Spotfire and Hadoop: Interactive Analysis on Big Data

WHY HADOOP?
Hadoop provides an inexpensive and robust way of storing large amounts of diverse data on commodity hardware and has in recent years become the default choice for organizations to store both structured and unstructured data. Hadoop has spawned a rich ecosystem of open source projects, technical approaches, and commercial vendors. Using the powerful combination of visual and advanced analytics, TIBCO helps organizations leverage and analyze data from Hadoop, enabling users of all levels to derive business value from big data, irrespective of their Hadoop technical skills.

TIBCO Spotfire® enables business users with an easy to use interface to access and analyze their Hadoop data. With TIBCO Spotfire, any complexity regarding where data resides and how it is orchestrated is dealt with behind the scenes, so users can focus on the business questions they want to answer to bring value to their organizations.

TIBCO offers two main types of native Hadoop integration: TIBCO Spotfire native data connectors, and TERR-Hadoop integration. The first allows uber fast interactive visualizations. The latter enables data calculations of any complexity on Hadoop data. Together, they offer a powerful combination of visual and advanced analytics that is, most importantly, easy to use.

Spotfire Hadoop connections facilitate the creation of rich insightful visualizations of Hadoop data. Analytic workflows, dashboards, and reports can be built, shared, reused, and consumed across organizations to virtually any user’s device. Extensive geo analytic support within Spotfire also makes it easy to generate insights from geographical data. Figure 1 shows an example of a Spotfire dashboard built on top of Hadoop data.

VISUALIZE DATA IN HADOOP
Table 1 shows a list of data connectors that allow Spotfire to access Hadoop data. Spotfire supports a number of modes for accessing data: in-database (data source), in-memory, on-demand, or any combination of these. In-database enables extremely fast analyses of huge data sets by minimizing data migration and pushing aggregations into Hadoop. In-database master visualizations can be combined with detailed visualizations where row level data slices are extracted from Hadoop on-demand and loaded into memory. These processes do not require any scripting or manual query editing. However, the ability to create custom queries is still available, if needed.
HADOOP | NOTES
--- | ---
Apache Hive | Supported via the Hortonworks connector
Apache Spark SQL | Certified native connectivity with support for Kerberos SSO and LDAP with SSL
Cloudera Hive | Certified native connectivity with support for Kerberos SSO and LDAP with SSL
Cloudera Impala | Certified native connectivity with support for Kerberos SSO and LDAP with SSL
Databricks Cloud | Certified native connectivity
Hortonworks | Certified connectivity through Hive
MapR Drill | Certified connectivity via the TIBCO Spotfire Server
Pivotal HAWQ | Native connectivity through Pivotal SQL query engine

TIBCO Spotfire connector for HPE Vertica provides deep integration of Vertica SQL engine with Cloudera, Hortonworks, MapR and other Hadoop distributions.

TIBCO Spotfire also has connectors for Amazon Redshift, IBM DB2, IBM Netezza, Oracle, Oracle Essbase, Pivotal Greenplum, SAP HANA, SAP NetWeaver Business Warehouse (SAP BW), Teradata and Teradata Aster.

Table 1. Hadoop data connectors.

**TERR AND HADOOP FOR ADVANCED ANALYTICS**

Combining Spotfire, TERR, and Hadoop enables deeper analyses of Hadoop data. In just a few mouse clicks, less technical users can run relevant calculations on huge Hadoop data sets. For example, they can run price optimization models per country, or models to identify which clients are likely to accept an offer, or which are likely to churn; which transactions may be fraudulent; or which machines are likely to break. They can later choose to deploy these models in batch or in real time.

Running TERR on Hadoop provides the wealth of R analytical tools. Faster, more scalable, and more robust than open source R, TERR inherently enables faster and more reliable Hadoop processing. Several alternatives can be dynamically combined to define how Hadoop and TERR interact:

**USE SPOTFIRE FOR DATA SELECTION, TERR FOR ADVANCED ANALYSIS**

Spotfire can exploit the power of Hadoop for in-database aggregations, presenting highly responsive overview visualizations of aggregated data, and then allowing users to extract selected slices of the data for deeper analysis. The detailed data can be sent to TERR for advanced analysis, and the entire workflow can be encapsulated and shared as a best practice within Spotfire.

**TERR GETS DATA FROM HADOOP DIRECTLY**

When passing data slices via Spotfire is prohibitive, TERR can reuse an existing connection to Hadoop to extract and analyze data directly. This enables routine transformation and analysis of big data.
MAP/REDUCE WITH TERR
TERR can execute MapReduce calls directly on a Hadoop cluster. Figure 2 shows a Spotfire dashboard with a simple input dialog that allows the user to choose variable roles (map step) and select the advanced analytics algorithm to run on them (reducer step). TERR algorithms run in parallel with the Hadoop data, with results passed to Spotfire visualizations. Figure 3 shows what happens under the hood.

As an illustration, a user may want to run a separate price optimization calculation for each country and product combination. They would select Country and Product as keys in the map step and select Price Optimization for the reducer step. Another example is running an ensemble model, where separate predictive models are created in parallel across the Hadoop cluster, and then combined to provide the best predictive model for the application.

SPOTFIRE, TERR, AND H2O
MapReduce involves making separate calculations on subsets of the data. However, certain use cases may require a single large calculation that runs on all Hadoop data at once. For example: predicting whether a piece of equipment will fail given its sensor readings. MapReduce can solve this with the ensemble models discussed above. Additionally, H2O provides a set of machine learning algorithms for distributed data sources. A point-click Spotfire dashboard can be set up that allows the end-user to select predictor variables and initiate H2O analyses.

TERR IN SPARKR
SparkR is an R package that provides a front end to the Apache Spark system for distributed computation. It allows R to invoke Spark jobs, which can then call R to perform computations on distributed worker nodes. The SparkR source can be modified to call the TERR engine rather than the R engine for extra computational power. A Spotfire dashboard can call out to prebuilt SparkR jobs or provide a familiar interface for the user to create new ones.

PUTTING IT ALL TOGETHER
Combining these powerful features means that very sophisticated and powerful analytic use cases can be encapsulated in easy-to-use Spotfire dashboards, enabling business users to visualize, analyze, and share the results of analytic applications without worrying about the details of the Hadoop architecture.

Figure 2. Example of a Spotfire template for running a MapReduce job using TERR. Spotfire dashboards can be customized for the target user, use case, and technologies. The image is an example of a dashboard that a business user can interact with to define a Hadoop MapReduce job. The code for the mapper and reducer jobs is generated behind the scenes and run on the data nodes. The user can choose to initially run calculations locally on a sample of data, before launching them onto Hadoop. Results can be returned to the Hadoop file system (HDFS) or loaded straight into Spotfire. The choice of models is made available in a drop-down menu. If relevant, the reduce R/TERR code can be exposed for the user to edit.
Figure 3. Under the hood of TERR MapReduce. The Spotfire point-and-click environment creates TERR scripts for Mapper and Reducer. These are run through the Hadoop streaming interface on TERR data nodes. Results are returned to HDFS and picked up in Spotfire for detailed drill-through analysis. This workflow enables complex statistical models to be run in parallel on the nodes via point-and-click in Spotfire, thus supplying the power of TERR and the scale of Hadoop to novice as well as expert users.

SPOTFIRE DATA CONNECTORS
In addition to many Hadoop data sources, Spotfire and TERR can connect to many other enterprise data sources, enabling you to combine all your data and tackle analytic challenges. Refer to the full list of supported data sources: http://spotfire.tibco.com/resources/spotfire-data-sources