Introduction to Analytics and Big Data - Hadoop

Thomas Rivera
Hitachi Data Systems
SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Data Challenges
Why Hadoop
Customer Challenges: The Data Deluge

IN 2010 THE DIGITAL UNIVERSE WAS 1.2 ZETTABYTES

IN A DECADE THE DIGITAL UNIVERSE WILL BE 35 ZETTABYTES

90% OF THE DIGITAL UNIVERSE IS UNSTRUCTURED

IN 2011 THE DIGITAL UNIVERSE IS 300 QUADRILLION FILES

The Economist, Feb 25, 2010
Big Data: Different than Business Intelligence

“BIG DATA ANALYTICS”
- Experimental, Ad Hoc
- Mostly Semi-Structured
- External + Operational
- 10s of TB to 100’s of PB’s

“TRADITIONAL BI”
- Repetitive
- Structured
- Operational
- GBs to 10s of TBs

© 2014 Storage Networking Industry Association. All Rights Reserved.
Questions from Businesses will Vary

Past

What happened?
Reporting, Dashboards
Why did it happen?
Forensics & Data Mining

What is happening?
Real-Time Analytics

Real-Time Data Mining

What is likely to happen?
Predictive Analytics

Prescriptive Analytics

Future
Web 2.0 is “Data-Driven”

“The future is here, it’s just not evenly distributed yet.”
William Gibson
The World of Data-Driven Applications

Introduction to Analytics and Big Data - Hadoop

© 2014 Storage Networking Industry Association. All Rights Reserved.
Attributes of Big Data

Volume:
- Terabytes
- Transactions
- Tables
- Records
- Files

Velocity:
- Batch
- Near Time
- Real Time
- Streams

Variety:
- Structured
- Unstructured
- Semistructured

Introduction to Analytics and Big Data - Hadoop
© 2014 Storage Networking Industry Association. All Rights Reserved.
Ten Common Big Data Problems

1. Modeling true risk
2. Customer churn analysis
3. Recommendation engine
4. Ad targeting
5. PoS transaction analysis
6. Analyzing network data to predict failure
7. Threat analysis
8. Trade surveillance
9. Search quality
10. Data “sandbox”
# The Big Data Opportunity

<table>
<thead>
<tr>
<th>Financial Services</th>
<th>Healthcare</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Financial Services" /></td>
<td><img src="image2" alt="Healthcare" /></td>
</tr>
<tr>
<td>Retail</td>
<td>Web/Social/Mobile</td>
</tr>
<tr>
<td><img src="image3" alt="Retail" /></td>
<td><img src="image4" alt="Web/Social/Mobile" /></td>
</tr>
<tr>
<td>Manufacturing</td>
<td>Government</td>
</tr>
<tr>
<td><img src="image5" alt="Manufacturing" /></td>
<td><img src="image6" alt="Government" /></td>
</tr>
</tbody>
</table>
Industries Are Embracing Big Data

Retail
- CRM – Customer Scoring
- Store Siting and Layout
- Fraud Detection / Prevention
- Supply Chain Optimization

Advertising & Public Relations
- Demand Signaling
- Ad Targeting
- Sentiment Analysis
- Customer Acquisition

Financial Services
- Algorithmic Trading
- Risk Analysis
- Fraud Detection
- Portfolio Analysis

Media & Telecommunications
- Network Optimization
- Customer Scoring
- Churn Prevention
- Fraud Prevention

Manufacturing
- Product Research
- Engineering Analytics
- Process & Quality Analysis
- Distribution Optimization

Energy
- Smart Grid
- Exploration

Government
- Market Governance
- Counter-Terrorism
- Econometrics
- Health Informatics

Healthcare & Life Sciences
- Pharmaco-Genomics
- Bio-Informatics
- Pharmaceutical Research
- Clinical Outcomes Research
Why Hadoop?

Big Data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to address business and technology trends that are disrupting traditional data management and processing.

Enterprises can gain a competitive advantage by being early adopters of big data analytics.
## Storage & Memory B/W lagging CPU

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>DRAM</th>
<th>LAN</th>
<th>Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual bandwidth improvement (all milestones)</td>
<td>1.5</td>
<td>1.27</td>
<td>1.39</td>
<td>1.28</td>
</tr>
<tr>
<td>Annual latency improvement (all milestones)</td>
<td>1.17</td>
<td>1.07</td>
<td>1.12</td>
<td>1.11</td>
</tr>
</tbody>
</table>

- CPU B/W requirements out-pacing memory and storage
- Disk & memory getting “further” away from CPU
- Large sequential transfers better for both memory & disk

Introduction to Analytics and Big Data - Hadoop
© 2014 Storage Networking Industry Association. All Rights Reserved.
Commodity Hardware Economics

For $1000
One computer can

Process
~32GB

Store
~15TB

99.9%
Of data is Underutilized
Enterprise + Big Data = Big Opportunity
Hadoop Adoption
HDFS
MapReduce
Ecosystem Projects
Hadoop Adoption in the Industry

Source: Hadoop Summit Presentations
What is Hadoop?

- A scalable fault-tolerant distributed system for data storage and processing
- Core Hadoop has two main components
  - Hadoop Distributed File System (HDFS): self-healing, high-bandwidth clustered storage
    - Reliable, redundant, distributed file system optimized for large files
  - MapReduce: fault-tolerant distributed processing
    - Programming model for processing sets of data
    - Mapping inputs to outputs and reducing the output of multiple Mappers to one (or a few) answer(s)
- Operates on unstructured and structured data
- A large and active ecosystem
- Open source under the friendly Apache License
The Data Set System
HDFS Concepts

- Sits on top of a native (ext3, xfs, etc..) file system
- Performs best with a ‘modest’ number of large files
- Files in HDFS are ‘write once’
- HDFS is optimized for large, streaming reads of files
- Hadoop Distributed File System
  - Data is organized into files & directories
  - Files are divided into blocks, distributed across cluster nodes
  - Block placement known at runtime by map-reduce = computation co-located with data
  - Blocks replicated to handle failure
  - Checksums used to ensure data integrity

- Replication: one and only strategy for error handling, recovery and fault tolerance
  - Self Healing
  - Make multiple copies
HDFS File Write Operation

Introduction to Analytics and Big Data - Hadoop
© 2014 Storage Networking Industry Association. All Rights Reserved.
HDFS File Read Operation

1. Open
2. Get block locations
3. Read
4. Read
5. Read
6. Close
Functional Programming meets Distributed Processing
What is MapReduce?

- A method for distributing a task across multiple nodes
- Each node processes data stored on that node
- Consists of two developer-created phases
  1. Map
  2. Reduce
- In between Map and Reduce is the Shuffle and Sort
MapReduce Provides

- Automatic parallelization and distribution
- Fault Tolerance
- Status and Monitoring Tools
- A clean abstraction for programmers
- [Google Technology RoundTable: MapReduce](#)
Key MapReduce Terminology Concepts

- A user runs a client program on a client computer
- The client program submits a job to Hadoop
- The job is sent to the JobTracker process on the Master Node
- Each Slave Node runs a process called the TaskTracker
- The JobTracker instructs TaskTrackers to run and monitor tasks
- A task attempt is an instance of a task running on a slave node
- There will be at least as many task attempts as there are tasks which need to be performed
MapReduce: Basic Concepts

- Each Mapper processes single input split from HDFS
- Hadoop passes developer’s Map code one record at a time
- Each record has a key and a value
- Intermediate data written by the Mapper to local disk
- During shuffle and sort phase, all values associated with same intermediate key are transferred to same Reducer
- Reducer is passed each key and a list of all its values
- Output from Reducers is written to HDFS
MapReduce Operation

What was the max/min temperature for the last century?
The requirement:

- you need to find out grouped by type of customer how many of each type are in each country with the name of the country listed in the `countries.dat` in the final result (and not the 2 digit country name). Each record has a key and a value.

To do this you need to:

- Join the data sets
- Key on country
- Count type of customer per country
- Output the results
MapReduce Paradigm

Input Map Shuffle and Sort Reduce Output

cat grep sort uniq output
**Problem**: Count the number of times that each word appears in the following paragraph:

John has a red car, which has no radio. Mary has a red bicycle. Bill has no car or bicycle.

<table>
<thead>
<tr>
<th>Server 1: John has a red car, which has no radio.</th>
</tr>
</thead>
<tbody>
<tr>
<td>John: 1</td>
</tr>
<tr>
<td>has: 2</td>
</tr>
<tr>
<td>a: 1</td>
</tr>
<tr>
<td>red: 1</td>
</tr>
<tr>
<td>car: 1</td>
</tr>
<tr>
<td>which: 1</td>
</tr>
<tr>
<td>no: 1</td>
</tr>
<tr>
<td>radio: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 2: Mary has a red bicycle.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary: 1</td>
</tr>
<tr>
<td>has: 1</td>
</tr>
<tr>
<td>a: 1</td>
</tr>
<tr>
<td>red: 1</td>
</tr>
<tr>
<td>bicycle: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 3: Bill has no car or bicycle.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill: 1</td>
</tr>
<tr>
<td>has: 1</td>
</tr>
<tr>
<td>no: 1</td>
</tr>
<tr>
<td>car: 1</td>
</tr>
<tr>
<td>or: 1</td>
</tr>
<tr>
<td>bicycle: 1</td>
</tr>
</tbody>
</table>

**Map**

<table>
<thead>
<tr>
<th>Server 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>John: 1</td>
</tr>
<tr>
<td>has: 2</td>
</tr>
<tr>
<td>a: 1</td>
</tr>
<tr>
<td>red: 1</td>
</tr>
<tr>
<td>car: 1</td>
</tr>
<tr>
<td>which: 1</td>
</tr>
<tr>
<td>no: 1</td>
</tr>
<tr>
<td>radio: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>car: 1</td>
</tr>
<tr>
<td>bicycle: 1</td>
</tr>
<tr>
<td>Bill: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill: 1</td>
</tr>
<tr>
<td>car: 2</td>
</tr>
<tr>
<td>a: 2</td>
</tr>
<tr>
<td>no: 2</td>
</tr>
<tr>
<td>red: 2</td>
</tr>
<tr>
<td>radio: 1</td>
</tr>
<tr>
<td>Mary: 1</td>
</tr>
</tbody>
</table>

**Reduce**

<table>
<thead>
<tr>
<th>Server 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>John: 1</td>
</tr>
<tr>
<td>has: 4</td>
</tr>
<tr>
<td>a: 2</td>
</tr>
<tr>
<td>no: 2</td>
</tr>
<tr>
<td>red: 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>car: 2</td>
</tr>
<tr>
<td>Bill: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Server 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>bicycle: 2</td>
</tr>
<tr>
<td>Bill: 1</td>
</tr>
</tbody>
</table>

---

Introduction to Analytics and Big Data - Hadoop
© 2014 Storage Networking Industry Association. All Rights Reserved.
Putting it all Together: MapReduce and HDFS

Client/Dev

Map Job
Reduce Job

Large Data Set
(Log files, Sensor Data)

Hadoop Distributed File System (HDFS)

Task Tracker

Map Job
Reduce Job

Task Tracker

Map Job
Reduce Job

Task Tracker

Map Job
Reduce Job

Job Tracker

Map Job
Reduce Job
Hadoop Ecosystem Projects

• Hadoop is a ‘top-level’ Apache project
  • Created and managed under the auspices of the Apache Software Foundation

• Several other projects exist that rely on some or all of Hadoop
  • Typically either both HDFS and MapReduce, or just HDFS

• Ecosystem Projects Include
  • Hive
  • Pig
  • HBase
  • Many more…..
## Hadoop, SQL & MPP Systems

<table>
<thead>
<tr>
<th>Hadoop</th>
<th>Traditional SQL Systems</th>
<th>MPP Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale-Out</td>
<td>Scale-Up</td>
<td>Scale-Out</td>
</tr>
<tr>
<td>Key/Value Pairs</td>
<td>Relational Tables</td>
<td>Relational Tables</td>
</tr>
<tr>
<td>Functional Programming</td>
<td>Declarative Queries</td>
<td>Declarative Queries</td>
</tr>
<tr>
<td>Offline Batch Processing</td>
<td>Online Transactions</td>
<td>Online Transactions</td>
</tr>
</tbody>
</table>
## Comparing RDBMS and MapReduce

<table>
<thead>
<tr>
<th></th>
<th>Traditional RDBMS</th>
<th>MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Size</strong></td>
<td>Gigabytes <em>(Terabytes)</em></td>
<td>Petabytes <em>(Exabytes)</em></td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Interactive and Batch</td>
<td>Batch</td>
</tr>
<tr>
<td><strong>Updates</strong></td>
<td>Read / Write many times</td>
<td>Write once, Read many times</td>
</tr>
<tr>
<td><strong>Structure</strong></td>
<td>Static Schema</td>
<td>Dynamic Schema</td>
</tr>
<tr>
<td><strong>Integrity</strong></td>
<td>High (ACID)</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Scaling</strong></td>
<td>Nonlinear</td>
<td>Linear</td>
</tr>
<tr>
<td><strong>DBA Ratio</strong></td>
<td>1:40</td>
<td>1:3000</td>
</tr>
</tbody>
</table>

*Reference: Tom White’s Hadoop: The Definitive Guide*
Diagnostics and Customer Churn

Issues
- What make and model systems are deployed?
- Are certain set top boxes in need of replacement based on system diagnostic data?
- Is there a correlation between make, model or vintage of set top box and customer churn?
- What are the most expensive boxes to maintain?
- Which systems should we proactively replace to keep customers happy?

Big Data Solution
- Collect unstructured data from set top boxes—multiple terabytes
- Analyze system data in Hadoop in near real time
- Pull data into Hive for interactive query and modeling
- Analytics with Hadoop increases customer satisfaction
Pay Per View Advertising

Issues

- Fixed inventory of ad space is provided by national content providers. For example, 100 ads offered to provider for 1 month of programming
- Provider can use this space to advertise its products and services, such as pay per view
- Do we advertise “The Longest Yard” in the middle of a football game or in the middle of a romantic comedy?
- 10% increase in pay per view movie rentals = $10M in incremental revenue

Big Data Solution

- Collect programming data and viewer rental data in a large data repository
- Develop models to correlate proclivity to rent to programming format
- Find the most productive time slots and programs to advertise pay per view inventory
- Improve ad placement and pay-per-view conversion with Hadoop
Risk Modeling

- Bank had customer data across multiple lines of business and needed to develop a better risk picture of its customers. i.e., if direct deposits stop coming into checking acct, it’s likely that customer lost his/her job, which impacts creditworthiness for other products (CC, mortgage, etc.)
- Data existing in silos across multiple LOB’s and acquired bank systems
- Data size approached 1 petabyte

Why do this in Hadoop?
- Ability to cost-effectively integrate + 1 PB of data from multiple data sources: data warehouse, call center, chat and email
- Platform for more analysis with poly-structured data sources; i.e., combining bank data with credit bureau data; Twitter, etc.
- Offload intensive computation from DW
Sentiment Analysis

- Sentiment Analysis
  - Hadoop used frequently to monitor what customers think of company’s products or services
  - Data loaded from social media sources (Twitter, blogs, Facebook, emails, chats, etc.) into Hadoop cluster
  - Map/Reduce jobs run continuously to identify sentiment (i.e., Acme Company’s rates are “outrageous” or “rip off”)
  - Negative/positive comments can be acted upon (special offer, coupon, etc.)

- Why Hadoop
  - Social media/web data is unstructured
  - Amount of data is immense
  - New data sources arise weekly
Resources: The Big Data Conversation

- World Economic Forum: “Personal Data: The Emergence of a New Asset Class” 2011
- McKinsey Global Institute: Big Data: The next frontier for innovation, competition, and productivity
- Big Data: Harnessing a game-changing asset
- IDC: 2011 Digital Universe Study: Extracting Value from Chaos
- The Economist: Data, Data Everywhere
- Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field
- O’Reilly – What is Data Science?
- O’Reilly – Building Data Science Teams?
- O’Reilly – Data for the public good
- Obama Administration “Big Data Research and Development Initiative.”
The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial.

Authorship History
Original Author: Rob Peglar, September 2012
Updates: Thomas Rivera, March 2014

Additional Contributors
Denis Guyadeen
Thomas Rivera

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org