

# Test Results for S3 NDP

## Abstract

This document contains content material for Rob's and Peter's sprint demo. Functionality and testing described in this document was implemented and tested by our shared effort. It is VERY difficult to test the performance of NDP server without Spark support. And it is also VERY difficult to test Spark performance without storage support.

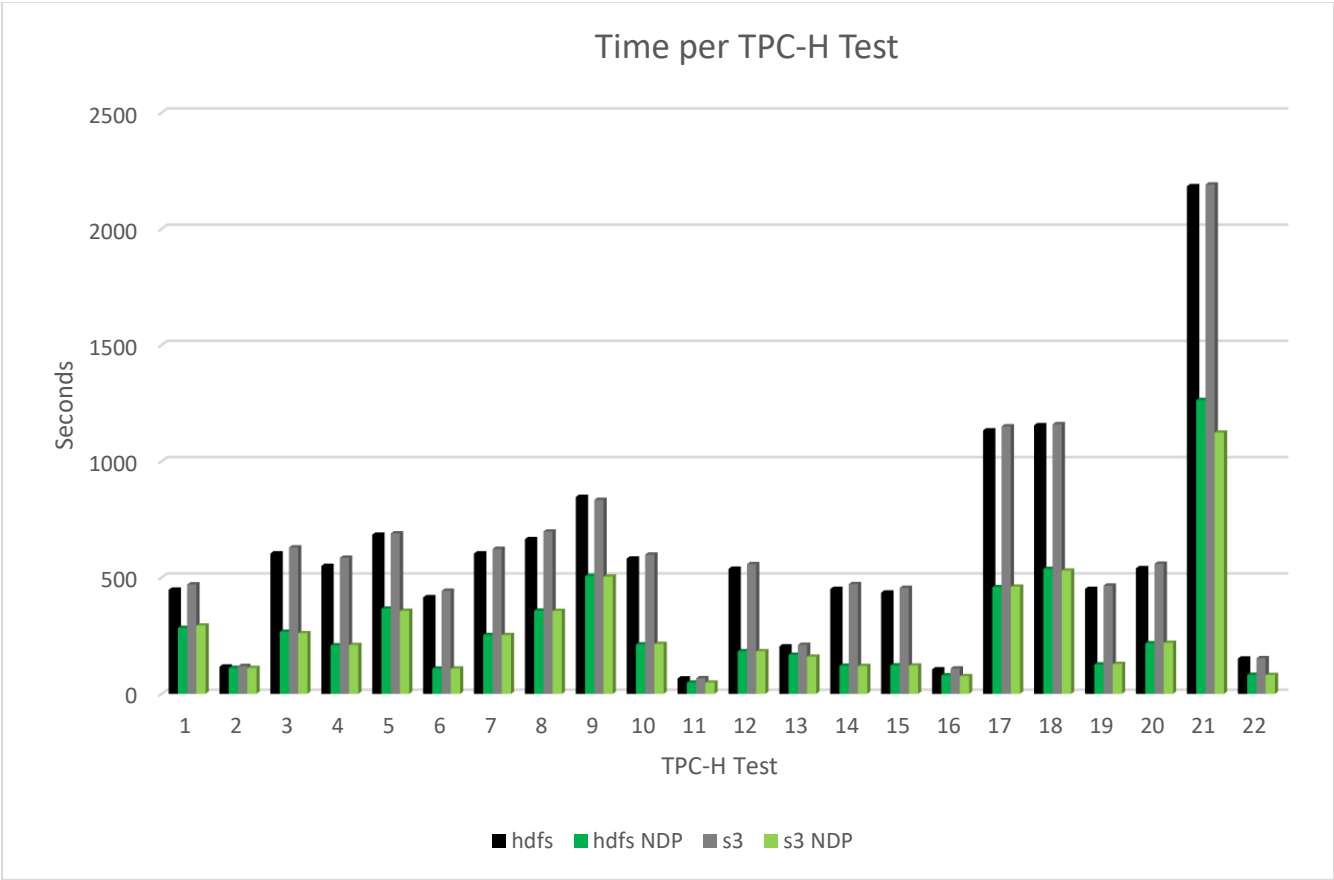
## Assumptions

The code being tested is from caerus-dike, which is our open source repository. Our goal is to demonstrate performance improvements for tpch benchmark. We are working with tpch 100GB dataset generated by `./dbgen -s 100`. The cases below use our V2 data source. The cases below use CSV files. All tests use 4 spark workers.

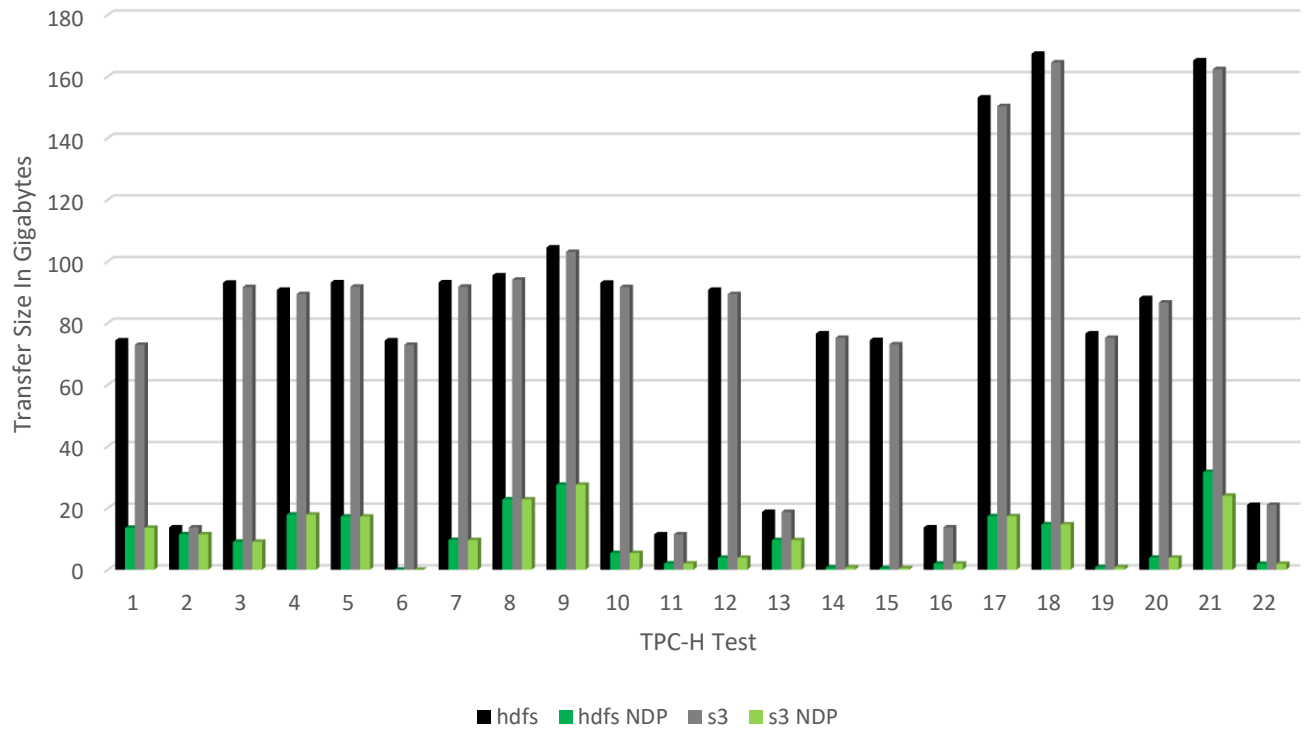
## What is new in this Demo:

- Dike S3 NDP on HDFS.
- Datasource support of minio (S3 compatibility) on HDFS.
- Support for CSV files in S3 and HDFS
- Gathering additional data
  - Spark compute node utilization
  - Throughput

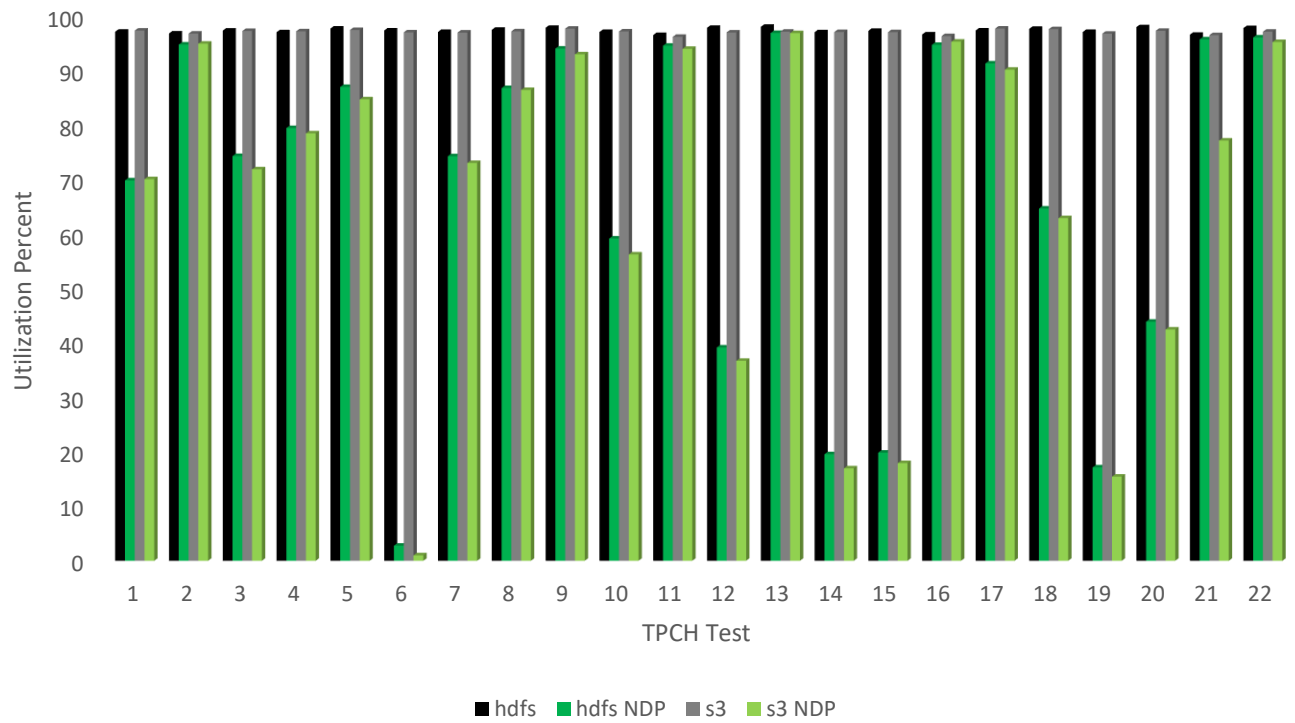
Performance Results for S3 vs HDFS



### Bytes Transferred per TPC-H test



### Compute Node Utilization per TPCH Test



Throughput per TPC-H Test

