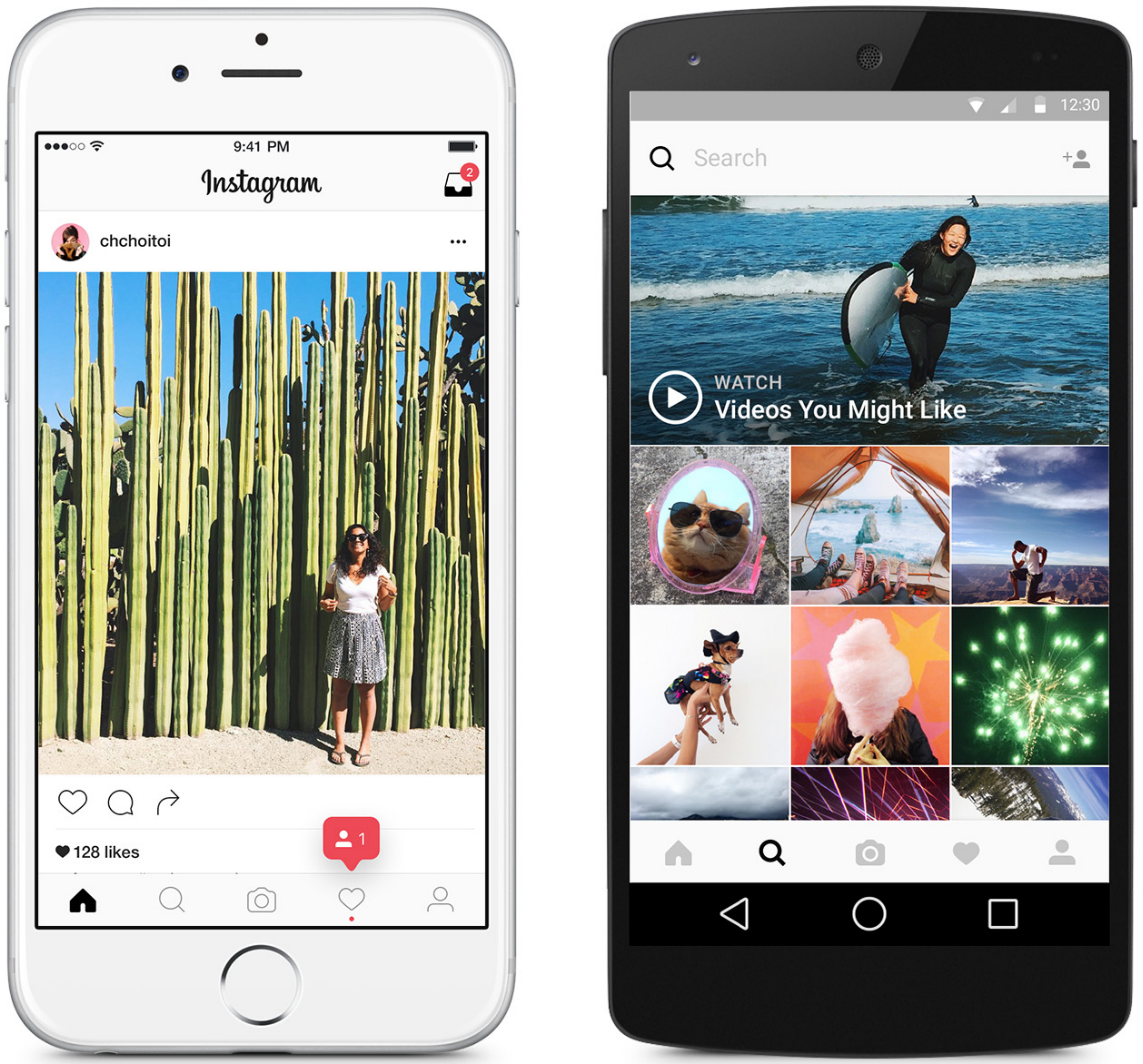
LESSONS LEARNED DEVELOPING AND MANAGING MASSIVE (300TB+) APACHE SPARK PIPELINES IN PRODUCTION

Brandon Carl


9:41 PM
Instagram
2
chchoitoi
128 likes
1
12:30
Search
WATCH
Videos You Might Like

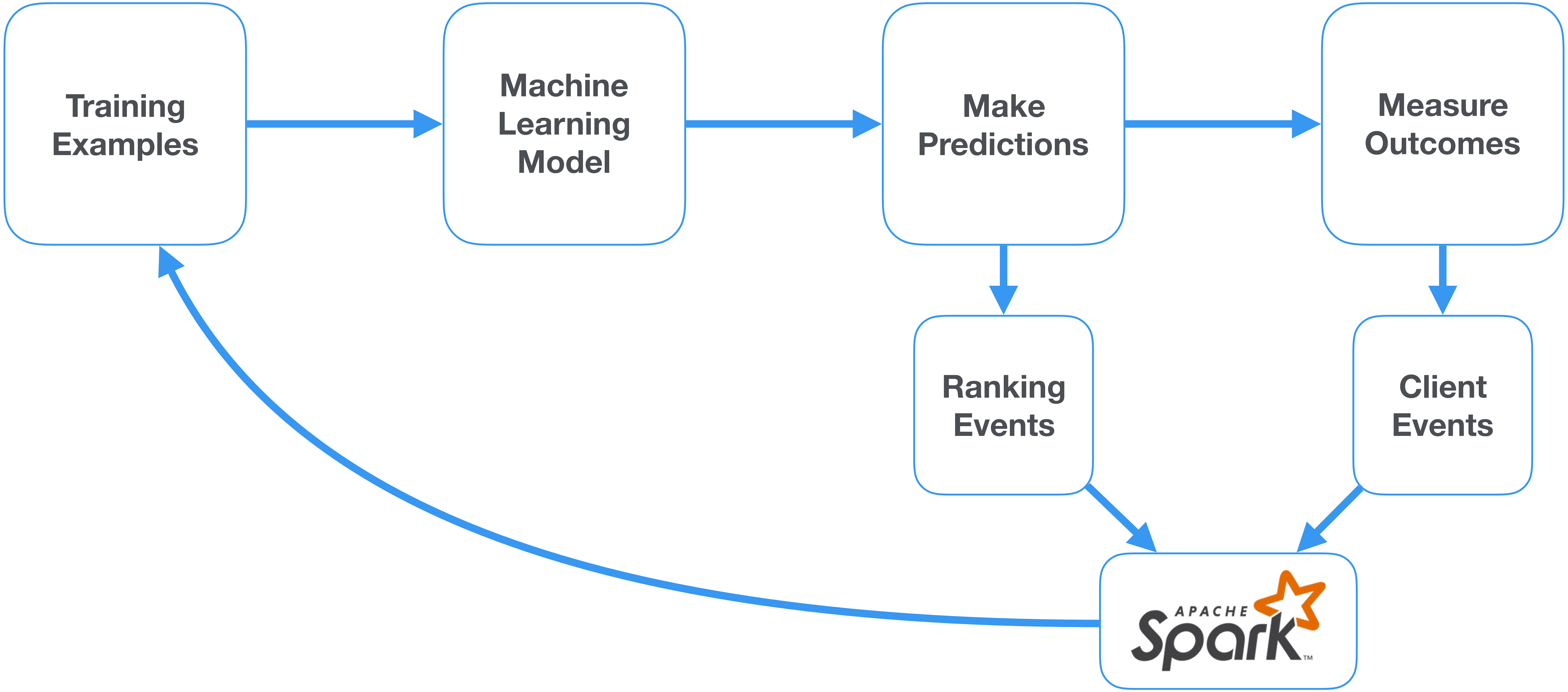
"SEE THE MOMENTS YOU CARE ABOUT FIRST"

MARCH 15, 2016

800
000
000

MACHINE LEARNING

MACHINE LEARNING LIFECYCLE



WHY SPARK?

- Performance
- Testability
- Modularity
- Serialized Logging

SERIALIZED LOGGING

```
{
  "id": 123,
  "scores": {
    "modelA": 0.2345,
    "modelB": 0.0012
  },
  "features": {
    1001: 0.9934,
    1002: 0.1923
  }
}
```

SERIALIZED LOGGING

```
struct Candidate {
  1: i64 id;
  2: map<string, double> scores;
  3: map<i64, double> features;
}
```

```
new Candidate()
  .setId(id)
  .setScores(scores)
  .setFeatures(features)
```

CHANGES OVER TIME

CHANGES OVER TIME

- RDD
- Dataset
- Training Data Joiner

Training Data Joiner

```
class MyTrainingDataJoiner(spark: SparkSession) extends TrainingDataJoiner {
  val labels: Map[String, LabelFunction] = ???
}

case class Output(id: Long, label_value: Double)
```

MANAGING MASSIVE SCALE

MANAGING MASSIVE SCALE - PEOPLE

AUTOMATE EVERYTHING

SIMPLE INTERFACE

SIMPLE INTERFACE

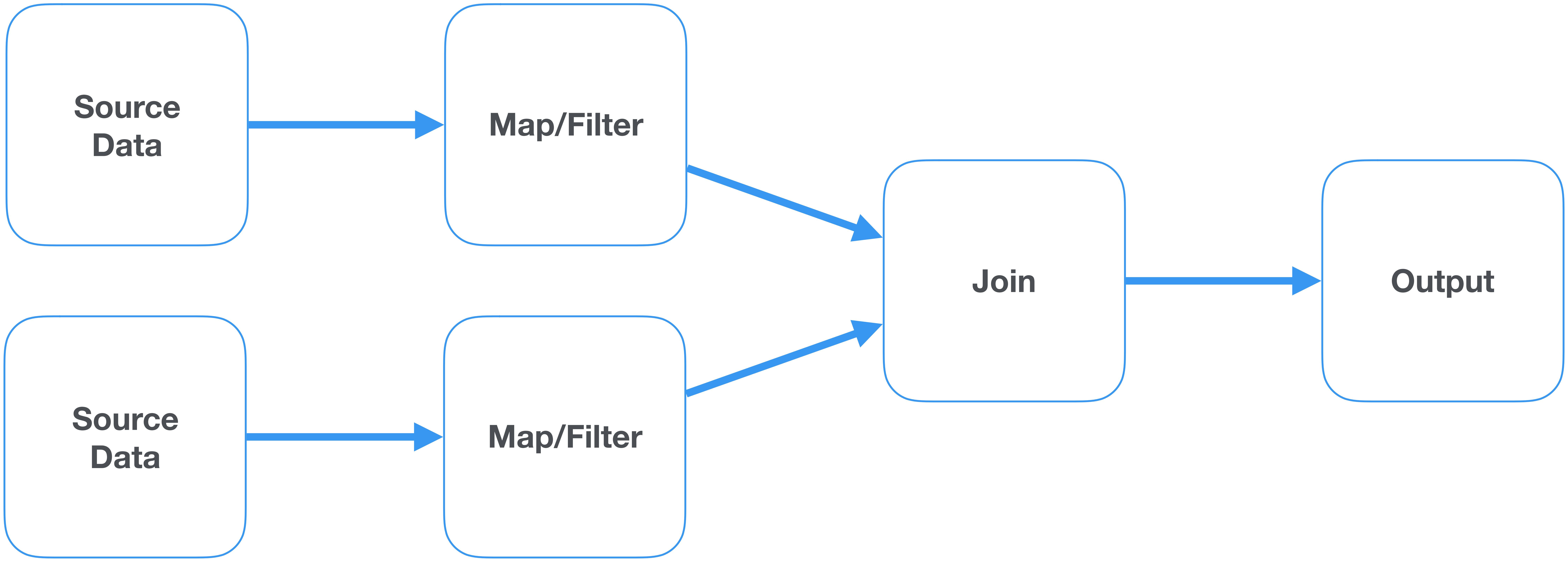
```
RankingEvent
  .read('input_table', '2017-10-25')
  .filter(...)
  .map(...)
  .write('output_table', '2017-10-25')
```

MANAGING MASSIVE SCALE - DATA

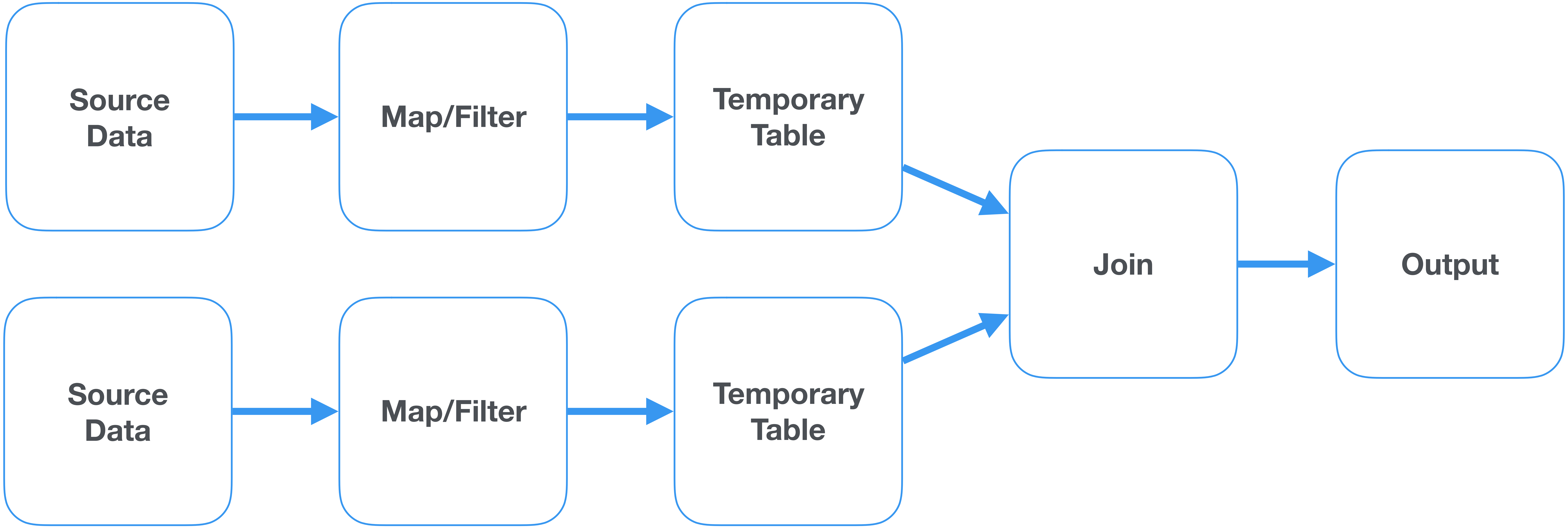
PLAN FOR GROWTH

PERSIST TO HDFS

PERSIST TO HDFS



PERSIST TO HDFS



KRYO SERIALIZATION

KRYO SERIALIZATION

```
new SparkConf()
 .set("spark.serializer", "org.apache.spark.serializer.KryoSerializer")
 .set("spark.kryo.registrationRequired", "true")
 .registerKryoClasses(Array(classOf[...], ...))
```

BIG-O MATTERS

BIG-O MATTERS

```
final def withName(s: String): Value =
  values
    .find(_.toString == s)
    .getOrElse(throw new NoSuchElementException(...))
```

BIG-O MATTERS

```
final def withName(s: String): Value =
  values
    .find(_.toString == s)
    .getOrElse(throw new NoSuchElementException(...))
```

DATA STRUCTURES MATTER

DATA STRUCTURES MATTER

- AnyRefMap
- IntMap
- LongMap
- fastutil (http://fastutil.di.unimi.it)

DATA SKEW MATTERS

TEST ON SAMPLED DATA

800
000
000