Data, Information, and Knowledge Management

John Harkins
LAS
jgharkin@ncsu.edu

Opportunity

Data production continues to accelerate:
2009: Google processed ~ 24 Petabytes/day (1 PB = 10^{15} Bytes)
2011: Internet data ~ 1 Exabyte (1 EB = 10^{16} Bytes)

Present:
Global data produced in 10-20 minutes > than previous 100 years

2025:
Global data produced in 5s > 10-20 min production in present day

Problem

How to manage data in a way that optimizes our ability to acquire actionable knowledge from the data given limited resources?
- What questions should the data answer?
- Which data model enables finding those answers?
- Which analytics/visualizations will transform data into information and information into knowledge efficiently using that model?

“DIKAR” Data Progression

Data → Information → Knowledge → Action → Result

Functional Considerations

Implemented system should support analyses, visualizations, and answer questions such as:
- Properties of and relationships between entities at any point in space and time
- Sequencing of data in absolute and relative terms (before, during, after)
- Exposing relationships between entities by context/semantics/ontology
- Interactive, dynamically scalable, geo-temporal data visualizations
- Network Analysis
- Complex Event Processing
- Traffic Analysis
- Trend Analysis
- Supervised/Unsupervised Machine Learning
- Provenance Log

Research

Managing “Big” Data Scale
- Differential data updates
- Partial & fast compression techniques
- Natural data decay rates; e.g. exponential functions of time, resource availability, and relevance (access time)

Graph Models
- All to all vertices
- Multi-dimensional edge weightings versus
- Multi-graph, one-dimensional models
- Query and transaction operation
- Entity affinities computed w.r.t. life-spans and proximity

Search

Exploit Existing Technologies
- Von Neumann based architectures
- GPUs
- RDBMS, NoSQL, Graph Stores
- RDF, JASON-LD

Data Model

Desire a data model that can be applied to many domains, ingest data in many forms (e.g. structured and unstructured), and be used to answer many questions (e.g. geo-temporal queries).

Conceptual/Logical Model
- Entities/Subtypes
- Semantics/Ontologies
- Attributes/Metadata
- Relationships
- Access Rules
- Compliance Rules
- Integrity Rules

Physical Model
- Tables
- Columns/Rows
- Keys/Indices
- Tuples
- Graphs

Visualization Examples: Network & Geo-Temporal

Visualizations include:
- Network Analysis
- Geo-Temporal Visualization
- Trend Analysis
- Supervised/Unsupervised Machine Learning
- Provenance Log

Graph Models
- All to all vertices
- Multi-dimensional edge weightings versus
- Multi-graph, one-dimensional models
- Query and transaction operation
- Entity affinities computed w.r.t. life-spans and proximity

Search

Exploit Existing Technologies
- Von Neumann based architectures
- GPUs
- RDBMS, NoSQL, Graph Stores
- RDF, JASON-LD

Created by Nicholas Jenkins, Elijah Mosko and Scott Muter | © 2013 Nicholas Jenkins, Published by Stanford University Libraries