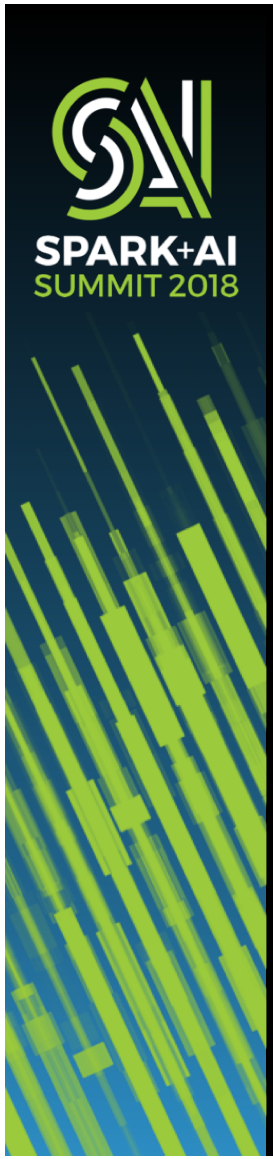


Which Data Broke My Code?

Inspecting Spark Transformations

Vinod K. Nair, Director of Product Management @ Pepperdata

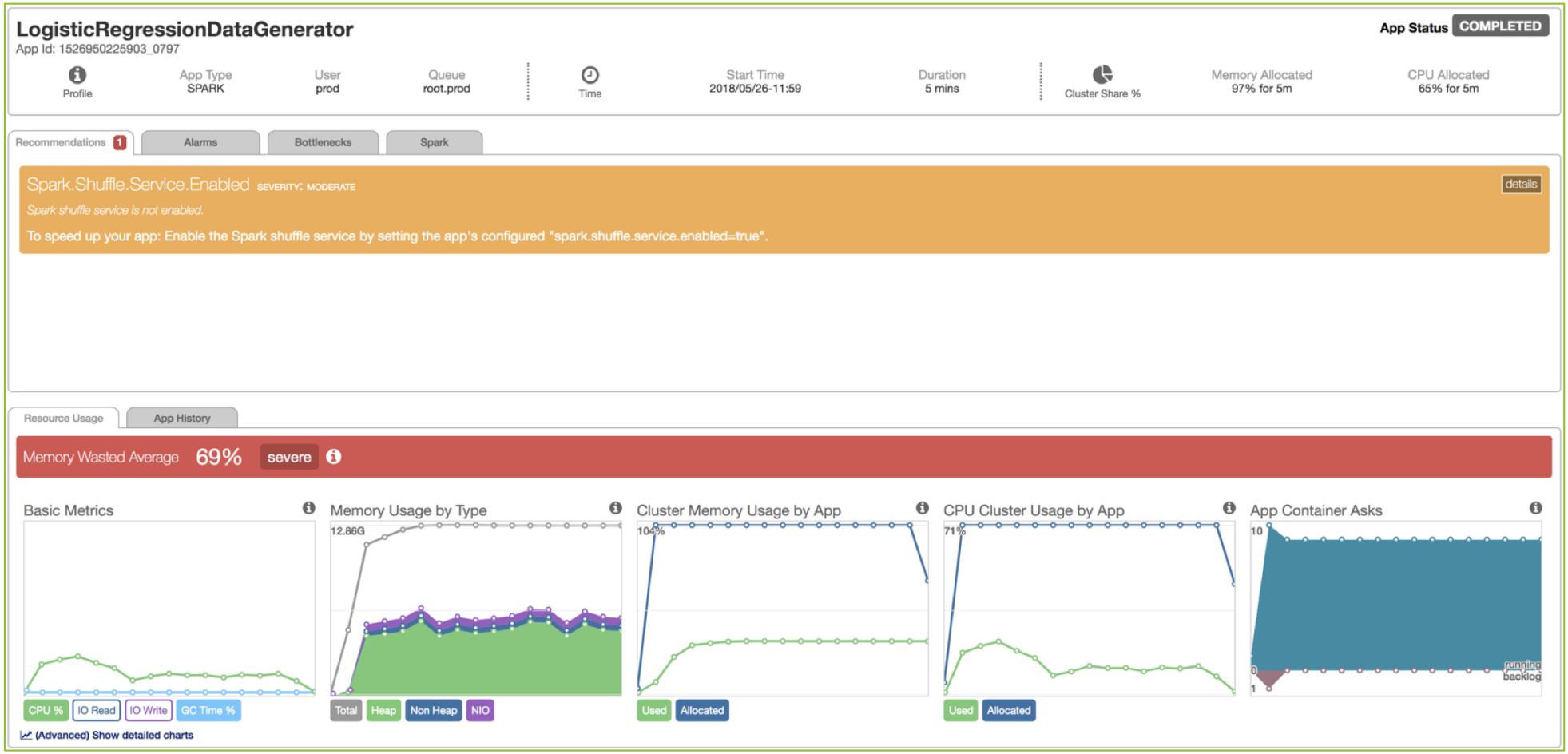
#DevSAIS12

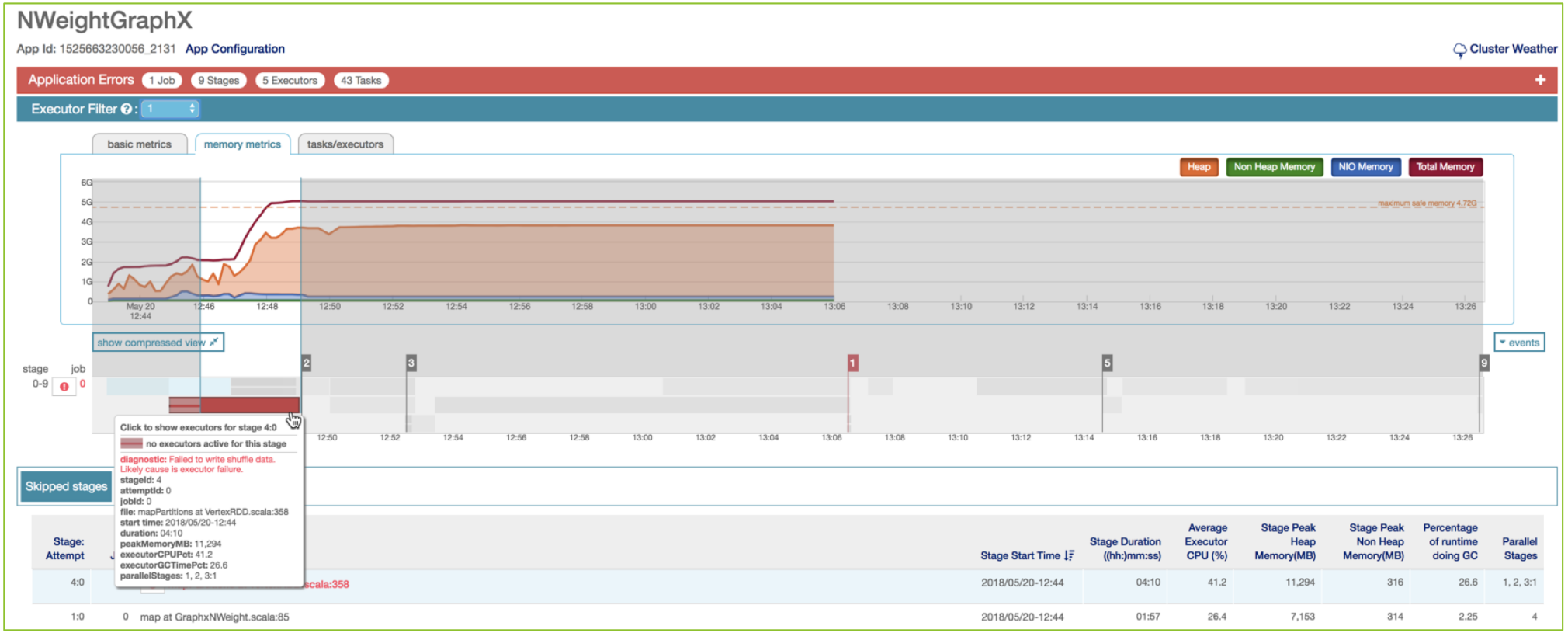
Talk Outline

- Introduction
- Problem: ‘laziness’ makes debugging hard
- Solution: interactive inspection of RDDs
- Demo
- Q & A

Introduction to Pepperdata

Application Performance Management (APM) for Spark (& Hadoop)


LogisticRegressionDataGenerator
App Id: 1526950225903_0797
App Status COMPLETED
Profile
App Type SPARK
User prod
Queue root.prod
Time
Start Time 2018/05/26-11:59
Duration 5 mins
Cluster Share %
Memory Allocated 97% for 5m
CPU Allocated 65% for 5m
Recommendations 1
Alarms
Bottlenecks
Spark
Spark.Shuffle.Service.Enabled SEVERITY: MODERATE
details
Spark shuffle service is not enabled.
To speed up your app: Enable the Spark shuffle service by setting the app's configured "spark.shuffle.service.enabled=true".
Resource Usage
App History
Memory Wasted Average 69% severe
Basic Metrics
CPU % IO Read IO Write GC Time %
Memory Usage by Type
12.86G
Total Heap Non Heap NIO
Cluster Memory Usage by App
104%
Used Allocated
CPU Cluster Usage by App
71%
Used Allocated
App Container Asks
10
0
1
running
backlog
(Advanced) Show detailed charts


NWeightGraphX
App Id: 1525663230056_2131 App Configuration
Cluster Weather
Application Errors 1 Job 9 Stages 5 Executors 43 Tasks
Executor Filter : 1
basic metrics
memory metrics
tasks/executors
Heap
Non Heap Memory
NIO Memory
Total Memory
maximum safe memory 4.72G
show compressed view
events
stage
job
Click to show executors for stage 4:0
no executors active for this stage
diagnostic: Failed to write shuffle data. Likely cause is executor failure.
stageId: 4
attemptId: 0
jobId: 0
file: mapPartitions at VertexRDD.scala:358
start time: 2018/05/20-12:44
duration: 04:10
peakMemoryMB: 11,294
executorCPUPct: 41.2
executorGCTimePct: 26.6
parallelStages: 1, 2, 3:1
Skipped stages
Stage: Attempt
Stage Start Time
Stage Duration ((hh:)mm:ss)
Average Executor CPU (%)
Stage Peak Heap Memory(MB)
Stage Peak Non Heap Memory(MB)
Percentage of runtime doing GC
Parallel Stages
4:0
2018/05/20-12:44
04:10
41.2
11,294
316
26.6
1, 2, 3:1
1:0
0
map at GraphxNWeight.scala:85
2018/05/20-12:44
01:57
26.4
7,153
314
2.25
4

Problem: ‘laziness’ makes debugging hard

RDD data unavailable until an ‘action’ triggers execution

Transformations are invisible

RDDs support two types of operations:

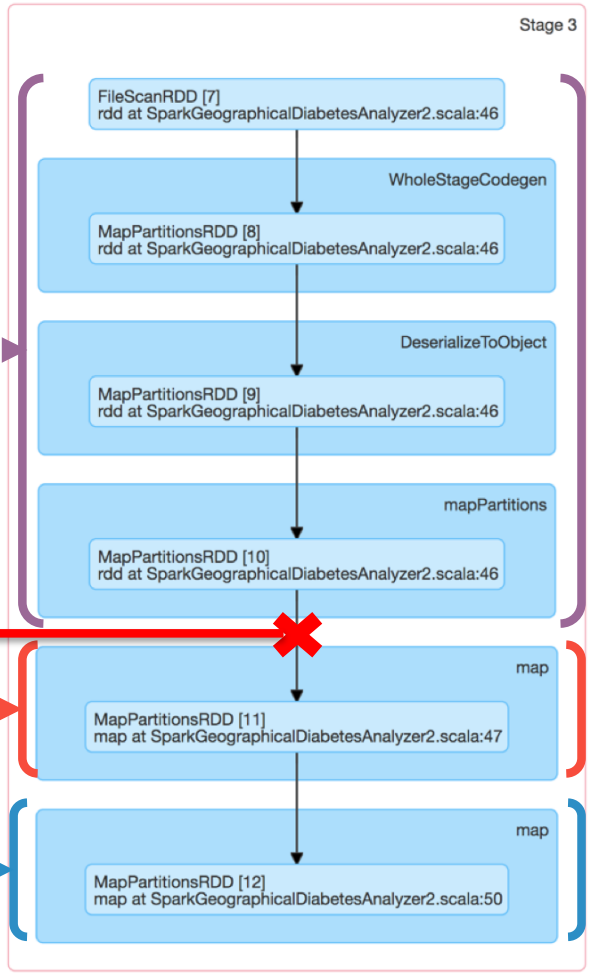
1. **transformations**, which create a new dataset from an existing one
2. **actions**, which return a value to the driver program after running a computation on the dataset

Transformations in Spark are ***lazy***. They are only computed when an action requires a result to be returned to the driver program.

- https://spark.apache.org/docs/latest/rdd-programming-guide.html

```
36  val healthData = spark.read
37    .option("header", "true") // Use first line of all files as header
38    .option("inferSchema", "true") // Automatically infer data types
39    .csv("hdfs:///tmp/health_data.csv")
40
41  // Filter to the relevant data aggregated by city, discard unused columns
42  val columnsToKeep = Seq("StateDesc", "Data_Value", "PopulationCount")
43  val filteredData = healthData.filter(healthData("GeographicLevel").===("City"))
44    .filter(healthData("DataValueTypeID").===("CrdPrv"))
45    .filter(healthData("MeasureId").===("DIABETES"))
46
47
48
49  // Map the state to geographical region and convert the percentage to count of impacted people
50  val regionMappedImpactedData = filteredData.map(x => (getMappedRegionOrUnknown(x._1), x._2))
51  val reducedRegionMappedImpactedData = regionMappedImpactedData.reduceByKey(_ + _).sortByKey()
52
```

filtererdData.take(10).foreach(println)



Solution available today

- Sprinkle your code with print statements
- ‘hopefully’ catch the right transformation causing the problem
- If you don’t catch it – repeat process

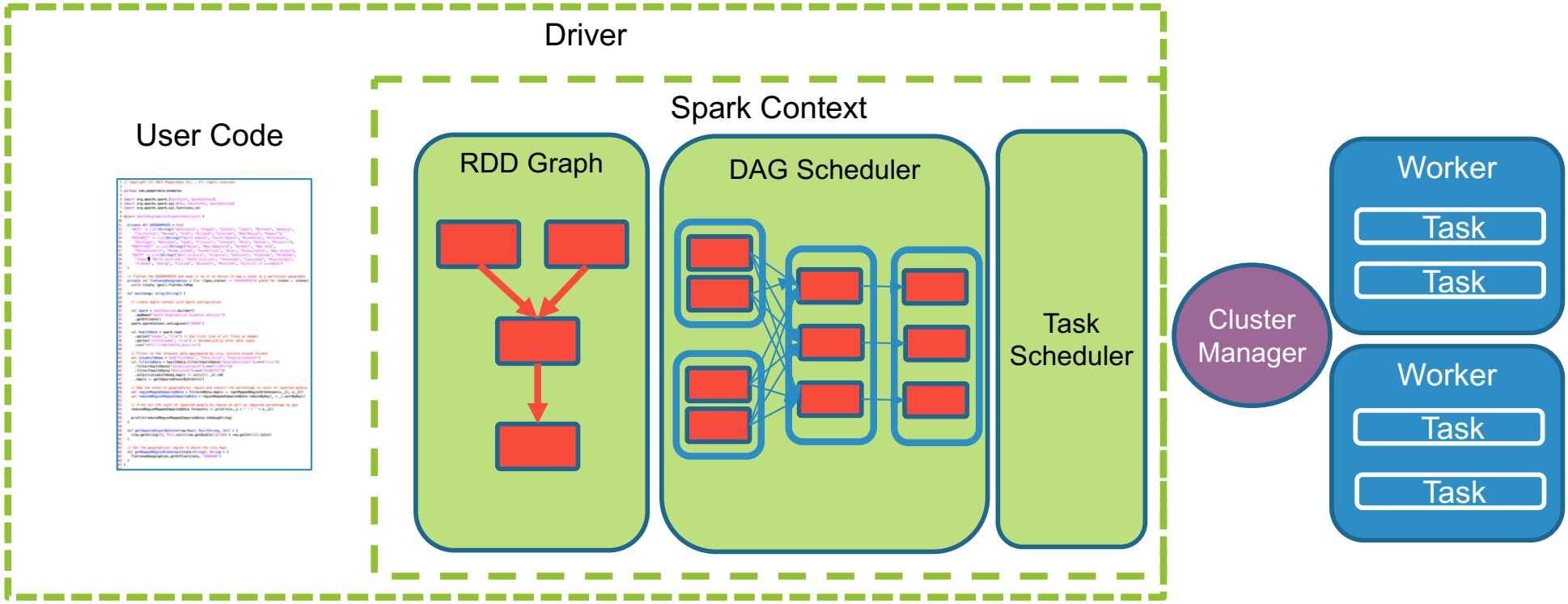
Our solution: interactive inspection of data in flight

Trigger an ‘action’ to enable inspection of any RDD in the DAG

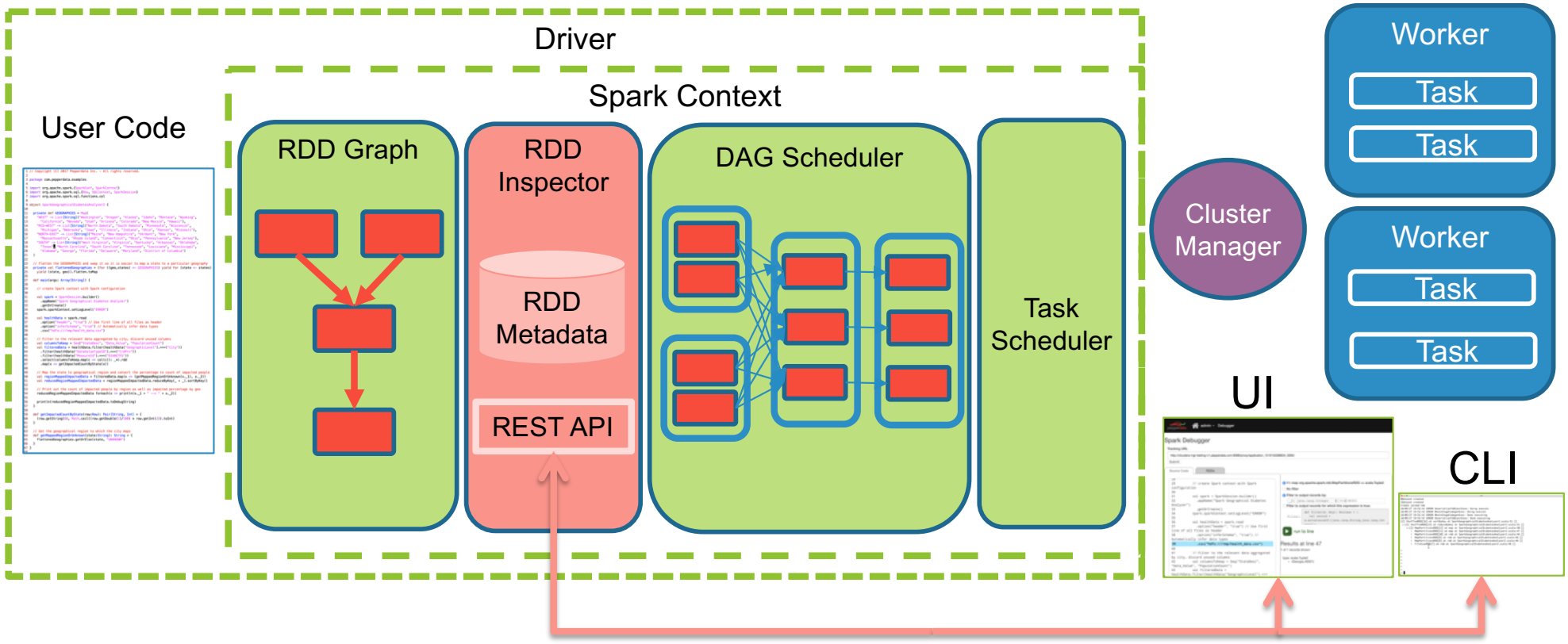
Solution requirements

- No user code changes required
- Work with any standard Spark distribution
- Provide a familiar interactive debugger interface

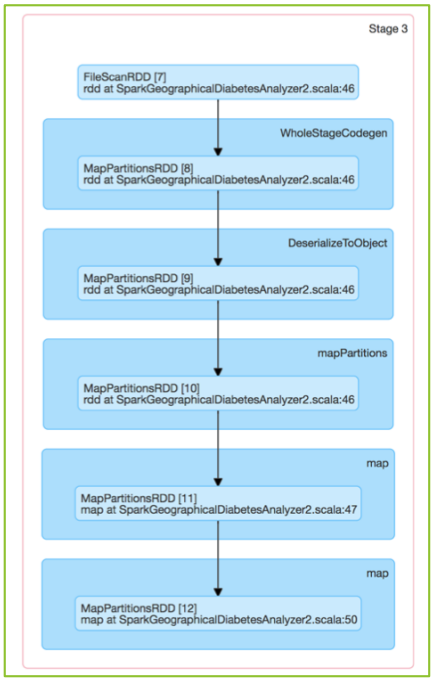
Solution overview



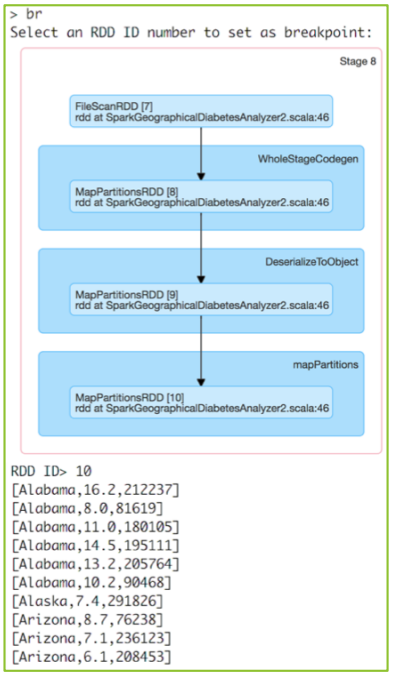
Solution overview



New Spark job to display RDD[10]

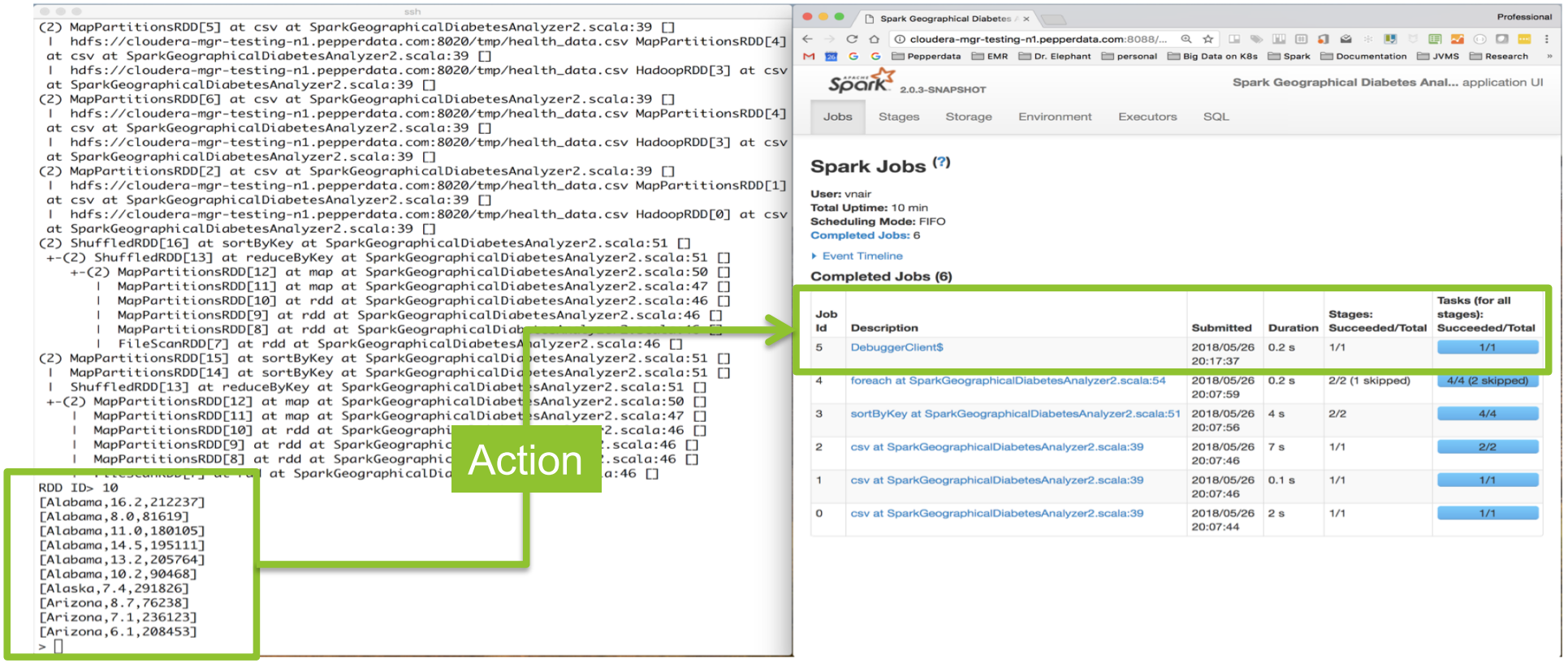


Original stage with RDD [10] as an intermediate transformation



Debug stage with RDD [10] as output

CLI command to inspect RDD[10]

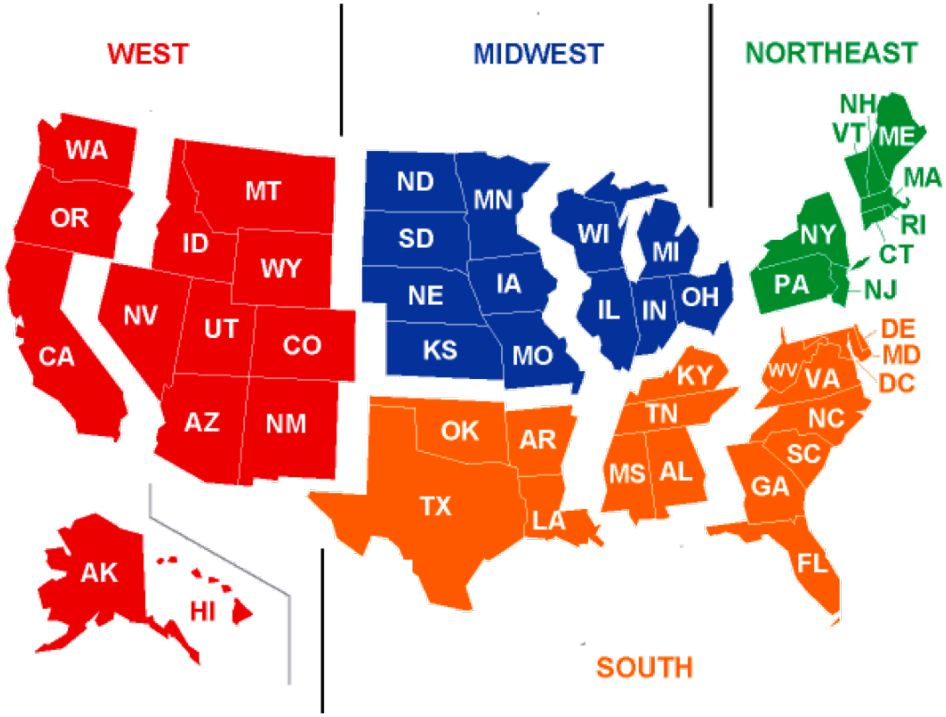


Demo

Interactive ‘debugger’ experience to inspect any RDD in the DAG

Analyze CDC's census data

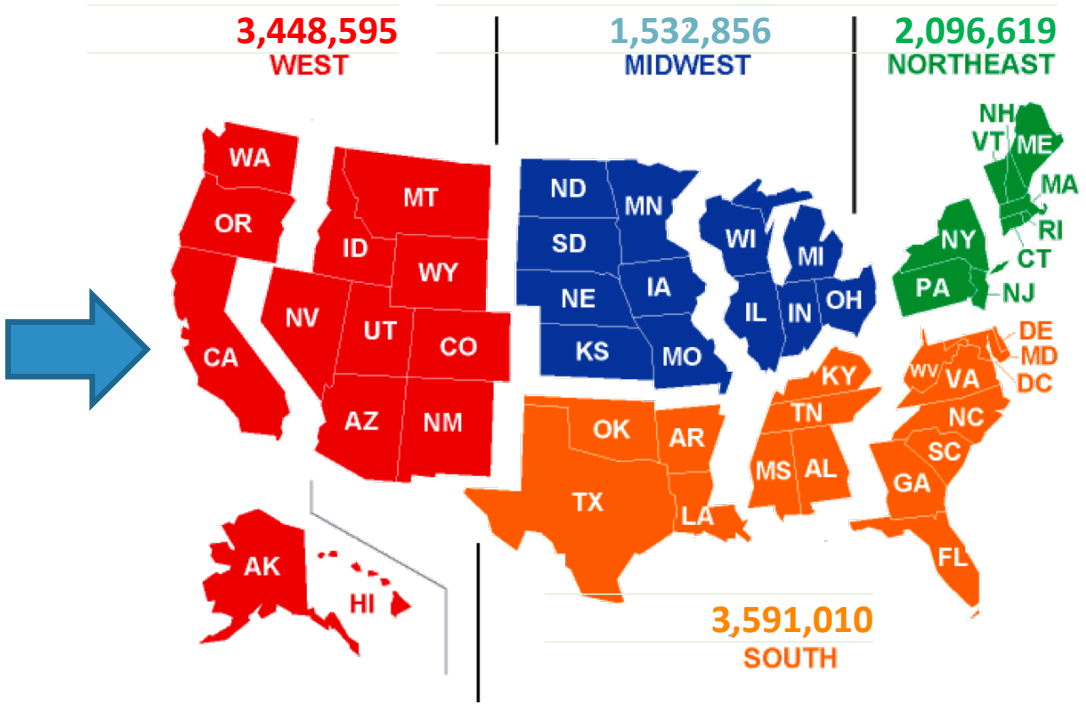
- CDC's 500 cities project
  (data.gov: 500 cities local data for better health)
- Input:
  Major chronic diseases by city
- Output:
  Diabetes in adults by US region



Distribution of diabetics by region

500_Cities__Local_Data_for_Better_Health.csv

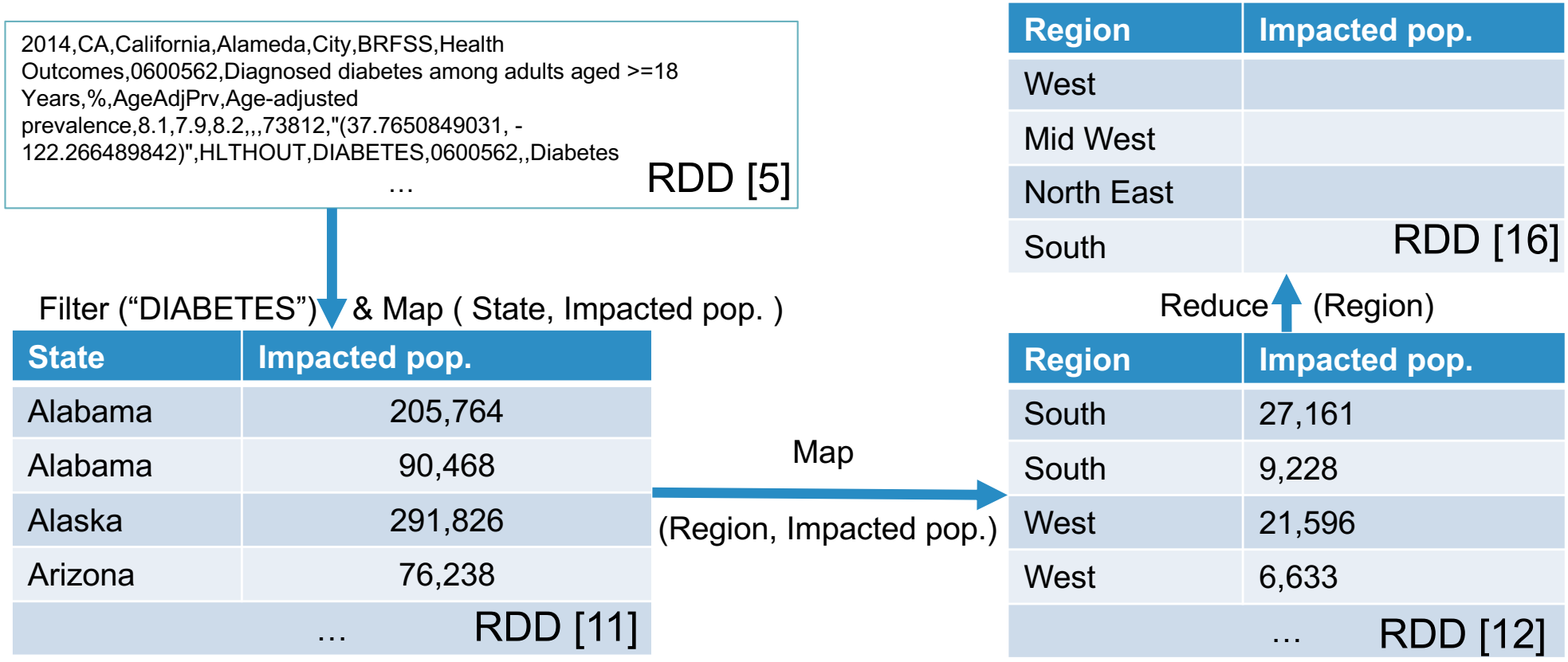
```
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010101,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,,,,*,Estimates suppressed for population less than 50,3,"(41.60035282, -90.4980
671216)",HLTHOUT,CANCER,1919000,19163010101,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010202,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,,,,*,Estimates suppressed for population less than 50,23,"(41.6051687821, -90.5
841494896)",HLTHOUT,CANCER,1919000,19163010202,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010300,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,,,,*,Estimates suppressed for population less than 50,9,"(41.6040401047, -90.67
77509264)",HLTHOUT,CANCER,1919000,19163010300,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010600,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,3.7,3.6,3.9,,,2422,"(41.5268094476, -90.5656367434)",HLTHOUT,CANCER,1919000,191
63010600,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010700,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,4.5,4.4,4.7,,,1527,"(41.5293018757, -90.5806944595)",HLTHOUT,CANCER,1919000,191
63010700,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010800,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,6.4,6.1,6.7,,,3277,"(41.5280925473, -90.5944178646)",HLTHOUT,CANCER,1919000,191
63010800,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163010900,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,4.9,4.7,5.1,,,1970,"(41.5194046979, -90.5858947174)",HLTHOUT,CANCER,1919000,191
63010900,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163011000,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,5.7,5.4,6.0,,,2725,"(41.5047261649, -90.6115509526)",HLTHOUT,CANCER,1919000,191
63011000,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163011100,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,6.9,6.5,7.3,,,3272,"(41.5355053826, -90.6078471834)",HLTHOUT,CANCER,1919000,191
63011100,Cancer (except skin)
2014,IA,Iowa,Davenport,Census Tract,BRFSS,Health Outcomes,1919000-19163011200,Can
cer (excluding skin cancer) among adults aged >=18 Years,%,CrdPrv,Crude prevalenc
e,4.8,4.6,5.1,,,2281,"(41.5367571624, -90.5962736008)",HLTHOUT,CANCER,1919000,191
63011200,Cancer (except
```



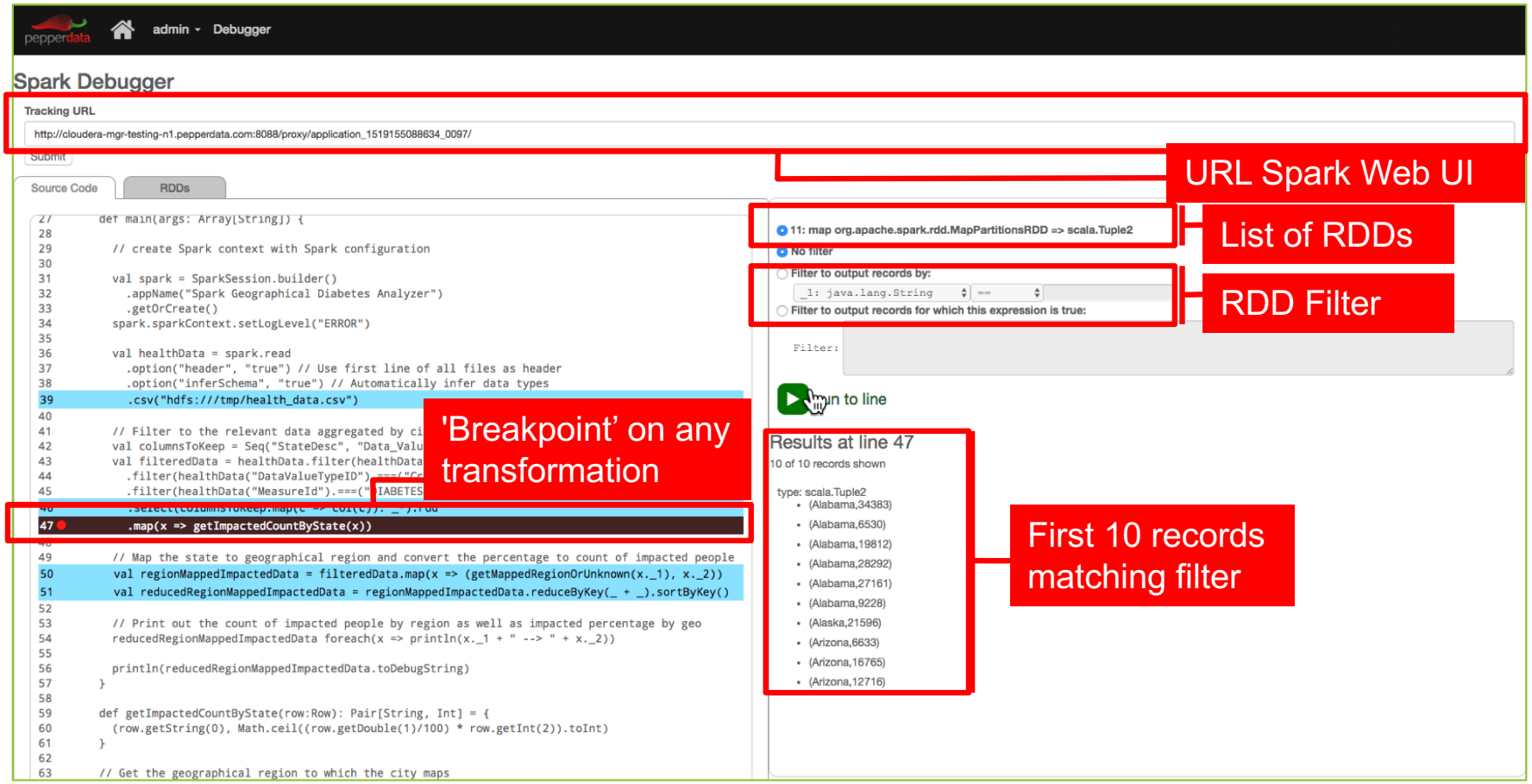
```
ssh
-bash-4.2$ /usr/local/spark/bin/spark-submit --executor-memory 512m --deploy-mode client --class com.pepperdata.exampl
es.SparkGeographicalDiabetesAnalyzer2 ~/cityaggregator-0.2.jar
```

Region	Adult Population with Diabetes
WEST	3,448,595
MID WEST	1,532,856
NORTH EAST	2,096,619
SOUTH	3,407,528
UNKNOWN	183,482

RDD transformations through the app



Spark Debugger UI



Spark Debugger Demo

To recap, you can…

- View data sets as they are transformed in your Spark App
  - no code changes are required
  - it works with any Spark distribution, and
  - it uses a familiar debugger interface to ‘set’ breakpoints and view RDDs in flight

What's next ?

Roadmap for Pepperdata's Spark Debugger

Areas of focus going forward

- UX Improvements
- Attach to a running app (streaming use case)
- Pause a job on hitting a condition
- Spark SQL support

To learn more visit the 

booth (#407)