Modernize your EDW into Big Data Warehouse

Feb 2016
• Sanjay Sharma
  – Principal Architect  @IMPETUS
  – 18+ years in Software, 6+ with Hadoop
  – Started Hadoop User Group in 2009
  – Hortonworks, Cloudera, MapR certified

• Email - ssharma@impetus.com
• Tweet - @indoos

IMPETUS

• Headquartered in Los Gatos; established in 1996; 1700+ employees; presences in Silicon Valley, Atlanta, NYC, and offshore.
• Unique products accelerated, innovation and R&D driven, vendor neutral consulting and solution services
FIRST CAR

Early 1900s

FLYING CAR

~2015

Source: Google images
Some Timelines

1970s
RDBMS became a recognized term
MS SQL Server, Sybase, Wang’s PACE, and Britton-Lee
SQL/DS, DB2, Allbase, Oracle, and Non-Stop

A normalized data model is designed first. Then the dimensional data marts, which contain data required for specific business processes or specific departments are created from the data warehouse.

http://searchbusinessintelligence.techtarget.in/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse
The data marts facilitating reports and analysis are created first; these are then combined together to create a broad data warehouse.
Logical Data Warehouse Reference Framework

DQ = data quality; Gov. = governance; MDM = master data management
Source: Gartner (June 2012)
Data Warehouse needs?

- ETL?
- Streaming?
- Ingest Any
- Data Cleansing?
- Data Integration Access (Physical/Virtualization)?
- Data Model?
- Intelligent Data Discovery?
- CDC/SCD?
- Metadata/Reference
- Intelligent Data Identification?
- Real Time Feedback Interface
- Reports
- Dashboards
- Adhoc Analytics
- Advanced Analytics

Data Governance & Security
(Authentication/Authorization/Audit/Data at Rest)
(Lineage, ILM)
Data Warehouse - Ingest Side

- Mainframe/OLAP DW?
- OLTP/ESB/Messaging Queue/SOA
- Social Media/Textual Data

Any source ETL?
Streaming Ready?
Easy Metadata Identification?
Faster ETL?
Easy Data Modelling?
CDC/SCD?
Metadata/Reference
Lineage?

Data Governance & Security
(Authentication/Authorization/Audit/Data at Rest)
(Lineage, ILM)
Data Warehouse- Storage?

Data Store

- Data Model?
- CDC/SCD?
- Metadata/Reference
- Data Preparation Processing?
- Metadata Management (Ref data)?
- DR/Backup/ Business Continuity
- Data Governance & Security (Authentication/Authorization/Audit/Data at Rest) (Lineage, ILM)

Intelligent Data Identification?
Data Warehouse - BI/ Analytics/ Integration?

- Data Integration Access (Physical/Virtualization)?
- Big Data Queries/Scalable Processing?
- Intelligent Data Discovery?
- Data Model?
- Derive Value
- JDBC/ODBC/ReST Integration?
- Metadata/Schema Discovery?
- Search?
- Fast Queries?
- Data Science?
- Real Time Feedback Interface
- Reports
- Dashboards
- Adhoc Analytics
- Advanced Analytics

Data Governance & Security
(Authentication/Authorization/Audit/Data at Rest)
(Lineage, ILM)
<table>
<thead>
<tr>
<th>Integrated 1. Big Data Platform</th>
<th>Data Warehouse Appliance</th>
<th>Active Enterprise Data Warehouse</th>
<th>Aster Big Analytics Appliance</th>
<th>Appliance for Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3.8K/TB</td>
<td>$34K/TB</td>
<td>$34K-$69K/TB</td>
<td>$8K/ TB</td>
<td>$2K/ TB</td>
</tr>
</tbody>
</table>

Source: http://blogs.teradata.com/data-points/how-illy-is-cost-per-terabyte/
Enterprise Hadoop

YARN: Data Operating System

HDFS (Hadoop Distributed File System)

Source: Hortonworks.com
MODERN DATA WAREHOUSE STRATEGY

BI DASHBOARDS

REPORTING

Limited ROI

Legacy EDW/RDBMS

PLANNED ANALYTICS

LIMITED ANALYSIS

COST EFFICIENCY

LINEAR SCALABILITY

HADOOP + ETL/Analytics Offloading

LIMITLESS DECOMOCATIZED ANALYTICS

ALL FORMS OF DATA

Structured
Unstructured
Semi-structured

Transformation for scale

Expanding ROI
**Traditional EDW Augmentation Example**

**EDW Key Challenges**
- Storage and processing scalability costs
- Static non-agile BI/Analytics
- Processing capacity hogged resulting in slow analytics

**Transition Challenges solved by Workload Migration**
- Manual process and so error prone
- Time taking and risky
- Migration Validation/ QA problem
- Hadoop best practices not known/missing
- Know what/how to offload
- Real time data

**BDW Advantages**
- Cost effective scalable storage and processing (10X-100X)
- Agile Self Service BI/Analytics
- Offload resources from Netezza/EDW and SAS for freeing analytical capabilities
- Exploratory environment and Swiss knife toolsets for Machine learning driven analytics
- In time ETL, BI and Analytics
- Streaming and Real time Analytics
- Unstructured data Analytics
Big Data Analytics Triple Play

Big Data Analytics - Streaming + Real-time + Batch

- **Fraud Detection in Real time**
  - Near-term
  - "What is happening"
  - Insight

- **Historical Transaction Analysis**
  - Historical
  - "What happened"
  - Hindsight

- **Fraud Analytics based on Historical Data**
  - Inferential
  - "What may happen"
  - Foresight

- **Personalized/Adaptable Fraud Detection in Real time**
  - Influential
  - "Make it happen"
  - Intelligent

Enhance Static Rules

Intelligent input for Dynamic Rules

BATCH

Information Discovery/Search

Traditional BI/Analytics

REAL TIME

1 3

2 4

BASIC ANALYTICS

ADVANCED ANALYTICS

Lambda/Kapda Architectures
EDW/RDBMS Modernization - Key Areas

- Scalable cost effective Storage and Query Engine
- Streaming/ real time capability
- Unstructured/semi structured with Structured mashup

- Schema on Fly
- ETL Offloading
- Analytics Offloading

- Democratized access
  - Self service BI/Analytics/ELTL
- Information Discovery/Search
- Security, Governance, Compliance
Scalable cost effective Storage and Query Engine

Unstructured/semi structured with Structured mashup

Schema on Fly
Scalable cost effective Storage and Query Engine

Streaming/ real time capability

kafka

STORM

Flink

Spark Streaming

STREAMANALYTIX™
ETL Offloading

Hive
Spark
Alteryx
Talend
Informatica

Big data
Big data Edition

© 2016 Impetus Technologies - Confidential
Analytics (SAS/SPSS)
Democratized access
Self service BI/
Analytics/ ELTL
Augmenting EDW with Hadoop (Reference Arch.)

Use Hadoop as a foundational data warehouse solution along with a relational big data warehouse for handling both structured and unstructured data analytics.
## Impetus Big Data Services & Products

<table>
<thead>
<tr>
<th>Big Data Services</th>
<th>Big Data Products/Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big Data Advisory, Strategy &amp; Training</strong></td>
<td>Real-time ingestion and analytics</td>
</tr>
<tr>
<td><strong>Big Data Labs, POV, Benchmarking, Training</strong></td>
<td>STREAMANALYTIX™</td>
</tr>
<tr>
<td><strong>Technology Evaluation &amp; Piloting</strong></td>
<td>Universal BI on Big-Data</td>
</tr>
<tr>
<td><strong>Big Data QA Services</strong></td>
<td>EDW Modernization</td>
</tr>
<tr>
<td></td>
<td>WORKLOAD MIGRATION</td>
</tr>
<tr>
<td></td>
<td>DATA BLENDING</td>
</tr>
<tr>
<td></td>
<td>DATA ACCESS</td>
</tr>
<tr>
<td></td>
<td>METADATA MANAGEMENT &amp; DISCOVERY</td>
</tr>
<tr>
<td></td>
<td>Hadoop Management/Integration</td>
</tr>
<tr>
<td></td>
<td>jumbune</td>
</tr>
<tr>
<td></td>
<td>Kundera</td>
</tr>
<tr>
<td></td>
<td>Ankush</td>
</tr>
</tbody>
</table>

Impetus Big Data Services & Products

<table>
<thead>
<tr>
<th>Big Data Services</th>
<th>Big Data Products/Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big Data Advisory, Strategy &amp; Training</strong></td>
<td>Real-time ingestion and analytics</td>
</tr>
<tr>
<td><strong>Big Data Labs, POV, Benchmarking, Training</strong></td>
<td>STREAMANALYTIX™</td>
</tr>
<tr>
<td><strong>Technology Evaluation &amp; Piloting</strong></td>
<td>Universal BI on Big-Data</td>
</tr>
<tr>
<td><strong>Big Data QA Services</strong></td>
<td>EDW Modernization</td>
</tr>
<tr>
<td></td>
<td>WORKLOAD MIGRATION</td>
</tr>
<tr>
<td></td>
<td>DATA BLENDING</td>
</tr>
<tr>
<td></td>
<td>DATA ACCESS</td>
</tr>
<tr>
<td></td>
<td>METADATA MANAGEMENT &amp; DISCOVERY</td>
</tr>
<tr>
<td></td>
<td>Hadoop Management/Integration</td>
</tr>
<tr>
<td></td>
<td>jumbune</td>
</tr>
<tr>
<td></td>
<td>Kundera</td>
</tr>
<tr>
<td></td>
<td>Ankush</td>
</tr>
</tbody>
</table>

Impetus Big Data Services & Products
ELTL Workload Migration: Automation Steps

The workload migration consists for the following major tasks

- **Identification of ideal workload**
  - Usage based tables classification
  - Workload based tables classification

- **Migration of Schema from current RDBMS to HDFS**
  - Recommendation for partitioning, clustering and number of buckets
  - Migrate role based security

- **Query/SQL Logic Migration**
  - Impetus UDF Library
  - Automatic conversation of SQL and Stored Procedure scripts

- **Query/SQL Logic Execution**
  - Schedule migrated code

- **Data Quality Enhancements**
  - Patented Machine Learning Algorithms

- **Data Restoration**
  - BI/Analytical data can be synced back with EDW if required

The above tasks have flexible implementation sequence to facilitate the nature of the workload and customer’s implementation strategy.

60%-80% - Automation!
30%-60% - Reduced Manual Time!
Reduced Risk!
Impetus Workload Migration Architecture

Step 1: Identification

Step 2: Schema, Security and Data Migration

Step 3: Logic Migration

Step 4: Logic Execution

Step 5: Data Quality Enhancement

Step 6: Analytical Data Moved to EDW Datamart

Reduced Time!
Reduced Risk!
Automation!
Thank you.

Questions?

WE ARE HIRING -  http://www.impetus.com/careers
ETL Offloading to Hadoop- Healthcare Payer Case Study

The Challenge
- Optimize space usage & processing on Teradata
- Schema evolution
- Versioning of data to enable capability of adding attributes to base
- Immunity to ‘data load sequence’
- Enable UDW equivalent security
- Delta detection for snapshot acquisition

Impetus Contributions
- Implement data provisioning for UDW and non-UDW databases
- Provide parametric security for Data Fabric platform
- Proposed reference architecture to handle schema evolution, versions management, data consolidation & transformation
- Metadata Management & Data Lineage

Benefits Realized
- Performance for on-line and batch processes
- Near real time processing of input
- Provisioning through web-services
ETL Offloading to Hadoop - Healthcare Payer  

- Source Data:
  - X12 837/835 CIC
  - HL7 V3 OCIE
  - Custom ECG
  - SMTP Email

- Data Fabric/Hub:
  - Hadoop + HBase + Hive/PIG

- Data Consumers:
  - Teradata EDW
  - Unified Format

- ETL Offload

- Cloud based Provisioning - DaaS

- Targeting 30%-60% reduction in Teradata Cost
- Schema evolution and version management challenges
- Compliance strategy

Close to $50 million savings!!!