MacroBase: Efficient Explanation On Big Data

Peter Kraft,
Firas Abuzaid, Sahaana Suri, Edward Gan,
Peter Bailis, Matei Zaharia, and many others

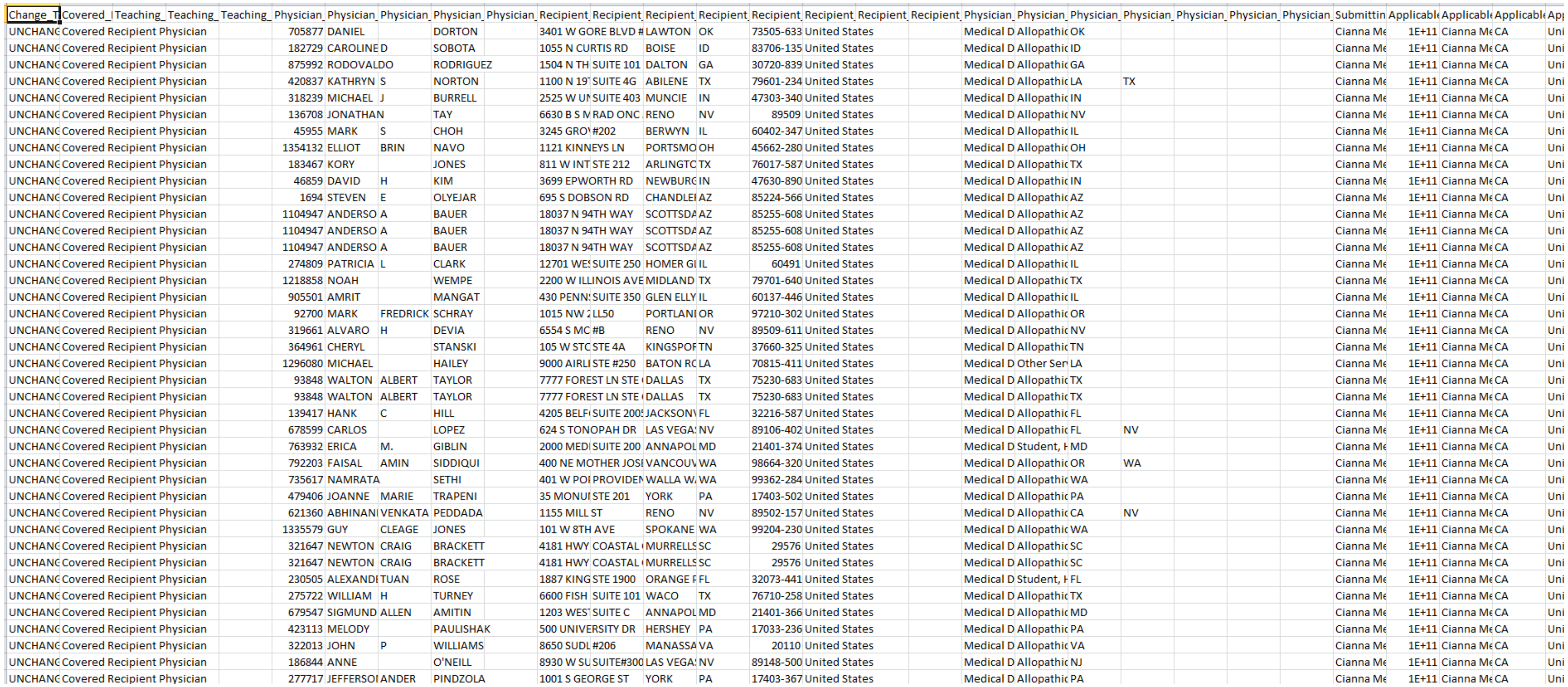
Explaining your data

Problem:
My users are complaining my app crashes a lot.

Explaining your data

Better question:
Which factors in my logs seem related to the crashes?

Explaining your Data



Change_T	Covered_	Teaching_	Teaching_	Teaching_	Physician_	Physician_	Physician_	Physician_	Physician_	Recipient_	Recipient_	Recipient_	Recipient_	Recipient_	Recipient_	Recipient_	Recipient_	Physician_	Physician_	Physician_	Physician_	Physician_	Physician_	Physician_	Submittin	Applicable	Applicable	Applicable	App
UNCHANG	Covered Recipient Physician				705877	DANIEL		DORTON		3401 W GORE BLVD #		LAWTON	OK	73505-633	United States			Medical D	Allopathic	OK					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				182729	CAROLINE	D	SOBOTA		1055 N CURTIS RD		BOISE	ID	83706-135	United States			Medical D	Allopathic	ID					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				875992	RODOVALDO		RODRIGUEZ		1504 N TH	SUITE 101	DALTON	GA	30720-839	United States			Medical D	Allopathic	GA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				420837	KATHRYN	S	NORTON		1100 N 19T	SUITE 4G	ABILENE	TX	79601-234	United States			Medical D	Allopathic	LA	TX				Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				318239	MICHAEL	J	BURRELL		2525 W UI	SUITE 403	MUNCIE	IN	47303-340	United States			Medical D	Allopathic	IN					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				136708	JONATHAN		TAY		6630 B S M	RAD ONC .	RENO	NV	89509	United States			Medical D	Allopathic	NV					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				45955	MARK	S	CHOH		3245 GRO	#202	BERWYN	IL	60402-347	United States			Medical D	Allopathic	IL					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1354132	ELLIOT	BRIN	NAVO		1121 KINNEYS LN		PORTSMO	OH	45662-280	United States			Medical D	Allopathic	OH					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				183467	KORY		JONES		811 W INT	STE 212	ARLINGTC	TX	76017-587	United States			Medical D	Allopathic	TX					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				46859	DAVID	H	KIM		3699 EPWORTH RD		NEWBURG	IN	47630-890	United States			Medical D	Allopathic	IN					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1694	STEVEN	E	OLYEJAR		695 S DOBSON RD		CHANDLEI	AZ	85224-566	United States			Medical D	Allopathic	AZ					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1104947	ANDERSO	A	BAUER		18037 N 94TH WAY		SCOTTSDA	AZ	85255-608	United States			Medical D	Allopathic	AZ					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1104947	ANDERSO	A	BAUER		18037 N 94TH WAY		SCOTTSDA	AZ	85255-608	United States			Medical D	Allopathic	AZ					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1104947	ANDERSO	A	BAUER		18037 N 94TH WAY		SCOTTSDA	AZ	85255-608	United States			Medical D	Allopathic	AZ					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				274809	PATRICIA	L	CLARK		12701 WE	SUITE 250	HOMER GI	IL	60491	United States			Medical D	Allopathic	IL					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1218858	NOAH		WEMPE		2200 W ILLINOIS AVE		MIDLAND	TX	79701-640	United States			Medical D	Allopathic	TX					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				905501	AMRIT		MANGAT		430 PENN	SUITE 350	GLEN ELLY	IL	60137-446	United States			Medical D	Allopathic	IL					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				92700	MARK	FREDRICK	SCHRAY		1015 NW 2	LL50	PORTLANI	OR	97210-302	United States			Medical D	Allopathic	OR					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				319661	ALVARO	H	DEVIA		6554 S MC	#B	RENO	NV	89509-611	United States			Medical D	Allopathic	NV					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				364961	CHERYL		STANSKI		105 W STC	STE 4A	KINGSPOF	TN	37660-325	United States			Medical D	Allopathic	TN					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1296080	MICHAEL		HAILEY		9000 AIRLI	STE #250	BATON RC	LA	70815-411	United States			Medical D	Other Ser	LA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				93848	WALTON	ALBERT	TAYLOR		7777 FOREST LN STE (		DALLAS	TX	75230-683	United States			Medical D	Allopathic	TX					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				93848	WALTON	ALBERT	TAYLOR		7777 FOREST LN STE (		DALLAS	TX	75230-683	United States			Medical D	Allopathic	TX					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				139417	HANK	C	HILL		4205 BELF	SUITE 200	JACKSONV	FL	32216-587	United States			Medical D	Allopathic	FL					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				678599	CARLOS		LOPEZ		624 S TONOPAH DR		LAS VEGA	NV	89106-402	United States			Medical D	Allopathic	FL	NV				Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				763932	ERICA	M.	GIBLIN		2000 MEDI	SUITE 200	ANNAPOL	MD	21401-374	United States			Medical D	Student, H	MD					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				792203	FAISAL	AMIN	SIDDIQUI		400 NE MOTHER JOSI		VANCOUV	WA	98664-320	United States			Medical D	Allopathic	OR	WA				Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				735617	NAMRATA		SETHI		401 W POI	PROVIDEN	WALLA W.	WA	99362-284	United States			Medical D	Allopathic	WA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				479406	JOANNE	MARIE	TRAPENI		35 MONUI	STE 201	YORK	PA	17403-502	United States			Medical D	Allopathic	PA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				621360	ABHINANI	VENKATA	PEDDADA		1155 MILL ST		RENO	NV	89502-157	United States			Medical D	Allopathic	CA	NV				Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				1335579	GUY	CLEAGE	JONES		101 W 8TH AVE		SPOKANE	WA	99204-230	United States			Medical D	Allopathic	WA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				321647	NEWTON	CRAIG	BRACKETT		4181 HWY	COASTAL (	MURRELLS	SC	29576	United States			Medical D	Allopathic	SC					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				321647	NEWTON	CRAIG	BRACKETT		4181 HWY	COASTAL (	MURRELLS	SC	29576	United States			Medical D	Allopathic	SC					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				230505	ALEXANDI	TUAN	ROSE		1887 KING	STE 1900	ORANGE P	FL	32073-441	United States			Medical D	Student, H	FL					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				275722	WILLIAM	H	TURNEY		6600 FISH	SUITE 101	WACO	TX	76710-258	United States			Medical D	Allopathic	TX					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				679547	SIGMUND	ALLEN	AMITIN		1203 WEST	SUITE C	ANNAPOL	MD	21401-366	United States			Medical D	Allopathic	MD					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				423113	MELODY		PAULISHAK		500 UNIVERSITY DR		HERSHEY	PA	17033-236	United States			Medical D	Allopathic	PA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				322013	JOHN	P	WILLIAMS		8650 SUDL	#206	MANASSA	VA	20110	United States			Medical D	Allopathic	VA					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				186844	ANNE		O'NEILL		8930 W SU	SUITE#300	LAS VEGA	NV	89148-500	United States			Medical D	Allopathic	NJ					Cianna Me	1E+11	Cianna Me	CA	Uni
UNCHANG	Covered Recipient Physician				277717	JEFFERSOI	ANDER	PINDZOLA		1001 S GEORGE ST		YORK	PA	17403-367	United States			Medical D	Allopathic	PA					Cianna Me	1E+11	Cianna Me	CA	Uni

I don't know what's going on here, and neither do you.

Explaining your Data

Version	Country	Support	Ratio
4.1	France	0.50	4.0

Here, Version 4.1 in French is the biggest difference between crashing and successful logs.

Explanations

Attribute 1	Attribute 2	Attribute 3	Metric 1	Metric 2
Value 1	Value 2	Value 3	X.XX	Y.YY
…	…	…	…	…

MacroBase returns explanations, sets of attributes that explain difference between two datasets.

What is MacroBase?

- MacroBase introduces a new SQL operator, DIFF.
- DIFF helps you find differences between data.
- DIFF is implemented using Spark and Spark-SQL.

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

Here are some logs. Let's figure out what's causing the crashes

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

???

Version	Country	Support	Ratio
4.1	France	0.5	4.0

We want an *explanation*—some attributes that are correlated with the crashes and metrics to tell us how.

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

DIFF

Let's make a new SQL operator that finds explanations!

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

```
DIFF TableCrash, TableSuccess
 ON Version, Carrier, Country
```

DIFF compares the crashing and successful logs. But how does it know how?

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

```
DIFF TableCrash, TableSuccess
ON Version, Carrier, Country
COMPARE BY
```

We need to give DIFF a list of rules.

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

```
DIFF TableCrash, TableSuccess
 ON Version, Carrier, Country
COMPARE BY
risk_ratio(COUNT(*)) > 3.0
```

We want attributes that occur much more often in crashing runs than in non-crashing runs

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

```
DIFF TableCrash, TableSuccess
 ON Version, Carrier, Country
        COMPARE BY
risk_ratio(COUNT(*)) > 3.0
 support(COUNT(*)) > 0.05;
```

We also want those attributes to have high sample sizes, or we'll just get low-sample size noise

DIFF operator

TableSuccess

Version	Carrier	Country
4.0	Verizon	USA
4.0	Verizon	USA

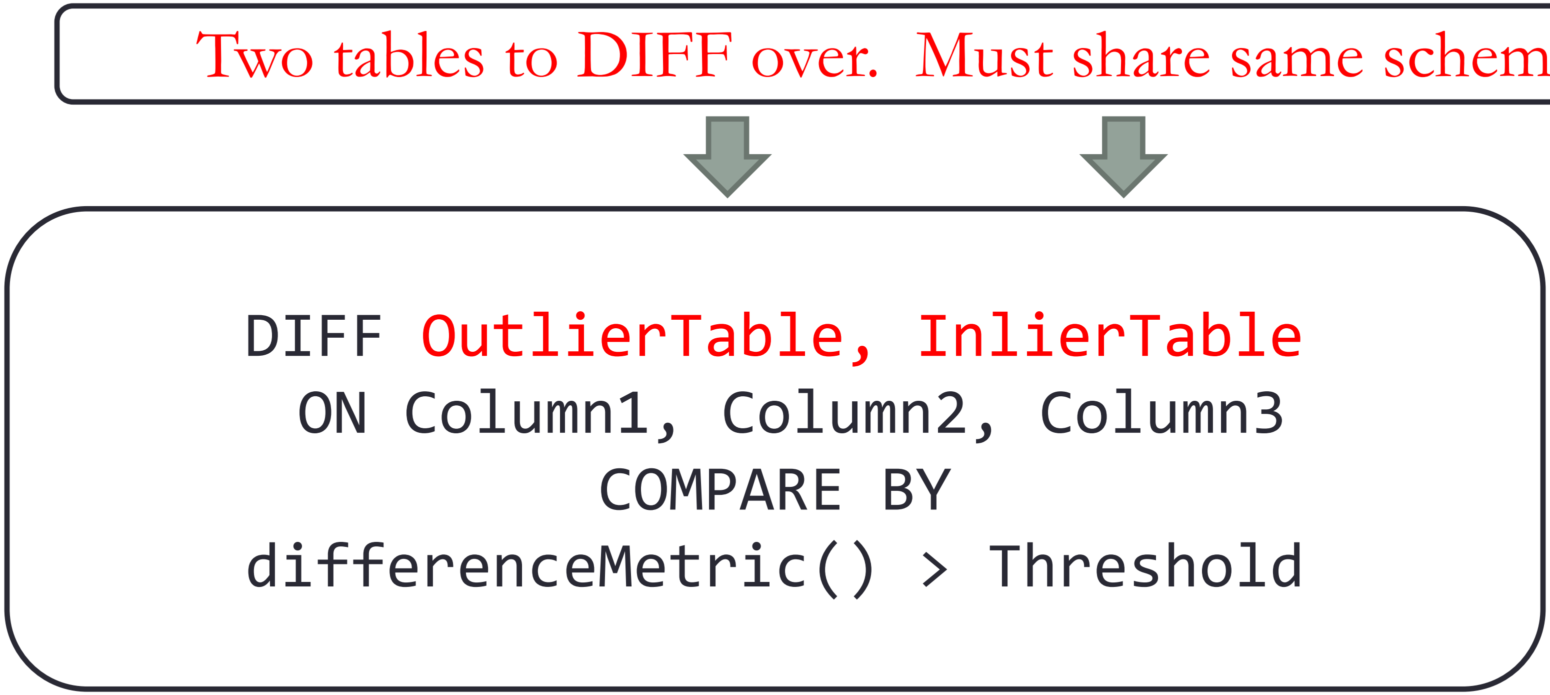
TableCrash

Version	Carrier	Country
4.1	Verizon	France
4.0	Verizon	USA

```
DIFF TableCrash, TableSuccess
ON Version, Carrier, Country
COMPARE BY
risk_ratio(COUNT(*)) > 3.0
support(COUNT(*)) > 0.05;
```

Version	Country	Support	Ratio
4.1	France	0.50	4.0

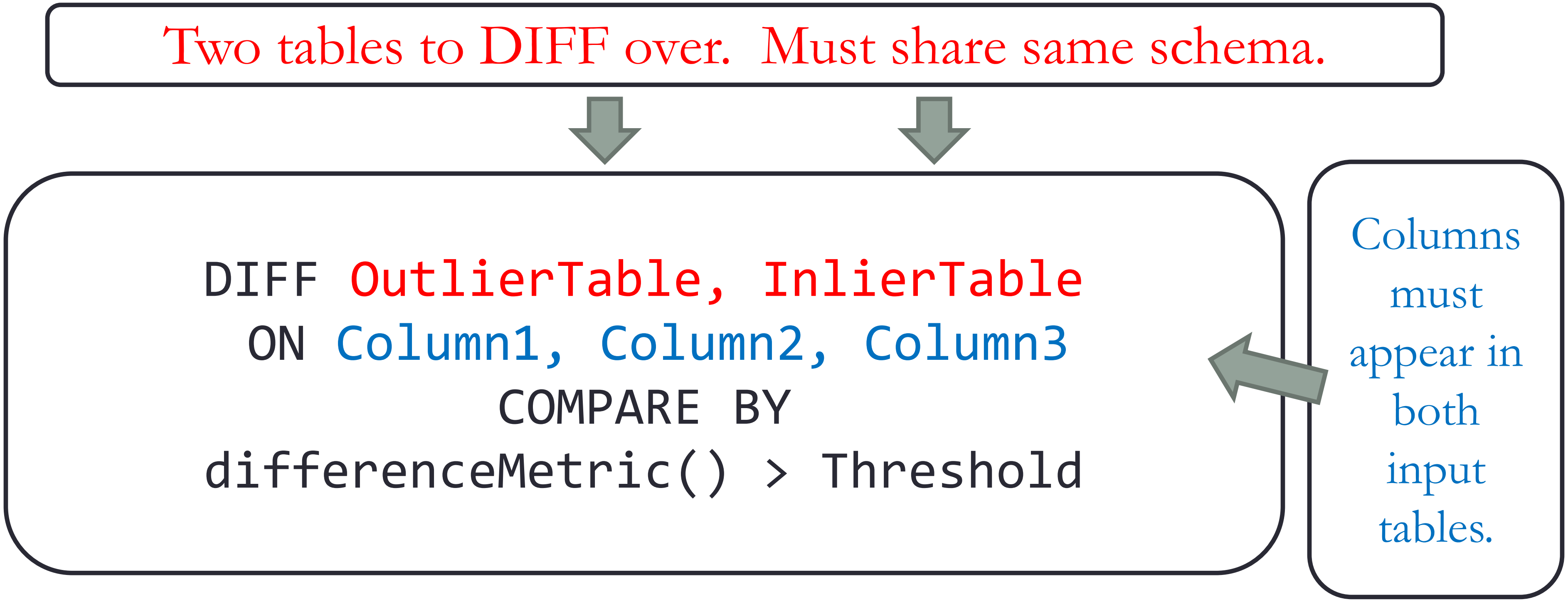
Anatomy of DIFF



Two tables to DIFF over. Must share same schema.

```
DIFF OutlierTable, InlierTable
  ON Column1, Column2, Column3
            COMPARE BY
differenceMetric() > Threshold
```

Anatomy of DIFF

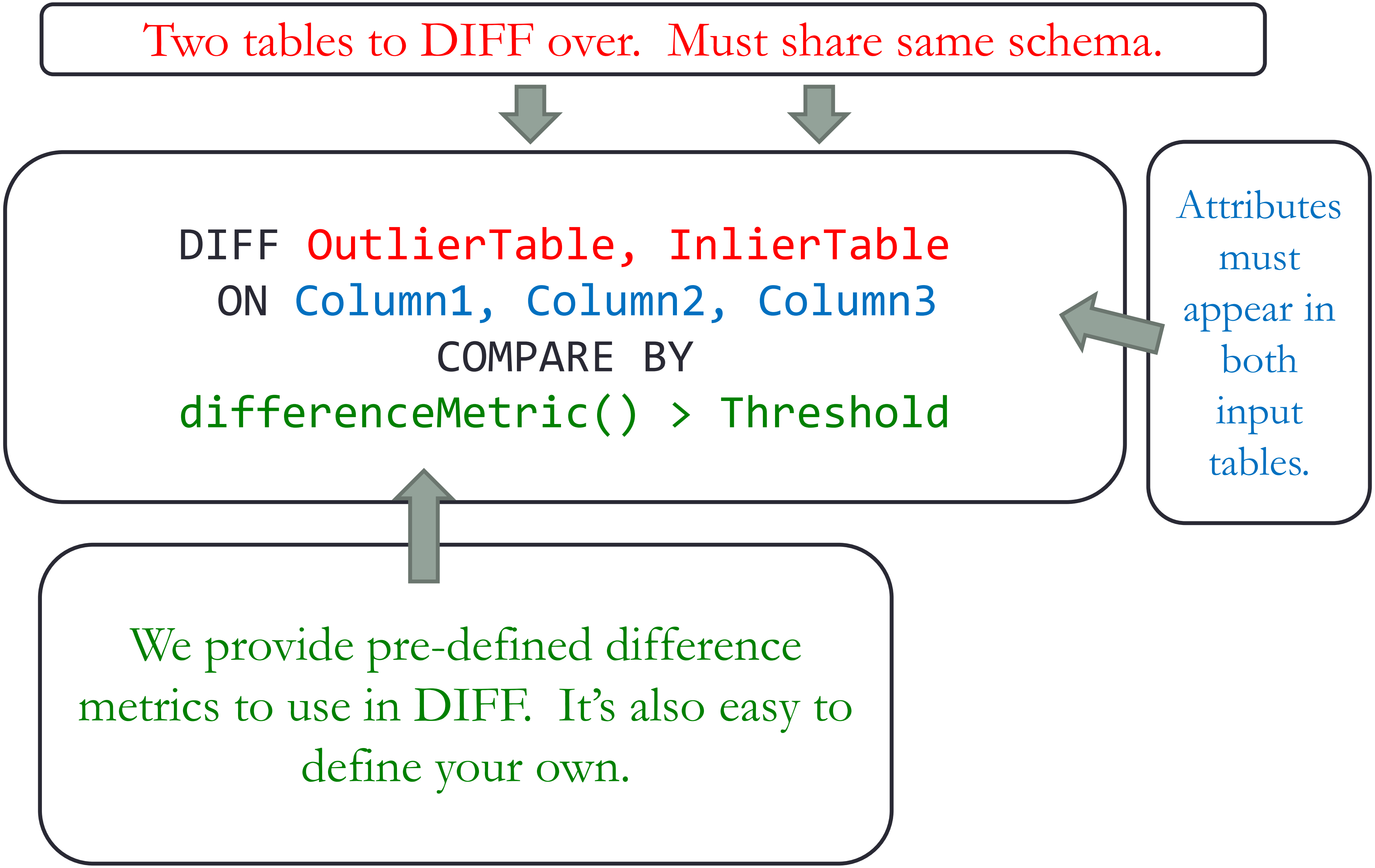


Two tables to DIFF over. Must share same schema.

```
DIFF OutlierTable, InlierTable
  ON Column1, Column2, Column3
           COMPARE BY
differenceMetric() > Threshold
```

Columns must appear in both input tables.

Anatomy of DIFF



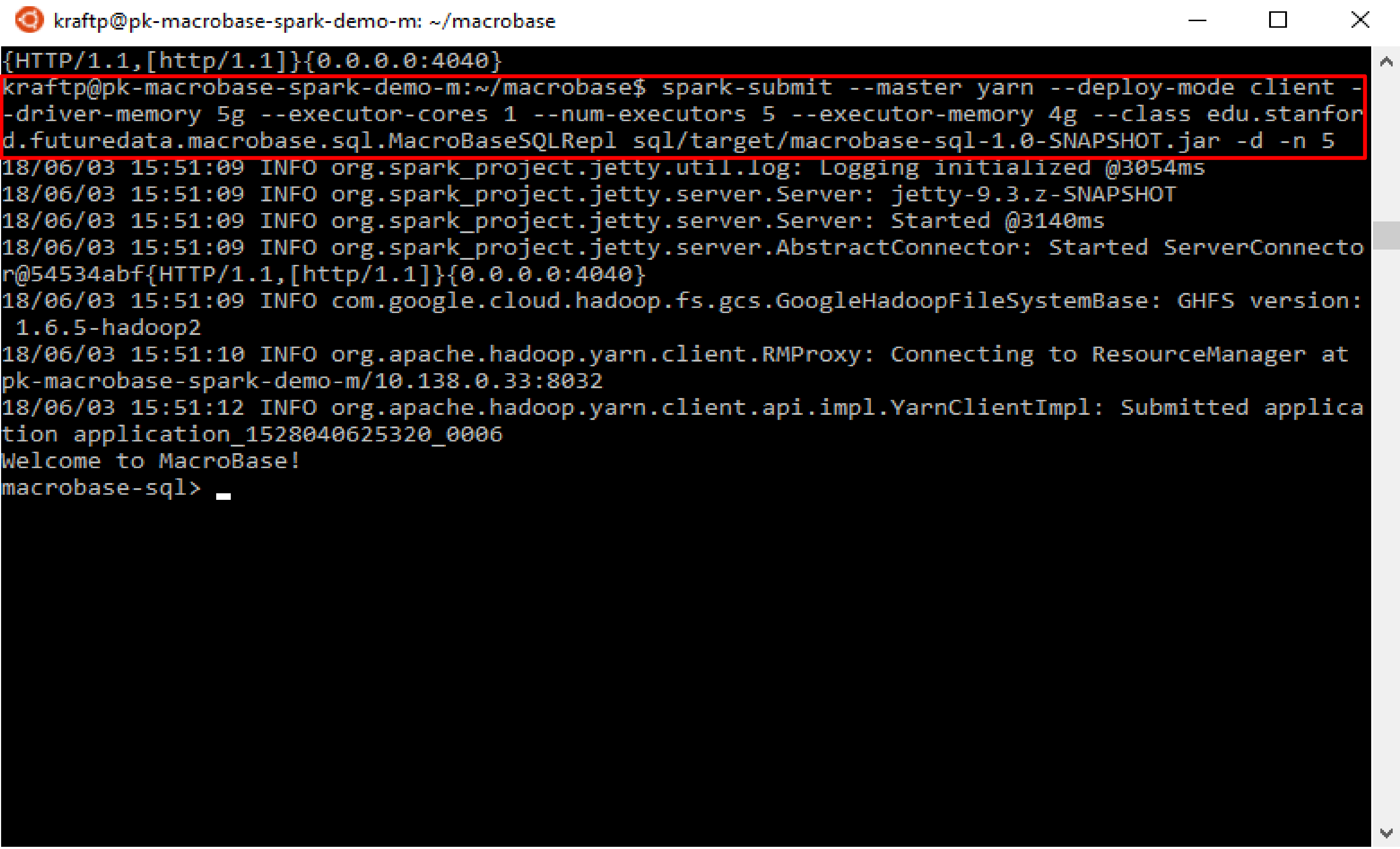
Two tables to DIFF over. Must share same schema.

```
DIFF OutlierTable, InlierTable
  ON Column1, Column2, Column3
          COMPARE BY
differenceMetric() > Threshold
```

Attributes must appear in both input tables.

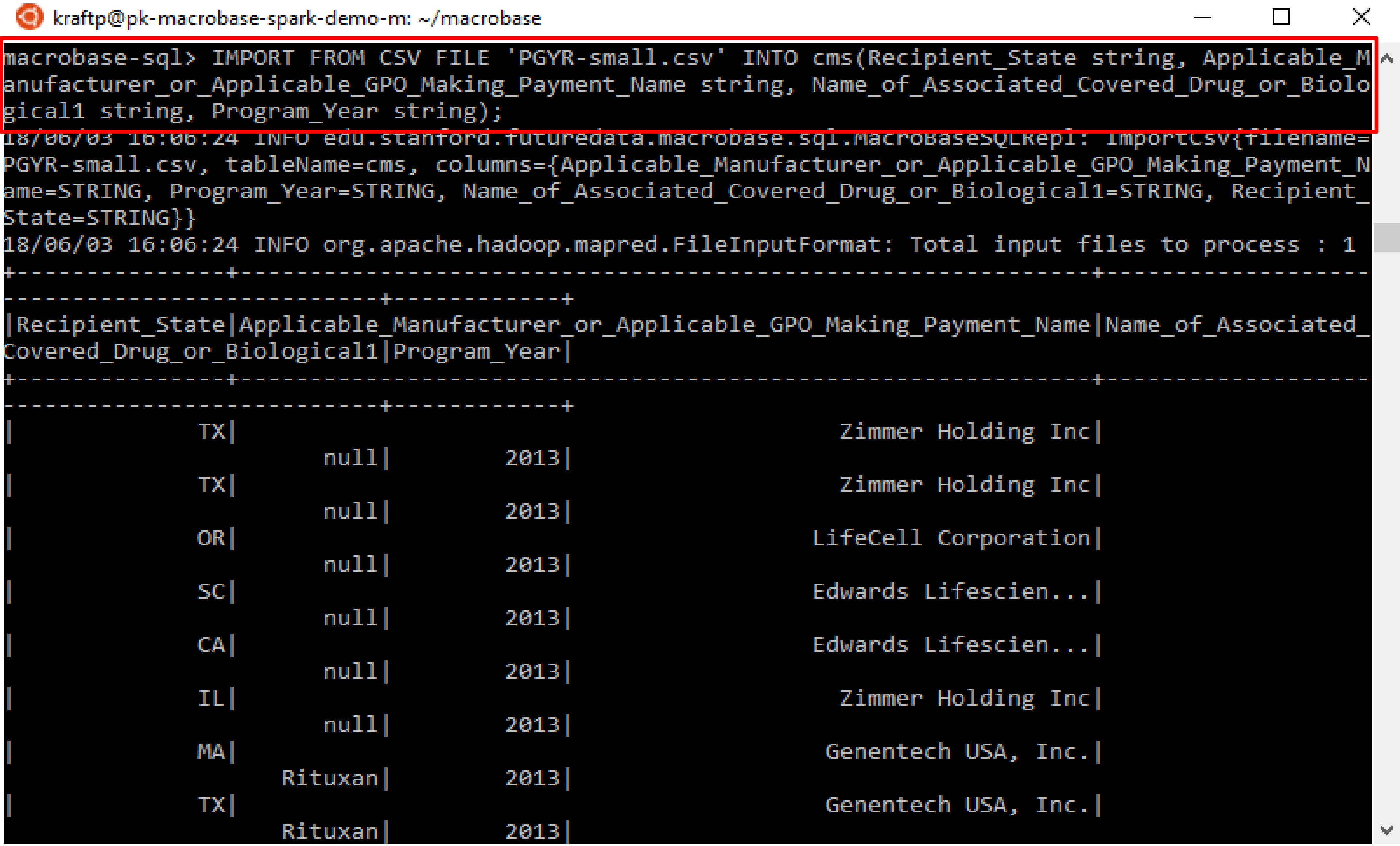
We provide pre-defined difference metrics to use in DIFF. It's also easy to define your own.

Demo - Startup



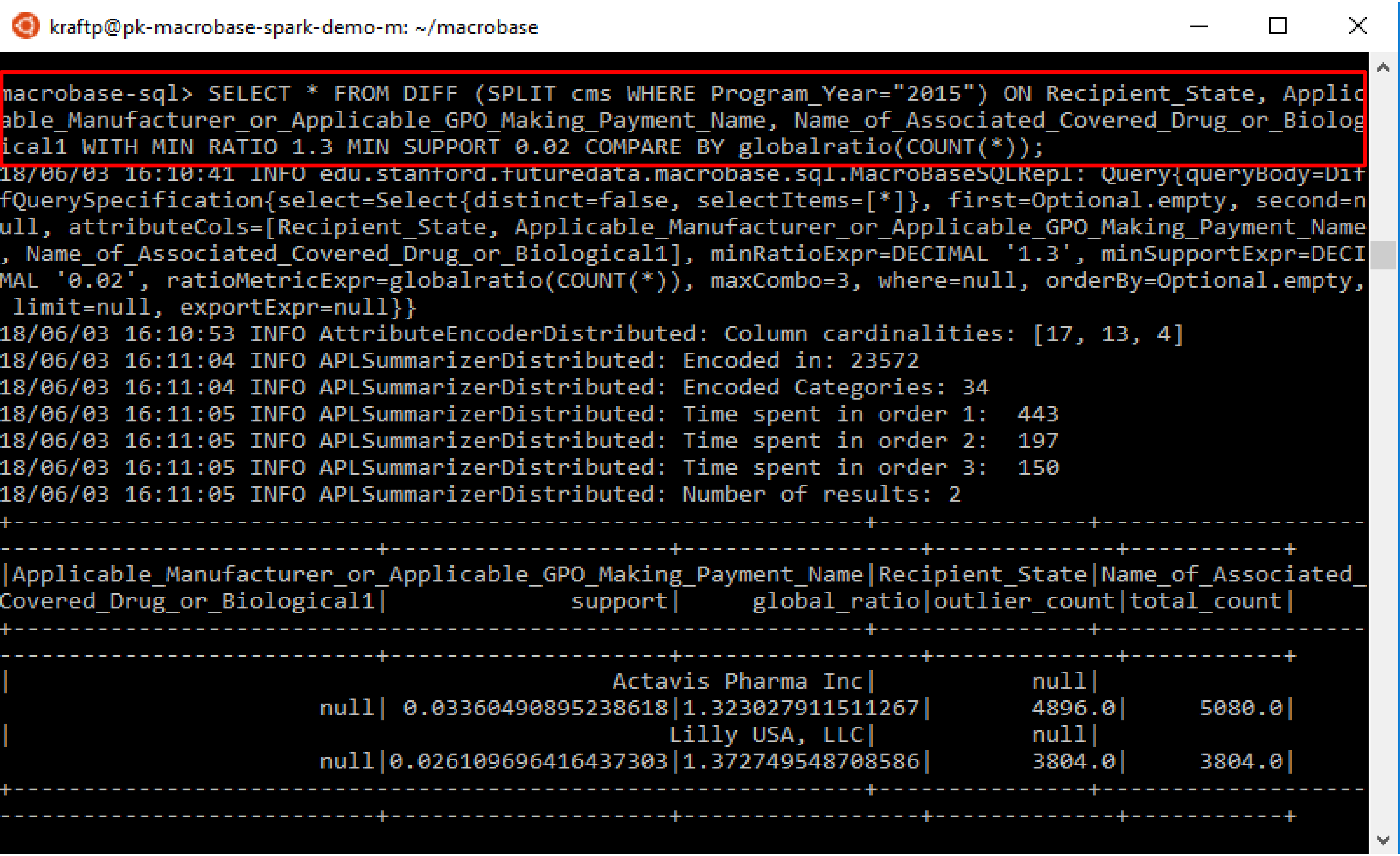
```
{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
kraftp@pk-macrobase-spark-demo-m:~/macrobase$ spark-submit --master yarn --deploy-mode client -
-driver-memory 5g --executor-cores 1 --num-executors 5 --executor-memory 4g --class edu.stanfor
d.futuredata.macrobase.sql.MacroBaseSQLRepl sql/target/macrobase-sql-1.0-SNAPSHOT.jar -d -n 5
18/06/03 15:51:09 INFO org.spark_project.jetty.util.log: Logging initialized @3054ms
18/06/03 15:51:09 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT
18/06/03 15:51:09 INFO org.spark_project.jetty.server.Server: Started @3140ms
18/06/03 15:51:09 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnecto
r@54534abf{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/06/03 15:51:09 INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystemBase: GHFS version:
 1.6.5-hadoop2
18/06/03 15:51:10 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at
pk-macrobase-spark-demo-m/10.138.0.33:8032
18/06/03 15:51:12 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted applica
tion application_1528040625320_0006
Welcome to MacroBase!
macrobase-sql> _
```

Demo - Ingest



```
macrobase-sql> IMPORT FROM CSV FILE 'PGYR-small.csv' INTO cms(Recipient_State string, Applicable_M
anufacturer_or_Applicable_GPO_Making_Payment_Name string, Name_of_Associated_Covered_Drug_or_Biolo
gical1 string, Program_Year string);
18/06/03 16:06:24 INFO edu.stanford.futuredata.macrobase.sql.MacroBaseSQLRepl: importCsv{filename=
PGYR-small.csv, tableName=cms, columns={Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_N
ame=STRING, Program_Year=STRING, Name_of_Associated_Covered_Drug_or_Biological1=STRING, Recipient_
State=STRING}}
18/06/03 16:06:24 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
+---------------+-------------------------------------------------------+-------------------
--------------------------+------------+
|Recipient_State|Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name|Name_of_Associated_
Covered_Drug_or_Biological1|Program_Year|
+---------------+-------------------------------------------------------+-------------------
--------------------------+------------+
|             TX|                                        Zimmer Holding Inc|
         null|        2013|
|             TX|                                        Zimmer Holding Inc|
         null|        2013|
|             OR|                                      LifeCell Corporation|
         null|        2013|
|             SC|                                      Edwards Lifescien...|
         null|        2013|
|             CA|                                      Edwards Lifescien...|
         null|        2013|
|             IL|                                        Zimmer Holding Inc|
         null|        2013|
|             MA|                                       Genentech USA, Inc.|
      Rituxan|        2013|
|             TX|                                       Genentech USA, Inc.|
      Rituxan|        2013|
```

Demo – DIFF



```
macrobase-sql> SELECT * FROM DIFF (SPLIT cms WHERE Program_Year="2015") ON Recipient_State, Applic
able_Manufacturer_or_Applicable_GPO_Making_Payment_Name, Name_of_Associated_Covered_Drug_or_Biolog
ical1 WITH MIN RATIO 1.3 MIN SUPPORT 0.02 COMPARE BY globalratio(COUNT(*));
18/06/03 16:10:41 INFO edu.stanford.futuredata.macrobase.sql.MacroBaseSQLRepl: Query{queryBody=Dif
fQuerySpecification{select=Select{distinct=false, selectItems=[*]}, first=Optional.empty, second=n
ull, attributeCols=[Recipient_State, Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name
, Name_of_Associated_Covered_Drug_or_Biological1], minRatioExpr=DECIMAL '1.3', minSupportExpr=DECI
MAL '0.02', ratioMetricExpr=globalratio(COUNT(*)), maxCombo=3, where=null, orderBy=Optional.empty,
 limit=null, exportExpr=null}}
18/06/03 16:10:53 INFO AttributeEncoderDistributed: Column cardinalities: [17, 13, 4]
18/06/03 16:11:04 INFO APLSummarizerDistributed: Encoded in: 23572
18/06/03 16:11:04 INFO APLSummarizerDistributed: Encoded Categories: 34
18/06/03 16:11:05 INFO APLSummarizerDistributed: Time spent in order 1:  443
18/06/03 16:11:05 INFO APLSummarizerDistributed: Time spent in order 2:  197
18/06/03 16:11:05 INFO APLSummarizerDistributed: Time spent in order 3:  150
18/06/03 16:11:05 INFO APLSummarizerDistributed: Number of results: 2
+-------------------------------------------------------------+---------------+-------------------
---------------------------+--------------------+-----------------+-------------+-----------+
|Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name|Recipient_State|Name_of_Associated_
Covered_Drug_or_Biological1|             support|     global_ratio|outlier_count|total_count|
+-------------------------------------------------------------+---------------+-------------------
---------------------------+--------------------+-----------------+-------------+-----------+
|                                           Actavis Pharma Inc|           null|
                       null| 0.03360490895238618|1.323027911511267|       4896.0|     5080.0|
|                                               Lilly USA, LLC|           null|
                       null|0.026109696416437303|1.372749548708586|       3804.0|     3804.0|
+-------------------------------------------------------------+---------------+-------------------
---------------------------+--------------------+-----------------+-------------+-----------+
```

How DIFF works: High Level

- DIFF calculates *difference* between two tables
- Quantify difference with difference metrics
- Return all sets of attributes that pass difference metrics

How DIFF works: Generalized Apriori

First, iterate through singleton attributes and store their frequency counts (or other aggregates)

Version	Carrier	Country	Crash
4.0	Verizon	USA	False
4.1	Verizon	USA	False
4.0	Verizon	France	True

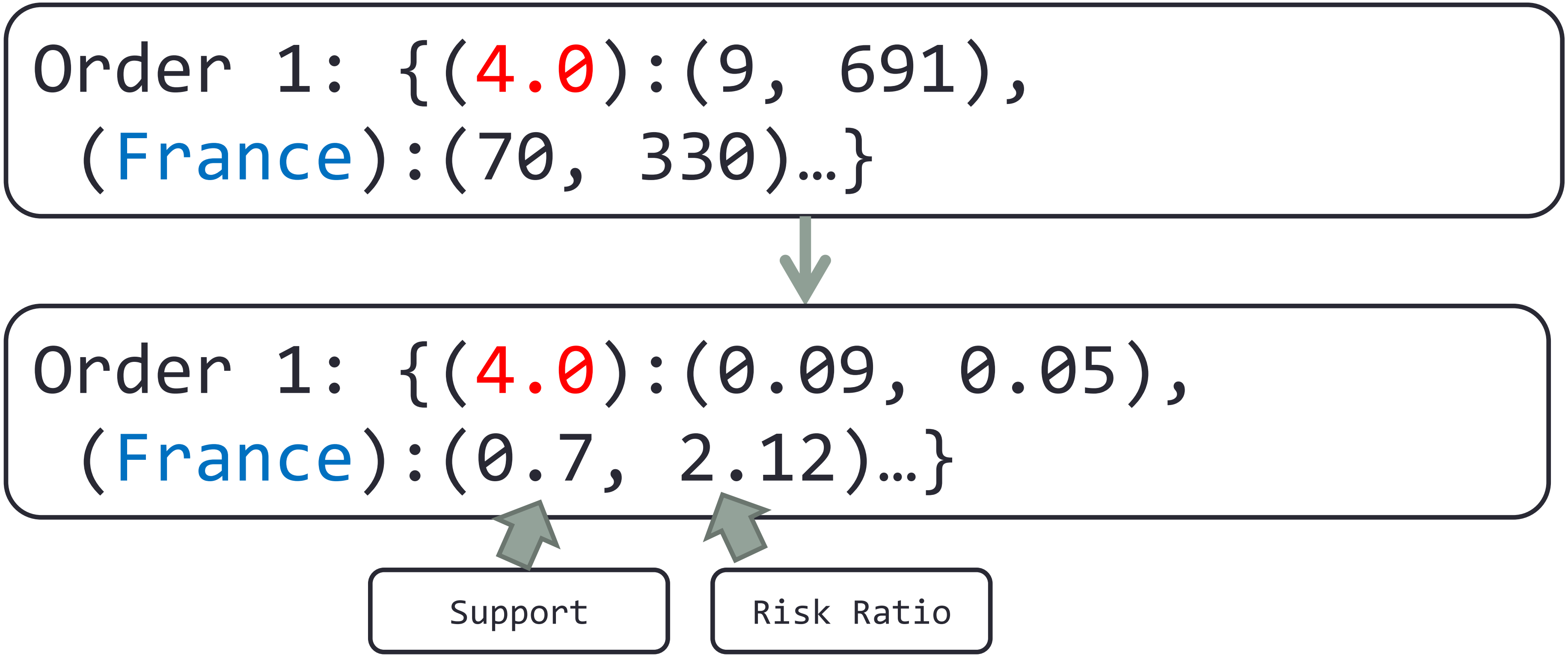
```
Order 1: {(4.0): (1, 1),
  (France): (1, 0)…}
```

Crashing Frequency Count

Success Frequency Count

How DIFF works: Generalized Apriori

Next, compute difference metrics—here, support and risk ratio—from aggregates. Assume 1000 logs, 100 of which crashed.



```
Order 1: {(4.0):(9, 691),
 (France):(70, 330)…}
```

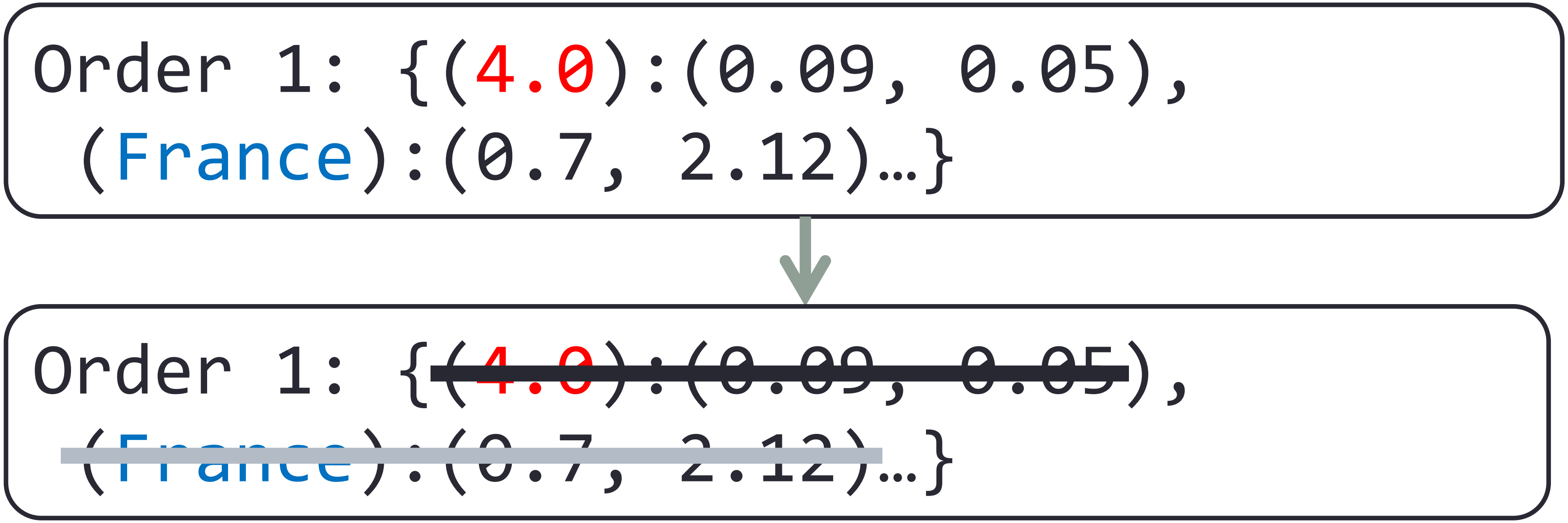
```
Order 1: {(4.0):(0.09, 0.05),
 (France):(0.7, 2.12)…}
```

Support

Risk Ratio

How DIFF works: Generalized Apriori

Then, prune with support (0.1) and risk ratio (3.0) thresholds. Keep track of who passed support threshold even if they failed risk ratio.



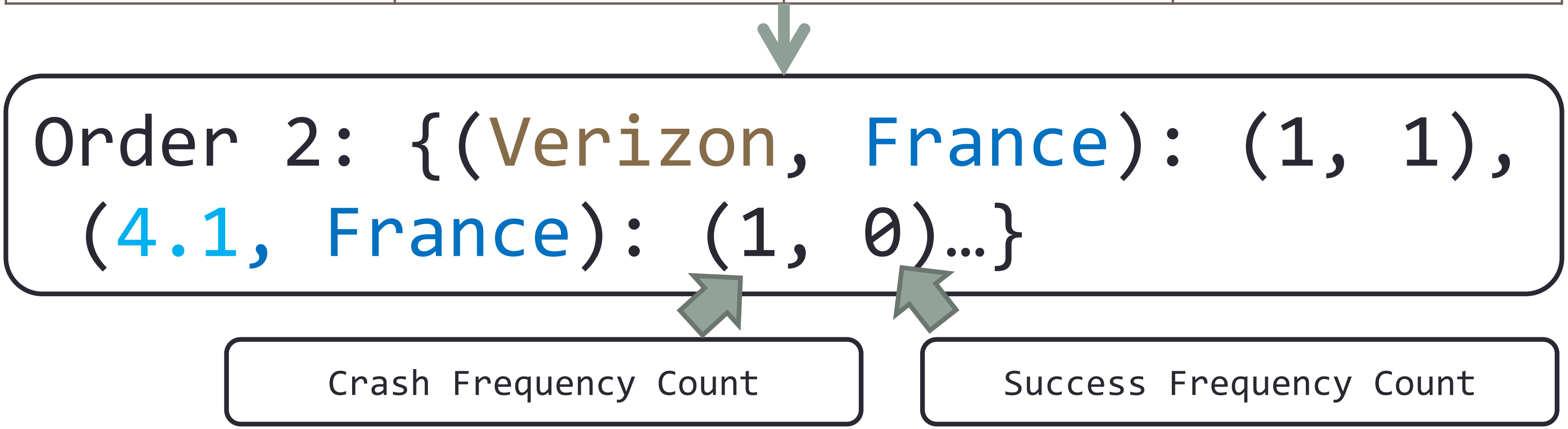
Order 1: {(4.0):(0.09, 0.05),
(France):(0.7, 2.12)…}

Order 1: {~~(4.0):(0.09, 0.05~~),
~~(France):(0.7, 2.12)~~…}

How DIFF works: Generalized Apriori

Now repeat for pairs. Iterate through pairs of attributes where both passed support threshold, even if they failed risk ratio.

Version	Carrier	Country	Crash
4.0	Verizon	USA	False
4.1	Verizon	USA	False
4.1	Verizon	France	True



```
Order 2: {(Verizon, France): (1, 1),
  (4.1, France): (1, 0)…}
```

Crash Frequency Count

Success Frequency Count

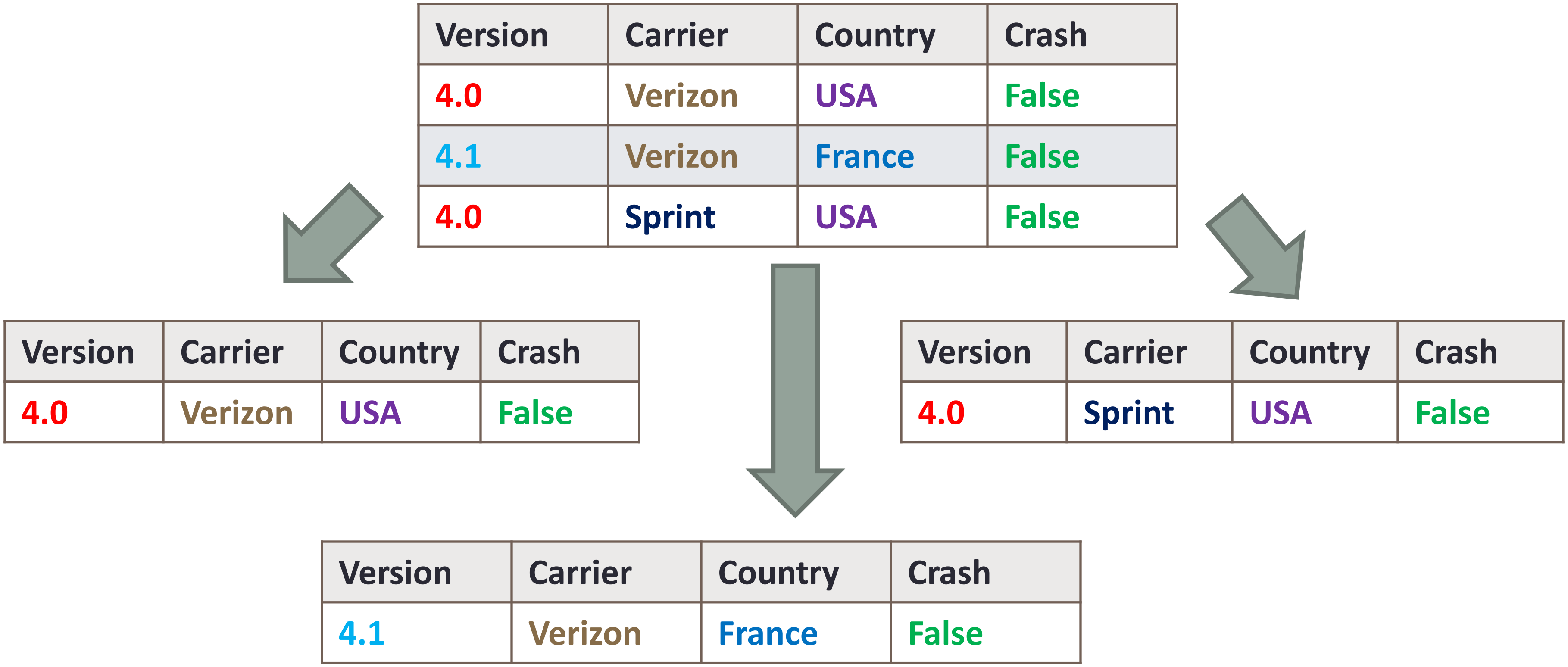
How DIFF works: Generalized Apriori

Repeat the same steps and we'll get our answer from before!

```
Order 2: {(4.1, France):(0.5, 4.0)…}
```

Version	Country	Support	Ratio
4.1	France	0.50	4.0

Distribute With Spark-Partitioning



Version	Carrier	Country	Crash
4.0	Verizon	USA	False
4.1	Verizon	France	False
4.0	Sprint	USA	False

Version	Carrier	Country	Crash
4.0	Verizon	USA	False

Version	Carrier	Country	Crash
4.0	Sprint	USA	False

Version	Carrier	Country	Crash
4.1	Verizon	France	False

First, partition your data by row to different machines.

Distribute With Spark-Mapping

Version	Carrier	Country	Crash
4.0	Verizon	USA	False

```
Order 1: {(4.0):(11, 522),
 (4.1): (50, 80)…}
```

Version	Carrier	Country	Crash
4.1	Verizon	France	True

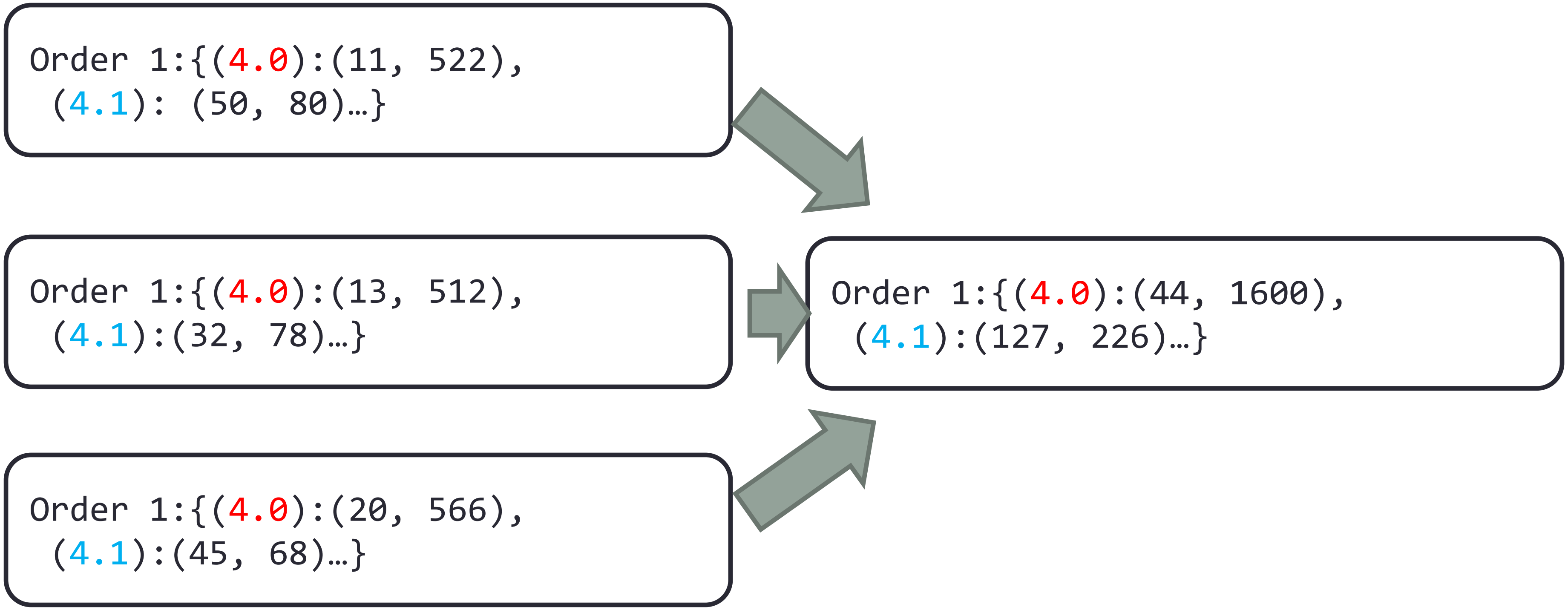
```
Order 1: {(4.0): (13, 512),
 (4.1): (32, 78)…}
```

Version	Carrier	Country	Crash
4.0	Sprint	USA	False

```
Order 1: {(4.0): (20, 566),
 (4.1): (45, 68)…}
```

Then, at each order, make a generalized Apriori map per partition. This is Spark's map step.

Distribute With Spark-Reducing



```
Order 1:{(4.0):(11, 522),
 (4.1): (50, 80)…}
```

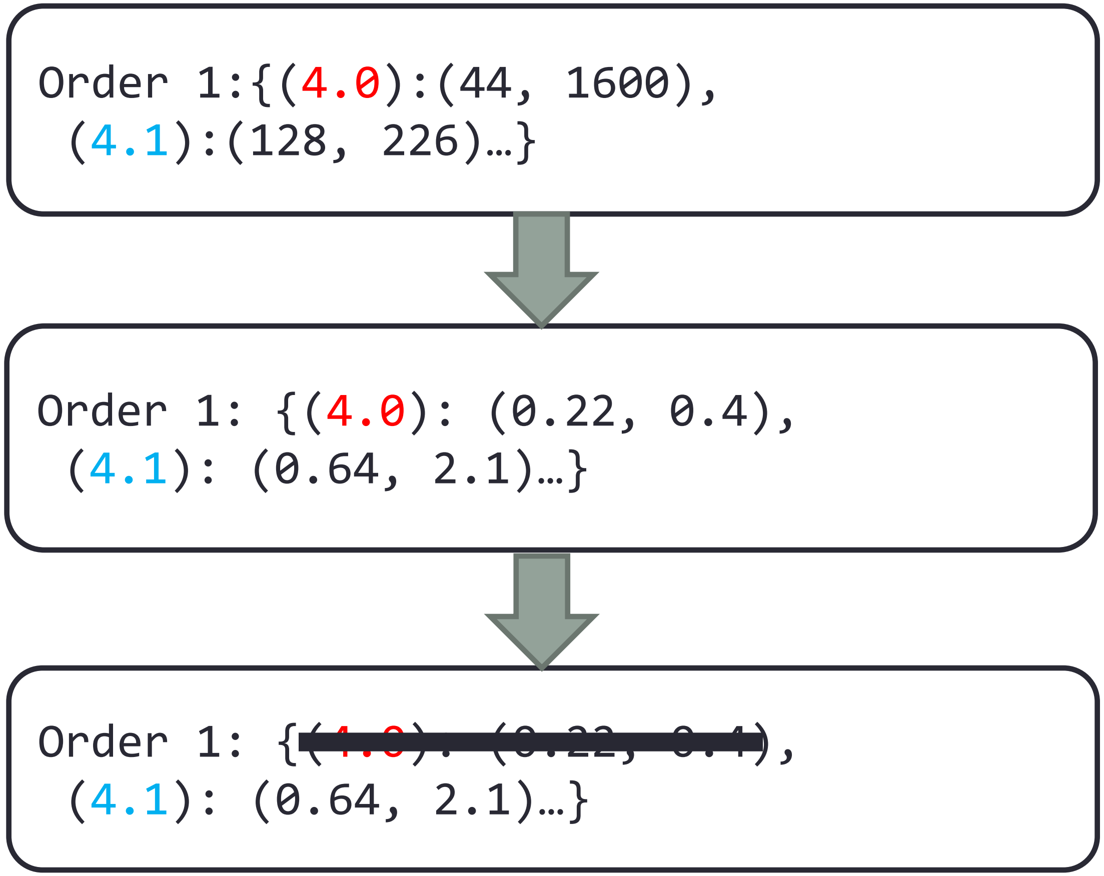
```
Order 1:{(4.0):(13, 512),
 (4.1):(32, 78)…}
```

```
Order 1:{(4.0):(20, 566),
 (4.1):(45, 68)…}
```

```
Order 1:{(4.0):(44, 1600),
 (4.1):(127, 226)…}
```

Next, combine all the maps onto a single machine. This is Spark's reduce step.

Distribute With Spark-Finishing



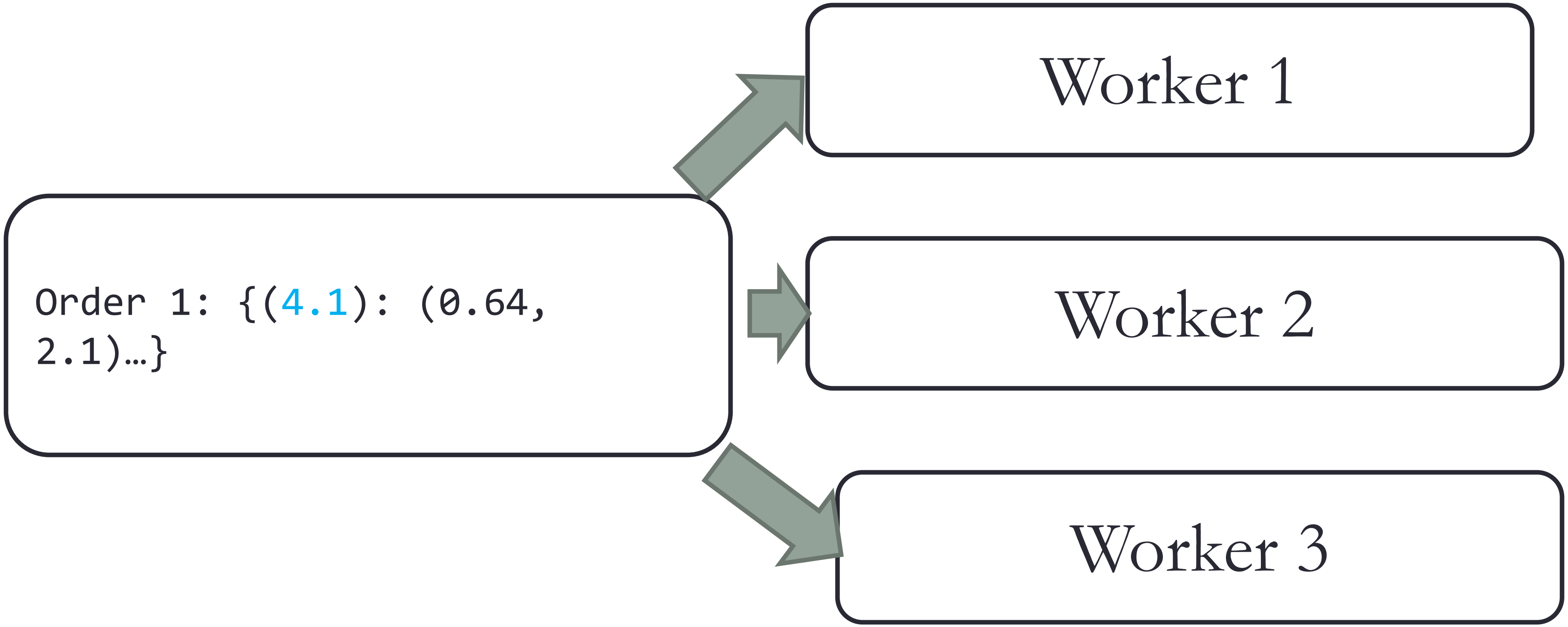
Order 1:{(4.0):(44, 1600),
 (4.1):(128, 226)…}

Order 1: {(4.0): (0.22, 0.4),
 (4.1): (0.64, 2.1)…}

Order 1: {~~(4.0): (0.22, 0.4)~~),
 (4.1): (0.64, 2.1)…}

Now that the map is on a single machine, compute difference metrics and prune as normal.

Distribute With Spark-Reducing



Order 1: {(4.1): (0.64, 2.1)…}

Worker 1

Worker 2

Worker 3

Next, combine all the maps onto a single machine. This is Spark's reduce step.

Use Macrobase!

- All our code is open-sourced on GitHub:
- https://github.com/stanford-futuredata/macrobase
- We also have tutorials and guides up on our website:
- https://macrobase.stanford.edu/
- https://macrobase.stanford.edu/docs/sql/spark/
- This includes info for both single-node MacroBase and MacroBase-Spark
- As long as you can get your data into a CSV (and onto HDFS for MB-Spark), you can run MacroBase on it!