# Assignment 1&2 Soluation


## Name:Majed Hameed Alodhaylah

## ID:432105522


**Q1: What is Big Data? Why is it required?**
Big Data refers to extremely large and complex datasets that traditional processing tools cannot handle efficiently. It is required to analyze patterns, trends, and insights for better decision-making and innovation.

**Q2: Describe the critical success factors in Analytical Process model requirements.**
The critical success factors include data quality, clear objectives, appropriate analytical tools, skilled personnel, and continuous evaluation to ensure accurate and useful insights.

**Q3: Describe the steps involved in the Analytics Process Model with a diagram.**
The steps in the Analytics Process Model include:

1. **Data Collection** – Gathering data from various sources.
2. **Data Processing** – Cleaning and transforming raw data.
3. **Data Analysis** – Applying statistical and machine learning techniques.
4. **Interpretation** – Extracting insights and patterns.
5. **Decision Making** – Using insights to make informed decisions.
   *(Diagram to be inserted in the document)*

**Q4: Give an example where Big Data analytics plays a crucial role in fields like marketing, medicine, or public health.**
In medicine, Big Data analytics helps in predicting disease outbreaks by analyzing patient records, environmental data, and social media trends. This allows healthcare providers to take preventive actions.

**Q5: Explain what are the sources of Big Data.**
The main sources of Big Data include social media, IoT devices, transaction records, sensors, search engines, and business applications.

**Q6: What are the different kinds of 6V's that form the pillars of Big Data? Explain the characteristics of each.**

1. **Volume** – Large amounts of data generated daily.
2. **Velocity** – Speed at which data is created and processed.
3. **Variety** – Different formats of data (structured, unstructured).
4. **Veracity** – Quality and reliability of data.
5. **Value** – Insights extracted from data.
6. **Variability** – Inconsistencies in data over time.

**Q7: Describe the major issues or challenges in handling this huge data that is generated today.**
Challenges include data storage, processing speed, security risks, data integration, privacy concerns, and ensuring data quality.

**Q8: What are the major challenges of mining a huge amount of data in comparison with mining a small dataset?**
Mining large datasets requires high computational power, efficient algorithms, better storage, and handling of missing/incomplete data, unlike small datasets, which are easier to process.

**Q9: Can Big Data and Data Warehouse co-exist together? Give your view briefly.**
Yes, Big Data and Data Warehouses can coexist, where the Data Warehouse is used for structured historical data, and Big Data solutions handle unstructured, real-time data for advanced analytics.

**Q10: Give examples of various technologies that are popularly used in Industry to handle Big Data.**
Popular technologies include Hadoop, Spark, NoSQL databases (MongoDB, Cassandra), Apache Kafka, and cloud platforms (AWS, Google Cloud, Azure).

**Q11: What are the different types of data sets that can be used for Big Data?**
Data sets include structured, semi-structured, and unstructured data, such as relational databases, XML/JSON files, and multimedia content.

**Q12: List the important properties of attribute values.**
Properties include distinctiveness, order, interval, and meaningful zero values.

**Q13: What is an attribute? List its different types with an example of each.**
An attribute is a characteristic of data. Types include:

1. **Nominal** – Categories (e.g., Gender: Male/Female).
2. **Ordinal** – Ordered categories (e.g., Education Level: High School, Bachelor's).
3. **Interval** – Numeric without a true zero (e.g., Temperature in Celsius).
4. **Ratio** – Numeric with a true zero (e.g., Age, Weight).

**Q14: Differentiate between discrete and continuous attributes.**

- **Discrete** – Countable values (e.g., Number of students in a class).
- **Continuous** – Measurable values (e.g., Height, Temperature).

**Q15: List the important characteristics of data.**
Characteristics include accuracy, completeness, consistency, timeliness, and accessibility.

**Q16: What are the problems with poor data quality?**
Issues include incorrect analysis, poor decision-making, increased costs, and inefficiency in operations.

**Q17: Explain the different methods of data preprocessing.**
Methods include data cleaning, integration, transformation, and reduction to improve data quality for analysis.

**Q18: What is sampling? Explain the types of sampling.**
Sampling is selecting a subset of data for analysis. Types include:

1. **Random Sampling** – Each data point has an equal chance.
2. **Stratified Sampling** – Dividing data into groups before sampling.
3. **Systematic Sampling** – Selecting every nth item from a dataset.
4. **Cluster Sampling** – Selecting entire groups randomly.