Architecting Your First Big Data Implementation

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

Who is Mass Street?

- Boutique data consultancy
- We work on data problems big or “small”
- We have a focus on helping organizations move from a batch to a real time paradigm.
- 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

Follow Me!

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

Upcoming MeetUps

October: Joint MeetUp with Data Science KC. Reproducible Research with R, The Tidyverse, Notebooks, and Spark

November: Joint MeetUp with Kansas City Apache Spark MeetUp. Practicing IoT skills with Satori. (Still in the works.)

December: Survey of common data science algorithms.

This Evening's Learning Objectives

- Tonight is a rare lecture with no hands on.
  - I didn't have time to put together a show and tell.
- General Info
  - What is Big Data?
  - What is Hadoop?
  - What kind of problems does Hadoop help me solve?
- What do I need to build big data competency?
  - Engineers
  - Linux experts
  - Data Analyst (Internal!)
- Big data software and tools
  - All the stuff!
- Use cases

Mass Street Analytics

Nerd alert!

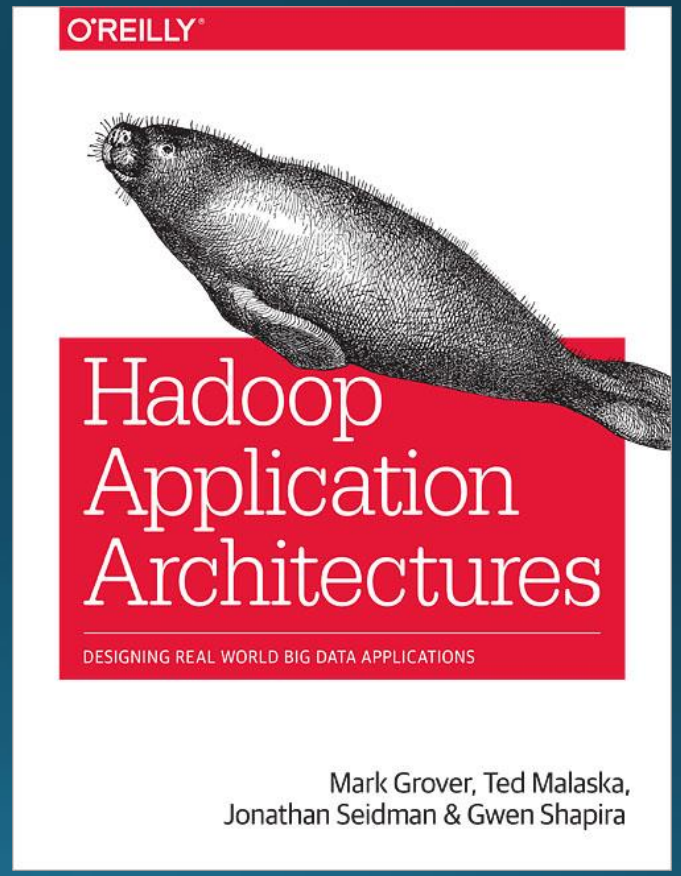
- Every single individual topic we're going to talk about could be:
  - A book unto itself
  - Several volumes of books unto itself
  - A three credit hour undergrad college course

- Watch out for soapbox moments!

Mass Street Analytics

Resources For This Evening's Material: Books

- Hadoop Application Architectures
- (it's a little dated)



Mass Street Analytics

Resources For This Evening's Material: Online Classes

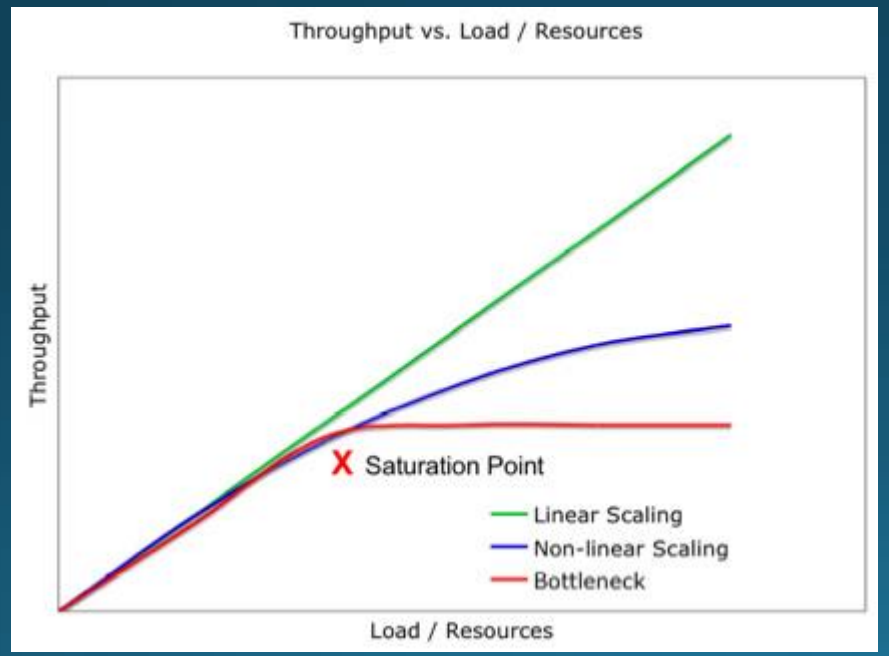
- Architect and Build Big Data Applications
- Analytic Data Storage in Hadoop

Buzzword definitions

- Distributed – many computers working on the same problem
- Commodity Hardware – This does not mean cheap. Means Cheaper.
- Schema-on-read – structure of data imposed at processing time.

Buzzword definitions

- Linear Scale – Adding more boxes gets you performance = mx+b vs. diminishing returns



Mass Street Analytics

What is Big Data?

Definition v1:
volume, velocity, variety

Mass Street Analytics

What is Big Data?

Definition v2:

When you have so much data you can't analyze it by traditional means.

What is Big Data?

```
SELECT address, COUNT(*)
FROM customers
GROUP BY address
```

What is Big Data?

Definition v3:

When you have so much data, you can't get things done in a reasonable amount of time.

What Is Hadoop?

1. Distributed computing for the masses.
2. Is the backbone to many distributed applications.
3. Can mean a lot of things depending on context.
    1. Could mean Hadoop core.
    2. Could mean the entire set of Big Data tools.
    3. Could mean the MapReduce framework.

What is Hadoop?

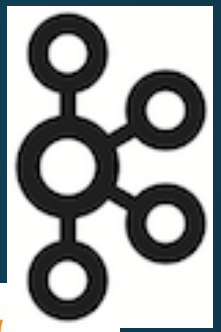
- Provides distributed fault tolerant data storage
- Provides linear scalability on commodity hardware
- Translation: Take all of your data, throw it across a bunch of cheap machines, and analyze it. Get more data? Add more machines


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!

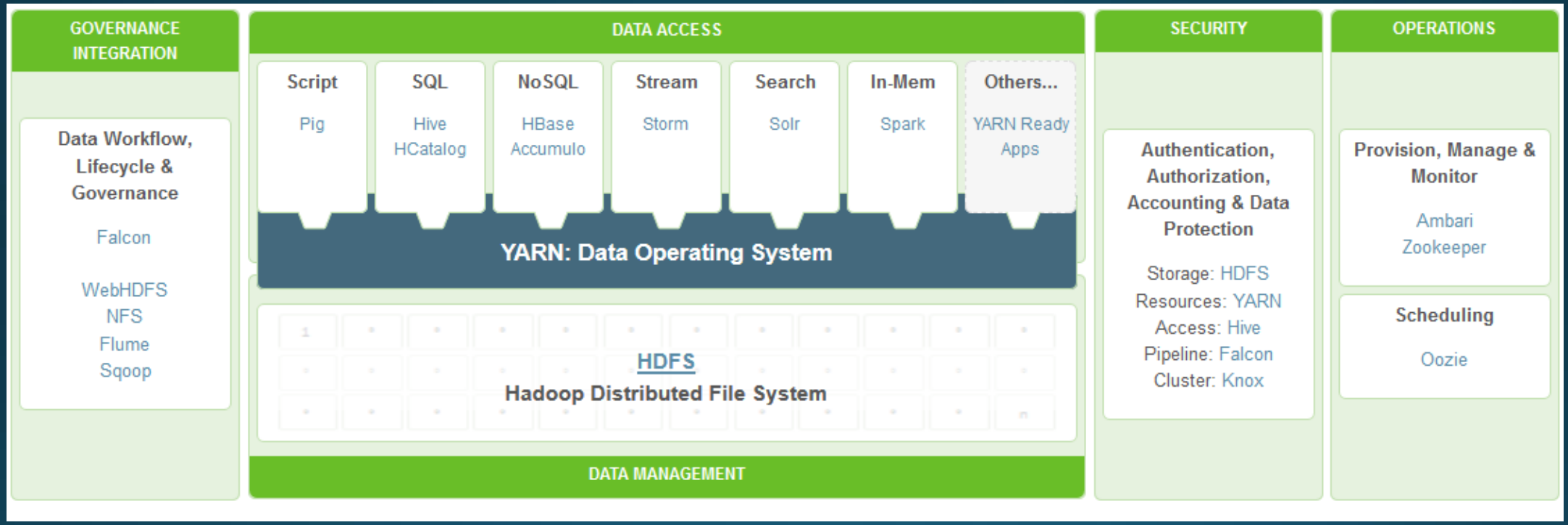


Photo Credit: Hortonworks website

What kind of problems does Hadoop help me solve?

- Easier analysis of big data
- Lower TCO of data
- Offloading ETL workloads
- Transition from batch to real time
- It lets you see the future.

What do I need to build big data competency?

- Engineers
  - Things aren't very point and clicky right now.
  - Java vs. Python vs. Scala
- Linux experts
- Data Analyst
  - These folks should be internal.
  - You can train a data scientist.
  - Python vs. R

Mass Street Analytics

Some questions to ask before you start your first big data project.

- Use these questions to drive your architecture.
- What kind of data do I have?
  - Highly organized vs. super messy.
  - Critical data that needs to move fast vs. data that can wait.
- How fast do I want to implement a solution?
  - Is there low hanging fruit?
- Who do I have vs. who do I need to hire?
  - Who is trainable?
  - Should I get outside help?

Some questions to ask before you start your first big data project.

- What is my current tech stack?
  - Windows vs. Linux
  - On Prim Vs. Cloud
    - If you're on prim do you have rack space?
  - What are my current BI Tools?
    - Excel vs. Microstrategy (free now!) vs. Bime vs. Tableau
- What is managing my application security?
  - LDAP vs. Active Directory

Mass Street Analytics

Big Data Software and Tools

- Most of it is free open source.
- Despite being open source, it's high quality.
- A lot of projects are part of the Apache Software Foundation.
- Several projects have dozens of committers.
- A cottage industry has sprung up around open source.

Cloud Solutions

- **AWS (S3)**
- **MS Azure**
- **Google Cloud (Big Query, Big Table)**
- **Databricks**
- **Confluent (Kafka Saas)**

Mass Street Analytics

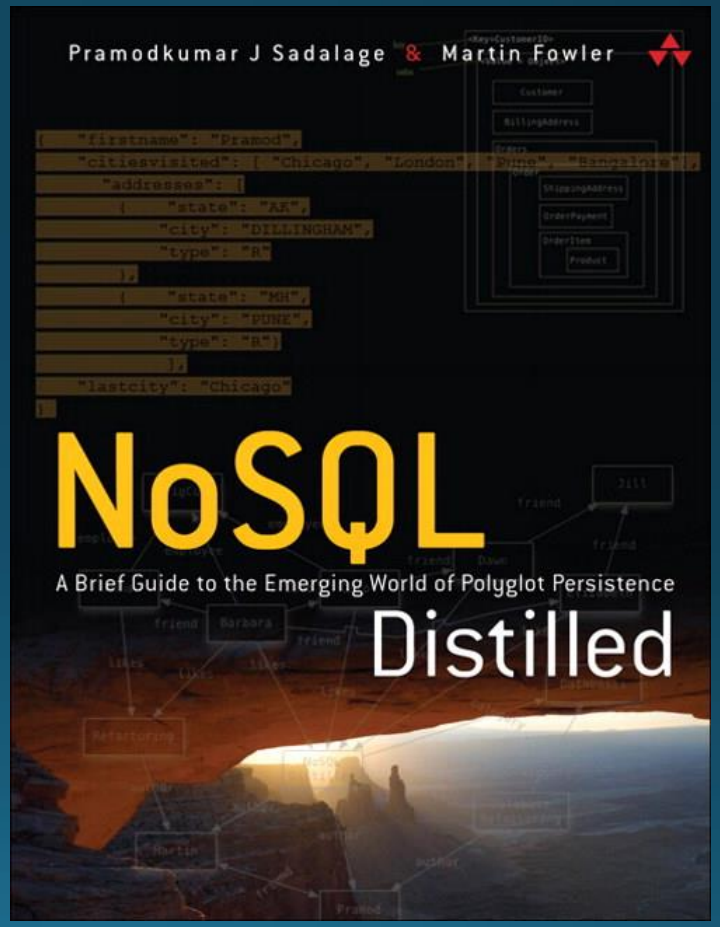
Cloud Solutions Bob's Thoughts

- **Great for tactical solutions.**
- **If you're a smaller organization, cloud offers a managed solution.**
- **Can be a hassle. You need top flight network engineers.**
- **Moving to the cloud is not something to do lightly. It requires long term thinking.**
- **I have no idea why anybody waste time with on prim servers anymore.**

Mass Street Analytics

NoSQL Databases

If you're interested in learning more about NoSQL, I suggest: NoSQL Distilled by Martin Fowler and Pramod J. Sadalage



Mass Street Analytics

NoSQL Databases

- Easy ETL
- Easy schema evolution
- Impedance mismatch solved

Types of NoSQL Databases

- Key Value – A simple has table, primary used when all access to the database is via primary key.
- Document – The database stores and retrieves documents, which can be XML, JSON, BSON
- Column-family - Data is stored in columns
- Graph – allows you to store relationships between entities

Issues with NoSQL Databases

- They do NOT work like relational databases
  - The guarantees that you are used to aren't there
  - CAP theory (consistency, availability, partition tolerance)

- Each database has it's own query language
  - That language will frequently look NOTHING like SQL

Examples of NoSQL Databases

Document	Key Value	Column Store	Graph
Mongo DB	Apache Accumulo	Apache Accumulo	Neo4J
	Redis (in memory)	Druid	
		Hive?	
		Cassandra (DataStax)	
		Hbase	

Mass Street Analytics

New SQL Databases

- Distributed versions of databases you're used to
- New concept Hybrid Transactional/Analytical Processing (HTAP)
- Generally two types
  - In Memory
  - Massively Parallel Processing (MPP)

New SQL Databases

- In Memory Databases
  - MemSQL
    - In Memory Databases: A Real Time Analytics Solution
    - MassStreet.net -> YouTube
  - VoltDB
  - NuoDB

New SQL Databases

- MPP Databases
- More suited to analytics
- Examples
  - Greenplumb (has gone open source!)
  - SQL Server PDW
  - MySQL Cluster CGE

Hadoop Distributions

- Get. A. Distro!
- What is a distro?
  - Popular big data software bundled together and installed as a total package.
- Popular distros:
  - Hortonworks
  - Cloudera
  - MapR
  - Various Others
- Most important difference is the support package
  - You can download and use distros without buying a support package.

Hadoop Core

- **Hadoop Distributed File System (HDFS)** – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- **Hadoop YARN** – a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications.
- **Hadoop MapReduce** – a programming model for large scale data processing.

Hadoop Core

- **Apache ZooKeeper** – A highly available system for coordinating distributed processes. Distributed applications use ZooKeeper to store and mediate updates to important configuration information.

- **Apache 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.

File Formats

- You can't just drop CSVs onto HDFS
  - That's super inefficient
- Pick a file format with the following requirements
  - Compressible
  - Splitable
- You have options on compression codecs
  - Snappy is popular
  - Ultimately an engineering decision

File Formats

- **Apache Avro**
  - Language neutral data serialization
  - Serializes data in a compact binary format
  - Fairly powerful tool
  - Useful for schema evolution

- **Apache Parquet**
  - General purpose storage format
  - Column oriented
  - Stores metadata thus self documenting

File Formats

- Avro is good for storing transactional workloads
- Parquet is good for storing analytical workloads
- When in doubt, use Avro
- Both can be used
  - Read and write Parquet with Avro APIs

IMPORTANT!

- Data in HDFS is IMMUTABLE
  - There is no random access
  - If you need that, you'll have to use a database technology
  - A lot of big data databases use HDFS to store the actual data file.

Batch Data Flow Tools

- **Apache Pig –** A platform for processing and analyzing large data sets. Pig consists of a high-level language (Pig Latin) for expressing data analysis programs paired with the MapReduce framework for processing these programs.

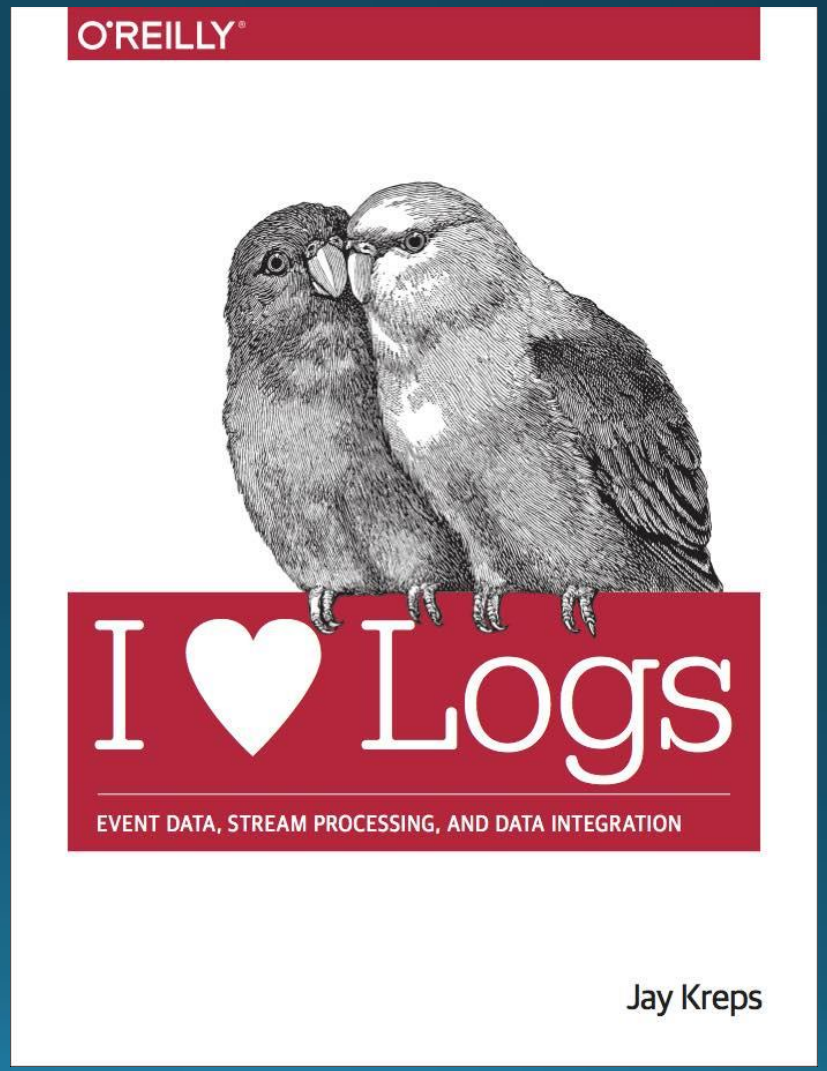
- **Cascading** - software abstraction layer for Apache Hadoop and Apache Flink. Cascading is used to create and execute complex data processing workflows on a Hadoop cluster using any JVM-based language (Java, JRuby, Clojure, etc.), hiding the underlying complexity of MapReduce jobs. It is open source and available under the Apache License. Commercial support is available from Driven, Inc.

- **Apache Sqoop –** Sqoop is a tool that speeds and eases movement of data in and out of Hadoop. It provides a reliable parallel load for various, popular enterprise data sources.

Real Time Processing

- **THIS IS THE QUANTUM LEAP GAME CHANGER IN YOUR DATA MANAGEMENT STRATEGY!!!!!**

- As a first step, read I (*heart*) Logs by Jay Kreps



Mass Street Analytics

Central Concepts

- The log as a unifying abstraction
- Everything in your org is an event that can be logged, captured, and transported

Real Time Processing Considerations

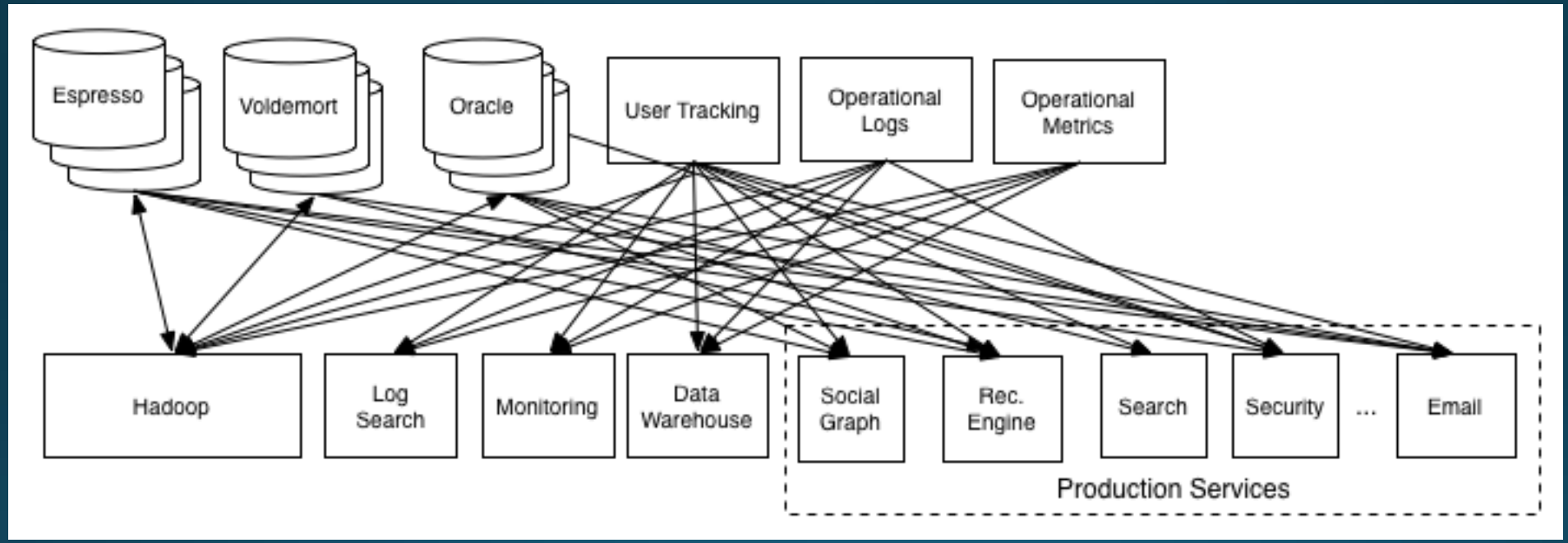
- Consider a radically different data architecture
  - Microservices?
- Apache Kafka is critical in any streaming architecture
- True Streaming vs. Microbatch
  - I don't think microbatch is a thing anymore
- Delivery guarantees
  - At least once
  - At most once
  - Exactly once

Real Time Processing

- **Apache Kafka** – Kafka is a fast and scalable publish-subscribe messaging system that is often used in place of traditional message brokers because of its higher throughput, replication, and fault tolerance.

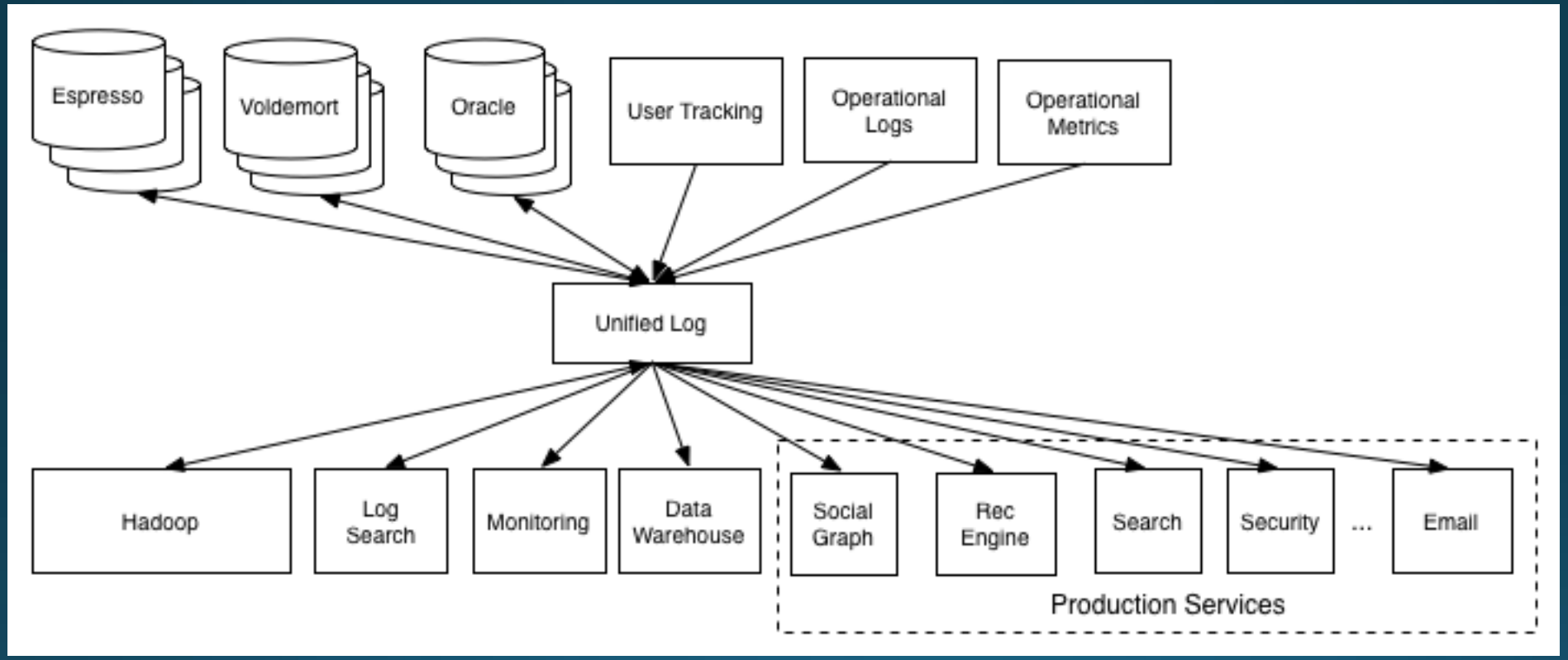
- Commercially backed by Confluent
  - Open source and Enterprise Versions

Radically Different Architecture



Mass Street Analytics

Radically Different Architecture



Mass Street Analytics

Real Time Processing

- **Apache Spark Structured Streaming** - scalable and fault-tolerant stream processing engine built on the Spark SQL engine.

- **Apache Storm** – Storm is a distributed real-time computation system for processing fast, large streams of data adding reliable real-time data processing capabilities to Apache Hadoop 2.x.

- **Apache Flume** – Flume allows you to efficiently aggregate and move large amounts of log data from many different sources to Hadoop.

Real Time Processing

- Other real time processing frameworks
  - Apache Flink
  - Apache Samza

- They all have benefits and drawbacks

Interesting Point and Click Tools

- Talend
  - Kafka/Spark streaming product
- Streamsets
  - Connector to SQL Server CDC
- NiFi

Data Access and Processing

- **Apache Spark** – Spark is ideal for in-memory data processing. It allows data scientists to implement fast, iterative algorithms for advanced analytics such as clustering and classification of datasets.

- **Apache Mahout** – Mahout provides scalable machine learning algorithms for Hadoop which aids with data science for clustering, classification and batch based collaborative filtering.

- **Hue** – Hadoop User Experience. Open source interface to the Hadoop ecosystem.
  - Provides interfaces and code editors for various projects.
  - Prefer it for interacting with data over Ambari.
  - If you're not using Cloudera, it's a little hard to install.

- **Apache Zeppelin** – really powerful notebook
  - Comes with Hortonworks

- **Jupyter Notebooks** – less powerful but more universal

Mass Street Analytics

BI Tools

- Apache Superset
- Microstrategy
- Things Board
- Grafana

Orchestration (Job Scheduler)

- Apache Oozie
- Apache Airflow

Security and Governance

- **Apache Knox** – The Knox Gateway ("Knox") provides a single point of authentication and access for Apache Hadoop services in a cluster. The goal of the project is to simplify Hadoop security for users who access the cluster data and execute jobs, and for operators who control access to the cluster.

- **Apache Ranger** – Apache Ranger delivers a comprehensive approach to security for a Hadoop cluster. It provides central security policy administration across the core enterprise security requirements of authorization, accounting and data protection.

Use Cases

- The use cases are only limited by your imagination and creativity.
- There are a few common use cases.
- We'll discuss two types:
  - Use cases you can find in a book.
  - My personal pet use cases

Use Case Assumptions

- All use cases require a cluster of machines
  - On prim
  - Or in the cloud
- All use cases will require a security/access plan
- You've had conversations with your engineers about the best approach for your specific situation

Use Case 1: Get started analyzing big data

Problem: We have large amounts of data and it takes a long time to analyze it.

Solution: Move that data into Hive.

Use Case 1: Get started analyzing big data

- Required tech:
  - Hadoop Core
  - Apache Hive
  - ETL tool of your choice
  - Data access tool of your choice
    - Recommend Hue
    - Ambari interface not that impressive for data access

Use Case 1: Get started analyzing big data

- Can be as simple as an ad-hoc process.
- Can be as complicated as a nightly run.
- Design
  - Identify data sets that need analyzed.
  - Think of it as creating views.
  - Create tables in Hive.
  - Dump data into those tables as necessary.

Use Case 2: Real Time Data Warehouse

Problem: Executives need faster access to historical data.

Solution: Re-engineer your ETL for a real time paradigm

Use Case 2: Real Time Data Warehouse

- Required tech:
  - Hadoop Core
  - Streaming ETL tool of your choice
    - Recommend StreamSets.
  - Apache Kafka
    - Recommend Enterprise Kafka through Confluent.
  - MPP database
    - Recommend Greenplumb

Use Case 2: Real Time Data Warehouse

- Required tech:
  - BI Tool of your choice
    - You should be able to plug your current BI tools right into Greenplumb

Use Case 2: Real Time Data Warehouse

- Does not require “Big Data”
- Will require a total rethink about how you move data around your organization.
- I highly recommend you read I (*heart*) Logs by Jay Kreps before you attempt this.
- I recommend that you create a POC.
  - Use cloud resources and the cost of your POC will be low.

Use Case 3: Breaking down data silos

Problem: We have data located in many different data sources being used differently by different parts of the enterprise.

Solution: Create a data lake/data hub.

Use Case 3: Breaking down data silos

- Required tech:
  - Hadoop Core
  - ETL tool of your choice
  - Data access tool of your choice
  - BI tool of your choice

Use Case 3: Breaking down data silos

- This is a batch process.
  - You can stream in data if you want.
- The goal is to drop everything into Hadoop in its NATIVE FORMAT.
  - Still use a file format
- Store it now. Figure out what to do with it later.

Mass Street Analytics

Use Case 3: Breaking down data silos

- Implementing a data lake is a presentation all its own.

- Sounds simple. Actually pretty complicated.
  - Recommend: Architecting Data Lakes by Alice LaPlante and Ben Sharma

Use Case 3: Breaking down data silos

- Data lakes solve a lot of problems at once.
  - Really cuts down the time on ETL development
  - Removes need to put some virtualization solution in place
  - Removes the need to put some federated data solution in place.
  - Implements a self serve model
  - Gives everybody access to all the data
    - Facilitates 360 degree analysis

Mass Street Analytics

Use Case 3: Breaking down data silos

- Data governance here is key
  - Data dictionary has to be on point.
  - Stuff needs to be clean.
  - All datasets need clear provenance.

Mass Street Analytics

Other Use Cases

- Clickstream Analysis
  - Basically analyze weblogs
  - Can be done in batch
  - More interesting to do it in real time
  - https://hortonworks.com/tutorial/visualize-website-clickstream-data/

Other Use Cases

- Fraud Detection
- Network Intrusion detection
- Same approach
  - Pull in data in real time and make decisions on it.
  - Requires data science tools to create models of behavior.
- Aren't there commercial solutions?
  - Pretty generic use case.
  - Build vs. buy.

Mass Street Analytics

Obligatory Q & A Period