Open-Source In-Memory Platforms

Benefits of Coming Out of the Closet

Dr. Konstantin Boudnik
VP Open Source Development
WANdisco
Short bio:

Got addicted to Linux back in 1994
Member of Linux Foundation
Member of Apache Software Foundation
  Apache Bigtop project founder; Apache Hadoop committer
  Committer, PMC member, contributor to many ASF projects
  Mentor of multiple Apache Incubator projects
    Ignite, Geode, Groovy, Zeppelin
Background in compilers, JVM, distributed computing, system integration & architecture
2015: got invited to IMCS to talk about FOSS
Open-source / Open community

• Open source is easy
  • Jump to a “social development” site (Bitbucket, Github)
  • Pick up a license you like: (L)GPL, ASL, MIT, BSD
  • while true; do <code>; done
• You might be lucky to recruit a lot of volunteers
• A selected group mostly owns the project road map
• Adoption might be an issue as the future is unknown
Open-source / Open community

• Open community is way harder
  • Place to collaborate; meritocracy
  • Consensus building / Conflicts resolution
  • Continuity: avoiding 'hit by the bus' situations
  • Protection of project brand (under some licenses)
  • Legal shielding and takeover protection
  • Infrastructure management
  • Projects cross-polination
• “Community over code”
In-memory: what's FOSS'ing?

- Was sorta quiet up to pretty much 2012
  - Spark appeared & gained momentum quickly
    - Great improvement on MapReduce
    - Solved many shortcomings of MR
- Then nothing spectacular was happening until
  - 2014: Apache Ignite (incubating) from GridGain
  - 2015: Apache Geode (incubating) from Pivotal
open-source or open community

• FOSS foundations facilitate open communities
• Spark: from a relatively small GitHub project to the most active Apache BigData project in 2 years
• Apache Ignite: doubling committer base in 5 months; quadrupling the user base
• Apache Geode: check the talk @IMCS!
Apache Bigtop:
From #BigData to #FastData

Coming out of the closet: benefits of FOSS
Solving the complexity

#BigData

Coming out of the closet: benefits of FOSS
Apache Bigtop primer

• A project, environment, and a phylosophy to:
  • Define and create software stacks (think Debian)
  • Deploy and validate actual software in the real world
  • Configuration management
• Guarantees of consistency and compatiblity
• Empirical vs Rational
  • don't rely on someone's hearsay
  • don't assume an environment: contol it

One stack to rule them all
Apache Bigdata stack

• Bigtop is the cutting edge of Apache Bigdata stack
• Delivers:
  • A ready data processing stack
  • Dev. env. for anyone to create their own
  • Framework for easy integration/deployment/validation
  • “It works on my laptop” isn't cool anymore
• 0.x release series was focused on Hadoop ecosystem
10K view of Bigdata

• There's more than just Hadoop
• Hadoop is mere 5-10% of all Bigdata usecases
  • Good for processing data in parallel
  • Analytics and ML
• But it is NOT ideal...
  • Suboptimal resource scheduling
  • Batch oriented (mostly)
What's missing

• Hadoop is all about batch
  • MR is slow and heavily IO-bound
  • 2\textsuperscript{nd} generation of tools might be a bit more interactive
• SQL is the most popular data access interface
  • yet immature in Hadoop ecosystem
• Supporting transactions is very hard
• Almost everything is HDFS-bound
  • Performance... performance... performance
• Scarce In-Memory Computing presence
IMC: what is that?

- technically, any computing gets done in memory, but...

  "IMC: middleware software that stores data in RAM, across a cluster of computers, and process it in parallel"

- Why In-Memory Computing?
  - RAM is about 5,000 faster than HDD
  - RAM is about 1,500-2,000 faster than SSD
Apache In-Memory Computing

#FastData

Coming out of the closet: benefits of FOSS
Let's get serious about IMC

- Bigtop boards more & more IMC(-like) components
- Provides transitional tech for legacy MR-based users
  - HDFS acceleration
- MR acceleration
- Uses RAM as inter-component data media
  - Crossing component boundaries w/o leaving RAM
  - Advanced clustering and service models
Connecting the stack

- Bigtop Data Fabric Core:
  - Works with HDFS/RDBMS/MR/Hive/Hbase/Spark/Storm/SQL
- Cluster memory is a natural media to exchange data
- A probable usecase:
  - Kafka --> Data Fabric --> HBase --> Data Fabric --> SQL querying --> Spark --> A service Singlethon --> Data Fabric --> RDBMS or FS
Coming out of the closet: benefits of FOSS

Data Fabric: what is that?
Coming out of the closet: benefits of FOSS
Data Fabric: ... some more

Coming out of the closet: benefits of FOSS
Transitory legacy support

Data Grid  Compute Grid  Service Grid  Streaming
Advanced Clustering  File System  Messaging  Events  Data Structures

Hadoop Acceleration

Coming out of the closet: benefits of FOSS
Coming out of the closet: benefits of FOSS
ML and NoSQL on fabric

Coming out of the closet: benefits of FOSS
Analysing w/ 3rd party tools

Data Grid
Compute Grid
Service Grid
Streaming
Advanced Clustering
File System
Messaging
Event

Data Structures

Tableau, etc.

Coming out of the closet: benefits of FOSS
Deploy nodes everywhere
Connecting the ...
Live Demo

• Deploy Apache Ignite (incubating)
• Run MR Pi on YARN
• Run **same** MR Pi against Data Fabric:
  • Only client config needs to be changed
• Gasp at the difference
Final recap

• Build your project in the open
• Open community helps in many ways
• Find a good foundation to be your home
• Be inclusive and welcoming
  • a developer from a competitor can be a great contributor and a friend
• There's no “boss” in open source
• Keep coding: your code is your best resume!
Coming out of the closet: benefits of FOSS
Dr. Konstantin Boudnik
@c0sin
cos@apache.org

Coming out of the closet: benefits of FOSS