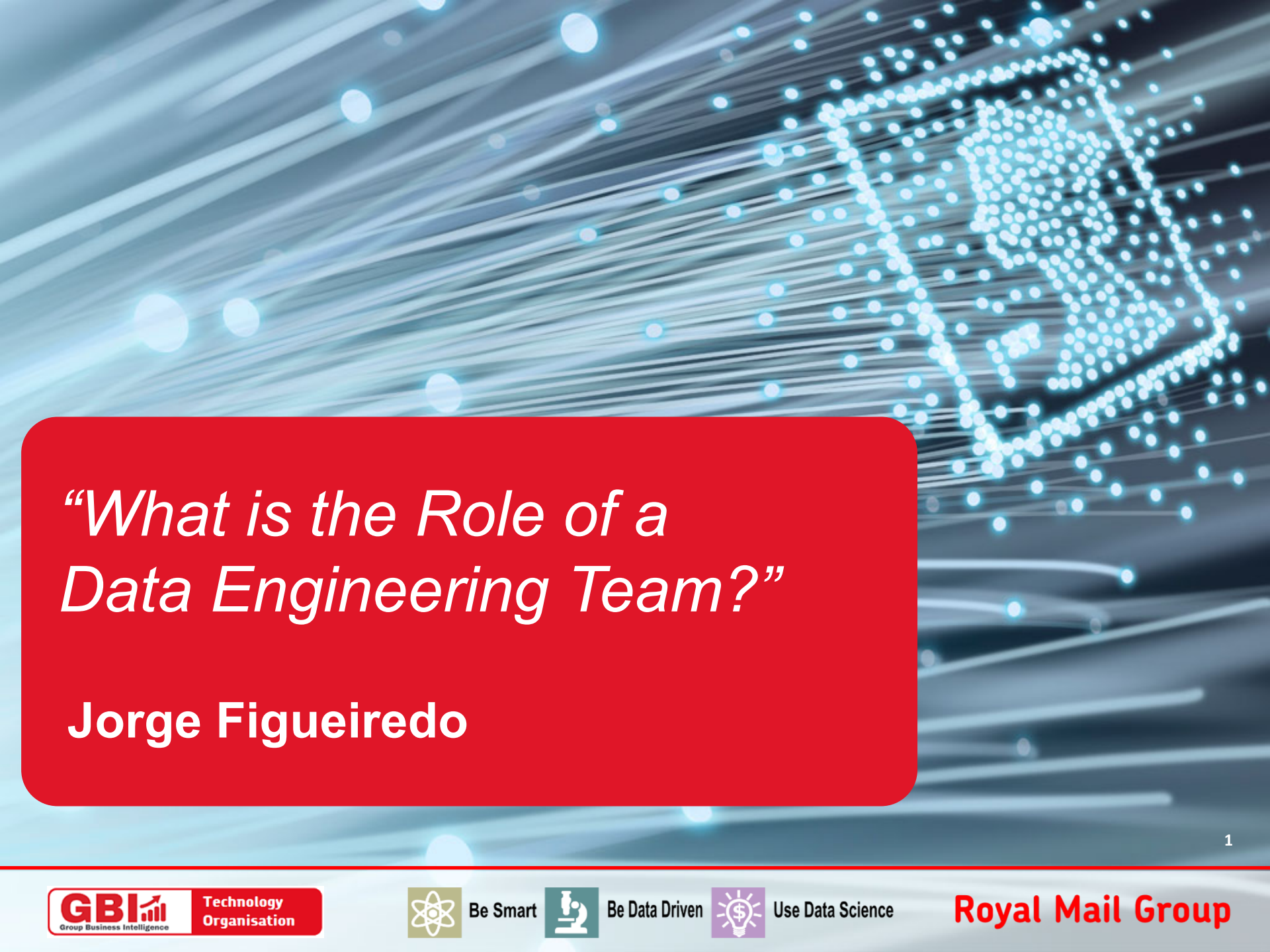
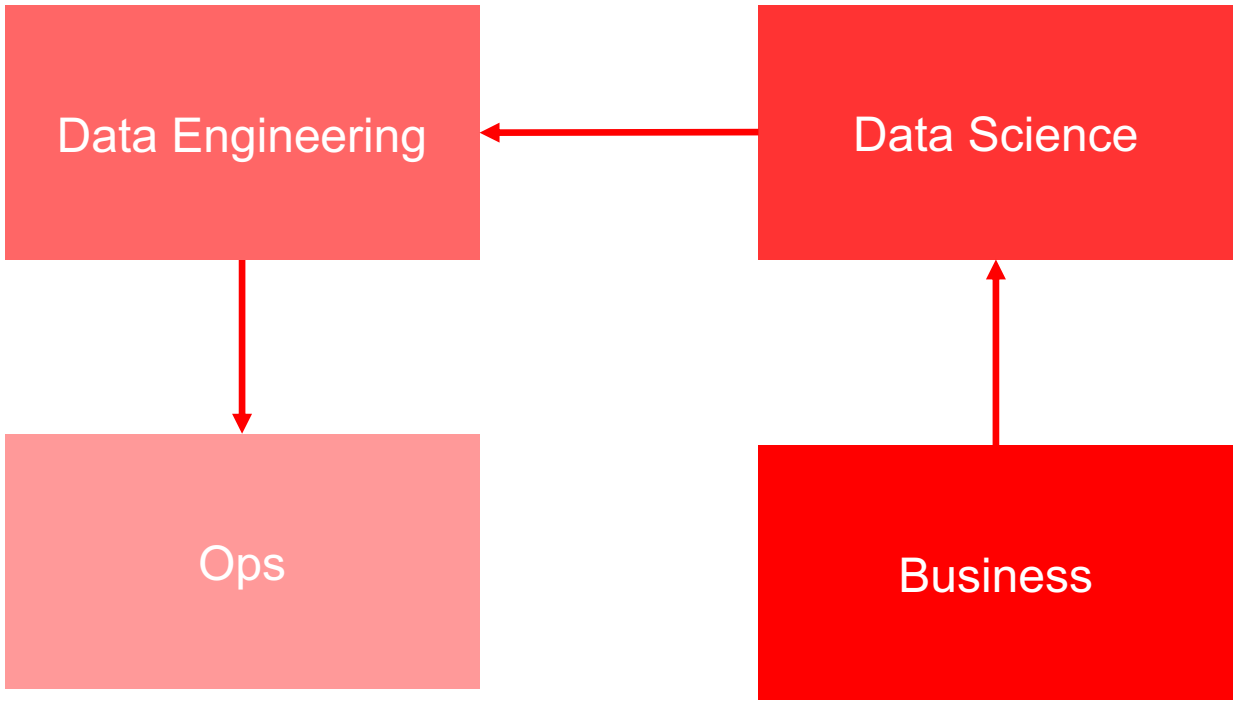


"What is the Role of a Data Engineering Team?"

Jorge Figueiredo

What is the problem?

What organizations do in this data age?



Data Scientist vs Data Engineer

Data Scientist	Overlapping Skills	Data Engineer
• Advanced Math/Statistics • AI/ML • Advanced Analytics	• Programming • Analysis • Big Data	• Advanced Programming • Distributed Systems • Data Pipelines

What have we used for ETL in Data Engineering?

- JVM driven ETL
- SQL driven ETL

What are the options for Data Engineering?

- JVM (Java/Scala) ETL pipelines

```
package com.twitter.scalding.examples

import com.twitter.scalding._
import com.twitter.scalding.source.TypedText

class WordCountJob(args: Args) extends Job(args) {
  TypedPipe.from(TextLine(args("input")))
    .flatMap { line => tokenize(line) }
    .groupBy { word => word } // use each word for a key
    .size // in each group, get the size
    .write(TypedText.tsv[(String, Long)](args("output")))

  // Split a piece of text into individual words.
  def tokenize(text: String): Array[String] = {
    // Lowercase each word and remove punctuation.
    text.toLowerCase.replaceAll("[^a-zA-Z0-9\\s]", "").split("\\s+")
  }
}
```

What are the options for Data Engineering?

- SQL (Hive) ETL pipelines

```
-- Dropping Database
DROP DATABASE IF EXISTS ipl_data CASCADE;
DROP DATABASE IF EXISTS ipl_stats CASCADE;

-- Creating a Database
CREATE DATABASE ipl_data ;  -- Database for Loading CSV Input File
CREATE DATABASE ipl_stats ; -- Database for Loading stats tables

-- Creating two tables to load input files ( matches and deliveries )
CREATE TABLE ipl_data.matches(
  `id` int,
  `season` int,
  `city` string,
  `date` date,
  `team1` string,
  `team2` string,
  `toss_winner` string,
  `toss_decision` string,
  `result` string,
  `dl_applied` int,
  `winner` string,
  `win_by_runs` int,
  `win_by_wickets` int,
  `player_of_match` string,
  `venue` string,
  `umpire1` string,
  `umpire2` string,
  `umpire3` string)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = ",",
"quoteChar" = "\""
)
STORED AS TEXTFILE
TBLPROPERTIES (
    'serialization.null.format' = '',
    'skip.header.line.count' = '1');
```

What are the options for Data Engineering?

- Bash ”pipelines”

```
#!/bin/bash
# Counting the number of lines in a list of files
# function version

# function storing list of all files in variable files
get_files () {
  files="`ls *.[ch]`"
}

# function counting the number of lines in a file
count_lines () {
  local f=$1  # 1st argument is filename
  l=`wc -l $f | sed 's/^\([0-9]*\).*$/\1/'` # number of lines
}

# the script should be called without arguments
if [ $# -ge 1 ]
then
  echo "Usage: $0 "
  exit 1
fi

# split by newline
IFS=$'\012'

echo "$0 counts the lines of code"
# don't forget to initialise!
```

What is the other way?

What is the other way?

- APIs
- No-Ops
- Self-service

APIs

DE

Kubernetes

IaaS

API

Dev Ops

PaaS
(GCP Data Flow)

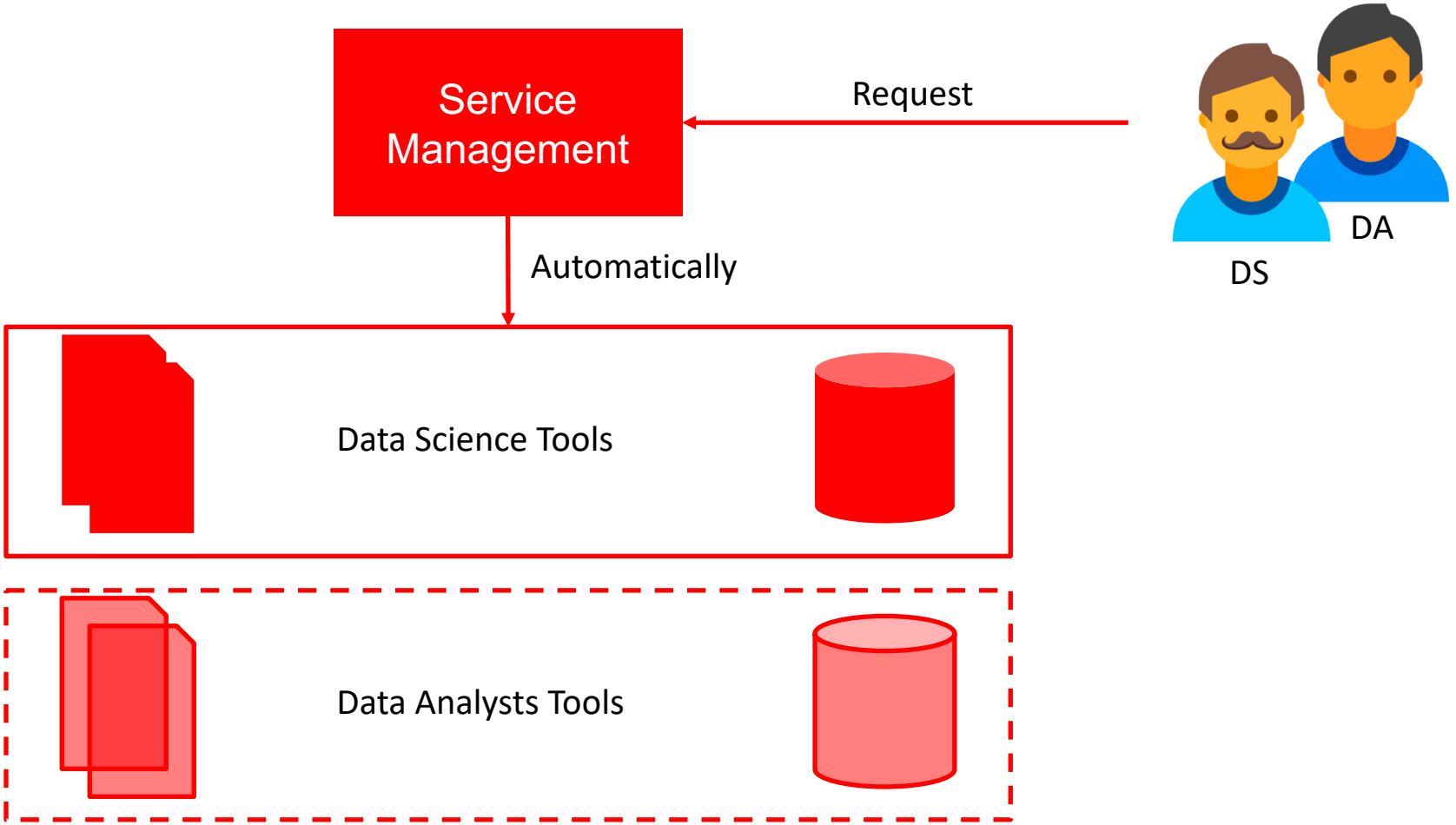
PaaS
(GCP BigQuery)

DS

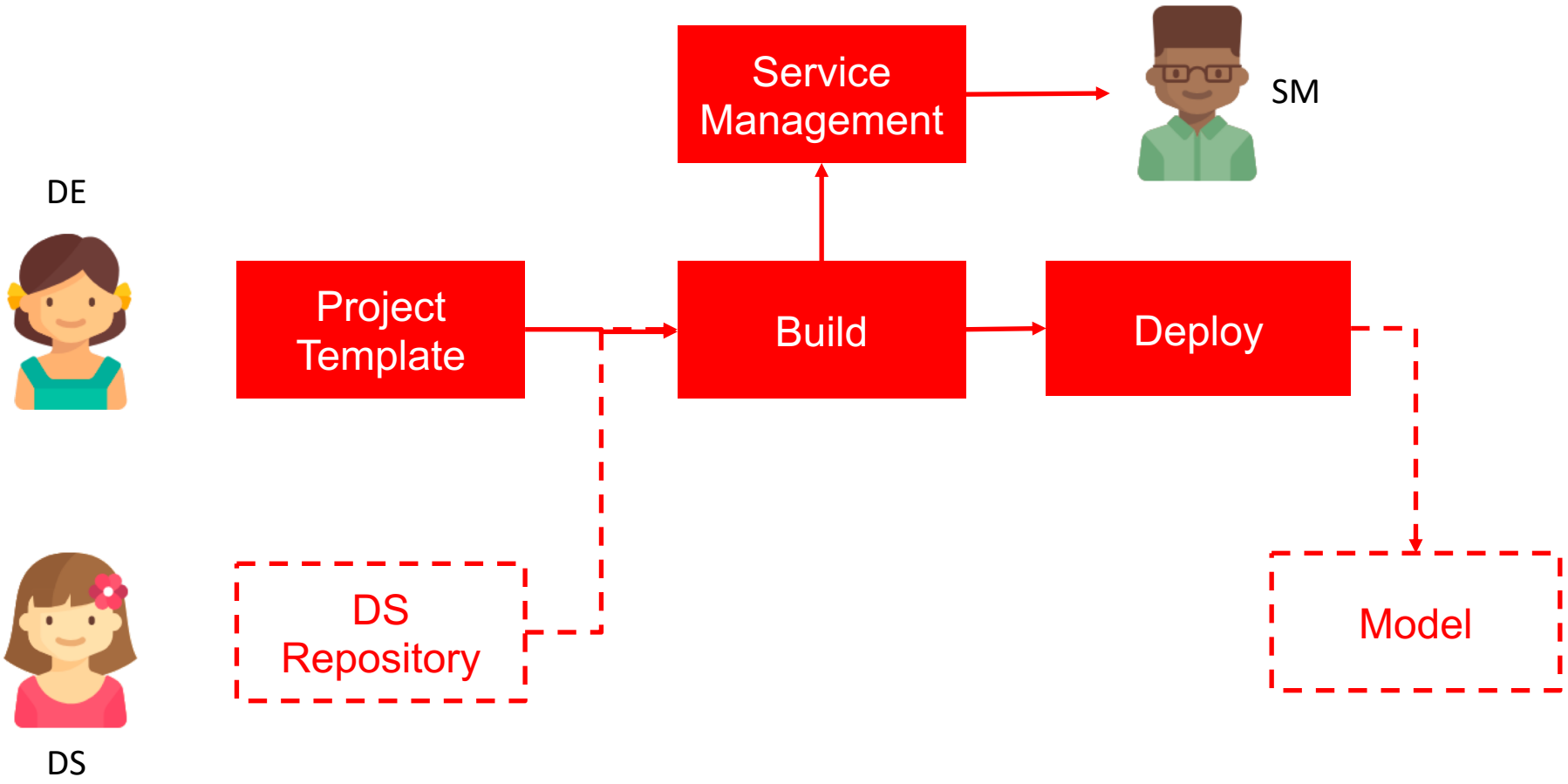
NoOps

DevOps	NoOps
Development and operations work together.	Development and operations never need to interact.
Dev and Ops work together to select, build, monitor and maintain self-service solutions.	Self-service infrastructure, continuous integration and deployment solutions.
Based on principles and practices.	Only possible through use of specific software solutions.

Technical patterns – Self-service Infrastructure toolkit



Technical patterns – Project Template



Technical patterns – Project template

Branch feature/GBIDE-524

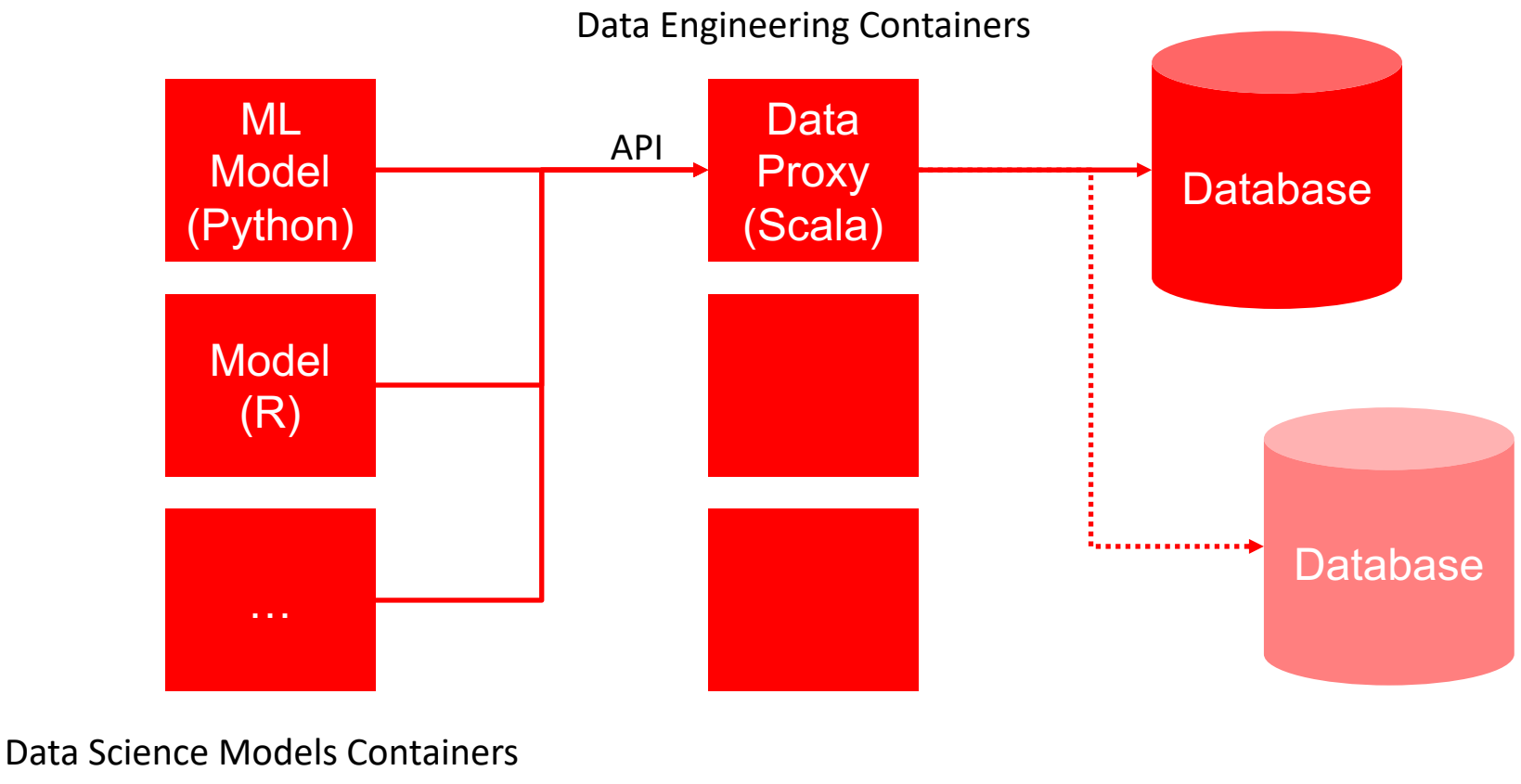
Full project name: scala-template/feature%2FGBIDE-524

Recent Changes

Stage View

	Checkout repository	Compile	Test	Publish	Build Images	Push Images
Average stage times: (Average full run time: ~4min 50s)	9s	1min 31s	17s	52s	31s	38s
#48 Feb 20 10:37 — 1 commit	8s	1min 13s	26s	54s	46s	54s
#47 Feb 20 10:31 — No Changes	9s	1min 16s	25s	57s	45s	22s
#46 Feb 20 10:20 — No Changes	8s	1min 18s	24s (paused for 11s)	51s	3s failed	

Technical patterns – Containers



In the end what should be the focus of a Data Engineer?

1. Create building blocks or abstractions that the Data Scientists can use autonomously.

2. Looking after the end to end development cycle of data applications.

3. Consulting to Data Scientists in case of more extreme cases.

Anti Patterns

1. Ratios of Data Engineers compared with Data Scientists.
2. Are your Data Scientist doing work that is Data Engineering or Operations?
3. Are you sure that you have Big Data? Scalability problems?

Questions