Flash Memory Summit - 2019

Leveraging NVMe, SDS and commodity hardware to double query performance for Oracle RAC systems

Mosaic SD NVMe

Brian Dougherty, CMA
Chief Technology Officer

CMA

Excelero

Mosaic SD NVMe

Agenda

- Customer Environment
- Solution Description
- Evolution
- Proof of Performance
- Questions

CMA

Mosaic SD NVMe

Excelero

Customer Environment

- One of the largest Medicaid recipient States in the US
- Providing analytics against millions of health care recipients with 20 years of claim data (20+ billion rows of claim data)
- Processing Terabytes of new claim data per year
- Servicing thousands of analytic users running thousands of queries/day

Mosaic SD NVMe

Current state-of-the-art platform

- A purpose built database platform – Pre-calibrated/pre-configured Oracle platform
- Architectural components - Compute / Storage / Interconnect
- Functional Components included with the solution:
  - A built in database Migration product – Mosaic DART
  - Web based Console providing real time inventory, health, and alerts of HW / SW / Database components

CMA

Excelero

Mosaic SD NVMe

Features

- Ultra high performance with very low latency
- Simple, yet elegant
- Rapid deployment
- Elastic scale-out
- Pre-installed, Pre-configured Oracle RAC cluster in a “box”
- Point and click migrations
- Rack, stacked and wired – ready to ship and plug into a data center
- Calibrated and scalable configurations - start small and grow quickly
- Fully redundant platform
- Compelling economics price/performance

CMA

Mosaic SD NVMe

Excelero

How did we get here?

- Evolved over 10 years using state-of-the-art technology at the time

- Why did we build the product?
  - Eliminate risk of long build process and suboptimal platform – R&D/Configuration is already done to minimize deployment time
  - Customers demanding better Performance
  - Data volumes continue to scale exponentially

CMA

Excelero

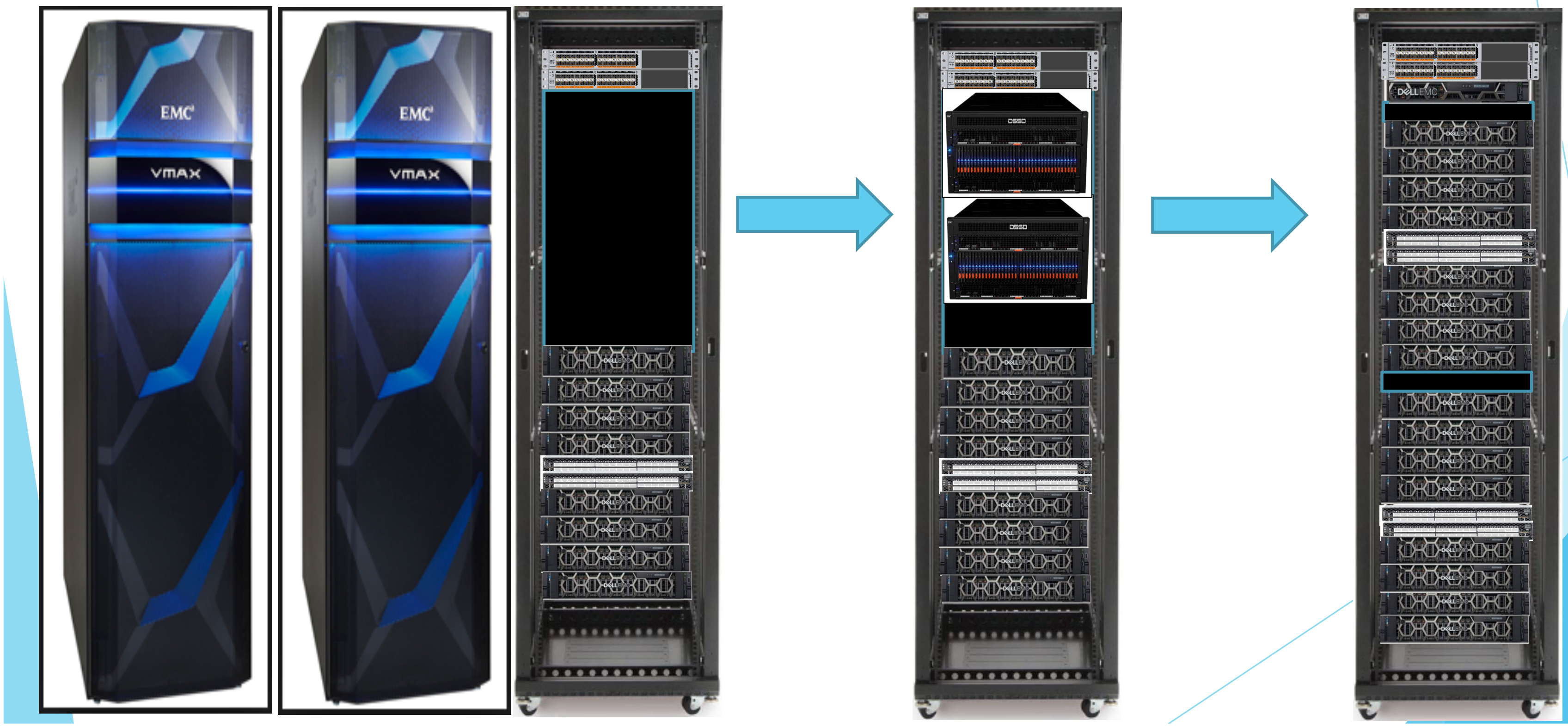
Mosaic SD NVMe

Evolution of the Customer Solution

Gen 1 - VMAX 400K

Gen 2 - DSSD 2xD5

Gen 3 - Mosaic SD NVMe



CMA

Mosaic SD NVMe

Excelero

Evolution of the Product Matrix

Generation	1	2	3
Technology:	VMAX – 400K – 2 Bay	DSSD - NVMe	Mosaic SD NVMe
Storage Protocol:	Fibre Channel	NVMe-PCIe attached	NVMeoF
Form Factor:	3 - 42U Racks	1 - 42U Rack	1 - 42U Rack
Scalability:	Difficult – Limited by monolithic storage technology, Very Expensive - Proprietary HW	Expensive and limited - Proprietary HW and firmware	Inexpensive and easily scales by adding 100s of commodity storage nodes
Performance - Bandwidth (Storage nodes):	34GB/s (6 Engine)	130GB/s (2 – D5)	154GB/s (8 – Dell 740XD)
Power Consumption:	20.6 kW	4.6 kW	3.6 kW
Cost **(Storage HW & SW)**	$2,000,000 +	$1,000,000 +	~ $500K

Mosaic NVMe SD Technology

Compute	
Optimized Kernel	(Oracle RAC / NVMe RDMA) Optimized Client Kernel
	Configure and verify Oracle required/recommended kernel parameters (shared memory, semaphores, network, RDMA, fileslystem)
	Configure Excelero-related kernel parameters (CPU Interrupt affinity, IRQ balancing, Mellanox affinity, performance profile)
I/O Interconnect Drivers (IB, SDS, Card)	Excelero Drivers
	Configure volumes to maximize drive and network bandwidth; configure volumes with appropriate RAID levels
	Provision volumes to clients as block devices
	Configure Excelero management software and supporting metadata database
	Topology manager monitors and manages errors in the fabric (failed drives, connectivity issues, etc)
Adapter Firmware	ConnectX5 firmware configuration
	Configure Mellanox firmware for maximum throughput and efficiency
	Configure target Mellanox firmware to support Excelero RDDA protocol
	Configure optimized PCI read-ahead

CMA

Excelero

Mosaic NVMe SD Technology

Compute Cont.	
Server	DELL R740 2U Servers
	Memory: 384G; 24 sockets (12 sockets/CPU); each set of 12 is organized into 6 channels for maximum bandwidth
	CPU: 2-socket, 22 cores/socket = 44 cores; Intel (R) Xeon(R) Gold 6152 CPU @ 2.10GHz
	PCIe Gen3: 2U server allows for ample PCIe Gen3 x16 slots to support massive storage traffic
I/O Adapters	ConnectX5 (X16 v X8)
	Multiple x16 slots support the bandwidth-hungry ConnectX-5 storage network adapters
	Queue pairs optimized to support RDDA protocol on target nodes
	I/O adapters balanced across NUMA nodes
Oracle NVMe Optimized Settings	Oracle NVMe RDMA optimized configuration
	Configure to be NUMA aware
	Configure Oracle to maximize physical i/o efficiency
Oracle RAC Software	Optimized Oracle Configuration
	Install/configure Oracle version 12c / 18c
	Pre-configured Oracle Real Application Cluster (RAC)
	Pre-configured Oracle Automatic Storage Management (ASM)
	Enable Container databases
	Pre-created diskgroups based on customer requirements
	Pre-created databases with TPCH sample data

CMA

Excelero

Mosaic NVMe SD Technology

Storage	
Storage Switches	Mellanox SB7800 Switches (EDR -100Gbps)
Storage Networking	NVMe client drivers/RDMA IB transport/IB network layer/IB link layer
Storage Drives	Industry optimized NVMe SSD drives
Storage Server	DELL R740XD 2U Servers (24 NVMe Chassis / 32 PCI Gen 3 Lanes) Cost effective CPU and memory configuration
Storage Target Drivers	Excelero Target and Management software
Storage Optimized Kernel / OS	(Oracle RAC / NVMe RDMA) Optimized Storage kernel
Storage Adapters	Mellanox ConnectX5 VPI
Storage Fabric Cabling	Mellanox-branded redundant cabling

CMA

Excelero

Mosaic NVMe SD Technology

Mosaic NVMe Management	
Software	
Mosaic NVMe SD Console	Version 2.1
	Provides integrated/comprehensive real time view of the Mosaic NVMe hardware and software.
	View inventory and health of all hardware and software components.
	View details and state of all of the Oracle databases on the platform.
NVMe Console Hardware	DELL R640 1U Server
Software Infrastructure	
Wildfly	Version 15.0.0
KeyCloak	Version 5.0.0
389 - LDAP	Version 1.3.8.4
Postgress	Version 11.3

CMA

Mosaic SD NVMe

Excelero

Proof of Performance

- Atomic Operations
  - Create tablespace
  - Direct path load
  - Index builds
  - Max sustainable scan rate
  - Aggregate sustainable scan rate

- OBIEE query benchmarks

CMA

Mosaic SD NVMe

Excelero

Atomic Operations – Comparison with entry level solution

Benchmark	VMAX 400K	NVMe DSSD	Mosaic SD NVMe (2x4)
	(Oracle 8 node RAC cluster with 6 engine EMC VMAX)	(Oracle 8 node RAC cluster with 2 D5s)	(Oracle 2 node RAC cluster with 4 Dell R740xd)
Create 10TB tablespace	111 mins	82 mins	31 mins
Parallel direct path insert of 8 billion rows*	156 mins	101 Mins	77 mins
Create unique global partitioned index*	11 mins	4.6 mins	4 mins
Create local bitmap index*	92 mins	12 mins	3 mins
Query single table scan time*	27 mins	1.6 mins	2.5 mins
Aggregate sustainable scan rate	34 GB/Sec	130 GB/Sec	75 GB/Sec
Maximum sustainable scan rate per RAC node	**8 GB/Sec**	**24.5 GB/Sec**	**37 GB/Sec**

CMA

Mosaic SD NVMe

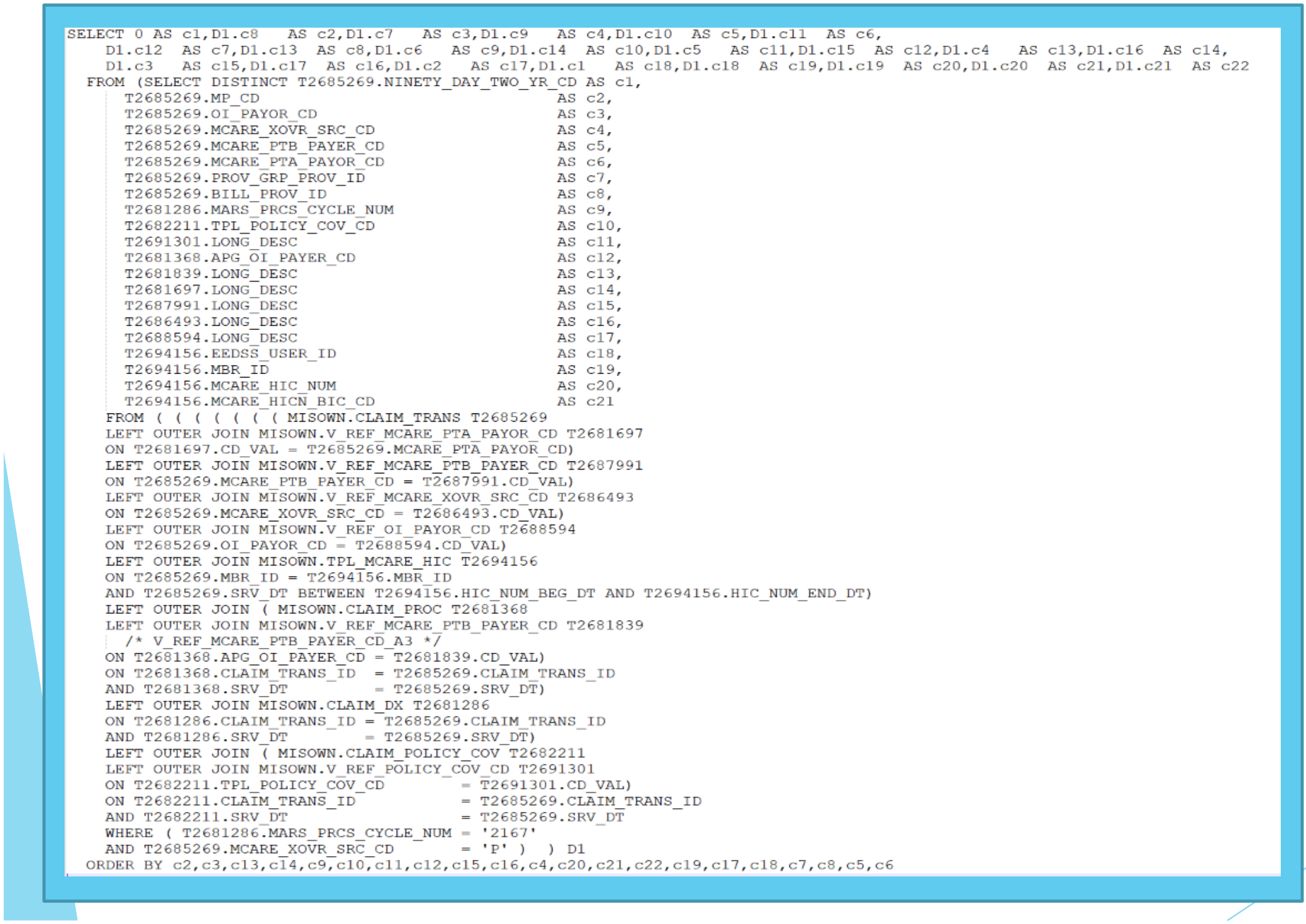
Excelero

Performance - Mosaic SD NVMe Large (8x8)

Customer Benchmarks	Mosaic SD NVMe (Large: 8x8)
General Benchmarks	(Oracle 8 node RAC cluster with 8 Dell R740xd - ASM Mirroring)
Average direct path read time per 8k block	< 100 microseconds
Total bytes scanned each hour by customer	137 TB (685 TB / day)
Total rows scanned each hour of day	.5 Trillion rows (2.5 Trillion rows/day)
Atomic Operations	
Create unique global partitioned index on 8+ billion rows table	2.3 mins
Parallel direct path partition to partition insert on 8+ billion row table (simultaneous read/write of 11.5TB of data)	60 mins
Query single table(11.5TB) scan time	1.4 mins
Maximum sustainable scan rate per RAC node (*3 – 8 lane PLX cards)	20 GB/Sec*
Aggregate sustainable scan rate	154 GB/Sec

Mosaic SD NVMe

Typical Customer Query



```
SELECT 0 AS c1,D1.c8   AS c2,D1.c7   AS c3,D1.c9   AS c4,D1.c10  AS c5,D1.c11  AS c6,
    D1.c12  AS c7,D1.c13  AS c8,D1.c6   AS c9,D1.c14  AS c10,D1.c5   AS c11,D1.c15  AS c12,D1.c4   AS c13,D1.c16  AS c14,
    D1.c3   AS c15,D1.c17  AS c16,D1.c2   AS c17,D1.c1   AS c18,D1.c18  AS c19,D1.c19  AS c20,D1.c20  AS c21,D1.c21  AS c22
  FROM (SELECT DISTINCT T2685269.NINETY_DAY_TWO_YR_CD AS c1,
      T2685269.MP_CD                               AS c2,
      T2685269.OI_PAYOR_CD                         AS c3,
      T2685269.MCARE_XOVR_SRC_CD                   AS c4,
      T2685269.MCARE_PTB_PAYER_CD                  AS c5,
      T2685269.MCARE_PTA_PAYOR_CD                  AS c6,
      T2685269.PROV_GRP_PROV_ID                    AS c7,
      T2685269.BILL_PROV_ID                        AS c8,
      T2681286.MARS_PRCS_CYCLE_NUM                 AS c9,
      T2682211.TPL_POLICY_COV_CD                   AS c10,
      T2691301.LONG_DESC                           AS c11,
      T2681368.APG_OI_PAYER_CD                     AS c12,
      T2681839.LONG_DESC                           AS c13,
      T2681697.LONG_DESC                           AS c14,
      T2687991.LONG_DESC                           AS c15,
      T2686493.LONG_DESC                           AS c16,
      T2688594.LONG_DESC                           AS c17,
      T2694156.EEDSS_USER_ID                       AS c18,
      T2694156.MBR_ID                              AS c19,
      T2694156.MCARE_HIC_NUM                       AS c20,
      T2694156.MCARE_HICN_BIC_CD                   AS c21
    FROM ( ( ( ( ( ( ( MISOWN.CLAIM_TRANS T2685269
    LEFT OUTER JOIN MISOWN.V_REF_MCARE_PTA_PAYOR_CD T2681697
    ON T2681697.CD_VAL = T2685269.MCARE_PTA_PAYOR_CD)
    LEFT OUTER JOIN MISOWN.V_REF_MCARE_PTB_PAYER_CD T2687991
    ON T2685269.MCARE_PTB_PAYER_CD = T2687991.CD_VAL)
    LEFT OUTER JOIN MISOWN.V_REF_MCARE_XOVR_SRC_CD T2686493
    ON T2685269.MCARE_XOVR_SRC_CD = T2686493.CD_VAL)
    LEFT OUTER JOIN MISOWN.V_REF_OI_PAYOR_CD T2688594
    ON T2685269.OI_PAYOR_CD = T2688594.CD_VAL)
    LEFT OUTER JOIN MISOWN.TPL_MCARE_HIC T2694156
    ON T2685269.MBR_ID = T2694156.MBR_ID
    AND T2685269.SRV_DT BETWEEN T2694156.HIC_NUM_BEG_DT AND T2694156.HIC_NUM_END_DT)
    LEFT OUTER JOIN ( MISOWN.CLAIM_PROC T2681368
    LEFT OUTER JOIN MISOWN.V_REF_MCARE_PTB_PAYER_CD T2681839
      /* V_REF_MCARE_PTB_PAYER_CD_A3 */
    ON T2681368.APG_OI_PAYER_CD = T2681839.CD_VAL)
    ON T2681368.CLAIM_TRANS_ID  = T2685269.CLAIM_TRANS_ID
    AND T2681368.SRV_DT         = T2685269.SRV_DT)
    LEFT OUTER JOIN MISOWN.CLAIM_DX T2681286
    ON T2681286.CLAIM_TRANS_ID = T2685269.CLAIM_TRANS_ID
    AND T2681286.SRV_DT        = T2685269.SRV_DT)
    LEFT OUTER JOIN ( MISOWN.CLAIM_POLICY_COV T2682211
    LEFT OUTER JOIN MISOWN.V_REF_POLICY_COV_CD T2691301
    ON T2682211.TPL_POLICY_COV_CD         = T2691301.CD_VAL)
    ON T2682211.CLAIM_TRANS_ID            = T2685269.CLAIM_TRANS_ID
    AND T2682211.SRV_DT                   = T2685269.SRV_DT
    WHERE ( T2681286.MARS_PRCS_CYCLE_NUM  = '2167'
    AND T2685269.MCARE_XOVR_SRC_CD        = 'P' )   ) D1
  ORDER BY c2,c3,c13,c14,c9,c10,c11,c12,c15,c16,c4,c20,c21,c22,c19,c17,c18,c7,c8,c5,c6
```

Fact Tables

Claim_Trans

- 13.5 TB
- 9.7 Billion rows

Claim_Proc

- 1.1 TB
- 5.9 Billion rows

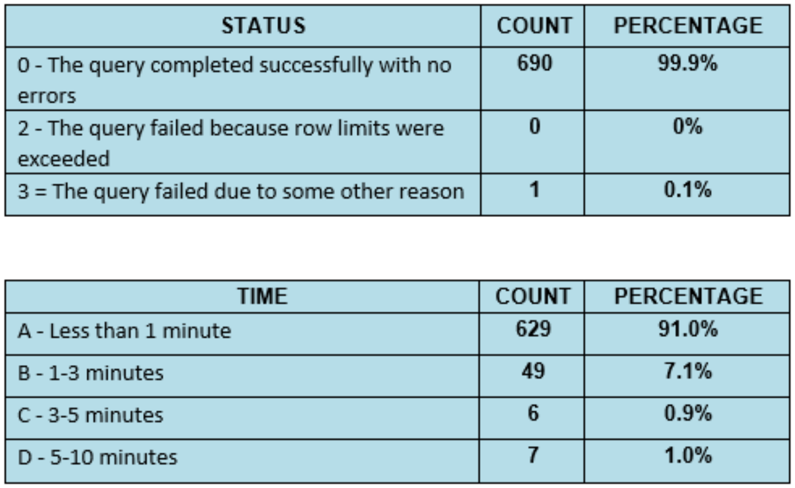
Claim_Dx

- 1.4 TB
- 16.1 Billion rows

Mosaic SD NVMe

Customer OBIEE Usage Statistics

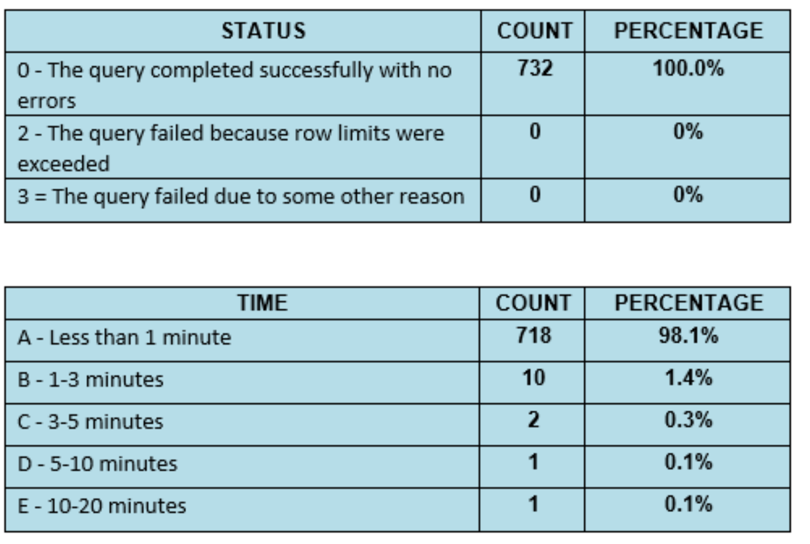
Day 1



STATUS	COUNT	PERCENTAGE
0 - The query completed successfully with no errors	690	99.9%
2 - The query failed because row limits were exceeded	0	0%
3 = The query failed due to some other reason	1	0.1%

TIME	COUNT	PERCENTAGE
A - Less than 1 minute	629	91.0%
B - 1-3 minutes	49	7.1%
C - 3-5 minutes	6	0.9%
D - 5-10 minutes	7	1.0%

Day 2



STATUS	COUNT	PERCENTAGE
0 - The query completed successfully with no errors	732	100.0%
2 - The query failed because row limits were exceeded	0	0%
3 = The query failed due to some other reason	0	0%

TIME	COUNT	PERCENTAGE
A - Less than 1 minute	718	98.1%
B - 1-3 minutes	10	1.4%
C - 3-5 minutes	2	0.3%
D - 5-10 minutes	1	0.1%
E - 10-20 minutes	1	0.1%

Mosaic SD NVMe

Questions?