Data Flow for Analytics and Machine Learning

**Sources**
- **Structured**
  - Databases
  - Data generated by processes
  - APIs
- **Unstructured**
  - Text documents
  - Social media feeds
  - Voice recordings
  - Images

**Pipelining**
- **Batch Load**
  - Data is ingested from various sources and transformed into raw data
  - Ingests data from different data sources
- **Streaming**
  - Data is processed in real-time from multiple sources
  - Data is continuously generated and ingested into a Data Lake

**Data Lake**
- **Prepared Data**
  - Data is transformed and organized for analysis
- **Enriched Data**
  - Data is enriched with identifiers from third-party sources

**Data Science and AI**
- **Model Development**
  - Jupyter notebooks for data exploration
- **Model Training**
  - Trains models using various algorithms

**Business Intelligence**
- **Data Warehouses**
  - Stores data for long-term retention
- **Data Lake Access**
  - Data is made accessible for analysis

**Managed Services**
- **Managed Endpoint**
  - Provides a scalable and managed environment for deployment

**Intelligent Devices**
- **IoT Fleet Manager**
  - Manages device fleet and collects data

**Glue**
- AWS Glue is a serverless data integration service for data warehousing and data lakes

**Rekognition**
- AWS Rekognition is a service that enables developers to add image and video analysis capabilities to their applications

**SageMaker**
- AWS SageMaker is a fully managed service for building, training, and hosting machine learning and deep learning models

**Transform**
- Data is transformed for analysis

**Transformed Data**
- Data is transformed for analysis

**Model Training**
- Models are trained using various algorithms

**Deploy**
- Models are deployed for use in production

**Managed Infrastructure**
- Data is managed and secured

**Third-party Services**
- Third-party services used for data enrichment

**Real-time Analytics**
- Data is analyzed in real-time

**Cloud-scale Infrastrucure**
- Data is processed on cloud-scale infrastructure

**Data Enrichment**
- Data is enriched with identifiers from third-party sources

**Enrichment**
- Data is enriched with identifiers from third-party sources

**Data Preparation**
- Data is prepared for analysis

**Data Lakes**
- Data Lakes store large amounts of data that can be accessed and analyzed

**Data Warehouse**
- Data Warehouse provide a central repository for data

**Data Science**
- Data Science is the application of statistical and probabilistic techniques to derive insights

**Machine Learning**
- Machine Learning is the application of statistical and probabilistic techniques to derive insights

**IoT**
- Internet of Things devices

**Real-time Streaming**
- Data is processed in real-time

**Semantic Search**
- Data is searched for meaning

**Big Data**
- Big Data is high-volume data that cannot be handled by traditional systems

**Unstructured Data**
- Data that does not conform to a predefined schema

**Structured Data**
- Data that conforms to a predefined schema

**Semistructured Data**
- Data that does not conform to a predefined schema

**Enrichment**
- Data is enriched with identifiers from third-party sources

**Managed Services**
- Managed Services provide a scalable and managed environment for deployment

**Intelligent Devices**
- Intelligent Devices are devices that can process data

**Data Science at Enterprise Scale**
- Data Science is leveraged across all aspects of the business