Business of Big Data

About Smoking Hand

Training and consulting company focused on Big Data, NoSQL and The Cloud

- Founded in 2008

We help companies with their Big Data transformations

- Use case evaluation
- Management training
- Technical training
- Architecture reviews
- Live and email programming support

About Me

Current: Instructor, Thought Leader, Monkey Tamer

Previously:

- Curriculum Developer and Instructor @ Cloudera
- Senior Software Engineer @ Intuit

Covered, Conferences and Published In:

- GigaOM, ArsTecnica, Pragmatic Programmers, Strata, OSCON, Wall Street Journal, CNN, BBC, NPR

See Me On:

- @jessetanderson
- http://tiny.smokinghand.com/linkedin
- http://tiny.smokinghand.com/youtube

About You

Your experience as a developer, analyst or administrator

Which language(s) you use

Experience with Hadoop, Big Data or NoSQL

Expectations from this class

About Workshops

Questions and discussion encouraged

Focus on technology as a solution to a problem

- Not a focus on the implementation

Slides are a starting point for discussion

Themes and projects are the basis for your ideas

Class Logistics

Start at 9 AM

- Please show up on time

Full-day class

- 15 minute breaks at ~10 AM and ~2 PM
- Hour long lunch at 12 PM

Downloads

- Slides (recommended)

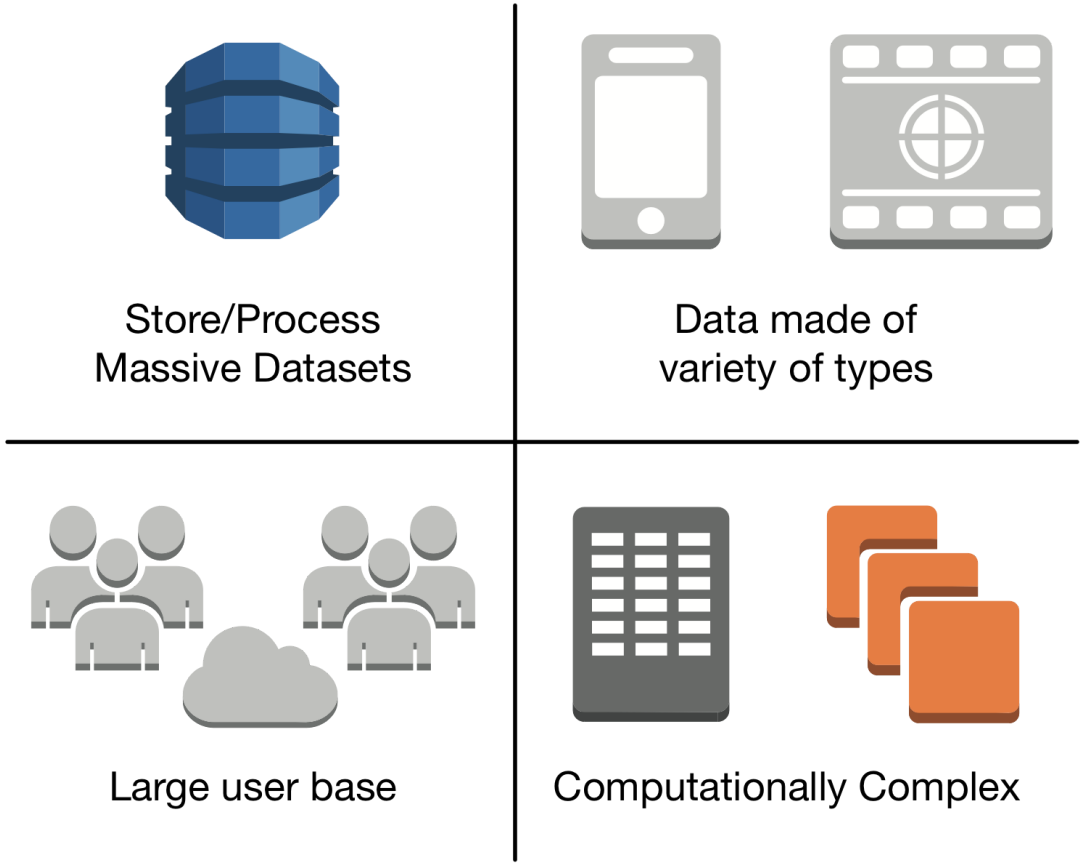
Course Chapters

- **Thinking in Big Data**
- Engineering Big Data Solutions
- Doing Data Science on the NFL Play by Play Dataset
- Conclusion
- Q & A

Thinking in Big Data

Introducing Big Data

What is Big Data?



Massive Storage

Google stores 2 EB

LinkedIn stores 176 TB per day in Kafka

- 800 billion events per day

Facebook's largest cluster stores 100 PB

Large Hadron Collider stores 25 PB per year

Processing Vast Amounts Of Data

500 million tweets are sent per day

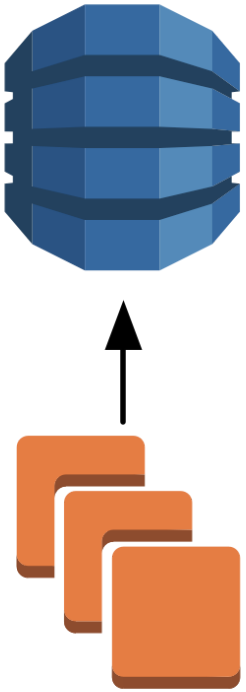
4.5 billion Facebook likes per day

Google processes 40,000 searches per second

Amazon.com in the US sells 232 million different products

LinkedIn processes 1.1 trillion messages per day

FINRA processes up to 75 billion events per day or about 6 TB



Wide Variety of Data

Two types of data

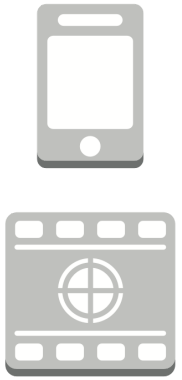
- Structured and unstructured

Structured data fits nicely into spreadsheets and RDBMS

Unstructured data is variable

- Human generated
- Audio or video files

80% of data is unstructured



Huge Userbase

Facebook has 1.35 billion monthly active users

LinkedIn has 380 million registered members in 200 countries

Amazon.com has 244 million active customers

Twitter has 284 million active users

Serving The Thoughts of the World

Humans cause great fluctuations in Big Data

618,725 tweets per minute were posted during the World Cup Final

Tweets and likes are integrated into most web pages

How to scale to variable demand efficiently?

Existing Systems Have Difficulty

Our existing systems cannot scale to store or process the data

Can a SAN scale to 100 PB?

How can you run sentiment analysis on 500 million tweets?

Can an RDBMS scale to 1 PB?

Distributed systems split up tasks on multiple computers

- Enable previously impossible things by splitting them up
- Often used with Big Data

Increases complexity and failure scenarios

Decreases overall costs by removing monolithic architecture

Removes single points of failure

What is Hadoop?

GFS Paper

Google needed to handle Big Data from the beginning

It needed to store and search the entire internet

Had to be done cost effectively

Google published a paper in 2003 on how they stored data

- Titled: The Google File System

Google MapReduce Paper

Google needed to process every web page on the entire internet

Had to be done in a time-efficient way

Google published a paper in 2004 on how they processed data

- Titled: MapReduce: Simplified Data Processing on Large Clusters

Apache Hadoop

Hadoop is a distributed system built for Big Data

Handles both the storage and processing of Big Data

Originally based on the Google papers

- Slowly deviated and improved

Basic Components

Hadoop is made up of two components

The component to store data

- Called HDFS

The component to process data

- Called MapReduce

HDFS Component

HDFS is a distributed file system

Automatically handles loss of data

Highly available

Can scale to hundreds of petabytes

Uses commodity hardware

- Usually off-the-shelf SATA drives

MapReduce Component

MapReduce is a distributed processing engine

Elegantly handles distributed coordination and errors

Can scale to process hundreds of petabytes

Uses commodity hardware

Hadoop Distributions

Hadoop distributions help standardize installs

- Gives a full stack software version that code is written and tested on

Similar to Linux distributions

- Standardizes support
- Removes compilation and associated issues

Several companies offer their own Hadoop distributions

- Cloudera, Hortonworks, MapR

The Ecosystem

Apache Hive

Apache Hive is a Data Warehousing infrastructure built on top of Hadoop

Uses a SQL-like language called HiveQL

- HiveQL is translated to a MapReduce job

Originally created by Facebook to process their data

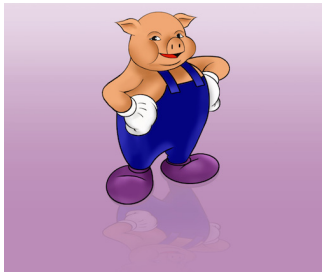
- Many people had SQL skills but no Java skills

Apache Pig

Apache Pig is a platform for analyzing large data sets

Uses a high-level data flow language called PigLatin

- PigLatin is translated to a MapReduce job by the Pig compiler



Originally created by Yahoo

- Many people had data flow language skills but no Java skills

Apache Crunch and Cascading

Many programmers want an API that works differently from the MapReduce API

Crunch gives programmers a simple Java API

- Allows creation of pipelines for processing

Cascading is closer to a functional programming API

- Allows separation of business logic and integration logic

Apache Oozie

Apache Oozie is a workflow scheduler system to manage Apache Hadoop jobs

Supports the automation of the majority of the Hadoop ecosystem

Uses XML to define jobs and settings

- Uses JSTL for control and logic

Hue

Hue is web-based way to access Hadoop

Can create, interact and run commands from many Hadoop ecosystem projects

- Run MapReduce jobs and view results
- View and interact with files in HDFS
- Schedule workflows to run with drag and drop interface

Apache Solr and Search

Clear text searching at scale is a common problem

Apache Solr and Apache Lucene are technologies to perform text searches

Solr exposes a REST interface for queries and creating advanced queries

Cloudera Search integrates Solr, Lucene and Hadoop

Apache HBase

Apache HBase is a column family oriented NoSQL database

Based on the Google BigTable paper

Uses HDFS for storage

Data can be retrieved quickly or batch processed with MapReduce

Apache Spark

Spark is a processing engine for Big Data

It offers a richer API for processing

Spark was originally created at AMPLab at UC Berkeley

An Apache top level project

Rapidly evolving

Processes in real-time using micro batches

Being pushed as the going forward technology for batch

Other Next Generation Processing Engines

Apache Storm works on streams and micro-batches

- Twitter has created a Storm successor called Heron

Apache NiFi works on streams, windows, and full batch mode

Kafka Streams works on streams and windows

This landscape is changing significantly

- The winner(s) are not clear
- Placing a bet may require rewrites

Real-time Queries

Many companies have extensive SQL skills

- But lack the Java and MapReduce skills

Impala and Presto allow data querying with SQL

- The queries can still operate over large amounts of data

Commonly used with BI and Analysis tools

Queries can be returned in milliseconds to minutes

Apache Kafka

Apache Kafka is a distributed publish subscribe system

It uses a commit log to track changes

Kafka was originally created at LinkedIn

- Open sourced in 2011
- Graduated to a top-level Apache project in 2012

Many Big Data projects are open source implementations of closed source products

- Unlike Hadoop, HBase or Cassandra, Kafka actually isn't a clone of an existing closed source product

Monitoring Hadoop Clusters

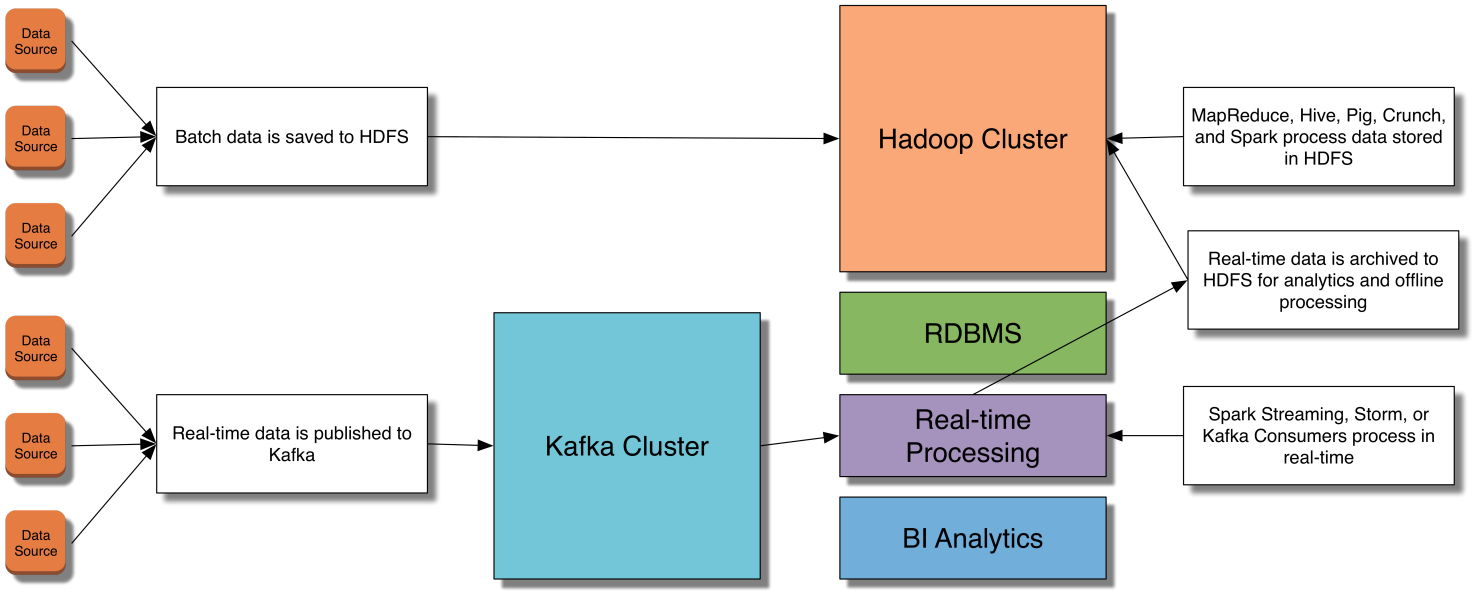
Hadoop clusters are often made up of hundreds of nodes

- Monitoring and configuring become a consistent problem

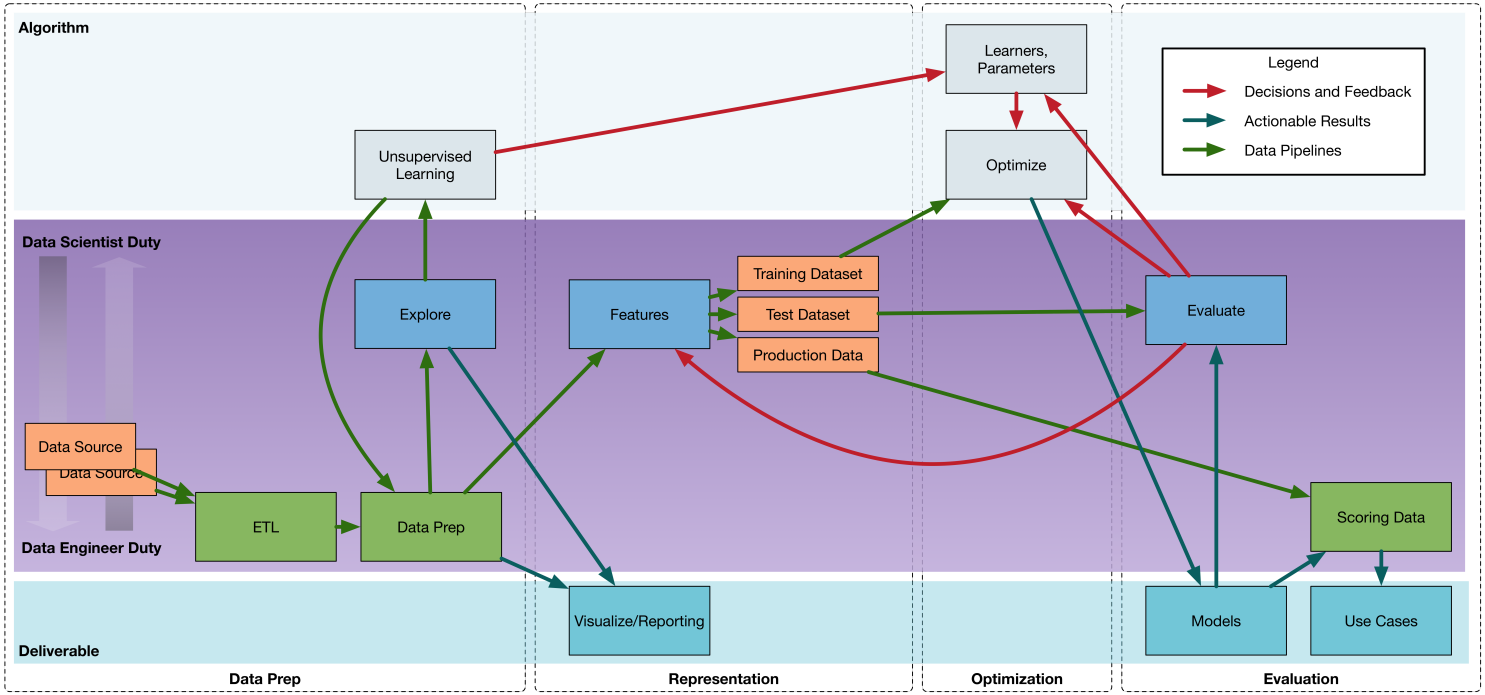
Cloudera Manager and Apache Ambari are purpose-built for Hadoop clusters

- Make cluster installation easier
- Monitor service status and issues
- GUI-based service configuration

General Architecture Diagram



Data Products



Introduction to HDFS

What Is HDFS?

HDFS is Hadoop's file system

- Stands for Hadoop Distributed File System

Purpose-built to store Big Data efficiently

Entire file system is accessible from any client

Not a general purpose file system

- Files are immutable

DEMO: HDFS With Legos

We will now demonstrate how HDFS works with Legos

Concepts shown:

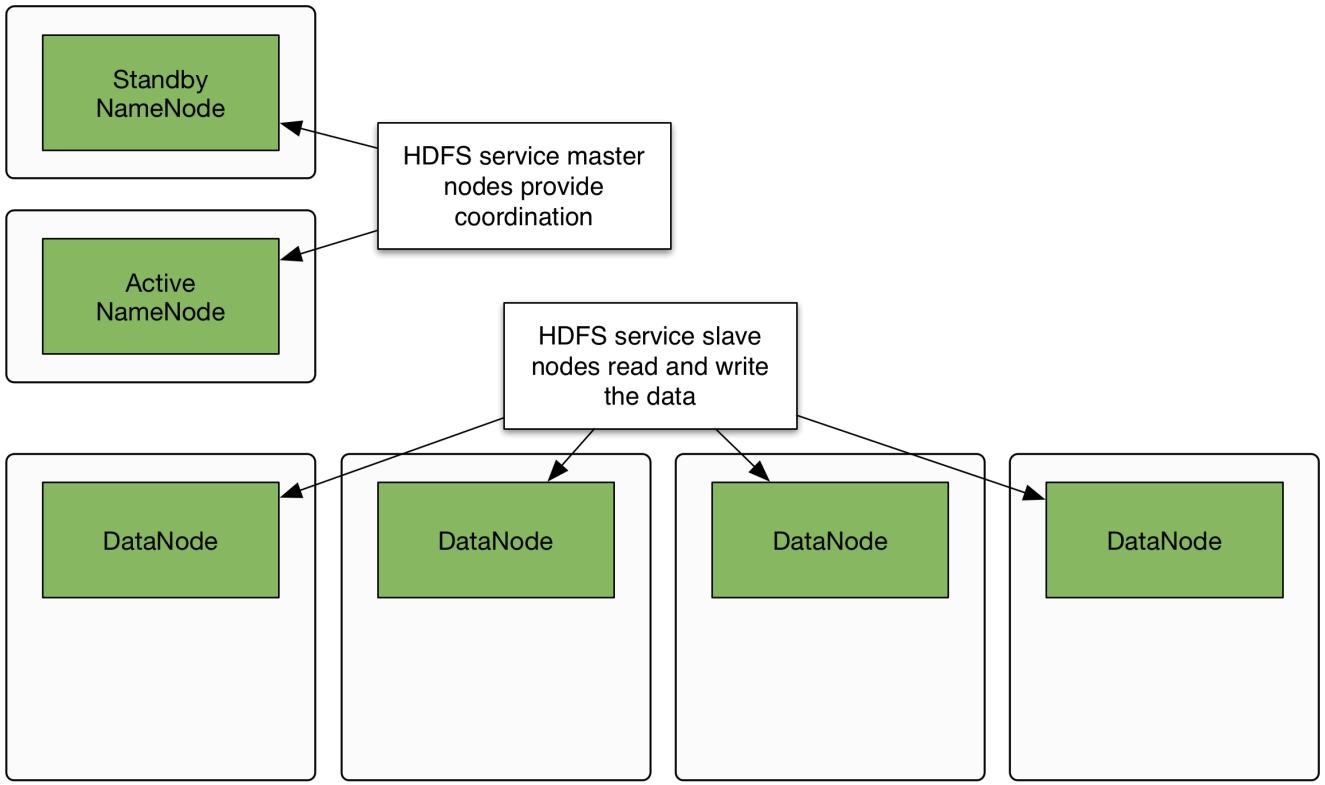
- HDFS Blocks
- Block replicas and rack awareness
- Write Path
- Read Path
- Failure handling

Lines:

1. I want to write a block where should I write it?
2. I want to read a file where can I find it?

Video Link

HDFS Daemons



Using HDFS

HDFS is a userspace file system

- Cannot use built-in Linux commands
- Have to use Hadoop's own commands or the API to access

HDFS is not a general purpose file system

- Shouldn't store many small files
- Files cannot be overwritten and are immutable

The local filesystem exist apart from each other

- Files will need to be moved from local filesystem to HDFS

Help and Ls Commands

Get a listing of the possible HDFS commands

```
$ hadoop fs
```

Do a file listing in HDFS of the user's home directory

```
$ hadoop fs -ls
```

`hadoop fs` commands have no current working directory and must be fully qualified

```
$ hadoop fs -ls /user/vmuser/data
```

Put/Get Commands

To copy a file from your local filesystem to HDFS

```
$ hadoop fs -put data.txt
```

```
$ hadoop fs -put /home/vmuser/data.txt /user/vmuser/data.txt
```

To copy a file from HDFS to your local filesystem

```
$ hadoop fs -get output.txt
```

```
$ hadoop fs -get /user/vmuser/output.txt /home/vmuser/output.txt
```

Mkdir/Rm commands

To create a directory

```
$ hadoop fs -mkdir newdirectory
```

To remove a file

```
$ hadoop fs -rm /user/vmuser/file.txt
```

To recursively remove a directory

```
$ hadoop fs -rm -r /user/vmuser/olddirectory
```

Hadoop has a trash concept, but it is not turned on by default

Other Commands

Copy files

```
$ hadoop fs -cp /user/vmuser/file.txt /user/vmuser/file2.txt
```

Move files

```
$ hadoop fs -mv /user/vmuser/file.txt /user/vmuser/file_moved.txt
```

Cat files (output to screen)

```
$ hadoop fs -cat /user/vmuser/file.txt
```

Tail files (output end of file to screen)

```
$ hadoop fs -tail /user/vmuser/file.txt
```

Admin Commands

HDFS is a POSIX-style file system with permissions

Add read permissions for all users to a directory

```
$ hadoop fs -chmod a+r output
```

Change the owner of a directory

```
$ hadoop fs -chown vmuser:vmuser output
```

Get the disk usage and free space for HDFS

```
$ hadoop fs -df
```

Get the disk usage for a directory

```
$ hadoop fs -du output
```

Introduction to MapReduce

What is MapReduce?

MapReduce is Hadoop's distributed processing engine

Makes use of data stored in HDFS

Automatically handles many error conditions of processing

Provides a simple API to create MapReduce jobs

DEMO: MapReduce With Playing Cards

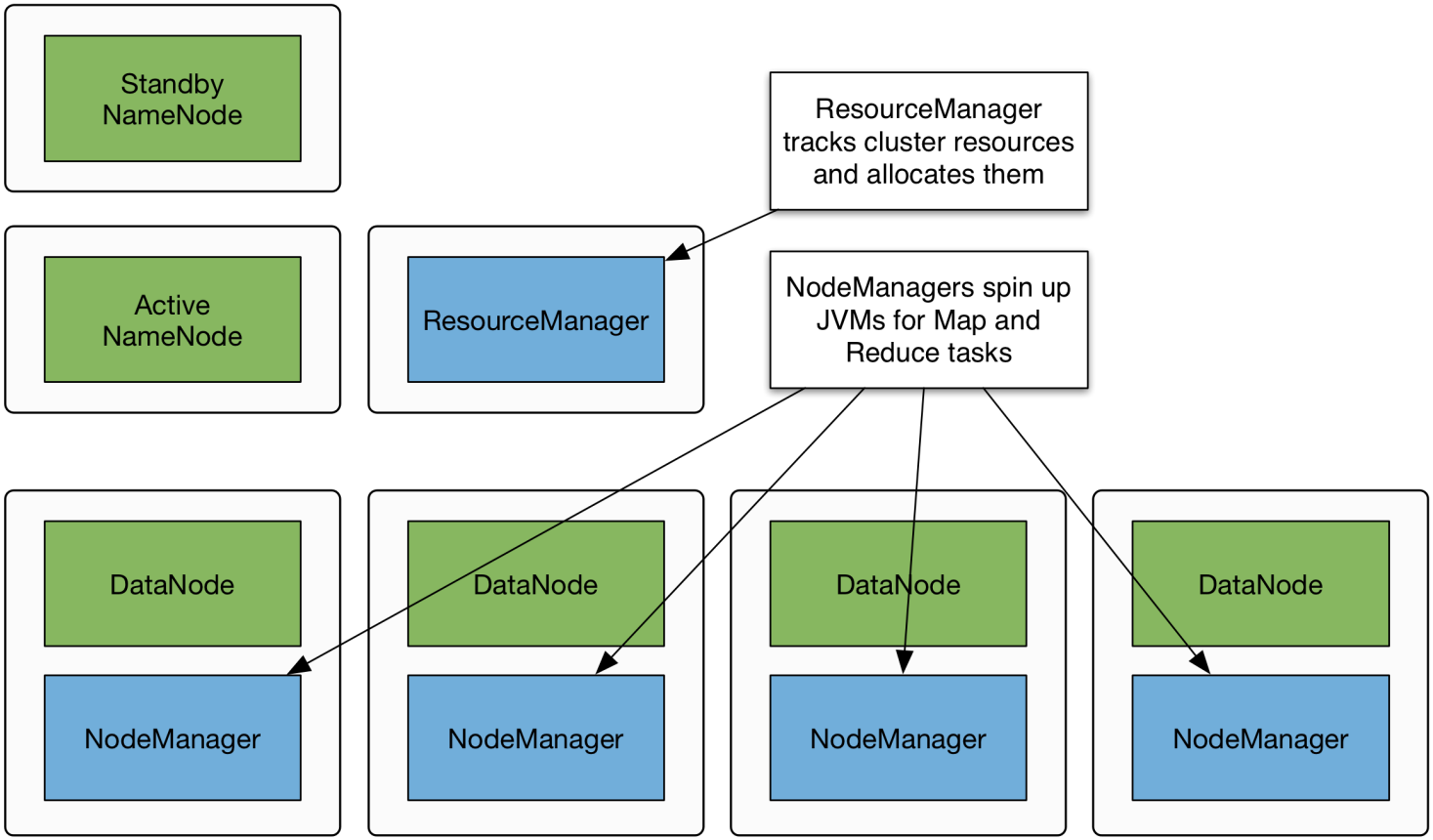
We will now demonstrate how MapReduce works with playing cards

Concepts shown:

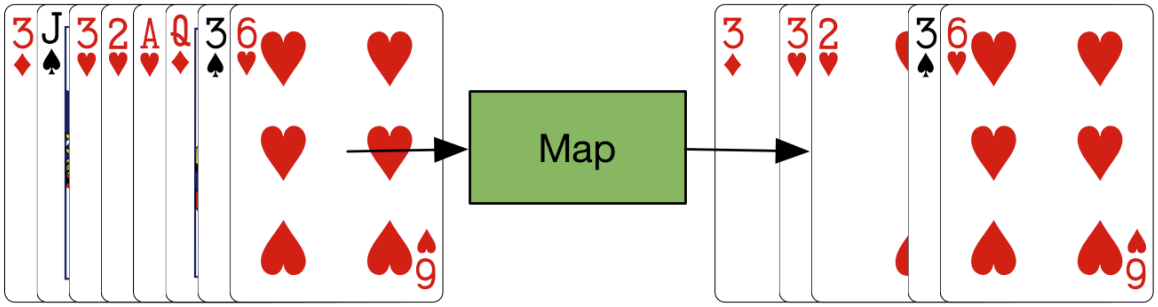
- Map and reduce phases
- Shuffle sort
- Shared nothing architecture
- Data locality
- Failure handling

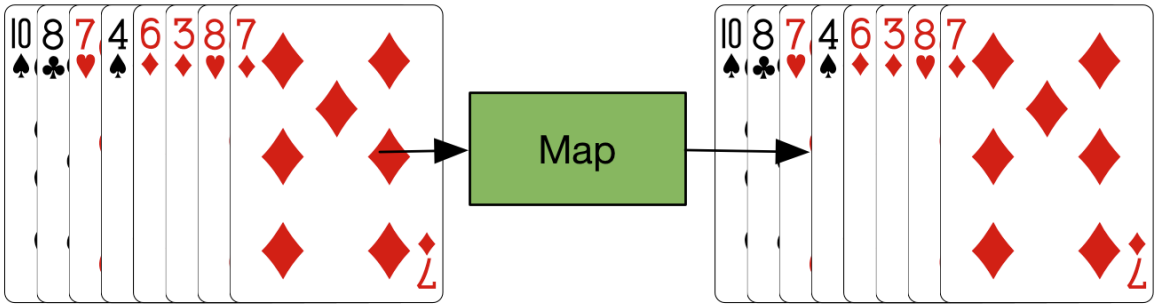
Video Link

YARN Daemons

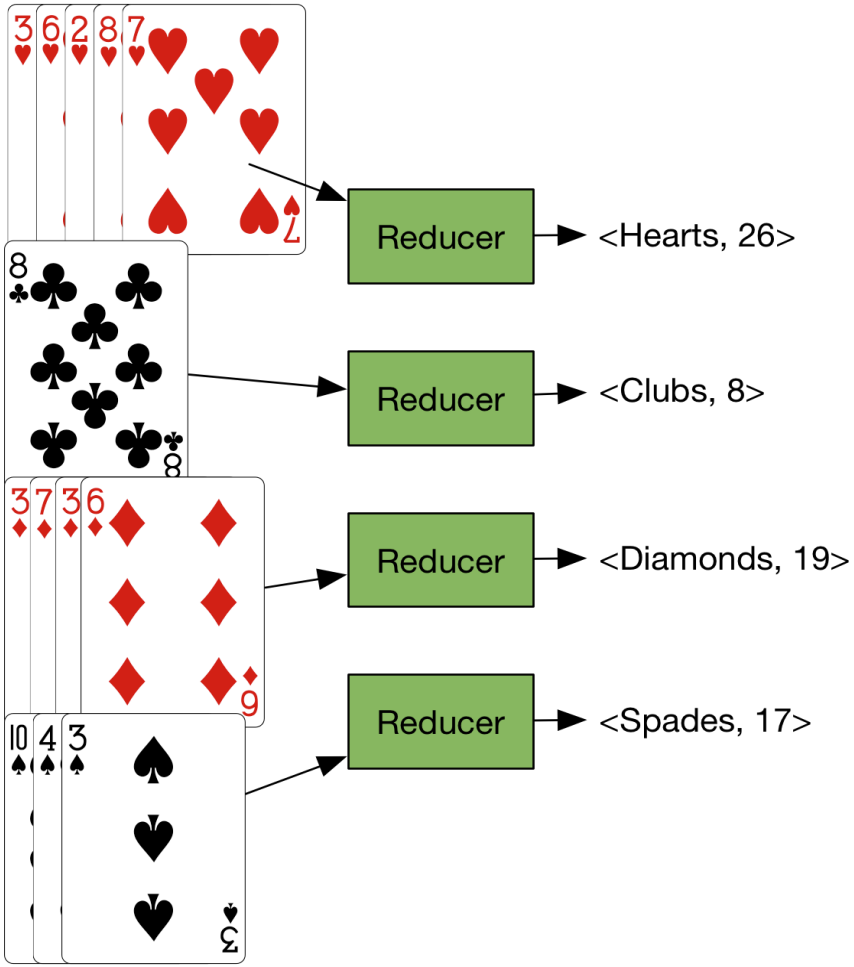


Map Phase

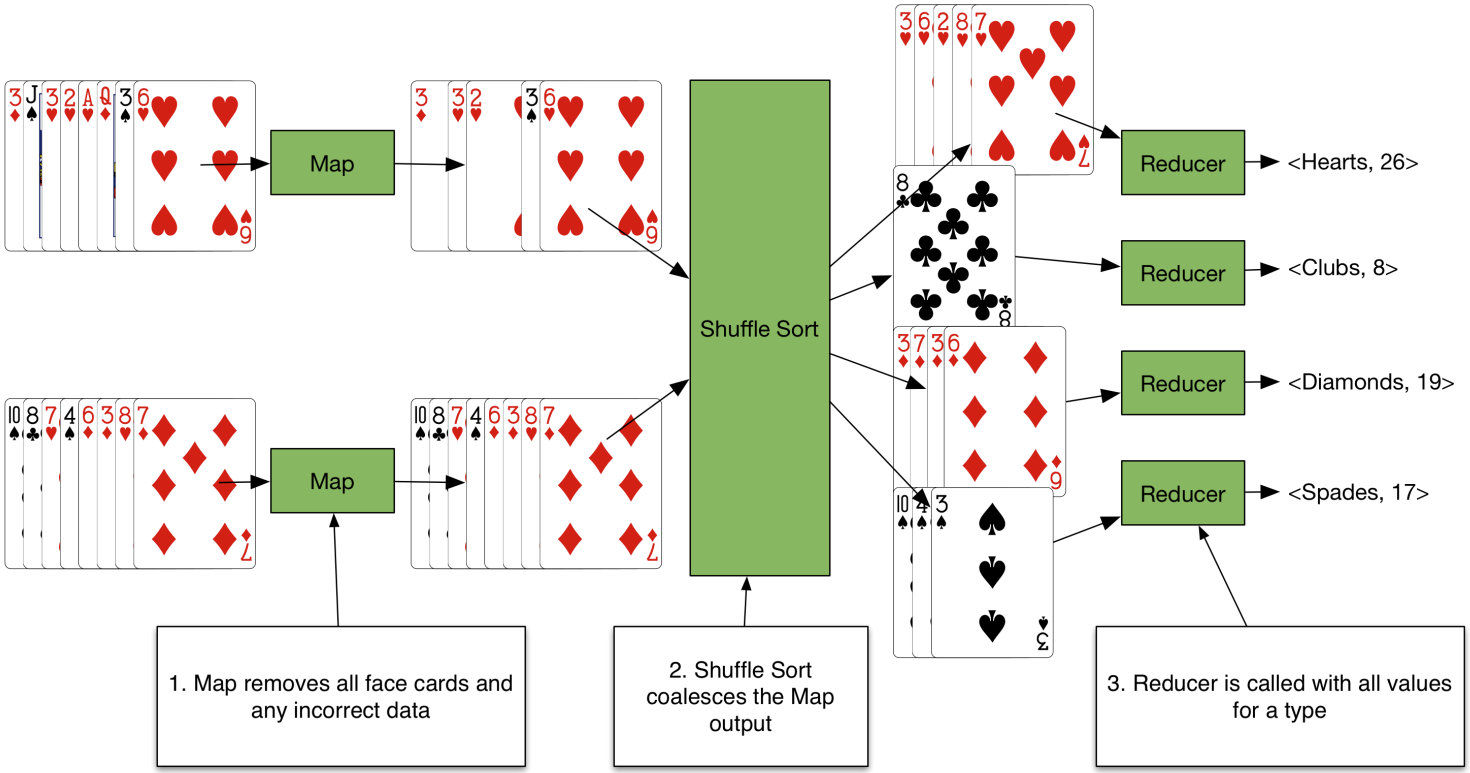




Reduce Phase



Full Flow



EXERCISE: Teams and Technologies

In this exercise, you will evaluate your use case and team with Big Data technologies.

Time: 60 minutes

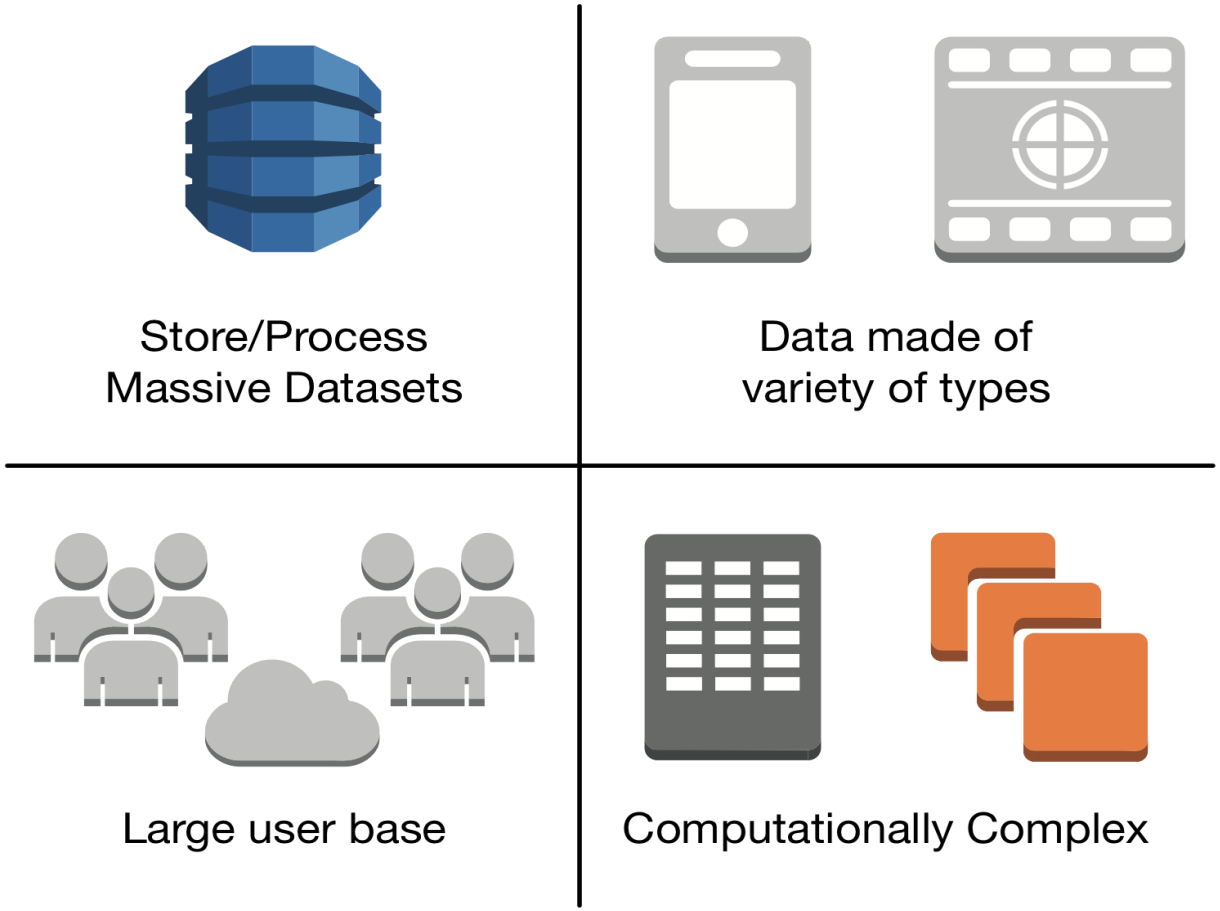
Course Chapters

- Thinking in Big Data
- **Engineering Big Data Solutions**
- Doing Data Science on the NFL Play by Play Dataset
- Conclusion
- Q & A

Engineering Big Data Solutions

Where Do We Start?

What Is Big Data?



Is It Big Data?

If it isn't Big Data look elsewhere

- Will it become Big Data?

Reasons for checking if it is Big Data

- There are other, better tools
- Less costly engineers

If it is Big Data you're in the right place

Identify Your Goals

What are you trying to accomplish?

Are others in the industry using Hadoop the same way?

Crawl, Walk, Run

- Crawl = Easily attainable goal
- Walk = Building on the attainable goal
- Run = Gaining maximum value from Data

What Is The Business Value?

What difference will it make in the company?

What value is there in your data?

How are you going to augment your decisions with data?

How can this data benefit your customers?

Technology Solutions

Which solution(s) to use?

- Hadoop and its ecosystem
- NoSQL solutions
- Other Big Data solutions

What do you need to do?

- Process data
- Storing data

How fast do you need it?

- Real-time
- Batch

Creating The Solution

Developers will need new skills and tools

QA will need to learn the new system

Project Managers will need to know what's possible

There will be a skills gap

- Should you train or hire?
- Do you outsource?
- Get architecture reviews?

Choosing Training

Training is the secret to project success

Simply providing the resources is not enough

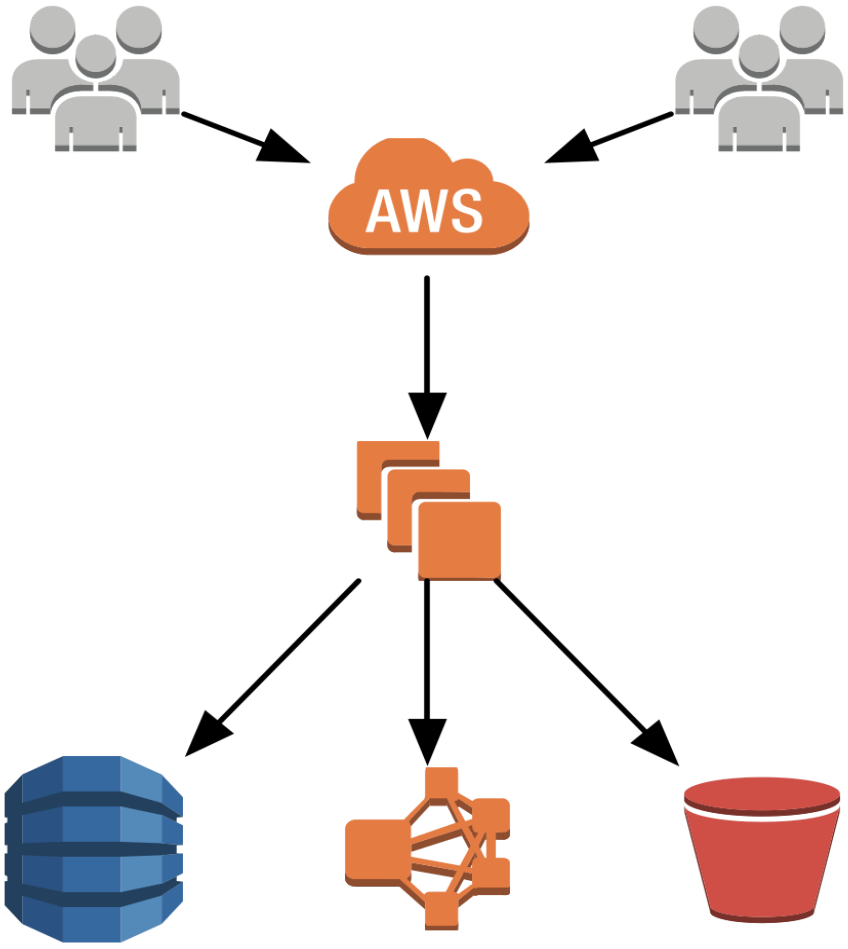
- Shiny new objects don't get used

Training will save the project money

Onsite training is personalized and focused on your use case

- A variety of skills and languages may be needed
- Hadoop uses Java and Linux

Creating A Proof Of Concept



Big Data Solutions Steps

1. Figure out if it is Big Data problem
2. Identify business needs
3. Get Training
4. Choose the right technologies
5. Write the prototype code
6. Create cluster in The Cloud
7. Evaluate. Was it a success? Did you gain value from data?
8. Repeat for crawl, walk, run

Time Frames

Choosing the time to switch from small data to Big Data is important

- There is an inherent cost of the cut over

A small company may have less to switch over

- Fewer resources to do it

A big company may have more legacy code and projects to switch over

- More resources to do it

I've seen it happen both ways

- Companies delay the switch too long and have vast technical debt
- Small companies make it big and can't switch fast enough
- Eventually you have to pay the piper

Course Chapters

- Thinking in Big Data
- Engineering Big Data Solutions
- **Doing Data Science on the NFL Play by Play Dataset**
- Conclusion
- Q & A

Doing Data Science on the NFL Play by Play Dataset

Why Football?

Football data is a metaphor for your data and your own workflow

- Use this as a starting point for ideas

You will need to take a dataset and:

- Clean it
- Analyze it
- Extract value
- Augment it

Plays

Advanced NFL stats released all Play by Play since 2002 season

2,898 total games

471,392 plays

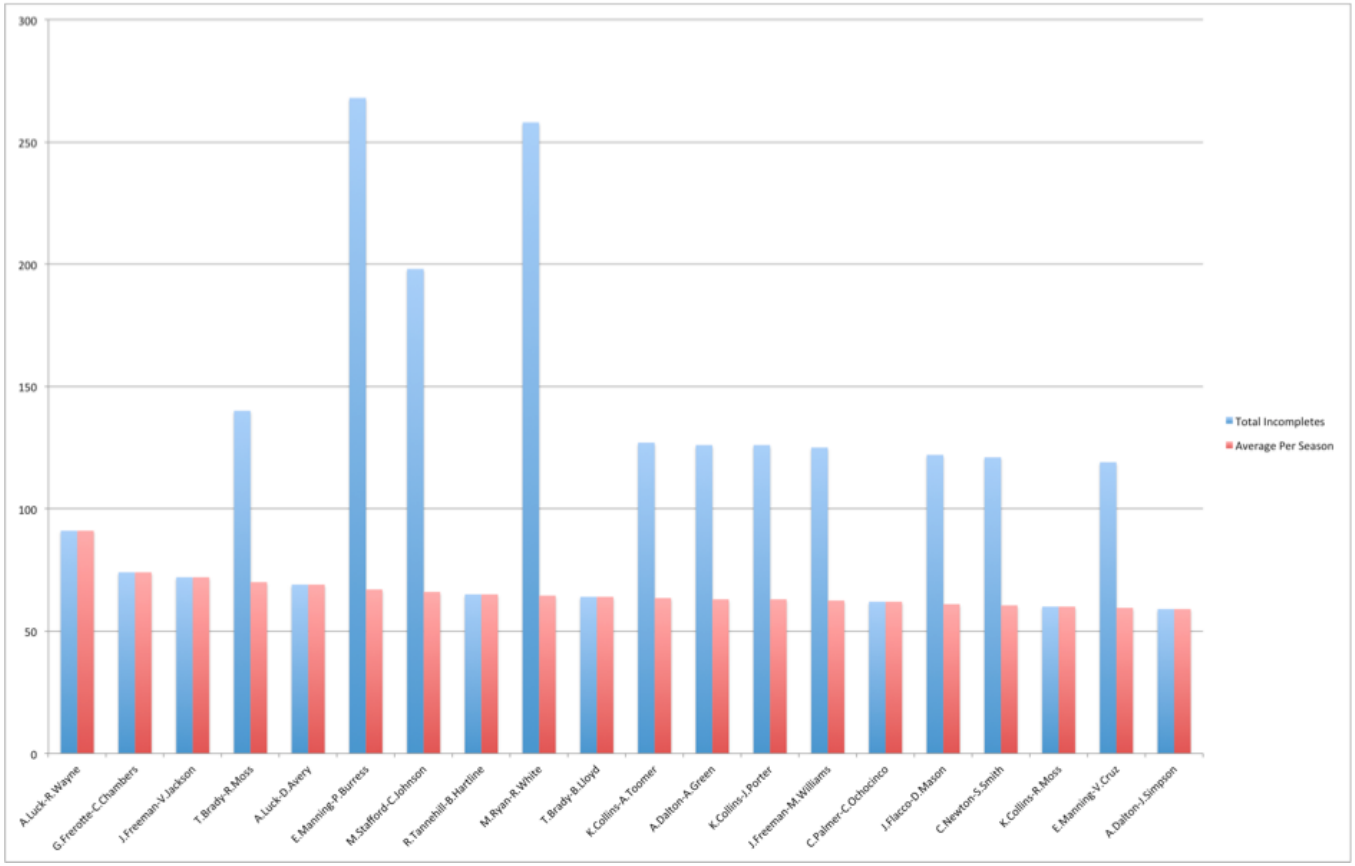
Full Play Entry

20121119_CHI@SF, 3,17,48,SF, CHI,3,2,76,20,0, (2:48) C.Kaepernick pass short right to M.Crabtree to SF 25 for 1 yard (C.Tillman). Caught at SF 25. 0-yds YAC,0,3,0,27,7 ,2012

Play Description

(2:48) C.Kaepernick pass short right to M.Crabtree to SF 25 for 1 yard (C.Tillman). Caught at SF 25. 0-yds YAC

There's A Chart for That



Incomplete passes to a receiver by a quarterback averaged over seasons together

There's A Custom MapReduce Behind That

```
public class IncompletesMapper extends
  Mapper<LongWritable, Text, Text, PassWritable> {
  @Override
  public void map(LongWritable key, Text value, Context context)
    throws IOException, InterruptedException {
    String line = value.toString();

    if (line.contains("incomplete")) {
      Matcher matcher = incompletePass.matcher(line);

      if (matcher.find()) {
        context.write(new Text(matcher.group(1) +
          "-" + matcher.group(2)),
          new PassWritable(1,Integer.parseInt(matcher.group(3))));

        // Lots more...
```

Enter the Query - The Hive Story

Queryable Data

Give me every run play by New Orleans in the 2010 season

From the Data: Fourth Downs

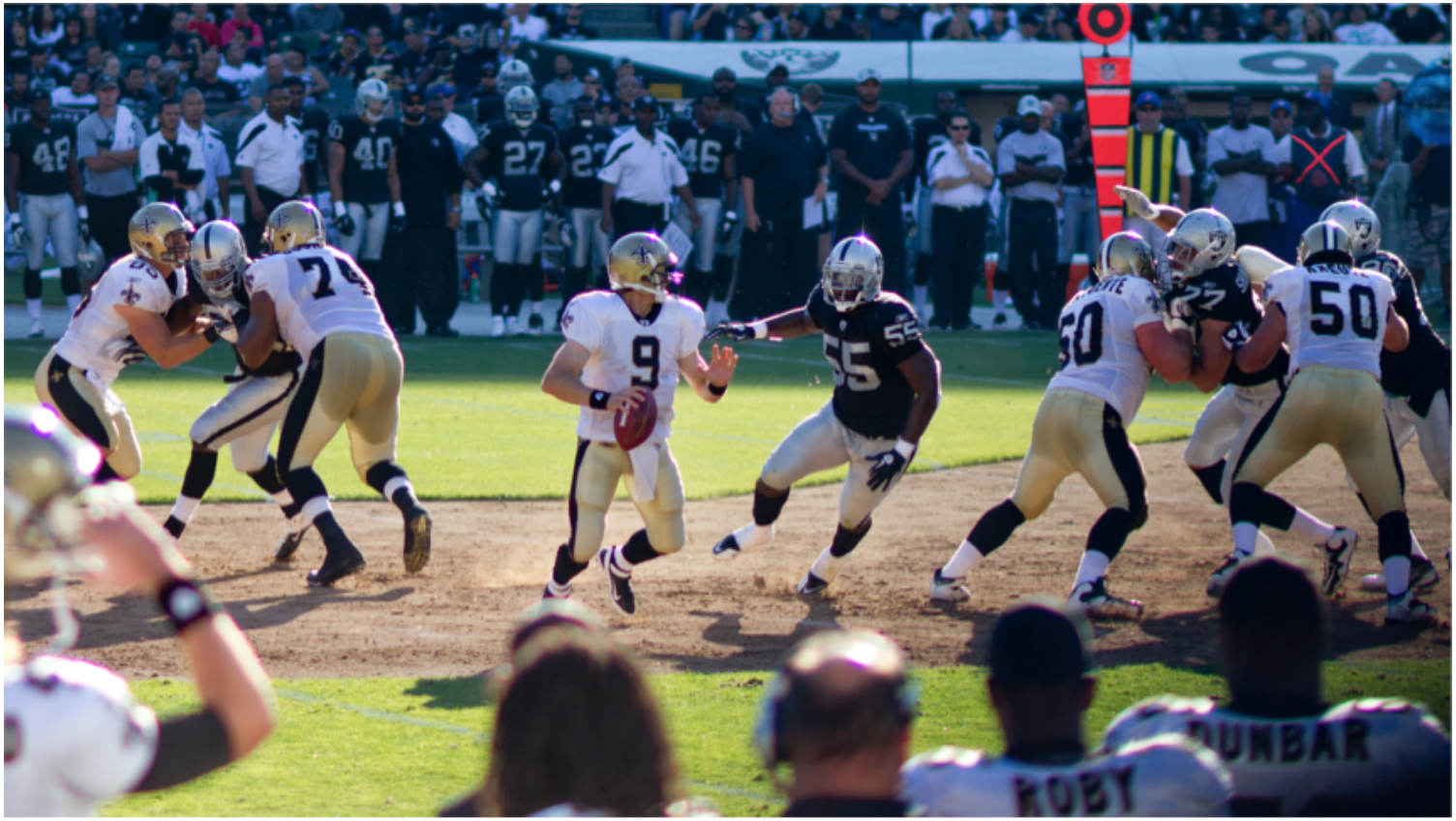
15% of 4th down plays weren't kicks

Play by Play Pieces

(2:48)
C.Kaepernick
pass short right
to **M.Crabtree** to
SF 25 for 1 yard
(**C.Tillman**).
Caught at SF 25.
0-yds YAC

From the Data: Sacks

QB sacks and scrambles double on 3rd downs



Hive

Abstraction on top of MapReduce

Allows queries using a SQL-like language

Hive Query

Give me every run by New Orleans in the 2010 season:

```
SELECT *
FROM playbyplay
WHERE
playtype = "RUN"
and year = 2010
and game like
"%NO%";
```

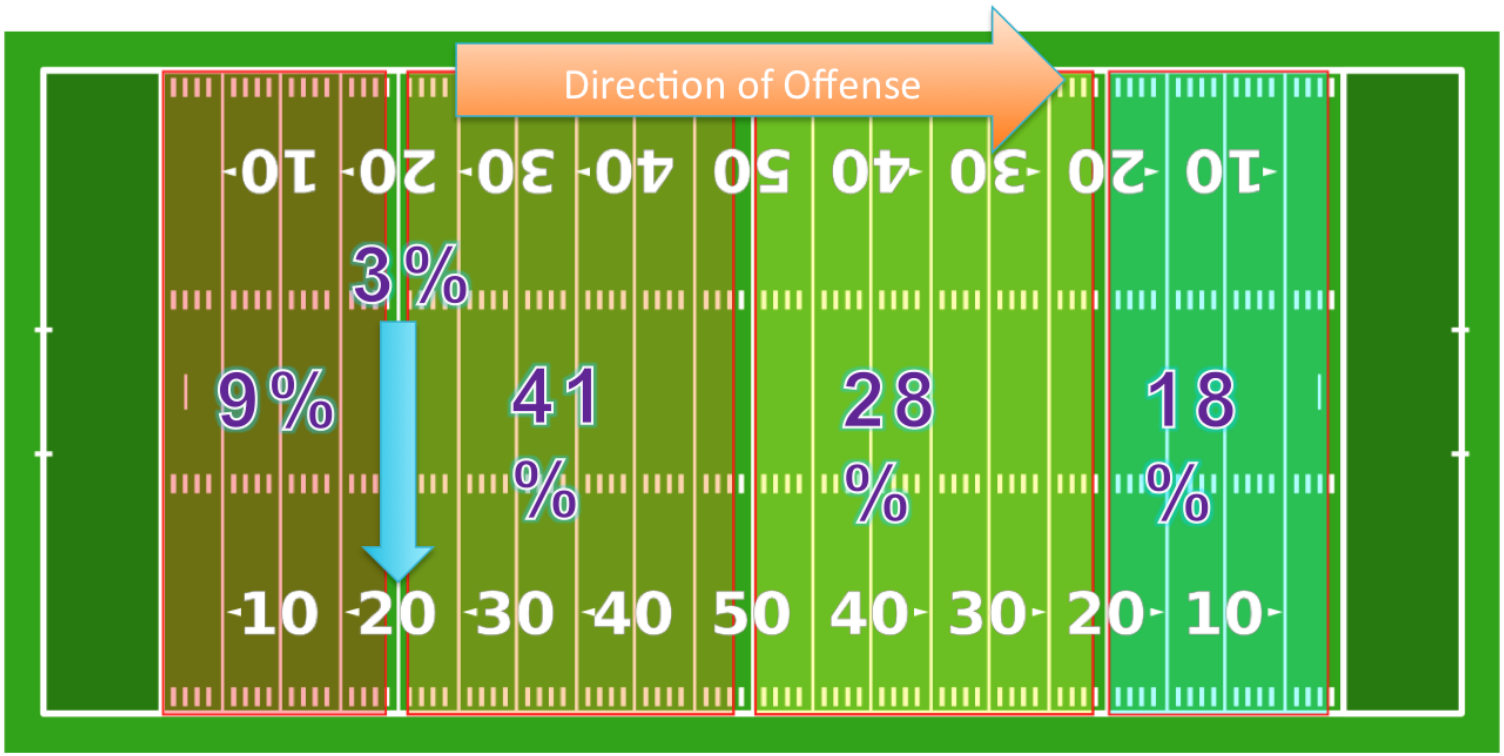
From the Data: Yards to Go

With 1 yard to go, 65% of plays are runs

Algorithms Alone - Lost in data

Data Janitorial

From the Data: Number of Plays By Yard Line



Stadium

Figuring Out Stadium

20121119_CHI@SF

Date Played

Away Team

Home Team

From the Data: Stadium Attendance

Stadiums with the smallest capacities average the best scores 20.55-17.79

Stadium Data

Field	Description
Stadium	The capacity of the stadium
Expanded Capacity	The expanded capacity of the stadium
Location	The location of the stadium
Playing Surface	The type of grass, etc that the stadium has
Is Artificial	Is the playing surface artificial
Team	The name of the team that plays at the stadium
Roof Type	The type of roof in the stadium (None, Retractable, Dome)
Elevation	The elevation of the stadium

From the Data: Stadium Elevation

There is a 1% increase in passes at Mile High versus others

Weather

1,015 games had weather

From the Data: Fumble

93% of games with weather have a fumble, 56% without

From the Data: Field Goals

Weather only increases misses by %1

- 14% of Field Goals are missed
- 21% of Field Goals are missed 30-39 MPH average winds

Weather Data

Field	Description
STATION	Station identifier
STATION NAME	Station location name
READING DATE	Date of reading
PRCP	Precipitation
AWND	Average daily wind speed
WV20	Fog, ice fog, or freezing fog (may include heavy fog)
TMAX	Maximum temperature
TMIN	Minimum temperature

From the Data: Weather

Wind had the most effect on games

At calm winds 41% pass and 37% run

At >30 MPH 34% pass and 46% run

From the Data: Home Field Advantage

Baltimore has the biggest weather advantage 22-14

Arrests

Arrest Data

Field	Description
Season	Season player arrested in (February to February)
Team	Team person played on
Player	Name of player arrested
Player Arrested	Was a player in the play arrested that season
Offense Player Arrested	Offense had player arrested in season
Defense Player Arrested	Defense had player arrested in season
Home Team Player Arrested	Home team had player arrested in season
Away Team Player Arrested	Away team had player arrested in season

Arrest = Win?

Whenever there are arrests either in the home team, away team or both, the home team:

Wins 57%

From 2002 to 2012, each team had many arrests. From to a low in 2002 of 56% to a

High of 91%

The data didn't bear out some of my preconceived notions

If we had every piece of data about a game, could we determine its outcome?

Course Chapters

- Thinking in Big Data
- Engineering Big Data Solutions
- Doing Data Science on the NFL Play by Play Dataset
- **Conclusion**
- Q & A

Conclusion

Summary

In this workshop you have learned:

- What is Hadoop
- What technologies exist in the Hadoop ecosystem
- What steps to take when creating a Big Data solution
- How data is valuable
- How data contradicts our preconceived notions

Course Chapters

- Thinking in Big Data
- Engineering Big Data Solutions
- Doing Data Science on the NFL Play by Play Dataset
- Conclusion
- **Q & A**

Q & A

Any questions?

About Me

Current: Instructor, Thought Leader, Monkey Tamer

Previously:

- Curriculum Developer and Instructor @ Cloudera
- Senior Software Engineer @ Intuit

Covered, Conferences and Published In:

- GigaOM, ArsTecnica, Pragmatic Programmers, Strata, OSCON, Wall Street Journal, CNN, BBC, NPR

See Me On:

- @jessetanderson
- http://tiny.smokinghand.com/linkedin
- http://tiny.smokinghand.com/youtube