Big Data Technologies And Why They Matter To R Users

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

Mass Street Analytics

No Deep Dives Today

It’s all 10,000 ft and the speed of heat

Mass Street Analytics

Motivation For Today's Presentation

- If I have more data can't I just get a bigger machine?

- How does Spark relate to Hadoop?

Mass Street Analytics

Today’s Take Away

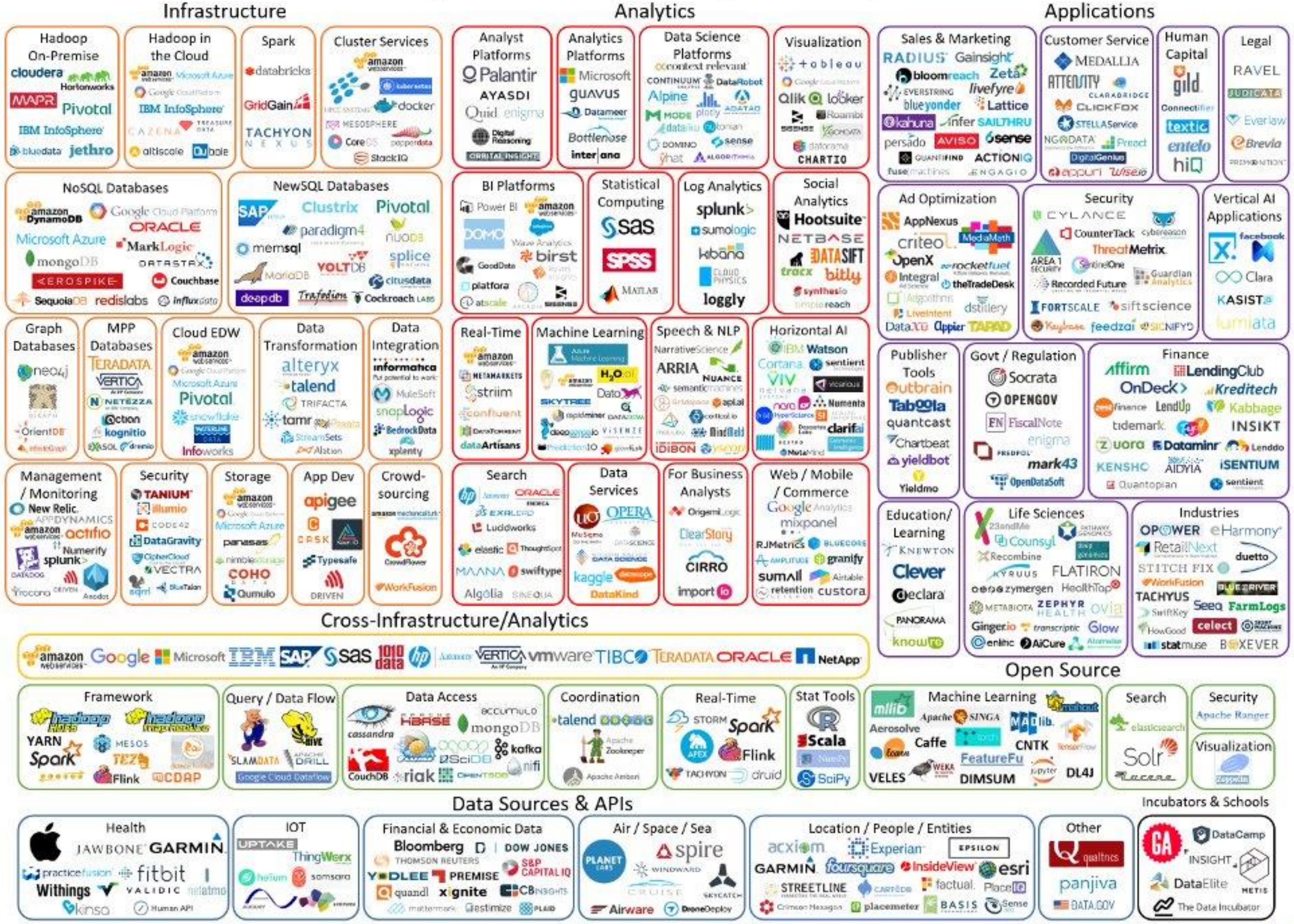
- The way you access data in the future is going to radically change. Be aware and ready

Intended Audience

- People getting started in data
- Academics that work with data
- (If you’re already working in industry, you might be bored

What exactly have you been studying for the past four years?!

Big Data Landscape 2016 (Version 2.0)



Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
amazon webservices, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon webservices, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, NuoDB, memsql, splice machine, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, GRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio

Cloud EDW
amazon webservices, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

Management / Monitoring
New Relic, APPDYNAMICS, amazon webservices, actifio, Numerify, splunk, DATADOG, DRIVEN, Ascdot, Rocana

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon webservices, Google Cloud Platform, Microsoft Azure, panasas, nimble storage, COHO DATA, Qumulo

App Dev
apigee, CASK, DRIVEN, Typesafe

Crowdsourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, Bottlenose, interana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, plotly, ADATAO, MODE, dataiku, Yhat, DOMINO, sense, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Qlik, looker, SISENSE, Roambi, ZOOMDATA, datorama, CHARTIO

BI Platforms
Power BI, amazon webservices, DOMO, salesforce Wave Analytics, birst, GoodData, Zoomdata, platfora, ARCADIA, SISENSE, atscale

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach

Real-Time
amazon webservices, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon, H2O.ai, Dato, SKYTREE, DATARPM, rapidminer, DATAROBOT, deepsense.io, VISENZE, PredictionIO, glowfish

Speech & NLP
NarrativeScience, ARRIA, NUANCE, semantic machines, Gridspace, api.ai, cortical.io, maluuba, MindMeld, IDIBON, Yseop

Horizontal AI
IBM Watson, Cortana, sentient, VIV, vicarious, nervana, nara, Numenta, HyperScience, Scaled Inference, Dextro, Deepomatic Labs, clarifai, MetaMind

Search
hp, Attivio, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
UO, OPERA, Mu Sigma, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumAll, Airtable, retention science, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, Infer, SAILTHRU, persado, AVISO, 6sense, QUANTIFIND, ACTIONIQ, fuse machines, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
AppNexus, criteo, MediaMath, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, Adgorithms, dstillery, LiveIntent, DataXu, Appier, TAPAD

Security
CYLANCE, cybereason, CounterTack, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Guardian Analytics, Recorded Future, FORTSCALE, siftscience, Kaybase, feedzai, SIGNIFYD

Vertical AI Applications
x.ai, facebook M, Clara, KASISTO, lumiata

Publisher Tools
Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
Affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient

Education/Learning
KNEWTON, Clever, declara, PANORAMA, knewre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, Seeq, FarmLogs, SwiftKey, HowGood, celect, Sight Machine, statmuse, BOXEVER

Cross-Infrastructure/Analytics
amazon webservices, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Apache Beam, Flink, CDAP

Query / Data Flow
Pig, HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, HBASE, accumulo, mongoDB, CouchDB, Hadoop, SciDB, kafka, riak, OPENTSDB, nifi

Coordination
talend, OOZIE, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mllib, mahout, Apache SINGA, MADlib, Aerosolve, torch, Caffe, CNTK, TensorFlow, Theano, FeatureFu, WEKA, VELES, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
zeppelin

Data Sources & APIs

Health
Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AVANT, ECHELON

Financial & Economic Data
Bloomberg, D&B, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CB INSIGHTS, mattermark, Estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
acxiom, Experian, EPSILON, GARMIN, foursquare, InsideView, esri, STREETLINE, CARTODB, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS, Sense360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, DataCamp, INSIGHT, METIS, DataElite, The Data Incubator

Last Updated 2/12/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

We're Experiencing A Modern Tech Renaissance

Hadoop -> SQL On Hadoop -> NoSQL ->

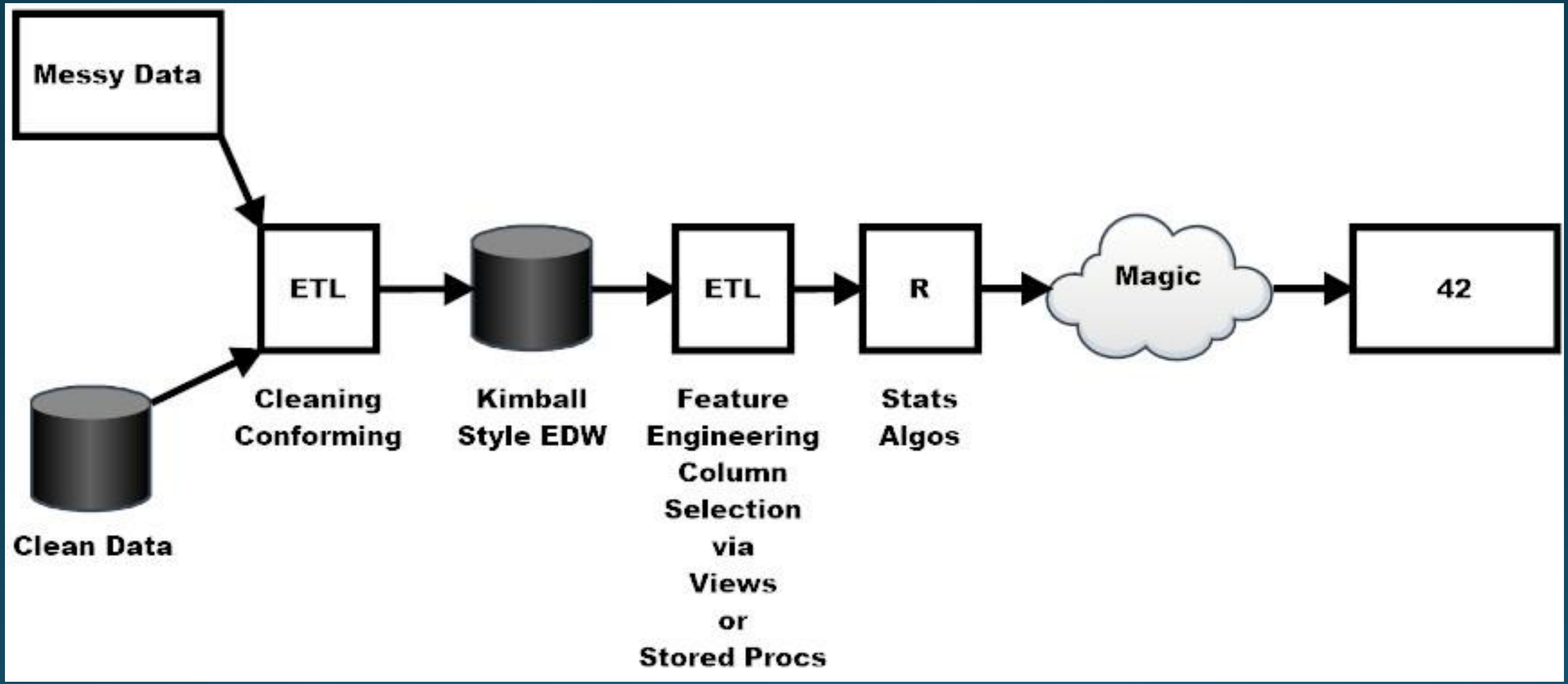
In Memory Databases ->MPP -> Spark 1.0 ->

Spark 2.0 -> Deep Learning -> Blockchain

Podcast Plug

- Not So Standard Deviations
  - Roger Peng, PhD Biostatistician Johns Hopkins School Of Public Health
  - Hillary Parker, PhD Data Scientist Stitch Fix

My Workflow



Mass Street Analytics

Data Retrieval And Storage
What is Hadoop?

- Can mean a lot of things depending on context.
  - Could mean Hadoop core.
  - Could mean the entire set of Big Data tools.
  - Could mean the MapReduce framework.

Data Retrieval And Storage
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

Mass Street Analytics

Data Retrieval And Storage
Cloud Storage

- AWS
  - S3 is the gold standard
- MS Azure
- Google Cloud

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

Data Retrieval And Storage

What Is Hive?

- Hadoop warehouse solution
- SQLesque language called Hive Query Language
- Adds structure to unstructured data
- Provides a window into HDFS
- You can connect through ODBC/JDBC
  - Which means you can work with Hive using standard BI tools

Mass Street Analytics

Data Retrieval And Storage NoSQL Databases

- Key Value – A simple hash table, primarily 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

Mass Street Analytics

Data Retrieval And Storage 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

Data Retrieval And Storage
Examples of NoSQL Databases

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

Mass Street Analytics

Data Retrieval And Storage
New SQL Databases

- Distributed versions of databases you're used to
- New concept Hybrid Transactional/Analytical Processing (HTAP)

Data Retrieval And Storage
New SQL Databases

- In Memory Databases
  - MemSQL
  - VoltDB
  - NuoDB

Data Retrieval And Storage
MPP Databases

- More suited to analytics
- Examples
  - Greenplumb
  - SQL Server PDW
  - MySQL Cluster CGE

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.

Mass Street Analytics

Real Time Processing

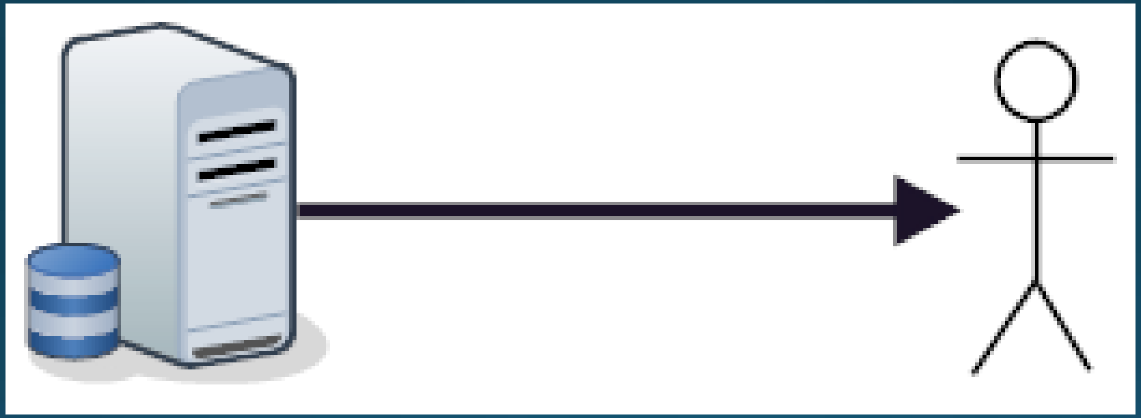
- Other real time processing frameworks
    - Apache Flink
    - Apache Samza

- They all have benefits and drawbacks

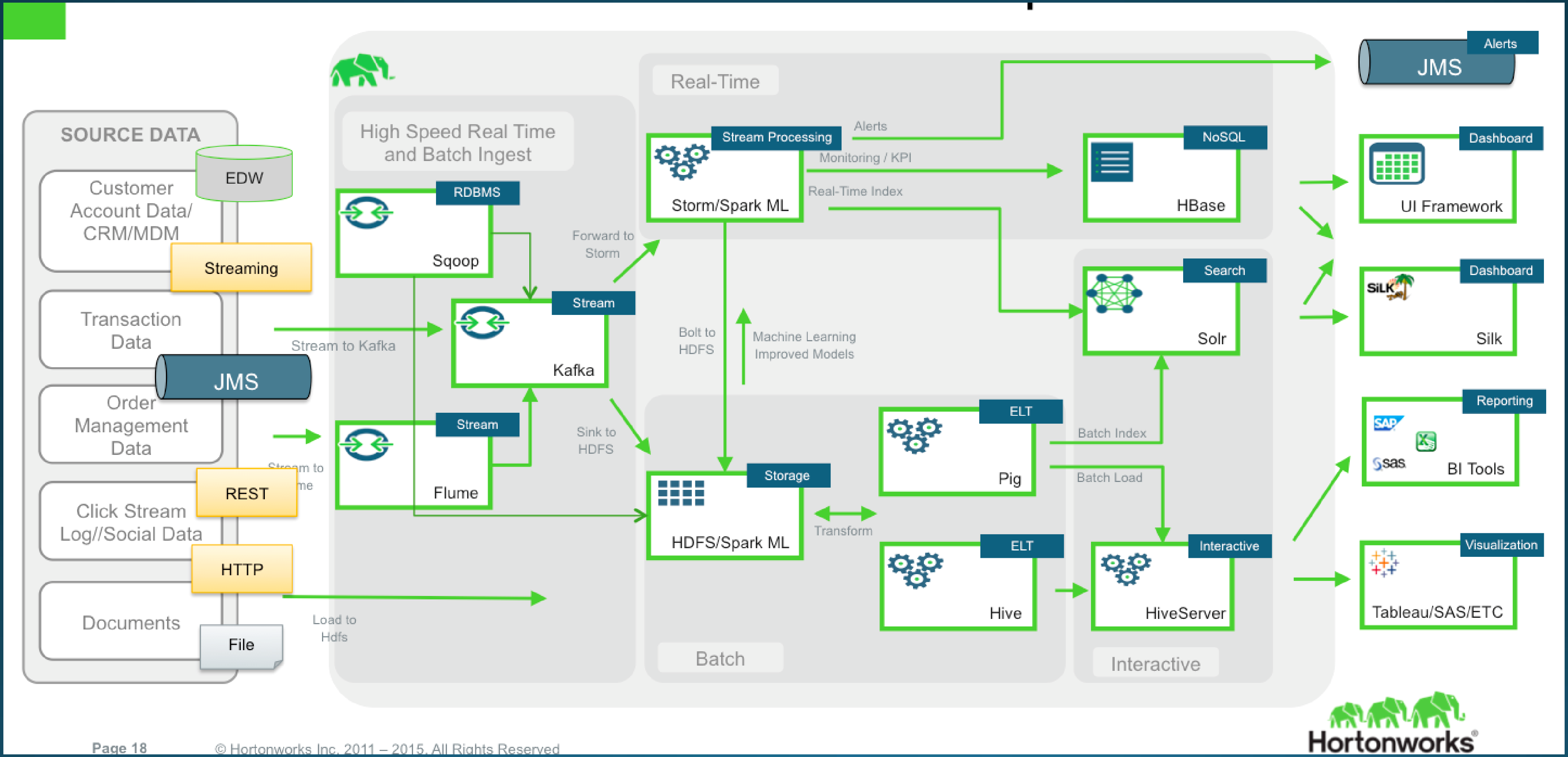
What Is Spark

- Distributed in memory processing framework
- Rapidly replacing MapReduce as a means to crunch data.
- Many ways to interact with Spark
  - R, Java, Scala, Python
  - Sparkr, Sparklyr, H2O with Sparkling Water
- Several APIs
  - RDDs, Dataset/Dataframe, Spark Streaming, Spark SQL, Spark Streaming, Strucutred Streaming

How You Work Now



What You Might Have To Deal With



Q&A

Mass Street Analytics