README for Fidelity
July 25, 2016

1. The following data sets can be used for performance testing of the MADLib logistic regression algorithm:

- higgs_1e7_1e2
- higgs_1e6_1e2
- higgs_1e8_1e2
- Higgs_1e7_1e1

For example, higgs_1e7_1e2 means: 1e7 (10M) rows and 1e2 (100) features.

2. In order to create the tables in the database, please do:

- gunzip -c higgs_1e6_1e2.sql.gz | psql

3. The query to run is logreg_perf.sql and is located on the same Google drive as the data sets.

4. The following results are for MADlib logistic regression on a Pivotal DCA v1 half-rack with GPDB 4.2.7.1 with 8 nodes and 6 segments per node:

GPDB	**#/rows**	**#/features**	**#/groups**	**Runtime (sec)**
Baseline	10M	100	1	70.6
#rows (small)	1M	100	1	11.2
#rows (large)	100M	100	1	704.9
#features (small)	10M	10	1	8.5

When these tests were run, there were no other workloads in the DCA.

Frank McQuillan
MADlib Product Manager
fmcquillan@pivotal.io