Quality Assurance of Big Data Security product using automated query generation through R

Client Overview
- The client is based out of San Francisco Bay area.
- The client is a leading provider of data-centric security for next-gen data platforms.
- The client keeps enterprises in control of their data by allowing them to give users access to the data they need.

Business Requirements
- The client wanted to test how their product fared in various Big Data tools environments viz. Hive, Impala, SparkQL, Cassandra, HAWQ etc.
- The client was looking for an out of the box solution for query generation which would cover all the use-cases and generate queries in an automated fashion.

Key Challenges
- The challenge was that there were multitude of use-cases (join, subqueries, case when etc.) which needed to be tested.
- This further led to millions of queries for each tool.
Our Approach

- Our team of Data Scientists and Big Data Engineers devised combinatorial algorithms inspired from apriori algorithm to automate the query generation process.
- The algorithm was modified to generate meaningful queries for all the Big Data tools environment in which client’s product worked.

Results

- Product QA became:
  - 10X faster
  - Replicable at each product deployment (client started shipping the algorithm code as well as sample query set to test the product deployment)
  - Comprehensive (covered all the use-cases)
  - Flexible with version changes (minimal user to generate new set of queries amenable to a new version)

Our Solutions

- The algorithm was written in R and packaged into modules to generate automated queries for all or one use-case.
- The algorithm could take any data model (set of tables) as input and generate meaningful queries (outputs non-trivial results) in an automated manner. The user just had to point the algorithm to the data model.