Standards:
The Common Language of Trusted Data

Jay Hollingsworth

CTO, Energistics

Agenda

- » About Energistics
- » The Four Standards – WITSML, PRODML, RESQML, ETP
- » Cloud Analytics
  - Data in Motion
  - Data at Rest
- » Data Needs a common language
- » Trusted Data
  - Quality Data

Who are we? (Hint: we are not a vendor…)

- Energistics is a global, non-profit, membership consortium focused on developing open data exchange standards for the upstream oil and gas industry
- Evolving from POSC, we have served the industry for more than 25 yrs
- Around 110 member companies, representing E&P operators, oilfield service companies, software vendors, system integrators, regulatory agencies and the global standards community
- Our standards are developed by workgroups made up of industry experts from our member companies
- In short, the standards are created **by the industry** and **for the industry**

energistics®
Energy Standards

Global Influence

accenture
High performance. Delivered.
Arizona Geological Survey AZGS
BAKER HUGHES a GE company
BARDASZ
bp
CNSOPB Canada-Nova Scotia Offshore Petroleum Board
CGG
Chevron
CDA Oil & Gas UK

CMG Computer Modelling Group Ltd.
CoreLogiQ
DataHorizon
U.S. Department of the Interior
devon
DIGIDRILL
DYNAMIC GRAPHICS, INC.
E&Tech Developments
energySYS
EPIM

ExxonMobil
F2i Consulting
GEOLOG Surface Logging
Geologix
GEOSIRIS
GROUNDWATER PROTECTION COUNCIL
HALLIBURTON
HASHMAP BIG DATA ACCELERATORS
HRH geological services
ifp Energies nouvelles

IDS Independent Data Services
Infosys
INT
KAPPA
KONGSBERG
L&T Infotech
Logtek
Australian Government National Offshore Petroleum Titles Administrator
NATIONAL OILWELL VARCO

Navita Origo solutions
NEW ZEALAND PETROLEUM & MINERALS
noah consulting
NORWEGIAN PETROLEUM DIRECTORATE
OFS Portal
Oil & Gas Authority
FI INDIA
OILWARE
OSIsoft.

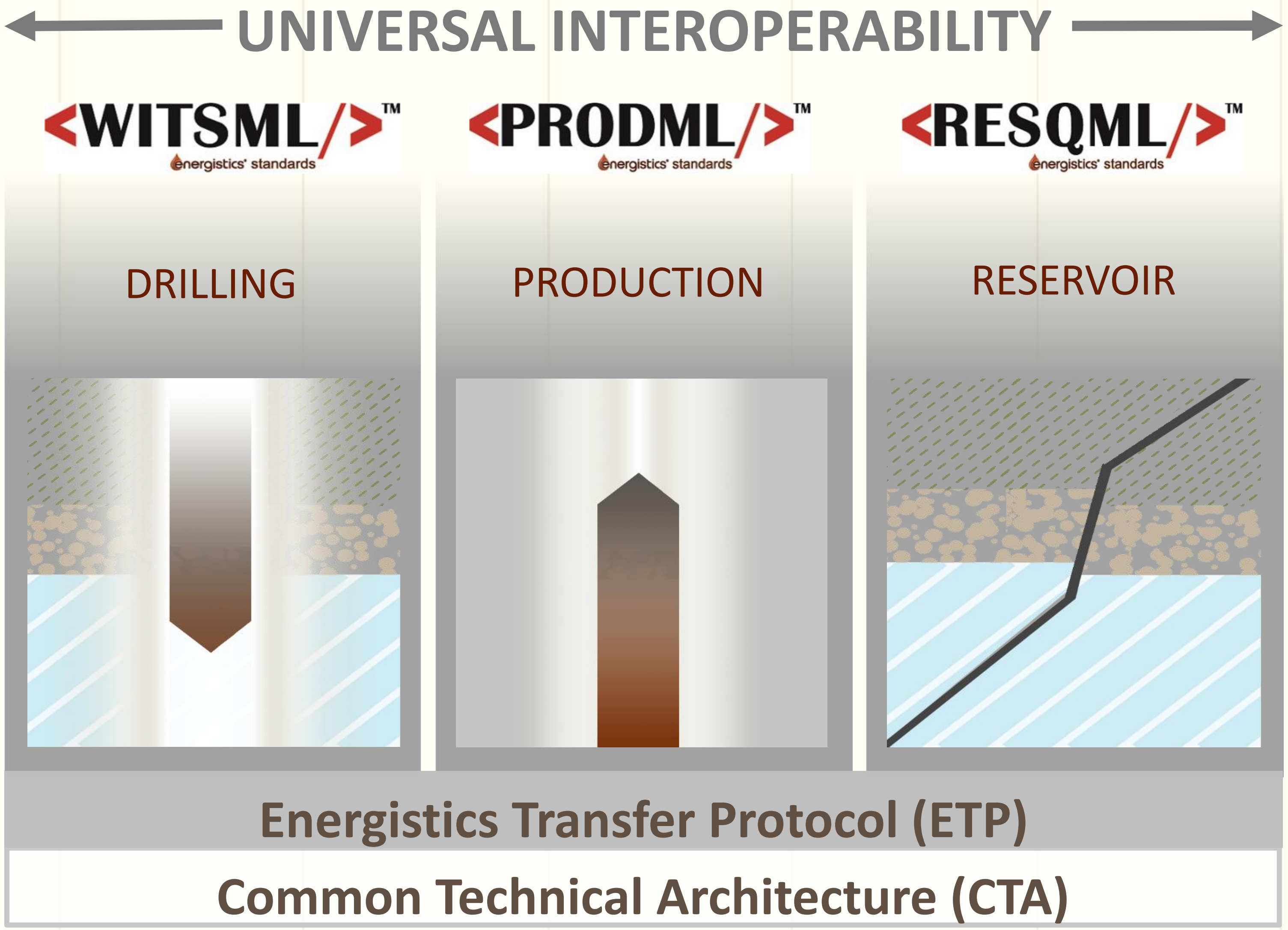
P2
Paradigm
pason
peloton well focused
PetraBYTES
BR PETROBRAS
Petroleum Agency SA
Petroleum Commission Ghana
شركة تنمية نفط عمان
Petroleum Development Oman
petrolink

PETRONAS
pds petrotechnical
PIONEER NATURAL RESOURCES
Polaris Guidance Systems LLC
proNova by TDE
Radmond Open platform for mobile collaboration
redhat
CNPC BGP Richfit
roxar

ارامكو السعودية
Saudi Aramco
Schlumberger
Scientific Drilling
skkmigas
SSG
STAATSOLIE
Statoil
energistics Energy Standards

Tech Mahindra
tecplot
TIBCO
TNO innovation for life
TOTAL
WellSite SOFTWARE
wipro
woodside

Energistics' Spectrum of Standards



UNIVERSAL INTEROPERABILITY

<WITSML/>™ energistics' standards

<PRODML/>™ energistics' standards

<RESQML/>™ energistics' standards

DRILLING

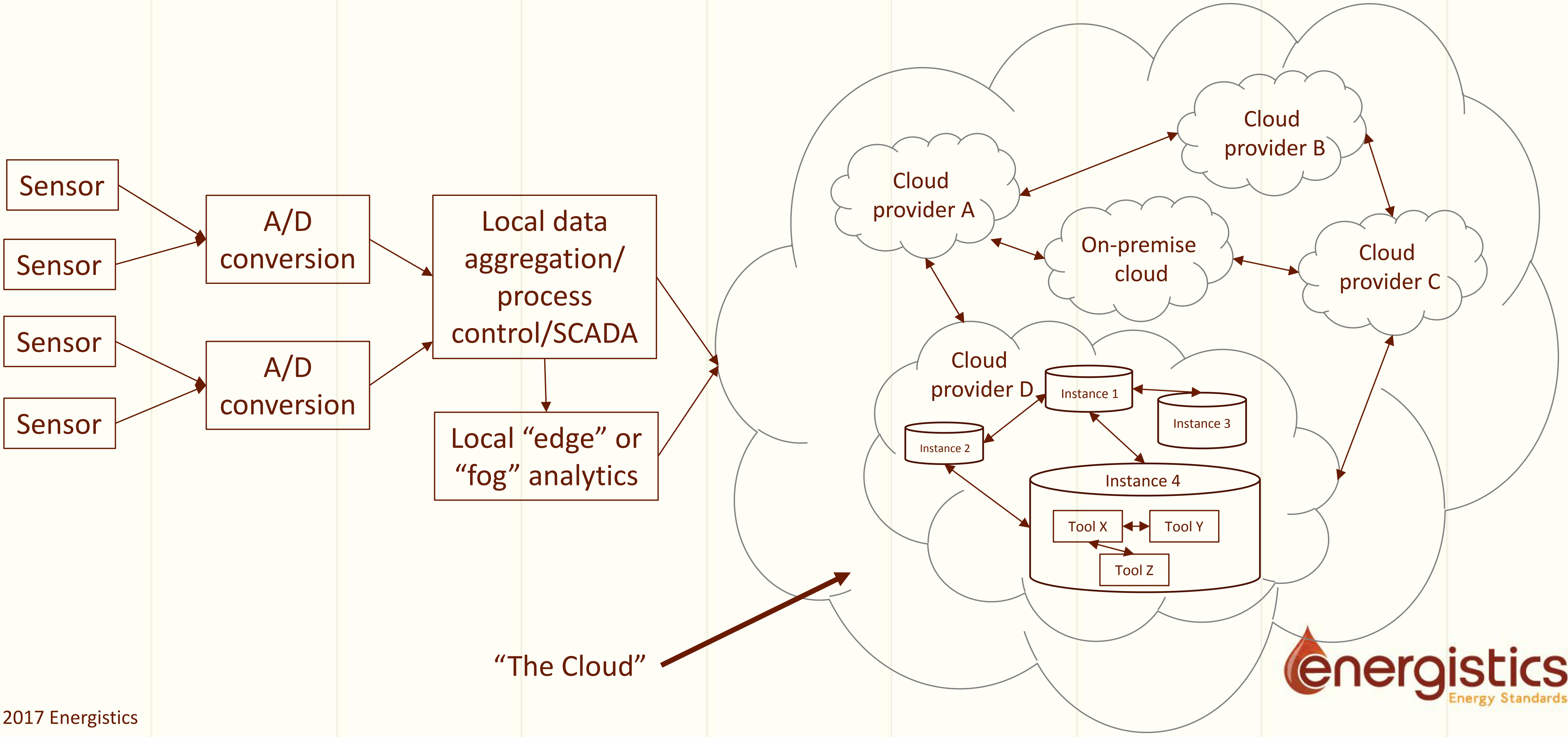
PRODUCTION

RESERVOIR

Energistics Transfer Protocol (ETP)

Common Technical Architecture (CTA)

Typical “Cloud” Data Analytics



Sensor

Sensor

Sensor

Sensor

A/D conversion

A/D conversion

Local data aggregation/ process control/SCADA

Local “edge” or “fog” analytics

Cloud provider A

Cloud provider B

On-premise cloud

Cloud provider C

Cloud provider D

Instance 1

Instance 2

Instance 3

Instance 4

Tool X

Tool Y

Tool Z

“The Cloud”

Analyzing Data in Motion

- For analyzing data in motion the need for transfer standards is obvious
  - Multiple service vendors will be feeding in a real-time analytics platform
  - Real-time streams will be forked off to feed many different partners
  - Commercial environments will be cheaper if there is little/no configuration
    - Pre-built standard interfaces will be the easiest
    - Would logically match the structure of the incoming data
- For analyzing data at rest the need for standards is in lake connections

Analyzing Data at Rest

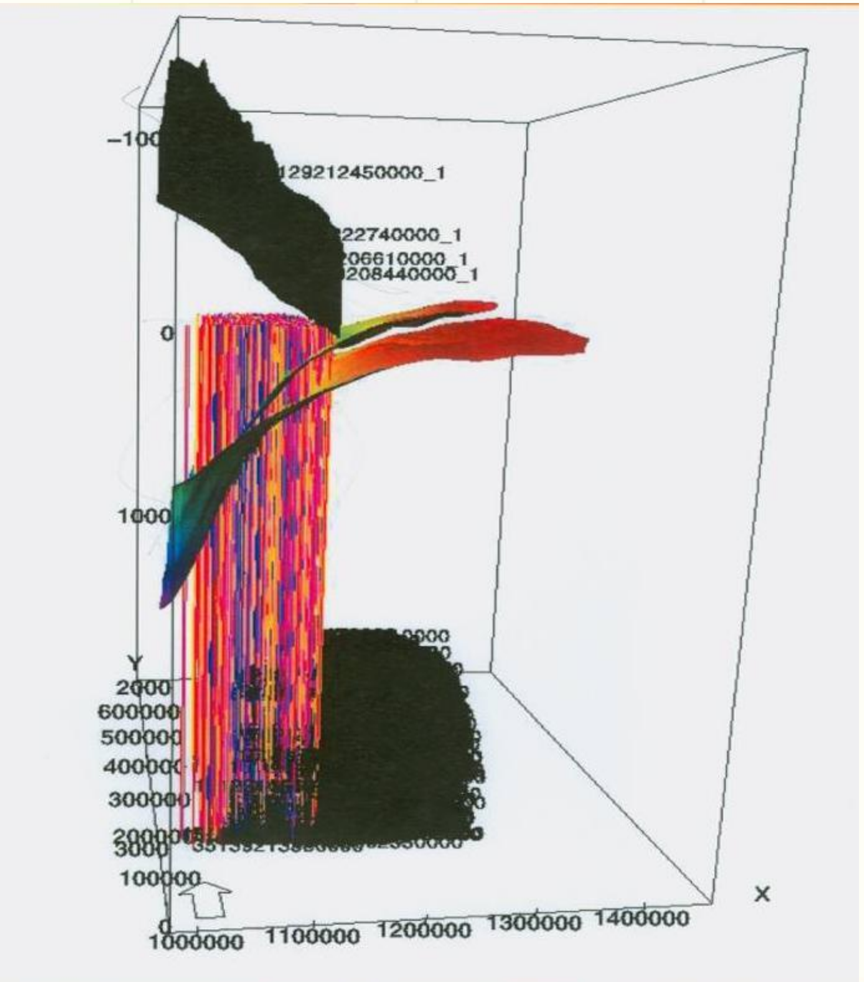
- To analyze data at rest there are 5 connections that need standards
  1. Ingestion – clearly at the loading point a data lake needs a known format
  2. Lake-to-lake – few companies (none?) will have only one data lake, so transfers among them will need to be of a known format
     - Technical, safety, trading, ERP, operations are likely to be different lakes
  3. Internal to the lake – the data in a lake is normally in the original format
     - If it matches the ingestion schema then the transfer standard is the storage format
  4. Lake-to-analysis – analysis is unlikely to be done on the lake directly
     - Analysis will likely be in a different running instance in the cloud (except for Hadoop/YARN)
       - A few select full-time data scientists may use it, but not thousands of general users
       - See https://sonra.io/2017/08/08/are-data-lakes-fake-news/
     - Analysis could be on-premise, like a local Spotfire or SAS or R or Spark
  5. Internal to analysis – microservices and others need live standardized conn's

Every Connection Needs a Language

- Words need common meaning
  - If data is going to speak to you
- There are open groups doing this
  - Energistics - upstream technical data
  - PIDX - upstream business data
  - SEG - seismic data
  - OPC - process control
  - IOGP – mapping
- How can you trust something you don't understand

When Official Sources of Information Fail Us

- Trust is lost when data looks like this or this
- Knowledge workers **will** fix their data
- They may abandon use of official sources
- And create their own databases
  - DW guys call these “rogue data marts”
- These private databases often gain users
- In the end, there is so much waste
  - Of time in creating rogue databases
  - In incorrect results because the “fixes” are bad



Trusted Data

- What gives us trust in data?
  - Assuming no *a priori* reason to doubt it
- We trust data when we are aware of its level of trustworthiness
  - Even if a source of data is less than trustworthy, our knowledge of that level of trustworthiness allows us to use it appropriately
  - Latest Energistics' standards allow us to describe trustworthiness
- Trustworthiness in data derives from three characteristic dimensions
  1. We trust the source of the data
  2. We trust the handling the data has experienced
  3. We trust that the data has not been tampered with since it left its source

From Data Transfer to Data Assurance: Trusted Data is the Key for Good Decisions –
2017 PNEC Conference Proceedings

energistics®
Energy Standards

Trust in the Source of the Data

Trust in a source of the data covers a number of facets

1. Believability – Does the data conform to sensible rules? That is, is it complete, within a reasonable range, the right datatype, etc.?
2. Reputation – Is the source considered to be a good source by a community? Is it the “official” source one is supposed to use? Any bad experiences?
3. Objectivity – Could there be a bias in the source? Are there known to be competing perspectives on the data which could cause it to be suspect?
4. Reliability – Does the source normally give good information? Are there known issues which reduce reliability (e.g., sensor too hot, calibration)?

energistics®
Energy Standards

Trust in the Handling of the Data

Has the data been handled in a trustworthy way?

1. Origination – Can the original source be identified?
2. Traceability – Is the lineage of the data available for inspection? Has the data been processed, and if so what algorithms and parameters were used?
3. Stewardship Status – Were the data captured automatically or manually?
4. Originating Software Inputs and Parameters – For calculated data, what software, settings and users were involved in creating it?

Trust in the Security of the Data

Has the data been handled securely

1. Authentication – Is the source of the data protected so that only access is limited to a knowledgeable few? Or could anyone (or thing) have created it?
2. Authorization – Is there a process for authenticated users to be identified and authorized, or do users self-register themselves?
3. Roles Policy – Can all users in the source system write or delete data?
4. Auditing Policy – Does the source system keep track of who made changes?
5. Licensing Policy – If this data is licensed, can I legally use it?
6. Disclosure Policy – Am I allowed to use data if the results are made public?

Data Assurance vs Data Quality

- When you know the “quality” of your data, you trust it
- “Data quality” too hard to universally define
- Data Quality workgroup focused on Data Assurance
- High priority use cases
  - Conformance to policies/rules
  - Traceability of data through users, servers, versions
  - Traceability through measurement, derivation, parameters
  - Calibration – when, how

energistics®
Energy Standards

Data Quality in the New Generation of MLs

- » Data quality measures can be transferred
  - “Quality” is not calculated or guaranteed by the transfer
- » The quality conveyance mechanisms introduced are
  - Arbitrarily complex “metadata” about a measurement or other data
  - The DataAssurance object – a special explicit carrier for data quality measures
  - Use of the Activity model to describe in detail how an data value was
    - Created (who did it using what program with what settings using what input data)
    - Inferred from a sensor via calibration (including the calibration activity, apparatus, etc.)

Q&A

Thanks for your attention

Jay Hollingsworth
jay.hollingsworth@energistics.org

energistics®
Energy Standards