Interview Questions - Data Ingestion with Apache Sqoop and Apache Flume

1. **What is the DistCp command?**

In Hadoop, Distcp allows simultaneously copying data to and within the inter/intra Hadoop filesystem. Distcp is implemented using the MapReduce job where the data is copied by the map-only jobs that run in parallel across the cluster.

The following command is used to copy data from dir1 to dir2.

```
$ hadoop distcp dir1 dir2
```

A very common use case for Distcp is in transferring data efficiently from one Hadoop cluster to another.

```
$ hadoop distcp hdfs://namenode1/dir1 hdfs://namenode2/dir2
```

Where

namenode1 - It is the public IP of the namenode of the source cluster.

namenode2 - It is the public IP of the namenode of the target cluster.

2. **Which is the reliable channel in Flume to ensure that there is no data loss?**

The File Channel is the most reliable channel to ensure no data is lost.

3. **Explain the replicating and multiplexing selectors in Flume.**

On the basis of the Flume header value, an event can be written just to a single channel or to multiple channels. If you do not explicitly define a channel selector, it will be the replicating selector by default.

Replicating selector sends the same event to all the channels, while a multiplexing selector sends different events to different channels.

4. **Why is there a need for data ingestion?**

Organisations usually need a common data lake for processes/departments/organisations to access the data received from multiple sources and in multiple formats so that they can use that data to carry out various business functions such as business analytics, modelling, decision making, among others. Data ingestion helps in transferring data to a common data lake.

Data ingestion also helps organise the data available from different sources of records in a standardised format. This can be done at an enterprise level, where data is ingested in a

common enterprise data lake (data warehouse) or at the project level, where a data mart or mini-warehouse can be created.

5. **What are the different types of data ingestion?**

Data ingestion is mainly of two types:

- Real-time ingestion: It includes log capture, transaction data, and event data.
- Batch ingestion: It includes end-of-the-day files coming from a system of records, once-in-a-month process-generated data, use-case data, etc.

6. **How is data stored after ingestion?**

The ingested data is stored on HDFS, HBase, and Hive. Nowadays, cloud-based storage services are also in use, such as Azure, AWS, and Google Cloud. Sometimes, project-specific data ingestion leads to data being stored on the above-mentioned storage tools in arbitrary data formats that are understood by project-specific applications such as CSV, RCFile, ORC, and Parquet.

7. **How would you compress data while ingesting data to HDFS?**

Data can be compressed using the compression options in MapReduce/Pig, Sqoop, etc. Examples of compression formats are Snappy, Deflate, .gz/gzip, etc. These compression algorithms compress files up to one-fifth of their size, and thus help save space.

8. **Explain the importance of using –split-by clause in Sqoop.**

The **--split-by** clause is used to specify the columns of a table that help generate splits for data imports while importing the data into the Hadoop cluster. This clause specifies the columns and helps improve the performance through increased parallelism. It also helps specify the column having an even distribution of data to create splits while importing data..

9. **What is Sqoop metastore?**

Sqoop metastore is a shared metadata repository for remote users to define and execute saved jobs created using the Sqoop job defined in the Sqoop metastore. The Sqoop –site.xml should be configured to connect to the Sqoop metastore.

10. **I have a website and I want to capture logs of the web server. Which channel should I use — memory or file channel?**

If the agent goes down, the Flume state can be restored. Based on this situation, let's consider the following scenarios.

File channel:
Let's say the agent went down and the source of the agent was reading from a database. Now, if you start the agent on another machine, it can resume processing the events from where it had left off.

Memory channel:
In the memory channel, the event state cannot persist in the channel and it will be lost if the agent goes down. However, the memory channel works fast in terms of performance, with the

caveat that if the agent goes down, it can't resume from the point where it had left off and the events are lost.

For web server logs, you should use the memory channel because if the agent goes down, some of the logs can be skipped. This is because it is not as important as transaction data, but if you use a file channel, it will keep pushing the local disk space; thus, it will be filled continuously as logs will be streamed without any fail. So, it is better to use a memory channel.

11. How is Flume fault-tolerant and how does it handle transient spikes in the data generation rate?

Flume uses the channel-based transactions with acknowledgements and is committed for error-free guaranteed data delivery; thus, it uses the Producer-Consumer model to handle situations where the rate of data generation increases suddenly.

12. How can you tune Flume for better performance?

Tiered data collection, choosing the right channel, batch size, channel capacity, and channel transaction capacity are some of the important factors for tuning the Flume and achieving better performance.

13. Is it possible to create and save queries in Sqoop?

Yes, this can be done using the Sqoop job command, which lets you create jobs and run them later.

14. You use –split-by clause but it still does not give optimal performance. How can you then improve the performance further?

In such situations, the –boundary-query clause can be used. Generally, Sqoop uses the SQL query select min(), max() from to determine the boundary values for creating splits. However, if this query is not optimal, then using the –boundary-query argument any random query can be written to generate two numeric columns.

15. How can you run a free form SQL query in Sqoop to import the rows in a sequential manner?

This can be done using the –num-mappers 1 option in the Sqoop import command. It will create only one MapReduce task which will then import rows in a serial order.