SQL in Big Data Technologies

Student's Name

Institutional Affiliation

Course Code

Instructor

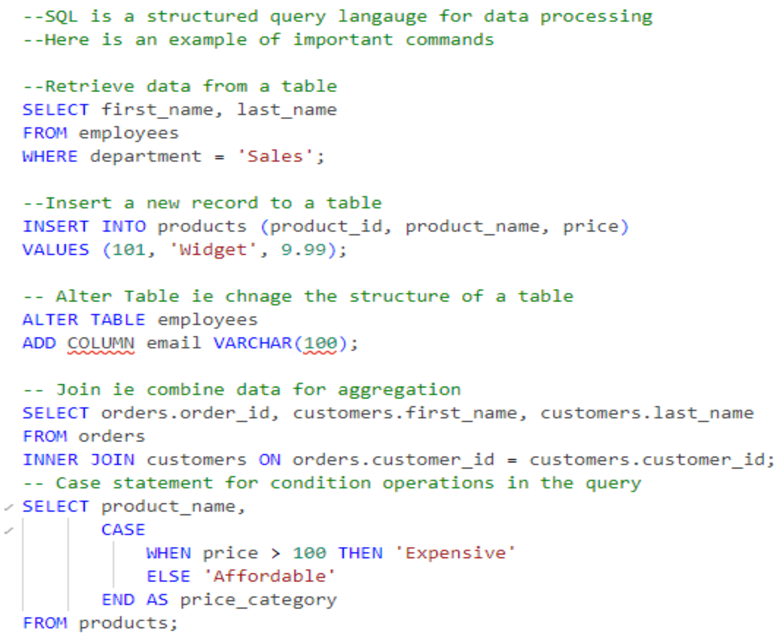
Date

SQL in Big Data Technologies

Introduction

Traditional relational database management systems like Oracle, MySQL, and PostgreSQL have long been used for data manipulation, but with the rise in the adoption of internet technology in the operations of businesses and organizations, there has been a rapid growth of data in terms of the volume and veracity of databases. This data encompasses a wide range of information including sales, finances, health data, social media, web information and much more. The current technology demands dynamic tools that have the capacity to handle unstructured data, are easy to use, and which can store, retrieve and manipulate large volumes of databases. In this article, we explore the new SQl tools in big data management through sampled SQL functions.

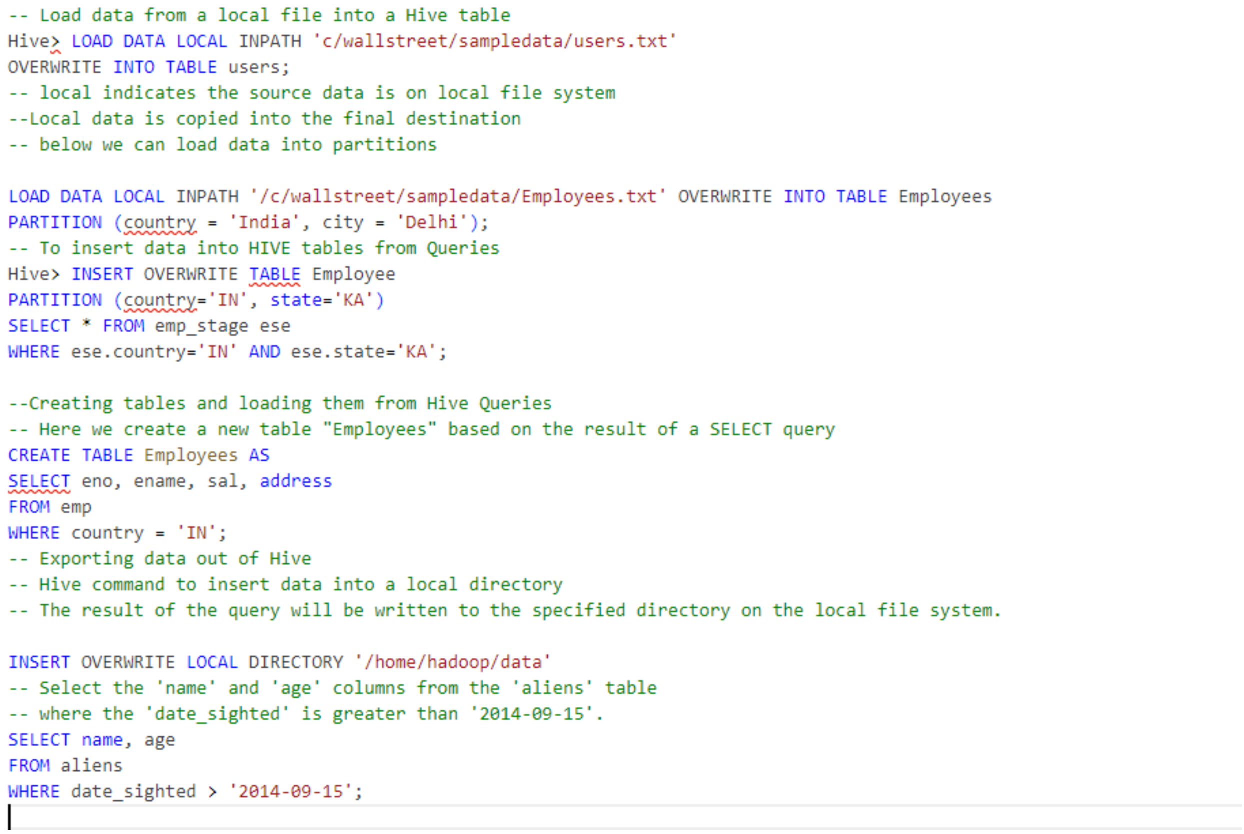
SQL (structured query language) has long been used for relational databases and exists in tabular form. Table 1 below shows the basic commands used to process data (SQL, 2023).



```
--SQL is a structured query langauge for data processing
--Here is an example of important commands

--Retrieve data from a table
SELECT first_name, last_name
FROM employees
WHERE department = 'Sales';

--Insert a new record to a table
INSERT INTO products (product_id, product_name, price)
VALUES (101, 'Widget', 9.99);

-- Alter Table ie chnage the structure of a table
ALTER TABLE employees
ADD COLUMN email VARCHAR(100);

-- Join ie combine data for aggregation
SELECT orders.order_id, customers.first_name, customers.last_name
FROM orders
INNER JOIN customers ON orders.customer_id = customers.customer_id;
-- Case statement for condition operations in the query
SELECT product_name,
       CASE
           WHEN price > 100 THEN 'Expensive'
           ELSE 'Affordable'
       END AS price_category
FROM products;
```

Table 1: Sample SQL commands, Source (Author)

In Big Data technologies, such as Hadoop, an SQL-like query language known as HIVEQL is used for processing and handling data that is stored in the Hadoop Distributed File Systems (HDFS). For data manipulation in HIVEQL, unlike in typical SQL, the row-level insert, update, and delete commands are not allowed. Some of the commands involved in HIVEQL include those demonstrated in Table 2 below.

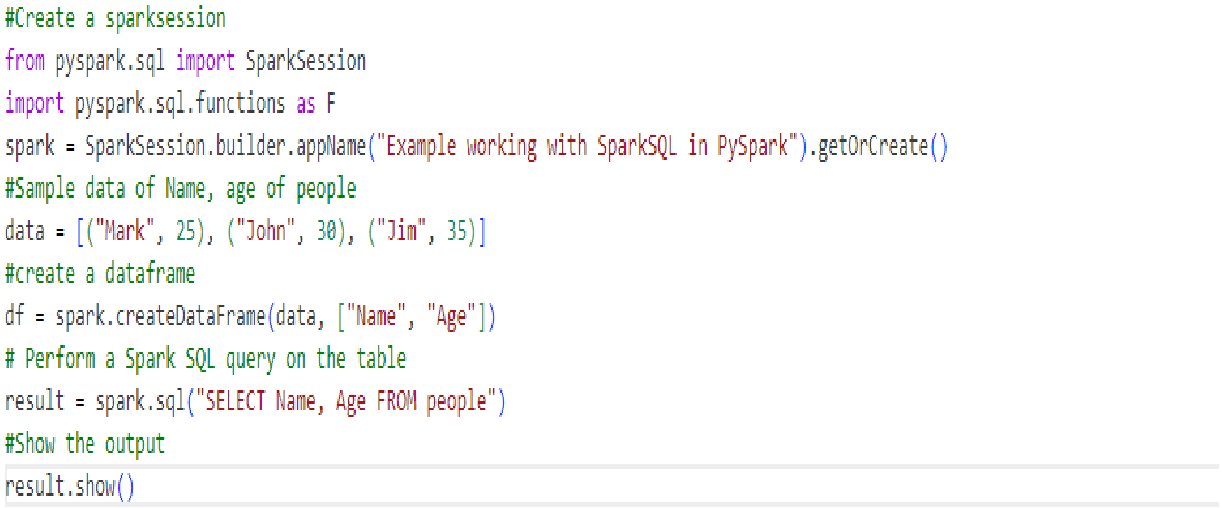


```
-- Load data from a local file into a Hive table
Hive> LOAD DATA LOCAL INPATH 'c/wallstreet/sampledata/users.txt'
OVERWRITE INTO TABLE users;
-- local indicates the source data is on local file system
--Local data is copied into the final destination
-- below we can load data into partitions

LOAD DATA LOCAL INPATH '/c/wallstreet/sampledata/Employees.txt' OVERWRITE INTO TABLE Employees
PARTITION (country = 'India', city = 'Delhi');
-- To insert data into HIVE tables from Queries
Hive> INSERT OVERWRITE TABLE Employee
PARTITION (country='IN', state='KA')
SELECT * FROM emp_stage ese
WHERE ese.country='IN' AND ese.state='KA';

--Creating tables and loading them from Hive Queries
-- Here we create a new table "Employees" based on the result of a SELECT query
CREATE TABLE Employees AS
SELECT eno, ename, sal, address
FROM emp
WHERE country = 'IN';
-- Exporting data out of Hive
-- Hive command to insert data into a local directory
-- The result of the query will be written to the specified directory on the local file system.

INSERT OVERWRITE LOCAL DIRECTORY '/home/hadoop/data'
-- Select the 'name' and 'age' columns from the 'aliens' table
-- where the 'date_sighted' is greater than '2014-09-15'.
SELECT name, age
FROM aliens
WHERE date_sighted > '2014-09-15';
```

Table 2: Data manipulation using HIVEQL, Source (Author)

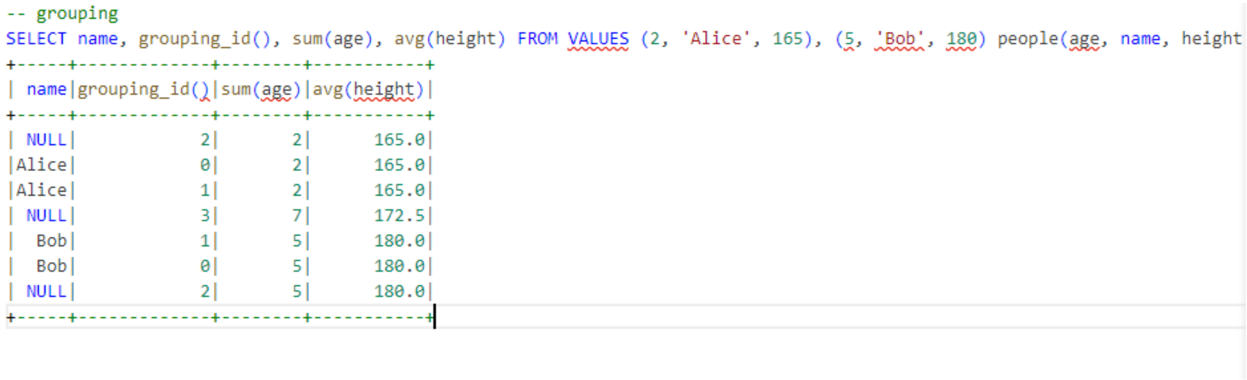
Apache Spark SQL is module-based in Apache Spark that uses both a programing interface and SQL query tool for processing stored data. It allows the querying of data using SQL syntax and acts as a distributed SQL query tool with the capacity to run up to 100x faster on deployments in big data. Spark SQL is a big data technology with a simple interface that makes it easier for users to work with large volumes of datasets, allowing querying just like in SQL. It supports a wide range of data sources such as Apache Hive, Avro, and JSON etc., making it among the best technologies to work with when dealing with a variety of data types (Spark, 2023).

A demonstration of the syntax in Spark SQL, with an example:



```
#Create a sparksession
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
spark = SparkSession.builder.appName("Example working with SparkSQL in PySpark").getOrCreate()
#Sample data of Name, age of people
data = [("Mark", 25), ("John", 30), ("Jim", 35)]
#create a dataframe
df = spark.createDataFrame(data, ["Name", "Age"])
# Perform a Spark SQL query on the table
result = spark.sql("SELECT Name, Age FROM people")
#Show the output
result.show()
```

Table 3: Data manipulation using SparkSQL, Source (Author)



```
-- grouping
SELECT name, grouping_id(), sum(age), avg(height) FROM VALUES (2, 'Alice', 165), (5, 'Bob', 180) people(age, name, height
+-----+-------------+--------+-----------+
| name|grouping_id()|sum(age)|avg(height)|
+-----+-------------+--------+-----------+
| NULL|            2|       2|      165.0|
|Alice|            0|       2|      165.0|
|Alice|            1|       2|      165.0|
| NULL|            3|       7|      172.5|
|  Bob|            1|       5|      180.0|
|  Bob|            0|       5|      180.0|
| NULL|            2|       5|      180.0|
+-----+-------------+--------+-----------+
```

Table 4: Grouping data using SparkSQL, Source (Author)

In conclusion, the world of Big Data tech is rapidly evolving. This means new ways of handling and managing data using structured query techniques. As demonstrated using the HiveQL and Apache Spark SQL technologies above, SQL-like commands continue to play a vital role in the utilization of tools for data management and analysis, whether working with structured, unstructured, or even cloud-based databases, such as Amazon RedShift or Google BigQuerry. The continued usefulness of SQL syntax in preprocessing, loading, extracting, and performing other forms of data manipulation is important for the generation of insights from data.

References

Paul, S. (2018). SQL on Big Data, Technology, Architecture, and Innovation. In S. Paul.

Spark. (2023, November 1). *Spark SQL is Apache Spark's module for working with structured data.* Retrieved from https://spark.apache.org/sql/

Sridhar, K. T. (2019). Big Data Analytics Using SQL. *11th International Conference on Research and Practical Issues of Enterprise Information Systems*, 2.

SQL, M. (2023, Novemebr 1). *Introducing SQL Server 2022*. Retrieved from https://www.microsoft.com/en-us/sql-server