Data Lakes and Data Mining

Information Governance in the Big Data Universe

John C. Montaña, J.D, FAI, FIIM

MONTAÑA
& ASSOCIATES

The 21st Century – A Data Intensive Era

- The daily total?
  - 2.5 quintillion bytes
  - Each and every day
- Data Collection is:
  - Ongoing
  - Pervasive
  - Growing
- Virtually every electronic tool we use collects data
- Every business collects vast amounts of data about everything it does

MONTAÑA & ASSOCIATES

How Much is That?

- A terabyte is:
  - A trillion bytes
  - 1,000,000 word documents
  - 1900 hours of video
  - 200 DVD movies
- A petabyte is 1,000 terabytes
  - A quadrillion bytes
- 2.5 Quintillion bytes = 2,500 petabytes
- That’s a lot of stuff
  - In thousands of formats, billions of sources

MONTAÑA & ASSOCIATES

All That’s Great, But What do We do With It?

- We spend a great deal of time and resources collecting it
- We have some known uses
- We’d like to find additional value, but:
- Even though we own only a little slice of the overall pile, using it is still finding a needle in a haystack
  - Too many sources
  - Too many formats
  - Too much stuff to look through

MONTAÑA
& ASSOCIATES

The Traditional Approach – The Data Warehouse

- Purpose-built repository
  - Pre-defined schema
  - Inputs from pre-defined sources
  - Outputs answer pre-defined question
- Upsides
  - Much better than nothing
  - Great for the questions it was built to answer
- Downsides
  - Cost
  - Effort
  - Limits to utility
  - Limited scalability

MONTAÑA & ASSOCIATES

Now – Data Lakes

- The goal: to make all of your data available and useful – and therefore valuable
  - Platform links multiple sources and repositories
  - Unlimited data types
  - Unlimited data volumes
  - No pre-defined questions or needs
- The upside
  - It’s huge – vast data volumes
- The downside
  - It’s huge – vast data volumes

MONTAÑA
& ASSOCIATES

The Concept

- Raw data is brought in – in vast quantities
  - Basic QC check
  - No attempt to sort, filter, organize
- Data is cleansed and validated
  - PII, PCI, etc.
- Sensitive data is protected
- Organization and structure imposed after fact
  - Users rate and rank data
- Data is made available for analysis
- Useful discoveries are monetized

MONTAÑA
& ASSOCIATES

The Tool – Apache Hadoop

- A software framework that enables:
  - Large scale utilization of commodity hardware for:
    - Distributed storage
    - Distributed, parallel processing
- The Result
  - Very large, essentially unified data sets
  - Massive collective processing power
  - The ability to do analysis/answer questions that cannot be answered otherwise

MONTAÑA
& ASSOCIATES

How is That Different from a Warehouse?

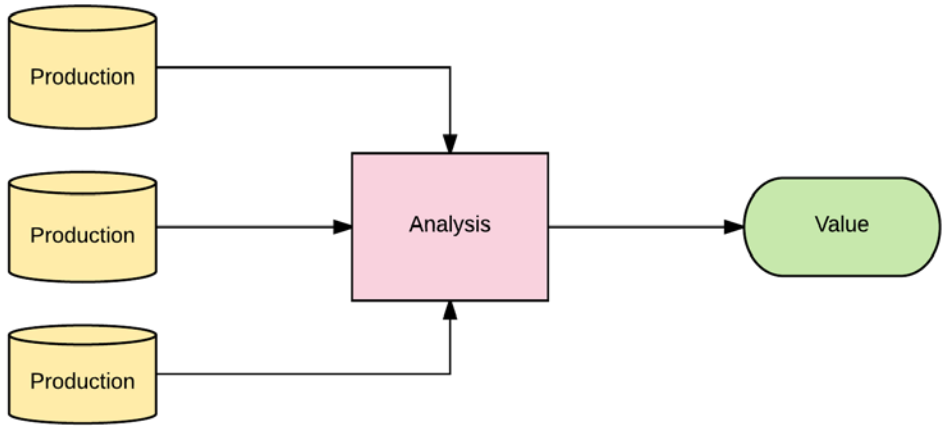
- No pre-defined data schema
- No pre-defined set of questions to answer
- Much denser, hence cheaper, storage
- Potentially available to many more users
- Potentially much more powerful analytics

IF it all works as planned

MONTAÑA
& ASSOCIATES

The Goal of a Data Lake

- Multiple data inputs from multiple business systems in any business area
- New correlations and additional value from comparing data from disparate sources
- Maximum researcher freedom to explore
- Ultimately, new value from new information and knowledge



MONTAÑA & ASSOCIATES

Keys to Making it Work

- Organization – data does not organize itself
- Management – data does not manage itself
- Retention – most data is not indefinitely useful

MONTAÑA
& ASSOCIATES

Organization

- Hadoop is not the next easy button
  - It does not auto-classify or auto-organize your data
  - It does not automatically do analytics
- Garbage in – garbage out still applies, maybe more so:
  - You're bringing in pretty much everything, and not even bothering to organize it up front
  - The ingestion process must be organized and managed
- You must ultimately impose some organization and structure
  - Researchers use various tools to query, analyze and organize data for their individual use cases

Management

- Compliance - still an issue
  - Privacy - PII, PCI, general data privacy
  - Data localization
  - Regulatory reporting
- Security, access control
  - Not every researcher is entitled to see all data
  - This data is subject to breach like all other data

- Value
  - Value consists of:
    - Development of analytical tools
    - Valuable/useful results
    - Data supporting the above
  - Value only accrues when these are:
    - Captured
    - Protected
    - Pushed to business units for monetization

MONTAÑA
& ASSOCIATES

Categories of Data in a Data Lake

- Data falls into three groups
  - Input data
  - Analytical tools
  - Output data and conclusions
- Each data type should be subject to management rules that controls costs and increase utility and value

MONTAÑA
& ASSOCIATES

Input Data

- Duplicate or duplicate excerpt from production environment data
- Used as the basis for further analysis
  - Alone
  - In aggregate with data from different sources
- May or may not prove useful or interesting

MONTAÑA
& ASSOCIATES

Tools

- Algorithms, programs, queries, scripts, etc. developed and used to analyze input data – may be valuable as:
  - Intellectual property, regardless of the value of any one use of them
  - In combination with input and output data, evidence of a new insight or correlation of value
- May also have no ongoing value at all

MONTAÑA & ASSOCIATES

Output Data

- The results of analysis of input data by tools
- May be valuable for:
  - Monetization
  - Regulatory compliance
  - Business process improvement
- May also have no ongoing value at all

MONTAÑA & ASSOCIATES

Challenges

- Ever-increasing data volumes create challenges:
  - Increasing storage costs
  - Decreasing value of data as it becomes comingled with later data
  - Potential regulatory and litigation issues

MONTAÑA
& ASSOCIATES

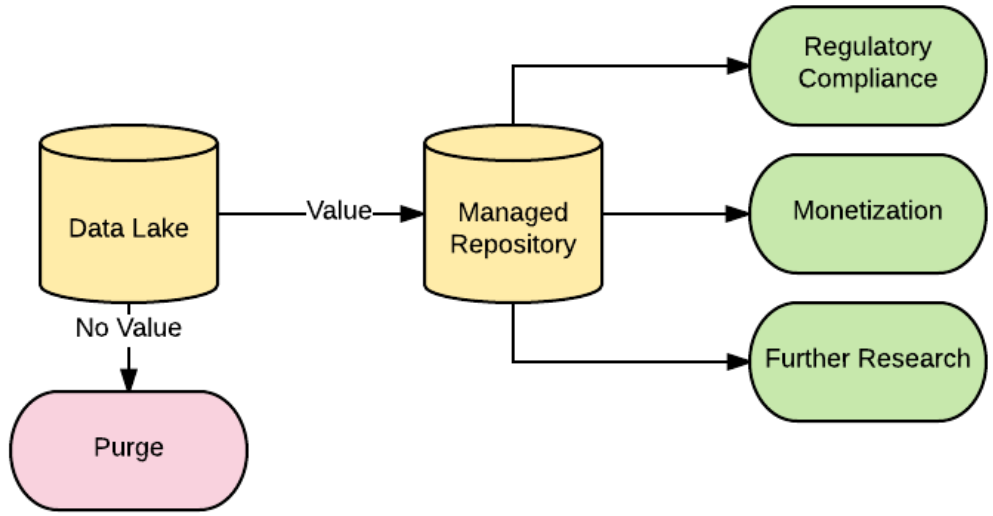
Legal Considerations

- Data in data lakes is potentially discoverable
  - The sheer volume of information makes this problematic and potentially very costly
- Discoveries may have regulatory significance
  - These will become records and must be managed in accordance with policies and procedures

MONTAÑA
& ASSOCIATES

How to Manage the Issues?

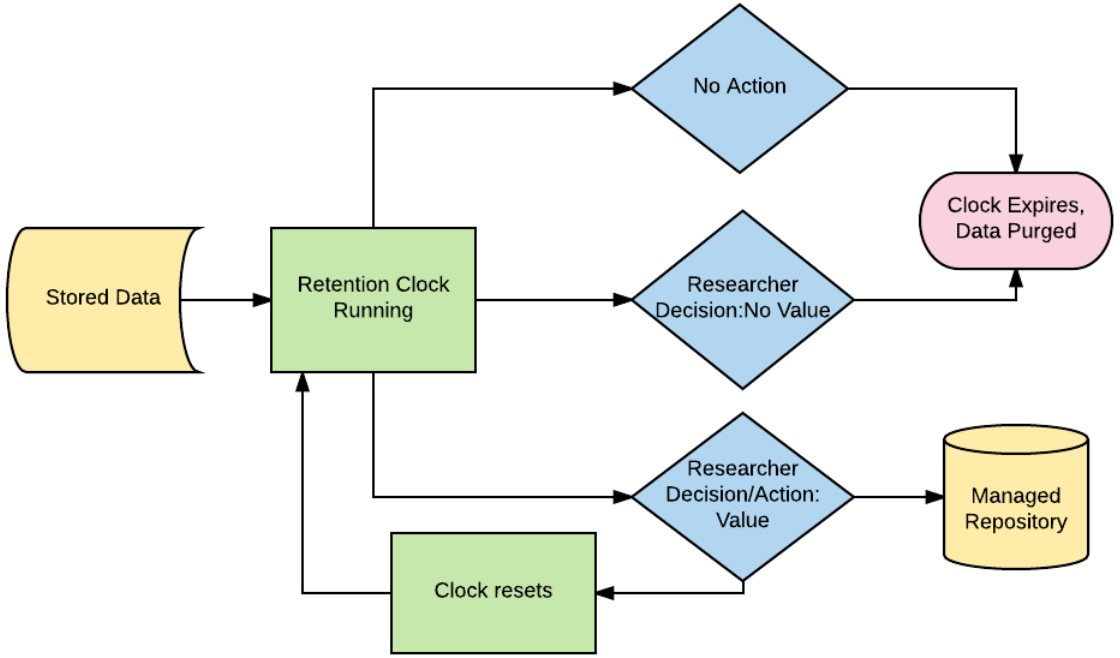
- Determine continuing value of a data set
- Data, tools, scripts, etc. that have ongoing value are:
  - Segregated
  - Formally managed – policies and procedures, retention schedule, etc.
  - Made available to other researchers, business process owners, etc.
- The trick is determining ongoing value



MONTAÑA & ASSOCIATES

Determining Ongoing Value

- Input data
  - Lack of use after X period of time = no ongoing value (note: thereafter, data can always be reimported from production environment); or
  - Affirmative researcher decision of no value
- Tools and output data
  - Researcher decision – “this is interesting/valuable”; else
  - No decision or affirmative “not interesting” decision = no ongoing value



MONTAÑA & ASSOCIATES

Value Management

- Discoveries that cannot be later found cannot be protected/used/monetized
- The value of the data lake is not in the discoveries alone, but in the subsequent availability and utilization of the discoveries
- Effective management of discoveries is the key to this

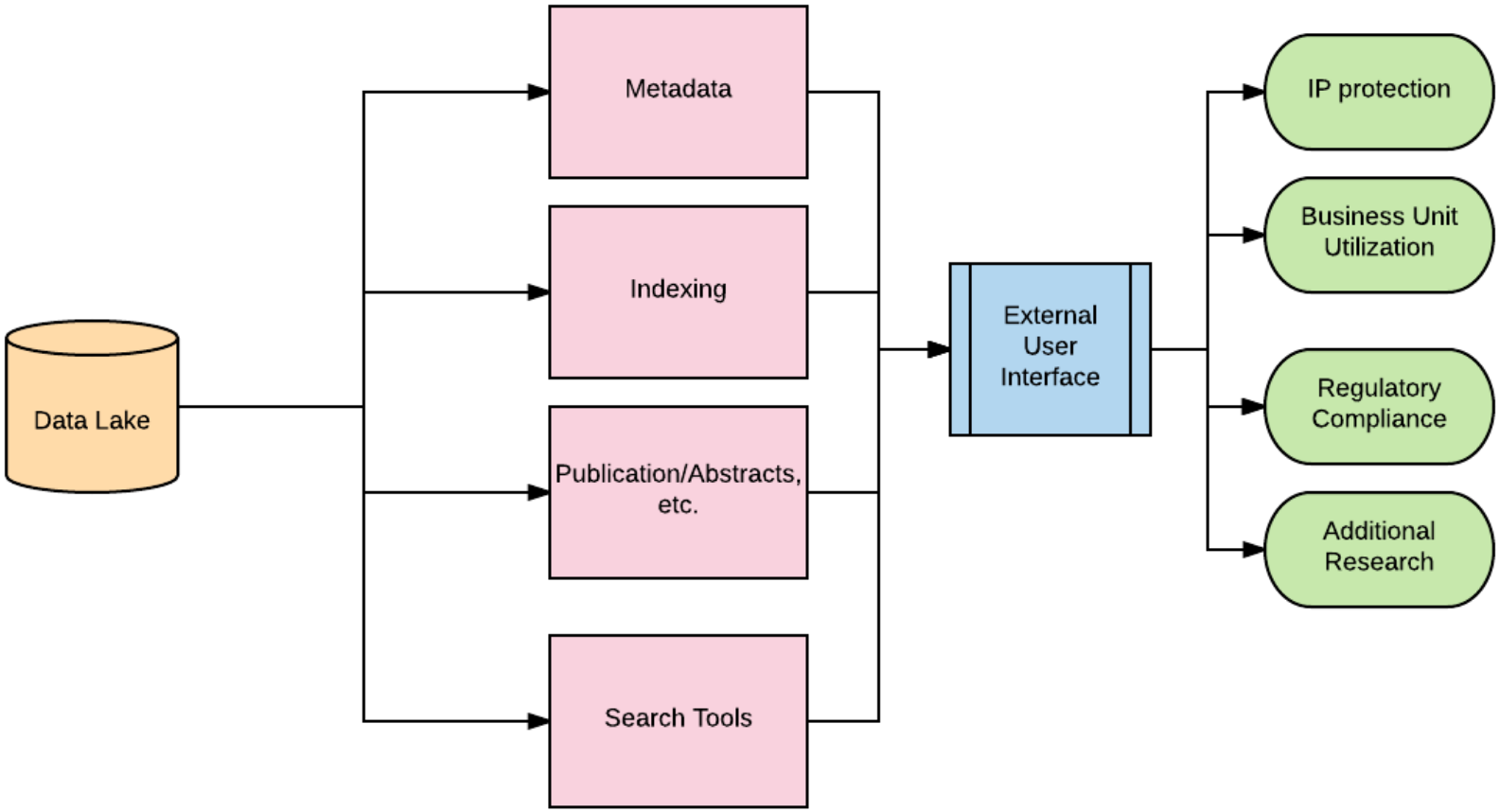
MONTAÑA
& ASSOCIATES

Downstream Utilization

- Once useful data is placed in managed repositories, it is available for:
  - Further research
  - IP protection
  - Regulatory compliance and reporting
  - Process improvement
  - Monetization

MONTAÑA & ASSOCIATES

The Value Stream



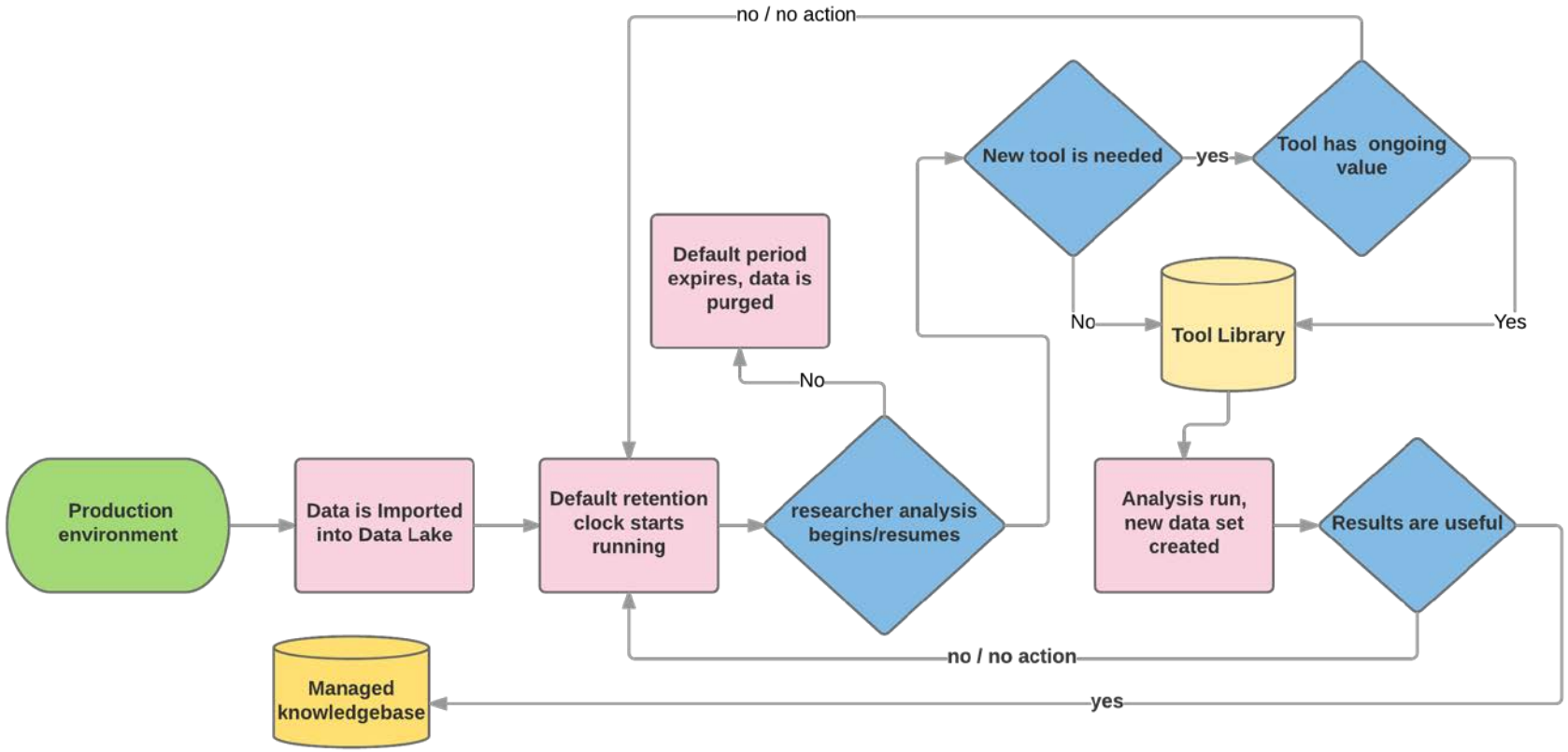
MONTAÑA & ASSOCIATES

Process

- Input data arrives
  - Retention clock commences running
  - If not used by x, data is purged; if data is used, clock resets and begins running again
- Analysis
  - Upon creation, retention clock commences running
  - If researcher determines ongoing value/interest, entire data set(input, tool, output) duplicated in separate repository for management
  - If no affirmative researcher input, clock starts and data is purged after x
- Useful tools can always be archived, regardless of interest/value of research results

MONTAÑA & ASSOCIATES

The Data Lake Management Process



MONTAÑA & ASSOCIATES

Retention - Most Data is not Indefinitely Valuable

- Space
  - Not unlimited
  - Not free
- Computing power
  - Not unlimited
    - more data to crunch = more processors spending more time to achieve the same result
    - You may have a lot of processors, but they are finite in number, finite in speed
    - Time spent crunching valueless data at best wastes CPU cycles, at worst skews results
- And remember, we are talking very large data sets here

So, you must impose some sort of retention rules on the data

MONTAÑA
& ASSOCIATES

What Does it All Add up To?

- Big Data =
  - Big Opportunities
  - Big Money spent
  - Big Value in return
  - Big Headaches if you mess it up

- From a management perspective, Big Data is no different than other data
  - Tools can’t self-organize it
  - Out-of-control growth balloons costs, reduces its value

- Management tools are getting better, but:
  - Concepts are still the same:
    - Organization
    - Management
    - Retention/Disposition

MONTAÑA
& ASSOCIATES

Bottom Line

- Plan your Big Data environment carefully
  - You don’t know what’s coming in and you don’t know what’s going out, but you still need to plan
  - Management processes must be robust and flexible enough to accommodate any input and any output
- Active central management is required
  - Researchers will not take the time to sweep the floor when they are done
  - Many activities should take place by automated rule implementation – e.g., stale data purge
  - Researchers must be provided with simple robust tools for monetization
  - Input staff and researchers must be aware of any compliance obligations

MONTAÑA
& ASSOCIATES