Data Exploration and Analysis in the Age of Big Data:
Getting Results Faster Than You Thought Possible

Philip Russom

TDWI Research Director for Data Management

June 25, 2015

Sponsor

Speakers

Philip Russom
TDWI Research Director,
Data Management

Dustin Smith
Community Manager,
Tableau

Agenda

- Why analyze big data?
- A four-step analytic process
  - *For big data, exploration, discovery, and visualization*
- A technology stack for exploratory analytics with data in Hadoop
- Process and tool details
  - *Big data, as managed in Hadoop*
  - *Data exploration*
  - *Advanced analytics*
  - *Advanced data visualization*
- Real-world use cases
- Advantages and caveats
- Summary and conclusions

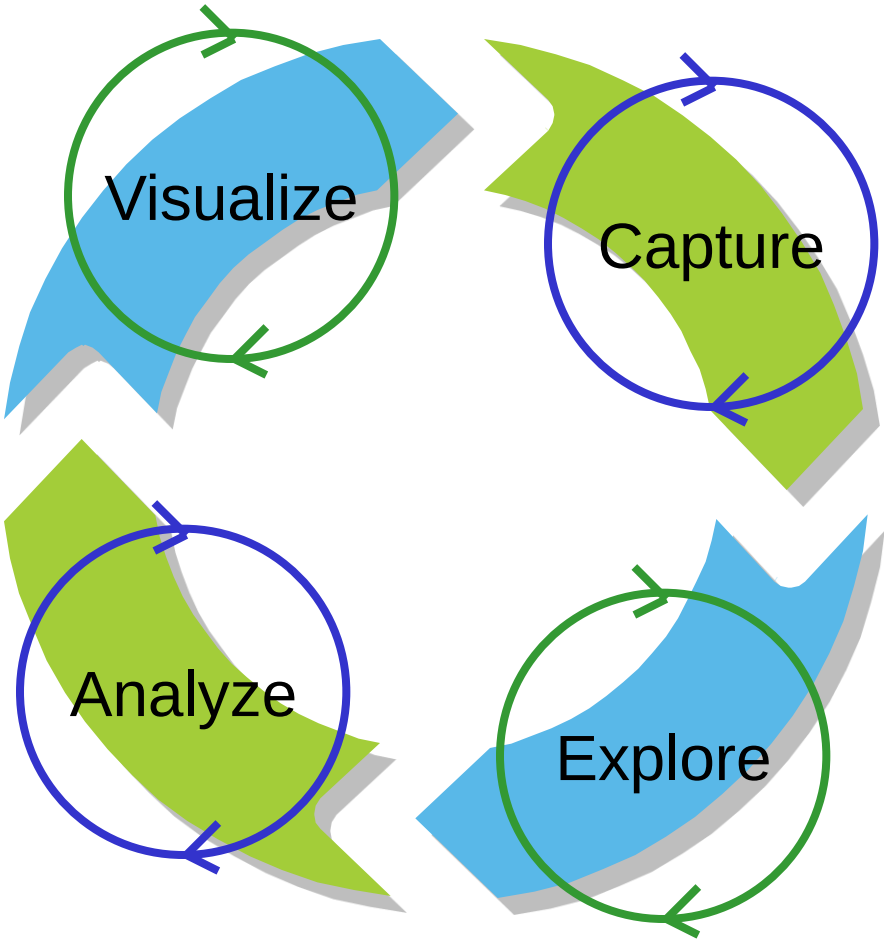
PLEASE TWEET --
@pRussom, @Tableau, #TDWI, #Hadoop, #Analytics, #BigData

Why analyze big data?

- Big data is a valuable resource
  - *Leverage it for business value*
  - *Never be content to merely manage big data as a cost center*
- Get value from big data by analyzing it
  - *Advanced forms of analysis are the main pathways to business value from big data*
  - *Self service makes analytics attainable*
- Most big data is also new data
  - *New sources – machines, sensors, vehicles, facilities, surveillance, devices, "The Internet of Things"…*
  - *Social media, Web apps, mobile apps…*
  - *New data from new sources leads to new insights via analytics*
- Big data provides bigger data samples
  - *Extend the life and value of older analytic applications for risk, fraud, and customer base segmentation*
- Big data increases breadth of older apps
  - *More attributes for complete customer view*
  - *More data points for customer sentiment*

ITERATIVE, FOUR-STEP PROCESS FOR

Exploratory Analytics with Big Data



SIMPLE TECHNOLOGY STACK FOR

Exploratory Analytics with Hadoop Data

FOUR STEPS

- Data Visualization
- Advanced Analytics
- Data Exploration
- Big Data Capture & Mgt

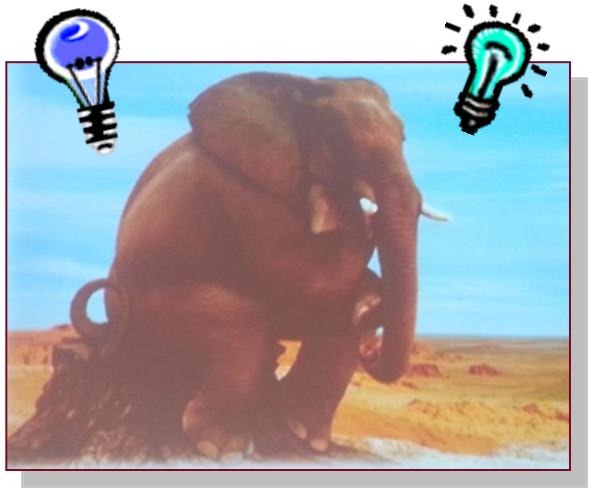
TECH STACK

Tool for
Exploratory Analytics
w/Data Visualization

w/Hadoop Distributed File System (HDFS), MapReduce, and Hadoop Data

CAPTURE BIG DATA

Why Hadoop now?



- Organizations want more business value from big data
  - *Analyzing big data yields value*
  - *Hadoop is built for big data analytics at massive scale*
  - *Also built for new data types, structures, and sources*
- Hadoop complements DW, DI, older Analytics
  - *Hadoop expands the biz value of these traditional platforms*
- Hadoop cracks the nuts that challenge traditional platforms
  - *Text, unstructured data, email archives, audio, video*
  - *Schema-free data, multi-structured data, NoSQL processing*
  - *File-based data: logs, dumps, XML, JSON, CSV, etc…*
  - *New data from: machines, sensors, social media, etc…*
  - *Algorithmic analytics: data mining, statistics, AI, NLP*

Importance of Data Exploration

- Data is the business.
  - *Data keeps a record of organizational activity and performance.*
  - *To know the business, you must know the data.*
- You have to start somewhere.
  - *Poking around in data gives you a sense of what happened*
  - *So you can start building a data set or model that represents a root cause, trend, or other analytic outcome.*
- Browsing data can be inspirational.
  - *This is how you discover new sources*
  - *Or determine which sources of data are appropriate to a specific report or analysis.*
- Exploring data is a prerequisite to analyzing data.
  - *By its natural, analysis makes correlations across data of diverse sources, structures, subjects, and vintages.*
  - *Finding just the right combination for successful analysis depends on data exploration as a first step.*

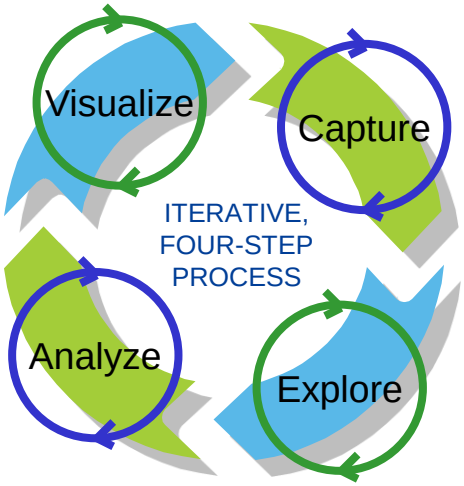
TECHNOLOGY REQUIREMENTS FOR

Data Exploration

- Search technology for exploring diverse data types.
  - *Data exploration should be as easy as Google*
  - *Parse data of many formats and structures*
  - *Allow any question; not confined to a predefined data model*
- Query capabilities in support of data exploration.
  - *Both business and technical users depend on query capabilities*
  - *Find just the right data; structure the result set for immediate use*
- High ease of use for user productivity.
  - *Some users are biz people who need to see data for themselves*
  - *They need a business friendly view*
  - *Ease of use accelerates technical developers' productivity, too*
- Support for all major data platforms, from relational to Hadoop.
  - *A modern data exploration tool needs to go where the data lives.*
- As you explore big data, you also:
  - *Extract data, model the result set, index big data*
  - *Perform these tasks as you go, not ahead of time, for greater agility*

A FEW REQUIREMENTS FOR

Advanced Analytics

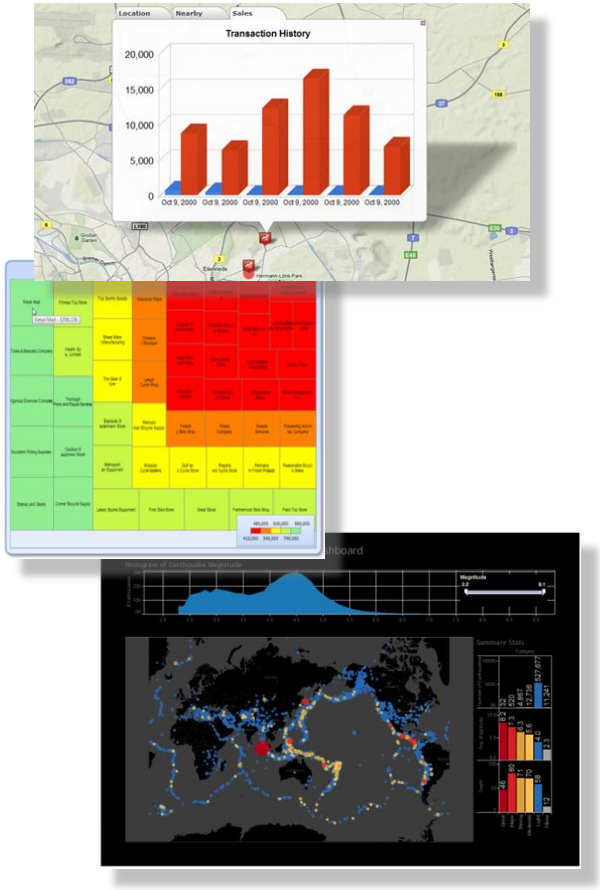


- Seamless integration
  - *In one tool environment, all functions for exploration, analysis, and visualization*
  - *The iterative, four-step process of exploratory analytics demands tight tool integration*
- Advanced forms of analytics
  - *Mining, predictive, statistics, NLP (not OLAP)*
  - *Algorithmic, as well as query based*
- Both canned and home-grown algorithms
  - *Tool should include library of pre-built algorithms*
  - *Tool should also help you write your own*
- High ease-of-use for broad collaboration
  - *Functions for both technical and business users*
  - *Both develop analytic apps and consume them*
  - *Assume that many user types will share their work*

THE IMPORTANCE OF

Data Visualization

- Critical to reaching your audience
  - *Data viz makes your analyses and datasets more consumable for more user types*
  - *Visual appeal makes big data analytics compelling*
- Viz's high ease-of-use empowers more user types
  - *Democratizes big data and analytics*
  - *Speed to insight for shorter decision cycles*
  - *The visualization _is_ the user interface*
  - *Enables visual discovery, customization*
- Seeing data relationships
  - *Critical for users to digest complex big data*
  - *Layering multiple sources, spotting patterns*
  - *Drag-and-drop reveals more relationships*
- Management Dashboards
  - *This is what most users want and need*
  - *Viz makes dashboards more mature, with more visual options and deeper data interaction*



Common Use Cases

For Exploratory Analytics with Hadoop Data

- Web site visitor behavior
- Price optimization in eCommerce
- Customer base segmentation
- Social media sentiment or pattern
- Medical research: DNA, outcomes…
- Quality assurance in manufacturing
- Fraud detection
- Risk calculations
- Facility monitoring & surveillance
- Capacity planning for grids, utilities, networks, facilities…
- Mobile asset management

SPECIAL USE CASE

Seize the many business opportunities of machine data.

- Machine data can be unique
  - *E.g., most robots are one of a kind, generating proprietary data forms*
- Some machine data is generated intermittently, not 24x7
  - *E.g., railcars are commonly fitted with sensors, but these are only read at rail yards or stations*
- GPS data is an important form of machine data
  - *Analyze where your customer makes certain purchases, which of your trucks is nearest the location where one is needed, what route products took from your plant to retail shelves, etc.*
- Machine data contributes to 360-degree views for a more complete and up-to-date picture
  - *Many new customer touch points generate machine data, like mobile apps, Web apps, social media*
  - *Correlate these with other data points for better views*

GETTING STARTED WITH

Exploratory Analytics with Hadoop Data

- Look for a problem to solve or an opportunity to leverage
  - *Finally get biz value from hoarded big data*
  - *Leverage big data from new customer touch points*
- Involve business people in defining applications for Hadoop
  - *Data stewards and data governors*
  - *Biz people affected by big data: CIO, CTO, marketing, sales, Web*
- Indentify how Hadoop data can integrate with other enterprise data
  - *More complete 360-degree views*
  - *Larger data samples for analytic apps for fraud, risk, segmentation*
- Consider a simple two-part technology stack; avoid “big bang”
  - *Integrated tool for data exploration, analytics, visualization*
  - *Hadoop as the data management platform for diverse big data*

ADVANTAGES OF

Exploratory Analytics with Hadoop Data

- Simple technology stack
  - *Just Hadoop Distributed File System (HDFS), MapReduce, and an analytic tool*
- Simple data preparation
  - *Capture raw source data in HDFS*
  - *Extract, model, & index data on-the-fly*
- Short time to use; fast development
  - *Due to simple technology and data preparation*
  - *Due to user-friendly analytic tool*
- Easy access to big data
  - *Productivity for technical developer*
  - *Visibility into business entities and process for end user*
- Leverage untapped big data for organizational advantage

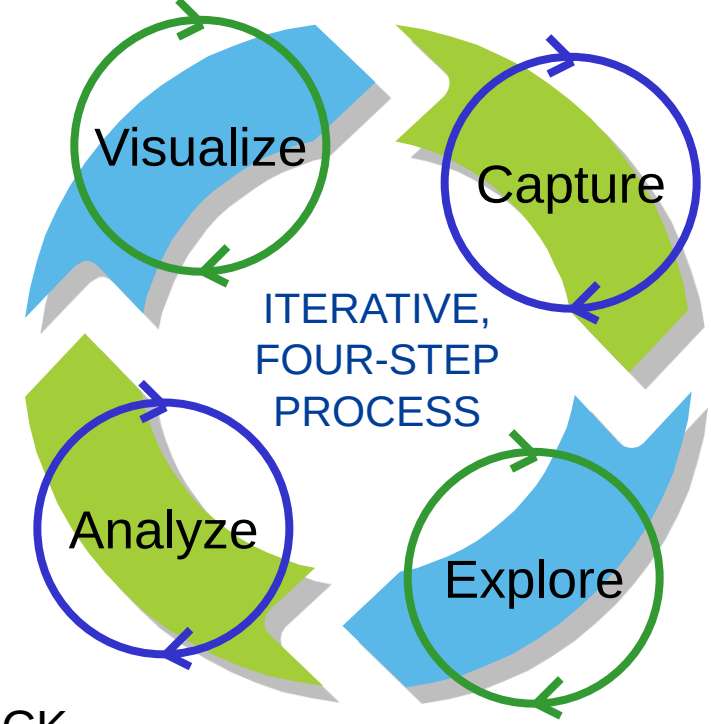
A FEW CAVEATS CONCERNING

Exploratory Analytics with Hadoop Data

- Assumes you have already deployed an HDFS cluster and populated it with big data
- This practice is mostly for ad hoc queries and algorithmic advanced analytics
  - *Rarely for scheduled reporting or real-time monitoring*
- Hadoop won't replace your data warehouse
  - *You still need your DW for standards reports, dashboards, OLAP, performance mgt, functions that require relational data, highly accurate or governed reports, etc…*
  - *Hadoop complements a DW by handling data that few DWs were designed to handle:*
    - Multi-structured data, unstructured data, file-based data, machine data, raw source data, massive data volumes, relatively low cost

IN CLOSING, LET'S REVIEW THE PROCESS STEPS and TECHNOLOGY STACK FOR

Exploratory Analytics with Big Data



FOUR STEPS enabled by a SIMPLE TECH STACK

- Data Visualization
- Advanced Analytics
- Data Exploration
- Big Data Capture & Mgt

Tool for
Exploratory Analytics
w/Data Visualization

hadoop
w/Hadoop Distributed File System (HDFS), MapReduce, and Hadoop Data

Questions?

Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

Dustin Smith, Tableau
dsmith@tableau.com