1. What are the major sources of Big Data?

Enterprise Data (Data for large enterprises like Google/Facebook)

Social Media (Like Twitter/Facebook/LinkedIn)

Public Data (Demographic data)

Transaction Data (Huge number of online/actual transaction in Amazon, Walmart, etc.)

Sensor data (Weather sensor, Traffic sensor, IoT, etc.)

Log Data (Medical diagnosis Logs/Customer Call Centre logs)

Image Video/Audio (Facebook photos/audios)

2. What is the criteria to decide whether the data is Big Data?

Volume, velocity, variety

3. Give three examples for unstructured data?

GPS Tracking, Medical Diagnosis Log, Social media data

4. What is streaming data?

Streaming data is meant to process and analyse large and constantly moving volumes of data that continuously arrives and needs to be processed before the next interval or its expiry

5. What is the difference between Batch and Streaming data?

|  |  |
| --- | --- |
| BATCH DATA | STREAMING DATA |
| Considerable lag between storage and processing | Negligible lag between storage and processing |
| Batch processing - Not real time | Real Time |
| Requirement of high storage capacity | No Requirement of high storage capacity |
| Persist able | Transient, cannot be persisted |
| Gap between two iterations of data arrival | No/negligible gap between two iterations of data arrival |
| Need to be processed in large volumes/entire volume | Frame wise processing mechanism |