Big data essentials

Big data, cloud computing, data warehousing

1. **Fundamentals of Big Data:**
   - Data storage, batch processing, introduction to distributed computation
2. **Big Data in Decision Analytics and Scalability:**
   - Importance of data in decisions, frameworks for analytics and AI, resource elasticity
3. **Big Data vs Data Science:**
   - Differences in focus between Big Data and Data Science
4. **Technology Evolution and Job Roles:**
   - Distributed computing revolution, roles like Data Engineers, Data Scientists, and DevOps
5. **Governance in Big Data:**
   - Security, confidentiality, and legal ethics
6. **Roles in Big Data Environment:**
   - Responsibilities of Data Engineers, Data Scientists, ML Engineers, Data Analysts (BI), and DevOps
7. **The 5 V's of Big Data:**
   - Variety, Volume, Value, Velocity, Veracity
8. **Challenges in Big Data:**
   - Handling IoT evolution, big data volume, velocity, and variety .
9. **Cloud Computing and Big Data Architecture:**
   - Data Lake vs Data Warehouse, scalable storage solutions, security, cost-efficiency
10. **Batch Processing and ETL vs ELT:**
    - Definition, use-cases, and benefits of batch processing, differences between ETL and ELT
11. **Distributed Computing Principles:**
    - Parallel processing, fault tolerance, scalability (scale in vs scale out), MapReduce, resource managemen
12. **Spark Framework:**
    - Advantages over Hadoop/HDFS, architecture, features like in-memory analytics, fault tolerance, integration
13. **Data Formats and Technologies:**
    - Columnar storage, parquet, ACID transactions in data management
14. **Databricks Platform:**
    - Unified analytics platform, features, and basic operations
15. **Glossary and Essential Concepts:**
    - Definitions of key terms like BI, Data Lake, Data Warehouse, Batch vs Real-time processing, Distributed computing, Node/Worker, MapReduce

Big data essentials

Data streaming

1. **Data Streaming Introduction:**

   Real-time processing, continuous data flow from multiple sources, timeliness, request/response vs event-driven models.
2. **Batch vs Streaming:**
   - Differences in processing methods, continuous vs periodic data handling
3. **Characteristics and Challenges of Data Streams:**
   - Size, velocity, volume, variety, scalability, and veracity.
4. **Use Cases in Various Industries:**
   - Applications in finance, health, media, retail
5. **Data Streaming Methods:**
   - RESTful API, event-driven approaches, pros and cons.
6. **Apache Kafka Overview:**
   - Introduction, architecture, key features like scalability, fault tolerance, real-time processing

Big data essentials

Data mining

1. **Introduction to Data Mining:**
    - Defining the problem, examples of data mining applications
2. **Data Sources and Preprocessing:**
    - Importance of quantitative variables, data cleaning, normalization, transformation, missing value imputation, category coding, dimensionality reduction.
3. **Pandas crash course**
4. **Data Analysis Techniques:**
    - Supervised machine learning, classification problems, generalization, overfitting vs underfitting.
5. **Descriptive Analysis Applications:**
    - Factors influencing price in different contexts like investment, mortgage lending, insurance, urban planning
6. **Handling Categorical Variables:**
    - Techniques like ordinal mapping, dummy coding, and their impact on model compatibility and interpretability
7. **Understanding Correlation:**
    - Overview, coefficient values, positive vs negative correlation, correlation vs causation