Data Mining (Rapid Miner)

**Oral Questions LP-II**

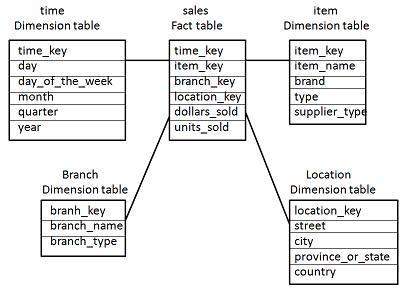
1. What are the different Data Warehousing Schemas?

**ANS:-**

**Schema is a logical description of the entire database**. **It includes the name and description of records** of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema

## Star Schema

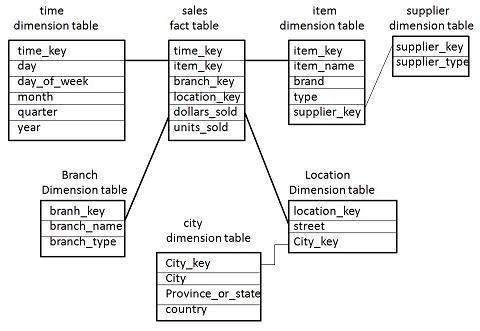
* **Each dimension in a star schema is represented with only one-dimension table.**
* This dimension table contains the set of attributes.
* The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.



* There is a fact table at the center. It contains the keys to each of four dimensions.
* The fact table also contains the attributes, namely dollars sold and units sold.

## Snowflake Schema

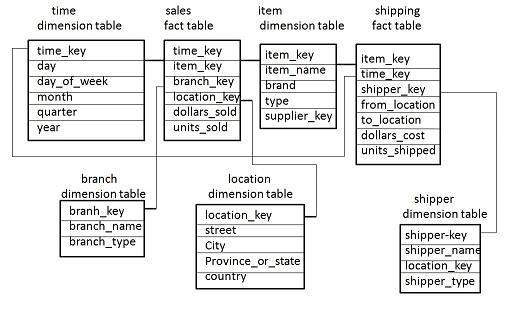
* **Some dimension tables in the Snowflake schema are normalized.**
* The normalization splits up the data into additional tables.
* Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



* Now the item dimension table contains the attributes item\_key, item\_name, type, brand, and supplier-key.
* The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier\_key and supplier\_type.

## Fact Constellation Schema

* **A fact constellation has multiple fact tables. It is also known as galaxy schema.**
* The following diagram shows two fact tables, namely sales and shipping.



* The sales fact table is same as that in the star schema.
* The shipping fact table has the five dimensions, namely item\_key, time\_key, shipper\_key, from\_location, to\_location.
* The shipping fact table also contains two measures, namely dollars sold and units sold.
* It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

2. Explain Star Schema vs. Snowflake Schema

|  |  |  |
| --- | --- | --- |
| S.NO | Star Schema | Snowflake Schema |
| 1. | In star schema, The fact tables and the dimension tables are contained. | While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. |
| 2. | Star schema is a top-down model. | While it is a bottom-up model. |
| 3. | Star schema uses more space. | While it uses less space. |
| 4. | It takes less time for the execution of queries. | While it takes more time than star schema for the execution of queries. |
| 5. | In star schema, Normalization is not used. | While in this, Both normalization and denormalization are used. |
| 6. | It’s design is very simple. | While it’s design is complex. |
| 7. | The query complexity of star schema is low. | While the query complexity of snowflake schema is higher than star schema. |
| 8. | It’s understanding is very simple. | While it’s understanding is difficult. |
| 9. | It has less number of foreign keys. | While it has more number of foreign keys. |
| 10. | It has high data redundancy. | While it has low data redundancy. |

3. Mention what is the responsibility of a Data analyst?

* **Collecting and interpreting data**
* **Analysing results**
* **Reporting the results back to the relevant members of the business**
* **Identifying patterns and trends in data sets**
* [Working alongside teams](https://www.roberthalf.com.au/career-advice/career-development/teamwork-skills) within the business or the management team to establish business needs
* **Defining new data collection and analysis processes**

4. List out some of the best practices for data cleaning?

Convert Data Types

Take Care of Missing Values

### Remove Irrelevant Values

The first and foremost thing you should do is remove useless pieces of data from your system. Any useless or irrelevant data is the one you don’t need. It might not fit the context of your issue.

You might only have to measure the average age of your sales staff. Then their email address wouldn’t be required. Another example is you might be checking to see how many customers you contacted in a month. In this case, you wouldn’t need the data of people you reached in a prior month.

However, before you remove a particular piece of data, make sure that it is irrelevant because you might need it to check its correlated values later on (for checking the consistency). And if you can get a second opinion from a more experienced expert before removing data, feel free to do so.

You wouldn’t want to delete some values and regret the decision later on. But once you’re assured that the data is irrelevant, get rid of it.

### 2. Get Rid of Duplicate Values

Duplicates are similar to useless values – You don’t need them. They only increase the amount of data you have and waste your time. You can get rid of them with simple searches. Duplicate values could be present in your system for several reasons.

Maybe you combined the data of multiple sources. Or, perhaps the person submitting the data repeated a value mistakingly. Some user clicked twice on ‘enter’ when they were filling an online form. You should remove the duplicates as soon as you find them.

### 3. Avoid Typos (and similar errors)

Typos are a result of human error and can be present anywhere. You can fix typos through multiple algorithms and techniques. You can map the values and convert them into the correct spelling. Typos are essential to fix because models treat different values differently. Strings rely a lot on their spellings and cases.

‘George’ is different from ‘george’ even though they have the same spelling. Similarly ‘Mike’ and ‘Mice’ are different from each other, also though they have the same number of characters. You’ll need to look for typos such as this and fix them appropriately.

Another error similar to typos is of strings’ size. You might need to pad them to keep them in the same format. For example, your dataset might require you to have 5-digit numbers only. So if you have any value which only has four digits such as ‘3994’ you can add a zero in the beginning to increase its number of digits.

Its value would remain the same as ‘03994’, but it’ll keep your data uniform. An additional error with strings is of white spaces. Make sure you remove them from your strings to keep them consistent.

### 4. Convert Data Types

Data types should be uniform across your dataset. A string can’t be numeric nor can a numeric be a boolean. There are several things you should keep in mind when it comes to converting data types:

* Keep numeric values as numerics
* Check whether a numeric is a string or not. If you entered it as a string, it would be incorrect.
* If you can’t convert a specific data value, you should enter ‘NA value’ or something of this sort. Make sure you add a warning as well to show that this particular value is wrong.

### 5. Take Care of Missing Values

There would always be a piece of missing data. You can’t avoid it. So you should know how to handle them to keep your data clean and free from errors. A particular column in your dataset may have too many missing values. In that case, it would be wise to get rid of the entire column because it doesn’t have enough data to work with.

Point to note: You shouldn’t ignore missing values.

5. Mention what is data cleansing?

**Data cleansing** or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](https://en.wikipedia.org/wiki/Storage_record) from a record set, [table](https://en.wikipedia.org/wiki/Table_(database)), or [database](https://en.wikipedia.org/wiki/Database) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](https://en.wikipedia.org/wiki/Dirty_data) or coarse data

6. List out some common problems faced by data analyst?

Some of the common problems faced by data analyst are

* Common misspelling
* **Duplicate entries**
* **Missing values**
* **Illegal values**
* Varying value representations
* Identifying overlapping data

7. List of some best tools that can be useful for data-analysis?

* [Tableau](https://career.guru99.com/top-10-tableau-interview-questions/)
* **RapidMiner**
* OpenRefine
* KNIME
* **Google Search Operators**
* **Solver**
* **NodeXL**
* **io**
* Wolfram Alpha’s
* Google Fusion tables

7. What is difference between Supervised and Unsupervised Learning?

|  |  |  |
| --- | --- | --- |
| **BASIS FOR COMPARISON** | **SUPERVISED LEARNING** | **UNSUPERVISED LEARNING** |
| Basic | Deals with labelled data. | Handles unlabeled data. |
| Computational complexity | High | Low |
| Analyzation | Offline | Real-time |
| Accuracy | Produces accurate results | Generates moderate results |
| Sub-domains | Classification and regression | Clustering and Association rule mining |

8. What are different similarities between Kmeans and KNN Algorithm?

Both methods involve computing distances in input space and assigning data points to a **set of nearest 'prototype points**

9. What is Euclidean distance? Explain with Suitable example?

**Distances are normally used to measure the similarity or dissimilarity between two data objects.** ■ The dissimilarity (or similarity) between the objects described by interval-scaled variables is typically computed based on the distance between each pair of objects. ■ The most popular distance measure is Euclidean distance, which is defined as: ■ where i = (xi1 , xi2 , . . . , xin ) and j = (x j1 , x j2 , . . . , x jn ) are two n-dimensional data objects.

10. What is hamming distance? Explain with Suitable example?

**Hamming distance is a metric for comparing two binary data strings**. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different.

The Hamming distance between two strings, a and b is denoted as d(a,b).

**It is used for error detection or error correction when data is transmitted over computer networks**. It is also using in coding theory for comparing equal length data words.

## Calculation of Hamming Distance

In order to calculate the Hamming distance between two strings, and , we perform their XOR operation, (a⊕ b), and then count the total number of 1s in the resultant string.

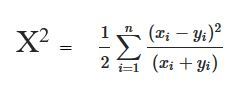
### Example

Suppose there are two strings 1101 1001 and 1001 1101.

11011001 ⊕ 10011101 = 01000100. Since, this contains two 1s, the Hamming distance, d(11011001, 10011101) = 2.

11. What is Chi Square Distance? Explain with Suitable example?

[**Chi-square**](https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/) **distance** calculation is a statistical method, **generally measures similarity between 2 feature matrices**. Such distance is generally used in many applications like similar image retrieval, **image texture, feature extractions etc.** The Chi-square distance of 2 arrays ‘x’ and ‘y’ with ‘n’ dimension is mathematically calculated using below formula :



In this article, we will learn how to calculate Chi-square distance using Python. Below given 2 different methods for calculating Chi-square Distance. Let’s see both of them with examples.

**Method #1:** Calculating Chi – square distance manually using above formula.

|  |
| --- |
| # importing numpy library  import numpy as np    # Function to calculate Chi-distace  def chi2\_distance(A, B):    # compute the chi-squared distance using above formula  chi = 0.5 \* np.sum([((a - b) \*\* 2) / (a + b)  for (a, b) in zip(A, B)])    return chi    # main function  if \_\_name\_\_== "\_\_main\_\_":  a = [1, 2, 13, 5, 45, 23]  b = [67, 90, 18, 79, 24, 98]    result = chi2\_distance(a, b)  print("The Chi-square distance is :", result) |

**Input :** a = [1, 2, 13, 5, 45, 23]

b = [67, 90, 18, 79, 24, 98]

**Output :** The Chi-square distance is : 133.55428601494035

**Input :** a = [91, 900, 78, 30, 602, 813]

b = [57, 49, 36, 759, 234, 928]

**Output :** The Chi-square distance is : 814.776999405035

12. What are different types of Clustering?

1. **Connectivity-based Clustering (Hierarchical clustering)**
2. **Centroids-based Clustering (Partitioning methods)**
3. **Distribution-based Clustering**
4. Density-based Clustering (Model-based methods)
5. Fuzzy Clustering
6. **Constraint-based (Supervised Clustering)**

13. What is Weka Tool? Explain the Step to Perform Clustering on Sample data set?

**Weka** contains a collection of visualization **tools** and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.

14. Explain Association Rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is **Market Based Analysis.**

Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

15. What is the Application of A-Priori algorithm?

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.
2. **In the Medical field: For example Analysis of the patient’s database.**
3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.
4. Apriori is used by many companies like
5. **Amazon in the Recommender System and by Google for the auto-complete feature.**

16. What is Market Basket Analysis? Explain with suitable example?

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.

17. Who propose A-Priori algorithm?

R. Agrawal

* Apriori is an algorithm for frequent item set mining and association rule learning over relational databases.
* It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

18. What is minimum support and minimum confidence?

1. A minimum support threshold is applied to find all *frequent itemsets* in a database.
2. A minimum confidence constraint is applied to these frequent itemsets in order to form rules.

19. What is use of Tokenize operator?

**This operator splits the text of a document into a sequence of tokens.** There are several options how to specify the splitting points. The default setting is non-letter that will result in tokens consisting of one single word and it‟s frequency of occurrence. Other modes available for Tokenize are specifying character, regular expression, linguistic sentences and linguistic tokens.

20. What are different modes of Tokenize operator

21. How to use Read Document operator?

If you scroll down, there is a "Read Documents" operator. Select this operator and enter it into your Main Process window by dragging it. When you select the Read Documents operator in the Main Process window, you should see a file uploader in the right-hand pane.

22. Why we use Filter token and Filter stop word?

23. How to use Filter Class operator?

STQA Questions and Answers

**Q #1) What is the difference between Quality Assurance, Quality Control and testing?**

**Quality Assurance is the process of planning and defining the way of monitoring and implementing the quality (test) processes within a team and organization**. This method basically defines and sets the quality standards of the projects.

**Quality Control is the process of finding defects and providing suggestions to improve the quality of the software**. The methods used by Quality Control are usually established by the quality assurance.

**Q #2) When do you think QA activities should start?**

**QA activity should start from the beginning of the project**. The more early it starts the more benefit it is to set the standards for achieving the quality.

The cost, time and efforts are very challenging in case the QA activities gets delayed.

**Q #3) What is the**[**difference between Test Plan and Test Strategy**](https://www.softwaretestinghelp.com/difference-between-test-plan-test-strategy-test-case-test-script-test-scenario-and-test-condition/)

**Test Strategy is** at a higher level, mostly **created by the Project Manager** which **demonstrates** the overall **approach of the testing for the entire project**, whereas **Test plan** basically depicts the **how the testing should be performed for a particular application**, falling under a project.

**Q #4) Can you explain the software testing life cycle?**

[Software Testing Life Cycle](https://www.softwaretestinghelp.com/what-is-software-testing-life-cycle-stlc/) refers to a testing process which has specific steps to be executed in a definite sequence to ensure that the quality goals have been met.

Please refer to the below link to know more:

**Q #5) How do you define a**[**format of writing a good test case**](https://www.softwaretestinghelp.com/how-to-write-effective-test-cases-test-cases-procedures-and-definitions/)**?**

A test case has the below format:

* **Test case ID,**
* **Test case description**
* Severity
* **Priority**
* **Environment**
* **Build version**
* **Steps to execute**
* **Expected results**
* **Actual results**

**Q #6) What is a good test case?**

In simple words**, a good test case is one which finds a defect.** But all test case will not find defects, so A good test case can also be one which has all the prescribed details and coverage.

**Q #7) What would you do if you have a large suit to execute in a very less time?**

**In case we have less time, and have to execute larger volume of test cases, we should prioritize the test case** at first instant and execute the high priority test cases first and then move on to the lower priority ones.

**Q #8) Do you think QA’s can also participate to resolve production issues?**

Definitely!! It would be a good learning curve for QA’s to participate in resolving production issues. Many a time production issues could be resolved by clearing the logs or making some registry settings or by restarting the services.

These kind of environmental issues could be very well fixed by the QA team.

Also If QAs have an insight on resolving the production issues, they may also include them while writing the test cases, and this way they can contribute to improve quality and try to minimize the production defects.

**Q #9) Suppose you find a bug in production, how would you make sure that the same bug is not introduced again?**

**Best way is to immediately write a test case for the production defect and include it in the regression suite. This way we ensure that the bug does not get introduced again.**

Also many a time we can also think of alternate test cases or similar kind of test case and include them in our planned execution.

**Q #10) What is the difference between functional and nonfunctional testing?**

**Functional testing** basically deals with the functional aspect of the application. **This technique tests that the system is behaving as per the requirement and specification.**

These are directly linked with customer requirement. We validate the test cases against the specified requirement and make the test pass or failed accordingly.

Examples include **regression, integration**, system, smoke etc…

[Nonfunctional testing](https://www.softwaretestinghelp.com/what-is-non-functional-testing/) – on the other hand tests the Nonfunctional aspect of the application. It tests NOT the requirement, **but the environmental factors like performance, load and stress**.

These are not explicitly specified in the requirement but are prescribed in the quality standards. So as QA we have to make sure that these testing are also given sufficient time and priority.

**Q #11) What is negative testing? How is it different from positive testing?**

Negative testing is a technique which validates that the system behaves gracefully in case of any invalid inputs.

**For example, in case user enters any invalid data in a text box, system should display a proper message instead of technical message which the user does not understands.**

[Negative testing](https://www.softwaretestinghelp.com/what-is-negative-testing/) is different from positive testing in a way **that positive testing validates that our system works as expected and compares the test results with the expected results.**

**Q #12) How would you ensure that your testing is complete and has good coverage?**

**Requirement traceability matrix and Test coverage matrices will help us to determine that our test cases have good coverage.**

Requirement traceability matrices will help us to determine that the test conditions are enough so that all the requirements are covered.

Coverage matrices will help us to determine that the test cases are enough to satisfy all the identified test conditions in RTM.

**Q #13) What are the different artifacts you refer when you write the test cases?**

The main artifacts used are:

* **Functional requirement specification**
* **Requirement understanding document**
* **Use Cases**
* Wireframes
* **User Stories**
* **Acceptance criteria**
* Many a time UAT test cases

**Q #14) Have you ever managed writing the test cases without having any documents?**

Yes, many a time we have a situation where we have to write test cases without having any concrete documents. In that case, best way is to

* Collaborate with the BA and development team.
* Dig into mails which have some information.
* Dig into older test cases / regression suite
* If the feature is new, try to read the wiki pages or help of the application to have an idea
* Sit with the developer and try to understand the changes being made.
* Based on your understanding, identify the test condition and send it to BA or stakeholders to review them.

**Q #15) What is meant by**[**Verification and Validation**](https://www.softwaretestinghelp.com/what-is-verification-and-validation/)**?**

**Validation is the process of evaluating the final product to check whether the software meets the business needs.**

The test execution which we do in our day to day life are actually the validation activity which includes smoke testing, functional testing, regression testing, systems testing etc…

**Verification is a process of evaluating the intermediary work products of a software development lifecycle to check if we are in the right track of creating the final product.**

**Q #16) What are the different verification techniques you know?**

Verification techniques are static in nature. There are 3 verification techniques:

Review, Inspection and walkthrough.

**1) Review – Is a method by which the code / test cases are examined by the individual other than the author who has produced it.**

**2) Inspection**– **Is a technical way to examine and correct the defects in the test artifact or code**

**3) Walkthrough – Is a process in which the author of the code reads the content and gets the feedback.**

**Q #17) What is the difference between**[**Load and Stress testing**](https://www.softwaretestinghelp.com/what-is-performance-testing-load-testing-stress-testing/)**?**

**Stress Testing is a technique which validates the behavior of the system when it executes under stress.** To explain, we reduce the resources and check the behavior of the system.

We first understand the upper limit of the system and gradually reduce the resources and check the system behavior.

**In Load testing we validate the system behavior under the expected load. The load can be of concurrent user or resources accessing the system at the same time.**

**Q #18) In case you have any doubts regarding your project, how do you approach?**

In case of any doubts, first try to get it clear by reading the available artifacts / application help. In case of doubts still persisting, ask immediate supervisor or the senior member of your team.

BA’s would also be a good choice to ask the doubts. We can also touch base with the development team in case of any doubts. The last option would be to follow up with the manager and finally to the stakeholders.

**Q #19) Have you used any Automation tools?**

*The answer to this question is very much exclusive to individual. Reply with all the tools and strategy of automation that you have used in your project.*

**Q #20) How do you determine which piece of software require how much testing?**

We can know this factor by finding out the [Cyclomatic Complexity](https://www.softwaretestinghelp.com/cyclomatic-complexity/).

The technique helps to identify the below 3 questions for the programs / features

* Is the feature / program testable?
* Is the feature/ program understood by every one?
* Is the feature / program reliable enough?

As a QA we can use this technique to identify the “level” of our testing.

It is a practice that if the result of cyclomatic complexity is more or a bigger number, we consider that piece of functionality to be of complex nature and hence we conclude as a tester; that the piece of code / functionality requires an in-depth testing.

On the other hand if the result of the Cyclomatic Complexity is a smaller number, we conclude as QA that the functionality is of less complexity and decide the scope accordingly.

As a QA its very important that we understand the entire testing lifecycle and should be able to suggest changes in our process if required.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* THE END\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*