## 1.11 Spark Core & SQL

### 1 Spark有几种部署方式？请分别简要论述

### 2 Spark任务使用什么进行提交，JavaEE界面还是脚本

### 3 Spark提交作业参数（重点）

### 4 简述Spark的架构与作业提交流程（画图讲解，注明各个部分的作用）（重点）

### 5 如何理解Spark中的血统概念（RDD）（笔试重点）

### 6 简述Spark的宽窄依赖，以及Spark如何划分stage，每个stage又根据什么决定task个数? （笔试重点）

### 7 请列举Spark的transformation算子（不少于8个），并简述功能（重点）

### 8 请列举Spark的action算子（不少于6个），并简述功能（重点）

### 9 请列举会引起Shuffle过程的Spark算子，并简述功能。

### 10 简述Spark的两种核心Shuffle（HashShuffle与SortShuffle）的工作流程（包括未优化的HashShuffle、优化的HashShuffle、普通的SortShuffle与bypass的SortShuffle）（重点）

### 1.11.11 Spark常用算子reduceByKey与groupByKey的区别，哪一种更具优势？（重点）

### 1.11.12 Repartition和Coalesce关系与区别

### 1.11.13 分别简述Spark中的缓存机制（cache和persist）与checkpoint机制，并指出两者的区别与联系

### 1.11.14 简述Spark中共享变量（广播变量和累加器）的基本原理与用途。（重点）

### 1.11.15 当Spark涉及到数据库的操作时，如何减少Spark运行中的数据库连接数？

### 1.11.16 如何使用Spark实现TopN的获取（描述思路或使用伪代码）（重点）

### 1.11.17 京东：调优之前与调优之后性能的详细对比（例如调整map个数，map个数之前多少、之后多少，有什么提升）

### 1.11.18 简述SparkSQL中RDD、DataFrame、DataSet三者的区别与联系? （笔试重点）

### 1.11.19 append和overwrite的区别

### 1.11.20 coalesce和repartition的区别

### 1.11.21 cache缓存级别

### 1.11.22 释放缓存和缓存

### 1.11.23 Spark Shuffle默认并行度

### 1.11.24 kryo序列化

### 1.11.25 创建临时表和全局临时表

### 1.11.26 BroadCast join 广播join

### 1.11.27 控制Spark reduce缓存 调优shuffle

### 1.11.28 注册UDF函数

### 1.11.29 SparkSQL中join操作与left join操作的区别？

## 1.12 Spark Streaming

### 1.12.1 Spark Streaming第一次运行不丢失数据

### 1.12.2 Spark Streaming精准一次消费

### 1.12.3 Spark Streaming控制每秒消费数据的速度

### 1.12.4 Spark Streaming背压机制

### 1.12.5 Spark Streaming 一个stage耗时

### 1.12.6 Spark Streaming 优雅关闭

### 1.12.7 Spark Streaming 默认分区个数

### 1.12.8 SparkStreaming有哪几种方式消费Kafka中的数据，它们之间的区别是什么？

### 1.12.9 简述SparkStreaming窗口函数的原理（重点）