Controlling script processes using threads and nodes

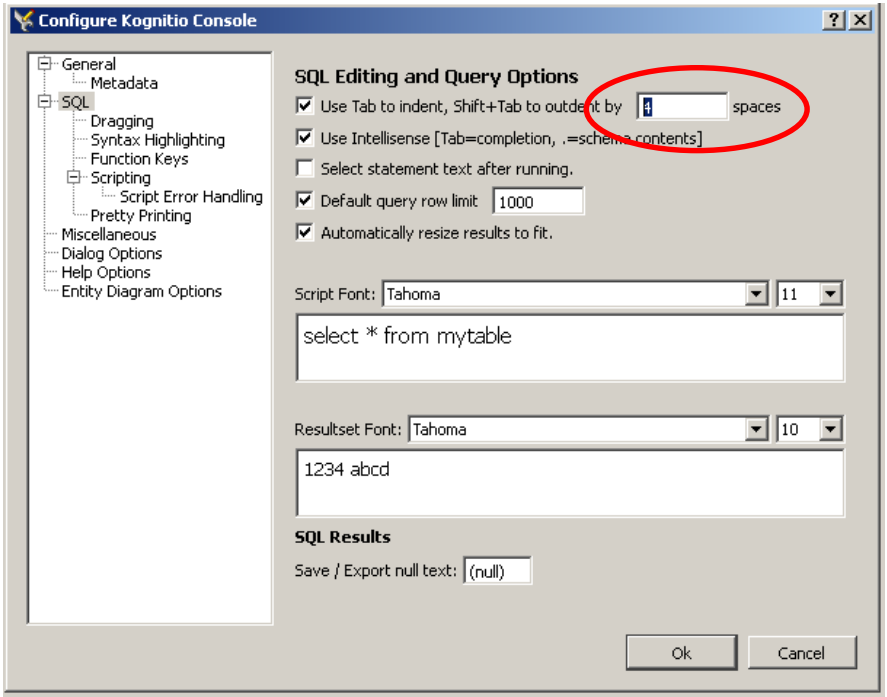
(Part 2 of External script technical training)

Click here to download the supporting SQL script

External Script Training Overview

This is the python version

1. Introduction to external script interface
2. Controlling Script Processes – threads and nodes
3. Controlling Data Processing – partitions
4. Passing parameters through to scripts



Python is strict about code indentation as it is used to define code within “for”, “if/else” statements etc.

When using Kognitio console set the tab indentation to 4 to match this requirement.

Tools->Configure->SQL

Kognitio

Controlling script processes:
threads and nodes

These examples illustrate functionality but are not the best way to approach getting the overall average, why?

```
create external script Kog_P_AvgOverAllRows
                    environment python
receives(Part_ID int, value float)
sends(result float)
limit 1 threads
script S'EOF(
… <identical python script> …
)EOF';
```

Limit the external script to 1 thread only. All data streams through a single invocation yielding 1 average over all the data

Parallelism is lost and the performance that comes from it

```
create external script Kog_P_AvgOverAllRows
                    environment python
receives(Part_ID int, value float)
sends(result float)
limit 1 threads per node
run on node ('nodename')
script S'EOF(
… <identical python script> …
)EOF';
```

The same single result is obtained using a combination of limiting threads per node and specifying which node to run the script on

As well as loosing parallelism the script becomes tied to the named node. Script is no longer portable between systems. Kognitio automatically manages where to execute processes so generally not required

Kognitio

Controlling script processes:
when to limit threads and nodes

```
create external script Kog_P_AvgOverAllRows
                    environment python
receives(Part_ID int, value float)
sends(result float)
limit 1 threads
script S'EOF(
… <identical python script> …
)EOF';
```

There are some processes that cannot be parallelised or broken down into parallel steps. Limiting 1 threads allows users to run these scripts on Kognitio

```
create external script Kog_P_AvgOverAllRows
                    environment python
receives(Part_ID int, value float)
sends(result float)
limit 1 threads per node
run on node '<nodename>'
script S'EOF(
… <identical python script> …
)EOF';
```

Limiting threads per node is very useful for processes requiring larger amounts of RAM. The parallelism is cut but not lost. The script invocations do not compete for RAM within the node. This is particularly useful if your processing makes use of multiple processors.

For use when the external executable is only installed on a subset of nodes. Due to licensing costs for example.

Kognitio

What is the best way to code avg in Kognitio?

Controlling script processes:
making use of parallelism

```
create external script Kog_P_SumOverAllRows
                       environment python
receives(Part_ID int, value float)
sends(sum1 float, count1 float)
script S'EOF(
import csv,sys
input = csv.reader(sys.stdin)
result = csv.writer(sys.stdout)
sum1=0
count1=0
for row in input:
    if row[0]!='':
        sum1=sum1+float(row[1])
        count1=count1+1
result.writerow([sum1,count1])
)EOF';

select sum(sum1)/sum(count1) Avg1
from ( external script Kog_P_SumOverAllRows
         from (select PRODNO, PRICE/100.00
               from DEMO_RET.V_RET_SALE
               where storeno between 1 and 5)
       ) sq1;
```

Design your scripts so that you make full use of the parallelism whenever possible. If you can code the problem in SQL it will generally be faster.

!This code emulates how Kognitio breaks down and parallelises the SQL avg function.

It is easy to utilise Kognitio to further process external script output; using SQL (like here) or passing it into another external script. This allows data to be streamed through a series of scripts to perform very complex tasks

As well as controlling the script processes in Kognitio it is easy to control the data flow. See part 3 for details

Kognitio