Five Things...
(we wish we had known)

British Gas Connected Homes

Josep Casals - Lead Data Engineer
Jim Anning - Head of Data & Analytics

hive

Waking Up
EVERY HOME
IN Britain by
2020

Heating

HERO
OF
the Week
ROB

RUN CLUB
WHEN? WEDNESDAYS
12.30pm
WHERE? REGENT'S PARK
AND BACK (5K)

SMEG

Gresse Street

Rathbone Place

Lakeside West

British

JOINING THE DOTS

MY SMART METER
LETS ME SEE
HOW MUCH I'M
Spending
ON MY ENERGY...

SUPPORTED BY MY
SMART ENERGY REPORT
WHICH HELPS ME UNDERSTA
WHAT AFFECTS my
energy
usage ...

...I THEN GET A
bill
BASED ON ACCURATE
&

1st

Hive : Control Your Heating from your Phone

Connected Boiler: Proactive Maintenance

MyEnergy: Understand your Energy Usage


Heating control
Schedule
Manual
Off
18.5°
inside now
20.5°
20°
19.5°
NW1
5°
heating schedule
next 20:00 - 22:00
hot water
control
hive
SUN 10:26
18.0°C
Hotter
Cooler
hive
hive
Central heating

ok

e

min

max

9

12 h

15

6

18

3

24 h

21

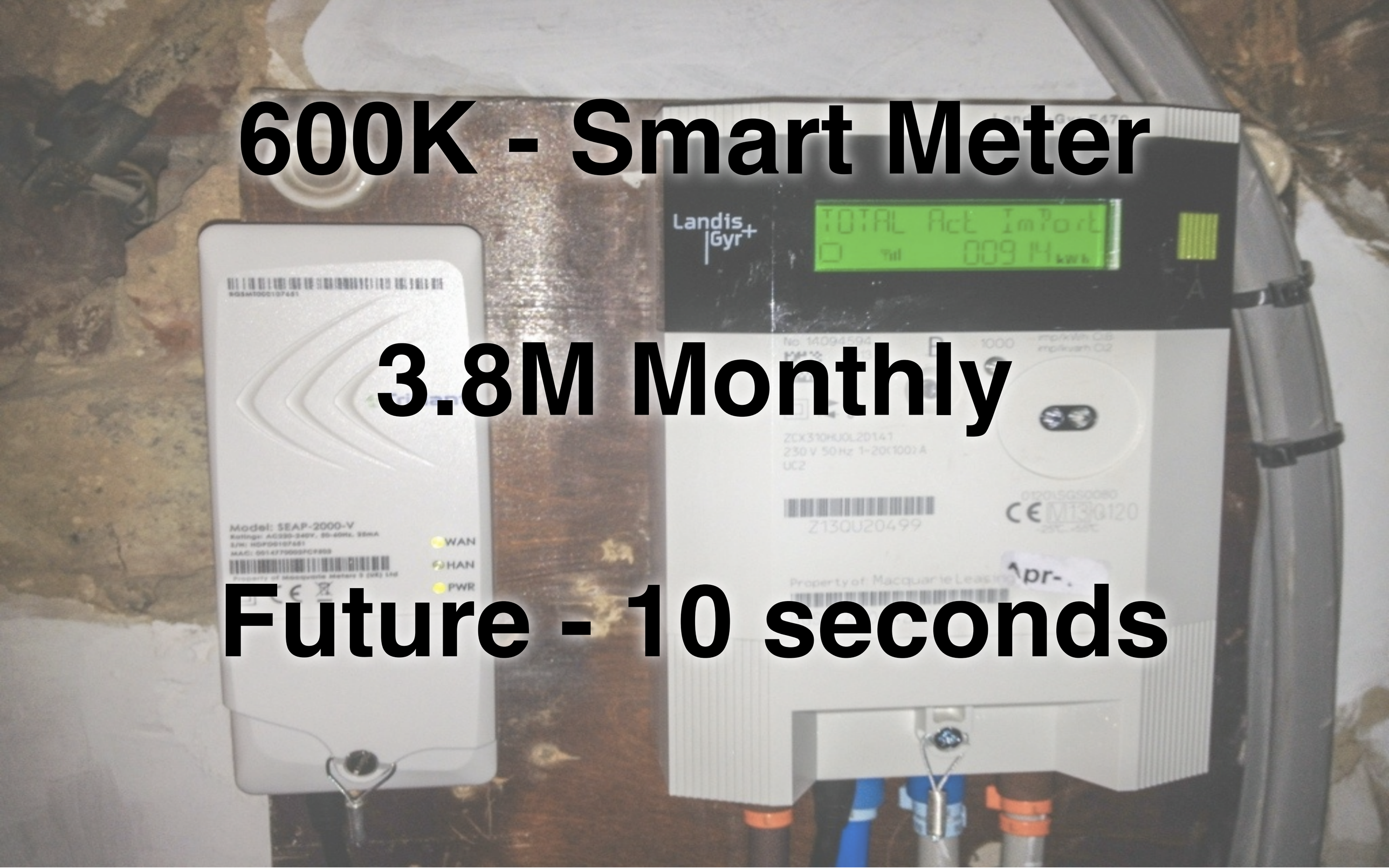
advance

advance

11

min

max


600K - Smart Meter
3.8M Monthly
Future - 10 seconds
Landis Gyr+
TOTAL Act ImPort
009 M kWh
Model: SEAP-2000-V
WAN
HAN
PWR
ZCX310HJ0L20141
230 V 50 Hz 1-20(100) A
UC2
Z130U20499
Property of Macquarie Leasing

British Gas
Looking after your world

At Home | Business | Corporate

About us | Contact us | Emergencies

Search

Products & services | Smarter living | Help & advice | Your account

› Your account › **my energy**

- **my energy**
- My usage
- Energy saving advice

my energy

‹ **January 2015 £367.17** ›

View graph

Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Highest day
£15.15
Sat 17th

Average day
£11.84

Lowest day
£8.54
Sat 10th

Send us a message

Give feedback or ask for help

Include a screenshot of this page

Next

60000

45000

30000

15000

0

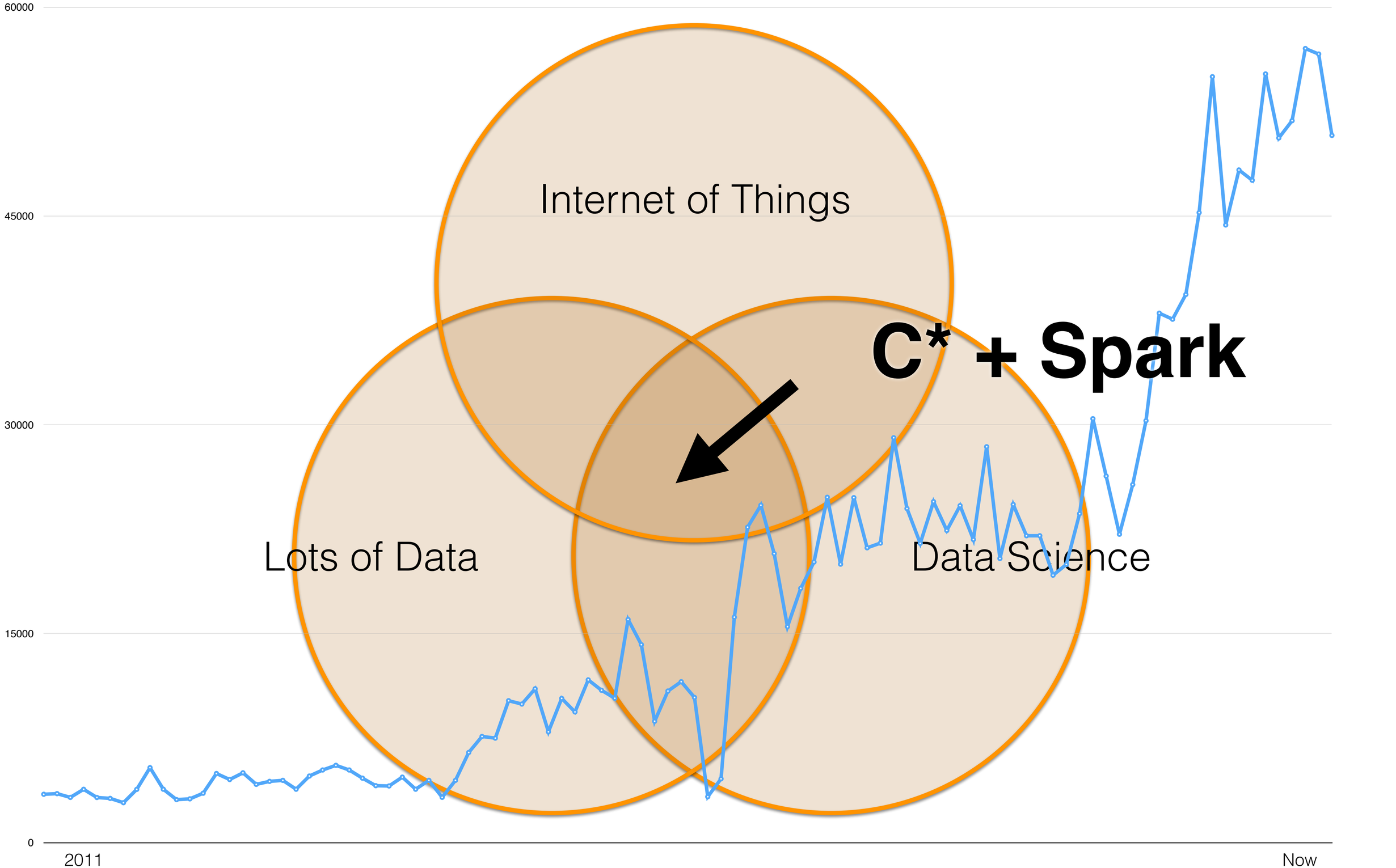
2011

Now

Internet of Things

Lots of Data

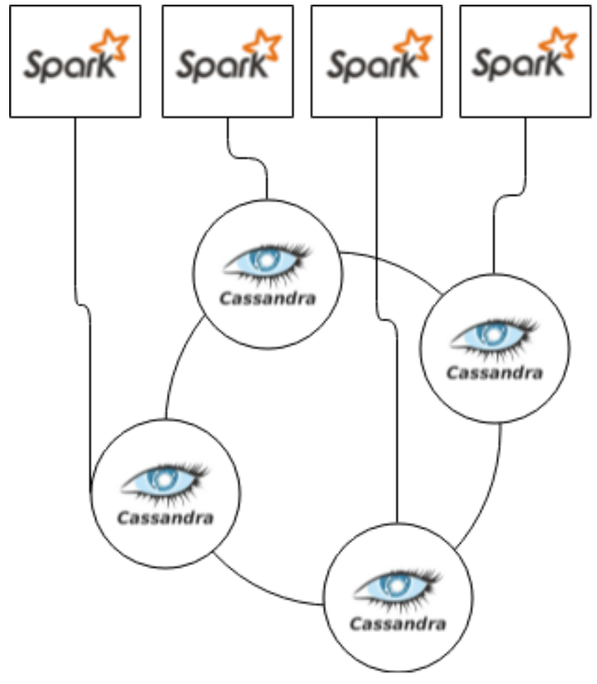
Data Science

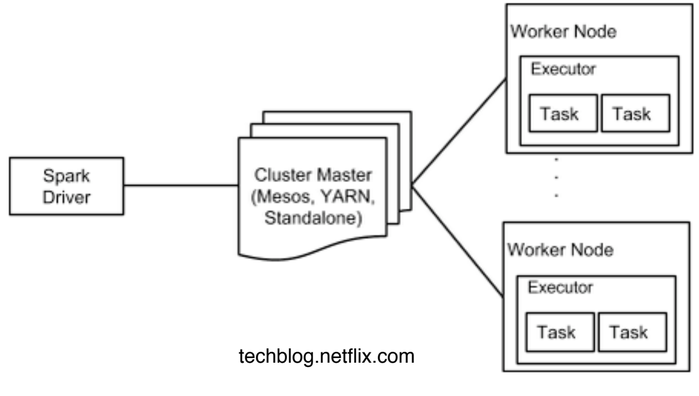

60000
45000
30000
15000
0
2011
Now
Internet of Things
Lots of Data
Data Science
C* + Spark

Lesson 1 : Not to race against bicycles

Spark is for parallel execution

- Makes sense when we have jobs that can't run on a single machine
- The Spark master needs to distribute the job to workers
- If the job shuffles all data to one single node, parallelism is lost
- For small tasks, many times a simple script is better

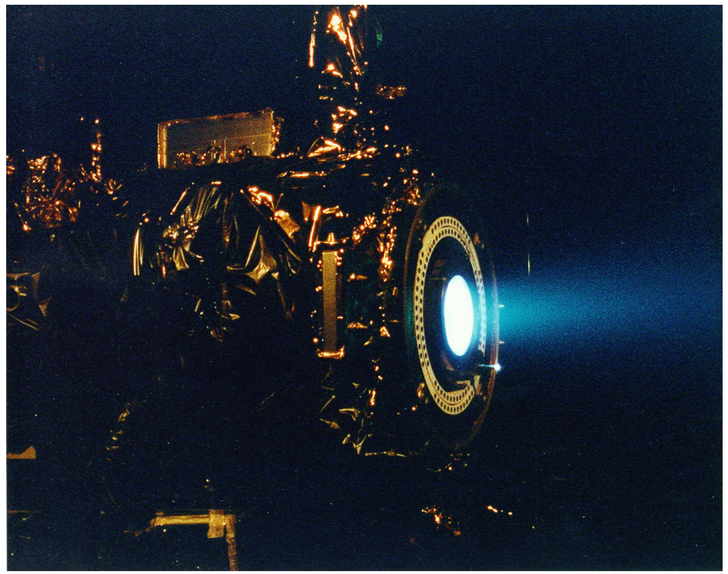




techblog.netflix.com

Things that look like a Spark / C* cluster

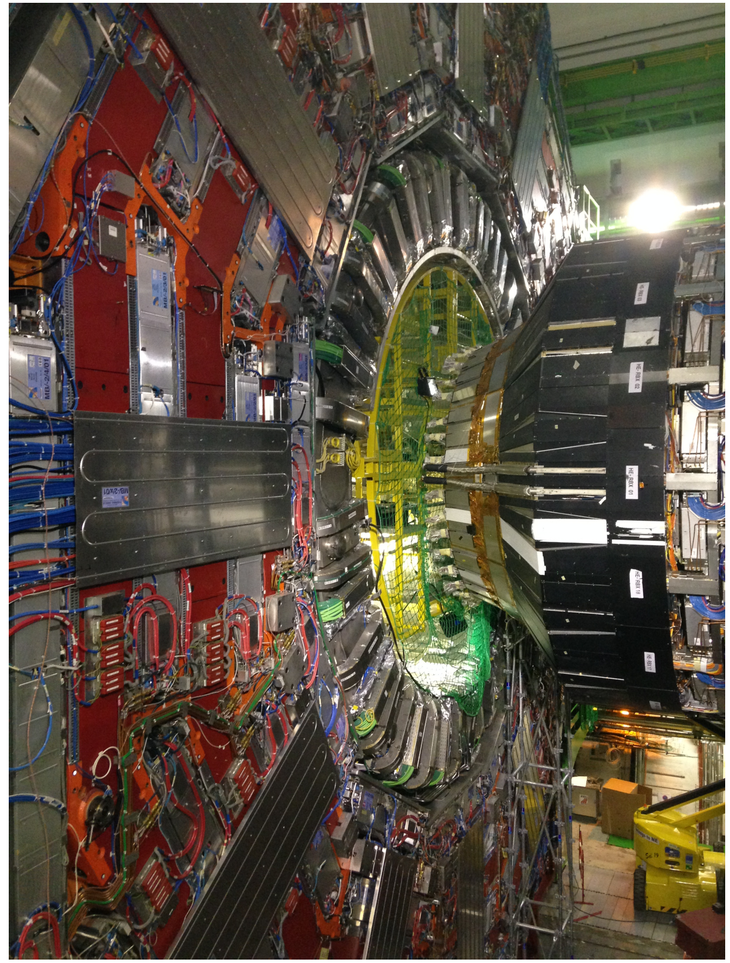
A Ion Thrust Engine



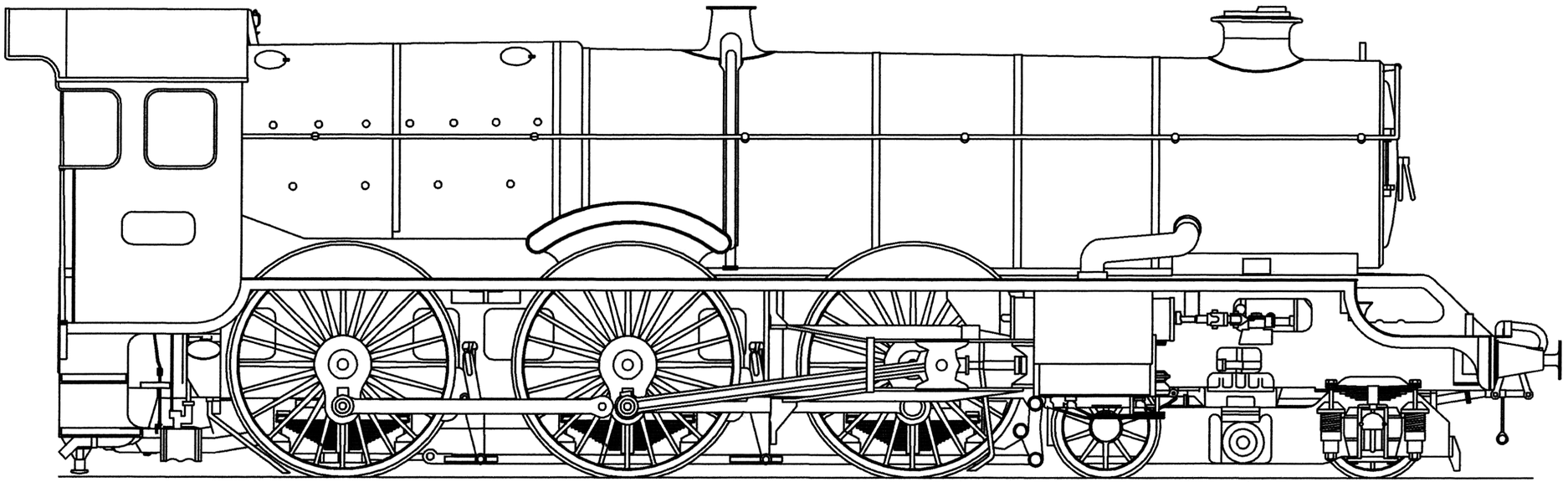
- It starts slow but in the long run goes very fast

- It can achieve big energies
- It takes a lot of fine tuning

A Large Hadron Collider



Who wins?

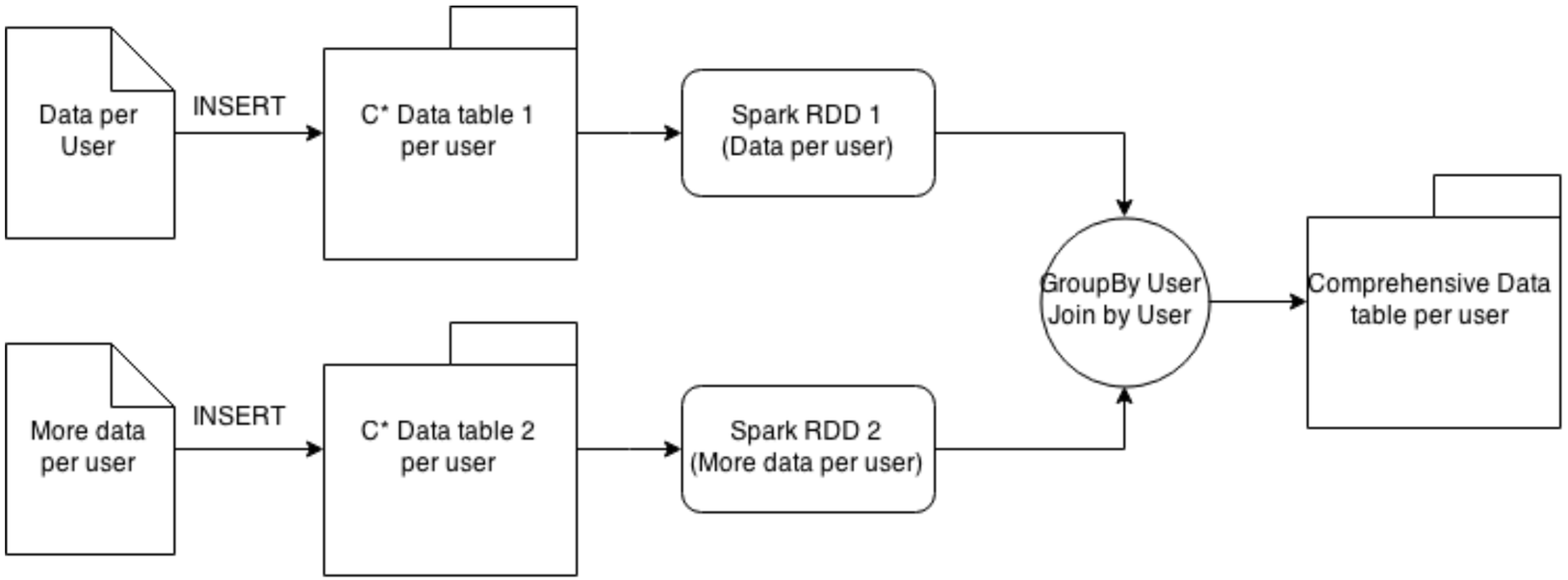


It depends on how far you go…

Lesson 2 : Not to use Spark too much

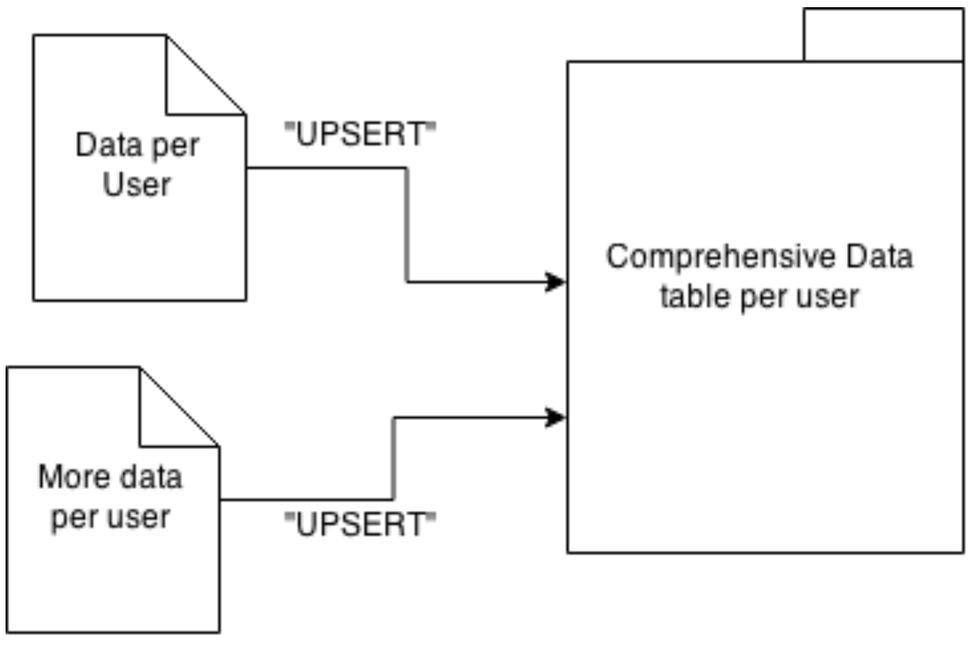
Joining data from multiple sources

Think twice when you do that



Upserting data from multiple sources

Do that if possible



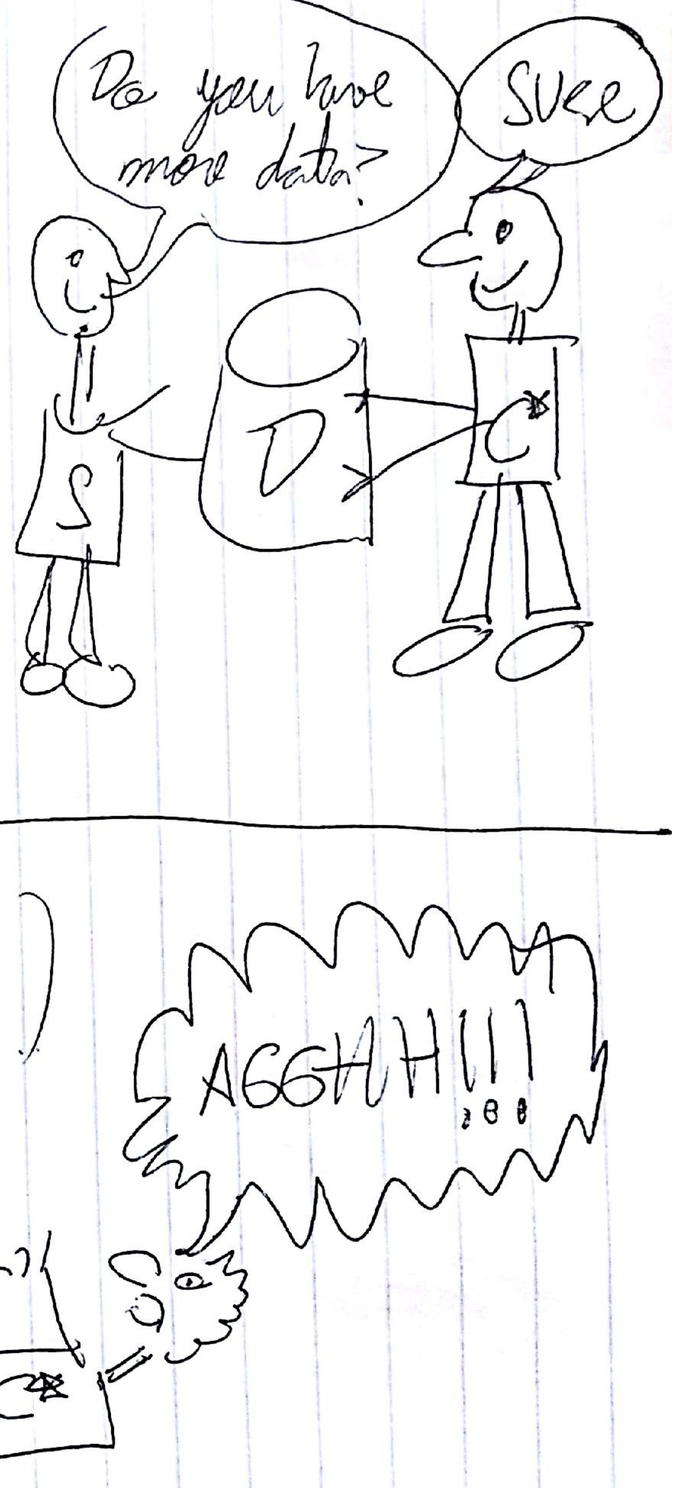
Upserting data from multiple sources

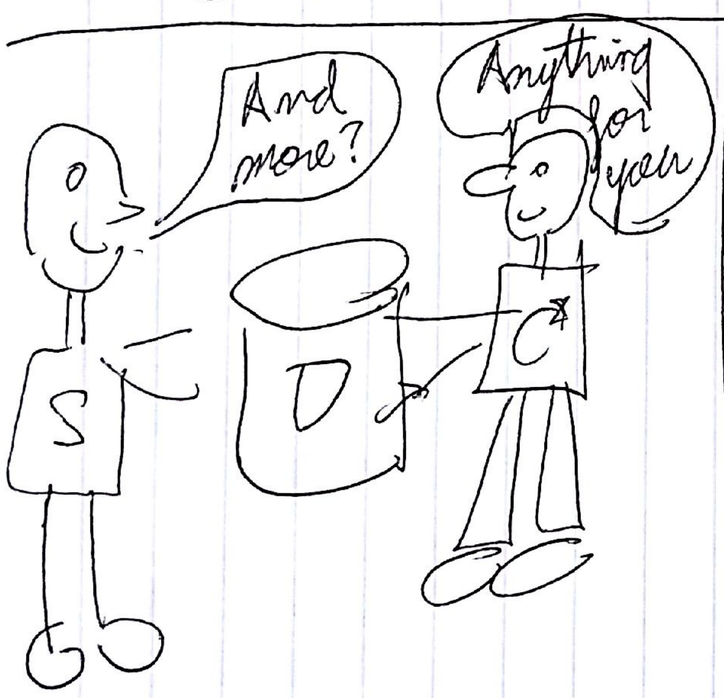
```
CREATE KEYSPACE meter_data
WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };

CREATE TABLE meter_data.consumption (
    meter_id text,
    elec_read_kwh bigint,
    gas_read_m3 bigint,
    read_date timestamp,
PRIMARY KEY ((meter_id),read_date));

INSERT INTO meter_data.consumption (meter_id, elec_read_kwh, read_date) VALUES ('10293847856', 2567,'2015-04-20');

INSERT INTO meter_data.consumption (meter_id, gas_read_m3, read_date) VALUES ('10293847856', 18363,'2015-04-20');

SELECT * FROM meter_data.consumption ;

 meter_id    | read_date                | elec_read_kwh | gas_read_m3
-------------+--------------------------+---------------+-------------
 10293847856 | 2015-04-20 00:00:00+0100 |          2567 |       18363
```

Lesson 3 : Spark is stronger than Cassandra


Hey, do you have some data?
Sure...
S
C


Thanks
Here you go
S
D


Do you have more data?
Sure
S
D
C


And more?
Anything for you
S
D
C

I'm finished!!
S
D

AGGHH!!!
C

Spark Properties & Cassandra-specific properties tuning

Write properties

You can set the following properties in SparkConf to fine tune the saving process.

spark.cassandra.output.batch.size.bytes

Default = auto. Number of bytes per single batch. The default, auto, means the connector adjusts the number of bytes based on the amount of data.

spark.cassandra.output.consistency.level

Default = LOCAL_ONE. Consistency level to use when writing.

spark.cassandra.output.concurrent.writes

Default = 5. Maximum number of batches executed in parallel by a single Spark task.

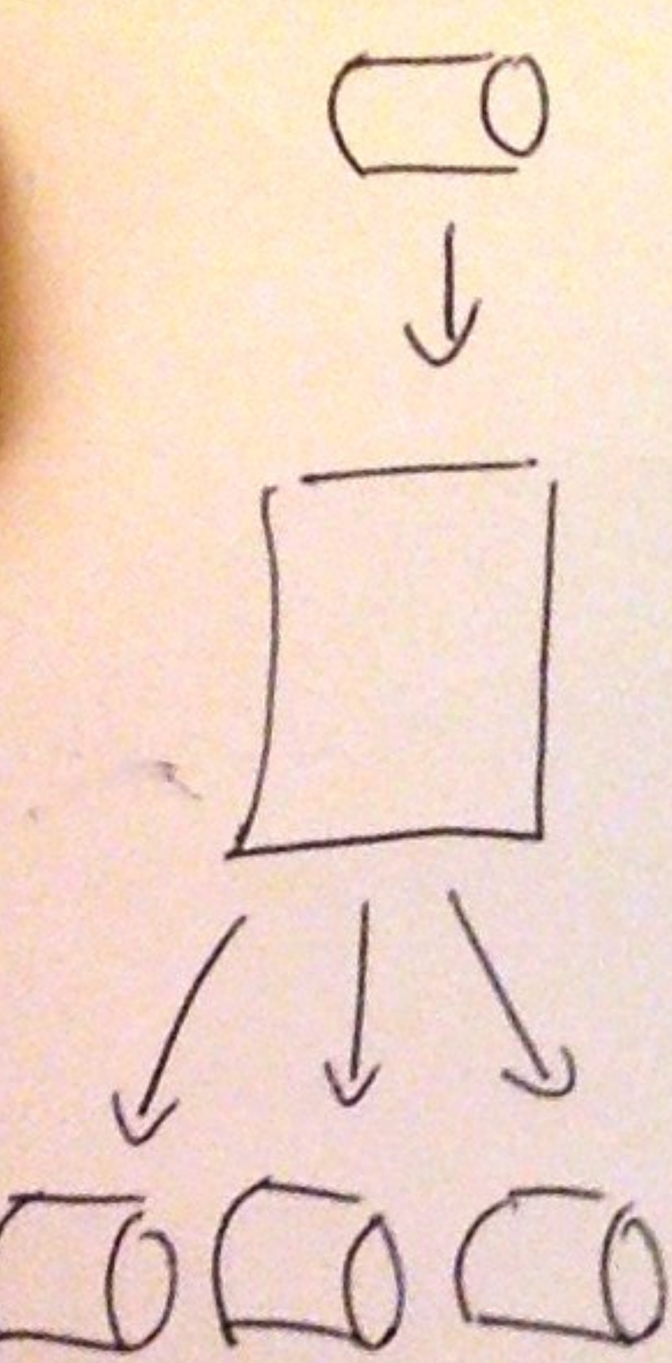
spark.cassandra.output.batch.size.rows

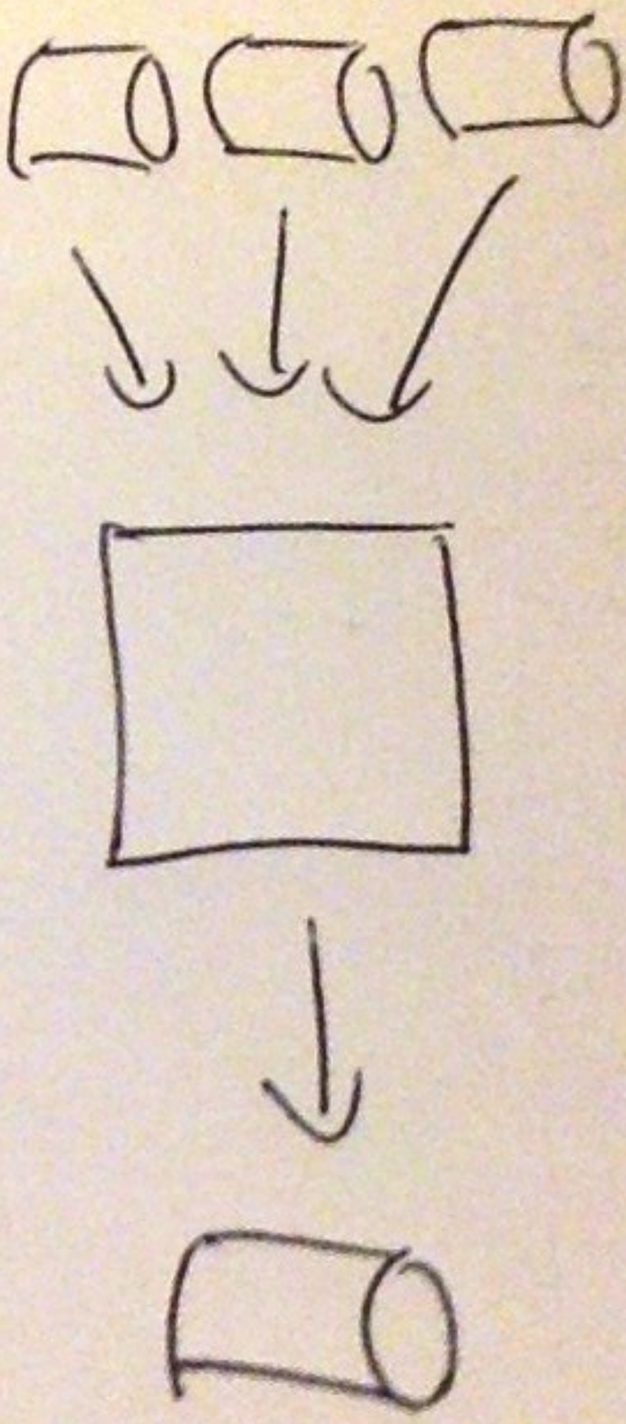
Default = 64K. The maximum total size of the batch in bytes.

Lesson 4 : Mindset

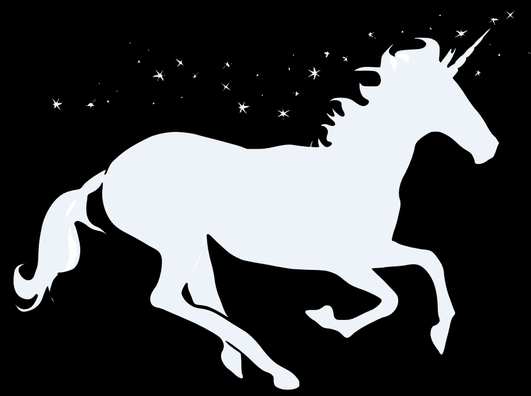
Fairmont
HOTELS & RESORTS

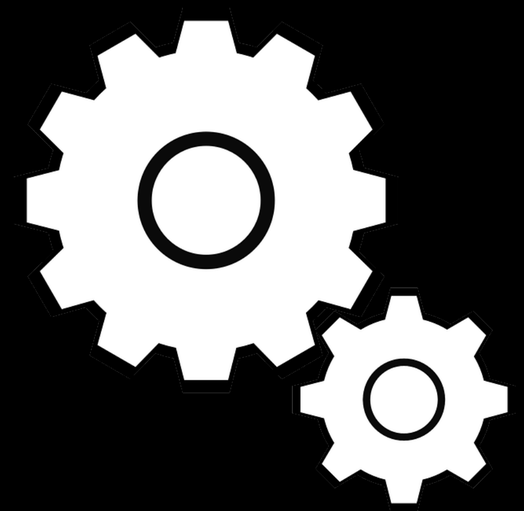
fairmont.com





Lesson 5 : Velocity





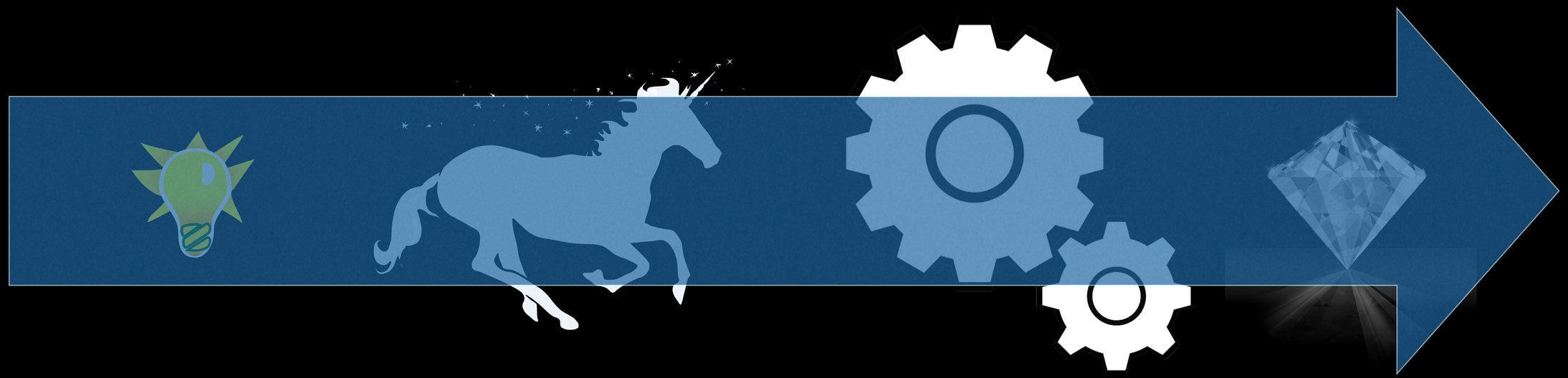
Idea

Data Science

Data Engineering

Data Operations

Value



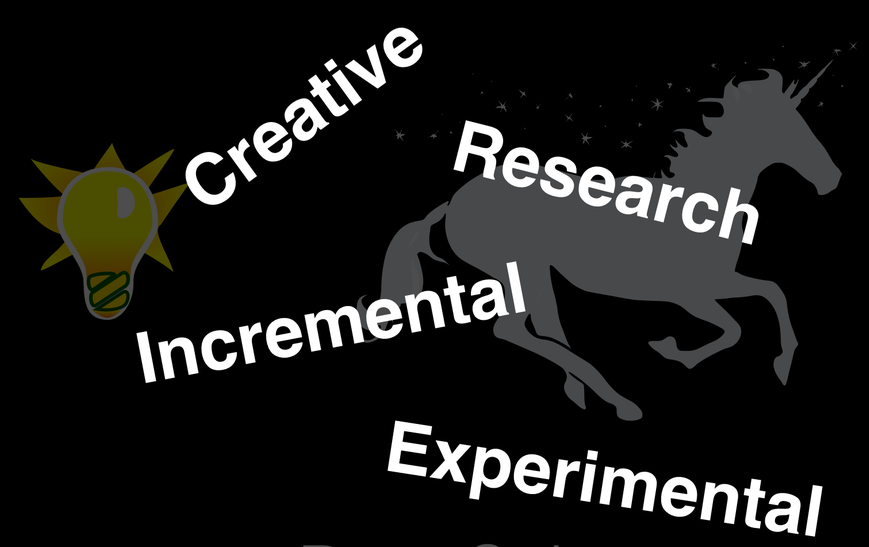
Idea

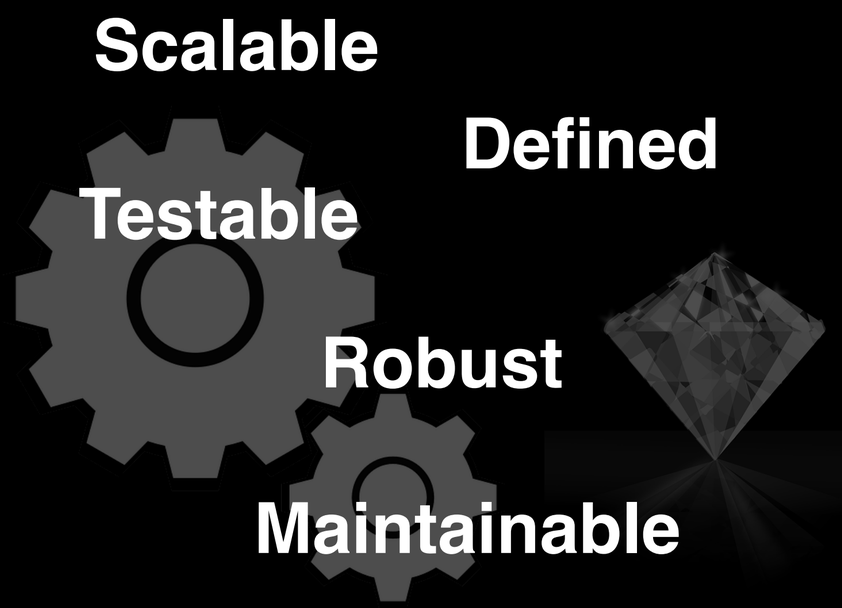
Data Science

Data Engineering

Data Operations

Value





Idea

Data Science

Data Engineering

Data Operations

Value

R

Python

Java

Scala

Creative

Research

Incremental

Experimental

Small Datasets

Idea

Data Science

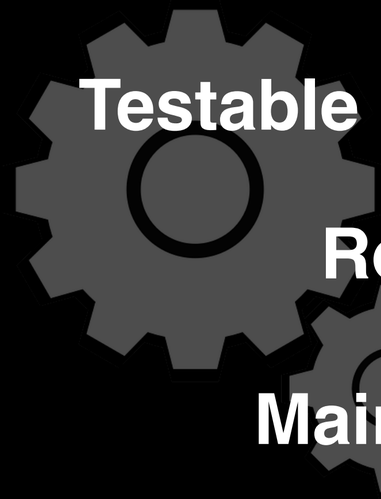
Offline

Single Machine

Scalable

Defined

Testable



Robust

Maintainable

Data Engineering

#BigData

Value

Data Operations

Clustered

Realtime

R

Python

Creative

Research

Incremental

Experimental

Small Datasets

Idea

Data Science

Offline

Single Machine

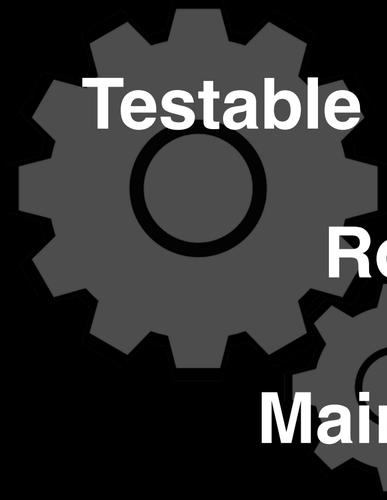
Java

Scala

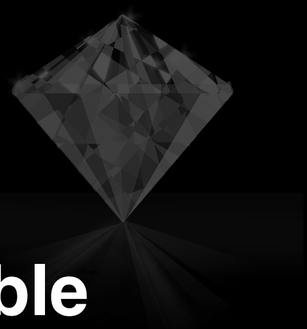
Scalable

Defined

Testable



Robust



Maintainable

Data Engineering

Value

#BigData

Data Operations

Clustered

Realtime

R

Python

Java

Scala

Creative

Research

Incremental

Scalable

Defined

Robust

Maintainable

Experimental

Small Datasets

Idea

Data Science

Data Engineering

Value

Offline

#BigData

Data Operations

Clustered

Single Machine

Realtime

Thankyou

@JimAnning : jim.anning@bgch.co.uk
@Jcasals : josep.casals@bgch.co.uk