The Faculty.. DAP,Spark and beyond

Isabela Breton

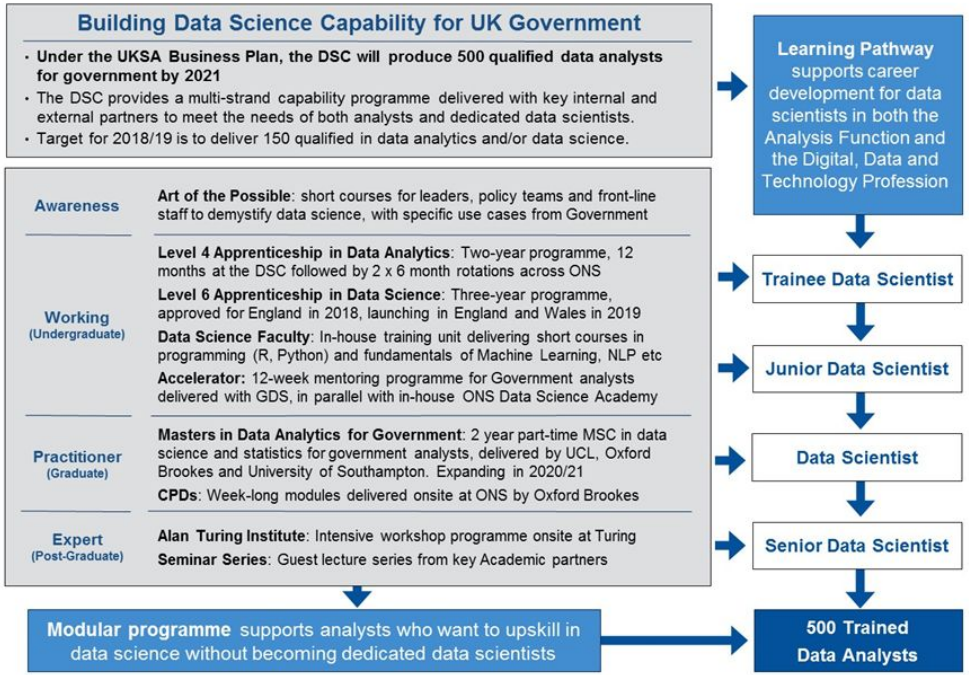
16 April, 2019

- Who we are..
- What we do..
- Who we are working with to support DAP..
- What we are doing to support DAP..
- Spark.. A very brief Introduction..

What is the Faculty?

Data Science Campus

The Faculty: Data science capability unit…




Data Science
Campus


GOV.UK


Department
for Work &
Pensions


Office for
National Statistics


HM Treasury


United Nations
Economic Commission
for Africa

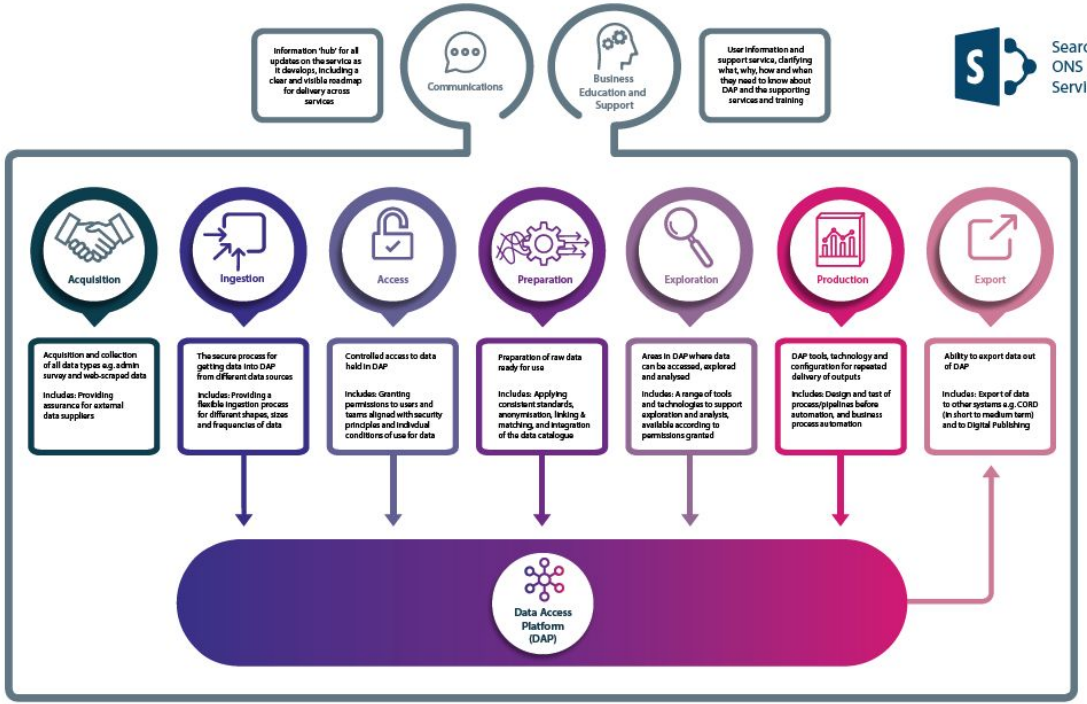
Data Science Campus

Specifically, we are working to support DAP, with:

- **DST (DAP-CATS)**
- **Learning Academy**
- **DAAS**
- **Departments**

Office for National Statistics

ONS **Data Service**



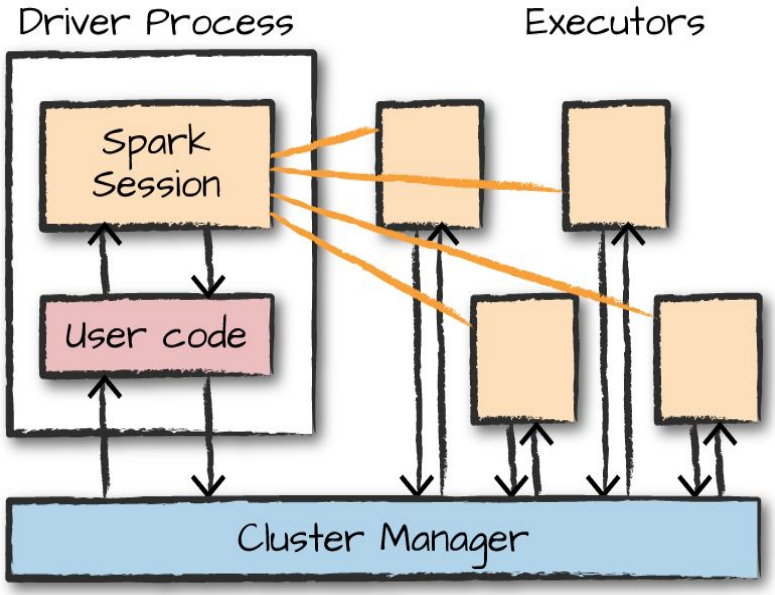
Search SharePoint: ONS Data Service Hub

onsdataservice@ons.gov.uk

What is Spark? And why is it important?

Data Science Campus

Spark is a **Distributed Computing Framework**.. used for **processing, querying and analysing Big Data**



Fundamentally **Spark** makes **processing, querying and analysing Big Data easier**

```
Item_Price_DataFrame.show(6)
```

```
## +-----------+----+-----+---------------+--------+
## |       Item|Code|Price|Last Year Price|Quantity|
## +-----------+----+-----+---------------+--------+
## |Black Chair|  22|  100|             70|      10|
## |White Table|   3|  500|            350|      50|
## | Floor Lamp|  16|   60|             50|       1|
## |White Table|   3|  500|            499|      20|
## |      Couch|  12| 1000|            900|       5|
## |White Table|   3|  500|            499|      20|
## +-----------+----+-----+---------------+--------+
```

How many languages do you need to know to communicate with Spark?

Just one! Python…pyspark

(With a little help from SQL)

Pyspark is **not really python**..

But close… what's the pandas equivalent?

```
Item_Price_DataFrame.show(6)
```

```
## +-----------+----+-----+---------------+--------+
## |       Item|Code|Price|Last Year Price|Quantity|
## +-----------+----+-----+---------------+--------+
## |Black Chair|  22|  100|             70|      10|
## |White Table|   3|  500|            350|      50|
## | Floor Lamp|  16|   60|             50|       1|
## |White Table|   3|  500|            499|      20|
## |      Couch|  12| 1000|            900|       5|
## |White Table|   3|  500|            499|      20|
## +-----------+----+-----+---------------+--------+
```

SQL.. yes you can use it with pyspark.. Spark

```
from pyspark import sql
from pyspark.sql import functions
Item_Price_DataFrame.filter('Price > 600').show() #SQL based command
#Below will have the same effect.
#Item_Price_DataFrame.filter(Item_Price_DataFrame['Price'] > 600).show()
#Item_Price_DataFrame.filter(Item_Price_DataFrame.Price > 600).show()
```

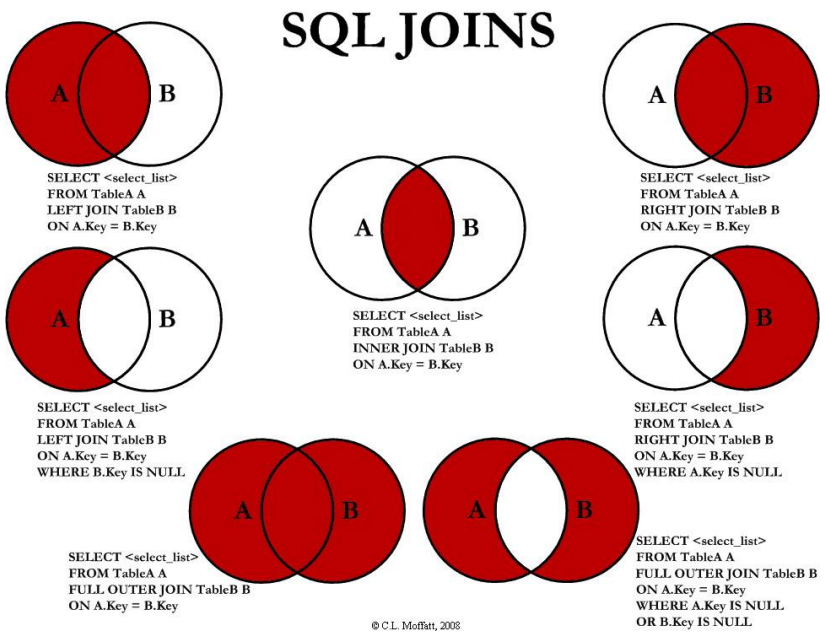
```
## +-----+----+-----+---------------+--------+
## | Item|Code|Price|Last Year Price|Quantity|
## +-----+----+-----+---------------+--------+
## |Couch|  12| 1000|            900|       5|
## +-----+----+-----+---------------+--------+
```

What uses are there for Spark?

Data Science Campus

Data manipulation: Joins

…joining, broadcasting and appending



Data Science Campus

Data wrangling ...

Machine learning…..

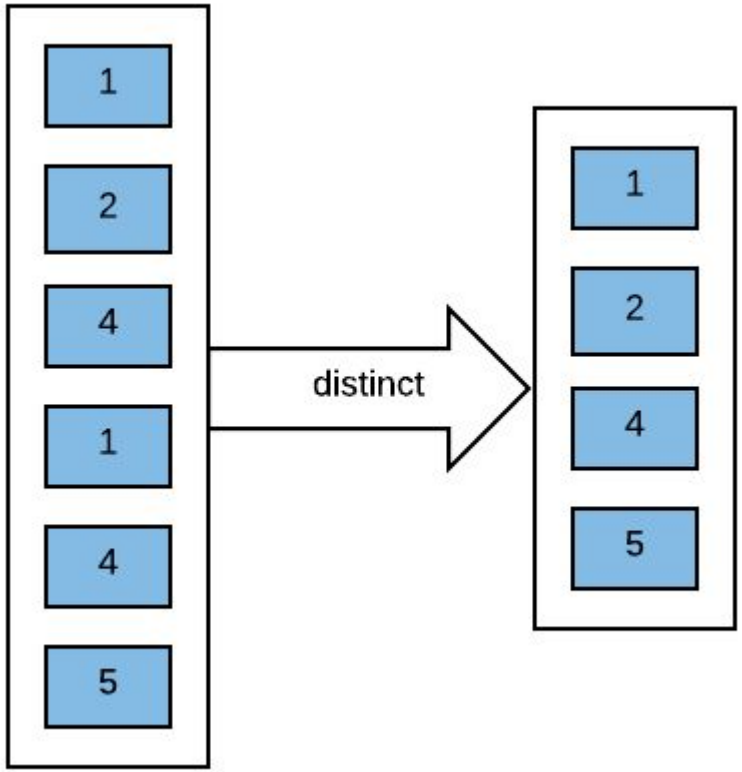
Simple classifier.. More info see the official guide https://spark.apache.org/docs/latest/ml-guide.html

```python
from pyspark.ml.classification import LogisticRegression

# Load training data
training = spark.read.format("libsvm").load("data/mllib/sample_libsvm_data.txt")

lr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8)

# Fit the model
lrModel = lr.fit(training)

# Print the coefficients and intercept for logistic regression
print("Coefficients: " + str(lrModel.coefficients))
print("Intercept: " + str(lrModel.intercept))

# We can also use the multinomial family for binary classification
mlr = LogisticRegression(maxIter=10, regParam=0.3, elasticNetParam=0.8, family="multinomial")

# Fit the model
mlrModel = mlr.fit(training)

# Print the coefficients and intercepts for logistic regression with multinomial family
print("Multinomial coefficients: " + str(mlrModel.coefficientMatrix))
print("Multinomial intercepts: " + str(mlrModel.interceptVector))
```

Data Science Campus

Find out more methods/applications..Api documentation:

https://spark.apache.org/docs/latest/sql-programming-guide.html



The faculty runs a Python in Spark course, which we advertise on eventbrite. https://www.eventbrite.co.uk/e/python-in-spark-30th-and-31st-may-2019-newport-tickets-60299273751

We have limited spaces so we prioritise:

- Staff which want to become mentors/champions
- Key business areas

If you are interested in the course email and a suitable date is not available email: datacampus@ons.gov.uk

- Who we are.. The Faculty
- What we do.. Data Science skills in government
- Who we are working with to support DAP
- What we are doing to support DAP
- Spark.. A very brief Introduction
- Where to find our course.. Eventbrite or email