TEXAS ADVANCED COMPUTING CENTER

WWW.TACC.UTEXAS.EDU

TEXAS
The University of Texas at Austin

R and Python Package Collection and Filtering

Robert McLay

December 15, 2022

XALT: Outline

- XALT can record package usage by R and Python (optionally Matlab)
- This talk centers on Python with a little on others
- Special code is required to use allow for tracking

R tracking requires a special package to turn on tracking

- Package developed at IU by James McCombs and Michael Scott
- I need to get one of them for docs on how to install it.

Python tracking

- I developed XALT's xalt_sitecustomize.py (installed as sitecustomize.py)
- This was with help from Riccardo Murri (with an intro from Kenneth Hoste!)
- To track python imports add $XALT_DIR/site_packages to $PYTHONPATH
- This puts XALT's sitecustomize.py to ALWAYS read by every python execution

XALT package tracking interface

- All package tracking “hooks” calls the program xalt_record_pkg to interface with XALT
- It can be found in $XALT_DIR/libexec
- In Python <NAME> or <PATH> might be marked as the string ’<unknown>’.
- No tracking if both are unknown.

```
../xalt_record_pkg program          <PROGRAM> \
                   xalt_run_uuid    <UUID>    \
                   package_name     <NAME>     \
                   package_version  <VERSION>  \
                   package_path     <PATH>
```

How XALT handles package records

- Since XALT won't know if the PKG prog will be tracked.
- All PKG programs will produce records but!
- All pkg records written to /dev/shmem or /tmp
- Non sampled runs remove records at end
- Sampled are send on "wire" (i.e. saved to db) via your transport method
- This way pkg records are written on the "wire" only when necessary

Save every python/R package?

- Probably not
- There are two kinds of optional filtering
- Python only (part of xalt_sitecustomize.py)
- All package filtering, available when accepting records to write to DB

The block python_pkg_pattern

```
python_pkg_patterns = [
   'k_s' : 'SKIP', 'kind' : 'path', 'patt' : r"^[^/]"      ,
   'k_s' : 'SKIP', 'kind' : 'name', 'patt' : r"^_"         ,
]
```

- A full example can be found in Config/rtm_config.py

Here are some of the rules we use

```
path : r"^[^/]"       # SKIP all built-in packages
name : r"^_"          # SKIP names that start with a underscore
name : r".*\."        # SKIP all names that are divided with periods: a.b
path :  r".*/.local/" # KEEP all packages installed by users
```

- Since this is part of sitecustomize.py
- No records are even written to /dev/shmem

All Package Filtering

- Filtering can by XALTdb.py (when loading into DB).
- It uses a flex routine to filter both Python and R programs
- This block of code goes in your Config.py
- I should have just provided this version first but ...

```
pkg_patterns = [
  ["SKIP",  r'^R:name:stats'],          # SKIP the R pkg named stats
  ["SKIP",  r'^R:name:base'],           # SKIP the R pkg named base
  ["SKIP",  r'^R:name:methods'],        # SKIP the R pkg named methods
 ["SKIP", r'^python:name:_.*'],     # SKIP all python name that start with an underscore
  ["SKIP",  r'^python:path:[^/].*'],    # SKIP all python built-in packages
  ["SKIP",  r'^python:path:home'],    # SKIP all python package in user locations
]
```

Suggested Rules for using either or both filter systems

- Use the Python only filtering for a few rules
- As this filtering happens for every import
- Save complicated filtering for the total filtering system

Filtering opportunity

- Currently all filtering is one path or name at a time
- Since all pkgs for are a single job are available at end
- It could be possible to provide an array of records before transmitting
- I'm looking for feedback.

Future Topics?

- I'm looking for Topics.
- Next Meeting will be on Jan. 19, 2023 at 10:00 am U.S. Central (16:00 UTC)