Data Warehouse – Mock Client Interview Q&A (Freshers)

Coverage: Architecture · Modeling · ETL/ELT · Cloud DWH · SQL · Real-World Scenarios

SECTION 1: Data Warehouse Architecture

Q1. What is a Data Warehouse and why do organizations need it? (Client Scenario)

Scenario:
A retail company has sales data in Oracle, customer data in CRM, and product data in Excel. Reports are slow and inconsistent.

Answer:
A **Data Warehouse (DWH)** is a centralized system designed for **analytical reporting and decision-making**.
Organizations use it to:

- Integrate data from multiple source systems
- Maintain historical data
- Enable fast, consistent BI reporting

Real-world usage:
Management dashboards, trend analysis, forecasting.

Q2. Explain the basic Data Warehouse architecture layers.

Answer:
A typical DWH has **three layers**:

1. **Staging Layer**
    - Raw data loaded from sources
    - Minimal or no transformations
2. **Integration Layer**
    - Cleaned, transformed, business-ready data

    - Fact and dimension tables exist here
3. **Access Layer**
    - Used by BI tools (Power BI, Tableau, Qlik)
    - Optimized for reporting and analytics

Client relevance:
Ensures clean separation between raw data and reporting data.

Q3. EDW vs Data Mart – When would you use each?

Scenario:
A global company wants company-wide reporting, while the finance team wants quick finance-only analytics.

Answer:

EDW	**Data Mart**
Enterprise-wide	Department-specific
Large scope	Smaller scope
Single source of truth	Faster delivery

Usage:

- EDW for executive dashboards
- Data Mart for Finance, HR, Sales teams

Q4. ETL vs ELT – What is the difference in real projects?

Answer:

ETL	**ELT**
Transform before loading	Transform after loading
Used in on-prem systems	Used in cloud DWH
Tool-heavy	SQL-driven

Real-world example:

- **ETL:** Informatica → Oracle DWH
- **ELT:** Fivetran → Snowflake → SQL transformations

Q5. Real-time vs Batch Data Warehousing

Scenario:
A stock trading platform vs monthly finance reporting.

Answer:

- **Batch:**
  - Scheduled loads (daily, hourly)
  - Used for finance, compliance
- **Real-time:**
  - Near-real-time ingestion
  - Used for fraud detection, monitoring

SECTION 2: Data Modeling

Q6. What is dimensional modeling and why is it preferred?

Answer:
Dimensional modeling organizes data into:

- **Fact tables** (measures)
- **Dimension tables** (context)

It is preferred because:

- Easy to understand
- Faster query performance
- BI-friendly

Q7. Explain Star Schema vs Snowflake Schema with use cases.

Answer:

Star Schema	**Snowflake Schema**
Denormalized dimensions	Normalized dimensions
Simple queries	Complex joins
Better performance	Storage efficient

Client usage:

- Star → dashboards
- Snowflake → complex enterprise models

Q8. What is a Fact table and what does it contain?

Answer:
A Fact table contains:

- **Measures:** Sales, Quantity, Revenue
- **Foreign Keys:** CustomerID, ProductID, DateID

Example:

```
Fact_Sales(Sales_Amount, Quantity, Customer_Key, Product_Key)
```

Q9. What are Dimension tables?

Answer:
Dimension tables store **descriptive attributes**.

Examples:

- Customer Dimension: Name, City, Segment
- Product Dimension: Category, Brand

Q10. What are Slowly Changing Dimensions (SCD)?

Scenario:
Customer changes address.

Answer:

Type	Behavior
SCD 1	Overwrite old data
SCD 2	Maintain history (new row)
SCD 3	Limited history (new column)

Most used: SCD Type 2

Q11. What is Granularity and why is it important?

Answer:
Granularity defines the **lowest level of detail**.

Example:

- Order-level vs Daily sales

Impact:
Wrong granularity leads to incorrect analysis.

Q12. What are Surrogate Keys and why are they needed?

Answer:
Surrogate keys are **system-generated keys**.

Why needed:

- Handle SCD Type 2
- Avoid dependency on business keys
- Improve joins

SECTION 3: ETL / ELT Processes

Q13. Explain a typical ETL pipeline.

Answer:

1. Extract from sources
2. Transform (clean, join, validate)
3. Load into DWH

Q14. How do you handle missing or duplicate data?

Answer:

- Missing data → default values, null handling
- Duplicates → DISTINCT, business rules

Q15. What is error handling and logging in ETL?

Answer:
Tracks:

- Failed records
- Load timestamps
- Error messages

Why important:
Auditing and debugging.

SECTION 4: Cloud Data Warehousing

Q16. Why are companies moving to Cloud DWH?

Answer:

- Scalability
- Pay-as-you-use
- Minimal infrastructure management

Q17. Compare Snowflake, BigQuery, Redshift, Databricks.

Answer:

Platform	Strength
Snowflake	Simplicity, performance
BigQuery	Serverless analytics
Redshift	AWS ecosystem
Databricks	Big data + ML

Q18. What is scalability in cloud DWH?

Answer:
Ability to scale compute and storage independently based on workload.

SECTION 5: SQL & Multi-Database Concepts

Q19. Why is SQL important in DWH projects?

Answer:
SQL is used for:

- Transformations
- Aggregations
- Data validation
- Reporting views

Q20. How do fact and dimension tables work together in SQL?

Answer:
Joined using surrogate keys to create analytical datasets.

SECTION 6: Real-World Project Scenarios

Q21. Explain an end-to-end retail DWH project.

Answer:

1. Sources: POS, CRM, Excel
2. Load to staging
3. Transform into star schema
4. Build BI dashboards

Q22. How does BI connect to a Data Warehouse?

Answer:
BI tools connect to:

- Access layer
- Optimized views or marts

Q23. Common challenges in DWH projects?

Answer:

- Data quality issues
- Changing business rules
- Performance tuning
- Late arriving data

INTERVIEW TIP FOR FRESHERS

When answering:

- Start with **business context**
- Explain **technical concept**
- End with **real-world usage**

Additional Mock Client Interview Questions & Answers (Freshers)

SECTION 7: Architecture – Real Implementation Scenarios

Q51. Why do companies use a staging layer instead of loading directly into fact tables?

Answer:
The staging layer:

- Isolates raw data from business logic
- Helps reprocess data without touching production tables
- Supports audit and reconciliation

Real-world usage:
Used when source systems resend corrected data.

Q52. What happens if staging is skipped in a DWH project?

Answer:

- Difficult debugging
- No rollback mechanism
- Risk of corrupt analytical data

Q53. Can a data warehouse have multiple staging layers?

Answer:
Yes.
Large enterprises use:

- **Raw staging**
- **Cleansed staging**

This improves traceability and compliance.

Q54. How do you decide the refresh frequency of a warehouse?

Answer:
Based on:

- Business requirement
- Source system capability
- Cost

Example:
Sales → hourly
Finance → daily

Q55. What is an Operational Data Store (ODS)?

Answer:
ODS sits between source systems and DWH:

- Near real-time
- Limited history
- Used for operational reporting

SECTION 8: Data Modeling – Practical Client Questions

Q56. Why should dimensions not contain measures?

Answer:
Dimensions describe context; measures belong in fact tables.
Mixing them breaks aggregation logic.

Q57. What happens if fact tables are over-normalized?

Answer:

- Complex joins
- Poor query performance
- Difficult BI usage

Q58. What is a role-playing dimension?

Answer:
Same dimension used multiple times.

Example:
Date → Order Date, Ship Date, Delivery Date

Q59. How do you handle multiple currencies in fact tables?

Answer:

- Store transaction currency
- Store converted amount
- Maintain exchange rate dimension

Q60. Why is Date dimension mandatory in most DWH projects?

Answer:
Because time-based analysis is core to analytics:

- YTD, MTD, YoY
- Trends and forecasting

SECTION 9: SCD – Real Client Use Cases

Q61. Why is SCD Type 2 preferred in analytics projects?

Answer:
It preserves historical changes, enabling:

- Customer behavior tracking
- Compliance and audits

Q62. What is the disadvantage of SCD Type 2?

Answer:

- Increased storage
- More complex queries

Q63. How do you identify the current active record in SCD2?

Answer:
Using:

- `Is_Current_Flag`
- `End_Date IS NULL`

Q64. Can a dimension have multiple SCD types?

Answer:
Yes.
Example:

- Address → Type 2
- Phone Number → Type 1

Q65. How do you handle late-arriving dimensions?

Answer:

- Create dummy records
- Update dimension later
- Reprocess facts if required

SECTION 10: ETL / ELT – Execution & Debugging

Q66. What is incremental loading and why is it used?

Answer:
Loads only new or changed records.
Used to:

- Reduce load time
- Minimize system impact

Q67. How do you detect changed records for incremental load?

Answer:

- Timestamps
- CDC (Change Data Capture)
- Hash comparison

Q68. What is full load and when is it used?

Answer:
Reloads entire dataset.
Used:

- Initial loads
- Small reference tables

Q69. What is idempotency in ETL pipelines?

Answer:
Running the same job multiple times produces the same result.

Q70. How do you recover from ETL job failure?

Answer:

- Restart from last successful checkpoint
- Reload failed partitions only

SECTION 11: Data Quality & Governance

Q71. What is data profiling and why is it important?

Answer:
Analyzes data structure, patterns, and anomalies before loading.

Q72. How do you handle null values in analytics?

Answer:

- Business default values
- Separate “Unknown” dimension records

Q73. What is data lineage?

Answer:
Tracks data flow from source to report.

Client value:
Trust and auditability.

Q74. What is metadata management?

Answer:
Managing:

- Table definitions
- Column meaning
- Data ownership

Q75. Why is data validation critical before BI consumption?

Answer:
Incorrect data leads to:

- Wrong decisions
- Loss of business trust

SECTION 12: Cloud DWH – Practical Scenarios

Q76. What does “separation of compute and storage” mean?

Answer:
Compute and storage scale independently.

Example:
Snowflake virtual warehouses.

Q77. How does cloud DWH reduce infrastructure cost?

Answer:

- Pay-per-use
- Auto-scaling
- No server maintenance

Q78. What is auto-suspend and auto-resume?

Answer:
Compute shuts down when idle and resumes automatically.

Q79. What is data sharing in cloud DWH?

Answer:
Sharing data without copying it between accounts.

Q80. How do cloud warehouses handle concurrency?

Answer:
By scaling compute resources dynamically.

SECTION 13: SQL & Performance Tuning

Q81. Why should large fact tables be partitioned or clustered?

Answer:
Improves query performance and reduces scan cost.

Q82. What is a surrogate key join advantage over natural keys?

Answer:
Faster joins and stable relationships.

Q83. What is query pruning?

Answer:
Skipping irrelevant data blocks during query execution.

Q84. Why avoid SELECT * in DWH queries?

Answer:

- Poor performance
- Higher cost
- Unnecessary data scan

Q85. How do aggregates improve BI performance?

Answer:
Pre-calculated summaries reduce query execution time.

SECTION 14: BI & Business Consumption

Q86. Why should BI tools connect to curated layers only?

Answer:
Avoids exposing raw or inconsistent data.

Q87. What happens if BI directly connects to staging tables?

Answer:

- Inconsistent reports
- Performance issues
- Business confusion

Q88. How do dimensions improve dashboard usability?

Answer:
They enable slicing, filtering, and drill-downs.

Q89. What is a semantic layer?

Answer:
A business-friendly abstraction over raw data.

Q90. Why do executives prefer dashboards over raw reports?

Answer:

- Quick insights
- Visual trends
- Decision-ready information

SECTION 15: End-to-End Project & Client Interaction

Q91. What questions should you ask a client before building a DWH?

Answer:

- Business KPIs
- Data sources
- Refresh frequency
- Security needs

Q92. How do you handle changing business rules?

Answer:

- Versioned transformations
- Historical tracking
- Clear documentation

Q93. What is a proof of concept (POC) in DWH?

Answer:
Small-scale implementation to validate approach and tools.

Q94. How do you validate DWH data with business users?

Answer:

- Reconciliation reports
- Parallel run with legacy reports

Q95. What defines success for a data warehouse project?

Answer:

- Trusted data
- Performance
- Business adoption
