**Agenda**

- A brief review of Big Data solutions
- From yesterday to today
  - New technologies
  - Emerging technologies
- Commercial solutions
- Even more use cases
- The impact on business and society as a whole
- Q&A
Big Data Review
What is **Big Data** why are we discussing it?

• Megabytes, gigabytes, terabytes, petabytes, exabytes, zettabytes, *yottabytes*, and now, *gigantabytes*

• Data, data, and more data
  
  – Facebook with 800+ million users generating something like 1.5 TB of data per day, storing over 70 PB of data
  – Google indexes the Web
  – 2 ZB stored by end of 2012!
  – 1+ ZB IP traffic annually by 2016
What is **Big Data** why are we discussing it?

**Big data—a growing torrent**

- **$600** to buy a disk drive that can store all of the world’s music
- **5 billion** mobile phones in use in 2010
- **30 billion** pieces of content shared on Facebook every month
- **40%** projected growth in global data generated per year vs. **5%** growth in global IT spending
- **235 terabytes** data collected by the US Library of Congress by April 2011
- **15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress

**Big data—capturing its value**

- **$300 billion** potential annual value to US health care—more than double the total annual health care spending in Spain
- **€250 billion** potential annual value to Europe’s public sector administration—more than GDP of Greece
- **$600 billion** potential annual consumer surplus from using personal location data globally
- **60%** potential increase in retailers’ operating margins possible with big data
- **140,000–190,000** more deep analytical talent positions, and
- **1.5 million** more data-savvy managers needed to take full advantage of big data in the United States
What is *Big Data* why are we discussing it?

- **Biomedical Big Data** is being driven from multiple areas and disciplines:
  - Genomics (genotyping, gene expression, sequencing)
    - 4 TB per person, now at the rate of 1,000s of people per month!
    - The computational complexity here is very high
  - Molecular pathology and predictive analytics
  - Payer/provider data (Electronic Medical Records, insurance records, pharmacy, etc.)
  - Information monetization
    - Making usage of the massive amounts of data already accumulated in electronic format
  - Predictive technologies are very important
What is *Big Data* why are we discussing it?

- So we now know that adoption is occurring at a staggering pace.
  - Web scale corporate adoption began almost 8 years ago.
  - Corporate adoption was slower, but has now become mainstream.

- Big Data solutions now provide us with the means to do this analysis and at a *price point* that is obtainable by many
  - The value in delivering such solutions is very large

- Big Data research now permeates biomedical data analysis, logistics, financial services, retail – really just about everything
What is *Big Data* why are we discussing it?

• So, we are seeing three main drivers here:
  – The sheer volume of data that must be processed
  – The rate at which the data arrives
  – The variation in data and the computational complexity of the required analytics

• We often talk about the “3 Vs” - *volume, velocity, and variety*.

• A year ago, these three items characterized Big Data. Today we also recognize that *Complexity* is another driver.
  – Google and pandemic prediction – generated 450 million models run against 5.5 quadrillion inquiries, and achieved amazing accuracy
What is *Big Data* why are we discussing it?

- Why else? Because it works!
  - Nate Silver
  - NYC Manhole cover explosions
  - The Obama campaign

- Is this a new paradigm?
  - N=ALL analysis
  - Yielding models that are developed prior to the theory
  - This is the real rise of machine learning
  - **We will return to this shortly....**
From ETL to Analytics

Potential Use Cases for Big Data Analytics

Real time
- Credit & Market Risk in Banks
- Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks (including Social Network Analysis)
- Event-based Marketing in Financial Services and Telecoms
- Markdown Optimization in Retail
- Claims and Tax Fraud in Public Sector

Data Velocity
- Predictive Maintenance in Aerospace
- Social Media Sentiment Analysis
- Demand Forecasting in Manufacturing
- Disease Analysis on Electronic Health Records
- Traditional Data Warehousing
- Text Mining
- Video Surveillance/Analysis

Batch
- Structured
- Semi-structured
- Unstructured

Data Variety

And Volume!

And Complexity!!
Innovative Use Cases

• Natural Language Processing
  – Sentiment analysis
  – Computational linguistics - Watson, the Jeopardy champion - new

• Earth sciences
  – Grid management
  – Discovery – sensor data
  – BP Oil spill response
  – Traffic control and logistics - new

• Marketing
  – Attribution Marketing - new
    • Golden steps
  – Recommendation engines – cluster analysis
    • The Target debacle
  – Social network analysis
    • The Facebook bigamist

• Biomedical
  – Computational geometry
  – Genetic/genomic research
  – Outcomes analysis - new
  – Disease epidemics – Google H1N1 - new

• Financial
  – Stress testing and back testing
  – Monte Carlo simulation
  – Risk management
  – Rogue trading
  – Market segment failure impact - new

• Fraud and abuse
  – Cyber-security
  – Financial fraud
  – Pharmaceutical fraud and drug diversion
Hadoop Technology and Solution Sets
Big Data Processing Systems

• **Hadoop – Java based Platform from Apache**
  – Distributed File System from Yahoo! based on the Google Distributed Files System, and BigTable
  – MapReduce – Google’s famous parallelization architecture. A core component of Hadoop. Language independent.

• Now an emergent open source community with significant contributions from LinkedIn, Amazon.com, Twitter, Facebook, Google, and many, many others
  – We must mention Cloudera, Hortonworks, and MapR

• And later we will dive into the proprietary market
Big Data Processing Systems

Lines of Code Contributed Since 2006, Cloudera Method

- hortonworks
- yahoo
- cloudera
- facebook
- linkedin
- ebay
- ibm
- inmobi
- apple
- twitter
- amazon
- other
The Technology of Hadoop

• The Hadoop parallel processing architecture
  – Understanding MapReduce
  – MapReduce v2 overview and goals
  – The Hadoop family of technologies
  – Other MapReduce Frameworks

• NoSQL Databases
  – HBase
  – MongoDB, Cassandra, and CouchDB
  – Vendor specific Big Data products
    • IBM, Vertica, and Greenplum

• Pig, Hive, Sqoop, Flume, Zookeeper, Mahout, and more
The Technology of Big Data

- Map Reduce in three images

Data is processed in parallel
The Technology of Big Data

- Map Reduce in three images

Data is processed in parallel

Data is sorted and assembled
The Technology of Big Data

- Map Reduce in three images
Some MapReduce Notes

• MapReduce is really just a standard Java process that works against partitioned data blocks.
  – The JVM is created by the Hadoop MapReduce framework and the Java code library is supplied to a target machine and started in this JVM.
  – The framework also carves up the data blocks into units to be processed by each unit.
  – Only Map phases are mandatory. Reduces are optional.
    • Note that presence of one or more Reducers implies that a sort will occur.
• When coded in Java, MapReduce can use standard Java libraries, specialized libraries (JSON, Google, others)
• MapReduce jobs can use JDBC to access databases, or can also use a ORM framework (like hibernate, JPA, etc.)
• Creative usage of Map -> Reduce, sort keys, and partitioner design patterns are being defined. They must be studied and learned.
MapReduce v2.0 - YARN

In the new version 2.0 architecture, the JobTracker is split into two areas of responsibility:
• Resource Management
• Job Scheduling and Monitoring.
MapReduce v2.0 - YARN

In the new version 2.0 architecture, the JobTracker is split into two areas of responsibility:

• Resource Management
• Job Scheduling and Monitoring.

The Resource Manager and its slave Node Managers are responsible for all computational activities. The RM consists of a Scheduler and an Application Manager.
MapReduce v2.0 - YARN

In the new version 2.0 architecture, the JobTracker is split into two areas of responsibility:

- Resource Management
- Job Scheduling and Monitoring.

Scheduling is now CPU, disk, memory, and network aware.

The Resource Manager and its slave Node Managers are responsible for all computational activities. The RM consists of a Scheduler and an Application Manager.
MapReduce v2.0 - YARN

In the new version 2.0 architecture, the JobTracker is split into two areas of responsibility:
• Resource Management
• Job Scheduling and Monitoring.

Scheduling is now CPU, disk, memory, and network aware.

The Resource Manager and its slave Node Managers are responsible for all computational activities. The RM consists of a Scheduler and an Application Manager.

PaaS is the obvious intended direction
Other MapReduce Frameworks

• Cascading from Concurrent
  – Uses source taps, pipes, and sinks to represent data flow
  – Implements a great library of common (SQL-like) connectors
Other MapReduce Frameworks

• Data streaming frameworks
  – Utilizes STDIN and STDOUT (console I/O) for Map and Reduce functions
  – Wrappers exist for Python, Ruby (Wukong, MRToolkit), Javascript, Perl, etc.

```python
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    print '%s\t%s' % (word, 1)
```

```ruby
#!/usr/bin/ruby
count = 0
ARGF.each do |line|
    count += 1
end
puts "#{count}"
The Hadoop Family

- **HBase** – a NoSQL database built on top of HDFS
- **Hive** – a “SQL” queriable “database” of sorts that produces M/R “jobs”
- **Pig** – a scripting language that produces M/R “jobs”
- **Zookeeper** for distributed systems management
- **Flume** and **Sqoop** for data acquisition
  - Flume accesses Web logs
  - Sqoop accesses relational databases
- **Oozie** for workflows
- **Impala**
- **Mahout** for machine learning
- **Graph databases**
The Hadoop Family - HBase

“Sharding” of records
The Pig shell is named Grunt. It isn’t good enough to be called Oink!
The Hadoop Family – Pig and Hive

**Pig Latin**

countries = load '/user/gharrison/PIG COUNTRIES' AS (country_id, country_name, country_subregion, region);
customers= load '/user/gharrison/PIG CUSTOMERS' AS (cust_id, first_name, last_name, gender, job, marital, postcode, city, country_id);
asiacountries = filter countries by region matches 'Asia';
joined = join customers by country_id, asiacountries by country_id;
grouped = group joined by country_name;
agged = foreach grouped generate group, COUNT(joined.customers::cust_id);
morethan500cust = filter agged by $1 > 500;
ordered = order morethan500cust by $1 desc;
dump ordered;

**SQL or Hive QL**

```
SELECT country_name, COUNT(cust_id) AS cust_count
FROM countries co
JOIN customers cu ON (co.country_id = cu.country_id)
WHERE country_region = 'Asia'
GROUP BY country_name
HAVING COUNT(cust_id) > 500
ORDER BY cust_count DESC
```
Sqoop

- Designed to easily move data between Hadoop cluster (HDFS, HBase, or other) and relational databases
  - Relies on JDBC driver
  - Can gather metadata
  - Can store data in many formats (text, Avro, SequenceFile, etc.)
  - Supports a form of “check-pointed” data acquisition
  - Can build persistent cataloged jobs
Flume

- A specialized utility to load data into HDFS from log files
  - Web logs (W3C format, and other readers available)
  - Syslog
  - Jobtracker, Namenode, Hadoop job logs
  - Twitter
  - Custom

- Uses a highly adaptable source -> channel -> sink data flow paradigm
• Zookeeper is an essential element in Hadoop. Zookeeper is a distributed storage that provides the following guarantees:
  – Sequential Consistency - Updates from a client will be applied in the order that they were sent.
  – Atomicity - Updates either succeed or fail. No partial results.
  – Single System Image - A client will see the same view of the service regardless of the server that it connects to.
  – Reliability - Once an update has been applied, it will persist from that time forward until a client overwrites the update.
  – Timeliness - The clients view of the system is guaranteed to be up-to-date within a certain time bound.
### Real Time Big Data

Table 1: Google Gives Us A Map (source: Strata + Hadoop World 2012 Keynote: Beyond Batch - Doug Cutting)

<table>
<thead>
<tr>
<th>Google Publication</th>
<th>Hadoop</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigTable (2006)</td>
<td>HBase (2008)</td>
<td>Online key/value</td>
</tr>
</tbody>
</table>
Impala

- Real-time SQL query against HDFS files
- Uses distributed servers to process requests in parallel
- Phenomenal performance against petabyte scale datasets
Impala

Other benefits include:

- Interface to other applications through standard JDBC/ODBC bridges
- Reuse of existing (Hive/HCatalog) metadata
- Seamless access to either HDFS or HBase resident data
Machine Learning

- Statistical Machine Learning
  - Designed to be better than rules
  - Statistical analysis of the data
  - Highly adaptive
  - But, no real “understanding” of the underlying problem
- Mahout is the Hadoop answer (one of)
  - It is highly scalable and contains many advanced features
  - It contains a command line interface that can be good to prototype
  - It is amazingly difficult to use when you must dig into the API
  - It is not well documented – in many cases, be prepared to look at the underlying Java code
Machine Learning

- Classification, clustering, recommendation
  - K-NN, K-Means, Canopy methods, Decision Trees
  - Naïve Bayes, CNB, Logistic Regression, Support Vector Machines, Latent Dirichlet Allocation
  - Full featured math libraries (Colt) to ease many tasks
    - Singular Value Decomposition
- Can be used with or without Hadoop
  - Some algorithms are parallel, others are not
  - Most scale well. Most work well. Some serious refactoring is coming!
  - An augmented “roll your own” approach works well.
A brief social discussion:

- In the past, theory came first, then we sought evidence to support the theory.
- Today, we collect statistical “evidence” and then derive a theory to match it.
  - Correlation versus causality
  - Impact on privacy
- Can statistical machine learning supplant theory?
- Is this a major or minor paradigm shift, or is it just more hype?
  - How do we prevent getting caught in this trap?
Big Data and Graph Theory

- Graph theory uses vertices and edges to represent knowledge
  - Social networking (and other disciplines) have brought graph theory back to the forefront
  - Many other aspects of real-life are well modeled by graphs
- Giraph is an effort to utilize Hadoop to conduct sophisticated graph search algorithms
  - Open Source version of Pregel
  - Uses Bulk Synchronous Parallel methods
  - Efficiency issues exist
- HAMA is now full Apache project – uses BSP methods, but much more general and extendable
- Many scalable (big data) databases have emerged – Titan, Neo4J, and many more
Graph Theory in Cyber Security

- What kind of graphs are interesting (or suspicious)?
  - Clique – highly inter-related entities
  - N-partite – sets of related entities
  - Centrality in a graph
Other NoSQL Alternatives

• MongoDB – 10gen
  – A document (collection) oriented NoSQL database
  – Data is stored in BSON (Binary Serialized Object Notation)
  – Allows for multiple indices
  – No concurrency, but automatic updates on write if record exists
  – Written in C++
Other NoSQL Alternatives

• Cassandra – open source from Facebook
  – High availability (fault tolerant)
  – Incremental scalability
  – Eventually consistent (durable)
  – Tunable tradeoffs between consistency and latency
  – Minimal administration
  – No Single Point of Failure
  – No JOIN, No ORDER BY
Other NoSQL Alternatives

• CouchDB
  – Very “Web oriented”
  – Document database
    • Lucene text search integration
  – Strong Hadoop MapReduce integration
  – Heavily JSON and Javascript centric
  – Written in Erlang
Vendor Specific Products
Vendor Specific Products

- BigInsights – from IBM
  - Hadoop-based Analytics
  - Data Warehousing and ETL integration
  - Visualization and Discovery
  - Text Analytics
  - Systems management
  - Security
Vendor Specific Products

• Vertica – from HP
  – Hadoop integration through Cloudera
  – SQL and column oriented
  – ETL integration
Vendor Specific Products

• Greenplum – from EMC
  – Industrial strength product integration
  – Rich toolset
  – Role centric
  – Hadoop MapR
Vendor Specific Products

- **Teradata Aster**
  - Leverages Teradata massively parallel warehouse
    - Not really Hadoop, but newest version is tightly integrated
  - Proprietary SQL interface on a proprietary appliance
  - Deep analytics toolset
    - Pattern matching
    - Text mining
    - Clustering
    - Simulation
    - Spatial analysis
    - Linear algebra
  - Significant data mgmt.
    - Data adapters
    - Data transformers
  - Now integrating Hadoop as first-class citizen

Teradata Analytic Platform Solutions

October 7, 2013
Proprietary and Confidential
Microsoft

• Microsoft has partnered with Hortonworks
  – HDInsight, SQL Server 2012
  – Hadoop delivered as part of Azure application fabric and as a standalone Server 2012 subsystem
    • PolyBase and Parallel Data Warehouse
  – Azure will deliver C#, F#, Javascript, and other .NET languages (and libraries) very soon now as streaming languages
  – See also Isotope and Hybrid IT
Vendor Specific Products

• You might be asking yourself, “Where is ORACLE®?”
  – Strategic Exadata appliance
    • Exalytics in-memory
  – Integrating with Cloudera
  – Have created their own NoSQL database that is based on BerkelyDB open source database
    • Can also use Cloudera’s HBase version
  – Promoting the R language
Hadoop Impact
Hadoop Administration

• Clusters, services, instances, and roles
  – Running multiple clusters
  – HDFS namespaces
  – Common worker configurations
  – Heat maps and the Host Inspector

• Reliability, Availability, Serviceability
  – System wide restarts
  – Rolling restarts
  – High availability and auto-failover
    • Secondary namenode with ability to manually or automatically failover
    • NFS usage
  – Dealing with maintenance and upgrades
    • Parcels – download, distribute, activate

• Commissioning and decommissioning nodes
Hadoop Administration

• Hardware
  – Commoditized hardware does not mean “low end”
  – Ratio of parallelization to CPU cores
    – Original idea: 2.00 : 1
    – Current idea: 0.75 : 1 ?????
  • Should I virtualize?
    – Jobtracker and Namenode – this is okay
      » Dual homing is also recommended
    – Tasktrackers and Datanodes – probably not
    – Note: VMWare’s Serengeti project is adding VM aware extensions
  – Memory requirements
    • 64-bit OS with as much memory as you can afford
Hadoop Administration

• Hardware
  – How big can it get?
    • Yahoo! runs a 4000 node cluster, moving to 10,000 nodes after Hadoop 2.0 is implemented
      – 42,000 servers in 1,200 racks at 4 data centers!
    • Facebook now has a single 100 PB cluster in a single HDFS
      – Created AvatarNode for real-time reliability

  – Note: Yahoo! Is making a strategic investment that is being called “science at scale”
    • The idea is to create an infrastructure so large and fast that real-time predictive analytics are achievable
Hadoop Administration

• Hardware
  - “Rack aware” planning
  - I/O capacity is critical
    • Should I use NFS?
      - Probably not, but look at MapR and how they are using it
    • Should I use a SAN?
      - Must consider the switching fabric – too small pipes will not yield expected results
      - Some next generation solutions look very good
        » Arista, Mellanox
      - 1 Gb will run out of gas in large cluster
      - 10 Gb is much better
Hadoop Administration
Hadoop Administration

• Hardware
  – External vendor options
    • Amazon Elastic Computing Cloud (EC2)
    • Rackspace, Heroku
    • Microsoft Azure, Isotope, and Hybrid IT
    • IBM, EMC, HP, Dell
    • Infochimps, Mortar Data, Continuity...
Hadoop Administration

• Are Big Data solutions using Hadoop Cloud-ready? – *Emphatically Yes!*
  – Used for Development pooling and sandboxing
  – Used for testing – but beware of load testing!
  – Used for overflow and peak to average usage surges.
Hadoop Administration

• Software
  – Version planning is critical
    • Not all versions of all components work together
    • New versions break existing software with high frequency
    • Plan out your upgrade roadmap – feature lists exist to help you do this
    • Look at Cloudera, MapR, and Hortonworks and their roadmaps
  – The “safe” configuration
    • Hadoop 0.20.205 and HBase 0.90
    • Old versus new API
  – The next best safe configuration
    • Hadoop 1.0 and HBase 0.94
  – If you are really brave
    • Hadoop 2.0 and HBase 0.94.1
  – Developing regression testing suites are highly recommended
Hadoop Administration

- Development
  - Skill sets
    - Java development skills are essential – there is really no way around this at this time
    - Training – get your initial team trained – development and administration
  - Eclipse is the primary development tool
    - Several add-ins/plug-ins aid in development, but many DO NOT WORK!
    - Hadoop project types
    - HDFS and HBase browsers
    - KarmaSphere development suites show great promise, but they are having difficulty keeping up with versions
  - Consider using frameworks
    - Cascading is conceptually much easier to use
    - Pig is also very good, but its level of abstraction may be “too high”
    - Hive can be very useful for data scientists and ad hoc queries, but be wary of developing a production dependency – also now must look at Impala
  - Plan for production job management
    - Oozie is a nightmare!!!!
Hadoop Administration

• Algorithms really matter
  – Data partitioning strategies and key construction
  – Mapper, Partitioner, Combiner, and Reducer complexities
    • Input and Output formatters, Multi-reducers, etc.
  – Even simple aggregation tasks can be pretty complex
  – Code bloat – can create very complex stockpiles of code libraries very quickly
  – Don’t reinvent the wheel – do some research on what is already available – GitHub, PiggyBank, etc.
Hadoop Administration

• Testing matters
  – Systems such as this are inherently difficult to test
    • Large complex job streams
    • Test data and results
    • Load testing, Failover testing, system stability
  – Tools have not yet caught up. There are log files everywhere!
    • KarmaSphere shows great promise
    • Other Eclipse add-ins are very substandard
    • Cloudera, Hortonworks, MapR etc. are closing this gap quickly

– I love sed, awk and grep, really I do!
  • It’s like it’s 1975 all over again

October 7, 2013
Proprietary and Confidential
Hadoop Administration

• Plan for systems management
  – All cluster members must be running the same software
    • OS, Java, and Hadoop components
    • Look at:
      – Big Top Apache packaging
      – Puppet
      – Chef
      – Dell Crowbar
      – Cloudera Distribution platform
        » Note also MapR, Greenplum, Hortonworks, IBM, Dell, and the list goes on and on....

  – Nodes will fail, disks will fail, and capacity will be realigned
    • Must plan for these routine events
Hadoop Administration

• Plan for systems management
  – System monitoring is critical
    • Hung components
    • Failover of Namenode is now handled by newer releases (Hadoop 1.0 and 2.0)
    • Linux HA and Hadoop HA
Hadoop Administration

• How do I plan for disaster recovery?
  – Some say simply, “You don’t.” - **WRONG ANSWER!**

• So, what are the options?
  – Use built in redundancy
  – Archive and rebuild – what goes in, and what comes out
  – DistCp and replication
  – SAN based storage
  – Redundant hot (or warm) site
  – **Cloudera Navigator and Cloudera Backup and Disaster Recovery (BDR)**

• And don’t forget software upgrades...
  – New Cloudera Manager 4.5 allows rolling upgrades.
Hadoop Administration

• As soon as you deploy,
  – No longer “knowable” workloads. Mixed workloads are introduced by user community and often cannot be analyzed until they are running.
  – Typical to see “word-of-mouth” sharing of resources
  – Hadoop has over 175 tunable parameters. Tuning is often “gut feel”. And is also highly workload dependent.
  – Train systems administrators
  – Integrate with existing monitors
Questions and Answers

Michael.Covert@AnalyticsInside.us
http://www.AnalyticsInside.us