YMC

# Building Scalable Big Data Pipelines

NOSQL SEARCH ROADSHOW ZURICH

**Christian Gügi, Solution Architect**

19.09.2013

# AGENDA

- Opportunities & Challenges
- Integrating Hadoop
- Lambda Architecture
- Lambda in Practice
- Recommendations

# ABOUT ME

- Solution Architect @ YMC
- Founder and organizer Swiss Big Data User Group
  - http://www.bigdata-usergroup.ch/
- Contact
  - christian.guegi@ymc.ch
  - http://about.me/cguegi
  - @chrisgugi

## ABOUT YMC

- Founded in 2001
- Based in Kreuzlingen, Switzerland
- Big Data Analytics, Web Solutions and Mobile Applications
- 24 experts
  - Consulting, creation, engineering

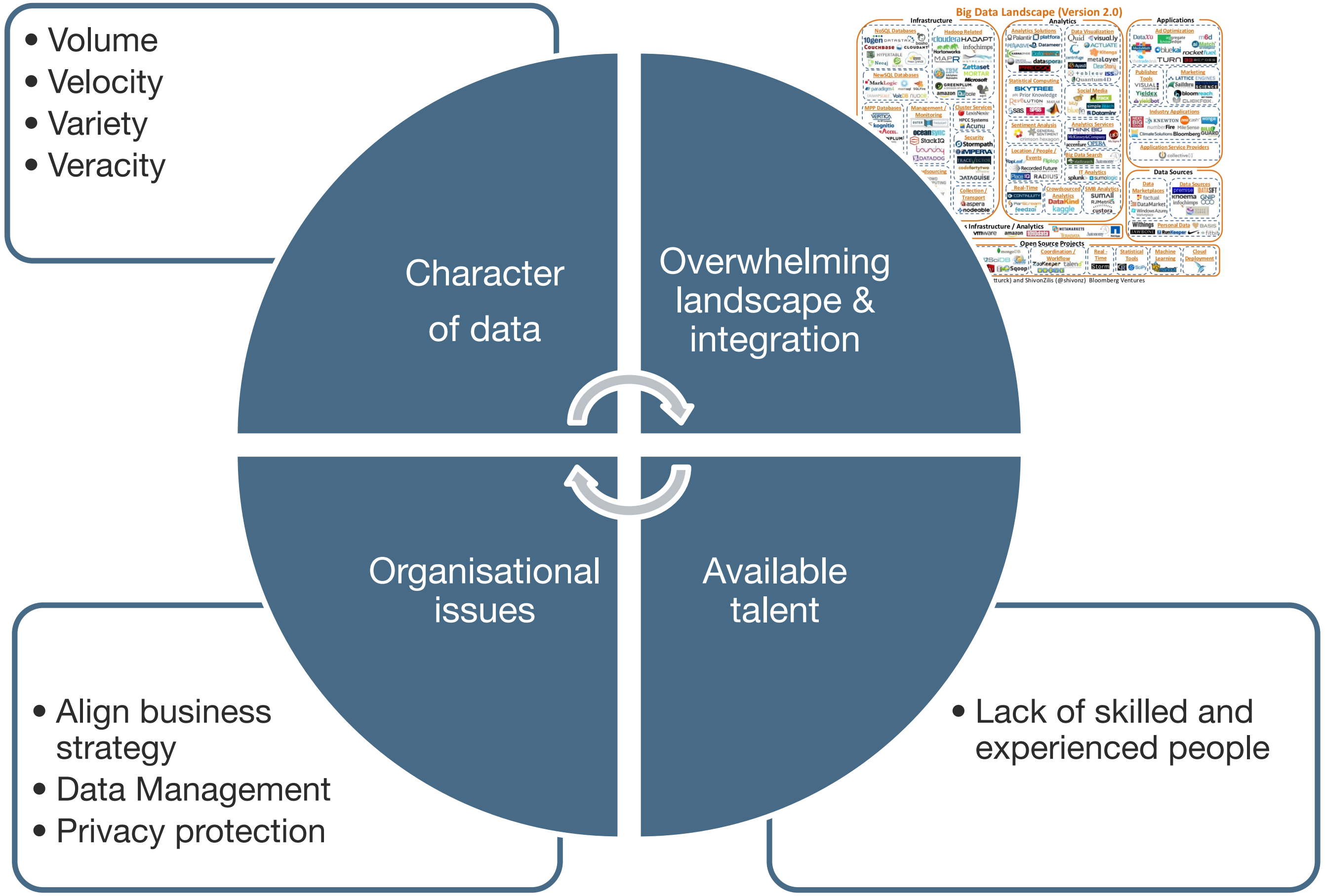
# OPPORTUNITIES & CHALLENGES

# BIG DATA – WHAT IS THE BIG DEAL?

A. New sources and types from *inside* & *outside* organisations
   - "Internet of things", sensors, RFID, intelligent devices, etc.
   - Unstructured information – documents, web logs, email, social media, etc.
   - Trusted 3rd party sources – industry provider & aggregators, governments "Open Data", weather, etc.

B. Technology innovations to exploit new world of data
   - Low cost storage and process power (cloud, on-premise & hybrid)
   - New software patterns to handle speed & volume, structured and unstructured (In-memory computation, Hadoop, Mapreduce, etc.)
   - Revolution in user experience, analytics, recommendations

# BIG DATA – CHALLENGES



- Volume
- Velocity
- Variety
- Veracity

Character of data

Overwhelming landscape & integration

Organisational issues

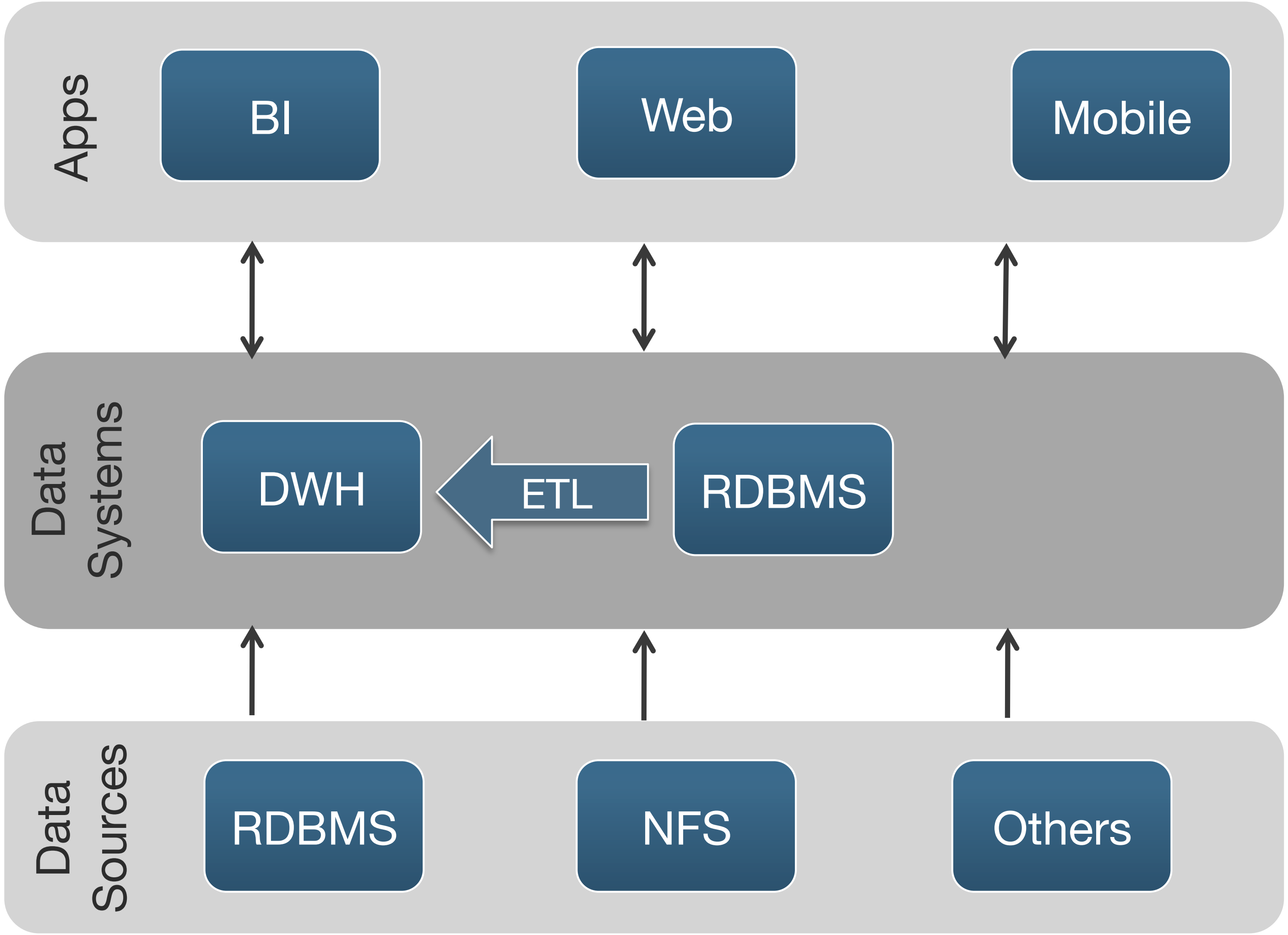
Available talent

- Align business strategy
- Data Management
- Privacy protection

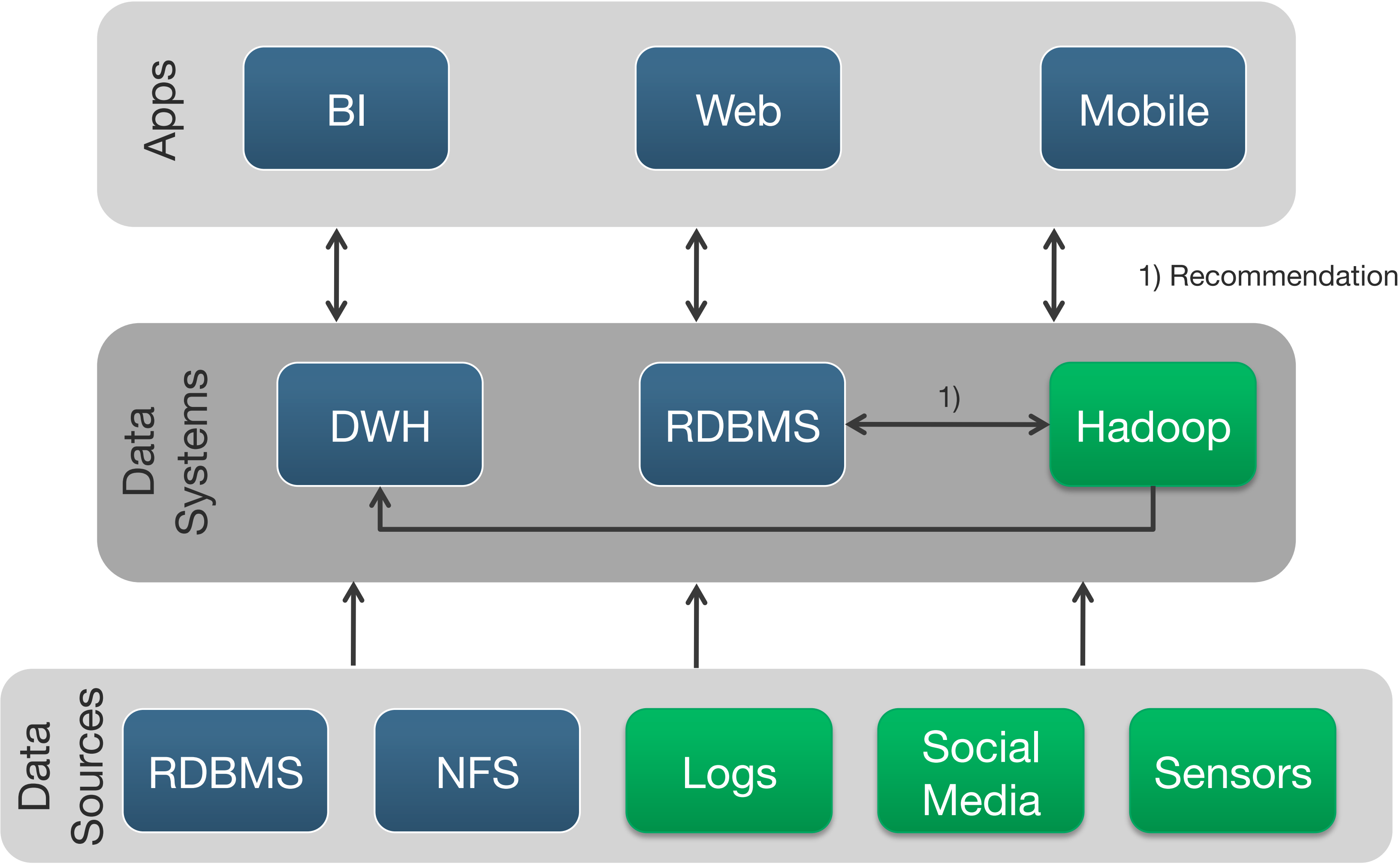
- Lack of skilled and experienced people

# INTEGRATING HADOOP

# TYPICAL RDBMS SZENARIO

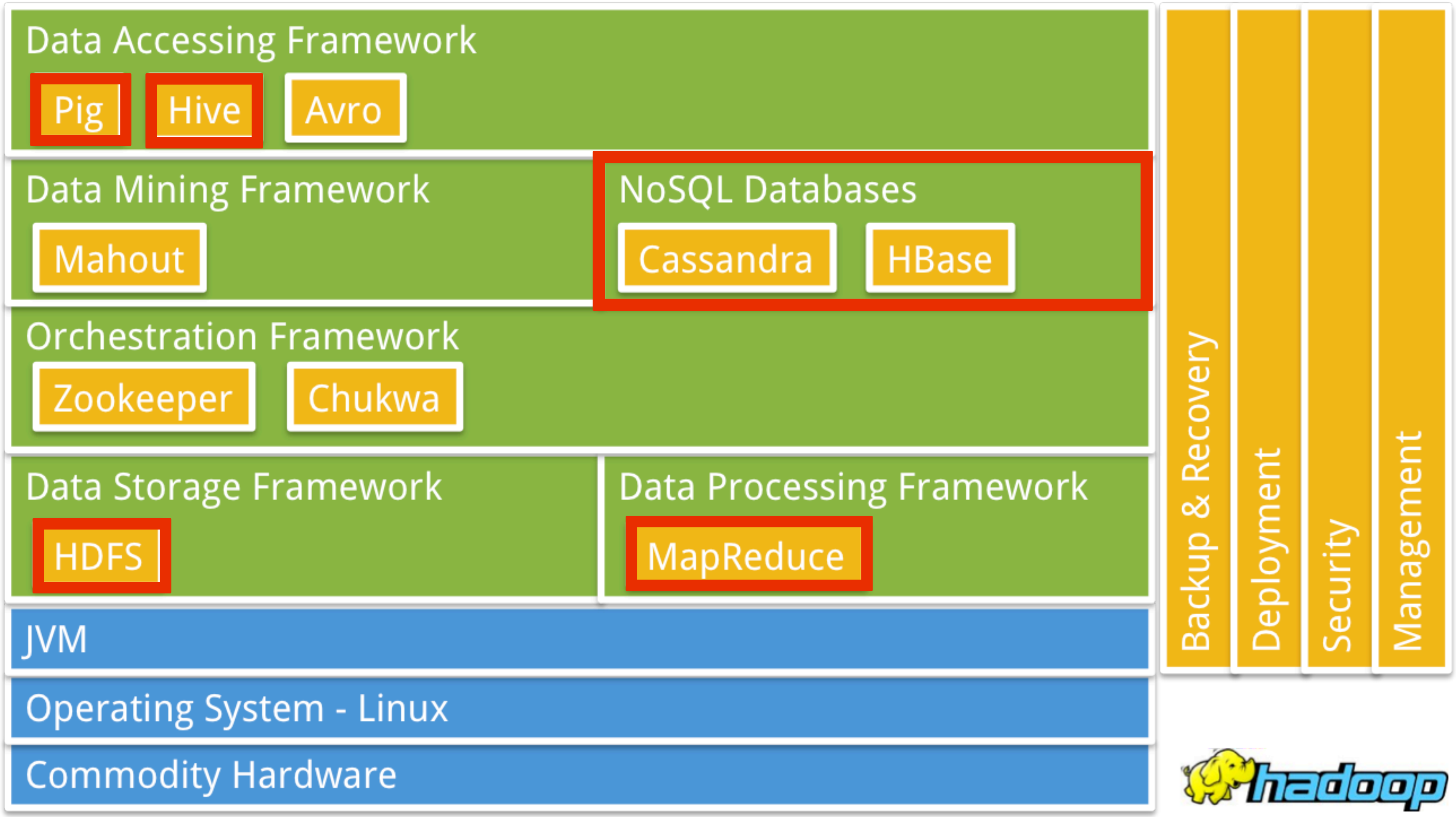


# BIG DATA SZENARIO



1) Recommendations, etc.

# HADOOP ECOSYSTEM



Data Accessing Framework

Pig

Hive

Avro

Data Mining Framework

Mahout

NoSQL Databases

Cassandra

HBase

Orchestration Framework

Zookeeper

Chukwa

Data Storage Framework

HDFS

Data Processing Framework

MapReduce

JVM

Operating System - Linux

Commodity Hardware

Backup & Recovery

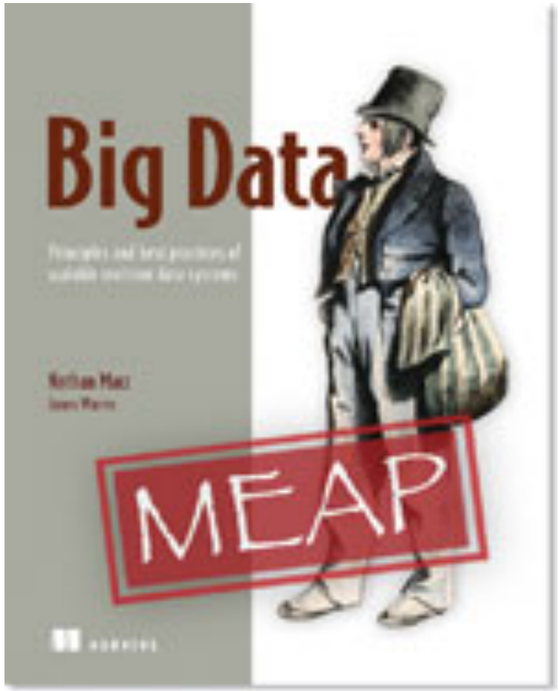
Deployment

Security

Management

# LAMBDA ARCHITECTURE

# LAMBDA ARCHITECTURE



- Credits Nathan Marz
- Former Engineer at Twitter
- Storm, Cascalog, ElephantDB

http://www.manning.com/marz/

# DESIGN PRINCIPLES

Lambda Architecture

- Human fault-tolerance
- Data immutability
- Re-computation

# HUMAN FAULT-TOLERANCE

Lambda Architecture

- Design for human error
    - Bugs in code
    - Accidental data loss
    - Data corruption
- Protect good data, so you can always fix what went wrong

# DATA IMMUTABILIY

Lambda Architecture

- Store data in it’s rawest form
- Create and read but no *update*
- No data can be lost
  - To fix the system just delete bad data
  - Can always revert to a true state

# DATA IMMUTABILIY

Lambda Architecture

## Capturing change traditionally (mutability)

| Name | Location |
|---|---|
| Alice | Zurich |
| Bob | Lucerne |
| Tom | Bern |

→

| Name | Location |
|---|---|
| **Alice** | **Basel** |
| Bob | Lucerne |
| Tom | Bern |

## Capturing change (immutability)

| Name | Location | Time |
|---|---|---|
| Alice | Zurich | 2009/03/29 |
| Bob | Lucerne | 2012/04/12 |
| Tom | Bern | 2010/04/09 |

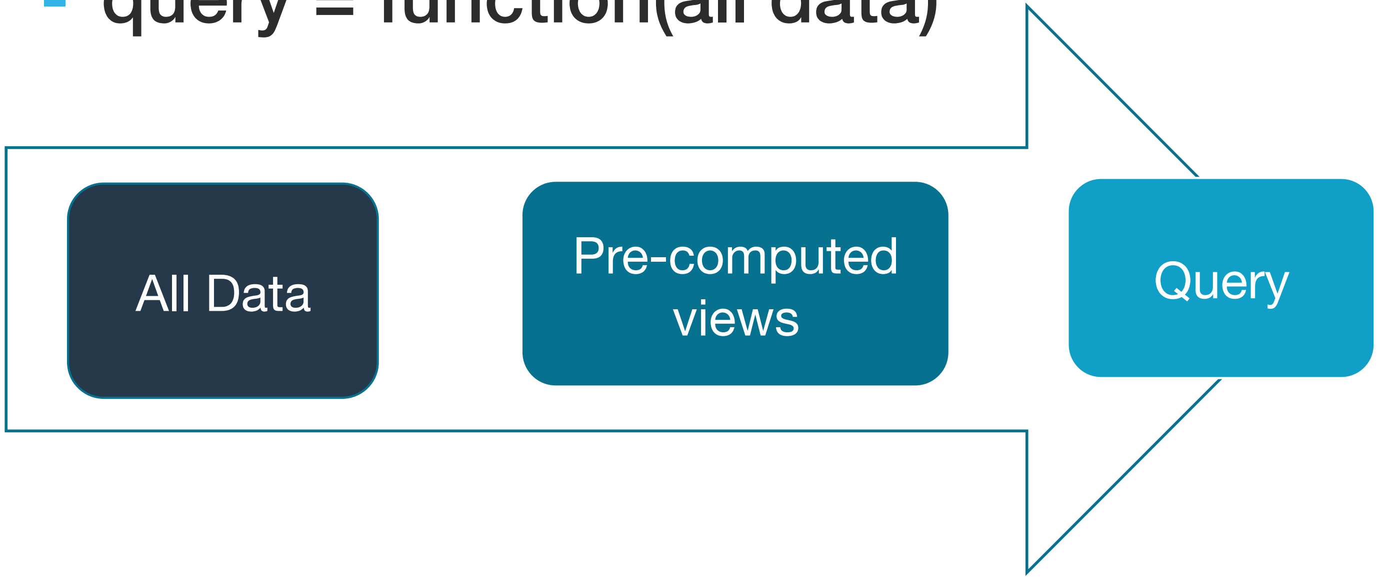
→

| Name | Location | Time |
|---|---|---|
| **Alice** | **Zurich** | **2009/03/29** |
| Bob | Lucerne | 2012/04/12 |
| Tom | Bern | 2010/04/09 |
| **Alice** | **Basel** | **2013/08/20** |

# RE-COMPUTATION

Lambda Architecture

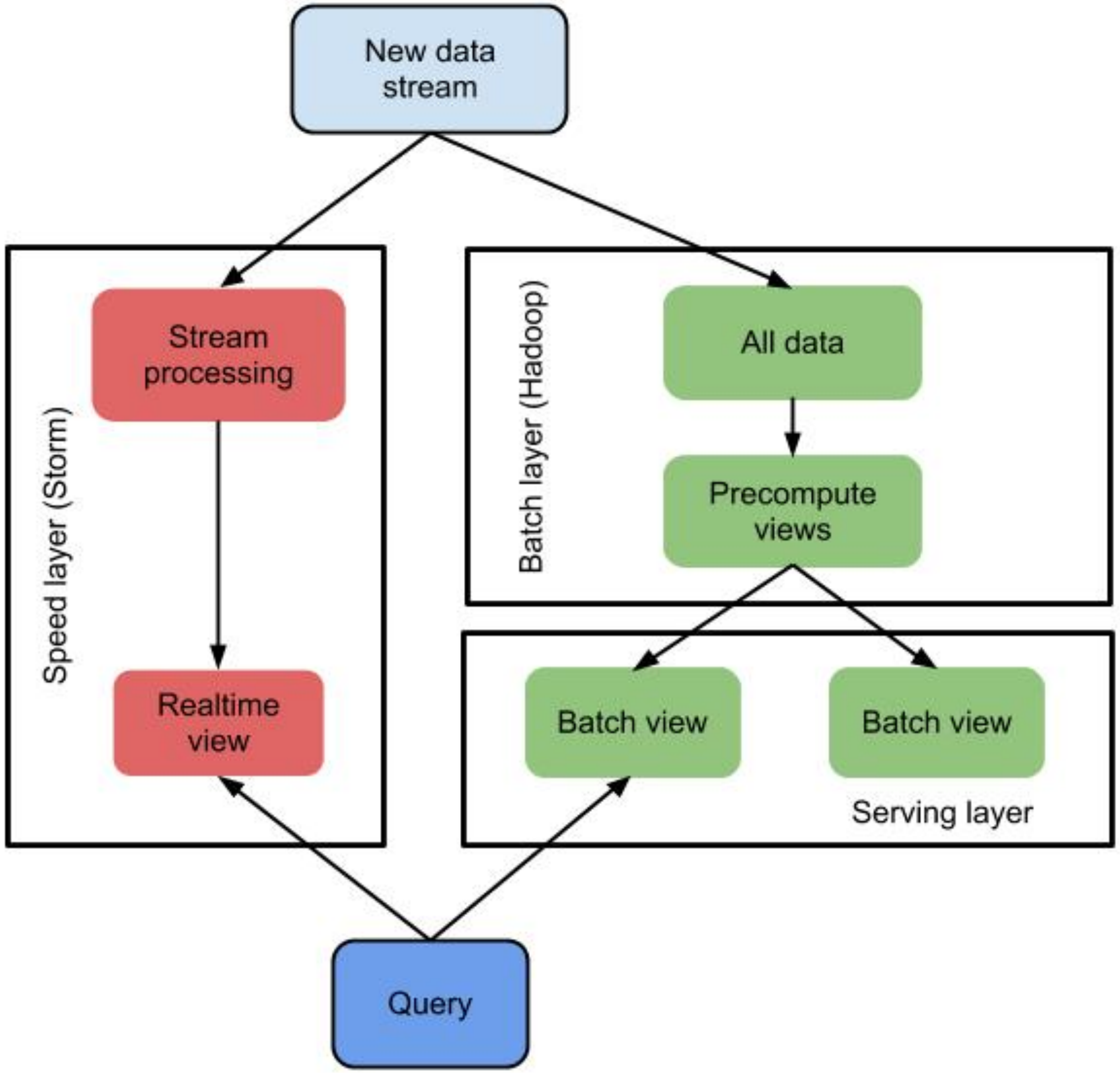
- Always able to re-compute from historical data
- Basis for all data systems
  - query = function(all data)



YMC

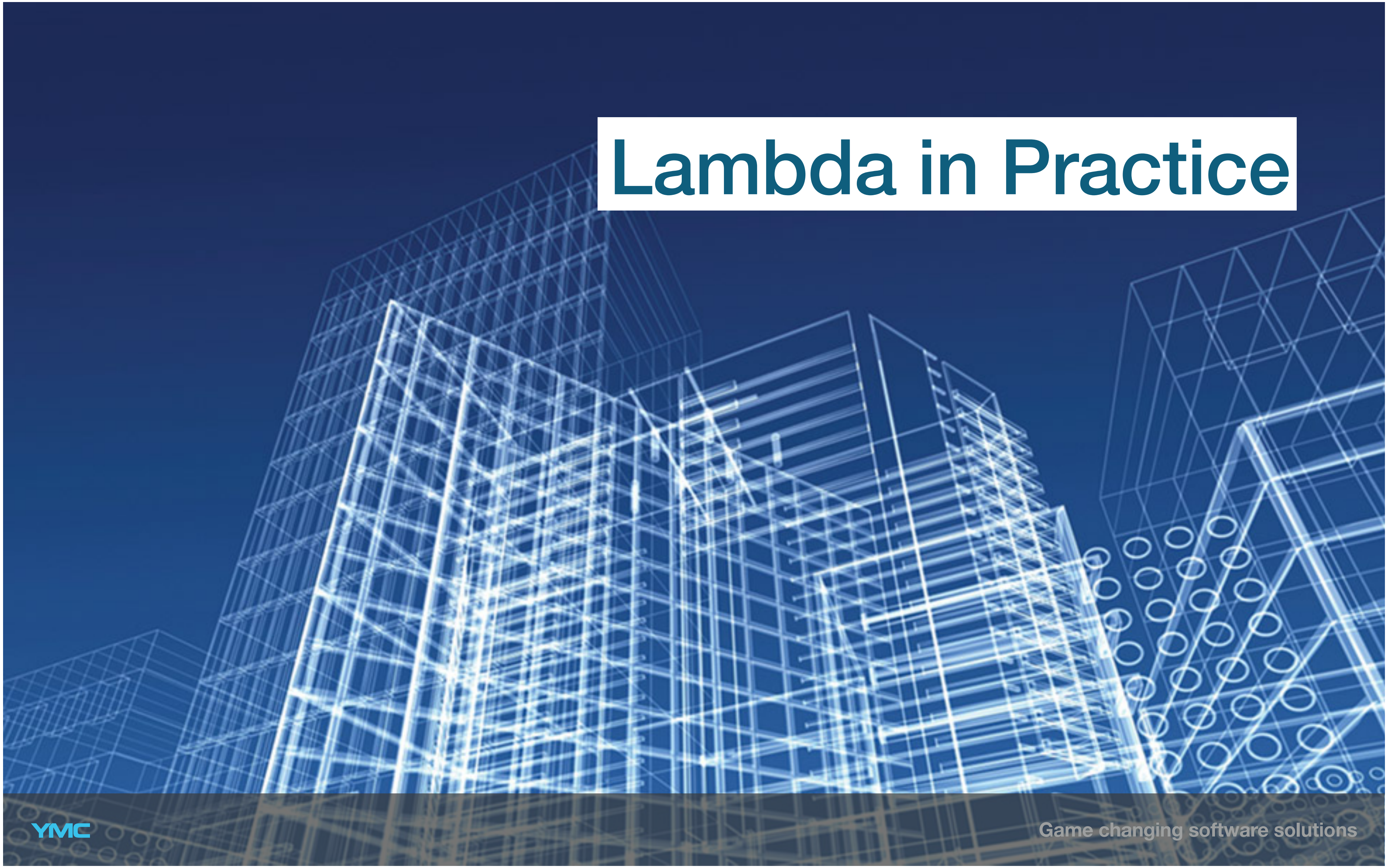
# LAYERS

**Lambda Architecture**



http://www.ymc.ch/en/lambda-architecture-part-1

# Lambda in Practice



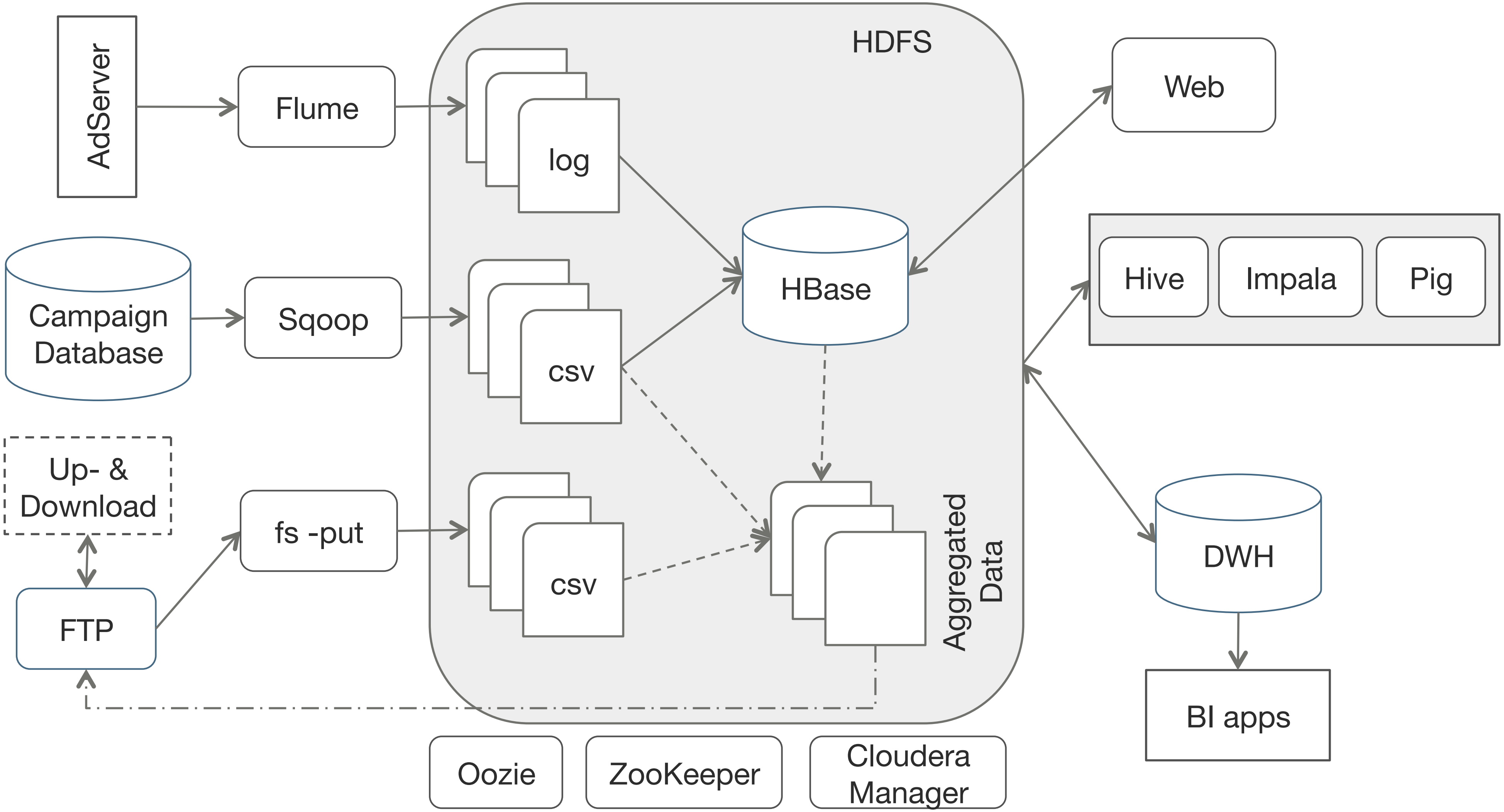
YMC

Game changing software solutions

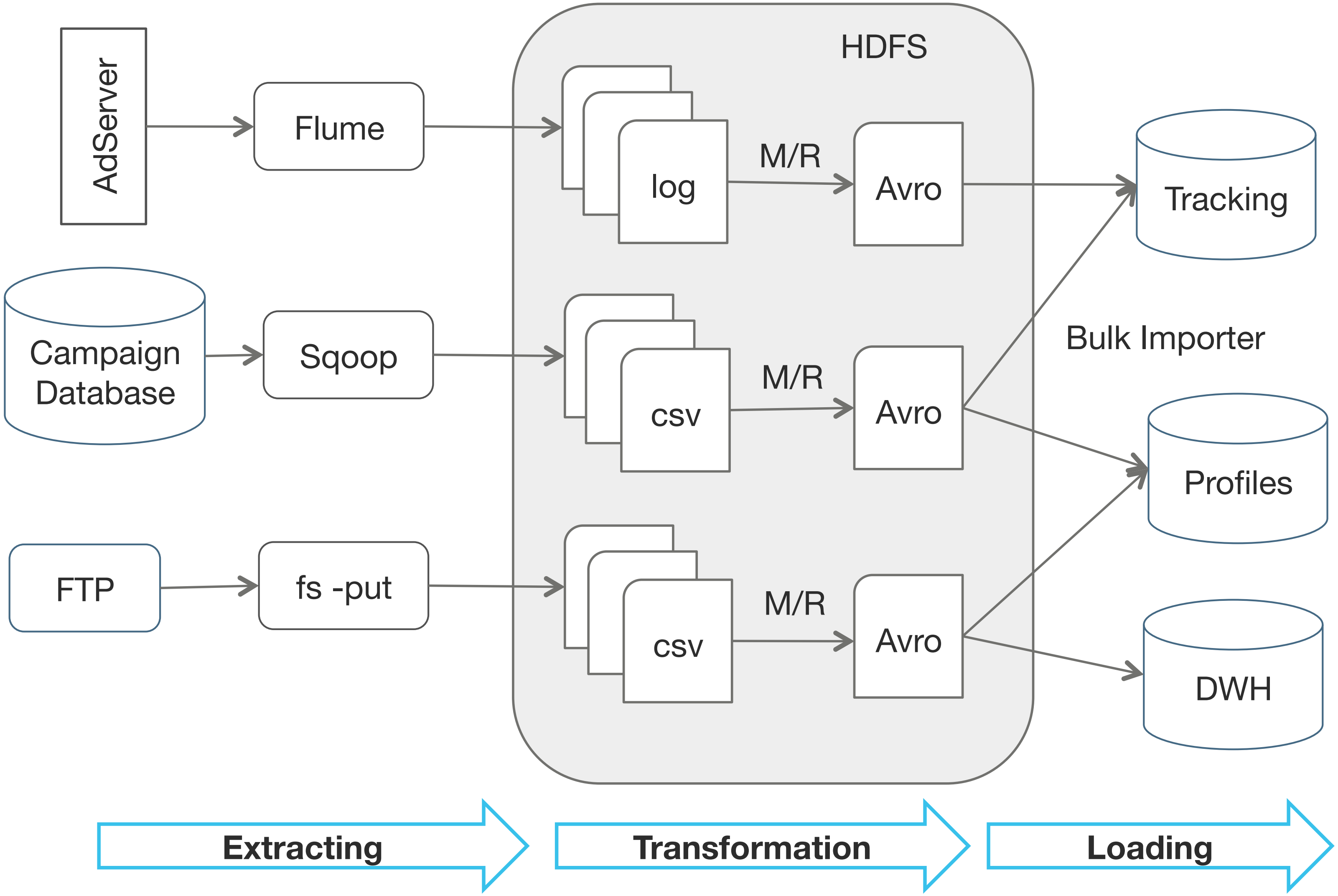
## ONLINE MARKETING

- Tracking and analytics solution
- Improve customer targeting and segmentation
- Various reports
- Real-time not required

## OVERVIEW


AdServer
Flume
Campaign Database
Sqoop
Up- & Download
FTP
fs -put
HDFS
log
csv
csv
HBase
Aggregated Data
Web
Hive
Impala
Pig
DWH
BI apps
Oozie
ZooKeeper
Cloudera Manager

# DATA PIPELINE



# ADVANTAGES

- Extensible – easily add speed layer later on
- Complements existing DWH/BI system
- ETL phases are decoupled
- Reliable
  - Infrastructure
  - Each step can be replayed
- Scalable
  - Storage
  - Processing
- Highly available
- Ad-hoc analysis right from the beginning

# RECOMMENDATIONS

# RECOMMENDATIONS

- Not a fixed, one-size-fits-all approach
  - Adopt to your needs/requirements
- Hadoop complements existing systems
- How real-time do I need to be?
- Immutability and pre-computation are just good ideas!
  - Store information in rawest format possible
  - Use a serialization framework (Avro, Thrift, Protocol Buffers)

# THANK YOU!

YMC

Game changing software solutions

# CONTACT

christian.guegi@ymc.ch

Tel. +41 (0)71 508 24 76

www.ymc.ch

@chrisgugi

Photo Credits:
Slide 05: Success opportunity achieve by Stephen McCulloch
Slide 08: Matrix by Gamaliel Espinoza Macedo.
Slide 12: Layers by Katelyn Leblanc
Slide 20: Mining For Information by JD Hancock
Slide 27: Warning Question by longzijun