Enterprise-Grade, Self-Service Data Prep for Hortonworks Connected Data Platforms

Big Data’s promise is to fuel better, smarter and faster decision-making. Today, Hortonworks makes it possible to collect, access and store unfathomable amounts of data from any source. The trick is turning all of that data into clean, contextual, consumable and accurate information that is worth analyzing. Most analysts call this “data preparation” and admit they spend 80% of their time cleaning, organizing and shaping data, leaving very little time for valuable decision-making.

Paxata was built from the ground-up to flip that ratio, with the first enterprise-grade data preparation platform that streamlines how quickly and easily analysts, data scientists and IT teams explore and transform raw data from Hortonworks Connected Data Platforms into clean, connected and trustworthy AnswerSets™ ready for any use case, all packaged applications and business analytics tools.

Paxata breaks the logjam of an IT-constrained model to a business-empowered one, pivoting from the traditional ETL, Data Quality and MDM model to agile data preparation within an environment that addresses IT requirements for administration, security, governance and collaboration. Powered by Apache Spark and machine learning, Paxata is the only self-service data preparation platform designed to work interactively at scale, delivering the results business teams need within the enterprise data management platform IT demands.

About Paxata

Paxata is the only enterprise-grade, self-service data preparation platform certified for Hortonworks Connected Data.

Paxata is designed for IT and business teams who need to explore, clean, shape, organize and combine data into information, regardless of source, format or volume.

Solution Highlights

• Eliminate data restrictions – work with all data interactively regardless of volume or variety
• Accelerate analytic workflows and improve decision-making by removing data preparation delays
• Available as multi-tenant cloud, VPC and on premise deployments
Paxata for Hortonworks HDP

Reduce time and friction of data preparation
Jumpstart your analytics process and get to insights faster with user efficiencies across the data preparation process.

Our customers now have the freedom to prepare data on their own or work with peers in a shared environment as they import, explore, enrich, combine, and share complete and accurate AnswerSets ready to publish to the ad-hoc analysis tool they choose.

Eliminate data restrictions
Whether already in HDP or in desktop files, Paxata eliminates the lag between needing more data and getting it integrated into the work already done. That means anyone who works with data can get the bigger picture, or add context – iterating through their analysis – without scheduling time with a data scientist or kicking off a nine-month data warehouse project.

Bring HDP to the business
Paxata leverages standard HDP components to deliver a complete, seamless raw data-to-insight data preparation platform on top of existing HDP components including HDFS, Spark, Hive and YARN that is 100% designed for the business analyst. Together, Paxata and the HDP provide a complete infrastructure that can address the challenges of Big Data while delivering rapid turnaround and analytics agility for the business.

Benefits of Paxata
• **Simple**: Excel-like interface, dynamic guidance, multi-user collaboration and simultaneous editing
• **Managed**: library for sharing data, one-stop shop for uploaded data and published AnswerSets
• **Automated**: schedule, review and reuse data prep projects
• **Open**: connectivity to data sources and BI tools with ODBC/JDBC, native connectivity to HDFS and Hive
• **Auditible**: transparent governance with time-stamping and versioning for every step performed, full replay of data prep actions, reordering or modifying of steps
• **Powerful**: data prep over a large variety and volumes of structured and unstructured data in real-time
• **Smart**: IntelliFusion™ runs proprietary machine learning, natural language processing, latent semantic indexing and statistical pattern recognition
Benefits of Hortonworks

- **Powerful**: store, process, and analyze all your data to drive competitive advantage
- **Comprehensive**: connected Data Platforms including HDP for data-at-rest and HDF for data-in-motion
- **Efficient**: Hadoop unifies compute and data to improve operational efficiency
- **Open**: 100% open source: HDP is the world’s leading open source distribution powered by Apache Hadoop
- **Simple**: easy to deploy and operate with centralized administration
- **Compatible**: leverage your existing investments for rapid adoption and lower TCO
- **Economical**: rethink the economics of data management with an open source platform on industry standard hardware - up to 90% more cost effective than traditional solutions
- **Enterprise Ready**: equipped with critical capabilities to support mission-critical operations

Transform data into information – no coding, no scripting, no waiting

Paxata’s Adaptive Data Preparation platform combines an intuitive, visually interactive data preparation user experience with an enterprise platform that dramatically accelerates time to analytics, and increases productivity of every analyst in the face of increasing volumes, variety and velocity of data. With Paxata, anyone working with data can streamline data quality, profiling, integration and shaping work in an easy-to-use Excel-like interface – no coding, no scripting, no schemas and no sampling required.

Analysts adapt and enrich data sets on the fly, as well as dynamically capture the steps involved in data prep projects. Paxata brings together data from enterprise applications, managed databases, HDFS, 3rd-party sources and local data including Excel, CSV, JSON, XML and Avro files. Paxata automatically detects data types and provides simple wizards for homogenizing and loading data sets into Paxata’s Data library within Hortonworks HDP. The Paxata Data Library compresses datasets as Parquet files which provides a fully governed and efficient landing zone for data within HDP.

Paxata highlights data quality issues including completeness, validity, consistency, timeliness and accuracy issues within data via easy-to-use, full-text search, interactive visual summaries of data values, interactive filters and visual data quality heat maps. Analysts can remediate errors, add data and make changes to entire columns or single fields without any coding, scripting or sampling. Data can be pivoted or de-pivoted, columns can be split and aggregations can be created in just a click.

Paxata automatically recommends how to connect multiple raw source data sets via machine learning and text analytic approaches. Paxata can identify single and multi-column relationships between data sets with fully configurable fuzzy matching logic.
Data sets prepared with Paxata are clean, contextually relevant and ready for analysis. These AnswerSets are published to the Paxata HDFS-backed data library and can be accessed directly via Hive or HDP HDFS, enabling a wide range of analytic tools to query large prepared data sets at scale.

Paxata’s Step Editor transparently records every action performed in a data preparation project. Paxata’s end-to-end governance model allows for replay (see what the data looked like at every step), reusability (apply previous data preparation steps to new data sets), reordering (run previous data preparation steps in a different sequence) and manage workloads (run data preparation projects in interactive or batch mode).

About Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the open source platform for storing, managing and analyzing Big Data. Hortonworks Data Platform, Hortonwork’s distribution of Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy Big Data solutions.

Hortonworks is the trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop an enterprise data platform. Hortonworks provides unmatched technical support, training and certification programs for enterprises, systems integrators and technology vendors. www.hortonworks.com

About Paxata

Paxata is the first enterprise-grade, self-service data preparation platform built for information-driven organizations who want to make data worth analyzing. Business analysts, data scientists, developers, data curators and IT teams use Paxata to accelerate the cleansing, shaping, transforming and integration of all data into rich AnswerSets™ which power ad hoc, operational, predictive and packaged analytics.

Paxata’s platform, built on Hadoop and optimized to run on Apache Spark, delivers unparalleled scalability and a unified environment that promotes transparent governance and collaboration. Paxata customers engage with an interactive, self-service platform powered by machine learning and designed to eliminate the need for coding, scripting and sampling. The solution can be deployed on premise or in public or private cloud, or in a hybrid environment.

Paxata partners with industry-leading companies such as Amazon Web Services (AWS), Hortonworks, In-Q-Tel and Carahsoft, and seamlessly connects to BI tools to greatly accelerate the time to actionable business insights. For more information, visit paxata.com.