livelessons
video instruction from technology experts

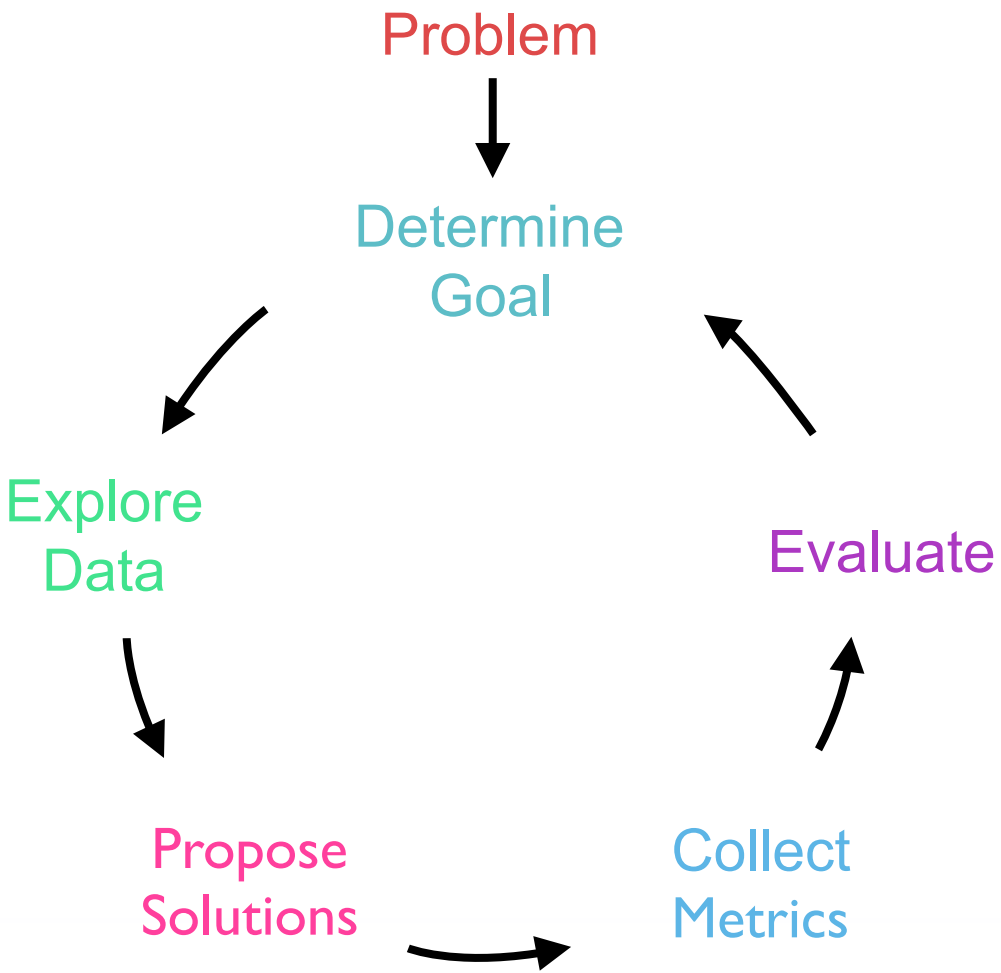
Lesson 3: Your First Spark Application

3.4 Data Quality Checks with Accumulators

Common Questions

- How many records in total are there?
- How many uniques values does each column contain?
- How many missing (or null) values are there?
- For numeric columns, what are its summary statistics
- What are values appear most often in each column?
- How are the values of each column distributed?

Data Science Process



Unique Values

- rdd.distinct()
- rdd.countApproxDistinct(relative_accuracy)

```
# HyperLogLog
rdd_dict.map(lambda row: row['_schoolid']).countApproxDistinct()
```

```
62989L
```

```
rdd_dict.map(lambda row: row['_schoolid']).distinct().count()
```

```
61402
```

http://dx.doi.org/10.1145/2452376.2452456

http://content.research.neustar.biz/blog/hll.html

Missing Values

- `column.isNull()`
- `dataframe.dropna(column_name)`
- `dataframe.fillna()`
- `dataframe.replace(to_replace, value)`
- `pyspark.accumulators`

https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrameNaFunctions

Missing Values

- `column.isNull()`
- `dataframe.fillna()`

```
df.filter(df_dates['students_reached'].isNull()).select('students_reached', 'funding_status').collect()
```

```
[Row(students_reached=None, funding_status=u'completed'),
 Row(students_reached=None, funding_status=u'completed'),
 Row(students_reached=None, funding_status=u'expired'),
 Row(students_reached=None, funding_status=u'completed'),
                           ...
```

```
df_no_null = df.fillna(0, ['students_reached'])
```

https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrameNaFunctions

```
accum = sc.accumulator(0)
```

```
from collections import Counter
```

```
class CounterAccumulatorParam(ps.accumulators.AccumulatorParam):
    def zero(self, initialValue):
        return Counter()

    def addInPlace(self, v1, v2):
        v1 += v2
        return v1
```

```
accum = sc.accumulator(Counter(), CounterAccumulatorParam())
```

```
def count_null(record):
    global accum

    c = Counter()

    for key, value in record.items():
            if value == '':
                c[key] += 1

    accum.add(c)
```

```
rdd_dict.foreach(count_null)
```

```
accum.value
```

```
Counter({u'date_thank_you_packet_mailed': 305004, u'date_completed': 253218, u'secondary_focus_subject': 238255, u'se
condary_focus_area': 238255, u'school_metro': 95180, u'school_ncesid': 50973, u'payment_processing_charges': 35082,
u'vendor_shipping_charges': 35082, u'fulfillment_labor_materials': 35082, u'sales_tax': 35082, u'school_district': 91
8, u'date_expiration': 208, u'students_reached': 150, u'resource_type': 48, u'primary_focus_subject': 42, u'primary_f
ocus_area': 42, u'grade_level': 33, u'teacher_prefix': 20, u'school_county': 18, u'school_zip': 4})
```