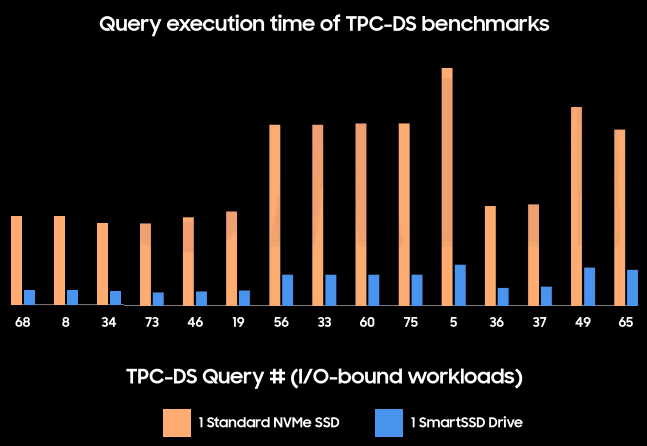
Solution Brief

Samsung SmartSSD® Drives and Bigstream for Spark Acceleration

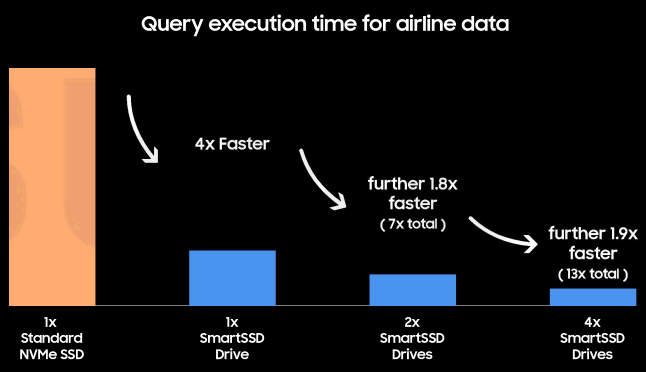
Accelerate SparkSQL Queries with Samsung SmartSSD Computational Storage Drives and Bigstream

- Faster Time to Insight - quickly trial and refine queries
- Higher Analyst Productivity - better utilize the time of scarce data scientists
- No Code Changes Needed - data scientists stay focused on their domain, rather than microarchitecture optimization
- Scale performance without increasing server count - reduce CPU load, infrastructure TCO, networking complexity, floor space sprawl

5x - 6x Faster Time to Insight



Scalable Acceleration

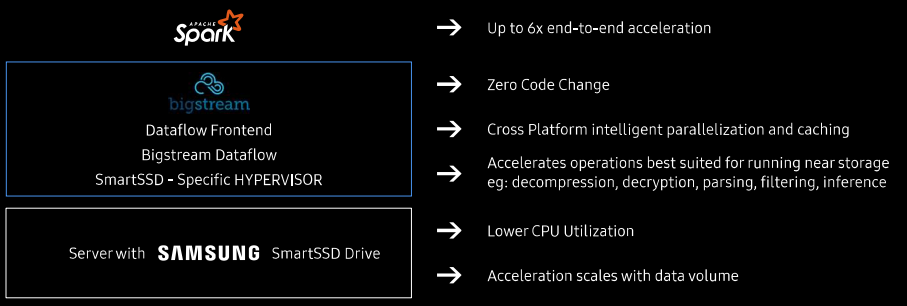


The Bottleneck: Massive Data Movement

The proliferation of machine-generated and user-generated content provides tremendous opportunity for business insights, but the volume of data creates challenges for analysis. Traditional analysis architectures rely on **moving large amounts of data from storage devices to a host CPU for analysis.** The host CPU becomes a system bottleneck with limited I/O connectivity (especially, PCI-Express lanes) and limited processing capability. These limitations prevent full and timely data analysis, and lead to costly server sprawl in the datacenter.

The Solution: Process Data Where It Is Stored

Partners Samsung and Bigstream have developed an end-to-end solution **to reduce unnecessary movement of data and speed up data processing.** The solution combines the SmartSSD advanced computational storage device from Samsung and a complete software stack from Bigstream for integration into existing analysis workflows. Queries are automatically optimized to allow processing of data directly on the SmartSSD drive, before it reaches the host CPU. This local processing is performed efficiently by a Xilinx field-programmable gate array (FPGA) on the SmartSSD drive, thus avoiding large data movement between storage and CPU and speeding up time-to-insight.



How It Works

The Bigstream Hyper-acceleration Layer analyzes the output of the Apache Spark Catalyst query optimizer, producing an alternate set of Spark tasks accelerated on the SmartSSD drive's Xilinx FPGA. The optimizing compiler optimizes key steps of the big data pipeline, including extract-transform-and-load (ETL), and SQL analytics (Spark SQL). Bigstream software optimizes specifically for the acceleration engine inside the Samsung SmartSSD drive: parsing, filtering, decompression, decryption, inference and other functions are performed directly in the SmartSSD drive, delivering only relevant data to the CPU.

How It Scales

A single server can contain multiple SmartSSD drives (up to 24 U.2 SmartSSD drives in a typical 2U server). Each SmartSSD drive can run query acceleration in parallel, producing almost linear speedup, even on a low-end or highly over-subscribed host CPU. Most enterprise SSDs are limited to only four lanes of PCI-Express by the U.2 drive connector, creating a bottleneck. By processing data directly on the storage layer, the SmartSSD drive can perform operations using wider interfaces than the U.2 connector supports. Unlike CPU-attached accelerators, the SmartSSD drive connects using the existing SSD PCI-Express lanes. This configuration frees up PCI-Express lanes for other uses, such as networking, or additional storage throughput.

Get Started Today

Get started with Bigstream and Samsung SmartSSD drive-accelerated Apache Spark for free today: trial the full hardware and software stack performance with a dataset of your own on Nimbix public cloud. Visit https://samsungsemiconductor-us.com/smartssd/ to learn how.

Solution Specification

bigstream Software	Data Analysis Software	Apache Spark 2.1.1 - 2.4.x
	Host Operating System	Ubuntu 16.04
	Required Drivers	OpenCL and Xilinx Runtime (XRT)
	Data File formats supported for acceleration	Row: CSV and JSON, with up to 64 columns/table and up to 50MB/file Coming soon: Parquet with up to 16 columns/table and up to 16 pages/row group
SAMSUNG SmartSSD® Computational Storage Drive	Form Factor	2.5" (U.2)
	Capacity	3.84TB
	Host Interface	PCIe Gen 3 x4
	Spec Compliance	NVMe spec rev. 1.3, PCIe base specification rev. 3.0
	NAND flash memory	Samsung V-NAND
	Acceleration Engine	Xilinx Kintex UltraScale+ FPGA
	Power Consumption	Dynamic power management and throttling
	Physical Dimensions	69 x 100 x 15 mm
	Weight	400 grams