Big Data for Government Symposium

http://www.ttcus.com

Linkedin/Groups:
Technology Training Corporation

@TECHTrain
Big Data For the Mission Side of the House

Leveraging the right infrastructure

Nancy Grady, SAIC Technical Fellow, Data Science
November, 2013
Big Data Across the Clouds

- Big Data
- Data Science
- Data Lifecycle
- Big Data Engineering
- Hybrid Infrastructure
- Mission Focus
Big Data

• The **Big Data Paradigm** consists of the distribution of data systems across horizontally-coupled resources to achieve the scalability needed for the efficient processing of extensive datasets.

• **Big Data** consists of extensive datasets, primarily in the characteristics of volume, velocity and variety, that require a scalable architecture for efficient storage, manipulation, and analysis.

• **Big Data Engineering** is the storage and data manipulation technologies that leverage a collection of horizontally coupled resources to achieve near linear scalability.

• **Big Data Models** refers to a non-relational logical data model for the storage and manipulation of data across horizontally scaled resources, including techniques categorized at a high level as name-value, big table, document or graphical.

Note: from upcoming NIST Big Data Definitions subgroup
It’s Not Just the Volume

Data is growing faster than “Moore’s Law” on all dimensions

- **Volume** – ever increasing challenge
- **Velocity** – rapid response needs
- **Variety** – combining data across domains

“Business benefits are frequently higher when addressing the variety of data than when addressing volume”

Mark Beyer and Doug Laney, Gartner, 2012
“The Importance of Big Data: A Definition”
It’s Not Just the V’s

Engineering
• Volume
• Velocity
• Variability
• Variety

Mission
• Value

Science
• Veracity
• Complexity
• Cleanliness
• Completeness
• Consistency
• Latency
• Provenance

Analytics by Type
• Structured
• Semi-structured
• Unstructured
Data Science
Analytics Trends

• Statistics
  – Computer design of experiments
  – Clean data
  – Deterministic analysis (necessary and sufficient)

• Data Mining (needle in haystack) or Modeling
  – Re-purposed data
  – Cleansed sample data
  – Approximate causal analysis

• Data Science = Fourth Paradigm*
  – Re-purposed data
  – Near system-population size sampling
  – Approximate correlation analysis

* Term coined by Jim Gray
Data Science

- It’s not a black box
- It’s not started in a vacuum
- It’s not new analytics
- It’s not a miracle

I think you should be more explicit in step 2

Then Data Science
• **Data Science** is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.

• A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming and systems engineering expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.
Data Science is a Team Sport

Mission Expertise

- Domain Expertise
- Research
- Data Science
- Analytic Systems
- Engineering Skills
- Algorithms
- Statistics
- Data Mining
Historical Data Analysis Emphasis

TRADITIONAL EMPHASIS
- ALGORITHMS
- SYSTEM
- DATA

DATA ENGINEERING
- ALGORITHMS
- SYSTEM
- DATA

DATA SCIENCE
- ALGORITHMS
- SYSTEM
- DATA

© SAIC. All rights reserved.
The Data Life Cycle
"Ten crates of data and one little envelope of information. Sign here."
Our business is to address the mission needs through the creation of operational analytics, including the time scale for action.

**Military Planning**
- Assess
- Contingency
- Situational Awareness
- Formulate COA’s
- Assess course of actions
- Decide
- Execute
- Monitor

**Engagement**
- Observe
- Orient
- Decide
- Act
Analysis

PRESCRIPTION
- What's the best that can happen?
- Recommendation

Maximization
- What will happen if I take this action?
- Modeling

OPTIMIZATION
- What will happen next?
- Forecasting

PREDICTION
- Data Mining
- Correlation

CONFIRMATION
- What am I missing?
- Why might this be happening?

DISCOVERY
- Alerting
- What is happening now?

DESCRIPTION
- Situational Awareness
- What has just happened?

HINDSIGHT
INSIGHT
FORESIGHT

Complexity
Data Complexity

- Is a function of the inter-relatedness of different datasets
- In Defense, the data fits along 4 axis
  - Temporal: past, present, future
  - Assets: friendly, neutral, hostile
  - Component: air, ground, sea, cyber
  - W’s: who, what, where, when, why, how
- With greater complexity the ability to move “up” the predictive analytics chain becomes more difficult
- The data lifecycle shrinks based on time-to-decision
  - Months: strategic planning
  - 24 hours: crisis management
  - 5 minutes: Real-time engagement
Big Data Engineering
Enterprise Data Lifecycle Services

Data Services: Architect, Model, Policy, Document, Monitor, Measure, ...

Big Data

internal data

data

Engineering Services: Dev/Test, Store, Backup, Recover, Archive, Transmit, ...

Infrastructure Services: Security, Cloud Brokerage, Data Center Operations, ...

Need

COLLECT

Benefit

KNOWLEDGE

ANALYZE

INFORMATION

CURATE

data

external data
Traditional Data Life Cycle

**COLLECT**
- ETL
- Staging
- Domain

**CURATE**
- Cleanse
- Transform
- Warehouse

**ANALYZE**
- Algorithm
- Analytic Mart
- Summarized Data

**ACT**
- Action

ETL = extract, transform, load
Note: Variability means spawning more collection servers
Big Volume Engineering

COLLECT  CURATE  ANALYZE  ACT

Volume

Raw Data Cluster

Map/Reduce

Cleanse
Transform
Analyze

Model Building
Model Analytics

Mart

Data Product

Domain

Complexity
Big Variety Engineering

COLLECT

CURATE

ANALYZE

ACT

Variety

Complexity

Map/Reduce

Query

Fused Data

Analyze

© SAIC. All rights reserved.
Hybrid Infrastructure for the Mission
New Options

Public Cloud

Government Cloud

Private Environment

Services

Data Store

Services

Data Store

Data Store

Services
Hybrid Infrastructure

**FROM DATA**
- External Data

**TO INFORMATION**
- Ingestion, Fusion, Enrichment, Alerting

**TO EXPLORATION**
- Enriched Data Storage

**TO KNOWLEDGE**
- Query, Modeling, Characterization, Prediction

**TO INSIGHT**
- Cloud Environment

**DATA**
- Data
- Enrichment Sources
- Query Analysis
- Models

**Big Data Ingestion**

**Big Data Analytics**

**Custom Enrichment**

**Browsing Modules**

**Custom Analytics**

**Analyst Tools**

SAIC.com

© SAIC. All rights reserved.
Data Analysis Use Case

COLLECT  CURATE  ANALYZE  ACT

Government or Private Cloud

ETL = extract, transform, load
Data Analysis - Sharing Use Case

COLLECT

CURATE

Integration

ANALYZE

Community Cloud

ACT

Data-as-a-Service

Agency 1 Internal

Agency 2 Internal

Data Stores

Data Stores

Analyze

Analyze

ACT
Dev/Test Use Case

ETL = extract, transform, load
Data Ingestion Use Case – Velocity

COLLECT

CURATE

ANALYZE

ACT

Public Cloud

Velocity

ingest

ETL

Data Store

Metrics

Internal Environment
Data Ingestion & Publication Use Case

COLLECT | CURATE | ANALYZE | ACT
---------|--------|---------|-----
Submit   | ingest | NoSQL Store | Analyze | ingest
ETL      | Data Store | Analyze | Transmit
Cloud Environment
Internal Environment
Data Silo Use Case – Variety

COLLECT

Silo1
Access
Department 1

Silo2
Access
Department 2

Silo3
Access
Department 3

CURATE

Logical Data Store

ANALYZE

Optional Integration

ACT

Private Cloud

Silo1

Silo2

Silo3

Access

Department 1

Department 2

Department 3

© SAIC. All rights reserved.
Mission Focus
It's about the Mission Stupid!

We have this awesome data on [insert mouth-watering description of data]! We cleaned it up and we're running sophisticated analysis on it. We see {story about fascinating patterns}. Isn't that cool?!

Booyah! That sounds like so much fun! Why are you doing it?

We're not sure yet, but imagine the possibilities! This has to be valuable!

http://datascopeanalytics.com/
Traditional Data Silos

Silo1: Sensor

Silo2: Satellite

Silo3: Modeling

Agency

Publish / Subscribe Services
Mission Focus: Logical Data Warehouse

COLLECT → CURATE → ANALYZE → ACT

Data Virtualization Layer

External Data

Fused Data

Act

Variety

Complexity

SAIC.com
Focus for Mission Support

- Reduction in data movement and copies across networks
  - Bring the algorithms to the data, not the data to the algorithms
- Better visualization for the human-in-the-loop
  - We must feed the cognition of decision makers
- The experts may not be in your organization
  - Data Science teams may need to combine government, universities and commercial experts to leverage all the data
- Entrepreneurs are leveraging weather and climate data to build previously impossible business
  - For example, The Climate Corporate selling bad weather insurance to farmers
Some Big Data Thoughts

- You have existing analytics infrastructures
- You want to add in big data and have scalable systems
- You need to cross data silos
- You need to increase efficiencies and reduce cost
- Your analysts want to use their familiar tools

- We want to avoid data migration, if possible
- Use services and semantic technologies (scalability)
  - No more point-to-point integrations
- Use the Logical Data Warehouse
Converged Solutions are the Future

Leverage big data

- Open Source and Internal Data
- Compute, Storage, VDI, and Apps

Leverage social analytics

- Collaboration and Visualization
- Mobile Synchronization

Integrate with cloud

Ever-changing platforms
Big Data Technologies Bring Big Opportunities

The ground rules have changed. Big Data is upon us...

How can we best
• Leverage the technology
• Connect across domains for new capabilities
• Construct business models based on the new economics

"Eureka! More information!"

Note: Drawing is from 1995

Used by permission Ted Goff

© 1995 Ted Goff www.tedgoff.com
Nancy Grady, Ph.D.
SAIC Technical Fellow, Data Science
Cyber, Cloud, and Data Science Solutions

nancy.w.grady@saic.com
865-604-6733