UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

# BIG DATA and AI
## for business

### Introduction to Big Data and AI

**Decisions, Operations & Information Technologies**
**Robert H. Smith School of Business**
**Fall, 2020**

# Business Value of Big Data and AI

## Facebook: Why our 'next-gen' comms ditched MySQL • The ...

www.theregister.co.uk/2010/12/.../**facebook**_messages_tech/ ▼ The Register ▼

Dec 17, 2010 - So Seligstein and crew **mocked** up a multifaceted messaging prototype, ... **Facebook** was already using **MySQL** for message storage, the open source ... **Originally** built by Powerset – a semantic search outfit now owned by ... with the average message **going** to multiple people, and as the new system adds ...

## Twitter growth prompts switch from MySQL to 'NoSQL ...

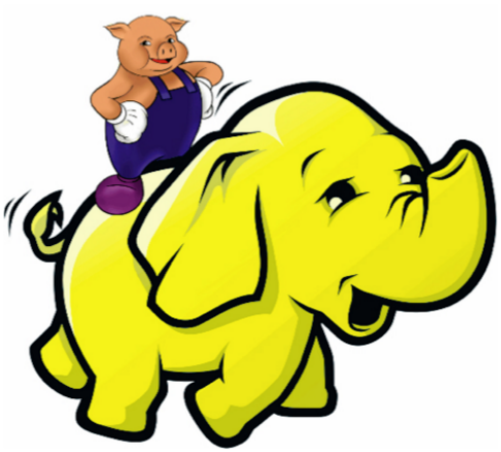
www.computerworld.com/.../**database**.../twitter-growth-p... ▼ Computerworld ▼

Feb 23, 2010 - Interested in **going** to one of the best colleges or universities to study technology? ... First developed by **Facebook** to augment its **MySQL** installation, ... Like **Facebook** and Twitter, Digg is also moving from **MySQL**, the **initial** ...

# Salesforce Hacker

03 NOVEMBER 2014

## Hadoop and Pig come to the Salesforce Platform with Data Pipelines



Event Log Files is big - really, really big. This isn't your everyday CRM data where you may have hundreds of thousands of records or even a few million here and there. One organization I work with does approximately twenty million rows of event data per day using Event Log Files. That's approximately 600 million rows per month or 3.6 billion every half year.

Because the size of the data does matter, we need tools that can orchestrate and process this data for a variety of use cases. For instance, one best practice when working with Event Log Files is to de-normalize Ids into Name fields. Rather than

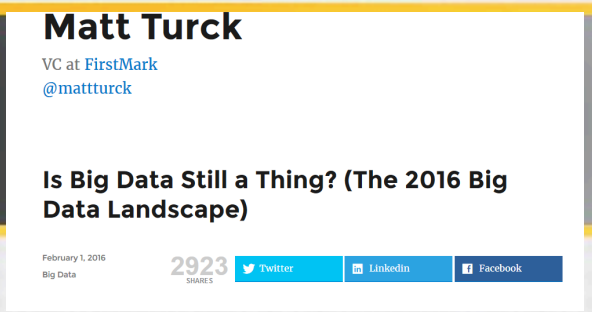
http://www.salesforcehacker.com/2014/11/hadoop-and-pig-come-to-salesforce.html

airbnb

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

"Airpal is a Web-based data-exploration and SQL query interface that runs on Presto, the in-memory SQL-on-Hadoop query technology that Facebook donated to Apache open source in late 2013. Airbnb invented Airpal because it needed a tool that would be more accessible to data analysts and even business users, not just the 23-person Airbnb data science team that handles Hive and Presto queries."

----Airbnb 3/5/2015 11:35 AM




UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- "VC investment in the space remains vibrant and the first few weeks of 2016 saw a flurry of announcements of big founding rounds for late stage Big Data startups: DataDog ($94M), BloomReach ($56M), Qubole ($30M), PlaceIQ ($25M), etc. Big Data startups received $6.64B in venture capital investment in 2015, 11% of total tech VC."

**Zoomdata raises $25 million for its real-time data visualization tool**

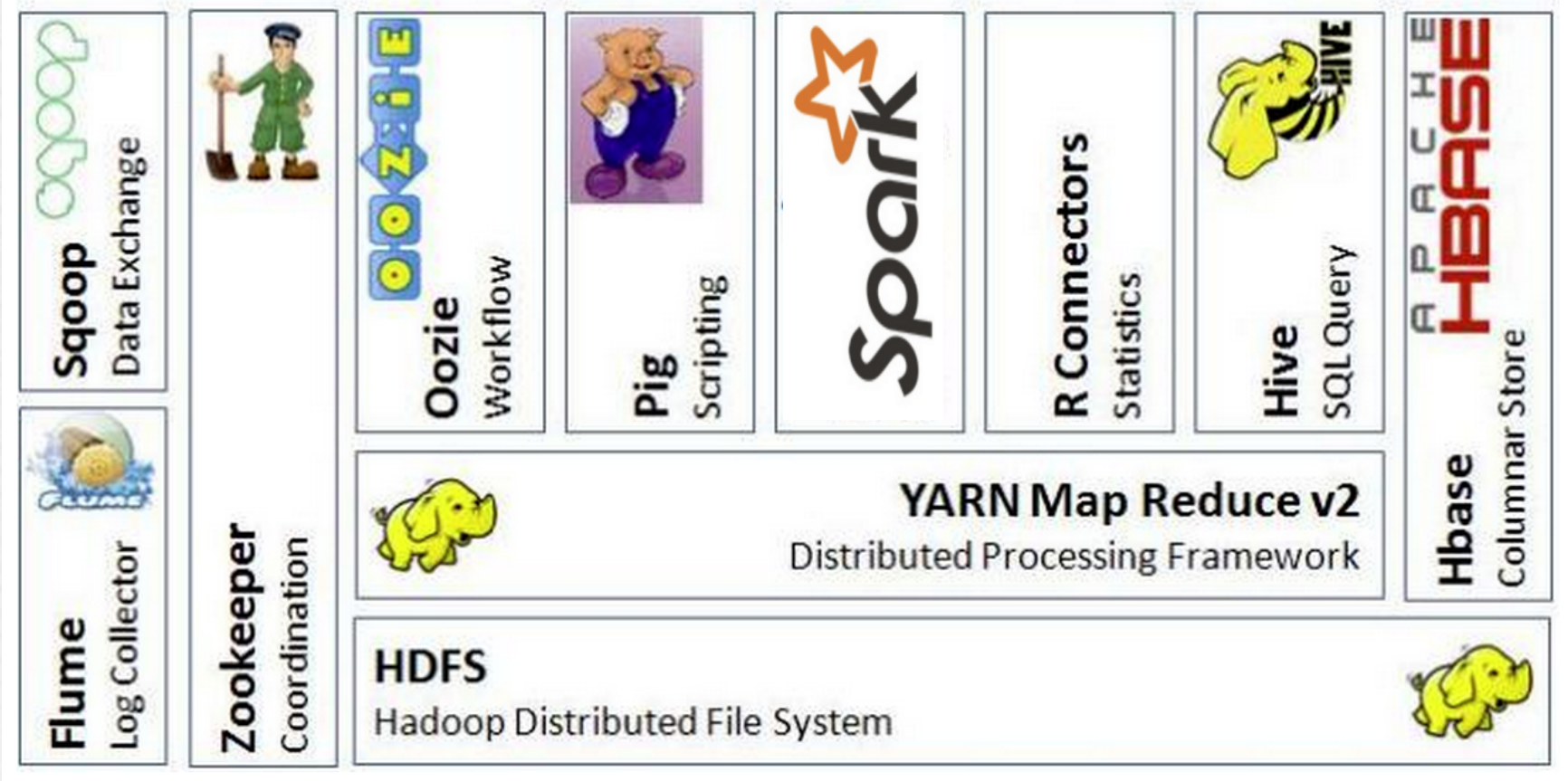
by Maria Deutscher | Feb 17, 2016 | 0 comments

Goldman Sachs

http://www.goldmansachs.com/our-thinking/pages/big-data.html

# Big Data ecosystem





# The open source community



- Yahoo!
  - Hadoop, Pig
  - Pig hides Java programming
- Facebook
  - Hive: provides SQL type functions for Hadoop files
- Netflix
  - Hbase: massage big data to be like a database
- UC Berkeley
  - Spark: in-memory processing to avoid the low disk I/O
- Twitter
  - Storm: near real-time streaming data

# Technology is still evolving rapidly



FUTURE OF WORK

## Why Cloudera is saying 'Goodbye, MapReduce' and 'Hello, Spark'

Derrick Harris
Sep 09, 2015

# And the al-mighty AI!



- 2012 Matlab
- 2013 Caffe
- 2014 Theano
- 2015 Torch
- 2016/7 TensorFlow
- 2018 PyTorch
- 2019 ???

- CNN, RNN, VAE, GANs …

- Sergey Brin @ 2017 Davos World Economic Forum
  - https://www.youtube.com/watch?v=jYuCVcGxtNM

# So, what's going on?

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- You need critical thinking to not get lost

# How does data generate value?

# Big data processes

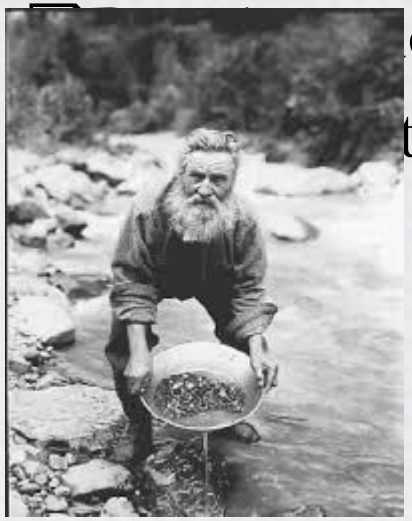


- Load data
- Clean up data
- Transform data
- Query data
- Machine learning/deep learning

# Realizing the benefits of Big Data

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Setting up Hadoop is just the beginning!
  - It just means that you are enabled to handle the big data
  - ot gua
  - te your



# The easy ones



- Faster and cheaper
  - In late 2007, the New York Times wanted to make available over the web its entire archive of articles, 11 million in all, dating back to 1851. Four-terabyte pile of images in TIFF format needed to translate that four-terabyte pile of TIFFs into more web-friendly PDF files.
    - Not a particularly complicated but large computing chore,
      - requiring a whole lot of computer processing time.

- a software programmer at the Times, Derek Gottfrid,
  - playing around with Amazon Web Services, Elastic Compute Cloud (EC2),
    - uploaded the four terabytes of TIFF data into Amazon's Simple Storage System (S3)
    - In less than 24 hours, 11 millions PDFs, all stored neatly in S3 and ready to be served up to visitors to the Times site.

- The total cost for the computing job? $240
  - 10 cents per computer-hour times 100 computers times 24 hours

# How to make data "actionable"

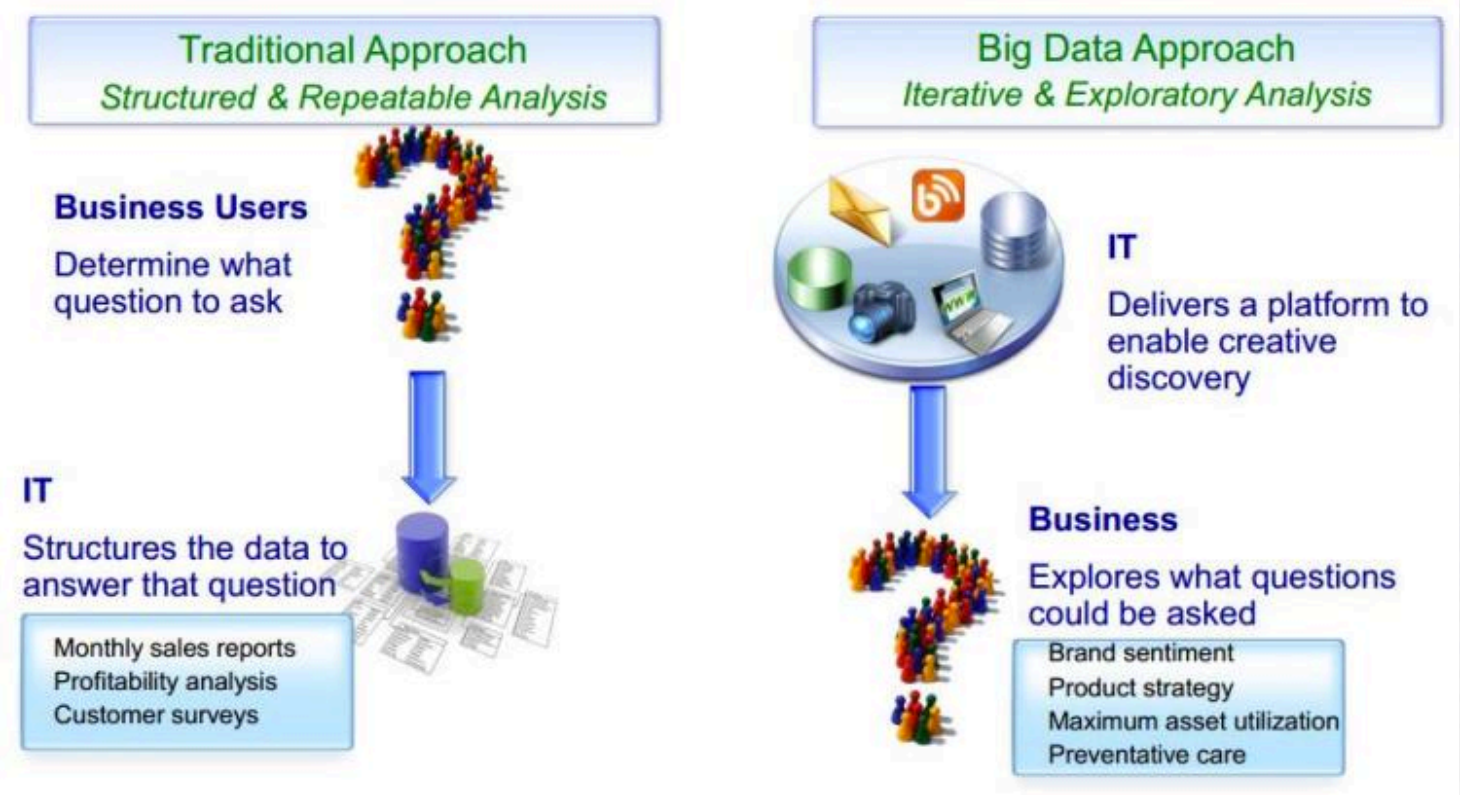
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- D-D-P-P

  - ❑**D**escriptive: what happened?
  - ❑**D**iagnostic:  why did it happen?
  - ❑**P**redictive: what is likely to happen?
  - ❑**P**rescriptive: what is the best course of action?

Courtesy of Cupid Chan

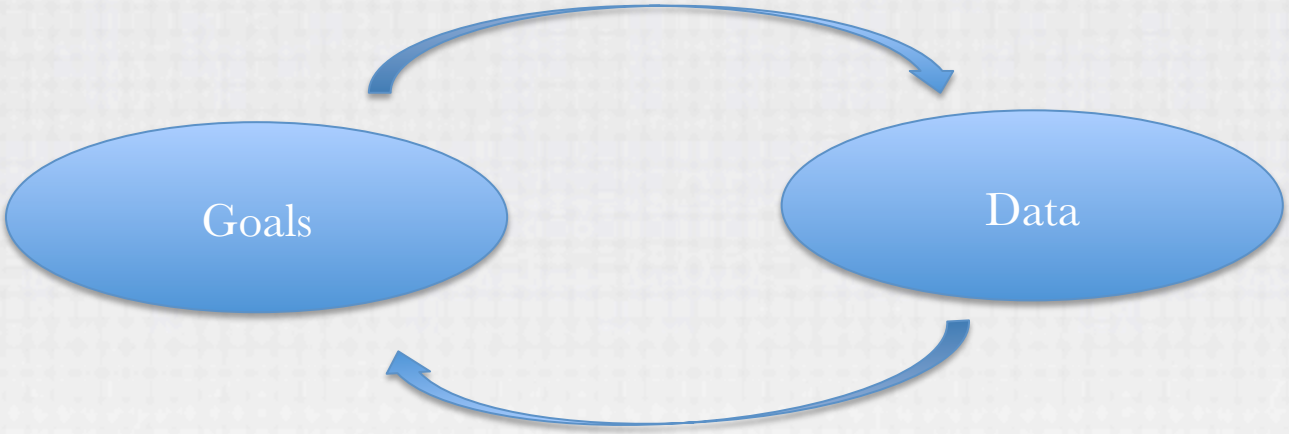
# Traditional vs. Big Data Approach



# A dynamic process



- What are the business goals and critical issues?
- What data do you have?
- What data can you potentially capture?
- What analytical tools could be applied?




**Goal: find business questions that can harness the power of big data**