Applying Process Mining to IT Big Data

Richard F. Eng

PRINCE2, PMP, CSQE, CRE, CQE

14 March 2014

The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

MITRE

Agenda

- Biography
- Big Data Definitions
- Process Mining
- Quantitative Analysis Stages and Steps
- Applying Process Mining
- Lessons Learned

Biography

- **Principal Software Systems Engineer at the MITRE Corporation**
- **Over 20 years of industry experience in telecommunications and software systems**
- **Areas of interests**
  - Solving systems level problems
  - Applying quantitative methods to improve business , IT processes, and software quality
- **Education:**
  - M.S. in Data Analytics, University of Maryland (Expected 2015)
  - MBA, Georgetown University
  - M.S. in Bioengineering, Brooklyn Polytechnic Institute
  - B.S. in Chemistry, Brooklyn Polytechnic Institute

Big Data Definitions

Big data is data which "exceed(s) the capacity or capability of current or conventional methods and systems." In other words, the notion of "big" is relative to the current standard of computation.

The National Institute of Standards and Technology

"… the increasing size of data, the increasing rate at which it is produced and the increasing range of formats and representations employed. This report predated the term "big data" but proposed a three-fold definition encompassing the "three Vs": Volume, Velocity and Variety. This idea has since become popular and sometimes includes a fourth V: Veracity, to cover questions of trust and uncertainty."

Gartner. In 2001, a Meta (now Gartner) report

Big Data Definitions (continued)

"Big data is the term increasingly used to describe the process of applying serious computing power—the latest in machine learning and artificial intelligence — to seriously massive and often highly complex sets of information."

Microsoft

"Big data opportunities emerge in organizations generating a median of 300 terabytes of data a week. The most common forms of data analyzed in this way are business transactions stored in relational databases, followed by documents, e-mail, sensor data, blogs, and social media.

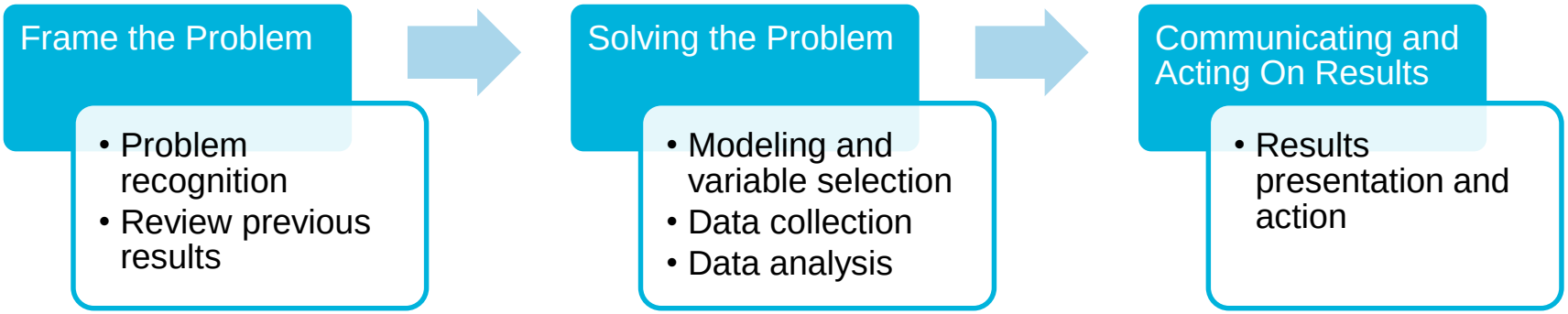
Intel

The Method for an Integrated Knowledge Environment open-source project. The MIKE project argues that big data is not a function of the size of a data set but its complexity. Consequently, it is the high degree of permutations and interactions within a data set that defines big data.

Process Mining

- **Data analytics technique**
- **System agnostic**
- **Examines large amounts of process data**
- **Provides the ability to**
  - Evaluate and understand actual process flows
  - Compare actual against expected process flows from a
    - Policy, procedures, and/or Information Technology perspective
- **Impact of visualizing actual process flows**
  - Detection of process patterns
  - Identification of process inefficiencies
  - Identification of anomalous behavior

Quantitative Analysis 3 Stages and 6 Steps



Source: Keeping up with the Quants: Your Guide to Understanding and Using Analytics – Davenport, T. and Kim, J. (2013).

MITRE

Frame the Problem

- **Problem recognition**
  - Figure out where the problems are with the **SYSTEM** & fix them fast
  - E-Commerce site experiencing significant operational availability issues
    - Site performance degrades and crashes as more customers access services
    - No current documentation (e.g., Systems requirements, design specifications, applications software, test procedures, test scripts, etc.)
      - We were Agile. We didn't have time to …
- **Review of previous results**
  - Ad hoc problem solving huddles
    - It's not my system. It must be your …
    - Try this, it should/might fix it …
  - No tracing of business processes to software applications to IT infrastructure
  - Manual review of system logs

"Furious activity is no substitute for understanding."

- H. H. Williams

Solving the Problem

- **Modeling and variable selection**
  - Apply process mining to model end user transactions through the E-commerce site
  - Variables needed
    - Unique Transaction ID, Start and Stop Time Stamp, and Activity
- **Data collection**
  - Gain access to system log files
    - Normal operations
    - During outages and low operational availability
- **Data analysis**
  - Run process model
  - Review results
  - Provide feedback to stakeholders that own the processes, application software, and IT infrastructure

Communicating and Acting on Results

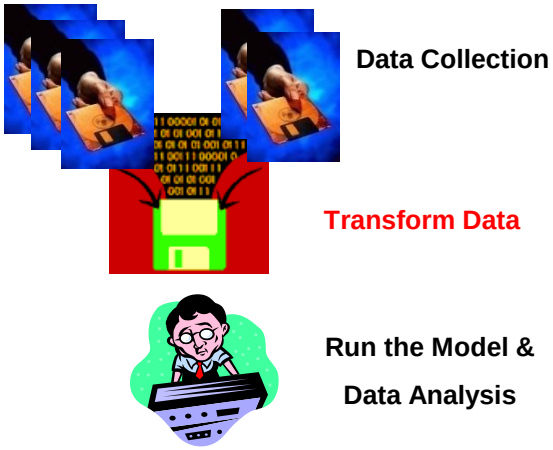
- **Results presentation and action**
  - Demonstrate process mining results to stakeholders
  - Capture process mining visualization as a video and e-mail to stakeholders so they can see the process bottlenecks and deviations
  - Explain the findings

Applying Process Mining to IT Big Data

- **Modeling and variable selection**
  - Process mining
    - Visualize normal and anomalous operations
      - Transaction process flows
      - Application process flows
      - IT Infrastructure process flows
  - Variables needed
    - Unique Transaction ID
    - Start and End Time Stamps
    - Business Process
    - Application Software
    - IT Infrastructure
    - IP Address

Applying Process Mining to IT Big Data (Continued)

- **Data collection**
  - Talk with system administrators to capture log data
  - Explain the data fields and formats you want
  - Transform the log files into the format that the process mining system needs
- **Data analysis**
  - Load the data into the process mining model
  - Run the process mining models
  - Visualize and analyze data
    - Transaction business process flows
    - Application software process flows
    - IT infrastructure process flows
  - Review the results



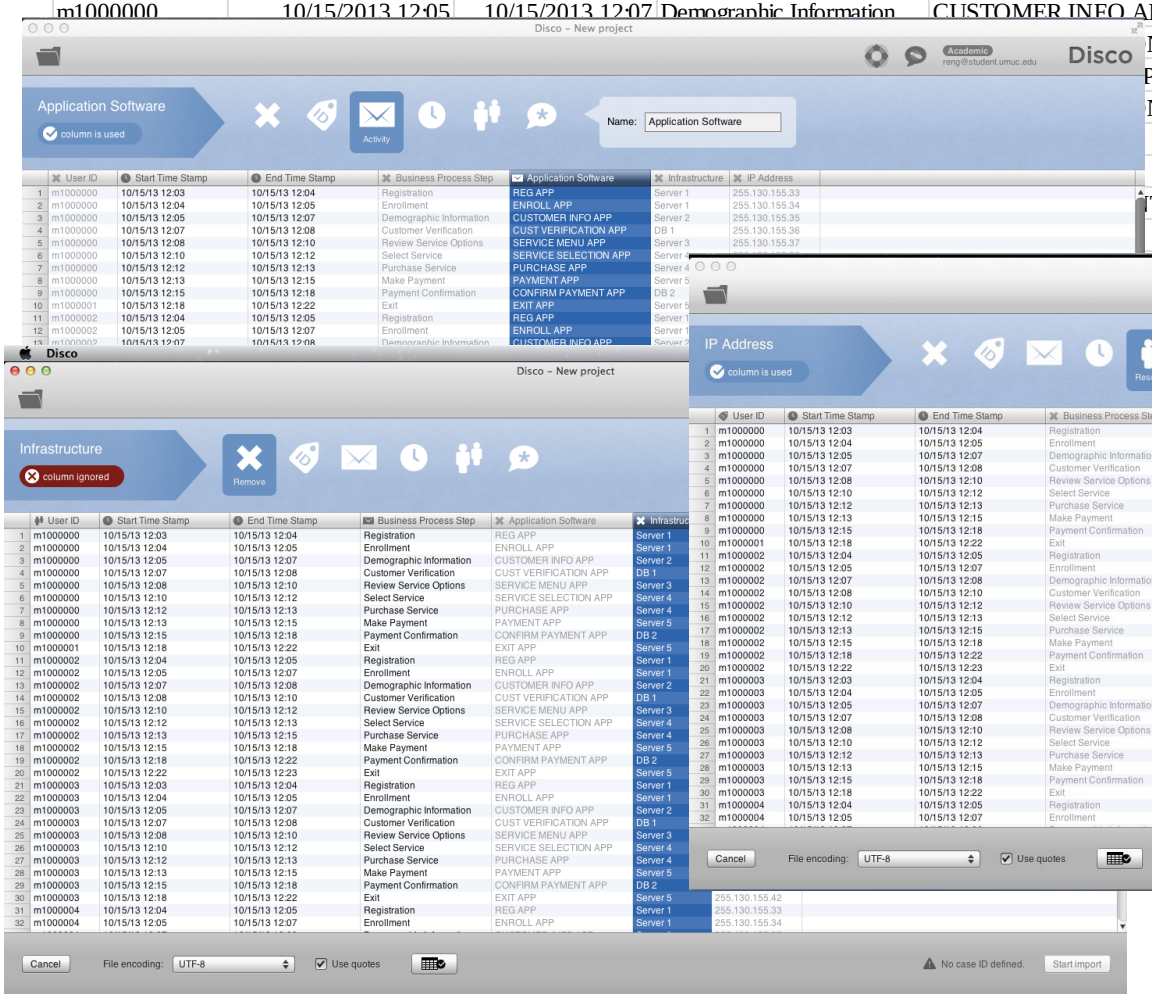
Applying Process Mining to IT Big Data (Continued)

- **Communicate results**
  - Provide feedback to stakeholders
    - Process owners
    - Application software team
    - IT Infrastructure team
- **Benefits**
  - Visualization of the actual process flows
  - Comparison of normal and anomalous operations - rather than manually reviewing individual system logs
  - More rapidly identify spurious processes

Illustration of Process Mining Data

DUMMY DATA

User ID	Start Time Stamp	End Time Stamp	Business Process Step	Application Software	Infrastructure	IP Address
m1000000	10/15/2013 12:03	10/15/2013 12:04	Registration	REG APP	Server 1	255.130.155.33
m1000000	10/15/2013 12:04	10/15/2013 12:05	Enrollment	ENROLL APP	Server 1	255.130.155.34
m1000000	10/15/2013 12:05	10/15/2013 12:07	Demographic Information	CUSTOMER INFO APP	Server 2	255.130.155.35
				N APP	DB 1	255.130.155.36
				P	Server 3	255.130.155.37
				N APP	Server 4	255.130.155.38
					Server 4	255.130.155.39
					Server 5	255.130.155.40
				T APP	DB 2	255.130.155.41
					Server 5	255.130.155.42



MITRE

Process Mining Demo (Using DUMMY DATA)

- **Transaction Business Process Flow**
- **Application Software Process Flow**
- **IT Infrastructure and IP Address Process Flow**

Lessons Learned

- **Good news**
  - Process mining concept works
- **Bad news**
  - Not all of the IT infrastructure was instrumented to collect log data
  - In some cases the log data was aggregate rather than per transaction
  - Some logs lacked unique transaction IDs
  - System time was not synchronized across the IT infrastructure
  - IaaS provider never contracted to provide system log data
- **Parting thoughts**
  - Make sure the network and system time for all of your infrastructure are synchronized
  - Require all XaaS providers to provide system log data
  - Plan for process mining at the start of the project

Should You Consider Process Mining?

- **“The most difficult subjects can be explained to the most slow-witted man if he has not formed any idea of them already; but the simplest thing cannot be made clear to the most intelligent man if he is firmly persuaded that he knows already, without a shadow of doubt, what is laid before him.” – Leo Tolstoy**

Contact Information

- **Richard F. Eng**

  reng@mitre.org

  reng@student.umuc.edu

  703-201-9112