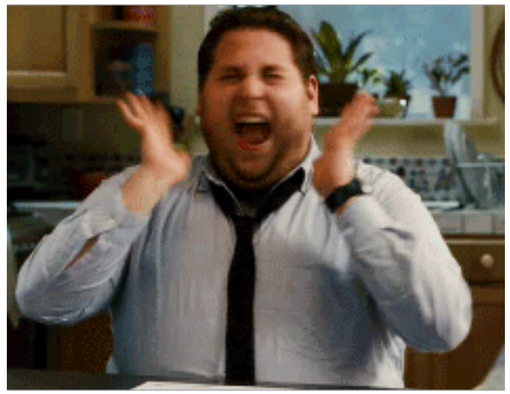
DOING BIG DATA WITH SPARK

data.miami - July 2018



WHO ARE YOU?

AARON RICHTER

- Data Scientist / Engineer @ Modernizing Medicine
- PhD Candidate @ FAU
- Spark Certified Developer
- @rikturr

- What is big data?
- Hadoop history lesson
- Spark is cool
- How to Spark with R

BIG DATA ?

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

The FOUR V's of Big

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

2005
2020

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Variety
DIFFERENT
S OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

acity
RTAINTY
DATA

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

BIG DATA

(in our context)

- Data or workloads that require non-traditional distributed processing
  - The same data may be “small” or “big” depending on what you want to do with it

BIG DATA

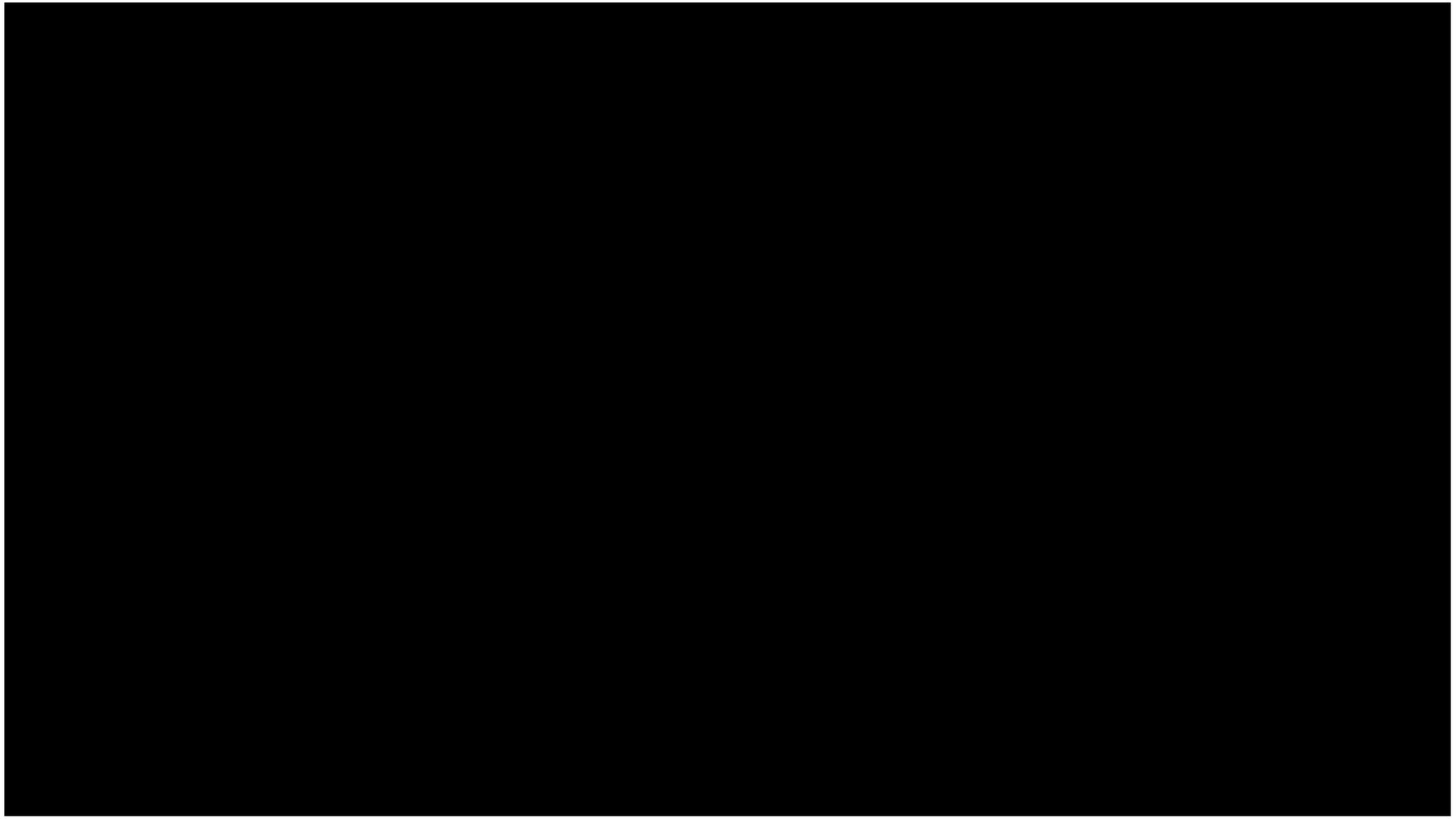
How to handle big data?

- Distributed processing
  - Distribute storage
  - Distribute computations
  - Cluster of machines
- Problems
  - Machine failures
  - Distributed programming is difficult
  - Allocating resources
  - Concurrency
  - Sharing data

HADOOP

HISTORY LESSON

- Framework for distributed processing across a cluster of machines
- Storage - Hadoop Distributed File System (HDFS)
- Computation - MapReduce, YARN
- “Hadoop Ecosystem”



https://www.youtube.com/watch?v=ebgXN7VaIZA

TL;DR

HADOOP / MAPREDUCE

- Framework for cluster computing
- Important to know about (widely used last ~10 years)
- You probably don’t need it


APACHE
Spark
TM

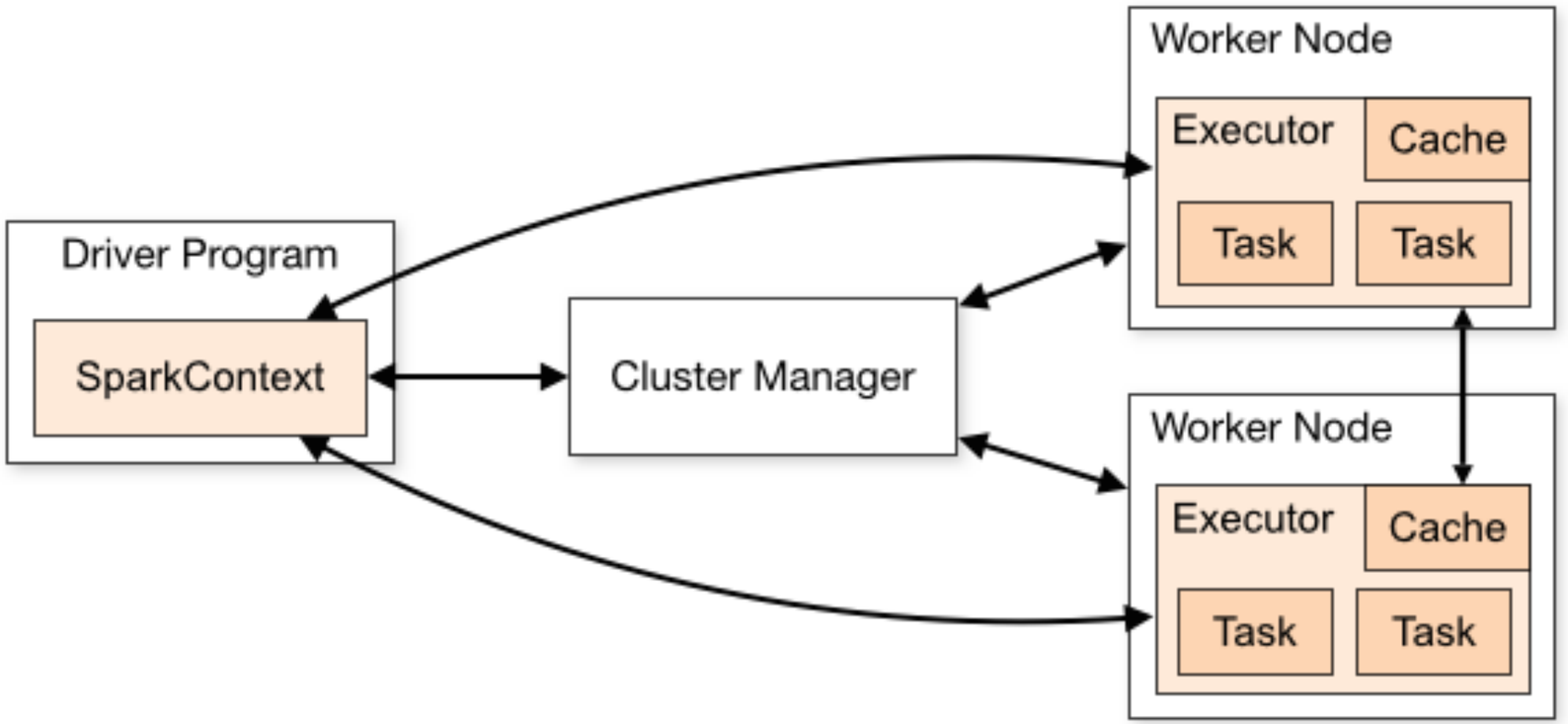
SPARK

WHY IS IT COOL?

- Distributed data processing engine
- Developed at UC Berkeley, now Apache project
- APIs for Scala , Java, Python, R, SQL
- DataFrames, Streaming, ML, Graph
- Its really fast

KEY CONSIDERATIONS

- Spark is for *processing* data, not *storing* data
- Can read just about any data source
  - CSV, JSON, databases, parquet
- Same code will run on a single machine or a cluster


Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task

WHY SHOULD I CARE?

- When R can’t handle your data
- Lots of companies use Spark
- If you want to be a cool kid

RSTUDIO TO THE RESCUE!

SPARKLYR

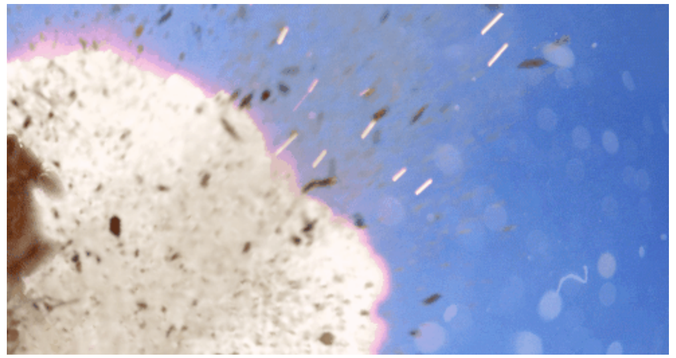
- R package for Spark
- Developed by RStudio
- Tidy syntax (uses dplyr)
- Supports DataFrames and ML

http://spark.rstudio.com

(an aside)

Your data fits in RAM

CODE



VIDEOS FOR DAYS

SPARK+AI SUMMIT

- https://databricks.com/sparkaisummit/sessions
- Virtualizing Analytics with Spark
  - How Spark unifies analytics for the enterprise
- R and Spark
    - Talk from RStudio about sparklyr and H2O
- Dynamic Healthcare Dataset Generation with PySpark
    - Shameless plug (talk by yours truly)

THANK YOU!

AARON RICHTER

- Data Scientist / Engineer @ Modernizing Medicine
- PhD Candidate @ FAU
- Spark Certified Developer
- @rikturr