Big Data Itself is Being Monetized

- Executives see the short path from data insights to revenue and profit
  - Big data often illuminates behavior
  - Analytic sandbox results taken directly to management
- Data is becoming an asset on the balance sheet
- Value big data by
  - cost to produce, including new technologies and storage
  - cost to replace if it is lost, perhaps impossible
  - revenue & profit opportunity, often immediate/ephemeral
  - revenue or profit loss if insights fall into competitors hands
Seismic Shifts in the Data Warehouse

- Low latency operational data mixed with history
  - New sources, new users
  - Mixed workloads, micro-batch or streaming loads

- Customer behavior data
  - Web traffic, need to analyze sessions
  - Significant integration: 50+ sources common

- Analysis of big data
  - Many new data formats and data types: social media; device sensors
  - Huge distributed data sets: full data set scans
  - Non-relational processing: complex branching & procedural processing

- Mission unchanged: Publish the Right Data

Extreme Integration: Tackling Data Silos at Cisco

Busting Silos with Enterprise Social Graph

Big Data Analytic Use Cases

- Behavior tracking
  - Search ranking
  - Ad tracking
  - Location and proximity tracking
  - Causal factor discovery
  - Social CRM
    - Share of voice, audience engagement, conversation reach, active advocates, advocate influence, advocacy impact, resolution rate, resolution time, satisfaction score, topic trends, sentiment ratio, and idea impact
  - Financial account fraud detection/intervention
  - System hacking detection/intervention
  - On line game gesture tracking

More Big Data Use Cases

- Non-numeric data and unique algorithms
  - Document similarity testing
  - Genomics analysis
  - Cohort group discovery
  - Satellite image comparison
  - CAT scan comparisons
  - Big science data collection

- Complex numeric data
  - Smart utility meters
  - Building sensors
  - In flight aircraft status

- Data bags – name/value pairs with ad hoc content
Houston: We Have a Problem

- The traditional pure relational data warehouse architecture can't handle ANY of these use cases.
- We need:
  - Non-scalar data: vectors, arrays, data bags, structured text, free text, images, waveforms
  - Iterative logic, complex branching, advanced statistics
  - Petabyte data sources loaded at gigabytes/second
  - Analysis in place across thousands of distributed processors, data often not in database format, full data scans often needed
  - Data loaded before structure is understood
  - Analysis while loading

Two Architectures to the Rescue: Zero Sum or Hybrid Coexistence?

- Extended relational
  - Extend the current formidable RDBMS legacy
  - Add features/functions to address big data analytics

- MapReduce/Hadoop
  - Build new architecture for big data analytics
  - Open source top level Apache project
  - Thousands of participants
Existing RDBMS Based Data Warehouse

Extended Relational Data Warehouse with Big Data Additions
MapReduce/Hadoop

Figure 2.3 from Tom White’s book, Hadoop, The Definitive Guide, 2nd Edition, (O’Reilly, 2010)

Hadoop Distributed File System (HDFS)

- HDFS is the distributed file system that supports a number of Hadoop projects:
  - Low level MapReduce programming in Java, Ruby, Python, C++
  - High level “MapReduce compilers” used for DW projects
    - Pig – “Pig Latin” procedural interface, leverages DBMS skills
    - Hive – SQL interface, described as a data warehouse
  - Hbase – open source, non-relational, column oriented database running directly on HDFS
Hybrid Possibilities

- Iterative Processing, Complex Logic, Non Scalar Sources in MapReduce/Hadoop (HDFS)
- Structured Querying in Relational DBMS EDW
- Standard BI Tools

Comparing the Two Architectures

- **Relational DBMSs**
  - Proprietary, mostly
  - Expensive
  - Data requires structuring
  - Great for speedy indexed lookups
  - Deep support for relational semantic
  - Indirect support for complex data structures
  - Indirect support for iteration, complex branching
  - Deep support for transaction processing

- **MapReduce/Hadoop**
  - Open source
  - Less expensive
  - Data does not require structuring
  - Great for massive full data scans
  - Indirect support for relational semantics, e.g. Hive
  - Deep support for complex data structures
  - Deep support for iteration, complex branching
  - Little or no support for transaction processing
Data Warehouse Disruptions

- The rise of the independent analyst
  - "newly discovered patterns have the most disruptive potential, and insights from them lead to the highest returns on investment."
- Sandboxes
- Demand for low latency
- Thirst for exquisite detail
- Light touch data waits for relevance to be exposed
- Simple analysis of all the data trumps sophisticated analysis of some of the data
- Declare the structures at query time, not load time

Whither the Data Warehouse?

- Corral and embrace the big data analysts
  - Analysts and IT must meet each other half way
  - Insist on using shared data warehouse resources
    - Conformed dimensions → integrate now, avoid data silos!
    - Virtualized data sets → quick prototypes, migrate to prod.
- Build a cross department analytic community with IT partnership
- Ditch the waterfall approach, go agile
  - Led by business – but sophistication required
  - Closely spaced deliverables, midcourse corrections
The Kimball Group Resource

- www.kimballgroup.com
- Best selling data warehouse books
  NEW BOOK! The Kimball Group Reader
- In depth data warehouse classes taught by primary authors
  - Dimensional modeling (Ralph/Margy)
  - Data warehouse lifecycle (Margy/Warren)
  - ETL architecture (Ralph/Bob)
- Dimensional design reviews and consulting by Kimball Group principals
- Informatica White Papers on Integration, Data Quality, and Big Data Analytics