The Modern Data Platform
Enterprise Data Hub Implementations

What is happening with Hadoop adoption
Why is workload moving to Cloud - Trends and Considerations

DAMA Chicago

Dr. Phil Shelley
The Modern Data Platform
The Enterprise Data Hub

What do we mean by MDP and EDH?

• A single point of truth for data
• Analytics separated from data integration and storage
• Real-time data integration where needed
• All workload referencing a single data source
• Layered and secured data architecture
• Data lineage visibility with audit capabilities
• A dynamic and flexible infrastructure
• Analytics tools brought to the Data (not moving/copying Data to the tools)
• Tool agnostic
• Future proofed architecture
• Designed to avoid software and vendor lock-in

Bring tools to the Data - NOT move and copy data to the tools
What is an EDH (Enterprise Data Hub)

- A Hybrid Environment - Adaptable for all Enterprises
- Supports important concepts
  - Data Hub – "Source-Once and Reuse" approach
  - Transformation - ETL → ETL
  - Heavy batch processing
  - Real-time data integration
  - Real-time Analytics
- Legacy Migration/Modernization
- Low-cost Intelligent Archive

- Source-Once and Reuse
- Drives efficiency
- Reduces data silos
- Reduces latency
- Faster time to value
- Improves analytics
- Greatly reduces costs
Big Data – Layered Architecture

- Defines data layers based on usage patterns
- Defines clear data separation model
What is the Status of Hadoop

Status

• More than a decade old
• Mature and stable
• In widespread use
• Deployments still growing

Changes/Trends

• New workload is using less “pure” Hadoop
• Hadoop is being used for:
  • Data ingestion
  • Profiling
  • Data quality
  • ETL replacement
• It is being used less for BI and Analytics
• Hadoop is being mixed with Cloud BI tools
• Growth in EMR and Apache Hadoop to control costs
• Cloud storage, especially S3 and GCP are taking significant share away from Hadoop storage
Analytics and Hadoop on Cloud - Why not

Why Not

- Cloud has (some) ongoing stigma
  - Related to putting data into a shared environment
  - Lacking granular control of user activity and access
  - Cloud can be expensive if not well managed
  - Enterprise security teams have concerns about the use of cloud
  - Some perceive that cloud adds complexity
  - Support costs and complexity are a concern
  - Some like to own everything within their own “walls”
  - Pay-for-use is uncomfortable for some IT executives who like to control their spending limits
- Some say “we already made huge investments in Data Warehouses and analytics tools, we need to use them”
- Copying data to another environment takes time, bandwidth, people, tools and added cost
- Some say they they want fewer tools and technologies, not more
- Skill shortage is a concern
- Audit groups feel that they will loses control and visibility
- Architects say that data Lineage becomes more complex
Cloud Investment

Analysts and vendors agree - Seeing a massive migration of dollars and data into "cloud"

**Gartner**, says:
- Worldwide public cloud services will grow 18% this year to become a $247 billion business
- Cloud will easily account for the majority of analytics purchases by 2020

**Forrester**
- Cloud growth at 19%
- Says it will be a $162 billion business in three years

This projected growth appears to come, in part, at the expense of on-premise Hadoop and Data Warehouse deployments

The installed base of Hadoop clusters in 2020 will not be as big as the industry widely believed during the peak Hadoop years of 2012 to 2014
Analytics and Hadoop on Cloud - Why

Why

● Cloud has fully “come of age”
● All tools are available
● Scale is unlimited
● Performance is equivalent and better in most cases
● Powerful cost advantages - Caution
● Security is fully matured and tested
● Simplify environment - e.g. Google BigQuery
● Reduce support costs and complexity of upgrades
● Agility - Spin-up and down platforms and BI tools in minutes
● Pay for only what is used
● Set usage limits and see the projected cost in advance
● Focus on analytics to drive value, not infrastructure
● Reduce or eliminate traditional software licensing
● Eliminate most traditional ETL - Reduce cost and data latency
● Easily adopt new tools as they become available
Analytics on Cloud - Costs

Cost

- Hadoop was considered an economical solution for both structured and unstructured data
  - It is designed to handle almost any workload workloads
- Deploying Hadoop on-premise was the standard over the past 10 years
- Cloud options were immature and expensive

Not any more....

- Organizations deploying BI/Analytics in the cloud
- Saving of 20% to 60% over on-premises infrastructure cost
  - The cost of cloud storage is as low as 1 to 3 cents/Gb/month
  - Separate storage and compute - A revolutionary change from legacy Hadoop
  - Eliminate most or all capital planning
  - Cluster capacity planning is gone
    - Even the concept of servers, nodes or clusters is fading
  - Buying capacity ahead of the need is gone
  - Short-term increases in capacity is now a routine capability
  - Infrastructure upgrade costs are mostly eliminated
  - Changes in hardware become irrelevant
Analytics and Hadoop on Cloud - Agility

Agility

● Analysts have anything at any time (within security profile restrictions)
  ○ Any data
  ○ Any BI tool
  ○ Any scale
  ○ Any time

● A well governed Enterprise Data Hub (EDH) eliminates up to 50% of analysts time to gather data
  ○ Faster advanced analytics
  ○ Analysts are much more productive
  ○ Faster time to dashboards with full historical data and full fidelity
  ○ Focus on business performance KPI analytics

● Tools can be tested and used dynamically
  ○ Advanced analytics and machine learning to build predictive models
  ○ Easily move into text analytics to analyze unstructured data
  ○ Leverage real-time streaming analytics

● Development teams can have their own clusters
  ○ Segregate the compute function for each team
  ○ Share read-only MDP/EDH data
  ○ Stress test in isolated environments
  ○ Geographically dispersed

● Backup and geographical distribution is much simpler, faster and simpler
Traditional Data Warehouse costs remain high are leading to this fast moving trend

Architecture, Accelerators and Cloud optimization are important to reduce risk, mistakes and speed modernization

Automation and tools
- Greatly speed the migration of Data Warehouse workload to Cloud
- Convert data, jobs and queries from Teradata, Netezza, Greenplum, Mainframe and more
- Mature cloud architectures reduce time to production
- Data matching and validation tools cross-platform, cloud and on-premise reduce errors and mistakes
- Automated Data model migration
- Automated SQL migration
- Automated migration project sequencing
- Test the migration plan in advance to avoid production issues
Analytics on Cloud - Simplify

Simplify

- A well-design MDP/EDH greatly simplifies the data landscape
- Sourcing data directly from service buses or transactional systems, removes an ETL layer
- Infrastructure is vastly simplified and much faster
  - Deploying an on-premise data Warehouse or Hadoop cluster takes months
  - Purchasing, installing and configuring new hardware takes weeks
- Training staff is reduced, particularly infrastructure and operations
- Dev/Test/QA environments are simple and flexible
  - Have as many as needed, just for as long as needed
  - Development teams can have isolated environments and access to full data sets
  - Stress testing is simple and fast in separate compute environments
- Capital planning is greatly simplified
- Chargeback to businesses is simpler and can be automated
- Tools are rich for automation of processes
- Reduced software licensing complexity as many tools are pay-per-use
- Accelerators speed the migration to Cloud and reduce errors
Why is workload moving to Cloud?

.... I think you can see why!
### About Us

| DataMetica was created and is led by some of the strongest experts in Cloud and Hadoop |
| From Ideation to Support – we are market leaders in Big Data, Cloud migration and Analytics |
| We come from a background of implementing large Big Data solutions in numerous global companies, our accelerators driving Data and Analytics project delivery time down by typically 50% |
| Expert technical know-how and proven implementation methodologies in Big Data & Cloud technologies - Including modern data platforms, operational model and governance processes |
| Help clients implement their Big Data Analytics, Cloud, BI and Performance Management priorities |
| Global Delivery Centre in India, offering cost-effective access to leading Big Data resources |
Example of Enterprise Data Hub build
Angie’s List

Presented by Bryan Tabiadon
Analytics Mission

Empower Angie’s List Employees, Members and Service Providers to make better decisions with actionable insight from data
Data Management Mission

Organize and Centralize Angies's List data by:
   Make it universally accessible
   Provide data enablement services
   Standardize Business Rules
   Provide the highest quality of data
Angie’s List Big Data Objectives

Key deliverables:
Provide high quality, real-time data, analytics and self-serve reporting.
Enterprise Data Hub on AWS-Conceptual

**Data Integration**
- Batch Processing: Sqoop/Flume
- Real Time Processing: Flume/Storm Spark

**Data Processing-AWS ERM**
- Batch Processing: Pig / Data Pipeline
- In Memory / Real Time: Analytical SQL
- Hive/Spark SQL: 3rd Party Data

**Data Management –Hadoop Platform**
- Metadata Management: Business Rules, Data Profiling etc.
- Hadoop Operating System / Workload / Data Pipeline
- Storage: Amazon S3

**Security**
- Access Management: AWS IAM/Jenkins Jobs

**Operations**
- Scheduling Workflow: AWS Data Pipeline
- Manage & Monitor: AWS Console & CloudWatch

**Enterprise Data Hub**

**Angie’s List Customer 360 Data Model**
- Schemas: Webmetrics, Sales, Ecom, Leadfeed, SP Cockpit, CEO, SEC

**Data Sources**
- AL 4.0
- Legacy
- Segment
- HR
- SFMC
- IVR
- Vindicia
- BrightNest
- DMP-Krux
- Bluebook

**3rd Party Partners**
- Complete
- In Progress
- Not Started
Enterprise Data Hub
Logical Architecture built on AWS

Source System

Provisioning Layer

Data Serving

Red Shift/Dynamo DB

Spark/Hive/Presto

Adhoc Querying

Processing Layer

Batch Processing

EMR

Adhoc Processing

On Demand Infrastructure

Storage Layer

Cloud

S3

Enterprise Data Hub

Incoming Layer

Basic File Level Validation
Persisted
Full History/Detail
Concatenated
Data Quality
Data Profiling

Gold Layer

Single Point of Truth
De-normalized
Concatenated
Segmented
User defined marts

Presentation Layer

Persisted Marts
Temporary Marts
Exports Marts
Security Driven Marts
Batch Working Area

Data Validation
Data Cleaning

Data Processing

Metadata Layer

Hcatalog

Hive Metastore

Management

Security Layer

IAM (Identity & Access Management)
AWS Directory Service (LDAP)

AWS Key Management Service

On premise/Cloud

Error Logging

Audit

Exceptions

Statistics

Alerts

Scheduling

Users

Low Latency Queries/Dashboard

Power BI Tableau

Adhoc Queries/Deep Analytics
Conceptual-Customer Model- Hive Schema

Customer Hub
Members/Non-Members/Service Providers

- Products/Contracts
- Sales (Transactions)
- Reviews
- Financials
- Omni Channel
- Geo/Households
- Customer Behavior (Web Metrics, Email etc.)
- Marketing/Social/DMP
- Organization (Sales/Marketing/Ads)
Enterprise Data Hub Physical Data Flow

Sources
- AL 4.0
- Legacy
- DMP
- Segment.io
- SFMC
- IVR
- Vindicia
- HR Oracle
- BrightNest
- Dunn & Bradstreet
- Bluebook

S3 Ingestion
- Source
- S3 Ingestion
- S3 Raw
- EDH Ingestion
- EDH Raw
- PIG Scripts
- AL 4.0
- Legacy
- DMP
- Segment.io
- SFMC
- Vindicia
- Oracle
- brightNest
- Dunn & Bradstreet
- Bluebook

GitHub Scripts maintained:
- Data Quality Rules
- Error logging
- Auditing
- Deduplication
- Data Lineage
- Data Profiling
- Data Dictionary
- Master Data Management

EDH Golden
- AL 4.0
- Legacy
- DDM
- Segment.io
- SFMC
- IVR
- Vindicia
- Oracle
- brightNest
- Dunn & Bradstreet
- Bluebook

Hive Schemas
- Web Metrics
- Sales
- Leads
- Event Dataa
- KPIs
- CEO

Data Quality
- Audit
- Error
- Lineage
- eCat glossary

Shared Dimension Tables

Data Governance

Delivery
- BI Self Service
- Marketing Platform Segmentation & Activation
- Partnerships
- Corporate Data Architecture Governance Strategy
- 3rd Party Vendor Data Exchange, D&B, Bluebook, BrightNest
- Application Delivery, AL 4.0 Tools, Salesforce
- Customers BYO
## Enterprise Data Hub Data Attribution Inventory

<table>
<thead>
<tr>
<th>Source</th>
<th>Total Tables</th>
<th>Total Dev/Staging/UAT</th>
<th>EDH Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL 4.0</td>
<td>496</td>
<td>190</td>
<td>70</td>
</tr>
<tr>
<td>Bluebook</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Brightnest</td>
<td>190</td>
<td>160</td>
<td>58</td>
</tr>
<tr>
<td>Drupal Content</td>
<td>610</td>
<td>420</td>
<td>126</td>
</tr>
<tr>
<td>Drupal Impressions</td>
<td>12,456</td>
<td>1,604</td>
<td>745</td>
</tr>
<tr>
<td>IVR</td>
<td>294</td>
<td>250</td>
<td>90</td>
</tr>
<tr>
<td>DMP-Krux</td>
<td>456</td>
<td>290</td>
<td>200</td>
</tr>
<tr>
<td>Legacy</td>
<td>8,409</td>
<td>925</td>
<td>57</td>
</tr>
<tr>
<td>Oracle</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Redshift</td>
<td>74</td>
<td>67</td>
<td>19</td>
</tr>
<tr>
<td>Segment (Raw JSON)</td>
<td>All</td>
<td>All</td>
<td>-</td>
</tr>
<tr>
<td>SalesForce Dot Com</td>
<td>340</td>
<td>340</td>
<td>178</td>
</tr>
<tr>
<td>SalesForce Marketing Cloud</td>
<td>34</td>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>Sizemek</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Vindicia</td>
<td>161</td>
<td>161</td>
<td>30</td>
</tr>
<tr>
<td>Home Survey</td>
<td>300</td>
<td>100</td>
<td>68</td>
</tr>
<tr>
<td>Data.Com</td>
<td>605</td>
<td>130</td>
<td>98</td>
</tr>
<tr>
<td>Dunn &amp; Bradstreet</td>
<td>200</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>24,629</strong></td>
<td><strong>4,740</strong></td>
<td><strong>1,813</strong></td>
</tr>
</tbody>
</table>
Enterprise Data Hub-Hadoop Nodes by Workload Clusters
Thank you!

Questions