SCALA FOR BIG DATA: THE BIG PICTURE

BIG DATA / AZURE / AWS / DATA LAKES

Anna-Maria Wykes
Data Engineering & Cloud Consultant

WHO AM I?

Senior Advancing Analytics Consultant

Data Engineering & Cloud

Over 14 years' experience working in Software & Data Engineering, most recently working with Scala, Kafka and various cloud tech

BSc in Multimedia Computing & Business, and a HND in Visual Communication

@annawykes

anna-maria-wykes-31939454

WHAT MAKES ME TICK?

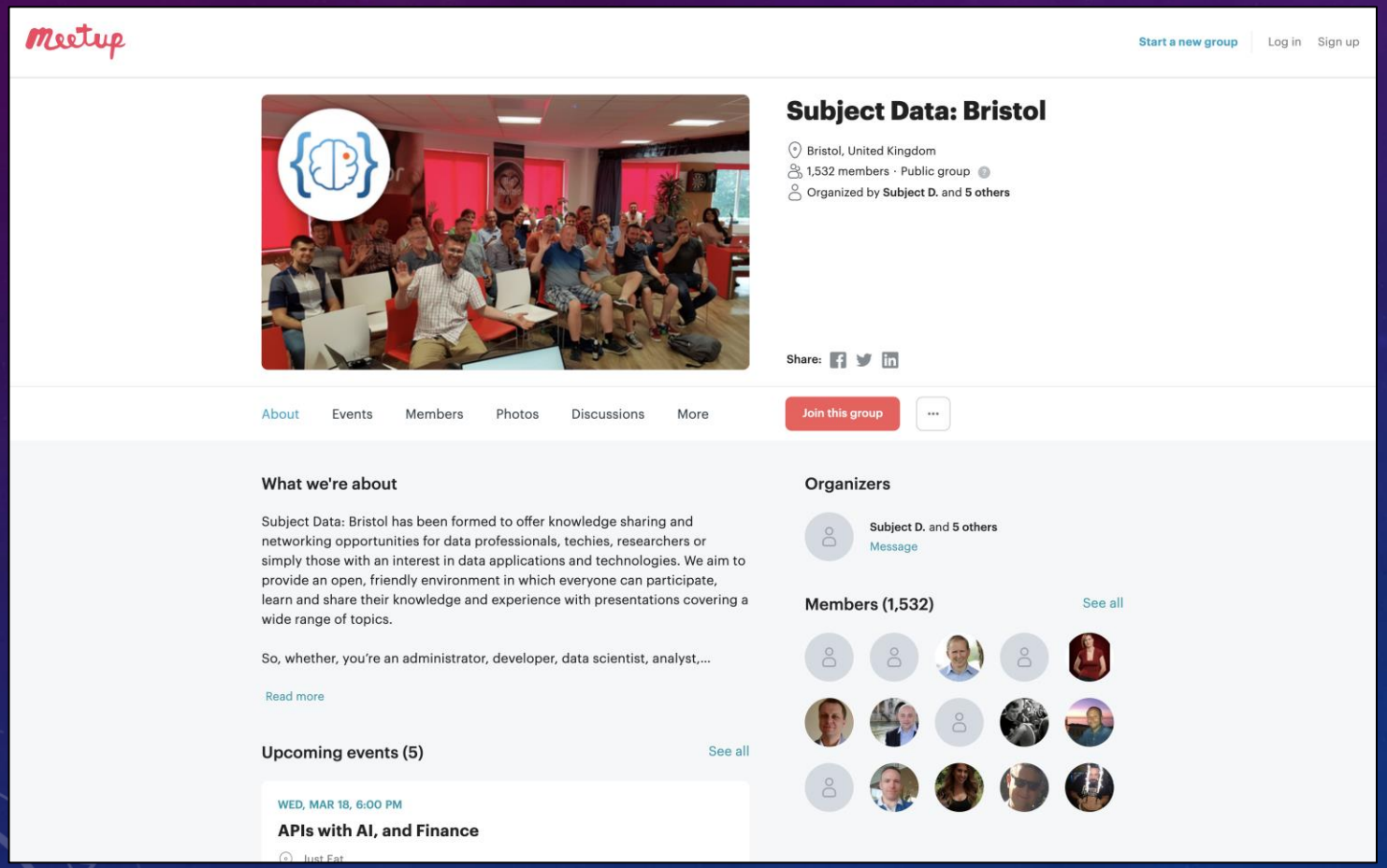
Passion for Data and strive to bring the worlds of Software Development and Data Science closer together.

Helped to organize/run local Code Clubs

Organize and volunteer at local events

Other areas of interest include UX, and Agile methodologies

MEETUPS AND COMMUNITY GROUPS



AGENDA: WHAT ARE WE GOING TO DO……

A BIT OF THEORY

Scala overview

Functional Programming

PRACTICAL

XML transformation to JSON

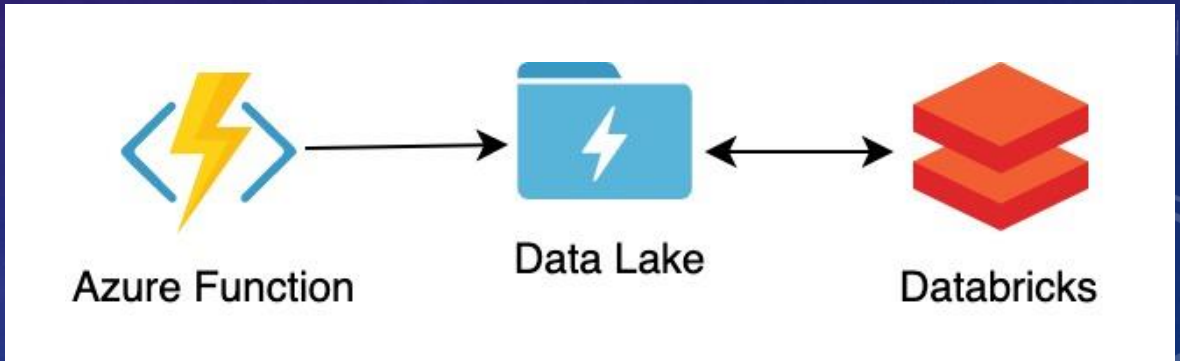
Write JSON Data to Data Lake Gen2

Refine data using Azure Databricks

THE (REAL WORLD) PROBLEM

Third party API providing old school XML responses that need to be transformed into JSON and then:

1) Written to an Event Queue
2) Stored in Data Lake



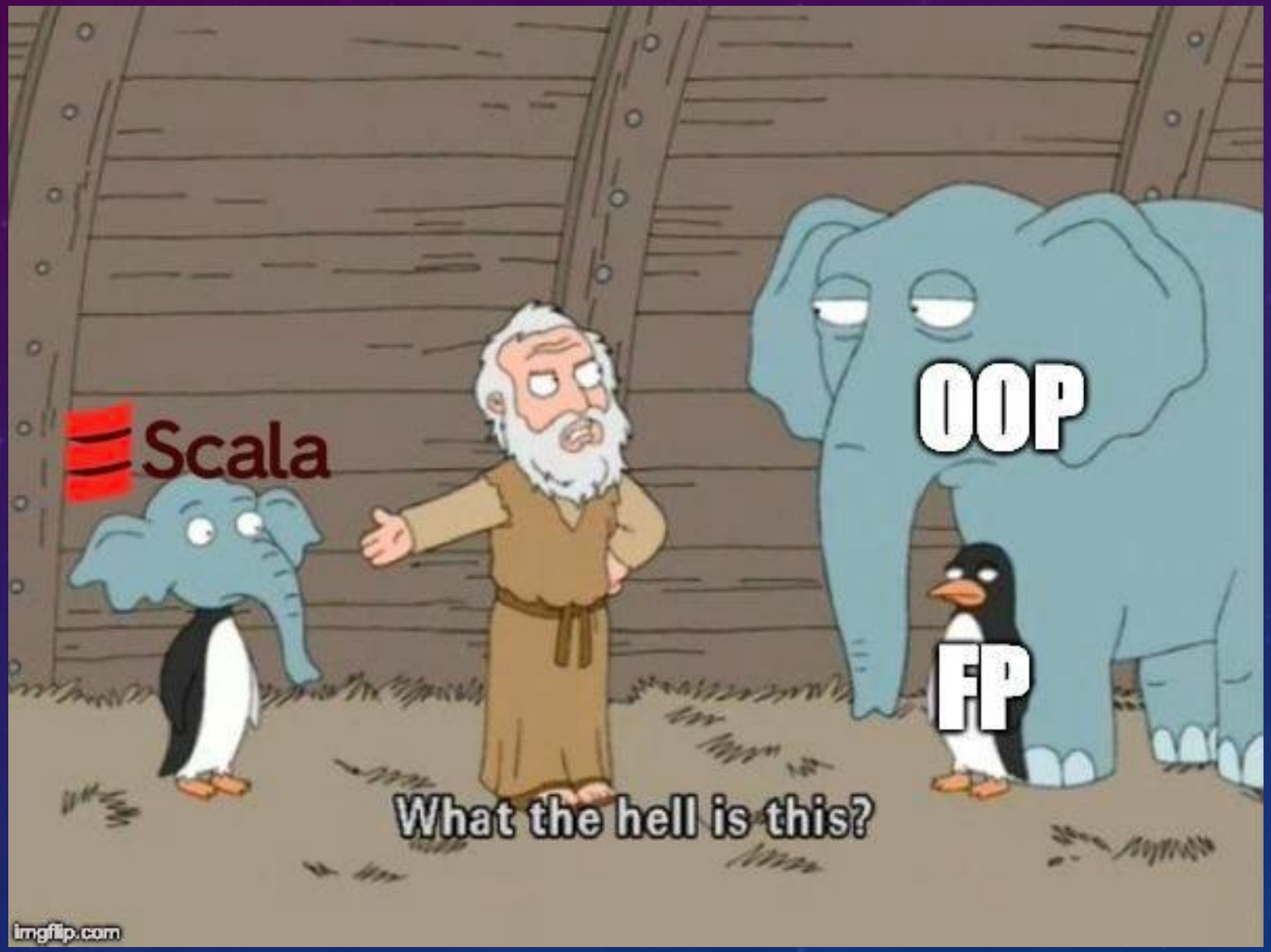
https://github.com/AnnaWykes/scala-for-big-data

WHAT IS SCALA?

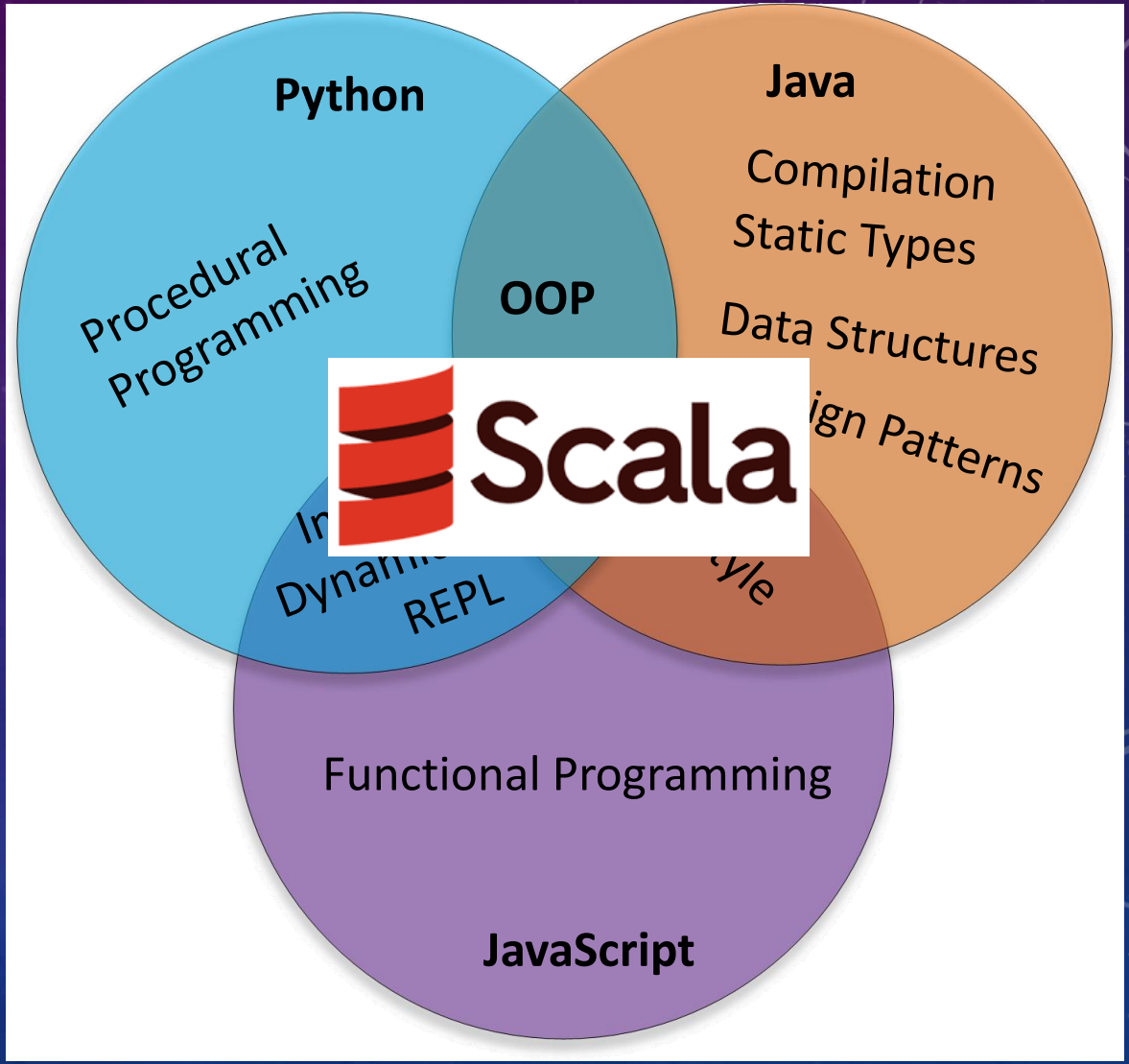
Object Orientated and **Functional** programming paradigms

What is **Scala used for**? A lot of things, ranging from **Machine Learning** to **Web Apps**

The name Scala stands for "scalable language." The language is so named because it was designed to grow with the demands of its users


Scala
OOP
FP
What the hell is this?
imgflip.com

THE ULTIMATE TEACHING LANGUAGE?



WHO USES SCALA

NETFLIX

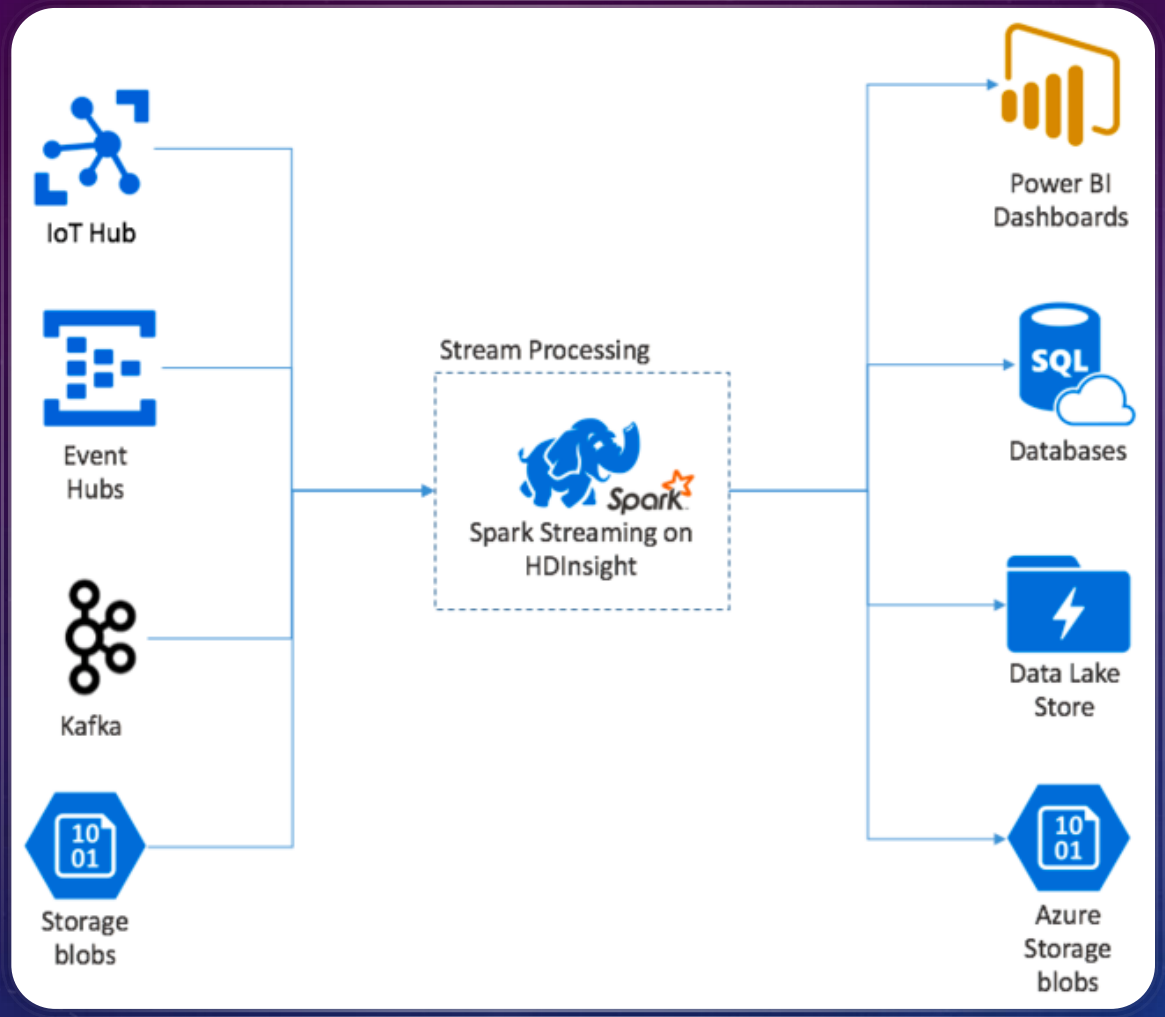
airbnb

Disney+

FOURSQUARE

SONY

SOFTWARE WRITTEN IN SCALA



SCALA AND ~~BIG~~ FAST DATA

"Scala has taken over the world of "Fast" Data"

Which is what some are calling the next wave of computation engines that rely more on the speed of data processing rather than the size of the batch, and the ability to process event streams in real-time.

Several prominent examples of that movement are Spark, Scalding, Kafka (including Kafka Streams), and Samza, which are rapidly gaining awareness and use

BUT STILL, WHY BOTHER….

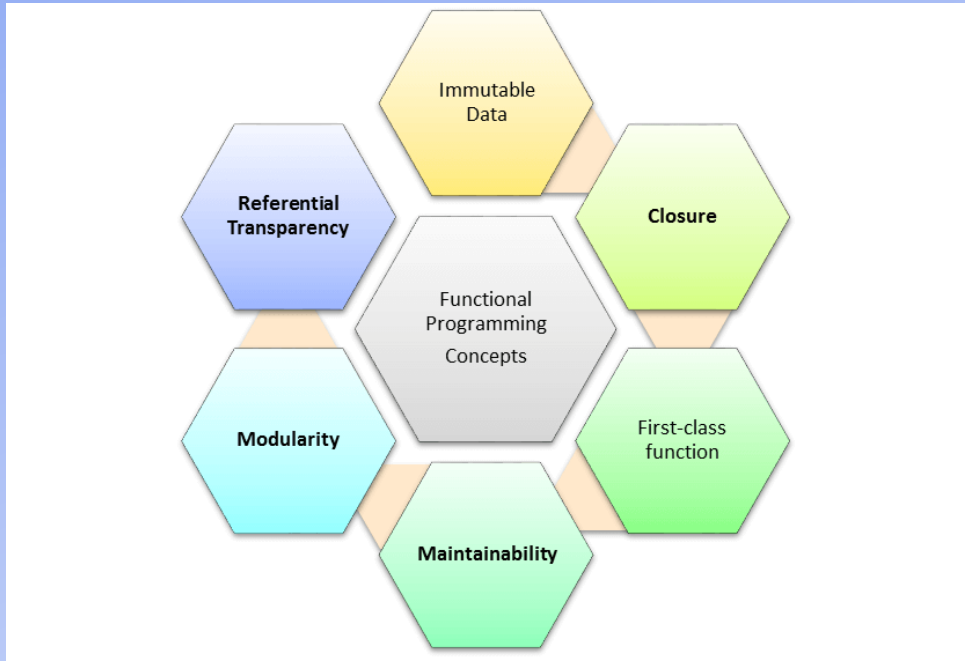
£

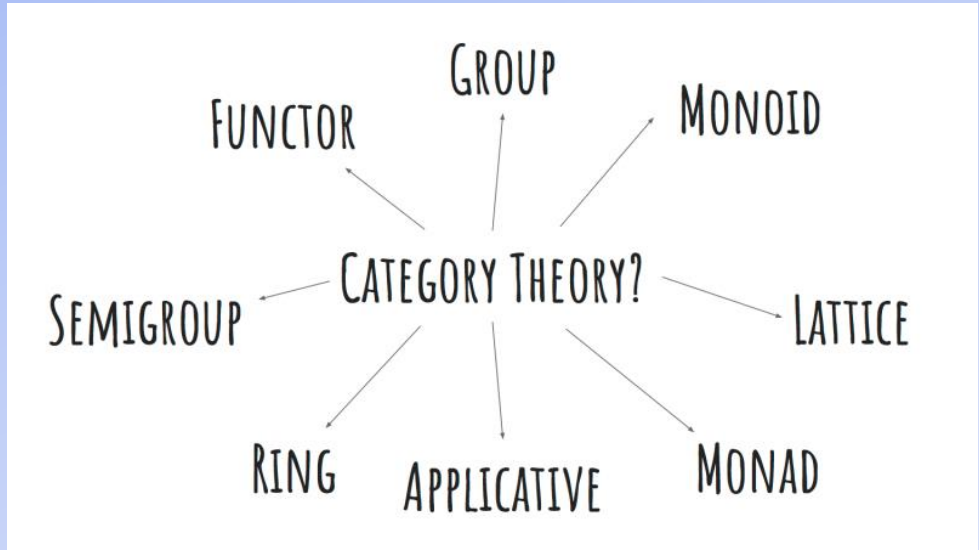
TYPE SAFETY

Type safety means that the compiler will validate types while compiling, and throw an error if you try to assign the wrong type to a variable

Types safety means you can’t turn a cat into a dog

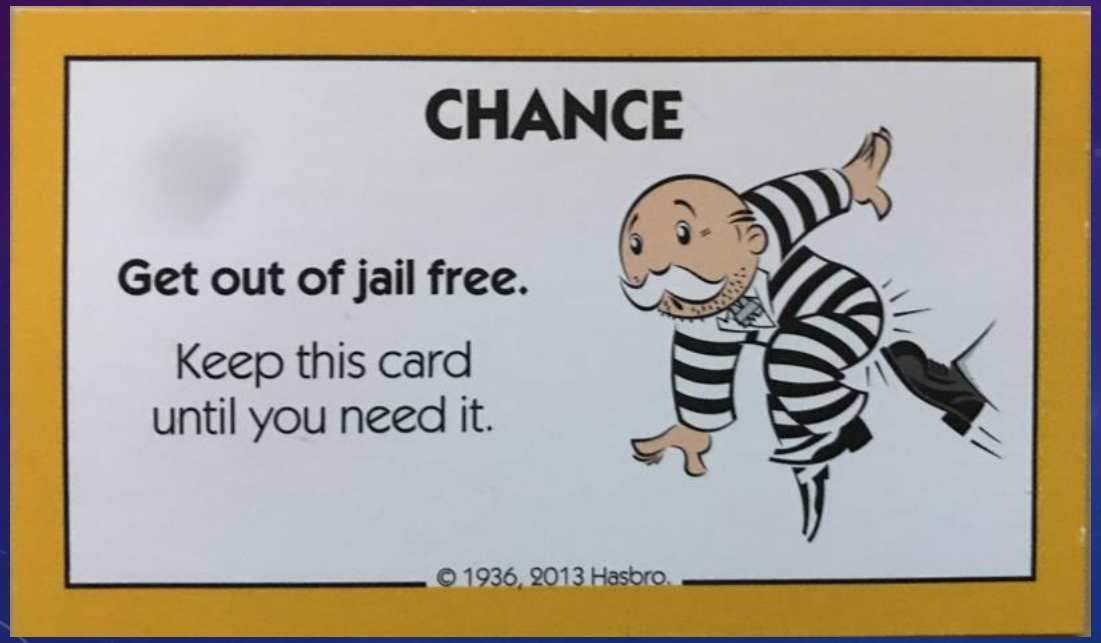
FUNCTIONAL PROGRAMMING

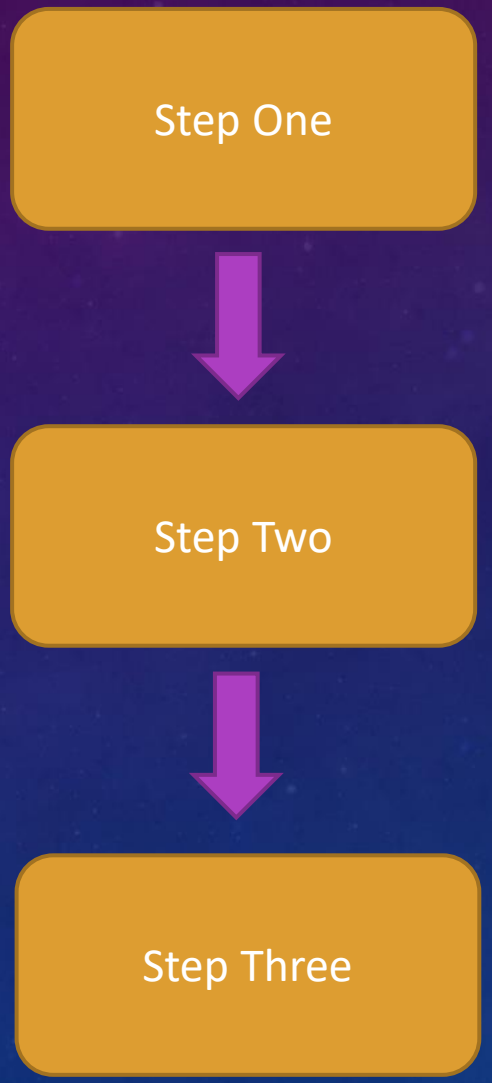




FUNCTIONAL PROGRAMMING: MONADS

A Monad is a sequence of events with a get out of jail card

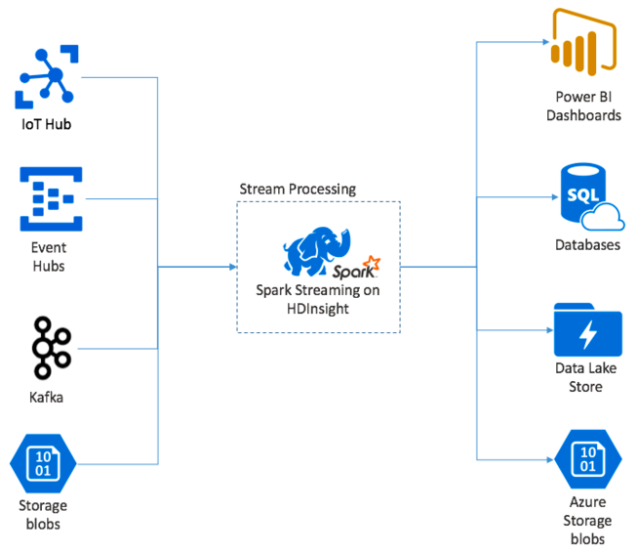


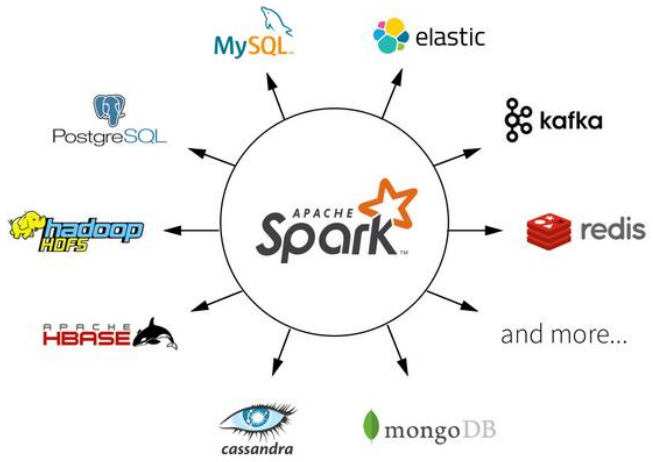


APACHE Spark™

databricks

APACHE SPARK



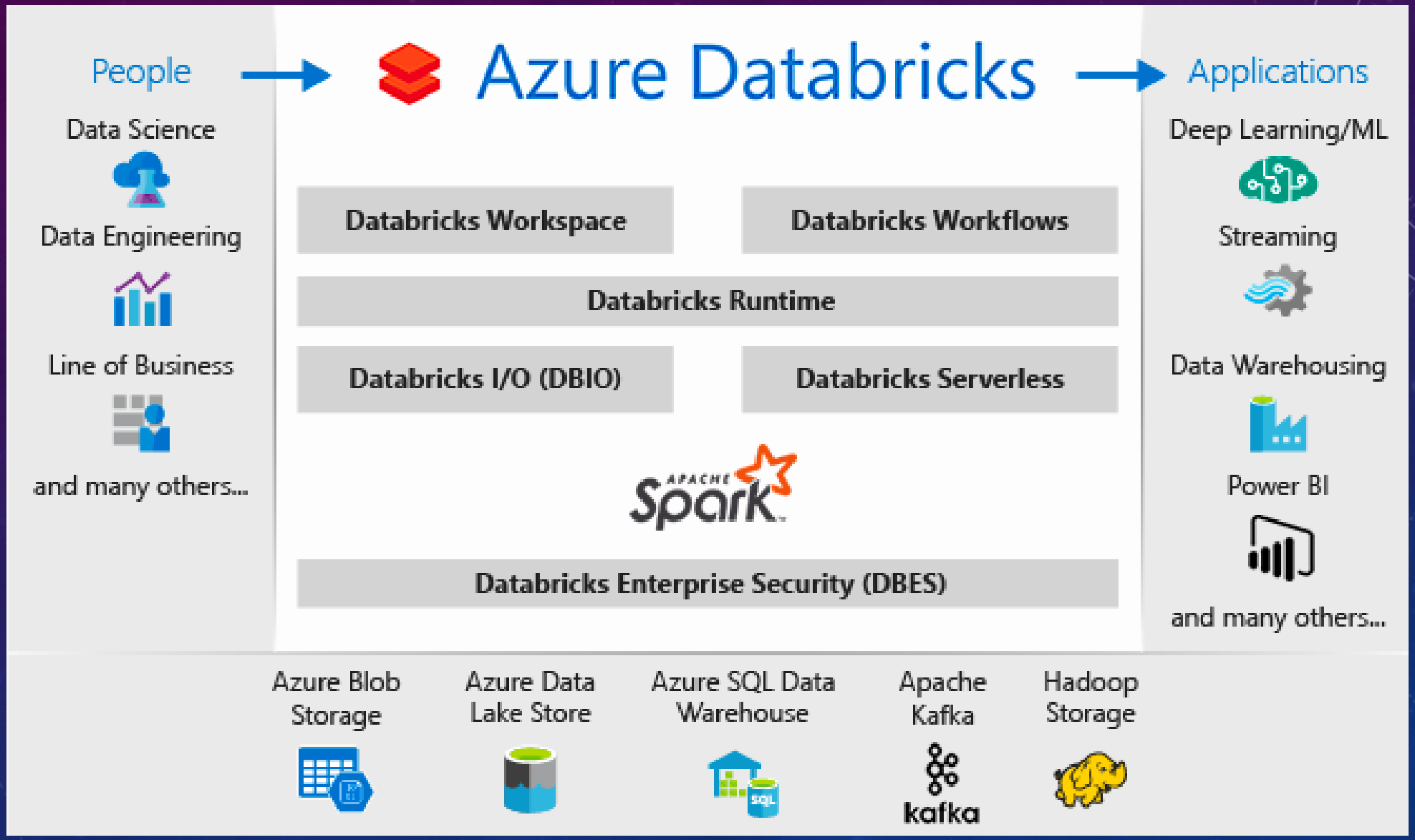


databricks

Databricks is an organization and big data processing platform founded by the creators of Apache Spark.

Databricks was founded to provide an alternative to the MapReduce system and provides a just-in-time cloud-based platform for big data processing clients.

Databricks was created for data scientists, engineers and analysts to help users integrate the fields of data science, engineering and the business behind them across the machine learning lifecycle. This integration helps to ease the processes from data preparation to experimentation and machine learning application deployment.


People
Data Science
Data Engineering
Line of Business
and many others...
Azure Databricks
Databricks Workspace
Databricks Workflows
Databricks Runtime
Databricks I/O (DBIO)
Databricks Serverless
APACHE Spark
Databricks Enterprise Security (DBES)
Applications
Deep Learning/ML
Streaming
Data Warehousing
Power BI
and many others...
Azure Blob Storage
Azure Data Lake Store
Azure SQL Data Warehouse
Apache Kafka
kafka
Hadoop Storage

BASICS OF SCALA

WHAT ARE WE GOING TO LOOK AT?

- Case Classes
- Option
- Pattern matching
- Map & FlatMap
- Reduce and Filter
- For Comprehensions

CASE CLASSES

Good for modelling immutable data Act like blueprints

CASE CLASSES

OPTIONS

Handles empty values without error

THE NULL MISTAKE

"I call it my billion-dollar mistake. It was the invention of the null reference in 1965" *— Tony Hoare, 2009*

HOW DOES SCALA HANDLE NULLS: **OPTIONS**

A **Scala Option** holds zero or one element of a type. This means that it is either a Some[T] or a none object. One place we get an **Option** value is through the get() method for a Map

SOME = Something

NONE = Nothing

MAP

Opens the box of data and transforms it accordingly

map
explained with emoji 😂

```
map([🐮, 🥔, 🐔, 🌽], cook)
=> [🍔, 🍟, 🍗, 🍿]
```

FLATMAP

Opens the box of data, transforms it, and flattens it accordingly

When you forget to use flatmap

REDUCE AND FILTER

Sum everything up Only get the data you need

filter, and reduce explained with emoji 😂

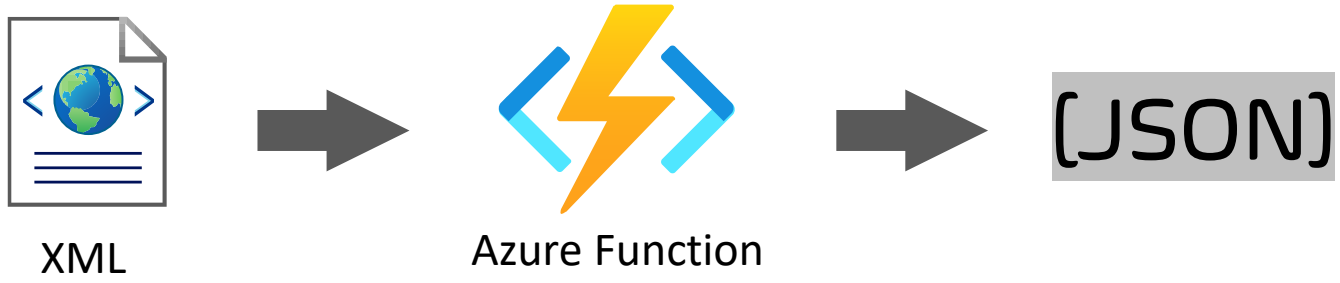
```
filter([🍔, 🍟, 🍗, 🍿], isVegetarian)
=> [🍟, 🍿]

reduce([🍔, 🍟, 🍗, 🍿], eat)
=> 💩
```

FOR COMPREHENSION

Monads

DEMO TIME:

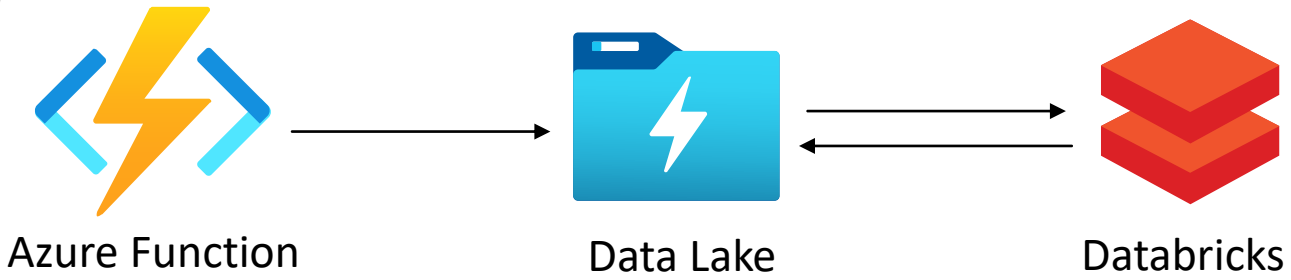
AZURE FUNCTION TO TRANSFORM DATA



Monad (Functional Programming style)

TIME TO REFINE OUR DATA

MOVE DATA FROM RAW FOLDER IN DATA LAKE INTO REFINED FOLDER



USEFUL LINKS

Learning

https://www.scala-exercises.org/

https://scala.epfl.ch/

https://www.coursera.org/specializations/scala

https://typelevel.org/cats/

https://medium.com/disney-streaming/tagged/thisweekinscala

Conferences

https://scaladays.org/

https://scala.world/

THANK YOU

https://github.com/AnnaWykes/scala-for-big-data

anna-maria-wykes-31939454

@annawykes