CSCI316 Big Data Mining - Introduction Question Bank

Assessment Format Guidelines

- 3-hour duration, 50 marks total
- 6 questions with multiple sub-parts
- Mark allocation: 5-12 marks per question
- Mix of theoretical explanations and practical coding tasks

Question 1: Big Data Fundamentals (8 marks)

Part A (3 marks)

Define Big Data and explain why traditional databases and tools are inadequate for processing Big Data. Provide three specific examples of data sources that generate Big Data.

Part B (3 marks)

List and briefly explain the four alternative terms commonly used to refer to Big Data. Discuss why these terms are sometimes used interchangeably.

Part C (2 marks)

A company collects 2TB of sensor data daily from their manufacturing equipment. Explain whether this constitutes Big Data, justifying your answer with reference to the fundamental characteristics of Big Data.

Question 2: The Four Vs of Big Data (10 marks)

Part A (2 marks each = 8 marks)

For each of the four Vs (Volume, Variety, Velocity, Veracity), provide:

- A clear definition
- Two real-world examples demonstrating this characteristic
- One challenge this characteristic presents for data processing

Part B (2 marks)

Calculate the following based on the 2016 statistics provided:

- Total data generated by Facebook, Google searches, and SMS combined per hour
- Show your detailed calculations and express the final answer in appropriate units

---

Question 3: The Fifth V - Value (7 marks)

Part A (4 marks)

Explain the concept of "Value" as the fifth V of Big Data. Discuss the four types of value mentioned in the materials:

- Academic value
- Statistical value
- Business value
- Correlational value

Provide one example for each type.

Part B (3 marks)

Analyze the statement: "Big data only makes sense when there is value associated with it." Explain how this differentiates Value from the other four Vs, and discuss why some organizations might struggle to extract value from their Big Data initiatives.

Question 4: Data Collection Methods (9 marks)

Part A (3 marks)

Compare and contrast the three main approaches to data collection in Big Data:

- Direct download datasets
- API-based collection
- Web scraping

Include advantages and limitations of each approach.

Part B (4 marks)

Coding Task: Write a Python function that simulates data collection from multiple sources. The function should:

```python
def collect_big_data_stats():
    # Your implementation here
    pass
```

- Calculate total data generated per minute across all platforms mentioned (YouTube, Facebook, Google, SMS, Instagram)

- Return a dictionary with individual and total statistics
- Handle the calculation showing step-by-step process

Part C (2 marks)

Explain why certain data sources (Amazon reviews, Google Scholar) are not available via APIs and require web scraping. Discuss the ethical considerations involved in web scraping.

Question 5: Data Scientist Roles and Skills (8 marks)

Part A (4 marks)

Describe the four main responsibilities of a Data Scientist:

- Data Collection
- Data Preprocessing
- Data Visualization
- Data Analytics and Application

For each responsibility, provide a specific example of tasks involved and tools that might be used.

Part B (2 marks)

Differentiate between a Data Scientist and a Data Engineer. Create a table showing their distinct responsibilities, required skills, and typical deliverables.

Part C (2 marks)

Analyze the five key skills needed to be a Data Scientist. Rank them in order of importance for a Big Data environment and justify your ranking with specific examples.

Question 6: Big Data Implementation Challenges (8 marks)

Part A (3 marks)

Scenario: A social media company processes 100 million posts per day, each containing text, images, and metadata. The data includes multiple languages, various media formats, and real-time user interactions.

Identify which of the four Vs this scenario primarily demonstrates and explain the specific challenges each V presents for this company.

Part B (3 marks)

Problem-Solving Exercise: Given the following data growth statistics:

- 2016: 44 billion GB per day
- 2025 prediction: 463 billion GB per day

Calculate:

1. The compound annual growth rate (CAGR) of data generation
2. Expected daily data generation for 2020 (midpoint)
3. Show all calculation steps clearly

Part C (2 marks)

Discuss the "junk in = junk out" principle in the context of Big Data. Explain how this principle affects each stage of the data science workflow and provide strategies to mitigate this issue.

Additional Practice Questions

Short Answer Questions (5 marks each)

Q7: Explain how the concept of data "Velocity" applies differently to streaming social media data versus satellite imagery data. Provide specific examples and discuss the technological requirements for each.

Q8: A healthcare organization wants to analyze patient records, real-time monitoring data, and medical imaging. Categorize each data type according to the "Variety" dimension and explain the integration challenges.

Q9: Design a data collection strategy for analyzing customer satisfaction across multiple platforms (surveys, social media, reviews, support tickets). Address how you would handle each of the four Vs.

Coding Challenges (7-10 marks each)

Q10: Implement a Python class `BigDataAnalyzer` that:

- Stores data generation statistics for different platforms
- Calculates growth rates and projections
- Identifies which platforms contribute most to each of the four Vs
- Includes error handling and data validation

Q11: Create a data preprocessing pipeline that addresses Veracity issues:

- Function to detect and handle missing values
- Methods to identify and remove duplicate entries
- Data quality scoring mechanism
- Documentation of each step's impact on data integrity

Case Study Questions (12 marks each)

Q12: **Netflix Data Analysis Case** Netflix processes viewing data from 200+ million subscribers globally, including:

- Real-time streaming behavior
- Content metadata (genres, ratings, cast)
- User profiles and preferences
- Geographic and temporal patterns

Analyze this scenario by:
a) Identifying how each of the four Vs manifests in Netflix's data (4 marks)
b) Designing a data collection architecture addressing scalability and real-time requirements (4 marks)
c) Proposing value extraction strategies for content recommendation and business intelligence (4 marks)

Q13: **Smart City Traffic Management** A city implements IoT sensors across 10,000 intersections, collecting:

- Vehicle counts and speeds (every 30 seconds)
- Weather and environmental data
- Traffic light status and timing
- Emergency vehicle locations
- Public transport schedules and delays

Design a comprehensive Big Data solution: a) Categorize the data types and their characteristics using the 4 Vs framework (3 marks) b) Propose data collection, storage, and processing architecture (4 marks)
c) Identify potential value outcomes and their societal impact (3 marks) d) Address data quality and privacy concerns (2 marks)

Assessment Rubric Guidelines

Excellent (90-100%)

- Comprehensive understanding of all Big Data concepts
- Clear articulation of theoretical principles with practical examples
- Accurate calculations with detailed step-by-step solutions
- Well-structured code with proper documentation
- Critical analysis and evaluation of scenarios

Good (75-89%)

- Solid grasp of most Big Data concepts
- Generally accurate explanations with some examples
- Mostly correct calculations with minor errors
- Functional code with basic documentation
- Some analytical thinking demonstrated

Satisfactory (60-74%)

- Basic understanding of core concepts
- Simple explanations with limited examples
- Calculation attempts with some errors
- Code that works but lacks optimization
- Limited critical analysis

Needs Improvement (Below 60%)

- Incomplete understanding of fundamental concepts
- Vague or incorrect explanations
- Significant calculation errors
- Non-functional or missing code
- Lack of analytical depth

Study Tips for Students

1. **Master the 4 Vs**: Ensure you can explain each V with multiple examples and understand their interconnections
2. **Practice Calculations**: Work through data volume calculations and growth rate problems
3. **Code Implementation**: Practice writing functions for data processing and analysis scenarios
4. **Real-world Applications**: Study current Big Data applications in various industries
5. **Integration Skills**: Understand how theoretical concepts apply to practical implementations