Data Science with Apache Spark

Kirill Pavlov

Data Science Team, Asia Miles Limited

April 19, 2016

Quick Questionnaire

- How many people have attended previous Spark talks?

Quick Questionnaire

- How many people have attended previous Spark talks?
- How many people are currently working with Spark?

Quick Questionnaire

- How many people have attended previous Spark talks?
- How many people are currently working with Spark?
- How many people are familiar with Scala?

About the Presenter

- MS degree from Moscow Institute of Physics and Technology with distinction.
- 8+ years of data science and machine learning experience.
- Worked in Yandex (Russian Google) on search and on-line contextual ads ranking algorithms.
- Developed and consulted start-ups in digital marketing, healthcare, real estate and home automation areas.
- Open-source contributor, full-stack engineer and data mining evangelist.
- Now data scientist in Asia Miles.

Table of content

1. Overview

2. Data transformation and analysis with Spark

3. Data mining with Spark

4. Conclusion

1. Overview

2. Data transformation and analysis with Spark

3. Data mining with Spark

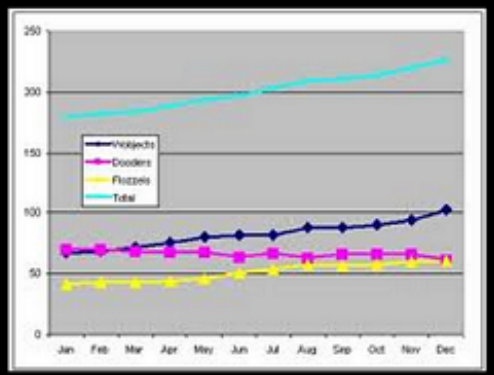
4. Conclusion

Data Scientist

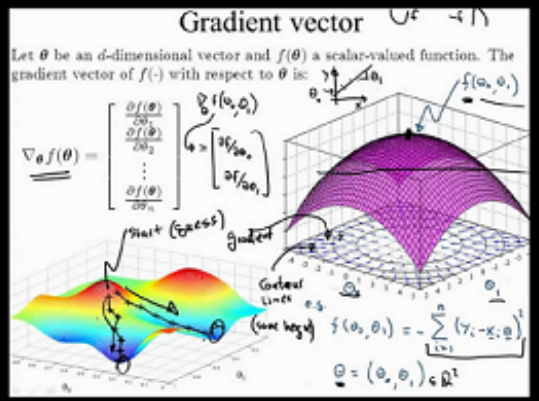
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

What I actually do

source: www.sintetia.com

Overview

- Data processing and cleaning takes **80%** of time.

Environment

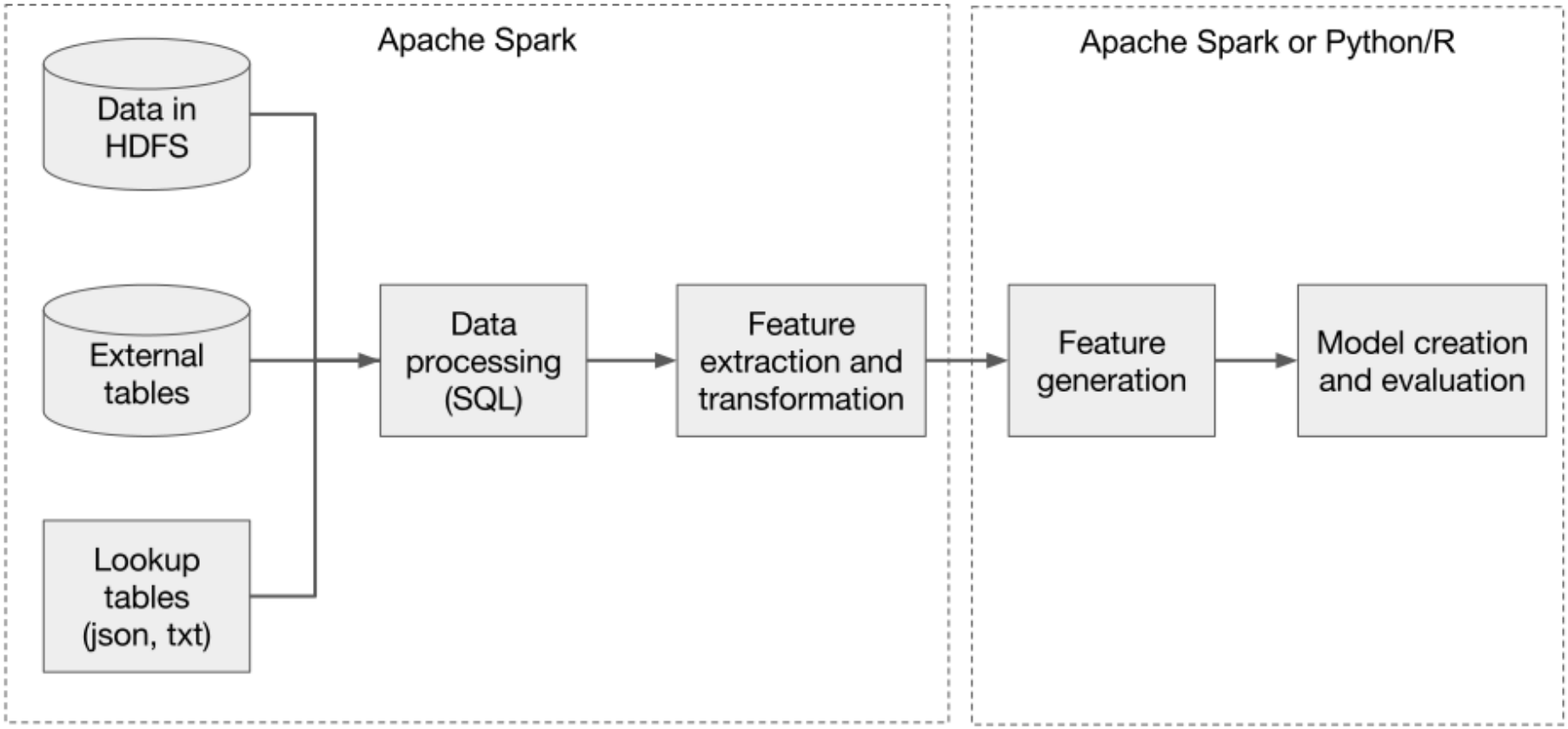
visualization

* Data visualization part (we use excel/external 3rd party tools). Sometimes d3js. * Data visualization split data generation from data plot. Automate genration part.

Example: Estimators and Transformers

```
1 // Define indexers and encoders
2 val fieldsToIndex = Array("gender", "curr_tier_type",  "prefred_lang")
3 val indexers = fieldsToIndex.map(f => new StringIndexer()
4   .setInputCol(f).setOutputCol(f + "_index"))
5
6 val fieldsToEncode = Array("gender", "prefred_lang")
7 val oneHotEncoders = fieldsToEncode.map(f => new OneHotEncoder()
8   .setInputCol(f + "_index").setOutputCol(f + "_flags"))
9
10 // Combine stages into pipeline
11 val pipeline = new Pipeline().setStages(indexers ++ oneHotEncoders)
```

Workflow


Apache Spark
Apache Spark or Python/R
Data in HDFS
External tables
Lookup tables (json, txt)
Data processing (SQL)
Feature extraction and transformation
Feature generation
Model creation and evaluation

Recap

Quiz

We are hiring data scientists

Thank you!

Kirill Pavlov, Data Science Team, Asia Miles Limited
kirill_pavlov@asiamiles.com