The Rules of Subsurface Analytics*

Jane McConnell¹ and Duncan Irving¹

Search and Discovery Article #70373 (2018)**
Posted February 11, 2019

*Adapted from oral presentation given at 2018 AAPG Middle East Region GTW, Digital Subsurface Transformation, Dubai, UAE, May 7-8, 2018
**Datapages © 2018 Serial rights given by author. For all other rights contact author directly. DOI:10.1306/70373McConnell2018

¹Teradata, Edinburgh, United Kingdom (Jane.McConnell@Teradata.com)

Abstract

Oil and Gas companies are increasingly embracing big data analytics and data driven approaches in the drive to optimize development and production costs, increase recovery factors, and ultimately better understand and quantify uncertainties in their workflows. Most companies now have a digitalization program in place and are taking steps towards this data-driven future. From the projects that Teradata has conducted in the Oil and Gas industry, we believe that implementing a successful analytics program in subsurface involves following a few key rules. Firstly, they require bringing together the right people. Ideally what we refer to as “T-shaped” people – people with deep knowledge in one or more areas, but wide (if shallow) knowledge of the whole process, and who are open to trying new approaches. Secondly, the right data platform. Subsurface data certainly meets the Big Data definition of volume, velocity, variety, and veracity. Performing analytics on deep and wide datasets requires thinking about parallelism and performance – while also thinking about storage costs. Ensuring that analytics projects provide measurable business value requires us to take an agile approach to project management, and to repeatedly check the business alignment to ensure that the analytical results we are delivering are in some way actionable. Companies do not make or save money by running analytics projects – that only happens when they can take the learnings from the analytics projects and put them to use. In analytics projects, a vast proportion of time is spent on locating and preparing data. The required data may be available only in application databases, only as original files, or spread around various systems. We take an approach we refer to as “good enough data management” when building an analytical data platform, where structure and quality are applied in a just-in-time manner to meet the needs of the analytics. We will illustrate these key rules using case studies and anecdotes from past projects in Norway, UK, US, and South East Asia.
References Cited


Websites Cited


THE RULES OF SUBSURFACE ANALYTICS

Jane McConnell, Teradata
Duncan Irving, Teradata
Subsurface analytics is different

The Rules
- Rule 1: Right People
- Rule 2: Right Platform
- Rule 3: “Good Enough” Data Management
- Rule 4: Agile Approach
- Rule 5: Business Buy-in

Recap
SUBSURFACE ANALYTICS IS DIFFERENT
1940’s business computing: When it all started for business analytics

Figure 5. Leo
Level set on business intelligence

Image copyright: http://www.leo-computers.org.uk/
Figure 6. A lot has been written about the IT/OT divide over the last few years – mainly by Gartner… The normal examples given are SCADA systems, historians, PLCs – the technology of factories and of production operations. We have similar specialist technologies in subsurface too. Think about seismic processing, about interpretation suites like OpenWorks or Petrel.
When we want to start doing analytics on this data – we need to start bridging the divide - which means training IT on what we have been doing for the last decade or two, so that they can start helping us with what they know – business intelligence, analytics and machine learning, and the data platform.
Figure 7. A lot has been written about the IT/OT divide over the last few years – mainly by Gartner… The normal examples given are SCADA systems, historians, PLCs – the technology of factories and of production operations. We have similar specialist technologies in subsurface too. Think about seismic processing, about interpretation suites like OpenWorks or Petrel.

When we want to start doing analytics on this data – we need to start bridging the divide - which means training IT on what we have been doing for the last decade or two, so that they can start helping us with what they know – business intelligence, analytics and machine learning, and the data platform.
RULE 1: RIGHT PEOPLE
➢ Too many disciplines for any one person to know it all
➢ “T-shaped people” who go wide across many disciplines but deep into their specific domain
➢ Need outstanding data management and data engineering skills (and culture)
➢ Need platform expertise for sustainability and deployment
➢ Need Subject Matter Expertise

Right People, plural. And T-shaped.
Analytics / data science workflow

Ingest  Curate  Analyse  Visualise
How we are working with a Norwegian operator

- Working as one team, hand-in-hand with the customer
  - Subject matter expertise
  - Source system expertise
  - Data management skills
  - Data platform skills
  - Coding skills
  - Data science skills
  - Frontend/visualisation skills
RULE 2: RIGHT PLATFORM
“Knowledge development” applications come with import filters for specific file types and specific tasks.

Data is modelled logically for well-defined (and hence brittle) processes that may not reflect all (or even any!) use cases.

Only “perfect” data can be imported into applications or schemas.

New data types, or new combinations, don’t work very well in this old world.
➢ If we don’t provide a platform for analytics, we will be in Desktop/Excel Hell.

➢ Build a platform that
  – Accepts data from any discipline
  – Makes it easy for data scientists to use their tools – R, Python etc
  – Provides the right level of governance and data quality
  – Provides parallelism and scale

Build a new platform that all disciplines can use
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley, gholt, dgg, edavydov, toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crepuso, Dan Dennison
{ebner, vchaudhary, myoung, jcrepso, dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of technical debt, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration.

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Static Analysis of Data Dependencies. In traditional code, compilers and build systems perform static analysis of dependency graphs. Tools for static analysis of data dependencies are far less common, but are essential for error checking, tracking down consumers, and enforcing migration and updates. One such tool is the automated feature management system described in [12], which enables data sources and features to be annotated. Automated checks can then be run to ensure that all dependencies have the appropriate annotations, and dependency trees can be fully resolved. This kind of tooling can make migration and deletion much safer in practice.
What a Machine Learning system really looks like

Hidden Technical Debt in Machine Learning Systems

Configuration
- Feature Extraction
- Data Collection

ML Code
- Analysis Tools
- Process Management Tools

Data Verification
- Machine Resource Management

Serving Infrastructure

Monitoring

© 2018 Teradata
RULE 3: “GOOD ENOUGH” DATA MANAGEMENT
“Good Enough” Data Management

- Curators mentor and support “citizen data management”
- Everyone cares about the data and its quality
- Everyone can do something about it when they find bad data
- Data governance is a function of data value

“Good Enough” means:

- **Good**: don’t compromise on quality
- **Enough**: don’t boil the ocean
Historically, we’ve stripped all the context away from each measure and observation for the sake of more storage
- Parse out the measurement data
- Link it through time and space
- Relate using metadata and master data
- Pause – until you know how you want to access it
Historically, we’ve stripped all the context away from each measure and observation for the sake of more storage

- Parse out the measurement data
- Link it through time and space
- Relate using metadata and master data
- Pause – until you know how you want to access it
Text

➢ Language
➢ Typos
➢ Consistency
➢ Quality

Use simple characterisation tools to understand what is in the data

Don’t try to build a whole text input and cleansing framework, if you don’t need it
Use simple characterisation tools to understand what is in the data
Don’t try to build a whole text input and cleansing framework, if you don’t need it
➢ Storage is cheap
  – But still, sometimes we have a LOT of data with very low information density – eg passive seismic
➢ If the data is still to large to handle then profile and decimate (it’s better than never using it!)
➢ Storage is cheap
  – But still, sometimes we have a LOT of data with very low information density – eg passive seismic
➢ If the data is still to large to handle then profile and decimate (it’s better than never!)
RULE 4: AGILE APPROACH
Agile, Scrum, DevOps, AnalyticOps, Interactive Visualisation
What not to do

Common resource-sinks:
- Point solutions
- Technology projects
- Waterfalls
- Brittle data modelling
- ML/AI-driven project

© xkcd.com
Iterate.
One project at a time.
Deliver value often.

Prepare
Contextualise and plan
Form problem statement
Prioritise by impact
Communicate analysis plan and responsibilities

Review
Create concise business story
Highlight overall business impact
Include assumptions and sources
Follow up with business on the actions

Assimilate
Store well-commented SQL
Document in wiki
Train BU in tool usage

Execute
Build on prior work
Validate data
Recheck hypotheses
Drive insights and recommendations

Document
Post code to repository
Stakeholder contacts
Final presentations

Assimilate
Store well-commented SQL
Document in wiki
Train BU in tool usage
Agility needs the right mindset

- Working together
- Willing to take risks
- Proactive
- Speed / Time-to-insight
RULE 5: BUSINESS BUY-IN
“We created an Analytics COE. We hired some data scientists. We installed Hadoop. We’re ready to Machine Learn something now. Do you have any use cases?”
Business-aligned data science

- What financial, operational or environmental impact are you delivering?
- What is the ACTION that you can take?

- What techniques, functions, workflows and skills are required?

- What data is required and in what form?
- Do we even have the needed data?
➢ Embed data ownership in the business units
➢ Engage with business leadership to plan, budget and deliver data-driven initiatives
➢ Define and drive data exploitation strategy
➢ Understand data value and leverage high value data for business impact

...do we need a **Chief Data Officer?**

*Source: Andrew White, Gartner 2017*
IN SUMMARY
The Rules

Right People
Right Platform
Good Enough Data Management
Agile Approach
Business Buy-in
### Data Management 2.0

**Stop doing**
- Brittle data management
- Silos
- Disposable data science
- Transfer and analysis in Excel

**Keep doing**
- Applying domain expertise
- High levels of governance
- Driving data quality
- Learning

**Start doing**
- Aligning with business
- Applying context
- Data profiling
- Enriching data
- Applying critical thinking
Jane McConnell
Practice Partner O&G, Industrial IoT Group
Jane.mcconnell@teradata.com
+44 (0) 7936 703343

My blog on Forbes
My blog on Teradata.com
Follow me on Twitter @jane_mcconnell
My profile