DIAGNOSING OPEN-SOURCE COMMUNITY HEALTH WITH SPARK

William Benton

Red Hat Emerging Technology

willb@redhat.com

BACKGROUND
FEDORA MESSAGING
BULK DATA INGEST
ANALYZING COMMUNITY
NEXT STEPS

BACKGROUND

FEDORA MESSAGING

BULK DATA INGEST

ANALYZING COMMUNITY

NEXT STEPS

WHAT MAKES A GREAT OPEN-SOURCE COMMUNITY?


fedora
TM
TM

QUICK FACTS ABOUT FEDORA

- Seven architectures
- ~17k packages (in Fedora 22)
- Tens of thousands of contributors
- Millions of users
- Many moving parts in Fedora infrastructure

BACKGROUND

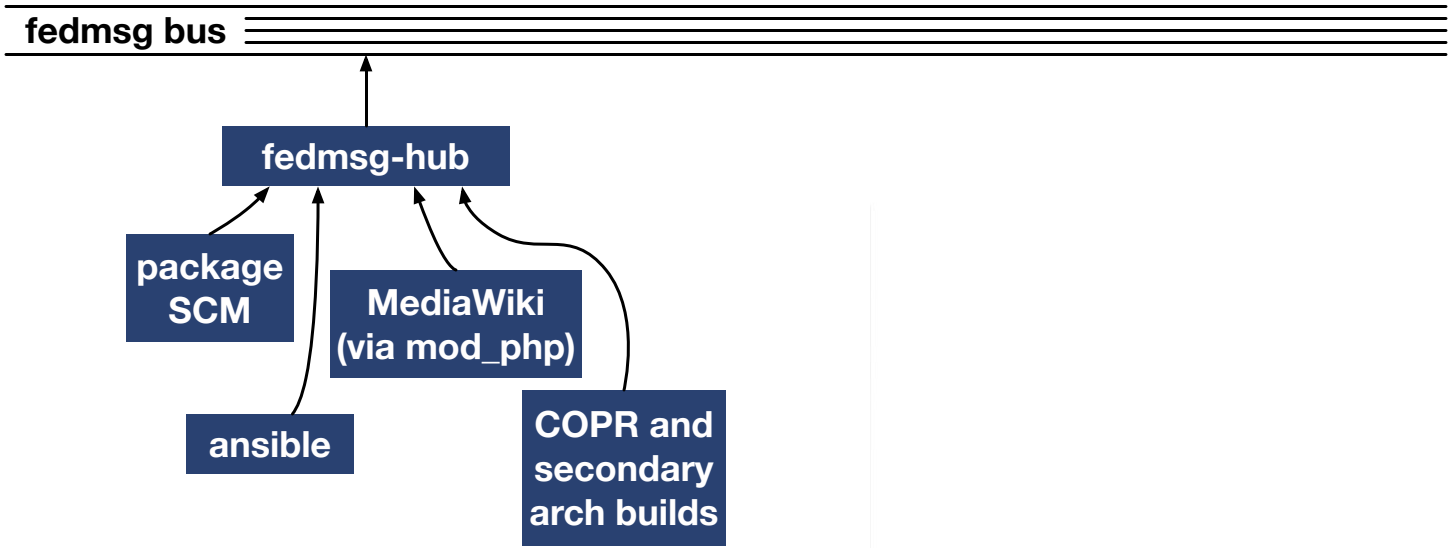
FEDORA MESSAGING

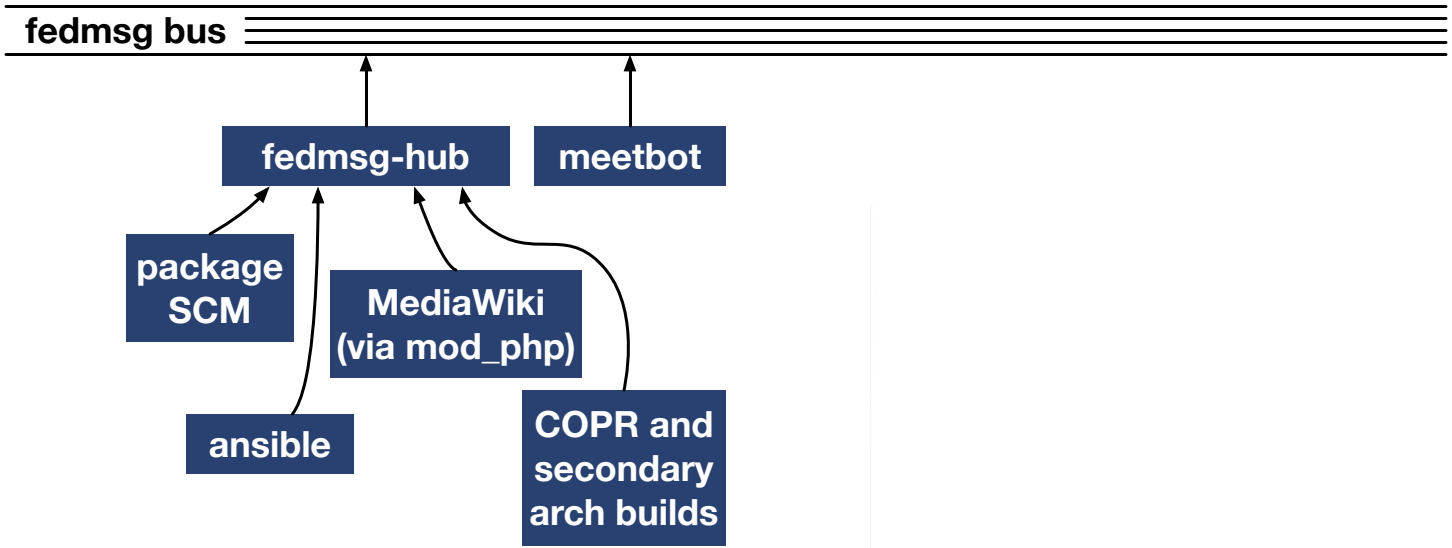
BULK DATA INGEST

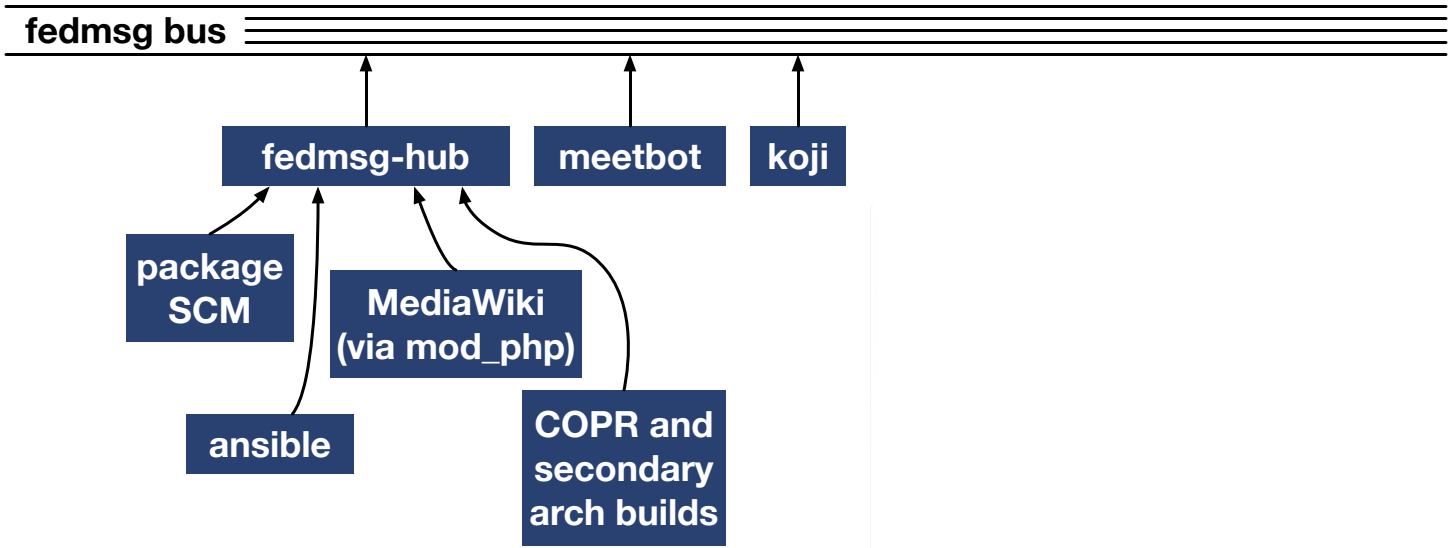
ANALYZING COMMUNITY

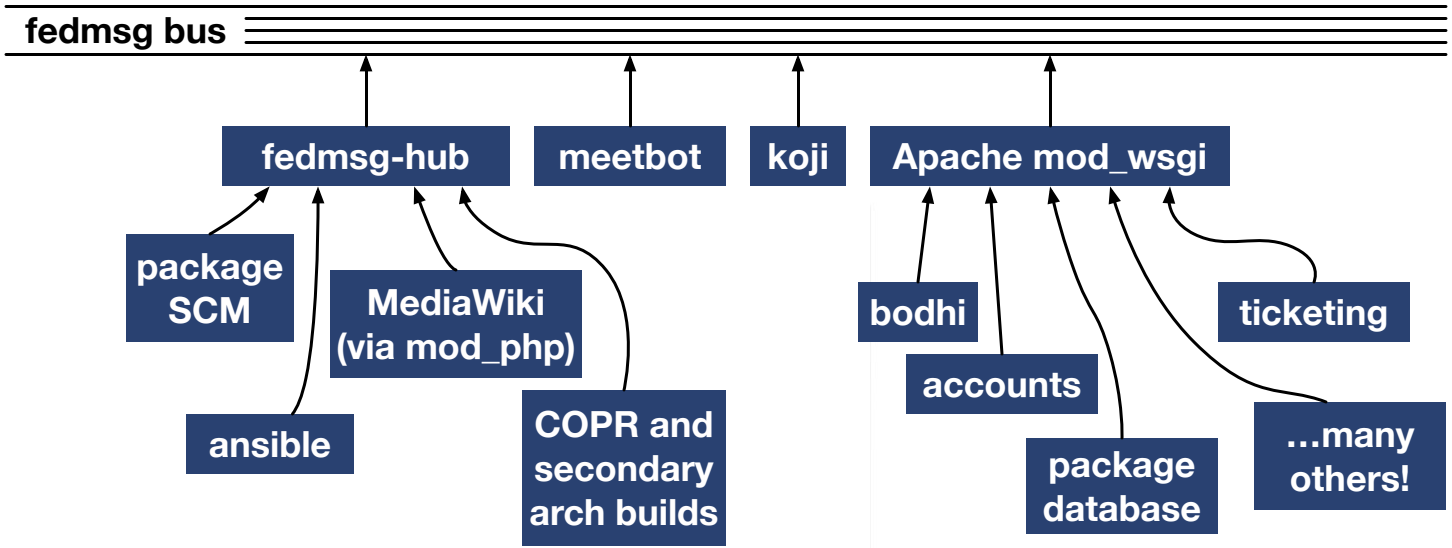
NEXT STEPS

fedmsg bus

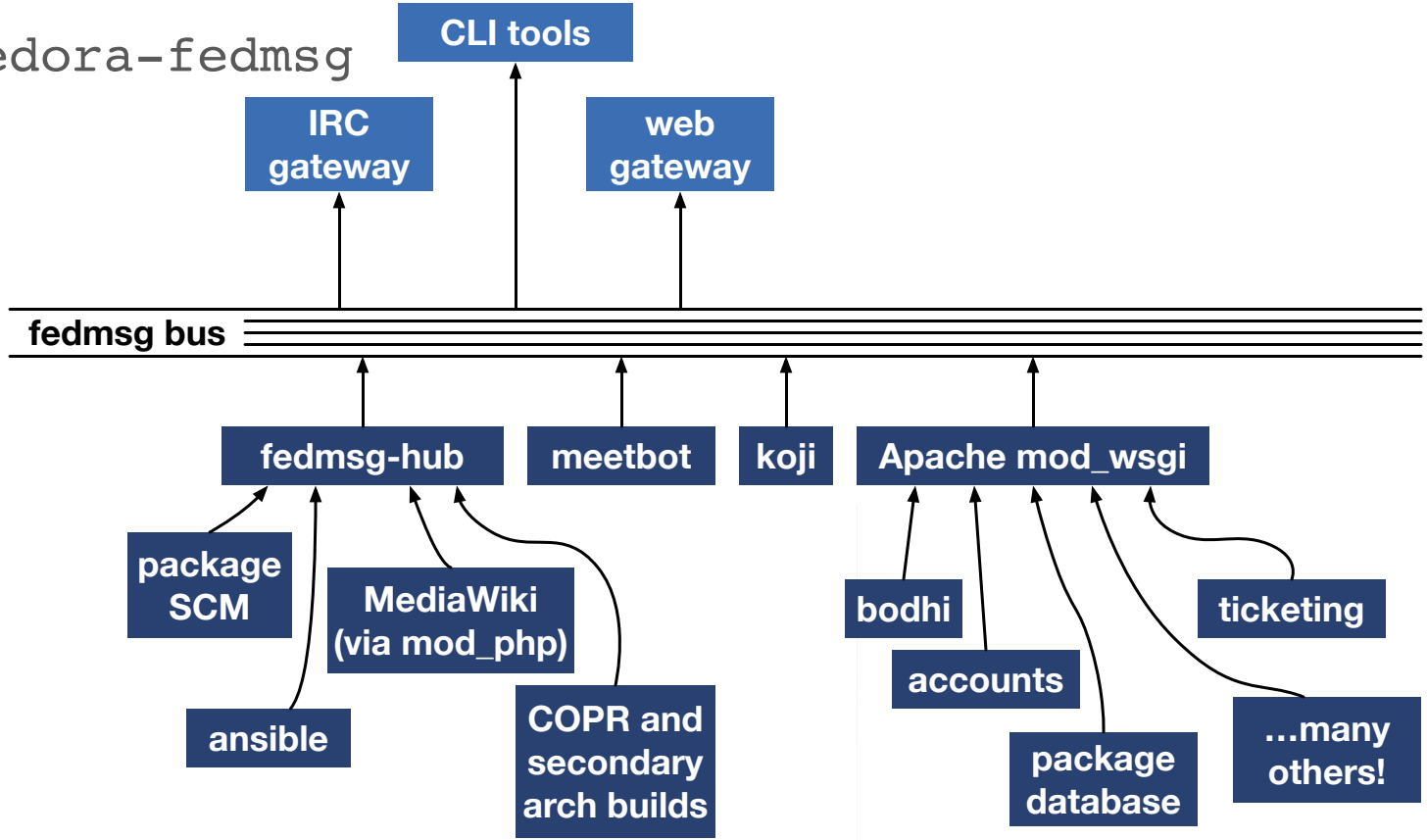

fedmsg bus
fedmsg-hub
package SCM
MediaWiki (via mod_php)
ansible
COPR and secondary arch builds


fedmsg bus
fedmsg-hub
meetbot
package SCM
MediaWiki (via mod_php)
ansible
COPR and secondary arch builds

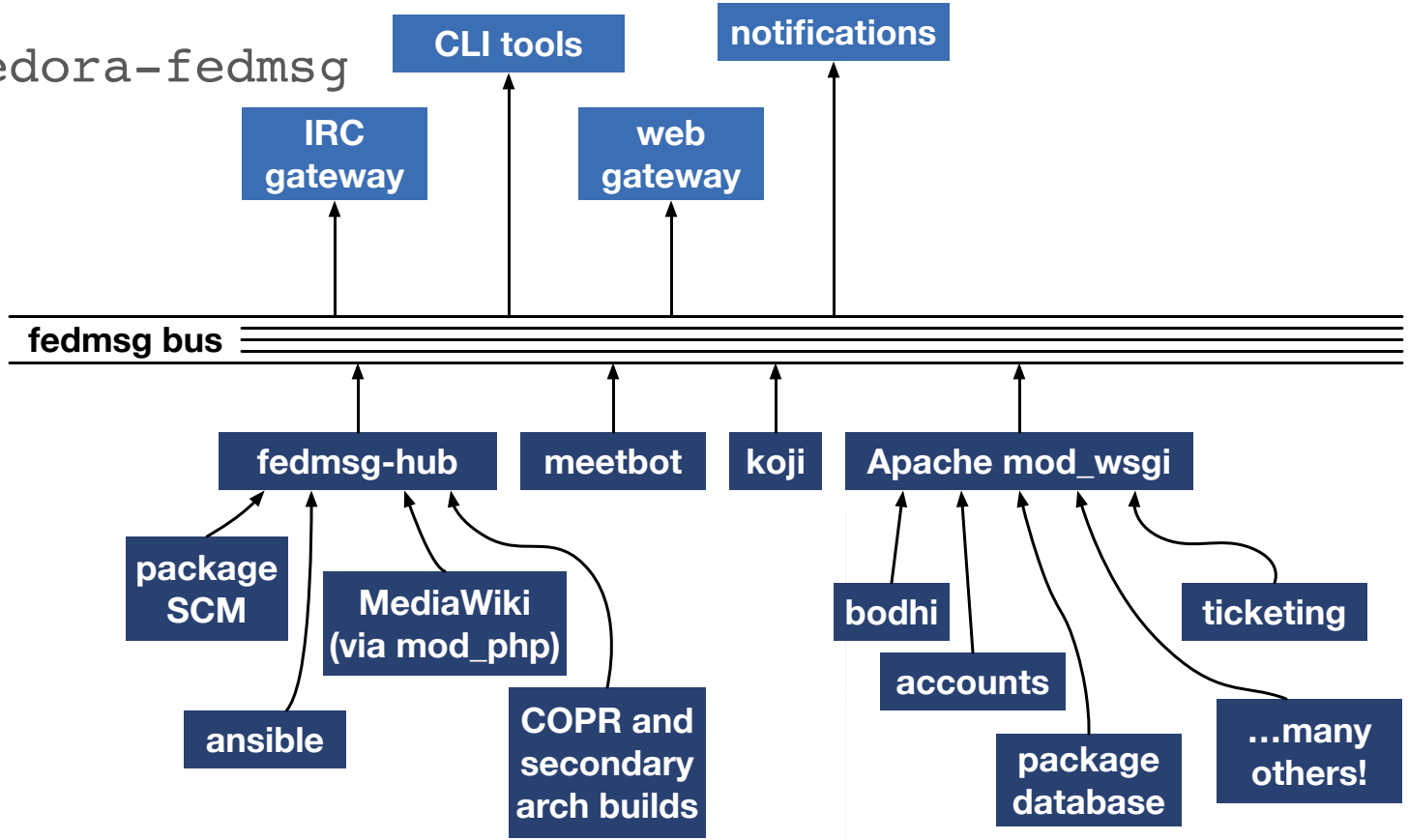

fedmsg bus
fedmsg-hub
meetbot
koji
package SCM
MediaWiki (via mod_php)
ansible
COPR and secondary arch builds


fedmsg bus
fedmsg-hub
meetbot
koji
Apache mod_wsgi
package SCM
MediaWiki (via mod_php)
ansible
COPR and secondary arch builds
bodhi
accounts
package database
ticketing
…many others!

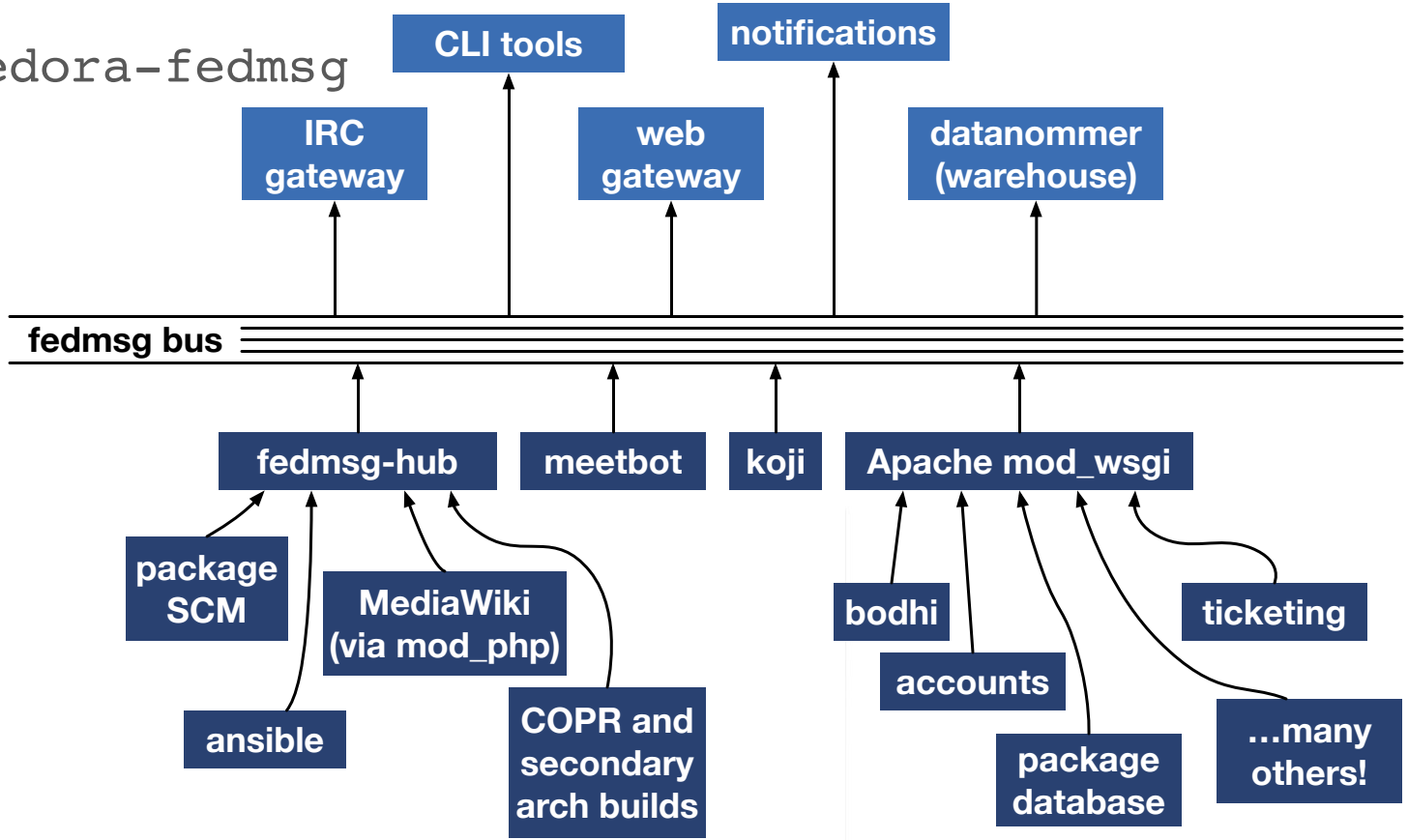
#fedora-fedmsg



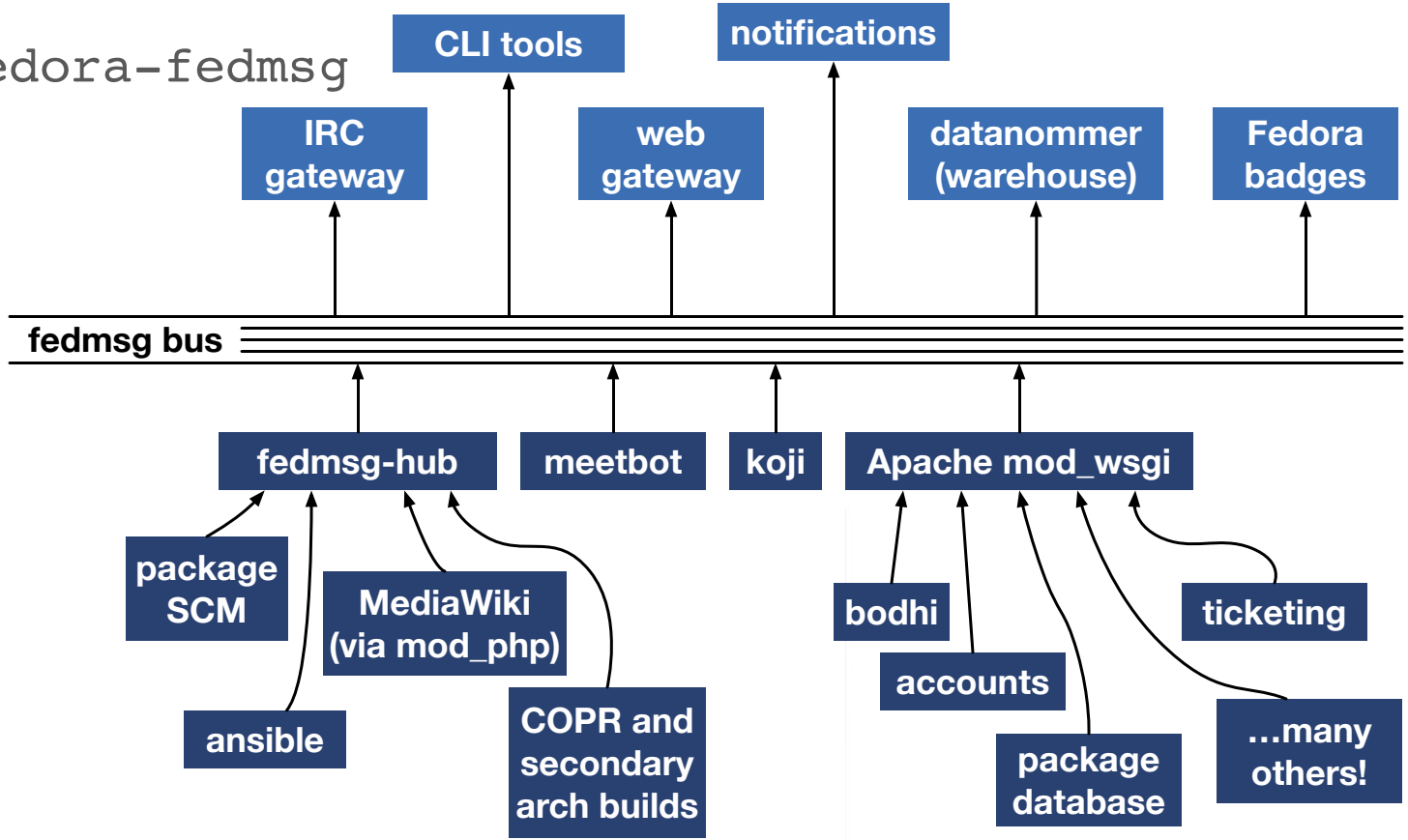
#fedora-fedmsg



#fedora-fedmsg



#fedora-fedmsg



fedora badges

About Explore Leaderboard Login

User Info

willb

Arrived on 2013-08-16.
Ranked 135 out of 15445 ranked users (top 0.9%).
Website: http://chapeau.freevariable.com
View user as: JSON, RSS, RDF

History


received on 2015-04-21 for this activity


received on 2015-04-21 for this activity


received on 2015-04-07 for this activity

Badges Earned

willb has earned **54** badges (**20.1%** of total).

Content Badges

Development Badges

Community Badges

Spark summit 2015

BACKGROUND

FEDORA MESSAGING

BULK DATA INGEST

ANALYZING COMMUNITY

NEXT STEPS

MESSAGE SOURCES

- Bulk historical data (PostgreSQL dump)
- Specific historical records (REST/JSON)
- Streaming messages (⌀MQ)

```
CREATE TABLE messages (
    id integer NOT NULL,
    i integer NOT NULL,
    "timestamp" timestamp NOT NULL,
    certificate text,
    signature text,
    topic text,
    _msg text NOT NULL,
    category text,
    source_name text,
    source_version text,
    msg_id text
);
```

```
CREATE TABLE

    i

    certificate
    signature
    topic
    _msg text NOT NULL,
    category
    source_name
    source_version
    msg_id
);
```

“I’LL JUST USE JSON!”

```
{
   "timestamp": "now",
   "you-have":
      [{"problems": 2}]
}
```

JSON AND SCHEMA INFERENCE

```
{
  "animals" : {
    "duck" :
      { "mammal": false, "bill": true, "eggs": true },
    "zebra" :
      { "mammal": true, "bill": false, "eggs": false },
    "platypus" :
      { "mammal": true, "bill": true, "eggs": true }
  }
}
```

JSON AND SCHEMA INFERENCE

```
{
  "animals" : [
    {"k": "duck",
     "v": { "mammal": false,
            "bill": true, "eggs": true } },
    {"k": "zebra",
     "v": { "mammal": true,
            "bill": false, "eggs": false },
    /* ... */
  ]
}
```

DIVERGING SCHEMAS

```
{ "branches" : ["f22", "master"] }

{ "branches" :
  [ { "name": "f22", "commit" : "05afffa1" } ]
}
```

PREPROCESSING WITH JSON4S

```
// use json4s and partial functions
// see the Silex library for an easy interface!
def renameBranches(msg: JValue) = {
 msg transformField {
  case JField("branches", v@JArray(JString(_)::_)) =>
   ("pkg_branches", v)
  case JField("branches", v@JArray(JObject(_)::_)) =>
   ("git_branches", v)
 }
}
```

BACKGROUND

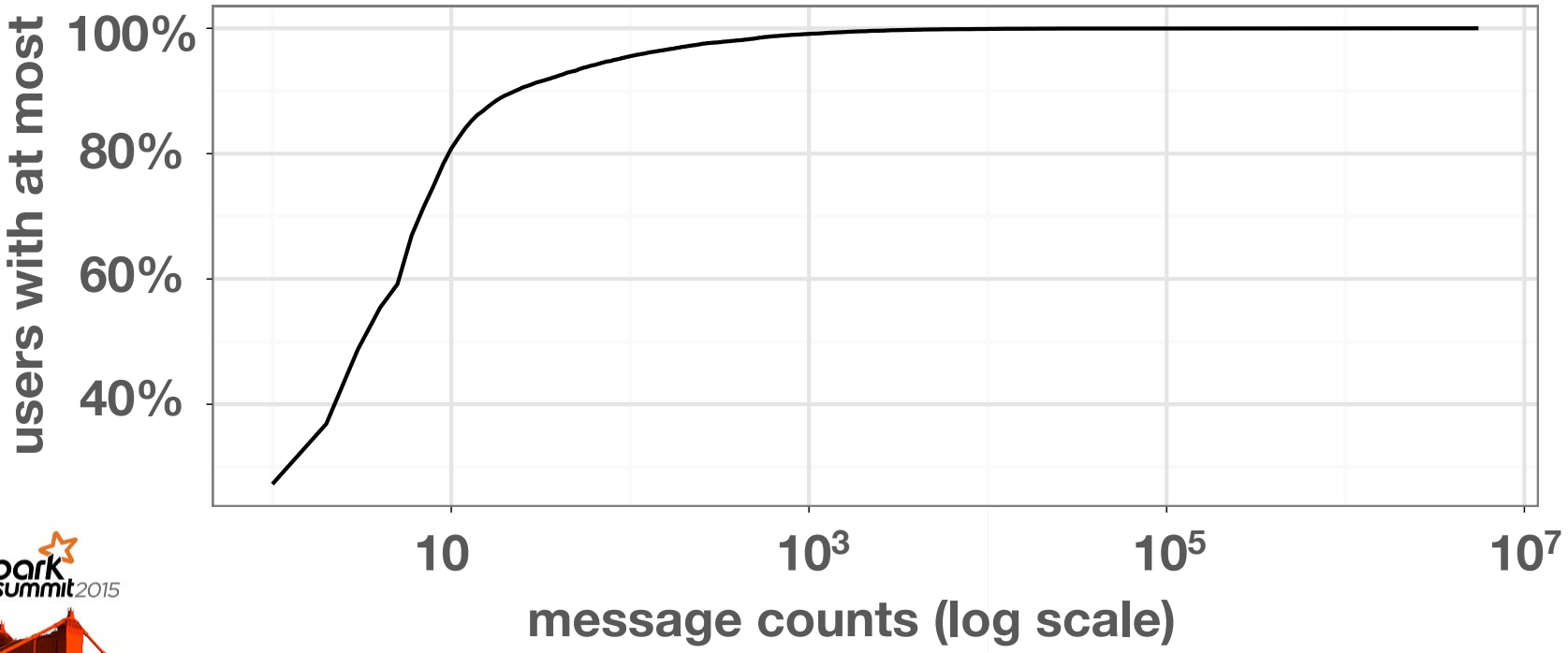
FEDORA MESSAGING

BULK DATA INGEST

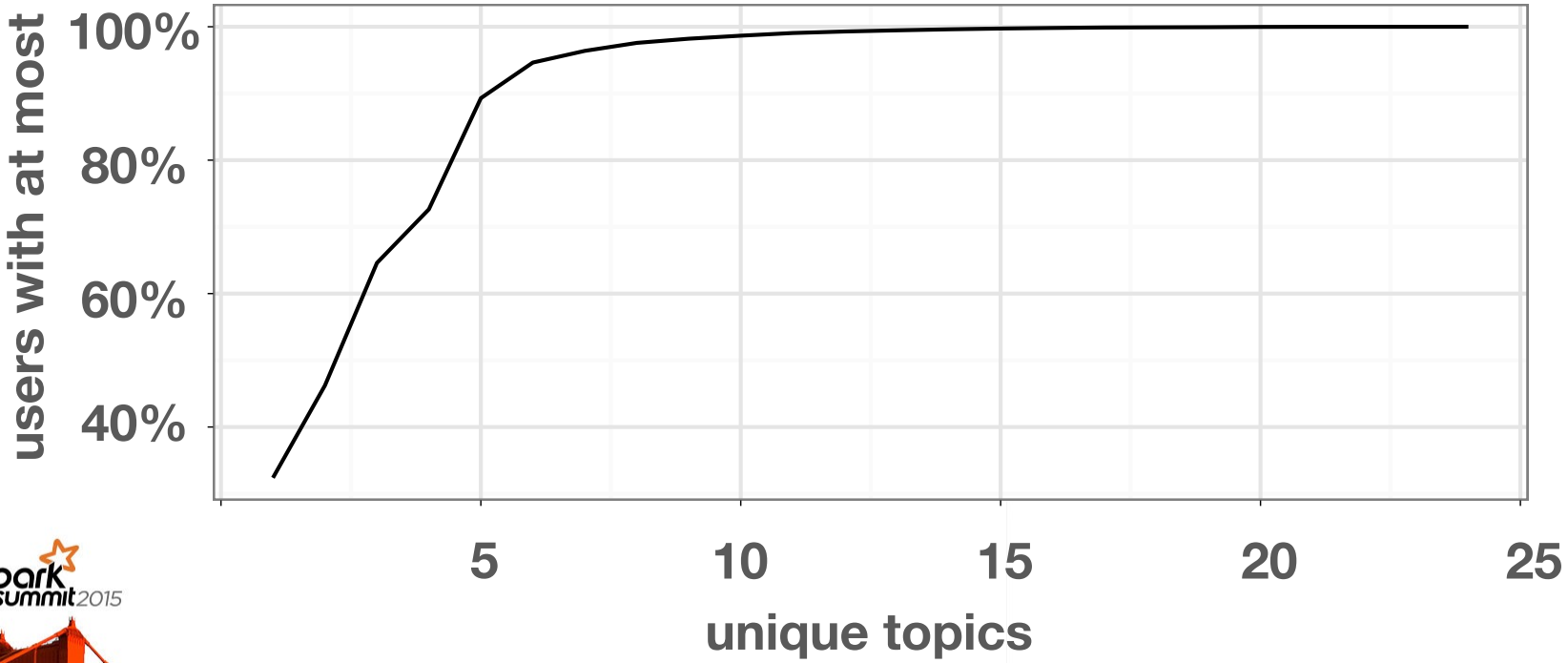
ANALYZING COMMUNITY

NEXT STEPS

MESSAGE COUNT DISTRIBUTION



UNIQUE TOPIC DISTRIBUTION



TOPIC SETS AS FEATURES

- Idea: characterize users by the sets of unique message topics affecting them
- Use ML Pipeline transformers to go from sets (modeled as arrays) to feature vectors

TOPIC SETS AS FEATURES

```
// expose this Spark-private DataType
package org.apache.spark.hacks {
   type VectorType =
     org.apache.spark.mllib.linalg.VectorUDT
}
```

TOPIC SETS AS FEATURES

```
trait SVParams extends Params {
  val inputCol =
    new Param[String](this, "inputCol", "...")
  val outputCol =
    new Param[String](this, "outputCol", "...")
  val vecSize =
    new IntParam(this, "vecSize", "...")

  // ...
}
```

TOPIC SETS AS FEATURES

```
trait SVParams extends Params {
 // ...

  def pvals(pm: ParamMap) = ( // 1.3-specific
    paramMap.get(inputCol).getOrElse("topicSet"),
    paramMap.get(outputCol).getOrElse("features"),
    paramMap.get(vecSize).getOrElse(128)
  )
}
```

```
class SetVectorizer(override val uid: String)
    extends Transformer with SVParams {
  // hopefully VectorUDT will be public soon!
  final val VT = new org.apache.spark.hacks.VectorType()

  def transformSchema(schema: StructType, params: ParamMap) = {
    val outc = paramMap.get(outputCol).getOrElse("features")
    StructType(schema.fields ++ Seq(StructField(outc, VT, true)))
  }

  def transform(df: DataFrame, params: ParamMap) = {
    val (inc, outc, vs) = pvals(paramMap)
    df.withColumn(outc, callUDF({ a: Seq[Int] =>
      Vectors.sparse(vs, a.toArray, Array.fill(a.size)(1.0))
    }, VT, df(inc)))
  }
}
```

CLASSIFYING FEDORA USERS

- Problem: can we identify Fedora packager sponsors by their message topic sets?
- Logistic regression was easy to try
- ~1% prediction error rate…but only ~10% of actual sponsors were predicted as such!

A BETTER APPROACH

Building a *decision tree* identified the following message topics as likely to predict that someone was a Fedora packager sponsor:

- `org.fedora.prod.fas.role.update`
- `org.fedora.prod.fas.user.update`

BACKGROUND

FEDORA MESSAGING

BULK DATA INGEST

ANALYZING COMMUNITY

NEXT STEPS

WHAT'S NEXT

- More about fedmsg: http://fedmsg.com
- Our Silex library includes helpers for JSON data ingest: http://silex.freevariable.com
- More details about this project: http://chapeau.freevariable.com[1]

[1] http://chapeau.freevariable.com/2015/06/summit-fedmsg.html

THANKS!