WHITEPAPER

Trifacta Data Wrangling for Hadoop: Accelerating Business Adoption While Ensuring Security & Governance

Introduction

By now, the majority of organizations understand the importance of big data and its required infrastructure investments—73 percent of all organizations reported that they had already invested or planned to invest in big data[1] by 2016. As data production continues to accelerate, organizations understand that they must accurately leverage these sources of information in order to stay ahead of the competition. For a typical Fortune 1000 company, just a 10% increase in data accessibility will result in more than $65 million additional net income[2].

It didn't take long, then, before the key driver of big data projects moved beyond infrastructure cost savings—such as data warehouse and ETL offloading—and toward exploratory analysis projects. Today's organizations want to do more than just save money; they want to drive innovation.

Data Lake Adoption

One common pattern within these big data initiatives is the utilization of a data lake (a.k.a. enterprise data hub, data reservoir) where a variety of data sources within the organization, as well as compelling data from external or third party sources, are stored in their respective native formats in a single environment. Because of Hadoop's theoretically limitless data storage and processing capacity at a fraction of the cost of traditional data warehouse technology, companies tend to store as much information as possible in order to, eventually, use it to inform how their business is run.

However, Hadoop has often been intimidating to organizations because only their most technologically capable (and scarce) data science resources could truly take advantage of the technology. Now, that barrier is breaking down.

The Self-Service Data Preparation Solution

At Trifacta, we're focused on helping Hadoop reach mainstream business adoption, which enables organizations to successfully deliver on the promise of big data initiatives. Trifacta's data wrangling solution empowers the individuals that know the data best—but are not necessary the most technically skilled—to access this variety of raw information and effectively explore, profile, transform, cleanse and publish it into the appropriate format so it can be consumed in downstream applications or business processes.

> "This self-service approach to data preparation—or data wrangling—allows business users to leverage the growing variety of data, which ultimately produces better business outcomes."

While data wrangling relieves enterprise IT organizations from being exclusively responsible for preparing data for their business counterparts, self-service access has introduced new challenges. IT must adapt their processes to collaborate with empowered business entities, while also complying with ever-growing constraining regulation and security issues.

In this paper, we'll provide guidance for enterprise IT organizations on deploying Trifacta's data wrangling solution with a special focus on the organizational security and data governance capabilities supported by Trifacta. To do so, we'll cover:

- The Data Lake Concept
- Data Wrangling on the Data Lake

Sources

[1] http://www.gartner.com/newsroom/id/2848718

[2] http://www.baselinemag.com/analytics-big-data/slideshows/surprising-statistics-about-big-data.html

- Trifacta's position within a Hadoop environment
- Trifacta's scalability in an Hadoop infrastructure
- Data security policy enforcement
- Metadata lineage as part of overall Data Governance
- Operationalization of Trifacta jobs

The Data Lake Concept

The majority of analysis initiatives leveraging Hadoop involve using the Hadoop platform as a shared data lake. There are as many definitions and perspectives of data lake architecture as there are expected outcomes for customers implementing a data lake.

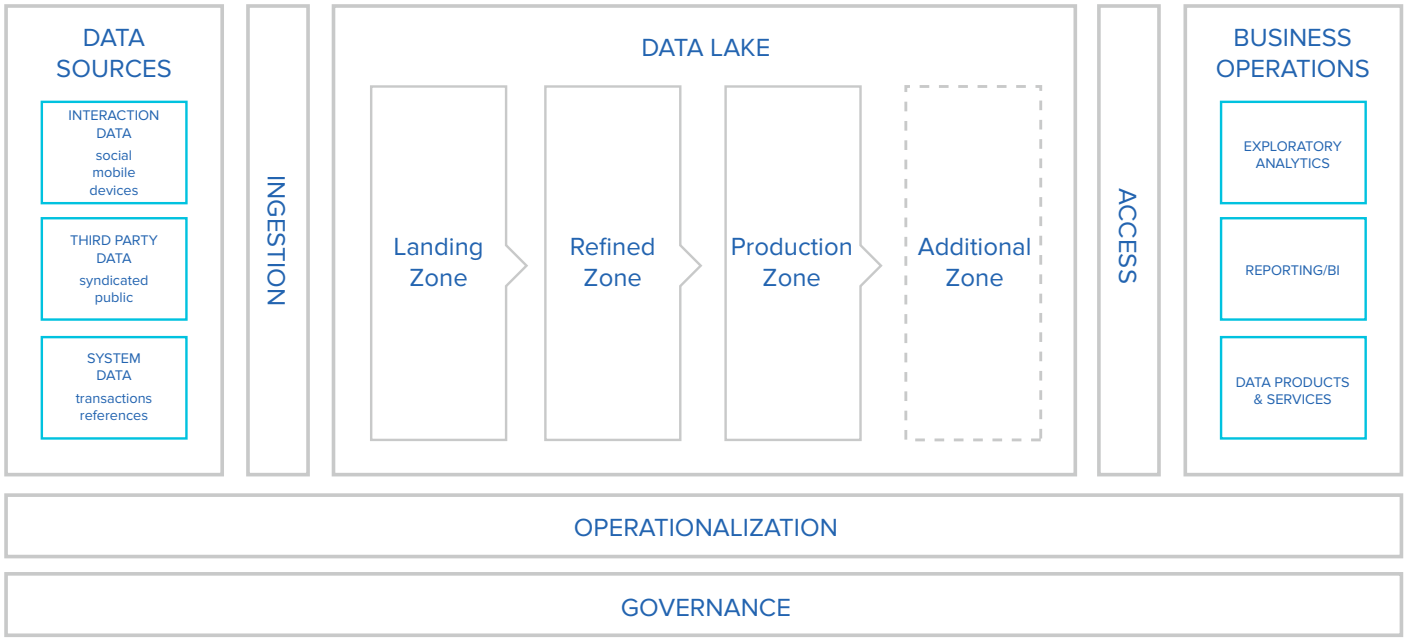
Disparate Data Sources

However, the most basic function of a data lake is that it enables the storage of disparate structured and unstructured data sets (e.g. transactional, interactive, social) in their native formats for later parsing and processing, rather than enforcing all-or-nothing integration/structuring up front as required in a traditional data warehouse.

Preserving the native form of the data also helps maintain data provenance and fidelity so that different analyses can be performed based on different contexts. By breaking the rigidity of traditional data warehouses, data lakes enable a new set of users to access information and perform exploratory analysis that was never possible before.

Data Lake Components

By extension, data lakes are often seen as a logical concept that complements traditional analytical technologies such as RDBMS, ETL and BI tools.



In the data lake architecture highlighted in the figure on the previous page, customers often combine these core components.

Data sources:

- Business Systems Data: CRM, ERP, etc.
- Third party data: Weather, Demographic, etc.
- Interaction data: Product usage, weblogs, etc.

Ingestion:
Technology to access, transfer and load data in real-time and batch into the data lake.

Data lake:
A mature data lake typically includes at least three separate refinement zones supporting different stages of analytics work:

- **Landing zone**
  This type of zone will contain the source data as is, with no transformation, such as a raw log file or a binary files coming from a mainframe.

  In the landing zone, some sub-zones can co-exist, such as a zone with pre-processed raw data into human readable format. The initial landing zone is often managed by the IT organization which automates the data lake ingestion process.

- **Refined zone**
  This is the place where data can be discovered, explored and experimented with for hypothesis validation and tests.

  It usually includes private zones for each individual user and a shared zone for team collaboration. It is often seen as a sandbox with minimal security constraints where end users can access and process the data they want with light automation.

Production zone:
This is where clean, well structured data is stored in the optimal format to inform critical business decisions and drive efficient operations.

It often includes an operational data store that feeds traditional data warehouses and data marts. This is a zone that has strict security restrictions for data access and automated provisioning of data where end users only have a read access.

Consumption:
This is a component that publishes the data proactively (e.g. alerts, next best offer) or on-demand (e.g. SQL, file export) to the business insight layer, optimized for the consuming application.

Business Operations:
Composed of the BI tools or statistical tools to consume the data for direct insight of for the creation of data products and services, it's another location where users can explore processed data for hypothesis validation.

Operationalization:
With data lakes, a large place is given to self-service exploration and analytics. However, there comes a time when IT has to automate and operationalize their users' work. Operationalization ensures report updates and analytics delivery processes are repeatable.

Governance:
Includes the processes and the technology to ensure that the data is classified, used securely and complies with applicable regulations.

Data Wrangling for the Data Lake

The term “data wrangling” is often used to describe the data preparation work done by non-technical users, such as business and data analysts. However, as we will see in the description of the various usages of data wrangling, data engineers are increasingly adopting this technology to facilitate their work and improve collaboration with business users.

Self-service data preparation, or what we call “data wrangling,” is the process of converting diverse data from their raw formats into a structured and consumable format for business intelligence, statistical modeling tools or machine learning.

The Trifacta Experience

What is unique with Trifacta’s solution is that this overall experience is built for non-technical users, making data wrangling intuitive, efficient and even enjoyable. Trifacta automatically discovers the data, structures it in a familiar grid interface, identifies potential invalid data and suggests the best ways to clean and transform the data.

Trifacta learns from the user’s interaction, providing immediate feedback to the user, to better guide him through structuring, enriching and validating the data at scale so it can be published with confidence to the next stage of the analytical process.

Situating Trifacta in Hadoop

Trifacta sits between the data storage and processing layer such as Hadoop and the visualization or statistical applications used downstream in the process.

As pictured in the figure below, data wrangling happens in various places within an analytical workflow.

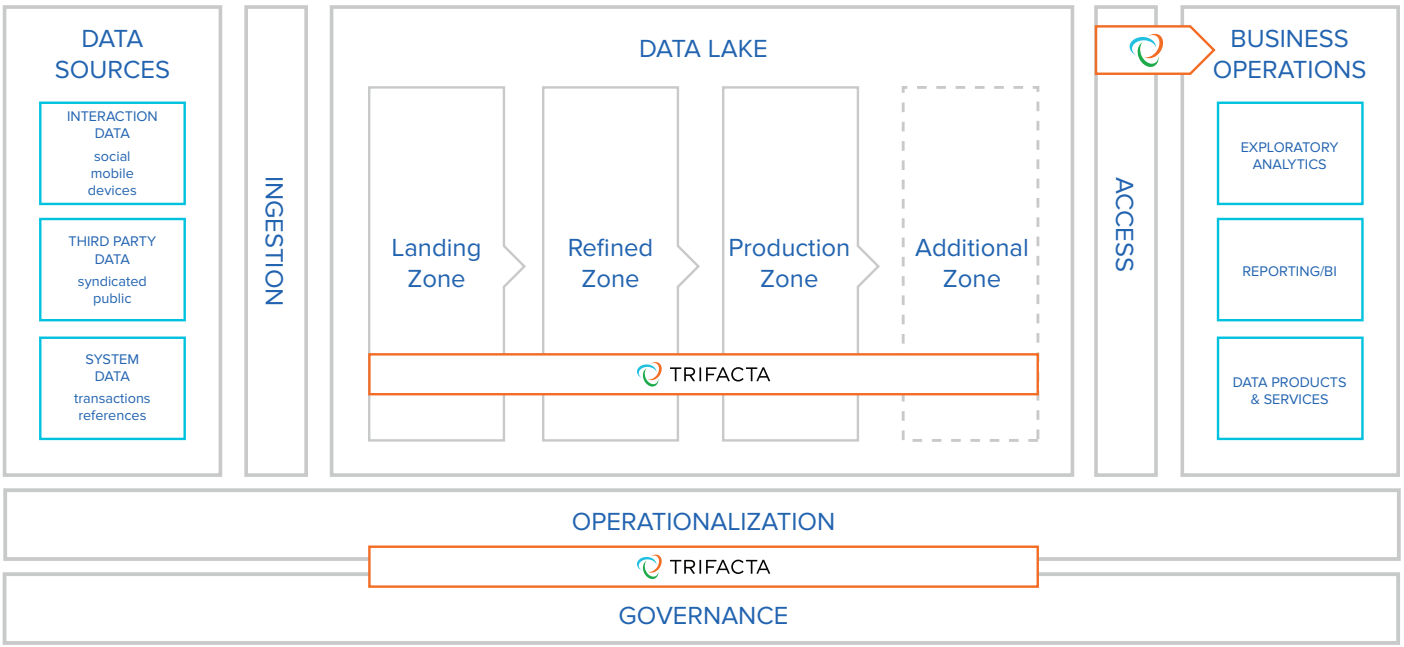
Between zones:
Data wrangling in the data lake typically occurs within a zone or is the process for moving between zones. Users may access raw and refined data to combine and structure it for their exploratory work or for defining new transformation rules they want to automate on a regular basis.

Trifacta can also be used for lightweight ingestion bringing external data sources (e.g. Excel, and relational data) to augment data already in the data lake for the purpose of exploration and cleansing.

Consumption:
Wrangling often occurs in the production zone to deliver data to the business insight layer. This can occur using SQL-based BI tools, or by exporting the data in a file format (eg. CSV, JSON, Tableau Data Extract format) for further usage with analytics tools including Tableau, SAS or R.

Operationalization:
In addition to the actual work of transforming data, Trifacta is also used within the operationalization layer where teams can regularly schedule wrangling jobs and control their execution.

Governance:
Trifacta augments existing data governance processes on the data lake by logging all data access, transformation, and interaction within the solution and makes that data available to data lineage tools so administrators can understand the provenance of data.



Common Use Cases for Data Wrangling

Trifacta's usage can be classified into three major scenarios (with variances), which may require a form of automation that can be organized by a business team or an IT team. Here is a summary of three common use cases for data wrangling.

1. Use-Case: Self-Service Automation

For these types of initiatives, the business teams manage the analytical process, from initial data ingestion to the eventual data consumption, including the data preparation process. Often, their end-goal is to create a "master report" for compliance purposes or to aggregate disparate data. In these initiatives, the IT organization is responsible for setting up the data lake and data ingestion so that, from there, the business team can handle their data requirements and schedule data preparation tasks without IT involvement.

Customer Example
PepsiCo needed to optimize their retail sales forecasts, which combine retailer point-of-sales (POS) data with internal transaction information. For PepsiCo, the major challenge is that each retailer provides sales information in different formats through automated report generations or emails attachments. With Trifacta, the business analyst team assumes ownership over the ingestion of the retailer data into PepsiCo's data lake, can explore and define how the data should be transformed, and execute jobs on-demand or by scheduling to deliver a consumable outcome to their downstream applications or processes.

2. Use-Case: Preparation for IT Operationalization

In this scenario, the data specialists—usually data analysts or data engineers—would design the preparation work themselves, test, validate, and run the rules at scale toward the desired outcome. Once operational workflows are created by end-users, they are able work with the IT organization to take their work and integrate it into enterprise workloads using Trifacta's job scheduling capabilities. The IT team will integrate within a broader process taking care of dependencies, on-going operations and monitoring.

Customer Example
A large European bank needed to extract chat logs from their website to improve customer service, as well as analyze product and service needs. The bank used Trifacta to transform these complex formats into discrete attributes for a broader customer 360 initiative, which incorporated additional data channels. In this case, the teams provides their IT organization with the data wrangling rules they've created so that IT can combine the various data flows consistently.

3. Use-Case: Exploratory Analytics

This use case, by definition, doesn't mandate to operationalize the outcome of the wrangling process. Trifacta is used on an ad-hoc basis for data exploration, case investigation, third party data discovery, to validate hypothesis, investigate data to discover patterns, or generate data sets for data scientist modeling.

Customer Example
A well-established marketing analytics provider aggregates and examines client data to provide analytic outputs for their customers to measure, predict and optimize the impacts of marketing efforts. Each customer has different data sources and formats, but Trifacta helps expedite the discovery and transformation process of their client data to create structured and clean datasets.

Trifacta's Integration with Hadoop

Since its day of inception, Trifacta was engineered for Hadoop, leveraging each single aspect of the technology to scale with the platform and comply with its security components. This strategy is reflected in our close partnerships with Hadoop vendors to support the latest distributions and improvements, the most recent being the authentication and resource usage authorization, in addition to leveraging native storage techniques and formats, as well as the processing power of Hadoop.

Trifacta obscures the complexity of Hadoop to maximize Hadoop adoption, while also enabling IT teams with seamless integration to their existing Hadoop relative management processes.

Trifacta integrates with Hadoop through standard and native Hadoop interfaces to ensure:

- Secure Authentication using Kerberos with Secure Impersonation;
- Role-based Authorization conforming to Sentry and Ranger policies.
- Physical data access through HDFS API and Hive JDBC driver.
- Scalable execution leveraging Spark and Pig on YARN.

"Trifacta obscures the complexity of Hadoop to maximize Hadoop adoption, while also enabling IT teams with a seamless integration to their existing Hadoop relative management processes."

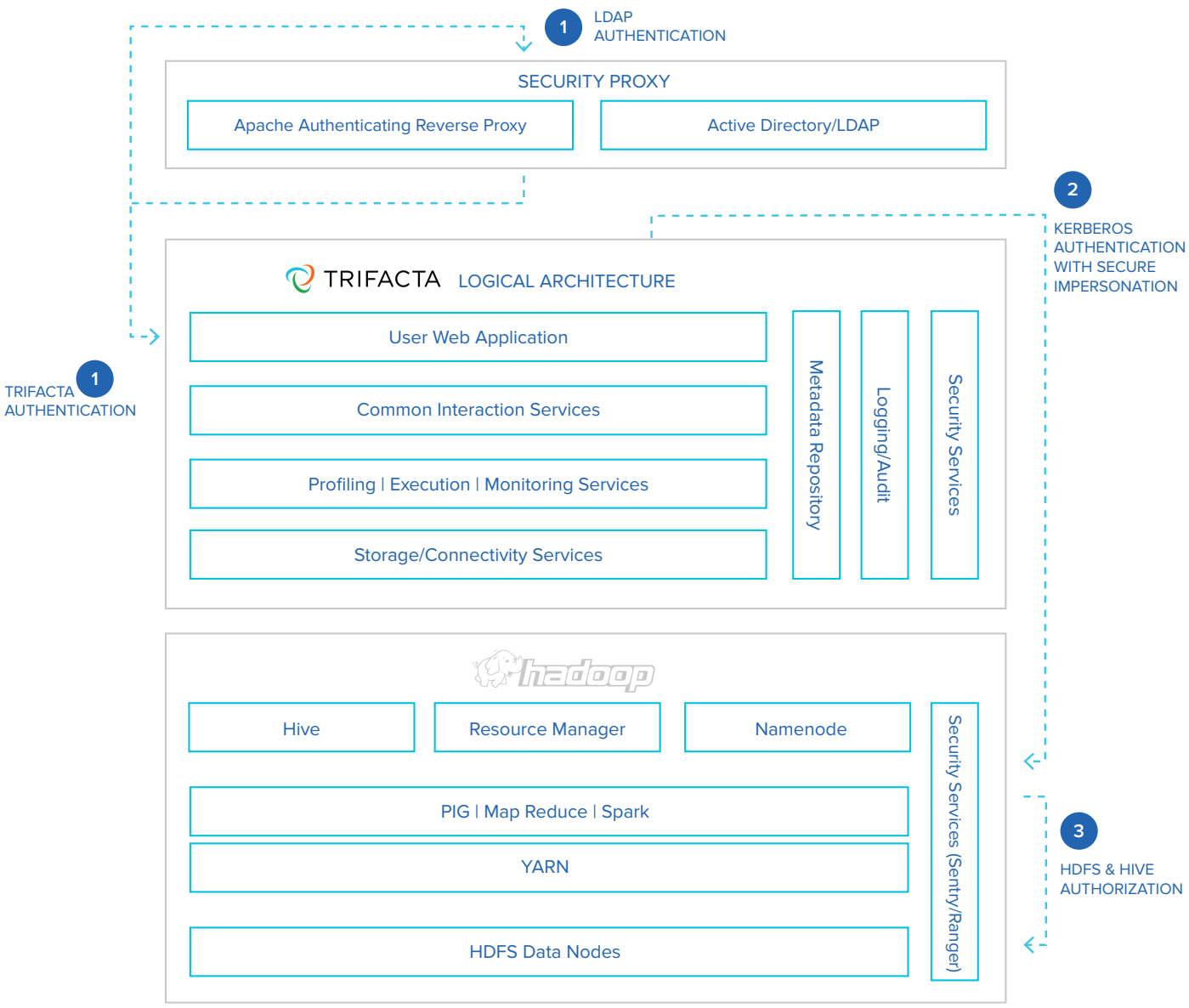


FIG 3. HIGH-LEVEL OVERVIEW OF AUTHENTICATION AND ACCESS CONTROL IN TRIFACTA

Ensuring a Fluid, Consistent User Experience

By definition, big data projects often mean processing gigantic datasets. At such volumes, a balance must be struck between providing a consistent and fluid user experience, as well as enabling the preparation of massive data sets at scale.

Representative Sampling

To reach this balance, Trifacta has built upon the analytics best practice of leveraging representative sample data sets in the interactive design phase, followed by the application of transformation rules to the whole data set at runtime.

The key benefit of this approach is the interface's immediate responsiveness in showing users the impact of their proposed data transformation. This feedback comes in two forms: first, we can show the resulting data values from the transformation (preview). Second, we show the distributional statistics of the output values (both in terms of type validity and in terms of distributional aspects like cardinality). As the user goes through the preparation process, the sample data can be regularly refreshed, ensuring that they have an accurate representation of the entire data set.

Transformation Logic

Ultimately, when the data transformation rules comply to the user's requirements, the wrangling script can be applied to the entire data set. For small-to-medium sized datasets, transformation logic can be executed in Trifacta's single node engine. For intermediate-to-large datasets, transformation logic can be compiled in Hadoop using Spark or MapReduce to process data across an entire cluster. After the data has been transformed at scale, Trifacta provides the option to generate distributional statistics for each column, providing feedback about data distribution, types, missing values, and errors. Data records containing errors and missing values are shown in this at-scale, distributional profiling interface and the user may bring these records into the interactive design interface as a new sample to further refine the transformation rules.

Connectivity

For the modern data analyst, data exists everywhere: in Hadoop, relational databases, the cloud, and on their desktops. With Trifacta, analysts can be productive with their data regardless of where it lives. To enable productivity with data that lives in multiple places, Trifacta has built support for connecting to enterprise sources such as Hive, S3, and relational sources through JDBC. In addition, Trifacta works with semi-structured sources such as JSON and XML and can work with desktop-based data in Excel, CSV and TXT among other formats.

Publishing Data

After wrangling, the structured outputs can be brought into BI tools for consumption and further analysis. Trifacta can store the results of the data preparation processes into Hadoop Hive tables using Avro or Parquet, which can also be accessed with BI tools through JDBC connectivity.

In addition, the result can be exported as a CSV file, a Tableau TDE (Tableau Data Exports) file, or in JSON format. Trifacta also supports the ability to publish wrangling outputs set to Redshift, enabling users to create a new Redshift table and upload the results from the web interface. JDBC connectivity enables Trifacta to publish outputs to the RDBMS of choice.

Operationalization

Trifacta offers a Command Line Interface (CLI) enabling programmatic control over a variety of operations on the platform. The command line can be used to manage the following tasks:

- Run a job
- Check job status
- Publish a completed job result to Hive or Redshift

The CLI is often used within operating system scripting languages and integrated with enterprise schedulers to integrate and automate certain Trifacta tasks within a bigger IT operationalization process.

For organizations that leverage enterprise schedulers such as Chronos or Tidal, Trifacta can support scheduling jobs through those frameworks.

Performance and Scalability Considerations

Trifacta is built on top of Hadoop's open and scalable platform, enabling the solution to support a variety of different scenarios. For different wrangling and data prep needs, it's often required to tune the system for maximum performance and scalability.

There are usually a few key considerations when planning out a performant environment to support enterprise data wrangling:

- Volume of data
- Data access interface (eg. Files, Hive, APIs)
- Number of concurrent users
- Type of wrangling operations and use cases (eg. structuring, blending, profiling)
- Hardware configuration (number of nodes, network setup)

Governance and Security for Data Wrangling

As part of an overall governance solution, Trifacta focuses on two key areas: secure data access, and metadata/data lineage management. In each of these capabilities Trifacta follows the approach of embracing the existing Hadoop ecosystem by integrating with the enterprise standard frameworks and products.

Data Security

Granting analysts with access to diverse data sources further up the process empowers them to become more informed and share those insights with others. However, this does create a new set of challenges for platform administrators to effectively manage the security and accessibility of the data.

Major Hadoop distributions have a focused on ensuring enterprise-grade security for user authentication and data access authorization. Trifacta's approach leverages native Hadoop security authorization so that the level of security that a team has configured on Hadoop is automatically inherited by Trifacta, restraining a user to only access the data that Hadoop authorizes. This is accomplished through Trifacta's support for Kerberos authentication[3] and single-sign-on through LDAP[4].

For data access authorization, Trifacta supports the robust security frameworks found in the Hadoop ecosystem, namely Apache Sentry[5] and Ranger[6]. When Trifacta users access HDFS or Hive data, the access request is validated and adheres to the policies defined in Sentry or Ranger. In addition, when needed data generated by Trifacta jobs is encrypted using the HDFS encryption services now standard in Hadoop distributions. This seamless integration ensures that user access to data is consistent and secure, simplifies security management for both the Hadoop and Trifacta administrator, and provides a hassle-free experience for the end user.

Sources

[3] http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SecureMode.html
[4] https://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol
[5] http://sentry.apache.org/
[6] http://ranger.incubator.apache.org/

Metadata Lineage

In a data lake approach, a common piece in a governance solution is to provide an end-to-end view of how users access, transform and publish wrangled data outputs.

To solve traceability into the transformations applied, Trifacta leverages Wrangle, a domain specific language for data transformation to track the lineage of each transformation step applied to the data.

```
splitrows col: column1 on: '\n'
split col: column1 on: '\t' limit: 3
extract col: column2 after: `<` before: `/`
drop col: column2
split col: column3 on: '\/' limit: 2
split col: column4 on: ':' limit: 2
split col: column5 on: '\/' limit: 2
header
split col: DATETTIME on: `{upper}`
```

FIG 4. WRANGLED CDR DATA IN TRIFACTA,
TRIFACTA TRANSFORMATION STEPS TO PREPARE CDR DATASET

Data Wrangling Lineage

Hadoop distributions provide a variety of metadata frameworks to record system and end-user touchpoints tracking how data is accessed and transformed in Hadoop. Popular examples include the Atlas project developed by Hortonworks and Cloudera Navigator. These solutions can be augmented with Trifacta's wrangling metadata to provide comprehensive lineage of data as it passes through the wrangling process.

Trifacta offers an API to integrate its metadata with these frameworks. For example, Trifacta and Cloudera have collaborated to integrate bi-directionally metadata with Cloudera Navigator. When a data analyst wrangles data in Trifacta, he can easily publish wrangle metadata to Cloudera Navigator to augment the metadata already within the framework. From within Cloudera Navigator, users can then search for Trifacta metadata and use Cloudera Navigator's lineage view to easily see Trifacta's wrangling steps associated with a variety of transformation input and output datasets in a Hadoop cluster.

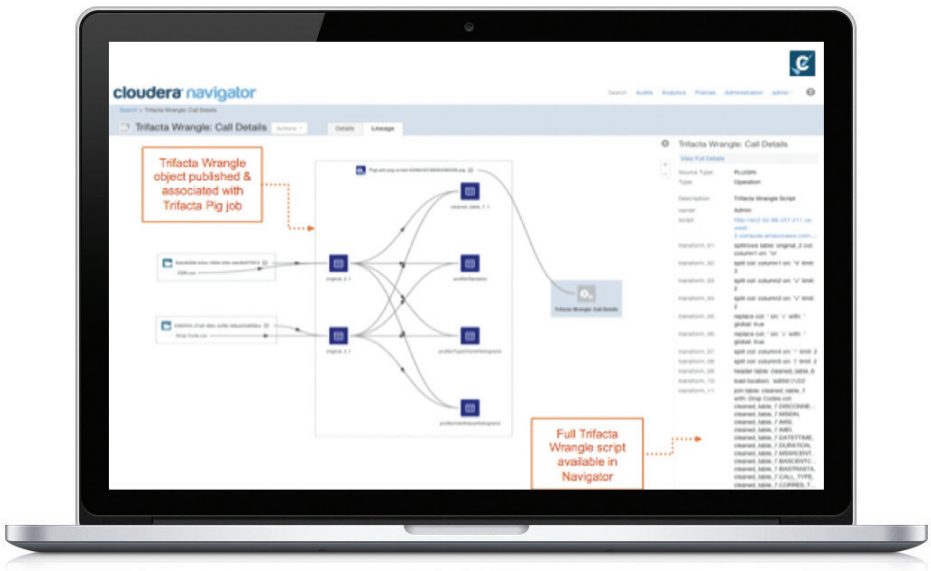


FIG 5. CLOUDERA NAVIGATOR

Data analysts greatly benefit from this integration because they have transparent access to data quality and data wrangling logic automatically captured and linked to the technical metadata that's already stored within Cloudera Navigator—without ever having to leave the Trifacta interface. System architects benefit from being able to search, trace and annotate all of the transformation jobs, HDFS files, Hive tables touched by Trifacta users with a human readable description of the transformation logic.

Conclusion

Trifacta's data wrangling solution is critical to accelerating business adoption of Hadoop by empowering non-technical analysts to explore and prepare diverse data in a self-service fashion. For seamless deployment, Trifacta is tightly integrated with the various Hadoop components to best leverage the platform's storage and processing power while ensuring enterprise grade support for data security and governance requirement.

Trifacta has been designed from the ground-up with an enterprise readiness perspective in-mind and therefore can be deployed with confidence in any organization's Hadoop environment.

About Trifacta

Trifacta, the leading data wrangling solution for exploratory analytics, significantly enhances the value of an enterprise's big data by enabling users to easily transform and enrich raw, complex data into clean and structured formats for analysis. Leveraging decades of innovative work in human-computer interaction, scalable data management and machine learning, Trifacta's unique technology creates a partnership between user and machine, with each side learning from the other and becoming smarter with experience. Trifacta is backed by Accel Partners, Greylock Partners and Ignition Partners.