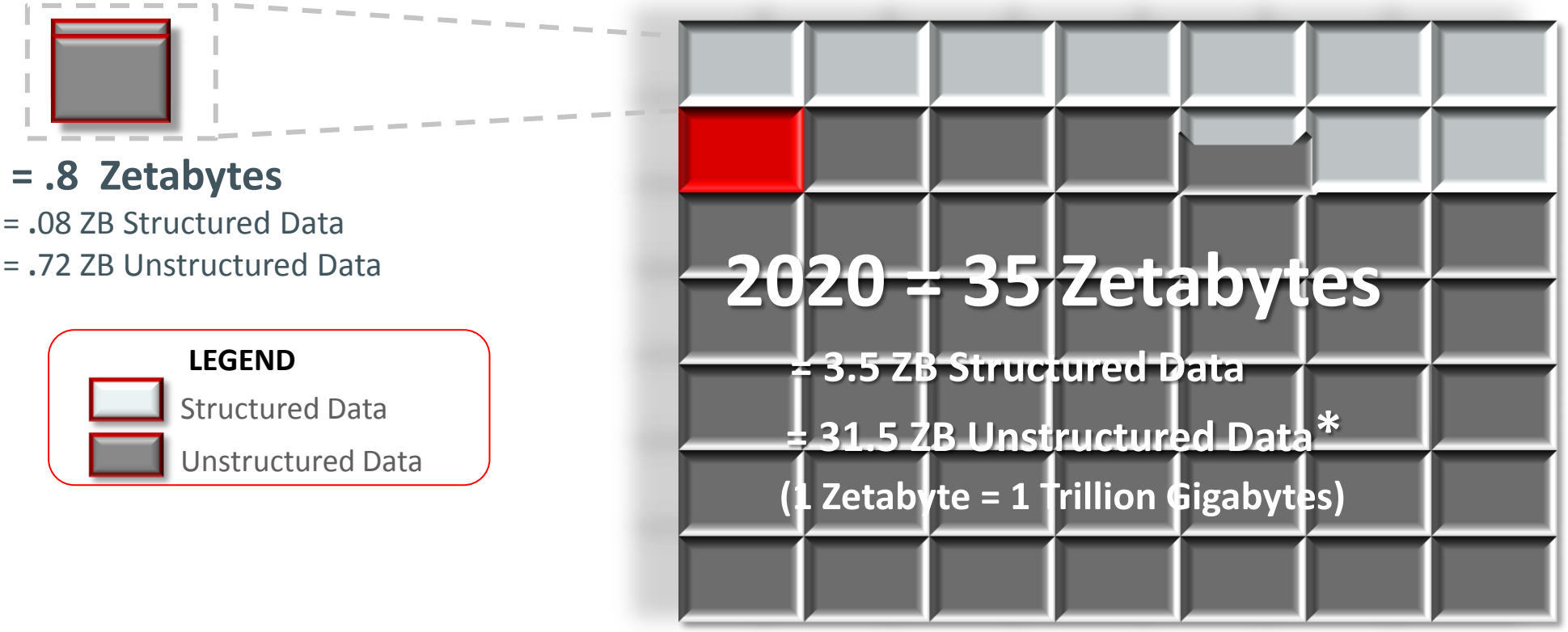
Big Data for Official Statistics

Processing Big and Fast Data
Optimizing Results with a Multi-Model Database

Steven Hagan
Vice President
Oracle Database Server Technologies
October, 2015

Global Digital Data Growth: Exceeds Storage Mfg

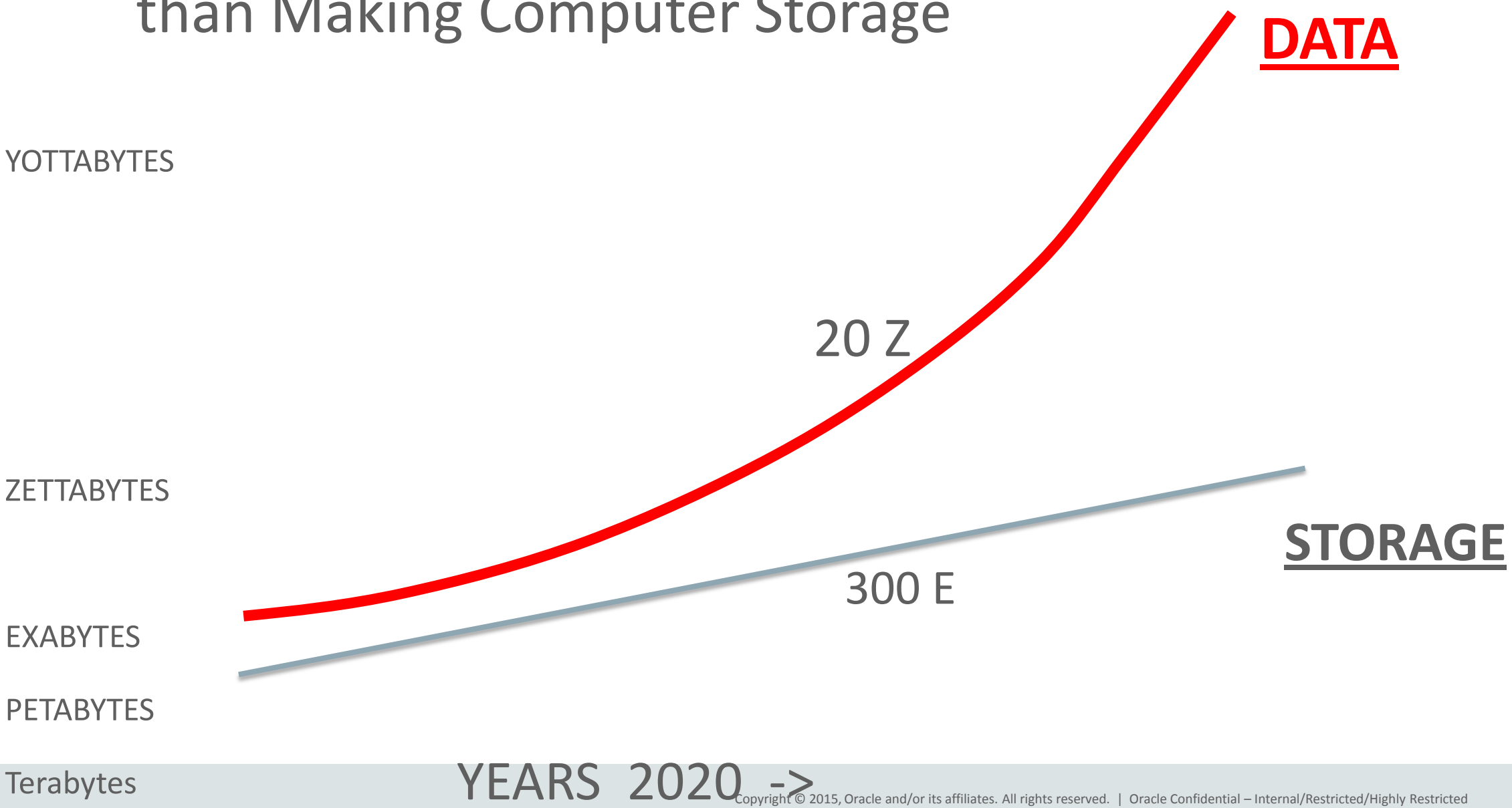
Growing leaps and bounds by 40+% YoY!



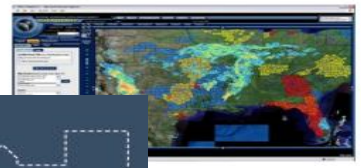
- Chart conservatively assumes a constant 9:1 ratio of unstructured data vs. structured data (based upon IDC's estimate that 90% of all digital data is unstructured).
- Chart does not reflect IDC's projection that unstructured data is currently growing twice as fast as structured data at the rate of 63.7% vs. 32.3% CAGR.

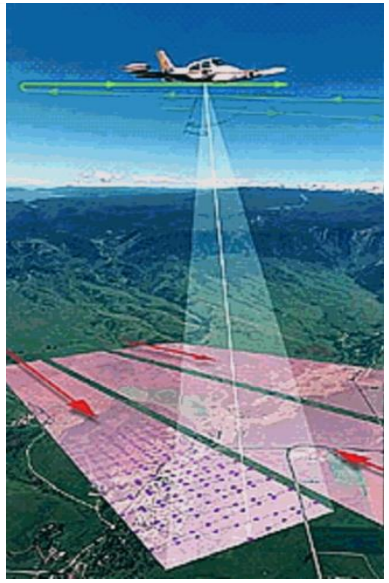
Source: IDC Digital Universe Study, A Digital Universe Decade – Are Your Ready?, 2010

Human Race is Generating Data Vastly faster than Making Computer Storage



Data Volume & Variety Generation Explosion Continues – Terabytes, Petabytes, Exabytes, Zettabytes

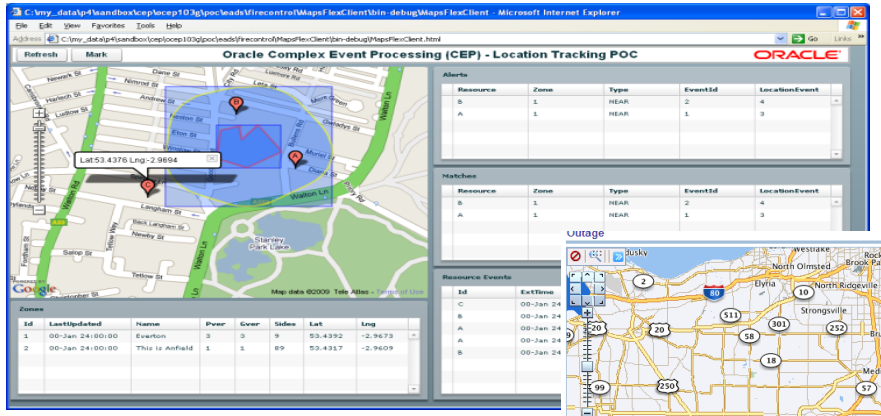




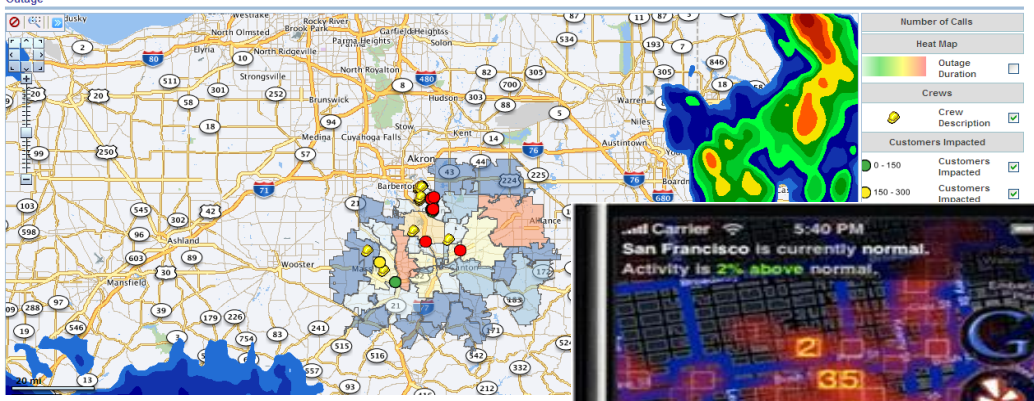
- VIDEO: UAVs, DRONES, SURVEILLANCE
- IMAGERY/Raster: (Satellites, Planes)
- Sensors (IOT), LIDAR, 3D, RFID
- Social Media, Web Scraping, Mobile Phones
- New data products for: Land and Water mgmt, Agriculture, Environment Transportation, Terrain and City Models, SDIs for planning, maintenance, Emergency response, Defense, Intelligence, Consumers **Location is a Powerful Organizing Principle**
- Semantics , Ontologies --
- Wearable Technologies
- Genomics (DNA Sequencing) , Astronomy
- **MULTIPLE VERSIONS OF THE ABOVE**

Data Velocity: Real-Time Spatially-aware Streams / Events / Sensors /"Internet of Things" (EVERYTHING)

Track / Monitor Moving Objects –UAVs, Drones, cars



Real-Time Business Intelligence



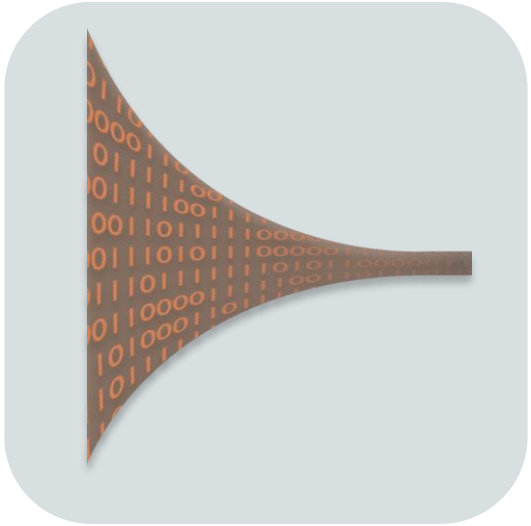
Real-Time Pattern Detection

- Ultra-high throughput
- (1 million/sec++) and microsecond latency
- Sensors on Aircraft Turbine Blades
- Filtering, correlation, and aggregation across event sources
- Detect patterns in the flow of events and message payloads, Complex Event Processing (CEP)
- Business Intelligence in Real Time
- Mobile Phones
- Self-Driving Cars


IN 60 SECONDS_
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING COMPANY
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com

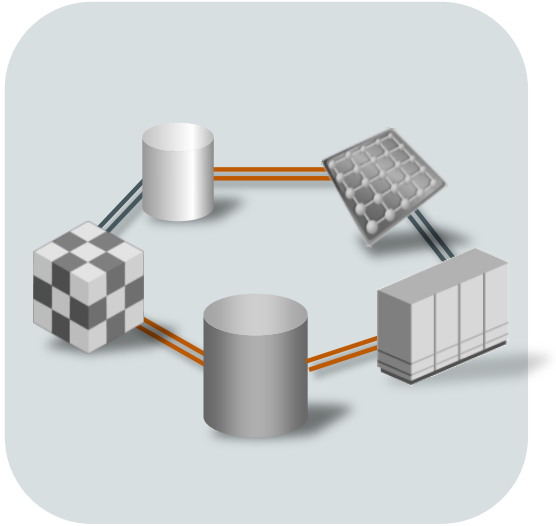
Processing Big & Fast Data: Video, Imagery, Sensors, Social, Mobile, …

Filter, Move, Transform, Analyze, Act - at High Velocity



FILTER CORRELATE AGGREGATE

Oracle Event Processing
Oracle Spatial



ENRICH & TRANSFORM

Oracle Coherence
Oracle GoldenGate
Oracle Data Integrator

ANALYZE

Oracle BAM
Oracle Mapviewer
Oracle Business Intelligence
Oracle Information Discovery

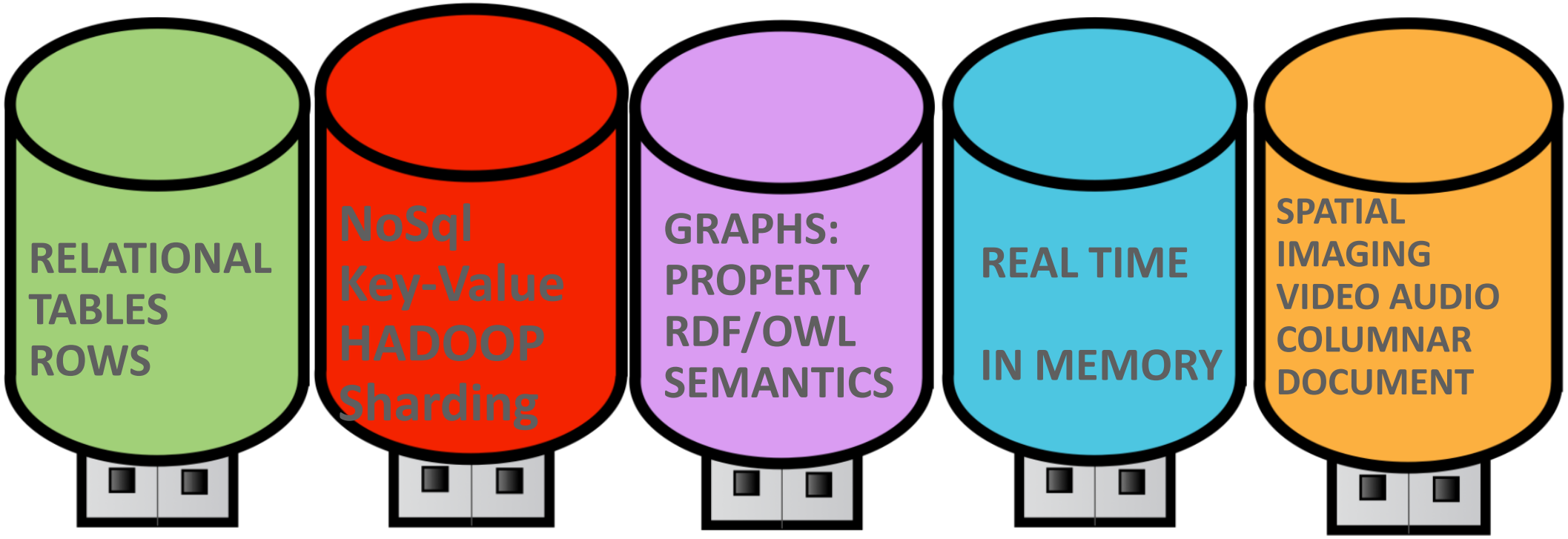
ACT

STORE / SAVE / ARCHIVE ??
THE RESULTS

TRENDS: Next 5 years or so

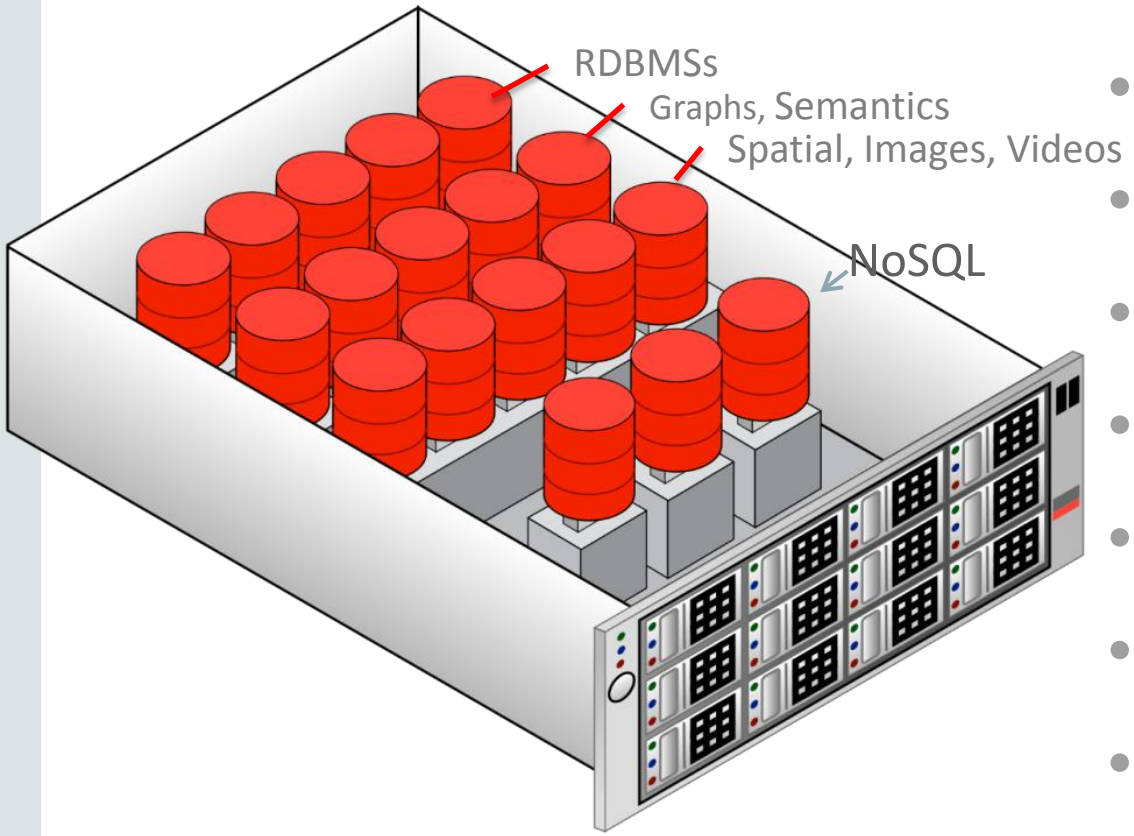
- Computer System Performance –
  - Hardware - Evolutionary – Moore's law still holding
  - New possibilities at Research Level – not yet proven
    - DNA for Storage; 3D Glass, Holography; Carbon Nanotubes, Graphene
  - Software – Disruptive – **Parallelism** enables clusters of 10,000+ computers, CLOUD
- Software is Supporting many Data types - **FLEXIBILITY**
  - Databases/persistent stores can handle all types of data – **Polyglot Persistence**
  - Software – Graph Storage, Semantics – Add all types of data and build new relationships
    - Without disruptive upgrades / schema changes
  - Stream data arriving; Filter the data; Keep what matches your requirements; aggregate it
  - Deletions: immediately/gradually
  - **NOTE: TEXT AND NUMBERS ARE NOT THE SPACE PROBLEM!**

SPECIAL DATA TYPES: SEVERAL POPULAR DATA MODELS:
But Unique separate persistent stores results in:
MANY databases to secure &manage



For National / UN Statistics: MULTI-MODEL Database is Best - Many Different Data Models Supported as Native Data Types in

ONE SHARED STORE



- Parallel Database Server has multiple models
- Unified Security Approach
- Highly Available
- Disaster Tolerant
- Shares Main Memory; more efficient
- Shares Disks, Flash Storage: more efficient
- Managed as a single entity: more efficient
- (ORACLE HAS THIS TODAY)

National Statistics: one Multi-Model Store

External Data Sources:
Transactional &
Operational Systems
Contents Repository
Databases
Mobile Devices, Web resources
Blogs, Mails, news
Satellite Imagery, UAVs

Real-time Data Streams

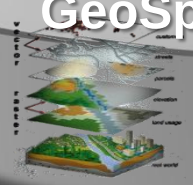
CEP

Search, Presentation, Report, Visualization, Query

Multi-Model Data Management Infrastructure

Secured

GeoSpatial



Historical Records

POIs
Demographics
Customer Data
Call Records

Documents

Automatic Responses and Publishing

SMS

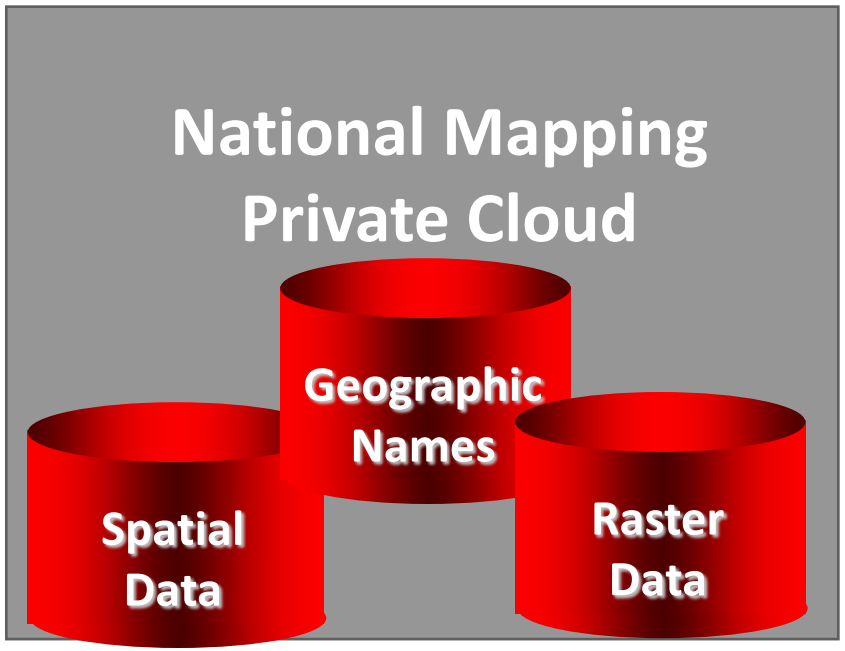
Console Alerts

EV Grid Management

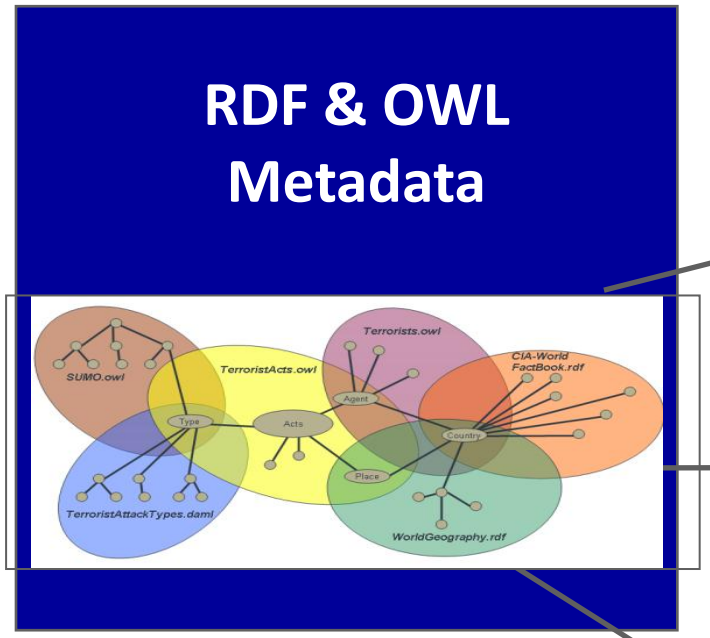
Workflow Initiation

Real-time Dashboards

Statistics Data Repurposing: Ontology-driven Enable Shared, Actionable Knowledge

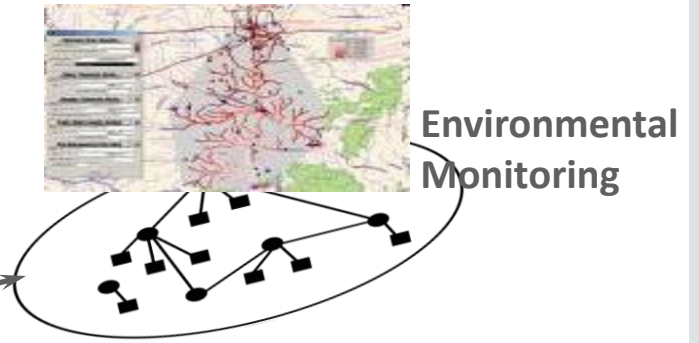


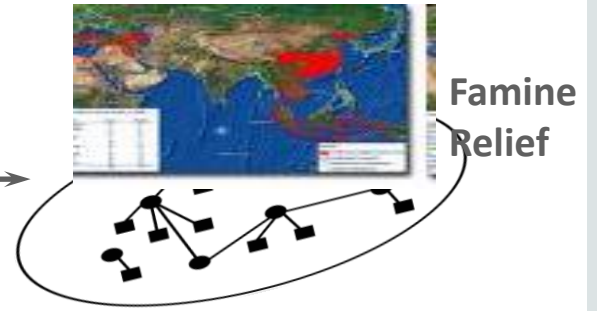
- Simple Features
- GeoRaster
- Topology
- Networks
- Gazetteers
- …



- Data Integration
- National Map schemas
- Geographic names
- Temporal
- Naïve Geography
- …

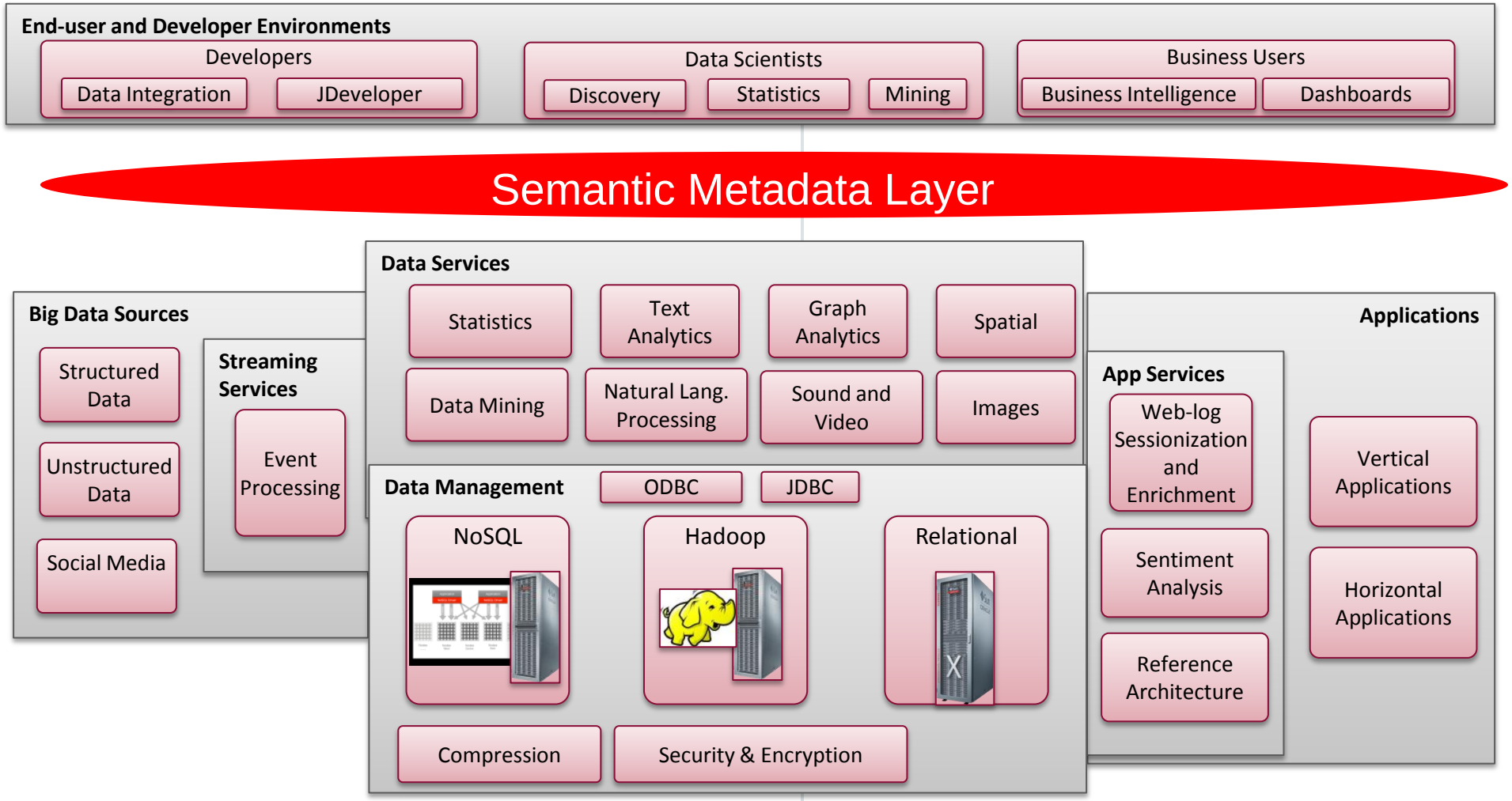
Application Ontologies





Support Breadth of National & UN Data ABOVE STOVEPIPES

Data arrives, is filtered, stored data is available to all Statistics Organizations



GUIDANCE: THIS IS AN ARCHITECTURE TO SUPPORT ONE SHARED MULTIPURPOSE NATIONAL STORE

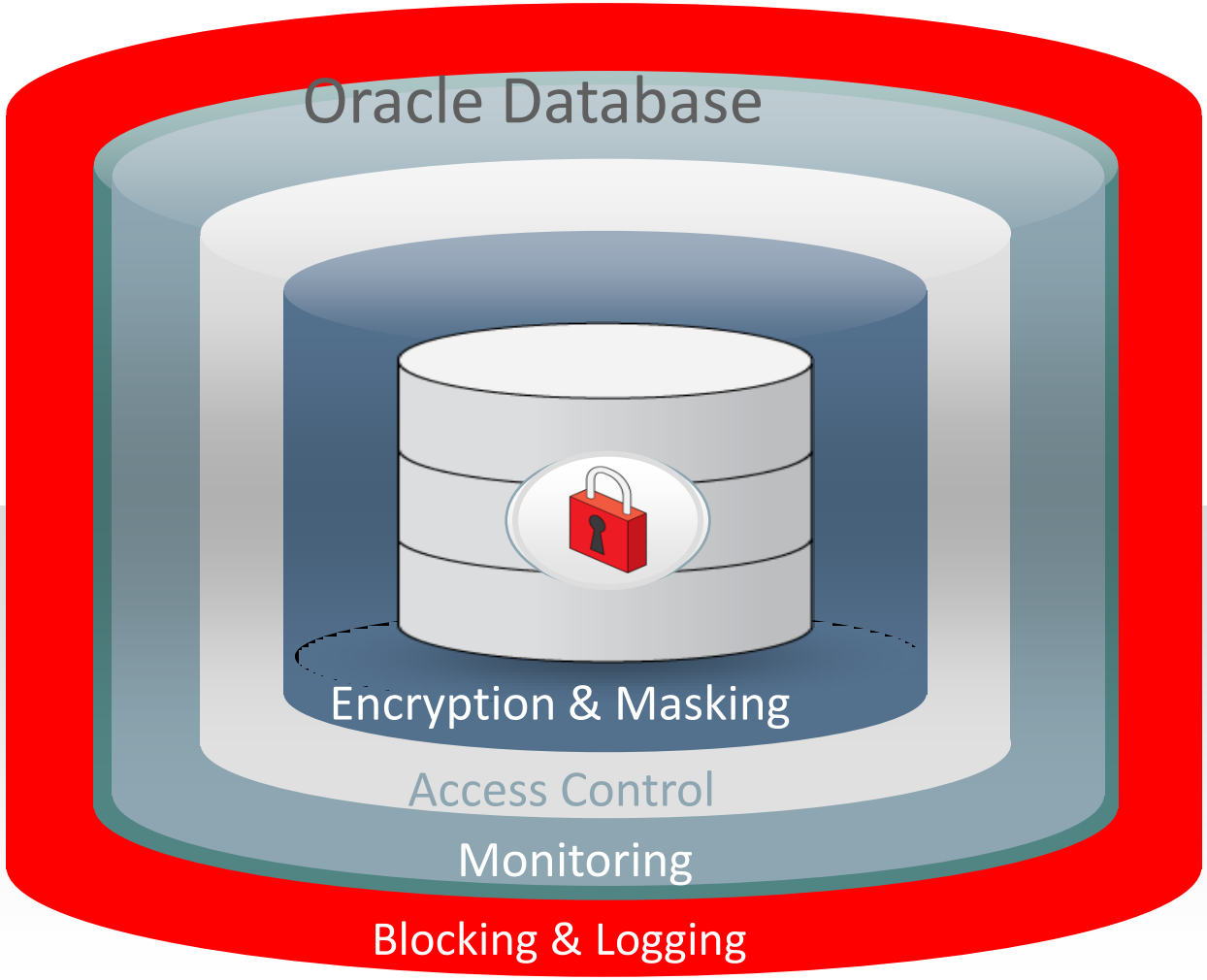
Semantic & Graph Technology What terms to look for: Buzzwords For Apps & Workflows using

- Semantic Web
- W3C RDF/OWL/SPARQL
- Graph Data Management
- Social Network Analysis (SNA)
- Knowledge Discovery
- Knowledge Mining
- Big Data
- Schema-less Data
- Property Graphs
- Taxonomy/Terminology Mgmt
- Faceted Search
- Inferencing / Reasoning
- Sentiment Analysis
- Text Mining
- NoSQL Database

Oracle: Graph (Linked Open Data) support: On-premise or in the Cloud

- Highly scalable, secure triple store based on RDF
  - ***1 TRILLION TRIPLE BENCHMARK**, leading Triple Store:W3.org*
  - 1.13 million triples per second query performance
- SPARQL and GeoSPARQL in SQL support
  - Apache Jena and OpenRDF Sesame pre-integrated
  - SPARQL endpoint enhanced with query control
  - GeoSPARQL support (classes, properties, datatypes, query functions)
- Forward-chaining based inferencing engine in the database
  - Various native rulebases (RDFS, OWL2 RL, SKOS, ...), integration with OWL2 reasonsers (TrOWL, Pellet)
- RDB to RDF mapping on relational data aligned with RDB2RDF standard

Accessible Shared Data: **CYBERSECURITY** is Major Challenge
Requires Information Security and Privacy



Monitoring
- Configuration Management
- Audit Vault
- Total Recall

Access Control
- Database Vault
- Label Security

Encryption & Masking
- Advanced Security
- Secure Backup
- Data Masking

United Nation Analysis – September 2013

Initiative on Global GeoSpatial Information Management

Future Trends

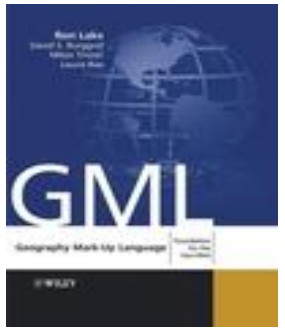
- Technology Trends in Data Creation, Maintenance, and Management
- Reliance on '*big data*' technologies
- The *right* information at the *right time*
- Machine-processable descriptions of data.
- Semantic technologies will play an important role
- Skills and Training: train the individuals is at least five years

Requirement for enhanced Data Management Systems

You Enhance Innovation & Statistics By Using **STANDARDS**

e.g. – The Spatial Data Domain

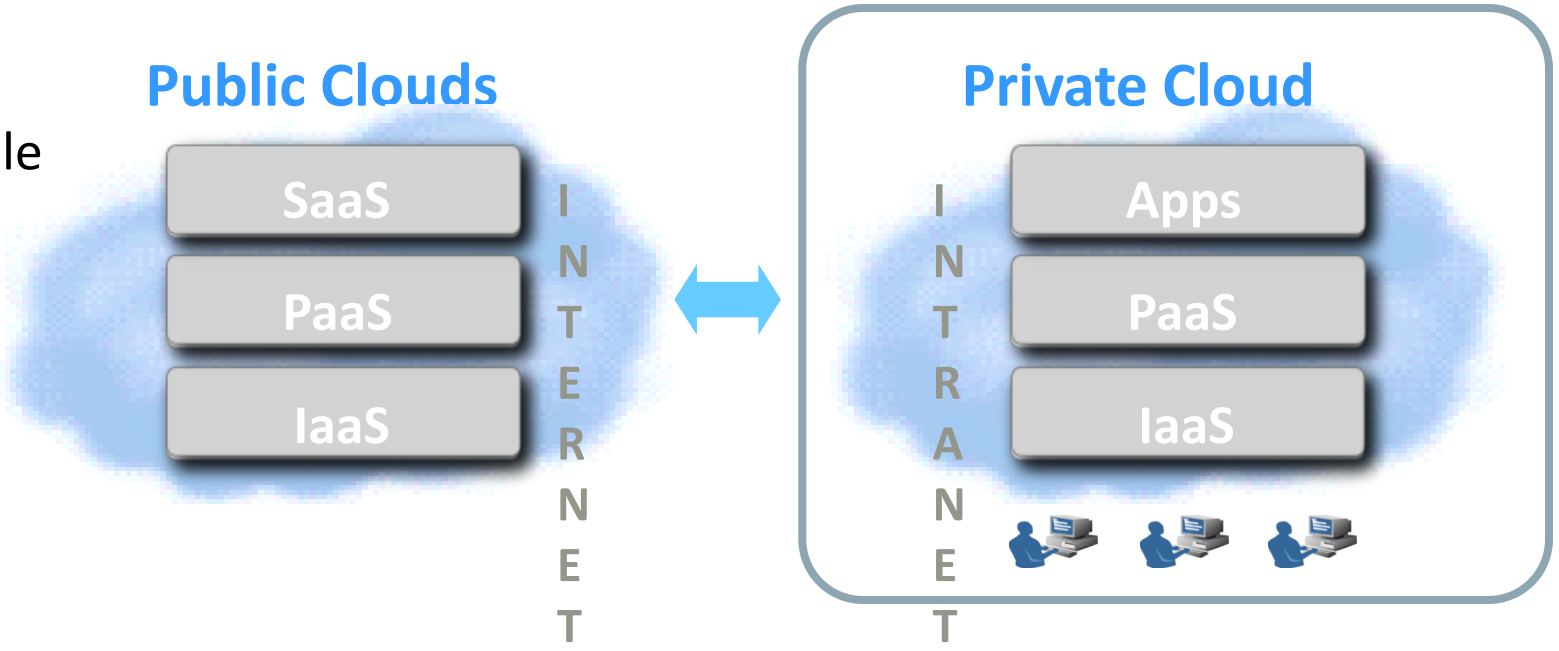
- **ISO**
  - **TC 211; TC 204**
- **Open Geospatial Consortium**
  - **Simple Features; GML; Web Services**
- **De-facto Standards**
  - **SHP, MGE, DXF, KML**
- **Professional Standards**
  - **ISPRS, FIG, WMO**
- **Java, .NET, Flash**
- **W3C: RDF,OWL, SPARQL, GeoSPARQL**
- **TAGGED METADATA – agree on tags**



SQL3/MM Spatial

Public Clouds, Private Clouds: Statistics Platforms

- Used by multiple tenants on a shared basis
- Hosted and managed by cloud service provider



- Exclusively used by a single organization
- Controlled and managed by in-house IT

Trade-offs

Lower *upfront* costs ⟷ Lower *total* costs

Outsourced management ⟷ Greater control over security, compliance, QoS

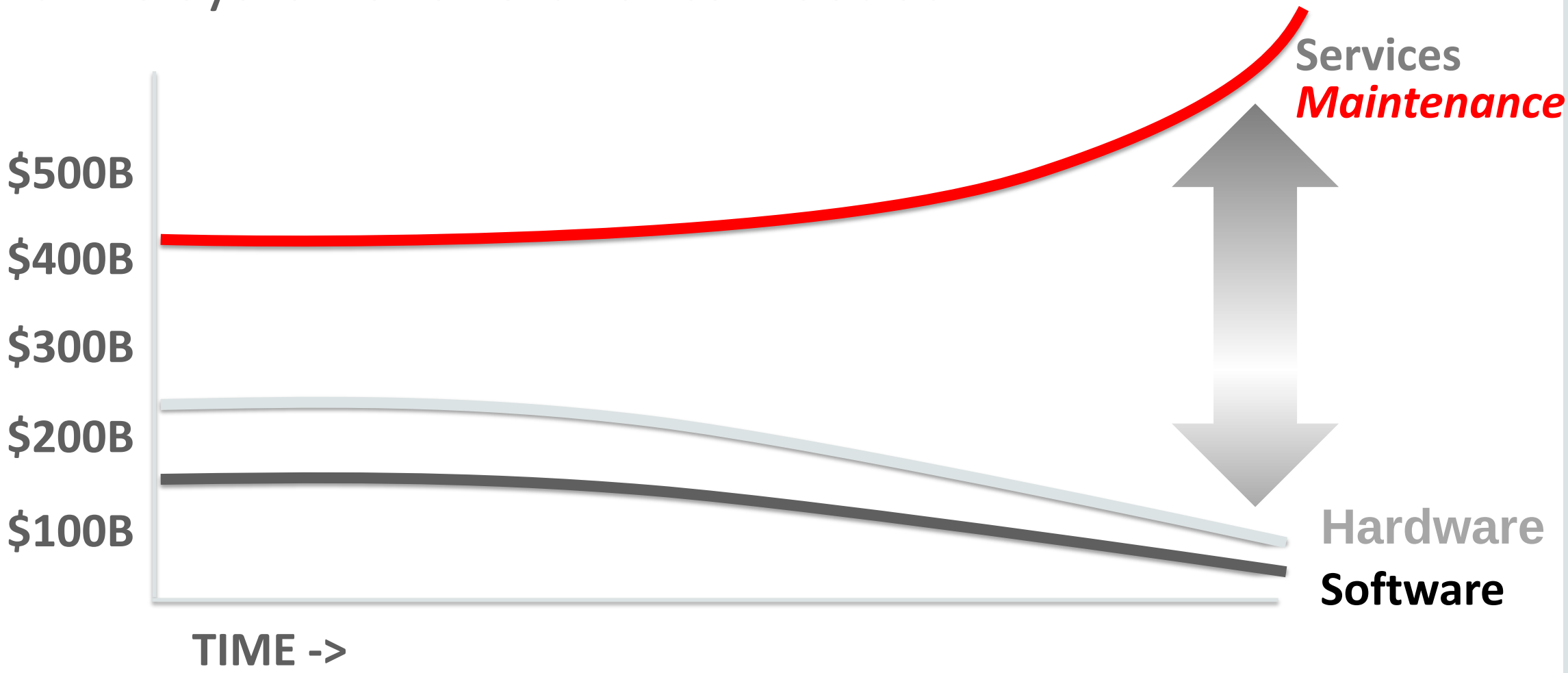
OpEx ⟷ CapEx & OpEx

ELASTICITY is key value of Clouds

YOU MAY NEED A CLOUD IN EACH COUNTRY ---DEPENDS ON THEIR LAWS

Today: More HW/SW Efficiencies: But Labor Costs Growing

Innovative Systems for Statistics Needed



Guidance: Do Not Build Your Statistics Solutions From Scratch

Long Term Cost of Ownership rises with custom construction & Open Source

Time to Build

Optimizations

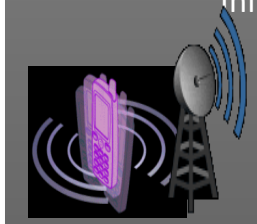
Maintenance

UN-GGIM: “train the individuals is at least five years”

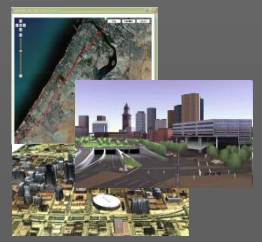
Guidance: Big Data for Official Statistics: Success Enhanced with MULTI-MODEL DATABASE PLATFORM

Big & Fast Data

Volunteered Geographic Statistical Information



Sensors Streaming Data



Geo-referenced Video, 3D, LiDAR Satellites

Simplify Statistics IT

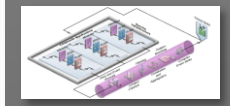
Support for Open Standards

Spatial Database, Application Server, BI, tools

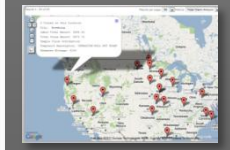
Support by Leading Partner solutions

Multi-Model Engineered Systems

Deep Analytics



Real-time Complex Event Processing



Dense Visualization

Spatial Analysis Graph Analytics

On Premise, On Cloud, Shared Services

Shared GeoSpatial Services
Location Aware Everything

Fully Parallel and Secure