**ASSIGNMENT 29.3**

**Problem Statement:**

Explain Brief of the following in brief

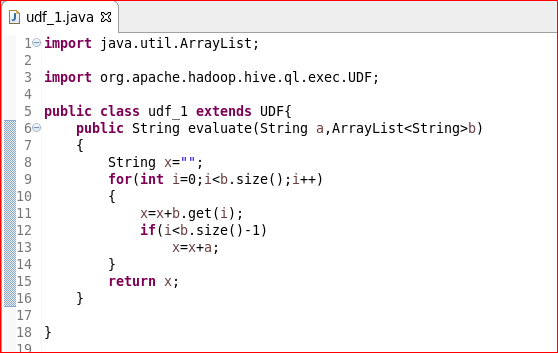
1. Hive UDF
2. Hive UDAF
3. Hive UDTF
4. Thrift server

**Solution:**

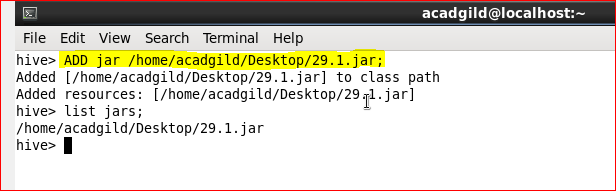
1. **UDF:**

UDF is a user-defined function that takes a single input value and produces a single output value. When used in a query, we can call it once for each row in the result set.

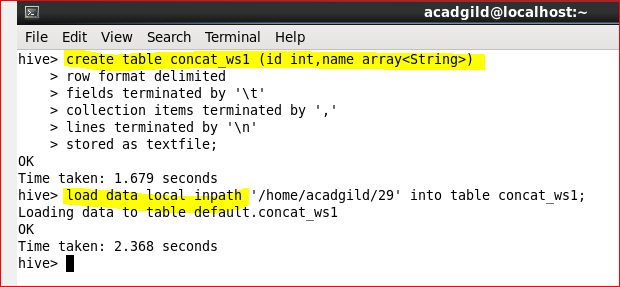
**UDF class:**

****

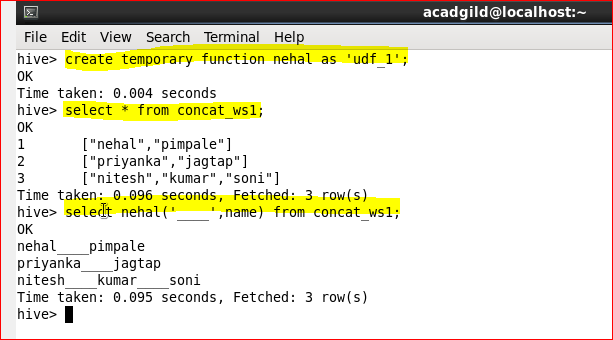
**Adding jar:**

****

**Creating table and loading data into the table:**

****

**Creating temporary function and using that UDF:**

****

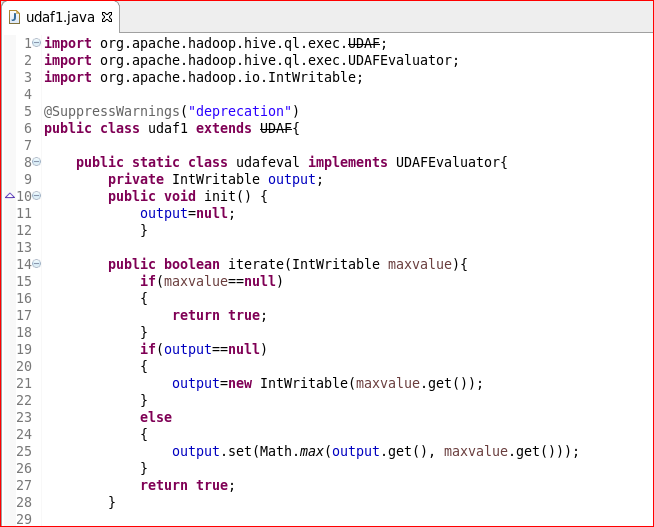
1. **Hive UDAF:**

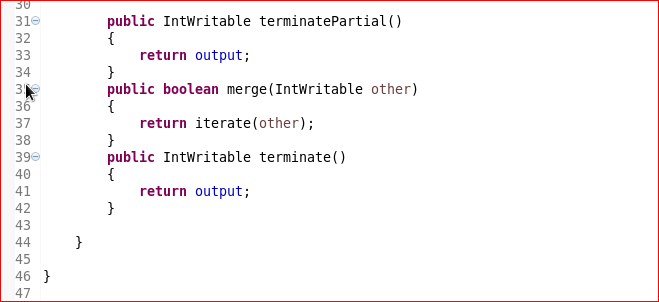
User-Defined Aggregation Functions (UDAFs) are an exceptional way to integrate advanced data-processing into Hive. Aggregate functions perform a calculation on a set of values and return a single value.

An aggregate function is more difficult to write than a regular UDF. Values are aggregated in chunks (potentially across many tasks), so the implementation has to be capable of combining partial aggregations into a final result.

We are writing UDAF to find the largest Integer from the input file.

The code to achieve this is explained in the below example, we need to make a jar file of the below source code and then use that jar file while executing hive scripts

****

****

**Let’s see now the steps for UDAF Execution.**

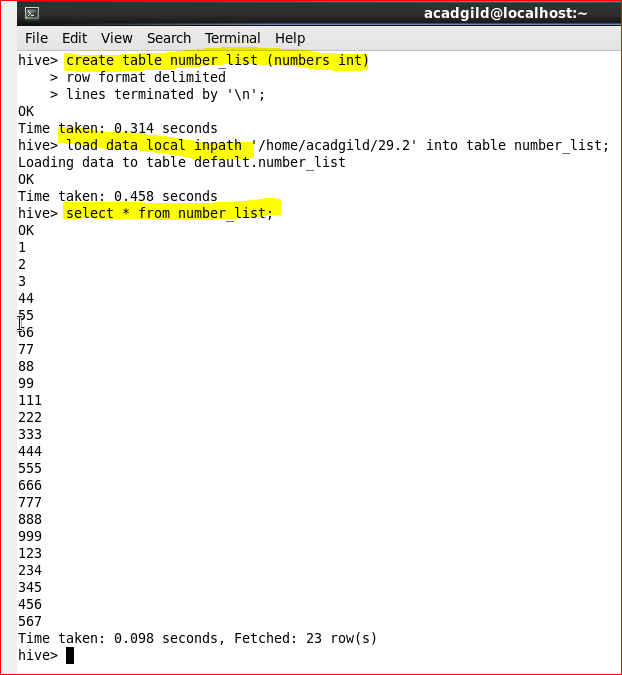
* **Creating a new Input Dataset:**

We need a input dataset to execute the above example. The Dataset that will be used for demonstration is Numbers\_List. It has one column, which contains List of Integer values.

* **Create a new table  and load the input dataset:**

In the below screenshot we have a created a new table Num\_list with only one field (column) Numbers.

Next, we have loaded the input dataset Numbers\_List contents into the table Num\_List.

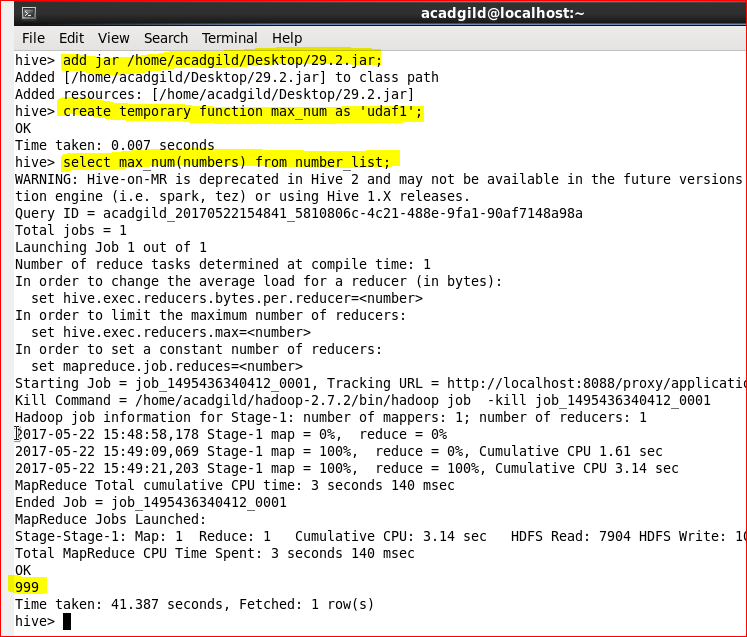
****

* **Add the Jar file in hive with complete path (Jar file made from source code need to be added)**
* **Create temporary function as shown below**

**Let us create a temporary function max for newly created UDAF.**

* **Use the select statement to find the largest number from the table Num\_List**

**After, successfully following the above steps we can see use the Select statement command to find the largest number in the table.**

****

1. **Hive UDTF:**

UDTF is a User Defined Table Generating Function that operates on a single row and produces multiple rows a table as output.

**Source Code:**

We can create a custom Hive UDTF by extending the GenericUDTF abstract class and then implementing the initialize, process, and possibly close methods.

**Initialize ():**

The Hive calls the initialize method to notify the UDTF the argument types to expect. The UDTF must then return an object inspector corresponding to the row objects that the UDTF will generate.

**Process ():**

Once initialize () method has been called, Hive will give rows to the UDTF using the process () method. While in process () function, the UDTF can produce and forward rows to other operators by calling forward () method.

**Close ():**

Finally, Hive will call the close () method when all the rows have passed to the UDTF. This function allows for any cleanup that is necessary before returning from the User Defined Table Generating Function. It is important to note that we cannot write any records from this function.

So far, from our above example, no data is required which needs to be cleaned up.

1. **Thrift server:**

As we know, Apache Hive is a data warehouse software that facilitates reading, writing and managing large data sets residing in distributed storage using SQL.

Let’s consider a scenario, where the user is looking forward to performing an operation on Hive server, and the Hadoop cluster or Hive software setup is not installed in his/her system. The solution for the above scenario is that the user can write codes in other languages and access Hive server using Apache Thrift interface.

Apache Thrift is a software framework for scalable cross-language services development, which combines a software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, Ruby, Perl, C#, JavaScript, Node.js and other languages.

Thrift can be used when developing a web service that uses a service developed in one language access that is in another language.

HiveServer is a service that allows a remote client to submit requests to Hive, using a variety of programming languages, and retrieve results. It is built on Apache Thrift, therefore it is sometimes called as the Thrift server.

In the context of Hive, Java language can be used to access Hive server. The Thrift interface acts as a bridge, allowing other languages to access Hive, using a Thrift server that interacts with the Java client.