Business Requirements Document: Advanced Data Cleaning Platform

Executive Summary

This Business Requirements Document outlines the specifications for an advanced, highly scalable data cleaning platform designed to process billions of rows with AI-powered capabilities. The platform targets a $27.28 billion market opportunity by 2033, addressing critical gaps in current solutions through simplified pricing, superior performance, and industry-specific AI models. IMARC Dataintelo With a token-based pricing model and three-tier cleaning levels, the platform aims to capture mid-market enterprises currently underserved by expensive, complex competitors.

1. Core Platform Requirements

1.1 Scalability and Performance

The platform must deliver unlimited scalability for processing millions to billions of rows through a distributed computing architecture:

Processing Capabilities:

- **Batch Processing**: Apache Spark 3.5+ for handling up to 100 billion rows per job chaosgenius
- **Real-time Processing**: Apache Flink for sub-second data quality validation chaosgenius +2
- **Performance Targets**: 2-5 second query times for billion-record aggregations ChaosGenius +2
- **Throughput**: 15 million messages per second via Apache Kafka orchestration Confluent +3

Infrastructure Requirements:

- Kubernetes-based microservices architecture with Linkerd service mesh Medium Linkerd
- KEDA event-driven autoscaling with scale-to-zero capabilities SpectroCloud Devtron
- Multi-cloud support (AWS, Azure, GCP) with 99.9% availability SLA
- Horizontal scaling with automatic resource allocation

1.2 Token-Based Pricing System

Three-Tier Cleaning Levels:

1. **Basic Tier** (1 token per million rows):
   - Standard deduplication
   - Data type validation

- Missing value handling
- Format standardization

2. **Advanced Tier** (3 tokens per million rows):
   - AI-powered anomaly detection
   - Fuzzy matching with Jaro-Winkler algorithms
   - Statistical outlier detection
   - Smart column mapping
3. **AI-Powered Tier** (5 tokens per million rows):
   - GPT-4 based data correction Numerous.ai
   - Industry-specific ML models
   - Predictive data quality assessment
   - Synthetic data generation for gaps

Token Management:

- Prepaid token packages with volume discounts
- Real-time usage tracking and alerts
- Automatic top-up options
- Usage analytics dashboard

1.3 Industry-Specific AI Models

Healthcare:

- FHIR data cleaning with BioBERT/ClinicalBERT models Medium
- ICD-10 code validation using MedCodER framework (0.60 F1 score) Medium PubMed Central
- PHI handling with HIPAA-compliant encryption
- Clinical data normalization with 99% accuracy

Finance:

- Anti-money laundering pattern detection with Graph Neural Networks Medium Silenteight
- Transaction data cleaning with 62% reduction in false positives Medium
- Regulatory reporting format compliance (SOX, Basel III)
- Real-time fraud detection using ensemble methods

Insurance:

- Claims data processing with 50-70% time reduction Medium
- Actuarial data cleaning with NLP techniques Simplesolve
- Risk assessment data validation
- Computer vision for damage assessment SmartDev

Retail:

- Product catalog deduplication with 40% effort reduction Express Analytics
- Customer data standardization
- Inventory data normalization
- Multi-channel data harmonization

2. Advanced Backend Architecture

2.1 System Architecture

Ingestion Layer:

- Apache Kafka for event streaming (15x faster than RabbitMQ) Confluent +2
- TUS protocol for resumable file uploads Tus Cloudflare (50MB chunks)
- Schema Registry for data contract management
- Support for 90+ data source connectors

Processing Layer:

- Apache Spark for batch processing (primary engine) chaosgenius
- Apache Flink for real-time validation chaosgenius +2
- Ray for ML-intensive workloads Domino Data Lab Onehouse
- dbt for data transformations Fivetran

Storage Layer:

- ClickHouse for analytical queries (millisecond response times) ChaosGenius +2
- Apache Pinot for real-time monitoring Startree RisingWave
- S3/Azure Blob/GCS for raw data archival
- PostgreSQL with Citus for transactional data

Orchestration Layer:

- Kubernetes with KEDA autoscaling Devtron

- Airflow for workflow management
- Circuit breakers for fault tolerance
- Bulkhead pattern for workload isolation

2.2 Database Technologies

Primary Storage (ClickHouse):

- Columnar storage with advanced compression Medium
- 2-5 second queries on billion-row datasets ChaosGenius +2
- Distributed architecture for horizontal scaling ChaosGenius
- Native SQL support with advanced analytics Medium

Real-time Analytics (Apache Pinot):

- Sub-second query responses Startree +2
- Support for 1M unique topics Startree startree
- Optimized for user-facing dashboards RisingWave
- Low-latency data ingestion Startree RisingWave

2.3 Monitoring and Observability

Technology Stack:

- OpenTelemetry for standardized telemetry collection Grafana Uptrace
- Prometheus for metrics storage Grafana Uptrace
- Grafana for visualization Grafana Grafana
- Jaeger for distributed tracing Uptrace SigNoz

Key Metrics:

- Processing latency: <200ms (p95)
- System availability: >99.9%
- Error rate: <0.1%
- Resource utilization: >80% efficiency

3. UI/UX Specifications

3.1 Frontend Technology Stack

Core Framework:

- Next.js 15 with React 19 RC support (Next.js)
- TypeScript for type safety
- Zustand for state management (Swansoftwaresolutions) (Medium)
- Tailwind CSS for styling

Visualization Libraries:

- Apache ECharts for large dataset visualization (Apache +2)
- AG-Grid for data table interactions (AG Grid +3)
- D3.js for custom visualizations (D3)
- Plotly for interactive dashboards (NPM Compare)

3.2 Key Interface Components

Data Upload Interface:

- Drag-and-drop file upload with TUS protocol
- Multi-gigabyte file support with chunked uploads
- Real-time progress tracking
- Resume capability for interrupted uploads

Data Preview and Profiling:

- Virtual scrolling for million-row datasets (AG Grid +4)
- Intelligent sampling with statistical summaries
- Interactive column profiling with histograms
- Real-time data quality indicators

Visual Workflow Builder:

- Node-based drag-and-drop interface (Alteryx)
- 300+ pre-built transformation tools (Alteryx)
- Real-time execution preview
- Version control for workflows

Processing Monitor Dashboard:

- Live execution progress with stage breakdown

- Resource utilization metrics
- Error tracking with contextual debugging
- Performance optimization recommendations

3.3 Mobile and Accessibility

Progressive Web App Features:

- Responsive design across all devices Mozilla
- Offline capability with service workers Mozilla
- Touch-optimized controls (48px minimum)
- WCAG 2.1 AA compliance Innowise DemoUp Cliplister

Accessibility Requirements:

- Full keyboard navigation support
- Screen reader compatibility
- 4.5:1 color contrast ratio Innowise
- ARIA labels for complex interactions

4. Advanced Data Processing Features

4.1 AI/ML Capabilities

Data Cleaning Algorithms:

- Transformer-based entity resolution using BERT H2O.ai Numerous.ai
- GPT-4 powered data correction with CoT prompting Numerous.ai
- Isolation forests for anomaly detection PubMed Central
- LSTM autoencoders for time-series cleaning Medium

Smart Features:

- Automatic schema inference with deep learning
- Semantic type detection using transformers
- Intelligent column mapping with 95% accuracy
- Predictive data quality scoring

Deduplication and Matching:

- Fuzzy matching with Levenshtein and Jaro-Winkler SoftwareReviews +4
- MinHash and LSH for scalable similarity search
- Deep learning entity resolution with 95% accuracy
- Probabilistic record linkage

4.2 Data Quality Assessment

Automated Profiling:

- Statistical analysis for data distribution
- Automated constraint generation (Great Expectations) Atlan Telmai
- Data quality scoring algorithms
- Anomaly detection in data profiles PubMed Central

Validation Rules:

- Pre-built industry-specific rules
- Custom rule builder with visual interface
- ML-powered rule suggestions
- Real-time validation feedback

5. Enterprise Features and Security

5.1 Security Architecture

Zero Trust Implementation:

- Continuous verification for all access Microsoft
- Identity-based perimeter with MFA Microsoft
- Least-privilege access controls Microsoft
- Risk-based authentication Government Technology CrowdStrike

Encryption Standards:

- AES-256 for data at rest
- TLS 1.3 for data in transit
- End-to-end encryption for sensitive data
- Key rotation with AWS KMS/Azure Key Vault

5.2 Compliance Frameworks

Regulatory Compliance:

- GDPR with right-to-be-forgotten implementation (Fortra +2)
- HIPAA for healthcare data (PHI handling) (DataSunrise +2)
- SOX for financial reporting integrity (DataSunrise)
- ISO 27001 and SOC 2 Type II certification

Audit and Governance:

- Comprehensive audit logging (6-year retention)
- Data lineage tracking with Apache Atlas
- Change management with approval workflows
- Automated compliance reporting

5.3 Multi-Tenant Architecture

Isolation Strategies:

- Schema-per-tenant for data isolation
- Resource quotas per tenant tier
- Network segmentation with VPCs
- Performance isolation with QoS tiers (Microsoft Learn +2)

Access Control:

- Role-based access control (RBAC)
- Attribute-based access control (ABAC)
- Fine-grained permissions at resource level
- Just-in-time access for privileged operations (Enterpriseready +3)

5.4 API Integration

RESTful API:

- OpenAPI 3.0 specification
- JWT-based authentication (owasp)
- Rate limiting with token bucket algorithm (owasp)
- Versioning with 12-month deprecation notice (OWASP Cheat Sheet Series +2)

GraphQL Support:

- Query depth limiting for security owasp
- Field-level authorization owasp
- Complexity scoring for rate limiting
- Subscription support for real-time updates

6. Implementation Roadmap

Phase 1: Foundation (Months 1-2)

Core Infrastructure:

- Kubernetes cluster deployment
- Apache Kafka implementation
- ClickHouse setup for storage
- Basic authentication system

Deliverables:

- Basic file upload capability
- Simple data preview interface
- User authentication and registration
- Initial API framework

Phase 2: Processing Engine (Months 2-3)

Data Processing:

- Apache Spark integration
- Apache Flink for real-time processing ChaosGenius Medium
- Basic cleaning algorithms
- Workflow orchestration with Airflow

Deliverables:

- Batch processing capability
- Basic deduplication features
- Data profiling dashboard
- Processing monitoring interface

Phase 3: AI Integration (Months 3-4)

AI/ML Features:

- GPT-4 integration for data correction Numerous.ai
- BERT models for entity resolution Numerous.ai
- Anomaly detection algorithms
- Industry-specific model deployment

Deliverables:

- AI-powered cleaning tier
- Smart data type detection
- Automated quality scoring
- ML-based recommendations

Phase 4: Enterprise Features (Months 4-5)

Security and Compliance:

- Zero Trust architecture implementation
- GDPR/HIPAA compliance features Frontiers PubMed Central
- Multi-tenant architecture
- Advanced access controls

Deliverables:

- Enterprise authentication (SAML/OIDC)
- Audit logging system
- Compliance reporting
- Role-based access control

Phase 5: Advanced Features (Months 5-6)

Platform Enhancement:

- Apache Pinot for real-time analytics
- Advanced visualization components
- API marketplace integration
- Performance optimization

Deliverables:

- Real-time monitoring dashboard
- Advanced workflow builder
- API documentation and SDK
- Performance benchmarks

Phase 6: Market Launch (Month 6)

Go-to-Market:

- Production deployment
- Customer onboarding system
- Support infrastructure
- Marketing website launch

Deliverables:

- Production platform
- Documentation and tutorials
- Customer support portal
- Billing system integration

7. Competitive Positioning

Market Opportunity

The data preparation market is projected to reach $27.28 billion by 2033 with a 16.4% CAGR, presenting significant growth opportunities for innovative solutions. IMARC Dataintelo

Key Differentiators

Versus Alteryx ($5,195/user/year):

- 70% lower cost with usage-based pricing Mammoth Analytics
- Simplified interface with shorter learning curve
- AI-powered features included in base tiers
- No annual contract requirements Mammoth Analytics

Versus Informatica ($100,000+ annually):

- Transparent, predictable pricing (Atlan)
- Self-service model without sales process
- Faster implementation (days vs months)
- Modern cloud-native architecture

Versus Open Source (OpenRefine):

- Enterprise-grade scalability
- Professional support and SLAs
- Advanced AI/ML capabilities
- Compliance and security features (TrustRadius) (SoftwareWorld)

Target Market Segments

Primary Targets:

- Mid-market companies ($10M-1B revenue) (GetApp)
- Data-driven startups and scale-ups
- Healthcare and financial services organizations
- Companies migrating from legacy solutions

Use Case Focus:

- High-volume data migration projects
- Regulatory compliance reporting
- Customer data consolidation
- Real-time data quality monitoring

8. Success Metrics and KPIs

Technical Performance

- Processing speed: 1 billion rows in <5 minutes
- System availability: 99.9% uptime
- API response time: <200ms (p95)
- Data accuracy: >99% after cleaning

Business Metrics

- Customer acquisition: 100 enterprises in Year 1

- Revenue target: $5M ARR by end of Year 1
- Token utilization: 70% of purchased tokens used
- Customer retention: >90% annual retention

User Experience

- Onboarding time: <30 minutes to first value
- Support ticket resolution: <4 hours
- User satisfaction: >4.5/5 rating
- Feature adoption: >60% using AI features

9. Risk Mitigation

Technical Risks

- **Scalability challenges**: Mitigated through proven distributed architectures
- **AI model accuracy**: Continuous training and human-in-the-loop validation
- **Security breaches**: Zero Trust architecture and regular security audits
- **Performance degradation**: Auto-scaling and performance monitoring

Business Risks

- **Competitive response**: Focus on continuous innovation and customer success
- **Market adoption**: Freemium tier and strong partner ecosystem
- **Regulatory changes**: Flexible compliance framework and legal monitoring
- **Talent acquisition**: Competitive compensation and remote-first culture

10. Budget and Resource Requirements

Development Team (20-25 people)

- 8 Backend Engineers
- 4 Frontend Engineers
- 3 ML Engineers
- 2 DevOps Engineers
- 2 QA Engineers
- 1 Product Manager
- 2 UX/UI Designers

- 3 Customer Success

Infrastructure Costs (Monthly)

- Cloud infrastructure: $50,000-100,000
- Third-party services: $10,000-20,000
- Security and compliance tools: $5,000-10,000
- Development tools: $5,000

Total Investment Required

- Year 1: $5-7 million
- Break-even: Month 18-24
- ROI: 300% by Year 3

Conclusion

This advanced data cleaning platform addresses critical market needs through innovative AI-powered features, transparent pricing, and enterprise-grade capabilities. By focusing on ease of use, scalability, and industry-specific solutions, Numerous.ai the platform is positioned to capture significant market share in the rapidly growing data preparation market. The comprehensive technical architecture, combined with a clear go-to-market strategy and competitive differentiation, provides a solid foundation for building a successful, scalable business that delivers exceptional value to customers while maintaining operational excellence and regulatory compliance.