bigstream

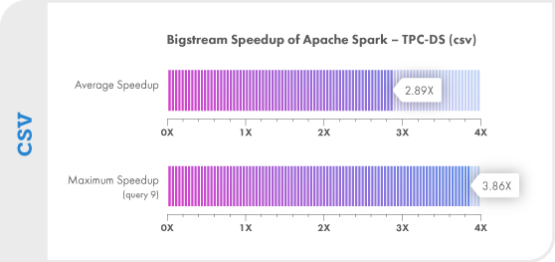
# Bigstream Benchmark Report

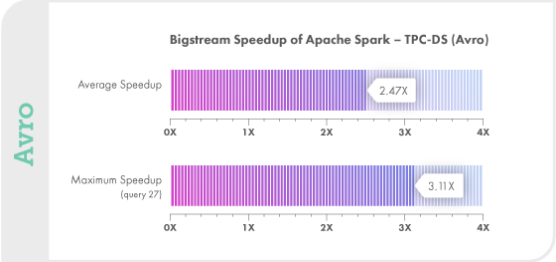
## Apache Spark - June 2017

## Overview

For this month, Bigstream is publishing recent results from running the TPC-DS benchmark against the baseline open source Apache Spark v2.1.1 versus Spark v2.1.1 with Bigstream Hyper-acceleration. The results show a nearly 300% performance improvement over baseline Apache Spark running on identical hardware.

Here is a summary of this month's results running Apache Spark with the Bigstream Hyper-acceleration Layer on Amazon EMR versus unaccelerated Apache Spark on the same configuration:




The TPC-DS benchmark is composed of a series of decision support queries and is commonly used by Big Data companies (Databricks, MapR, Cloudera, etc.) to measure performance.

Details of the results are below, but the figures above show a 2.89X and 2.47X speedup for csv and Avro data respectively. Top acceleration factors for individual queries were 3.86X acceleration of TPC-DS Query 9 for csv data, and 3.11X acceleration of Query 27 for Avro data. Future results will reflect Parquet and JSON data types.

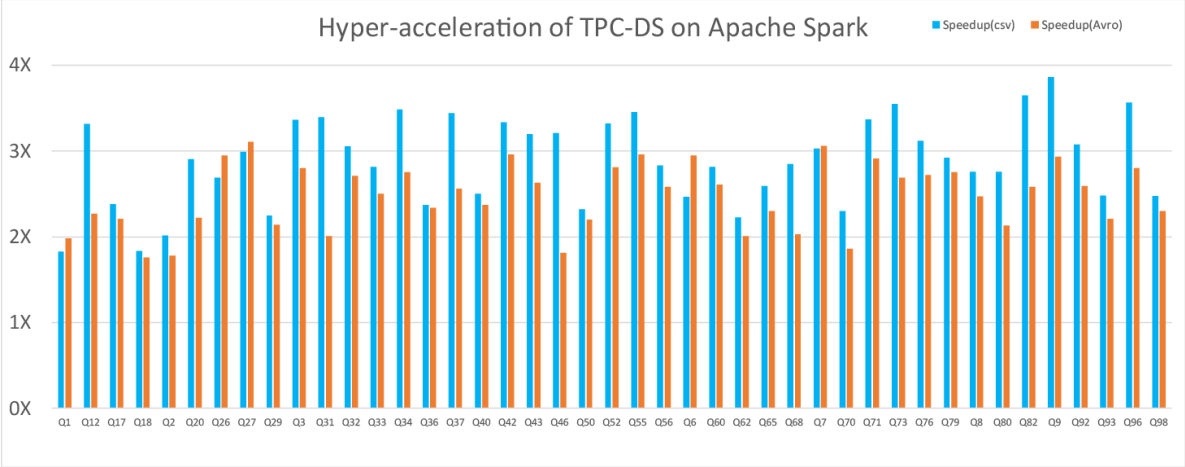
Given these results, very rough arithmetic shows a monthly savings of 54% on Amazon EC2/EMR given the 2.89X average acceleration and 46% given the 2.47X speedup.



2570 W. El Camino Real, Ste 510, Mountain View, CA 94040
info@bigstream.co | sales@bigstream.co | www.bigstream.co

# Detailed Results

This month, the Bigstream team ran the TPC-DS benchmarks using both CSV and Avro data on Amazon EMR cluster of four M4.4XL instances. Here is a graph of the results.




As you can see by the graph, speedups versus standard Apache Spark range from a little below 2X to nearly 4x. Here is a quick drill down into some of the better results:

| Bigstream TPC-DS Query Results (csv) | | |
|---|---|---|
| Average Speedup | 2.89X | |
| Maximum Speedup | 3.86X | Q9 |
| Top Performing | 3.86X | Q9 |
| | 3.65X | Q82 |
| | 3.57X | Q96 |
| | 3.55X | Q73 |
| | 3.49X | Q34 |
| | 3.46X | Q55 |
| | 3.44X | Q37 |
| | 3.40X | Q31 |
| | 3.37X | Q71 |
| | 3.36X | Q3 |

| Bigstream TPC-DS Query Results (Avro) | | |
|---|---|---|
| Average Speedup | 2.47X | |
| Maximum Speedup | 3.11X | Q27 |
| Top Performing | 3.11X | Q27 |
| | 3.06X | Q7 |
| | 2.96X | Q55 |
| | 2.96X | Q42 |
| | 2.95X | Q26 |
| | 2.95X | Q6 |
| | 2.93X | Q9 |
| | 2.91X | Q71 |
| | 2.80X | Q52 |
| | 2.80X | Q96 |



2570 W. El Camino Real, Ste 510, Mountain View, CA 94040
info@bigstream.co | sales@bigstream.co | www.bigstream.co

# Observations

## Average Cost Savings

| Monthly Cost Savings<br>TPC-DS Results – May 2017 | | |
|---|---|---|
| **Average Speedup (csv)** | **2.89X** | |
| Average monthly savings* | 54% | |
| Top monthly savings (csv) | P2.16xlarge<br>X1.32xlarge<br>P2.8xlarge | $65,089<br>$56,103<br>$33,180 |
| **Average Acceleration (Avro)** | **2.47X** | |
| Average monthly savings | 46% | |
| Top monthly savings (Avro) | P2.16xlarge<br>X1.32xlarge<br>P2.8xlarge | $58,197<br>$48,982<br>$29,677 |
| *Based on an Amazon EC2/EMR 10-node cluster, on demand instances. Details below | | |

An average speedup of 2.89X for csv data yields an average TCO savings of 54% for on demand instances, and 48% for reserved instances. The following instance types showed 60% or better TCO savings:

> **hi1.4xlarge, hs1.8xlarge, i2.xlarge, i2.2xlarge, i2.4xlarge, i2.8xlarge, d2.xlarge, d2.2xlarge, d2.4xlarge, d2.8xlarge, cg1.4xlarge, P2.xlarge, p2.8xlarge, P2.16xlarge, c1.xlarge**

See a description of Amazon EC2 instance types here.

As a point of reference, for a 10 node cluster using this range of instances, the average monthly savings would be between $2764 (c1.xlarge) - $65,089 (P2.16xlarge)!

Of course, cost reduction/avoidance is not the only reason people like performance. The key benefits our customer are looking for also include faster time to insight, reducing latencies for interactive or real-time applications, reducing burdensome batch processing windows, and gaining the ability to do more sophisticated computations, data science models and advanced analytics.

## CSV versus Avro Data Parsing

The average speed for csv is about 25% faster than for Avro. This is most likely a reflection of how Spark handles csv parsing. Besides working on accelerating Spark SQL/Dataframes, the Bigstream development team has focused a lot on document parsing of csv, Avro, JSON and Parquet data, as well as compression and decompression.



2570 W. El Camino Real, Ste 510, Mountain View, CA 94040
info@bigstream.co | sales@bigstream.co | www.bigstream.co

# Testing Environment

The current results were achieved on a 4 node Amazon EMR cluster using m4.4XL instances. To refresh your memory, the M4.4XL instances feature 16 vCPUs running on 2.3 GHz Intel Xeon® E5-2686 v4 (Broadwell) processors or 2.4 GHz Intel Xeon® E5-2676 v3 (Haswell) processors), and 64 GB of memory. The baseline open source Apache Spark v2.1.1 is downloaded and built from the Apache Spark website.

# Cost Savings Calculation

Amazon AWS/EMR cost savings are based on current pricing from the Amazon site and a fairly simple TCO model we have developed that looks like this:

**SP = (BX* (EC2 + EMR)) - (EC2 + EMR +BP) / (BX * (EC2+EMR))**
**Monthly savings = SP * (EC2 + EMR) * 24 hours * 30 days**

BX = Bigstream speedup factor
EC2 = hourly per instance pricing
EMR = hourly per node pricing
BP = Bigstream pricing
SP = savings percentage

# Disclaimer and Next Steps

Chances are, if you are using Spark SQL/Dataframes, and ingesting and processing JSON, CSV, Parquet or Avro data from S3, HDFS, or Kafka. You are going to see significant performance gains with Bigstream.

Benchmark results are performed for the purposes of assessing how Bigstream hyper-accelerates Apache Spark and to help us test our product. The results are useful in understanding and assessing how our product is performing on a wide array of decision support tasks. But hyper-acceleration will bring different benefits to different environments and application architectures. Luckily, it is fairly easy to test Bigstream in your environment - especially if you are using Amazon AWS.

Cost savings are extrapolated from a 4 node EMR cluster benchmarking run. These calculations are meant to give you a rough estimate, clearly your mileage may vary. Also, you may realize cost savings by electing to use accelerated performance to reduce the number of servers in your configuration, thereby reducing capital or cloud expenses.

To get optimal performance, we can work with your team to get more precise performance and savings numbers for your environment.



2570 W. El Camino Real, Ste 510, Mountain View, CA 94040
info@bigstream.co | sales@bigstream.co | www.bigstream.co