**Big Data Essentials – Key Terms & Definitions**

**1. Big Data Analytics**

* **Big Data**: Large, complex datasets that traditional systems cannot handle efficiently.
* **Big Data Analytics**: The process of examining big data to uncover patterns, correlations, and trends for decision-making.

**2. Big Data Controllers**

* **Big Data Controllers**: Systems or individuals responsible for managing and processing big data.
* **Data Governance**: Policies and procedures to ensure data quality, privacy, and security.

**3. Big Data Problems**

* **Data Volume**: Massive amounts of data generated daily (e.g., social media, IoT).
* **Data Variety**: Different types of data (structured, unstructured, semi-structured).
* **Data Velocity**: The speed at which data is generated and processed.
* **Data Veracity**: The reliability and accuracy of big data.

**4. Big Data Challenges**

* **Storage Issues**: Managing large datasets efficiently.
* **Processing Power**: High computational resources required for big data analytics.
* **Security & Privacy**: Protecting sensitive data from unauthorized access.
* **Scalability**: Ensuring systems can handle growing data sizes.

**5. Big Data Solutions**

* **Cloud Computing**: Storing and processing data over the internet instead of local computers.
* **Distributed Systems**: Spreading data across multiple servers for efficient processing (e.g., Hadoop).
* **Machine Learning**: Using AI models to analyze big data and predict trends.

**6. Big Data Classification**

* **Classification**: Categorizing big data into meaningful groups for analysis.
* **Supervised Learning**: Using labeled data to train machine learning models.
* **Unsupervised Learning**: Finding patterns in data without predefined labels.

**7. Representation Learning**

* **Feature Extraction**: Automatically identifying key data attributes for better classification.
* **Dimensionality Reduction**: Reducing large datasets while keeping essential features (e.g., PCA).

**8. Distributed File Systems**

* **Hadoop Distributed File System (HDFS)**: A system designed for storing and processing big data across multiple machines.
* **MapReduce**: A programming model for processing large datasets using parallel computing.

**9. Classification Modeling**

* **Decision Trees**: A tree-like model used for making predictions based on data attributes.
* **Random Forest**: Multiple decision trees combined to improve prediction accuracy.
* **Deep Learning**: Neural networks with multiple layers that can learn complex patterns.

**10. Big Data Scalability**

* **High-Dimensional Systems**: Data with many features (e.g., image recognition, genetics).
* **Low-Dimensional Structures**: Data with fewer features but still meaningful for analysis.