(Not Just Another) Overview of Apache Hadoop

Bob Wakefield
Principal
bob@MassStreet.net
Twitter:
@BobLovesData

Who is Mass Street?

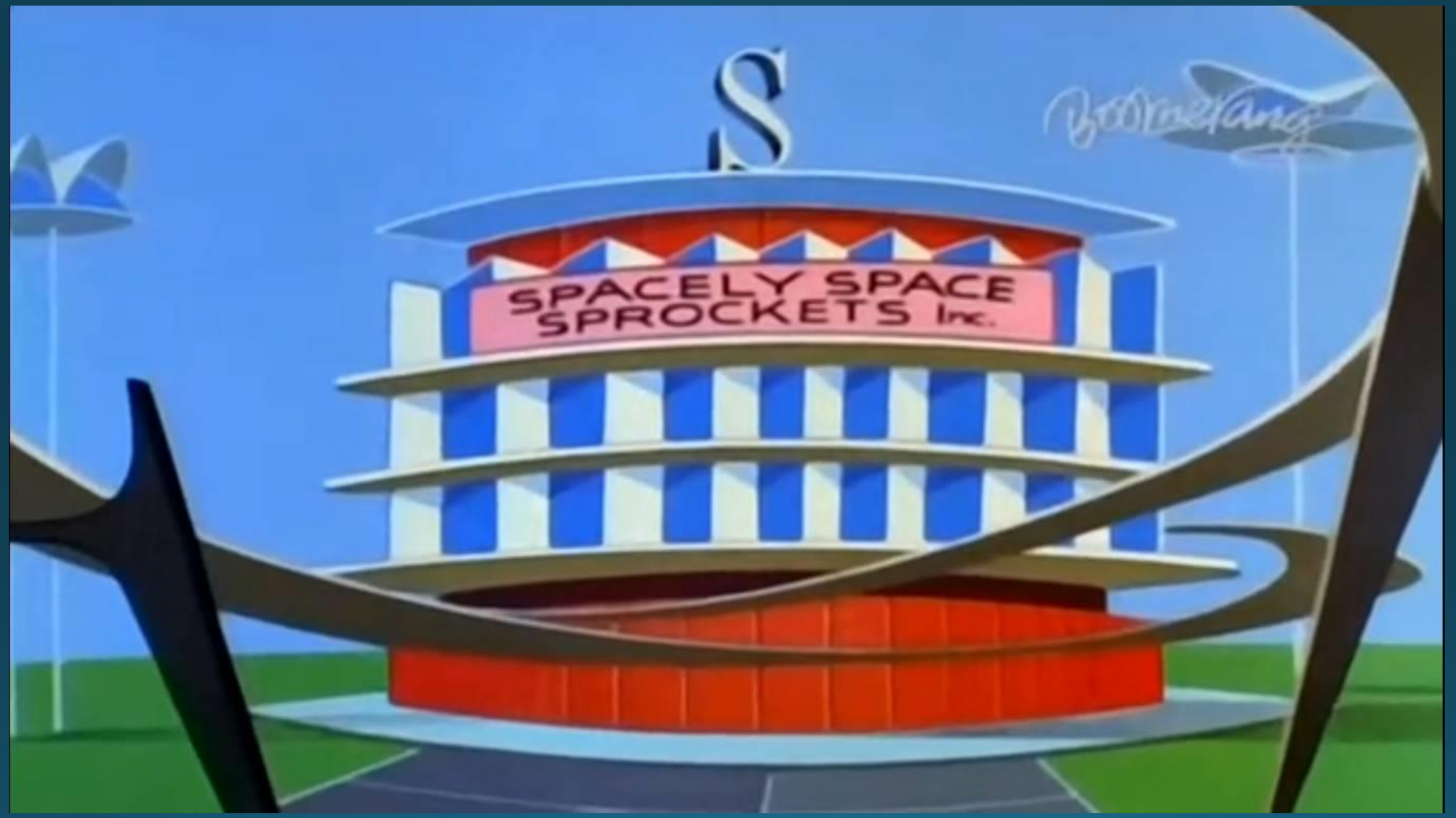
- Sole proprietor data consultancy
- Will start providing Big Data solutions in the near future
- Looking for partners
  - Especially Hadoop engineers

This evening's objectives

1. Convince you that Hadoop is awesome!
2. Convince you to convince your boss that Hadoop is awesome!

Yo Bob! I thought we were going to learn how to do stuff with MongoDB tonight!

There is a disconnect between the hype (and reality) of “Big Data” and the number of organizations that are ready to DO “Big Data”.


SPACELY SPACE
SPROCKETS Inc.

How do I know if I should be taking a look at Hadoop?

If you have to make hard choices about how much historical data you are going to store... you might need Hadoop.

If your analyst are spending more time fixing clunky ETL processes that look like they were designed by Rube Goldberg instead of delivering results to decision makers...

you might need Hadoop.

If you are doing crazy inappropriate things with your warehouse load to get closer to real time analytics... you might need Hadoop.

If you're trying to shove unstructured data into a RDBMS... you might need Hadoop.

If you're even thinking about doing a data warehouse project... you might need Hadoop.

If you're spending hundreds of thousands of dollars on data storage solutions... you might need Hadoop.

EDW = $15,000 - $80,000 per TB
Hadoop = $2,000 - $6,000 per TB
Source: Santosh Chitakki, VP of Appfluent

If you are having a hard time crunching numbers with the resources at your disposal... you might need Hadoop.

What is Hadoop?

- The savior of us all!
- More mature than you think!
  - Been around since 2006
- Hadoop is for everybody!

What is Hadoop?

- It's a paradigm
- It's a framework
- It's a collection of software
- It's a partridge in a pear tree

No Really! What is Hadoop?

- Provides distributed fault tolerant data storage
- Provides linear scalability on commodity hardware
- Translation: Take all of your data, throw it across a bunch of cheap machines, and analyze it. Get more data? Add more machines

So you're going to try to explain Hadoop to your boss? Just say no to technobabble.

Top Reasons to Implement Hadoop

Reason #1: Hadoop makes money

Reason #2: Hadoop saves money

Reason #1: Hadoop makes money

- Cottage industry growing up around Big Data
- Turns data into a potential source of revenue
- Enables the kind of wiz bang analysis you're always hearing about

Reason #2: Hadoop saves money

- Drastically reduces the cost of storing data
- Eliminates a lot of the ETL intensive work found in the old world

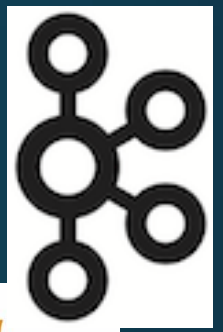
OK Bob. You've piqued my interes
Where do I go to get started with this stuff?


Apache Ambari
http://incubator.apache.org/ambari


APACHE FALCON


Spark


oozie




hadoop


TEZ


AVRO
TM


HIVE


APACHE
HBASE


Apache
Solr


mahout


chukwa


Cassandra

It's like a child's erector set!

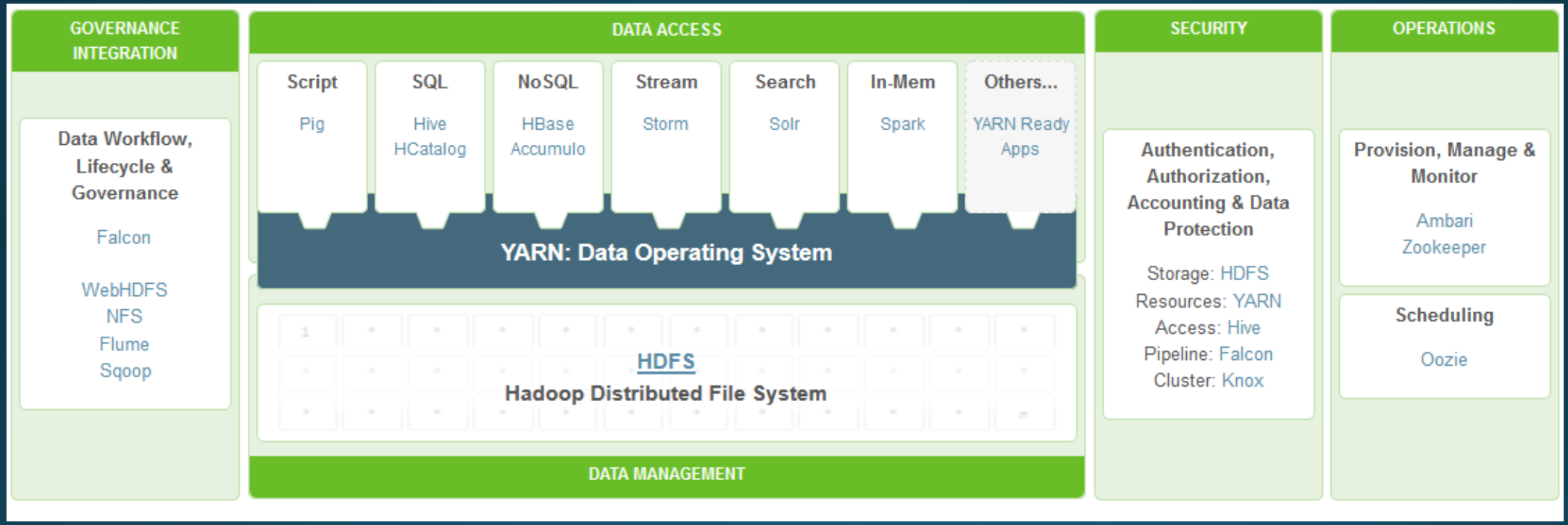


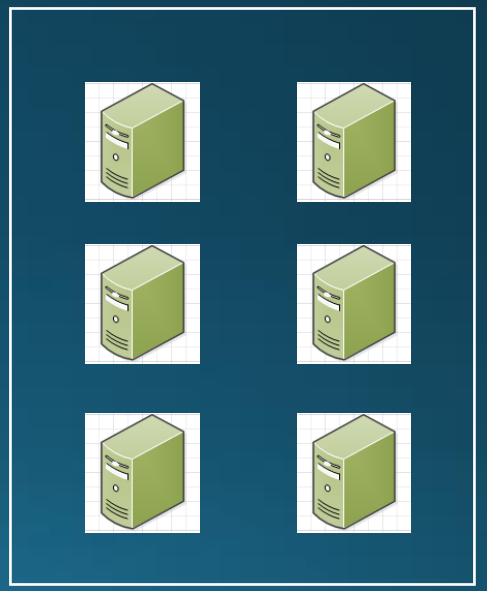
Photo Credit: Hortonworks website

Node

- A single computer

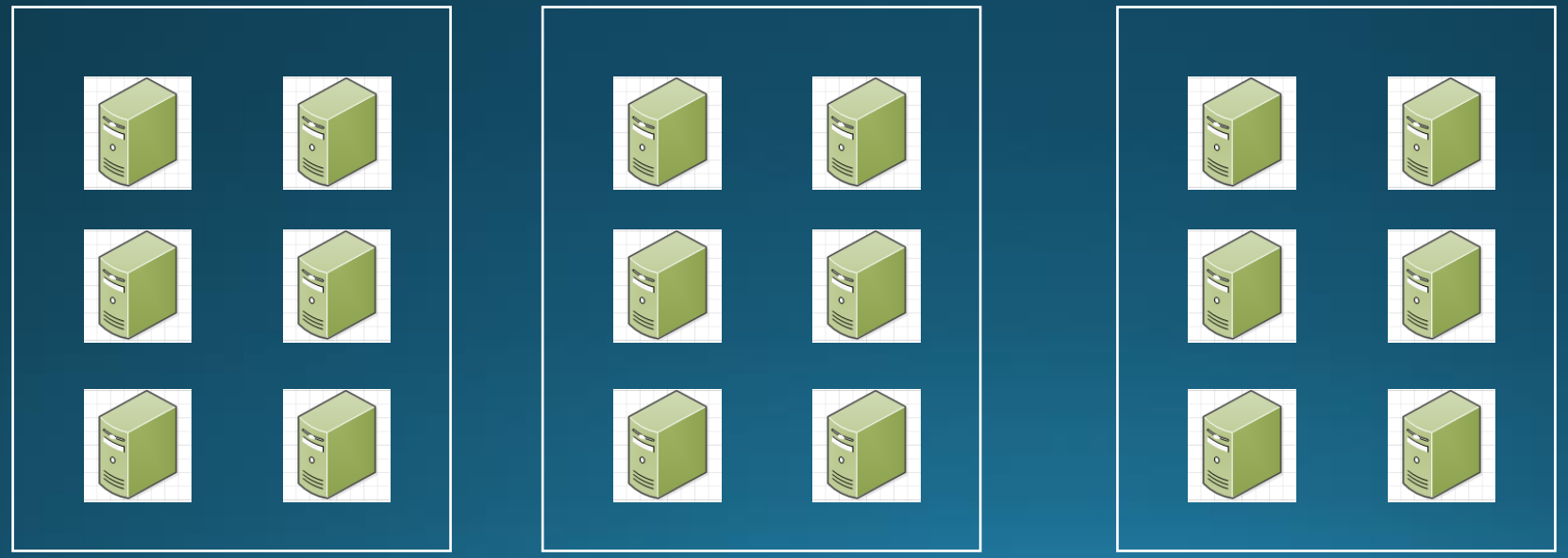
Rack

- Collection of nodes
- All nodes connected by single switch
- Stored close together
- High bandwidth



Cluster

- Collection of racks
- Cluster can consist of a single node
- Rack awareness



HDFS

- Hadoop Distributed File System
  - The data operating system!
  - Manages nodes in the cluster
  - Scalable and highly fault tolerant

Mass Street Analytics

HDFS

- Mechanics
  - Cuts up data into blocks and spreads across nodes
  - Replicates blocks across nodes
  - Process optimization

Mass Street Analytics

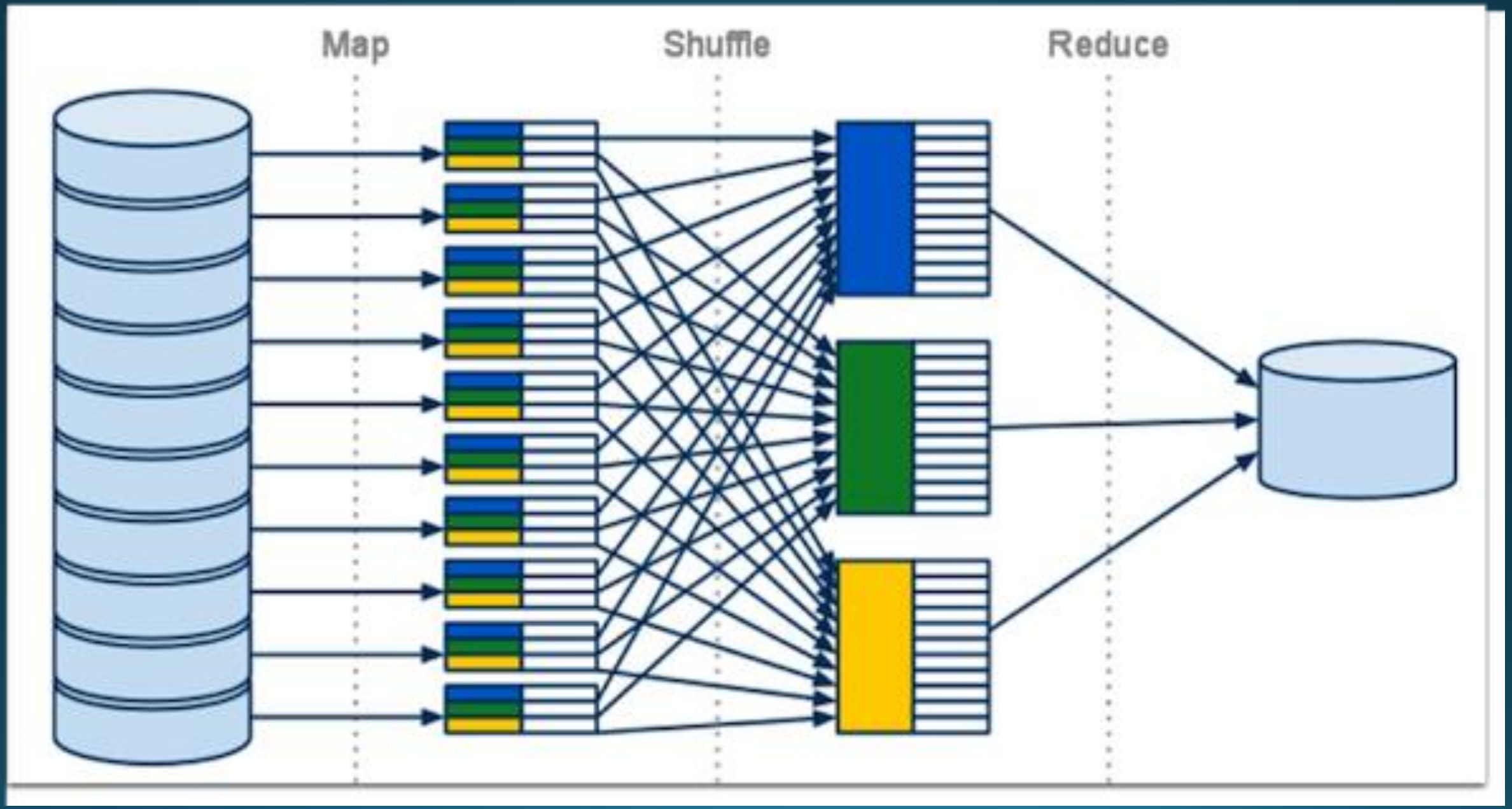
HDFS

- Components
  - Name node
  - Data node
  - A bunch of other nodes

Mass Street Analytics

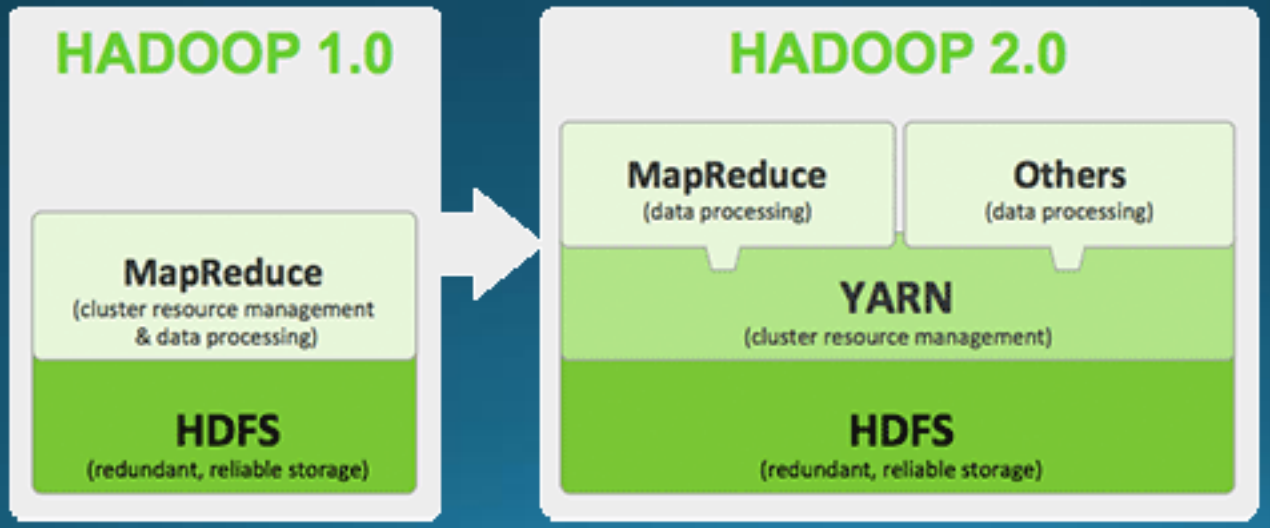
MapReduce

- This is how Google indexes the web
- It's a low level programming framework for pulling data out of the cluster
- Communicates with HDFS
- Designed for batch processing
- Can use any language to write MapReduce jobs
- How does MR work? Pfffffft!


Map
Shuffle
Reduce

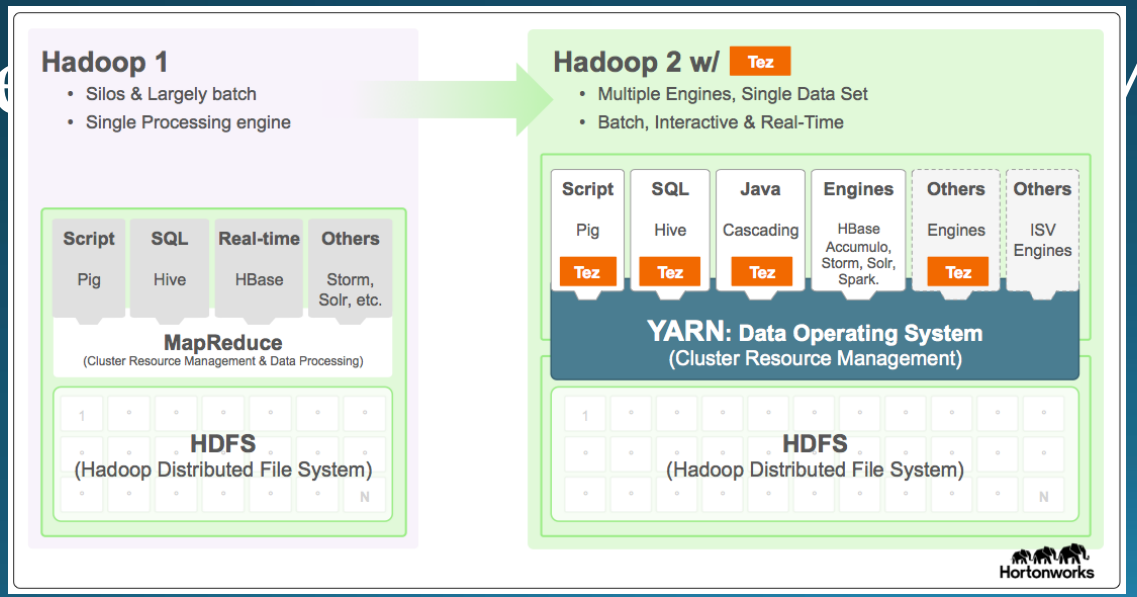
YARN

- Decouples HDFS from MapReduce
- Allows you to run other apps besides MapReduce



TEZ

- Distributed execution framework
- Replaces MapReduce
- Written for other frameworks like Hive and Pig
- Huge pe MapReduce



Hive

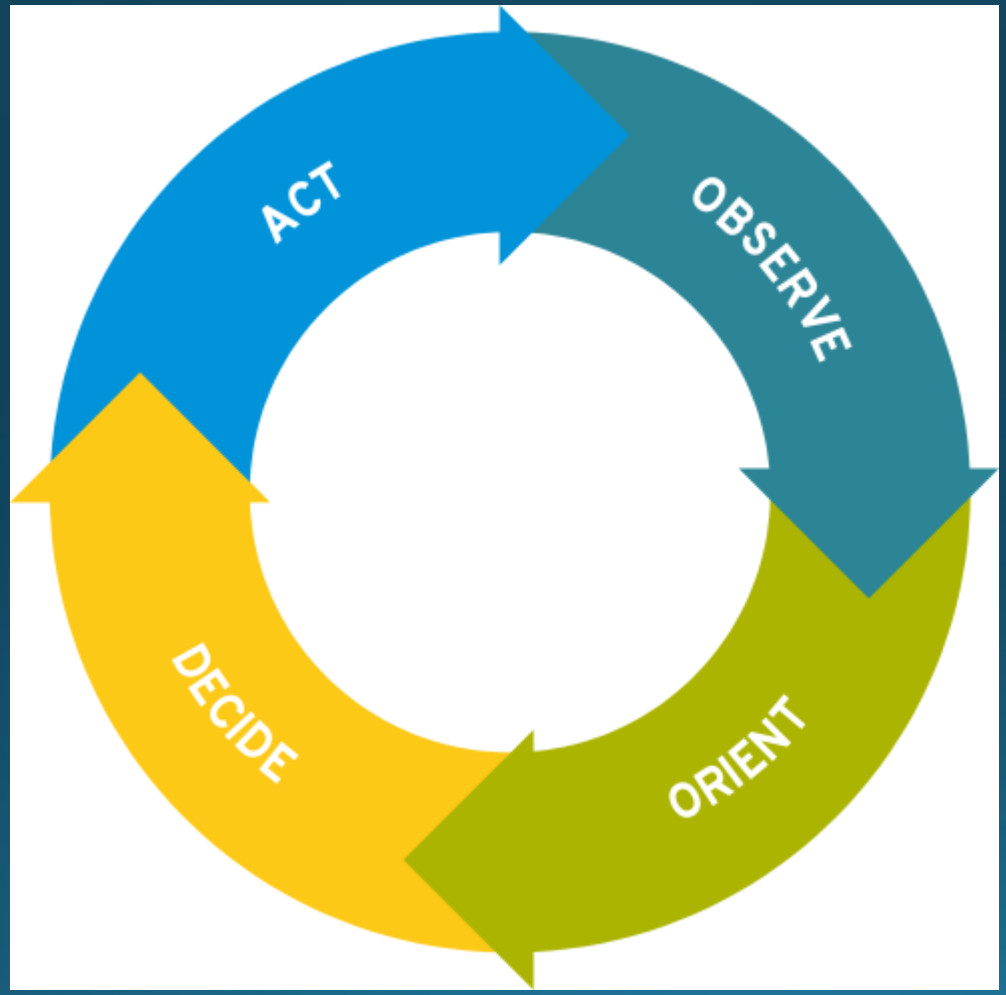
- Hadoop warehouse solution
- SQLesque language called Hive Query Language
- Adds structure to unstructured data
- Provides a window into HDFS

Mass Street Analytics

HBASE / Cassandra

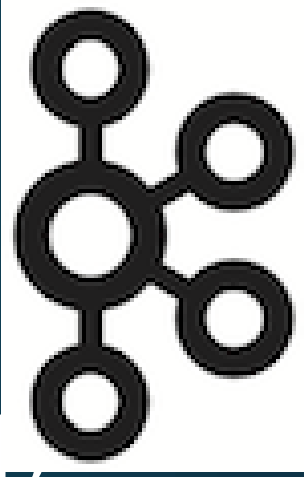
- Both column family NoSQL databases
- There is a difference in how they store data
- Helps solve the append only “problem”.

OODA Loop



Mass Street Analytics

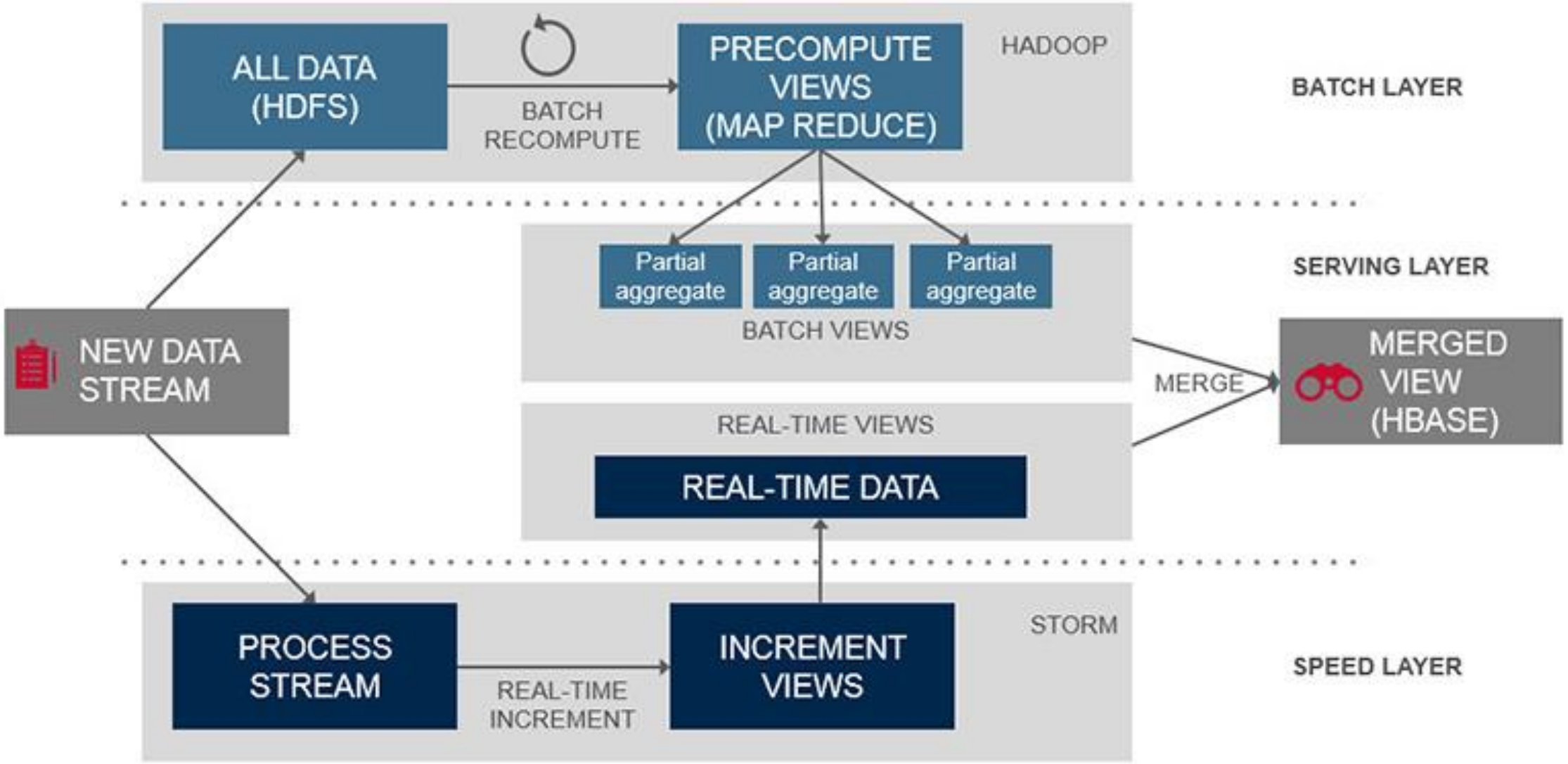
Kafka / Storm / Trident



- Kafka – an open source distributed pub/sub messaging system
- Storm – real time computation framework
- Both are distributed and designed for horizontal scale
- Guarantees at least once processing
- Batch + Real Time = Lambda Architecture

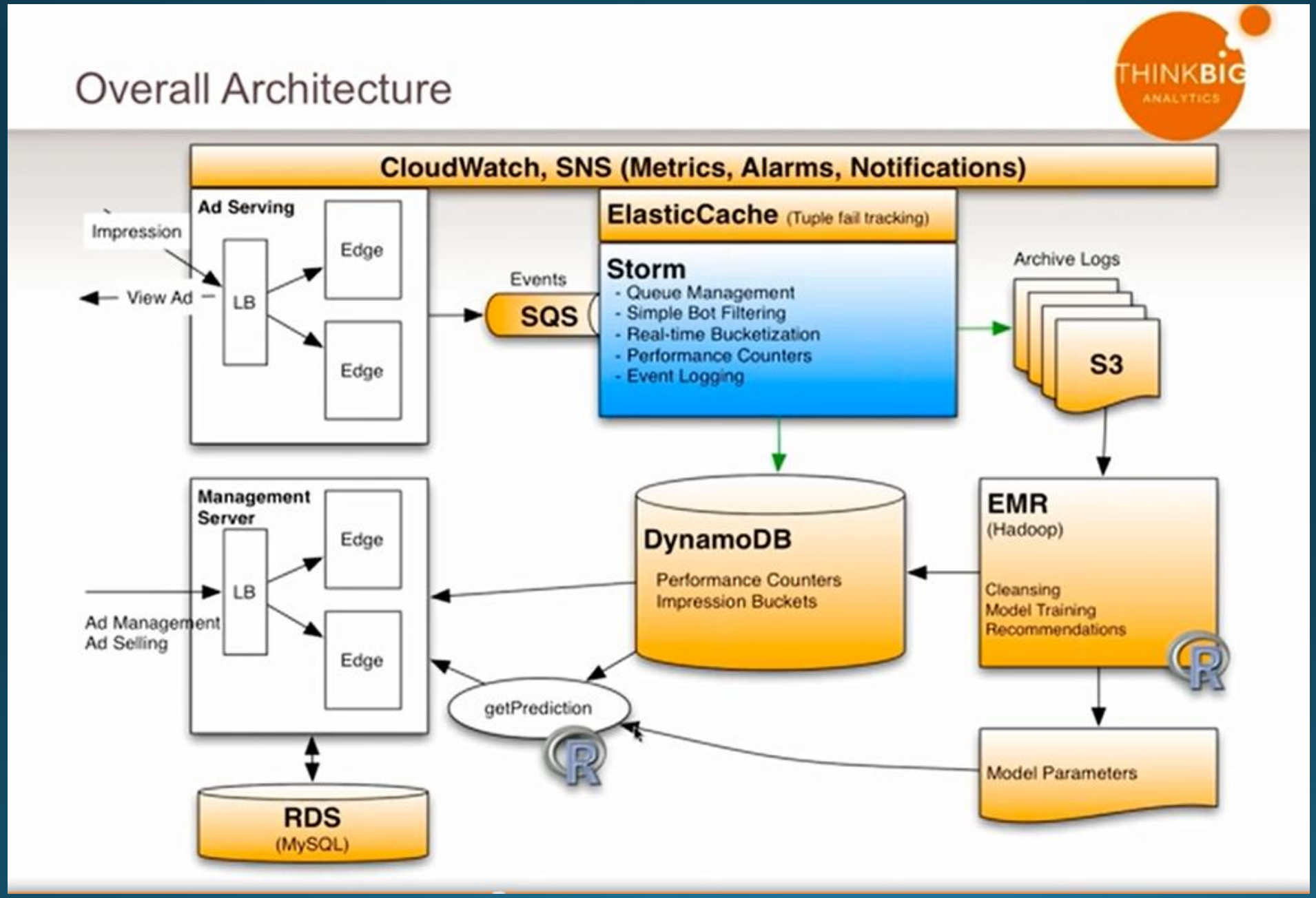
Mass Street Analytics

Lambda Architecture



Slide Credit: MapR

Overall Architecture



Slide Credit: YouTube presentation "Predictive Analytics using Storm, Hadoop, R and AWS"

Honorable Mention

- Sqoop – ETL tool
- Pig – data wrangling tool
- Drill – legit SQL
- Mahout – Java machine learning library
- HCatalog – HDFS abstraction
- SAMOA – real time machine learning

Getting Started

- Hortonworks
  - Sandbox
- YouTube
- Elastic MapReduce
  - Misnomer!!
- Kansas City Data Engineering at Scale Meetup

Possible Future Topics

- Building a real time analytics solution step by step
- Streaming machine learning with SAMOA

Story time/Case study?