TRIFACTA

575 Market St, 11th Floor
San Francisco, CA 94105

trifacta.com
844.332.2821

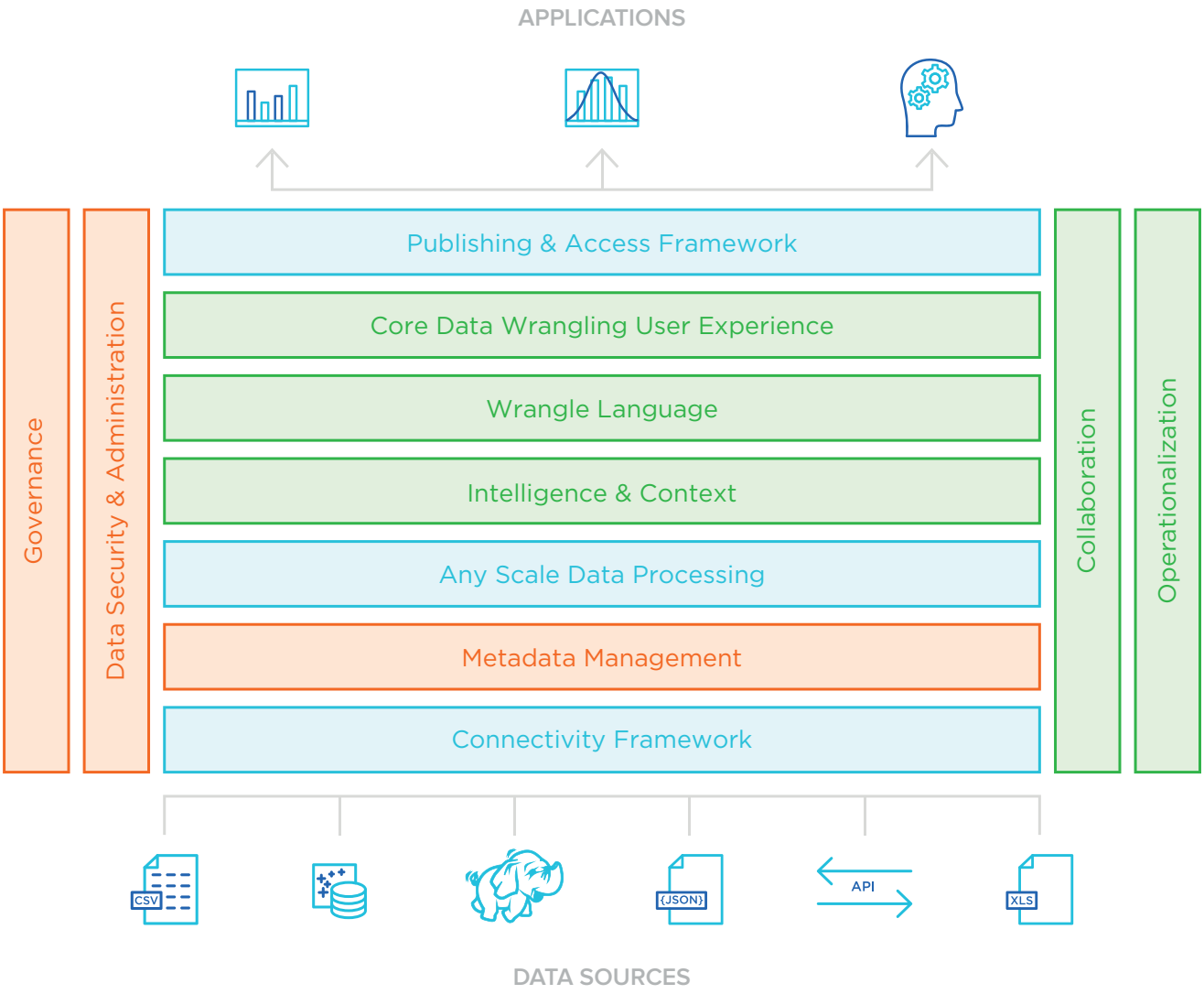
WHITEPAPER

Trifacta Architecture: An Intelligent & Interoperable Data Wrangling Platform

Trifacta was born out of the belief that the people who know the data best should be able to wrangle it themselves. Instead of a rigid and highly-technical data transformation process, we sought out to build an intuitive user experience and workflow for every user within an organization. This required years of research at UC Berkeley and Stanford across machine learning, human-computer interaction and parallel processing, a deep consideration for the analytics ecosystem in its entirety, and an enduring empathy for end-user analysts, which eventually produced the world's leading data platform. Today, Trifacta's architecture is validated by leading industry analysts and in production at industry-leading organizations around the world.

Trifacta sits between data storage and processing environments and the visualization, statistical or machine learning tools used downstream. As a best-in-breed technology, Trifacta has been architected to be open and adaptable so as the technologies upstream and downstream change, the investments and logic created in Trifacta are able to utilize those innovations.

What follows are foundational aspects of Trifacta.



Ecosystem & Extensibility

Connectivity Framework

Trifacta maintains a robust connectivity and API framework, enabling users to access live data without requiring them to pre-load or create a copy of the data separate from the source data system. This framework includes connecting to various Hadoop sources, Cloud services, Files (CSV, TXT, JSON, XML, etc.) and relational databases. All of these connectors support governance and security features—roles and permissions, SSL, Kerberos Auth (SSO) and impersonation.

Any Scale Data Processing

Using Trifacta's Intelligent Execution Engine, every transformation step defined in the user interface automatically compiles down into the best-fit processing framework based on data scale. Trifacta can transform the data on-the-fly in the application or compile down to Spark, Google DataFlow, or our in-memory engine, Photon. The platform natively supports all major Hadoop on-premise and cloud platforms. With this model, Trifacta can handle any scale.

Publishing & Access Framework

Trifacta maintains a robust Publishing and Access framework. Outputs of wrangling jobs are able to be published to a variety of downstream file systems, databases, analytical tools, file and compression formats. Trifacta has deep API and bi-directional metadata sharing with a variety of analytics, data catalog and data governance applications. This enables users to share context and work between Trifacta and the external applications they're leveraging through native integration.

User Experience

Core Data Wrangling User Experience

Trifacta leverages the latest techniques in data visualization, machine learning and human-computer interaction to guide users through the process of exploring data and preparing data. Interactive Exploration presents automated visualizations of data based upon its content in the most compelling profile. Predictive Transformation converts every click or select within Trifacta into a prediction—the system intelligently assesses the data at hand to recommend a ranked list of suggested transformation for users to evaluate and edit.

Intelligence & Context

Trifacta learns from data registered into the platform and how users interact with it. Common tasks are automated and users are prompted with suggestions to speed their wrangling. The platform supports fuzzy matching, enabling end users to join data sets with non-exact matching attributes. Data registered in Trifacta are inferred to identify formats, data elements, schemas, relationships and metadata. The platform provides visibility into the context and lineage of data—both inside and outside of Trifacta.

Wrangle Language

Core to Trifacta's differentiation is the platform's Domain Specific Language Wrangle enabling users to abstract the data wrangling logic they're creating in the application from the underlying data processing of that logic. Advanced users can create more complex wrangling tasks including window functions, user defined functions. Every step defined in Trifacta's Wrangle language makes up a data preparation recipe or set of steps created in Trifacta that can be set into a repeatable pipeline.

Collaboration

Within Trifacta, users can share reusable data preparation logic and dataset relationships, which lets them leverage and build upon each other's efforts. Multiple users can contribute to a single project, which parallelizes workflows, allows different degrees of participation, and speeds up time to completion. Datasets and data preparation steps can also be integrated with 3rd party applications through Trifacta's API. Additionally, preparation steps can be exported and shared outside Trifacta.

Operationalization

Trifacta's operationalization features introduce the ability for data analysts to schedule and monitor workflows that run jobs at scale in production, while still providing the traceability and access control for IT. Every data preparation recipe or set of steps created in Trifacta can be set into a repeatable pipeline according to hourly, daily, weekly schedules or the time period defined by the user. Individual recipes can makeup broader pipelines that make up multiple datasets and recipes.

Enterprise Data Governance & Security

Metadata Management

Trifacta has support for enriching data with geographic, demographic, census and other common types of reference data. Common taxonomies and ontologies are automatically recognized, such as geographic and time-based content, as well as data format taxonomies for nested data structures like JSON and XML. The platform is also open/extensible through APIs, giving customers and partners the ability to seamlessly integrate additional data sources and targets.

Governance

Collaborative Data Governance refers to features within Trifacta that provide extensive support for open source and vendor-specific security, metadata management and governance frameworks. This approach gives organizations the visibility and administration over the data wrangling users are performing. Trifacta supports user hierarchies across roles determining data access and user functionality within the application. Administrators and data stewards are able to manage platform authentication and security at various user hierarchy levels.

Data Security & Administration

Trifacta provides end-to-end secure data access and clear auditability that comply with the stringent requirements of enterprise IT. The platform provides support for encryption, authentication, access control and masking. Trifacta's differentiated approach to security focuses on providing enterprise functionality (such as SSO, impersonation, roles and permissions) while balancing extensive security framework integration with existing policies. Customers can integrate Trifacta into what's already working for them without having to support a separate security policy.

About Trifacta

Trifacta, the global leader in data wrangling software, 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 with self-service data preparation. 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, Cathay Innovation, Greylock Partners and Ignition Partners.

For more information on Trifacta's architecture and how organizations are generating new sources of business value through data wrangling, please visit **trifacta.com**.