# Bluemetrix Data Manager automates the Ingest, Transformation and Governance of all data onto a Hadoop Cluster

bluemetrix
the data control company

## Accelerate the Journey to an Operational Hadoop Cluster

Bluemetrix Data Manager (BDM) has been developed to allow a non-technical resource to build, schedule, transform, ingest and manage data pipelines inside Hadoop without having to write any code or know the underlying Hadoop environment. It applies automation to a range of different tasks so that the necessary code and commands are created and deployed as required. BDM fully compliments the Hadoop eco-system and creates no proprietary code.

It works exclusively on the Spark environment within Hadoop

BDM is a framework for the Ingestion, Masking, Translation, Transformation, Governance, Validation, Management and Quality Assurance of Data on Hadoop

## Data Ingest

- Simple template-based Connector system for all data sources
- Multiple Connectors available
- No need to develop any ingest code or select appropriate Hadoop components
- New data sources can be deployed in hours rather than weeks or months
- Storage can be selected to suit the data type and processing requirement i.e. HIVE, HBase, etc.
- No extra code is developed, reducing the code release cycle time and complexity

## Data Masking/Tokenization

- Data Masking is available on ingest to the cluster;
- It can be carried out on a column or table basis
- Stateful and Stateless Tokenization solutions are available
- Different masking algorithms can be applied to suit the data i.e.
  - Complete removal of selected columns
  - Replace values with random data
  - Add a random value to each row in the table
  - Categorize data e.g. exact salary replaced with a range
  - Geolocation data – apply rotation methods to mask the data

## Data Quality & Validation

- Data Consistency is guaranteed by applying checksums and other controls on the data
- Data Integrity is provided by Regular Expression and ML algorithms
- All quality data is accessible through a dashboard which will provide a snapshot of the health of the data on the cluster

## Data Translation

- All source schemas are translated into Hadoop compatible schema's
- Control characters removed and changed as appropriate
- Data cleansed, formatted and factored

## Data Transformation

- Data transformations are coded and stored in a custom library deployed in Spark
- Data maps/flows can be created using a drag and drop interface
- Dramatic reduction in code developed and deployed
- Dramatic reduction in scripts developed
- No requirement for SQL skills or HIVE knowledge to transform the data
- No requirement for Spark expertise to create transformations
- An API can be provided to the Spark library allowing client developers create and deploy their own Spark transformations

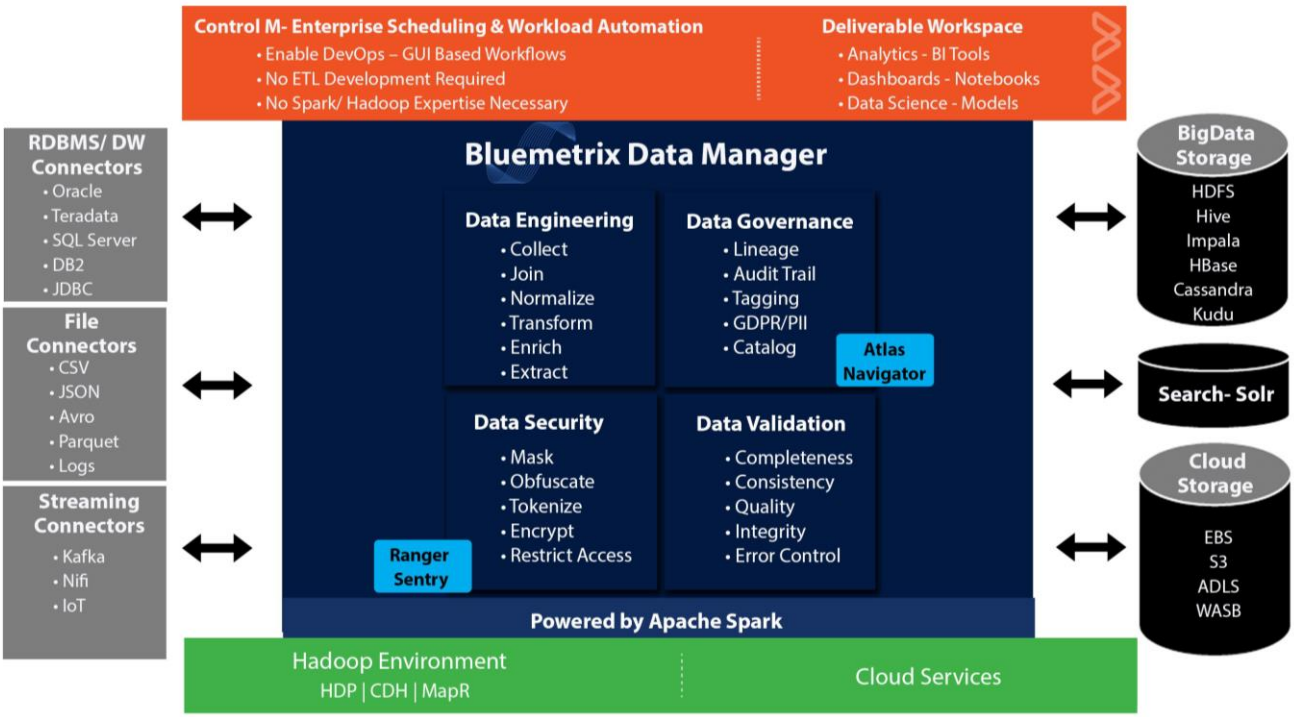
**"Bluemetrix Data Manager (BDM) allows you ingest data, replicate business workflows, validate data, measure data quality and automatically capture all governance of data – and without writing any Hadoop code"**

## Data Catalogue and Auto-Tagging

- All data is automatically detected and catalogued on the system
- ACL's -including tag based policies – implemented and integrated with underlying Hadoop Stack
- Data is stored on Solr and provides full text search capability across all data storage on the cluster
- Tags can be either technical or business
- Auto-Tagging based on Machine Learning Algorithms is available i.e. data can be automatically tagged or the system can suggest tags to be validated by the user.

## Data Governance & Lineage

- All data governance capabilities – Audit, Change Tracking, etc. – are built into Atlas
- Governance functionality can be easily customized to add new data and features i.e. addition of new GDPR compliance tags, etc.
- Process is completely independent of the end user and happens in the background
- Only solution with end-to-end data governance enabled on Atlas available in the market today




## Why Bluemetrix?

**Experience**: We have been working with Hadoop since 2009 and have experience in all areas of the Stack – Architecture, Infrastructure, Security, Application Development, Deployment, Operations and Data Science
**Guaranteed Delivery:** We guarantee delivery on all Hadoop projects we undertake
**Innovation & Automation**: We are leaders in developing and deploying innovative solutions to deal with problems on the Hadoop Stack, with a focus on developing real-world automation solutions that removes the need for Hadoop expertise.

bluemetrix
the data control company

## Contact us.


Info@bluemetrix.com
+353 21 4212223
Unit 5C River House, Blackpool Retail Park,
Blackpool, T23 R5TF Cork,
Ireland.