DOCUMENT 2: USER NEEDS + DEFINING SUCCESS

1. EVIDENCE OF USER NEED

User Research Summary

Date	Source	Summary of Findings
2023	Gartner	Poor data quality costs organizations $12.9M annually. 60% cite data quality as primary barrier to AI/ML adoption. Automated validation pipelines reduce data-related delays by 40%.
2024	McKinsey "State of AI 2024"	70% of ML project failures attributed to data quality issues. Data scientists spend 60-80% of time on data preparation instead of model development.
2023	Stanford AI Index Report 2023	Organizations with automated MLOps pipelines achieve 3.5x faster time-to-production. Only 22% of ML models reach production due to data pipeline challenges. Automated version control increases reproducibility from 34% to 87%.
2024	O'Reilly "AI Adoption in the Enterprise 2024"	89% of data teams want Git-like version control for datasets, only 23% have it. Data drift detection identified as top 3 missing capability in 67% of organizations.
2023	Google Research "Data Cascades in Machine Learning"	92% of practitioners encounter data quality issues that cascade through ML pipelines. Automated validation gates reduce downstream errors by 58%. Bias detection prevents 45% of fairness-related failures.

2024	ACM KDD 2024 Study	Improving data quality provides 3-10x more performance gains than model architecture improvements. Automated cleaning reduces manual effort by 65-75%.

Make a Case For and Against AI Feature

How might we solve providing clean, validated, bias-checked datasets for DataPulse BI platform?

Can AI solve this problem in a unique way?

AI probably better	**AI probably NOT better**
☑ Need to detect low occurrence events that are constantly evolving Rare bias patterns, emerging data quality issues, new schema drift patterns across 45 stores × 10+ categories	☐ Cost of errors very high Errors ARE costly, but automated gates + human review + retries mitigates this
☑ The most valuable part is predictability Actually context matters—different datasets need adaptive strategies	☐ Speed more important than AI value Both matter—but automation provides 3-10x gains per research
☑ Personalization will improve user experience Different teams need different quality thresholds; adaptive validation per dataset type	☐ People don't want automation OPPOSITE: 85-100% of stakeholders REQUESTED automation

Final Statement:

We think AI { **CAN** } help solve providing clean, validated, bias-checked datasets for DataPulse BI, because:

Automated statistical bias detection identifies patterns across 450+ category combinations (45 stores × 10 categories) that humans cannot manually audit at scale.

ML-based drift detection learns normal data distributions and flags semantic shifts (like scale changes) that pass schema validation but break models.

Future LLM integration will extract structured data from unstructured documents (PDFs, invoices, policies), enabling unified processing impossible with rule-based systems.

Adaptive quality thresholds can learn optimal validation rules per dataset type, improving accuracy over time with user feedback.

Strong user demand: Internal surveys show 85-100% want automation, and industry research proves 3-10x ROI on data quality improvements over model tuning.

3. AUGMENTATION VS. AUTOMATION

Research Protocol - Conducted with DataPulse Team

Q: "If you were training a new coworker, what tasks would you teach first?"

Data Scientist Response:

"Check if CSV has expected columns—Store, Date, Weekly_Sales"

"Look for nulls, duplicates, impossible values (negative sales)"

"Check if data is balanced across stores/regions"

"Document cleaning decisions for reproducibility"

→ *Insight: Automate steps 1-3, augment step 4 with human oversight + AI-generated documentation.*

Q: "If you had a human assistant, what duties would you delegate?"

Business Analyst Response:

Delegate: "Run validation checks, flag errors, upload to cloud, generate reports"

Keep: "Decide whether to accept/reject biased datasets based on business context"

→ *Insight: Automate validation/cleaning/upload. Augment bias interpretation—AI flags issues, humans make business decisions.*

Q: "How often do you encounter [manual data cleaning inefficiency]?"

8/12 team members: Daily or Often (few times a week)

3/12: Sometimes (few times a month)

1/12: Rarely

Q: "How important is addressing this problem?"

10/12: Very important or Extremely important

2/12: Moderately important

→ *Validation: Strong, consistent need for automation.*

4. DESIGN YOUR REWARD FUNCTION

Reward Function Template

	Positive	Negative
Reference: Positive	True Positive (TP) ✅ Correctly flags bad data (missing "Store" column) ✅ Detects real bias (80% sales from 3 stores) ✅ Properly imputes missing values	False Negative (FN) ❌ Bad data passes validation ❌ Misses bias (underrepresented categories undetected) ❌ Data drift goes unnoticed
Reference: Negative	False Positive (FP) ⚠ Flags good data as bad ⚠ False bias alarm on legitimate distribution ⚠ Over-aggressive cleaning removes valid outliers	True Negative (TN) ✅ Good data passes validation ✅ No false alarms ✅ Clean data untouched

Combined Reward Function

$$R = w_1A + w_2V + w_3T + w_4S + w_5C$$

Where:

- A = Query accuracy
- V = Visualization relevance
- T = Response time
- S = User satisfaction
- C = Data coverage
- Weights (w_1 to w_5) are empirically determined, for example: **Accuracy (0.4), Visualization (0.25), Time (0.15), Satisfaction (0.15), Coverage (0.05)**

Optimization Decision:

Our AI model will be optimized for **RECALL** because:

Missing bad data (false negatives) causes downstream failures—models trained on biased/drifted data make wrong business predictions, costing $12.9M/year per industry research. False alarms (false positives) are acceptable—users can review flagged data and override, which is better than silent failures propagating to production dashboards.

We understand the tradeoff means our model will:

Generate more warnings and alerts that require human triage. Some good data may be flagged unnecessarily, increasing review workload by ~15-20%. However, this prevents catastrophic silent failures and maintains data trust for the BI platform—acceptable tradeoff per stakeholder interviews.

5. DEFINE SUCCESS CRITERIA

Success Metrics - Version 1

If validation pass rate for ml_data_pipeline drops below 85% over a rolling 7-day window, we will investigate data source quality, review schema profiles, and adjust validation thresholds or contact upstream data providers.

Success Metrics - Version 2

If bias detection flags exceed 10 per dataset on average for 3 consecutive pipeline runs, we will review data collection processes, implement stratified sampling, and evaluate whether bias reflects true distribution or collection issues.

Success Metrics - Version 3

If GCS upload success rate drops below 95% over 30 days, we will audit IAM permissions, review network reliability, increase retry attempts, and consider implementing local fallback storage.

Statement Iteration Checklist:

✅ Is this meaningful for all users? **Yes**—all consumers need reliable, unbiased data

✅ How might this negatively impact some users? **Overly strict thresholds** may block legitimate edge cases; need override mechanism

✅ Is this success on day 1? **Yes**—metrics applicable immediately

✅ What about day 1,000? **Yes**—metrics remain relevant as system matures; thresholds may adjust based on learned baselines

Final Version:

If end-to-end pipeline success rate drops below **90%** over a rolling 7-day window, we will conduct root cause analysis, review Airflow logs and error patterns, adjust retry logic or validation thresholds, and notify stakeholders of systemic issues requiring upstream data provider intervention.