Self-Service Data Prep for Cloudera Enterprise Data Hub

Organizations are overwhelmed by the amount and variety of data the business is demanding, in part due to the wide adoption of self-service analytic tools consumed by businesses who want to be information-driven. As customers build strategies for managing and centralizing data access into the Enterprise Data Hub and provide business teams with tools from Tableau and Qlik, they will also need to rethink their approach to data preparation and delivery.

Today, legacy data preparation and delivery models represent 80% of the work and time spent in any analytic exercise. Whether relying on traditional ETL (Extract, Transform, Load) processes or a team of SQL or R coders, the business can no longer wait for data developers and IT to respond to every request for data integration, quality, enrichment, cleansing and shaping.

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 Adaptive Data Preparation. Paxata is a truly self-service data preparation platform designed to work interactively at scale, delivering the results businesses need with the governance IT demands and the ease of use required by business analysts, technical data analysts, data architects and data scientists who need to address a wide set of analytic use cases.

Get data ready for analytics with no coding, scripting or sampling

Paxata’s Adaptive Data Preparation combines an intuitive, visually interactive data preparation application with an enterprise platform that dramatically accelerates time to analytics, and increases productivity of every analyst in the face of increasing volumes and variety 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.
Solution Highlights

- Self-service enabled Enterprise Data Hub, empowering business analysts to select, prepare and analyze data on their schedule and to meet their needs with full governance and repeatability
- Eliminate data restrictions – work with all data interactively regardless of volume or variety
- Accelerate analytic workflow and improve decision-making by removing data preparation delays

Paxata brings together data from enterprise, 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 your Enterprise Data Hub. The Paxata Data Library compresses datasets as Parquet files which provides a fully governed and efficient landing zone for data within EDH.

Paxata data preparation application highlights 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 or scripting. Data can also 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 Impala, 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 (take the same data preparation steps in a different sequence) and operationalization (run data preparation projects in interactive or batch mode at scale leveraging the scaling and performance of Spark.

Paxata and Cloudera Solution Benefits

Reduce time and friction of data preparation
Jumpstart your analytics process and get to insights faster with user efficiencies across the data prep 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 resident in the Enterprise Data Hub 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
Benefits of Cloudera

- **Powerful** – Store, process, and analyze all your data to drive competitive advantage
- **Efficient** – Hadoop unifies compute and data to improve operational efficiency
- **Open** – 100% open source: CDH is the world’s most popular 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

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 EDH to the business

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

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Founded in 2008, Cloudera was the first and is still today the leading provider and supporter of Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security and search. With over 15,000 individuals trained, Cloudera is a leading educator of data professionals, offering the industry’s broadest array of Hadoop training and certification programs. Cloudera works with over 700 hardware, software and services partners to meet customers’ big data goals. Leading organizations in every industry run Cloudera in production, including finance, telecommunications, retail, internet, utilities, oil and gas, healthcare, biopharmaceuticals, networking and media, plus top public sector organizations globally.

www.cloudera.com
Benefits of Partner Product

- **Simple** – Dynamic guidance, multi-user collaboration and simultaneous editing
- **Managed** – Repository for sharing data sets, one-stop shop for uploaded data and published AnswerSets
- **Automated** – Schedule / reuse data prep projects through REST APIs
- **Open** – Connectivity to data sources and BI tools with ODBC/JDBC
- **Auditable** – Transparent governance with time-stamping and versioning for every operation performed, ability to do full replay of data prep actions, reordering or modifying of steps
- **Powerful** – Preparation over a large variety and volumes of structured and unstructured data in real-time
- **Smart** – IntelliFusion™ runs proprietary machine learning, latent semantic indexing, statistical pattern recognition and text analytics techniques
- **Flexible** - Cloud or on-premise deployments

About Partner

Paxata is the first Adaptive Data Preparation solution 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 application powered by IntelliFusion™ and designed to eliminate the need for coding, scripting and sampling. The solution can be deployed on-premise or in the cloud.

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

© Paxata, Inc. All rights reserved. The Paxata logo and brand trademarks used herein are owned by Paxata. Other company and product names used herein may be trademarks of their respective owners.