Storage Implications of Cognitive Computing

Balint Fleischer & Jian Li
Huawei
6/13/2016
What do we do with all the Data?

Some examples of Unstructured Data
- Financial and legal documents
- Email, Blog
- Research papers
- Video and Social Network postings
- Patient records
- Industry reports, Market Studies
- Regulatory publication
What do we do with all the Data?

Store it

Some examples of Unstructured Data
- Financial and legal documents
- Email, Blog
- Research papers
- Video and Social Network postings
- Patient records
- Industry reports, Market Studies
- Regulatory publication

Build a Highly Scalable Storage
What do we do with all the Data?

Some examples of Unstructured Data
- Financial and legal documents
- Email, Blog
- Research papers
- Video and Social Network postings
- Patient records
- Industry reports, Market Studies
- Regulatory publication

Process it with an Analytic Platform

- Analyze it
- Make predictions

Content
Sentiment
What do we do with all the Data?

Some examples of Unstructured Data
- Financial and legal documents
- Email, Blog
- Research papers
- Video and Social Network postings
- Patient records
- Industry reports, Market Studies
- Regulatory publication

Utilize it on a Cognitive Computing Platform

Integrate with our knowledge base
## Cognitive Computing (AI System)

Augmenting (Enhancing, Scaling and Accelerating) human expertise & 
Transforming human <-> Computer interaction

### Business Benefits

- Identify connections between events, people and trends
- Discovery of new insights, uncover breakthroughs and predict trends through real time understanding of current and historical data
- Enabling new customer experience via service personalization
- Reinvention of business models and operations

### Supporting Functionality

- Evolve with goals and respond to changes
- Participate in the shared discovery process and problem refinement iteration
- Understand meaning, goal, syntax, regulation, time, etc.
- Utilizes real time sensory and behavioral inputs as well as contextual data

A growing number of use cases benefiting from Interactive Cognitive Systems
Analytics vs. Cognitive Computing

Based on past history:
What restaurant will I choose today?

Responsive & Accurate

Everyone will find a choice
Ample parking
The wait is short.

Big Data
Cognitive Computing

Information
Knowledge

Predict
Classify

Analytics Systems
- Reactive (post transactional)
- Static, Empirical, some self-learning
- Batch/Streaming
- Stateless
- Highly accurate on sparse data
- Transactional, Unstructured
- SQL, ML, some DL
- NoSQL data store

Cognitive Systems
- Proactive (pre decision)
- Adaptive, Autonomous learning
- Streaming/Interactive
- Stateful
- Contextual
- Complex, varied, sensor data
- Deep Learning
- Complex, including associative memory

Gapingvoid art
Example use case:
Enhancing Mobile User Experience

New customer experiences
Understanding intentions
Anticipating needs

Personal Sensors
Tracking Actions & Surroundings
- Location
- Usage
- Health
- Image
- Voice
- Movement

Mobile User

Mobile User Experience

Cognitive Systems
- Social Networking
- Messaging

Cognitive Applications

Personal Assistants
- Health
- Direction
- Communication
- Shopping
- Entertainment
- Transportation

Smart Connections
- Waze
- Google Fit
- Google Now
- Apple Siri
- Skype Translator
- Skype
- Sina Weibo
- WhatsApp
- Facebook
Example use case:
Future Enterprise Management System

Interactive sessions

Enterprise Cognitive Hub
Query processing
Content Discovery
Visual Data Processing
Semantic Extraction
Sensor Data Processing
Probabilistic Reasoning

Cognitive Computing Platform

Data Sources

eComm CRMM ECM Trouble Ticket Products KMS
Why is Cognitive Computing is different?

- Focusing on Human-Computing Interaction
- Computationally intensive
- Cognitive Computing typically has a response time constraint (QoS)
- Working on Large Data sets
- Data access latency is critical to performance
- Data Access Patterns differ from Analytics (Sparse matrix, Graph etc.)
- Context (state) matters!
- Data Needs to be sharable across applications implementing pipelines
- Multi tenancy with QoS is key to achieve economics
Huawei’s Vision for Cognitive Computing

Highly Optimized for Interactive usages
AI based architecture developed for Human-Computer Symbiosis.
Tight integration of Event and Context Data
Processing engines tuned for performance and real time execution

Extensible Libraries
Product recommendation, Media Sentiment Analytics,
Fraud detection, Ranking systems, NLP, Speech to Text and Text to Speech, Tradeoff analytics, Visual Recognition, Cognitive insight, Etc
Cognitive Computing Platform

Operational Data Store

- High Performance, Low latency
- Cognitive Computing Optimized
- Scale out, Resilient, Elastic
- Sharable across Pipeline stages and Apps
# Compressing the Storage Hierarchy

## Storage Tiers

### On-die Data
- CPU L1-L3 cache
- Very Low Latency (<100ns), Extreme High BW

### In Memory Data Store
- DRAM, Persistent Memory (3D Xpoint, etc.)
- Low Latency (<5 µsec), Very High BW

### Near Memory Data Store
- NVMe drives, NVDIMM-F
- Medium Latency (~100µsec), Medium BW

### Network Attached Data Store
- Network attached storage (Block/Object/Files)
- Large Latency (>1msec), Limited BW

## Implementation

- Access Latency Reduction
- Storage Stack bypass
- Enabling Specific Semantics

## Moving Data
- Closer to Processing
Further Reducing Application to Data Latency

In Memory Computing Storage API

Near Data Processing Main Memory based

Near Data Processing HMC etc based

Co located Near Data Processing

Medium Latency <50usec
Medium BW
Data Sharing
Storage API
“Storage” functions

Low Latency <5usec
Medium/High BW
Direct Data Access
Application integrated*
General Purpose CPU

Low Latency <1usec
High BW
Direct Data Access
Application integrated*
General Purpose CPU

Very Low Latency <<1 usec
Extreme High BW
Direct Data Access
Application integrated*
Embedded CPU
Storage for Cognitive Computing

OLTP/OLAP
- RDBMS
- Structured Data Store Tables

Big Data Systems
- Batch & Streaming Analytics
- Unstructured Data Store Files, Objects

Cognitive Systems
- Deep Learning Systems Complex Pipelines
- Cognitive Data Store Objects, Graphs, Matrices

Brain
- Knowledge, Recognition, Intuition, Approximation, Etc.
- Associative Memory

Memory and information processing in neuromorphic systems Giacomo Indiveri and Shih-Chii Liu Senior, IEEE Proceedings 2015
Saffron Technology
Cognitive Computing Optimized Storage Stack

<table>
<thead>
<tr>
<th>Model Data Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cluster wide, Resilient and Elastic</td>
</tr>
<tr>
<td>• Very low latency to support random access*</td>
</tr>
<tr>
<td>• Pipeline data into on die Engine Data Store</td>
</tr>
<tr>
<td>• Direct Data Access by applications</td>
</tr>
<tr>
<td>• Mapped onto main memory and next gen fast NVME drives</td>
</tr>
<tr>
<td>• Data Parallel and Model Parallel modes</td>
</tr>
<tr>
<td>• Assist functions for Objects, Graphs, Matrices</td>
</tr>
<tr>
<td>• Fast access to Context Store</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context Data Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cluster wide, Persistent, Resilient and Elastic</td>
</tr>
<tr>
<td>• Stores Pre computed data and Context info</td>
</tr>
<tr>
<td>• Mapped onto direct attached NVMe drives</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Historical Data Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data Center wide, Network Attached, Persistent, Resilient and Elastic</td>
</tr>
<tr>
<td>• Stores Historical and Reference data</td>
</tr>
</tbody>
</table>
Mapping to Classic Storage Hierarchy

Storage Tiers

- On-die Data
- In Memory Data Store
- Near Memory Data Store
- Network Attached Data Store

Implementation

- CPU L1-L3 cache
  - Very Low Latency <100nsec, Extreme High BW
- DRAM, Persistent Memory
  - (3D Xpoint, Etc.)
  - Low Latency <5 usec, Very High BW
- NVME drives, NVDIMM-F
  - Medium Latency ~100usec, Medium BW
- Network attached storage
  - (Block/Object/FILE)
  - Large Latency >1ms, Limited BW

Cognitive Data Store Data Tiers

- Engine Data Store
- Models, Working Data
- Context Knowledge Indices (Persistent)
- Historical Data (Persistent)
Enabling Cognitive Computing with Cloud Services
Huawei’s Data Function Virtualization Platform Vision

**DFV Data Container model**
- Data Container per application
- Application can be stand alone or clustered
- Attributes are application specific
- Locality management, Storage semantics,
- Tiering policy, Resiliency Policy,
- Security, Performance,
- Sharing semantics, Etc.

**DFV Control Plane functions**
- Pool management, Allocation, Data Services,
- Data Pool classes, Migration, Staging,
- Recovery Assist, Elasticity management,
- Global replication, Etc.

* Note, Analytics and Cognitive Platforms maybe running in containers
Summary

- Augmenting human expertise yield great business value
- New technologies, such as new Memory, high core count CPUs, fast Fabrics and various accelerators are critical HW ingredients of Cognitive Computing Platforms
- New Software innovations coupled built on Near Data Processing enables the delivery these High Performance, Responsive Cognitive Platforms
- Re architecting of the storage stack will make it possible to scale cloud architectures to support high performance solutions
- Huawei is developing a comprehensive vision addressing these changes