**Ingestion Exercises**

1. ABC Corporation has archived data stored on Amazon S3 and also incremental data to be loaded from Oracle database. Suggest the tools that can be used to ingest this data into the Hadoop big data system.

**Answer:** Distcp for loading data from S3 and Sqoop for loading data from a relational database.

1. XYZ Health Analytics is setting up a big data cluster to host their data analytics platform. They are setting up data ingestion using Nifi. Nifi connectors for Kafka and HDFS are used for ingesting and storing the data. Specify the input and output for each of the Nifi connectors.

**Answer:**

Nifi connector 1: Input – Kafka with a Kafka topic, output: connector 2

connector 2: Input: connector 1, output: HDFS

1. An auto manufacturer is setting up a trend analysis system based on twitter data. They are using Flume for twitter data ingestion into a Hadoop cluster. You have seen that Flume uses a source-sync architecture. Define the source, sync and channel for this system so that data is ingested in a fault-tolerant way.

Hint: Channel can be made fault tolerant by using the file system for intermediate storage.

**Answer:** Source: Twitter (Specify twitter authentication), Channel: File (Specify file path) , sync: HDFS (Spcify master node, path)

1. A Telecom enterprise wants to take advantage of real-time big data technology for their data analytics. Sample data is coming like below and they want to process out of order data using Flink. The data analyst wants to process data in **one minute** batches and use the **call end time** for batch processing. Identify the batches in which the data will be processed. The times are in HH:MM:SS format.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Ingestion Time** | **Source Number** | **Destination Number** | **Call start time** | **Call end time** | **Batch** |
| 1 | 10:04:01 | 1112223331 | 1113333331 | 10:01:01 | 10:03:01 | Batch 1 |
| 2 | 10:04:01 | 1112223332 | 1113333332 | 10:01:01 | 10:03:04 | Batch 1 |
| 3 | 10:05:06 | 1112223333 | 1113333333 | 10:01:06 | 10:04:06 | Batch 2 |
| 4 | 10:06:03 | 1112223334 | 1113333334 | 10:02:03 | 10:04:03 | Batch 2 |
| 5 | 10:06:23 | 1112223335 | 1113333335 | 10:02:23 | 10:03:23 | Batch 1 |
| 6 | 10:07:05 | 1112223336 | 1113333336 | 10:03:05 | 10:05:05 | Batch 3 |
| 7 | 10:07:45 | 1112223337 | 1113333337 | 10:03:45 | 10:06:45 | Batch 4 |
| 8 | 10:08:17 | 1112223338 | 1113333338 | 10:04:17 | 10:06:17 | Batch 4 |
| 9 | 10:08:28 | 1112223339 | 1113333339 | 10:04:28 | 10:05:28 | Batch 3 |
| 10 | 10:08:34 | 1112223340 | 1113333340 | 10:03:34 | 10:04:34 | Batch 2 |

**Answer:** Filled in above excel. Out of order data is marked in red.

1. A global financial enterprise has offices in three continents, North America, Europe and Asia. They want to set up real time data processing using Cassandra. They have data coming from ten different sources and the data needs to be stored into different tables in Cassandra. Suggest the real time data tool to use for data ingestion and how the different sources can be handled.

**Answer:** Kafka can be used to ingest data from multiple sources in real time. Different sources can be handled by creating a topic for each source in Kafka.