



## **DataForge — Data Engineering Platform**

**AI-assisted data engineering that turns raw operational data into analytics-ready, governed datasets.**

DataForge is a data engineering platform designed to make pipelines repeatable, transparent, and trustworthy. It ingests raw source files and operational data in multiple formats, refines them through a structured bronze → silver → gold (medallion) pipeline, and produces clean, modeled, analytics-ready datasets — with full lineage and documentation for every transformation.

The platform combines automated data cleaning, schema and contract enforcement, semantic modeling, and optional AI LLM assistance. It is built for teams that need more than ad-hoc scripts and spreadsheets: they need traceable lineage, version history, and governed data they can defend and reuse.

### **Designed for:**

Data engineering teams, quality and operations analysts, manufacturing data teams, BI and reporting teams, governance and compliance teams, and digital transformation teams.



## AI Context Builder

Start with a plain business goal

Gemini

### Business goal

Analyze production orders to identify which supplier material lots, labor teams, and shift configurations correlate most heavily with quality non-conformances and cycle time delays.

Generate Context

### Business Problem

### Scope & KPIs

### Who uses this?

Pipeline Mode:  Deterministic (Demo Mode)  Dynamic LLM (AI Generated)

New Session

### LLM Settings

gemini

API key

Connect

### Ingestion



Upload data files

Quality

--

Awaiting gold output

Workspace Schema Map Visual Insights Knowledge Graph User Guide

Active Rows Gold Columns Steps Contract

#### LINEAGE & LOGIC

### What happened to the uploaded files?

Single active-file output

Upload files to see whether the app is processing one active file or building a unified Gold dataset.

#### ACTIVE RAW FILE

### Bronze Column Review

Propose

#### SILVER PROCESSING

### Transformation Queue

Apply

#### GOLD OUTPUT

### Active File Preview

Contract Export CSV Excel

No rows loaded

# Set up Business goal, problem, Scope & KPI

LLM Settings

openai

.....

Connect

Ingestion

04\_labor\_hours\_TIMESHEET\_sensitive.csv active

01\_production\_orders... 3235 rows / 15 cols

02\_cycle\_time\_EVENT... 18149 rows / 16 cols

03\_downtime\_events... 5482 rows / 16 cols

04\_labor\_hours\_TIME... 42150 rows / 16 cols

05\_quality\_NC\_mixed... 2370 rows / 13 cols

06\_supplier\_lots\_mater... 995 rows / 11 cols

07\_maintenance\_work... 1600 rows / 11 cols

08\_sensor\_iot\_reading... 50000 rows / 8 cols

09\_asset\_master\_DUPLIC... 79 rows / 6 cols

10\_shift\_calendar\_dime... 540 rows / 5 cols

Quality

---

Awaiting gold output

Production Oee High impact downtime hours by reason and mould Mould + shift + week across uploaded production data Edit

**Business Context Applied** Used in workflow intent, dashboard focus, insight prompts, contract metadata, and pipeline export.

<b>PROBLEM</b> Downtime is increasing cycle time and reducing production availability. Factory teams need to identify the recurring downtime drivers across reason, unit, shift, and week.	<b>BUSINESS IMPACT</b> High	<b>SCOPE</b> Mould + shift + week across uploaded production data / Single Site	<b>PRIMARY KPI</b> downtime hours by reason and mould
<b>SECONDARY KPIS</b> Top downtime reason by reason, Downtime hours by unit, Downtime trend by period	<b>CURRENT STATE</b> Downtime events are captured, but recurring loss drivers are not prioritized consistently.	<b>TARGET STATE</b> Top downtime drivers are quantified weekly by mould, reason, and shift so improvement actions can be assigned.	<b>UPDATE FREQUENCY</b> Weekly
<b>PRIMARY USER</b> Factory Manager	<b>USER ACTION</b> prioritize the largest recurring downtime reasons and assign corrective actions to responsible teams	<b>EXPECTED OUTCOME</b> reduce cycle-time loss and improve production availability by focusing resources on the highest-impact downtime patterns	

Workspace Schema Map New Visual Insights Ready Knowledge Graph Semantic User Guide

Active Rows 42150 Gold Columns 66 Steps 19 Contract --

**LINEAGE & LOGIC** Generic multi-file Gold workflow

What happened to the uploaded files?

<b>ACTIVE RAW PREVIEW</b> 04_labor_hours_TIMESHEET_sensitive.csv Bronze Profile still follows the active raw file selected in the left file list.	<b>GOLD OUTPUT</b> generic_multi_file_gold.csv Gold output is produced by Generic Multi-File Gold Workflow, not by the active file alone.	<b>LLM USAGE</b> Used Dynamic LLM mode is active (provider=openai); deterministic fallbacks remain available.
<b>GOLD GRAIN</b> one row per production orde...	<b>SUGGESTED INTENT</b> Selected explicit entity + scope + period grain from identifier, scope dimension, and time-window columns.	

Sources

<b>PARENT_FACT</b> 01_production_orders_CORE... 3235 rows / 15 cols / key: Production Order	<b>CHILD_TRANSACTION</b> 02_cycle_time_EVENTS.csv 18149 rows / 16 cols	<b>CHILD_TRANSACTION</b> 03_downtime_events_SEMICO... 5482 rows / 16 cols	<b>CHILD_EVENT</b> 04_labor_hours_TIMESHEET_se... 42150 rows / 16 cols
<b>CHILD_TRANSACTION</b> 05_quality_NC_mixed_keys.csv 2370 rows / 13 cols	<b>SUPPLIER_BATCH_TRANSACTION</b> 06_supplier_lots_materials.csv 995 rows / 11 cols	<b>CHILD_TRANSACTION</b> 07_maintenance_work_orders... 1600 rows / 11 cols	<b>TIME_SERIES</b> 08_sensor_iot_readings_large.csv 50000 rows / 8 cols
<b>DIMENSION</b> 09_asset_master_DUPLICATES.... 79 rows / 6 cols	<b>CALENDAR_REFERENCE</b> 10_shift_calendar_dimension.csv 540 rows / 5 cols		

Bring in your uncleaned CSV, Excel, or JSON files.

## Unified Gold Preview

Contract

Export

CSV

Excel

PRODUCTION_ORDER	BLADE_ID	SITE	FACTORY_ID	MOULD	AREA	SHIFT	TEAM
PO-WST-100000	700000	WST	F002	M02	Trim		D
PO-LHR-100001	700001	LHR	F003	M06	Assembly	Weekend	D
PO-WST-100002	700002	WST	F002	m05	Paint	Weekend	2
PO-WST-100003	700003	WST	F002	M04	Assembly	Day	1
PO-LHR-100004	700004	LHR	F003	M05	Infusion	Night	B
PO-NAK-100005	700005	NAK	F001	M11	Assembly	Weekend	2
PO-LHR-100006	700006	LHR	F003	M06	Infusion		3
PO-WST-100007	700007	WST	F002	M03	Trim		2
PO-WST-100008	700008	WST	F002	m06	Infusion	Day	B
PO-NAK-100009	700009	NAK	F001	M04	Infusion	Day	3
PO-NAK-100010	700010	NAK	F001	M09	Trim	Day	B
PO-LHR-100011	700011	LHR	F003	M11	Paint		1

## Data Governance

Auto-scanned on upload. Click to re-scan or run deeper analysis.

PII: HIGH

Re-scan PII

Check Schema Drift

Run Quality Rules

## PII Scan

RISK: HIGH 3 finding(s)

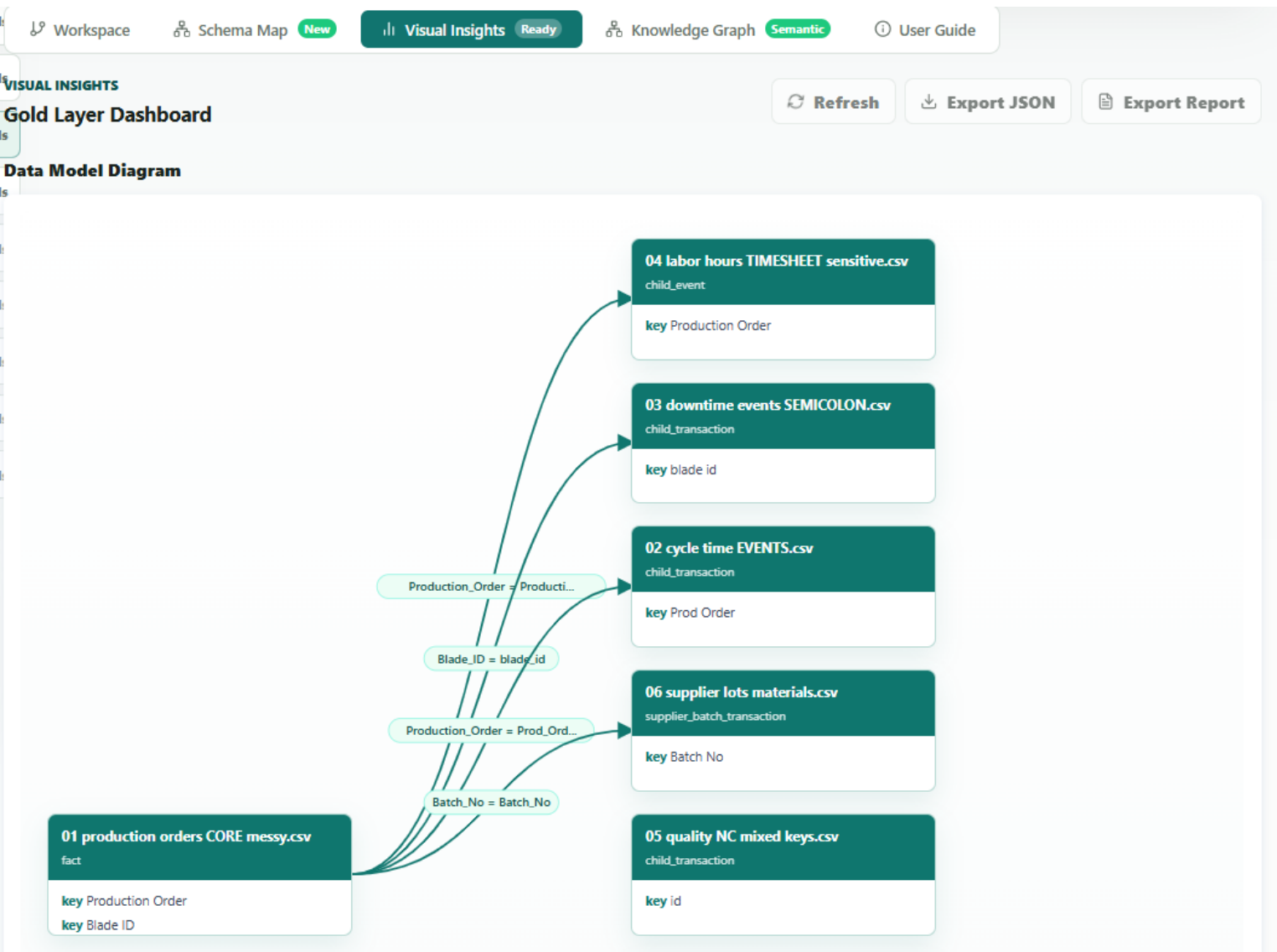
PII Columns: Employee Email, ct\_event\_id

COLUMN	TYPE	SEVERITY	METHOD
ct_event_id	passport	medium	value_pattern
Employee Email	email	high	column_name
Employee Email	email	high	value_pattern

## Schema Drift

No schema drift detected

The AI automatically detects column types, identifies errors, and applies deterministic or LLM cleaning rules.



The system discovers relationships across multiple uploaded datasets and recommends safe joins.

Executive



Use LLM if available (openai)

Regenerate

Generated by: llm Provider: openai Model: gpt-4o-mini

Executive Summary

Downtime is significantly impacting cycle time and production availability, necessitating a focused analysis on recurring downtime drivers. Prioritizing these drivers will enable targeted corrective actions to enhance operational efficiency.

- Make shorter
- Make more executive
- Copy summary
- Export insight JSON

Key Findings

**FINDING**  
**Top Downtime Reason Identified**

The leading cause of downtime is 'Quality hold', accounting for 643,725 minutes of downtime.

**FINDING**  
**Mould-Specific Downtime**

Mould 'M12' has the highest downtime at 308,797 minutes, indicating a need for targeted analysis.

**FINDING**  
**Shift Impact on Downtime**

The first shift accounts for 1,618,005 minutes of downtime, representing 41.89% of total downtime.

Risks

**RISK**  
**Inconsistent Prioritization of Downtime Drivers**

Failure to consistently prioritize downtime drivers may lead to unresolved issues and continued production inefficiencies.

**RISK**  
**Potential for Increased Cycle Time**

If downtime drivers are not addressed, cycle time may continue to increase, further reducing production availability.

**RISK**  
**Resource Misallocation**

Without clear identification of downtime drivers, resources may be misallocated, hindering effective corrective actions.

Recommended Actions

**RECOMMENDATION**  
**Prioritize Downtime Reasons**

Establish a systematic approach to prioritize and address the top downtime reasons weekly.

**RECOMMENDATION**  
**Assign Corrective Actions**

Assign specific corrective actions to responsible teams based on the prioritized downtime drivers.

**RECOMMENDATION**  
**Monitor Downtime Trends**

Regularly monitor downtime trends to identify emerging issues and adjust strategies accordingly.

Data Quality Notes

**DATA\_QUALITY**  
**Data Completeness**

The dataset contains 3,200 rows, providing a comprehensive view of production data.

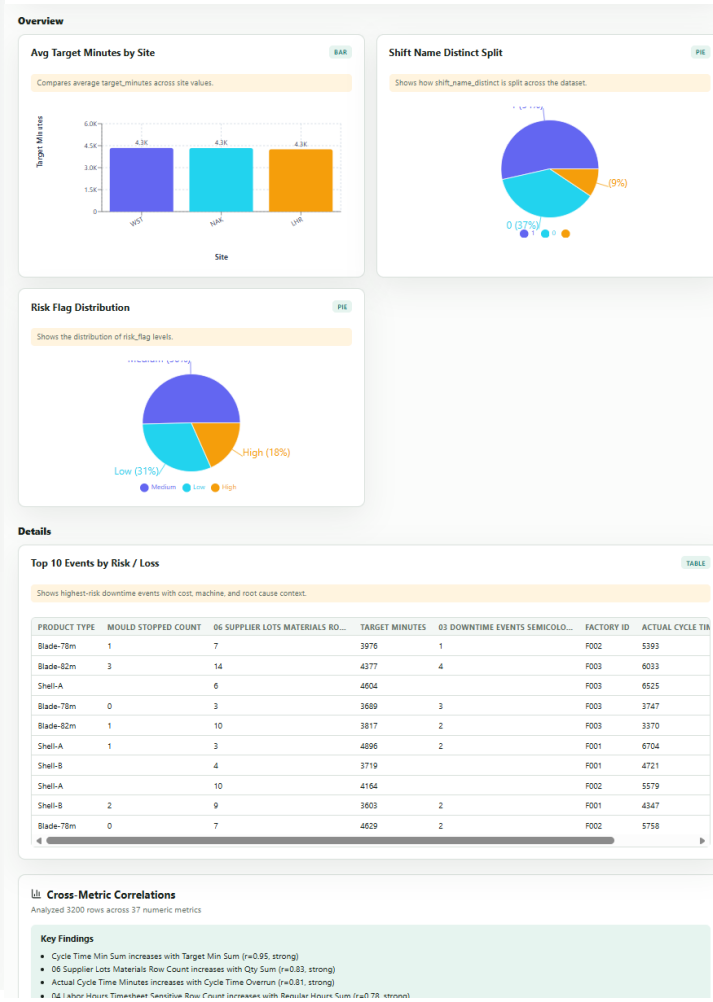
**DATA\_QUALITY**  
**High Confidence in Metrics**

Key metrics have a confidence level of 0.98, indicating reliable data for analysis.

**DATA\_QUALITY**  
**Potential Data Gaps**

Ensure ongoing data collection processes are robust to prevent gaps in future analyses.

Build a unified dataset, auto-generate visual dashboards, and create AI Insights.



SEMANTIC INTELLIGENCE LAYER

Knowledge Graph

Ready

- Ontology
- Knowledge Graph**
- Query
- Business Rules

Build Knowledge Graph

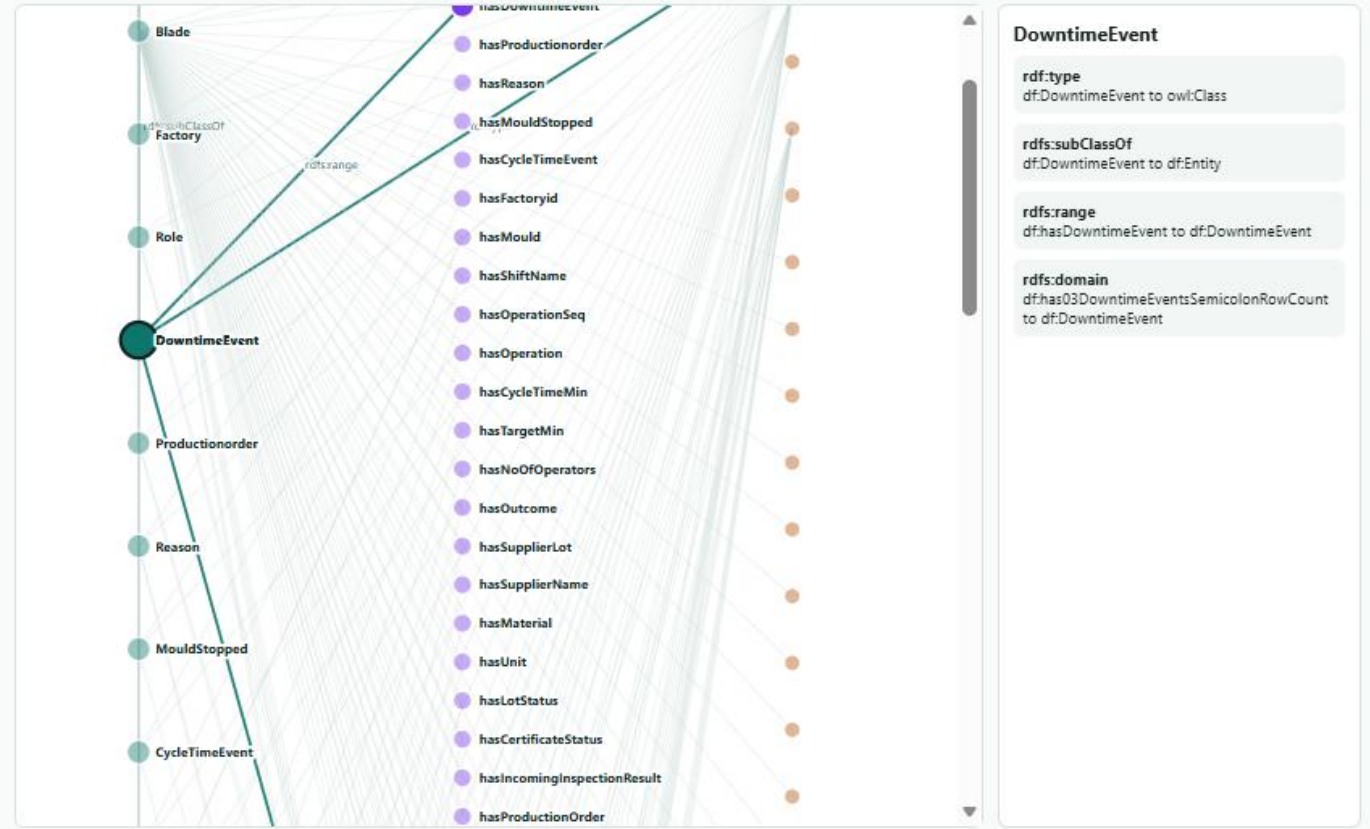
EXPLICIT TRIPLES <b>111724</b>	INFERRED TRIPLES <b>0</b>	ENTITIES <b>2890</b>
-----------------------------------	------------------------------	-------------------------

Blade x 3000

- Graph JSON
- Nodes CSV
- Links CSV
- SVG
- Ontology TTL

Classes Properties Entities

Showing a readable sample of 150 nodes and 258 links. Click a node to inspect its connected triples.



Generates ontology, Knowledge graph, Query and Business rules

> Platform Overview

- > Getting Started
- > Workspace Walkthrough
- > File Upload & Profiling
- > Cleaning & Transformations
- > Multi-File Workflows
- > Gold Dataset
- > Data Contracts
- > Quality Scorecard
- > Export & Downloads
- > Visual Insights Dashboard
- > Knowledge Graph
- > LLM / AI Integration
- > Industry Scenarios
- > Pipeline Modes
- > Data Governance
- > Platform Features
- > Persistent State
- > Additional Workflows
- > Limitations
- > Troubleshooting
- > Glossary
- > FAQ



## Platform Overview

DataForge AI is an automated data engineering platform that transforms raw CSV, Excel, JSON, and TSV uploads into clean, analytics-ready **Gold datasets** — complete with quality scores, data contracts, and interactive dashboards.



### Upload Any Format

CSV, TSV, JSON, XLSX — files are parsed, standardized, and profiled automatically.



### Smart Cleaning

Deterministic transformations — date normalization, null filling, deduplication, unit conversion — proposed and applied in one click.



### Multi-File Intelligence

Upload multiple files and DataForge auto-detects relationships, joins tables, aggregates child data, and builds one unified Gold dataset.



### Data Contracts

Auto-generated YAML contracts with typed columns, quality rules, and DuckDB-compatible SQL checks.



### Visual Insights

Auto-generated dashboards with KPIs, bar charts, line charts, pie charts, and AI-narrated findings.



### Knowledge Graph

Generate or upload an ontology, build semantic triples from the Gold dataset, query with SPARQL, apply rules, and export graph files.



### Optional AI

Works fully without API keys. Optionally connect OpenAI, Anthropic, or Gemini for dynamic SQL, text classification, and insight narratives.

### ⚡ No black boxes.

Every transformation shows its SQL template. Every join shows its logic. Every quality score shows its formula. The pipeline JSON records every step for audit.

# Detailed User Guide