Supply Chain Data: Pains, Pitfalls, & Some Rays of Hope

Dr. Mamnoon Jamil, IBM Chief Information Office
(mamnoon@us.ibm.com)

Dr. Cheranellore(Vasu) Vasudevan, IBM Chief Data Office
(vasuv@us.ibm.com)
Stories of a Frustrated Data Scientist

❖ Who doesn’t like stories?
❖ Grandma’s stories
  ▪ They were often repeated
  ▪ No rational questions – stories are simply stories !!!
  ▪ But, always, they are interesting to hear and their ending is invariably good

❖ Data Stories
  ▪ Characters in our stories are different
    - They are simply data, data, and data in different colors, shapes, and costumes at different locations
“Data Data Everywhere…..”

➢ There are about two dozens client data repositories
➢ We have no dearth of data
➢ 80% or more world data is generated in last two years (blah.. blah... half of it is garbage !!!)
➢ How do we identify the garbage? (You have to see it, touch it, smell it, and/or brave enough to taste it !!!)
➢ Can we process / clean them to make it useful?
➢ How can we separate the garbage and gems hidden in it?
Incorrect Data

- It never happens (you wish!)
- Nobody can explain why this happens (even worse)
- Look for anomalies by defining ranges of accepted values, values that are inconsistent, etc.

Example (not real) from data repository!!!

<table>
<thead>
<tr>
<th>Geographic identifier code</th>
<th>Geographic area name</th>
<th>Products and services code</th>
<th>Meaning of Products and services code</th>
<th>Year</th>
<th>Products shipments value ($1,000)</th>
<th>Relative standard error for estimate of product shipments value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3339997</td>
<td>All other miscellaneous general industrial machinery code</td>
<td>2015</td>
<td>6469773</td>
<td>2.3</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3339999A</td>
<td>All other miscellaneous machinery products, exc. - machinery, code</td>
<td>2016</td>
<td>1093753</td>
<td>7.6</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3339999A</td>
<td>All other miscellaneous machinery products, exc. - machinery, code</td>
<td>2051</td>
<td>1097953</td>
<td>7.8</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3339999W</td>
<td>All other miscellaneous general purpose machinery code</td>
<td>2016</td>
<td>1343216</td>
<td>3.8</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3339999W</td>
<td>All other miscellaneous general purpose machinery code</td>
<td>2015</td>
<td>1346917</td>
<td>38</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341111</td>
<td>Electronic computer manufacturing code</td>
<td>2016</td>
<td>8950521</td>
<td>0.5</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341111</td>
<td>Electronic computer manufacturing code</td>
<td>2015</td>
<td>9172030</td>
<td>0.4</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341111</td>
<td>Host computers, mainframes, super computers code</td>
<td>2016</td>
<td>5371882</td>
<td>0.1</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341111</td>
<td>Host computers, mainframes, super computers code</td>
<td>2015</td>
<td>5724332</td>
<td>0.1</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341117</td>
<td>Single user computers, microprocessor-based, code</td>
<td>2016</td>
<td>2678449</td>
<td>0.6</td>
</tr>
<tr>
<td>0100000US</td>
<td>United States</td>
<td>3341117</td>
<td>Single user computers, microprocessor-based, code</td>
<td>2015</td>
<td>25565691</td>
<td>0.6</td>
</tr>
</tbody>
</table>
**Improperly Defined Data**

Products which are withdrawn are still selling !!!

An example from a live order execution data repository!!

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Product #</td>
<td>SALES_DOC_ORDER NC</td>
<td>SALES REPORT DA</td>
<td>ANNOUNCE DAT</td>
</tr>
<tr>
<td>3</td>
<td>63CR12</td>
<td>******0894</td>
<td>1/12/2013</td>
<td>8/17/2010</td>
</tr>
<tr>
<td>7</td>
<td>9527X</td>
<td>******8340</td>
<td>1/18/2014</td>
<td>4/11/2012</td>
</tr>
<tr>
<td>8</td>
<td>98E4C</td>
<td>******1546</td>
<td>7/14/2013</td>
<td>4/2/2008</td>
</tr>
</tbody>
</table>
Wrong Data – the Helplessness!!!

USE THE CRS DATA-BASE TO SIZE THE MARKET.

THAT DATA IS WRONG.

THEN USE THE SIBS DATA-BASE.

THAT DATA IS ALSO WRONG.

CAN YOU AVERAGE THEM?

SURE. I CAN MULTIPLY THEM TOO.
Data that is Fragmented / Incomplete

- Data segmented by Geo/countries
- Data fragmented by product brands / sub-brands
- No data for certain types of orders

- More than 20 sources for customer data in typical large enterprises
- Healthcare data is a good example of data fragmentation – drug data, patient data, clinical test data, med insurance data, etc.
- Duplication and mismatches from different pieces
- Differences in schema, access, and governance techniques
Data that is Expired (old)

▪ Data that was correct (once upon a time !)
▪ No updates made
▪ Updates are made elsewhere
▪ Don’t you know that we migrated to a new warehouse with our “moon shot” Transformation Initiative.......)

▪ Overage Inventory Data
▪ Did some “smart” data transformation by the data modeler who is a student in a class
▪ Got “near perfect” results – who cares if the data was fitted to the model!!!
▪ $R^2$ value of 0.98 looked suspicious in a “doubting professor’s” head
▪ A “can of worms” - what else to expect from rotten stuff

“...DJIW topped 20000.....” (it was in 2016 !!!)
Expired Data: Example Inventory Data

- Data is updated only when an event happens
- Previous Data is replaced with current data
- Data is NOT transactional
- Dynamics of the data changes are not preserved
- Data definition has changed

<table>
<thead>
<tr>
<th>LAST_UPDATE</th>
<th>PLANT_ID</th>
<th>MATERIAL_ID</th>
<th>STORAGE</th>
<th>UNRESTRICTED_STK</th>
<th>TRANSFER_STK</th>
<th>CONSIGNED_STK</th>
<th>QUALITY_STK</th>
<th>BLOCK_STK</th>
<th>RESTRICT_USE_STK</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-Oct-15</td>
<td>ab02</td>
<td>***1060</td>
<td>PGEN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28-Oct-15</td>
<td>ab02</td>
<td>***1407</td>
<td>PGEN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30-Oct-15</td>
<td>ab02</td>
<td>***1176</td>
<td>BMES</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Data that is Smudged (not clear)

Issues:
• Missing values, inconsistent information, missing documentation, possible duplication, mixed text and numeric data
• Data is not clearly categorized - New Orders Vs. Renewal/Upgrade orders
Too Much Data

- Confusing – not sure which one to use among 5/6 price fields
- 2 Months of intense investigation with no definite answer
### Other Data Issues

**Data with no ownership / lineage:**
- No catalog
- No labels / description
- No ownership
- Not sure which is the source

**Data that is over-protected:**
- Need special access
- Restrictions in use / publishing
- Government / Legal Regulations

**Data with no tools to process:**
- Difficult to retrieve
- No Tools to process
- Sealed / Remote
Data Issues – Lessons Learned

- Need to understand the business model/process and dynamics of data before any transformation is attempted
- Data collectors may not understand what data is significant
- Data in warehouses are often created with “silo” objectives
- Important pieces as well as flags/characteristics are missed out, often beyond repair / reconstruction and hence limits its use

Some Silver Bullets?
- Metadata / Data Catalog
- Data Lake
- Cognitive search
Literally, “data about data”, data that defines and describes the characteristics of other data,

Metadata provides the description of the structure and content of data including the usage and relationships within different data elements.

Metadata includes

- where the data is originated (lineage),
- who owns it,
- where the data is stored,
- how it is stored / organized,
- how frequently it is changing,
- what it represents,
- how accurate it is,
- What are the business rules / policies,
- how to use,
- details data type, relationships, hierarchy, structure, etc.
Metadata / Catalog – the Inherent Power

- Helps to understand the data
- Identifies the key pieces of data for different usage scenarios
- Provides multiple views of the data without explicitly creating them
- Leads to easier maintenance of data by both data owners/users
- Cuts down the analysis project life cycle significantly
- Avoids several “silly” data analysis that may be wasteful
- Provides an early indication of relevance and accuracy of analysis
What is a Data Lake?

A Data Lake is…

• a generic **ecosystem of data repositories** that are

• **managed under a single information governance program** and

• together **offer a data distribution and self-service data access capability** for analytics and other big data use cases.

Source: Chessell, Mandy et al., "Designing and Operating a Data Reservoir," IBM Redbook, May 26, 2015, SG24-8274-00.
How Metadata fits in Data Lake Repository
What makes Data Lake Different from Warehouse/Marts

• **Availability of disparate data sources/types**
  • 360° view /analysis of business operations through cross correlation which is not possible with silo warehouses
  • optimize cross-domain activities

• **Availability of catalogs / curated data**
  • reduce the amount of time spent on cleaning/understanding the data (accounts for over 70% of typical analytics projects)

• **Common data access interface / governance**
  • Need user-friendly tools (some “fishing nets”, “gears”, “boats”, “divers”...) around the Data Lake such as Search functions, intelligent agents, Mobile apps, establishing data access interfaces with source data repositories, etc.
• Use Watson APIs or similar algorithms to understand the "types of data/terms" the user is looking for based on the following dynamic information:
  • Relationships to previously searched and accessed terms
  • Associated terms explicitly included as part of metadata
  • Frequency of terms used in search
  • Most recently searched terms
  • Time spent in reading details of an opened search item
  • Depth of search starting from original list
  • Prior searched history and results (including sequence of searched terms)
  • Profile of the user (data scientist from marketing vs. BI dashboard developer from Logistics)
  • Natural language analysis of Long descriptions (to understand the nature of data)
What Can We Do Further?

- Need to have 100x power and speed to manage the “onslaught” of data and some suggested approaches are
  - Annotate data
  - Socialize with data
  - Design automatic weed eaters – objective-based data filtering/prioritization

- Need new kind of magnets to pick up the needles as many of them are rusted, discolored, and too sharp to pick up
  - Automated cataloging / auto-generation of metadata (e.g., IBM AMG)
  - Cognitive Search
  - Metadata Analytics
  - Distributed analysis – spread the intelligence - need crowd analysis
"Every story ends with the brave prince marrying the beautiful princess destroying the evil villain. And, they lived happily ever after...... “

- Data scientists/users should be brave enough to understand the domain
  - not just the data types and data ranges alone....go beyond the schema, numbers, and chars.
  - understand the dynamics of data, explore how the data is created, collected, (manipulated), and maintained before touching it
- Don’t blindly trust the self-proclaimed domain experts
  - there are many “El Niño” effects on data. Try to become an “expert” yourself. Maybe you can teach few lessons back
- Are these data issues new? - not at all
  - maybe it is magnified because of the huge explosion of data

✓ Data cataloging is an excellent beginning – wrap the needle !!!
✓ Fortunately, better and more powerful tools are available for pre-processing the data
Thank You
Characteristics of Unstructured Data

Some Make it 4V’s

**Volume**
- Data at Rest
  - Terabytes to exabytes of existing data to process

**Velocity**
- Data in Motion
  - Streaming data, milliseconds to seconds to respond

**Variety**
- Data in Many Forms
  - Structured, unstructured, text, multimedia

**Veracity**
- Data in Doubt
  - Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations
Big Data / Dark Data Sources

>10 Billion networked sensors in industry supply chains

>3.2 Billion Internet users

>80% of world’s data is unstructured

>10 Billion
Mobile Devices World Wide
Most smartphones with GPS and camera

Twitter processes more than 20 terabytes of data every day - > 500 million tweets per day

Every day, the New York Stock Exchange captures info about 3.6 billion transactions

World Data Centre for Climate
- 220 Terabytes of Web data
- 9 Petabytes of additional data

Facebook processes > 40 terabytes of data every day
300 million photo’s uploaded per day

July 17, 2019
Institute of Industrial and Systems Engineers Webcast
Examples of Big Data Applications

**Financial Services**
- Fraud detection
- Risk management
- 360° View of the Customer

**Utilities**
- Weather impact analysis on power generation
- Transmission monitoring
- Smart grid management

**Transportation**
- Weather and traffic impact on logistics and fuel consumption

**Health & Life Sciences**
- Epidemic early warning system
- ICU monitoring
- Remote healthcare monitoring

**Telecommunications**
- CDR processing
- Churn prediction
- Geomapping
- Marketing
- Network monitoring

**Law Enforcement**
- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

**Retail**
- 360° View of the Customer
- Click-stream analysis
- Real-time promotions

**IT**
- Transition log analysis for multiple transactional systems
  - Cybersecurity
Example Tools for Data Cleansing

- IBM Infosphere Quality Stage / Information Analyzer
- Cloudingo
- Data Ladder
- Reifer
- TIBCO Clarity
- Winpure
- Drake
- IBM Cognos
- Talend Data Quality
Example Tools for Data Transportation

- IBM Data Movement Tool
- Centerprice Data Integrator
- Informatica Power Center
- Clover ETL
- Appache NiFi
- IBM Infosphere
- Talend Data Integration
- Aloomala
- StichData
- SnapLogic
Example Tools for Data Analysis / Staging

- IBM – Infosphere Information Server
- Informatica Data Staging Tool / Power Center
- HEVO
- Talend Open Studio
- CloverDX
- Pentaho Data Integration
- Panoply
- Amazon RedShift
- TeraData
- IBM Cognos
- IBM Watson Analytics
Analytics of Unstructured Data - Texts

• **What is the approach?**
  - Parsing and interpretation of sentences (Deep Parsing)
  - Learning the annotation from patterns (Machine Learning)
  - Domain specific rule based extraction (Annotation)

• **Current Technology / tools**
  - Watson Natural Language APIs
  - TensorFlow
  - System T – AQL
  - Watson Knowledge Studio

• **Use cases**
  - Question/Answering
  - Better interface to machines
  - Summarization
  - Translation
  - Rule generation / Information Extraction
Analytics of Unstructured Data - Audio

- **What is the approach?**
  - Cleansing
  - Annotation
  - Indexing

- **Current Technology /tools**
  - Sonic Annotator
  - Sonic Visualizer
  - Vamp Plugins
  - DBTune

- **Use cases**
  - Search by Audio Features
  - Question/Answering
  - Transformation (music)
  - Translation to Text
  - Translation between languages
Analytics of Unstructured Data - Image / Video

• **What is the approach?**
  - Quantitative Image Analysis
  - Object-based Image Analysis
  - Image Tagging
  - Image segmentation,
  - motion detection e.g. Single particle tracking,
  - video tracking,

• **Current Technology / tools**
  - Google’s Cloud Vision APIs
  - ImageJ
  - Aiforia Image Analysis Tool
  - MatLab Image Analysis Tools
  - IBM Watson APIs for Image Analysis

• **Use cases**
  - Face Recognition
  - Reconnaissance
  - Object Identification
  - Navigation
  - Machine vision, such as to automatically count items
  - Medical Image Analysis and diagnosis
  - Optical character recognition, such as automatic license plate detection