QUESTION 1 (40 Marks)

1.1 Data Warehouse Design for Ordering System (25 Marks)

Logical Assumptions:

- All transactions are recorded daily
- Each order can have multiple items
- Customers can make multiple orders
- Products have standard pricing
- Need to track vendor performance

Star Schema Design:

Fact Table: Fact_Sales

- Sales_SK (Primary Key)
- Date_SK (FK)
- Customer_SK (FK)
- Product_SK (FK)
- Vendor_SK (FK)
- Order_Number
- Quantity (Measure)
- Unit_Price (Measure)
- Total_Amount (Measure)
- Tax_Amount (Measure)
- Discount_Amount (Measure)

Dimension Tables:

Dim_Date

- Date_SK (PK)
- Date
- Day
- Month
- Quarter
- Year
- Day_of_Week
- Is_Weekend
- Is_Holiday
- Fiscal_Period

Dim_Customer

- Customer_SK (PK)
- Customer_ID
- Customer_Name
- Address
- City
- State
- Country
- Customer_Type

Dim_Product

- Product_SK (PK)
- Product_ID
- Product_Name
- Category
- Sub_Category
- Brand
- Vendor_ID

Dim_Vendor

- Vendor_SK (PK)

- Vendor_ID
- Vendor_Name
- Contact_Person
- Address
- Rating

Hierarchies:

1. Time Hierarchy: Year → Quarter → Month → Day
2. Product Hierarchy: Category → Sub_Category → Product

1.2 Why Computed Columns are Better Suited (5 Marks)

Computed columns are better in analytical systems because:

1. **Performance Optimization**: Pre-calculated values reduce query processing time
2. **Consistency**: Ensures uniform calculations across all queries
3. **Reduced Complexity**: Simplifies report writing and ad-hoc queries
4. **Storage Trade-off**: While they increase storage, the read performance gain is worth it in DW
5. **Example**: Total_Amount = Quantity × Unit_Price - Discount_Amount (pre-calculated during ETL)

1.3 Usage of Surrogate Keys (5 Marks)

Surrogate keys are used for:

1. **Independence from Source Systems**: Protects DW from changes in operational systems
2. **Historical Tracking**: Enables tracking of slowly changing dimensions
3. **Performance**: Integer keys are faster for joins than natural keys
4. **Integration**: Allows merging data from multiple sources with different key formats
5. **Data Quality**: Handles missing or duplicate natural keys

1.4 Why De-normalized Structures are Preferred (5 Marks)

De-normalization is preferred because:

1. **Query Performance**: Fewer joins mean faster query execution
2. **Simplicity**: Easier for business users to understand and query
3. **Aggregation Efficiency**: Pre-joined data speeds up analytical queries
4. **Read-Optimized**: DW is optimized for reading, not writing
5. **Predictable Performance**: Query performance is more consistent

QUESTION 2 (15 Marks)

2.1 What "Data is New Oil" Means (2 Marks)

This statement means:

- Data is a valuable resource that drives modern economy
- Like oil, data needs to be refined (processed) to be useful
- It's a strategic asset for competitive advantage

2.2 Important Challenges in "V's of Data" (4 Marks)

Example: E-commerce Platform

1. **Volume**: Millions of transactions daily requiring massive storage
2. **Velocity**: Real-time inventory updates and order processing
3. **Variety**: Structured (orders), semi-structured (logs), unstructured (reviews)
4. **Veracity**: Ensuring data accuracy from multiple channels

2.3 Why Veracity is Important (3 Marks)

Veracity is crucial because:

1. **Decision Quality**: Poor data leads to poor decisions
2. **Trust**: Stakeholders lose confidence in inaccurate reports
3. **Compliance**: Regulatory requirements demand accurate data

2.4 Importance of Teams in Big Data Projects (3 Marks)

Teams are essential for:

1. **Diverse Skills**: Combining technical, business, and analytical expertise
2. **Scalability**: Large projects need collaborative effort
3. **Knowledge Sharing**: Cross-functional understanding improves outcomes

2.5 Important Factors in Big Data Strategy (3 Marks)

1. **Infrastructure**: Scalable storage and processing capabilities
2. **Data Governance**: Policies for quality, security, and privacy

3. **Skills Gap**: Training and hiring appropriate talent
4. **Integration**: Connecting disparate data sources
5. **ROI Measurement**: Clear business value metrics

QUESTION 3 (15 Marks)

3.1 Why Web Content Mining is Challenging (4 Marks)

Challenges compared to Big Data Vs:

1. **Unstructured Nature**: Web content lacks consistent format
2. **Dynamic Content**: Pages change frequently
3. **Noise**: Advertisements, navigation elements interfere
4. **Scale**: Billions of pages to process

3.2 Why Tokenization is Important (3 Marks)

Tokenization is crucial for:

1. **Text Processing**: Breaks text into analyzable units
2. **Feature Extraction**: Creates input for machine learning
3. **Language Understanding**: Identifies meaningful elements

3.3 Classification Techniques in Text Mining (3 Marks)

Techniques include:

1. **Naive Bayes**: For spam detection
2. **SVM**: For sentiment analysis
3. **Decision Trees**: For topic categorization

3.4 Use in Recommender Systems (3 Marks)

- **Transactions**: User purchase history
- **Customers**: User profiles and preferences
- **Products**: Item features and categories
- Combined to create collaborative and content-based recommendations

3.5 Practical Applications (2 Marks)

1. **Sentiment Analysis**: Brand monitoring
2. **Customer Service**: Automated ticket classification
3. **Content Categorization**: News article classification
4. **Fraud Detection**: Analyzing communication patterns

QUESTION 4 (16 Marks)

4.1 Difference Between Predictive and Prescriptive Analytics (3 Marks)

Predictive Analytics: Forecasts what will happen

- Example: Predicting customer churn probability

Prescriptive Analytics: Recommends actions to take

- Example: Suggesting retention strategies for high-risk customers

4.2 Default Borrower Analysis (6 Marks)

From the data:

- Default rate: 10% (1 out of 10)
- Pattern: Lower income correlates with default
- Married status shows mixed results

Prediction for new customer:

- Based on married status and 120K income
- Similar to row 4 (married, 120K, no default)
- Likely prediction: No default

4.3 Using Predictive Analytics for Spam (3 Marks)

1. **Feature Extraction**: Keywords, sender patterns, frequency
2. **Training Model**: Use labeled spam/ham emails
3. **Classification**: Apply model to incoming emails
4. **Continuous Learning**: Update model with new patterns

4.4 Confusion Matrix Advantages (3 Marks)

1. **Detailed Performance**: Shows true/false positives and negatives
2. **Multiple Metrics**: Enables calculation of precision, recall, F1-score
3. **Class Imbalance**: Reveals performance on minority classes

4. **Error Analysis**: Identifies specific misclassification patterns

QUESTION 5 (15 Marks)

5.1 Need for Special Date Dimension (4 Marks)

Examples:

1. **Retail**: Analyze holiday vs. regular day sales
2. **Banking**: Month-end vs. mid-month transactions
3. **Manufacturing**: Weekday vs. weekend production
4. **Seasonality**: Identify quarterly patterns

5.2 Multi-lingual Date Dimension Design (4 Marks)

Include columns:

- Month_Name_English
- Month_Name_Local
- Day_Name_English
- Day_Name_Local
- Holiday_Name_Multi
- Use locale codes for systematic organization

5.3 Date Hierarchies Examples (4 Marks)

1. **Calendar**: Year → Quarter → Month → Week → Day
2. **Fiscal**: Fiscal_Year → Fiscal_Quarter → Fiscal_Month
3. **Academic**: Academic_Year → Semester → Month
4. **Retail**: Season → Month → Week

5.4 Role-Playing Dimension (3 Marks)

A role-playing dimension is when the same dimension is used multiple times in a fact table with different meanings.

Example: Date dimension used as:

- Order_Date
- Ship_Date
- Payment_Date
- Return_Date

QUESTION 6 (15 Marks)

6.1 Usage of Separate Date Dimensions (3 Marks)

Separate date dimensions are used when:

- Different calendar systems (fiscal vs. calendar)
- Different granularities (daily vs. hourly)
- Specific business requirements

6.2 Diagnostic vs. Descriptive Analytics (3 Marks)

Descriptive: What happened?

- Example: Last month's sales were $1M

Diagnostic: Why did it happen?

- Example: Sales increased due to promotional campaign

6.3 Time Series Analysis Challenges (3 Marks)

1. **Seasonality**: Identifying cyclic patterns
2. **Missing Values**: Handling gaps in data
3. **Trend Detection**: Separating trend from noise
4. **External Factors**: Accounting for holidays, events

6.4 Association Rule Implementation Areas (3 Marks)

1. **Retail**: Market basket analysis
2. **Healthcare**: Treatment pattern discovery
3. **Web Analytics**: Clickstream analysis
4. **Fraud Detection**: Unusual transaction patterns

6.5 SCD in Data Analytics Design (3 Marks)

SCD (Slowly Changing Dimensions) handles changes in dimension attributes over time:

- Type 1: Overwrite (no history) Type 2: Add new row (full history)Type 3: Add columns (limited history) Type 4: Mini-dimensions Type 6: Hybrid approach This ensures historical accuracy in analytical reports.

DIGITAL SIGNATURE

Document Information

Document: paper 21.pdf

Company: Tech Solutions Lanka (Pvt) Ltd

Signature Details

Signed by: David

Print Name: David

Email: david.anderson@gmail.com

Date: 2025-07-17

IP Address: 127.0.0.1

Timestamp: 2025-07-17 00:53:14 UTC