



# DATAOPS:

# S.M.A.C.K.ING THE INSIGHTS FROM YOUR DATA!

FILIPE COELHO

BLIP

DSPT#53 & DEVOPS PORTO#27

16 MAY 2019

# AND NOW FOR SOME ADVERTISING…

# FILIPE COELHO

- PHD, INFORMATICS ENGINEERING @ FEUP
- DATA SCIENTIST | RESEARCH @ INESC TEC
- BIG DATA ENGINEER | E-COMMERCE
  @ FARFETCH, PROZIS & HOSTELWORLD

fil.coelho@gmail.com
@DataPlumbR
linkedin.com/in/dataplumbr

Universidade do Porto
FEUP Faculdade de Engenharia


FARFETCH

PROZIS

HOSTELWORLD

# TECH COMMUNITIES

## PORTODATA / SQL SATURDAY


- "HDINSIGHT: NEW TRICKS FOR THE AZURE ELEPHANT" (HADOOP)
- "POLYBASE: UNLEASH THE POWER OF BIG DATA IN SQL SERVER 2016" (POLYBASE)
- "NOTEBOOKS, NOTEBOOKS EVERYWHERE!" (JUPYTER NOTEBOOKS)

## DATA SCIENCE PORTUGAL


- DATAOPS: THE NEXT LEVEL (AGILE DATA SCIENCE TEAMS)
- DATAOPS: S.M.A.C.K.ING THE INSIGHTS FROM YOUR DATA

# … AND WE'RE BACK TO OUR REGULAR SCHEDULE!

# OVERVIEW

## THE "WHEN"

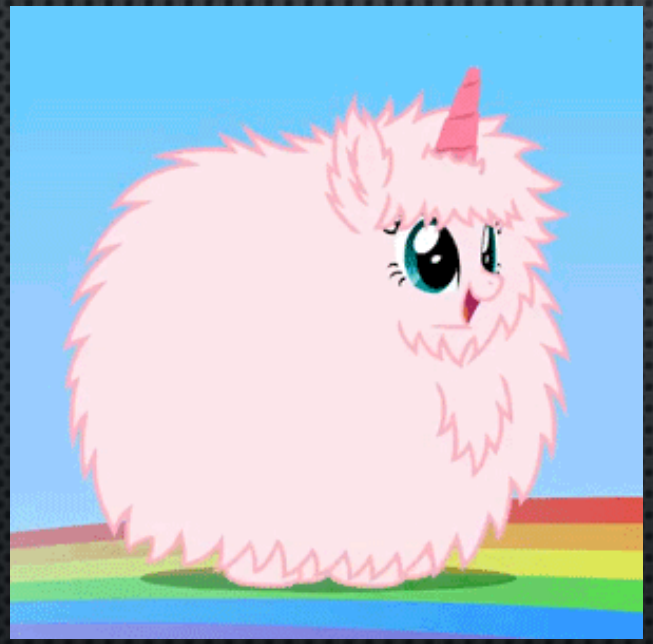
## THE "WHAT"

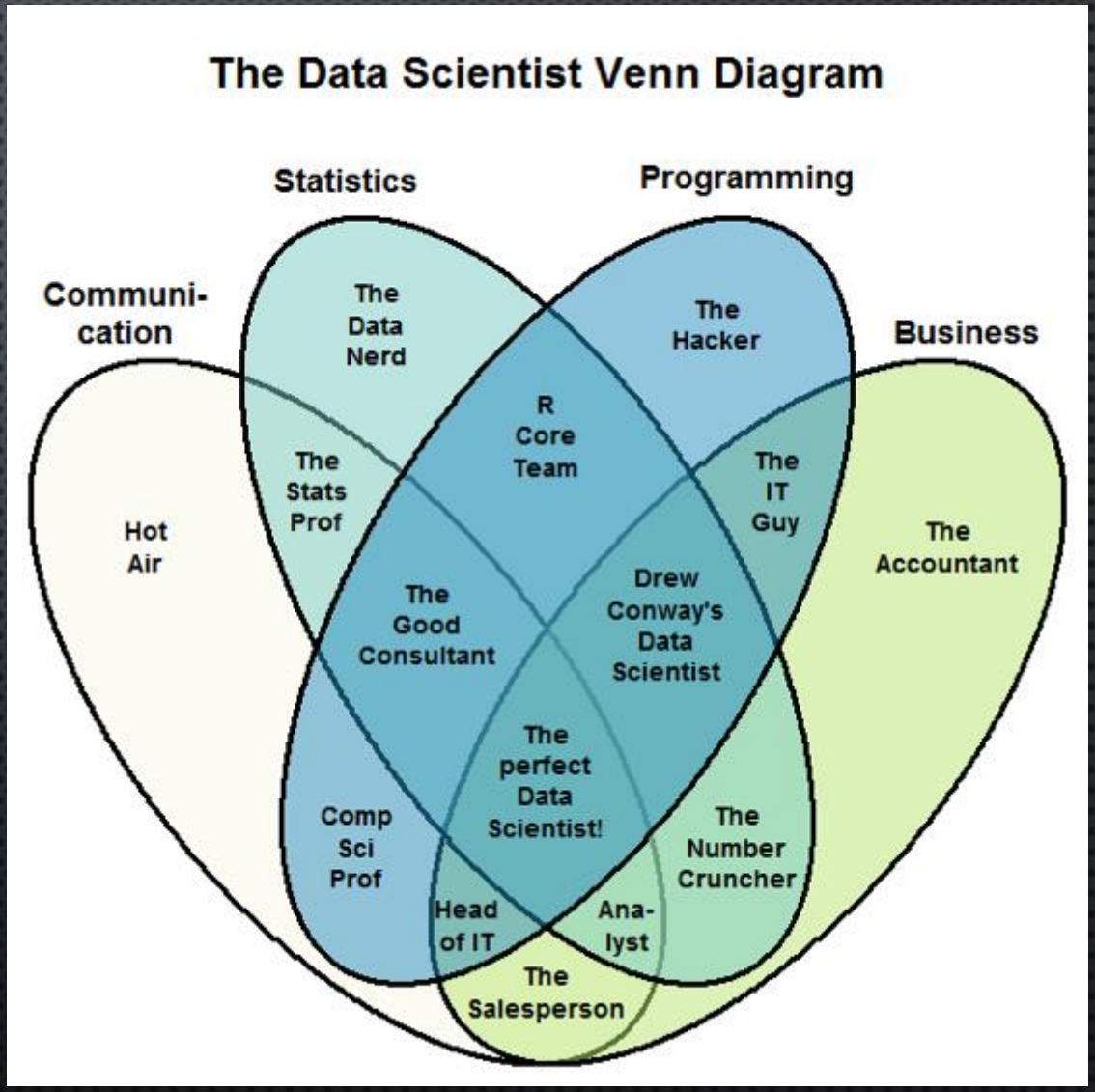

## THE "HOW"


# THE “WHEN”

# IN THE BEGINNING, THERE WAS JUST "BIG DATA"…

# … AND THE "UNICORN"!





## MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

## PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

## DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

## COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

## Who am i?

**Peadar Coyle**

## Mailing List

You may want to follow my newsletter!

## Top Posts & Pages

Adding value as Data Scientists
Why Probabilistic Programming is the next big thing in Data Science
How to use AWS Lambda to build a tweetbot

# Avoiding being a 'trophy' data scientist

July 23, 2017

Recently I've been speaking to a number of data scientists about the challenges of adding value to companies. This isn't an argument that data science doesn't have positive ROI, but that there needs to be an understanding of the 'team sport' and organisational maturity to take advantage of these skills.

The biggest anti-pattern I've experienced personally as an individual contributor has been a lack of 'leadership' for data science. I've seen organisations without the budgetary support, the right champions or clear alignment of data science with their organisational goals. These are some of the anti-patterns I've seen, it's non-exhaustive so I provide it.

# Engineers Shouldn't Write ETL: A Guide to Building a High Functioning Data Science Department

**JEFF MAGNUSSON**

March 16, 2016 - San Francisco, CA

Tweet this post! Post on LinkedIn

"What is the relationship like between your team and the data scientists?" This is, without a doubt, the question I'm most frequently asked when conducting interviews for data platform engineers. It's a fine question – one that, given the state of engineering jobs in the data space, is essential to ask as part of doing due diligence in evaluating new opportunities. I'm always happy to answer. But I wish I didn't have to, because this a question that is motivated by skepticism and fear.

# DATA Engineer

DataCamp
Learn Data Science By Doing

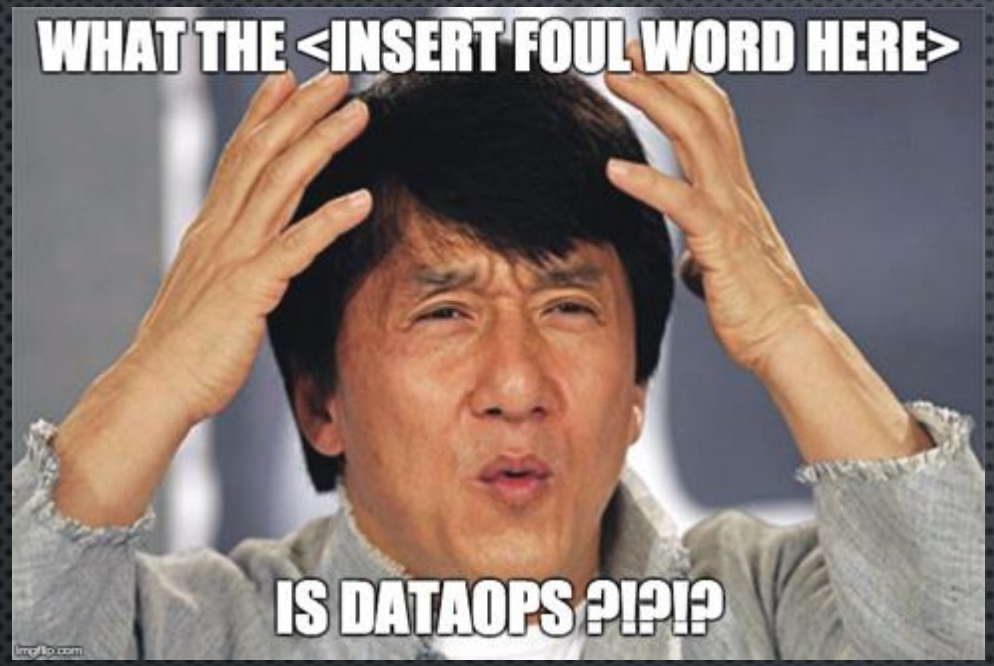
# DATA Scientist

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Data Engineers & Data Scientists work together to wrangle Big Data and provide insights to business critical decisions.

While their skills may widely overlap, the two positions are becoming more and more distinct.



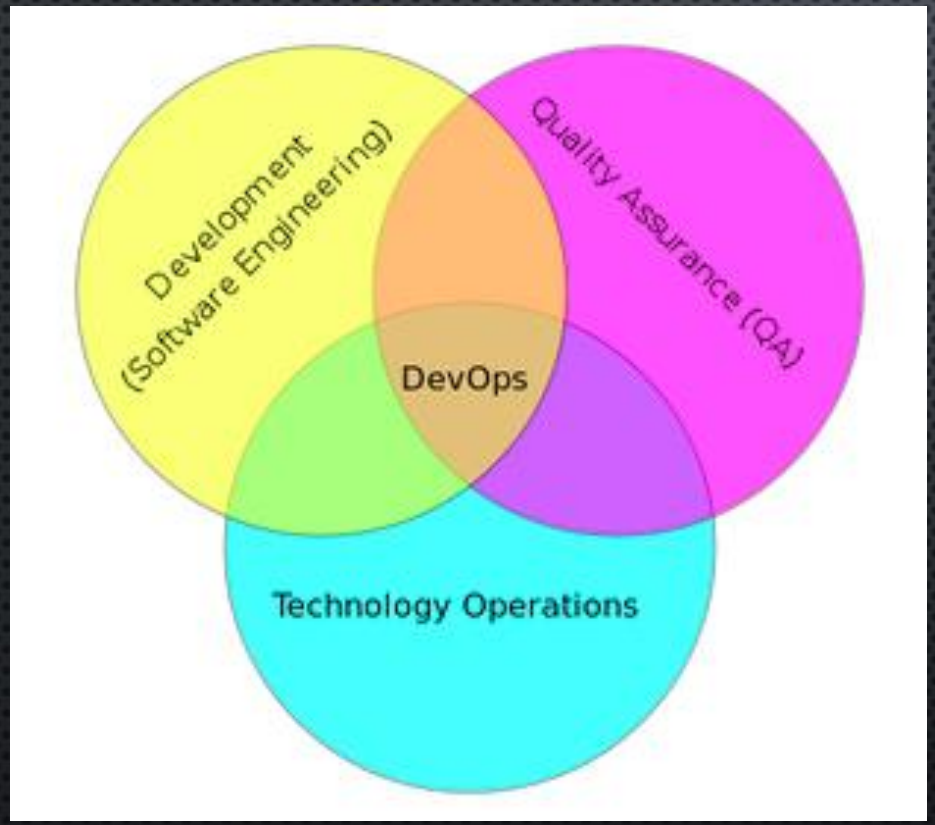

# THE "WHAT"

# "DATAOPS": LMGTFY...

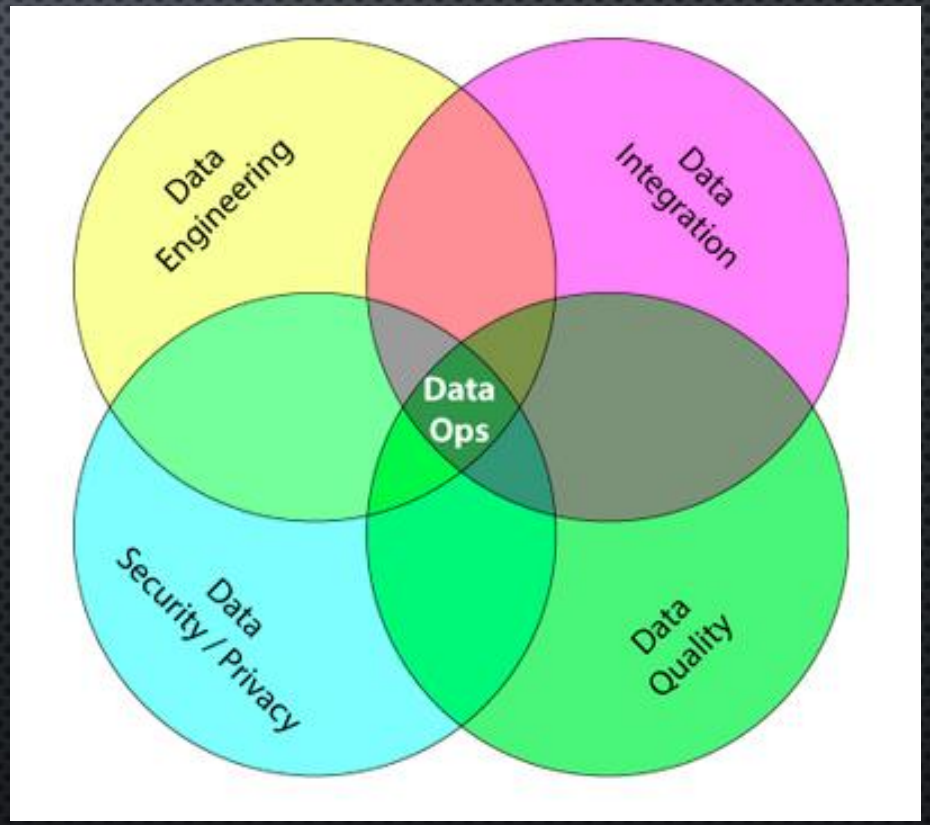
Let me take a stab at **definition**: **DataOps** is a data management method that emphasizes communication, collaboration, integration, automation and measurement of cooperation between data engineers, data scientists and other data professionals. May 7, 2015


From DevOps to DataOps, By Andy Palmer - Tamr Inc.
www.tamr.com/from-devops-to-**dataops**-by-andy-palmer/

# FROM DEVOPS…




# …TO DATAOPS

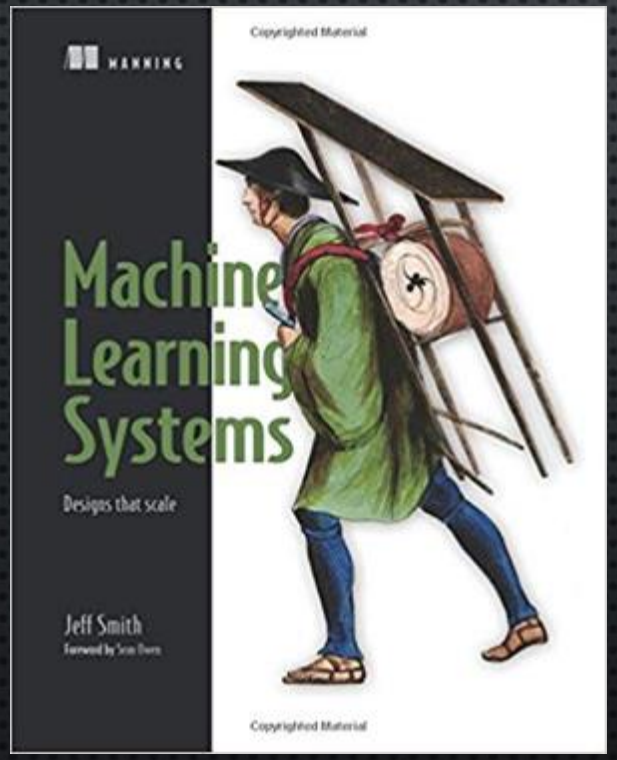


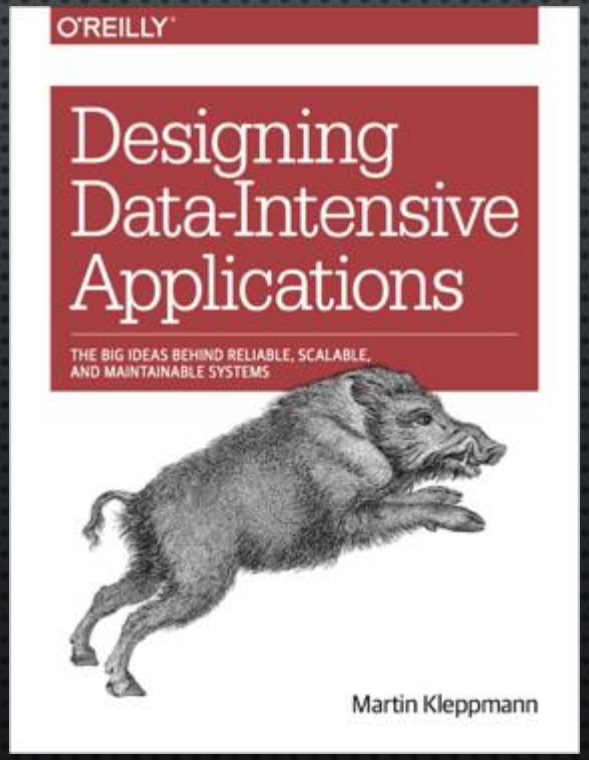
# 13 Steps to Better Data Science: A Joel Test of Data Science Maturity

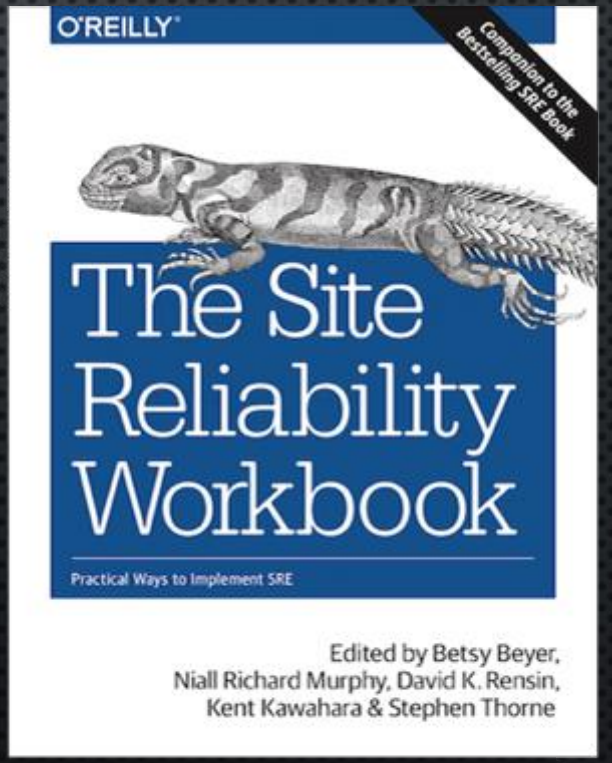
Jan 13, 2018

1. Are results reproducible?
2. Do you use source control?
3. Do you create a data pipeline that you can rebuild with one command?
4. Do you manage delivery to a schedule?
5. Do you capture your objectives (scientific hypotheses)?
6. Do you rebuild pipelines frequently?
7. Do you track bugs in your models and your pipeline code?
8. Do you analyse the robustness of your models?
9. Do you translate model performance to commercial KPIs?
10. Do new candidates write code at interview?
11. Do you have access to scalable compute and storage?
12. Can Data Scientists install libraries and packages without intervention by IT?
13. Can Data Scientists deploy their models with minimal dependencies on engineering and infrastructure?

# FOOD FOR THOUGHT…



# Two sides of the same coin

## Data Products

- Data Scientists (A/B Testing, Visualization)
- Data Science Engineers (ML, embedded)

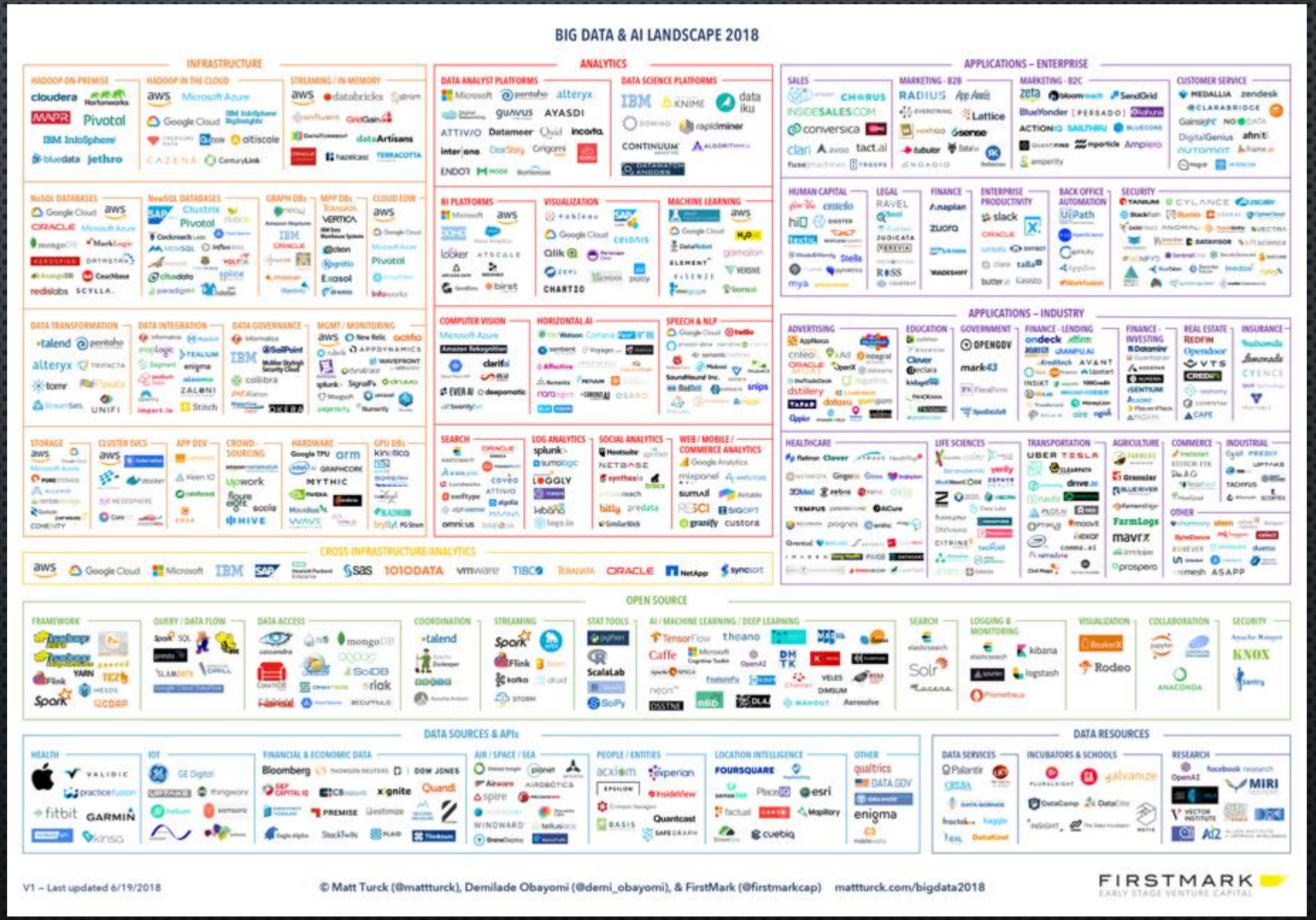
## Data Platform

- Data Developers (DataFlow, Pipelines)
- Data Platform Engineers (Infra, SRE)


# THE "HOW"

# … NOWADAYS!

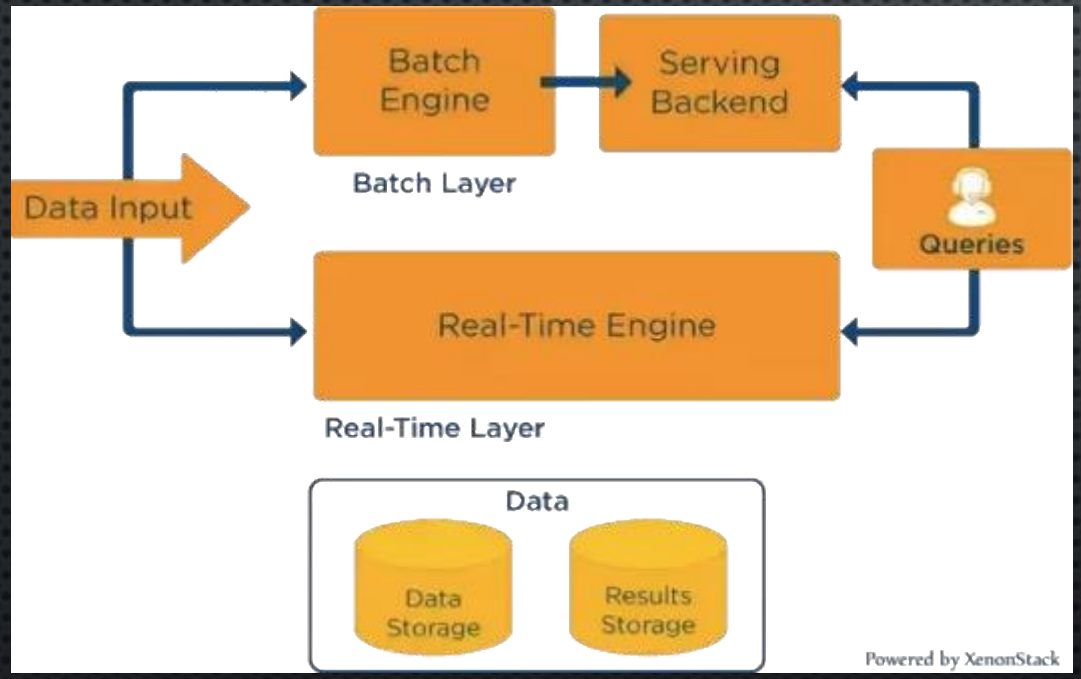

MY HADOOP IS
BIGGER
THAN YOURS...


BIG DATA & AI LANDSCAPE 2018
INFRASTRUCTURE
ANALYTICS
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
CROSS-INFRASTRUCTURE/ANALYTICS
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
V1 – Last updated 6/19/2018
© Matt Turck (@mattturck), Demilade Obayomi (@demi_obayomi), & FirstMark (@firstmarkcap) mattturck.com/bigdata2018
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

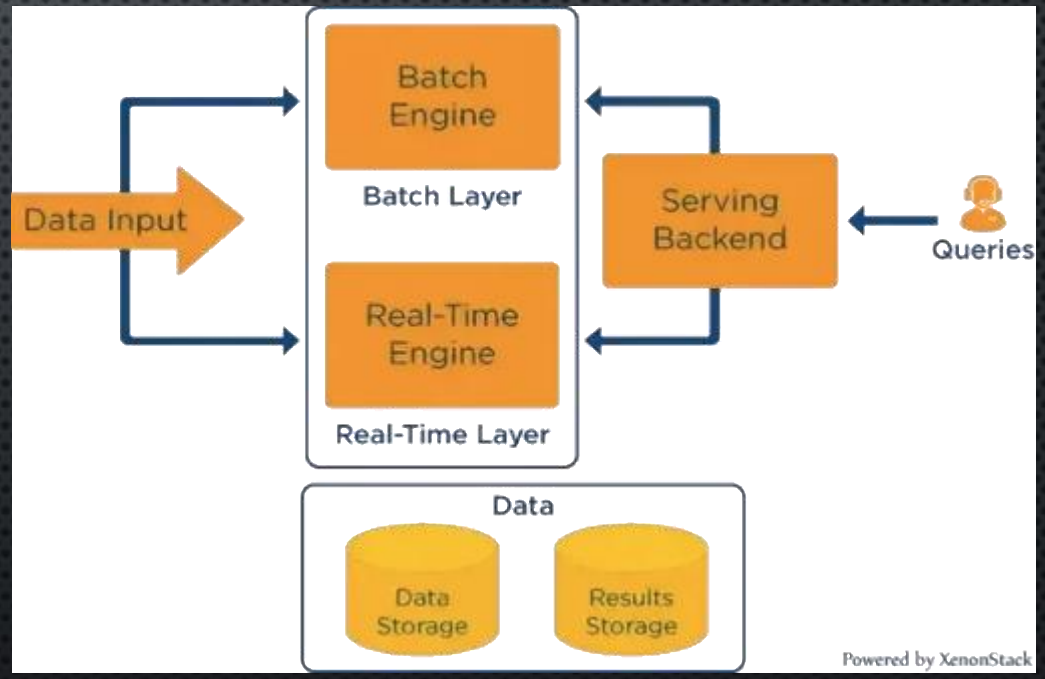

WE NEED TO GO
DEEPER
troll-face.ru


INFRASTRUCTURE
HADOOP ON-PREMISE
cloudera
Hortonworks
MAPR
Pivotal
IBM InfoSphere
bluedata
jethro
HADOOP IN THE CLOUD
aws
Microsoft Azure
Google Cloud
IBM InfoSphere BigInsights
TREASURE DATA
Qubole
altiscale
CAZENA
CenturyLink
STREAMING / IN-MEMORY
aws
databricks
Striim
confluent
GridGain
DataTorrent
dataArtisans
ORACLE
hazelcast
TERRACOTTA
NoSQL DATABASES
Google Cloud
aws
ORACLE
Microsoft Azure
mongoDB
MarkLogic
AEROSPIKE
DATASTAX
ArangoDB
Couchbase
redislabs
SCYLLA.
NewSQL DATABASES
SAP
Clustrix
Pivotal
nuodb
Cockroach LABS
MEMSQL
influxdata
VOLTDB
citusdata
splice
paradigm4
GRAPH DBs
neo4j
Amazon Neptune
IBM
ORACLE
OrientDB
InfiniteGraph
Objectivity
MPP DBs
TERADATA
VERTICA
IBM Data Warehouse Systems
actian
Kognitio
Exasol
dremio
CLOUD EDW
aws
Google Cloud
Microsoft Azure
Pivotal
snowflake
Infoworks
DATA TRANSFORMATION
talend
pentaho
alteryx
TRIFACTA
tamr
Paxata
StreamSets
UNIFI
DATA INTEGRATION
Informatica
MuleSoft
snaplogic
TEALIUM
Segment
enigma
podium data
alooma
ZALONI
import.io
Stitch
DATA GOVERNANCE
Informatica
IBM
SailPoint
McAfee Skyhigh Security Cloud
collibra
Alation
Waterline Data
OKERA
MGMT / MONITORING
aws
New Relic
octifio
rubrik
APPDYNAMICS
WAVEFRONT by VMWARE
DATADOG
dynatrace
splunk>
SignalFx
Moogsoft
unravel
pagerduty
Numerify
STORAGE
aws
Google Cloud
Microsoft Azure
PURESTORAGE
ALLUXIO
nimblestorage
Qumulo
panasas
COHESITY
CLUSTER SVCS
aws
kubernetes
docker
MESOSPHERE
CoreOS
piperdata
APP DEV
Lightbend
Keen IO
rainforest
COSK
CROWD-SOURCING
amazon mechanical turk
Upwork
figure eight
scale
HIVE
HARDWARE
Google TPU
arm
intel AI
GRAPHCORE
MYTHIC
NVIDIA
Cerebras
Movidius
WAVE COMPUTING
HABANA
GPU DBs
kinetica
MAPD
SQREAM
BLAZINGDB
brytlyt
PG Strom

# From Lambda…

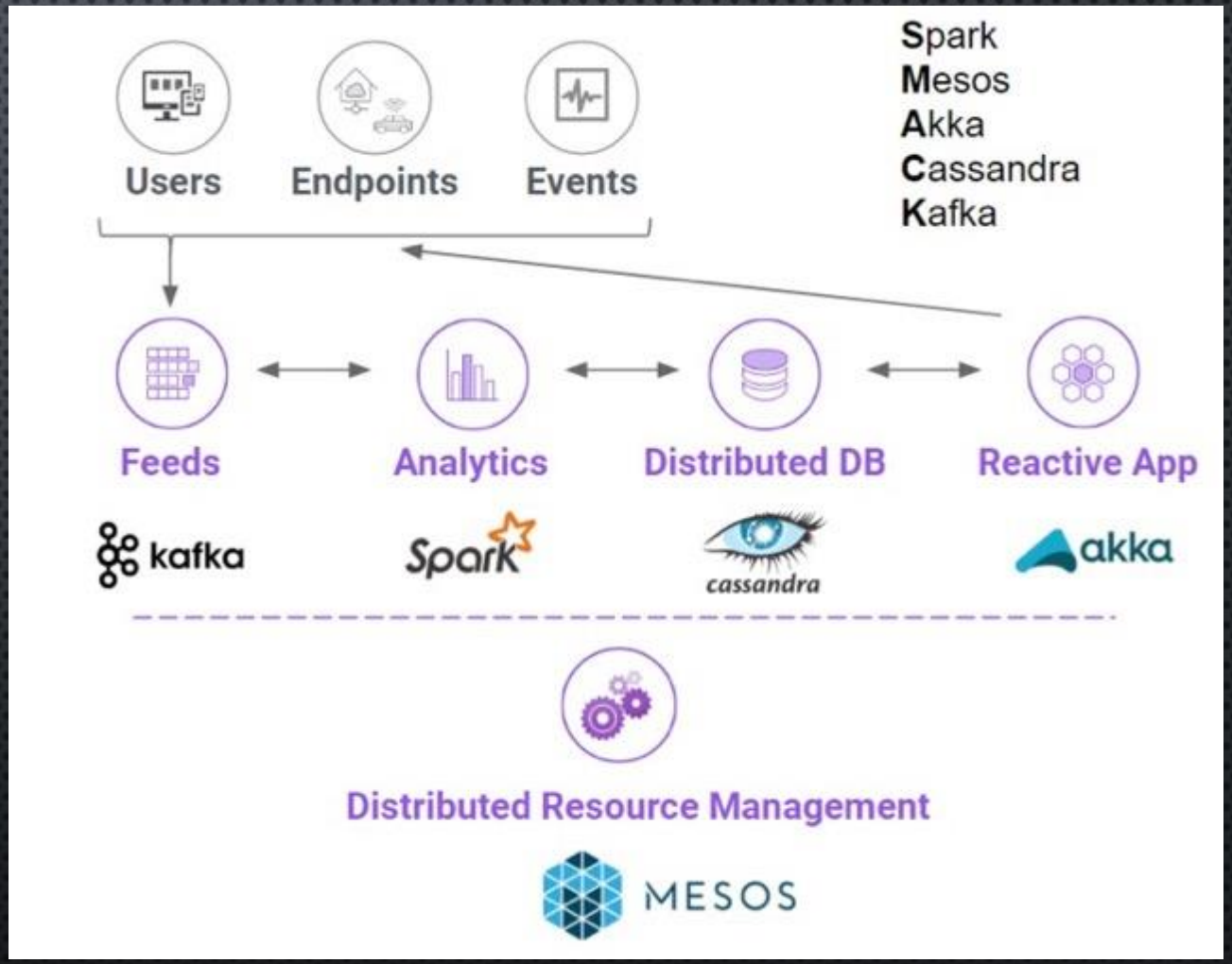



# … to Kappa Architecture!

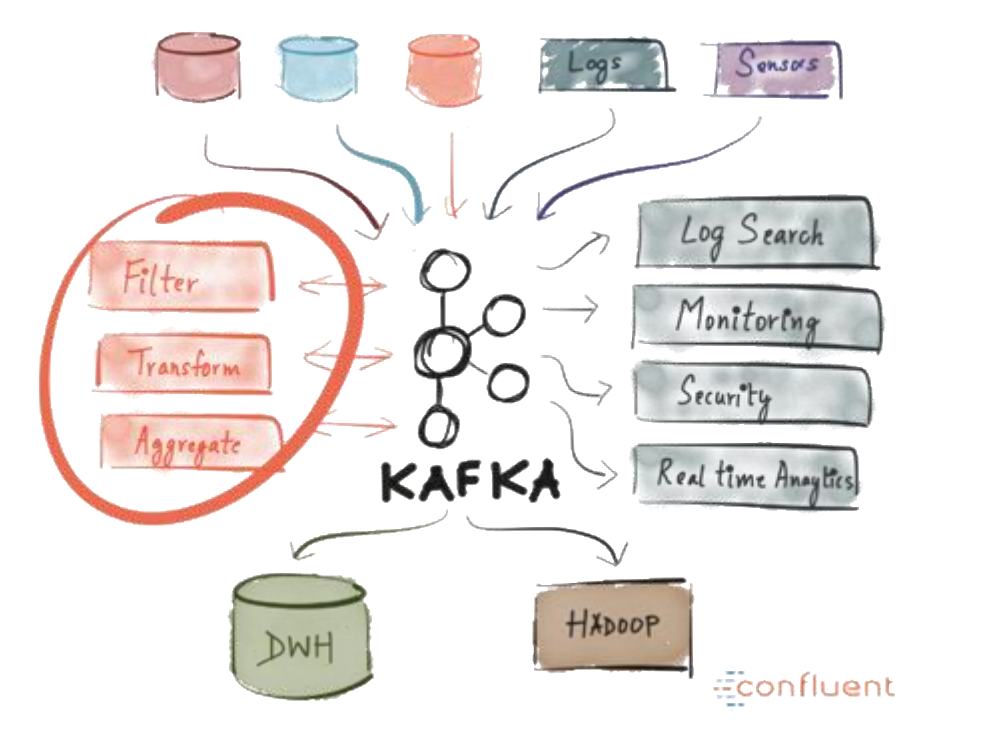


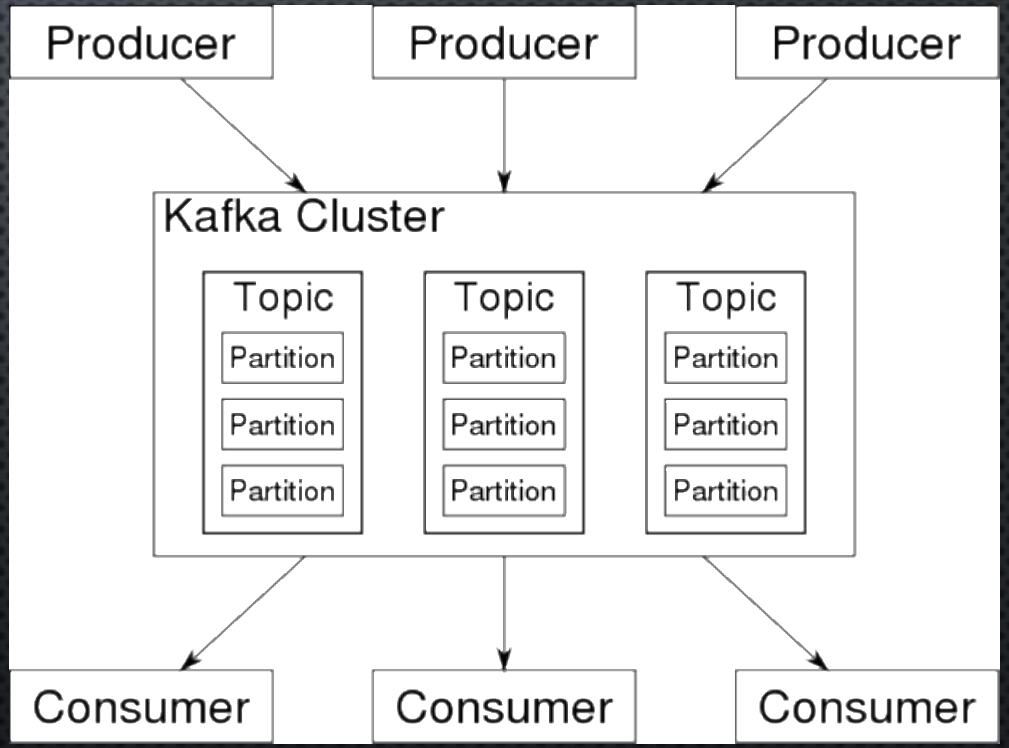
# THE S.M.A.C.K. STACK

- A CONCISE TOOLBOX THAT CAN DEAL WITH A WIDE VARIETY OF DATA PROCESSING SCENARIOS
- COMPOSED OF PROVEN, BATTLE TESTED AND WIDELY USED OPENSOURCE SOFTWARE COMPONENTS
- EASILY SCALABLE AND REPLICATION OF DATA HAPPENS WHILE STILL PRESERVING LOW LATENCIES

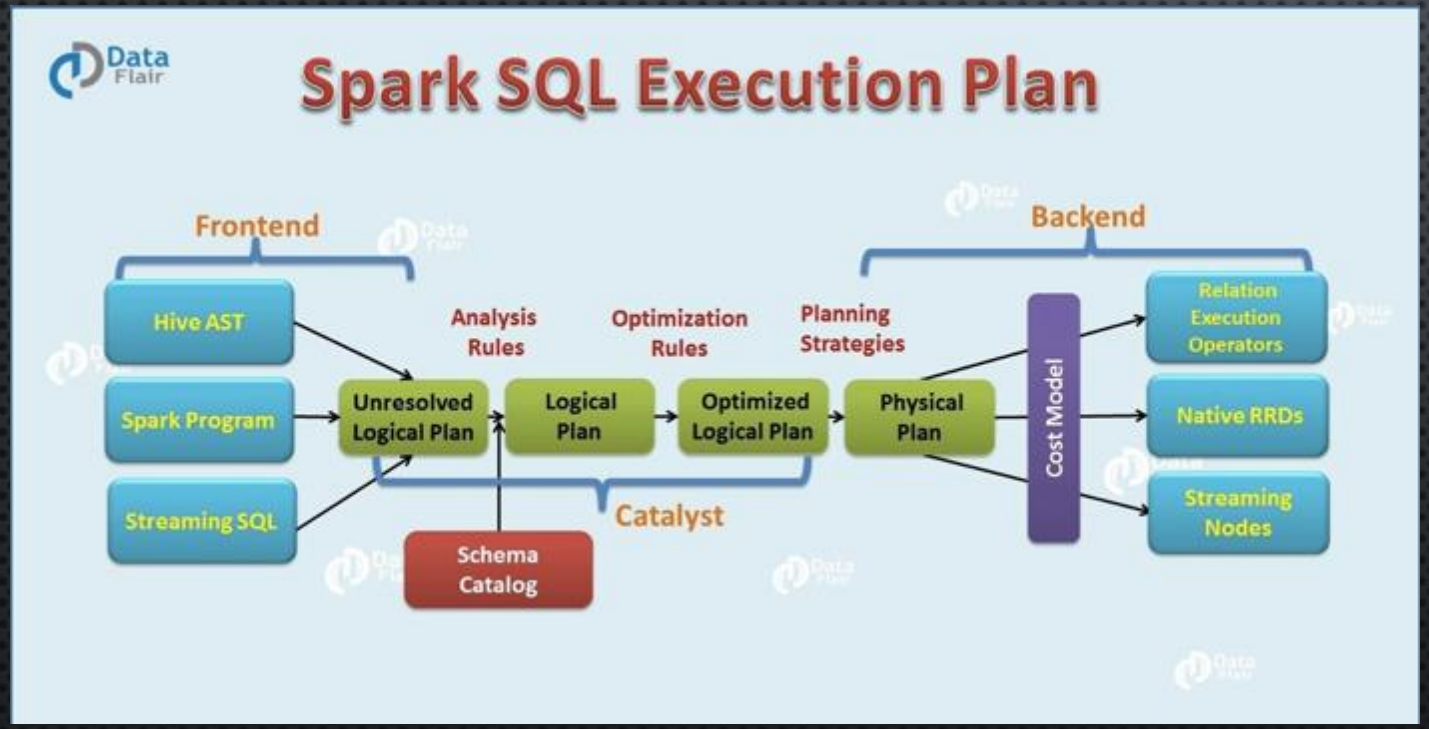


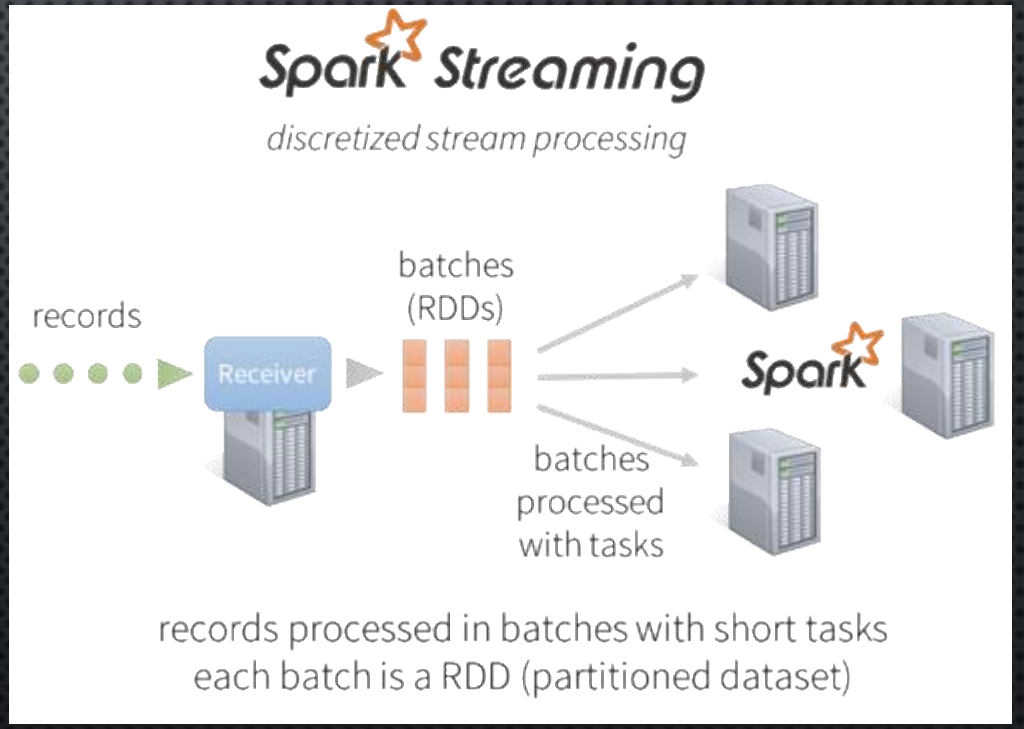
# APACHE KAFKA



# APACHE SPARK

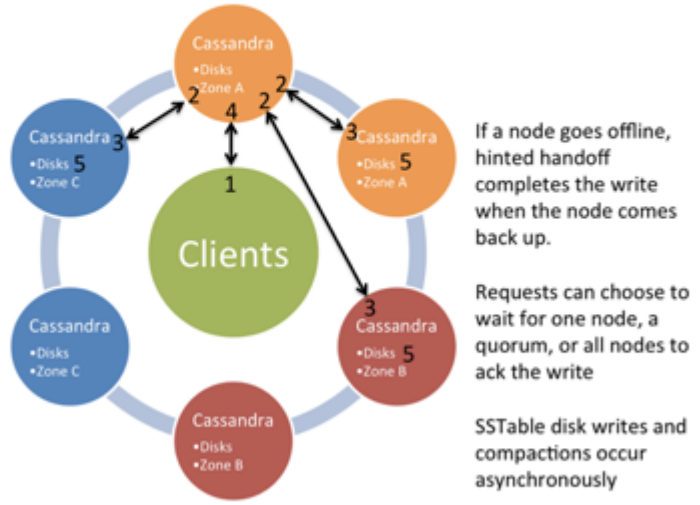


# APACHE CASSANDRA

## Cassandra Write Data Flows

### Single Region, Multiple Availability Zone




1. Client Writes to any Cassandra Node
2. Coordinator Node replicates to nodes and Zones
3. Nodes return ack to coordinator
4. Coordinator returns ack to client
5. Data written to internal commit log disk

If a node goes offline, hinted handoff completes the write when the node comes back up.

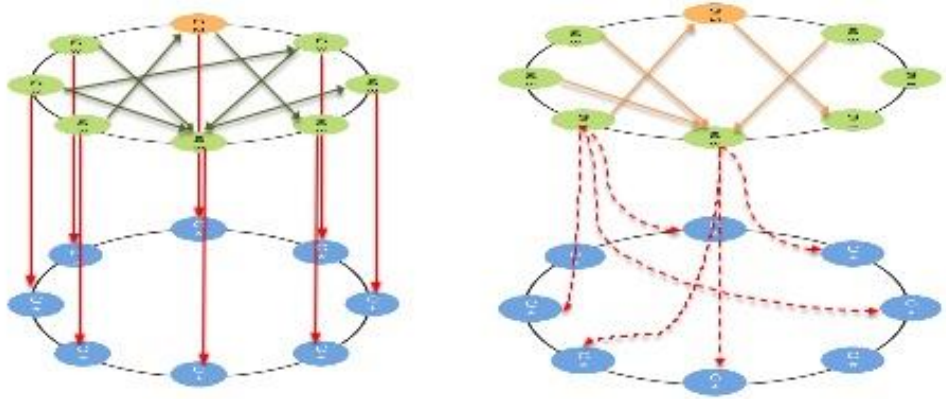
Requests can choose to wait for one node, a quorum, or all nodes to ack the write

SSTable disk writes and compactions occur asynchronously

NETFLIX

## Write data locality

- either stream data with Spark using *repartitionByCassandraReplica()*
- or flush data to Cassandra by async batches
- in any case, **there will be data movement on network** (sorry no magic)



# TL;DR

# TRY TO STAY OUT OF THE HYPE TRAIN…

# …BECAUSE MORE THAN TECH, IT'S ALL ABOUT THE ***MINDSET***!

"COVER YOUR ASS" (CYA)


"#TODOSJUNTOS!" ♥

# Q&A TIME !!

# (AND THANKS FOR WATCHING! ☺)