Data Science for Managerial Decisions (MB 511)

Building Machine Learning Solutions using Cloud

Instructor
Anant Prakash Awasthi

EMBA Program
MB-511

Introduction to Cloud and It’s Types

Cloud services refer to a wide range of services delivered over the internet that provide on-demand access to computing resources, such as storage, databases, servers, and software applications, without the need for direct management of physical infrastructure.

These services are typically offered by cloud providers (like AWS, Microsoft Azure, and Google Cloud) and allow businesses and individuals to access powerful computing capabilities with flexibility and scalability. Cloud services are designed to support diverse needs, from data storage and processing to machine learning and artificial intelligence, with payment models based on usage rather than upfront investment.

Cloud Services

Object Storage

Stores large volumes of unstructured data, like images, videos, logs, and datasets. Commonly used by data scientists to store raw data and model outputs. Object storage is highly scalable, making it ideal for storing datasets that grow over time.

- Examples: AWS S3, Azure Blob Storage, Google Cloud Storage.

Data Warehouse

Central repository optimized for storing structured, historical data, often used for analytics and business intelligence. Data scientists use data warehouses to run SQL queries on large datasets to extract insights, analyze trends, and prepare data for modeling.

- Examples: Amazon Redshift, Azure Synapse Analytics, Google BigQuery.

Cloud Services

Machine Learning Platform

Provides tools and environments for building, training, and deploying machine learning models. These platforms support end-to-end workflows, including data preprocessing, model training, hyperparameter tuning, and deployment, which are essential for data scientists working with machine learning.

- Examples: AWS SageMaker, Azure Machine Learning, Google AI Platform.

Cloud Machine Learning

Managed Apache Spark

Managed Spark environments for big data processing. Apache Spark is widely used by data scientists to handle large-scale data processing tasks, especially for data preparation, feature engineering, and iterative model training on big data.

- Examples: AWS EMR, Azure Databricks, Google Dataproc.

Dataproc

Cloud Services

Big Data Processing (Hadoop)

Primarily used for batch processing large datasets, often in combination with Hadoop Distributed File System (HDFS) and MapReduce. Data scientists use these tools for large-scale data cleaning, transformation, and distributed data processing.

- Examples: AWS EMR, Azure HDInsight, Google Dataproc.

Dataproc

Data Lake

Stores raw and semi-structured data at any scale. Data lakes allow data scientists to store massive amounts of data, including structured and unstructured types, in one central repository, making it easy to perform analysis and training on diverse datasets.

- Examples: AWS Lake Formation, Azure Data Lake Storage, Google Cloud Storage.

Cloud Storage

Cloud Services

Data Integration / ETL

Tools for Extracting, Transforming, and Loading (ETL) data from various sources into a data warehouse or lake. ETL is essential for data scientists as it helps prepare, clean, and transform raw data into structured formats suitable for analysis and model training.

- Examples: AWS Glue, Azure Data Factory, Google Dataflow.

NoSQL Database

Stores non-relational data, such as JSON or key-value pairs. NoSQL databases are ideal for semi-structured data, flexible schemas, and high-speed data ingestion, making them useful for handling large volumes of unstructured data or data from real-time applications.

- Examples: Amazon DynamoDB, Azure Cosmos DB, Google Cloud Firestore/Bigtable.

Cloud Services

Relational Database

Stores structured data in tables with predefined schemas. Data scientists use relational databases for organized, structured data storage and for querying data using SQL, which is helpful in data exploration, analysis, and integrating structured data with other sources.

- Examples: Amazon RDS, Azure SQL Database, Google Cloud SQL.

Data Labeling

Provides tools for labeling data, such as images or text, which is essential for supervised machine learning. Data scientists use these services to create labeled datasets for training models, especially in tasks like image classification, object detection, and NLP.

- Examples: AWS SageMaker Ground Truth, Azure Machine Learning Data Labeling, Google Cloud Data Labeling.

Cloud Services

Monitoring

Tracks the performance, health, and metrics of applications and infrastructure, including machine learning models in production. Monitoring helps data scientists ensure that models are running as expected and can alert them to performance or accuracy drift.

- Examples: AWS CloudWatch, Azure Monitor, Google Cloud Monitoring.

Logging

Records events and actions across applications and infrastructure, useful for debugging, auditing, and performance tracking. For data scientists, logging is important for understanding data pipeline operations, model training, and performance in production.

- Examples: AWS CloudTrail/CloudWatch Logs, Azure Log Analytics, Google Cloud Logging.

Cloud Services – Provider View

Service Type	AWS	Azure	GCP
Object Storage	Amazon S3	Azure Blob Storage	Google Cloud Storage
Data Warehouse	Amazon Redshift	Azure Synapse Analytics	Google BigQuery
Machine Learning Platform	Amazon SageMaker	Azure Machine Learning	Google AI Platform
Managed Apache Spark	Amazon EMR	Azure Databricks	Google Dataproc
Big Data Processing (Hadoop)	Amazon EMR	Azure HDInsight	Google Dataproc
Data Lake	AWS Lake Formation / S3	Azure Data Lake Storage	Google Cloud Storage / BigQuery
Event-Driven Messaging	Amazon SNS / SQS	Azure Event Grid / Service Bus	Google Cloud Pub/Sub
Data Integration / ETL	AWS Glue	Azure Data Factory	Google Dataflow
NoSQL Database	Amazon DynamoDB	Azure Cosmos DB	Google Cloud Firestore / Bigtable
Relational Database	Amazon RDS	Azure SQL Database	Cloud SQL
Data Labeling	Amazon SageMaker Ground Truth	Azure Machine Learning Data Labeling	Google Cloud Data Labeling
Monitoring	Amazon CloudWatch	Azure Monitor	Google Cloud Monitoring
Logging	AWS CloudTrail / CloudWatch Logs	Azure Log Analytics	Google Cloud Logging

Cloud Services - Implementation

Before we say good bye!!

"As you complete MB511, remember that this is just the beginning of an exciting journey in data science. You've gained the tools, perspectives, and resilience to tackle real-world challenges and make impactful discoveries. Data science is a field that constantly evolves, and your curiosity, adaptability, and dedication will be your greatest assets.

Wishing each of you success, growth, and fulfilment in your careers as data scientists/data science manager. Go out there and make an impact! Data Science is a small world, our path will cross soon“

By the time, we are meeting again. take care!!

LEARNING NEVER STOPS KEEP UPGRADING YOURSELF!
Eivoling data science upgrading yourself—

Have a question?

Feel Free to Reach out at

- +91-88846-92929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)