Introduction to Machine Learning with Apache Spark

Module 3: Data Engineering for Machine Learning using Apache Spark

Welcome! This alphabetized glossary contains many terms you will find in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

Terms	Definition	Video
Checkpointing	A mechanism for recovering query progress in case of node failures and writing stream data to disk	Spark Structured Streaming
Code snippet	A small section of code that demonstrates a specific functionality or task.	Spark SQL
Data warehouse	A central repository that stores large amounts of data from various sources for analysis and reporting	ETL Workloads with Apache Spark
Distributed computing	Spark can handle large-scale data processing and machine learning tasks by distributing computations across multiple nodes in a cluster.	Machine Learning Pipelines Using Spark
End-to-end latency	The time is required for data to process from the source to the sink	Spark Structured Streaming
Factor analysis	A statistical method used to identify latent factors underlying observed variables in a dataset	Feature Extraction and Transformation
Header	The file's first row contains the column names or labels in a tabular data format.	Spark SQL
IDF	Inverse Document Frequency (IDF) measures the rarity or uniqueness of a term across a collection of documents. You can calculate it by using the logarithm of the ratio between the total number of documents in the collection and the number of documents that contain the term. The IDF value is higher for terms that appear in fewer documents, indicating that these terms are more discriminative and provide more information about the documents in which they appear.	Feature extraction and transformation
Inference	Applying a trained machine learning model to make predictions or draw conclusions on new, unseen data	Feature Extraction and Transformation
JDBC (Java Database Connectivity)	A Java API that allows interaction with relational databases using SQL queries.	Spark SQL
Machine learning pipeline	A structured approach to building and deploying machine learning models, encompassing the entire process from data ingestion to model deployment	Machine Learning Pipelines Using Spark

MaxAbsScaler	A function in Spark for scaling numerical features by their maximum absolute value	Feature Extraction and Transformation
MinMaxScaler	A function in Spark for scaling numerical features to a specified range, typically between 0 and 1	Feature Extraction and Transformation
Model persistence	The process of saving a trained machine learning model to disk for future use, enabling its reuse, sharing, and deployment in various applications.	Model Persistence.
Model selection and training	The step in the pipeline is where the appropriate machine learning model is selected and trained using the preprocessed data.	Machine Learning Pipelines Using Spark
One-hot encoding	A technique that converts categorical features into numerical features suitable for machine learning	Feature Extraction and Transformation
Portability	The capability of loading a saved model on different computing environments or infrastructures allows for easy sharing, collaboration, and integration into various projects.	Model Persistence.
Principal Component Analysis (PCA)	A dimensionality reduction technique to identify a smaller set of features that can explain the variance in a dataset	Feature Extraction and Transformation
Reproducibility	The ability to replicate and verify machine learning experiments or projects using the same saved model ensures consistency in findings and facilitates knowledge exchange.	Model Persistence.
Result set	The output generated from executing an SQL query on a Data Frame contains the selected data based on the query criteria.	Spark SQL
Scalability	The feature of easily deploying and scaling saved models to handle large volumes of data, supporting efficient processing of big data and real-time predictions	Model Persistence.
Scaling and normalization	Techniques for transforming numerical features into a common scale to prevent biases in data analysis	Feature Extraction and Transformation
SQL queries	Statements written in SQL syntax to retrieve, manipulate, and analyze data stored in a data frame.	Spark SQL
StandardScaler	A function in Spark for scaling numerical features to have zero mean and unit variance	Feature Extraction and Transformation
Streaming data	Continuously generated data often comes from multiple sources, requiring incremental processing due to its continuous nature.	Spark Structured Streaming
TF	Term Frequency (TF) measures the frequency of a term within a document. It indicates how often a term appears in a document relative to its total number of terms. The idea behind TF is that terms that appear more frequently within a document are	Feature extraction and transformation

	likely to be more important or relevant to that document's content.	
TF-IDF	TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used in information retrieval and text mining to measure the importance of a term (or word) within a document or a collection of documents. TF-IDF combines term frequency (TF) and inverse document frequency (IDF).	Feature extraction and transformation
Tokenization	Tokenization is the process of breaking down a sequence of text into smaller units called tokens. These tokens can be individual words, phrases, sentences, or even characters, depending on the specific requirements of the task at hand. Tokenization is fundamental in natural language processing (NLP) and text analysis tasks.	Feature extraction and transformation
Unified programming interface	Spark's consistent and integrated interface allows data scientists to work seamlessly with different data sources and machine learning algorithms.	Machine Learning Pipelines Using Spark
Watermarking	A process that manages late data in streaming, including late-arriving data and updating results after initial processing	Spark Structured Streaming
Word2Vec	A technique that represents words as vectors in a high-dimensional space, capturing semantic relationships between words.	Feature Extraction and Transformation