April 2016

Spark Multi-user Benchmark

M. Genkin, Z. Hu, S. Warren, J. Nguyen

Feb. 12/2016

Spark Multi-user Benchmark

- Benchmark Objectives
- Use Cases
- SMB Stage 1 Description
  - Theory
  - Implementation
  - Metrics and analysis

Spark Multi-user Benchmark Objective

The Spark Multi-user Benchmark (SMB) is designed to **measure resource manager performance under multi-user conditions**:

- Multiple users run jobs on the systems, managed by the resource manager, concurrently.
- Each user submits a sequence of jobs. The total number of jobs is the same for every user.
- The total number of users running jobs on the system is increased until a desired number is reached.
- The system reaches and retains steady state.
- As user job sequences complete, the overall system utilization decreases.
- SMB measures resource manager's scheduling efficiency and ability to maintain QoS for users.

Use Case 1: End-of-Year/Quarter/Month Analytics

Major bank needs to publish end-of-year/quarter/month report

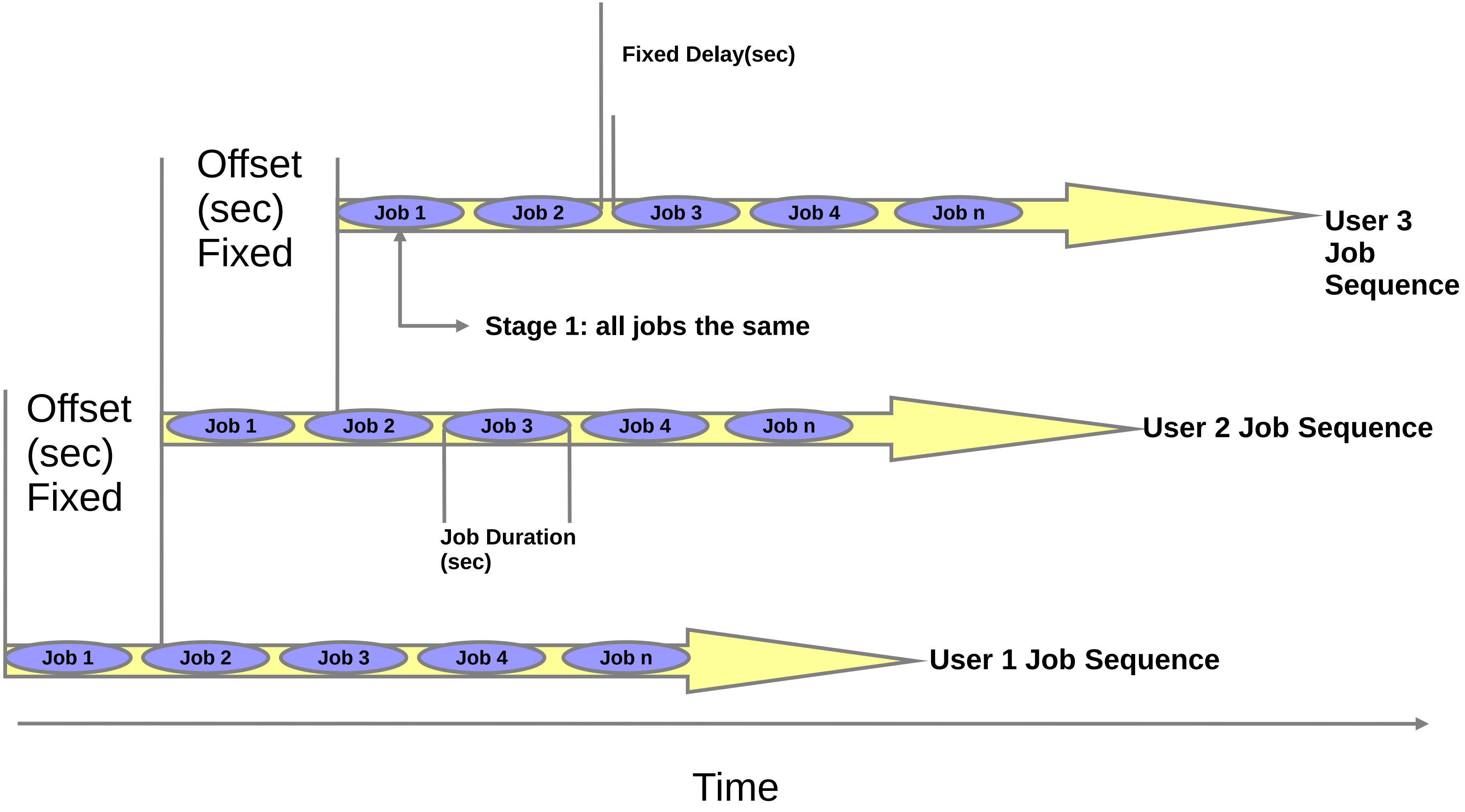
- Analysts run multiple types of analytic jobs to analyze sales performance.
  - Reports by product category
  - Reports by geography
  - Reports by customer demographic
- As the deadline approaches, the number of analysts running jobs on the cluster increases.
- As the deadline passes, the number of analysts running jobs on the cluster decreases
- At peak, the cluster is heavily utilized and in steady-state.

Use Case 2: Online Trading Site

Major brokerage runs an online trading site, with Spark analytics

- Spark analytics are used to analyze trader's profile and search or sort applicable trades.
- During high-volume trading days, the number of traders on the site increases until the analytic cluster is fully utilized.
- The cluster remains in steady-state heavy operation until the high-volume trading day (for example, triple-witching day) is over, and the load on the analytic cluster gradually decreases.

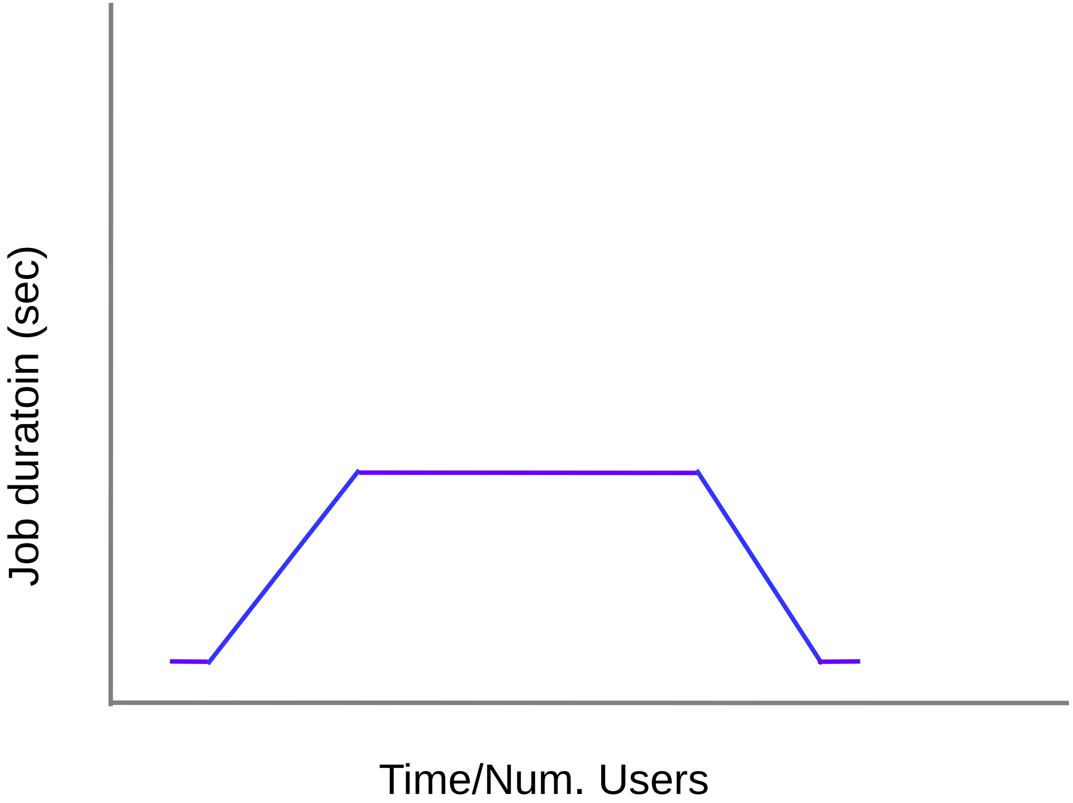
SMB-1 Benchmark



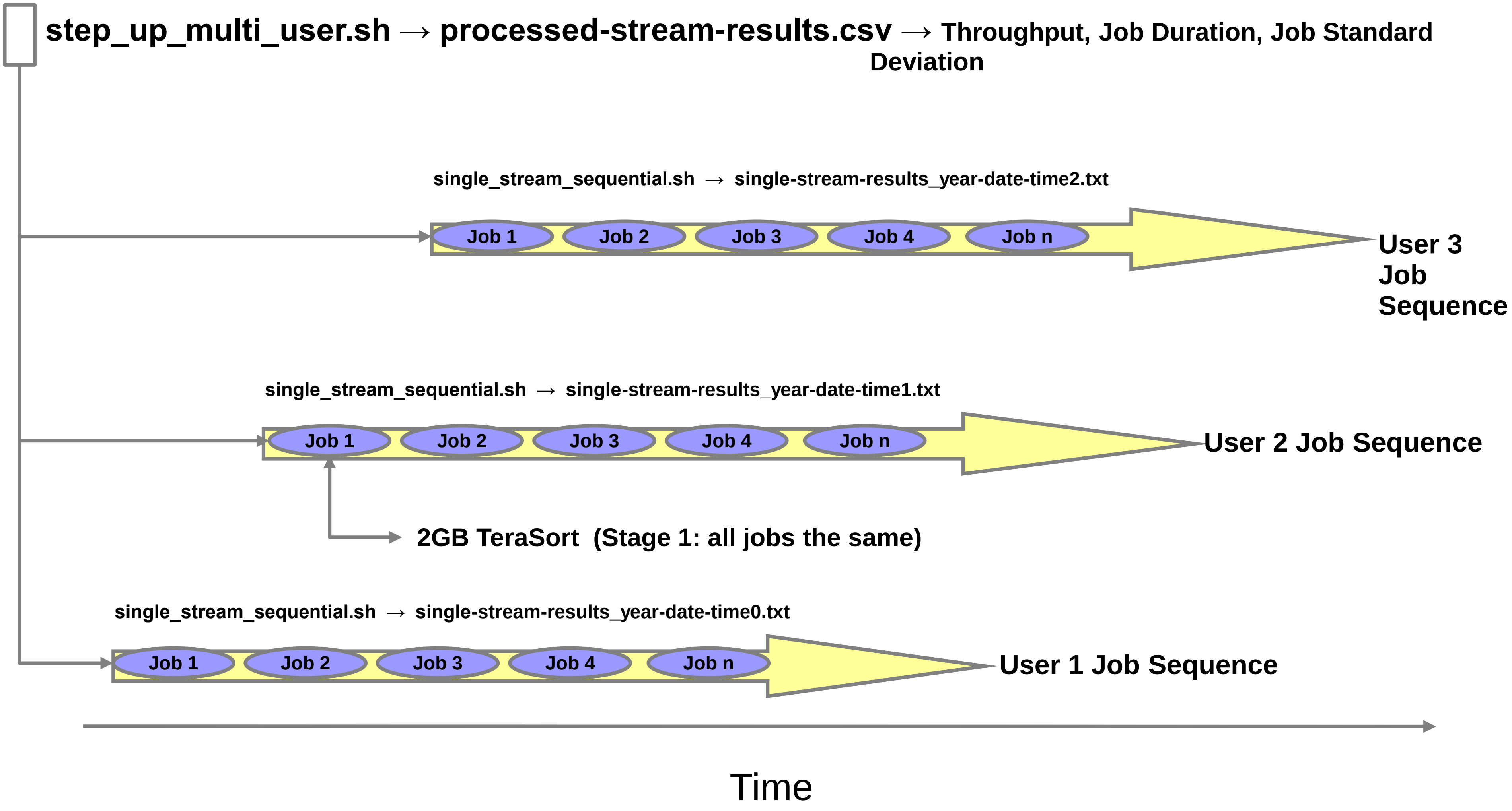
SMB Benchmark Theory

- Duration of each job executed is proportional to resources allocated by the resource manager
- Plot of job duration for all jobs vs. test duration should show a pattern similar to figure on the right
- Job duration data can be used to calculate key metrics related to resource manager efficiency:

  1) Throughput
  2) Job duration
  3) Job duration variance



SMB-1 Benchmark Implementation



step_up_multi_user.sh → processed-stream-results.csv → Throughput, Job Duration, Job Standard Deviation

single_stream_sequential.sh → single-stream-results_year-date-time2.txt

Job 1 Job 2 Job 3 Job 4 Job n

User 3 Job Sequence

single_stream_sequential.sh → single-stream-results_year-date-time1.txt

Job 1 Job 2 Job 3 Job 4 Job n

User 2 Job Sequence

2GB TeraSort (Stage 1: all jobs the same)

single_stream_sequential.sh → single-stream-results_year-date-time0.txt

Job 1 Job 2 Job 3 Job 4 Job n

User 1 Job Sequence

Time

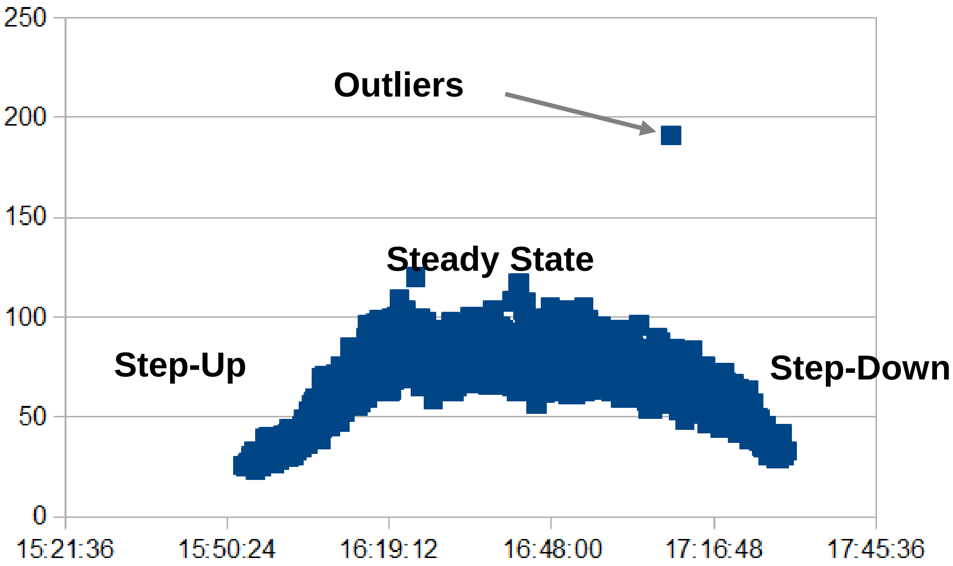
SMB-1 Example Benchmark Environment

SMB-1 environment specs:

- 1 master node + 10 compute/data nodes in the cluster
- Each node is a IBM System x3630 M4 server with Intel Xeon Processor E5-2450 at 2.10GHz, 32 vcores (2 CPU, 8 physical cores per CPU, 2 hyperthreads per core), 96 GB RAM
- RHEL 7.1 on all nodes
- The master node has 1 local disk for OS and software install.
- Each compute/data node uses 12 local disks, 1 for OS and software installs, 11 for data disks of Spark, HDFS, and YARN
- 10 GbE network
- NFS for Spark history log

SMB-1 Benchmark Metrics And Analysis

- **Throughput:**
  - Measured in jobs/hr
  - All jobs which successfully completed during the step-up, steady-state and step-down phases are counted
- **Job duration:**
  - Measures 90th percentile job duration in sec
  - All jobs which successfully completed during the step-up, steady-state and step-down phases are counted
- **Job standard deviation:**
  - Measure of variance, or scatter of the data
  - Measures differences in job duration in sec
  - All jobs which successfully completed during the step-up, steady-state and step-down phases are counted



Plot of the job duration data points shows how fairly the resource manager distributes resources among jobs