Spark和Flink的区别

可回答：1）Spark Streaming和Flink的区别

问过的一些公司：杰创智能科技(2022.11)，阿里蚂蚁(2022.11)，阿里云(2022.10)(2019.03)，携程(2022.10)，银联(2022.10)，顺丰(2022.09)(2022.05)，贝壳(2022.09)，美团(2022.09)，字节(2022.08)x2(2022.05)(2022.04)(2021.10)(2021.08)，兴金数金(2022.08)，星环科技(2022.07)，西安华为实习(2022.05)，小红书(2022.04)，欢聚(2021.09)，蔚来(2021.09)，百度提前批(2021.08)，网易严选(2021.08)(2019.08)，字节社招(2021.05)，字节实习(2021.03)，中信银行信用卡中心(2020.11)，招银网络(2020.09)，转转(2020.09)，美团优选社招(2020.09)x2，触宝(2020.09)，京东(2020.08)，网易(2020.08)(2018.11)，竞技世界(2020.08)，趋势科技(2020.08)，美团实习(2020.04)，华为实习(2020.04)，美团(2020.04)，快手(2020.03)，爱奇艺(2020.03)，蘑菇街实习(2020.03)，嘉云数据(2020.01)，360社招(2020.01)，阿里(2018.11)

参考答案：

1、编程模型方面

Spark基于**批处理模型，将连续的数据流划分成一系列的微批处理（batch）**，并在每个微批处理中执行Spark RDD操作。因此，它采用了与Spark相同的编程模型，允许开发人员使用Scala、Java或Python进行编程。

Flink则基于**数据流模型**，数据以流的形式输入和输出，支持连续数据处理和有限数据处理。开发人员可以使用Flink提供的DataStream API编写处理逻辑，这些API提供了类似于Spark RDD的转换和操作。

2、数据处理模式方面

Spark**将数据流划分成微批处理，并在每个微批处理中执行一组操作，因此它是一个基于“微批”（micro-batch）的引擎。**这意味着在处理每个微批处理时，Spark Streaming会等待所有数据到达批处理，因此会存在一定的延迟，延迟是秒级。

Flink则是一个**基于“事件时间”（event time）的引擎。**它支持流式处理和批处理，可以根据事件时间对数据进行有序处理，避免了由于乱序数据引起的问题。因此，Flink处理数据时可以保证更低的延迟和更高的准确性，延迟能够达到毫秒级。

3、架构模型方面

Spark Streaming在运行时的主要角色包括：Master、Worker、Driver、Executor， Flink 在运行时主要包：Jobmanager、Taskmanager 和 Slot。

4、任务调度

Spark Streaming连续不断的生成微小的数据批次，构建有向无环图DAG，Spark Streaming会依次创DStreamGraph、JobGenerator、JobScheduler。

Flink 根据用户提交的代码生成 StreamGraph，经过优化生成 JobGraph，然后提交给JobManager 进行处理，JobManager 会根据 JobGraph 生成 ExecutionGraph，ExecutionGraph 是 Flink 调度最核心的数据结构，JobManager 根据 ExecutionGraph 对 Job 进行调度。

5、时间机制

Spark Streaming支持的时间机制有限，只支持处理时间。

Flink支持了流处理程序在时间上的三个定义：处理时间、事件时间、注入时间。同时也支持watermark机制来处理滞后数据。

6、容错机制

对于Spark Streaming任务，可以设置Checkpoint，然后假如发生故障并重启，可以从上次Checkpoint之处恢复，但是这个行为只能使得数据不丢失，可能会重复处理，不能做到恰好一次处理语义。利用Spark Streaming的direct方式与Kafka可以保证数据输入源的，处理过程，输出过程符合Exactly Once。

Flink则使用两阶段提交协议来保证Exactly Once。

7、数据方面

在Flink的世界观中，一切都是由流组成的，离线数据是有界限的流，实时数据是一个没有界限的流，这就是所谓的有界流和无界流。流处理的特点是无界、实时, 无需针对整个数据集执行操作，而是对通过系统传输的每个数据项执行操作，一般用于实时统计。

在Spark的世界观中，一切都是由批次组成的，离线数据是一个大批次，而实时数据是由一个一个无限的小批次组成的。批处理的特点是有界、持久、大量，非常适合需要访问全套记录才能完成的计算工作，一般用于离线统计。

8、应用场景方面

Flink的延迟是毫秒级别，而Spark Streaming的延迟是秒级延迟。

Flink更适合实时流数据处理和事件驱动应用。它是专门设计用于流式数据处理的框架，可以对实时数据流进行高效的计算和处理。

Spark最初是为批处理而设计的，它非常适合对大规模的数据集进行批处理分析，还通过其SQL查询功能提供快速的交互式查询。

Spark Structured Streaming支持实时流处理，但相对于Flink，在处理延迟和状态管理方面可能稍逊一些。