MANGO SOLUTIONS

Spark, R and Sparklyr

Doug Ashton – Senior Data Scientist

Aimee Gott – Senior Data Scientist

Mark Sellors – Head of Data Engineering

@MangoTheCat

info@mango-solutions.com

About Mango

www.mango-solutions.com
@MangoTheCat

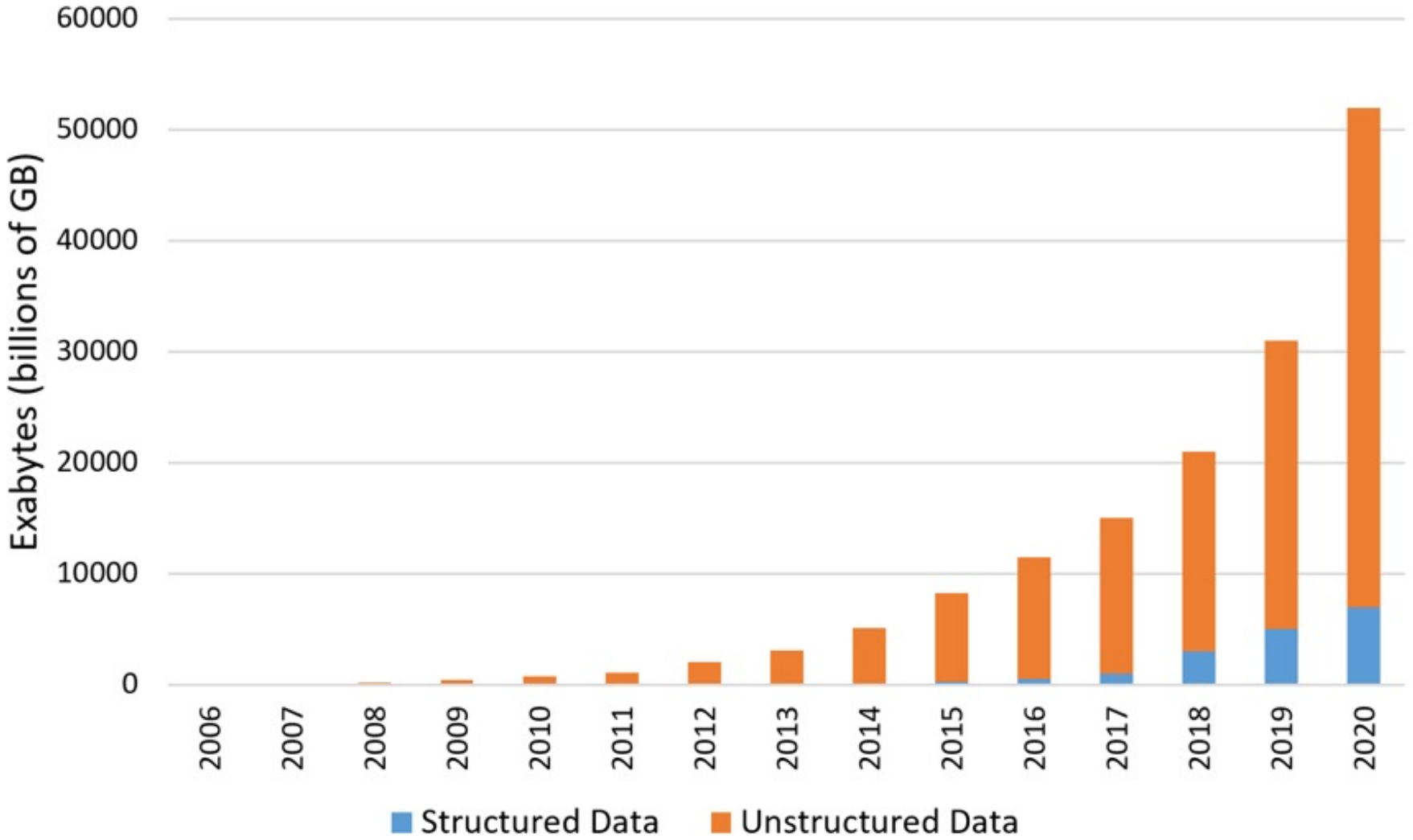

EARL17
ENTERPRISE APPLICATIONS OF THE R LANGUAGE
LONDON
12-14 September
The Tower Hotel
SAN FRANCISCO
5-7 June
Holiday Inn Golden Gateway
BOSTON
November


The Rise of
Big Data

Lots of Vs

- Created by Doug Laney in 2001
- Describe data characteristics
  - Volume
  - Variety
  - Velocity
  - (Veracity)

The Cambrian Explosion...of Data



Source: Patrick Cheesman

The Boeing 787 produces over 500GB of data during every flight

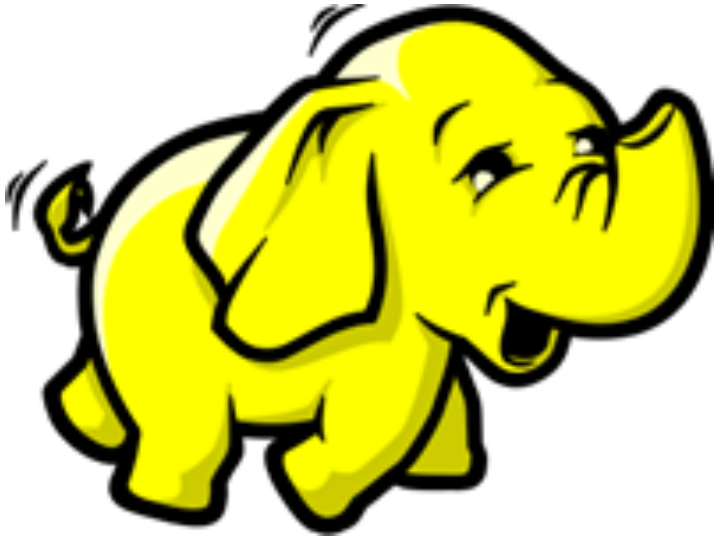
BY MATTHEW HUMPHRIES
03.07.2013 :: 7:26AM EDT @MTHWGEEK

With the introduction of the 787 Dreamliner (**battery and electrical problems aside**), Boeing produced a commercial jet that's packed full of tech. The windows have an electrical dimming system, engine noise has been reduced with a clever wave pattern design around each jet engine exhaust, and the aircraft even employs **accelerometers to counteract turbulence**.


The Yellow
Elephant

Apache Hadoop

- Grew out of Google File System 2003
- Attributed to Doug Cutting @ Yahoo!
- Named after a toy elephant
- Released in 2006



Apache Hadoop

- HDFS (Hadoop Distributed File system)
  - Distributed, scalable file system
  - Store (and access) large files across machines
  - Written in Java
- MadReduce
  - Programming Model for processing parallelizable problems
  - Map > Shuffle > Reduce steps

Hadoop Deployments

- Many (most?) hadoop deployments do not achieve an acceptable ROI
- An answer looking for a question
- Skillsets needed to get the most out of the infrastructure are extremely difficult to recruit for
- IT & Data Silos

MapReduce Woes

- Disc bound and expensive
- Not suited to modern analytics
- Can be slow
- Java is only native API
- Scarcity of resources

R & Big Data

What is R?

- The most popular analytic language in the world
- Programming language for modelling
- Aligned with the Data Science movement

- But …
  - Single Threaded
  - Memory Bound


Spark

Spark

- Open source cluster computing framework
- Relies on in memory processing
- One of the most contributed-to big data projects of the past few years
- Started in the AMPLab at UC Berkeley in 2009

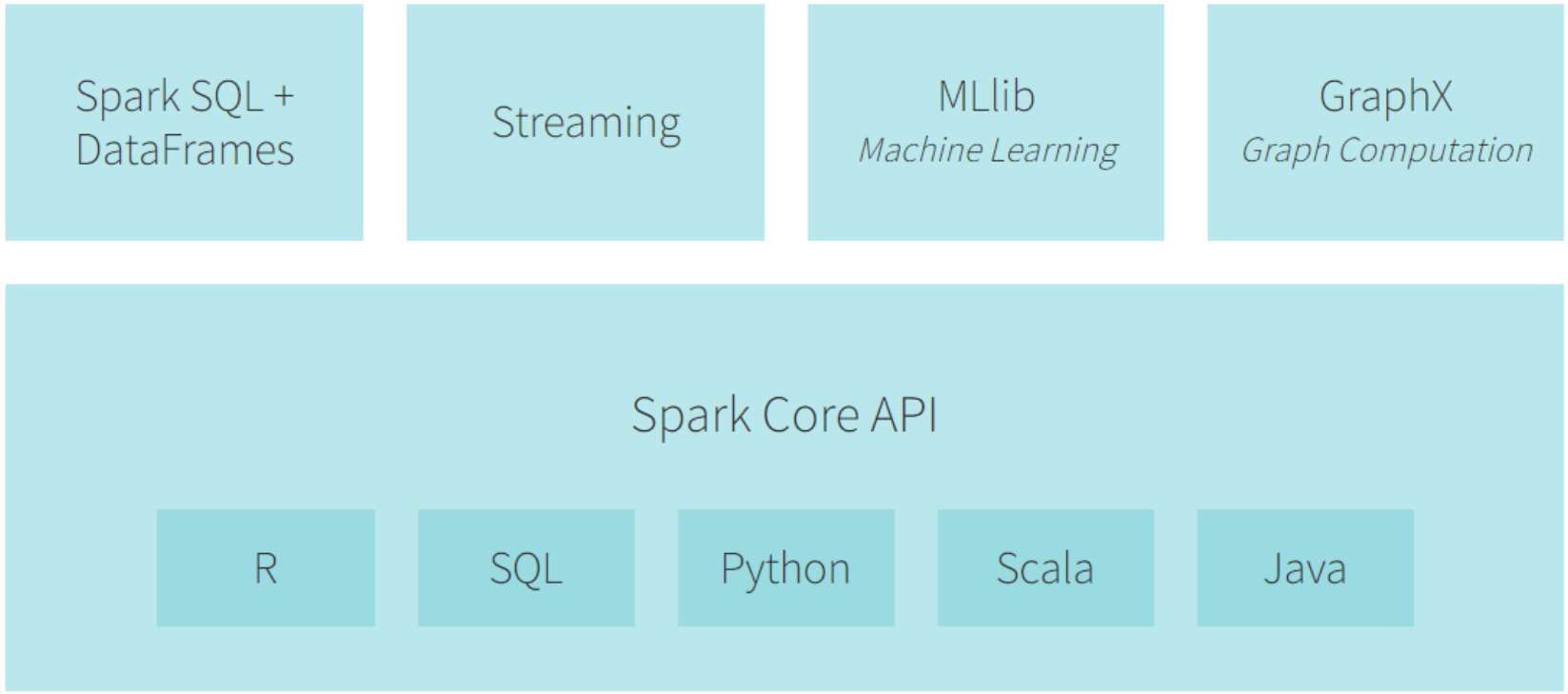
What problem does it solve?

- In memory makes for very fast data processing
- Minimal disk IO
- High level programming abstraction
- Reduces the amount of code
- In turn makes it more suitable for exploratory work

How does it do it?

- Provides a core programming abstraction called RDD (Resilient Distributed Dataset)
- Data split over memory on multiple machines
- The RDD API has been extended to include DataFrames
- Can deploy ad-hoc processing clusters as well as integrate with HDFS/Hadoop

Spark Ecosystem



Spark and Hadoop

- Very complementary technologies
- Spark is already included in all the major Hadoop distributions
- Easier to use and faster than Map/Reduce
- Suitable to exploratory work which used to be very difficult


Spark and R

Spark and R

- Originally supported languages Scala, Java and Python
- SparkR was a separate project, integrated into Spark as of v1.4
- Lets you create and work with MASSIVE data frames over a cluster of machines

SparkR

- Limited capability
- Unfamiliar code structures
- New implementations (e.g. of lm)

Sparklyr

- Designed by Rstudio
- Based on the dplyr patterns so familiar to users
- Still under very active development
- Can set up and install spark locally itself for learning/experimentation
- Filter and aggregate Spark datasets then bring them into R for analysis and visualization
- Use Spark distributed ML library from R

What’s in it for you?

- If you’ve ever run out of memory in R or need to interface with a Hadoop cluster, Spark and Sparklyr may be the solution for you.
- Allows you to create and work with vast datasets
- You can work interactively or create batch jobs, all using a familiar R/dplyr syntax

Workshop Environment

- Username: user<n>
- Password: …
- Host: http://spark2.mangodatalabs.com
- Host: http://spark3.mangodatalabs.com

spark.mangodatalabs.com:8787

Apps Bookmarks G + Pocket GBookmark Home Mango tools MIT - OCW | Comp... 7 Essential Lessons... GitHub · Where soft... Mango URLs Other Bookmarks

File Edit Code View Plots Session Build Debug Profile Tools Help user75

Go to file/function Addins Project: (None)

Console ~/

```
R version 3.3.3 (2017-03-06) -- "Another Canoe"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Environment History Spark

Import Dataset List

Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Home

Name	Size	Modified
R		

Creating a spark context

- library(sparklyr)
- config <- spark_config()
- config$spark.executor.cores <- 2
- config$spark.executor.memory <- “4G”
- config$spark.ui.port <- “<portnumber>”
- sc <- spark_connect(master = “spark://…:7077”, config = config,
  app_name = “<your_name>”)