Hands On: Introduction to the Hadoop Ecosystem

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

Who is Mass Street?

- Boutique data consultancy
- We work on data problems big or "small"
- Free Big Data training

Mass Street Analytics

Mass Street Partnerships and Capability

- Hortonworks Partner
- Confluent Partner
- ARG Back Office

Mass Street Analytics

Bob's Background

- IT professional 16 years
- Currently working as a Data Engineer
- Education
  - BS Business Admin (MIS) from KState
  - MBA (finance concentration) from KU
  - Coursework in Mathematics at Washburn
  - Graduate certificate Data Science from Rockhurst
- Addicted to everything data

Mass Street Analytics

My Experience With Hadoop

- Trained by MetaScale. (Rip MetaScale)
- Running a seven node lab cluster for three years
- Built because sandboxes are limited and tired of waiting for AWS
- Used for clients and R&D
- Cluster specs:
  - CentOS 7
  - Hortonworks Data Platform 2.6.1
  - 1 Name Node
  - 3 Data Nodes
  - 3 Node Kafka cluster
  - Total HD space: 19TB
  - Total RAM: 196 GB
- Total cost ~ $3,000

Follow Me!

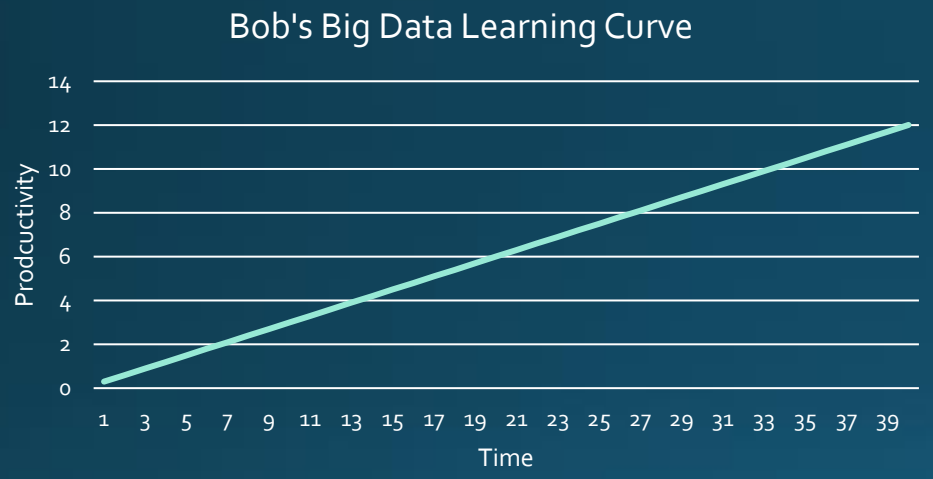
- Personal Twitter: @BobLovesData
- Company Twitter: @MassStreet
- Blog: DataDrivenPerspectives.com
- Website: www.MassStreet.net
- Facebook: @MassStreetAnalyticsLLC

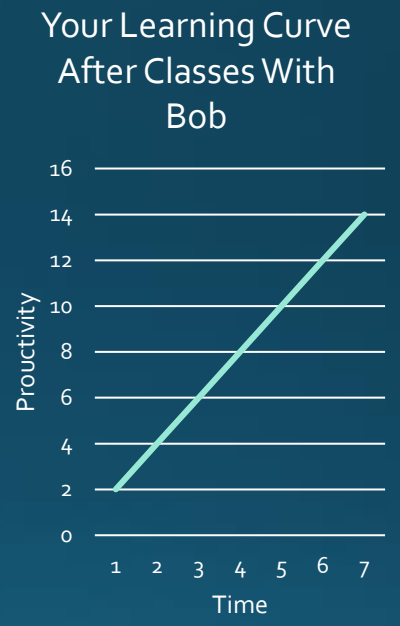
KC Learn Big Data Objectives

- Educate people about what you can do with all the new technology surrounding data.
- Grow the big data career field.
- Teach skills not products

Mass Street Analytics

KC Learn Big Data Objectives





Mass Street Analytics

This Evening's Learning Objectives

- Get familiar with Ambari
- Get familiar with Hive
- Demonstrate BI for Big Data
- If there is time, go over a real world use case.
- Always expect crazy at lab time!

Mass Street Analytics

Logistics

- Modalities of material delivery.
- We need a better meeting place. I'm open to ideas!

Mass Street Analytics

Upcoming MeetUps

September: Architecting your first Hadoop implementation

October: Joint MeetUp with Data Science KC. Up and Running with Spark and R

November: Joint MeetUp with Kansas City Apache Spark MeetUp. Practicing IoT skills with Satori

December: Survey of common data science algorithms.

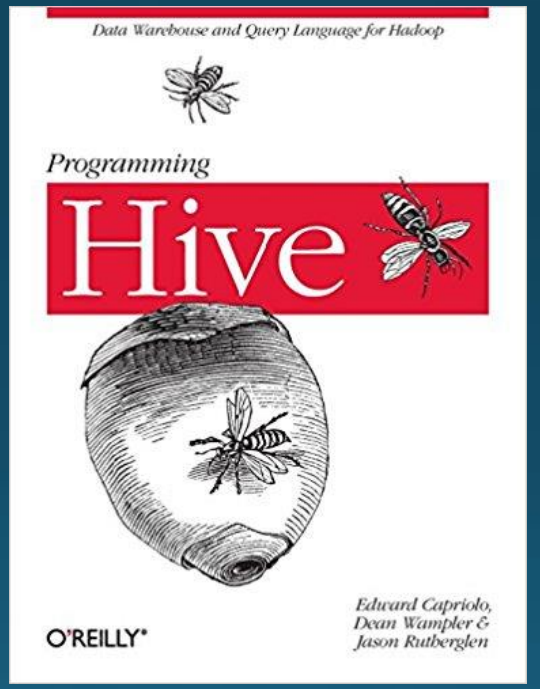
Mass Street Analytics

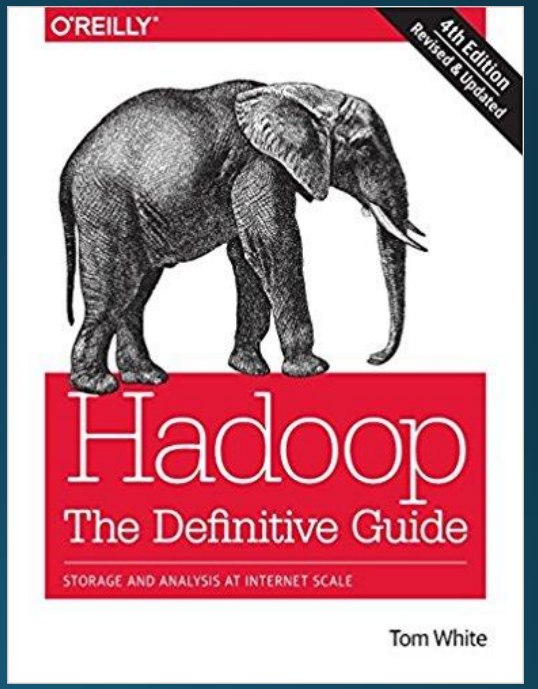
This Evenings Objectives

1. Hadoop ecosystem.
2. MapReduce Concepts.
3. Basic understanding of Data Science Principals.

Resources For This Evening's Material: Books

- Programming Hive: Data Warehouse and Query Language for Hadoop
- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale





Mass Street Analytics

Resources For This Evening's Material: Online Classes

- Udemy course: Comprehensive Course Hadoop Analytic Tool: Apache Hive
- No good Hadoop admin classes online
- Links to resources can be found in the Resources folder of the GitHub repo for this class.
- https://github.com/MassStreetAnalytics/Hands-On-Hadoop

I want to do big data work. What do I need to know?

1. Basic understanding of Hadoop ecosystem.
2. Basic understanding of MapReduce Concepts.
3. Basic understanding of Data Science Principals.

Pick A Path

1. Data Science
2. Data Engineering
3. Data Interpreter?

Basic Skills For Each Role

Data Scientist

1. Full understanding of DS/ML concepts
2. R or Python

Data Engineer

1. Java or Scala or Python or R
2. Build tools like Maven, Ant, or Gradle
3. Source Control preferably GitHub
4. Full understanding of MapReduce
5. Full understanding of Hadoop Ecosystem
6. Full understanding of NoSQL
7. Full understanding of MPP DBs

Mass Street Analytics

DATA Engineer

DataCamp
Learn Data Science By Doing

DATA Scientist

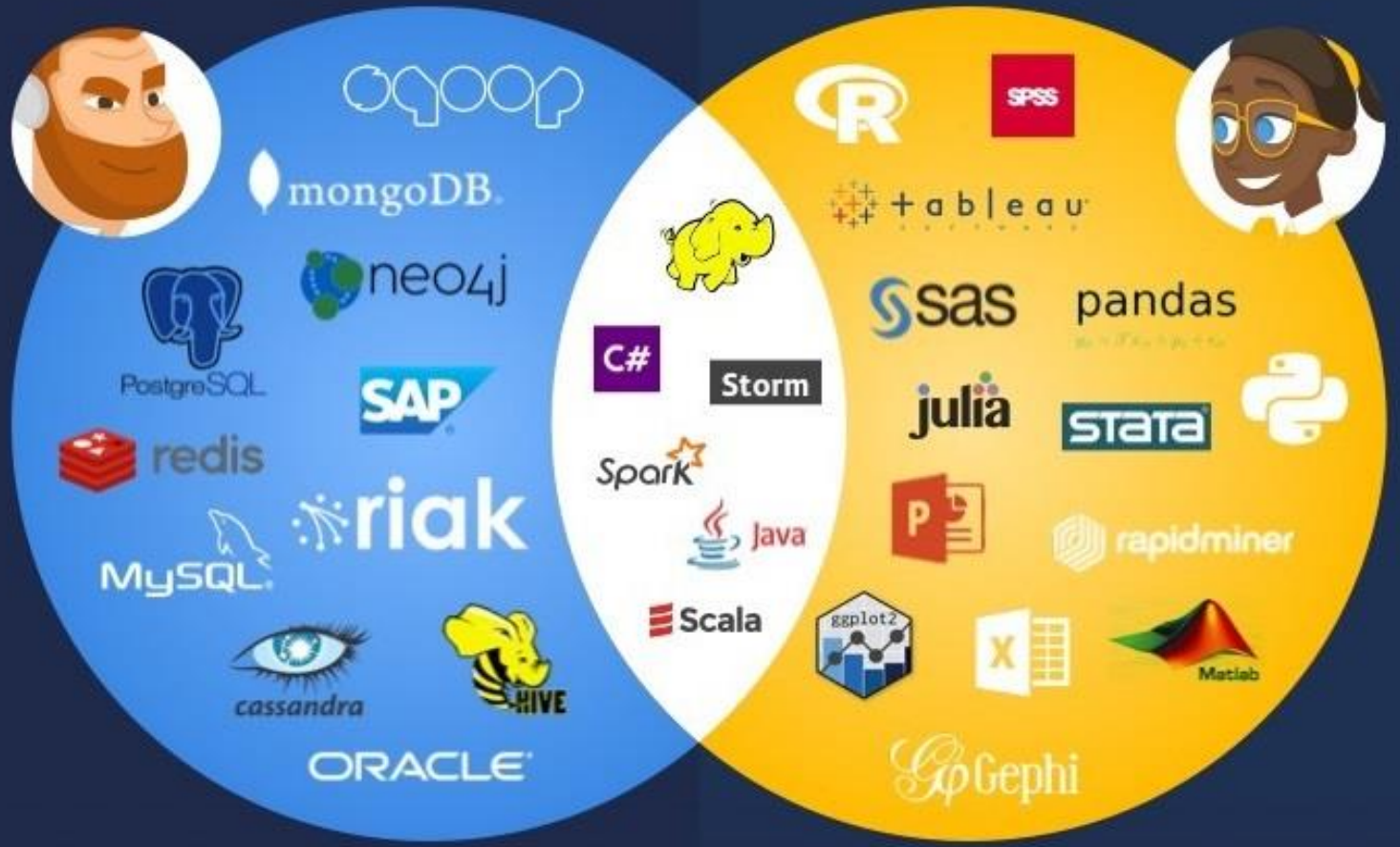
Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

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.

Languages, Tools & Software




Salaries
BY PAYSA.COM
$124,000
MEDIUM MARKET
MIN $34k
MAX $341k
$135,000
MEDIUM MARKET
MIN $43k
MAX $364k

Responsibilities...

ON DAILY BASIS

Develop, construct, test, and maintain architectures (such as databases and large-scale processing systems)

Ensure architecture will support the requirements of the business

Discover opportunities for **data acquisition**

Develop data set processes for data modeling, mining and production

Employ a variety of languages and tools (e.g. scripting languages) to **marry systems together**

Recommend ways to **improve data** reliability, efficiency and quality

Conduct research to **answer industry and business questions**

Leverage large volumes of data from internal and external sources to answer that business

Employ sophisticated analytics programs, machine learning and statistical methods to **prepare data for use in predictive and prescriptive modeling**

Explore and examine data to **find hidden patterns**

Automate work through the use of predictive and prescriptive analytics

Tell stories to key stakeholders based on their analysis

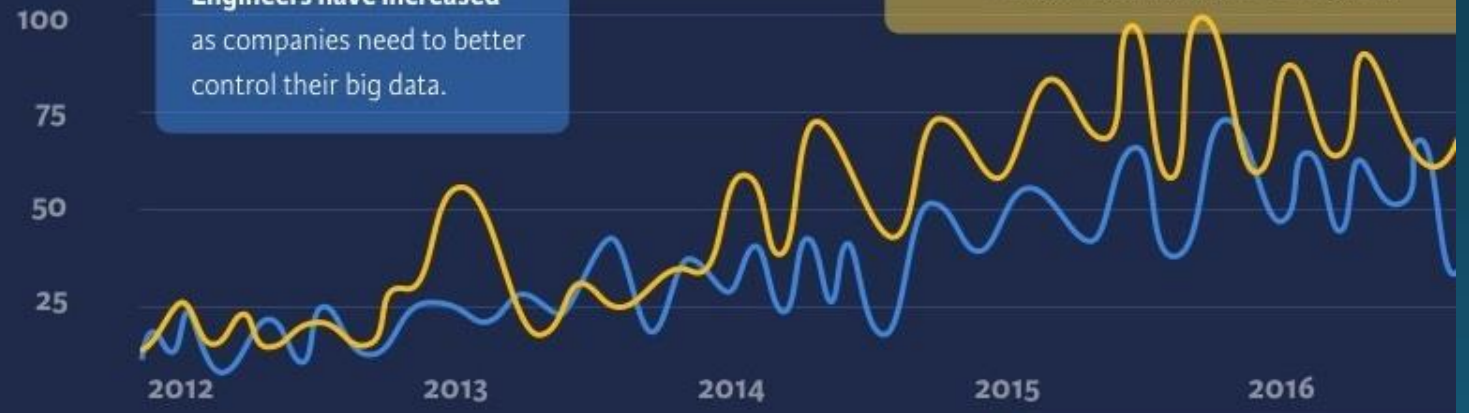
Job outlook in the USA

BASED ON GOOGLE TRENDS DATA

> In 2018 the US could face a **shortage of 140,000 to 190,000 people with deep analytic skills and 1.5 million** managers and analysts with the know-how to use the analysis of big data to make effective decisions.
>
> *McKinsey, 2011*

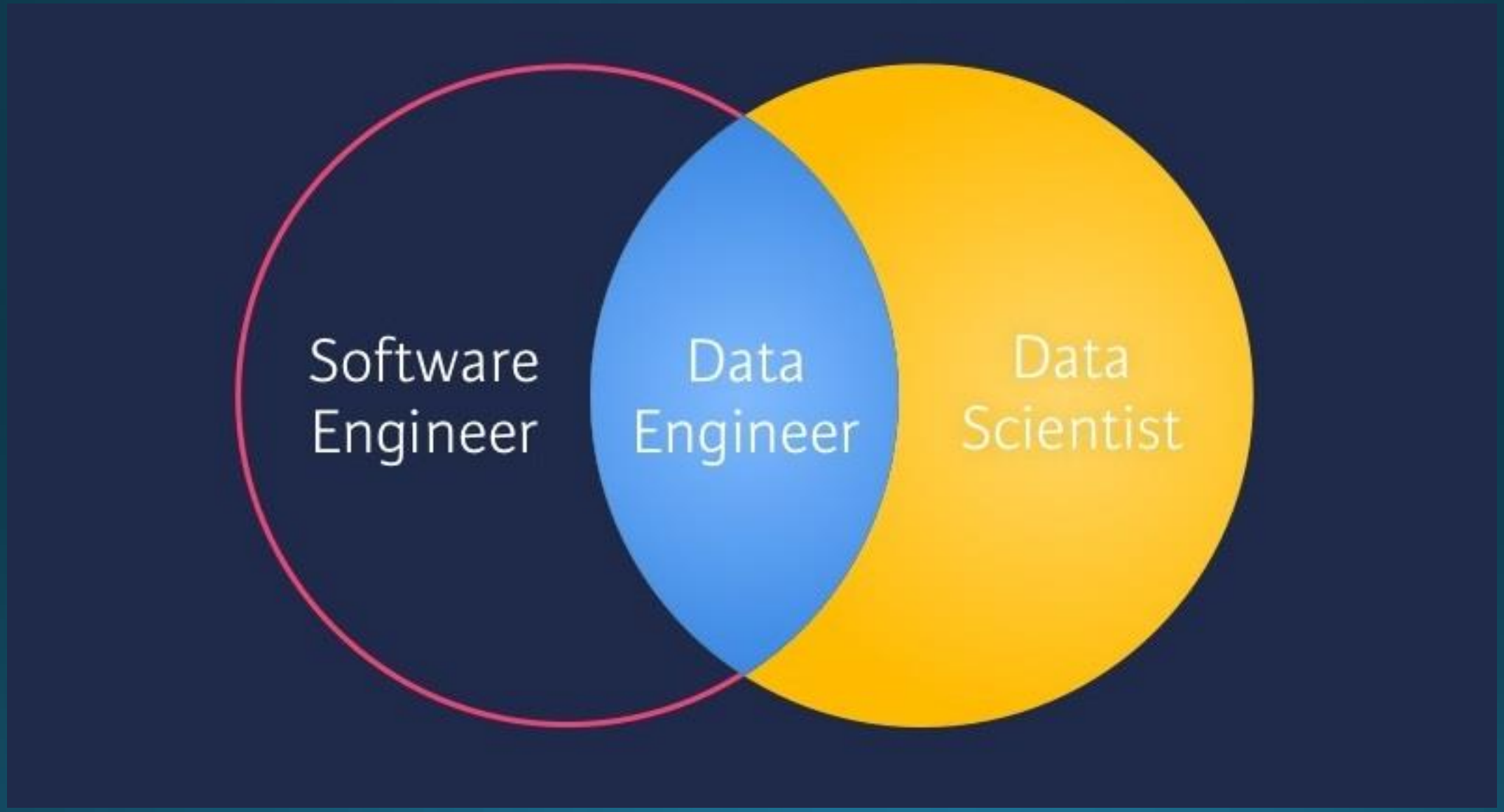
Beginning in 2014, job postings **looking for Data Engineers have increased** as companies need to better control their big data.

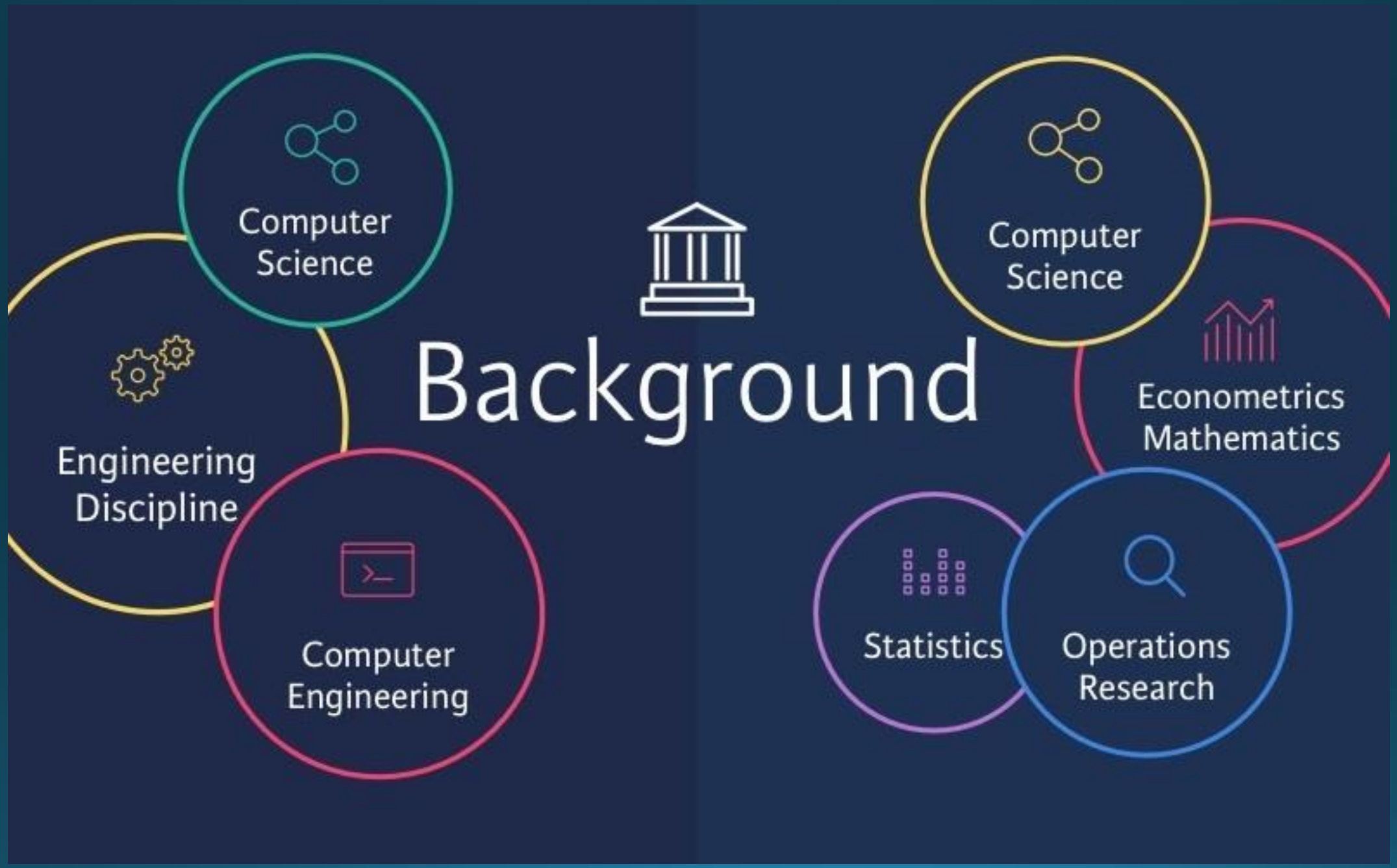
Data science is **beginning to merge with more traditional roles** (i.e. business analysts and intelligence analysts).



Companies need data pipeline builders as they move to the Cloud and as they develop "data lakes".

While the "bubble" of data scientist hiring may burst, **the demand for big experts will remain high.**


Software Engineer
Data Engineer
Data Scientist


Computer Science
Engineering Discipline
Computer Engineering
Background
Computer Science
Econometrics Mathematics
Statistics
Operations Research

Courses to get started

INTRODUCTION TO DATA ENGINEERING BY UNIVERSITY OF WASHINGTON

IMPORTING DATA IN PYTHON BY DATACAMP

DATA SCIENCE WORKSHOPS BY GALVANIZE

INSIGHT DATA ENGINEERING FELLOWS PROGRAM

IMPORTING DATA IN R BY DATACAMP

DATA SCIENCE AND ENGINEERING WITH APACHE® SPARK™ ON EDX

COURSERA BIG DATA SPECIALIZATION

EXPLORATORY DATA ANALYSIS BY DATACAMP

MACHINE LEARNING ENGINEER NANODEGREE ON UDACITY

INTRODUCTION TO PYTHON FOR DATA SCIENCE ON EDX

MACHINE LEARNING TOOLBOX BY DATACAMP

INTRODUCTION TO R FOR DATA SCIENCE ON EDX

MACHINE LEARNING ON COURSERA

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

Common Hadoop Use Cases

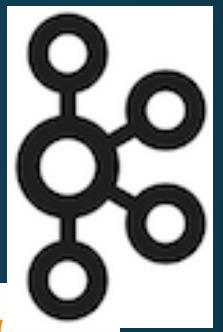
- Easier analysis of big data
- Lower TCO of data
- Offloading ETL workloads
- Data warehousing
- Data lakes
- Backbone of various advanced data systems


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


APACHE FALCON


Spark


oozie




hadoop


TEZ


AVRO
TM


HIVE


APACHE
HBASE


Apache
Solr


mahout


chukwa


Cassandra

What these things do

- Sqoop – ETL tool
- Pig – data flow language
- Drill/HAWQ – SQL on Hadoop
- Knox/Ranger – Security
- Zookeeper – Coordination
- Spark – In memory computing
- Kafka – Distributed Pub/Sub messaging

It's like a child's erector set!

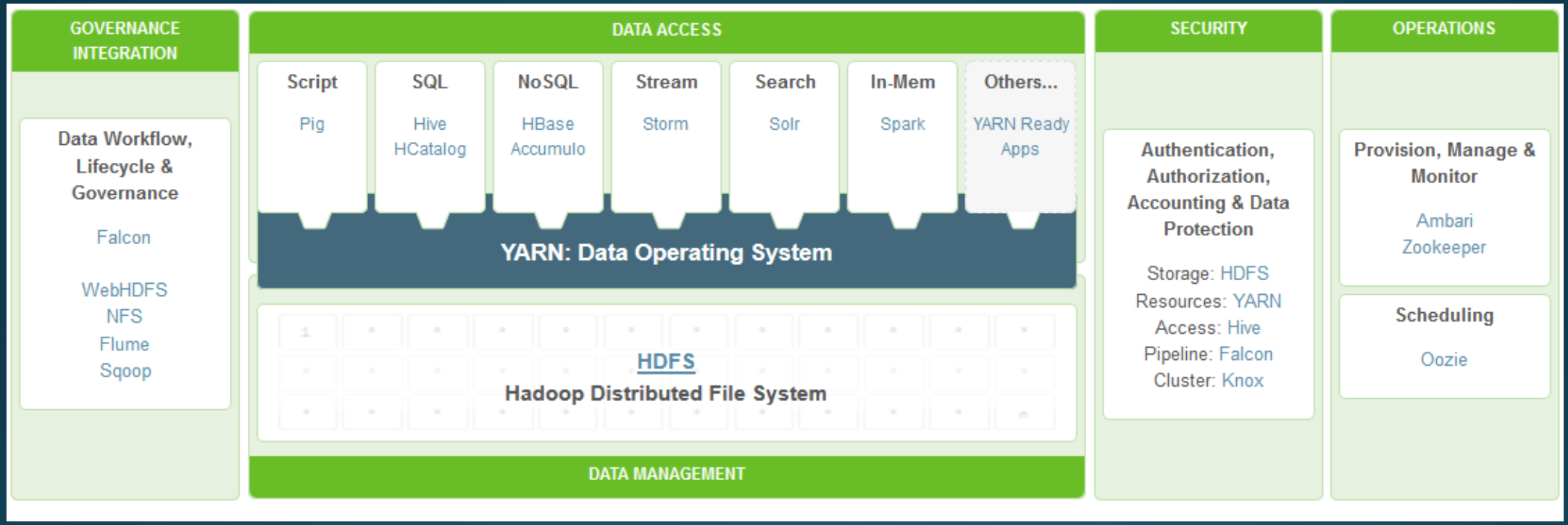


Photo Credit: Hortonworks website

HDFS

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

Mass Street Analytics

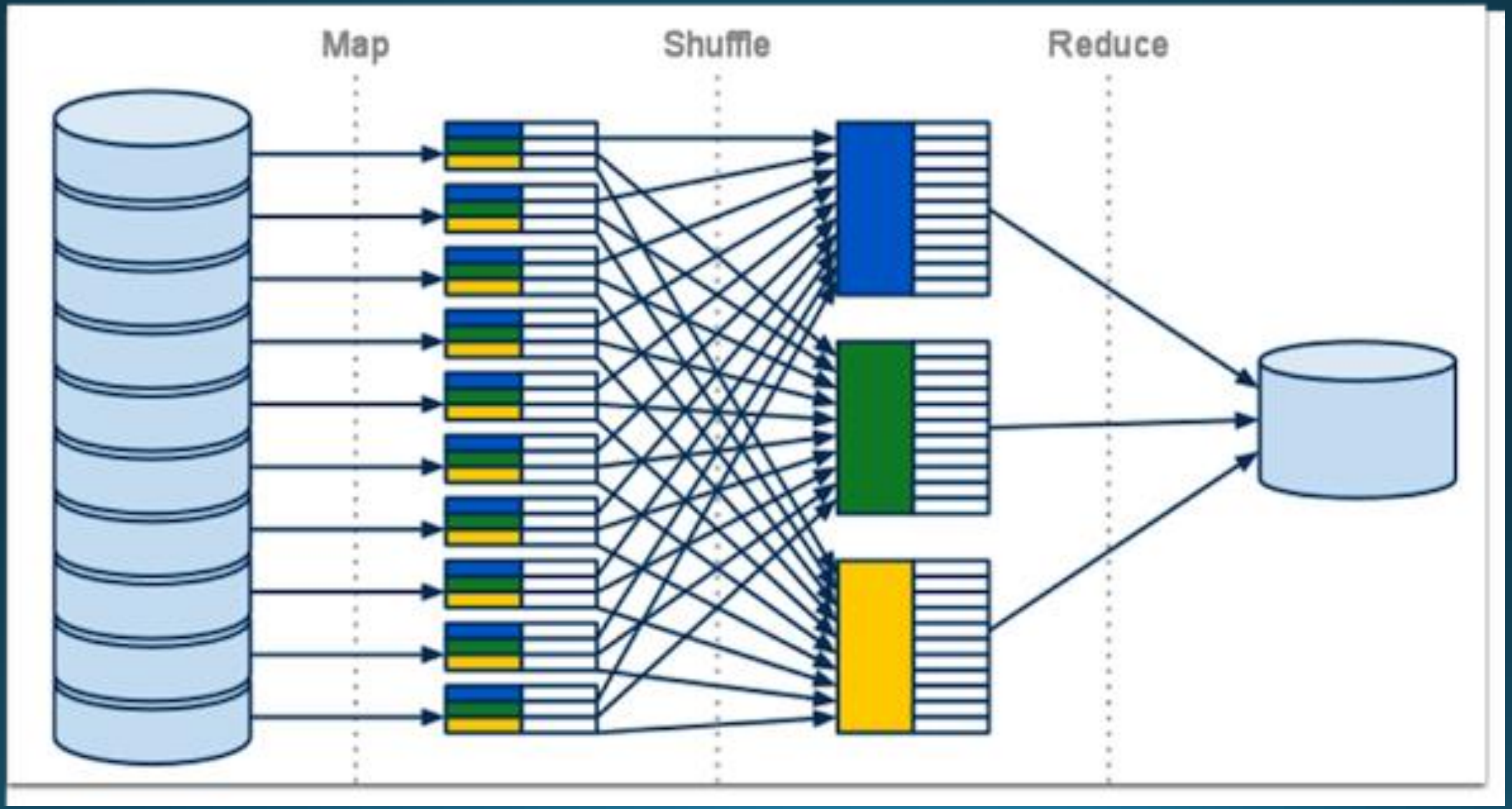
HDFS

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

Mass Street Analytics

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

Hive

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

Mass Street Analytics

Hive

- Writes MapReduce behind the scenes (sorta)
- Should be used for OLAP task
- Queries aren't fast, just faster than normal
- You can connect through ODBC/JDBC
  - Which means you can work with Hive using standard BI tools

Tez

- Tez generalizes the MapReduce paradigm to a more powerful framework for executing a complex DAG (directed acyclic graph) of tasks for near real-time big data processing.

Bob's Super Crashy Crash Course In SQL

- SQL – Structured Query Language
- SQL is how you retrieve data from relational databases
- Each statement consist of a minimum of three elements
  - SELECT
  - FROM
  - WHERE

Bob's Super Crashy Crash Course In SQL

- In the old world, data is stored in two dimensional tables; rows and columns
- There are rules about how this information is stored.
- Data is usually modeled as a set of relationships
- These relationships can be thought of in terms of parent child.

Bob's Super Crashy Crash Course In SQL

- There are all kinds of relationships.
  - Many to one. Simplest and most common.
  - One to one.
  - Many to many.

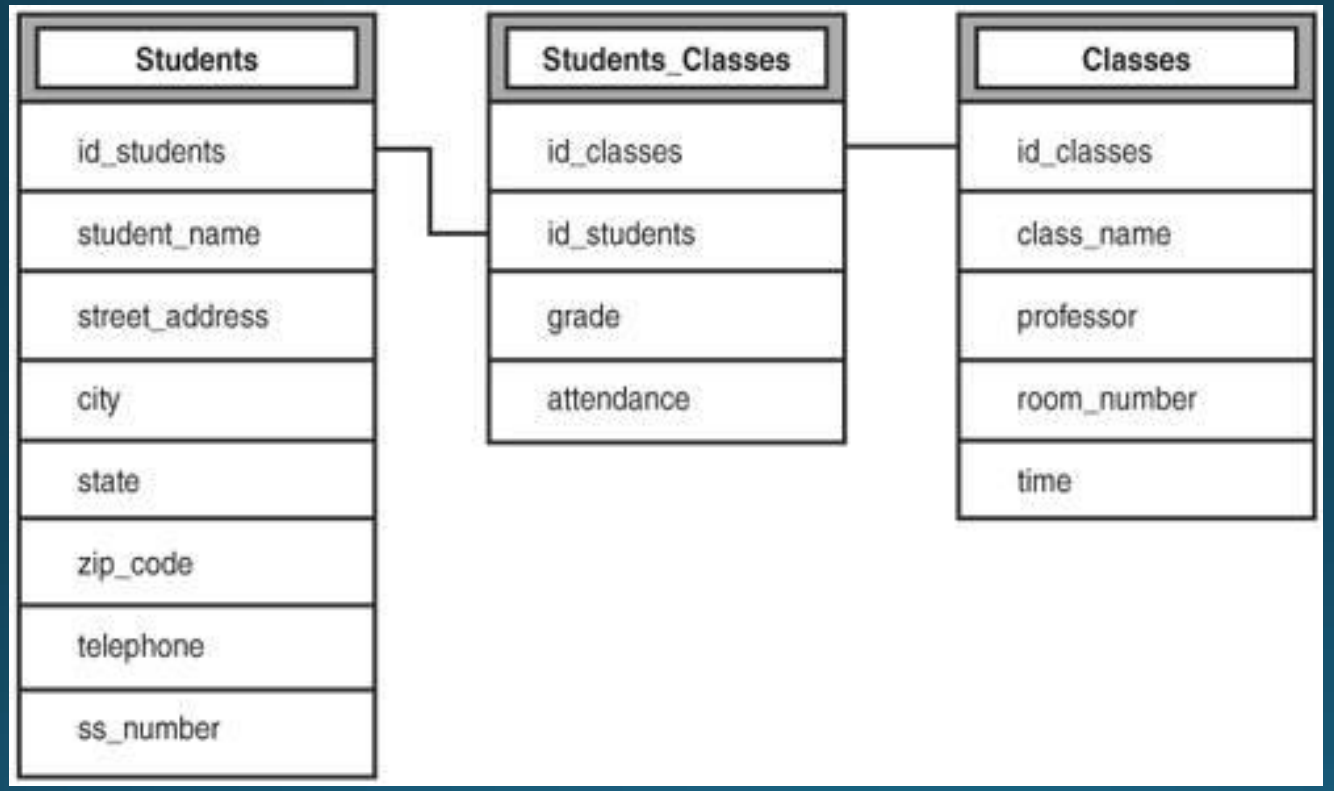
Bob's Super Crashy Crash Course In SQL

- Normally, each record is identified by some value called a key.
- The are all kinds of keys but the most important are:
  - Primary Keys
  - Foreign Keys

Bob's Super Crashy Crash Course In SQL

- Primary Keys uniquely identify a record
- Foreign Keys are the primary key of a parent record.

Bob's Super Crashy Crash Course In SQL



Mass Street Analytics

A brief history of SQL on Hadoop

- Data on HDFS is not stored in a relational style.
- NoSQL
- Data retrieval very difficult
- Attempts to make HDFS talk SQL
  - HAWQ
  - Drill
- Opinion: Efforts are floundering because of the rise of MPP databases

Required Resources

- Azure Cloud Account
- Putty
- Files from GitHub: https://github.com/MassStreetAnalytics/Hands-On-Hadoop

Tonight's Scenario

You work in a large organization with a lot of data but it is stored in old world relational databases. You need to analyze this data but there are over a billion records and your queries are incredibly slow. You convince your boss to let you try Hadoop. You spin up a cluster and get to work.

Analyst Task

1. Get your data into HDFS.
2. Get your data into Hive.
3. Run some initial queries on the data.
4. Share your analysis.
5. (Make sure you destroy your cloud instance!)