Hadoop gets Groovy

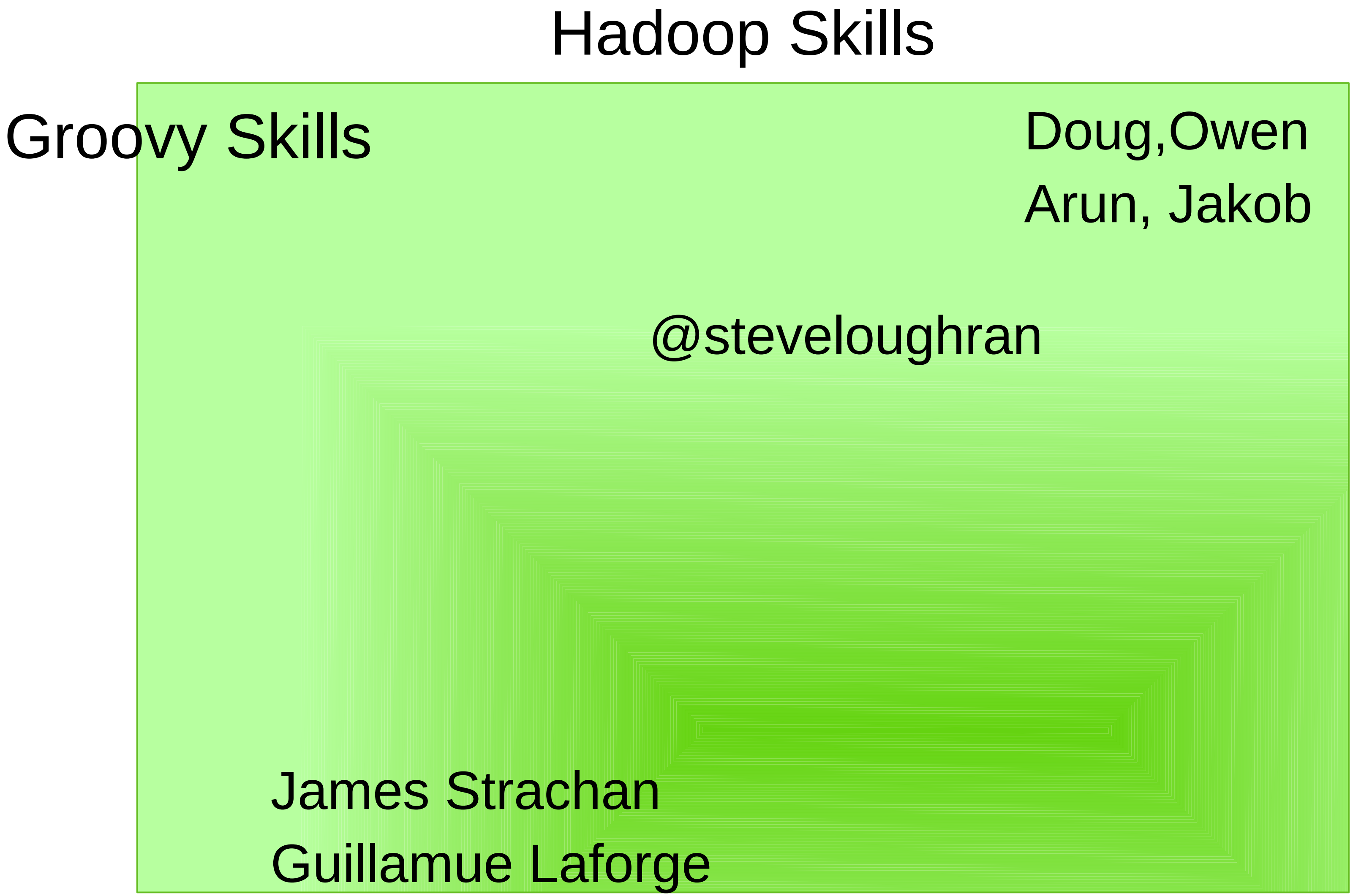
Steve Loughran– Hortonworks

stevel at hortonworks.com

@steveloughran

Berlin, June 2012

you in this diagram?



Hadoop Skills

Groovy Skills

Doug,Owen
Arun, Jakob

@steveloughran

James Strachan
Guillamue Laforge

Grumpy : Groovy Hadoop Library

Something lightweight for testing

Wanted to play in the M/R layer

Already using Groovy

Liked: JVM integration, tooling, libraries, IntelliJ IDEA, Books…

git@github.com:steveloughran/grumpy.git

What is Groovy?

A dynamic language within the JVM

Java++

- Maps, lists, tuples, Closures

Flavours of Ruby and Python

- 'Duck' typing, Grails, (Scripting)

A way to do things in the JVM that Sun didn't imagine

Can use & subclass java classes:

```
class LineCountMapper
  extends Mapper<LongWritable, Text, Text, IntWritable> {

static final def emitKey = new Text("lines")
static final def one = new IntWritable(1)

void map(LongWritable key,
           Text value,
           Mapper.Context context) {
    context.write(emitKey, one)
  }
}
```

Closures & lists

```
class CountReducer2 extends Reducer {

  def reduce(Text k,
            Iterable values,
            Reducer.Context ctx) {

    def sum = values.collect() {it.get() }.sum()

    ctx.write(k, new IntWritable(sum));
  }

}
```

Closures & lists

```
values.collect() {
    it.get()
  }.sum()
```

List<values> -> List<int> -> int

Result: MR jobs in Groovy

In:

```
gate1,b46cca4d3f5f313176e50a0e38e7fde3,,2006-10-30,16:06:17,Fleurball

gate1,f1191b79236083ce59981e049d863604,,2006-10-30,16:06:20,vklaptop

gate1,b45c7795f5be038dda8615ab44676872,,2006-10-30,16:06:21,Franky Panky

gate1,02e73779c77fcd4e9f90a193c4f3e7ff,,2006-10-30,16:06:23,

gate1,eef1836efddf8dbfe5e2a3cd5c13745f,,2006-10-30,16:06:24,Vas

gate1,b46cca4d3f5f313176e50a0e38e7fde3,,2006-10-30,16:06:32,Fleurball

gate1,f1191b79236083ce59981e049d863604,,2006-10-30,16:06:36,vklaptop

gate1,b45c7795f5be038dda8615ab44676872,,2006-10-30,16:06:37,Franky Panky

gate1,eef1836efddf8dbfe5e2a3cd5c13745f,,2006-10-30,16:06:38,Vas

gate1,02e73779c77fcd4e9f90a193c4f3e7ff,,2006-10-30,16:06:43,

gate1,2afaf990ce75f0a7208f7f012c8d12ad,,2006-10-30,16:06:54,Smiley
```

Out: 163,198,223 device sightings!

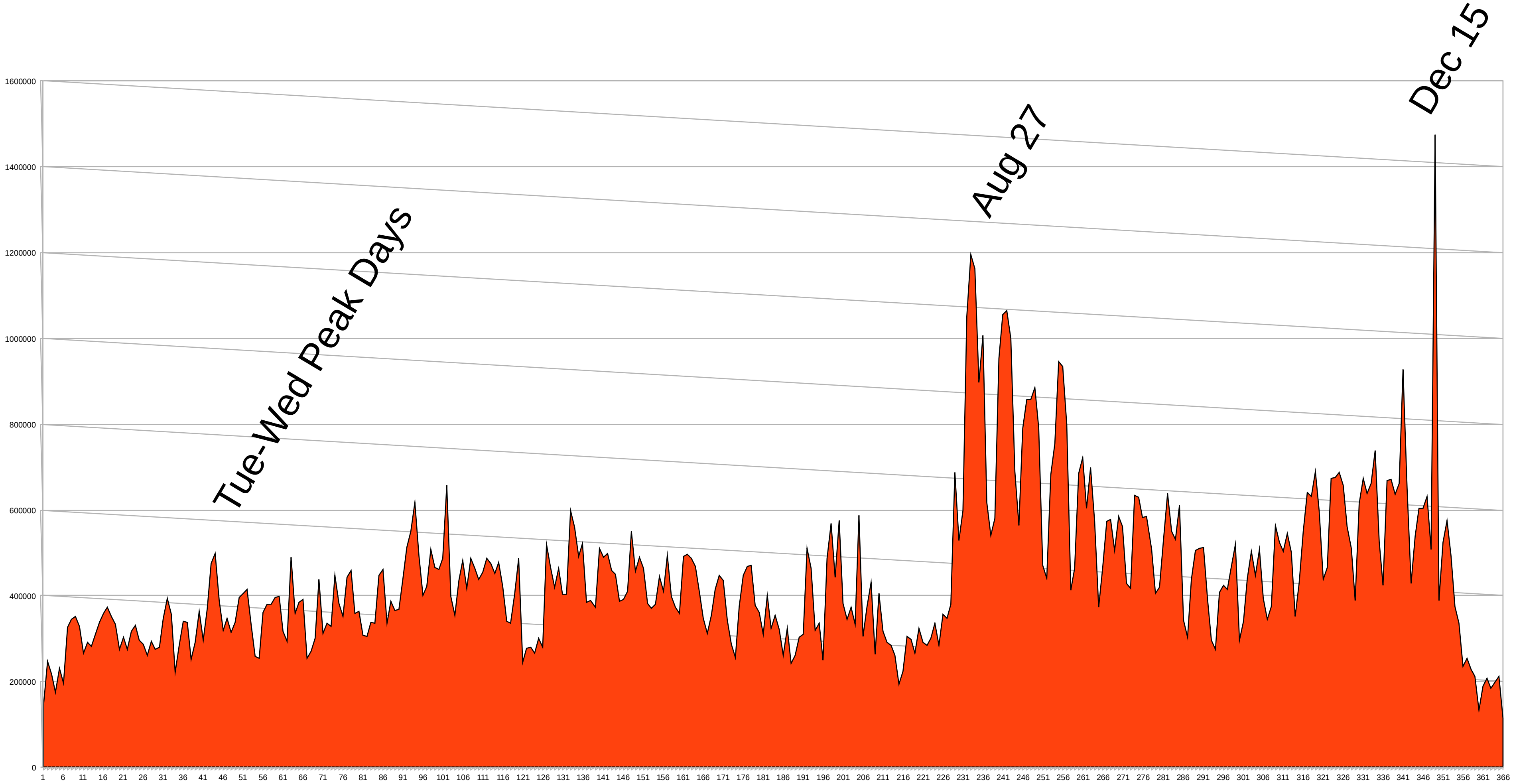
why no Pig? Sliding Window Debounce

```
void map(LongWritable key, BlueEvent event,
           Mapper.Context context) {

  BlueEvent ev2 = window.insert(event)
  List<BlueEvent> expired = window.purgeExpired(event)
  expired.each { evt ->
      emit(context, evt)
  }
}

void cleanup(Mapper.Context context) {
  window.each { evt ->
    emit(context, evt)
  }
}
```

Device sightings by day for 2007



Tue-Wed Peak Days

Aug 27

Dec 15

Improving Hadoop APIs

```
Configuration.metaClass.setAt = { key, val ->
 set(key.toString(), val.toString())
}

Configuration.metaClass.getAt = { key ->
  get(key)
}

Configuration.metaClass.add = {map ->
  map.each {elt ->
    set((elt.key).toString(),
        (elt.value).toString() )
}
```

& Configuration gets better

```
conf['mapscript'] = new File(src).text

String scriptText = conf['mapscript']

conf.add([
  window:60000,
  'redscript':reduceScript
  ])
```

Extending to Job class trickier –subclassing better

New today! script driven MR jobs!

```
protected void setup(Mapper.Context ctx) {
  this.ctx = ctx
  this.conf = ctx.configuration
  ScriptCompiler comp = new ScriptCompiler(conf)
  String scriptText = conf['mapscript']
  map = comp.parse(scriptText, this, ctx)
 }

 protected void map(Writable key, Writable value,
   Mapper.Context ctx) {
  map.setProperty('key',key)
  map.setProperty('value',value)
  map.run()
 }
```

Things to consider

Performance: Groovy 2 on Java7

'False friends' -Types, if(), exceptions

If you can use Pig, use it.

Use Groovy for testing, extending Hadoop classes (output formatter, etc)

Play with YARN and Giraph with it

Questions?

hortonworks.com

hortonworks.com

Performance?

Groovy 1 over-introspects

HLL hides a lot of overhead

If your work is I/O bound, less important

Speed of development vs execution

Need to benchmark on Java 7