Big Data and Data Science: Behind the Buzz Words

Peggy Brinkmann, FCAS, MAAA
Actuary
Milliman, Inc.

April 1, 2014

Contents

- Big data: from hype to value
- Deconstructing data science
- Managing big data
- Analyzing big data

Big data: from hype to value

“Show me the money.”

- *Jerry Maguire* (1996)

The real issue

- Data that you can’t process and use quickly enough with the technology you have
- Possible reasons for this
  - Volume
  - Velocity
  - Variety (diverse/unstructured formats)
- Not a new problem, but new data sources are increasing the amount of challenging data

Sources of challenging data

- Transactions
- Web log files
- Mobile
- Voice, images, text, video from web and other sources
- Sensors
- Genomic

New data management solutions

- Need to handle larger volumes, unstructured formats, and/or real-time processing have driven new technologies
- Can lower costs, increase processing speeds for data that can’t be handled well with relational databases and/or single servers

Opportunities from big data

- Cost reduction
- Improve models/decisions with
  - new data
  - more data
  - faster cycle times
- New products and services

What about insurance?

- Product design
- Marketing
- Underwriting
- Pricing
- Sales management
- Claims
- IT

Develop a strategy

- What does your business need?
- What data do you have that is underutilized?
- What data are you missing that would be valuable?

Deconstructing data science

Mr. Maguire: “I just want to say one word to you, just one word.”

Ben: “Yes, sir.”

Mr. Maguire: “Are you listening?”

Ben: “Yes, I am.”

Mr. Maguire: “Plastics.”

- *The Graduate* (1967)

Some definitions of data scientist

- A data analyst in California
- A statistician under 35
- A developer of “data products”
- A practitioner of “data jujitsu”

Something new, or re-branding?

C. F. Jeff Wu (1998):

- Data collection
- Modeling and analysis
- Problem solving and decision making

William S. Cleveland (2001):

- Multidisciplinary investigation
- Models and methods
- Computing with data
- Tool evaluation

Some more recent attempts

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it

Combine the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data

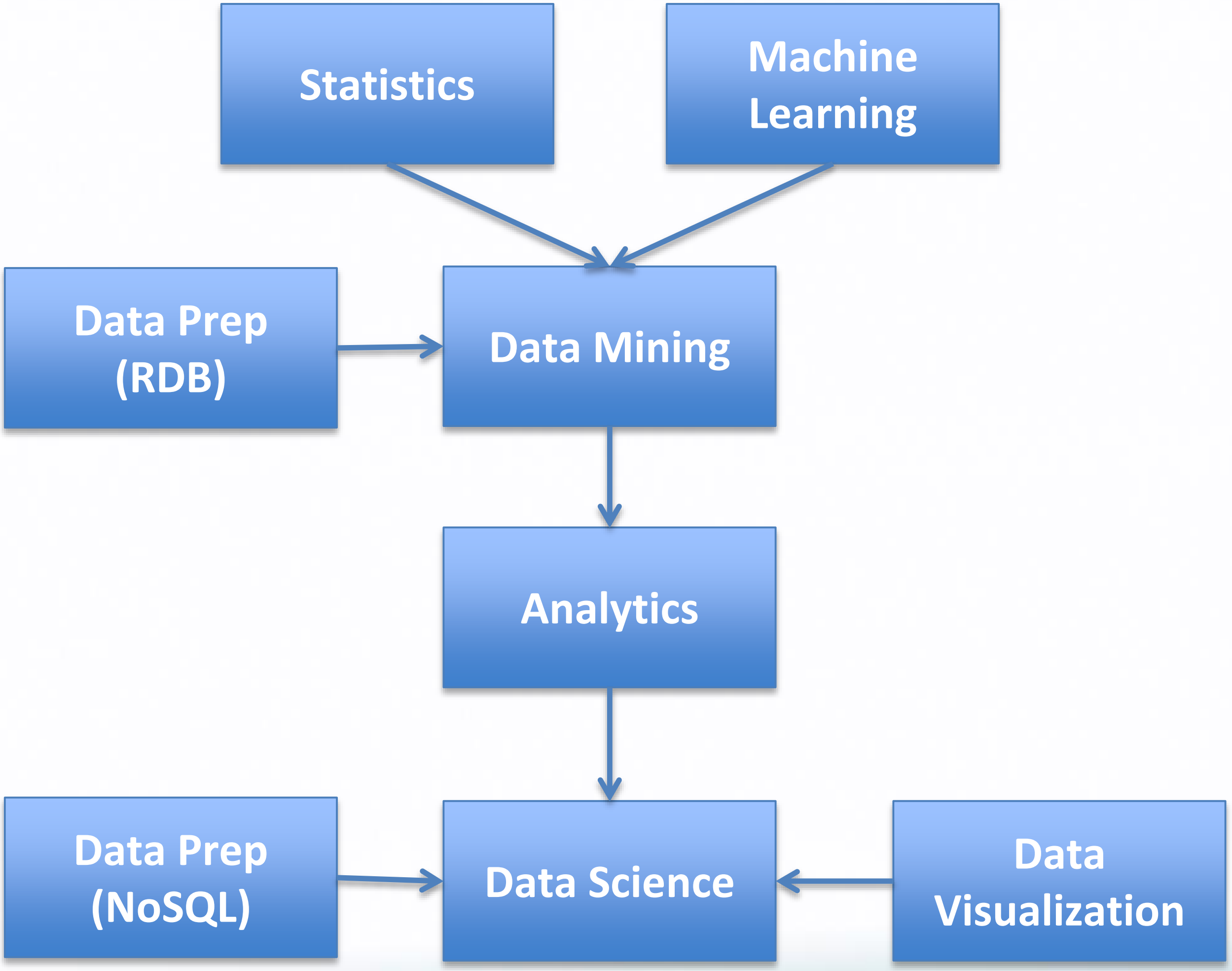
start by looking at what the data can tell them, and then picking interesting threads to follow, rather than the traditional scientist’s approach of choosing the problem first and then finding data to shed light on it

Extract information from large datasets and then present something of use to non-data experts

What seems different

- Using large datasets
- Hands-on, heavy data prep of unstructured data
- Coding with general purpose languages (Python, C++, Java)
- Starting with the data, not a question?
- Emphasis on storytelling/visualization

Family Tree


Statistics
Machine Learning
Data Prep (RDB)
Data Mining
Analytics
Data Prep (NoSQL)
Data Science
Data Visualization

Managing big data

“You’re gonna need a bigger boat.”

- *Jaws* (1975)

Managing big data

- Distribute data storage, data processing across multiple computers
- Can use cheaper, commodity hardware because data is duplicated on multiple machines – can be recovered when one fails
- Faster run times - use the parallel computing power of the machines where the data is stored, and avoid I/O of extracting data first

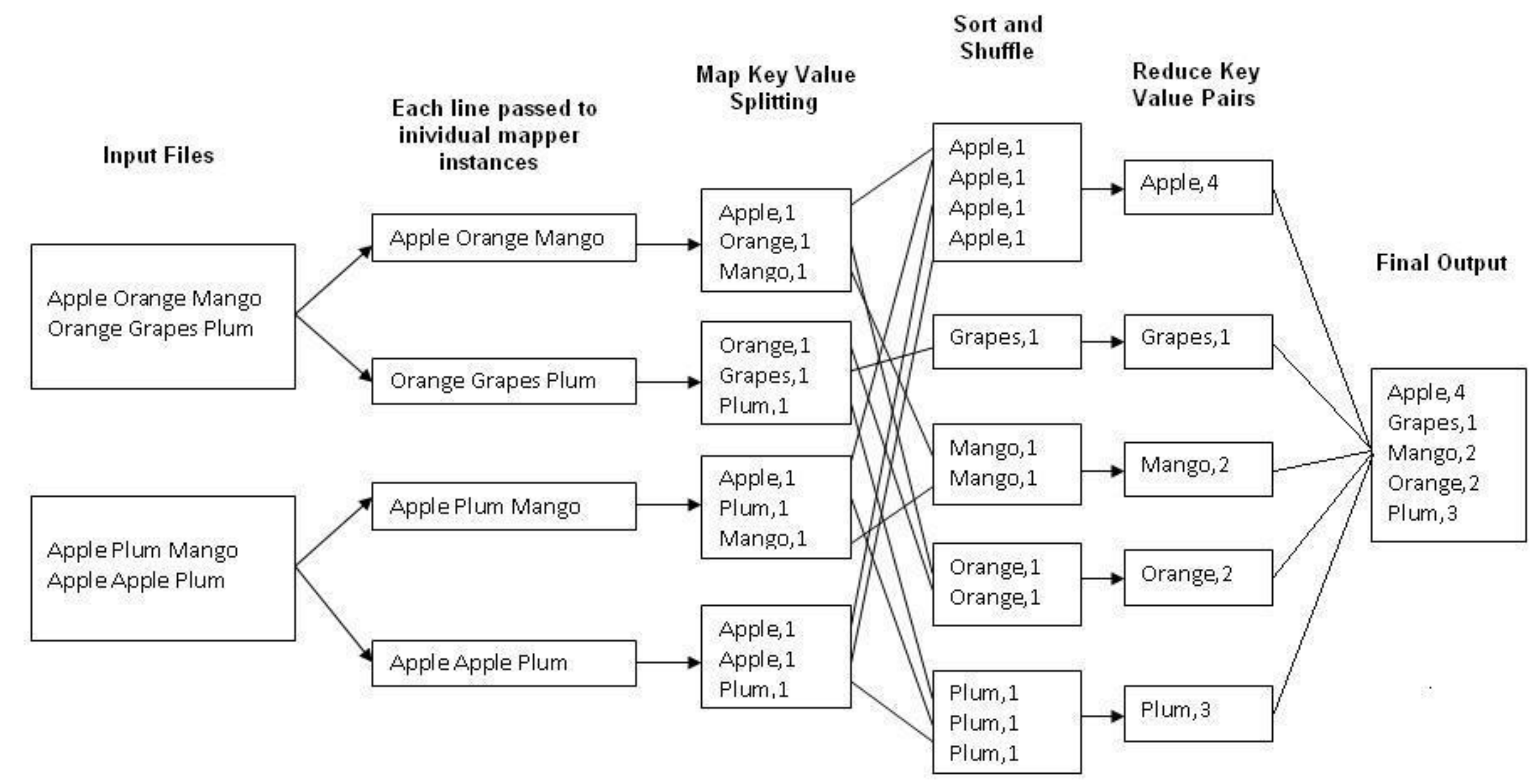
Let’s talk about the elephant in the room, Hadoop

- Software framework for storing and processing structured and unstructured data
- Distributes (and replicates) your data across multiple commodity machines (a “cluster”)
- File system (HDFS) keeps track of where the data is
- Programming framework (MapReduce) to process the data

Many Hadoop vendors

- Apache
- Cloudera
- Hortonworks
- IBM
- MapR (although technically a different file system)
- Microsoft
- Pivotal

What is MapReduce?



Source: http://kickstarthadoop.blogspot.com

Other Hadoop tools

- Hive – SQL-like query language
- Pig Latin – scripting language for creating MapReduce programs
- HBase – column-oriented database within Hadoop
- Mahout – Java machine learning library
- Sqoop – moves data between Hadoop and relational databases

“Not Only Hadoop”

Family	Category	Examples	Pros	Cons
Relational	Massively Parallel Processing (MPP)	Teradata, Netezza, Greenplum, Vertica, Oracle Exadata	Fast and familiar	Expensive Poor for unstructured data
“Not Only SQL”	Key-Value	Redis, Riak, Voldemort	Simple, fast I/O	Poor for complex data
	Column	Hbase, Hypertable, Cassandra	Good for unstructured data	Poor for interconnected data
	Document	CouchDB, MongoDB	Good for unstructured data	Poor for interconnected data
	Graph	Neo4j, InfiniteGraph	Certain types of problems	Not really scalable

Analyzing big data

“I feel the need – the need for speed!”

- *Top Gun* (1986)

First, it isn’t always as big as it seems

- Use big data tools to summarize it down, then apply the usual analysis software
- Do you really need every observation? Then sample it down

Intermediate steps

- Use software/algorithms that process outside of memory (bigGLM, Revolution R)
- Get more memory – a new machine, a big memory instance on a cloud

If you go for it . . .

Need analysis software that has been written to work in parallel

Product	Algorithms supported for distributed processing
SAS on Hadoop	C&RT, Time series, GLM, Logistic regression, Random Forest, Clustering
Revolution R Enterprise	Regression, Logistic regression, GLM, Clustering, Decision Trees, Random Forest
IBM SPSS Analytic Server	Linear regression, Neural Net, C&RT, CHAID
Mahout	Collaborative filtering, Naïve Bayes, Random Forest, Clustering, Principal Components
MapReduce	Write your own MapReduce directly or with an interface like RHadoop

THANK YOU

peggy.brinkmann@milliman.com