**AWS Data Engineering and Machine Learning Essentials**

**AWS SageMaker**

* **Purpose**: Develop, train, and deploy machine learning models in Jupyter notebooks on AWS.
* **Key Tools**:
  1. **SageMaker Studio**: All-in-one environment for model management.
  2. **SageMaker Autopilot**: Automated ML (AutoML) tool for model building.
  3. **SageMaker Experiments**: Tracks and organizes model versions.
  4. **SageMaker Debugger**: Identifies issues during training.
  5. **SageMaker Model Monitor**: Detects data drift and monitors model quality.

**Data Preparation**

**Data Gathering**

* **Sources**: Public repositories, web scraping.

**Data Cleaning (Handling Missing Values)**

* **Methods**:
  + **Do Nothing**: Ignore missing observations.
  + **Removal**: Delete records with missing data.
  + **Imputation**: Fill with Mode (categorical), Median/Mean (numerical), or Model-based (e.g., KNN, regression).
  + **Forward/Backward Fill**: Use surrounding values.
  + **Interpolation**: Estimate using other observations.

**Feature Engineering**

* **Feature Extraction**: Reducing the number of features (e.g., using PCA).
* **Encoding Categorical Data**:
  + **Binary Encoding**: For true/false or binary data.
  + **Label Encoding**: Categorical labels with ordinality.
  + **One-Hot Encoding**: Convert categorical data to binary columns.
* **Numeric Transformation**:
  + **Normalization**: Scale to [0,1].
  + **Standardization**: Center around mean 0, SD 1.
  + **Binning**: Group continuous data into categories.
* **Text Feature Engineering**:
  + **Bag of Words, N-Grams**: Tokenizes text for word frequency analysis.
  + **TF-IDF**: Weights words by importance.

**AWS Data Migration and ETL Tools**

* **Amazon Data Pipeline**: Automates regular data movement.
* **AWS Database Migration Service (DMS)**: Migrates databases (e.g., MySQL to DynamoDB).
* **AWS Glue**: Fully managed ETL service for cleaning and transforming data.
* **Amazon Athena**: SQL querying on S3 data.

**Data Streaming with AWS Kinesis**

* **Kinesis Data Streams**: Streams data from IoT, social media; uses shards.
* **Kinesis Data Firehose**: Transmits data to S3, Redshift, Elasticsearch (uses Lambda).
* **Kinesis Video Streams**: Processes video data (webcams, security cameras).
* **Kinesis Data Analytics**: SQL-based data processing from Kinesis Streams or Firehose.

**Big Data Processing and Storage**

* **Amazon EMR (Elastic MapReduce)**: Managed cluster for big data processing (supports Hadoop, Spark).
* **Amazon Managed Service for Apache Flink**: Java/Scala/SQL for real-time data analysis.
* **Amazon Redshift**: Cloud-based data warehouse for petabyte-scale analytics.
* **AWS Data Exchange**: Marketplace for third-party data.
* **Amazon QuickSight**: Data visualization service.

**Machine Learning Fundamentals**

**Modeling Types**

1. **Supervised Learning**: Predict target using labeled data (e.g., classification, regression).
   * **Examples**: Linear Regression, Logistic Regression, Decision Trees, SVM, Neural Networks.
2. **Unsupervised Learning**: Group data without labeled outcomes (e.g., clustering).
   * **Examples**: K-Means, PCA.
3. **Reinforcement Learning**: Decision-making based on environment feedback.
   * **Examples**: Q-Learning, Monte Carlo.

**Hyperparameters**

* **Definition**: Parameters set before training (e.g., learning rate).
* **Tuning Methods**:
  1. **Random Search**: Randomly tests parameter combinations.
  2. **Bayesian Search**: Optimizes parameters based on regression.

**Built-In Algorithms**

* **Regression**: Linear Regressor.
* **Clustering**: K-means.
* **Classification**: XGBoost, Linear Learner, K-Nearest Neighbors.
* **Image & Text Analysis**: ResNet, Word2Vec, Sequence-to-Sequence.
* **Anomaly Detection and Forecasting**: Specialized algorithms for trends.

**AWS Security for Machine Learning**

1. **IAM Policies**: Restrict access to SageMaker resources.
2. **VPC Configuration**: Isolate SageMaker within a VPC.
3. **Encryption**: Enable S3/EBS encryption.
4. **Access Control**: Use Security Groups and Network ACLs.
5. **Monitoring**: Use CloudWatch and CloudTrail for activity logs.
6. **Lifecycle Configurations**: Automate security setups at notebook startup.
7. **Model Monitor**: Tracks data drift and detects anomalies.

**Additional AWS Services for AI/ML**

* **Amazon Mechanical Turk**: Crowdsourcing human tasks that computers struggle with.
* **AWS IoT Greengrass**: Runs local data processing on IoT devices.
* **Amazon Connect**: Contact center service with AI-driven analytics and omnichannel support.