

Data Architecture for Data-driven Enterprises

Deepti Aggarwal, Rukma Talwadker ATG, NetApp May 25, 2018



© 2018 NetApp, Inc. All rights reserved.

Big Data and AI/ML for insights





Change in the Data Lifecycle

Warrants the need for a re-look at Data Architecture





Evolving Data Pipeline





Evolving Data Pipeline



Edge	Cloud





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 \rightarrow 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





NetApp



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
 - <u>https://arxiv.org/abs/1712.01208</u>
 - 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

Untrusted domain



Compliance governs data archival

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





Archival Tiering

- Reducing archival cost due to exponential growth in data
 - Tiering
 - 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



Putting it all together...

NetApp Data Fabric







- 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



NetApp

Thank You



Backup



Concluding Remarks

- Everything about data is changing
 - Types of data
 - Data flow
 - Data lifecycle
- IOT one of the most emerging upcoming trends
- Key data management challenges
 - Performance
 - Compliance







Light: Title Slide 2

Subtitle text placeholder

© 2018 NetApp, Inc. All rights reserved. — NETAPP CONFIDENTIAL —





Light: Title Slide 3

Subtitle text placeholder

© 2018 NetApp, Inc. All rights reserved. - NETAPP CONFIDENTIAL -

Other Data Management Challenges at the Edge



- First point of data entry
 - Perfect place for metadata enrichment
- Quality assessment
 - Completeness, Accuracy, Validity, Consistency
- Storage efficiency
 - In the light of data reduction
 - Quantifiable data approximation to reduce size
 - WAN transfer





Storage efficiency and Metadata Enrichment

- WAN transfer from the edge to the core/cloud
- Storage efficiency over the wire
 - 'Useful' data
 - 'New' data
 - 'Interesting' data
- All data keep everything and let Al worry about too much data
- Efficient anytime algorithm for identifying recurring patterns, anomalies, etc.
 - http://www.cs.ucr.edu/~eamonn/MatrixProfile.html

- First point of data entry into enterprise environment
- Perfect point for Metadata enrichment
 - Data Quality
 - Geo-location
 - Behavior identification
- Helps in various data workflows down the pipeline



Performance never been more critical!

GPUs read in parallel from storage; resource heavy AI/ML routines



PERFORMANCE



Agenda Slide

List major section of your presentation

- 1) First section/topic
- 2) Second section/topic
 - A brief description example
- 3) Third section/topic
- 4) Fourth section/topic





Agenda Slide

List major section of your presentation

- 1) First section/topic
- 2) Second section/topic
 - A brief description example
- 3) Third section/topic
- 4) Fourth section/topic

5) Fifth section/topic

- 6) Sixth section/topic
 - A brief description example



Closing Thoughts

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce ultrices felis eget enim tincidunt faucibus. Praesent vitae orci vestibulum, ultricies arcu in, sagittis tortor. Suspendisse convallis ligula a gravida tincidunt.



Closing Thoughts

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce ultrices felis eget enim tincidunt faucibus. Praesent vitae orci vestibulum, ultricies arcu in, sagittis tortor. Suspendisse convallis ligula a gravida tincidunt. Fusce posuere ipsum at finibus euismod. Vestibulum interdum tincidunt ligula vel tincidunt.