**UNIT 3: Big Data and Data Analytics – Answers by Topic**

**1. Sources of Big Data**

**Questions:**

* What are the sources of Big Data?
* Explain any three sources of Big Data.

**Answer:**  
Big Data originates from various sources:

1. **Social Media**: Facebook, Twitter, LinkedIn generate huge volumes of unstructured data.
2. **Machine/Sensor Data**: IoT devices, traffic sensors, and industrial machines generate continuous data.
3. **Transaction Data**: E-commerce, banking, and retail generate structured data.
4. **Web Logs**: Server logs provide user interaction data.
5. **Public Data**: Government, research, and weather datasets.

Sources of Big Data:

+-----------------+ +-------------+ +-------------+ +------------+ +-----------+

| Social Media | | Sensors/IoT | | Transactions| | Web Logs | | Public Data|

+-----------------+ +-------------+ +-------------+ +------------+ +-----------+

**2. Data Analytics Lifecycle**

**Questions:**

* Draw the diagram of data analytics life cycle in big data and briefly explain its phases.
* Explain Data Analytics Cycle with suitable diagram and its phases.

**Answer:**  
Phases of the Data Analytics Lifecycle:

1. **Discovery** – Identify business problem and data sources.
2. **Data Preparation** – Clean, transform, format the data.
3. **Model Planning** – Select techniques and tools.
4. **Model Building** – Develop and test models.
5. **Evaluate** – Validate model accuracy and business fit.
6. **Deployment** – Integrate model into decision systems.
7. **Operationalize** – Deliver final reports or applications.

Data Analytics Lifecycle:

Discovery → Data Preparation → Model Planning → Model Building → Evaluate → Deployment → Operationalize

**3. Model Building Phase**

**Questions:**

* Explain model building phase with example.
* What is Model Building? Elaborate this phase with a suitable example.
* Explain in detail how the model building phase is built by team in data analytics life cycle?
* List out the activities to be carried out in model planning and model building phase.

**Answer:**  
In this phase:

* Algorithms are selected based on goals (e.g., classification or regression).
* Data is split into training and testing sets.
* Tools like Python (Scikit-learn), R, or RapidMiner are used.
* Models are trained, validated, and iteratively improved.  
  **Example:** Predicting house prices using Linear Regression.

Model Building Phase:

Raw Data

↓

Split into Training Set & Testing Set

↓

Training Set → Train Model → Model

↓

Testing Set → Validate Model

↺ (Feedback loop to tune model)

**4. Data Discovery Phase**

**Questions:**

* What is data discovery phase? Explain with example.
* List and Explain the various activities involved in identifying potential data resources.

**Answer:**  
Data Discovery involves:

* Understanding the business problem.
* Identifying internal and external data sources.
* Assessing data availability and quality.  
  **Example:** In healthcare, identifying patient data and hospital records.

Data Discovery Phase:

Business Problem

↓

Identify Data Sources

↓

Assess Data Quality & Relevance

↓

Access Data

**5. Data Preparation Phase**

**Questions:**

* List and explain the steps in data preparation phase.
* What is the data Preparation phase? What is Analytics Sandbox and ETLT?

**Answer:**  
Steps:

1. Data cleaning (missing value treatment)
2. Data transformation (normalization, encoding)
3. Feature engineering
4. Data formatting  
   **Analytics Sandbox**: Safe environment for analysis  
   **ETLT**: Extract, Transform, Load, Transform – enables refining after loading into sandbox.

Data Preparation:

Raw Data → Data Cleaning → Data Transformation → Feature Engineering → Prepared Data

(Analytics Sandbox for experimentation and testing transformations)

**6. Big Data Analytics Architecture**

**Question:**

* Explain big data analytics architecture with diagram.

**Answer:**  
Typical architecture:

1. **Data Sources** – Sensors, logs, databases
2. **Data Ingestion** – Kafka, Flume
3. **Storage** – HDFS, NoSQL
4. **Processing** – Hadoop, Spark
5. **Analytics** – ML models, SQL engines
6. **Visualization** – Tableau, Power BI

Big Data Analytics Architecture:

Data Sources → Data Ingestion → Storage Layer → Processing Layer → Analytics Layer → Visualization Layer

Examples:

Kafka HDFS Spark ML Models Tableau

**7. Business Intelligence vs Data Science**

**Questions:**

* What is data science? Differentiate BI and Data Science.
* Differentiate BI versus Data science.

**Answer:**  
**Data Science**: Uses advanced analytics, ML, and statistics to gain insights.  
**BI**: Focuses on reporting and dashboards using historical data.  
**Key Differences:**

* BI is descriptive; DS is predictive/prescriptive.
* BI uses structured data; DS uses all types.

| **Feature** | **Business Intelligence** | **Data Science** |
| --- | --- | --- |
| Data Type | Structured | Structured & Unstructured |
| Focus | Reporting, Dashboards | Predictive & Prescriptive |
| Techniques | SQL, OLAP | Machine Learning, Stats |
| Timeframe | Past & Present | Future Predictions |

**8. Driving Data Deluge**

**Question:**

* What is driving data deluge? Explain with one example.

**Answer:**  
The data deluge is driven by:

* Explosion of IoT, social media, sensors, mobile devices  
  **Example:** YouTube uploads 500+ hours/minute of video globally.

Sources of Data Deluge:

IoT Devices

↓

Social Media Platforms

↓

Mobile Apps

↓

Sensors

↓

Big Data Explosion (Cloud)

**9. Big Data Characteristics**

**Question:**

* What are the three characteristics of Big Data?
* What are the main considerations in processing Big Data?

**Answer:**  
**3 V's:**

1. Volume – Large data size
2. Velocity – Speed of generation
3. Variety – Structured & unstructured data  
   **Considerations:** Scalability, storage, fault tolerance, real-time processing

Volume

●

/ \

/ \

Variety Velocity

**10. Types of Analytics**

**Question:**

* Explain Descriptive, Diagnostic, Predictive analytics.

**Answer:**

1. **Descriptive**: Summarizes past data (e.g., sales report)
2. **Diagnostic**: Analyzes reasons behind results (e.g., why sales dropped)
3. **Predictive**: Uses ML to forecast future outcomes (e.g., customer churn prediction)

| **Analytics Type** | **Purpose** | **Example** |
| --- | --- | --- |
| Descriptive | What happened? | Sales Reports |
| Diagnostic | Why it happened? | Root Cause Analysis |
| Predictive | What will happen? | Customer Churn Prediction |

**11. Tools & Model Selection**

**Questions:**

* Write short notes on: i) Common tools for model building ii) Model selection for data analytics
* List out tools used for these phases.

**Answer:**  
**Tools**: Python, R, Weka, KNIME, SAS, Spark MLlib  
**Model Selection**: Depends on data type, size, accuracy required, and interpretability

Common Tools:

Python | R | RapidMiner | KNIME | SAS | Spark MLlib

**12. ETL and ETLT**

**Question:**

* Write short note on ETL.

**Answer:**  
**ETL**: Extract data from sources → Transform (clean, process) → Load into storage  
**ETLT**: Extra transform step after loading into data warehouse for flexible processing

ETL Process:

Extract → Transform → Load

ETLT Process:

Extract → Transform → Load → Transform

**13. Stakeholders & Roles**

**Questions:**

* List out stakeholders and what they expect at conclusion.
* List and explain key roles for successful analytics project.

**Answer:**  
**Stakeholders:**

* Business Sponsor: ROI and insights
* Data Scientist: Model accuracy
* Analyst: Reporting
* Engineer: Scalable deployment  
  **Expectations:** Working models, dashboards, clear recommendations

Stakeholder Hierarchy:

Business Sponsor

|

Data Scientist — Data Engineer — Analyst

|

Domain Experts

Expectations: ROI, Model Accuracy, Reports, Deployment

**14. Linear Regression**

**Question:**

* What is linear regression, and its objectives?
* Difference between simple and multiple linear regression?
* How to evaluate performance?

**Answer:**  
**Linear Regression**: Predicts dependent variable using independent variables

* **Simple**: One predictor
* **Multiple**: More than one predictor  
  **Evaluation**: R² (coefficient of determination), MSE, RMSE

Linear Regression:

Scatter Plot of Data Points

● ● ● ● ● ●

\ /

\ /

\\_\_\_\_/ Regression Line (y = mx + c)

Simple Linear Regression: One Predictor Variable

Multiple Linear Regression: Multiple Predictor Variables

Evaluation Metrics: R², RMSE

**1. Sources of Big Data**

| **Notes** | **Keywords** |
| --- | --- |
| Big Data originates from multiple sources such as social media platforms, IoT devices, transactional data, web logs, and public data. These provide structured and unstructured data streams. | Social Media, IoT, Transactions, Web Logs, Public Data, Structured, Unstructured, Data Streams, Sensors |

**2. Data Analytics Lifecycle**

| **Notes** | **Keywords** |
| --- | --- |
| The lifecycle includes Discovery, Data Preparation, Model Planning, Model Building, Evaluation, Deployment, and Operationalization. It guides how data is converted into actionable insights. | Discovery, Data Preparation, Model Planning, Model Building, Evaluation, Deployment, Operationalize, Insight |

**3. Model Building Phase**

| **Notes** | **Keywords** |
| --- | --- |
| Model building involves splitting data into training/testing sets, training models, validating them, and tuning based on results to improve accuracy. | Model Training, Training Set, Testing Set, Validation, Tuning, Accuracy, Feedback Loop, Prediction |

**4. Data Discovery Phase**

| **Notes** | **Keywords** |
| --- | --- |
| Focuses on understanding business problems, identifying relevant data sources, assessing data quality, and gaining access for analysis. | Business Problem, Data Sources, Quality Assessment, Data Relevance, Data Access, Exploration, Requirements |

**5. Data Preparation Phase**

| **Notes** | **Keywords** |
| --- | --- |
| Raw data is cleaned, transformed, and features are engineered. The Analytics Sandbox is a safe space for testing transformations before final use. | Data Cleaning, Transformation, Feature Engineering, Analytics Sandbox, Data Formatting, Preprocessing, Data Quality |

**6. Big Data Analytics Architecture**

| **Notes** | **Keywords** |
| --- | --- |
| Architecture layers: Data Sources → Ingestion → Storage → Processing → Analytics → Visualization. Tools like Kafka, HDFS, Spark, and Tableau are commonly used. | Architecture, Data Ingestion, Storage, Processing, Analytics, Visualization, Kafka, HDFS, Spark, Tableau |

**7. Business Intelligence vs Data Science**

| **Notes** | **Keywords** |
| --- | --- |
| BI focuses on structured data for reporting; Data Science deals with structured/unstructured data using predictive analytics and machine learning. | Business Intelligence, Data Science, Structured Data, Unstructured Data, Reporting, Predictive Analytics, Machine Learning |

**8. Driving Data Deluge**

| **Notes** | **Keywords** |
| --- | --- |
| Massive data growth driven by IoT, social media, mobile apps, and sensors creating exponential data generation. | Data Deluge, IoT, Social Media, Mobile Apps, Sensors, Data Explosion, Volume, Velocity |

**9. Big Data Characteristics (3 Vs)**

| **Notes** | **Keywords** |
| --- | --- |
| Volume (amount), Velocity (speed), Variety (types) define Big Data's unique challenges and opportunities. | Volume, Velocity, Variety, Big Data, Data Characteristics, Data Complexity |

**10. Types of Analytics**

| **Notes** | **Keywords** |
| --- | --- |
| Descriptive (what happened), Diagnostic (why happened), Predictive (what will happen) analytics serve different business needs. | Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Business Needs, Data Analysis |

**11. Tools & Model Selection**

| **Notes** | **Keywords** |
| --- | --- |
| Popular model building tools include Python, R, RapidMiner, KNIME, SAS, and Spark MLlib. Model selection depends on problem type and data. | Python, R, RapidMiner, KNIME, SAS, Spark MLlib, Model Selection, Data Analytics Tools |

**12. ETL / ETLT Process**

| **Notes** | **Keywords** |
| --- | --- |
| ETL: Extract, Transform, Load; ETLT adds a second transformation after loading to refine data further. | ETL, ETLT, Extract, Transform, Load, Data Pipeline, Data Processing |

**13. Stakeholders & Roles**

| **Notes** | **Keywords** |
| --- | --- |
| Key roles: Business Sponsor, Data Scientist, Data Engineer, Analyst, Domain Experts. Their expectations include ROI, model accuracy, and actionable insights. | Business Sponsor, Data Scientist, Data Engineer, Analyst, Domain Experts, ROI, Model Accuracy, Expectations |

**14. Linear Regression**

| **Notes** | **Keywords** |
| --- | --- |
| Linear regression models relationship between dependent and independent variables. Simple uses one predictor; multiple uses several. Evaluated using R² and RMSE. | Linear Regression, Simple, Multiple, Predictor Variables, Dependent Variable, R², RMSE, Model Evaluation |

**Absolutely! Here are common keywords and phrases that you can sprinkle into *any* Unit 3 answer to boost clarity, show your understanding, and gain extra marks:**

**Common Keywords for Unit 3 (Big Data & Data Analytics)**

* **Big Data**
* **Data Analytics Lifecycle**
* **Data Sources**
* **Data Preparation**
* **Model Building**
* **Data Discovery**
* **Data Quality**
* **Data Transformation**
* **Data Ingestion**
* **Storage Layer**
* **Processing Layer**
* **Business Intelligence (BI)**
* **Data Science**
* **3 Vs (Volume, Velocity, Variety)**
* **Descriptive / Diagnostic / Predictive Analytics**
* **ETL / ETLT Process**
* **Stakeholders**
* **Tools (Python, R, Spark, Tableau)**
* **Linear Regression**
* **Model Evaluation (R², RMSE)**
* **Analytics Sandbox**
* **Machine Learning**
* **Deployment**
* **Data Visualization**
* **Data Deluge**
* **Feedback Loop**
* **Feature Engineering**

**How to use these keywords:**

* **Start answers by defining Big Data and the importance of the Data Analytics Lifecycle.**
* **Mention data sources and the necessity of data preparation and data transformation.**
* **Highlight how model building fits in the lifecycle and uses tools like Python and Spark.**
* **Explain phases with terms like data ingestion, processing, and storage layers.**
* **When comparing concepts, use Business Intelligence vs Data Science keywords.**
* **Always talk about the 3 Vs to emphasize the scale and complexity of Big Data.**
* **Reference different types of analytics: descriptive, diagnostic, and predictive.**
* **Include the role of ETL/ETLT in data pipelines and mention stakeholders to show project awareness.**
* **If relevant, talk about linear regression and evaluation metrics like R² and RMSE.**
* **Wrap up answers by mentioning deployment and the importance of the feedback loop for continuous improvement.**

Sure! Here are **9 versatile points** you can include in *any* Unit 3 answer to ensure you cover important concepts and score well:

1. **Big Data originates from diverse sources** such as social media, IoT devices, transactional systems, and web logs, generating massive volumes of structured and unstructured data.
2. The **Data Analytics Lifecycle** includes phases like Discovery, Data Preparation, Model Planning, Model Building, Evaluation, and Deployment, ensuring systematic handling of data.
3. The **Data Discovery phase** involves identifying relevant data sources, assessing data quality, and understanding the business problem to prepare for analysis.
4. During **Data Preparation**, raw data is cleaned, transformed, and features are engineered to make it suitable for modeling and analysis.
5. **Model Building** is a crucial phase where algorithms are trained on data, validated, and tuned to create predictive or descriptive models.
6. The \*\*3 Vs of Big Data—Volume, Velocity, and Variety—\*\*define its complexity and guide the selection of appropriate tools and methods.
7. **ETL (Extract, Transform, Load)** and **ETLT** processes ensure efficient data ingestion and transformation to support analytics workflows.
8. Different types of analytics—**Descriptive, Diagnostic, and Predictive**—serve distinct purposes in interpreting past events, diagnosing causes, and forecasting future trends.
9. Successful analytics projects depend on **collaboration among stakeholders** such as business sponsors, data scientists, engineers, and domain experts to ensure actionable insights and ROI.

PREDICTED QUE AND FOR THIS YEAR FROM UNIT 3 Below are **detailed answers with diagrams and explanations** for all the **previous-year questions** related to the **predicted 2025 topics**:

**🔮 TOPIC 1: Data Analytics Lifecycle**

**📘 2022 Q2b:**

**Q:** *Explain big data analytics architecture with diagram. What is data discovery phase? Explain with example.*

**✅ Answer:**

**📊 Big Data Analytics Architecture Diagr** +-----------------------+

| Data Sources |

+-----------------------+

↓

+----------------+

| Data Ingestion |

+----------------+

↓

+------------------------+

| Distributed Storage | ← (e.g., HDFS)

+------------------------+

↓

+------------------------+

| Data Processing Layer | ← (e.g., Spark, MapReduce)

+------------------------+

↓

+----------------------+

| Analytics Layer | ← (ML, statistical models)

+----------------------+

↓

+----------------------+

| Visualization Layer | ← (Tableau, PowerBI)

+----------------------+

**🔍 Discovery Phase:**

The **Discovery Phase** is the first step in the Data Analytics Lifecycle.  
It involves:

* Understanding the problem
* Identifying business goals
* Determining relevant data sources
* Assessing tools, stakeholders, and constraints

**🧠 Example:**

For a bank predicting credit card fraud, the team may gather transaction logs, customer behavior, location data, etc.

**📘 2023 (P1) Q1a & 2023 (P3) Q1a:**

**Q:** *Draw the diagram of data analytics life cycle in big data and briefly explain its phases.*

**✅ Answer:**

**🔄 Data Analytics Lifecycle Diagram:**

+-----------+ +-----------------+ +--------------+

| Discovery | ---> | Data Preparation | --> | Model Planning |

+-----------+ +-----------------+ +--------------+

↓ ↓ ↓

+--------------+ +-----------------+ +-----------------+

| Model Building| <--- | Evaluation | --> | Deployment |

+--------------+ +-----------------+ +-----------------+

**📚 Phases:**

1. **Discovery**: Define objectives and success criteria
2. **Data Preparation**: Clean and organize data
3. **Model Planning**: Select modeling approach
4. **Model Building**: Train and validate models
5. **Evaluation**: Interpret results
6. **Deployment**: Implement solution for end use

**🔮 TOPIC 2: Discovery Phase / Data Identification**

**📘 2023 (P3) Q1b:**

**Q:** *List and explain the various activities involved in identifying potential data resources in discovery phase.*

**✅ Answer:**

**🧭 Key Activities in Discovery Phase:**

1. **Business Understanding**
   * Define business problem & analytics objective
2. **Stakeholder Interviews**
   * Understand goals, success metrics
3. **Initial Hypothesis Formation**
   * What might influence the outcome?
4. **Identify Data Sources**
   * Structured (DBs), semi-structured (XML), unstructured (text, video)
5. **Assess Data Availability**
   * Is the required data accessible, complete, and legal to use?
6. **Tools and Infrastructure**
   * Evaluate software, compute, and storage needed

**🔮 TOPIC 3: Model Planning & Model Building**

**📘 2022 Q2a:**

**Q:** *What are the sources of Big Data? Explain model building phase with example.*

**✅ Answer:**

**🌍 Sources of Big Data:**

1. **Social Media** (Facebook, Twitter)
2. **Sensor Data / IoT** (temperature, motion)
3. **Web Logs / Clickstreams**
4. **Financial Transactions**
5. **Surveillance Video / Audio Feeds**

**🔨 Model Building Phase:**

Steps:

1. Select model type (e.g., classification)
2. Split data into train/test sets
3. Train model
4. Tune hyperparameters
5. Evaluate with metrics (accuracy, recall)

**🧠 Example:**

In email spam detection, Naïve Bayes or Logistic Regression models are trained using labeled email datasets.

**📘 2023 (P1) Q1b:**

**Q:** *Explain in detail how the model building phase is built by the team in data analytics lifecycle?*

**✅ Answer:**

**🧑‍💻 Team Activities:**

* **Data Scientists**: Write training pipelines
* **Engineers**: Manage data infrastructure
* **Domain Experts**: Interpret features and results
* **ML Engineers**: Build deployable models

**🔁 Cycle:**

Train → Test → Evaluate → Tune → Repeat

**📘 2023 (P2) Q1a:**

**Q:** *What is Model Building? Elaborate with example.*

**✅ Answer:**

**Model building** is creating mathematical/ML models that can generalize from data.

**🧠 Example:**

A telecom company wants to predict customer churn. Features like usage patterns, complaints, and call history are used to train a classification model.

**📘 2023 (P3) Q2b & 2024 Q2a:**

**Q:**  
*Write short notes on:*

* i) Common Tools for Model Building
* ii) Model Selection for Data Analytics  
  *AND*  
  *List the activities for model planning & model building.*

**✅ Answer:**

**🛠️ Common Tools:**

* **Python**: Scikit-learn, TensorFlow, Keras
* **R**: caret, randomForest
* **SAS, RapidMiner, Weka**

**🧩 Model Selection Factors:**

* Type of problem: classification/regression/clustering
* Data size & quality
* Interpretability needs
* Speed vs. accuracy tradeoff

**🔮 TOPIC 4: Regression (Linear & Logistic)**

**📘 2022 Q4a:**

**Q:** *Explain the following:*  
i) Linear Regression  
ii) Logistic Regression

**✅ Answer:**

**🔷 i) Linear Regression:**

y=β0+β1x+εy = \beta\_0 + \beta\_1x + \varepsilon

* Predicts continuous values
* E.g., House price based on area

**📊 Diagram:**

Price

|

| .

| . .

| . .

|\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Area

**🔷 ii) Logistic Regression:**

σ(x)=11+e−x\sigma(x) = \frac{1}{1 + e^{-x}}

* Predicts probability (binary classification)
* Output between 0 and 1
* Threshold (e.g., 0.5) to classify

**📘 2024 Q2b:**

**Q:** *What is linear regression and its objectives? How does it differ from multiple linear regression? How to evaluate performance?*

**✅ Answer:**

**🔹 Linear vs Multiple Linear Regression:**

| **Feature** | **Linear** | **Multiple Linear** |
| --- | --- | --- |
| Inputs | 1 variable | Multiple variables |
| Equation | y=b0+b1xy = b\_0 + b\_1x | y=b0+b1x1+b2x2+...y = b\_0 + b\_1x\_1 + b\_2x\_2 + ... |
| Complexity | Low | Higher |

**📏 Evaluation Metrics:**

* Mean Squared Error (MSE)
* R-squared (R²)
* Root Mean Squared Error (RMSE)

**📘 2023 (P3) Q4a:**

**Q:** *Explain the need of logistic regression and its types.*

**✅ Answer:**

**🧪 Need:**

* When output is categorical (Yes/No, 0/1)
* Linear regression fails in classification tasks

**🔄 Types of Logistic Regression:**

* **Binary** (Yes/No)
* **Multinomial** (3+ categories, e.g., low/medium/high)
* **Ordinal** (ordered categories)

**📘 2023 (P2) Q4a:**

**Q:** *Explain the logistic function in logistic regression.*

**✅ Answer:**

**🔁 Sigmoid Function:**

σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}

Where z = β0 + β1x1 + β2x2 + ...

* Output is between 0 and 1
* Used to classify into class 1 or 0

**📉 Graph:**

| \*

| \* \*

| \* \*

| \* \*

|\_\*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\*\_\_\_\_

-6 0 +6

Would you like this full answer set in a downloadable **PDF format** with neat diagrams and headers?