SNIA | CLOUD STORAGE
CSTI | TECHNOLOGIES

Cloud Storage and Big Data, A Marriage Made in the Clouds

Live Webcast

October 21, 2021

10:00 am PT / 1:00 pm ET

Today's Presenters

Moderator:
Chip Maurer
Senior Principal Engineer
Dell Technologies

Presenter:
Andy Longworth
Solution Architect
HPE

Presenter:
Vincent Hsu
VP, IBM Fellow, and CTO for Storage and Software Defined Infrastructure
IBM

SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements Incorporate them into standards and programs

Collaborate with other industry associations

SNIA Legal Notice

The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.

Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

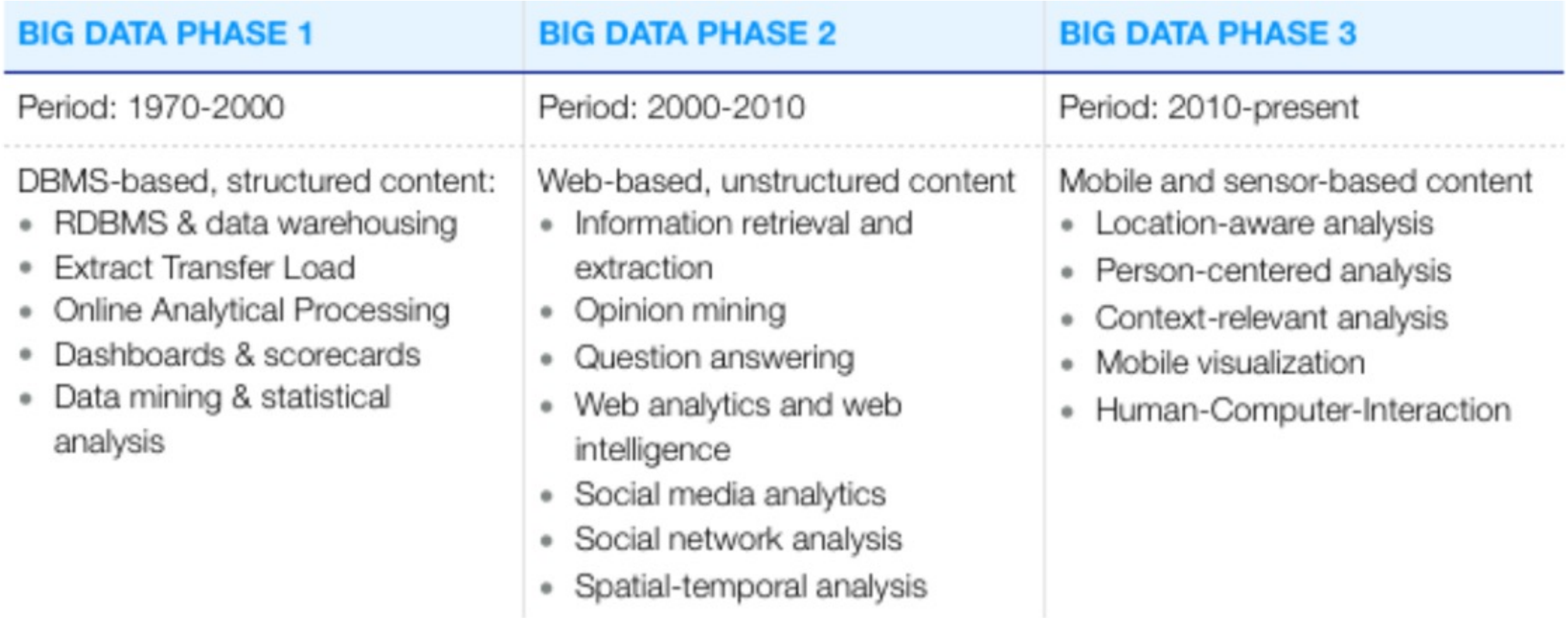
Agenda

- History of Big Data
- Current state
- Modernization challenges
- Evolving workloads, processing outside of data center
- Look towards the future

History of Big Data

History of Big Data

Enterprise Big Data Framework (https://www.bigdataframework.org/short-history-of-big-data/)



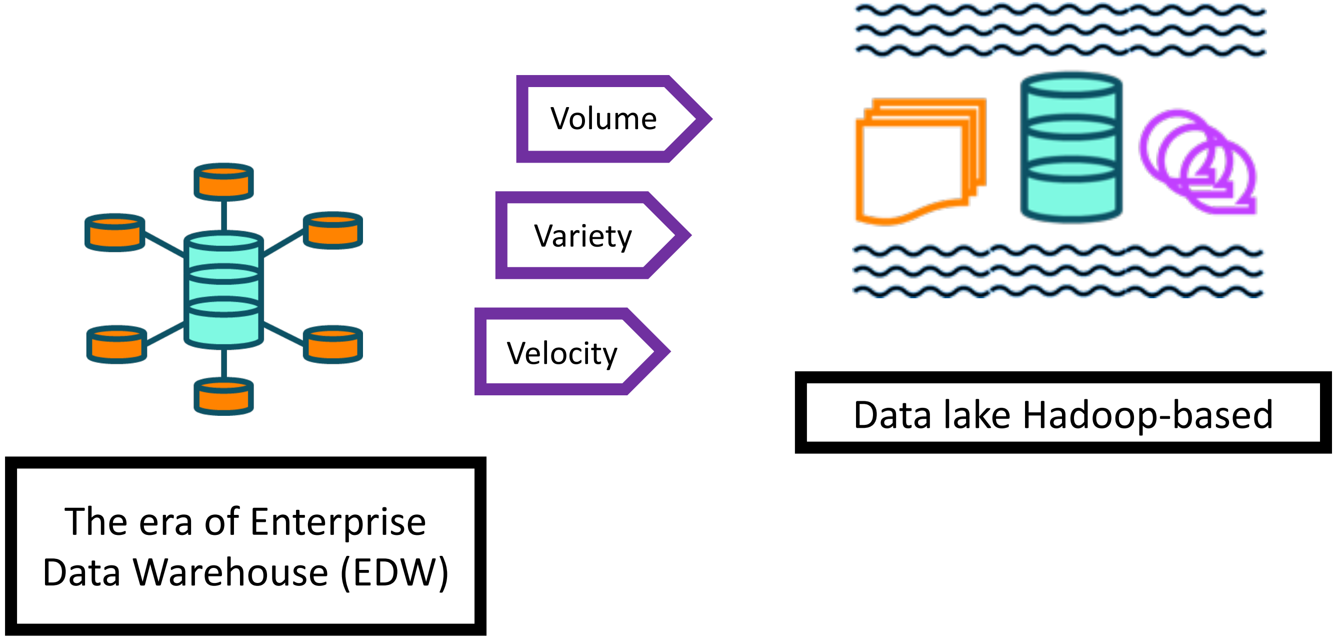
BIG DATA PHASE 1	BIG DATA PHASE 2	BIG DATA PHASE 3
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
DBMS-based, structured content: • RDBMS & data warehousing • Extract Transfer Load • Online Analytical Processing • Dashboards & scorecards • Data mining & statistical analysis	Web-based, unstructured content • Information retrieval and extraction • Opinion mining • Question answering • Web analytics and web intelligence • Social media analytics • Social network analysis • Spatial-temporal analysis	Mobile and sensor-based content • Location-aware analysis • Person-centered analysis • Context-relevant analysis • Mobile visualization • Human-Computer-Interaction

Block, NFS, POSIX **HDFS** **Unified storage (Object and File)**

Current State

Evolution of Data Platforms

- Everything from Enterprise Data Warehouses to Hadoop based Data Lakes
- No one size fits all
- Emergence of cloud services
- Picking the right system for the right workload
    - Structured vs Unstructured
    - Batch vs Real-time
    - On-premises vs Cloud



The era of Enterprise Data Warehouse (EDW)

Volume

Variety

Velocity

Data lake Hadoop-based

The Five Vs of Big Data

- Started as 3 Vs
  - Volume: the huge amount of data that is produced every day
  - Variety: diversity of data, both types and sources
  - Velocity: the speed with which the data is generated
- Additional Vs
  - Veracity: is the authenticity and credibility of data
  - Value: transforming data into value for the business

Is Hadoop Dead?

For

- Cost: running on commodity hardware
- Batch analytics
- Availability through fault tolerance
- Spark on Hadoop

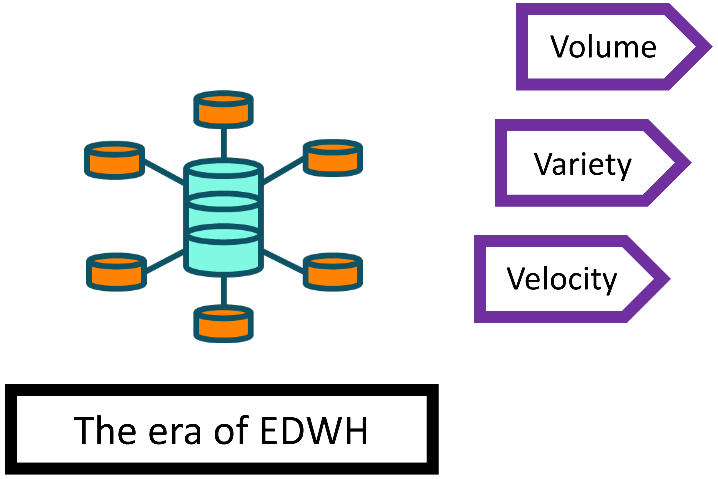
Against

- Inefficient for small datasets
- Real-time analytics
- Cloud alternatives
- Lack of integration with cloud services such as S3

- Merge of Hortonworks and Cloudera
- Cloudera Enterprise 6.2 & 6.3 EOL March 2022
- Hortonworks Data Platform 3.1 EOL December 2021

Modernization Challenges

Evolution of Data Platforms

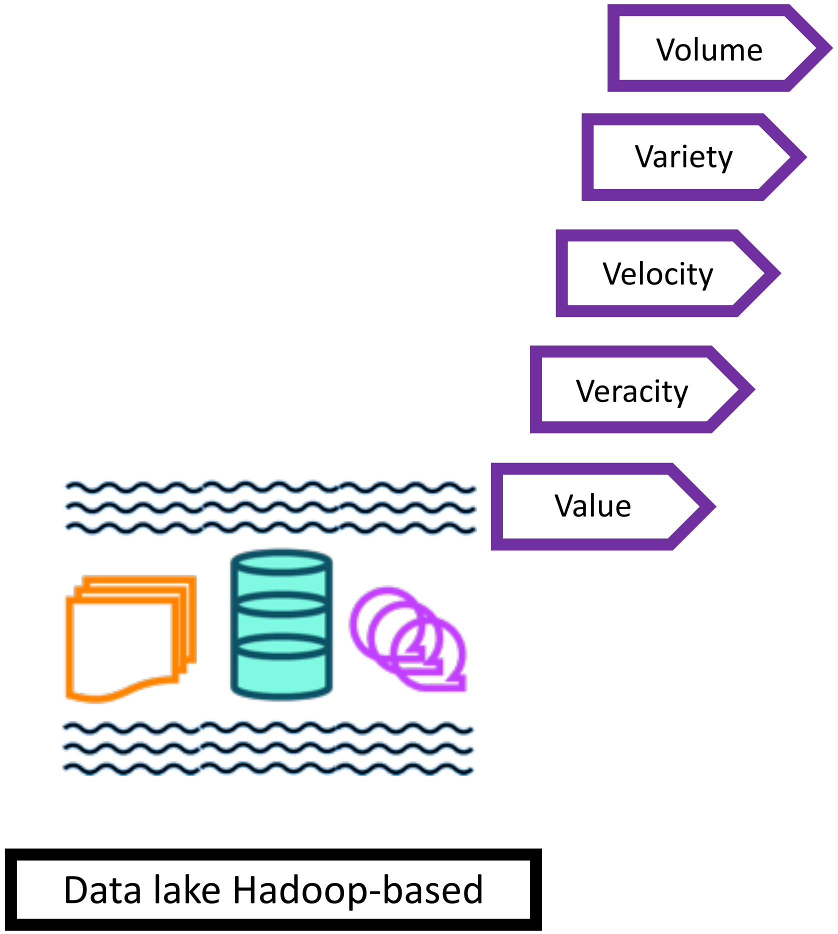


Volume

Variety

Velocity

The era of EDWH



Volume

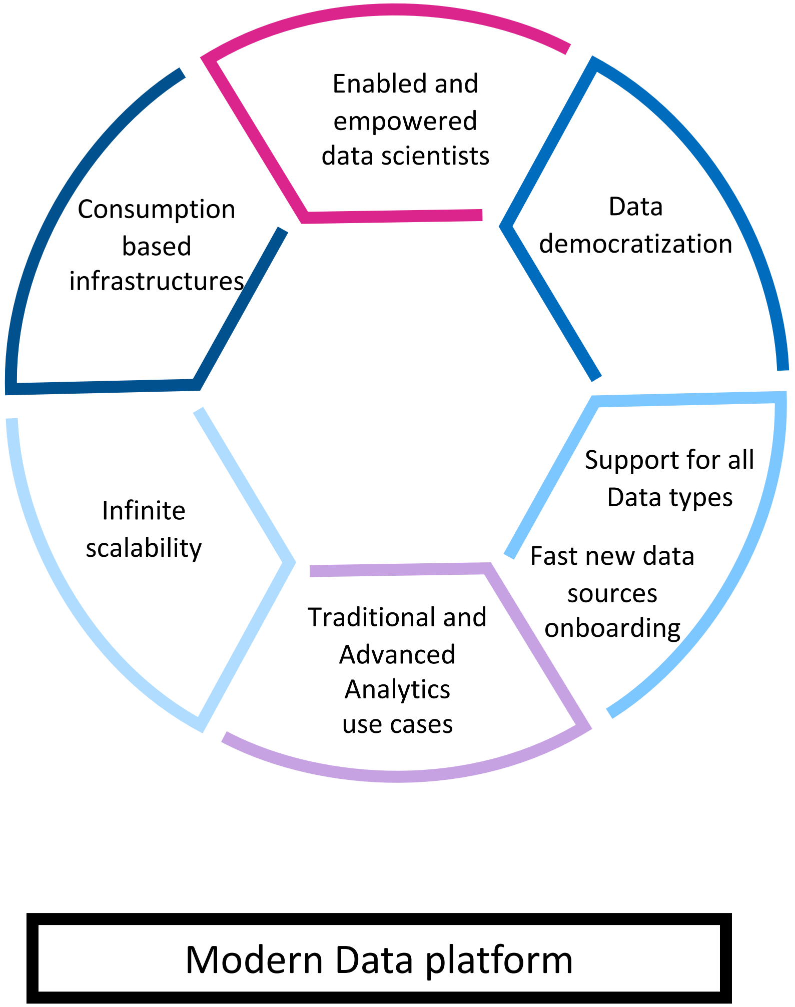
Variety

Velocity

Veracity

Value

Data lake Hadoop-based



Enabled and empowered data scientists

Data democratization

Support for all Data types

Fast new data sources onboarding

Traditional and Advanced Analytics use cases

Infinite scalability

Consumption based infrastructures

Modern Data platform

Challenges – Questions Need to be Answered

- What workloads do we need to support?
  - Batch vs Streaming
  - AI vs Traditional analytics
- What protocols need to be supported?
  - HDFS vs S3 vs …
- Where best to run your data platform?
  - On-premises vs Cloud vs Hybrid
- Data considerations
  - Gravity
  - Sovereignty
  - Compliance
  - Security

The Skills Challenge

- Bringing in new technologies
  - Selecting the right tools for the right workload out of the huge number of choices
  - Containerization
- How to get support for new tools and technologies?
  - Fast moving ecosystem
  - Many open source projects
- Where do we find the people for these platforms and workloads?
  - In demand skills
  - Upskilling existing teams

Building for the Future

- Can we futureproof your data platform?
- How to not make the same mistakes again?
- Does everything need to move to the cloud?

Modernization Challenges

Data sprawl

- Data is spread across multiple on premises and public cloud locations
- Data is accessible via multiple protocols (NFS, HDFS, S3)
- Finding relevant data
- Managing multiple copies of data

Data governance and data gravity

- Data classification
- Data sovereignty
- Regulatory compliance
- Not all data can move to public cloud--leverage data catalog to ensure compliant data movement and data placement
- Expensive lift and shift

Performance, scalability, and durability

- Bring data closer to compute -- long latencies when accessing data from data lake storage
- Cost prohibitive to keep all data in high performance storage tier
- With non-persistent cache, all data must be reloaded in the event of failure
- Elasticity and cloud bursting
- Computational storage

Data security

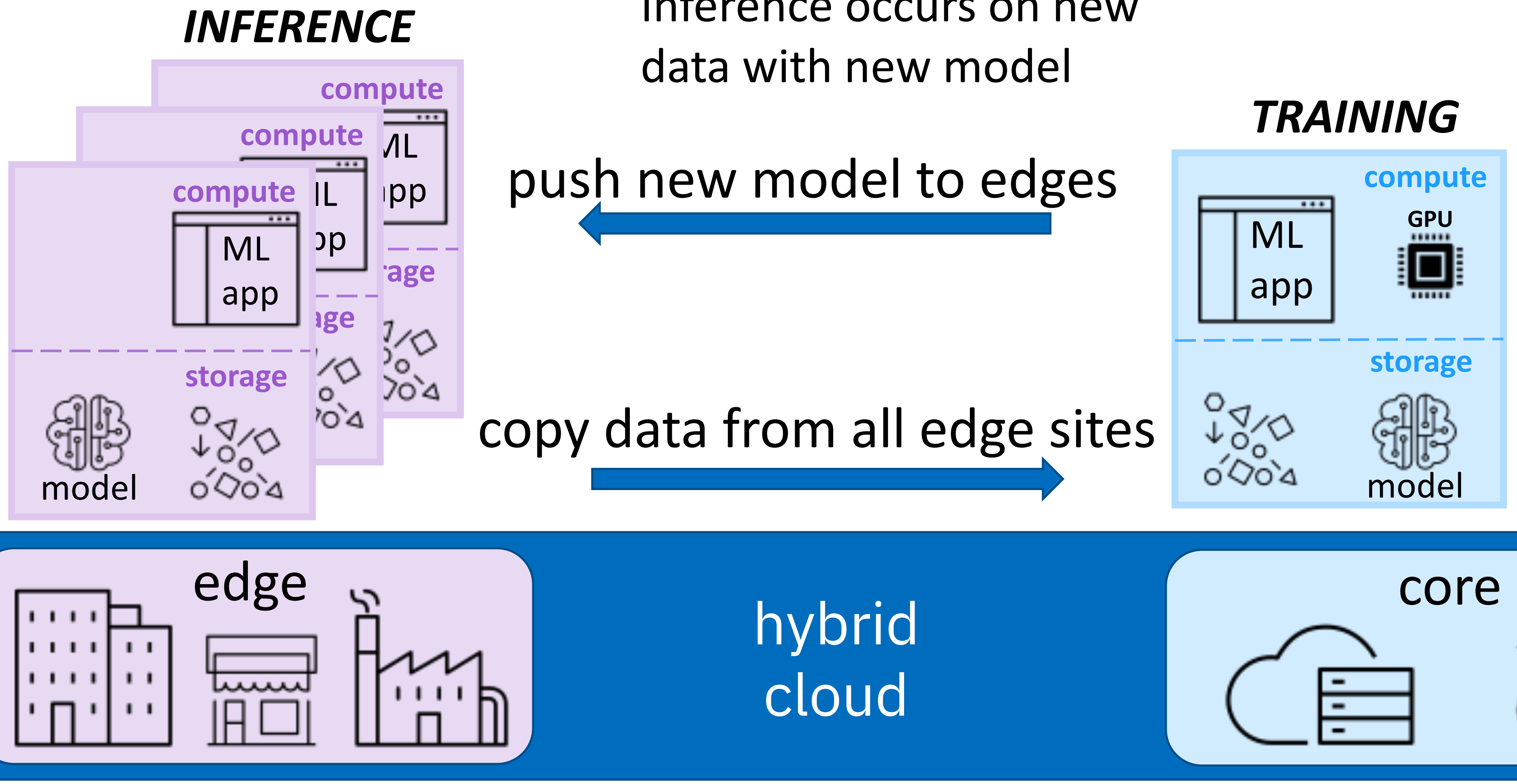
- Encryption of data in flight and at rest
- Hybrid key management
- Role based access control

Evolving Workloads

Machine Learning with Data from All Edge Sites

What usually happens today…

Inference occurs on new data with new model



INFERENCE

compute

ML app

storage

model

TRAINING

compute

ML app

GPU

storage

model

push new model to edges

copy data from all edge sites

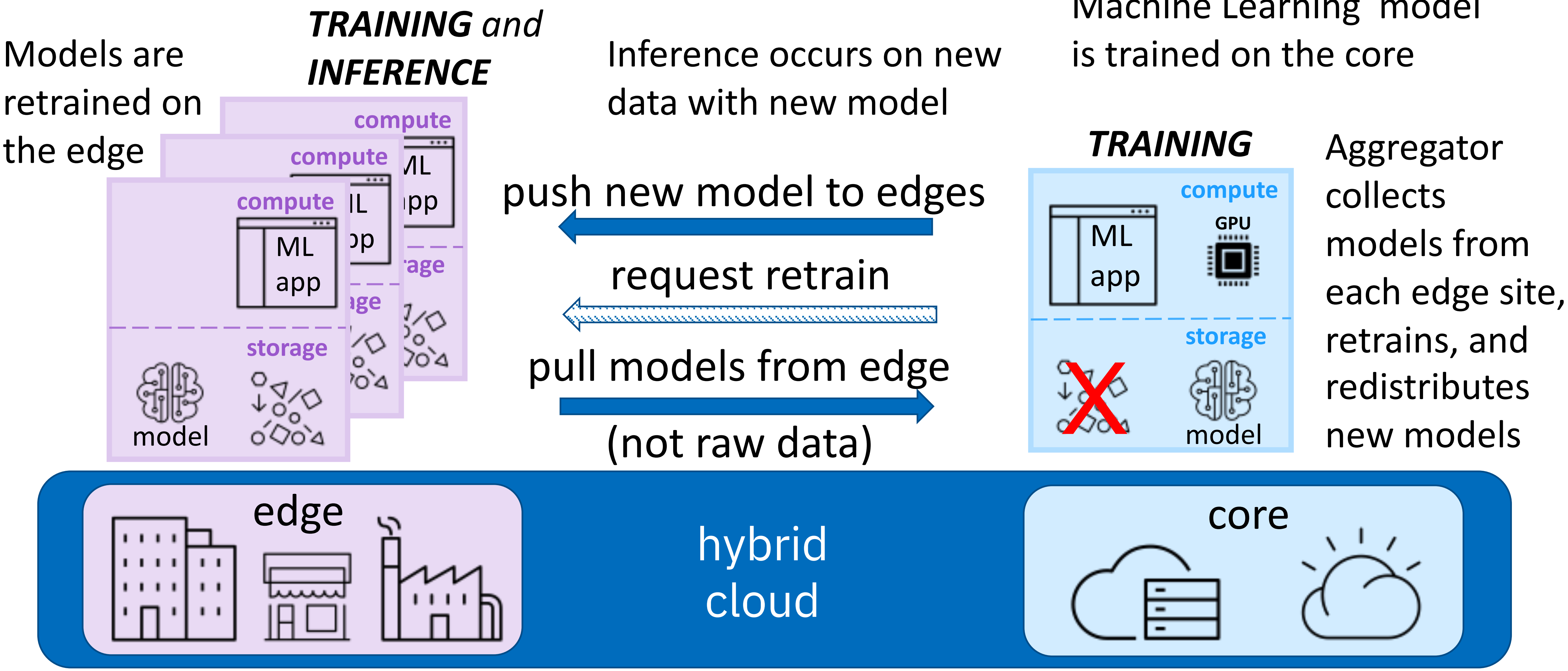
edge

hybrid cloud

core

Machine Learning model is trained at the core

Federated Learning



Models are retrained on the edge

TRAINING** and **INFERENCE

Inference occurs on new data with new model

Machine Learning model is trained on the core

TRAINING

Aggregator collects models from each edge site, retrains, and redistributes new models

push new model to edges

request retrain

pull models from edge (not raw data)

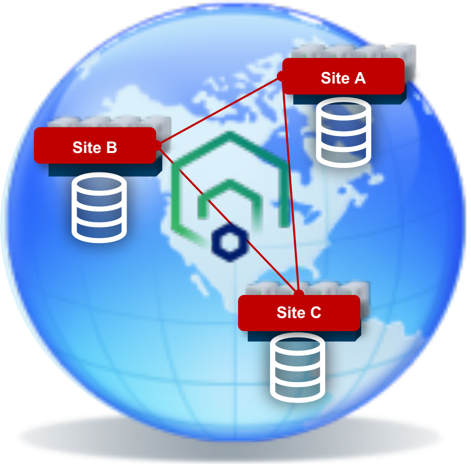
Value of Federated Learning

- Improve model training across locations
- Address data privacy, locality and security
- Adhere to regulatory compliance
- Tackle data volumes at lower cost and risk (e.g., minimize egress charges)

Looking Towards the Future

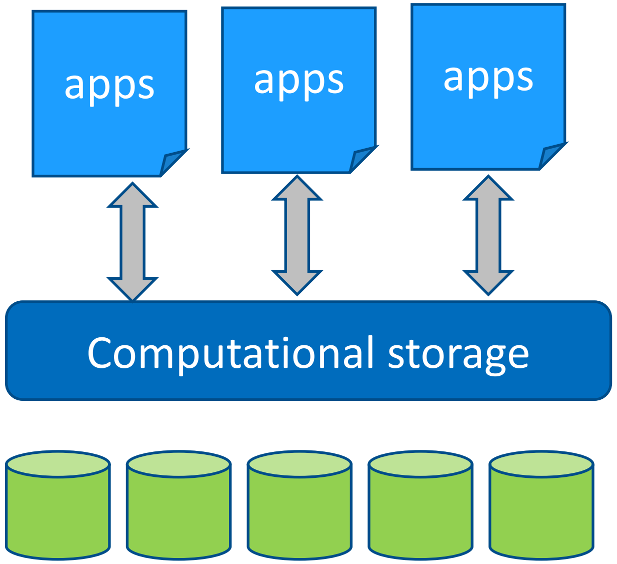
Look Towards the Future – Storage for the New Big Data

- True hybrid cloud data fabric



- Acceleration technology: FPGA, GPU, DPU, IPU,…

- Computational storage



Looking Towards the Future

- Data and Analytics as a Core Business Function
- Data and Analytics at the Edge
- Operationalization of AI
  - DevOps, AI Ops, ML Ops
- The Data Lakehouse
  - Bringing together the best of the data warehouse and data lakes

Thanks for Viewing this Webcast

Please rate the webcast and provide us with feedback

This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library

A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/

Follow us on Twitter @SNIACloud

Thank You

Questions?