Sentiment Analysis using Hadoop

Sponsored By Atlink Communications Inc
Capstone Project Group 1

Instructor : Dr.Sadegh Davari

Mentors : Dilhar De Silva, Rishita Khalathkar

Index

- Introduction
- Phase 1
  - Implementation of project
  - Demo
- Phase 2
  - Big data as a problem - Hadoop
  - Implementations of Hadoop
  - Infrastructure, Cost & Security in implementing Hadoop
  - Future scope in Hadoop
- Task Division
What is Sentiment Analysis?

- Sentiment analysis is the detection of attitudes
- Sentiments of any product can be Positive, Negative or Neutral

Overview of our Project

- Data Collection
- Data Feeding
- Data Manipulation
- Data Visualization
Hive and HiveQL

Hive: Data-warehouse infrastructure || HiveQL: Query language

Querying
- Analyze table
- Describe column
- Describe database
- Export table
- Import table

Managing
- Load data
- Show table extended
- Show indexes
- Show columns

Summarization

Analysis
Hortonworks and Microsoft have partnered to bring the benefits of Apache Hadoop to Windows

**HDP Sandbox**
Runs on VirtualBox or VMWare:

- HDP 2.0 with Red Hat 64 bit
- Host Operating Systems: Windows 8.1 of 64-bit machine
- Virtual Machine: Virtual Box
- 8 GB of RAM, with 4 GB to HDP
- Number of Processors: 2
Project Demo

- Slide containing Demo Video. Video Not attached due to high play Lags
Big Data as a Problem
Hadoop as a Solution
Big Data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. The diagram illustrates the four V's of big data:

- Volume: The sheer amount of data that needs to be processed.
- Velocity: The speed at which data is generated, shared, and accessed.
- Variety: The diversity of data types and sources.
- Veracity: The quality and accuracy of the data.

These concepts highlight the challenges and opportunities in handling big data.
Challenges of Big Data

1. Uncertainty of the Data Management Landscape
2. The Big Data Talent Gap
3. Getting Data into the Big Data Platform
4. Getting Useful Information out of the Big Data Platform
5. Synchronization across the Data Sources
Uncertainty of the Data Management Landscape

Challenge

Traditional RDBMS are challenged to CAPTURE, STORE, SEARCH, SHARE, ANALYZE and VISUALIZE data.

Cannot Handle 3 Vs

Solution

NoSql is Non-relational
Distributed database
Horizontally scaled out
Schema Free

Handles 3 V’s

NoSql + Hadoop helps to overcome Data Management Challenges
The Big Data Talent Gap

- Scarcity of experts
  - Experts are trained, not born

- Tools to transfer RDBMS to NoSql are available
  - ETL (Extract, transfer and Load) utilities are used to transfer RDBMS to NoSql

Image Courtesy: http://www.codeproject.com/Articles/279947/Migration-of-Relational-Data-structure-to-Cassandra
Data integration technologies

Getting Data into the Big Data Platform
- Accessing data stored in a variety of standard configurations (including XML, JSON, and BSON objects)

Synchronization across the Data Sources
- Employ push-down capabilities of a wide variety of data management systems (ranging from conventional RDBMS data stores to newer NoSQL approaches) to optimize data access

Getting Useful Information out of the Big Data Platform
- Relying on standard relational data access methods (such as ODBC/JDBC)
Hadoop and its Architecture

- Open-source Software Framework used to run Big Data Applications
- Hadoop framework supports technologies that help in solving Big Data challenges
Hadoop Components

Implementations of Hadoop, Technical Roles available & Skills Required
Distribution of Apache Hadoop

Hadoop distribution is concerned, the three companies that really stand out in the completion are [6]:

- Cloudera
- MapR
- Hortonworks
Comparison of Cloudera, Hortonworks and MapR

The comparison is made on the following [6]:

- Performance and Scalability
- Dependability
- Manageability
- Data Access
# Performance and Scalability

<table>
<thead>
<tr>
<th></th>
<th>Hortonworks</th>
<th>Cloudera</th>
<th>MapR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Ingest</td>
<td>Batch</td>
<td>Batch</td>
<td>Batch and streaming writes</td>
</tr>
<tr>
<td>Metadata Architecture</td>
<td>Centralized</td>
<td>Centralized</td>
<td>Distributed</td>
</tr>
<tr>
<td>HBase Performance</td>
<td>Latency spikes</td>
<td>Latency spikes</td>
<td>Consistent low latency</td>
</tr>
<tr>
<td>NoSQL Applications</td>
<td>Mainly batch</td>
<td>Mainly batch</td>
<td>Batch and online/real-time</td>
</tr>
<tr>
<td></td>
<td>application</td>
<td>application</td>
<td>applications</td>
</tr>
</tbody>
</table>
## Dependability

<table>
<thead>
<tr>
<th></th>
<th>Hortonworks</th>
<th>Cloudera</th>
<th>MapR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Availability</td>
<td>Single failure recovery</td>
<td>Single failure recovery</td>
<td>Self healing across multiple failures</td>
</tr>
<tr>
<td>MapReduce Job Tracker</td>
<td>Restart jobs</td>
<td>Restart jobs</td>
<td>Continuous without restart</td>
</tr>
<tr>
<td>Upgrading</td>
<td>Planned downtime</td>
<td>Rolling upgrades</td>
<td>Rolling updates</td>
</tr>
<tr>
<td>Replication</td>
<td>Data</td>
<td>Data</td>
<td>Data + metadata</td>
</tr>
<tr>
<td>Snapshots</td>
<td>Consistent only for closed files</td>
<td>Consistent only for closed files</td>
<td>Point-in-time consistency for all files and tables</td>
</tr>
<tr>
<td>Disaster Recovery</td>
<td>No</td>
<td>File copy Scheduling</td>
<td>Mirroring</td>
</tr>
</tbody>
</table>
## Manageability

<table>
<thead>
<tr>
<th></th>
<th>Hortonworks</th>
<th>Cloudera</th>
<th>MapR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management Tools</td>
<td>Amabari</td>
<td>Cloudera Manager</td>
<td>MapR control system</td>
</tr>
<tr>
<td>Volume Support</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Alarms, Alerts</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Integration with REST API</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data and Job Placement Control</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Hortonworks</td>
<td>Cloudera</td>
<td>MapR</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------</td>
<td>----------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>File System Access</td>
<td>HDFS, read-only NFS</td>
<td>HDFS, read-only NFS</td>
<td>HDFS, read/write NFS</td>
</tr>
<tr>
<td>File I/O</td>
<td>Append only</td>
<td>Append only</td>
<td>Read/write</td>
</tr>
<tr>
<td>Security: ACLs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Wire-level Authentication</td>
<td>Kerberos</td>
<td>Kerberos</td>
<td>Kerberos, Native</td>
</tr>
</tbody>
</table>
# Technical Roles Available in Hadoop

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Job Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop Developer</td>
<td>Develop MapReduce jobs, designs data warehouses</td>
</tr>
<tr>
<td>Hadoop Admin</td>
<td>Manages Hadoop cluster, designs data pipelines</td>
</tr>
<tr>
<td>Data Scientists</td>
<td>Data mining and figuring out hidden knowledge in data</td>
</tr>
<tr>
<td>Business Analyst</td>
<td>Analyzes data</td>
</tr>
</tbody>
</table>
Technical Skill Required for Hadoop

No SQL Skills
- Experience in handling Unstructured Data
- Ex: Cassandra, MongoDB

Map Reduce Skills
- Java programming skills
- Unix scripting skills
- Understanding of concepts of Hadoop/Map Reduce
- Data modelling skills
- Good analytical skills
Technical Skill Required for Hadoop

Linux Skills
- Unix commands & Unix based file systems, Perl /Bash Scripting
- Knowledge on Memory, OS storage, CPU, Networking
- Good Troubleshooting Skills
- Good understanding of system’s capacity and bottlenecks

Hadoop Training
- All this companies provide hadoop training in specific domain:
  - Cloudera
  - MapR
  - Hortonworks
Infrastructure, Cost and Security
Implementing Hadoop
Infrastructure

**Hadoop Installation Type**
- Deliver New Business Value
- Deliver Data Center Efficiency

**Data Processing Type**
- Continuously
- Frequently
- Historically
- Regulatorily

**Types of Data**
- Clickstream
- Sentiment
- Sensor/Machine
- Geo Tracking
- Server Logs
- Text

**Data Growing Type**
- Exponentially
- Regularly
Following is the hardware infrastructure to **Deliver New Business Value with Geo Tracking Data**, processed **Historically** at a **Regular Growth Rate** [8]

<table>
<thead>
<tr>
<th>Hardware estimate</th>
<th>Infrastructure estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master Nodes: 1</td>
<td>Floor Space: 7 ft²</td>
</tr>
<tr>
<td>Slave Nodes: 7</td>
<td>Cooling: 32800 BTU/Mo</td>
</tr>
<tr>
<td>Racks: 1</td>
<td>Power: 12 Kwh/Mo</td>
</tr>
<tr>
<td>Rows: 0</td>
<td>Network Bandwidth: 15 Mb/s</td>
</tr>
</tbody>
</table>

Image courtesy: http://hortonworks.com/cluster-sizing-guide/
Spectrum Of Hadoop Deployment Options

- **On-premise full custom**
  - Full control of the hadoop cluster

- **Hadoop appliance**
  - Preconfigured hadoop cluster to use

- **Hadoop hosting**
  - Rely on a service provider to deploy and operate

- **Hadoop-as-a-Service**
  - Access to hadoop clusters with a pay-per-use model

## Six Key Areas for Cloud Based Vs On Premise Hadoop

<table>
<thead>
<tr>
<th>Area</th>
<th>Cloud Based Hadoop</th>
<th>On Premise Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price Performance Ratio</td>
<td>More Price Efficient, But Slow</td>
<td>Less Price Efficient, But Fast</td>
</tr>
<tr>
<td>Data Enrichment</td>
<td>Multiple Datasets, Slow Connectivity</td>
<td>Multiple Datasets, Same Connectivity</td>
</tr>
<tr>
<td>Data Gravity</td>
<td>Easy Data Saturation and Migration Control</td>
<td>Costly and Time Consuming Data Handling</td>
</tr>
<tr>
<td>Enhance Team Collaboration and Productivity</td>
<td>More Collaboration More Productivity</td>
<td>Less Collaboration Less Productivity</td>
</tr>
<tr>
<td>Productivity Of Developers And Data Scientists</td>
<td>Partial or No Control</td>
<td>Easy and Full Control</td>
</tr>
<tr>
<td>Data Privacy</td>
<td>Offsite Data, Unreliable Security</td>
<td>Onsite Data, Reliable Security</td>
</tr>
</tbody>
</table>
Total Cost of Ownership (TCO) Comparison

<table>
<thead>
<tr>
<th>Bare-metal</th>
<th>Monthly TCO</th>
<th>Hadoop-as-a-Service</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$21,845.04</td>
<td></td>
</tr>
<tr>
<td>Staff for operation</td>
<td>$9,274.46</td>
<td>$3,091.49</td>
</tr>
<tr>
<td>Technical support</td>
<td>$6,656.00</td>
<td>$1,372.27</td>
</tr>
<tr>
<td>(third-party vendors)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data center facility</td>
<td>$2,914.58</td>
<td>$2,063.00</td>
</tr>
<tr>
<td>and electricity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Server hardware</td>
<td>$3,000.00</td>
<td>$15,318.28</td>
</tr>
</tbody>
</table>

Bare-metal Hadoop Cluster

- The cluster size with 24 nodes and 50 TB of HDFS capacity
- Hardware cost at $4,500 per node based on retail server hardware vendors.
- Server node assumes four 2 TB hard disk drives, 24 GB memory, and 12 CPU cores.

Hadoop-as-a-Service (Amazon EMR)

- $3,091.49 for cloud-related internal operation staff personnel cost, which is one-third of its bare-metal counterpart
- Using a service provider shifts a large portion of operational burden is reduced
- Large scale distribution and sharing service lowers service rates

Top Hadoop-as-a-Service (HDaaS) Providers...

TechNavio pinpointed top 16 companies offering hadoop-as-a-service that are expected increase Compound annual growth rate (CAGR) of **84.81 percent from 2014-2019**[12]

1. Amazon Web Services
2. EMC²
3. IBM
4. Microsoft
5. Altiscale
6. Cask Data
7. Cloudera
8. FICO
9. Google
10. Hortonworks
11. HP
12. Infochimps
13. MapR Technologies
14. Datadog (Mortar Data)
15. Pentaho
16. Teradata

Pricing Criteria to the Customers

Hourly or Monthly/Yearly Charges

- Microsoft, Amazon Web Services, HP Cloud

Memory, CPU, Shared Core, Hosting Nation

- Google

Supporting and Maintenance Charges

- HortonWorks
Future Scope With Hadoop
Key Points

- Which scripting language is best to use?
- Future with NoSQL.
- Importance of Hadoop Tester in future.
- Data visualization.

Most of the scripting languages like php, python, perl, ruby, pig latin, bash is good.

Pig is a high level scripting language that is used with Apache Hadoop.

It is a data flow language that translate scripts into MapReduce jobs and then executes those jobs.

Sql like language, no need to know java.

NO-SQL Databases

- It helps massive parallel processing engines developed to analyze relational data over the years are very fast.
- RDBMS fails while working and handling millions of rows.

Image courtesy: https://hadoop4korea.files.wordpress.com/2012/04/kb_nosql.png
• **Clover**\(^{[15]}\) provides the metrics you need to better balance the effort between writing code that does stuff, and code that tests stuff.

• Clover runs in your IDE or your continuous integration system, and includes test optimization to make your tests run faster, and fail more quickly.

• **JUnit**\(^{[16]}\) is a simple framework to write repeatable tests.
Data Visualization

- It is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly.
  - Interactively explore billions of rows of data in seconds
  - Put fast, powerful analytics in the hands of all users

Companies market share in Big Data

2011 Big Data Pure-Play Market Share

- Vertica, 18%
- Opera Solutions, 16%
- Mu Sigma, 12%
- Aster Data, 11%
- Splunk, 10%
- Greenplum, 9%
- 1010data, 4%
- Clouder, 4%
- Calpont, 3%
- Fractal Analytics, 2%
- Think Big Analytics, 2%
- MapR, 2%
- Digital Reasoning, 1%
- Hortonworks, 1%
- DataStax, 1%
- RainStor, 1%
- HPCC Systems, 0.5%
- Karmasphere, 0.5%
- Other, 2%

Datameer, 1%

Total 2011 Big Data Pure-Play Revenue: $480 million

Image courtesy: http://www.slideshare.net/machinepulse/managing-your-assets-with-big-data-tools-45931405
## Task Division – Phase 1

<table>
<thead>
<tr>
<th>Task</th>
<th>Done By</th>
</tr>
</thead>
<tbody>
<tr>
<td>UML use case, Analysis and Sequence models</td>
<td>Srijha Reddy, Ankur Uprit, Kiranmayi Ganti, Pinaki Ghosh</td>
</tr>
<tr>
<td>Implementing HortonWorks Data Platform</td>
<td>Pinaki Ghosh</td>
</tr>
<tr>
<td>Implementing Twitter Json file and Dictionary</td>
<td>Kiranmayi Ganti</td>
</tr>
<tr>
<td>Implementing and Managing Hive tables</td>
<td>Ankur Uprit</td>
</tr>
<tr>
<td>Analysis Twitter data using Hive</td>
<td>Ankur Uprit, Pinaki Ghosh</td>
</tr>
<tr>
<td>Data Visualization using BI Tools</td>
<td>Kiranmayi Ganti, Srijha Reddy</td>
</tr>
<tr>
<td>Designing Website and Maintenance</td>
<td>Pinaki Ghosh</td>
</tr>
<tr>
<td>Unit &amp; System Testing</td>
<td>Ankur Uprit, Pinaki Ghosh, Kiranmayi Ganti, Srijha Reddy</td>
</tr>
<tr>
<td>Task</td>
<td>Done By</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Big Data as a Problem Hadoop as a Solution</td>
<td>Srijha Reddy Reddy Gangidi</td>
</tr>
<tr>
<td>Implementations of Hadoop, Technical Roles available &amp; Skills Required</td>
<td>Kiranmayi Ganti Ganti</td>
</tr>
<tr>
<td>Infrastructure, Cost and Security Implementing Hadoop</td>
<td>Pinaki Ghosh Ranjan Ghosh</td>
</tr>
<tr>
<td>Future Scope With Hadoop</td>
<td>Ankur Uprit Uprit</td>
</tr>
</tbody>
</table>
References

15. https://www.atlassian.com/software/clover/overview
Thank You for your attention!

Any Queries ???

Sentiment Analysis using Hadoop
Sponsored By Atlink Communications Inc.
Capstone Project Group 1