The 5th 'V' of Big Data: Are You Generating Real 'Value' from your Big Data?

David Katz, Principal Consultant
Jagrata Minardi, Staff Solutions Consultant

#TIBCONOW
SAFE HARBOR DISCLOSURE

During the course of this presentation, TIBCO or its representatives may make forward-looking statements regarding future events, TIBCO’s future results or our future financial performance. Although we believe that the expectations reflected in the forward-looking statements contained in this presentation are reasonable, these expectations or any of the forward-looking statements could prove to be incorrect and actual results or financial performance could differ materially from those stated herein.

TIBCO could experience factors that could cause actual results or financial performance to differ materially from those contained in any forward-looking statement made in connection with this presentation. TIBCO does not undertake to update any forward-looking statements that may be made from time to time or on its behalf.
This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release and availability dates for TIBCO products and services. This document is provided for informational purposes only and its contents are subject to change without notice. TIBCO makes no warranties, express or implied, in or relating to this document or any information in it, including, without limitation, that this document, or any information in it, is error-free or meets any conditions of merchantability or fitness for a particular purpose. This document may not be reproduced or transmitted in any form or by any means without our prior written permission.

The material provided is for informational purposes only, and should not be relied on in making a purchasing decision. The information is not a commitment, promise or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any features or functionality described for our products remains at our sole discretion.

During the course of this presentation TIBCO or its representatives may make forward-looking statements regarding future events, TIBCO’s future results or our future financial performance. These statements are based on management’s current expectations. Although we believe that the expectations reflected in the forward-looking statements contained in this presentation are reasonable, these expectations or any of the forward-looking statements could prove to be incorrect and actual results or financial performance could differ materially from those stated herein. TIBCO does not undertake to update any forward-looking statement that may be made from time to time or on its behalf.
The following information is proprietary information of TIBCO Software Inc. Use, duplication, transmission, or republication for any purpose without the prior written consent of TIBCO is expressly prohibited.
Agenda

1. Introduction
2. Data Connectors
3. Map Reduce
4. Spark
5. Analytics
6. Conclusion
Introduction

Value from Data at TIBCO
Value from Big Data

TIBCO Spotfire® Value Proposition

• Interactive Visual Exploration
• Integrated Analytics
• Author/Consumer Model
• Guided Analysis
• Collaboration
What is Big Data?
What is Big Data

Too big for JAWS (just another workstation)

What is Big Data

Too big for Just Another Server

What is Big Data

“You're Gonna Need a Bigger Boat”
What is Big Data

Use a Cluster

Memory / Use Cases

Not all data needs be in memory at once:

• Select
• Aggregate
• Sample

All data should be in memory at once:

• Iterate to create analytic models
Memory / Use Cases

- In Workstation
- In Server
- In Cluster Memory
- Too Big for all of Cluster Memory
Memory / Use Cases / Spotfire

- In Workstation In-memory Engine
- In Server = WebPlayer
- In Cluster Memory
  - Cached tables for query
- Too Big for Memory
  - Convert to iii via aggregation or sampling
- In-database aka client-server
This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release and availability dates for TIBCO products and services. It is for informational purposes only and its contents are subject to change without notice.

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.
Data Connectors

- Seamlessly pull data from external data stores without custom configuration or coding
- Layer between Spotfire and a SQL-based data store
- Generates SQL transparently from visualization choices
  - Demo here
Data Connectors

Data connectors offer direct connectivity to big data

- Reduced need for IT cycles
- Uncover data issues to be addressed upstream
- Proceed with analysis
- Address great volume and variety
- Leverage processing power of the data store
Data Connectors

- Custom Queries
  - When complex SQL is required, it can easily be shared between users
  - And can be made flexible via business-user-supplied parameters
select a.*
from default.titems as a inner join default.titems as b
  on a.transactionId = b.transactionId
where b.categoryId = ?Basket
and a.categoryId <> ?Basket
# Hadoop Certified Data Connectors

<table>
<thead>
<tr>
<th>HADOOP</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Hive</td>
<td>Supported via the Hortonworks connector</td>
</tr>
<tr>
<td>Apache Spark SQL</td>
<td>Certified native connectivity with support for Kerberos SSO and LDAP with SSL</td>
</tr>
<tr>
<td>Cloudera Hive</td>
<td>Certified native connectivity with support for Kerberos SSO and LDAP with SSL</td>
</tr>
<tr>
<td>Cloudera Impala</td>
<td>Certified native connectivity with support for Kerberos SSO and LDAP with SSL</td>
</tr>
<tr>
<td>Databricks Cloud</td>
<td>Certified native connectivity</td>
</tr>
<tr>
<td>Hortonworks</td>
<td>Certified connectivity through Hive</td>
</tr>
<tr>
<td>MapR Drill</td>
<td>Certified connectivity via the TIBCO Spotfire Server</td>
</tr>
<tr>
<td>Pivotal HAWQ</td>
<td>Native connectivity through Pivotal SQL query engine</td>
</tr>
</tbody>
</table>
MapReduce
Job Processing through Spotfire
This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release and availability dates for TIBCO products and services. It is for informational purposes only and its contents are subject to change without notice.

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.

MapReduce

System Topology

Spotfire machine
- Windows
- TIBCO Spotfire® Analyst
- Spotfire Connector for Cloudera Impala

Master Node
- Apache Hadoop, Apache Hive, Cloudera Impala
- TIBCO Spotfire® Statistics Services
- TERR, R packages
- shell scripts used by MapReduce jobs

Core Nodes
- Hadoop
- TERR, R packages

Calling Sequence

Spotfire via Statistics Services
mapper.R via TERRscript
reducer.R via TERRscript

Hadoop Streaming
$ hadoop-streaming -map mapper.R -reduce reducer.R
-input 'inputfile' -output 'outputfile'

HDFS
Each Node Processes its own data using TERR
Data Node Data Node Data Node Data Node
MapReduce

Specifying a MapReduce Job

On this page, you specify a MapReduce job that will run on a Hadoop cluster. For all unique combinations of values in the specified key variables, a reducer task will run on the corresponding rows of the source data, modeling either `volmisecase` or `volmisecase`. The reducer task will have access to all columns not shown in the blacklist.

A Assign key variables and value variables. The choice of key variables, and data set, determines the number of groups to which the reducer is applied. This is summarized at right.

B Specify a reducer function from the list. Optionally, you may edit the reducer function. Then specify a location in which to execute the routine. Choose the data set size, and click Execute:

Reducer Function:

```
reduce_function <- function(df) {
    mod1 <- lm(volmisecase ~ volmisecase + volmisecase, data=df);
    y <- coef(mod1);
    return(y)
}
```

output details

last run: 2016-01-09 05:01:21 GMT
elapsed: 68.530 secs

Online
**Advanced Analytics Award at Strata 2014**

**Use Case: Manufacturing Process Analysis of Production Issues**

Multiple data streams from PLCs monitoring machine parameters

11,244 Files, 98,400 record values each ~1.1MM records

Extract history of 30 seconds previous to an event of interest plus the duration of the event.

Analyze each event to identify early warning signals.

---

**Map Reduce with TERR (TIBCO Enterprise Runtime for R) from Spotfire**

**Split-Apply-Combine**

A. Choose a role from the list box below, then select variables at right.

B. Select variables to which you wish to assign that role. Then click Set Role back in A.

C. Specify a Reducer Function in the text area, or choose one of the example models. Then specify a location in which to execute the routine. Click Execute.

**Reduce Function**:

```r
reduce_function <- function(df) {
  mod1 <- lm(rollShip ~ volMissCase, data=df);
  y <- coef(mod1);
  y[2] <- ifelse(is.na(y[2]), 1000, y[2]);
  return(y)
}
```

**Output Details**

Last run: 2014-06-02 15:38:26 PDT

Elapsed: 0.200 secs

---

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.

#TIBCONOW
• Convention: Positive is event identification, which is a failed process.

• We want a model where true positive rate change is very steep near zero.

• Identify a high number of true positives at a small cost of false positives (unnecessary process stop).
Spark

Distributed In-Memory Processing … via Spotfire
Spark Landscape

The Apache Spark project was started in the UC Berkeley AMPLab. It had two specific goals.

1. Extend the MapReduce model to better support iterative algorithms (machine learning, graphs) and interactive data mining

2. Enhance programmability by allowing interactive use from a Scala interpreter

https://spark.apache.org/talks/overview.pptx
Spark Landscape

What is the status?
This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release and availability dates for TIBCO products and services. It is for informational purposes only and its contents are subject to change without notice.

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.

https://weidongzhou.files.wordpress.com/2015/09/spark_engine.jpg
Integration with Spark

With the emphasis on interactive-latency analytics, where does TIBCO fit into the picture?

1. Querying data in big data stores or pre-cached in a Spark cluster

2. Launching an interactive-latency Spark job ... and back to 1.
Integration with Spark SQL

4 Easy Steps for Ultra-Fast Visualization of Big Data with Spotfire and Spark SQL

Launching Spark Jobs from Spotfire

Using the SparkR package (R API for Spark), business users access Spark *interactively* for both data management and modeling tasks.

Resulting DataFrames in Spark do not need to be “collected.” They can be queried from distributed memory using the Spark SQL Connector against the context of their creation.
Launching Spark Jobs from Spotfire

Spotfire and Spark are also used in a workflow that deploys models against **streaming** data.

Spark models are developed in a Spotfire-centric workflow, then deployed to a TIBCO StreamBase® application.

These models are periodically tuned in Spotfire and redeployed.

This is done, for example, in the TIBCO Accelerator for Apache Spark.
• TERR for divide and conquer
• H₂O
• MLLib
• Mahout
• Etc.....
H₂O – The Killer-App for Spark

<table>
<thead>
<tr>
<th></th>
<th>In-Memory</th>
<th>Community</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLlib</td>
<td>Big Data, Columnar</td>
<td>Devis, Data Science</td>
</tr>
<tr>
<td>H₂O</td>
<td>100x faster Algos</td>
<td></td>
</tr>
<tr>
<td>SQL</td>
<td>CRAN, API, fast engine</td>
<td></td>
</tr>
<tr>
<td>H₂ORDD</td>
<td>Spark API, Java MM</td>
<td></td>
</tr>
<tr>
<td>HDFS=DATA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.
This document (including, without limitation, any product roadmap or statement of direction data) illustrates the planned testing, release, and availability dates for TIBCO products and services. It is for informational purposes only and its contents are subject to change without notice.

© Copyright 2000-2016 TIBCO Software Inc. All rights reserved. TIBCO Confidential & Proprietary Information.
Conclusion
Takeaways

1. Spotfire exposes information in data of any size by using multiple strategies for managing resources.

2. TERR offers access to job processing on data of any size through in-database, HDFS-based, and distributed memory frameworks.

3. The combination of Spotfire and TERR offers ease of use for business users, and depth for data scientists who can become authors of self-service Spotfire interfaces.

4. TIBCO Analytics extends the real value of today’s rapidly evolving big data tools to a wide range of business users.