
D) It must be restored only from tape backup

Rationale: Derived data is redundant and can be regenerated from the original source.

Q34.
According to the chapter, whether a database is a system of record or a derived data system depends primarily on what?

A) Its vendor's licensing model
B) Whether it runs in the cloud
C) How the database is used in the application ✓ Correct
D) Whether it supports SQL joins

Rationale: The distinction depends on the role the database plays in the application, not on the tool itself.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q35.

What common rule of thumb is given for deciding whether to build in-house or buy from a vendor?

A) Always buy if the system stores regulated data
B) Build only if the software is open source
C) Do in-house what is core or strategically differentiating ✓ Correct
D) Always outsource infrastructure and analytics

Rationale: The chapter says core competencies or competitive advantages should generally be done in-house.

Q36.

What does the term on premises mean in the chapter?

A) Running only in a public cloud region
B) Using serverless functions for all workloads
C) Running on hardware you control, even if hosted in a rented datacenter ✓ Correct
D) Deploying software only on end-user devices

Rationale: On premises refers to hardware under your control, even if it is physically located in a rented datacenter.

Q37.

What does IaaS stand for?

A) Integrated application and storage
B) Infrastructure as a service ✓ Correct
C) Internet access and security
D) Internal analytics as software

Rationale: IaaS is the abbreviation for infrastructure as a service.

Q38.

Which risk is specifically associated with cloud services due to lack of standard APIs?

A) Increased SQL normalization effort
B) Vendor lock-in ✓ Correct
C) Loss of all local caching
D) Guaranteed lower throughput

Rationale: Nonstandard APIs can raise switching costs and create vendor lock-in.

Q39.

The term cloud native refers to an architecture designed primarily to do what?

A) Avoid all use of APIs
B) Take advantage of cloud services ✓ Correct
C) Run only on bare metal hardware
D) Store all data on local disks

Rationale: Cloud native describes architectures designed to exploit cloud services.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q40.
Which of the following is an example of object storage mentioned in the chapter?

A) Amazon S3 ✓ Correct
B) SQLite
C) Kubernetes
D) OpenTelemetry

Rationale: Amazon S3 is listed as an example of an object storage service.

Q41.
In cloud native systems, storage and compute are often described as what?

A) Monolithic
B) Virtualized only
C) Disaggregated ✓ Correct
D) Serialized

Rationale: The chapter describes cloud native architectures as separating, or disaggregating, storage and compute.

Q42.
What does multitenant mean in the context of cloud native systems?

A) Each customer receives a dedicated physical datacenter
B) The same shared hardware serves data and computation for multiple customers ✓ Correct
C) Every service must span multiple legal jurisdictions
D) Each query runs on a separate operating system kernel

Rationale: Multitenancy means multiple customers share the same underlying hardware and service.

Q43.
Which philosophy is described as integrating development and operations responsibilities into shared teams?

A) ETL
B) DevOps ✓ Correct
C) OLAP
D) RAID

Rationale: DevOps is the philosophy of combining development and operations responsibility.

Q44.
In the cloud era, traditional capacity planning often shifts toward what additional concern?

A) Compiler optimization
B) Cost optimization and financial planning ✓ Correct
C) Handwritten SQL authorization
D) Manual disk partitioning only

Rationale: With metered billing, capacity planning becomes closely tied to financial planning and cost optimization.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q45.
A system involving several machines communicating via a network is called what?

A) A distributed system ✓ Correct
B) A normalized system
C) A transactional schema
D) A stateless cache

Rationale: A distributed system is defined as several machines communicating over a network.

Q46.
In a distributed system, each participating process is called a what?

A) Shard
B) Node ✓ Correct
C) Bucket
D) Frame

Rationale: Each process participating in a distributed system is called a node.

Q47.
Which term refers to techniques for collecting and querying execution data to diagnose behavior in distributed systems?

A) Normalization
B) Observability ✓ Correct
C) Serialization
D) Tokenization

Rationale: Observability is the discipline of collecting and analyzing execution data for distributed systems.

Q48.
Which tool set is named as supporting distributed tracing?

A) MySQL, PostgreSQL, SQLite
B) Pandas, R, scikit-learn
C) OpenTelemetry, Zipkin, Jaeger ✓ Correct
D) S3, EBS, Aurora

Rationale: The chapter lists OpenTelemetry, Zipkin, and Jaeger as tracing tools.

Q49.
In a microservices architecture, what is considered good practice regarding databases?

A) All services should share one central database
B) Each service should typically have its own database ✓ Correct
C) No service should persist data directly
D) Only analytical services may use databases

Rationale: The chapter states that it is common and good practice for each service to have its own database.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q50.
What is the principle of data minimization?

A) Store all available data in case it becomes useful later
B) Replicate personal data across regions for resilience
C) Collect and retain only data necessary for specified purposes ✓ Correct
D) Convert all personal data into aggregate reports immediately

Rationale: Data minimization means collecting and keeping only the data needed for explicit, limited purposes.

Q51.
A hospital patient-portal team notices that nightly executive revenue reports are slowing appointment scheduling transactions during business hours. Which architectural change best addresses this operational problem?

A) Move reporting queries to a separate analytical warehouse populated from the operational system ✓ Correct
B) Allow analysts to run ad-hoc SQL directly on the scheduling database after indexing more columns
C) Replace the operational database with a larger single-node server and keep all workloads together
D) Store reports in the application frontend so the backend database is queried less often

Rationale: Separating OLTP from OLAP prevents expensive analytical queries from degrading low-latency transactional workloads.

Q52.
A healthcare startup wants clinicians to retrieve individual patient records quickly while data scientists explore longitudinal utilization trends across millions of encounters. Which pairing of systems is most appropriate?

A) A single OLAP warehouse for both workloads
B) An operational OLTP database plus a separate analytical system ✓ Correct
C) A cache-only architecture without a database
D) A frontend-local database on each clinician device

Rationale: Point lookups for current records fit OLTP, while large-scale trend analysis fits a separate analytical system.

Q53.
An analyst wants to ask, "What was the average emergency department wait time by hour across all facilities last quarter?" This query is best characterized as:

A) An OLTP point query
B) A cache invalidation request
C) An OLAP aggregate query ✓ Correct
D) A stateless HTTP session update

Rationale: The query scans many records and computes an aggregate over time, which is characteristic of OLAP.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q54.

A product team is building a fraud dashboard that must ingest events continuously and support low-latency aggregate queries visible to customer support within seconds. Which class of system best matches this requirement?

A) Traditional batch-oriented OLAP only

B) Product analytics or real-time analytics system ✓ Correct

C) Frontend browser storage

D) Standalone spreadsheet reporting

Rationale: Real-time analytics systems are designed for continuous ingestion and fast aggregate queries in user-facing products.

Q55.

A health system has separate operational databases for billing, pharmacy, laboratory, and staffing. Executives want one place to analyze cross-department performance. What is the primary architectural rationale for a data warehouse here?

A) To eliminate the need for any data transformation

B) To provide a single read-only analytical store that integrates data silos ✓ Correct

C) To let end users update records in all source systems from one interface

D) To replace all operational databases with one shared schema

Rationale: A data warehouse consolidates read-only copies from multiple operational systems so analysts can query across silos.

Q56.

A clinic uses a SaaS customer relationship management platform and wants to combine CRM outreach data with internal appointment data for analytics not supported by the vendor's reports. Which approach is most appropriate?

A) Run SQL directly on the SaaS vendor's production database

B) Use ETL through the SaaS API into the organization's warehouse ✓ Correct

C) Store screenshots of CRM dashboards in the data lake

D) Move all appointment scheduling into the CRM tool

Rationale: When direct database access is unavailable, ETL via the SaaS API enables integration into the organization's analytical environment.

Q57.

A data scientist needs to analyze radiology images, free-text notes, and tabular claims data together for model development. Which repository is most suitable as the central landing zone?

A) A normalized OLTP database only

B) A data lake holding heterogeneous file types ✓ Correct

C) A browser cache on analyst workstations

D) A message queue with no persistent storage

Rationale: A data lake is designed to store varied data types without imposing one schema or model.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q58.

An organization keeps raw event files in object storage and allows different teams to transform them for their own purposes rather than enforcing one warehouse schema at ingestion. This approach most directly reflects:

A) The sushi principle ✓ Correct

B) RAID mirroring

C) Distributed transactions

D) Shared-nothing OLTP

Rationale: The sushi principle refers to the idea that keeping raw data available can better serve diverse downstream uses.

Q59.

A machine learning team trains a readmission-risk model in the analytical environment and then exposes predictions inside the care-management application. This flow is an example of:

A) Normalization

B) Reverse ETL into operational systems ✓ Correct

C) Frontend session persistence

D) RAID recovery

Rationale: Reverse ETL describes analytical outputs being pushed back into operational systems for production use.

Q60.

In a medication-ordering application, the primary orders table is authoritative, while a search index and cache are rebuilt from it when needed. Which statement is most accurate?

A) The cache is the system of record because it serves reads faster

B) The search index is the source of truth because it is optimized for lookup

C) The orders table is the system of record and the others are derived data systems ✓ Correct

D) All three are equally authoritative because they store the same facts

Rationale: The first authoritative write goes to the orders table, while the cache and index are redundant derived representations.

Q61.

A health IT manager proposes storing the same patient allergy facts independently in three unrelated databases because it improves reporting speed. Which concept best clarifies why this design may create confusion?

A) Derived data should be explicitly traced back to a system of record ✓ Correct

B) All redundant data is prohibited in analytics

C) Only cloud services can support multiple copies of data

D) OLTP systems should never contain indexes or caches

Rationale: Architectures are clearer when one authoritative source is defined and redundant copies are recognized as derived data.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q62.
A payer already has a strong infrastructure team and a predictable steady workload for claims processing. Which option is most likely to be cost-effective according to the chapter's trade-off discussion?

A) Self-hosting the needed software on owned or controlled infrastructure ✓ Correct
B) Using only serverless functions for all data storage
C) Replacing the claims platform with a data lake
D) Running all analytics in the frontend browser

Rationale: If skills already exist and load is predictable, self-hosting can often be cheaper than cloud services.

Q63.
A small behavioral health company needs a new stream-processing tool but has no staff with operational expertise in that technology. What is the strongest argument for adopting a managed cloud service first?

A) It guarantees zero outages
B) It avoids any need for security oversight
C) It can reduce time to adoption and the cost of building specialized operational expertise ✓ Correct
D) It eliminates all vendor dependence

Rationale: Managed services are often easier and quicker when the organization lacks experience operating the system itself.

Q64.
A hospital's analytics workload spikes unpredictably when many researchers launch computationally intensive queries, then sits idle for long periods. Which deployment model best matches this pattern?

A) Fixed-capacity self-hosted servers sized only for average load
B) Cloud resources that can scale up and down with demand ✓ Correct
C) Frontend-only processing on researcher laptops
D) A single shared OLTP database for all research and clinical traffic

Rationale: Highly variable workloads benefit from elastic cloud scaling so resources need not remain idle between bursts.

Q65.
A compliance officer worries that a critical cloud database service lacks a required audit feature, and the engineering team cannot add it themselves. Which cloud trade-off does this illustrate?

A) Lower query latency than local function calls
B) Full customer control over service internals
C) Dependence on vendor feature roadmaps ✓ Correct
D) Guaranteed portability across providers

Rationale: A major downside of managed services is limited control when a needed feature is absent.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q66.
An organization is concerned that moving sensitive patient data to a foreign cloud provider could create access problems during geopolitical conflict. Which specific risk from the chapter is most relevant?

A) Cold-start latency
B) Sanctions or lockout due to cross-border political conflict ✓ Correct
C) Over-normalization of schemas
D) Shared-memory contention

Rationale: The chapter notes that using a provider in another country can create lockout risk if sanctions arise.

Q67.
A cloud-native analytics platform stores durable data in object storage and uses compute nodes mainly for execution, replacing them freely when load changes. This architecture most directly demonstrates:

A) Tight coupling of storage and compute
B) Separation of storage and compute ✓ Correct
C) Exclusive reliance on local disks for durability
D) A monolithic frontend design

Rationale: Cloud-native systems commonly disaggregate durable storage from compute resources.

Q68.
A database architect is deciding whether to build a service on virtual block disks attached to cloud VMs. Which concern from the chapter is most applicable?

A) Virtual disks remove all network dependence from I/O
B) Every block I/O becomes a network call, increasing sensitivity to glitches and overhead ✓ Correct
C) Virtual disks prevent any resizing of compute instances
D) Virtual disks are only available for object stores

Rationale: Cloud block devices emulate disks over the network, so I/O overhead and network sensitivity are important trade-offs.

Q69.
A managed analytics service runs workloads from many hospitals on shared hardware while aiming to isolate performance and security between customers. This design characteristic is called:

A) Checkpointing
B) Multitenancy ✓ Correct
C) Normalization
D) Feature engineering

Rationale: Multitenancy means multiple customers share underlying hardware within the same service.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q70.
A newly formed platform team in a cloud-heavy enterprise is told they no longer need operations staff because infrastructure is outsourced. Which response best reflects the chapter?

A) Correct, cloud adoption eliminates operations work entirely
B) Incorrect, operations still matter but shift toward integration, cost optimization, security, and service selection ✓ Correct
C) Correct, because only DBAs are needed in self-hosted environments
D) Incorrect, because cloud services cannot be monitored at all

Rationale: Cloud use changes operations responsibilities rather than removing the need for operations.

Q71.
An SRE team automates environment creation, uses short-lived instances, and performs post-incident learning reviews. These practices are most aligned with:

A) The DevOps/SRE philosophy ✓ Correct
B) Traditional HPC checkpointing
C) Manual capacity planning as the primary operations strategy
D) Exclusive use of shared databases across services

Rationale: Automation, ephemeral infrastructure, frequent updates, and learning from incidents are hallmark DevOps/SRE practices.

Q72.
A finance executive asks why cloud spending rose after a migration even though hardware purchases stopped. Which concept best explains the operational shift described in the chapter?

A) Capacity planning disappears and no planning is needed
B) Performance optimization is irrelevant in the cloud
C) Capacity planning becomes financial and cost optimization under metered billing ✓ Correct
D) Cloud costs are fixed regardless of usage

Rationale: With metered services, planning focuses on resource consumption and cost rather than buying fixed hardware.

Q73.
A telehealth application serves users in North America, Europe, and Asia and wants to reduce user-perceived delay during video session setup. Which reason for distribution best fits this decision?

A) Normalization
B) Latency reduction through geographically closer servers ✓ Correct
C) Elimination of all network failures
D) Avoidance of any compliance obligations

Rationale: Distributing servers geographically can reduce latency by serving users from nearby regions.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q74.
A genomics pipeline exceeds the memory and storage limits of one machine and must spread work across several nodes. Which motivation for distributed systems is most directly involved?

A) Scalability ✓ Correct
B) Derived data minimization
C) Frontend persistence
D) API versioning

Rationale: When one machine cannot handle the workload, distribution is used to scale capacity.

Q75.
A public health service must continue operating if one datacenter fails during a regional outage. Which distributed-system motivation is primary?

A) Feature engineering
B) Fault tolerance and high availability ✓ Correct
C) Ad-hoc SQL freedom
D) Object-store schema enforcement

Rationale: Redundancy across machines or sites is used so service can continue despite failures.

Q76.
A multinational insurer must keep certain citizen data stored and processed within each applicable country. Which reason for distribution is most relevant?

A) Sustainability
B) Legal compliance with data residency requirements ✓ Correct
C) Serverless cold starts
D) OLTP point query optimization

Rationale: Data residency laws can require geographically distributed storage and processing.

Q77.
A cloud architect proposes splitting storage, analytics, and machine learning across different specialized hardware profiles. Which distributed-system benefit does this most directly exploit?

A) Checkpoint recovery
B) Specialized hardware matched to workload ✓ Correct
C) Elimination of observability needs
D) Avoidance of all network transfer

Rationale: Different system components can be placed on hardware optimized for their specific workloads.

Q78.
A sustainability officer suggests scheduling nonurgent batch analytics in regions and times with abundant renewable energy. Which distributed-system motivation does this reflect?

A) Sustainability ✓ Correct
B) Schema normalization
C) API backward compatibility
D) Shared-memory acceleration

Rationale: The chapter notes that flexible placement and timing of jobs can reduce carbon emissions.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q79.

An engineer says, "If a request times out, we know the server never received it, so retrying is always safe." Which evaluation is most accurate?

A) Correct, because network timeouts prove the request was dropped before arrival

B) Correct, because HTTP guarantees at-most-once semantics

C) Incorrect, because a timeout may occur even when the server received the request, making blind retries unsafe ✓ Correct

D) Incorrect, but only for analytical systems and not operational ones

Rationale: A timeout leaves uncertainty about whether the request was processed, so retries may duplicate effects.

Q80.

A data engineering team moves large datasets over the network from storage nodes to separate compute nodes for every transformation step. Which counterpoint from the chapter should they consider?

A) Network calls are always faster than local function calls

B) Bringing computation to where data already resides can outperform moving large volumes of data ✓ Correct

C) Distributed systems remove the need for optimization

D) Single-threaded programs cannot process large datasets

Rationale: The chapter emphasizes that moving computation to data can be more efficient than extensive data transfer.

Q81.

A complex patient-engagement platform is slow, but no team can tell which downstream service is causing delays across many API calls. Which capability is most helpful?

A) Manual spreadsheet logs only

B) Observability with distributed tracing tools ✓ Correct

C) Replacing all services with one cache

D) Adding more frontend JavaScript

Rationale: Observability and tracing help identify where latency occurs across distributed service calls.

Q82.

A microservices team gives each service its own database rather than sharing one enterprise schema. What is the strongest architectural reason from the chapter?

A) Shared databases always violate legal requirements

B) Separate databases reduce coupling and let services evolve implementation details independently ✓ Correct

C) Separate databases eliminate all consistency problems

D) Shared databases cannot support SQL

Rationale: Owning separate databases keeps schema details out of another service's API and reduces inter-service coupling.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q83.

A startup with one small engineering team is considering breaking a simple scheduling application into 25 microservices. Based on the chapter, the best recommendation is to:

A) Proceed, because microservices are always superior to simpler designs

B) Avoid the split unless needed, because microservices mainly solve coordination problems in larger organizations ✓ Correct

C) Proceed only if every service shares one database

D) Avoid the split because HTTP cannot be used between services

Rationale: Microservices add overhead and are most valuable when many teams need to work independently.

Q84.

A development team struggles because changing one service API often breaks clients late in staging. Which practice best addresses the problem described in the chapter?

A) Rely on undocumented field changes to encourage flexibility

B) Use API description standards such as OpenAPI or gRPC to manage client-server contracts ✓ Correct

C) Store API responses only in object storage

D) Replace all APIs with manual database access

Rationale: Formal API descriptions help coordinate evolution of service interfaces and reduce surprise breakage.

Q85.

A cloud provider advertises function-as-a-service for lightweight event handlers that run only when requests arrive. What is the main economic advantage emphasized in the chapter?

A) You pay primarily for execution time rather than preprovisioned instances ✓ Correct

B) You never need to think about quotas or limits

C) Functions can run indefinitely without constraints

D) The provider guarantees zero startup delay

Rationale: Serverless shifts code execution toward metered billing based on actual runtime usage.

Q86.

A team moves a latency-sensitive workflow to serverless functions and then sees delays on the first request after idle periods. Which limitation best explains this?

A) RAID rebuild time

B) Cold-start behavior and runtime constraints in serverless platforms ✓ Correct

C) OLAP batch loading

D) Schema denormalization

Rationale: Serverless platforms may introduce slow initial invocations and other runtime limits.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q87.
A research computing center runs long climate simulations, periodically checkpoints state, and may stop the whole cluster to repair a failed node before restarting from the latest checkpoint. This pattern is most characteristic of:

A) Cloud-native OLTP services
B) High-performance computing or supercomputing ✓ Correct
C) Frontend web applications
D) Real-time product analytics

Rationale: Checkpointed batch jobs with whole-cluster restart behavior are typical of HPC environments.

Q88.
A cloud architect compares enterprise cloud systems with supercomputers. Which statement is most accurate?

A) Cloud systems usually assume all customers fully trust one another
B) Supercomputers commonly emphasize shared-memory or RDMA-style communication for trusted high-performance workloads ✓ Correct
C) Cloud systems never distribute nodes across regions
D) Supercomputers are designed mainly for user-facing high-availability web requests

Rationale: HPC environments often use high-performance communication mechanisms that assume a trusted setting.

Q89.
A privacy officer asks engineers to support deletion requests for personal data that may also exist in append-only logs and trained models. Why is this a significant architectural challenge?

A) Because GDPR requires all data to be retained forever
B) Because immutable and derived structures complicate selective erasure ✓ Correct
C) Because deletion is only a frontend issue
D) Because systems of record cannot contain personal data

Rationale: Right-to-erasure requirements are difficult when data has propagated into immutable logs and derived artifacts.

Q90.
A consumer health app wants to retain every location ping indefinitely because it might become useful later, even though the purpose is unclear. Which principle from the chapter most strongly argues against this?

A) Data minimization ✓ Correct
B) Shared-memory optimization
C) RAID transparency
D) Microservice independence

Rationale: Data minimization holds that personal data should be collected for explicit purposes and not kept longer than necessary.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q91.
A reproductive health platform debates whether to store precise location histories. Which risk highlighted in the chapter is most relevant to the decision?

A) Location data can expose sensitive or criminalized behaviors if disclosed or compelled ✓ Correct
B) Location data cannot be analyzed in a data lake
C) Location data is always anonymous under GDPR
D) Location data prevents use of OLTP databases

Rationale: The chapter notes that stored location traces can create serious safety risks by revealing sensitive behavior.

Q92.
A chief information officer says the only cost of storing user data is the monthly object-storage bill. Which response best aligns with the chapter?

A) Correct, because cloud pricing captures all relevant storage consequences
B) Incorrect, because legal liability, fines, reputational harm, and compelled disclosure risks also matter ✓ Correct
C) Correct, unless the data is used for analytics
D) Incorrect, but only because RAID hardware is expensive

Rationale: The chapter stresses that data storage costs include regulatory, legal, and societal risks beyond infrastructure charges.

Q93.
A payment processor undergoes recurring independent audits to verify adherence to card-security requirements. Which example from the chapter most closely matches this scenario?

A) PCI compliance ✓ Correct
B) Feature engineering
C) Distributed tracing
D) Reverse ETL

Rationale: Payment processors are specifically cited as adhering to PCI standards verified through audits.

Q94.
A software vendor seeking enterprise contracts is asked to demonstrate third-party audited controls over its operations and security practices. Which framework named in the chapter is most relevant?

A) SOC Type 2 ✓ Correct
B) OLAP Cube Certification
C) RDMA Trust Standard
D) ETL Maturity Index

Rationale: The chapter notes that many buyers require vendors to comply with SOC Type 2 standards.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q95.
A backend service handles one HTTP request, returns a response, and retains no memory of that request except what was written externally. This service is best described as:

A) Stateful with local persistence
B) Stateless ✓ Correct
C) An analytical warehouse
D) A system of record by definition

Rationale: A stateless service forgets request-specific information once processing is complete unless it is stored elsewhere.

Q96.
A mobile app stores a small amount of data on the device, but the organization's biggest data challenges remain on the server side. What is the best explanation?

A) Frontends never store any data at all
B) Backends manage data on behalf of all users, creating larger infrastructure challenges than per-user local storage ✓ Correct
C) Mobile apps cannot communicate over the internet
D) Backend systems are always analytical rather than operational

Rationale: The backend must manage shared data for all users, which creates greater scale and coordination challenges.

Q97.
A hospital wants one platform to support low-latency patient updates and also immediate large-scale fraud pattern scans on the same application data. Which option from the chapter is most directly suited to this mixed need?

A) A spreadsheet with macros
B) HTAP, while recognizing it does not eliminate the broader distinction between OLTP and analytics ✓ Correct
C) A cache without a database
D) A frontend-only architecture

Rationale: HTAP targets workloads that need both transactional updates and analytical scans within the same application context.

Q98.
An architect argues that because HTAP exists, the enterprise no longer needs a separate warehouse that combines data from many operational systems. Which response is best?

A) Correct, HTAP always replaces enterprise warehouses
B) Incorrect, because enterprises still often need a single analytical store to combine data from many separate operational databases ✓ Correct
C) Correct, but only for SaaS data sources
D) Incorrect, because HTAP supports only frontend workloads

Rationale: HTAP does not remove the need for a central warehouse used to integrate data across many operational systems.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q99.

A claims platform stores current claim status in its transactional database, while analysts need to study how claims changed over time. Which contrast best explains why the analytical copy is useful?

A) OLTP emphasizes latest state, whereas analytics often examines historical events over time ✓ Correct

B) OLTP supports only images, whereas analytics supports only text

C) OLTP allows arbitrary ad-hoc SQL by end users, whereas analytics forbids it

D) OLTP always has larger datasets than analytics

Rationale: Operational systems usually represent current state, while analytical systems are commonly used to analyze historical change.

Q100.

A health network is deciding whether to keep a new quality-reporting workload on one powerful machine or distribute it across a cluster. Based on the chapter, what is the best initial recommendation?

A) Distribute immediately, because more nodes are always faster

B) Prefer the simplest single-node approach if it meets requirements, since distributed systems add complexity and cost ✓ Correct

C) Use microservices regardless of workload size

D) Avoid databases and rely only on flat files

Rationale: The chapter advises not to rush into distribution when a single machine can handle the workload more simply and cheaply.

Q101.

A health system wants sub-second patient-chart updates for clinicians while also supporting enterprise-wide ad hoc quality analytics across years of encounters. Which architecture best reflects the chapter's trade-off analysis?

A) Use a single OLTP database for both workloads because SQL is flexible enough to eliminate architectural separation

B) Separate operational systems from analytical systems, moving data through ETL or streaming into a warehouse or lake ✓ Correct

C) Store all data only in a cache to reduce latency for both clinical and analytical users

D) Prohibit ad hoc analytical queries and require analysts to use only predefined application queries

Rationale: Operational and analytical workloads have different access patterns and are often best served by separate systems connected by ETL or streams.

Q102.

An organization labels its recommendation feature store, search index, and cache as "authoritative patient data sources." Based on the chapter, what is the strongest critique of this design description?

A) It confuses derived data systems with systems of record ✓ Correct

B) It assumes OLAP queries must always be written in Python instead of SQL

C) It ignores that frontends can store local data on devices

D) It incorrectly treats data lakes as relational warehouses

Rationale: Caches, indexes, and trained-model artifacts are derived data, whereas a system of record holds the authoritative version first written.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q103.

A startup is debating microservices for a product used by a single internal team of six engineers. Which recommendation is most consistent with the chapter?

A) Adopt microservices because distributed transactions are easier in microservices than in monoliths

B) Adopt microservices because every service should always own a separate database regardless of team size

C) Prefer the simplest implementation unless independent-team coordination is a major problem ✓ Correct

D) Prefer microservices because network calls are usually faster than in-process function calls

Rationale: The chapter frames microservices as primarily solving organizational coordination problems and warns that they may be unnecessary overhead for small teams.

Q104.

A public health agency needs fraud detection on claims using both low-latency per-claim updates and analytical scans over many recent events. Which conclusion best fits the text?

A) Traditional OLAP alone is sufficient because fraud detection never needs low-latency writes

B) HTAP may be useful because the same application mixes transactional and analytical access patterns ✓ Correct

C) A data lake alone is ideal because it imposes no schema and therefore guarantees low latency

D) A cache should replace the operational database because caches are canonical sources of truth

Rationale: HTAP is presented as useful when one application must both update individual records quickly and run analytical scans, such as fraud detection.

Q105.

An analytics team argues that because their warehouse already stores cleaned relational tables, a data lake is unnecessary. A lead data scientist disagrees. Which rationale most strongly supports the data scientist?

A) Data lakes are mandatory whenever GDPR applies

B) Data lakes better support arbitrary nonrelational artifacts such as images, text, feature vectors, and custom ML workflows ✓ Correct

C) Data lakes eliminate the need for ETL from operational systems

D) Data lakes are always more secure than warehouses because they use object storage

Rationale: The chapter emphasizes that data lakes provide flexible storage for many data types and workflows not well suited to relational warehouse schemas.

Q106.

A hospital stores raw device telemetry in object storage, then different teams transform it into warehouse tables, feature matrices, and dashboards. Which principle does this most closely illustrate?

A) The sushi principle that raw data is better for diverse downstream consumers ✓ Correct

B) The rule that all analytical data should remain normalized

C) The requirement that operational systems must be queried directly by analysts

D) The idea that frontends should own persistent organizational data

Rationale: Keeping raw data centrally available allows each downstream consumer to transform it into the form best suited to its own needs.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q107.

A CIO claims that moving to a cloud provider means the organization no longer needs operations staff. Which response is best supported by the chapter?

A) Correct, because cloud services fully eliminate monitoring, security, and troubleshooting work

B) Correct, because metered billing replaces all forms of capacity and cost planning

C) Incorrect, because operations shifts toward service selection, integration, cost optimization, security, and incident response ✓ Correct

D) Incorrect, but only because cloud services do not support automation

Rationale: The chapter states that operations remains essential in the cloud, though the focus changes from machine administration to service-level concerns.

Q108.

A data platform team is choosing between self-hosting a familiar database on fixed workloads versus adopting a managed cloud equivalent. Which decision argument is most aligned with the chapter?

A) Managed cloud is always cheaper because hardware ownership is inherently inefficient

B) Self-hosting may be cheaper when the team already has expertise and the workload is predictable ✓ Correct

C) Cloud services should be avoided whenever data volumes exceed a single machine

D) Self-hosting is always preferable because vendor lock-in can never be mitigated

Rationale: The text notes that if a team already knows how to operate a system and load is predictable, self-hosting is often cheaper.

Q109.

A compliance officer asks why analysts should not directly run exploratory queries on production scheduling and billing databases. Which justification is least central in the chapter?

A) Analytical queries may degrade OLTP performance for operational users

B) Data may be spread across silos that are hard to combine in one query

C) Operational schemas may be poorly suited for analytics

D) SQL cannot express aggregate analytical queries over many records ✓ Correct

Rationale: The chapter explicitly notes that SQL is flexible and works for both transaction processing and analytics, even if separate systems are often preferred.

Q110.

A cloud-native database stores durable data in object storage while using local instance disks mainly as temporary caches. What trade-off is this design exploiting?

A) It sacrifices all elasticity to maximize local disk durability

B) It separates storage from compute to improve scaling and recovery in cloud environments ✓ Correct

C) It avoids multitenancy because object stores require single-tenant isolation

D) It removes network transfer costs by keeping all computation on the storage nodes

Rationale: Cloud-native systems commonly disaggregate storage and compute, using durable cloud storage and treating local disks as ephemeral caches.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q111.

A team wants to place every workload on a distributed cluster by default because "more nodes are always faster." Which critique best reflects the chapter?

A) Distributed systems guarantee lower latency than single-node systems for all workloads

B) Single-node systems are obsolete because modern datasets always exceed one machine

C) Network overhead and coordination can make a simple single-node program outperform a large cluster ✓ Correct

D) Clusters should be used only for analytical systems, never for operational systems

Rationale: The chapter warns that distributed execution can be slower due to network costs and coordination, and that single-node systems may outperform clusters.

Q112.

A privacy review board asks whether storing years of detailed location logs is justified "just in case" they become useful later. Which principle from the chapter most directly challenges this reasoning?

A) Normalization

B) Data minimization ✓ Correct

C) Multitenancy

D) Observability

Rationale: Data minimization argues against retaining personal data speculatively when it is not necessary for an explicit purpose.

Q113.

A vendor promises a fully managed analytics service but offers no standard API and no export path for metadata or workloads. Which strategic risk from the chapter is most relevant?

A) RAID failure

B) Vendor lock-in ✓ Correct

C) Point-query amplification

D) Feature engineering drift

Rationale: Lack of compatible APIs and migration paths raises switching costs and creates vendor lock-in.

Q114.

A national telehealth platform must keep residents' records within each country's borders while still offering one unified service. Which reason for distribution is most directly implicated?

A) Specialized hardware

B) Elasticity

C) Legal compliance ✓ Correct

D) Batch checkpointing

Rationale: The chapter identifies data residency laws as a reason to distribute data and processing across geographic locations.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q115.

A product manager wants analysts to query the same production database used by the patient portal because "it avoids ETL complexity." Which evaluation is most balanced according to the chapter?

A) This is ideal because OLTP layouts are optimized for ad hoc analytics

B) This may work at small scale, but separation is often preferred as workloads become more demanding ✓ Correct

C) This is impossible because SQL cannot support both transactional and analytical queries

D) This is best only if the database is multitenant

Rationale: The chapter notes that using one database for both can work initially, but specialized separation often emerges as scale and workload demands grow.

Q116.

A team says its analytical warehouse is not redundant because it stores transformed historical data. Based on the chapter, how should this be interpreted?

A) It is still derived data because it can be recreated from source systems ✓ Correct

B) It is a system of record because historical data is always authoritative

C) It is not data infrastructure because analysts, not users, query it

D) It is a frontend component because it supports dashboards

Rationale: Analytical systems usually contain derived data copied and transformed from operational systems and can be recreated from those sources.

Q117.

A hospital uses a SaaS billing platform and wants richer cross-domain analytics than the vendor API supports. Which architectural move is most consistent with the chapter?

A) Run ad hoc SQL directly against the SaaS vendor's internal database

B) Use ETL connectors to bring SaaS data into the organization's own analytical environment ✓ Correct

C) Replace the SaaS product with a browser-based frontend cache

D) Require data scientists to use only the SaaS vendor's dashboard reports

Rationale: The chapter notes that ETL from SaaS APIs into an internal warehouse can enable analyses not possible through the SaaS interface alone.

Q118.

A team proposes putting all service data into one shared relational database to simplify joins across services. Which concern from the chapter most strongly argues against this?

A) Shared databases make API and schema evolution across services harder and can create performance interference ✓ Correct

B) Shared databases prevent the use of HTTP between services

C) Shared databases are incompatible with any form of security control

D) Shared databases eliminate the need for observability

Rationale: The chapter explains that a shared database effectively exposes schema as API and can couple services operationally and organizationally.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q119.

A machine-learning group complains that a warehouse table format is cumbersome for feature engineering and text processing. Which architectural response is most justified by the chapter?

A) Move all operational systems to OLAP engines

B) Introduce a data lake to store raw and varied data types for nonrelational analysis workflows ✓ Correct

C) Prohibit Python and require all transformations in SQL

D) Replace the warehouse with a frontend mobile database

Rationale: Data lakes are presented as better suited than relational warehouses for flexible ML and unstructured-data workflows.

Q120.

A finance executive asks why cloud cost management has become a core operational concern after migration from on-premises servers. Which explanation best fits the chapter?

A) Because cloud services remove all resource limits, spending becomes impossible to predict

B) Because metered billing shifts traditional capacity planning toward financial planning and cost optimization ✓ Correct

C) Because cloud providers forbid automation of resource allocation

D) Because self-hosted systems never required performance tuning

Rationale: The chapter states that cloud operations reframes capacity planning as financial planning and performance optimization as cost optimization.

Q121.

An engineering leader says, "Our backend is stateless, so we do not need persistent server-side data infrastructure." Which reply is most accurate?

A) Correct, because stateless services automatically persist request history in HTTP

B) Correct, because frontends always hold all durable data for all users

C) Incorrect, because any information that must persist across requests must be stored on the client or in server-side data infrastructure ✓ Correct

D) Incorrect, but only for analytical systems

Rationale: Stateless backend code forgets request-specific state after handling a request, so persistent information must live elsewhere.

Q122.

A data architect is comparing a warehouse and a data lake for an academic medical center. Which statement best synthesizes the distinction?

A) A warehouse is schema-free raw file storage, whereas a lake is curated relational SQL storage

B) A warehouse optimizes relational analytics, whereas a lake centralizes potentially useful data in flexible file-based forms ✓ Correct

C) A warehouse is only for operational systems, whereas a lake is only for frontends

D) A warehouse stores only current state, whereas a lake stores only immutable logs

Rationale: The chapter contrasts warehouses as relational analytical stores with lakes as flexible repositories for many file and data types.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q123.
A distributed tracing initiative is proposed after recurring latency spikes in a microservice-based patient scheduling platform. Which problem is tracing intended to address?

A) Reducing object storage costs through compression
B) Determining which service called which other service and how long each step took ✓ Correct
C) Eliminating the need for API versioning
D) Replacing ETL between operational and analytical systems

Rationale: Tracing tools support observability by showing request paths and timings across distributed services.

Q124.
A team plans to retry every timed-out network request automatically. Which evaluation is most faithful to the chapter's caution?

A) Safe, because a timeout proves the remote service never received the request
B) Unsafe in some cases, because after a timeout you may not know whether the request was received or acted upon ✓ Correct
C) Always required, because distributed systems guarantee idempotent handling
D) Unnecessary, because network failures are negligible in datacenters

Rationale: The chapter notes that timeouts create ambiguity about whether the remote service processed the request, making naive retries risky.

Q125.
A cloud architect recommends building on higher-level managed services whenever possible rather than assembling lower-level primitives. Which argument from the chapter best supports this?

A) Higher-level abstractions are always more general and customizable
B) If a higher-level service matches the use case, it can meet needs with much less hassle than building it yourself ✓ Correct
C) Lower-level components cannot be combined in cloud environments
D) Higher-level services eliminate all compliance obligations

Rationale: The chapter's rule of thumb is that higher-level abstractions are advantageous when their intended use cases align with your needs.

Q126.
An enterprise has 200 operational applications but one centralized warehouse used by analysts. Which explanation best reflects the chapter's reasoning?

A) Operational systems are commonly separated by application or service, while analysts benefit from one place to combine data across silos ✓ Correct
B) Warehouses can only ingest from exactly one source system
C) Operational databases must be shared among all services to preserve consistency
D) A single warehouse is required because OLAP cannot process historical data from multiple domains

Rationale: The chapter notes that operational systems are often independent, whereas one warehouse helps analysts query across many systems.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q127.

A cloud database team rejects virtual block storage in favor of object storage plus specialized services for smaller values. What is the strongest justification from the chapter?

A) Virtual disks are illegal in multitenant environments

B) Block-device emulation adds overhead and makes every I/O sensitive to network glitches ✓ Correct

C) Object storage supports byte-level updates more efficiently than block devices

D) Virtual disks prevent encryption at rest

Rationale: The chapter criticizes virtual disk emulation for overhead and network sensitivity, especially compared with cloud-native designs.

Q128.

A chief architect says, "Because our warehouse is read-only for analysts, it cannot influence operational products." Which counterexample from the chapter most directly refutes this claim?

A) Derived data can flow back to operational systems through reverse ETL or deployed ML models ✓ Correct

B) Read-only systems must never be connected to operational networks

C) Warehouses are always the system of record for recommendation engines

D) Analysts routinely write directly to production OLTP tables

Rationale: The chapter describes reverse ETL and production deployment of analytical outputs, showing analytical systems can influence operational behavior.

Q129.

A hospital wants to deploy a custom-built low-latency imaging inference service using GPUs and maintain fine-grained tuning of hardware and software. Which deployment choice is most defensible from the chapter?

A) A generic SaaS analytics dashboard, because cloud services are always the best fit for specialized workloads

B) Self-hosting or tightly controlled infrastructure, because specialist requirements may not be met by existing cloud services ✓ Correct

C) A shared spreadsheet, because small teams should avoid all infrastructure

D) A data warehouse, because warehouses are optimized for GPU workloads

Rationale: The chapter notes that specialist requirements, including highly specialized performance needs, may justify in-house systems.

Q130.

A systems engineer proposes a global supercomputer-style restart policy for a patient-facing cloud service: if one node fails, pause the whole cluster, repair it, and restart from checkpoint. Why is this a poor fit according to the chapter?

A) Cloud services generally need continuous availability, unlike many HPC batch workloads ✓ Correct

B) Checkpointing is impossible in distributed systems

C) Cloud systems cannot use more than one datacenter

D) Patient-facing services never use batch processing

Rationale: The chapter contrasts HPC restart-from-checkpoint approaches with cloud services that must keep serving users with minimal interruption.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q131.

A team wants to use serverless functions for bursty API-driven workloads. Which trade-off should they explicitly evaluate?

A) Serverless guarantees unlimited execution time and no cold starts

B) Serverless shifts provisioning to the provider but may impose runtime limits and slow starts ✓ Correct

C) Serverless eliminates the need for API integration with other services

D) Serverless requires dedicated hardware per customer

Rationale: The chapter highlights autoscaling and usage-based billing benefits of serverless, along with limits on execution time and possible cold starts.

Q132.

A privacy engineer asks how an append-only event log can coexist with GDPR erasure requests. What is the chapter's main point?

A) Immutable logs are automatically exempt from deletion requirements

B) The issue creates a genuine engineering challenge because legal principles do not prescribe a simple technical recipe ✓ Correct

C) The solution is always to convert the log into a cache

D) The problem disappears if the log is stored in a data lake

Rationale: The chapter emphasizes that compliance with deletion rights in systems using immutable constructs is difficult and not solved by a single prescribed technology.

Q133.

A product analytics team needs low-latency aggregate queries over event streams embedded in a user-facing application. Which option best fits the chapter's taxonomy?

A) Traditional batch-oriented OLAP warehouse only

B) Product analytics or real-time analytics system such as Pinot, Druid, or ClickHouse ✓ Correct

C) A pure OLTP database with no analytical optimization

D) A frontend local database on each user device

Rationale: The chapter identifies product or real-time analytics systems as optimized for low-latency analytical queries over streaming data in user-facing products.

Q134.

A CIO argues that because cloud providers specialize in operations, their managed services always outperform any self-hosted deployment. Which response best reflects the chapter's trade-off framing?

A) True, because specialization always beats workload-specific tuning

B) False, because self-hosting may allow custom configuration and tuning for a particular workload ✓ Correct

C) True, because cloud providers never share hardware among customers

D) False, because managed services cannot provide high availability

Rationale: The chapter notes that providers gain operational expertise, but self-hosting can still win when tailored closely to a specific workload.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q135.

An informatics team is deciding whether to centralize both storage and compute on the same machines or to disaggregate them. Which evaluation is most consistent with the chapter?

A) Disaggregation can improve elasticity and recovery, but it introduces network transfer between storage and computation ✓ Correct

B) Disaggregation removes all network dependence from the system

C) Keeping storage and compute together is mandatory in cloud-native architectures

D) Disaggregation is relevant only to frontends, not backends

Rationale: The chapter presents separation of storage and compute as advantageous in cloud-native systems while acknowledging the cost of moving data over the network.

Q136.

A quality analytics director asks why analysts are usually allowed to write arbitrary SQL while end users of a patient portal are not. Which explanation best matches the chapter?

A) Analysts are trusted to modify production data directly, whereas patients are not

B) Portal queries are usually fixed in application code to control permissions and avoid expensive ad hoc queries on operational systems ✓ Correct

C) End users cannot benefit from SQL because SQL is only for historical data

D) Analysts never use visualization tools to generate queries automatically

Rationale: The chapter explains that OLTP systems typically restrict users to predefined queries for security and performance reasons, unlike analytical systems.

Q137.

A data governance committee wants a single concept to clarify how data moves from intake forms to dashboards, indexes, and predictive models. Which conceptual pair from the chapter is most useful?

A) Frontend and backend

B) Point queries and aggregate queries

C) Systems of record and derived data systems ✓ Correct

D) Cloud and on-premises

Rationale: The distinction between systems of record and derived data clarifies which data is authoritative and which is transformed from other sources.

Q138.

A biomedical research platform must schedule compute jobs where renewable electricity is abundant and defer them when the grid is strained. Which reason for distribution does this illustrate?

A) Latency

B) Sustainability ✓ Correct

C) Multitenancy

D) Normalization

Rationale: The chapter explicitly identifies sustainability and renewable-energy-aware placement as a reason to distribute workloads.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q139.

A team says that because a cloud storage service hides individual machines, they can ignore service quotas and concurrency limits. Which response is most accurate?

A) Correct, because abstraction removes all practical limits

B) Incorrect, because cloud services still impose limits and quotas that require planning ✓ Correct

C) Correct, but only for object storage and not for compute services

D) Incorrect only if the service is self-hosted

Rationale: The chapter notes that cloud services still have quotas and resource limits that operations teams must understand and plan around.

Q140.

A hospital's data scientists need both daily rerunnable analyses and second-by-second reactions to suspicious account activity. Which synthesis best matches the chapter?

A) Use only file-based batch analysis because streams cannot support analytics

B) Use only stream processing because files cannot support reprocessing

C) Combine file-based analysis and stream processing according to time sensitivity and use case ✓ Correct

D) Avoid analytical systems because operational systems should own all logic

Rationale: The chapter contrasts periodic reruns on files with faster stream processing and suggests choosing based on application time sensitivity.

Q141.

A development team is struggling to test one microservice locally because it depends on many others. Which broader lesson from the chapter does this exemplify?

A) Microservices reduce all forms of complexity

B) Decomposition into many services can create operational and development complexity despite organizational benefits ✓ Correct

C) The problem occurs only when services use SQL databases

D) This means the system should use distributed transactions everywhere

Rationale: The chapter emphasizes that microservices can breed complexity in testing, deployment, monitoring, and dependency management.

Q142.

A chief data officer wants to know why large-scale systems tend to become more specialized over time. Which answer best reflects the chapter?

A) Because specialization is legally required once data reaches terabyte scale

B) Because general-purpose systems are comfortable at small scale, but demanding workloads push optimization toward specialized systems ✓ Correct

C) Because specialized systems eliminate the need for ETL and integration

D) Because specialized systems always reduce the number of technologies in use

Rationale: The chapter states that as workloads become more demanding, systems often specialize to optimize for particular access patterns and goals.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q143.

A hospital is considering putting its scheduling service, claims service, and pharmacy service in separate microservices with separate databases. Which challenge should leadership expect to own at the application level?

A) Maintaining cross-service data consistency without relying heavily on distributed transactions ✓ Correct

B) Eliminating all observability tooling because services are independent

C) Avoiding any need for ETL into analytical systems

D) Guaranteeing that network calls are as fast as local function calls

Rationale: The chapter notes that when each service owns its database, consistency across services becomes the application's problem and distributed transactions are rarely used.

Q144.

A security team is reviewing a multitenant cloud-native database. Which concern is most directly highlighted by the chapter?

A) Whether one customer's workload can affect another customer's performance or security ✓ Correct

B) Whether multitenancy prevents object storage from being used

C) Whether multitenancy makes SQL impossible to support

D) Whether multitenancy requires all customers to share a schema

Rationale: The chapter states that multitenancy requires careful engineering so one customer does not harm another's performance or security.

Q145.

A chief architect wants to know when a single-node design should be preferred over a distributed one. Which answer is most aligned with the chapter?

A) Only when the application has no users

B) Whenever possible, because single-node systems are often simpler and cheaper unless distribution is truly needed ✓ Correct

C) Never, because cloud-native architecture requires at least three nodes

D) Only for caches, not for databases or analytics

Rationale: The chapter advises not to rush into distribution and notes that single-machine solutions are often simpler and less costly.

Q146.

A legal counsel asks engineers to include reputational damage, fines, and compelled disclosure risks in data-retention decisions. Which principle from the chapter does this reflect?

A) Storage cost should be assessed only as infrastructure spending

B) The value of storing data must be weighed against broader legal, ethical, and safety risks ✓ Correct

C) Derived data is exempt from legal review

D) Only government agencies need to consider compelled disclosure

Rationale: The chapter broadens storage trade-offs beyond technical cost to include liability, compliance, safety, and reputational risks.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann

Q147.

An enterprise says its HTAP platform proves the operational-versus-analytical distinction is obsolete. Which evaluation best fits the chapter?

A) Correct, because HTAP removes all meaningful differences in workload design

B) Incorrect, because many HTAP systems still internally couple transactional and analytical components and the distinction remains useful ✓ Correct

C) Correct, because ETL is impossible in HTAP environments

D) Incorrect, but only because HTAP cannot support fraud detection

Rationale: The chapter notes that HTAP often still contains separate internal components and does not erase the conceptual distinction between OLTP and analytics.

Q148.

A cloud migration team claims that object storage is simply a drop-in replacement for a local filesystem. Which response is most accurate?

A) Correct, because object stores expose the same interface and semantics as POSIX filesystems

B) Incorrect, because object stores offer more limited APIs but hide physical machines and provide scalable durable storage ✓ Correct

C) Correct, because object stores support efficient small random updates like block devices

D) Incorrect, because object stores cannot store large files

Rationale: The chapter describes object storage as a simpler, more limited abstraction than a filesystem, with benefits in scale and durability.

Q149.

A health insurer is deciding whether to keep a fraud model's training data indefinitely. Which argument from the chapter most strongly supports a shorter retention policy?

A) Historical data can never improve analytical systems

B) Personal data should be kept only for explicit purposes and not longer than necessary ✓ Correct

C) Training datasets are not considered derived data

D) Deleting data always reduces model bias

Rationale: The chapter links privacy regulation and data minimization to the requirement that personal data be retained only as long as necessary for explicit purposes.

Q150.

A senior architect must recommend a data architecture for a growing digital health company. Which overall recommendation best synthesizes the chapter's message?

A) Standardize on one database and one deployment model because architectural consistency is more important than workload differences

B) Choose technologies by evaluating trade-offs among workload patterns, organizational structure, operational capability, compliance, and cost ✓ Correct

C) Prefer distributed cloud-native services in every case because modern systems must be elastic

D) Avoid derived data because redundancy is inherently poor design

Rationale: The chapter's core theme is that no single approach is best; architecture should be chosen by weighing trade-offs in context.

Access Complete Test Bank for Designing Data - Intensive Applications 2nd Kleppmann