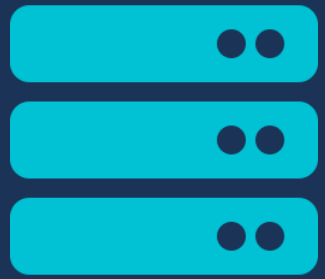
S2D
Storage Spaces Direct

Architecture, Use Cases & Considerations



Microsoft in partnership with DataON
June 2026

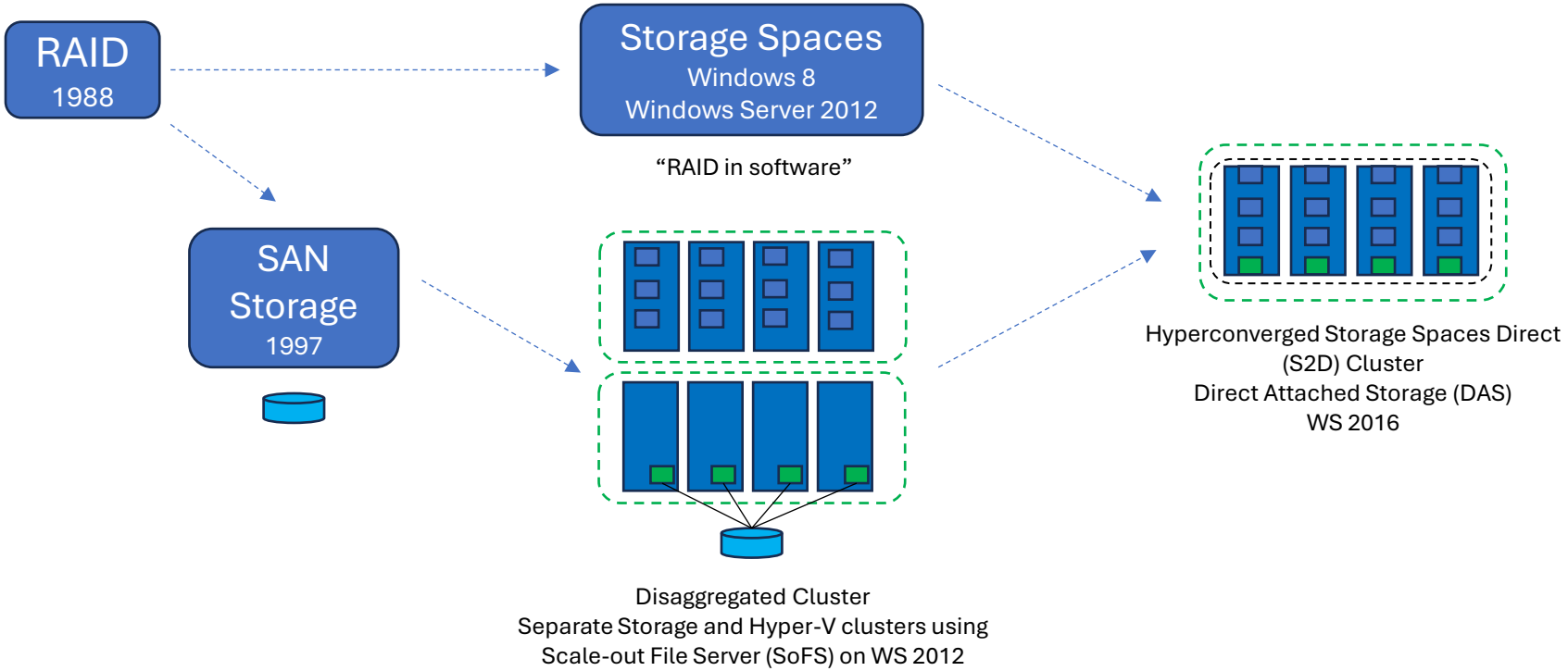
DataON | Microsoft Azure

Agenda

- What is S2D & where is it used?
- Architecture deep dive
- How nodes decide where to read/write
- Top 3 use cases
- Advantages
- Limitations
- Q&A

History of Storage Spaces Direct (S2D)

- Businesses use **stateful applications** (purchases, bank deposits, withdraws, reservations, etc.)
- **Stateful applications**: Hyper-V, SQL Server, File Server, Exchange, Search, User Preferences
- Highly available, redundant storage at low cost is highly desirable for **stateful applications**



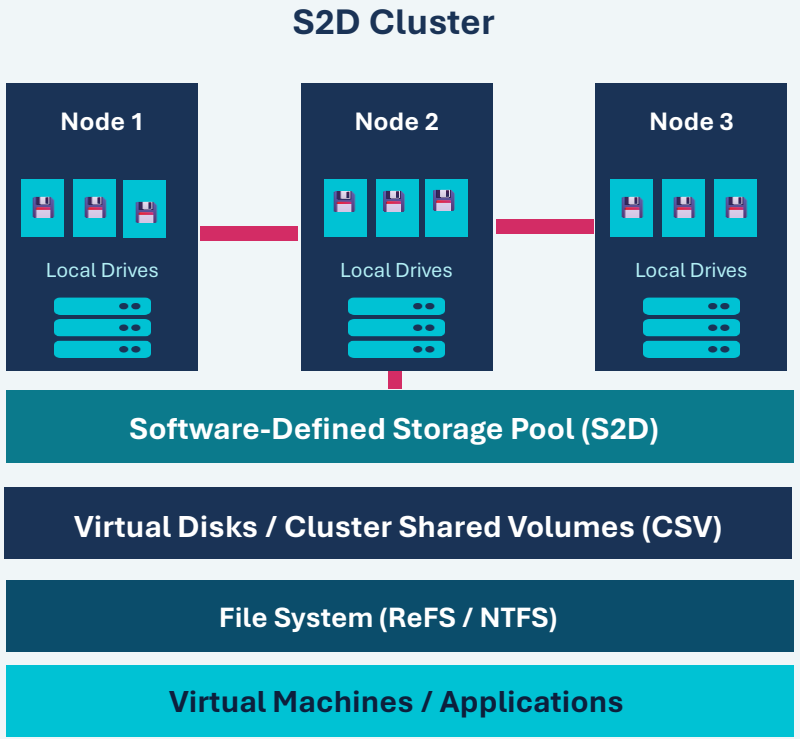
What is Storage Spaces Direct (S2D)?

Storage Spaces Direct (S2D) is a software-defined storage (SDS) technology built into Windows Server and Azure Local Medium HCI.

It uses industry-standard servers with local-attached drives to create highly available, scalable storage. **Works standalone or alongside existing SAN/NAS infrastructure.**

Where is S2D Used?

Azure Local Medium	**Windows Server 2016/2019/2022/2025**
VMs & Arc Connected Virtual Machines (VMs)	**VMs; SQL Server FCI; SAP…**



S2D Architecture

VM / Application Layer
Hyper-V, containers, workloads

Cluster Shared Volumes (CSV)
Unified namespace for clustered VMs

Cluster Health Service
Manage health of disks

File System (ReFS / NTFS)
Formats virtual disks; ReFS recommended for integrity streams & faster repair

Virtual Disks
Resiliency: Mirror, Parity, Mirror-Accelerated Parity

S2D

Storage Pool
Aggregated capacity from all local disks across all nodes

Storage Bus Layer
Presents all local disks cluster-wide; enables the pool

Physical Layer: NVMe / SSD / HDD
Cache tier (NVMe/SSD) + Capacity tier (HDD/SSD)

RDMA Networking

Low-latency, high-bandwidth node-to-node communication (RoCE v2 or iWARP)

Cache Tier

Mixed drives: faster media (NVMe/SSD) auto-caches HDD. All-flash: no cache layer needed; all drives in pool.

Resiliency

2-way or 3-way mirror, single/ dual parity — survives drive and node failures

Top 3 Use Cases for S2D

01

Hyper-Converged Infrastructure (HCI) for Virtualization

Run VMs, SQL, VDI, and mission-critical apps on the same cluster with built-in resiliency — no separate storage hardware required.

VDI | SQL Server | Mission-Critical

02

Enterprise Storage Modernization

Replace or augment traditional storage arrays with software on commodity hardware. S2D supports SAN coexistence, making it easy to migrate workloads at your own pace.

SAN Coexistence | Migration | Cost Reduction

03

Hybrid & Edge Infrastructure

Built for retail, manufacturing, and remote sites — local resiliency even in disconnected environments. Brings cloud-style storage on-prem with optional Azure Arc integration.

Edge | Distributed | Azure Arc

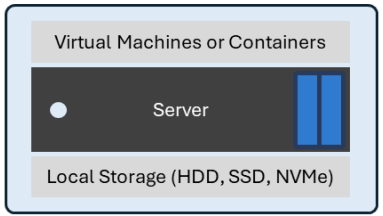
Windows Server Storage Architectures

Storage Area Network (SAN) & Network Attached Storage (NAS)

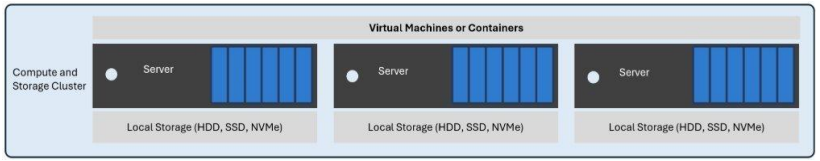
- Compute clusters accesses storage over network
- VMs highly available across compute nodes
- Storage and compute scale independently

Standalone

- Single server, local storage
- No automatic VM failover (non-clustered)
- Optional RAID for disk resiliency



Hyperconverged S2D

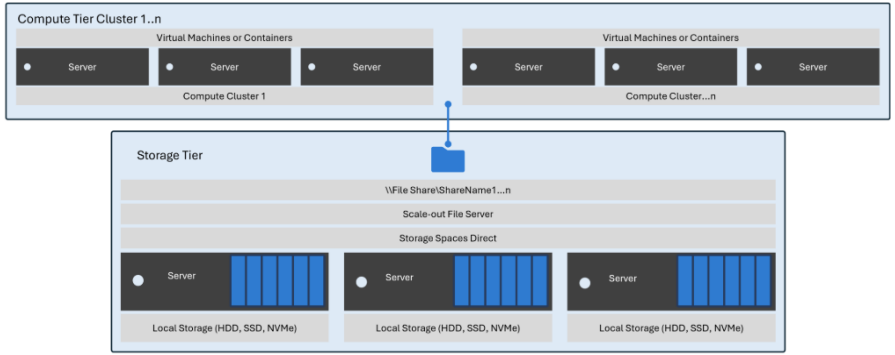


- Compute + storage on same nodes
- Data replicated across cluster
- Symmetric scale-out (1-16 nodes)

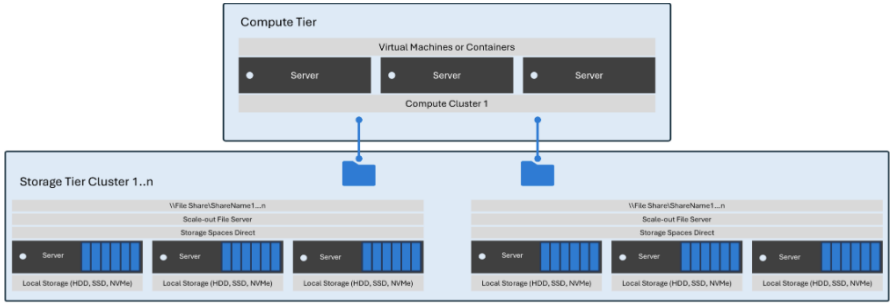
Windows Server Storage Architectures

Disaggregated S2D

- Separate compute & storage clusters
- Independent scaling (CPU vs capacity)
- Storage cluster: 1–16 nodes



Disaggregated with multiple compute clusters and a single storage cluster

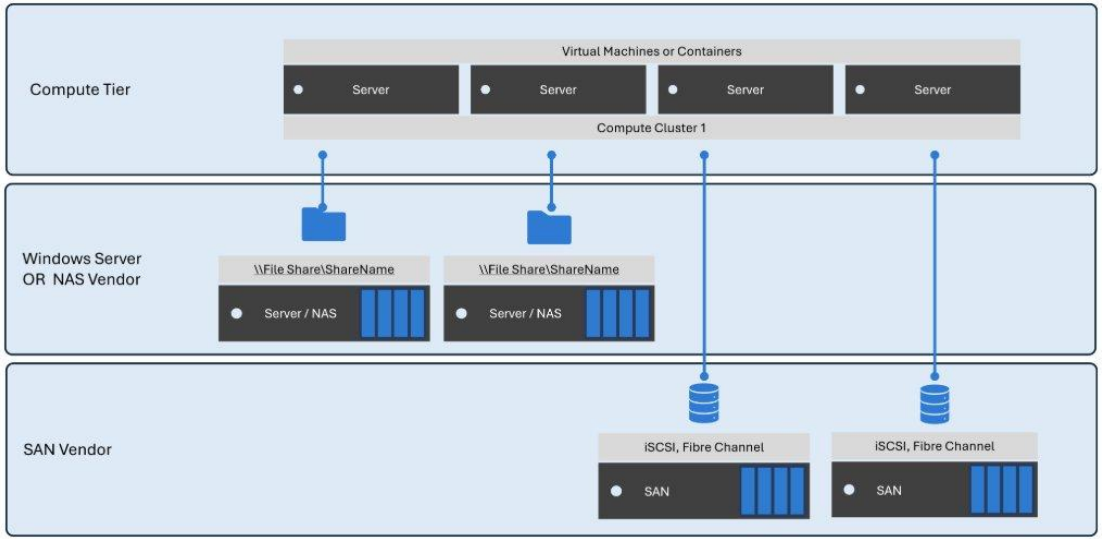


Disaggregated with single compute cluster and a multiple storage clusters

Windows Server Storage Architectures

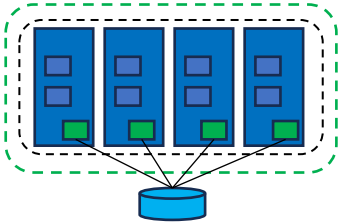
Mixed Support

- Disaggregated S2D + SAN/NAS
- Mixed storage tiers in one compute cluster
- Ideal for gradual adoption or hybrid scenarios

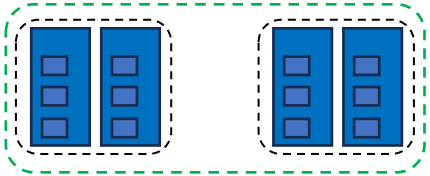


Variations of Storage Spaces Direct (S2D) Supported

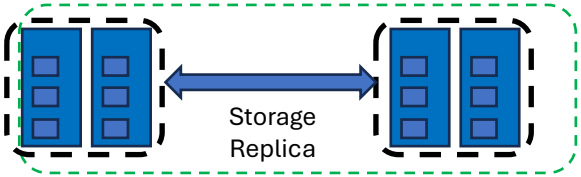
- S2D and SAN Coexistence in the same cluster (single rack)
- S2D Campus Cluster (AKA Azure Local Rack Aware Cluster)
- S2D Stretch Cluster using Storage Replica (Storage Pool in each Site) (coming to WS 2025 soon*)



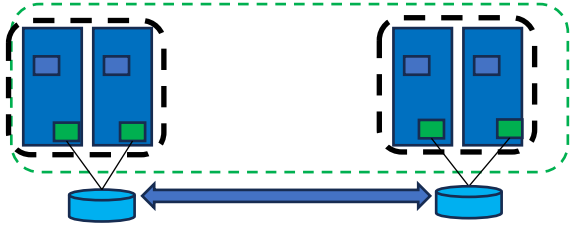
S2D and SAN (FC, iSCSI) Coexistence



S2D Campus Cluster (Single Pool)



S2D Stretch Cluster Each Site has a S2D Pool*



Stretch Cluster with SAN Replication

S2D Resiliency Types

2-Way Mirror

Tolerates 1 drive or node failure

Your data is written to 2 separate nodes simultaneously. If one node or drive fails, your data is still fully available on the other. Uses 2× your raw capacity.

Best for: Edge, branch, and ROBO deployments where cost matters more than maximum protection. Minimum 2 nodes.

Edge / ROBO | Low Cost | 2× Overhead

3-Way Mirror

Tolerates 2 simultaneous failures

Data is written to 3 separate nodes at once. You can lose two drives or nodes at the same time and keep running. Uses 3× your raw capacity.

Best for: Production VMs, SQL, and mission-critical workloads where uptime is non-negotiable. Minimum 3 nodes.

Production | Max Resilience | 3× Overhead

Single Parity

Tolerates 1 failure with less overhead

Similar to RAID-5 — data and parity information are spread across drives. More capacity-efficient than mirroring but slower to rebuild after a failure.

Best for: Large, sequential workloads like backups or archives where capacity efficiency matters more than raw performance.

Archival | Capacity Efficient | Lower Performance

Mirror-Accelerated Parity

Best of both worlds at scale

Hot (frequently accessed) data lives in a mirror tier for fast access. Cold data is automatically moved to a parity tier to save space. S2D manages the movement automatically.

Best for: Large-scale clusters with mixed workloads where you need performance AND capacity efficiency. Minimum 4 nodes.

Large Scale | Mixed Workloads | Most Efficient

How Nodes Decide Where to Read / Write

📖 Read Path — Prefer Local

- S2D prefers reading from local drives on the same node running the VM (owner node).
- Avoids unnecessary network hops — data is served directly from NVMe/SSD cache or HDD on the same host.
- CSV (Cluster Shared Volumes) redirector coordinates access; if data is local, reads bypass the network entirely.
- Cache-hit reads (NVMe/SSD) are served at full local bus speed — typically sub-millisecond latency.

🖍️ Write Path — Distributed for Resiliency

- Writes go through the cache tier first (write-back cache on NVMe/SSD).
- Data is simultaneously mirrored to drives on at least one (2-way) or two (3-way) other nodes over RDMA.
- Write is acknowledged only after the required mirror copies are confirmed — ensuring durability.
- RDMA (Remote Direct Memory Access) makes cross-node writes low-latency (~200–500µs typical).

💡 *Key insight: S2D combines local read performance with distributed write resiliency — delivering enterprise-grade throughput whether deployed standalone or as part of a broader storage strategy alongside SAN.*

Recommended Configurations by Scenario

Scenario	Topology	Drive Config	Resiliency	Notes
Branch / Edge / ROBO	2-Node	All-flash or HDD+SSD	2-way mirror	Cloud witness recommended; low cost, easy to manage
General Virtualization	3–4 Node	NVMe + SSD or HDD	3-way mirror	Sweet spot for most VDI and VM workloads
SQL / Mission-Critical	4–6 Node	All-NVMe	3-way mirror	Prioritize low latency; dedicated NVMe for SQL data files
Large-Scale / SOFS	8–16 Node	NVMe + HDD	Mirror-Accelerated Parity	Maximize usable capacity; rack-aware fault domains
SAN Coexistence / Migration	3–4 Node	Match existing + new	2 or 3-way mirror	Phased migration; S2D and SAN run concurrently
Hybrid / Azure-Connected	3+ Nodes	Flexible	2 or 3-way mirror	Arc-enabled; Azure monitoring and governance included

These are general guidance recommendations — actual configuration should be validated on workload requirements and environment.

Setup Hints

Before You Deploy	During Setup	Day 2 — Ongoing
Validate RDMA first Network misconfiguration (RoCE v2 or iWARP) is the #1 root cause of performance issues. good confirm before standing up the cluster	**Let S2D auto-assign tiers** Allow S2D to automatically designate cache vs. capacity drives (manual overrides are rarely needed and often cause problems)	**Monitor repair jobs actively** A stuck repair job can appear healthy and check for actual progress, not just active status. Look for Health Service actions, especially with maintenance mode.
Plan drive symmetry Mixed drive types across nodes cause pool expansion failures later. align your hardware across all nodes upfront	**Verify Storage Bus health first** Confirm the Storage Bus layer is healthy before creating the pool. don’t skip this validation step	**Don’t add drives mid-repair or rename nodes** Wait for the pool to fully stabilize after a failure before expanding capacity. Avoid renaming cluster nodes when using S2D storage.
Choose the witness type Cloud witness for edge and branch; file share witness for air-gapped environments	**Start with 3-way mirror** Use 3-way mirror for production workloads from day one. migrating from 2-way mirror later is disruptive	**Test failover before you need it** Run failover scenarios during a planned maintenance window and not during an incident

Pros: Why Choose S2D?

Flexible Deployment

Deploy standalone or alongside existing SAN/NAS. Modernize at your own pace without disrupting current investments — and reduce infrastructure costs as you scale.

Seamless Scalability

Start with 2–4 nodes, scale to 16+. Add nodes or drives without downtime — grow capacity and performance as your needs evolve.

Built-in Resiliency

2-way or 3-way mirror, single or dual parity. Rack-local reads and Failover Clustering keep workloads running through drive, node, or network events.

High Performance

NVMe cache tier delivers sub-ms latency. RDMA networking and rack-local reads enable high throughput with minimal cross-node traffic.

Native Windows Integration

Built into Windows Server & Azure Stack HCI — same tools, APIs, and management as any Windows environment.

Azure Arc Integration

Manage on-premises and edge clusters from the Azure portal. Enables hybrid monitoring, governance, and cloud billing across your entire footprint.

Cons: Challenges & Limitations

Complexity

Storage, networking, and clustering expertise required for configuration, tuning, and troubleshooting. Manageable with the right partner, but admins should plan for a learning curve.

Network Dependency

Performance relies on fast, low-latency RDMA networking. Misconfigured or congested networks can significantly impact throughput and latency.

Rebuild Time on Failure

When a drive or node fails, re-mirroring across nodes takes time. Large HDD-based clusters may see multi-hour rebuild windows during recovery.

Minimum Node Requirement

2 nodes minimum, 3 recommended for full resiliency. Scaling in fixed increments can mean some capacity planning overhead.

Drive Capacity Overhead

Resiliency takes space: 3-way mirror = 3× raw capacity needed. Efficiency tradeoff — not all raw storage is usable storage.

Workload Sensitivity

Mixed workloads (IOPS-heavy + throughput-heavy) can compete for cache. Noisy-neighbor effects require careful workload planning.

What to Watch For: Common S2D Support Scenarios

Virtual Disk Health

The most common support scenario — virtual disks becoming degraded, detached, or inaccessible following a disk or network event.

What you see: VMs go offline, repair jobs appear stuck, virtual disk shows ‘No Redundancy’ or ‘Detached’ status

What helps: Catch disk failures early with health monitoring; don’t ignore warnings. Act on alerts before they cascade.

Maintenance Mode & Disk Retirement

Placing nodes or disks into maintenance mode without following the correct sequence is a frequent trigger for cascading issues.

What you see: Disks stuck in ‘Removing from Pool’ or ‘Draining’ state; virtual disks become unhealthy; subsequent operations blocked

What helps: Always follow the documented maintenance sequence — never skip steps under time pressure. Validate health before and after each action.

Pool Expansion & Rebalancing

Adding nodes, drives, or volumes can fail when hardware symmetry requirements aren’t met, or when a repair job is already in flight.

What you see: Volume extension fails, ‘activity not supported’ errors, pool rebalance doesn’t complete, new disks not recognized

What helps: Ensure consistent hardware across nodes before expanding. Let repair jobs complete and the pool stabilize before making changes.

Performance & Latency

Persistent high latency and slow VM performance are often symptoms of an underlying configuration issue rather than a hardware limitation.

What you see: High disk queue lengths, slow VM response, IO wait warnings in event logs, background jobs impacting foreground workloads

What helps: Validate RDMA networking, confirm cache tier is active and healthy, and check whether background repair jobs are competing with production IO.

Areas of Focus for Window Server vNext

Areas of active attention to reduce operational burden — not product commitments or guarantees of future functionality.

Manageability & Day 2 Operations

- Clearer diagnostics and event visibility so admins understand what is happening — not just that the system is up
- Better guided recovery paths for disk, metadata, and cluster events to reduce reliance on escalation

Repair & Recovery Workflows

- Reducing time-to-resolution for degraded or detached virtual disks after node or network events
- Improving transparency and predictability of background repair jobs so admins can trust the system is making progress

Scalability & Growth

- Smoother pool expansion and volume creation, with clearer guidance when hardware symmetry requirements block growth
- More predictable rebalance behavior as environments scale and drive configurations evolve over time

⚠️ *These reflect patterns observed across support engagements. These are not product commitments, roadmap items, or guarantees of future functionality.*

Q&A

Questions & Discussion

Thank you for joining!

For more on DataON + S2D solutions:
www.DataON.io

DataON | Microsoft Azure