Mike Beckerle, Data Archeologist, Tresys Technology
mbeckerle at tresys.com or at apache.org
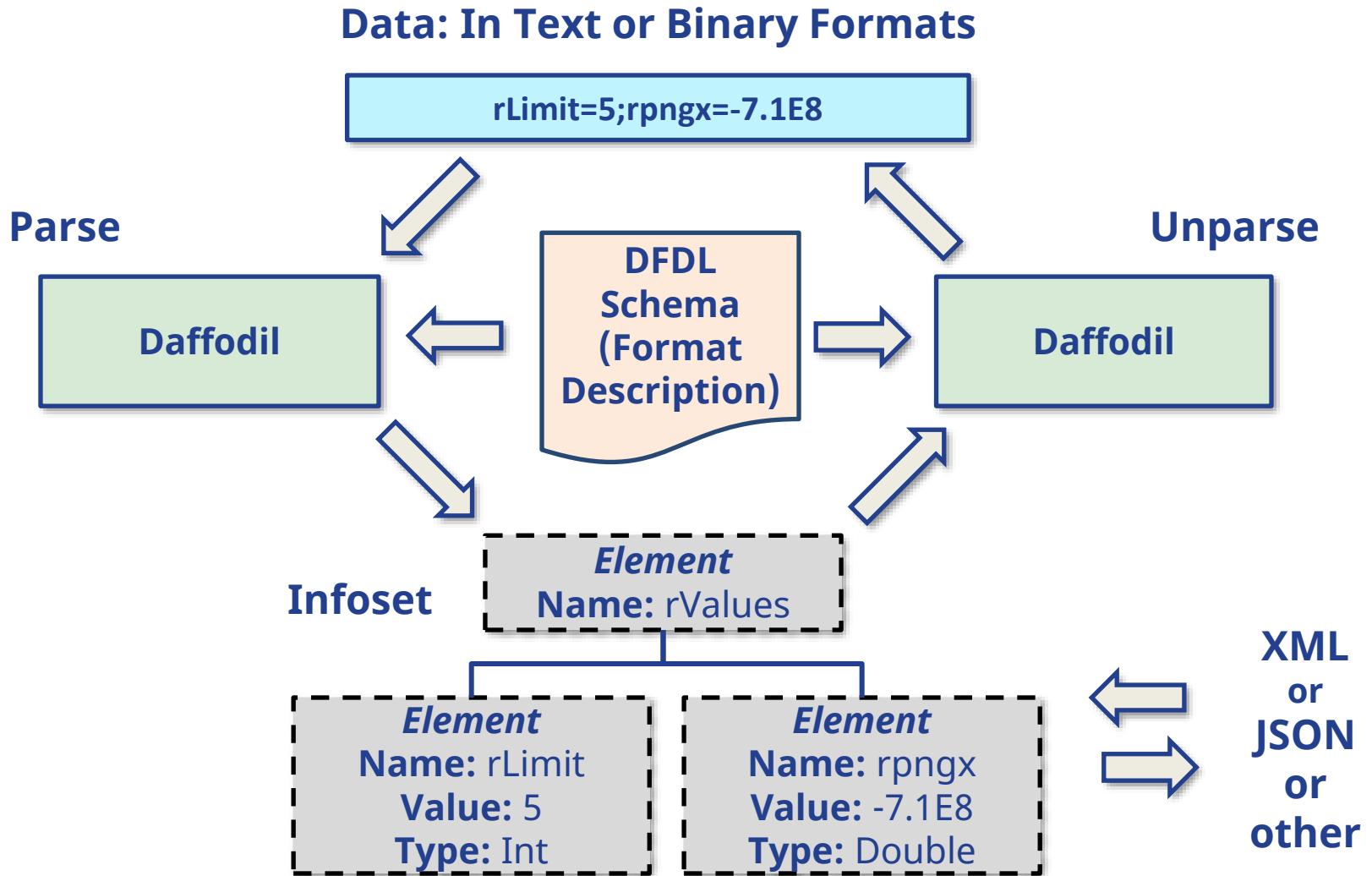
Kill the Data Format Problem

Data: In Text or Binary Formats

rLimit=5;rpngx=-7.1E8

Parse

Unparse

Daffodil

DFDL Schema (Format Description)

Daffodil

Infoset

Element
Name: rValues

Element
Name: rLimit
Value: 5
Type: Int

Element
Name: rpngx
Value: -7.1E8
Type: Double

XML
or
JSON
or
other

APACHE
Daffodil.

TRESYS
Deep.

Daffodil - Kill the Data Format Problem

- Podling since Sept 2017
- 4 Apache Releases - most recent July 2019
- Technology Status
  - Being deployed in military cybersecurity systems today
  - More than 40 data format schemas available. Breadth of formats covering:
    - FINSERV - Card Transactions ISO8583, NACHA, Point-of-Sale TLOG
    - Supply-Chain - EDIFACT
    - Health Care - HL7
    - Military messaging - Link16, VMF, USMTF
  - Preliminary Integrations with NiFi, Spark

- Status: Community Building
  - The bad news: Existing PMC all work for Tresys ☹
- What have we been doing about it?
  - Presented at ApacheCon NA 2018, 2019, Apache DC Roadshow
  - Ready for new developers
    - 32 "Beginner" JIRA Tickets

APACHE **Daffodil**™

TRESYS
Deep.

Proposal to Sharks
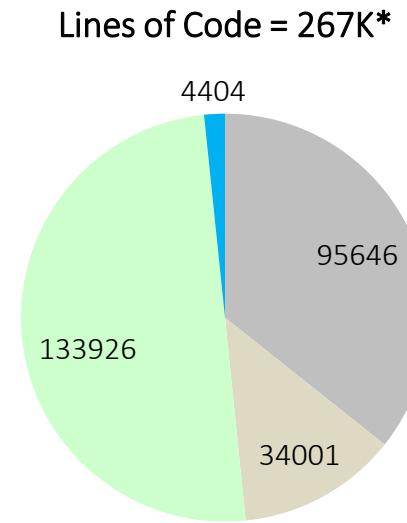
We are requesting

- 10 hours of devoted shark time to
  - Identify new developers
  - Spread the word about Daffodil to data-framework/data-centric Apache projects
- In exchange for a pledge of
  - Up to 40 total hours of virtual pair-programming hands-on training for new developers
  - Assistance to anyone in DFDL schema development for their data format
  - Eternal gratitude

Challenges in Recruiting Developers

- Daffodil is big and complex
  - Full-blown compiler for DFDL schemas
  - Efficient low-level runtime for parse/unparse
  - Learning curve is steep
- Know or Learn Scala
- Know or Learn XML Schema

- We need developers ....
  - undaunted by these challenges
  - motivated by the desire to *kill the data format problem once and for all.*

Lines of Code = 267K*

4404
95646
34001
133926

APACHE
Daffodil™

TRESYS
Deep.

Cool Daffodil Ideas…

- Add support for pointer-based formats like TIFF and ZIP.
- Extend Daffodil with SAX-style event streaming
- Hybrid Infoset
  - Daffodil directly constructs your framework-native data representation with tight metadata coupling
    - E.g., Spark Struct objects
  - Removes the XML/JSON conversion overheads
- Ultra-fast small-footprint backend for Daffodil
  - Generate C/C++
  - Generate FPGA for wire-speed parse/unparse
- Data-Format Debugger
  - Graphical display shows data and parsed tree and schema interactively