Data Science and Analytics

Warehouses, lakes and data safety

The notes for the lesson “Warehouses, lakes and data safety”: chapters 6.7 to 6.8, with their figures, worked examples, exam traps, quizzes and exam question bank, in one file.

2 chapters | Chapters 6.7 to 6.8

About these notes

These are the notes for the University free n easy video lesson **Warehouses, lakes and data safety** on Data Science and Analytics: chapters 6.7 to 6.8, with every figure, worked example, exam trap and quiz, and the long-answer exam questions on these chapters.

6.7 Data warehouses and data lakes

A **data warehouse** is a central store of cleaned, integrated, historical data, organised for analysis and reports and kept separate from the operational databases that run the business. A **data lake** stores raw data of every type at low cost in its original format, and applies a structure only when someone reads it. A warehouse answers known business questions fast and reliably (“sales by state by quarter”); a lake keeps everything for data science and for questions nobody has asked yet. Most large organisations run both, or a **lakehouse** that combines them.

You need: An earlier lesson (Data integration and ETL processes), Chapter 6.2 (HDFS) and Chapter 6.4 (Hive).

The warehouse: definition, OLTP and OLAP

Bill Inmon defined a data warehouse as a **subject-oriented, integrated, time-variant and non-volatile** collection of data that supports management decisions. Subject-oriented: organised around sales, customers and products, not around one application. Integrated: codes, names and units made consistent across sources (“TN”, “Tamil Nadu” and “33” become one value). Time-variant: history is kept and every fact carries a date. Non-volatile: data is loaded and read, not updated in place.

	OLTP (OPERATIONAL DATABASE)	OLAP (WAREHOUSE)
Purpose	Run the business: book a ticket, record a sale	Analyse the business: trends, comparisons
Operations	Many short reads and writes of single rows	Few long read-only queries over millions of rows
Data	Current, detailed	Historical, integrated, often summarised
Design	Normalised (3NF) to avoid update errors	Dimensional (star schema) for fast queries
Users	Clerks, apps, customers	Analysts, managers

Data flows from sources through **ETL** (extract, transform, load) in a staging area, into the warehouse, and from there into **data marts**, smaller subject-specific subsets such as a sales mart. Analysts reach it through an OLAP server and BI (business intelligence) tools. Inmon's approach is top-down (build the enterprise warehouse first, then marts); Ralph Kimball's is bottom-up (build dimensional marts that share common, "conformed" dimensions).

Dimensional modelling and OLAP operations

- A **fact table** holds the numeric **measures** of a business event (quantity, amount) and a foreign key to each dimension. Its **grain** says what one row means, such as one product on one bill line. It is the largest table.
- **Dimension tables** hold the descriptive attributes used to filter and group: date (day, month, quarter, year), store (city, state), product (brand, category).
- **Star schema**: one fact table joined directly to denormalised dimensions. One join per dimension, simple and fast.
- **Snowflake schema**: dimensions normalised into sub-tables (store → city → state). Less redundancy, more joins.
- **Fact constellation** (galaxy): several fact tables, such as sales and shipments, sharing dimensions.

The data can be pictured as a **cube** with one axis per dimension and a measure in each cell. The five OLAP operations are **roll-up** (aggregate up a hierarchy, city → state, or drop a dimension), **drill-down** (the reverse, quarter → month), **slice** (fix one dimension to one value), **dice** (select a sub-cube with conditions on two or more dimensions) and **pivot** (rotate the axes of the report). Measures may be **additive** (sales add across every dimension), **semi-additive** (a bank balance adds across accounts but not across days) or **non-additive** (a profit percentage).

sales, ₹ lakh

	Phones Q1	Phones Q2	Laptops Q1	Laptops Q2
Chennai (TN)	40	**45**	30	**35**
Coimbatore (TN)	25	20	10	12
Bengaluru (KA)	55	**60**	50	**52**
Mysuru (KA)	15	18	8	9

slice Q1: 135 + 98 = 233

dice: Chennai Q2 80, Bengaluru Q2 112

roll-up, phones: city → state

TN	Q1 40 + 25 = 65	Q2 45 + 20 = 65
KA	Q1 55 + 15 = 70	Q2 60 + 18 = 78

fact rows a year:
200 × 5,000 × 365 × 4%
= 14,600,000 rows ≈ 584 MB

Figure 6.7.1 The cube as a table: the blue Q1 columns are the slice (total ₹233 lakh), the green cells the dice for Chennai and Bengaluru in Q2, and adding cities by state is the roll-up.

Data lakes and lakehouses

A data lake lands data as it comes (CSV, JSON, logs, images, audio) in cheap storage, HDFS or a cloud object store, usually in zones: **raw** (exactly as received), **curated** (cleaned, deduplicated, in columnar formats such as Parquet) and **consumption** (ready for reports or models). It uses **schema-on-read**, the opposite of the warehouse's **schema-on-write**. This suits **ELT**: load first, transform later inside the lake with Spark or Hive. Without a catalogue, ownership and quality checks, a lake becomes a **data swamp** that nobody can search or trust (Chapter 6.8). A **lakehouse** adds open table formats (Delta Lake, Apache Iceberg, Apache Hudi) that give lake files ACID transactions, schemas and time travel, so one copy of the data serves both BI and machine learning.

	DATA WAREHOUSE	DATA LAKE
Data	Structured, cleaned, integrated	All types, raw and processed
Schema	On write, designed in advance	On read, applied by each user
Storage cost	Higher (database storage)	Low (files, object store)
Users	Analysts, managers	Data scientists, engineers
Strength	Fast, trusted answers to known questions	Flexibility, new questions, ML

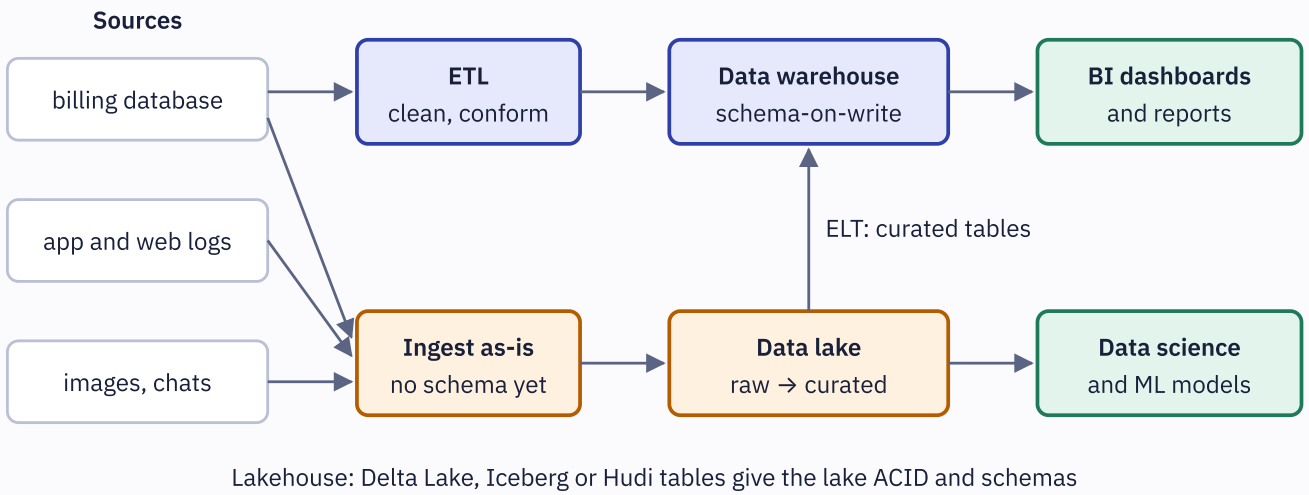


Figure 6.7.2 The two paths. Only structured data takes the top path through ETL into the warehouse; everything lands in the lake, and curated lake tables can also feed the warehouse (ELT). Notice that the images and chats reach only the lake.

WORKED EXAMPLE

Question. A retail chain's sales (₹ lakh) for two quarters are below. (a) Roll up phone sales from city to state. (b) Slice the cube for Q1 and give the total. (c) Dice for city ∈ {Chennai, Bengaluru}, quarter = Q2, both products. (d) The chain has 200 stores and 5,000 products, and on a given day 4% of store–product pairs sell something. At the grain "one product, one store, one day", how many fact rows and how many MB per year, at 40 bytes per row? (10 marks)

CITY (STATE)	PHONES Q1	PHONES Q2	LAPTOPS Q1	LAPTOPS Q2
Chennai (TN)	40	45	30	35
Coimbatore (TN)	25	20	10	12
Bengaluru (KA)	55	60	50	52
Mysuru (KA)	15	18	8	9

1. (a) Roll-up, phones: TN Q1 = 40 + 25 = **65**, TN Q2 = 45 + 20 = **65**; KA Q1 = 55 + 15 = **70**, KA Q2 = 60 + 18 = **78**. Four cities became two states.
2. (b) Slice quarter = Q1: a 4 × 2 table (city × product). Phones 40 + 25 + 55 + 15 = 135; laptops 30 + 10 + 50 + 8 = 98; Q1 total = **₹233 lakh**.
3. (c) Dice: Chennai Q2 = 45 + 35 = **80**; Bengaluru Q2 = 60 + 52 = **112**; together ₹192 lakh. Conditions on two dimensions (city and quarter) leave a smaller cube.
4. (d) Possible rows per year = 200 × 5,000 × 365 = 365,000,000. At 4%: 365,000,000 × 0.04 = **14,600,000 rows**. Size = 14,600,000 × 40 B = 584,000,000 B ≈ **584 MB per year**.

What it shows. OLAP operations are just GROUP BY and WHERE on the star schema. Even a fine grain gives a fact table of modest size, because only combinations that actually sell become rows; drill-down to months would need the month level in Dim_Date.

EXAM TRAPS

- Slice fixes one dimension to one value; dice puts conditions on two or more dimensions. Both give a sub-cube.
- In a star schema the dimensions are denormalised; in a snowflake they are normalised. Do not call the fact table "denormalised".
- A data lake is not a dump: without metadata and governance it becomes a swamp. Schema-on-read still needs a schema; it is applied later.
- ETL transforms before loading (warehouse); ELT loads first and transforms inside the target (lake, cloud warehouse).
- Favourite question: "Differentiate OLTP and OLAP", "Explain star, snowflake and fact constellation schemas" and "Explain OLAP operations with an example".

CHECK YOURSELF

– REMEMBER **Give Inmon's four properties of a data warehouse.** 2 marks

Subject-oriented, integrated, time-variant, non-volatile.

– REMEMBER **What is a data mart?** 1 mark

A smaller warehouse for one subject or department, such as sales, built from or alongside the enterprise warehouse.

– UNDERSTAND **Why is a warehouse kept separate from the OLTP database?** 3 marks

Long analytical queries would slow the short transactions that run the business; the warehouse also needs history and data integrated from many sources, and a dimensional design that would be wrong for frequent updates.

– UNDERSTAND **Explain schema-on-write and schema-on-read with one advantage of each.** 3 marks

Schema-on-write checks data against a design when it is loaded (warehouse): queries are fast and data is trusted. Schema-on-read stores raw data and applies a structure when read (lake): loading is quick and new uses need no redesign.

– APPLY **From the worked example, roll up laptops to state for Q2.** 2 marks

TN = 35 + 12 = **47**; KA = 52 + 9 = **61** (₹ lakh).

– APPLY **Design a star schema for IRCTC bookings: name the fact table, its measures and four dimensions.** 4 marks

Fact_Booking (grain: one passenger on one ticket) with measures fare, number of passengers, cancellation charge. Dimensions: Date (journey and booking), Train (number, type, zone), Station (from, to, as role-playing dimensions), Passenger (age band, gender, quota), Class (SL, 3A, 2A).

– ANALYZE **A hospital wants monthly bed-occupancy reports and also wants to train a model on X-ray images. Warehouse, lake or both?** 4 marks

Both, or a lakehouse. Occupancy is structured and asked every month: a warehouse star schema gives fast, trusted reports. X-rays are unstructured and need raw storage for ML: a lake. A lakehouse could serve both from one store, with access controls for patient data (Chapter 6.8).

BEYOND THE SYLLABUS

This is not for the exam. Cloud warehouses (Snowflake, Google BigQuery, Amazon Redshift) separate storage from compute and scale each on demand. dbt has made ELT with SQL models a standard practice, and data mesh argues for domain-owned data products instead of one central warehouse team. Search for “medallion architecture” (bronze, silver, gold zones) and “Apache Iceberg”.

6.8 Data security, privacy and governance

Big data platforms hold data about people: payments, locations, health records, browsing. **Security** keeps that data from being read, changed or destroyed by the wrong people. **Privacy** decides what personal data may be collected and used at all, and protects the people it describes. **Data governance** is the set of roles, policies and processes that keep data accurate, documented, secure and lawful through its whole life. India’s Digital Personal Data Protection Act, 2023 (DPDP Act) and the European Union’s GDPR turn these into legal duties with large penalties, so every data scientist needs the basics.

You need: Chapters 6.2 to 6.7 (where the data sits); an earlier lesson (Data sources and data collection).

Security: the CIA triad on a Hadoop cluster

Security aims at three properties: **confidentiality** (only authorised people read the data), **integrity** (data is not altered without detection) and **availability** (data can be used when needed). Out of the box, Hadoop trusts whatever user name a client claims, so a production cluster adds layers, each answering one question:

- **Perimeter**: the cluster sits on a private network behind a firewall; users enter through a gateway such as Apache Knox.
- **Authentication** (who are you?): Hadoop's secure mode uses **Kerberos**. A user logs in to the Key Distribution Centre once and receives a ticket; every service, NameNode and DataNodes included, checks that ticket instead of a password.
- **Authorization** (what may you do?): HDFS has Unix-style permissions (owner, group, others; read, write, execute) and access control lists (ACLs). **Apache Ranger** manages central policies down to a Hive column, with row filters and column masking ("analysts see only the last 4 digits of the phone number").
- **Encryption**: in transit with TLS; at rest with HDFS **encryption zones**, directories whose files are encrypted with keys held in a key management server (KMS), so a stolen disk is unreadable.
- **Auditing**: every access is logged (who, what, when, allowed or denied) so misuse can be found and proved.

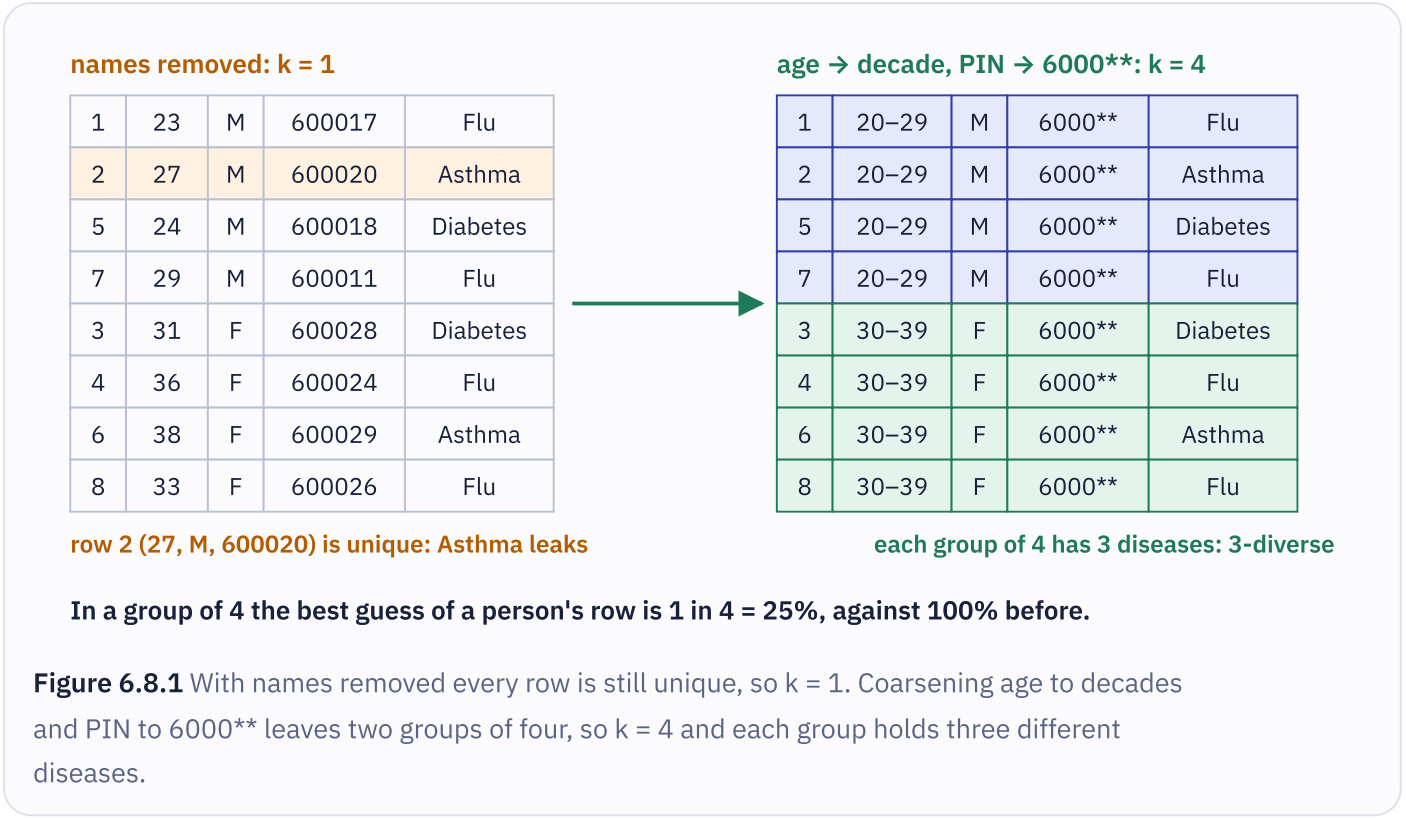


Figure 6.8.1 With names removed every row is still unique, so k = 1. Coarsening age to decades and PIN to 6000** leaves two groups of four, so k = 4 and each group holds three different diseases.

Privacy: personal data and how to protect it

Personal data is any data about an identifiable person: name, Aadhaar number, phone, email, face photo. **Direct identifiers** name a person by themselves. **Quasi-identifiers** such as age, gender and PIN code look harmless, but combined they often single out one person, so deleting the name column is not enough. The techniques, from weakest to strongest:

- **Data minimisation**: do not collect what the purpose does not need.
- **Masking**: hide part of a value for display or testing (XXXX-XXXX-4821).
- **Pseudonymisation** or **tokenisation**: replace an identifier with a token; whoever holds the mapping table can reverse it, so the data is still personal data.
- **Anonymisation**: make re-identification impossible in practice, for example by **generalisation** (age 23 → 20–29) and **suppression** (delete a value or a row). A table is **k-anonymous** if every combination of quasi-identifier values appears in at least k rows, so each person hides in a crowd of k. **l-diversity** further requires at least l different sensitive values in each such group.

Under the **DPDP Act, 2023**, the person is the **Data Principal**, the organisation that decides why and how data is processed is the **Data Fiduciary**, and a company processing on its behalf is a **Data Processor**. Processing needs a lawful purpose and, for most uses, consent that is free, specific, informed and unambiguous, given after a clear notice. Data Principals have rights to access, correction, erasure and grievance redressal. Fiduciaries must protect the data with reasonable security safeguards and report breaches to the Data Protection Board of India and to the people affected; failing to take those safeguards can cost up to ₹250 crore. The EU's **GDPR** (applied from May 2018) has similar ideas, adds rights such as data portability, and fines up to €20 million or 4% of worldwide annual turnover, whichever is higher.

Governance: roles, metadata and lineage

Governance assigns people and rules. The **data owner** (a business head) is accountable for a data set and approves access; the **data steward** defines terms and watches quality; the **data custodian** (the IT team) stores, secures and backs it up. Data is **classified** (public, internal, confidential, restricted) and each class gets its controls. **Data quality** is measured on accuracy, completeness, consistency, timeliness, validity and uniqueness. A **data catalogue** lists what data exists, what each column means and who owns it, and **lineage** records where each column came from and which jobs changed it (Apache Atlas does this for Hadoop). A **lifecycle** policy covers creation, storage, use, sharing, archiving and deletion, with retention periods. This is what keeps a lake from becoming a swamp (Chapter 6.7).

WORKED EXAMPLE

Question. A hospital wants to share the table below with researchers. (a) Classify the columns. (b) Find k after removing names. (c) Find k after generalising age to decades and PIN to five digits (60001*, 60002*). (d) Make the table at least 3-anonymous and check l-diversity. (e) What is the largest chance of re-identifying a person in your final table? (8 marks)

#	NAME	AGE	GENDER	PIN	DISEASE
1	Arun	23	M	600017	Flu
2	Bala	27	M	600020	Asthma
3	Chitra	31	F	600028	Diabetes
4	Divya	36	F	600024	Flu
5	Ezhil	24	M	600018	Diabetes
6	Farah	38	F	600029	Asthma
7	Gopal	29	M	600011	Flu
8	Hema	33	F	600026	Flu

1. (a) Name: direct identifier. Age, Gender, PIN: quasi-identifiers. Disease: sensitive attribute.
2. (b) With names removed, every (age, gender, PIN) triple is different, so each group has 1 row: **k = 1**. Anyone who knows Bala is a 27-year-old man from 600020 learns he has asthma.
3. (c) Groups: (20–29, M, 60001*) = rows 1, 5, 7 (3 rows); (20–29, M, 60002*) = row 2 alone (1 row); (30–39, F, 60002*) = rows 3, 4, 6, 8 (4 rows). k is the smallest group: **k = 1**, because Bala is still alone.
4. (d) Generalise PIN one more digit, to 6000**. Groups: (20–29, M, 6000**) = rows 1, 2, 5, 7; (30–39, F, 6000**) = rows 3, 4, 6, 8. Both have 4 rows: **k = 4**. Diseases: group 1 has Flu, Asthma, Diabetes, Flu; group 2 has Diabetes, Flu, Asthma, Flu. Each has 3 distinct values: **3-diverse**. (Another way: keep five-digit PINs and suppress row 2, giving k = 3 with 7 rows.)
5. (e) In a group of 4 an attacker can at best guess the right row with chance 1 ÷ 4 = **25%**, against 100% in (b).

What it shows. Every step of generalisation lowers the risk and also the detail researchers get (a PIN now covers a wider area). Check l-diversity too: if all four people in a group had flu, knowing that Arun is in the group would reveal his disease without finding his row.

EXAM TRAPS

- Privacy is not the same as security. A perfectly secured database still breaks privacy if it holds data collected without consent or uses it for another purpose.
- Removing names is not anonymisation; quasi-identifiers re-identify people. Pseudonymised data is still personal data.
- Authentication asks who you are (Kerberos); authorization asks what you may do (ACLs, Ranger).
- In the DPDP Act the person is the Data Principal and the company is the Data Fiduciary, not the other way round.
- Favourite question: “What are the security and privacy challenges of big data, and how does the Hadoop ecosystem address them?”

CHECK YOURSELF

– REMEMBER **Expand CIA in security and define each term.** 2 marks

Confidentiality: only authorised people read data. Integrity: data is not changed without detection. Availability: data is usable when needed.

– REMEMBER **Name the three roles of data governance and one duty of each.** 2 marks

Owner: accountable, approves access. Steward: definitions and quality. Custodian: storage, security, backup.

– UNDERSTAND **How does Kerberos authenticate a user to HDFS without sending a password to the NameNode?** 3 marks

The user proves identity once to the Key Distribution Centre and receives a ticket-granting ticket, then a service ticket for the NameNode. The NameNode checks the ticket, which is encrypted with a key it shares with the KDC; no password crosses the network to it.

– UNDERSTAND **Distinguish pseudonymisation from anonymisation.** 2 marks

Pseudonymisation replaces identifiers with tokens that can be reversed with a key, so it is still personal data. Anonymisation removes the link to a person irreversibly, so the data is no longer personal.

– APPLY **Quasi-identifier groups in a released table have sizes 5, 3, 7 and 2. What is k, and what is the worst re-identification risk?** 2 marks

k = smallest group = **2**; worst risk 1 ÷ 2 = **50%**.

– APPLY **A fintech app stores phone numbers, Aadhaar numbers and transaction amounts in Hive. Suggest one control for each.** 3 marks

Phone: Ranger column masking showing the last 4 digits to analysts. Aadhaar: tokenisation, with the mapping table in a restricted encryption zone. Amounts: visible to analysts but in an encryption zone at rest, with access audited.

– ANALYZE **A company encrypts all its data and claims it is therefore DPDP-compliant. Evaluate the claim.** 4 marks

Encryption covers only the security-safeguard duty. The Act also needs a lawful purpose and valid consent with notice, use limited to that purpose, honouring access, correction and erasure requests, breach reporting and deletion when the purpose is over. So the claim is incomplete.

BEYOND THE SYLLABUS

This is not for the exam. Differential privacy adds carefully measured random noise to query answers so that no single person's presence can be detected; the US Census Bureau used it for the 2020 census. Federated learning trains models on phones without collecting raw data, and the EU AI Act (2024) adds governance rules for AI systems built on data. Search for "differential privacy", "privacy-enhancing technologies" and "DPDP Rules".

Key terms of this lesson

TERM	MEANING OR FORMULA	CHAPTER
Warehouse (Inmon)	Subject-oriented, integrated, time-variant, non-volatile	6.7
OLAP operations	Roll-up, drill-down, slice, dice, pivot	6.7
Fact rows	Product of the dimension sizes at the grain × fraction that occurs	6.7
k-anonymity	Every quasi-identifier combination appears in ≥ k rows; risk ≤ 1 ÷ k	6.8
DPDP Act roles	Data Principal (person), Data Fiduciary (decides purpose), Data Processor	6.8

Exam question bank

Long-answer exam questions on these chapters, each with an outline of the points an examiner looks for. Answer one in full on paper before you read its outline.

Remember and understand

– UNDERSTAND **What is a data warehouse? Compare OLTP and OLAP, and explain the star, snowflake and fact constellation schemas with diagrams.** 13 marks

Inmon's four properties; OLTP versus OLAP table; ETL, staging, marts, BI; fact table, grain, measures, dimensions; star, snowflake, galaxy; Inmon top-down versus Kimball bottom-up.

Apply

– APPLY **Design a star schema for UPI transactions at a bank and write the OLAP operations for "amount by state per quarter", "only Q4" and "only Maharashtra and Karnataka in Q4".** 13 marks

Fact_Payment (grain: one transaction) with amount, fee, success flag; dimensions Date, Payer (age band, state), Merchant (category, city, state), App, Bank. Roll-up from city to state and from date to quarter; slice quarter = Q4; dice state ∈ {MH, KA} and quarter = Q4. Note semi-additive account balances if included.

– APPLY **Given a patient table with age, gender, PIN and diagnosis, make it 3-anonymous by generalisation and suppression, check l-diversity, and state the re-identification risk.** 8 marks

Classify columns; group by quasi-identifiers; generalise age to bands and PIN to fewer digits until every group has ≥ 3 rows, suppressing an outlier row if cheaper; count distinct diagnoses per group; risk ≤ 1 ÷ 3 ≈ 33% (method of the Chapter 6.8 example).

Analyze and evaluate

– ANALYZE **“HDFS makes HBase unnecessary, and a data lake makes a warehouse unnecessary.” Evaluate both claims.** 13 marks

HDFS gives streaming whole-file access only; HBase adds sorted, indexed, random row reads and writes in milliseconds on top of it, so both are needed. A lake stores everything cheaply but without curated schemas, quality and fast BI; the warehouse gives trusted, fast answers; a lakehouse narrows the gap but needs governance to avoid a swamp. Spans Chapters 6.2, 6.5 and 6.7.

– ANALYZE **A hospital plans to move patient records and scans to a Hadoop data lake for research. Evaluate the privacy and governance risks and propose controls.** 13 marks

Risks: re-identification through quasi-identifiers, insider misuse, purpose creep beyond consent, swamp with unknown copies, breach of unencrypted disks. Controls: classification, owners and stewards; Kerberos and Ranger with column masking; encryption zones; k-anonymity and l-diversity for research extracts; catalogue and lineage (Atlas); audit; consent and retention under the DPDP Act; breach reporting.

Glossary (terms used in this lesson)

BASE	Basically available, soft state, eventually consistent: the relaxed guarantee of many NoSQL stores.
DataNode	HDFS worker that stores blocks and serves reads and writes.
Data Fiduciary	Under the DPDP Act, the organisation that decides why and how personal data is processed.
Data lake	Low-cost store of raw data of every type, structured only when read.
Data warehouse	Subject-oriented, integrated, time-variant, non-volatile store for analysis.
Dice	OLAP operation selecting a sub-cube with conditions on two or more dimensions.
Fact table	Central table of a star schema holding measures and keys to the dimensions.
Lineage	The record of how data was derived: in Spark, the chain of transformations; in governance, where each column came from.
NameNode	HDFS master that keeps the namespace and block locations in memory.
Quasi-identifier	An attribute such as age or PIN that can identify a person when combined with others.
Schema-on-read	Applying structure to stored data only when it is queried.

Star schema	A fact table joined directly to denormalised dimension tables.

Sources and further reading

Textbooks

- Tom White, *Hadoop: The Definitive Guide*, 4th edition, O'Reilly, 2015: HDFS, MapReduce, YARN, Hive, Pig and HBase.
- Seema Acharya and Subhashini Chellappan, *Big Data and Analytics*, Wiley India: Hadoop, NoSQL, MongoDB, Cassandra, Hive and Pig, widely prescribed in Indian universities.
- Davy Cielen, Arno D. B. Meysman and Mohamed Ali, *Introducing Data Science*, Manning, 2016: handling large data, Hadoop and Spark, NoSQL.
- Jure Leskovec, Anand Rajaraman and Jeffrey D. Ullman, *Mining of Massive Datasets*: the chapter on MapReduce and distributed file systems.
- Bill Chambers and Matei Zaharia, *Spark: The Definitive Guide*, O'Reilly, 2018.
- Pramod J. Sadalage and Martin Fowler, *NoSQL Distilled*, Addison-Wesley, 2012.
- Martin Kleppmann, *Designing Data-Intensive Applications*, O'Reilly, 2017: replication, partitioning, consistency and stream processing.
- Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit*, 3rd edition, Wiley: dimensional modelling.
- Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann: the chapter on data warehousing and OLAP.

Open web resources

The open web could not be reached while this book was written, so these links were not opened for it. They are well-known free resources listed from memory; check that each still opens. The DPDP Act penalty ceiling (₹250 crore for failing to take reasonable security safeguards) was confirmed with a web search that returned summaries only.

- Apache Hadoop, "HDFS Architecture" guide: hadoop.apache.org, HdfsDesign.
- Apache Spark documentation, including the RDD programming guide and the Structured Streaming guide: spark.apache.org/docs/latest.
- Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", OSDI 2004: research.google.com archive, mapreduce-osdi04.pdf.
- Leskovec, Rajaraman and Ullman, *Mining of Massive Datasets*, free online edition and lecture slides: mmds.org.

- Apache Cassandra documentation, on data distribution, replication and consistency levels: cassandra.apache.org/doc/latest.
- Ministry of Electronics and Information Technology, Government of India, for the text of the Digital Personal Data Protection Act, 2023: meity.gov.in.

Syllabus: the University free n easy universal syllabus, built from top universities' syllabi.

Built for the University free n easy YouTube channel from the universal syllabus of Data Science and Analytics (Oct 2026). Written with top universities' syllabi and standard textbooks; check a formula against your prescribed book before an exam, and report mistakes on the channel.