# ASSIGNMENT-3

## 1) Design of HDFS:

* The Design of HDFS:-

→ When a dataset outgrows the storage capacity of single physical machine, it becomes necessary to partition it across a number of seperate machines.

→ File systems that manage the storage across a network of machines are called Distributed File Systems.

→ Hadoop comes with a distributed file system called HDFS, which stands for Hadoop distributed file system.

→ HDFS is a file system designed for storing very large files with streaming data accen patterns, running on clusters of commodity hardware.

• Very large files:-

→ "very large" in this context means files that are hundreds of megabytes, gigabytes, or terabytes in size. There are Hadoop clusters running today that store petabytes of data.

• Streaming Data access:-

→ HDFS is built around the idea the most efficient data processing pattern is a write-once, read many-times pattern. A dataset is typically generated or copied from source, then various analyses are performed on that dataset over time.

• Commodity Hardware:-

→ Hadoop doesn't require expensive, highly reliable hardware to run on. It's designed to run on clusters of commodity hardware for which the chance of node failure across the cluster is high, at least for large clusters. HDFS is designed to carry on working without a noticeable interruption to the user in the face of such failure.

These are areas where HDFS is not a good fit today:

- Low Latency data access:

→ Applications that require low-latency access to data, in the tens of milliseconds range, will not work well with HDFS.

- Lots of small files:-

→ Since the name node holds file system metadata in memory, the limit to the number of files in a file system is governed by the amount of memory on the name node.

- Multiple writers, arbitoary file modifications:-

→ Files in HDFS may be written to by a single writer. Writes are always made at the end of the file. there is no support for multiple writers, or for modifications at arbitary offsets in the file.

2) Explain the process of Data ingestion with flume and SQOOP.

* Data ingestion:-

→ Data ingestion is the process of obtaining and importing data or immediate use or storage in a database.

→ Data can be streamed in real time & ingested in batches.

* Data ingestion with flume:-

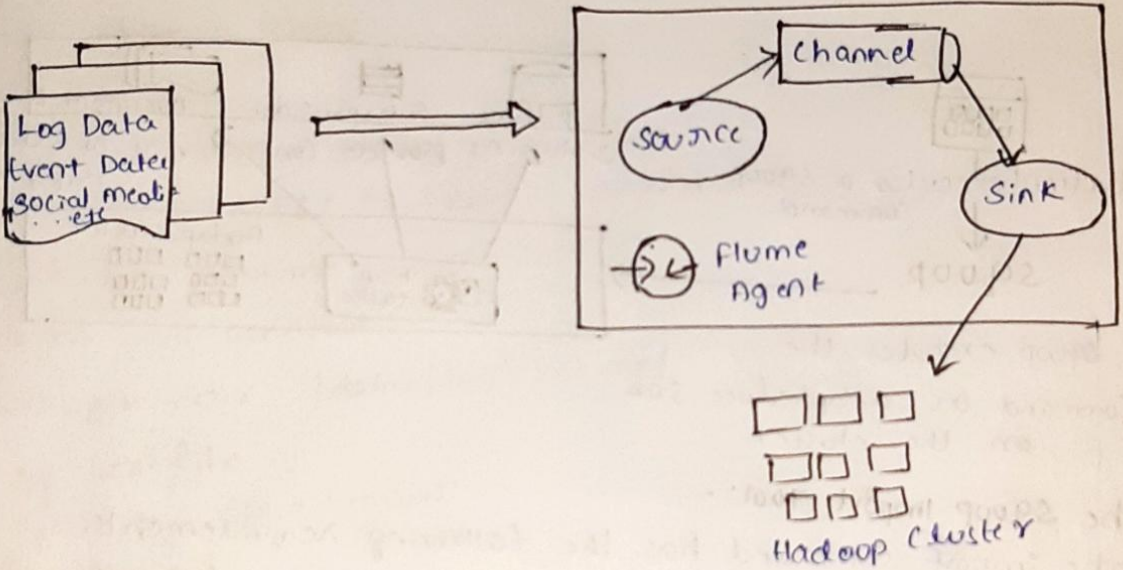
→ Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating and moving large amounts of log data into HDFS.

→ It has a simple and flexible architecture based on streaming data flows; and robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.

→ Enterprises use Flume's powerful streaming capabilities to land data from high-throughghput streams in the HDFS.

→ Typical sources of these streams are application logs, sensor and machine data, geolocation data and social media.

→ These different types of data can be landed in Hadoop for future analysis using interactive queries in Apache Hive. or they can feed business dashboards served ongoing data by Apache HBase.



Components:-

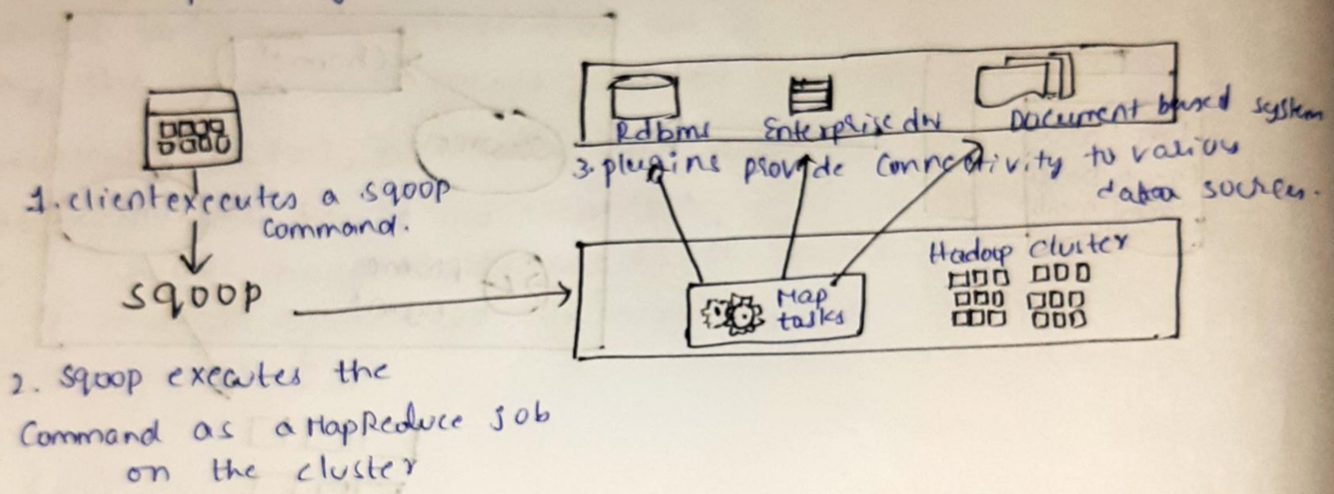
- Event: A singular unit of data that is transported by Flume.
- Source: The entity through which data enters into flume. Sources either actively poll for data or passively wait for data to be delivered to them.
- Sink: The entity that delivers the data to the destination. A variety of sinks allow data to be streamed to a range of destinations.
- Channel: The conduit between the source and sink. Source ingest events into the channel and the sink drains the channel.
- Agent: Any physical JVM running flume. It is a collection of sources, sinks and channels.
- Client: The entity that produces and transmits the event to the source operating within the Agent.

## * Data Ingestion with sqoop:-

→ Apache sqoop is a tool designed to efficiently transfer data between Hadoop and relational databases. We can use sqoop to import data from a relational database table into HDFS.

→ the import process is performed in parallel and thus generates multiple files in the format of delimited text.

→ Moreover, sqoop can export the data bact to the relational databases for consumption by external application to users.


Rdbms
Enterprise dw
Document based system
3. plugins provide connectivity to various data sources.
1. client executes a sqoop command.
sqoop
Hadoop Cluster
Map tasks
2. sqoop executes the command as a MapReduce job on the cluster

- The sqoop Import Tool:-

→ The import command has the following requirements:

- Must specify a connect string using the -connect argument.
- Credentials can be included in the connect string, so use the --username and --password arguments.
- Must specify either a table to import using --table or the result of an SQL query using --query.

Importing a table:-

```
sqoop import
-- connect jdbc:mysql://host/nyse
-- table stockprices
-- target-dir /data/stockprice/
--as-textfile.
```

importing specific columns:-

```
sqoop import
--connect jdbc:mysql://host/nyse
--table StockPrices
--columns StockSymbol, volume, High, ClosingPrice
--target-dir /data/dailyhighs/
--as-textfile
--split-by StockSymbol
-m 10.
```

Importing from a query:-

```
sqoop import
--connect jdbc:mysql://host/nyse
--query "SELECT * FROM StockPrices s
WHERE s.Volume >= 1000000
AND \$CONDITIONS"
--target-dir /data/highvolume/
--as-textfile
--split-by StockSymbol.
```

- The sqoop Export Tool:-
The export command transfers data from HDFS to a database:

→ The export command transfers data from HDFS to a database table.

-Use --table to specify the database table.

- Use --export-dir to specify the data to export.

→ Rows are appended to the table by default.

→ If you define --update-key, existing rows will be updated with the new data.

→ Use - Call to invoke a stored procedure.

Exporting a Table :

```
sqoop export
--connect jdbc:mysql://host/mylogs
--table logData
--export-dir /data/logfiles/
--input-fields-terminated-by "\t".
```

3) Define Biginsights and BigSheets of IBM.

*** Biginsights:-**

→ Infosphere Big insights is a Software platform designed to help firms discover and analyze business insights hidden in large volumes of a diverse range of data.

→ Example of such data include log records, click streams, Social media data, news feeds etc.

→ Biginsights incorporates several open source projects and a number of IBM-developed technologies.

→ Biginsights doesn't replace a relational database management system or a traditional dataware house. It isn't optimized for interactive queries over tabular data structures, online analytical processing, or OLTP applications.

*** BigSheets:-**

→ BigSheets is a Spreadsheet-style tool for business analysts Provided with IBM Infosphere BigInsights, Bigsheets enables non-programmers to iteratively explore, manipulate, and visualize data stored in your distributed file system.

→ Bigsheets translates user commands, expressed through a graphical interface, into Pig scripts executed against a subset of the underlying data.

→ When satisfied, the user can save and run the workbook, which causes Bigsheets to initiate MapReduce jobs over the full set of data, write the results to the distributed file system and display the contents of the new workbook.

→ Since BigSheets is a service running on big data cluster, user does not need to worry about connectivity.