livelessons
video instruction from technology experts

Lesson 2: Spark Programming APIs

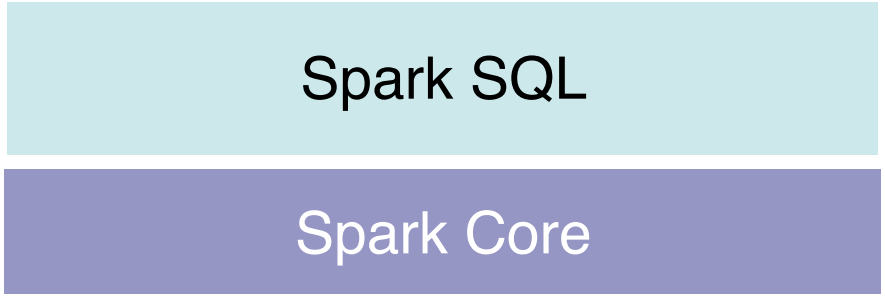
2.8 Why (Spark) SQL?

Integrated API

```
context = ps.HiveContext(sc)

# query with SQL
results = context.sql(
  "SELECT * FROM people”)

# apply Python transformation
names = results.map(lambda p: p.name)
```



Why SQL?

```
{'name': 'Amy', age: 18,  hobby: 'sports'}

{'name': 'Greg', age: 60,  hobby: 'fishing'}

{'name': 'Susan', age: 30}
```

Why SQL?

Query: Older than 18 with hobbies

```
rdd.filter(lambda d: d['age'] > 18) \
   .filter(lambda d: 'hobby' in d.keys()) \
   .map(lambda d: d['name'])
```

```
{'name': 'Amy', age: 18,  hobby: 'sports'}

{'name': 'Greg', age: 60,  hobby: 'fishing'}

{'name': 'Susan', age: 30}
```

Why SQL?

Query: Older than 18 with hobbies

```
SELECT name
WHERE age > 18
AND hobby IS NOT NULL
```

```
{'name': 'Amy', age: 18,  hobby: 'sports'}
{'name': 'Greg', age: 60,  hobby: 'fishing'}
{'name': 'Susan', age: 30}
```