Big data systems are popular for processing huge amounts of unstructured data from multiple data sources. The complexity of the big data system increases with each data source. Most of the business domains have different data types like marketing genes in healthcare, audio and video systems, telecom CDR, and social media. All these have diverse data sources and data from these sources is consistently produced on large scale.

The challenge is to leverage the resources available and manage the consistency of data. Data ingestion is complex in hadoop because processing is done in batch, stream or in real time which increases the management and complexity of data. Some of the common challenges with data ingestion in Hadoop are parallel processing, data quality, machine data on a higher scale of several gigabytes per minute, multiple source ingestion, real-time ingestion and scalability. Apache Sqoop and Apache Flume are two popular open source etl tools for hadoop that help organizations overcome the challenges encountered in data ingestion. If you are looking to find the answer to the question -**"What's the difference between Flume and Sqoop?"** then you are on the right page. The major difference between Sqoop and Flume is that Sqoop is used for loading data from relational databases into HDFS while Flume is used to capture a stream of moving data.

Sqoop vs Flume-Comparison of the two Best Data Ingestion Tools

Sqoop vs Flume

Comparison of Data Ingestion Tools

BASIC DIFFERENFCE

Apache Sqoop is an effective hadoop tool for importing data from RDBMS's.

Apache Flume is service designed for streaming logs into Hadoop environment.

Scoop works well with any kind of RDBMS that has JDBC connectivity.

Flume functions well for streaming data sources which are generated continuously in hadoop environment such as log files from multiple servers.

DATA FLOW

TYPE OF LOADING

Scoop
Not Event Driven

Flume
Event Driven

WHEN TO USE??

Sqoop is an ideal fit if the data is sitting in databases like Teradata, Oracle, MySQL Server, Postgres.

Flume is a better choice when moving bulk streaming data from various sources like JMS or Spooling directory.

LINK TO HDFS

HDFS is the destination for importing data.

Data flows from multiple channels into HDFS

What is Sqoop in Hadoop?

Apache Sqoop (SQL-to-Hadoop) is a lifesaver for anyone who is experiencing difficulties in moving data from the data warehouse into the Hadoop environment. Apache Sqoop is an effective hadoop tool used for importing data from RDBMS's like MySQL, Oracle, etc. into HBase, Hive or HDFS. Sqoop hadoop can also be used for exporting data from HDFS into RDBMS. Apache Sqoop is a command line interpreter i.e. the Sqoop commands are executed one at a time by the interpreter.

Need for Apache Sqoop

With increasing number of business organizations adopting Hadoop to analyse huge amounts of structured or unstructured data, there is a need for them to transfer petabytes or exabytes of data between their existing relational databases, data sources, data warehouses and the Hadoop environment. Accessing huge amounts of unstructured data directly from MapReduce applications running on large Hadoop clusters or loading it from production systems is a complex task because data transfer using scripts is often not effective and time consuming.

How Apache Sqoop works?

Sqoop is an effective hadoop tool for non-programmers which functions by looking at the databases that need to be imported and choosing a relevant import function for the source data. Once the input is recognized by Sqoop hadoop, the metadata for the table is read and a class definition is created for the input requirements. Hadoop Sqoop can be forced to function selectively by just getting the columns needed before input instead of importing the entire input and looking for the data in it. This saves considerable amount of time. In reality, the import from the database to HDFS is accomplished by a MapReduce job that is created in the background by Apache Sqoop.

Learn more on How Apache Sqoop works!

Features of Apache Sqoop

- Apache Sqoop supports bulk import i.e. it can import the complete database or individual tables into HDFS. The files will be stored in the HDFS file system and the data in built-in directories.
- Sqoop parallelizes data transfer for optimal system utilization and fast performance.
- Apache Sqoop provides direct input i.e. it can map relational databases and import directly into HBase and Hive.
- Sqoop makes data analysis efficient.
- Sqoop helps in mitigating the excessive loads to external systems.
- Sqoop provides data interaction programmatically by generating Java classes.

Companies Using Apache Sqoop

- The Apollo Group education company uses Sqoop to extract data from external databases and inject results of Hadoop jobs back into the RDBMS's.
- Coupons.com uses Sqoop tool for data transfer between its IBM Netezza data warehouse and the hadoop environment.

What is Flume in Hadoop?

Apache Flume is service designed for streaming logs into Hadoop environment. Flume is a distributed and reliable service for collecting and aggregating huge amounts of log data. With a simple and easy to use architecture based on streaming data flows, it also has tunable reliability mechanisms and several recovery and failover mechanisms.

Need for Flume

Logs are usually a source of stress and argument in most of the big data companies. Logs are one of the most painful resources to manage for the operations team as they take up huge amount of space. Logs are rarely

present at places on the disk where someone in the company can make effective use of them or hadoop developers can access them. Many big data companies wind up building tools and processes to collect logs from application servers, transfer them to some repository so that they can control the lifecycle without consuming unnecessary disk space.

This frustrates developers as the logs are often not present at the location where they can view them easily, they have limited number of tools available for processing logs and have confined capabilities in intelligently managing the lifecycle. Apache Flume is designed to address the difficulties of both operations group and developers by providing them an easy to use tool that can push logs from bunch of applications servers to various repositories via a highly configurable agent.

For the complete list of big data companies and their salaries-

How Apache Flume works?

Flume has a simple event driven pipeline architecture with 3 important roles- Source, Channel and Sink.

- Source defines where the data is coming from, for instance a message queue or a file.
- Sinks defined the destination of the data pipelined from various sources.
- Channels are pipes which establish connect between sources and sinks.

Apache flume works on two important concepts-

1. The master acts like a reliable configuration service which is used by nodes for retrieving their configuration.
2. If the configuration for a particular node changes on the master then it will dynamically be updated by the master.

Node is generally an event pipe in Hadoop Flume which reads from the source and writes to the Sink. The characteristics and role of a flume node is determine by the behaviour of source and sinks. Apache Flume is built with several source and sink options but if none of them fits in your requirements then developers can write their own. A flume node can also be configured with the help of a sink decorator which can interpret the event and transforms it as it passes through. With all these basic primitives, developers can create

different topologies to collect data on any application server and direct it to any log repository.

Features of Apache Flume

- Flume is a flexible tool as it allows to scale in environments with as low as five machines to as high as several thousands of machines.
- Apache Flume provides high throughput and low latency.
- Apache Flume has a declarative configuration but provides ease of extensibility.
- Flume in Hadoop is fault tolerant, linearly scalable and stream oriented.

Companies Using Apache Flume

- Goibibo uses Hadoop flume to transfer logs from the production systems into HDFS.
- Mozilla uses flume Hadoop for the BuildBot project along with Elastic Search.
- Capillary technologies uses Flume for aggregating logs from 25 machines in production.

Difference between Sqoop and Flume

- Apache Sqoop and Apache Flume work with various kinds of data sources. Flume functions well in streaming data sources which are generated continuously in hadoop environment such as log files from multiple servers whereas Apache Sqoop is designed to work well with any kind of relational database system that has JDBC connectivity. Sqoop can also import data from NoSQL databases like MongoDB or Cassandra and also allows direct data transfer or Hive or HDFS. For transferring data to Hive using Apache Sqoop tool, a table has to be created for which the schema is taken from the database itself.
- In Apache Flume data loading is event driven whereas in Apache Sqoop data load is not driven by events.
- Flume is a better choice when moving bulk streaming data from various sources like JMS or Spooling directory whereas Sqoop is an ideal fit if the data is sitting in databases like Teradata, Oracle, MySQL Server, Postgres or any other JDBC compatible database then it is best to use Apache Sqoop.
- In Apache Flume, data flows to HDFS through multiple channels whereas in Apache Sqoop HDFS is the destination for importing data.
- Apache Flume has agent based architecture i.e. the code written in flume is known as agent which is responsible for fetching data whereas in Apache Sqoop the architecture is based on connectors. The connectors in Sqoop know how to connect with the various data sources and fetch data accordingly.
- Lastly, Sqoop and Flume cannot be used achieve the same tasks as they are developed specifically to serve different purposes. Apache Flume agents are designed to fetch streaming data like tweets from Twitter or log

file from the web server whereas Sqoop connectors are designed to work only with structured data sources and fetch data from them.

- Apache Sqoop is mainly used for parallel data transfers, for data imports as it copies data quickly where Apache Flume is used for collecting and aggregating data because of its distributed, reliable nature and highly available backup routes.