Data Meaning

Big Data Analytics Approach
LET’S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
Agenda

• Welcome to Big Data
• Approach Implementation Tactics
• Working with Big Data
• Cheat Sheet to Analytics Over Big Data
Welcome to Big Data
According to Gartner:

Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight, discovery, and process optimization.
For our discussion today, we are going to look at approach implementation tactics. As such, to limit the possible branches of this discussion we will assume the following:

How We Got There
Alternative Workflow
Alternative Shaping
Alternative Loading
The Need for Life-Cycle Management
You cannot extract the entire source system(s) data because storing it on an RDBMS entails prohibitive COST/COMPLEXITY.

- You’d need to define the data structures of everything in the source system first before storing it. That means understanding it!

As the reporting system naturally grows, you have to revisit every layer along the solution path with each new field added.

- This makes your solution ridged, fragile and unchanging, but we lived with it!

The **cost of storing and structuring your data** in order to capture it IS THE CONSTRAINT that created this approach way back in the day. This is referred to as your TECHNICAL DEBT. You’ve been working around it ever since.
Working with Big Data: Where You Are Going

- The HADOOP collects data from all source systems ... all day ... every day.....ALL OF THE DATA!
  - This is referred to as Exhaust Data Collection, no need to model and determine what’s valuable. Its all valuable at some point. We just do not yet know its value ... Collect it before it is removed by the source system
- All ETL[Data Shaping] now happens on the Hadoop. Artifacts from that process(s) land on other platforms for Analysis
- There is a single starting point for all data journeys: the data lake
- Data shaped artifacts shipped to Analytics Tier(s), but all have a common source
- Analytics Tier can remain as it was; dependent on PDM
- LAMDA operations can be implemented to eliminate data drift (total replacement of shaped artifacts on Analytics Tier)
- TECHNICAL DEBT is recovered, eliminating the cost and complexity of doing E, T & L the old way and use of PDM
- PDM for capture/archive now eliminated
- Resources and cycles now freed on Analytic Tier (because aggregations being done on Hadoop)
- Life-Cycle Mgmt. Exposure only to Analytics Platforms
Working with Big Data: Shaping

HDFS: daily_sales  +  HDFS: addresses  +  HDFS: clients  +  HDFS: items

HDFS: part_1
HDFS: part_1
HDFS: part_1
HDFS: part_1
HDFS: part_1
HDFS: part_2
HDFS: part_2
HDFS: part_2
HDFS: part_2
HDFS: part_2
HDFS: part_1
HDFS: part_3

1 billion rows of sales data in a single folder
Smaller amounts of enumerate data in a single folders

HDFS: daily_sales_detail
HDFS: sale_month_yr = 08-2017
HDFS: item = item_1
HDFS: item = item_2
HDFS: sale_month_yr = 08-2017
HDFS: item = item_1

Re-shaped folder structures with files
Partitioned by:
Sales month-YR, ITEM

Spark/Scala Data Shaper

Hive 2
Working with Big Data: The Need for Life-Cycle Mgmt.

You'll benefit from suggesting to your Hadoop Data Shaper teams a means of life-cycling the data. Because Hadoop admin and data shaper teams are traditionally software developers and not database administrators, your suggestion about how they should manage the life-cycle during updating your analytics tier is critical. Without said suggestion, they will be updating records under your queries without any form of management, which leads to query disaster.
Working with Big Data: Workflow

- **DB: Expansion Request**
- **Dashboard Team**
  - Newly Added Tables
  - Existing Tables
- **Shaper Team**
  - Scala: Author shapers for each table(s)
  - Bundle: All shapes, all tables plus table life-cycle mgmt, scripts
- **Admin Team**
  - Workflow: Deploy scripts into Enterprise Zone for running
  - DATA shaped by physical sort order and folder hierarchies
  - Scheduled: Execution of bundle run at some interval

**Delivery**: Request for new table descriptions
**Delivery**: Relationship between new data
**Delivery**: Example, WHERE, GROUP, ORDER by clauses over data

**Dashboard Team**
- I need more data to make this work
  - Delivery: Request for new table descriptions
  - Delivery: Relationship between new data
  - Delivery: Example, WHERE, GROUP, ORDER by clauses over data

**Shaper Team**
- Where's the data?
- Here's the data you're looking for

**Admin Team**
- Need something like this

**New Tables**
- Newly Added Tables
- Existing Tables

**ODBC**
Analytics and Reality

Capture
- Stage 0: Data Capture
- Stage 1: Data Validation
- Stage 2: Archive
- Stage 3: Reshape Data

Data Lake Cluster
- Analytics Platforms
  - MicroStrategy
  - Snowflake

Archive

Analytics

©2018 Data Meaning Services Group Inc. All Rights Reserved. Contents of this presentation are proprietary and confidential.