Data Mining Important Questions PDF

The most important Data Mining and Warehousing questions are the KDD process, data preprocessing (cleaning, integration, transformation, reduction), data warehouse architecture, star, snowflake and fact constellation schemas, OLAP operations, OLAP versus OLTP, the Apriori and FP-growth algorithms with numericals, decision tree and Naive Bayes classification, k-means and hierarchical clustering, outlier analysis and applications of data mining.

Key takeaways

- Apriori, FP-growth and k-means numericals appear in almost every paper.
- Draw the star schema and the three-tier warehouse architecture neatly.
- Learn the OLAP operations (roll-up, drill-down, slice, dice, pivot) with one example each.
- These are practice questions, not past papers; check your own syllabus and official previous papers.

About these important questions

This page lists the most important Data Mining and Data Warehousing questions for B.Tech CSE and IT exams, arranged unit by unit. The units follow the standard syllabus that most Indian technical universities share, so the order may differ slightly from yours.

Please note: these are practice questions based on common syllabus topics. They are not verified past papers, and we do not claim that any question will appear in your exam. Check your own university syllabus first, then solve the official previous year papers. AKTU students can download them from the AKTU old question paper archive.

Unit 1: Data Warehousing Basics

Covers data warehouse concepts, architecture and design.

Short questions (2 marks)

- Define a data warehouse (Inmon's definition).
- Differentiate between OLTP and OLAP.
- What is a data mart?
- What is metadata in a data warehouse?

Long questions

1. Explain the three-tier data warehouse architecture with a diagram.
2. Explain the ETL process (extract, transform, load).

3. Compare star, snowflake and fact constellation schemas with diagrams for a sales example.
4. Explain the difference between a data warehouse and an operational database.
5. Explain the types of data warehouse: enterprise warehouse, data mart and virtual warehouse.
6. Explain slowly changing dimensions and the difference between fact tables and dimension tables, with a retail example.

Unit 2: OLAP and the Multidimensional Model

Covers data cubes, OLAP operations and OLAP servers.

Short questions (2 marks)

- What is a data cube?
- What is a concept hierarchy?
- Differentiate between ROLAP and MOLAP.

Long questions

1. Explain the OLAP operations roll-up, drill-down, slice, dice and pivot with a sales cube example.
2. Compare ROLAP, MOLAP and HOLAP.
3. Explain data cube computation and materialisation of views.
4. Explain indexing of OLAP data with bitmap and join indexes.
5. Explain measures: distributive, algebraic and holistic.
6. A sales cube has dimensions time, item and location. Show the result of rolling up location from city to country and slicing on time = Q1.

Unit 3: Data Mining and Data Preprocessing

Covers the KDD process, data mining functions and data preprocessing.

Short questions (2 marks)

- List the steps in the KDD process.
- What is noisy data?
- What is data discretisation?

Long questions

1. Explain the KDD process with a diagram.
2. Explain data mining functionalities: characterisation, association, classification, clustering and outlier analysis.
3. Explain methods to handle missing values and noisy data (binning, regression, clustering).
4. Normalise a set of values using min-max, z-score and decimal scaling.
5. Explain data reduction: dimensionality reduction, numerosity reduction and data compression.
6. Explain the major issues in data mining.

Unit 4: Association Rule Mining

Covers frequent patterns, Apriori and FP-growth. This is the main numerical unit.

Short questions (2 marks)

- Define support and confidence.
- What is a frequent itemset?
- What is the Apriori property?

Long questions

1. Find frequent itemsets for a given transaction table with minimum support (say 50%) using Apriori, and generate rules with minimum confidence.
2. Build an FP-tree for the same data and mine frequent patterns with FP-growth.
3. Compare Apriori and FP-growth.
4. Explain methods to improve the efficiency of Apriori.
5. Explain multilevel and multidimensional association rules, and the lift measure.
6. Explain closed and maximal frequent itemsets, and why they reduce the number of patterns you need to store.

Unit 5: Classification, Clustering and Applications

Covers classification, prediction, clustering and outlier detection.

Short questions (2 marks)

- Differentiate between classification and clustering.
- What is an outlier?

Long questions

1. Build a decision tree using information gain for a given data set.
2. Classify a new record using Naive Bayes.
3. Apply k-means to a set of points for two iterations.
4. Explain hierarchical clustering with a dendrogram, and DBSCAN.
5. Explain how to evaluate a classifier: accuracy, precision, recall and cross-validation.
6. Explain the applications of data mining in retail, banking, healthcare and fraud detection.

Sample answer outline: Apriori algorithm numerical

A typical question gives 5 to 9 transactions, a minimum support count and a minimum confidence. Answer like this:

1. **State the Apriori property.** Every subset of a frequent itemset must also be frequent. This lets you prune candidates early.
2. **Find C1 and L1.** Count every single item (C1). Keep items with count at least the minimum support (L1).
3. **Generate C2 from L1.** Join L1 with itself, count each pair in the transactions, and keep the frequent pairs as L2.
4. **Generate C3 and prune.** Join L2 with itself. Remove any 3-itemset that has an infrequent 2-item subset, then count the rest to get L3. Stop when no new frequent itemsets appear.

5. **Generate rules.** For each frequent itemset, form rules such as {A, B} → {C} and compute confidence = support(A, B, C) / support(A, B). Keep rules that meet the minimum confidence.
6. **Present the results** in tables at every step (C1, L1, C2, L2 and so on), and list the final strong rules.

The tables are where most marks are. Keep the counts neat and in the same item order throughout.

Marks weightage and how to write answers

A common pattern, used by AKTU and many other universities, has a compulsory short-answer section that covers every unit, followed by long-answer questions with a choice inside each unit. Marks per question change between universities and years, so confirm the scheme on a recent official paper.

- **Short answers (about 2 marks):** a crisp definition, one example or diagram, and nothing more. Three to five lines is enough.
- **Long answers (7 to 10 marks):** start with a definition, add a labelled diagram or table, explain the steps, and finish with an example, advantages or a comparison.
- For schema and architecture questions, draw the diagram first and label every table or tier.
- For Apriori and FP-growth, show the candidate and frequent itemset tables at every level.

A 7-day revision plan for Data Mining and Warehousing

1. **Days 1 to 5:** one unit a day. Read your notes, then write the long answers for that unit from memory and check them against your notes.
2. **Day 6:** redo one Apriori and one FP-growth problem, one normalisation problem, k-means, and draw the star schema and warehouse architecture.
3. **Day 7:** solve one full official previous year paper in three hours, then check which units cost you time.

Keep a single sheet of formulas, definitions and diagrams for the last hour before the exam.

Frequently asked questions

Which Data Mining and Warehousing topics are most important?

Data warehouse architecture, star and snowflake schemas, OLAP operations, the KDD process, data preprocessing, Apriori and FP-growth, decision trees, Naive Bayes and k-means are asked most often.

How do I avoid mistakes in Apriori numericals?

Write the minimum support as a count, not a percentage, before you start. Keep items in alphabetical order, prune candidates with infrequent subsets, and check each count against the transaction table.

Is this subject useful for data jobs?

Yes. Data warehousing, OLAP and preprocessing are the base of data analyst, BI and data engineering work, and the mining algorithms lead into machine learning.

Are these Data Mining and Warehousing questions from previous year papers?

No. They are practice questions built from topics that appear in most B.Tech Data Mining and Data Warehousing syllabi. Use them with your own university's official previous papers, for example the AKTU archive at https://aktu.ac.in/old-question-paper.html.

Is this PDF free?

Yes. Fill in the short form on this page and the PDF downloads right away.

Sources and further reading

- AKTU: old question papers (official archive)
- NPTEL: online courses from IITs (search for Data Mining)
- scikit-learn: clustering user guide

About this PDF

These are practice questions for revision, prepared by the Uncodemy trainer team from common B.Tech syllabus topics. They are not past papers, a certificate or a promise of marks. Spot an error? Write to info@uncodemy.com and the team will review it.

More free resources and courses: uncodemy.com | info@uncodemy.com | +91 98183 66550