理论习题答案

一、填空题

1.source　channel　sink

2.数据采集

3.hdfs

4.内存

二、选择题

1. C　　2. B　　3. C　　4. C

三、问答题

1.试谈谈对 flume 作用的理解。

答：Flume 是一个分布式、可靠且可扩展的大数据采集工具，它可以帮助用户高效地收集、聚合和移动大量的日志数据。Flume 的作用主要体现在以下几个方面：

（1）数据收集：Flume 可以从各种数据源（如日志文件、消息队列数据库等）中收集数据，并且具有高度可扩展性，能够应对大规模数据的采集需求。通过 Flume 的 Agent 组件，用户可以配置多个不同的数据源，实现并行的数据收集，提高采集效率。

（2）数据聚合：Flume 支持对收集到的数据进行聚合操作，将多个数据源的数据合并为一条记录，便于后续的处理和分析。通过 Flume 的 Channel 组件，用户可以将多个数据源的数据有序地聚合到一起保证数据的完整性和一致性。

（3）数据传输：Flume 可以将收集到的数据传输到目标系统，如 Hadoop、Hive、HBase 等，以便进行进一步的存储、处理和分析通过 Flume 的 Sink 组件，用户可以配置多个不同的目标系统，实现灵活的数据传输和存储策略。

（4）数据过滤：Flume 支持对收集到的数据进行过滤操作，可以根据用户定义的规则对数据进行筛选、转换或丢弃。通过 Flume 的 Interceptor 组件，用户可以自定义一系列的过滤规则，对数据进行实时处理，满足不同的业务需求。

（5）数据可靠性：Flume 具有高度可靠性的特点，能够保证数据的安全传输和可靠存储。Flume 通过可配置的重试机制和故障恢复机制能够在网络故障或系统崩溃的情况下，保证数据的完整性和一致性。

（6）数据监控：Flume 提供了丰富的监控功能，可以实时监控数据的采集、

传输和存储情况。通过 Flume 的监控工具和可视化界面，用户可以实时查看数据的流向、吞吐量、延迟等指标，帮助用户及时发现和解决问题。

Flume 作作为一款强大的大数据采集工具，可以助用户高效、可靠地收集和传输大量的日志数据，支持多种数据源和目标系统，具有灵活的配置和丰富的监控功能。通过使用 Flume，用户可以轻松应对大规模数据采集的需求，提高数据处理的效率和可靠性，从而为后续的数据分析和挖掘工作奠定基础。

2.论述 flume 中各组件的作用。

答：Client：Client 生产数据，运行在一个独立的线程。

Event：一个数据单元，消息头和消息体组成，(Event 可以是日志记录、avro 对象等)。

Flow：event 从源点到达目的点的迁移的抽务。

Agent：一个独立的 flume 进程，包含组件 source、channel、sink，(agent 使用 JVM 运行 flume.每台机器运行一个 aqent，但是可以在一个 agent 中包含多个 sources 和 sinks)。

Source:数据收集组件,(source 从 Client 收集数据,传递给 channel)Source 是数据的收集端，负责将数据捕获后进行特殊的格式化，将数据封装到事件(event)里，然后将事件堆入 Channel 中。Flume 提供了很多内置的 Source，支持 Avro，lg4i，svslog 和 htto postbody 为 ison 格式)。可以应用程序同已有的 Source 直接打交道,如 AvroSource 如果内置的 Source 无法满足需要,Flume 还支持自定义 Source。

Channel:中转 event 的一个临时存诸,保存由 source 组件传递过来的 event，(channel 连接 source 和 sink，这人有点像一个队列)。

Channel 是连接 Source 和 Sink 的组件，大家可以将它看做一个数据的缓冲区(数据队列)，它可以将事件暂存到内存中也可以持久化到本地磁盘上，直到 Sink 处理完该事件。个绍两个较为常用的 Channel，MemorvChannel 和 FileChannel。

Sink：从 channel 中读取并移除 event，将 event 传递到 flowPipline 中的下一个 agent(如果有的话 sink 从 channel 收集数据运行在一个独立的线程)。

Sink 从 Channel 中取出事件，然后将数据发到别处，可以向文件系统、数

据库、 hadoop 存数据，也可以是其他 agent 的 Source。在日志数据较少时，可以将数据存储在文件系统中，并且设定一定的时间间隔保存数据。

Flume 数据流：Flume 的核心是把数据从数据源收集过来，再送到目的地。为了保证输送一定成功，在送到目的地之前，会先缓存数据，待数据真正到达目的地后，删除自己缓存的数据 Flume 传输的数据的基本单位是 Event，如果是文本文件，通常是一行记录，这也是事务的基本单位。Event 从 Source，流向 Channel,再到 Sink,本身为一个 byte 数组，并可携带 headers 信息。Event 代表着一个数据流的最小完整单元，从外部数据源来，向外部的目的地去。

值得注意的是，Flume 提供了大量内置的 Source、Channe 和 Sink 类型，不同类型的 Source.Channe 和 Sink 可以自由组合。组合方式基于用户设置的配置文件，非常灵活。

比如：Channe 可以把事件暂存在内存里，也可以持久化到本地硬盘上。Sink 可以把日志写入 HDFS,HBase，甚至是另外一个 Source 等等。Flume 支持用户建立多级流，也就是说，多个 agent 可以协同工作。

3.简述 spoolding Directory 的作用。

答：Spooling Directory 是监听指定的目录，自动将目录中出现的新文件的内容进行收集如果不指定，默认情况下，一个文件被收集之后，会自动添加一个后缀.COMPLETED，通过通过属性 fileSuffix 来修改。

4.简述 agent 为 a1 的 flume 程序的启动命令。

答：bin/flume-ng agent -n a1 -c conf -f conf/exec.conf -Dflume.root.logger=INFO,console。