FAQ

1 - MADlib General

Q1-1 What is MADlib?

- MADlib is an open source library for scalable in-database analytics. It provides data-parallel implementations of mathematical, statistical, graph and machine learning methods for structured and unstructured data.
- MADlib is a Top Level Project in the Apache Software Foundation.
- MADlib home page http://madlib.apache.org/

Q1-2 Which database platforms does MADlib support and what is the upgrade matrix? Also which operating systems?

- Please see this page for supported databases and OS

Q1-3 What are the main advantages of MADlib?

- Use MPP architecture's full compute power
- Use MPP architecture's entire memory to process large data sets, so don't need to sample
- Familiar SQL interface

Q1-4 Who uses MADlib?

- Data scientists mostly
- Wide range of verticals including financial services, healthcare, retail, energy, manufacturing and government

Q1-5 What are the benefits of MADlib compared with products like R and scikit-learn?

- Performance
  - MADlib is a fully parallelized implementation on GPBD for large data sets, so it is far more performant than R or Python libraries for large data.
- Scalability
  - Add more nodes to achieve higher performance as your data scales. R and Python libraries are limited by the amount of data you can load into memory on a single node.
  - Using all data, not a sample, can improve accuracy
- Familiar, user friendly SQL interface
- Ease of data preparation
  - Supports commonly used database data formats

Q1-6 Where can I find the details of how existing MADlib algorithms have been implemented?

- Please refer to the algorithm technical design document

Q1-7 What is the MADlib security model?

- MADlib security model is the same as that of the underlying database. It runs in-database and does not do file level operations.
- Regarding specific roles and privileges, MADlib users need object privileges in the schema that they are working in. This is because MADlib creates temporary tables and output tables as part of regular execution of MADlib functions.
- For Greenplum database, see Table 2 for a list of required object privileges: http://gpdb.docs.pivotal.io/latest/admin_guide/roles_privs.html .
- For PostgreSQL, see the GRANT commandL https://www.postgresql.org/docs/current/static/sql-grant.html .

2 - MADlib Usage

Q2-1 What are my options if MADlib does not have the algorithm that I need?

- In this case, you might think about building separate models for different chunks of a larger dataset (partition by state, range of user IDs, product category, etc.)
- This is referred to data parallelism: break up the problem into a number of parallel tasks, if you can ensure there is no dependency (or communication) between those parallel tasks.
- Then you can use a procedural languages such as PL/Python or PL/R on each chunk and combine the results downstream.
- And of course, you can build and contribute the module to MADlib for the benefit of the community

Q2-2 What are the differences in functionality between the GPDB and HDB/HAWQ versions of MADlib?

- There are very few differences and they are listed below.
- K-means clustering
  - Can specify a user defined distance function for GPDB & PostgreSQL. For HAWQ, does not support UDFs for distance, so restricted to the built-in distance functions provided.
- “Deprecated modules” quartile and profile have some minor differences and limitations between HAWQ and GPDB. See the documentation for details.

Q2-3 Can I export models from MADlib to PMML?

- Yes. MADlib models can be exported in PMML format for use in scoring by a PMML evaluator.
- The following MADlib algorithms can be exported in PMML format:
  - Linear regression
  - Logistic regression
  - GLM
  - Multinomial regression
  - Ordinal regression
  - Decision trees
  - Random forest

Q2-4 What is PMML?

- The Predictive Model Markup Language (PMML) is an XML-based file format that provides a way for applications to describe and exchange models produced by data mining and machine learning algorithms.
- For more information, please see http://www.dmg.org/

Q2-5 What is JPMML?

- JPMML is an open source PMML evaluator available under GPL license.
- For more information, please see https://github.com/jpmml and https://github.com/jpmml/openscoring

Q2-6 Can I import models from PMML to MADlib?

- Not currently. You can only export from MADlib into PMML as described above.

Q2-7 Can I call MADlib from Python?

- There is a nascent Python wrapper called PyMADlib. It is based on an older version of MADlib and supports a very limited number of algorithms.
- For more information, please see http://pivotalsoftware.github.io/pymadlib/ and https://github.com/pivotalsoftware/pymadlib
- We want to build a full Python API for MADlib, so let us know if you would like to participate in this effort.

Q2-8 What are some examples of connecting MADlib with 3rd party products?

- MADlib and Tableau http://things-about-r.tumblr.com/post/98652993554/deep-down-below-using-in-database-analytics-from
- MADlib and Knime https://drive.google.com/a/pivotal.io/file/d/0B3Pw2DP_4X47S3FXU0duRG84X2M/view?usp=sharing

3 - PivotalR

Q3-1 What is PivotalR?

- PivotalR is a package that enables users of R, the most popular open source statistical programming language and environment, to interact with the GPDB, HAWQ and the open source database PostgreSQL on large data sets. It does so by providing an interface to the operations on tables /views in the database.

Q3-2 Where can I learn more about PivotalR?

Please refer to this PivotalR wiki page.

4 - Other

Q4-1 Your question here

- Let us know if you have a question and we will add it here.