Big Data Across the Clouds

Leveraging the right infrastructure

Nancy Grady, SAIC Technical Fellow, Data Science
March 5, 2013
Big Data Across the Clouds

- How We Got Here
- The Data Life Cycle
- Data Science
- Big Data Engineering
- Big Data Infrastructure
The Fifth Wave of Computing ... Meets ...

- Servers and dumb terminals (1950s-1960s)
- Personal computers (1970s)
- Internet (1990s)
- Cloud (2000s)
  - Infrastructure, applications, data, and analytics
- Virtualization (2010s)

We’ve come full cycle now with the push to virtual desktop infrastructure

- Mobile devices (2010s)
  - Carry them everywhere
  - Connectivity
  - Sensors
  - Location
  - Photographs
  - Oh, ... and phone service
### Data Analytics (The Fourth Paradigm)

<table>
<thead>
<tr>
<th>Science</th>
<th>Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Experiments</td>
<td>1. Statistics</td>
</tr>
<tr>
<td>2. Theory</td>
<td>• Design of experiments</td>
</tr>
<tr>
<td>3. Simulation</td>
<td>2. Data mining</td>
</tr>
<tr>
<td>– Big iron</td>
<td>• Repurposed data</td>
</tr>
<tr>
<td>4. Data science</td>
<td>3. Data science</td>
</tr>
<tr>
<td>– Big data</td>
<td>• Agile analytics</td>
</tr>
<tr>
<td>– Sensors</td>
<td></td>
</tr>
</tbody>
</table>

In “The Fourth Paradigm: Data-Intensive Scientific Discovery,” pioneering computer scientist Jim Gray refers to discovery based on data-intensive science
The Fuss Is About Data Scaling

We’re feeling the disruption of powers-of-ten scaling

• Compute power – growing according to Moore’s Law
  – Data volumes are growing faster

• Convergence in 2009 of Moore’s Law cycles for memory, networking, storage, processing

Every day we create as much information as we did from the dawn of civilization up to 2003.

- Eric Schmidt, 2010
**Terminology**

- **Big data**
  - Volume, Velocity, Variety, Value, Veracity
  - Complexity, Latency, Cleanliness, Completeness, Provenance

- **Data lifecycle**
  - From Raw data to Information to Knowledge
  - To a Data Product or Data-as-a-Service

- **Data science (Doing the Right Things)**
  - Data used as evidence through hypothesis and experiment
  - Opportunistic experiment versus designed experiment
  - End-to-end data life cycle

- **Data engineering (Doing Things Right)**
  - Design and construction of software systems
## Big Data Is Not Just About Volume

<table>
<thead>
<tr>
<th>Engineering</th>
<th>Science</th>
<th>Analytics by Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Volume</td>
<td>• Veracity</td>
<td>• Structured</td>
</tr>
<tr>
<td>• Velocity</td>
<td>• Complexity</td>
<td>• Semi-structured</td>
</tr>
<tr>
<td>• Variety</td>
<td>• Cleanliness</td>
<td>• Unstructured</td>
</tr>
</tbody>
</table>

**Business**
- Value

**Analytics by Type**
- Structured
- Semi-structured
- Unstructured

**Provenance**
The Data Life Cycle
We’re in the Data Business

"Ten crates of data and one little envelope of information. Sign here."

 Benefit
 Action
 Knowledge
 Information
 Raw Data
Data Life Cycle Processes

- Collect
- Curate
- Mission
- Analyze
- Monitor
- Visualize
- Act

SAIC Data Science Life Cycle

© SAIC. All rights reserved.
<table>
<thead>
<tr>
<th>Mine</th>
<th>Military</th>
<th>SAS®</th>
<th>CRISP-DM (data mining)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect</td>
<td>Observe</td>
<td>Sample</td>
<td>Business understanding</td>
</tr>
<tr>
<td>Curate</td>
<td>Orient</td>
<td>Explore</td>
<td>Data understanding</td>
</tr>
<tr>
<td>Analyze</td>
<td>Decide</td>
<td>Modify</td>
<td>Data preparation</td>
</tr>
<tr>
<td>Act</td>
<td>Act</td>
<td>Model</td>
<td>Modeling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assess</td>
<td>Evaluation</td>
</tr>
</tbody>
</table>

CRISP-DM = Cross Industry Standard Process for Data Mining
SAS is a registered trademark of SAS Institute Inc. in the U.S. and/or other countries.
Analysis

PRESCRIPTION
- What’s the best that can happen?
- How do I make it happen?
- Recommendation

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

OPTIMIZATION
- What will happen next?
- Forecasting

PREDICTION
- What am I missing?
- Mining

CONFIRMATION
- Why is this happening?
- Correlation

DISCOVERY
- What is happening now?
- Alerting

DESCRIPTION
- What has just happened?
- Reporting

Value

HINDSIGHT  INSIGHT  FORESIGHT

Complexity

© SAIC. All rights reserved.
SAIC.com
Data Science Is Going Agile
Historical Data Emphasis

TRADITIONAL EMPHASIS

ALGORITHMS

SYSTEM

DATA

DATA ENGINEERING

ALGORITHMS

SYSTEM

DATA

DATA SCIENCE

ALGORITHMS

SYSTEM

DATA
Big Data Engineering
Data Services Life Cycle

Data Services: Architect, Model, Cleanse, Query, Visualize, Monitor, Measure, ...

Big Data

internal data

external data

Engineering Services: Secure, Store, Backup, Cluster, Archive, Transmit, ...

Need

Collect

Collect

Curate

Curate

Inform

Information

Analyze

Analyze

Knowledge

Act

Act

Benefit

© SAIC. All rights reserved.
Traditional Data Life Cycle

**COLLECT**
- ETL
  - Staging

**CURATE**
- Cleanse Transform
  - Warehouse

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

**ACT**
- Action

ETL = extract, transform, load
Big Velocity Engineering

COLLECT

CURATE

ANALYZE

ACT

Velocity

Domain

Cleanse Transform

Enriched Data Cluster

Volume

Alerting
SAIC: High Velocity DigitalEdge™ Platform

Setup Tools
- Data Model Editor
- Table Manager
- System Builder

Runtime Tools
- Tenant Console
- System Monitor

Alerting Tools
- Fusion and Enrichment
- Real Time Analysis

Analysis Tools
- Hadoop/HDFS
- Long Term Threat Analytics

Query Tools
- Enrichment Sources

Sources
- Source 1
- Source 2
- Source 3
- Source X

Eucalyptus (In-House) or Amazon (Internet-Accessible)
Big Volume Engineering

COLLECT

CURATE

ANALYZE

ACT
SAIC Big Volume Scale2Insight™ Platform

S2i Model Analytics
- User defined analytic
- User defined analytic

S2i Model Building
- Mahout
- Graph
- User defined

Storm

Oozie

Elastic Search

HBase

Hadoop

HDFS

© SAIC. All rights reserved.
Scale2Insight and DigitalEdge™ are trademarks of Science Applications International Corporation in the U.S. and/or other countries.
Big Variety Engineering

COLLECT

CURATE

ANALYZE

ACT

Variety

Complexity
Big Data Engineering and Architectures
Why Clouds?

- Scalability
- Turn-around speed
- Analytics
- Shared services
- Cost
- Efficiency
- Security

- Public
  - Government
- Private
- Community
- Hybrid
Data Analysis Use Case

COLLECT | CURATE | ANALYZE | ACT

ETL = extract, transform, load

ETL = extract, transform, load
Traditional Data Silos

Silo1
Sensor
Access

Silo2
Satellite
Access

Silo3
Modeling
Access

Agency

Publish / Subscribe Services
Data Silo Use Case – Variety

COLLECT

Silo1
Access
Department 1

Silo2
Access
Department 2

Silo3
Access
Department 3

Optional Integration
Logical Data Store

Private Cloud

ANALYZE

ACT

Optional Integration

Analyse
Traditional Coupled Systems

- Silo1 (Sensor Access)
- Silo2 (Satellite Access)
- Silo3 (Modeling Access)
- Silo4 (Agency 2 Access)
- Publish / Subscribe Services
Data Sharing Use Case

COLLECT

CURATE

ANALYZE

ACT

Data Stores

Integration

Community Cloud

Data-as-a-Service

Internal Environment

SAIC.com

© SAIC. All rights reserved.
Data Publication Use Case – Volume

COLLECT

ETL

Data Store

CURATE

Integration

ANALYZE

Transmit

ACT

Velocity

Public Cloud

Internal Environment

SAIC.com
Data Ingestion Use Case – Velocity

COLLECT

PUBLIC CLOUD

PUBLIC CLOUD

Internal Environment

CURATE

ANALYZE

ACT

SAIC.com

© SAIC. All rights reserved.

ingest

ETL

Data Store

Metrics

Metrics
Data Synchronization Use Case

COLLECT

CURATE

A N A L Y Z E

ACT

Cloud Environment

Submit

ingest

NoSQL Store

Analyze

ingest

Query

ETL

Data Store

Analyze

Transmit

Internal Environment

© SAIC. All rights reserved.
Adding Big Data and Hybrid Clouds

FROM DATA

TO INFORMATION

Big Data Ingestion

External Data

Ingestion, Fusion, Enrichment, Alerting

Custom Enrichment

TO EXPLORATION

Big Data Analytics

Enriched Data Storage

Browsing Modules

TO KNOWLEDGE

Query, Modeling, Characterization, Prediction

Custom Analytics

TO INSIGHT

Cloud Environment

Internal Environment

Data

Enrichment Sources

Query Analysis

Models

Analyst Tools

SAIC.com

© SAIC. All rights reserved.
Some Big Data Thoughts

- We have existing analytics infrastructures
- We want to add in big data and have scalable systems
- We need to cross data silos
- We want to increase efficiencies and reduce cost
- Our 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 data services (to better include mobile)
Big Data Technologies Bring Mission Opportunity

You don’t need to worry about the quantity...

You just need to
• Engineer YOUR surfboard
• Learn the science of surfing!

"Eureka! More information!"
Nancy Grady, Ph.D.
SAIC Technical Fellow, Data Science
Homeland and Civilian Solutions

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