
# Challenges of Visualizing and Exploring Big Data

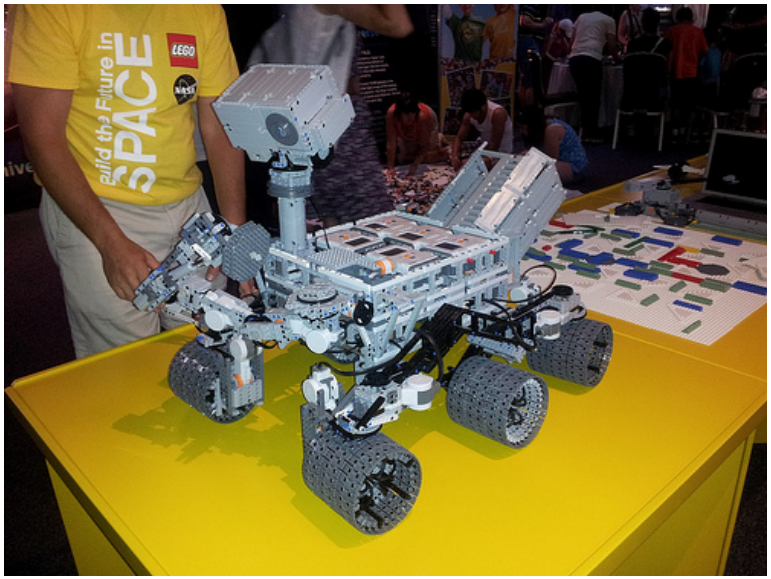
Will Gorman, Chief Architect, Pentaho

@wpgorman

April 23, 2013

# About Me

- Started Career at GE Research
- 6 years at Pentaho
- Hobbies include LEGO




# About Pentaho

Delivering the future of analytics: modern, unified data integration and business intelligence platform

- Full business analytics & data integration
- Native integration into big data ecosystem
- Embeddable, cloud-ready analytics

Open source development model enables **fast and broad innovation**

**Critical mass** achieved:

- Over 1,200 commercial customers
- Over 10,000 production deployments
- Over 185 countries

**One Download Every**
30 Seconds

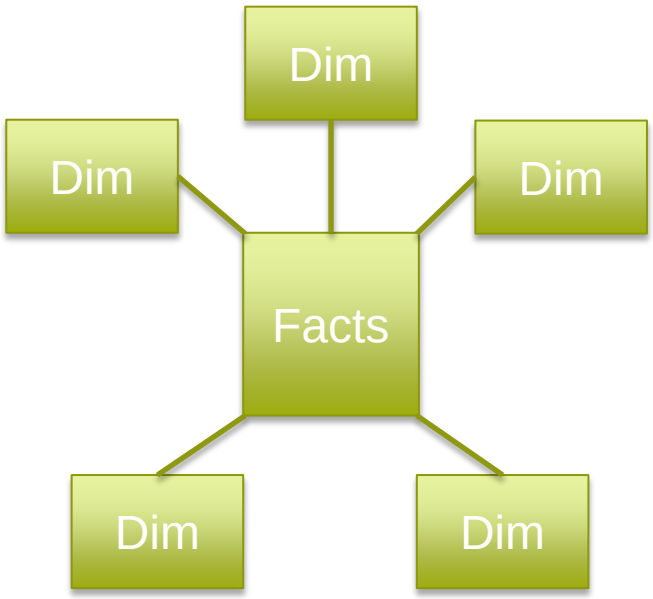


pentaho

# Why?

- Traditional Approaches to OLAP don’t cut it in Big Data
- There’s a need to analyze large amounts of data
- The data isn’t in a form designed for traditional analysis
- There’s an opportunity for software to solve the problem

# Traditional OLAP Stack




Visualization

- Crosstab, Bar, Line, Scatter
- Drilldown, Drill Through
- Filtering

Query Processing

- MDX, SQL
- MOLAP
- ROLAP
- MPP

Storage

- In-Memory
- Columnar
- Compression

# Challenges

- Big Data tends to be…
  - Nested
  - Schemaless
  - Unstructured
  - LARGE
- But we still want the types of experiences we had with regular data…
  - Near real time analytical querying

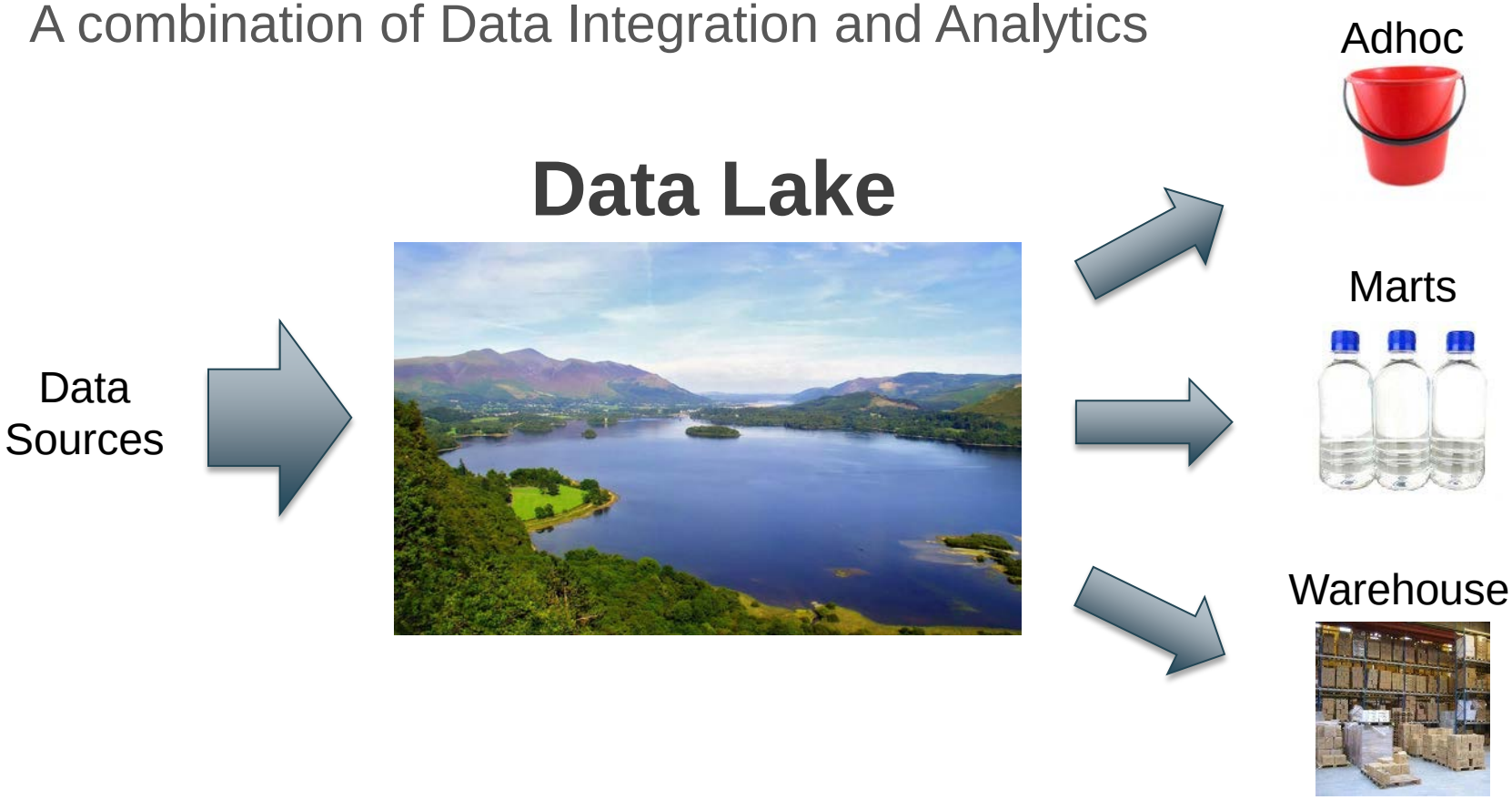
# MPP Databases

- Databases like Vertica and Greenplum give us a taste of Big Data
  - 10 to 100 times better performance than traditional systems
- Scale out architectures
  - Address volume and velocity, but not variety and value


# Approach with Current Technologies

Hybrid Architecture

- A combination of Data Integration and Analytics



# Pentaho Instaview

Instaview Components

Powered by Existing Technologies

**Template Definition**

**Configuration**
- Kettle Input Dialogs
- Parameterized Input

**Bulk Loading**
- Agile Mart Step

**Auto-Modeling**

**Visualization**

Spoon

Prompting Library

Transformation Engine

Agile BI Modeler

Mondrian - ROLAP

Analyzer

MonetDB



pentaho

# Limitations of Current Approach

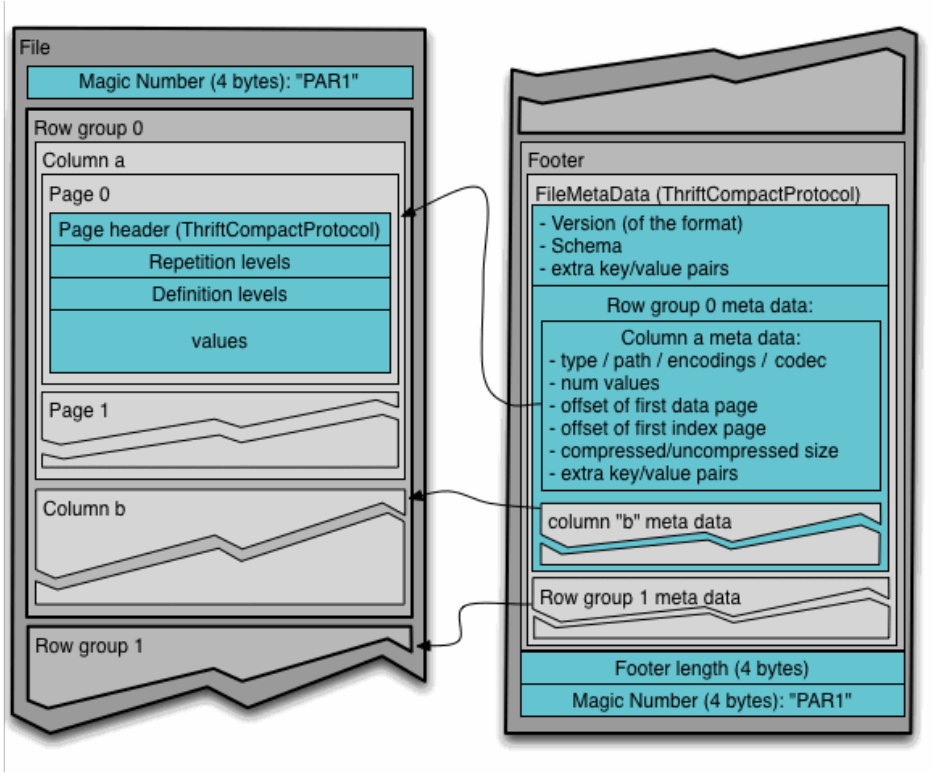
- Manually Intensive
  - Cleansing and Prep are done at a low level
- Analysis limited to traditional data sizes
- Expensive

# But folks are working to solve these problems

- Google Dremel / Big Query
  - “A letter from the future”
- Cloudera Impala
- Berkeley’s Spark and Shark
- Mapr’s Apache Drill
- MetaMarket’s Druid
- HortonWork’s Stinger Initiative

# With Some Interesting Algorithms and Datastructures

- Parquet – Dremel-based Columnar format for Hadoop
- Database Cracking – Continuous Physical Reorganization



# What about MDX?

- MDX is useful for representing business questions

```
select {[Customer].[All Customers].Children} on columns,
{[Measures].[Sales]} on rows
from [SalesFact] where {[Year].[2013]}
```

- Pentaho is investing in Mondrian, an Open Source ROLAP engine, to play nice with Big Data
  - Support for Impala and other big data query engines
  - Iterators over lists
  - Batch cell processing over recursion
  - Query planning for better native pushdown
  - Distributed Member Cache
  - Skunkworks: NOROLAP?

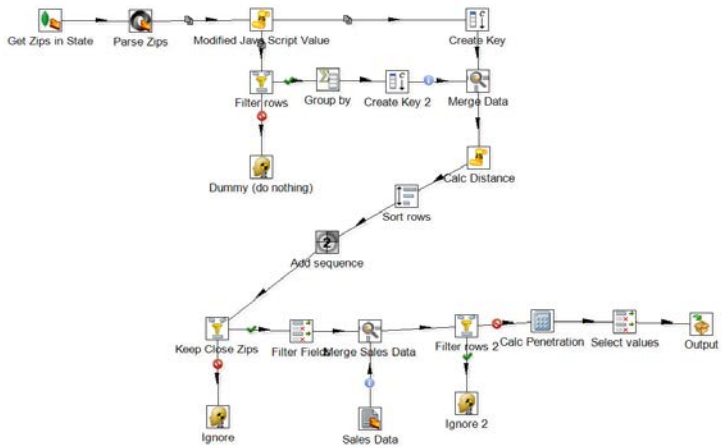
# But queries aren't everything

Once these systems can provide the OLAP performance that we've been wanting, there are still some issues…

- We still need to get data into these systems
- We still need to explore and visualize the data

# How do you Deal with a Petabyte of Dirty Data?

- Cleanse and augment data as it flows into the system
  - Streaming and Messaging Systems – Storm, Spark, Kafka
  - Descriptive and Predictive Systems – Mahout, WEKA
  - ETL – Kettle
  - Mashups – Bring data together
- Post process data via Map Reduce
- “Some Coding Required”

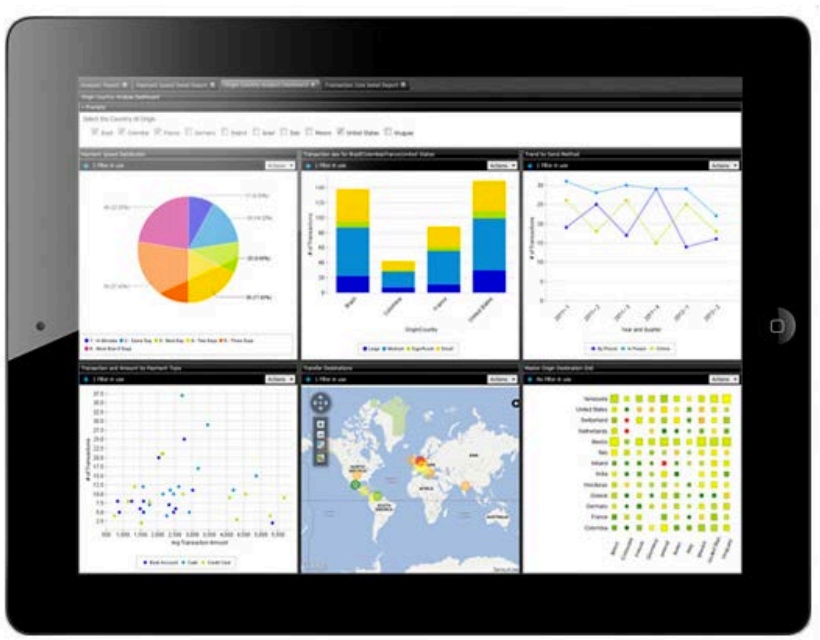





pentaho

# Visualizing Big Data

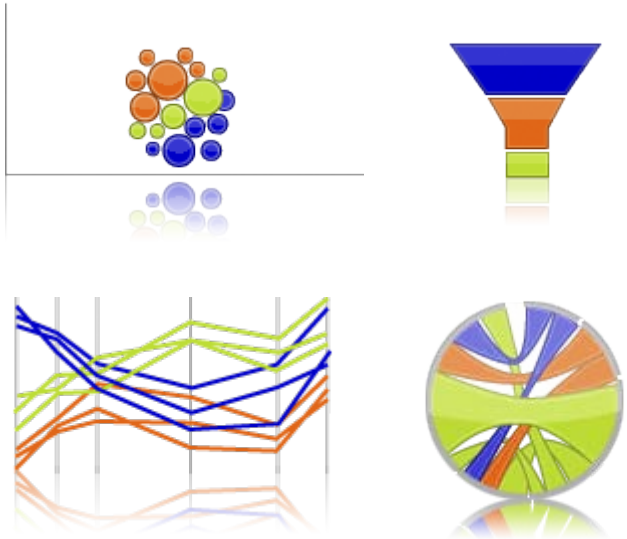
- Some things never change…
  - We still want to see high level aggregations like we always have



- But new visualizations and interactions will unlock insights into these unique types of datasets

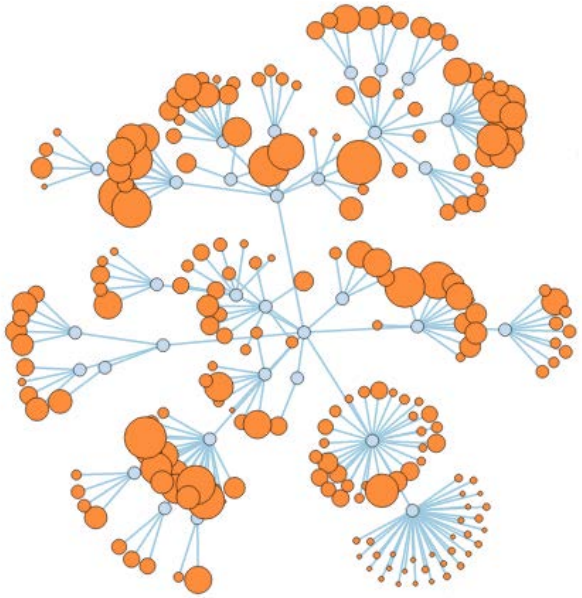
# We Need More Flexibility

- Visualization libraries like D3 are becoming more prevalent over traditional charting packages
- These tools work a layer below traditional charting, where developers work with SVG primitives
- “Some Coding Required”



# To Implement Unique Visualizations

- Graphs - Relationships
- Heatmaps
- Treemaps

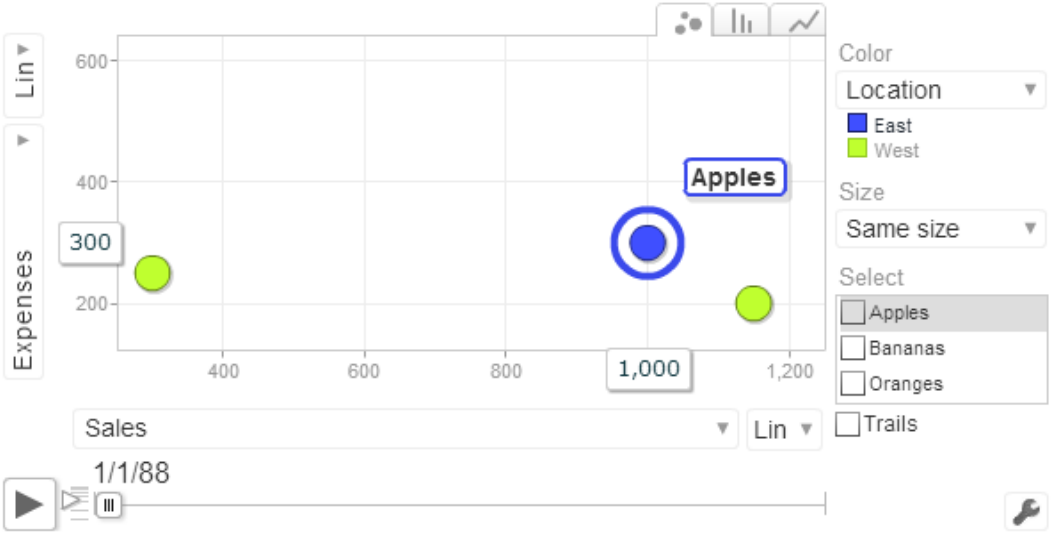


- Powered by WebDetails C-Tools
  - Community Chart Components
  - Community Graphics Generator
  - Based on Protovis

# And to Enable Big Data Interactions

Traditional Exploration

- Filter
- Pivot
- Drill Down




New Interactions

- Zoom
- Tiling / Paging
- Travel in Time – Google Motion Chart

# Join the Conversation…

irc.freenode.net ##pentaho

http://www.github.com/pentaho

http://forums.pentaho.com


# Thank You

Join the conversation. You can find us on:

 http://blog.pentaho.com

 @Pentaho

 Facebook.com/Pentaho

 Pentaho Business Analytics

# References


D3 - http://d3js.org/

Data Lake – http://blog.pentaho.com/2010/10/15/pentaho-hadoop-and-data-lakes/

Database Cracking – http://pdf.aminer.org/000/094/728/database_cracking.pdf

Dremel – http://research.google.com/pubs/pub36632.html

Drill – http://incubator.apache.org/drill/

Druid - https://github.com/metamx/druid/wiki

Google Big Query - https://developers.google.com/bigquery/

Google Motion Chart - https://developers.google.com/chart/interactive/docs/gallery/motionchart

Impala - http://www.cloudera.com/content/cloudera/en/products/cloudera-enterprise-core/cloudera-enterprise-RTQ.html

Parquet – http://parquet.io/

Spark - http://spark-project.org/

Stinger - http://hortonworks.com/blog/100x-faster-hive/

WebDetails C-Tools: http://www.webdetails.pt/ctools.html

Will's LEGO Models – http://www.battlebricks.com