MINISTRY OF EDUACATION AND TRAININ(

NATIONAL ECONOMIC UNIVERSITY

Faculty of Economic Mathematics

DSEB Program

MIDTERM EXAM (Practice)

....

Program: DSEB **Intake**: 63

Date: 09/11/2024 **Session: 1**

Time limit: 60 minutes

Distributed System and Apache Spark Fundamentals

1. What is a key characteristic of a distributed system?

A) Centralized control

B) Scalability

C) Single point of failure

D) Limited resource sharing

2. Which of the following is a common challenge in distributed systems?

A) Network latency

B) Data locality

C) Single-threaded processing

D) High memory usage

3. In Apache Spark, what does RDD stand for?

A) Resilient Data Distribution

B) Random Data Distribution

C) Resilient Distributed Dataset

D) Regular Data Definition

4. Which of the following is NOT a feature of Apache Spark?

A) In-memory processing

B) Lazy evaluation

C) Real-time data processing

D) Strict consistency

5. What is the purpose of the SparkContext in Spark applications?

A) To define the schema of data

B) To manage resources and coordinate the execution of tasks

C) To store data in memory

D) To read data from external sources

6. Which of the following storage formats can Spark read? (Select all that apply)

A) Parquet

B) JSON

C) CSV

D) XML

7. How does Spark achieve fault tolerance?

A) By using replication

B) By using RDD lineage

C) By storing all data on disk

D) By automatically restarting nodes

8. What does the map() transformation do in Spark?

A) It filters elements from an RDD.

B) It applies a function to each element and returns a new RDD

C) It reduces the number of partitions.

D) It combines two RDDs.

9. Which of the following components is part of the Spark ecosystem? (Select all that apply)

A) Spark SQL

B) Hadoop HDFS

C) Spark Streaming

D) Apache Kafka

10. What is a DataFrame in Apache Spark?

A) A distributed collection of data organized into named columns

B) A fixed-size array of data.

C) A single partition of data.

D) A collection of RDDs.

11. Which of the following methods can be used to persist data in Spark? (Select all that apply)

A) cache()

B) persist()

C) saveAsTextFile()

D) store()

12. How can you handle missing values in a DataFrame?

A) Using fillna()

B) Using dropna()

C) Ignoring them

D) Converting them to zero

13. What is the role of the Driver program in Spark?

A) To execute tasks on worker nodes.

B) To manage the execution flow and coordinate tasks

C) To store data in memory.

D) To read data from external sources.

14. Which of the following is a valid way to create an RDD from a collection?

A) spark.createRDD(collection)

B) spark.parallelize(collection)

C) spark.makeRDD(collection)

D) spark.newRDD(collection)

15. What does the reduceByKey() operation do?

A) It combines values with the same key using a specified function

B) It filters out keys based on a condition.

C) It sorts keys in ascending order.

D) It groups keys together without aggregation.

16. In which scenario would you use broadcast variables in Spark?

A) To send large amounts of data to all nodes efficiently

B) To store intermediate results.

C) To partition data across nodes.

D) To filter datasets.

17. Which of the following describes "lazy evaluation" in Spark?

A) Operations are executed immediately upon being called.

B) Transformations are not computed until an action is called

C) Data is stored on disk by default.

D) All computations happen in parallel.

18. What type of join does Spark perform by default when joining two DataFrames?

A) Inner join

B) Left join

C) Right join

D) Full outer join

19. How can you optimize performance in a Spark application? (Select all that apply)

A) Using partitioning effectively

B) Reducing the number of transformations

C) Increasing the number of partitions unnecessarily

D) Caching intermediate results

20. What is a common use case for Apache Spark?

A) Batch processing large datasets

B) Building web applications

C) Real-time user authentication

D) Static website hosting

Apache Spark DataFrames

1. What is the primary abstraction used for working with structured data in Spark?

A) RDD

B) DataFrame

C) Dataset

D) Table

2. Which of the following methods can be used to create a DataFrame from an existing RDD?

A) createDataFrame()

B) toDF()

C) fromRDD()

D) loadDataFrame()

3. How can you display the first 10 rows of a DataFrame?

A) df.show(10)

B) df.head(10)

C) df.first(10)

D) df.display(10)

4. Which method is used to rename a column in a DataFrame?

A) renameColumn()

B) withColumnRenamed("oldName", "newName")

C) changeColumnName()

D) setColumnName()

5. How do you filter rows in a DataFrame based on a condition?

A) df.filter(condition)

B) df.where(condition)

C) df.select(condition)

D) Both A and B

6. What is the default behavior of the dropDuplicates() method in a DataFrame?

A) It drops all rows.

B) It keeps the first occurrence of each duplicate row .

C) It drops all duplicates without keeping any.

D) It only drops duplicates based on specified columns.

7. Which of the following functions can be used to aggregate data in a DataFrame? (Select all that apply)

A) count()

B) sum()

C) avg()

D) concat()

8. What does the withColumn() method do in a DataFrame?

A) It adds a new column or replaces an existing column .

B) It filters rows based on a condition.

C) It renames an existing column.

D) It drops a column.

9. How can you convert a DataFrame to an RDD?

A) df.toRDD()

B) df.rdd

C) df.asRDD()

D) df.convertToRDD()

10. Which method can be used to read a CSV file into a DataFrame?

A) spark.read.csv("file.csv")

B) spark.loadCSV("file.csv")

C) spark.read.load("file.csv")

D) spark.importCSV("file.csv")

11. What is the purpose of the join() method in DataFrames?

A) To concatenate two DataFrames vertically.

B) To combine two DataFrames based on a common column .

C) To merge two DataFrames into one.

D) To filter rows from two DataFrames.

12. How do you display the schema of a DataFrame?

A) df.printSchema()

B) df.showSchema()

C) df.schema()

D) df.displaySchema()

13. In Spark, what does the cache() method do?

A) It permanently stores the DataFrame.

B) It optimizes the query plan.

C) It stores the DataFrame in memory for faster access .

D) It drops the DataFrame from memory.

14. Which of the following methods can be used to drop a column from a DataFrame?

A) drop("columnName")

B) remove("columnName")

C) delete("columnName")

D) exclude("columnName")

15. What does the distinct() method do in a DataFrame?

A) It sorts the DataFrame.

B) It removes duplicate rows .

C) It filters out null values.

D) It aggregates data.

16. How can you group data in a DataFrame and perform an aggregation?

A) df.groupBy("column").agg(sum("value"))

B) df.aggregate("column", sum("value"))

C) df.group("column").sum("value")

D) df.groupBy("column").aggregate(sum("value"))

17. Which of the following can be used to handle missing values in a DataFrame? (Select all that apply)

A) fillna(value)

B) dropna()

C) replaceNulls(value)

D) ignoreNulls()

18. What is the purpose of the orderBy() method in a DataFrame?

A) To filter rows based on conditions.

B) To sort the rows based on one or more columns .

C) To group rows together.

D) To aggregate data.

19. How can you save a DataFrame as a Parquet file?

A) df.write.parquet("output.parquet")

B) df.saveAsParquet("output.parquet")

C) df.writeToParquet("output.parquet")

D) df.saveParquet("output.parquet")

20. Which of the following statements about Spark DataFrames is true? (Select all that apply)

A) They are immutable .

B) They can contain mixed data types .

C) They can only contain numeric data types.

D) They are optimized for query execution .

21. What does the union() method do in Spark DataFrames?

A) Combines two DataFrames, retaining duplicates .

B) Combines two DataFrames, removing duplicates.

C) Merges two DataFrames based on keys.

D) Filters rows from two DataFrames.

22. Which method allows you to change the data type of a column in a DataFrame?

A) cast("newType")

B) changeType("newType")

C) convertType("newType")

D) modifyType("newType")

23. How do you perform an inner join between two DataFrames?

A) df1.join(df2, "key", "inner")

B) df1.innerJoin(df2, "key")

C) df1.join(df2, "key")

D) df1.joinInner(df2, "key")

24. What is the output of calling df.columns on a DataFrame?

A) The number of columns in the DataFrame.

B) The names of all columns in the DataFrame .

C) The data types of all columns.

D) The schema of the DataFrame.

25. Which function would you use to find the minimum value of a column in Spark DataFrames?

A) MIN_VALUE()

B) min()

C) lowest()

D) minimum()

26. How do you create a temporary view from a DataFrame for SQL queries?

A) df.createTempView("view_name")

B) df.registerTempTable("view_name")

C) df.createView("view_name")

D) df.createGlobalTempView("view_name")

27. What does the explode() function do in Spark DataFrames?

A) It flattens nested structures into separate rows .

B) It combines multiple columns into one.

C) It filters out null values.

D) It aggregates data.

28. Which of the following is not a valid way to read data into a Spark DataFrame?

A) spark.read.json("file.json")

B) spark.read.csv("file.csv")

C) spark.read.load("file.txt")

D) spark.read.textFile("file.txt")

29. How can you apply a user-defined function (UDF) to a column in a DataFrame?

A) df.apply(udf, "column")

B) df.withColumn("new_column", udf(df["column"]))

C) df.transform(udf, "column")

D) df.udf("column")

30. What does the coalesce() method do when applied to a DataFrame?

A) It increases the number of partitions.

B) It reduces the number of partitions .

C) It merges multiple DataFrames.

D) It filters out null values.

Apache Spark SQL

1. What is the primary abstraction used in Spark SQL?

A) DataFrame

B) RDD

C) Dataset

D) Table

2. Which of the following formats can Spark SQL read natively? (Select all that apply)

A) Parquet

B) JSON

C) CSV

D) XML

3. How can you register a DataFrame as a temporary view in Spark SQL?

A) df.createOrReplaceTempView("view_name")

B) df.registerTempTable("view_name")

C) df.createGlobalTempView("view_name")

D) df.createView("view_name")

4. What is the default behavior of Spark SQL when performing a join operation?

A) Inner join

B) Left outer join

C) Right outer join

D) Full outer join

5. When using Spark SQL, what is the purpose of the explain() method?

A) To execute the query

B) To display the physical plan for the query execution

C) To optimize the query

D) To show the schema of the DataFrame

6. Which of the following functions can be used to aggregate data in Spark SQL? (Select all that apply)

A) COUNT(*)

B) SUM(*)

C) AVG(*)

D) CONCAT

7. In Spark SQL, what is a common way to handle null values in a DataFrame?

A) Use the dropna() method

B) Ignore null values

C) Replace them with a constant value using fillna()

D) Convert nulls to zeros

8. What is the purpose of using the GROUP BY clause in Spark SQL?

A) To filter rows

B) To aggregate results based on one or more columns

C) To sort results

D) To join tables

9. Which of the following statements about DataFrames is true? (Select all that apply)

A) They are immutable

B) They can hold heterogeneous data types

C) They can only contain numeric data types

D) They are optimized for query execution

10. How do you perform a left outer join in Spark SQL?

A) df1.join(df2, "key", "outer")

B) df1.join(df2, "key", "left_outer")

C) df1.leftJoin(df2, "key")

D) df1.join(df2, "key", "left")

11. What is the effect of using distinct() on a DataFrame?

A) It removes duplicate rows from the DataFrame

B) It sorts the DataFrame

C) It adds unique identifiers to each row

D) It filters out null values

12. Which SQL function would you use to concatenate two strings in Spark SQL?

A) CONCATENATE()

B) JOIN()

C) CONCAT(*)

D) MERGE()

13. What is a common use case for window functions in Spark SQL?

A) To group data by categories

B) To perform calculations across a set of rows related to the current row

C) To filter data based on conditions

D) To create temporary views

14. In Spark SQL, how can you change the column names of a DataFrame?

A) Use withColumnRenamed() method

B) Use renameColumns() method

C) Use setColumnNames() method

D) Use changeColumnNames() method

15. Which of the following clauses is used to filter records in a Spark SQL query?

A) WHERE

B) HAVING

C) FILTER

D) SELECT

16. How can you create a permanent table in Spark SQL?

A) Using CREATE TEMPORARY TABLE

B) Using CREATE TABLE with a specified location

C) Using CREATE VIEW

D) Using CREATE GLOBAL TEMPORARY TABLE

17. What does the LIMIT clause do in a Spark SQL query?

A) It filters rows based on conditions.

B) It restricts the number of rows returned by the query

C) It sorts the result set.

D) It aggregates data.

18. Which method can be used to convert a DataFrame into an RDD?

A) df.toRDD()

B) df.rdd

C) df.asRDD()

D) df.convertToRDD()

19. How can you handle schema evolution when reading Parquet files in Spark SQL?

A) Ignore schema evolution

B) Use options like mergeSchema when reading

C) Always define a fixed schema

D) Convert files to CSV format first

20. What is the result of executing SELECT * FROM table WHERE column IS NULL?

A) All rows where 'column' has no value

B) All rows where 'column' has any value

C) An error will occur

D) No rows will be returned

21. Which of the following can be used to perform string manipulation in Spark SQL? (Select all that apply)

A) UPPER(*)

B) LOWER(*)

C) TRIM(*)

D) SPLIT()

22. What is the purpose of the HAVING clause in Spark SQL?

A) To filter records before aggregation

B) To filter records after aggregation

C) To sort records

D) To group records

23. How can you optimize query performance in Spark SQL? (Select all that apply)

A) Use partitioning on large tables

B) Avoid using too many joins

C) Always use non-optimized formats like CSV

D) Cache frequently accessed DataFrames

24. Which function would you use to find the maximum value of a column in Spark SQL?

A) MAX_VALUE()

B) MAX(*)

C) HIGHEST()

D) TOP()

25. In Spark SQL, what does the COALESCE function do?

A) It combines two DataFrames.

B) It returns the first non-null value among its arguments

C) It filters out null values.

D) It sorts values.

26. What is the purpose of using UNION in Spark SQL?

A) To combine rows from two or more queries, removing duplicates

B) To merge two DataFrames into one

C) To join tables based on keys

D) To aggregate results

27. Which of the following statements about DataFrames and Datasets is true? (Select all that apply)

A) DataFrames are untyped, while Datasets are typed

B) Both provide similar functionalities and optimizations

C) DataFrames can only contain numeric data types.

D) Datasets are slower than DataFrames.

28. How do you perform an aggregation with grouping in Spark SQL?

A) SELECT column, SUM(value_column FROM table GROUP BY column

B) SELECT SUM(value_column), GROUP BY column FROM table

C) SELECT column, COUNT(value_column FROM table GROUP BY column

D) SELECT GROUP(column), SUM(value_column FROM table

29. Which command would you use to drop a table in Spark SQL?

A) DELETE TABLE table_name

B) REMOVE TABLE table_name

C) DROP TABLE table_name

D) ERASE TABLE table_name

30. What does the CAST function do in Spark SQL?

A) Converts one data type into another

B) Filters records based on conditions

C) Joins two tables together

D) Aggregates data by type

Apache Spark – Best Practices

1. What is the recommended way to manage Spark application resources?

A) Use a single executor for all tasks

B) Allocate resources dynamically based on workload

C) Set a high number of cores per executor

D) Ignore resource allocation settings

2. When should you use DataFrames instead of RDDs in Spark?

A) When you need to perform low-level transformations

B) When you want better optimization and performance

C) When working with unstructured data

D) When you require complex data structures

3. Which of the following is the best practice for handling large datasets in Spark?

A) Load all data into memory at once

B) Use partitioning to distribute data efficiently

C) Avoid using caching or persistence

D) Read data from disk only once

4. How should you handle skewed data in Spark?

A) Ignore the skew and proceed with processing

B) Use salting techniques to distribute data evenly

C) Increase the number of partitions

D) Use only one partition for processing

5. What is the best practice for writing data back to storage in Spark?

A) Write data in a single format only

B) Use the Parquet format for its efficiency and compression

C) Write data as text files for simplicity

D) Write data without any partitioning

6. When using Spark SQL, what is the best practice for query optimization?

A) Avoid using filter conditions in queries

B) Use broadcast joins for small tables

C) Always use subqueries instead of joins

D) Write complex queries without analyzing execution plans

7. What should you do to avoid memory issues when processing large datasets?

A) Increase the driver memory limit

B) Use more shuffle partitions

C) Load all data into memory

D) Reduce the number of executors

8. Which of the following is the best practice for using UDFs (User Defined Functions)?

A) Use UDFs for all transformations

B) Minimize the use of UDFs and prefer built-in functions

C) Always define UDFs in Python only

D) Use UDFs to handle simple aggregations

9. What is a recommended approach for logging in Spark applications?

A) Use print statements for debugging

B) Implement structured logging with log levels

C) Ignore logging to improve performance

D) Log only errors and ignore other messages

10. How can you improve the performance of Spark jobs?

A) Increase the number of shuffle partitions to a very high number

B) Reduce the amount of data being shuffled

C) Avoid using caching entirely

D) Use more complex transformations

11. Which of the following is thebest practice when dealing with DataFrame operations?

A) Chain multiple operations together

B) Perform all transformations in one step

C) Avoid using DataFrames for small datasets

D) Convert DataFrames back to RDDs frequently

12. What should you do before running a Spark job on a production cluster?

A) Test the job locally with a small dataset

B) Run it directly on the production data

C) Skip testing if the code works on your local machine

D) Increase the number of executors without testing

13. When working with Spark Streaming, what is the best practice for managing stateful operations?

A) Keep all state in memory indefinitely

B) Periodically checkpoint state information

C) Ignore state management for simplicity

D) Use only stateless operations

14. How can you efficiently read data from external sources in Spark?

A) Read all data at once without filtering

B) Use pushdown predicates to minimize data transfer

C) Read data using multiple formats simultaneously

D) Read data without specifying schema

15. What is the best practice regarding Spark application configuration?

A) Use default configurations without changes

B) Tune configurations based on workload requirements

C) Hard-code configurations in your application code

D) Use separate configurations for every job without consistency

16. How should you manage dependencies in Spark applications?

A) Include all dependencies directly in your codebase

B) Use build tools like Maven or SBT to manage dependencies

C) Ignore dependency management to save time

D) Keep dependencies as loose files in your project directory

17. What is the best approach to monitor Spark applications?

A) Monitor logs only after job completion

B) Use the Spark UI and external monitoring tools

C) Ignore monitoring unless there are errors

D) Rely solely on system resource metrics

18. How can you handle schema evolution when reading data from sources like Parquet?

A) Ignore schema changes and proceed with processing

B) Define strict schemas that never change

C) Use options to handle schema evolution gracefully

D) Always convert data to a single schema format before reading

19. What is a recommended practice when using joins in Spark SQL?

A) Always use full outer joins for simplicity

B) Limit the size of the joined datasets when possible

C) Avoid filtering before joining datasets

D) Join multiple large datasets in a single operation

20. How should you handle exceptions in Spark applications?

A) Ignore exceptions and let the job fail silently

B) Implement structured error handling and retries

C) Only log exceptions without taking action

D) Always restart the entire application on failure