**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

# BÁO CÁO SEMINAR CHƯƠNG 3

## Môn học: Dữ liệu lớn

## Đề tài: APACHE SQOOP

**Mã lớp: IS405.O11.HTCL**

**Giảng viên hướng dẫn: ThS. Nguyễn Hồ Duy Tri**

**Sinh viên thực hiện:**

| | |
|---|---|
| Trần Phương Thảo | 20521938 |
| Phạm Thụy Ý Vy | 20522183 |
| Lưu Yến Vy | 20522180 |

**TP. Hồ Chí Minh, 2023**

# MỤC LỤC

# LỜI CẢM ƠN

Trước hết, chúng em xin gửi tới các thầy, cô khoa Hệ thống thông tin, thuộc trường Đại học Công nghệ thông tin – Đại học quốc gia TP. HCM lời cảm ơn vì đã tận tâm truyền đạt kiến thức, hướng dẫn, đặt nền tảng cơ bản cho chúng em có thể thực hiện đồ án này.

Đặc biệt, chúng em xin gửi lời cảm ơn chân thành đến **Thầy Nguyễn Hồ Duy Tri**

Để đồ án này được đạt kết quả tốt như hiện nay, chúng em đã nhận được rất nhiều sự hỗ trợ và hướng dẫn từ thầy.

Mặc dù đã nổ lực cố gắng hết sức nhưng do kiến thức còn nhiều mặt hạn chế, nên trong quá trình thực hiện không tránh khỏi những thiếu sót. Kính mong nhận được sự góp ý và giúp đỡ từ quý thầy cô để chúng em có thể hoàn thiện đồ án một cách trọn vẹn nhất.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện:

# NHẬN XÉT CỦA GIẢNG VIÊN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

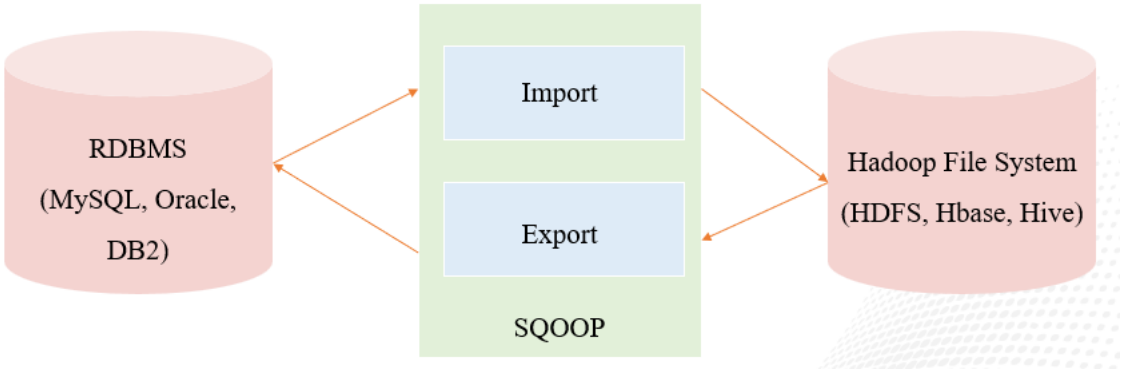
……

# CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ APACHE SQOOP

## 1.1 Tổng quan về Apache Sqoop

### 1.1.1 Khái niệm

Apache Sqoop là một công cụ mã nguồn mở, được sử dụng để chuyển đổi dữ liệu lớn giữa Hadoop và các kho dữ liệu có cấu trúc. Sqoop chuyển dữ liệu từ RDBMS (Hệ quản trị cơ sở dữ liệu quan hệ) như MySQL hay Oracle sang HDFS (Hệ thống lưu trữ phân tán Hadoop). Apache Sqoop cũng có thể được sử dụng để chuyển đổi dữ liệu trong Hadoop MapReduce và sau đó xuất dữ liệu đó sang RDBMS.




Lý do sử dụng SQOOP: Trước tiên, dữ liệu phải được đưa vào các cụm Hadoop từ nhiều nguồn khác nhau để được xử lý bằng Hadoop. Tuy nhiên, việc tải dữ liệu từ nhiều nguồn không đồng nhất là một nhiệm vụ đầy thách thức và bên cạnh đó xuất hiện các vấn đề cần phải xử lý bao gồm:

- Sự nhất quán của dữ liệu.
- Vấn đề quản lý tài nguyên hiệu quả.
- Không thể tải hàng loạt dữ liệu vào Hadoop.
- Việc tải dữ liệu không nhanh chóng.
- MapReduce không thể truy cập trực tiếp CSDL bên ngoài.

→ Giải quyết đơn giản bằng Sqoop, giúp truyền một lượng lớn dữ liệu từ RDBMS vào Hadoop trở nên đơn giản. Trong Sqoop, sử dụng nhập xuất dữ liệu bằng kiến trúc Mapreduce, giúp quá trình xử lý song song trở nên nhanh chóng, tiết kiệm chi phí.

### 1.1.2 Tính năng

Apache Sqoop cung cấp nhiều tính năng nổi bật như:

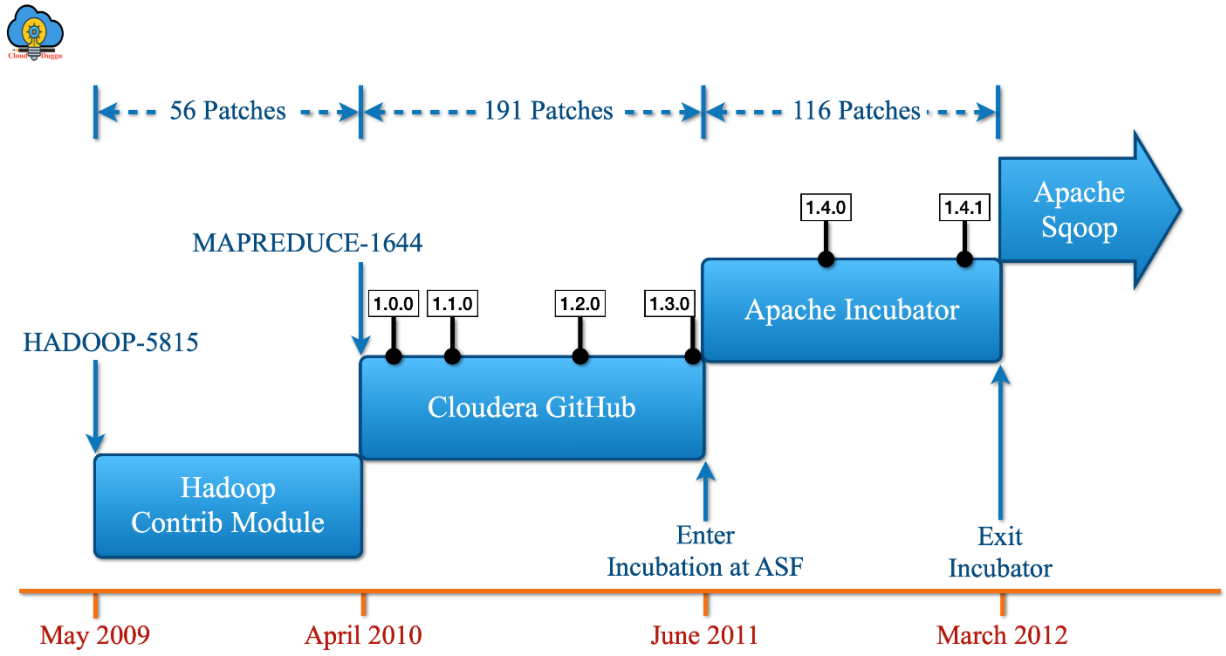
- *Full load/ Incremental load:* Apache Sqoop có thể tải toàn bộ các bảng từ cơ sở dữ liệu chỉ bằng 1 lệnh duy nhất, và có thể cập nhất dữ liệu bất cứ lúc nào.
- *Parallel import/ export:* Apache Sqoop sử dụng framework YARN để import/ export dữ liệu làm tăng khả năng chịu lỗi dựa trên quá trình song song.
- *Import results of SQL query:* Có thể import kết quả trả về từ truy vấn SQL vào HDFS.
- *Compression:* Có thể nén dữ liệu của mình bằng cách sử dụng thuật toán deflate (gzip) với đối số - compress hoặc bằng cách chỉ định đối số - compress -codec. Và cũng có thể tải bảng nén trong Apache Hive.
- *Connectors for all major RDBMS:* Apache Sqoop cung cấp trình kết nối cho hầu như tất cả cơ sở dữ liệu.
- *Kerberos security integration:* Apache Sqoop cung cấp trình kết nối cho hầu như tất cả cơ sở dữ liệu.
- *Load data directly into HIVE/ Hbase:* Có thể tải trực tiếp dữ liệu vào Apache Hive và cũng có thể xuất dữ liệu từ Hbase – một cơ sở dữ liệu NoSQL.

## 1.2 Lịch sử hình thành và phát triển

Apache Sqoop được tạo ra bởi Aaron Kimball vào năm 2009. Ban đầu, nó là một dự án tại Cloudera, một công ty chuyên về các giải pháp dành cho Hadoop và các công nghệ liên quan. Phiên bản đầu tiên của Sqoop, được gọi là Sqoop 1, được phát hành vào tháng 3 năm 2010.

Vào cuối năm 2011, Sqoop được chuyển giao cho Apache Software Foundation và trở thành một dự án Apache được gọi là "Apache Sqoop." Quá trình chuyển giao này đảm bảo tính mã nguồn mở và phát triển tiếp cận rộng rãi của dự án. Trong thời gian tiếp theo, Apache Sqoop 2 được ra đời và phát triển. Apache Sqoop 2 có kiến trúc hiện đại hơn và một loạt các cải tiến so với phiên bản trước.

Trong các thập kỷ sau, Apache Sqoop tiếp tục cung cấp một giải pháp quan trọng cho việc chuyển dữ liệu giữa các hệ thống lưu trữ dữ liệu phân tán và là một phần không thể thiếu trong quá trình xử lý và phân tích dữ liệu lớn trong môi trường Hadoop và Big Data hiện đại.



## 1.3 Các tình huống sử dụng

Apache Sqoop có thể được sử dụng trong các tình huống sau:

- *Sao lưu dữ liệu từ cơ sở dữ liệu quan hệ:* Sqoop có thể được sử dụng để sao lưu dữ liệu từ cơ sở dữ liệu quan hệ vào HDFS hoặc lưu trữ khác để đảm bảo tính sẵn sàng và khả năng phục hồi dữ liệu.
- *Integrate dữ liệu từ nhiều nguồn khác nhau:* Sử dụng Sqoop để tích hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, bao gồm các cơ sở dữ liệu quan hệ và hệ thống lưu trữ dữ liệu phân tán vào một hệ thống lưu trữ dữ liệu duy nhất như HDFS hoặc HBase.
- *Sử dụng dữ liệu trong các công cụ phân tích dữ liệu:* Sau khi dữ liệu đã được chuyển đổi và đưa vào Hadoop, Sqoop cho phép sử dụng dữ liệu này trong các công cụ phân tích dữ liệu như Apache Hive, Apache Pig, Apache Spark hoặc các ứng dụng khác để thực hiện xử lý và phân tích dữ liệu.
- *Tích hợp với lịch trình chạy tự động:* Sqoop có thể tích hợp vào các công cụ quản lý lịch trình như Apache Oozie hoặc cron jobs để thực hiện các quy trình chuyển đổi dữ liệu tự động và định kỳ.

## 1.4 Ứng dụng trong các doanh nghiệp

Hiện nay, Apache Sqoop đã được sử dụng rộng rãi tại hơn 627 công ty và tổ chức trên toàn thế giới để giải quyết các vấn đề liên quan đến xử lý dữ liệu lớn trong nhiều lĩnh vực. Một số các công ty lớn hiện nay đang ứng dụng Apache Sqoop như: Walmart (thuộc lĩnh vực bán lẻ), Citigroup (thuộc lĩnh vực tài chính), Charles River Laboratories (thuộc lĩnh vực sức khỏe), …

Online Marketer Coupons.com đã ứng dụng Apache Sqoop để chuyển dữ liệu giữa Hadoop và kho dữ liệu IBM Netezza.

Công ty chuyên về lĩnh vực giáo dục The Apollo Group đã sử dụng Apache Sqoop để trích xuất dữ liệu và truy vấn dữ liệu giữa Hadoop và các cơ sở dữ liệu quan hệ của họ.

# CHƯƠNG 2: CÁC ĐẶC ĐIỂM CHÍNH

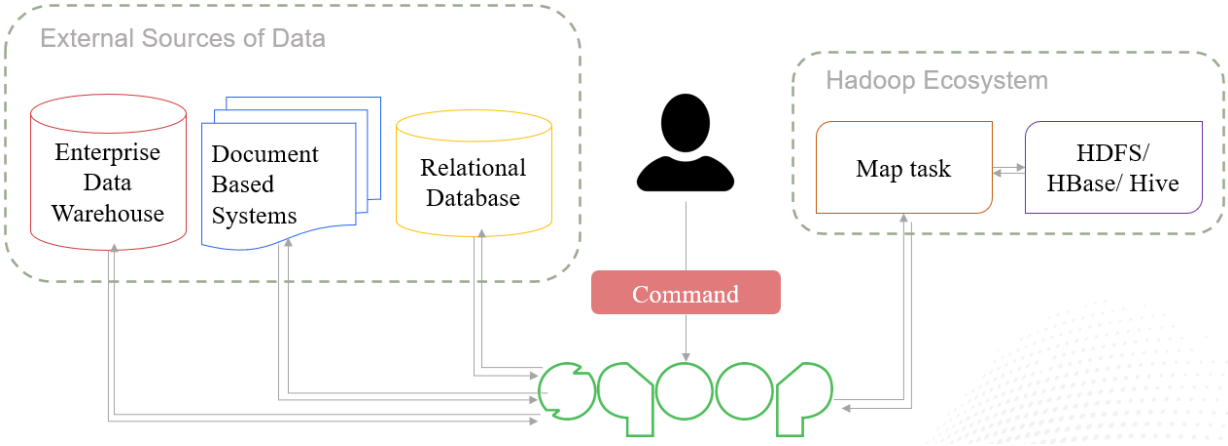
## 2.1 Kiến trúc

Về cơ bản, Apache Sqoop thực hiện các hoạt động chuyển dữ liệu giữa các hệ thống lưu trữ dữ liệu phân tán và cơ sở dữ liệu truyền thống một cách thân thiện với người dùng. Sqoop cung cấp một giao diện dòng lệnh cho người dùng để xác định các tác vụ nhập và xuất dữ liệu. Bên cạnh đó, Sqoop cũng hỗ trợ tương tác qua API Java để tương tác với người dùng.

Cụ thể hơn là:

- *Phân tích Argument:* Sqoop bắt đầu bằng việc phân tích các đối số do người dùng cung cấp trong giao diện dòng lệnh. Các đối số này được gửi đến bước xử lý tiếp theo.
- *Map Task:* Các đối số được chuyển cho công việc chỉ dành cho Map, và Sqoop sẽ tạo ra nhiều Map Task tùy thuộc vào số lượng mappers được xác định bởi người dùng. Các công việc này được phân chia để xử lý các phần dữ liệu tương ứng. Sqoop sử dụng xử lý song song để tối ưu hóa hiệu suất.
- *Kết nối Cơ sở dữ liệu:* Mỗi Map Task sẽ tạo một kết nối riêng với cơ sở dữ liệu bằng cách sử dụng mô hình kết nối cơ sở dữ liệu Java. Nó sau đó truy vấn cơ sở dữ liệu để trích xuất các phần dữ liệu cần thiết.
- *Truy xuất và Chuyển Dữ liệu:* Các Map Task tiến hành truy xuất và chuyển dữ liệu từ cơ sở dữ liệu nguồn vào hệ thống lưu trữ dữ liệu phân tán (HDFS, HBase, Hive) theo định dạng được xác định trong đối số dòng lệnh.
- *Quá trình Xuất Dữ liệu (Tùy chọn):* Nếu hoạt động là xuất dữ liệu, Sqoop cho phép tập hợp các tệp từ hệ thống phân tán Hadoop và xuất chúng đến cơ sở dữ liệu đích. Các tệp này thường là các bản ghi được cung cấp trong đầu vào của công việc xuất.

→ Sqoop đảm bảo quá trình nhập và xuất dữ liệu giữa các hệ thống một cách hiệu quả và tự động. Nó cung cấp sự linh hoạt cho người dùng để cấu hình các tác vụ chuyển dữ liệu và làm cho quá trình này trở nên thân thiện và tiện lợi.




## 2.2 Cách thức hoạt động

### 2.2.1 Nhập Dữ Liệu với Sqoop

Hoạt động nhập dữ liệu với Apache Sqoop cho phép người dùng chuyển dữ liệu từ hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS) vào hệ thống lưu trữ dữ liệu phân tán như Hadoop. Dưới đây là một số chi tiết về hoạt động này:

- *Trích xuất từ RDBMS:* Sqoop sử dụng lệnh nhập để trích xuất dữ liệu từ bảng hoặc cơ sở dữ liệu RDBMS. Người dùng chỉ cần xác định nguồn dữ liệu, cơ sở dữ liệu nguồn, và các thông tin kết nối liên quan.
- *Lưu trữ dữ liệu trong Hadoop:* Dữ liệu trích xuất từ RDBMS được lưu trữ trong Hadoop HDFS hoặc các hệ thống lưu trữ dữ liệu phân tán khác. Mỗi bản ghi từ RDBMS thường được lưu trữ trong một tệp văn bản riêng biệt.

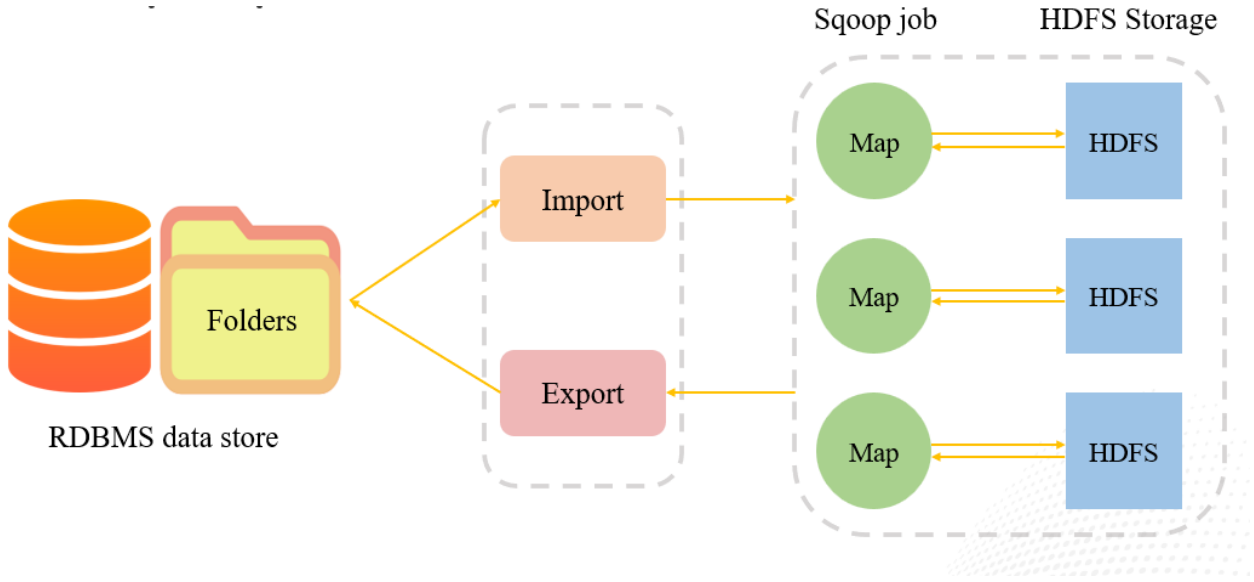
- *Tự động phân vùng:* Sqoop hỗ trợ tự động phân vùng dữ liệu để phân chia và quản lý dữ liệu trên các nút Hadoop. Điều này giúp tối ưu hóa hiệu suất và phân tải dữ liệu.
- *Hỗ trợ nhập dữ liệu tăng dần:* Sqoop cung cấp khả năng nhập dữ liệu tăng dần. Điều này có nghĩa là bạn có thể thêm dữ liệu mới vào Hadoop mà không cần phải nhập lại toàn bộ cơ sở dữ liệu. Sqoop sẽ chỉ nhập các hàng mới được thêm vào từ lần cuối cùng.

### 2.2.2 Xuất Dữ Liệu với Sqoop

Hoạt động xuất dữ liệu với Apache Sqoop cho phép chuyển dữ liệu từ hệ thống lưu trữ dữ liệu phân tán như Hadoop vào hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS). Dưới đây là một số chi tiết về hoạt động này:

- *Kiểm tra cơ sở dữ liệu đích:* Trước khi xuất dữ liệu, Sqoop thực hiện bước kiểm tra để xác minh cơ sở dữ liệu đích. Sqoop sẽ kiểm tra cơ sở dữ liệu và bảng đích để xác định cấu trúc dữ liệu.
- *Xử lý dữ liệu:* Dữ liệu được chuyển đổi thành hồ sơ (record) trước khi được xuất. Sqoop xử lý dữ liệu để đảm bảo đúng định dạng và cấu trúc cho cơ sở dữ liệu đích.
- *Di chuyển dữ liệu:* Dữ liệu đã được xử lý sẽ được di chuyển từ Hadoop hoặc hệ thống lưu trữ dữ liệu phân tán đến cơ sở dữ liệu RDBMS đích. Sqoop sẽ tạo và thực thi các câu truy vấn để chuyển dữ liệu.

→ Hoạt động nhập và xuất với Sqoop giúp người dùng tích hợp dữ liệu giữa các môi trường dữ liệu khác nhau một cách dễ dàng và hiệu quả.


Sqoop job
HDFS Storage
Map
HDFS
Map
HDFS
Map
HDFS
Import
Export
Folders
RDBMS data store

# CHƯƠNG 3: ƯU ĐIỂM VÀ NHƯỢC ĐIỂM

## 3.1 Ưu điểm

*Hỗ trợ nhiều nguồn dữ liệu:* Sqoop cho phép chuyển dữ liệu từ nhiều nguồn dữ liệu có cấu trúc khác nhau như các hệ quản trị cơ sỡ dữ liệu quan hệ (RDBMS) như Oracle, MySQL, PostgreSQL.

*Tối ưu hóa chi phí:* Do dữ liệu được chuyển và lưu trữ trong Hadoop, Sqoop hỗ trợ thực hiện các hoạt động ETL (Extract, Transform, Load) một cách nhanh chóng và hiệu quả về chi phí.

*Chuyển dữ liệu song song:* Sqoop có khả năng thực hiện chuyển dữ liệu song song, giúp tăng tốc độ chuyển dữ liệu và cải thiện hiệu suất tổng thể.

*Làm phong phú hệ sinh thái:* Apache Sqoop tích hợp tốt với nhiều dự án khác trong hệ sinh thái Apache như Apache Hive (cho việc truy vấn dữ liệu), Apache HBase (cho lưu trữ dữ liệu cột gia đình), và Apache Flume (cho việc thu thập dữ liệu). Điều này tạo ra khả năng tổ hợp và xử lý dữ liệu mạnh mẽ.

*Tạo ra các quy trình xử lý dữ liệu phân tán:* Bằng cách sử dụng Apache Sqoop để chuyển dữ liệu vào các hệ thống xử lý dữ liệu phân tán như Apache Hadoop và Apache Spark, chúng ta có thể tạo ra các quy trình xử lý dữ liệu phân tán. Điều này có nghĩa là dữ liệu được phân phối và xử lý trên nhiều nút máy tính cùng một lúc, giúp tận dụng sức mạnh tính toán của các hệ thống phân phối.

## 3.2 Nhược điểm

*Khó khăn trong việc xử lý lỗi:* Khi có lỗi xảy ra trong quá trình chuyển dữ liệu, Sqoop không cung cấp các công cụ mạnh mẽ để giúp bạn xác định và sửa lỗi một cách dễ dàng. Điều này có thể đòi hỏi sự can thiệp thủ công và làm cho việc sửa lỗi trở nên phức tạp, đặc biệt khi xử lý dữ liệu lớn.

*Thiếu tính năng quản lý công việc tự động:* Sqoop không cung cấp các công cụ mạnh mẽ cho việc lên lịch, theo dõi và quản lý các công việc chuyển dữ liệu một cách tự động. Điều này đồng nghĩa rằng bạn cần phải thực hiện nhiều công việc quản lý thủ công, như tạo lịch chạy, giám sát tiến trình và xử lý lỗi khi cần thiết.

## 3.3 Đặc điểm nổi bật của Apache Sqoop so với các sản phẩm cùng loại

Apache Sqoop, Kafka Connect, Apache Flume là các công cụ có khả năng chuyển dữ liệu, nhưng chúng hoạt động trong các ngữ cảnh và mục tiêu khác nhau. Dưới đây là một số khía cạnh so sánh giữa các công cụ này, cùng với một số đặc điểm làm nổi bật Apache Sqoop với hai công cụ còn lại.

### 3.3.1 Kafka Connect

- *Mục tiêu chính:* Kafka Connect được sử dụng để truyền tải dữ liệu thời gian thực giữa các hệ thống, đặc biệt là trong ngữ cảnh của xử lý dữ liệu dòng (streaming data).
- *Kiến trúc:* Kafka Connect được tích hợp trực tiếp vào hệ thống Kafka và sử dụng mô hình plugin để kết nối với các nguồn và đích khác nhau. Nó thường hoạt động trong thời gian thực và có khả năng xử lý dữ liệu liên tục.

❖ **Đặc điểm nổi bậc của Apache Sqoop so với Kafka Connect**

- *Khả năng chuyển dữ liệu:* Apache Sqoop là một công cụ mạnh mẽ cho việc chuyển dữ liệu từ và đến cơ sở dữ liệu truyền thống (SQL-based) và Hadoop. Nó cung cấp tích hợp sẵn để làm việc với các cơ sở dữ liệu phổ biến như MySQL, Oracle, PostgreSQL, và nhiều hệ thống SQL khác.

- *Tối ưu hóa cho việc sao chép lớn:* Sqoop được tối ưu hóa cho việc sao chép dữ liệu lớn. Nó có thể xử lý việc sao chép hàng triệu dòng dữ liệu một cách hiệu quả.

### 3.3.2 Apache Flume

- *Mục tiêu chính:* Apache Flume được thiết kế để thu thập, chuyển và đẩy dữ liệu từ nhiều nguồn khác nhau đến các hệ thống lưu trữ khác nhau. Nó chủ yếu là một công cụ cho xử lý dữ liệu thời gian thực.
- *Dữ liệu thời gian thực:* Flume là lựa chọn tốt khi bạn cần xử lý dữ liệu thời gian thực, như dữ liệu log, sự kiện, hoặc luồng dữ liệu không có cấu trúc từ nhiều nguồn đến một hệ thống lưu trữ (ví dụ: Kafka hoặc HDFS).
- *Luồng dữ liệu liên tục:* Flume chủ yếu là một công cụ xử lý luồng dữ liệu liên tục và cho phép bạn thực hiện xử lý dữ liệu trong thời gian thực khi nó chuyển đến hệ thống đích.

❖ **Đặc điểm nổi bậc của Apache Sqoop so với Apache Flume**

- *Hỗ trợ dữ liệu có cấu trúc:* Sqoop là lựa chọn hàng đầu khi cần chuyển dữ liệu có cấu trúc từ cơ sở dữ liệu SQL vào Hadoop hoặc ngược lại. Nó có tích hợp tốt với cơ sở dữ liệu quan hệ và được sử dụng rộng rãi trong các nhiệm vụ ETL (Extract, Transform, Load).
- *Batch Processing:* Sqoop phù hợp cho việc chuyển dữ liệu trong các quá trình xử lý lô, nơi dữ liệu được chuyển đổi và xử lý trong các lô dữ liệu lớn.

Tóm lại, Sqoop là một công cụ mạnh mẽ cho việc chuyển dữ liệu có cấu trúc giữa cơ sở dữ liệu SQL và Hadoop trong các quá trình xử lý lô, trong

khi Flume và Kafka thích hợp cho xử lý dữ liệu thời gian thực và luồng dữ liệu liên tục, với Kafka Connect làm một phần quan trọng trong hệ thống Kafka. Lựa chọn giữa chúng phụ thuộc vào nhu cầu cụ thể của dự án.

# CHƯƠNG 4: HƯỚNG DẪN CÀI ĐẶT VÀ CẤU HÌNH

## 4.1 Cài đặt Apache Sqoop

### 4.1.1 Tải Apache Sqoop

Môi trường mà nhóm đã cài đặt:

- Sqoop chạy trên môi trường Linux, khuyên dùng Ubuntu 20.04 trở lên.
- Java Development Kit (Phiên bản JDK 11.0.20).
- Hadoop (Phiên bản 3.3.6).

Để kiểm tra phiên bản Java bạn đã cài đặt, hãy chạy lệnh sau.

```
java -version
```

Tiếp theo tải xuống *phiên bản mới nhất* của Sqoop đồng thời giải nén nó.

### 4.1.2 Cấu hình Apache Sqoop

Vào file *.bashrc* và cấu hình các biến môi trường.

```
sudo gedit ~/.bashrc
```

Trong file *.bashrc*:

```
export HADOOP_HOME=/home/vp/hadoop-3.3.6

export HADOOP_INSTALL=$HADOOP_HOME

export HADOOP_MAPRED_HOME=$HADOOP_HOME

export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME
```

```
export YARN_HOME=$HADOOP_HOME
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS"-
Djava.library.path=$HADOOP_HOME/lib/nativ"


export SQOOP_HOME=/home/vp/sqoop-1.4.7.bin__hadoop-
2.6.0
export PATH=$PATH:$SQOOP_HOME/bin
```

Cập nhật và áp dụng các thay đổi cấu hình.

```
source ~/.bashrc
```

Vào thư mục */sqoop-1.4.7.bin__hadoop-2.6.0/conf* và cấu hình đường dẫn hadoop.

```
cd $SQOOP_HOME/conf
mv sqoop-env-template.sh sqoop-env.sh
sudo gedit sqoop-env.sh
```

Trong file *sqoop-env.sh*:

```
export HADOOP_COMMON_HOME=/home/vp/hadoop-3.3.6
export HADOOP_MAPRED_HOME=/home/vp/hadoop-3.3.6
```

Lưu file đã chỉnh sửa và kiểm tra phiên bản Apache Sqoop đã cài đặt.

```
sqoop version
```

## 4.2 Chuyển dữ liệu giữa MySQL và Apache Hadoop

Tải Driver JDBC MySQL (Phiên bản 8.0.21) *tại đây*

```
wget https://repo1.maven.org/maven2/mysql/mysql-
connector-java/8.0.21/mysql-connector-java-8.0.21.jar
```

Giải nén folder vừa tải về:

```
unzip mysql-connector-java-8.0.21.jar
```

Chuyển file *connector-java-8.0.21.jar* từ thư mục *mysql-connector-java-8.0.21* vào thư mục */$SQOOP/lib*

```
mv mysql-connector-java-8.0.21/mysql-connector-java-
8.0.21.jar /$SQOOP_HOME/lib
```

Tải các thư viện và tài nguyên cần thiết của Apache Sqoop để có thể kết nối và chuyển dữ liệu từ MySQL qua Hadoop *tại đây*

```
wget https://talend-
update.talend.com/nexus/content/repositories/libraries/
org/apache/sqoop/sqoop/1.4.7/sqoop-1.4.7.jar
```

### 4.2.1 Cài đặt MySQL

Thực hiện cài đặt theo câu lệnh

```
sudo apt install mysql-server
```

Cấu hình MySQL: Để cài lại mật khẩu root, thực hiện câu lệnh sau trong MySQL:

```
ALTER USER 'root'@'localhost' IDENTIFIED WITH
mysql_native_password BY '123456Vy!';
```

Cấu hình bảo mật cho MySQL:

```
sudo mysql_secure_installation
```

Set time zone:

```
SET GLOBAL time_zone = '+3:00';
```

Phân quyền cho *'root'@'localhost'*:

```
GRANT ALL PRIVILEGES ON *.* TO 'root'@'localhost' WITH
GRANT OPTION;

FLUSH PRIVILEGES;
```

### 4.2.2 Tạo Database trong MySQL

```
--Tạo database BANHANG:

CREATE DATABASE BANHANG;


-- Sử dụng database BANHANG

USE BANHANG;


-- Tạo bảng Customers

CREATE TABLE Customers (

    CustomerID INT PRIMARY KEY,

    CustomerName VARCHAR(255),

    ContactName VARCHAR(255),

    Country VARCHAR(255)
```

```
);
-- Tạo bảng Products
CREATE TABLE Products (
    ProductID INT PRIMARY KEY,
    ProductName VARCHAR(255),
    UnitPrice DECIMAL(10, 2),
    StockQuantity INT
);

-- Tạo bảng Orders
CREATE TABLE Orders (
    OrderID INT PRIMARY KEY,
    CustomerID INT,
    OrderDate DATE,
    TotalAmount DECIMAL(10, 2),
    FOREIGN KEY (CustomerID) REFERENCES
Customers(CustomerID)
);

-- Tạo bảng DetailOrders
CREATE TABLE DetailOrders (
    DetailOrderID INT PRIMARY KEY,
```

```
    OrderID INT,
    ProductID INT,
    Quantity INT,
    UnitPrice DECIMAL(10, 2),
    FOREIGN KEY (OrderID) REFERENCES Orders(OrderID),
    FOREIGN KEY (ProductID) REFERENCES
Products(ProductID)
);

-- Thêm dữ liệu vào bảng Customers
INSERT INTO Customers (CustomerID, CustomerName,
ContactName, Country)
VALUES
    (1, 'Customer 1', 'Contact 1', 'Country A'),
    (2, 'Customer 2', 'Contact 2', 'Country B'),
    (3, 'Customer 3', 'Contact 3', 'Country A');

-- Thêm dữ liệu vào bảng Products
INSERT INTO Products (ProductID, ProductName,
UnitPrice, StockQuantity)
VALUES
    (101, 'Product A', 10.99, 100),
```

```
    (102, 'Product B', 19.99, 50),
    (103, 'Product C', 5.49, 200);
```

-- Thêm dữ liệu vào bảng Orders

```
INSERT INTO Orders (OrderID, CustomerID, OrderDate,
TotalAmount)
VALUES
    (1001, 1, '2023-10-01', 50.00),
    (1002, 2, '2023-10-02', 75.50),
    (1003, 3, '2023-10-03', 30.25);
```

-- Thêm dữ liệu vào bảng DetailOrders

```
INSERT INTO DetailOrders (DetailOrderID, OrderID,
ProductID, Quantity, UnitPrice)
VALUES
    (2001, 1001, 101, 5, 10.99),
    (2002, 1001, 102, 3, 19.99),
    (2003, 1002, 102, 2, 19.99),
    (2004, 1002, 103, 4, 5.49),
    (2005, 1003, 101, 10, 10.99),
    (2006, 1003, 103, 2, 5.49);
```

### 4.2.3 Import và Export dữ liệu giữa HDFS và MySQL.

Trước đó, tiến hành thêm thư viện *commons-lang-2* vào Hadoop và Sqoop. Tải thư viện tại (file jar): *tại đây*

Sau đó tiến hành copy vào *hadoop-3.3.6/* và *sqoop-1.4.7.bin__hadoop-2.6.0/lib*

❖ **Import data từ HDFS sang MySQL.**

Import data from a table to HDFS (all rows and columns)

```
sqoop import --connect  jdbc:mysql://localhost/BANHANG
--username root --password 123456Vy! --table Customers
--m 1;
```

Import data from a table to HDFS (all rows but spescific colums)

```
sqoop import --connect  jdbc:mysql://localhost/BANHANG
--username root --password 123456Vy! --table Products -
-columns "ProductID","ProductName","UnitPrice" --m 1;
```

Import data from a table to HDFS (all columns, filter rows)

```
sqoop import --connect  jdbc:mysql://localhost/BANHANG
--username root --password 123456Vy! --table Customers
--where "Country='Country A'" --m 1;
```

Import data from a table to HDFS (specific columns, filter rows)

```
sqoop import --connect  jdbc:mysql://localhost/BANHANG
--username root --password 123456Vy! --table Customers
--columns "CustomerID","CustomerName" --where
"Country='Country A'" --m 1;
```

### ❖ Export data từ MySQL sang HDFS.

--Tạo bảng Customers_export:

```
CREATE TABLE Customers_export (
    CustomerID INT PRIMARY KEY,
    CustomerName VARCHAR(255),
    ContactName VARCHAR(255),
    Country VARCHAR(255)
);
```

--1. export data from HDFS to RDBMS (all rows and columns)

```
sqoop export --connect jdbc:mysql://localhost:/BANHANG
--username root --password 123456Vy! --table
Customers_export --export-dir /user/vy/Customers -m 1;
```

--2. export data from HDFS to RDBMS (all rows but spescific colums)

```
sqoop export --connect jdbc:mysql://localhost:/BANHANG
--username root --password 123456Vy! --table
Customers_export --export-dir /user/vy/Customers --
columns "CustomerID","CustomerName" -m 1;
```

# NGUỒN THAM KHẢO

[1] M. Baddeley, "Installing and Configuring Sqoop 1.4.7 to run on Hadoop 3.X," [Online]. Available: https://www.youtube.com/watch?v=L7TqAqJyCJ0&t=42s&ab_channel=MarshaBaddeley. [Accessed 7 11 2022].

[2] R. Digital, "Apache Sqoop Tutorial for Beginners | Data loading from RDBMS To HDFS | Hadoop Tutorial," [Online]. Available: https://www.youtube.com/watch?v=V1YowbzzqAo. [Accessed 2 12 2022].

[3] manasmohapatra, "Overview of SQOOP in Hadoop," [Online]. Available: https://www.geeksforgeeks.org/overview-of-sqoop-in-hadoop/. [Accessed 19 8 2021].

[4] Simplilearn, "Sqoop Hadoop Tutorial | Apache Sqoop Tutorial | Sqoop Import Data From MySQL to HDFS," [Online]. Available: https://www.youtube.com/watch?v=Lo1MoNKE-l8&ab_channel=Simplilearn. [Accessed 17 4 2019].