

ET — Emerging Technologies

Complete Book | Table of Contents | 20 chapters + 10 sections

Module A — Introduction to Data Analytics

CH-1	Information Value Chain and Sources of Data	2
CH-2	Data Storage, Extraction and Types of Analytics	7
CH-3	Reporting, Data Visualization, AR and VR	12
CH-4	Data Analytics Tools and Use of Data Analytics in Banking	16

Module B — Big Data

CH-5	Big Data, MapReduce Programming and the Hadoop Ecosystem	21
CH-6	SQL, Advanced SQL and In-database Analytics	25
CH-7	Apache Spark and Use of Big Data in Banking	29

Module C — Blockchain

CH-8	Blockchain Fundamentals: Byzantine Agreement, Dynamics and Types	34
CH-9	Forks, Side Chains and Distributed Ledger Technologies	39
CH-10	Smart Contracts, Trust Algorithms and Blockchain in Banking	43

Module D — Machine Learning – Supervised and Unsupervised Learning

CH-11	Machine Learning and its Mathematical Foundations	48
CH-12	Linear Models for Regression and Classification	53
CH-13	Decision Trees and Support Vector Machines	57
CH-14	Unsupervised Learning: Clustering, K-means, EM and Gaussian Mixtures	62
CH-15	Dimensionality Reduction: Factor Analysis, PCA and ICA	67
CH-16	Graphical Models, HMM, CRF and Use of Machine Learning in Banking	71

Module E — Artificial Intelligence

CH-17	AI Basics and Neural Network Fundamentals	76
CH-18	Convolutional Neural Networks, Image and Text Classification and Emerging Architectures	81
CH-19	Recurrent Neural Networks, LSTM and Time Series Forecasting	85
CH-20	NLP, Expert Systems, Knowledge Management and AI in Banking	90

Additional Resources

—	Case Studies — Module A: Introduction to Data Analytics	96
—	Case Studies — Module B: Big Data	98
—	Case Studies — Module C: Blockchain	101
—	Case Studies — Module D: Machine Learning – Supervised and Unsupervised Learning	103
—	Case Studies — Module E: Artificial Intelligence	108
—	Mega Tests — Mega Test 1 — 100 Questions (Full Syllabus)	112
—	Mega Tests — Mega Test 2 — 100 Questions (Full Syllabus)	121
—	Mega Tests — Mega Test 3 — 100 Questions (Full Syllabus)	131
—	Mega Tests — Mega Test 4 — 100 Questions (Full Syllabus)	140
—	Mega Tests — Mega Test 5 — 100 Questions (Full Syllabus)	149

CH-1

Information Value Chain and Sources of Data

Topic 1 — Information Value Chain — Meaning and Purpose

The **information value chain** is the sequence of activities that turns raw data about real-world events into decisions and actions that create business value.

- **Starting point** — Every chain begins with a **real-world event**, such as a deposit, a card swipe or a loan application.
- **Capture** — Source systems record the event and its **characteristics** (who, what, when, where, how much) as data.
- **Refining** — Data is stored, extracted and analysed to produce **information and insight**.
- **End point** — Value arises only when insight leads to a **decision and an action**, such as approving, pricing or blocking.
- **Feedback loop** — Every action creates new events and new data, so the chain runs as a **continuous cycle**.

Why the chain matters

- **Value sits at the end** — Stored data has little value until it changes a decision.
- **Weakest link** — An error or delay at any stage reduces the value of every later stage.
- **Question first** — Analysts begin with the business question and work backwards to the data needed.
- **Value-chain idea** — Like the business value chain, each stage must add value; here the product being refined is information.

Remember this

- **Information value chain:** Event → Data capture → Storage → Extraction → Analysis → Information → Decision → Action → Value, with feedback to new events.

Topic 2 — Stages of the Information Value Chain

Stage	What happens	Banking illustration
1. Real-world event	A business activity occurs	A customer pays a merchant with a debit card
2. Data capture	A source system records the attributes of the event	Card switch logs card token, amount, merchant code, time and location
3. Data storage	Data is kept in databases, warehouses or lakes	Transaction is posted in the card system and copied to the warehouse overnight
4. Data extraction	Relevant data is pulled, cleaned and combined	Analyst extracts six months of spends with the customer profile
5. Analysis	Patterns, trends and predictions are derived	Model finds heavy and regular travel spending
6. Information and reporting	Results are presented to decision-makers	Dashboard flags a "frequent traveller" segment
7. Decision	A choice is made among options	Product head decides to offer a travel credit card
8. Action and value	Decision is executed and the outcome measured	Offer sent; activations and fee income tracked

- **Capture quality** — A wrong merchant code at capture is inherited by every later stage (**garbage in, garbage out**).

- **Cheapest fix** — Errors cost least to correct at the point of capture and most after a decision is taken.
- **Latency** — Value decays with delay; a fraud alert is useful in seconds, a month-end report only explains the loss.
- **Ownership** — Business owns definitions, IT owns systems, analysts own models and management owns the action.

Example: Kaveri Bank processes 40 lakh card transactions a day. A pattern of small "test" debits followed by a large purchase is scored within seconds and the card is blocked, saving ₹1,85,000. The same data analysed at month-end would have produced only a loss report.

Remember this

- The value of information falls as **quality falls** or **delay rises**; the chain is only as strong as its weakest stage.

Topic 3 — Data, Information, Knowledge and Wisdom

The **DIKW hierarchy** explains how meaning and value are added as data moves up the information value chain.

Level	Meaning	Banking illustration
Data	Raw facts and figures without context	Debit of ₹45,000 at 02:10 hours
Information	Data processed and organised in context (who, what, when, where)	Third night-time overseas ATM withdrawal by one customer this week
Knowledge	Information combined with experience to explain patterns (how)	Such a pattern is typical of a skimmed or cloned card
Wisdom	Judgement applied to choose the right action (why, what is best)	Block the card, call the customer and inspect the ATM

Attributes of valuable information

- **Accurate** — Free from error and faithful to the source event.
- **Timely** — Available when the decision has to be taken.
- **Relevant** — Answers the question the decision-maker is actually asking.
- **Complete** — Covers all the facts needed, without gaps that mislead.
- **Consistent** — Uses the same definitions across branches, systems and periods.
- **Secure and accessible** — Reaches authorised users only, in a usable form.

Remember this

- **DIKW:** Data = raw facts; Information = data in context; Knowledge = understanding of patterns; Wisdom = judgement on the best action.

Topic 4 — Information Value Chain in Banking

Business area	Data captured	Decision and action
Credit	Application, credit report, bank statements, GST returns	Sanction, price or decline the loan
Fraud	Real-time transaction, device and location data	Allow, step up authentication or block
Marketing	Spending patterns and life events	Targeted cross-sell offer
Collections	Repayment behaviour and early-warning signals	Reminder, follow-up call or recovery step
Regulatory reporting	CBS and other source-system data	Returns filed on RBI's Centralised Information Management System (CIMS)



- **CIMS** — RBI's next-generation data warehouse, used by banks to submit returns and by RBI for analytics.
- **Governance** — The chain needs data owners, a data dictionary, lineage records and access controls.

Where value leaks and how to stop it

Stage	Typical leak	Remedy
Capture	Manual entry errors, missing fields	Mandatory fields and validation at source
Storage	Same customer under several IDs	Master data management; one Unique Customer Identification Code (UCIC)
Extraction	Wrong joins or stale extracts	Tested queries; refresh time-stamps on every report
Analysis	Biased or outdated models	Independent validation and drift monitoring
Action	Insights never acted upon	Named owners and follow-up KPIs

Topic 5 — Sources of Data — Meaning and Classification

A **data source** is any system, device, person or organisation from which data originates.

Basis	Types	Banking illustration
Origin	Internal / External	CBS and loan system (internal); credit report, GST data (external)
Collection	Primary / Secondary	Bank's own customer survey (primary); RBI statistical tables (secondary)
Generation	Human-generated / Machine-generated	Loan application form; ATM logs, server logs, sensor readings
Nature	Transactional / Master (reference)	Daily debits and credits; customer, branch and product masters
Timing	Batch / Real-time (streaming)	End-of-day general ledger feed; UPI transaction stream

- **Syllabus focus** — Three key sources are **Management Information Systems**, the **Internet of Things** and **social networks**.
- **Primary data** — Collected first-hand for the current purpose; costly but specific.
- **Secondary data** — Collected earlier by someone else for another purpose; cheap but may not fit exactly.

Other external sources for Indian banks

- **Credit information companies** — Four CICs are licensed by RBI under the Credit Information Companies (Regulation) Act, 2005.
- **Account Aggregators** — NBFC-AAs move data from Financial Information Providers to Financial Information Users only on explicit, revocable consent.
- **Data-blind** — An Account Aggregator cannot see, store or sell the data it carries.
- **Alternative data** — GST returns, utility and telecom payments and e-commerce records help assess thin-file borrowers.

Remember this

- **Internal vs external** = where the data originates; **primary vs secondary** = whether it was collected first-hand for the present purpose.

Topic 6 — Management Information Systems as a Source of Data

A **Management Information System (MIS)** collects, processes and presents data from operational systems to support planning, control and decisions.

System level	Main users	Role as a data source
Transaction Processing System (TPS)	Operational staff	Records day-to-day transactions; the raw data source
MIS	Middle management	Summarises TPS data into periodic and exception reports
Decision Support System (DSS)	Analysts and managers	Applies models to semi-structured decisions
Executive Information System (EIS)	Top management	Dashboards combining internal and external data

Types of MIS reports

- **Scheduled** — Produced at fixed intervals, such as a daily branch position or a monthly NPA report.
- **Demand (ad hoc)** — Produced on request, such as deposits mobilised under a new scheme in one region.
- **Exception** — Lists only items outside limits, such as overdrafts above the sanctioned limit.
- **Drill-down** — Lets the user move from a summary figure to the underlying detail.
- **Bank MIS feeds** — CBS, loan origination, treasury, cards, payments, HRMS and CRM systems.
- **Strength** — MIS data is structured, reconciled to the books and relatively reliable.
- **Limitation** — It mostly reflects **internal, past** activity and rarely captures sentiment or real-time context.
- **Automated data flow** — Pulling returns directly from source systems reduces manual intervention and errors.

Example: The MIS of **Narmada Co-operative Bank** shows 18% of new savings accounts with no transaction in 90 days. A drill-down by branch and channel traces most of them to one campaign in which debit cards were never activated.

Remember this

- **TPS** records → **MIS** summarises → **DSS** models → **EIS** gives executives the big picture.

Topic 7 — Internet of Things as a Source of Data

The **Internet of Things (IoT)** is a network of physical objects fitted with sensors, software and connectivity that collect and exchange data with little human involvement.

Component	Role
Devices and sensors	Sense temperature, location, motion, usage or status
Connectivity	Wi-Fi, Bluetooth, RFID/NFC, cellular (4G/5G) and low-power wide-area networks
Gateway / edge	Filters and pre-processes readings close to the device
Platform / cloud	Stores and manages streaming data at scale
Analytics and application	Turns readings into alerts, dashboards and automated actions

Characteristics of IoT data

- **High velocity** — Devices send readings continuously, often every few seconds.
- **Time-series** — Each reading carries a device identifier and a time-stamp.
- **Machine-generated** — Little manual entry, but faulty sensors create noise and gaps.
- **Edge computing** — Processing near the device cuts bandwidth use and response time.
- **Security risk** — Default passwords and unpatched firmware make devices easy entry points for attackers.



Banking use	How IoT data helps
ATM and branch monitoring	Sensors report cash levels, faults, door status and temperature; supports predictive maintenance and cash planning
Toll payments	FASTag RFID tags are read at toll plazas and the linked account or wallet is debited automatically
Vehicle and equipment loans	Telematics show usage and location of financed assets for monitoring and recovery
Motor insurance	Driving data supports usage-based (pay-as-you-drive) premiums
Agriculture lending	Soil, weather and satellite data support crop assessment and claims
Wearable payments	NFC-enabled rings, watches and bands make contactless payments

N1. [Numerical] A bank monitors 6,000 ATMs. Each ATM has 8 sensors, and each sensor sends one reading of 200 bytes every 30 seconds. Find the readings per day and the data generated per day and per 30 days (1 GB = $1,024^3$ bytes).

Item	Value
ATMs	6,000
Sensors per ATM	8
Reading interval	30 seconds
Size per reading	200 bytes

Solution:

Formula:

$$\text{Readings/day} = \text{Devices} \times \frac{86,400}{\text{Interval (s)}}$$

Readings per sensor per day = $86,400 \div 30 = 2,880$.

Readings = $6,000 \times 8 \times 2,880 = 13,82,40,000$

Bytes = $13,82,40,000 \times 200 = 27,64,80,00,000$

GB/day = $27,64,80,00,000 \div 1,024^3 = 25.75$

Per 30 days = $25.75 \times 30 = 772.5$ GB, about 0.75 TB.

Answer: 13.82 crore readings a day; about 25.75 GB a day and 772.5 GB ($\approx$ 0.75 TB) in 30 days.

Remember this

- IoT data is **machine-generated, high-velocity, time-stamped** streaming data; **edge computing** processes it near the source.

Topic 8 — Social Networks as a Source of Data

Social networks are online platforms where users create profiles, connect with others and share content, producing huge volumes of user-generated data.

Data type	Examples	Analytics use
Content	Posts, comments, reviews, images, videos	Sentiment analysis, complaint detection
Engagement	Likes, shares, mentions, hashtags	Campaign reach and brand health
Network (graph)	Followers, friends, groups	Influencer identification, ring detection
Profile and metadata	Stated location, interests, time of post	Segmentation and location trends

Social network analysis

- Nodes and edges** — Users or accounts are nodes; relationships or interactions are edges.
- Centrality** — Measures how connected or influential a node is within the network.
- Communities** — Clusters of closely linked nodes; the same graph method finds mule-account rings in transfer data.

Banking uses

- Social listening** — Tracks mentions of the bank to spot service failures early.
- Grievances** — Complaints raised on social media are logged into the bank's grievance redress system.
- Sentiment** — Natural language processing classifies posts as positive, negative or neutral.
- Impersonation** — Fake customer-care handles and phishing pages using the bank's name are detected and reported.

Limitations

- Unverified** — Content may be false, exaggerated or posted by bots.
- Unrepresentative** — Active users are not a fair sample of all customers.
- Hard to parse** — Slang, code-mixed Hinglish, sarcasm and emojis confuse text analytics.
- Privacy** — The DPDP Act, 2023 does not apply to personal data made publicly available by the person; platform terms and ethics still apply.

Example: Godavari Bank sees a spike in posts saying "UPI failed" between 10 and 11 a.m. Social listening alerts the operations team before call-centre volumes rise, and a server fault is fixed within the hour.

Topic 9 — Structured Data and Unstructured Data

Data is classified by how well it fits a predefined data model, or schema.

Feature	Structured	Semi-structured	Unstructured
Format	Fixed rows and columns	Tags or keys with flexible fields	No predefined format
Schema	Schema-on-write (defined before loading)	Self-describing	Schema-on-read (applied at analysis)
Storage	Relational databases, data warehouses	NoSQL document stores, files	Data lakes, object storage, file systems
Analysis	SQL, spreadsheets, BI tools	Parsers, NoSQL queries	NLP, speech, image and video analytics
Bank example	CBS account and loan tables, general ledger	JSON API responses, XML (ISO 20022) payment messages	Scanned KYC documents, call recordings, emails, CCTV footage

- Share** — Industry research estimates that about **80%** of the world's data is unstructured.
- Nature** — Structured data is mostly quantitative; unstructured data is mostly qualitative.
- Metadata** — Data about data (file name, size, author, time-stamp) gives an unstructured file some structure.
- Search** — Structured data is easy to query; unstructured data first needs indexing or extraction.

Remember this

- Structured** = fixed schema (SQL tables); **semi-structured** = tags or keys (JSON, XML); **unstructured** = no schema (text, audio, images, video).



LEARNING SESSIONS | ASHISH SIR

YOUTUBE: Learning Sessions

Free live classes & topic-wise videos

JAIIB • CAIIB • Certifications • Bank Promotion

**Topic 10 — Handling Unstructured Data**

Data	Technique	Structured output
Scanned KYC documents and cheques	OCR and intelligent document processing	Name, PAN, address and amount fields
Call-centre recordings	Speech-to-text followed by NLP	Complaint category and sentiment
Emails, chats, social posts	Text mining and NLP	Topic, intent and sentiment scores
Images and video	Computer vision	Face match, tampering flags, CCTV events

- **Conversion** — Most analysis first converts unstructured content into structured features such as fields, tags and scores.
- **Two-tier storage** — Raw files stay in a data lake or object store; extracted fields move to the warehouse.
- **Cost** — Unstructured data needs more storage, computing power and specialised skills.
- **Value** — It explains reasons and emotions, such as why a customer is leaving, that numbers alone cannot.

Example: Sutej Bank converts 2 lakh call recordings a month into text. Topic analysis shows "home loan EMI change" as the top complaint theme, which never appeared in structured complaint codes because agents had chosen "Others".

Topic 11 — Data Quality, Privacy and Governance of Data Sources

Quality dimension	Meaning	Typical check
Accuracy	Values correctly reflect the real event	Match with the source document
Completeness	All required fields are filled	% of mandatory fields populated
Consistency	The same value across systems	Mobile number in CBS vs CRM
Timeliness	Data is current when used	Refresh time-stamp
Validity	Values follow formats and rules	PAN pattern, permitted date range
Uniqueness	No duplicate records	One UCIC per customer

LEARNING SESSIONS | ASHISH SIR

JAIB • CAIB • Certifications • Bank Promotion

Complete Books • Live Classes • Mock Tests • Crash Courses

GET THE APP



iibf.info/app

YOUTUBE



Learning Sessions

WHATSAPP



8360944207

N2. [Numerical] A bank holds 2,50,000 customer records. Mobile number is missing in 7,500 records. PAN is captured in 2,40,000 records, of which 3,600 fail the format check. 1,250 customers hold two customer IDs. Compute mobile completeness, PAN validity and the duplicate rate.

Item	Value
Total records	2,50,000
Mobile missing	7,500
PAN captured	2,40,000
PAN invalid format	3,600
Customers with two IDs	1,250

Solution:

Formula:

$$\text{Completeness} = \frac{\text{Populated}}{\text{Total}} \times 100$$

$$\text{Validity} = \frac{\text{Valid}}{\text{Populated}} \times 100$$

$$\text{Mobile completeness} = \frac{2,42,500}{2,50,000} \times 100 = 97.0\%$$

$$\text{PAN validity} = \frac{2,36,400}{2,40,000} \times 100 = 98.5\%$$

$$\text{Duplicate rate} = \frac{1,250}{2,50,000} \times 100 = 0.5\%$$

Answer: Mobile completeness 97.0%; PAN validity 98.5% (PAN completeness is 96.0%); duplicate rate 0.5%.

Legal and regulatory guardrails

- **DPDP Act, 2023** — Governs digital personal data; consent must be free, specific, informed, unconditional and unambiguous.
- **DPDP Rules, 2025** — Notified on 13 November 2025; consent-manager rules apply from 13 November 2026 and the main duties from 13 May 2027 (both after the exam cut-off).
- **Payment data** — Under RBI's 2018 directive, the entire payment system data must be stored only in India.
- **Processing abroad** — Data processed outside India must be deleted there and brought back within one business day or 24 hours, whichever is earlier.
- **KYC records** — Transaction records are kept 5 years from the transaction; identity records 5 years after the relationship ends.

Remember this

- **Completeness** uses all records as the base; **validity** uses only populated values. Payment data: **stored only in India**; back within **1 business day or 24 hours**.

ET | Module A | CH-1 | MCQs

Q1. Kaveri Bank's analytics team builds a churn model that correctly flags 2,000 customers likely to close accounts. The retention team never contacts them, and most of them leave. In information value chain terms, where was value lost? [Ref: Topic 1, p. 2]

- At data capture, because the source systems recorded wrong attributes
- At data storage, because the warehouse retained too little history
- At analysis, because a churn model cannot predict account closure
- At decision and action, because the insight was never acted upon



Q2. Arrange the following stages of the information value chain in the correct order: 1. Analysis 2. Data capture 3. Data extraction 4. Decision 5. Data storage [Ref: Topic 2, p. 2]

- (a) 2, 3, 5, 1, 4 (b) 2, 5, 3, 1, 4
(c) 5, 2, 3, 1, 4 (d) 2, 5, 1, 3, 4

Q3. A fraud report reads: "Customer 4471 has made three night-time overseas ATM withdrawals this week." At which level of the DIKW hierarchy does this statement sit? [Ref: Topic 3, p. 2]

- (a) Data, because it consists only of raw transaction values
(b) Information, because data has been organised in context
(c) Knowledge, because it explains the cause of the pattern
(d) Wisdom, because it prescribes the best action to take

Q4. Two banks have identical fraud models. Bank P scores card transactions in real time; Bank Q runs the same model on a month-end batch. Which statement best explains why Bank P gains more value from the same data? [Ref: Topic 2, p. 2]

- (a) Real-time data is always more accurate than batch data
(b) Bank Q's data is unstructured while Bank P's data is structured
(c) The value of information decays with delay, so earlier action prevents loss
(d) Batch processing does not permit any analysis of transaction data

Q5. A merchant category code is wrongly captured at the card switch. Downstream reports, models and offers all use the wrong category. Which statement is correct? [Ref: Topic 2, p. 2]

- (a) The error is best corrected in the final dashboard by a manual override each month
(b) Only the storage stage can introduce errors that travel down the chain
(c) Errors at capture do not matter if the analytics model is sophisticated enough
(d) The error is cheapest to fix at capture, as every later stage inherits it

Q6. A bank analyst is tagging data sources. Which of the following is BOTH internal and machine-generated? [Ref: Topic 5, p. 3]

- (a) ATM switch logs generated by the bank's own ATM network
(b) Credit report received from a credit information company
(c) Loan application form filled in by a customer at a branch
(d) District-wise deposit statistics published by RBI

Q7. Which of the following statements about MIS as a source of data are correct? 1. MIS data is largely structured and reconciled with the books. 2. MIS captures customer sentiment in real time. 3. MIS mostly reflects internal and past activity. [Ref: Topic 6, p. 3]

- (a) 1 and 2 only (b) 2 and 3 only (c) 1 and 3 only (d) 1, 2 and 3 only

Q8. A regional manager receives a report listing only those cash-credit accounts in which the outstanding exceeds the sanctioned limit. This is an example of a/an: [Ref: Topic 6, p. 3]

- (a) Scheduled report produced at fixed intervals for all accounts
(b) Exception report highlighting items outside set limits
(c) Demand report produced once on an ad hoc request
(d) Drill-down report moving from a summary to detail

Q9. A bank monitors 2,500 ATMs, each with 6 sensors. Every sensor sends one 250-byte reading every 60 seconds. Approximately how much data is generated per day (1 GB = 1,024³ bytes)? [Ref: Topic 7, p. 3]

- (a) About 0.08 GB (b) About 50.29 GB
(c) About 301.75 GB (d) About 5.03 GB

Q10. Which of the following is NOT a typical characteristic of IoT data? [Ref: Topic 7, p. 3]

- (a) It arrives continuously at high velocity from many devices
(b) Each reading carries a device identifier and a time-stamp
(c) It is mostly keyed in manually by operational staff
(d) Faulty sensors can create noise and gaps in the stream

Q11. Vibration sensors in ATMs must trigger a tamper alarm within milliseconds, even when the network link to the data centre is slow. Which design best meets this need? [Ref: Topic 7, p. 3]

- (a) Process the readings on an edge gateway at or near the ATM
(b) Send all readings to the warehouse in a nightly batch file
(c) Store readings in the data lake and review them every month
(d) Reduce the sensor reading frequency to once every hour

Q12. In a graph of accounts linked by fund transfers, an account connected to an unusually large number of newly opened accounts can best be identified using which measure? [Ref: Topic 8, p. 4]

- (a) A sentiment score derived from customer posts
(b) A centrality measure from network analysis
(c) An optical character recognition confidence score
(d) A schema-on-write validation rule

Q13. Godavari Bank finds that 70% of social-media posts about it last month were positive. Which is the strongest reason NOT to treat this as its overall customer satisfaction? [Ref: Topic 8, p. 4]

- (a) Social-media data cannot be analysed with natural language processing tools
(b) Social-media posts are structured data and therefore lack any context
(c) Positive posts are always generated by bots and should be removed
(d) Active social-media users are not a representative sample of all customers

Q14. Which of the following is the best example of semi-structured data in a bank? [Ref: Topic 9, p. 4]

- (a) A loan account table in the core banking database
(b) A recorded customer complaint call in audio format
(c) An ISO 20022 payment message in XML format
(d) A scanned image of a cheque for clearing

Q15. A bank loads raw chat transcripts into a data lake without defining fields, and analysts apply structure only when they query the data. This approach is known as: [Ref: Topic 9, p. 4]

- (a) Schema-on-read (b) Schema-on-write
(c) Data normalisation (d) Master data management

Answer key

Q	Ans	Q	Ans	Q	Ans	Q	Ans	Q	Ans
1	d	4	c	7	c	10	c	13	d
2	b	5	d	8	b	11	a	14	c
3	b	6	a	9	d	12	b	15	a



LEARNING SESSIONS | ASHISH SIR
GET THE APP: iibf.info/app

Notes, mock tests & live classes on your phone
JAIIB • CAIIB • Certifications • Bank Promotion

CH-2

Data Storage, Extraction and Types of Analytics

Topic 1 — Data Storage — Meaning, Units and Storage Tiers

Data storage is the recording and retention of digital data so that it can be retrieved, processed and protected throughout its life cycle.

Unit	Size (binary convention)	Unit	Size (binary convention)
Byte	8 bits	Terabyte (TB)	1,024 GB
Kilobyte (KB)	1,024 bytes	Petabyte (PB)	1,024 TB
Megabyte (MB)	1,024 KB	Exabyte (EB)	1,024 PB
Gigabyte (GB)	1,024 MB	Zettabyte (ZB)	1,024 EB

- **Primary storage** — RAM and cache; fastest and costliest, but **volatile** (lost when power fails).
- **Secondary storage** — Solid-state drives and hard disks; non-volatile and used for live databases.
- **Tertiary (archival) storage** — Tape, optical media and cloud archive; cheapest and slowest, used for long retention.

Tier	Access pattern	Typical media	Banking illustration
Hot	Constant, millisecond access	In-memory, SSD	Live CBS balances, UPI transaction data
Warm	Occasional access	Hard disk, standard cloud storage	Last two years of statements
Cold	Rare access, long retention	Tape, archive cloud tier	KYC images of accounts closed four years ago

Remember this

- Speed and cost fall together from primary to archival storage; place data in a tier according to **how often it is accessed** and **how long it must be kept**.

Topic 2 — Data Storage Architectures — DAS, NAS, SAN and Object Storage

Architecture	How it works	Access level	Typical use
DAS (Direct-Attached Storage)	Disks attached directly to one server	Block	Small branch or single-application on servers
NAS (Network-Attached Storage)	File server shared over the office LAN	File (shared folders)	Scanned documents, shared drives
SAN (Storage Area Network)	Dedicated high-speed network (Fibre Channel or iSCSI) linking servers to pooled disks	Block	CBS and payment databases needing fast input-output

Storage type	Organisation	Best suited for
File storage	Files in a hierarchy of folders and paths	User documents and shared drives

Storage type	Organisation	Best suited for
Block storage	Data split into fixed-size blocks with addresses	Databases and virtual machines needing low latency
Object storage	Each object holds data, rich metadata and a unique ID in a flat namespace	Unstructured data, backups and data lakes at massive scale

RAID — protection against disk failure

- **RAID 1** — Mirrors data on two disks; survives one disk failure.
- **RAID 5** — Stripes data with distributed parity; survives one disk failure.
- **RAID 6** — Uses double parity; survives two disk failures.
- **RAID 10** — Combines mirroring and striping for speed and redundancy.
- **Not a backup** — RAID does not protect against deletion, corruption or ransomware; backups are still needed.

Remember this

- **NAS** = file-level over the LAN; **SAN** = block-level over a dedicated network; **object storage** = data + metadata + unique ID.

Topic 3 — Databases, Data Warehouses, Data Lakes and Lakehouses

- **Relational database (RDBMS)** — Stores data in related tables queried with SQL; ideal for transactions.
- **ACID** — Transactions are **Atomic** (all or nothing), **Consistent**, **Isolated** and **Durable**.
- **Atomicity in practice** — If the credit leg of a fund transfer fails, the debit leg is rolled back too.
- **NoSQL databases** — Flexible schemas built for scale and varied data; many trade strict consistency for availability.

NoSQL type	Structure	Banking use
Key-value	A unique key mapped to a value	Session data, caching of balances
Document	JSON-like documents with flexible fields	Customer 360 profiles, app content
Column-family	Wide tables grouped by column families	High-volume logs and time-series data
Graph	Nodes and relationships	Fraud-ring and mule-account detection

Store	Data held	Schema	Typical use
Data warehouse	Cleaned, integrated, mainly structured history	Schema-on-write	BI, MIS and regulatory reporting
Data mart	Subset of the warehouse for one function	Schema-on-write	Retail, treasury or risk department analysis
Data lake	Raw data of all types in native format	Schema-on-read	Data science, machine learning, exploration
Data lake house	Lake storage with warehouse-style management	Both	BI and machine learning on one governed platform

- **Data swamp** — A lake without metadata, quality checks and governance becomes an unusable data swamp.

Remember this

- **ACID**: Atomicity, Consistency, Isolation, Durability.
Warehouse = schema-on-write; **lake** = schema-on-read.