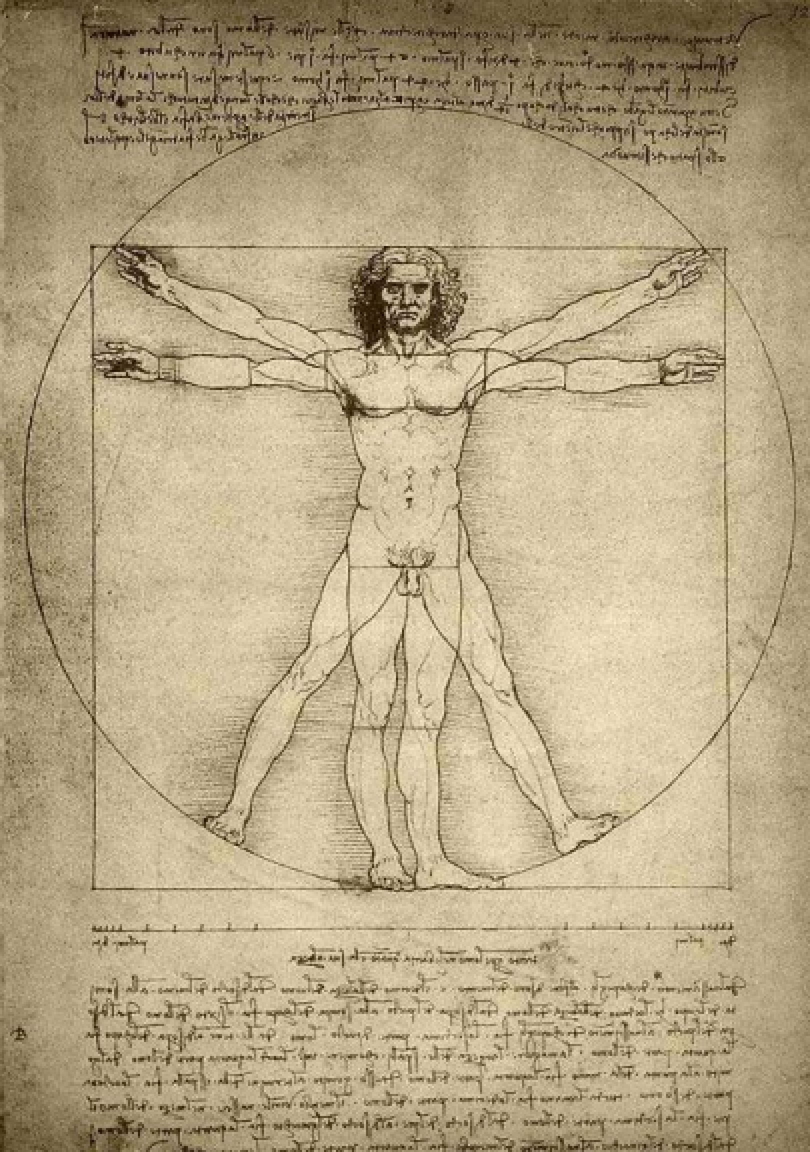
Data Engineering Demystified

Omid Vahdaty
Big Data Ninja

Welcome

Big Data Demystified Meetup

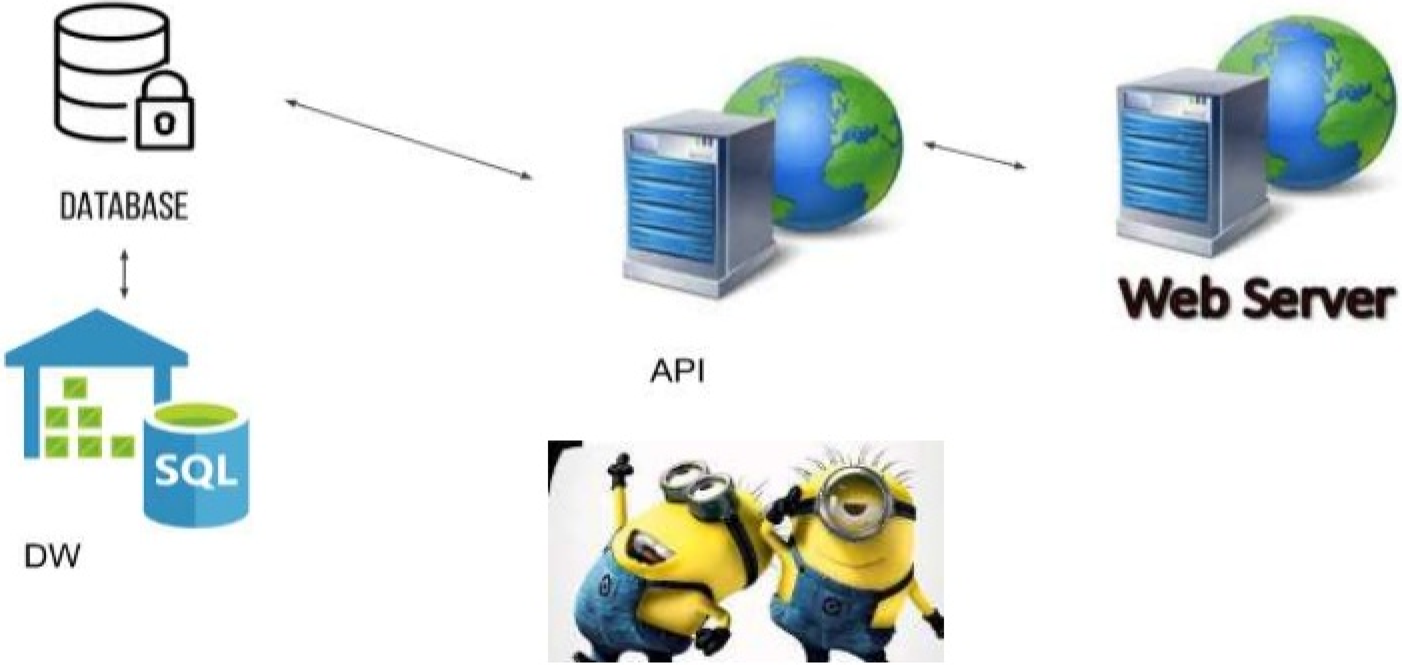


Disclaimer

- I am not the best, I simply love what I do VERY much.
- You are more than welcome to challenge me or anything I have to say as I could be wrong.

A long time ago
in a galaxy far, far away....

In the Past(web,api, ops db, data warehouse)



DATABASE

DW

API

Web Server

Then came Big Data...

Then came the cloud...

Then came the invoice ...

INFRASTRUCTURE
HADOOP ON PREMISE
HADOOP IN THE CLOUD
STREAMING / IN-MEMORY
ANALYTICS
DATA ANALYST PLATFORMS
DATA SCIENCE PLATFORMS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER SERVICE
NOSQL DATABASES
NEWSQL DATABASES
GRAPH DBS
MPP DBS
CLOUD EDW
BI PLATFORMS
VISUALIZATION
VERTICAL ANALYTICS
STATISTICAL COMPUTING
DATA SERVICES
HUMAN CAPITAL
LEGAL
FINANCE
ENTERPRISE PRODUCTIVITY
BACK OFFICE AUTOMATION
SECURITY
DATA TRANSFORMATION
DATA INTEGRATION
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
GOVERNMENT
FINANCE - LENDING
FINANCE - INVESTING
REAL ESTATE
INSURANCE
STORAGE
CLUSTER SERVICES
APP DEV
CROWDSOURCING
HARDWARE
SEARCH
LOG ANALYTICS
SOCIAL ANALYTICS
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
COMMERCE
OTHER
FRAMEWORK
QUERY / DATA FLOW
STREAMING
SECURITY
DATA SOURCES & APIS
HEALTH
IOT
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
INCUBATORS & SCHOOLS
RESEARCH

It keeps growing….

Solution?

+

=

Cloud

Big Data

Data Engineering

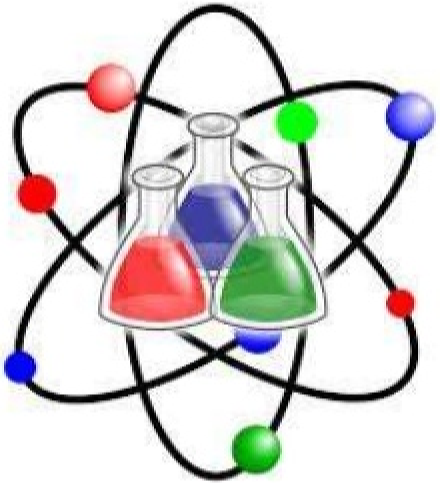
Part1

Jargon, Basic concepts
Basic questions

Data Engineering VS Data Science

- Architecture
- Data Platform scalability
  - Faster
  - Cheaper
  - Simpler
  - More secure
- Design ETL pipeline
- Network, Security & Regulation

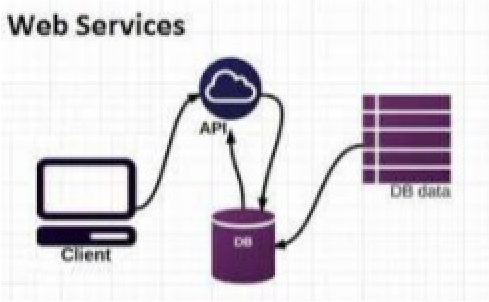
- Predictive analytics
  - Data
  - Recognition
  - User behaviour
  - NLP
- Recognition
  - Vision
  - Speech
  - Video



Data Science - API VS DS PaaS VS Hardcore DS

ML api

- General purpose algorithms
- Available in each cloud
- Speech recognition
- Image recognition
- Sentiment analysis
- Developer and Data engineering level.



Data Science as a service

- PaaS
- Notebook
- out of the box algorithm
- Data science pipeline from dev to production
- Scalable
- Zero devops
- Easy to get started even as data engineer

Amazon SageMaker

AutoML

Data Science Hardcore

- ML frameworks
- notebook
- Write your own neural networks
- Harder learning curve
- 100% data scientist

TensorFlow

mxnet

Cloud VS DC ?

Cloud

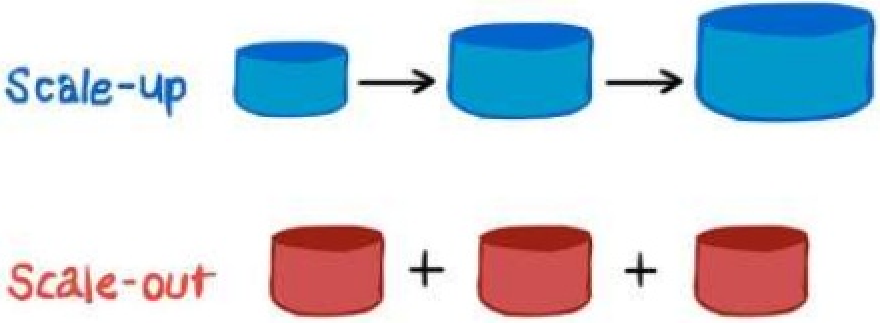
- Agile innovation
- Scalable
- Cheap to get started
- Easy to learn
- PaaS and managed services

Data Center

- Change require time
- Design for peek
- Costly to get started
- Harder to learn
- DIY

Which one is faster?
Which one is cheaper?
Which one is simpler?
Which one is more secure?

Scale Up VS Out



Scale Up

- Small cluster
- Usually active/passive
- Increase resources per machine
- Pros
  - Power Queries
  - Joins
- Cons
  - Parallelism

faster?
cheaper?
simpler?

Scale Out

- Add more servers
- Distributed : Each node can handle a fraction of the task
- Pros
  - Parallelism
- Cons
  - Power Queries
  - Joins

Fixed cost VS PayAsYouGo

amazon REDSHIFT

hadoop

Google Big Query

Amazon Athena

faster?

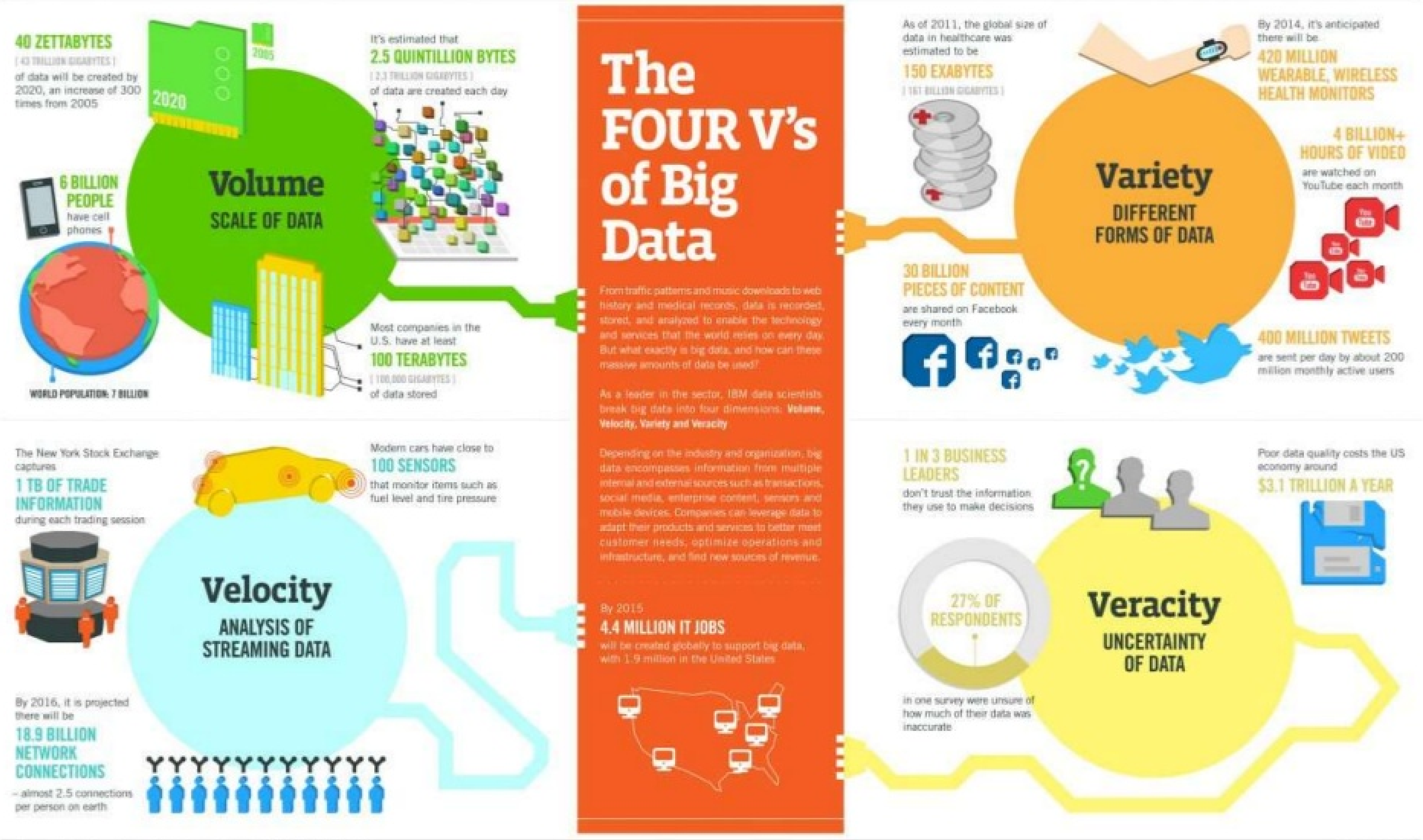
cheaper?

simpler?

Streaming VS batch Processing

the execution of a series of programs each on a set or "**batch**" of inputs, rather than a single input (which would instead be a custom job

Streaming Data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes)



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

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

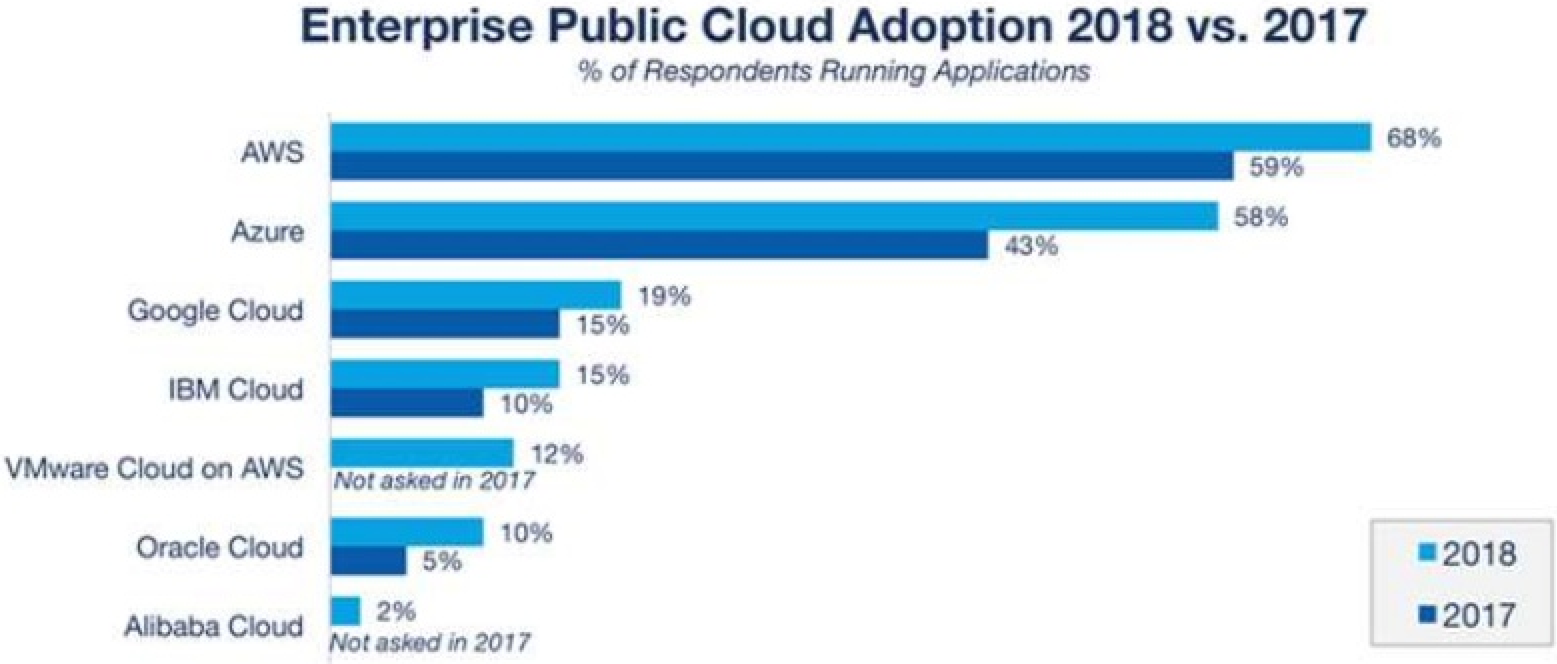
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

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

Part2

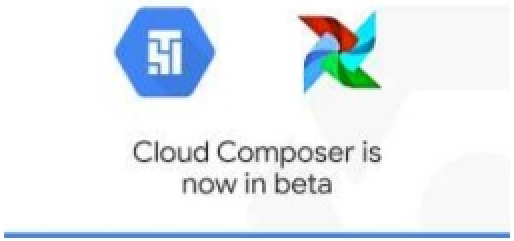
Big Data ?!
Big Questions?!

which Cloud?!


Enterprise Public Cloud Adoption 2018 vs. 2017
% of Respondents Running Applications
AWS
68%
59%
Azure
58%
43%
Google Cloud
19%
15%
IBM Cloud
15%
10%
VMware Cloud on AWS
12%
Not asked in 2017
Oracle Cloud
10%
5%
Alibaba Cloud
2%
Not asked in 2017
2018
2017

Data Engineering landscape @ GCP

Google Big Query


Cloud Composer is now in beta

Google Cloud Pub/Sub

BigTable

DataFlow

Cloud SQL

Data Engineering landscape @ AWS

amazon REDSHIFT

Spectrum

Amazon RDS

DynamoDB

Amazon Athena

AWS Glue

AWS SQS

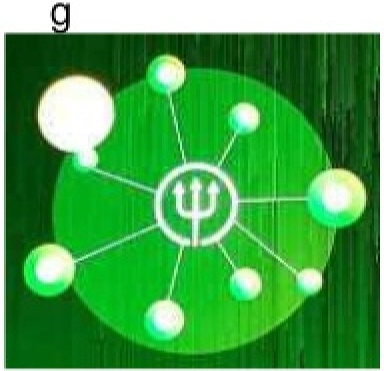
Amazon EMR

Amazon Kinesis Streams

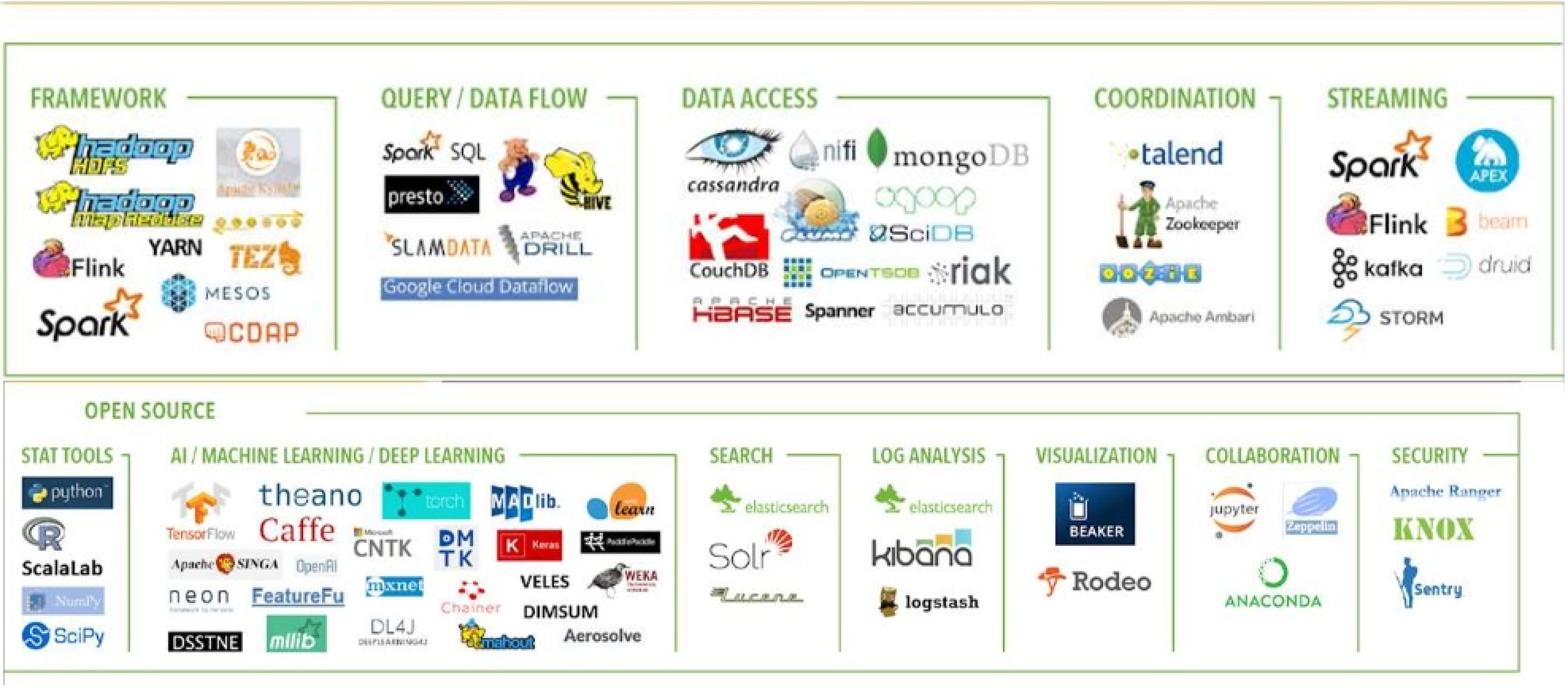
Amazon Kinesis Firehose

Amazon Kinesis Analytics

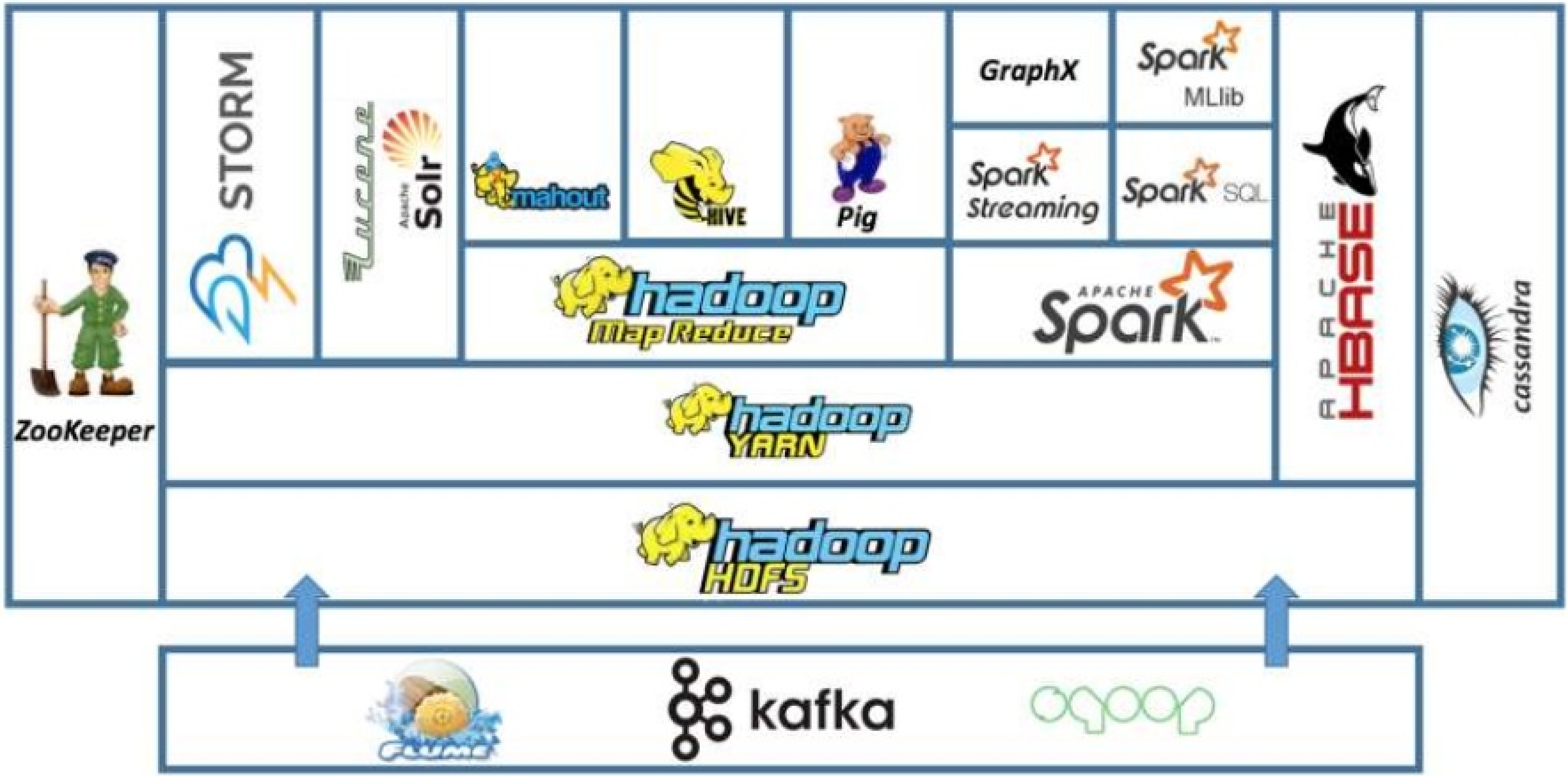
g



Data Engineering Landscape @ open source



Data Engineering Landscape @ Hadoop



DE challenges

- What is the company **use case** with data?
- **Where** should we build the data platform (**cloud or DC**)?
  - Which cloud? Which is one is cheaper?
- What **technologies** ?
  - Which new ones do we embrace why?
  - Which ones do we **depreciate** and why?
- Is the data **structured**? Semi structured? Unstructured?
- Is **SQL** good enough for the use case?
- How to build DE and DS cost effective **development pipeline**?
- How to **communicate** change in the company?
- How much time is spend on development (**query time/ wait time**)
- How much is going to **cost** me in the end of the month?
- How can we **simplify** the process of data development?
- **Regulation**?

Pop quiz, hotshot!

How much percent of the monthly infrastructure budget can saved by applying DE methodologies ?

Pop quiz, hotshot!

How much faster can your query run by applying DE methodologies ?

Pop quiz, hotshot!

How simple is it to use your data platform ?

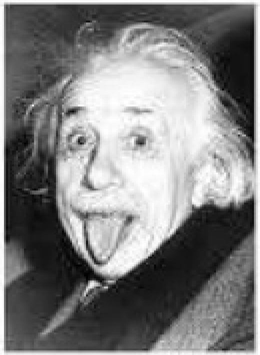
Summery

- If you have Big data problem you need a DE
  - Know your data use case
  - Choose your Cloud vendor carefully
  - Choose your tools that match use case
  - Big Data is not a buzzword it is an ecosystem
- Be sensitive to the COST
  - Understand underlying Infrastructure costs
  - Track Usage
  - Use PaaS to get started - get metrics
  - optimize as u go

Summary… Data Engineering is all about:

Faster

Cheaper



Simpler

"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

How to get started | Call for Action

Lectures: AWS Big data demystified lectures #1 until #4

AWS Big Data Demystified Meetup

Big Data Demystified meetup

My Next Meetups

GCP Big Data Demystified |

1. Investing.com Big Data Journey
2. BigQuery Demystified

Stay in touch...

- Omid Vahdaty
- +972-54-2384178
- https://big-data-demystified.ninja/
- Join our meetup, subscribe to youtube channels
  - https://www.meetup.com/AWS-Big-Data-Demystified/
  - https://www.meetup.com/Big-Data-Demystified/
  - Big Data Demystified YouTube
  - AWS Big Data Demystified YouTube
  - WhatApp group