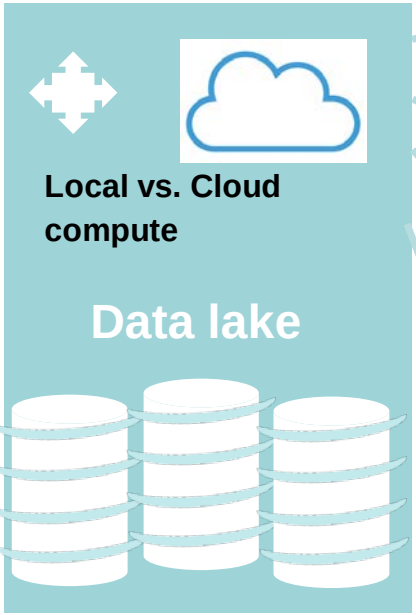
SDC 18
September 24-27, 2018
Santa Clara, CA
www.storagedeveloper.org

Data Architecture for Data-driven Enterprises

Deepti Aggarwal, Rukma Talwadker
NetApp, Inc.

Big Data and AI/ML for insights



Data Filtration/ cleaning

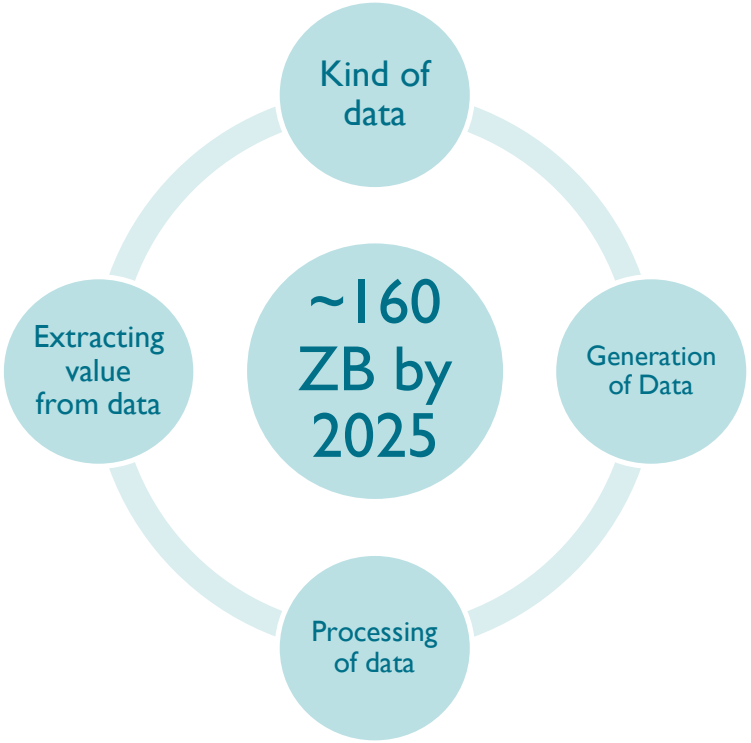
Data Processing
(real / near real / batch)

Data Analytics
Machine Learning
Deep Learning

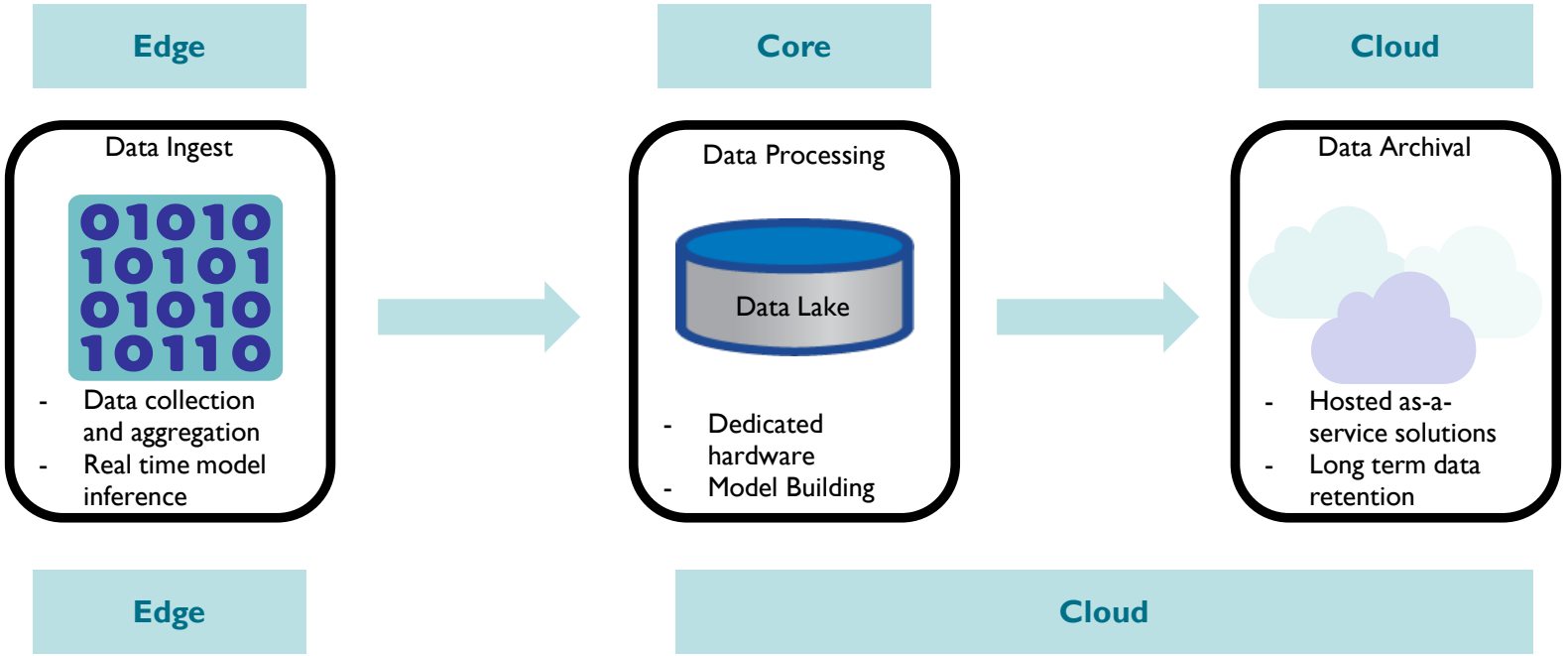
Data Visualization

Change in the Data Lifecycle

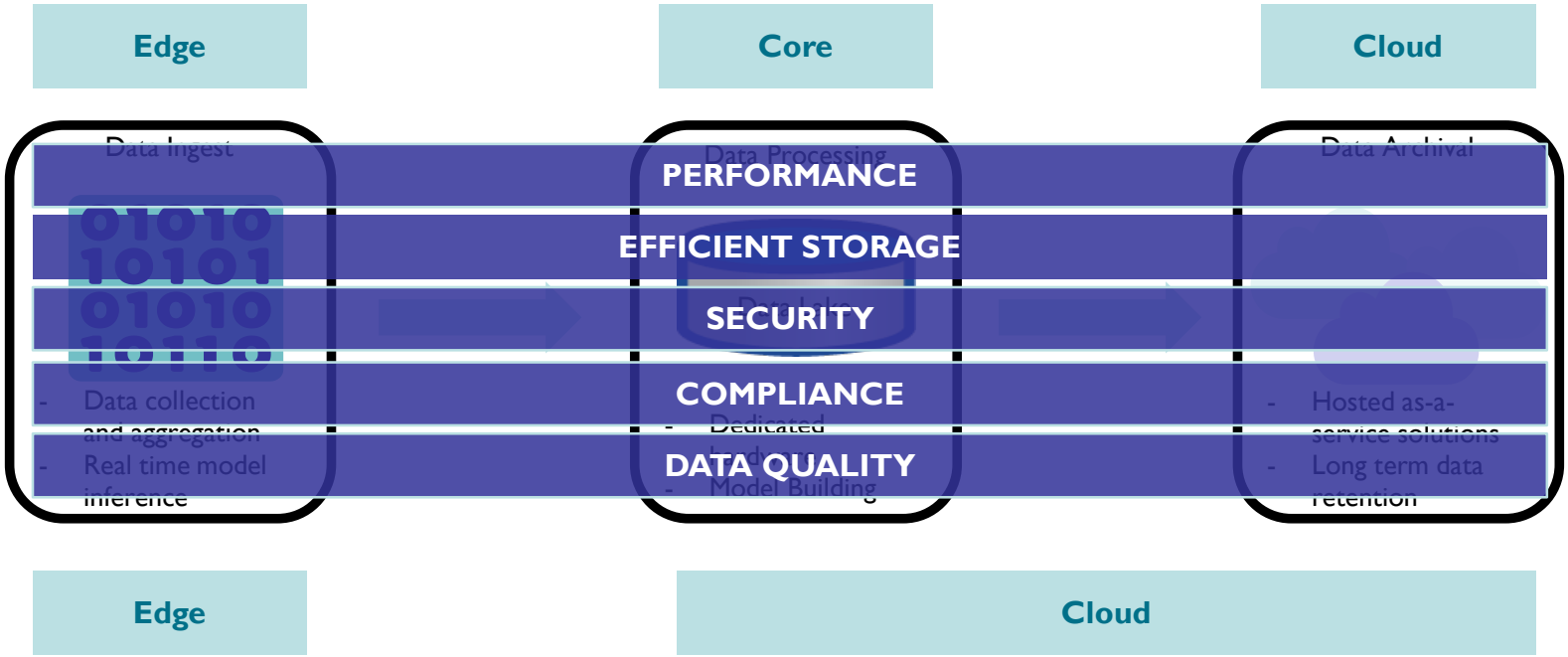
Warrants the need for a re-look at Data Architecture



Evolving Data Pipeline



Evolving Data Pipeline



Real Time Insights

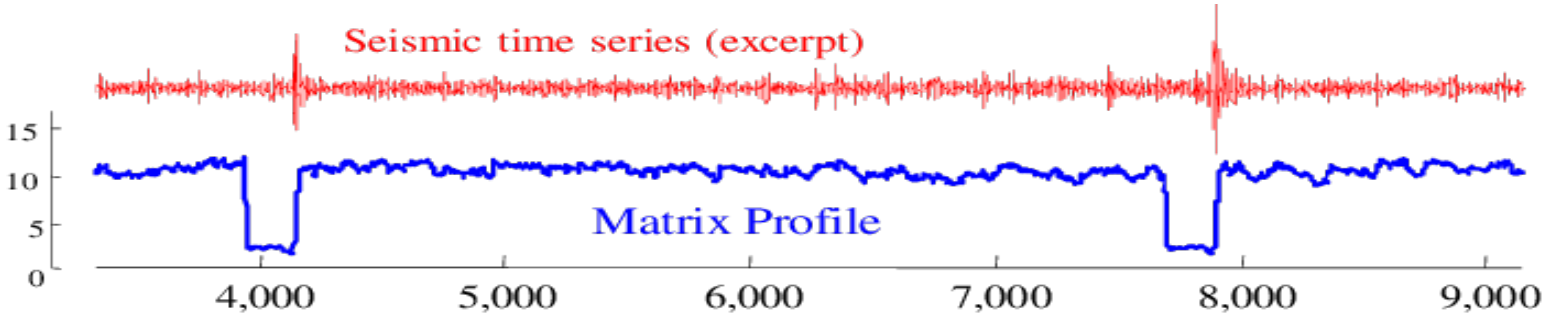
Data Ingestion & Transfer

Storage Efficiency over the Wire

- WAN transfer from the edge to the core/cloud
- Two views:
  - Transfer limited data
    - ‘Useful’ data
    - Deviations from ‘normal’
  - Transfer all data
    - Keep everything
    - Quantifiable data loss/approximation

Matrix Profiles: An anytime algorithm for time series analysis

http://www.cs.ucr.edu/~eamonn/MatrixProfile.html



- Anytime Algorithm
  - Consistent result at any iteration
- Scales with hardware
- Helps identify recurring sub-sequences, anomalies, change points, etc.
- Time series represented as a sequence of repeating patterns → data reduction

Metadata Enrichment: Quality, type, content, behavior….

- First point of data entry into the pipeline
- What tags to add?
  - Data Quality
    - Completeness, accuracy, timeliness
  - Kind of data
    - File extension based, content based
  - Behavior
    - Anomalous, normal
  - Lineage, auditing
- Aids data workflows down the pipeline
- Kafka added support for enriching streams with custom metadata

Security is a major concern at the edge

- Data privacy
- Access controls and perimeter security (firewalls)
- Too many end-points
- What if device gets hacked/stolen?
  - Means of authentication at the edge node
- Establish secure session based public key sharing
  - SSL Certificates

Private Cloud, Datacenter

Secure Domain

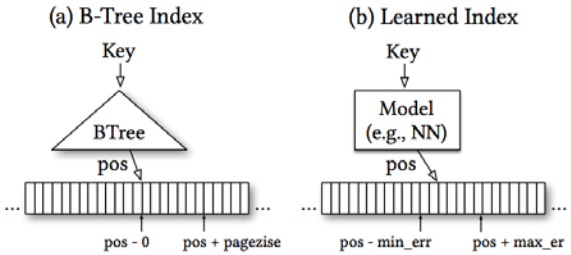
Parallel compute demands high performance from storage

- More advanced Neural Network variants
  - RNN, CNN etc
- Easily available platforms
  - Tensorflow, Keras, Theano…
- More and more data!!
- Operationalizing AI/ML workflows and prevalence of GPUs
- High, predictable performance from the underlying storage
- Model training stretching to days
  - A small factor slowdown in storage impacts cost heavily.

‘Smart’ Data Indexing to meet performance requirements

- Changing workloads
- New analytic applications
- Dynamic indexing based on query patterns
  - Which columns to index?
  - When to update the indices?
    - Cost of updating indexes is compute heavy
  - Which indices are obsolete?

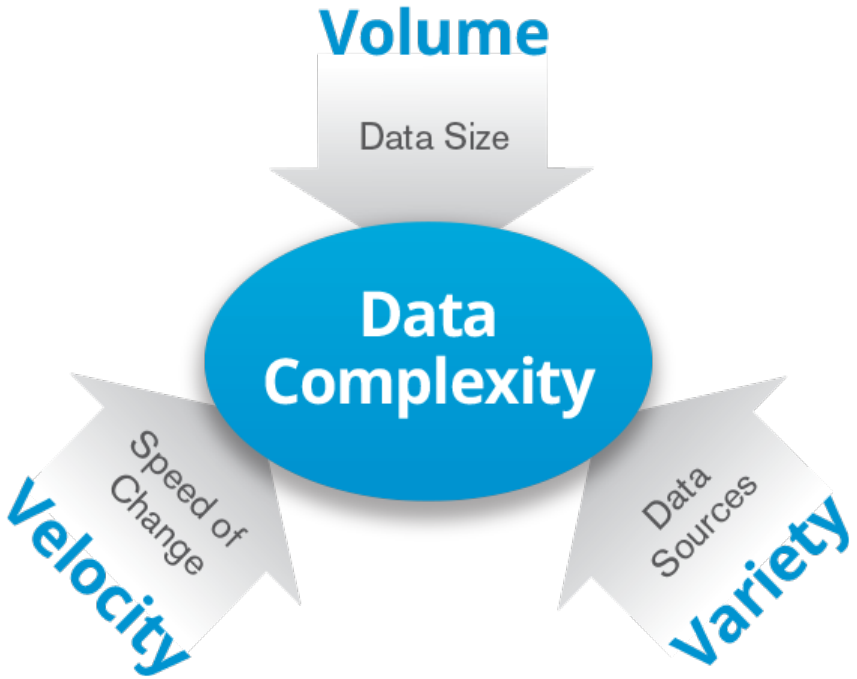
- Traditional indices do not take advantage of the data distribution or patterns in the data.
- ’Learned’ Indexes
  - https://arxiv.org/abs/1712.01208
  - Key idea: Structure of keys or sort order learnt, used to predict location of record



Futuristic scenario: AI/ML models replacing Core Data Management components

Aggregation and Preparation for transfer to Cloud

- Effective aggregation to present holistic view of data to applications
  - Raw data from millions of sensors
  - Continuous streams – transformations by analytic apps
  - Structured, unstructured, text, multimedia
- Data Anonymization
  - Data goes out of private domain
- Data preparation for transfer
  - Bundle which data together?
  - Encrypt and maintain Key Metadata Store
  - Compression



Data archival

Less trusted domain

Compliance governs data archival

- Compliance regulations dictate lifecycle of data
- Right to be forgotten/GDPR
  - Mandates need for efficient indexing
- Provenance/traceability
  - Metadata about every small modification
  - Data auditing
- Long term data retention
  - Cost
  - Unpredictable growth in data

Archival Tiering

- Reducing archival cost due to exponential growth in data
  - Erasure Coding over replication

- Archival Tiering
  - AWS archive tiers: S3, Glacier Expedite, Glacier Bulk
  - Tradeoff between retrieval times and $/GB

- Another way to minimize cost further: Choose different Erasure coding Schemes for different tiers
  - Tradeoff between storage overhead, read times, bandwidth and compute performance.
  - Coldest tier: minimal storage overhead, active archive: lower read times.

Cloud goes hand-in-hand with Data Security

- Store encrypted data in cloud with key metadata on-prem in protected environment
- VPNs – one mechanism to provide restricted access
- Enterprises' looking to run analytics on archives?
  - In presence of encryption?

- Ongoing work: encryption schemes which compromise on some security to preserve a particular aspect of data
  - Searchable encryption
  - Order-preserving encryption

Other Dimensions of interest

- Storage infrastructure suitable for each layer
  - Object stores v/s tradition NAS
- Implementation of Compliance regulations
- Global namespace across the pipeline

Key Takeaways

- Changing data lifecycle
- Data management issues remain the same – need a re-think in context of changing data workflows
- Performance, Compliance, Data Quality, Security and Efficient Storage are the key data management challenges that emerge