ORACLE DATA SHEET

Oracle Big Data Preparation Cloud Service (BDP)

ORACLE
Big Data Preparation Cloud Service

Preparing data for analysis at any scale is a notoriously time consuming and error prone process. It is estimated that up to 90% of the time spent on data analysis projects is spent on data preparation. The problem is that data originates from an ever growing number of sources, comes in a wide variety of complex formats, and can span the range from structured, semi-structured, and more often unstructured content. All this content is vast, inconsistent, incomplete, and often off topic. In this environment each dataset takes weeks or months of effort to process, frequently requiring programmers writing custom scripts. Accelerating and automating data preparation is the key to unlocking the potential of all your data.

KEY FEATURES

- Self service data preparation in the hands of business users
- Powerful recommendation driven process leveraging a unique combination of Machine Learning and semantic technologies.
- Advanced transformation capabilities such as classification and enrichment from internal/external sources
- Import and ingestion of structured, semi-structured, and unstructured datasets
- Statistical profiling engine identifies issues with data
- Operationalize data flows into ETL or Business Intelligence
- Monitoring and governance dashboard

KEY BENEFITS

- Data preparation costs and time reduced to a fraction of manual efforts
- Empowers business analysts to quickly extract value from data
- Governance dashboard enables users to monitor and solve issues with data curation workloads
- Greatly reduces risks of error prone manual curation efforts

The Oracle Solution: Big Data Preparation Cloud Service

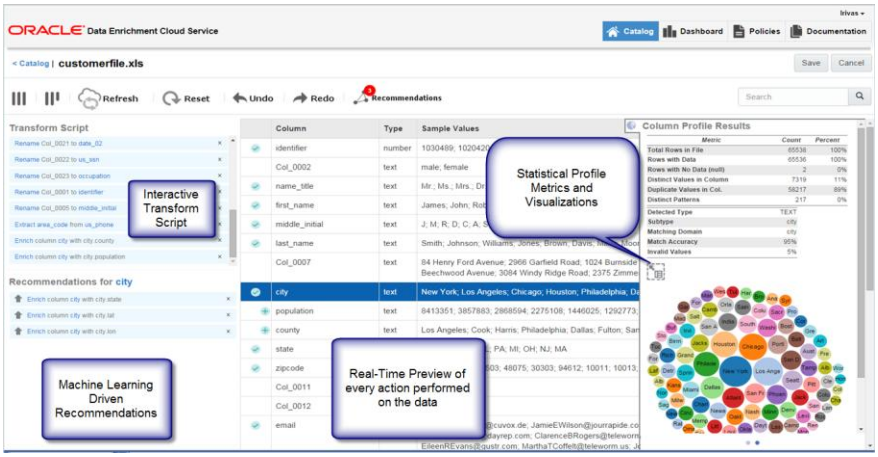

Figure 1. Transformation Authoring Screen assisted by Machine Learning Driven Recommendations

Big Data Preparation Cloud Service (BDP) provides a set of coordinated services that automate, streamline, and guide the process of data ingestion, preparation, enrichment, and governance without costly manual intervention. BDP is available in the Oracle Cloud and powered by Apache Spark and Hadoop. It provides a highly intuitive and interactive user experience, guiding business users with a rich set of recommendations, which results in a significant cost advantage in analytical and big data projects by reducing the amount of time and resources required to ingest and prepare datasets for multiple downstream processes. Typically complex operations are made easy; and, error-prone setup and configuration are resolved. In summary, Big Data Preparation Cloud Service renders the hardest parts of today's business data ecosystem simple, scalable, and automated via the Oracle Cloud reducing noise and boosting signal

quality that tremendously improves your data for downstream applications.

Oracle Big Data Preparation Cloud Service – Capabilities

Big Data Preparation Service provides a complete capability for data ingestion, preparation, enrichment, and publication. Other platforms only provide solutions for one or at most two of these three necessary modules. Each module is necessary for enabling full automation of the data preparation process.

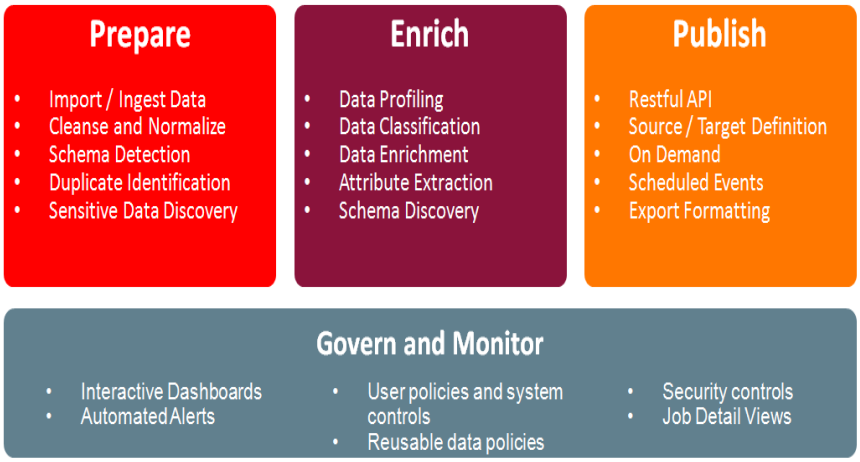

Figure 2. Core Data Preparation Lifecycle Features

Data Preparation and Repair

- **Statistical Profiling** – standard statistical analysis of numerical data and frequency and term analysis of text data.
- **Cleansing, Normalization** – removing non-essential characters, standardizing content such as dates.
- **Data Repair** – identifying and fixing where possible inconsistencies in the data.
- **Data Enrichment** – Knowledge Service based enrichments on related data.
- **Explicit Schema Detection** – identifying the schema/metadata that is explicitly defined in header, field, tag, or other information.
- **Duplicate Identification** – identifying duplicates in data.

Semantic Metadata Discovery, Enrichment, and Correlation

- **Classification, Attribute Extraction** – identifying categories in the data and identify characteristics of the data in terms of attributes, properties, schemata.
- **Implicit Schema Detection** – often it is possible to identify schema by the instances associated with the schema such as email address, postal address, name, date, time, etc. The service provides this out-of-the-box capability for many standard classes in structured and semi-structured data.

Monitoring and Governance

- **Dashboard** – a unified timeline of processed datasets provided via an operational analytics and metrics dashboard.
- **Email Alerts** – email notifications on job executions, completions, warnings, alerts, and errors.

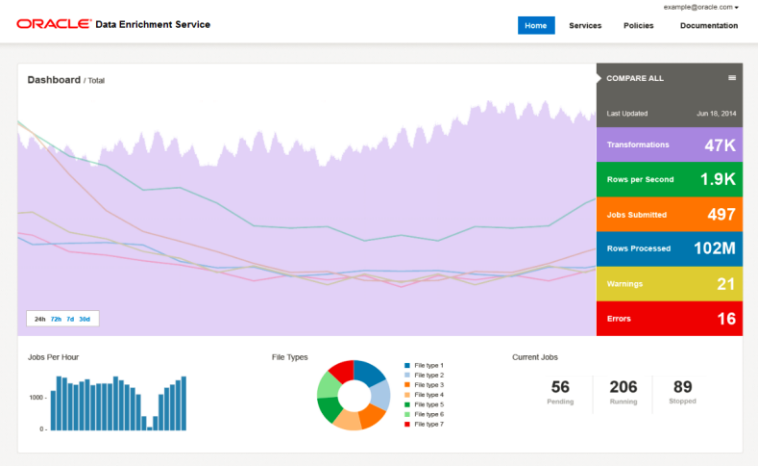

Figure 3. Monitoring and Governance Dashboard

Publishing

- **Automation Methods** – service executions can be automated via the built-in scheduler or a set of automation RestFul API's
- **Source/Targets** – the system supports a rich set of sources and targets including Oracle Storage Cloud, other external Cloud Stores, and URL sources.
- **Formats** – the service provides the ability to export the curated datasets to commonly used formats which enables downstream and on-premise BI, Analytics, and ETL processes.

ORACLE

CONTACT US

For more information about the Oracle Big Data Preparation Cloud Service, visit oracle.com or call +1.800.ORACLE1 to speak to an Oracle DIS Sales representative.

CONNECT WITH US

blogs.oracle.com/dataintegration

Oracle Data Integration

@OracleDI

oracle.com

Hardware and Software, Engineered to Work Together