COURSE GLOSSARY

Introduction to PySpark

Action: An operation (e.g., count, collect, show) that triggers execution of the pending transformations and returns a result or writes data

Apache Spark: An open-source distributed computing system for fast processing of large-scale data by splitting work across a cluster of machines

Broadcast join: An optimization that replicates a small dataset to all worker nodes so joins avoid expensive data shuffles and run more efficiently

Caching and persisting (cache, persist, unpersist): Techniques to store intermediate DataFrames in memory or on disk to avoid recomputation across multiple actions, with unpersist freeing the storage when done

CSV (Comma-Separated Values): A simple plain-text tabular data format that is widely compatible but lacks enforced schema and can lead to ambiguous data types when read into Spark

Data types (IntegerType, LongType, FloatType, DoubleType, StringType): Primitive column types in PySpark used to declare and enforce the type of values stored in DataFrame columns

DataFrame (PySpark DataFrame): A distributed, table-like collection of data with named columns and a schema, optimized for large-scale data processing and SQL-like operations

Large Language Model (LLM): A deep learning–based language model trained on very large text corpora that can perform a wide range of natural language tasks such as generation, summarization, translation, and question answering

describe: A pandas DataFrame method that computes summary statistics (count, mean, std, min, quartiles, max) for numeric columns to give a quick sense of distribution

Joins: Operations that combine rows from two DataFrames based on matching column values, supporting inner, left, right, and full outer join types

Master node: The cluster node that coordinates and schedules tasks, manages resources, and directs worker nodes in a Spark cluster

Pandas UDF: A vectorized UDF type that leverages Apache Arrow and pandas to operate on batches of data for much better performance on large datasets than row-wise Python UDFs

Parquet: A columnar, schema-enforcing file format optimized for efficient read-heavy queries and complex nested data, commonly used in big data workflows

PySpark: The official Python API for Apache Spark that lets you write Spark jobs and manipulate distributed data using Python constructs

Resilient Distributed Dataset (RDD): A low-level immutable distributed collection of elements that supports fault-tolerant parallel transformations and actions across a cluster

Schema: The structure of a DataFrame that specifies column names and data types, enabling Spark to optimize and validate operations

Spark cluster: A group of networked computers (nodes) configured to run Spark jobs in parallel, enabling distributed storage and computation

Spark SQL: The Spark module that lets you run SQL queries against DataFrames and registered views, returning results as DataFrames for further processing

SparkSession: The primary entry point for using Spark functionality in PySpark that manages the connection to a Spark cluster and provides APIs to read data, create DataFrames, and run SQL

StructType / StructField: PySpark types used to define a DataFrame schema, where StructType is the overall structure and StructField defines each column's name and data type

Temporary view (createOrReplaceTempView): A session-scoped named view of a DataFrame that allows querying it with Spark SQL for the duration of the SparkSession

Transformation: A lazy operation on an RDD or DataFrame (e.g., map, filter, select) that defines a new dataset without immediately executing computation

Union: An operation that stacks two DataFrames with identical schemas by appending rows from one DataFrame to another to create a single combined DataFrame

User-Defined Function (UDF): A custom function registered to Spark that applies arbitrary Python logic to DataFrame columns, useful for transformations not covered by built-in functions

Worker node: A cluster node that executes tasks and performs the actual data processing as assigned by the master