Essential Roadmap to Benchmarks/POCs For Executives

This is a high-level roadmap to executing a benchmark to drive business benefit at minimum cost.

Key Points:

1. Identify the key pain points of your existing environment or the key functionality that you want to achieve with the selection of a new platform. These form the basis of your success criteria. Some of these criteria should be measurable.
2. Select a subset of your data infrastructure that is representative of the environment you intend to have in the future. The more realistic a simulation of production that could occur, the better the benchmark will predict future capabilities.
3. Determine data ownership, confidentiality, and privacy issues for the selected data. Start working any legal issues early since this can take weeks or months to address. Sensitive data can be masked to eliminate concerns.
4. Choose a subset of tables and queries that represents the range of users, query complexity, and performance that the platform will need to support. Include some incremental updating in the tests since data warehouses aren't read-only. Avoid too large a scope because benchmark costs and duration go up for you and vendors with a larger scope.
5. Focus the testing on the data warehouse versus everything else by capturing the queries and processes executed by BI and ETL tools. These should be run by a query driver like TdBench. This eliminates variations introduced by shared servers and network. It also reduces license cost, potential delays and setup complexity and effort.
6. Transferring large amounts of data can be time consuming, expensive, and impacts production. Extrapolation can allow creating a data volume that mimics the requirements several years into the future.
7. The tests should be designed with the end decision in mind. Tests should clearly mimic the intended production environment and have straight-forward measurements. Tests should map to the success criteria.
8. Reduce variation across current and potential platforms to keep analysis meaningful. Ensure same data, same queries, same tests, same rules, and if multiple vendors involved, insist on similar size platforms.

Discussion:

A set of clearly documented success criteria is essential to a thorough, cost effective data warehouse benchmark. Failure to understand the minimal performance requirements for each type of usage could result in a selection of a platform that can only meet a portion of the future requirements or require dire compromises for users and implementation expense for IT. Failure to clearly understand available expense and capital budget could result in a fast platform that is unaffordable or requires lengthy funding battles before implementation. Don't be tricked by heavily discounted platforms now that will require expensive upgrades later after your applications are built.

Some companies choose too large a scope of data, resulting in excessive people and system costs to extract data, impacts to production systems, difficulty in writing data to transportable media, complexity in loading, and large numbers of mismatches between tables, views, procedures, reports, and queries. It is less expensive both for you and your vendor to take a smaller set of 25 to 75 tables that are representative of data in your company with several including your larger, most used tables. Provide table definitions, queries, row counts, and sample data to the vendor prior to actual testing to ensure they are provided a complete set of materials. This allows the vendor to enlist specialists for potential unique features you will be using. Remember that this is a test of the platform and not the breadth of skills of the vendor's benchmark team.

The set of queries chosen should represent different types of usage including analytic, ad hoc reporting and tactical retrieval for various business functions that would use the data warehouse. If a smaller set of queries is selected, a more robust set of tests can easily be constructed by adding parameters to the queries to simulate users with different interests. The tests should be constructed with the same ratio of query types expected in the production data warehouse. A rule of thumb is that with think/retrieval time, 10 logged on users results in one query in flight at any point in time. Therefore, a test with 15 batch query execution streams is a fair simulation of 150 logged on users. We recommend the following tests:

- **Test 1**: serial execution of queries and data load/data update processes,
- **Tests 2-5**: concurrent queries with 5, 10, 20, 40, and 80 streams (simulating 50 to 800 logged on users), and
- **Test 6**: mixed workload test with load/update processing and 10 or 20 streams of queries

Encourage hands-on participation by user organization and technology support staff. They will be impacted by the platform selection. Bringing them along early in the process will make it easier to achieve consensus when a decision is made. Even better: get some hands-on time to execute/measure 3-6 surprise queries and to get a feel of DBMS complexity for yourself.

Finally, there should be a thorough collection of results. For tables, that includes row counts and space after initial loading and before and after any insert/update/delete operations. For queries and update processes, measure the duration, CPU, and I/O resources consumed. For concurrent tests, calculate the platform's executions per hour and calculate total cost per query.

Next Step:

1. Make sure you have a complete overview of the vendor's capabilities with a thorough technology introductory session. That will provide ideas on areas of the technology that could improve your business and could be included in the tests.
2. Ask the vendor to conduct a half-day workshop to assist planning a fair and productive Benchmark/POC test.