**PYSPARK with EMR**

**1.** What does EMR stand for in AWS?

A. Elastic MapReduce  
B. Elastic Memory Resource  
C. Elastic Managed Runtime  
D. Enhanced MapReduce

**2.** Which service is commonly used as storage with EMR and PySpark?

A. EBS  
B. S3  
C. RDS  
D. CloudWatch

**3.** In PySpark, which method reads a CSV file into a DataFrame?

A. spark.read.json()  
B. spark.read.csv()  
C. spark.load.csv()  
D. spark.import.csv()

**4.** Which cluster component in EMR manages the Hadoop cluster metadata (like NameNode)?

A. Core nodes  
B. Master node  
C. Task nodes  
D. Driver node

**5.** Which EMR node type is **optional** and does not store data in HDFS?

A. Master node  
B. Core node  
C. Task node  
D. Metadata node

**6.** Which PySpark action triggers computation?

A. filter()  
B. map()  
C. collect()  
D. select()

**7.** Which file system interface does EMR use to connect Spark with S3?

A. HDFS  
B. EMRFS  
C. S3FS  
D. FSx

**8.** Which mode runs the Spark driver on your local machine when using EMR?

A. YARN cluster mode  
B. YARN client mode  
C. Local mode  
D. Mesos mode

**9.** Which PySpark method is preferred to avoid the **small files problem** in S3 output?

A. df.write.option("merge","true")  
B. df.coalesce(n)  
C. df.groupBy()  
D. df.persist()

**10.** What is the default file system when you spin up an EMR cluster without custom configuration?

A. S3  
B. EMRFS + HDFS  
C. Only local disk  
D. DynamoDB

**11.** Which configuration controls the number of shuffle partitions in Spark SQL?

A. spark.executor.memory  
B. spark.sql.shuffle.partitions  
C. spark.driver.memory  
D. spark.hadoop.fs.s3.maxConnections

**12.** Why is collect() discouraged on very large datasets in EMR?

A. It disables Spark caching  
B. It brings all data to the driver → Out of Memory risk  
C. It is not supported by EMRFS  
D. It deletes cached data after collection

**13.** Which service does EMRFS **consistent view** use internally?

A. S3 Inventory  
B. DynamoDB  
C. CloudTrail  
D. Athena

**14.** In YARN cluster mode on EMR, where does the Spark driver run?

A. On the EMR master node  
B. Inside YARN’s ApplicationMaster container  
C. On all core nodes  
D. In CloudWatch

**15.** Which of the following is a **performance optimization** for reading Parquet in PySpark on EMR?

A. Disable predicate pushdown  
B. Enable vectorized Parquet reader  
C. Convert all Parquet to CSV first  
D. Increase speculative execution

**16.** Which parameter best controls how many parallel tasks a single executor can run?

A. spark.executor.instances  
B. spark.executor.cores  
C. spark.sql.adaptive.enabled  
D. spark.executor.memoryOverhead

**17.** Why does Spark on EMR not achieve **true data locality** with S3?

A. S3 doesn’t support block-level storage  
B. EMRFS disables speculation  
C. HDFS blocks are not replicated  
D. Spark only works with RDS

**18.** If executors in EMR are running on Spot Instances, what is the main risk?

A. Executors can be preempted and terminated anytime  
B. Driver is also terminated  
C. Shuffle files cannot be created  
D. EMR disables EMRFS

**19.** Which optimization reduces shuffle file overhead in skewed joins?

A. Broadcast joins  
B. Salting keys  
C. Increasing executor memory  
D. Writing to local disk

**20.** Which configuration must be enabled along with **spark.dynamicAllocation.enabled=true** for Spark on EMR?

A. spark.speculation  
B. spark.shuffle.service.enabled  
C. spark.executor.memoryOverhead  
D. spark.hadoop.fs.s3.consistent