Hadoop: Big Data or Big Deal

Eduard Erwee
Introduction

- Eduard Erwee
- Data Soil Ltd (www.datasoil.uk)
- Background
  - Working with Microsoft data products over 20 years
  - MCSD VB6, SQL Server 7
  - 5 years as Microsoft Certified Trainer
  - 4 years as SQL Server PFE, Reading - UK
  - Today, clean data toilets for the highest bidder
  - No Linux / No Big Data (until 9 months ago)
Agenda

A) What is Big data?
   i) Origins
   ii) Technologies & Terminologies
   iii) The Players

B) How is Big Data Different?
   i) Philosophies

C) How to ride the Elephant?
   i) All about the tools
   ii) Sources of Inspiration

D) BI5 to the Future!
   i) Current Common Use-cases
   ii) Future Opportunities

E) Summary

F) Conclusion

G) Q&A
What is Big data?

i) Origins
- Nutch-to-Google-to-Yahoo and beyond
- Apache Who??

ii) Technologies & Terminologies
- Core Hadoop
- Hive
- HCatalog
- Pig
- Sqoop
- Oozie
- HUE (flavours-of)
- Mahout
- Loads of others
- Ha-dump!

iii) The Players
- The Big 3
- One to Watch : Cascading & Lingual
i) Origins

- Nutch-to-Google-to-Yahoo and beyond
- Apache Who??
Doug Cutting & Mike Cafarella starts working on Nutch (Open source web search engine based on Lucene and Java)

Google publishes GFS and MapReduce papers

Cutting adds DFS & MapReduce support to Nutch

Yahoo! hires Cutting, Hadoop spins off Nutch (named after Cutting’s Son’s Toy Elephant)

NY Times converts 4 TB of archives over 100 EC2’s

Web scale deployments at Y!, Facebook, Last.fm

Today Hadoop is -- Apache top-level project

April: Y! does fastest TB sort, 3.5 min over 910 nodes

May: Y! fastest TB sort, 62 seconds over 1460 nodes

May: Y! sorts PB, 16.25 hours over 3658 nodes

October: Yahoo Open-sources their Hadoop Production code
Apache Who??

The Apache Software Foundation

Community-led development since 1999.

- The Apache Software Foundation (http://www.apache.org/)
- The ASF is made up of nearly 150 Top Level Projects (Big Data and more)
  - Most of the Hadoop components we will discuss

All trademarks mentioned herein belong to their respective owners
ii) Technologies & Terminologies

- Core Hadoop
  - Hadoop Common:
  - Hadoop Distributed File System (HDFS™)
  - Hadoop MapReduce:
  - Hadoop YARN
- HUE (flavours-of)
- Hive
- HCatalog
- Pig
- Sqoop
- Oozie
- Mahout
- Loads of others
- Ha-dump!
Core Hadoop

- **Hadoop Common:**
  - The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):**
  - A distributed file system that provides high-throughput access to application data.
Core Hadoop

- Hadoop MapReduce

Take word counting as an example, something that Google does all of the time.
Hadoop MapReduce (continues):

- **MapReduce-V2**
  - A YARN-based system for parallel processing of large data sets.
  - Built on top of Tez

**Hadoop YARN (Yet Another Resource Negotiator):**

- A framework for job scheduling and cluster resource management.
HUE (flavours-of)

- Hue aggregates the most common Apache Hadoop components into a single UI.
- "Just use" Hadoop web based interface without worrying command line.
Hive

- Managing large datasets residing HDFS.
- Mechanism to query the data using a SQL-like language called HiveQL.
- Runs in HUE

```sql
select * from nyse_stocks 
where nyse_stocks.stock_symbol = 'IBM'
and date between '2001-12-01' and '2001-12-31'
order by date asc
```
HCatalog

- Built on top of the Hive metastore and incorporates Hive's DDL
- HCatalog’s table abstraction - presents relational view - of data in (HDFS)
- Removes worry about format their data is stored

- For me - Very similar to a set of views in SQL Server over staging feeds
- Exposed to Pig / Map Reduce / Hive
- Runs in HUE
HCatalog - Sample

File options
- Encoding: Unicode UTF8
- Delimiter: Tab (\t)
- Replace delimiter with: Tab (\t)
- Single line comment

Task preview

<table>
<thead>
<tr>
<th>Column name</th>
<th>Column name</th>
<th>Column name</th>
<th>Column name</th>
<th>Column name</th>
<th>Column name</th>
</tr>
</thead>
<tbody>
<tr>
<td>exchange</td>
<td>stock_symbol</td>
<td>date</td>
<td>stock_price_open</td>
<td>stock_price_high</td>
<td>stock_price_low</td>
</tr>
<tr>
<td>string</td>
<td>string</td>
<td>string</td>
<td>string</td>
<td>string</td>
<td>string</td>
</tr>
<tr>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
</tr>
<tr>
<td>12.55</td>
<td>12.5</td>
<td>12.8</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
</tr>
<tr>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
</tr>
<tr>
<td>12.59</td>
<td>12.59</td>
<td>12.5</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
</tr>
<tr>
<td>12.61</td>
<td>12.61</td>
<td>12.5</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
<td>NYSE</td>
<td>ASP</td>
</tr>
<tr>
<td>12.8</td>
<td>12.8</td>
<td>12.5</td>
<td>12.4</td>
<td>12.4</td>
<td>12.4</td>
</tr>
<tr>
<td><strong>timestamp</strong></td>
<td><strong>timestamp</strong></td>
<td>2001-12-19</td>
<td>12.42</td>
<td>12.35</td>
<td>12.4</td>
</tr>
<tr>
<td><strong>timestamp</strong></td>
<td><strong>timestamp</strong></td>
<td>2001-12-18</td>
<td>12.37</td>
<td>12.35</td>
<td>12.35</td>
</tr>
</tbody>
</table>
Pig

- Pig is a high-level platform used for creating MapReduce.
- The programming language is called Pig Latin
- Optimizer turns Pig into optimized Java Mapreduce.

```java
a = LOAD 'nyse_stocks' using org.apache.hcatalog.pig.HCatLoader();
b = filter a by stock_symbol == 'IBM';
c = group b all;
d = foreach c generate AVG(b.stock_price_open), AVG(b.stock_price_close);
dump d;
```

- Similar to M in Power Query
- It’s the VB.net Vs C++ debate all over again.
- Structure
  - Hive require data to be more structured
  - Pig allows you to work with unstructured data.
- Compatible with Hcatalog
- Runs in Hue
Apache Sqoop(TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.

- Runs in Hue
Oozie

- Workflow scheduler system to manage Apache Hadoop jobs.
- Oozie Coordinator jobs
  - Recurrent Oozie Workflow
  - Jobs triggered
    - by time (frequency)
    - data availability.
- Integrated with the rest of the Hadoop stack
- Scalable, reliable and extensible system.
- Available in HUE
Mahout

- Goal: scalable machine learning library.

- Examples of Mahout use cases:
  - Recommendation mining
    - takes users' behaviour and from that tries to find items users might like. (Netflix)
  - Clustering
    - Group documents, web pages and articles based on
      - contained topics
      - their related documents.
    - Most common use of this is search engines, which cluster pages based on keywords, page links, etc.
  - Classification
    - Based on prior categorization of documents
    - Evaluates new documents and determine best categories.
    - Filter new mail into INBOX
      - Auto-organize new content
      - flag potential spam comments.
Steaming pile of Data

Store

Inside the Elephant !?
iii) The Players

- The Big 3
- One to Watch: Cascading & Lingual
The Big 3

- Hortonworks claims to be the only fully open source distribution.
- Cloudera is close on their heals with everything based on open source but has some additional maintenance and installation functionality that is proprietary.
- MAP-R on the other hand re-wrote the storage engine from scratch to improve performance at the cost of being vendor specific.

My Opinion?

Benchmarking -- Altoros

Altoros did some significant benchmarking between the 3, and can be found here: [http://www.altoros.com/hadoop_benchmark.html](http://www.altoros.com/hadoop_benchmark.html)
One To Watch : Cascading & Lingual

- Developed by Chris Wensel & Team from Concurrent:
  - http://www.concurrentinc.com/
- Cascading is a development platform for building data applications on Hadoop
  - Developed on top of Cascading:
    - Lingual
      - Simplifies systems integration -- ANSI SQL compatibility -- JDBC driver
    - Pattern
      - Machine learning scoring algorithms through PMML compatibility
    - Scalding
      - Enables development with Scala, a powerful language for solving functional problems
    - Cascalog
      - Enables development with Clojure, a Lisp dialect
    - Driven
      - Understand data usage + accelerate Cascading application development and management
Driven -- Visualize Development of Flows

- Like SSMS Execution Plans
- Breaks up Query
- Shows Data flow
  - Drill down ....
Driven -- Application Insights

- Drill down into steps
  - Execution Time
  - Bottle-necks
  - Resource usage
Why Watch : Cascading & Lingual ?

- All 3 Big data platform vendors mentioned before
  - supports Cascading integration
  - investing in ensuring continued support for Cascading on their own platforms

- Used by
  <img src="twitter.png" width="50" height="50"> <img src="airbnb.png" width="100" height="100"> <img src="razorfish.png" width="100" height="100">

- Single platform to develop code on that evolves with changing big data landscape.
- Single JAR deployment.
- Ansi-92 interface via JDBC for moving data between systems / platforms
- All Open-Source (no vendor lock-in)
- Data Soil is contributing to develop the SQL Server Plug-in for Cascading & Lingual.
  - (see our blogs for getting into Cascading using Microsoft Technologies)
B) How is Big Data Different?

- Philosophies
  - Current Architecture vs Schema-On-Read
  - S-O-R : Advantages & Disadvantages
  - Integration with SQL Server & Windows
## Current BI Architecture vs Big Data BI Architecture

<table>
<thead>
<tr>
<th>Current BI Architecture</th>
<th>Big Data BI Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get Business Requirements and prioritize</td>
<td>Get Business Requirements and prioritize</td>
</tr>
<tr>
<td>Find / Collect all relevant data sources</td>
<td>All Data is already in the Ha-dump</td>
</tr>
<tr>
<td>Normalize / copy to staging / create structures / schemas / ETL</td>
<td>Create schema for question 1 / ETL</td>
</tr>
<tr>
<td>Create Warehouse / Cube</td>
<td>Send processing instructions to data</td>
</tr>
<tr>
<td>Start answering questions 1 / 2 / 3 / 4 / 5</td>
<td>Answer question 1 {&amp; Repeat}</td>
</tr>
</tbody>
</table>
S-O-R : Advantages & Disadvantages

- **Advantages**
  - Store first, ask questions later
    - Storage is cheap compare to high availability SAN
    - Format agnostic as not pre-normalization / conversion required
  - All data is available in a central place
  - High degree of parallel processing → speeds up large batch processing
  - Possible to start answering business questions quicker

- **Disadvantages**
  - New skillsets & training required
  - Company may not support new software stack
  - Creating new schemas for proprietary data can be difficult
Integration with SQL Server & Windows

- ODBC
  - Hortonworks / Cloudera / MAPR all have supported ODBC drivers
  - Create Linked Servers directly from SQL Server
  - SSIS integration
  - Pull Data directly into Excel (see Hortonworks Sandbox)

- JDBC & Other
  - Tableau / squirrel-sql / Revolution R / Business Objects ext.

- Other ETL Tools
  - Talend (to be discussed later)

- Local Install
  - Hortonworks Data Platform (HDP)
  - HDInsight Emulator
C) How to ride the Elephant?

i) All about the tools
   - Local VM platform providers
   - Online platform providers
   - Vagrant
   - Talend
   - Reuse of old machines

ii) Sources of Inspiration
   - Sandbox’s
   - The Apache Software Foundation
   - Github
i) All about the tools

- Local VM platform providers
- Online platform providers
- Vagrant
- Talend
- Pet Project : Reuse of old machines
Local VM platform providers

- Hyper-V (Microsoft)
  - Windows Server
  - Windows 8.1

- VMWARE
  - VMWARE Server Products
  - Workstation - On Windows
    - Personally, I absolutely LOVE Workstation 10.0
  - Fusion - On Mac

- Virtual Box (Oracle)
  - Runs on EVERYTHING
    - Close second favourite
    - Integrates extremely well with Vagrant (to be discussed)
Online platform providers

- **Azure & Big Data**
  - HD-Insight (Based on Hortonworks HDP platform)
    - Real World Big Data (SQL-Bits Session)
      - Adam Jorgensen / John Welch
      - Restored my confidence in MS Big Data Cloud Solutions

- **Amazon Cloud (AWS)**
  - EC2
  - Host of supporting services

All trademarks mentioned herein belong to their respective owners
Vagrant

- Vagrant provides
  - easy to configure,
  - reproducible,
  - and portable work environments built on industry standards.
- Spins up / Hibernates / Destroys complex development environments with one line of code
- Supports Virtualbox / VMWARE / Docker / Hyper-V / Custom Providers
- Ability to spin up environments locally or directly to Amazon EC2

All trademarks mentioned herein belong to their respective owners
Enterprise grade development environment for creating data integration across just about anything.

**Talend Open Studio for Big Data**
- **BASIC - Free**
  - Eclipse-Based Tooling
  - Hadoop 2.0 and YARN Support
  - Big Data ETL and ELT
  - HDFS, HBase, HCatalog, Hive, Pig, Sqoop Components
  - Job Designer
  - Apache License 2.0
  - Broadest NoSQL Support
  - Fully Open Source

http://www.talend.com/download

All trademarks mentioned herein belong to their respective owners
Talend (i)
Talend
## Supported Database & Data Source Connectivity

<table>
<thead>
<tr>
<th>Database &amp; Data Source</th>
<th>Supported Databases/Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon RDS</td>
<td>HIVE</td>
</tr>
<tr>
<td>Amazon Redshift</td>
<td>HSQLDB</td>
</tr>
<tr>
<td>Amazon S3</td>
<td>Informix</td>
</tr>
<tr>
<td>AS400</td>
<td>Ingres</td>
</tr>
<tr>
<td>DB2</td>
<td>InterBase</td>
</tr>
<tr>
<td>Derby DB</td>
<td>JavaDB</td>
</tr>
<tr>
<td>Exasol</td>
<td>JDBC</td>
</tr>
<tr>
<td>eXist-db</td>
<td>MaxDB</td>
</tr>
<tr>
<td>Firebird</td>
<td>Microsoft OLE-DB</td>
</tr>
<tr>
<td>Greenplum</td>
<td>MySQL</td>
</tr>
<tr>
<td>H2</td>
<td>Netezza</td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
</tr>
<tr>
<td>ParAccel</td>
<td></td>
</tr>
<tr>
<td>PostgresSQL</td>
<td></td>
</tr>
<tr>
<td>PostgresPlus</td>
<td></td>
</tr>
<tr>
<td>SAS</td>
<td></td>
</tr>
<tr>
<td>SQLite</td>
<td></td>
</tr>
<tr>
<td>Sybase</td>
<td></td>
</tr>
<tr>
<td>Teradata</td>
<td></td>
</tr>
<tr>
<td>VectorWise</td>
<td></td>
</tr>
<tr>
<td>Vertica</td>
<td></td>
</tr>
<tr>
<td>Windows Azure Blob Storage</td>
<td></td>
</tr>
</tbody>
</table>
Pet project: Reuse of old machines

- Challenge your manager
- If you can build a cluster from your old desktops that will outperform his current development server, he has to give you a raise!
- You’d be surprised what you can do with a pile of these!
ii) Sources of Inspiration

- Sandbox’s
- The Apache Software Foundation
- Github
Sandbox’s

- All three the Big Data Players have their pre-built Sandbox’s you can download and experiment with

- Hortonworks
  - Current Version 2.1
  - Supports: VirtualBox / VMWare / Hyper-V

- Cloudera
  - Current Version CDH 5.0.x
  - Cloudera Live online (beta)
  - Supports: VirtualBox / Vmware / Linux KVM (Kernel-based Virtual Machine)

- MAPR
  - Supports: VirtualBox / Vmware

- Cascading & Lingual
  - Vagrant Image that spins up 4 Node Cluster via GitHub
  - Supports: VirtualBox
The Apache Software Foundation

- Want to know about BIG future technologies
- Apache Incubator - (http://incubator.apache.org/)
  - Tez → Speed up MapReduce
  - Storm → high-performance realtime computation system
  - Optiq → SQL interface & advanced query optimization - non-RDBMS systems
  - Falcon → quickly onboard their data, associated processing & management tasks on Hadoop clusters
GitHub is a web-based hosting service based on Git.

Git a distributed revision control and source code management (SCM) system initially designed and developed by Linus Torvalds for Linux kernel development.

Great source of Vagrant-Based VM’s

- Cascading & Lingual Cluster (Get Vagrant & Virtual Box)
- [https://github.com/Cascading/vagrant-cascading-hadoop-cluster](https://github.com/Cascading/vagrant-cascading-hadoop-cluster)
D) BIG to the Future!

- i) Current Common Use-cases
- ii) Future Opportunities
i) Current Common Use-cases

- Sentiment (twitter feeds / wordpress scrapes / facebook likes)

- Recommendation Engines using Mahout / Other (Netflix)

- Anti Money Laundering ??
  - Live Transaction monitoring - not that big for some reason
  - Graph Databases seems to be doing better here.
ii) Future Opportunities

- Sensors
- Self-Contained Clusters
- Combination?
These days, sensors can be installed everywhere to monitor all aspects of life / business

- Temperature Sensors
- Pressure Sensors
- Gas Sensors
- Smoke Sensors

A better understanding of day to day happenings can save money and lives.
Self-Contained Clusters

Met these guys at the Hadoop Summit in Amsterdam 2014
(http://bigboards.io/)

5 data processing nodes
20 CPU cores and 5TB of raw storage
1GB ethernet to interlink everything
1 management console with technology and data library
Self-Contained Clusters + Sensors
Self-Contained Clusters + Sensors
Self-Contained Clusters + Sensors
E) Summary

- Big data does not replace random read and reporting capabilities of SQL Server.
- Big Data is not close to replacing our trusted, high volume, transaction safe, OLTP frameworks we built.
- Big data opens up opportunities for storing and processing data at a larger scale than we could never have dreamed of before.
F) Conclusion

- THE FUTURE is not going to be won by one OR the other...

...but by a **combination** of BOTH!
Q & A
Tools To Play With

- Hortonworks Sandbox
- Cloudera Sandbox
- MAPR Sandbox
- Cascading & Lingual Cluster (Get Vagrant & Virtual Box)
  - https://github.com/Cascading/vagrant-cascading-hadoop-cluster
- Vagrant
  - http://www.vagrantup.com/
- Virtual Box
  - https://www.virtualbox.org/
- Talend
  - http://www.talend.com/download
- VMWARE Workstation 10
  - https://my.vmware.com/web/vmware/info/slug/desktop_end_user_computing/vmware_workstation/10_0
- HDInsight Emulator
Appendix : References

**1) Hadoop : Distributed Data Processing [Amr Awadallah]
   - http://www.slideshare.net/cloudera/hadoop-distributed-data-processing

**2) Hadoop [K Subrahmanyam]
   - http://www.authorstream.com/Presentation/aSGuest129127-1356869-techseminar-on-hadoop-ppt/

**3) An Introduction to Apache Hadoop MapReduce [Mike Frampton]
   - http://www.powershow.com/view/3fdd1b-MGRkZ/An_Introduction_to_Apache_Hadoop_MapReduce_powerpoint_ppt_presentation

**4) Mahout Explained in 5 Minutes or Less [Josh Gertzen]
   - http://blog.credera.com/technology-insights/java/mahout-explained-5-minutes-less/

**5) What is Apache Tez? [Roopesh Shenoy]
Thank you - COPY OF SLIDES ON WEB!

- Eduard Erwee
- Data Soil Ltd
  - E-mail: eduard.erwee@datasoil.uk
  - Web Site: www.datasoil.uk
  - Blog: blog.datasoil.uk
  - Twitter: @datasoil
  - Facebook: www.facebook.com/datasoil

- Please Remember to do the feedback form online
  - http://www.sqlbits.com/SQLBitsXIIISaturday