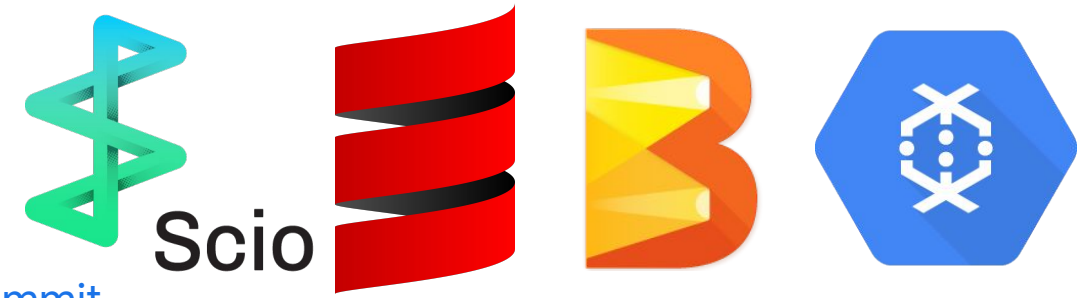
Can I use Scala with Apache Beam?

Israel Herraiz <ihr@google.com>
Strategic Cloud Engineer
Google Cloud
@herraiz on Twitter



https://github.com/iht/scio-scala-beam-summit

Scio,
the Scala
framework

https://github.com/iht/scio-scala-beam-summit

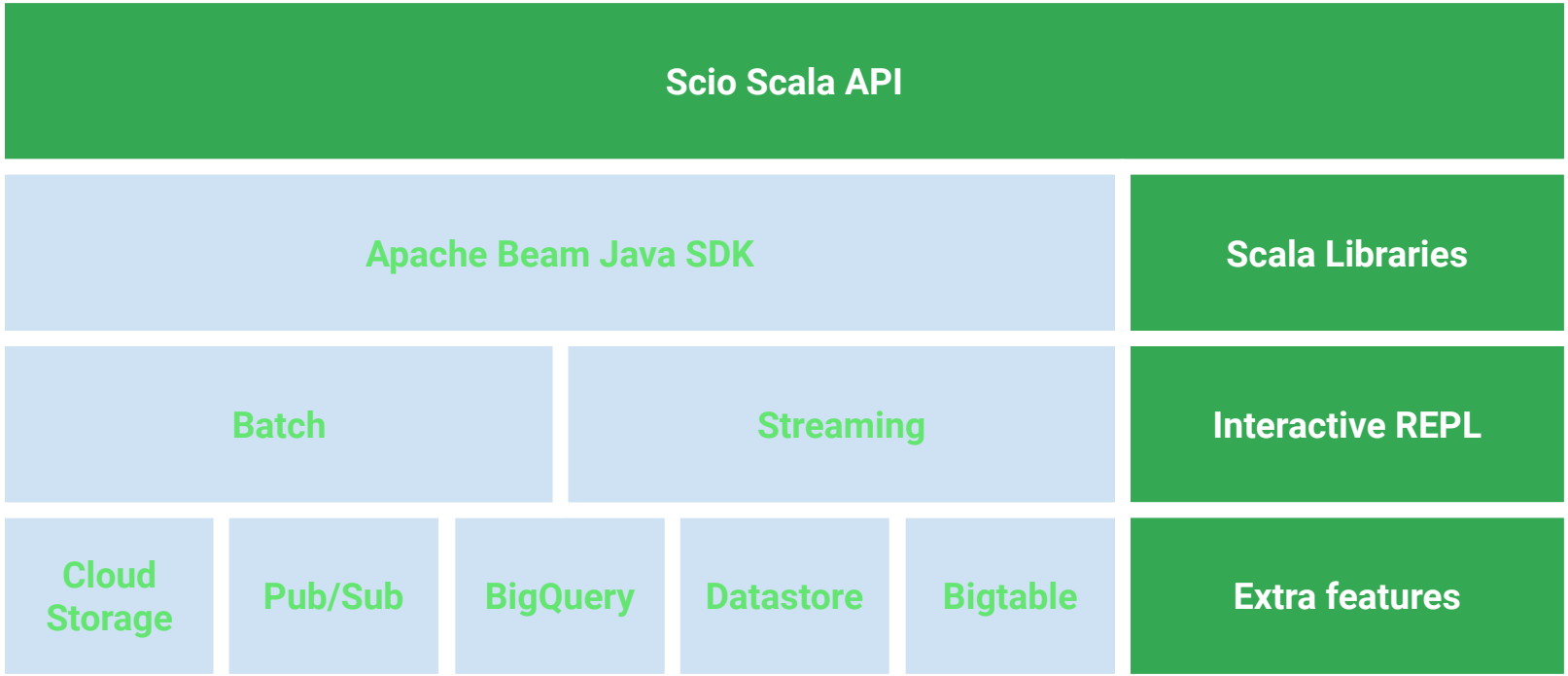
Beam and Scala: Scio

- High level DSL
- Familiarity with Scalding, Spark and Flink
- Functional programming natural fit for data
- Numerical libraries - Breeze, Algebird
- Macros & shapeless for code generation

https://github.com/iht/scio-scala-beam-summit

github.com/spotify/scio
Apache Licence 2.0

https://github.com/iht/scio-scala-beam-summit



https://github.com/iht/scio-scala-beam-summit

WordCount

```
val sc = ScioContext()
sc.textFile("shakespeare.txt")
  .flatMap { _
    .split("[^a-zA-Z']+")
    .filter(_.nonEmpty)
  }
  .countByValue
  .saveAsTextFile("wordcount.txt")
sc.close()
```

https://github.com/iht/scio-scala-beam-summit

Type safe BigQuery

```
@BigQuery.fromQuery("SELECT id, name FROM [users] WHERE ...")
class User  // look mom no code!
sc.typedBigQuery[User]().map(u => (u.id, u.name))

@BigQuery.toTable
case class Score(id: String, score: Double)
data.map(kv => Score(kv._1, kv._2)).saveAsTypedBigQuery("table")
```

https://github.com/iht/scio-scala-beam-summit

REPL

```
$ scio-repl
Welcome to
                     _____
    ________________(_)_____
    __  ___/  ___/_  /_  __ \
    _(__  )/ /__ _  / / /_/ /
    /____/ \___/ /_/  \____/   version 0.3.4

Using Scala version 2.11.11 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_121)

Type in expressions to have them evaluated.
Type :help for more information.

Using 'scio-test' as your BigQuery project.
BigQuery client available as 'bq'
Scio context available as 'sc'

scio> _
```

https://github.com/iht/scio-scala-beam-summit

How would you write Scio yourself?

WordCount example:

- In Beam Java:
  - https://github.com/nevillelyh/scio-deep-dive/blob/master/src/main/java/WordCount0.java
- Using Beam Java API in a Scala program:
  - https://github.com/nevillelyh/scio-deep-dive/blob/master/src/main/scala/WordCount1.scala
- Wrapping the Java PCollection in a Scala SCollection class:
  - https://github.com/nevillelyh/scio-deep-dive/blob/master/src/main/scala/WordCount15.scala

https://github.com/iht/scio-scala-beam-summit

Learn more

For this workshop: https://github.com/iht/scio-scala-beam-summit

About Scio: https://spotify.github.io/scio/

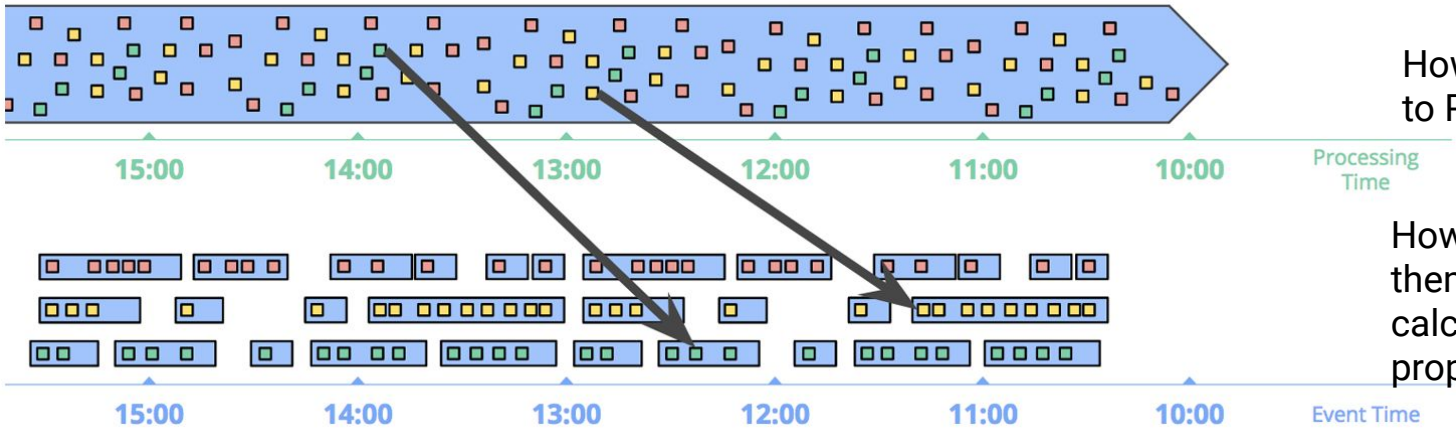
- Examples: https://spotify.github.io/scio/examples.html

Get started: https://cloud.google.com/free/

Pipeline for our workshop

https://github.com/iht/scio-scala-beam-summit

Taxi rides: reconstruction of sessions



How events arrive to PubSub

How we will group them to be able to calculate business properties

Each taxi ride sends a message every a few seconds.
For each *ride_id*, we will recover the sessions (group **and order**) all the messages, even if they are coming out of order.

We will calculate:

- Duration of session (*event* timestamp of most recent message, minus *event* timestamp of oldest message)
  - Most recent and oldest not in arriving order, but after ordering by event time
- Number of points in the session

https://github.com/iht/scio-scala-beam-summit

The data:

Attributes/metadata in Pubsub:

Pubsub is payload-agnostic

We need *attributes* to work in *event time**

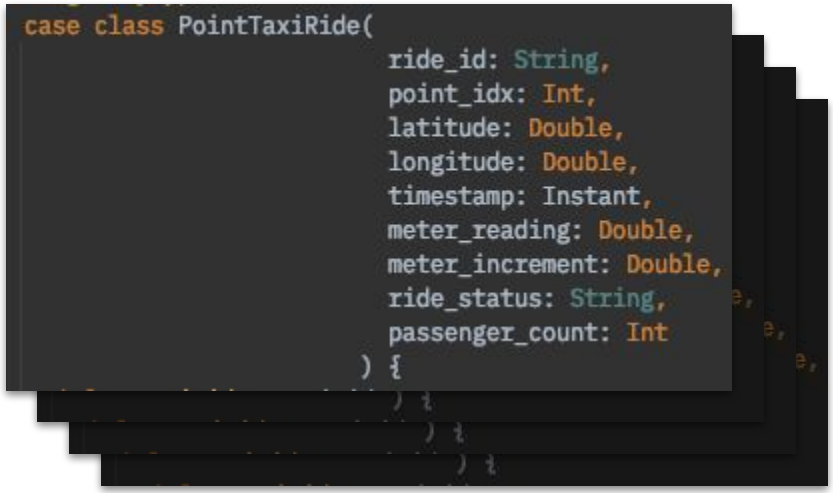
	MESSAGE_ID	ATTRIBUTES	DELIVERY_ATTEMPT
ıt":1}	1206646531578961	ts=2020-05-20T14:55:49.51875-04:00	

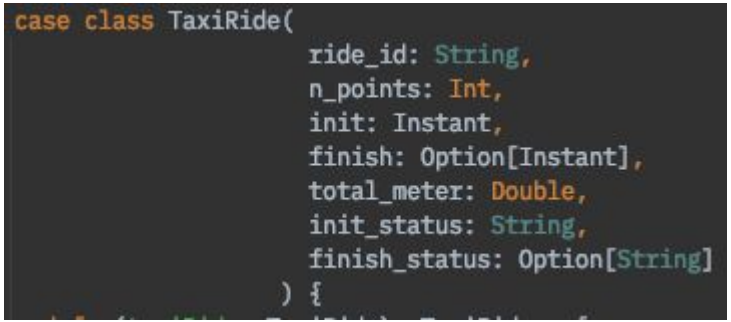
*or we could add timestamps to the data later on before applying the window


```
{
  "ride_id": "f172ff2a-b6fc-4345-b6df-d8e3c207b89c",
  "point_idx": 817,
  "latitude": 40.81768,
  "longitude": -73.94772,
  "timestamp": "2020-05-20T14:55:48.85921-04:00",
  "meter_reading": 17.08353,
  "meter_increment": 0.020910075,
  "ride_status": "enroute",
  "passenger_count": 1
}
{
  "ride_id": "c87a5a0e-4c75-4e12-8b8c-d3823cd31da2",
  "point_idx": 263,
  "latitude": 40.76919,
  "longitude": -73.91166000000001,
  "timestamp": "2020-05-20T14:55:48.86264-04:00",
  "meter_reading": 5.7470374,
  "meter_increment": 0.021851853,
  "ride_status": "enroute",
  "passenger_count": 1
}
{
  "ride_id": "2f9c21cb-b116-4219-a775-51edcc3c7531",
  "point_idx": 522,
  "latitude": 40.69608,
  "longitude": -73.98343000000001,
  "timestamp": "2020-05-20T14:55:48.86356-04:00",
  "meter_reading": 10.735218,
  "meter_increment": 0.020565553,
  "ride_status": "enroute",
  "passenger_count": 2
}
```


https://github.com/iht/scio-scala-beam-summit

Recovering the taxi sessions





https://github.com/iht/scio-scala-beam-summit