TIBCO Statistica Big Data Analytics

With version 13, TIBCO Statistica™ (currently in release 13.3) continues comprehensive support for popular open-source capabilities and tooling for handling unstructured data, big data, and in-memory, and in-database analytics. To be specific, Statistica supports:

**APACHE® HADOOP®/HDFS**
Statistica provides direct file import from HDFS, and data aggregation workflows can be made into a reusable template with Statistica Enterprise for repeatable analytic and data prep workflows. Statistica can also retrieve data from Apache Hive® and Hive-on-Spark® using Statistica Server and Statistica Monitoring and Alerting Server (MAS).

**STATISTICA BIG DATA PROCESSING ARCHITECTURE**
Statistica architecture for big data processing and in-memory computational frameworks is summarized in this graphic:

---

Statistica supports multiple workspace processing nodes that define computational pipelines for H2O (Sparkling Water) and/or Apache Spark™. In short, analytic workflows deployed to Spark can be executed in Statistica using data managed in Spark as resilient distributed datasets (RDDs). Results can then be retrieved by Spark and published again to Statistica (either TIBCO Statistica™ Server or client). This workflow has the advantage of enabling version control of analytic and data prep workflows and steps executed in Statistica, while deploying big-data analytics to Spark servers.
A large number of analytics are supported, both for H2O and Spark ML, as well as for deep learning.

Spark and H2O Nodes

Statistica supports the common interfaces and scripting languages that are used by data scientists today, including Python, R, Scala, and C#. Custom nodes and programs can also be scripted in these languages, so the Statistica installation can be configured or customized with proprietary analytics pipelines.
Two Data Science Workflows (top) Used as Reusable Template (bottom)

In addition, Statistica provides point-and-click interfaces and implementations to the most effective machine learning algorithms, making it the ideal platform for managing advanced analytics throughout the enterprise. Analytics can be performed on Statistica or Spark servers using open-source tools or Statistica algorithms. Analytic workflows can be validated and managed using best practices for version control, electronic signatures, and audit trails, and analytic and data prep templates can be deployed to groups of users in a role-based environment. Results can be infused into business processes and/or delivered to self-service visualization tools and dashboards for interactive exploration and drill-down analytics.