

# Welcome to the IBM TechXchange Community

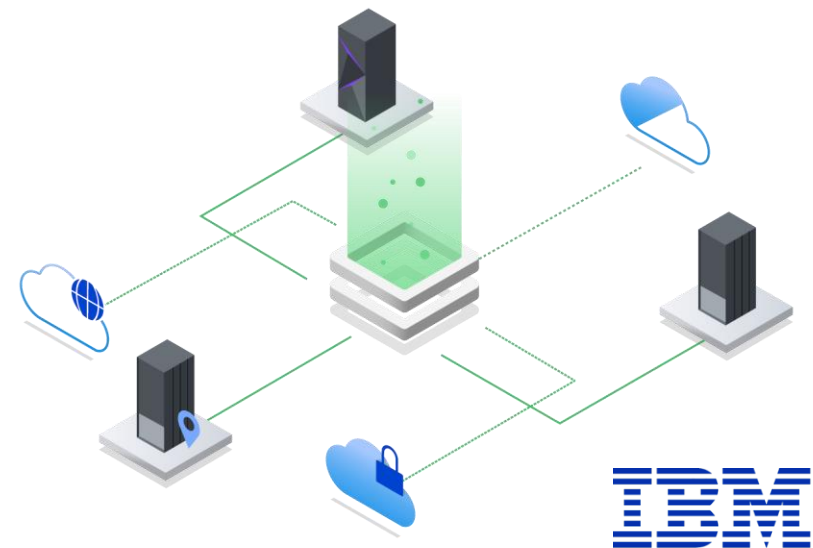
Connect and engage to get answers, discuss best practices, and continually learn more about IBM solutions.



## Planning Infrastructure for Data-Intensive Science



Chris Maestas  
IBM CTO, Data and AI Storage Solutions  
Chief Troublemaking Officer



## The world is changing ...



# 2005

Luca Bruno/AP



## The world is changing ...



# 2005

Michael Sohn/AP



# 2013



# IBM Storage Scale Technology Exchange

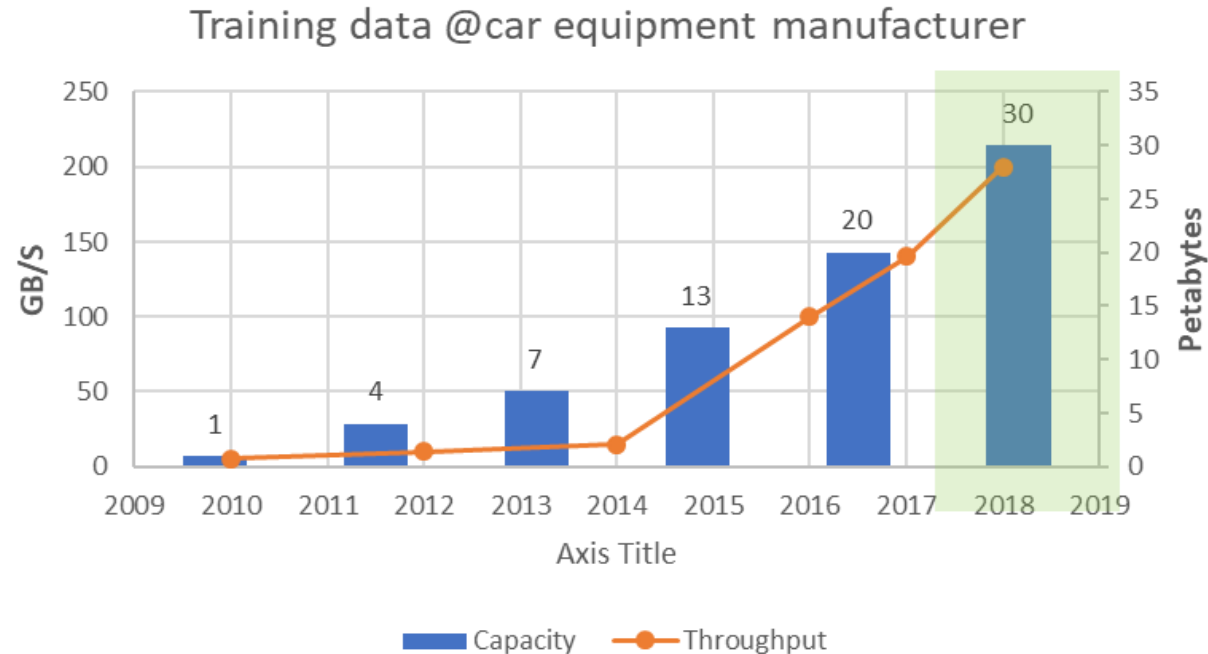
The world today ....

# 2025



# IBM Storage Scale Technology Exchange

## Scalable performance with enterprise functionality



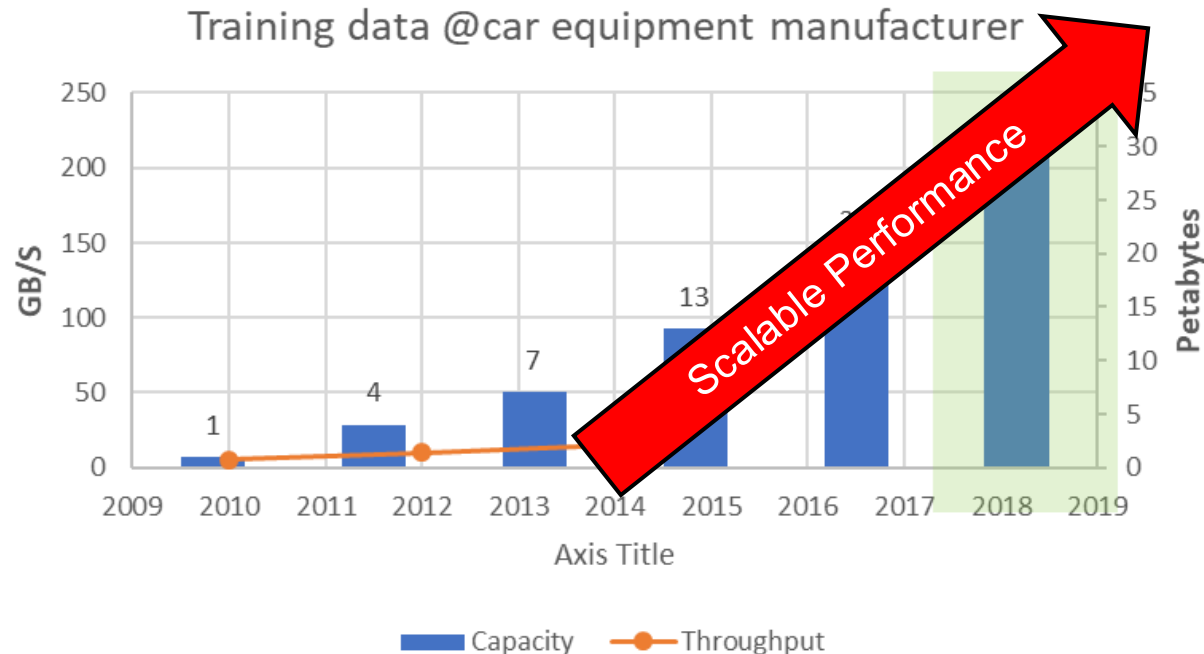
- **Scalable Capacity**

- Over the years, acquired data sums up to 10s or 100s PB of data.
- In the early years NAS/Scale-out NAS is capable to handle the required capacity and performance.

- ➔ Capacity: Data volumes are growing since decades.
- ➔ Throughput: With new workloads such as AI/ML/DL the growth of data volume implicitly triggers a growth of performance requiring scalable performance.

# IBM Storage Scale Technology Exchange

## Scalable performance with enterprise functionality



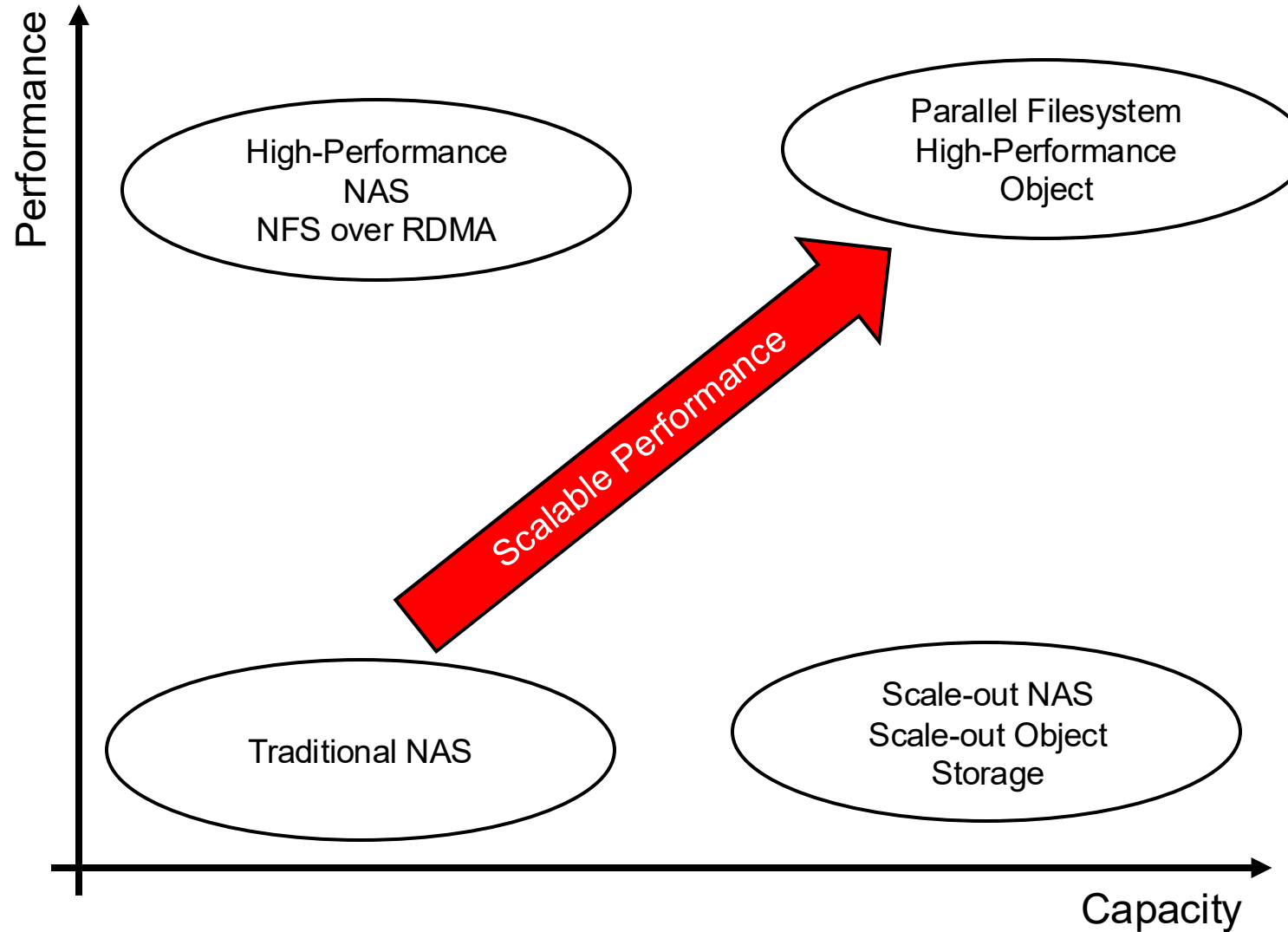
- ➔ Capacity: Data volumes are growing since decades.
- ➔ Throughput: With new workloads such as AI/ML/DL the growth of data volume implicitly triggers a growth of performance requiring scalable performance.

- **Scalable Capacity**
  - Over the years, acquired data sums up to 10s or 100s PB of data.
  - In the early years NAS/Scale-out NAS is capable to handle the required capacity and performance.
- **Scalable Performance**
  - Over the years the performance requirements typically grow because of
    - advancements in sensor technology increase data volumes and data rates,
    - the training of refined AI models requires high-speed access to all acquired data,
    - improved data processing pipelines run more ingest, analysis and training jobs in parallel.
  - At this stage customers need to transition from NAS/Scale-out NAS to a parallel filesystem with enterprise Data Management functionality to meet the growing performance requirements.



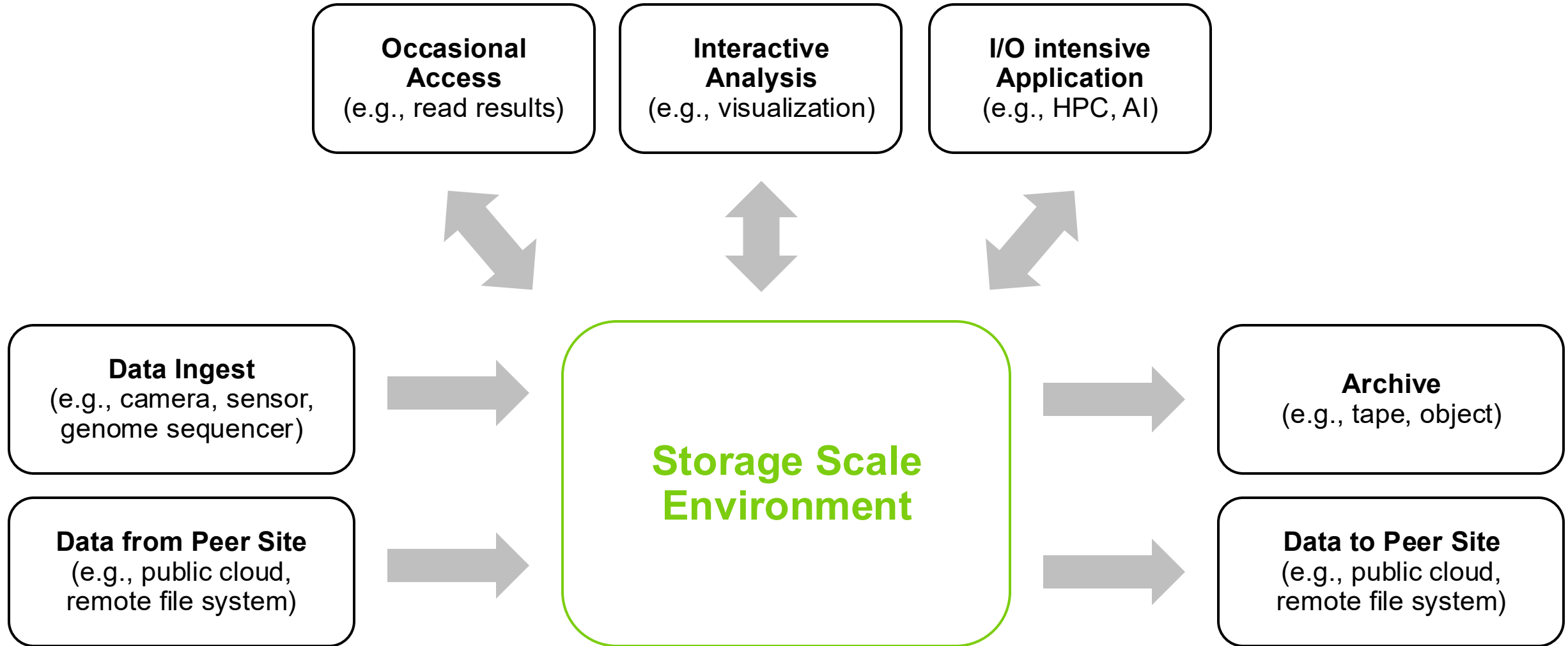
# IBM Storage Scale Technology Exchange

## Scalable Capacity vs. Scalable Performance



# IBM Storage Scale Technology Exchange

## Data Intensive Science and Engineering





## Skills

---

- More and more businesses require to gain insight from huge amounts of unstructured data.
- Respective data intensive Workflows trigger Workloads which force customers to adopt infrastructure and cloud-like operating models that they did not used in the past.
- In particular traditional block storage and NAS filers have limitations to meet the respective scaling and performance requirements.
- Customers need to plan for the respective system administration and end-to-end skills to architect, implement, operate and troubleshoot HPC, cloud and hybrid environments.
- Customers new to such environments should start with small environments and a limited set of features and then incrementally grow as they gain more experience in real production environments.



Skills

# IBM Storage Scale Technology Exchange

## Contrasting file-based environments

		Parallel File System (POSIX)	Network Attached Storage (NAS)
Workload	Applications	Broad range of scientific applications, big data and analytics, ML/DL, parallel applications	Broad range of office applications, roaming profiles, etc.
	Scalable Performance	High (large data sets, fast metadata operations, high throughput, low latency)	Medium/Low (average performance and scaling needs)
	Consistency	Strict (Node see updates from remote nodes immediately)	Eventual (Node may see updates from remote nodes after a delay)
Infrastructure / Features	Access to clients	Controlled (Limited number of privileged users)	Wild west (End user have root access to laptops, etc.)
	Client OS Interoperability	Limited (number of operating systems, number of versions, number of architectures)	Flexible (Broad range of different OS versions including very old OS versions and architectures)
	Predominant Client OS	Linux	Linux, Windows, macOS
	Protocol	Proprietary (e.g., Storage Scale NSD)	Standard (NFS, SMB)
	Number of clients	Thousands ( <16k for Storage Scale)	Tens of thousands
	Network	Private Cluster Network	Shared Data Center Network
Skills	Deployment Model	Software Defined Infrastructure	Hardware Appliance
	Client Software	Additional software package for access to parallel filesystem	S3, NFS and or SMB are included in the operating system
	Admin Skills	System administrators (Deep skills in Linux, networking, system software, etc.)	Storage administrators (Mostly management of storage appliances)

# IBM Storage Scale Technology Exchange

## Contrasting computing paradigms

	Traditional IT	Traditional HPC	Data Intensive Workloads & Workflows
Objective	Enables core business processes	Enables compute and data intensive research	Enables core business processes
Teams	Different teams for network, servers, storage, applications, infrastructure services, etc.	Typically dedicated team for HPC system in addition of team for Traditional IT	Integrated team for all aspects of the IT solution
Responsibilities	Each team is responsible for their component / application.	One team is responsible for HPC system	One team is responsible for software-defined infrastructure including the underlying hardware and the integration in corporate IT landscape
Skill	Component specific skill. Reaches-out to other teams for matters related to other components.	End-to-end skill for HPC system. Reaches out to Traditional IT for selected infrastructure services and remote access to HPC system.	End-to-end skill for corporate wide software-defined infrastructure. Reaches out to Traditional IT for selected infrastructure services and external networks.



# IBM Storage Scale Technology Exchange

Contrasting computing paradigms

	Traditional IT	Traditional HPC	Data Intensive Workloads & Workflows
Objective	Enables core business processes	Enables compute and data intensive research	Enables core business processes
Teams	Different teams for network, servers, storage, applications, infrastructure services, etc.	Typically dedicated team for HPC system in addition of team for Traditional IT	Integrated team for all aspects of the IT solution
Responsibilities	Each team is responsible for their component / application.	One team is responsible for HPC system	One team is responsible for software-defined infrastructure including the underlying hardware and the integration in corporate IT landscape
Skill	Component specific skill. Reaches-out to other teams for matters related to other components.	End-to-end skill for HPC system. Reaches out to Traditional IT for selected infrastructure services and remote access to HPC system.	End-to-end skill for corporate wide software-defined infrastructure. Reaches out to Traditional IT for selected infrastructure services and external networks.

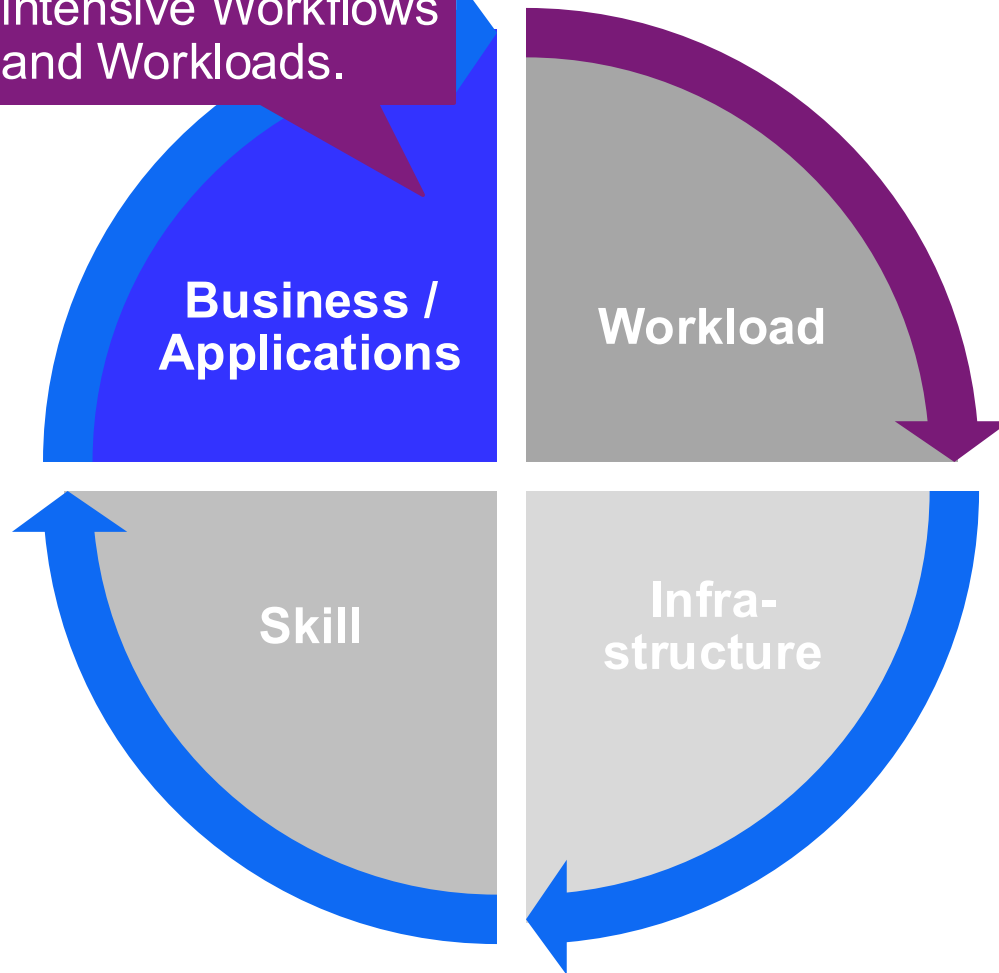
➔ Customers adopting Data Intensive Workflows and Workloads need to plan for acquiring the respective skills.

➔ Engage a partner who guides you on your journey to Data Intensive Science at scale.

# IBM Storage Scale Technology Exchange

## Choosing the right solution

Adoption of data intensive Workflows and Workloads.



- The **business requirements** determine the required **applications**.
- The **applications determine** the generated **workload**.
- The **workload determines** the required **infrastructure**.
- The **infrastructure determines** the required **skills**.
- The available **infrastructure and skills determine** the capability to support the **business**.

# IBM Storage Scale Technology Exchange

## Summary

---

- **Data Intensive Science** describes research and engineering efforts where the storage, the management and the analysis of acquired data requires special considerations to enable the scientific effort.
- Data Intensive Science includes but is not limited to **Analytics & AI**.
- **Storage Scale and Scale System** provides scalable Software Defined Storage which can meet the evolving workload requirements of Data Intensive Science.
- Some of them **required years** to transition to a new architecture. Now they have 10s of PBs.
- Organizations who are adopting data-intensive science have **acquired new skill sets**.



