

MINISTRY OF EDUCATION AND TRAINING

NATIONAL ECONOMIC UNIVERSITY



Faculty of Economic Mathematics

DSEB Program

FINAL EXAM (Practice)

.....

Program: DSEB Intake: 63

Date: 07/12/2024 Session: 1

Time limit: 60 minutes

Contents

General Knowledge.....	2
Distributed System Questions.....	2
Spark Cores Questions.....	3
Spark DataFrames Questions	4
Spark SQL Questions	5
Best Practices Questions.....	6
Apache Spark Machine Learning.....	7
Graph Analytics with Apache Spark.....	17
Streaming Analytics with Apache Spark – Spark Structured Streaming.....	22

General Knowledge

Distributed System Questions

1. **In a distributed system, which of the following is a common approach to achieve high availability?**
 - A) Data replication across multiple nodes
 - B) Using a single master node
 - C) Implementing a round-robin scheduling algorithm
 - D) Centralized data storage
2. **Which of the following consistency models allows for temporary inconsistencies in distributed systems?**
 - A) Strong consistency
 - B) Eventual consistency
 - C) Causal consistency
 - D) Linearizability
3. **What is the CAP theorem in the context of distributed systems?**
 - A) It states that you cannot achieve Consistency, Availability, and Partition Tolerance simultaneously
 - B) It describes the three types of distributed databases
 - C) It outlines the limitations of single-node systems
 - D) It defines the performance metrics for distributed applications
4. **In distributed systems, what is the purpose of a consensus algorithm?**
 - A) To ensure data is replicated across all nodes
 - B) To agree on a single value among distributed processes
 - C) To partition data evenly across nodes
 - D) To manage resource allocation

5. **Which of the following techniques is used to minimize latency in distributed systems?**
- A) Data sharding
 - B) Load balancing
 - C) Centralized processing
 - D) Synchronous communication

Spark Cores Questions

1. **What is the role of the Spark Driver in a Spark application?**
- A) To execute tasks on worker nodes
 - B) To manage job scheduling and resource allocation
 - C) To store data in memory
 - D) To handle user input and output
2. **Which transformation in Spark creates a new RDD by applying a function to each element of the original RDD?**
- A) filter()
 - B) map()
 - C) reduceByKey()
 - D) flatMap()
3. **What is the significance of RDD lineage in Spark?**
- A) It determines data partitioning strategy
 - B) It allows Spark to recompute lost data efficiently
 - C) It optimizes memory usage during execution
 - D) It tracks user queries for auditing purposes
4. **Which of the following statements about Spark Executors is true?**
- A) Executors are responsible for running tasks and storing data

- B) Executors can only run on the driver node
 - C) Executors are created at runtime and destroyed after task completion
 - D) Executors do not communicate with each other
5. **In Spark, what does the term "lazy evaluation" refer to?**
- A) Immediate execution of transformations upon their definition
 - B) Delaying execution until an action is called
 - C) Executing all transformations in parallel
 - D) Automatically optimizing RDDs for performance

Spark DataFrames Questions

1. **What is one advantage of using DataFrames over RDDs in Spark?**
 - A) DataFrames have a more complex API
 - B) DataFrames support schema evolution
 - C) DataFrames are less efficient for large datasets
 - D) DataFrames require more memory than RDDs
2. **Which method would you use to perform SQL-like operations on a DataFrame?**
 - A) executeSQL()
 - B) sqlContext()
 - C) selectExpr()
 - D) queryDataFrame()
3. **How can you optimize a DataFrame query using partitioning?**
 - A) By increasing the number of partitions without consideration
 - B) By partitioning based on frequently filtered columns
 - C) By avoiding partitioning altogether
 - D) By using random partitioning
4. **What happens when you call `df.cache()` on a DataFrame?**
 - A) The DataFrame is written to disk
 - B) The DataFrame will be stored in memory for faster access
 - C) The DataFrame is deleted from memory
 - D) The DataFrame's schema is optimized

5. **Which operation would you use to combine two DataFrames with different schemas?**
 - A) union()
 - B) join()
 - C) merge()
 - D) unionByName()
6. **How can you convert a DataFrame back into an RDD?**
 - A) df.toRDD()
 - B) df.rdd
 - C) convertToRDD()
 - D) fromDataFrame()
7. **Which method would you use to group data in a DataFrame and perform aggregations?**
 - A) aggregateByKey()
 - B) groupBy()
 - C) collect()
 - D) summarize()

Spark SQL Questions

1. **What does the WITH clause do in Spark SQL?**
 - A) It creates temporary views for subqueries
 - B) It defines global variables
 - C) It limits query execution time
 - D) It optimizes joins
2. **Which function allows you to define window specifications in Spark SQL for analytics?**
 - A) window()
 - B) over()
 - C) partitionBy()
 - D) groupBy()
3. **In Spark SQL, how can you optimize join operations between large tables?**
 - A) By increasing executor memory
 - B) By using broadcast joins when one table is small enough

- C) By performing joins sequentially
 - D) By avoiding joins altogether
4. **What does EXPLAIN do when prefixed to a SQL query in Spark SQL?**
- A) Executes the query without returning results
 - B) Provides execution plan details for optimization analysis
 - C) Displays runtime statistics
 - D) Creates an index on queried columns
5. **How does Spark SQL handle schema inference when reading JSON files?**
- A) It requires explicit schema definition
 - B) It infers schema based on sample data
 - C) It uses default data types
 - D) It ignores schema entirely
6. **What is the purpose of COALESCE in Spark SQL?**
- A) To merge multiple rows into one
 - B) To reduce the number of partitions
 - C) To create new columns
 - D) To filter null values
7. **Which command would you use to drop a column from a DataFrame before executing SQL queries on it?**
- A) dropColumn()
 - B) removeColumn()
 - C) drop()
 - D) deleteColumn()
8. **What type of join returns all records from both tables, matching where possible, and filling with NULLs where there are no matches?**
- A) Inner join
 - B) Left outer join
 - C) Full outer join
 - D) Right outer join

Best Practices Questions

1. **What strategy should be employed to minimize data shuffling in Spark applications?**
- A) Use wide transformations exclusively

- B) Use narrow transformations whenever possible
 - C) Increase executor memory
 - D) Reduce the number of partitions
2. **Which configuration setting can significantly improve performance when working with large datasets in Spark?**
- A) `spark.sql.shuffle.partitions = 1000`
 - B) `spark.default.parallelism = 10`
 - C) `spark.sql.autoBroadcastJoinThreshold = 10485760`
 - D) `spark.memory.fraction = 0.6`
3. **When should you consider using `persist()` over `cache()` for an RDD or DataFrame?**
- A) When you want to store it only temporarily
 - B) When you need to specify different storage levels
 - C) When using small datasets
 - D) When you want automatic eviction
4. **What is one best practice when writing UDFs (User Defined Functions)?**
- A) Write UDFs that operate on entire datasets at once
 - B) Ensure UDFs are stateless and avoid side effects
 - C) Use UDFs for all transformations regardless of built-in functions
 - D) Avoid testing UDFs before deployment
5. **In which scenario would it be beneficial to use broadcast variables in Spark applications?**
- A) When sharing large datasets across tasks
 - B) When sharing read-only variables efficiently among tasks
 - C) When modifying variables frequently across tasks
 - D) When avoiding serialization issues

Apache Spark Machine Learning

1. **What is the primary purpose of MLlib in Apache Spark?**
- A) Data storage
 - B) Machine learning library
 - C) Data visualization
 - D) Data cleaning

2. **Which of the following algorithms is used for classification in MLlib?**
 - A) K-Means Clustering
 - B) Logistic Regression
 - C) Principal Component Analysis
 - D) Linear Regression
3. **What does the fit() method do in a Spark ML pipeline?**
 - A) It evaluates the model
 - B) It trains the model on the dataset
 - C) It transforms the dataset
 - D) It saves the model to disk.
4. **Which of the following is NOT a component of a Spark ML pipeline?**
 - A) Estimator
 - B) Transformer
 - C) Action
 - D) Stage
5. **What is a transformer in Spark MLlib?**
 - A) An algorithm that produces a model
 - B) An algorithm that transforms one DataFrame into another DataFrame
 - C) A method for data storage
 - D) A tool for data visualization.
6. **Which function would you use to evaluate a regression model in Spark?**
 - A) evaluate()
 - B) score()
 - C) regressionMetrics()
 - D) assess()
7. **What does the crossValidator do in Spark MLlib?**
 - A) It splits data into training and testing sets
 - B) It performs hyperparameter tuning using cross-validation
 - C) It evaluates model performance on unseen data
 - D) It combines multiple models into one.
8. **Which of the following statements about RDDs and DataFrames is true?**
 - A) RDDs are more efficient than DataFrames for all operations
 - B) DataFrames provide optimizations through Catalyst and Tungsten
 - C) RDDs can only handle structured data
 - D) DataFrames cannot be created from RDDs.

9. **What is the purpose of feature extraction in machine learning?**
- A) To reduce the number of features in a dataset
 - B) To transform raw data into a format suitable for modeling
 - C) To evaluate model performance
 - D) To visualize data distributions.
10. **Which method can be used to handle missing values in a DataFrame? Select many**
- A) Fill with mean or median
 - B) Drop rows with missing values
 - C) Ignore missing values entirely
 - D) Replace with zeroes.
11. **What is the role of an estimator in Spark's MLlib?**
- A) To transform data into features
 - B) To fit a model to training data and produce a transformer
 - C) To evaluate model performance
 - D) To load datasets into memory
12. **Which algorithm is typically used for clustering in Spark MLlib?**
- A) Logistic Regression
 - B) K-Means Clustering
 - C) Decision Trees
 - D) Naive Bayes
13. **In PySpark, which function would you use to create a DataFrame from an existing RDD?**
- A) createDataFrame()
 - B) toDF()
 - C) fromRDD()
 - D) buildDataFrame()
14. **What does the term "overfitting" refer to in machine learning?**
- A) When a model performs poorly on training data
 - B) When a model performs well on training data but poorly on unseen data
 - C) When a model has too few parameters
 - D) When a model is too simple
15. **Which evaluation metric is commonly used for binary classification models? Select many**
- A) Precision

- B) Recall
- C) Mean Squared Error
- D) R-squared

16. What does VectorAssembler do in Spark MLlib?

- A) It splits features into separate columns
- B) It combines multiple columns into a single feature vector
- C) It normalizes feature values
- D) It evaluates feature importance

17. Which method would you use to save a trained model in Spark MLlib?

- A) saveModel()
- B) write().save()
- C) exportModel()
- D) persistModel()

18. In Spark, what does Pipeline represent?

- A) A series of transformations applied to data
- B) The entire workflow for machine learning tasks
- C) The storage location for models
- D) The method of evaluating models

19. What does StringIndexer do in PySpark's MLlib?

- A) Converts categorical variables into numerical indices
- B) Normalizes string values
- C) Tokenizes text into words
- D) Encodes strings as one-hot vectors

20. Which algorithm would you use for multi-class classification problems in Spark MLlib? Select many

- A) Decision Trees
- B) Logistic Regression (with One-vs-Rest strategy)
- C) K-Means Clustering
- D) Linear Regression

21. What is the purpose of StandardScaler in machine learning preprocessing?

- A) To reduce dimensionality
- B) To normalize features by removing mean and scaling to unit variance
- C) To encode categorical variables
- D) To fill missing values

22. **Which type of variable can be used to accumulate values across tasks in Spark?**
- A) Broadcast Variable
 - B) Shared Variable
 - C) Accumulator Variable
 - D) Global Variable
23. **In PySpark, which function allows you to split your dataset into training and test sets?**
- A) split()
 - B) randomSplit()
 - C) trainTestSplit()
 - D) partition()
24. **Which statement about hyperparameter tuning is true? Select many**
- A) It involves adjusting parameters before training the model
 - B) It should be done after training the model
 - C) Techniques include grid search and random search
 - D) Hyperparameters are learned from training data
25. **What does CrossValidator help with in machine learning workflows?**
- A) It performs feature selection
 - B) It tunes hyperparameters using cross-validation techniques
 - C) It evaluates model accuracy
 - D) It combines multiple models
26. **Which of the following methods can be used for feature selection in PySpark's MLlib? Select many**
- A) Chi-Squared Test
 - B) Recursive Feature Elimination
 - C) Feature Importance from Tree Models
 - D) Lasso Regression
27. **In which scenario would you prefer using DataFrame over RDD in Apache Spark? Select many**
- A) When working with structured data
 - B) When requiring complex transformations
 - C) When needing optimized execution plans through Catalyst optimizer
 - D) When handling unstructured data

28. Which function would you use to convert categorical features into numerical format in PySpark's MLlib? Select many

- A) StringIndexer
- B) OneHotEncoder
- C) VectorAssembler
- D) IndexToString.

29. What does PipelineModel represent in Spark's MLlib?

- A) The entire workflow for machine learning tasks
- B) The trained version of a pipeline that can be used for predictions
- C) The storage location for models
- D) The method of evaluating models.

30. What is the primary advantage of using ML Pipelines in Apache Spark?

- A) They simplify the workflow by chaining multiple algorithms together
- B) They provide better performance than RDDs
- C) They automatically handle missing values
- D) They require less memory than traditional methods.

31. What is the purpose of OneHotEncoder in PySpark's MLlib?

- A) To convert categorical variables into numerical indices
- B) To create binary columns for each category
- C) To normalize numerical features
- D) To handle missing values

32. Which of the following is a common technique for dimensionality reduction in Spark MLlib?

- A) K-Means Clustering
- B) Principal Component Analysis (PCA)
- C) Logistic Regression
- D) Decision Trees

33. What does the VectorAssembler class do in PySpark?

- A) It splits a vector into individual components
- B) It combines multiple feature columns into a single feature vector
- C) It normalizes feature values
- D) It evaluates model performance

34. Which of the following metrics is NOT typically used for evaluating regression models?

- A) Mean Absolute Error (MAE)

- B) Root Mean Squared Error (RMSE)
- C) F1 Score
- D) R-squared

35. What is the main advantage of using Pipeline in Spark MLlib?

- A) It allows for parallel processing of data
- B) It simplifies the workflow by chaining multiple data transformations and estimators
- C) It automatically handles missing values
- D) It increases model accuracy.

36. Which method would you use to perform hyperparameter tuning with cross-validation in Spark MLlib?

- A) CrossValidator
- B) ParamGridBuilder
- C) TrainValidationSplit
- D) HyperparameterTuner

37. What does the randomSplit() function do in PySpark?

- A) It splits a DataFrame into training and testing sets
- B) It combines multiple datasets into one
- C) It evaluates model performance on test data
- D) It aggregates data across partitions.

38. In Spark MLlib, what does ChiSquareTest help to evaluate?

- A) Model accuracy
- B) Feature importance based on categorical features
- C) The effectiveness of regression models
- D) The performance of clustering algorithms.

39. Which function can be used to convert a DataFrame column to a feature vector in PySpark?

- A) toVector()
- B) assemble()
- C) VectorAssembler().transform()
- D) createVector()

40. What type of learning does Reinforcement Learning refer to?

- A) Learning from labeled data
- B) Learning through trial and error to maximize rewards

- C) Learning from unlabeled data
- D) Learning by mimicking human behavior

41. Which algorithm is commonly used for anomaly detection in Spark MLlib?

- A) K-Means Clustering
- B) Isolation Forest
- C) Decision Trees
- D) Linear Regression

42. What does the StandardScaler class do in PySpark?

- A) It reduces dimensionality
- B) It scales features to have zero mean and unit variance
- C) It encodes categorical variables
- D) It fills missing values

43. Which method can be used to assess feature importance in tree-based models?

- A) FeatureSelector()
- B) featureImportances
- C) ImportanceEvaluator()
- D) ImportanceMetrics()

44. In PySpark, what is the purpose of IndexToString?

- A) To convert numeric indices back to original string labels
- B) To encode string labels into numeric indices
- C) To normalize string values
- D) To tokenize text into words

45. Which algorithm would you use for collaborative filtering in Spark MLlib?

- A) K-Means Clustering
- B) Alternating Least Squares (ALS)
- C) Logistic Regression
- D) Decision Trees

46. What is the primary purpose of using Accumulators in Spark?

- A) To collect metrics from multiple nodes
- B) To cache intermediate results in memory
- C) To broadcast variables across tasks
- D) To manage distributed datasets.

47. In PySpark, which function is used to create a pipeline?

- A) createPipeline()

- B) Pipeline() constructor
- C) buildPipeline()
- D) initPipeline()

48. What does TrainValidationSplit do in Spark MLlib?

- A) Splits data into training and validation sets for hyperparameter tuning
- B) Combines training and testing datasets
- C) Evaluates model performance on unseen data
- D) Performs cross-validation on multiple models.

49. Which method would you use to visualize decision boundaries of a trained model in PySpark?

- A) plotDecisionBoundary()
- B) visualizeModel()
- C) plot() with Matplotlib or Seaborn after collecting predictions from the model
- D) drawBoundary()

50. Which type of neural network is commonly used for image classification tasks?

- A) Recurrent Neural Network
- B) Convolutional Neural Network (CNN)
- C) Feedforward Neural Network
- D) Generative Adversarial Network

51. In PySpark, what does VectorSlicer do?

- A) It selects specific features from a vector column based on indices
- B) It normalizes vector values
- C) It combines multiple vectors into one
- D) It splits vectors into separate columns

52. Which function would you use to compute the confusion matrix for a classification model in PySpark?

- A) computeConfusionMatrix()
- B) MulticlassMetrics()
- C) evaluateConfusionMatrix()
- D) getConfusionMatrix()

53. What is the purpose of using StringIndexerModel after fitting a StringIndexer?

- A) To convert strings back to their original form
- B) To apply the same transformation to new data using the fitted model
- C) To evaluate string encoding accuracy
- D) To normalize string values

54. Which method can be used to save and load models in Spark MLlib? Select many

- A) write().save() / load()
- B) exportModel() / importModel()
- C) saveModel() / loadModel()
- D) persistModel() / retrieveModel().

55. What is the role of MinMaxScaler in machine learning preprocessing?

- A) To scale features to a specified range, usually [0, 1]
- B) To standardize features by removing mean and scaling to unit variance
- C) To normalize categorical variables
- D) To fill missing values

56. In which scenario would you use Random Forest over Decision Trees in Spark MLlib? Select many

- A) When needing better generalization and reduced overfitting risk
- B) When requiring faster training times
- C) When dealing with high-dimensional datasets
- D) When interpreting individual decision paths is crucial

57. What does the term "bagging" refer to in ensemble learning methods like Random Forests?

- A) Combining predictions from different algorithms
- B) Training multiple models on different subsets of data and averaging their predictions
- C) Using boosting techniques to improve weak learners
- D) Selecting features randomly for each model

58. How can you handle imbalanced datasets when training models in Spark MLlib? Select many

- A) Use oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique)
- B) Use undersampling techniques to reduce majority class samples
- C) Ignore class imbalance entirely
- D) Use algorithms that are robust to class imbalance

59. Which method would you use to evaluate multi-class classification models effectively?

- A) MulticlassClassificationEvaluator().evaluate()
- B) BinaryClassificationEvaluator().evaluate()

- C) accuracy(), precision(), recall(), F1 score metrics calculations directly
- D) confusionMatrix().

60. In PySpark, what does CrossValidatorModel represent after fitting a CrossValidator?

- A) The best hyperparameters found during cross-validation process
- B) The trained version of the entire pipeline with best parameters selected from cross-validation process
- C) The evaluation metrics calculated during cross-validation
- D) The storage location for trained models.

Graph Analytics with Apache Spark

1. What is GraphFrames in Apache Spark?

- A) A library for machine learning
- B) A package for DataFrame-based graphs
- C) A tool for data visualization
- D) A method for data storage

2. Which of the following operations can be performed using GraphFrames?
Select many

- A) Motif finding
- B) PageRank calculation
- C) Data cleaning
- D) Linear regression

3. What is the primary purpose of the vertices DataFrame in a GraphFrame?

- A) To store edge information
- B) To store node information
- C) To perform graph queries
- D) To visualize the graph

4. In GraphFrames, what does the edges DataFrame represent?

- A) The relationships between vertices
- B) The attributes of vertices
- C) The weights of edges
- D) The graph's metadata

5. Which method would you use to create a GraphFrame in PySpark?

- A) GraphFrame(vertices, edges)

- B) createGraph(vertices, edges)
- C) buildGraph(vertices, edges)
- D) initializeGraph(vertices, edges)

6. **What does the breadthFirstSearch() method do in GraphFrames?**

- A) It finds the shortest path between two vertices
- B) It performs a breadth-first search traversal of the graph
- C) It calculates the degree of each vertex
- D) It identifies connected components in the graph.

7. **Which algorithm is used to find connected components in a graph using GraphFrames?**

- A) PageRank
- B) Connected Components algorithm
- C) Breadth-First Search
- D) Dijkstra's algorithm

8. **What is a motif in the context of GraphFrames?**

- A) A single vertex in a graph
- B) A specific pattern of connections between vertices
- C) An edge connecting two vertices
- D) The overall structure of the graph.

9. **Which function would you use to compute PageRank in a GraphFrame?**

- A) computePageRank()
- B) pageRank()
- C) rankVertices()
- D) calculatePageRank()

10. **What does the filter() method do when applied to a GraphFrame?**

- A) It removes vertices from the graph
- B) It filters edges based on specified conditions
- C) It aggregates vertex attributes
- D) It transforms the graph structure.

11. **In PySpark, how can you visualize a GraphFrame?**

- A) Using Matplotlib directly on GraphFrames
- B) By converting it to an RDD
- C) Using built-in visualization tools in Databricks or external libraries like NetworkX
- D) Visualization is not supported for GraphFrames

12. Which of the following statements about GraphFrames is true? Select many

- A) GraphFrames can handle multiple types of relationships between vertices
- B) GraphFrames are limited to undirected graphs only
- C) They can be used to perform SQL-like queries on graphs
- D) GraphFrames cannot be used with Spark SQL.

13. What is the purpose of the agg() function in a GraphFrame?

- A) To aggregate vertex properties based on some criteria
- B) To filter edges from the graph
- C) To create new vertices in the graph
- D) To perform motif finding.

14. Which property can you set when creating a GraphFrame to define edge weights?

- A) weight
- B) score
- C) value
- D) strength

15. How can you add new vertices to an existing GraphFrame?

- A) By using addVertex() method
- B) By creating a new DataFrame and merging it with existing vertices
- C) By directly modifying the vertices DataFrame
- D) By using appendVertex() method

16. What does find() function do in a GraphFrame?

- A) Finds specific patterns or subgraphs within the main graph
- B) Aggregates vertex data across different motifs
- C) Filters out unwanted motifs from analysis
- D) Computes centrality measures for motifs

17. Which Spark library provides functionalities for working with graphs?

- A) Spark SQL
- B) MLlib
- C) GraphX
- D) Both GraphX and GraphFrames.

18. How do you convert a DataFrame into a format suitable for creating a GraphFrame? Select many

- A) Ensure it has appropriate columns for vertices and edges
- B) Convert it into an RDD first

- C) Use schema definitions that match vertex and edge requirements
- D) Flatten nested structures into single columns.

19. In which scenario would you use `aggregateMessages()` in a `GraphFrame`?

- A) To perform aggregation across all vertices
- B) To send messages along edges and aggregate results at destination vertices
- C) To filter messages based on conditions
- D) To count total messages sent between vertices

20. What does `inDegrees` property return in a `GraphFrame`?

- A) The total number of outgoing edges from each vertex
- B) The total number of incoming edges to each vertex
- C) The average degree of all vertices
- D) The maximum degree among all vertices

21. Which function would you use to find number of triangles that pass through each vertex in a graph using `GraphFrames`?

- A) `findTriangles()`
- B) `triangleCount()`
- C) `countTriangles()`
- D) `detectTriangles()`

22. What does `outDegrees` property return in a `GraphFrame`?

- A) The number of edges connected to each vertex
- B) The total number of incoming edges to each vertex
- C) The total number of outgoing edges from each vertex
- D) The average degree among all vertices

23. Which method allows you to run SQL queries on a `GraphFrame`? Select many

- A) `sqlQuery()`
- B) `runSQL()`
- C) `createOrReplaceTempView()` followed by `spark.sql()`
- D) `queryGraph()`

24. How can you efficiently handle large graphs in Spark using `GraphFrames`? Select many

- A) By partitioning data appropriately before creating graphs
- B) By using small datasets only
- C) By leveraging distributed computing capabilities of Spark
- D) By avoiding complex queries on large graphs.

25. **What is one limitation when using traditional RDDs compared to DataFrames for graph analytics in Spark?**
- A) RDDs cannot store structured data
 - B) RDDs lack optimization features like Catalyst and Tungsten found in DataFrames
 - C) RDDs cannot be used with machine learning algorithms
 - D) RDDs cannot handle large datasets
26. **Which type of query would be best suited for analyzing relationships in graphs using PySpark's Graph APIs?**
- A) JOIN operations on edge attributes
 - B) GROUP BY operations on vertex properties
 - C) Pattern matching queries using motifs API
 - D) Aggregation queries on edge weights
27. **In PySpark, how would you visualize relationships between nodes after performing analysis with GraphFrames? Select many**
- A) Use built-in visualization tools provided by Databricks or third-party libraries like NetworkX or Matplotlib
 - B) Directly visualize using Spark's native plotting functions
 - C) Export results as CSV and use external software like Gephi or Cytoscape for visualization
 - D) Visualization is not possible with GraphFrames.
28. **What is one advantage of using DataFrame-based graphs over RDD-based graphs in Apache Spark? Select many**
- A) Better performance due to optimizations like Catalyst and Tungsten
 - B) Ability to handle unstructured data
 - C) Easier integration with other Spark components like MLlib and Spark SQL
 - D) More complex API requiring deeper understanding.
29. **Which algorithm would you use for community detection in large graphs with GraphFrames?**
- A) K-Means Clustering
 - B) Label Propagation algorithm
 - C) Decision Trees
 - D) Logistic Regression
30. **What does triangleCount() return when applied to a GraphFrame?**
- A) The total number of triangles formed by three connected vertices
 - B) The average number of triangles per vertex

- C) The maximum triangle size found
- D) The minimum triangle size found.

Streaming Analytics with Apache Spark – Spark Structured Streaming

1. **What is Spark Structured Streaming?**
 - A) A library for batch processing
 - B) A stream processing engine built on Spark SQL
 - C) A tool for data visualization
 - D) A method for static data analysis
2. **Which of the following describes the processing model used by Structured Streaming?**
 - A) Continuous processing model
 - B) Micro-batch processing model
 - C) Batch processing model
 - D) Event-driven processing model
3. **What is the purpose of watermarks in Structured Streaming?**
 - A) To track the progress of data in a stream
 - B) To manage stateful operations
 - C) To filter out late data
 - D) To optimize query performance
4. **Which of the following sources are suitable for ingesting data into Structured Streaming? Select many**
 - A) Kafka
 - B) Flume
 - C) HDFS
 - D) Kinesis
5. **What does the writeStream method do in Spark Structured Streaming?**
 - A) It writes data to a database
 - B) It starts a streaming query to write output to a sink
 - C) It stops a running streaming query
 - D) It reads data from a source.
6. **Which output mode in Structured Streaming only outputs new rows as they arrive?**
 - A) Complete mode

- B) Append mode
- C) Update mode
- D) Batch mode

7. What is the role of foreachBatch in Structured Streaming?

- A) To process each batch of data as it arrives
- B) To aggregate results over time
- C) To join two streams together
- D) To filter out unwanted records.

8. What does the trigger option control in a streaming query?

- A) The frequency of data ingestion from sources
- B) The timing of when to execute the query
- C) The type of output sink used
- D) The format of the incoming data.

9. Which function allows you to perform aggregations on streaming data?

- A) groupBy()
- B) aggregate()
- C) count()
- D) sum()

10. In Spark Structured Streaming, what is a "checkpoint"?

- A) A temporary storage location for intermediate results
- B) A mechanism for fault tolerance that saves the state of a streaming application
- C) A method for optimizing query performance
- D) A way to visualize streaming data

11. What is the default output mode for streaming queries in Spark?

- A) Complete
- B) Append
- C) Update
- D) Batch

12. Which method would you use to stop a running streaming query gracefully?

- A) stopQuery()
- B) terminate()
- C) awaitTermination()
- D) stop().

13. Which of the following statements about stateful operations in Structured Streaming is true? Select many

- A) They require maintaining state across micro-batches
- B) They cannot be used with aggregations
- C) They allow for operations like windowed aggregations
- D) They are less efficient than stateless operations.

14. What does `outputMode("complete")` do in a streaming query?

- A) Outputs only new rows as they arrive
- B) Outputs the entire result set every time there is an update
- C) Outputs updated rows only
- D) Outputs no results at all.

15. What is the purpose of `trigger(ProcessingTime("10 seconds"))` in a streaming query?

- A) To set the batch interval for processing incoming data
- B) To define how often to write results to an output sink
- C) To control how long to wait before stopping the query
- D) To specify how often to check for new data in Kafka.

16. In Spark Structured Streaming, what does `selectExpr()` allow you to do?

- A) Perform complex transformations on DataFrames
- B) Execute SQL expressions directly on DataFrames
- C) Filter out specific records from a stream
- D) Aggregate results over time

17. Which function would you use to read from a Kafka topic in Spark Structured Streaming?

- A) `readKafka()`
- B) `readStream().format("kafka")`
- C) `streamFromKafka()`
- D) `loadKafkaStream()`

18. What does `writeStream.format("console")` do in a streaming application?

- A) Writes output to a file system
- B) Displays results on the console
- C) Sends output to a Kafka topic
- D) Stores results in HDFS

19. How can you handle late-arriving data in Structured Streaming? Select many

- A) By using watermarks
- B) By ignoring late data entirely

- C) By adjusting batch intervals
- D) By using stateful operations with event-time processing.

20. What does groupByKey() do in Spark Structured Streaming?

- A) Groups rows based on specified keys and allows aggregation
- B) Sorts rows by key values only
- C) Filters out non-matching keys
- D) Merges multiple streams into one

21. Which operation would you use to join two streaming DataFrames?

- A) join()
- B) merge()
- C) combine()
- D) union()

22. What is meant by "exactly-once" semantics in Spark Structured Streaming?

- A) Each record may be processed multiple times
- B) Each record is guaranteed to be processed exactly once
- C) Records can be lost during processing
- D) Records are processed at least once but not exactly once

23. In which scenario would you use Continuous Processing mode in Spark Structured Streaming? Select many

- A) When low latency is critical and you need response times under 1 millisecond
- B) When processing large batches of historical data
- C) When high throughput is required
- D) When you want at-least-once guarantees rather than exactly-once guarantees.

24. Which command would you use to monitor active streaming queries in Spark?

- A) spark.streams.active()
- B) monitorQueries()
- C) listQueries()
- D) showActiveStreams()

25. What does trigger(Once()) do in a structured streaming application?

- A) Processes all available data once and then stops
- B) Continuously processes incoming streams
- C) Triggers every 5 seconds
- D) Runs indefinitely until manually stopped.

26. Which method allows you to apply transformations on each RDD generated from a DStream?

- A) foreachRDD()
- B) mapRDD()
- C) transformRDD()
- D) applyToRDD().

27. What is the purpose of checkpointing in Spark Structured Streaming?

- A) To save intermediate results for later analysis
- B) To recover from failures and maintain state across restarts
- C) To optimize performance by caching data
- D) To visualize streaming data.

28. How can you enable checkpointing for a streaming query?

- A) Set checkpointLocation when starting the query
- B) EnableCheckpointing() method
- C) Use startCheckpointing() function
- D) Checkpointing cannot be enabled.

29. Which function would you use to aggregate counts over a sliding window in structured streaming?

- A) countByWindow()
- B) window() followed by groupByKey().agg(count())
- C) slidingWindowCount()
- D) aggregateWindow().

30. In Spark Structured Streaming, what does awaitTermination() do?

- A) Stops all running queries
- B) Waits indefinitely until the query is terminated manually or due to an error
- C) Automatically restarts failed queries
- D) Triggers all pending queries immediately.